UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

TÉCNICAS MONTE CARLO APLICADAS A LA ENSEÑANZA DE LA ESTADÍSTICA

REPORTE DE ACTIVIDAD DOCENTE

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIO

P R E S E N T A :

GUILLERMO CUAUHTEMOCTZIN GRANADOS GARCIA

DIRECTOR:
ACTUARIO FRANCISCO SÁNCHEZ VILLARREAL
2017

Hoja de Datos del Jurado

1. Datos del alumno
Granados
García
Guillermo Cuauhtemoctzin
46 22 62 91
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
306160136

2. Datos del tutor
Act.
Sánchez
Villarreal
Francisco

3. Datos del sinodal 1
M. en C.
José Salvador
Zamora
Muñoz

4. Datos del sinodal 2
Mat.
Margarita Elvira
Chávez
Cano

5. Datos del sinodal 3
Act.
Yurguen Hugo
Camargo
Serafín

6. Datos del sinodal 4
Act.
Harim
García
Lamont

7. Datos del trabajo escrito
Técnicas Monte Carlo aplicadas a la enseñanza de la estadística
369 p.
2017

ÍNDICE

Introducción *6*

Capítulo I: *12*

Generación de valores pseudoaleatorios por medio de técnicas Monte Carlo *12*

Métodos computacionales para generar números pseudoaleatorios **12**

Pruebas estadísticas para los números pseudoaleatorios **13**

Métodos generales para simular distribuciones de probabilidad **15**

Generación de números pseudoaleatorios distribuidos acorde a variables aleatorias comunes dentro de la estadística. **22**

Variables aleatorias discretas 22

Variables aleatorias continuas 53

Capítulo II: *110*

Enseñanza de métodos descriptivos a través del enfoque de simulación Monte Carlo. *110*

Concepto de población **111**

Análisis tabular **111**

Análisis gráfico **113**

Generación de muestras por simulación Monte Carlo **117**

Histogramas **119**

Curvas de frecuencias **130**

Medidas de tendencia central **131**

Media aritmética 131

Moda 135

Simulación de distribuciones multimodales 136

Mediana 140

Medidas de dispersión **145**

Amplitud total o Rango 145

Varianza muestral y poblacional 147

Rango intercuartil 156

Estadísticas Básicas **160**

Coeficiente de variación 160

Aplicación del coeficiente de variación en datos reales 161

Comparativo de las propuestas para el cálculo del coeficiente de variación empleando simulación Monte Carlo 170

Coeficiente de asimetría 171

Coeficiente de Kurtosis 176

Capítulo III: *183*

Intervalos de confianza *183*

Intervalo de confianza para la media de una población Normal con varianza conocida **184**

Intervalo de confianza para la media de una población Normal con varianza desconocida **188**

Simulación de intervalos de confianza para poblaciones Normales con varianza desconocida **191**

Intervalo de confianza para la varianza de una población Normal con media desconocida **193**

Análisis de los intervalos de confianza para la varianza de una población Normal por medio de simulación Monte Carlo **195**

Intervalo de confianza para la diferencia de medias de poblaciones normales con tamaños de muestra desigual **197**

Intervalo de confianza para la diferencia de proporciones **201**

Capítulo IV: *206*

Pruebas de hipótesis *206*

Hipótesis estadísticas **207**

Proceso para realizar pruebas de hipótesis **208**

Errores dentro de la conclusión sobre una hipótesis **209**

Prueba sobre la media de una población bajo el supuesto de normalidad **212**

Prueba sobre la varianza de una población bajo el supuesto de normalidad. **219**

Prueba sobre la diferencia de medias de dos muestras independientes con varianzas conocidas **230**

Capítulo V: *235*

Enseñanza del método de análisis de regresión lineal simple mediante simulación Monte Carlo *235*

Modelo de regresión lineal simple **236**

Generación del modelo lineal por simulación Monte Carlo **237**

Desarrollo de los estimadores para α y β por mínimos cuadrados. **240**

Ejemplo de cálculo y revisión de propiedades de los estimadores por mínimos cuadrados con muestras simuladas **242**

Pruebas de hipótesis e intervalos de confianza para los estimadores α ,β y σ^2 ..... **250**

Coeficiente de correlación y coeficiente de determinación **265**

Pruebas de hipótesis sobre el coeficiente de correlación ρ 271

Análisis de varianza en regresión 272

Estimación del valor medio y puntual de Y a partir del modelo ajustado 276

Intervalos de confianza para la predicción y el valor medio 277

Análisis de residuos 281

Simulación Monte Carlo de datos con heteroscedasticidad 282

Método gráfico para detectar heteroscedasticidad. 285

Prueba de Breusch-Pagan para la detección de Heteroscedasticidad 286

Prueba de Durbin-Watson para detectar residuos correlacionados 290

Simulación Monte Carlo de datos con errores correlacionados 292

Métodos gráficos para detectar correlación entre los errores 296

Prueba de Jarque-Bera para evaluar la normalidad de los errores 298

Gráfico de probabilidad Normal de residuos 303

Conclusiones y comentarios finales 306

Apéndice 307

Parte I 307

Pruebas de bondad de ajuste Ji-cuadrada, y Kolmogorov-Smirnov 307

Parte II 312

Códigos en lenguaje macros para realizar las simulaciones Monte Carlo 312

Codigos de programación en R® 349

Bibliografía 367

Recursos WEB 369

Introducción

Una de las herramientas de análisis que toma cada vez más importancia en el mundo es la estadística, por su versatilidad de técnicas para la resolución de problemas o en el análisis de un fenómeno simple hasta uno realmente complejo, por lo cual, ha sido incluida como parte del programa de estudio en diferentes carreras impartidas en la UNAM.

La transmisión del conocimiento acerca de la aplicación y entendimiento de las diversas técnicas estadísticas, usualmente involucra ciertos ejemplos prácticos que muestren el funcionamiento de tales técnicas, en los cuales, partiendo de ciertas bases matemáticas, se aprende la metodología y se visualiza la parte abstracta de la técnica. Posteriormente se muestra cuáles son sus limitaciones, o sus puntos débiles, para exponer cuándo pueden surgir los escenarios inusuales y recomendar posibles soluciones alternativas. Lo anterior permite una mejor interpretación de los resultados, que para el futuro analista, será una de sus principales fortalezas ante alguien que sólo tenga la capacidad de manejar algún potente software.

Para desarrollar estos ejemplos prácticos ahora, la cantidad de datos a la que alguien puede tener acceso es inmensa, sin embargo, esta información se encuentra con mayor facilidad en datos del sector público, por otra parte, los ejemplos prácticos para ciertas carreras requieren también explorar temas particulares relacionados con las empresas privadas, desafortunadamente estas empresas conservan su información bajo protección por diversas razones. En la práctica, cuando se tiene acceso a una base de datos, su estructura o su tamaño son tan complejos que toma tiempo aprender a interactuar con la base correctamente, para posteriormente poder analizarla. Para un curso introductorio de estadística a nivel licenciatura, el tiempo es una variable de gran importancia, como para añadir temas que son considerados como más avanzados.

Este trabajo presenta un enfoque para la enseñanza de tópicos dentro del currículo básico de un profesionista especializado en la aplicación de técnicas estadísticas, para llevar esto a cabo, se emplearon y desarrollaron diferentes herramientas en base a métodos de simulación Monte Carlo. Estos objetivos y enfoques específicos, que se desarrollaron a lo largo de los siguientes 5 capítulos se resumen a continuación:

En el primer capítulo, en el cual se trata el tema de simulación de distribuciones de probabilidad, se exploraron las distribuciones más comunes dentro de un curso de probabilidad, tratando tanto sus características principales como la manera de simular muestras por medio de técnicas Monte Carlo.

En cuanto al tema de estadística descriptiva, se trató la explicación de técnicas descriptivas básicas donde se desarrollaron ejemplos mediante simulación, junto con otros ejemplos basados en datos públicos, además del uso del software R®, con lo cual se relacionaron de forma visual y tabular, aspectos teóricos de los estadísticos revisados, con su comportamiento sobre muestras simuladas.

El tema sobre los Intervalos de Confianza, tuvo un enfoque orientado a ejemplificar el desarrollo de los intervalos de confianza desde una perspectiva más intuitiva y gráfica.

En el cuarto capítulo, se trató algunas de las técnicas de contraste de hipótesis estadísticas, desde sus fundamentos, hasta la práctica con ejemplos desarrollados por medio de simulación, donde se resalta su relación con los intervalos de confianza y se comprueba de igual manera sus comportamientos teóricos esperados.

El último capítulo mostró la técnica de análisis de regresión simple aplicada a diversas muestras simuladas para emplear métodos de estimación de parámetros. Posteriormente se desarrollaron ejemplos sobre el análisis y la evaluación del modelo lineal.

En los ejemplos que se generaron se menciona la automatización de la simulación por procesos Macro en lenguaje VBA, los cuales pueden ser consultados en la segunda parte del apéndice y se mantuvieron lo más simples posible. La intención de publicar el código es que se pueda copiar desde la versión digital y ejecutarlo en otra hoja de cálculo para generar ejemplos sobre otros subtemas o técnicas en particular.

Marco histórico de la simulación Monte Carlo

Como antecedente de cálculos relacionados con fenómenos aleatorios, en el siglo XII, se cuenta con el poema de Richard de Fournival *De Vetula* donde calcula los posibles resultados que se pueden obtener al lanzar tres dados. Más adelante en el Renacimiento, época en la que se intentó racionalizar la belleza del mundo, se planteó el problema de la repartición de ganancias, en un juego el cual se veía interrumpido, hubo intentos de resolver el problema por parte de Cardano, Pacioli y Tartaglia, los cuales ahora se sabe que están en un error. Fue hasta el siglo XVII con las cartas entre Pierre de Fermat y Blaise Pascal, que se retoma el problema y se le da solución. Lo anterior señala que, la historia de la teoría de la probabilidad parte de problemas cotidianos que generan interés en estudiarlos, resolverlos y difundir las soluciones, lo cual genera conocimiento formal.

La clase de problemas, basados en juegos y apuestas, en los que se tenía interés en ese entonces, también se les puede considerar como un experimento en el cual las reglas pactadas someten al juego a un ambiente controlado, en donde se espera como resultado, un ganador o un perdedor, y es repetido varias veces. La frecuencia con que se realizaban los juegos producía una idea intuitiva en los jugadores acerca de la predicción de los resultados, pero esta idea era de carácter cualitativa en lugar de cuantitativa.

Existen antecedentes de la segunda mitad del siglo XIX sobre gente que realizaba experimentos sobre el problema presentado por Georges-Louis Lecrerc conde de Buffón, que consistía en arrojar agujas sobre un piso marcado con líneas paralelas, para identificar cuáles tocaban las líneas. Teóricamente si la distancia entre las líneas paralelas es igual a la longitud de la aguja, entonces la probabilidad de contacto entre una línea y la aguja es $2/\pi$. Entonces como motivación, por medio de la reproducción del experimento se calculaba una estimación del valor de π.

A principios del siglo XX, en temas relacionados a la enseñanza de la estadística, las academias británicas creían que la verdadera comprensión de la estadística, podía lograrse si los estudiantes observaban de forma tangible el fenómeno y su

comportamiento aleatorio. Con base en lo anterior, en un laboratorio se realizaban experimentos en aras de comparar las predicciones teóricas con las observadas.

En Inglaterra, W. S. Gosset, mejor conocido en el ámbito estadístico por el seudónimo *Student,* nombre con el cual también se reconoce a la distribución de sus publicaciones. *Student* realizó estudios a través de la repetición de experimentos para determinar la distribución del coeficiente de correlación de Pearson r, partiendo de la función $(1-\alpha r^2)\beta$, donde α y β son constantes, concluyendo por medio de estimaciones que para una muestra de tamaño n, $\alpha = 1,\ \beta = \frac{1}{2}(n-4)$.

Fue con la publicación de Andrey Nikolaevich Kolmogorov en el año 1933 sobre la teoría de la probabilidad, que se pudo tener el rigor suficiente para considerar a la probabilidad como una rama de las matemáticas, a pesar de ser utilizada anteriormente por físicos, y de contar con resultados por parte de Markov, Liapunov y Chebyshev. Dicha teoría estaba basada en una serie de axiomas, que fueron a su vez influenciados por la teoría de la medida desarrollada en el siglo XIX. La forma abstracta que plantea Kolmogorov permitió entonces la formalización, y el antecedente clásico de la probabilidad pasó a ser una más de sus aplicaciones.

La necesidad de formalizar la teoría de la probabilidad se debió a que, las soluciones publicadas a diversos problemas, no funcionaban en otros aplicando el mismo razonamiento. Posterior al trabajo de Kolmogorov era posible abordar diversos problemas ya sea de modelación o de análisis estadístico con una teoría bien fundamentada, incluso cuando la naturaleza de los fenómenos observados es compleja, y los objetivos de la modelación ambiciosa.

Después de la segunda guerra mundial, el estudio de fenómenos más complejos necesitó de la experimentación por medio de simulaciones, al tratar con problemas de observación no factible, por ejemplo, en las estrategias militares o el manejo de la energía nuclear. Fue así que en diciembre de 1945, en los Estados Unidos, nació el proyecto Research and Development (RAND), con un contrato especial con la compañía de aeronáutica Douglas. El proyecto RAND también estuvo destinado a realizar estudios de índole social para determinar políticas públicas, además una de sus mayores aportaciones, siendo ya una institución independiente desde 1948, fue la publicación en el año de 1955 de la primera tabla de un millón de números aleatorios.

Los números aleatorios de la publicación se generaron por medio de una ruleta electrónica, en la cual se dejaba pasar un pulso de frecuencia aleatoria, que por medio de un sistema de circuitos se mandaba a cinco contadores binarios, para tener un total de 32 diferentes posiciones. El sistema desechaba 12 de las posibles posiciones, y el resultado era convertido al sistema decimal. Una serie de pruebas estadísticas fueron aplicadas a los números que se registraron, donde se identificaron problemas, a consecuencia del mantenimiento constante que la maquina necesitaba para obtener un resultado aceptable. Al obtener el millón de números aleatorios se les aplicó posteriormente la operación módulo 10, para registrar dígitos aleatorios. La tabla también contenía una serie de instrucciones sobre cómo conseguir un comportamiento aleatorio al realizar un experimento, ya que se sugería abrir el documento en una página de manera aleatoria y leer la tabla de una manera especial, además se

recomendaba abrir las páginas de manera inversa y realizar el procedimiento por más de una persona.

La publicación también contenía 100 000 números aleatorios de la distribución Normal con media cero y varianza unitaria, para generarlos se tomaron 5 dígitos formando un numero D para sustituir en la ecuación.

$$\frac{\boldsymbol{D + 0.5}}{\boldsymbol{10^5}} = \boldsymbol{F(x)}$$

Que se resuelve para x, donde F es la función de distribución de una Normal(0,1). El valor D se supone elegido entre 0 y 99999 aleatoriamente, y el factor 0.5 puede ser visto como un factor de corrección de continuidad. La solución de la ecuación se obtuvo por medio de una tabla de probabilidades de la distribución Normal.

El término “análisis de Monte Carlo” fue mencionado por primera vez por Von Neumann y Ulam en 1944, en referencia al famoso casino Monte Carlo de Mónaco. Técnica que se resume como un conjunto de técnicas que emplean números pseudoaleatorios para resolver algún problema. Su popularidad se debió a que no solo resolvía problemas de índole probabilística, como el modelaje de la energía atómica a través de las propiedades de incertidumbre de las partículas fundamentales, sino también problemas del tipo determinista en los cuales las soluciones son complicadas de encontrar o en algunos casos es imposible, como ejemplo está la estimación de Von Neumman en la solución de un sistema de ecuaciones lineales por medio de un enfoque de cadenas de Markov, y en la solución de ecuaciones diferenciales de orden superior.

Otros usos del análisis de Monte Carlo fueron notables en ciertos campos, por ejemplo, en modelos biológicos sobre la competencia de especies en un ecosistema, en tópicos sobre la Investigación de Operaciones con problemas relacionados a la teoría de colas y sistemas de inventarios, en aspectos relevantes a las teorías económicas las simulaciones permitieron pasar de modelos clásicos y estáticos a modelos dinámicos de varios participantes, se generaron predicciones sobre el nivel del rio Nilo en temporadas de lluvia para controlar posibles afectaciones, en astronomía se consiguió estimar el tiempo de vida de los cometas con base en sus trayectorias, y a las leyes de Kepler, en problemas físicos de percolación se modeló el paso de un fluido fijo sobre medios de propiedades aleatorias.

A su vez, el desarrollo de las computadoras digitales en los años 50´s permitió ver a la simulación como una forma eficiente de modelar y analizar sistemas que si se asemejan a los ejemplos anteriores, es posible que no haya una solución analítica o su observación se asociara a costos que no se estuviera dispuesto a asumir. De manera paradójica, su popularidad hizo que hubiera una revisión más minuciosa sobre estas técnicas desestimando su uso desmedido, pues se desviaba la atención de resolver los problemas de manera exacta, incluso en problemas donde la solución llegaba a obtenerse de manera menos eficiente con la simulación.

Otro de los factores que influyeron fuertemente en el avance de las técnicas de simulación fue la producción de software especializado en los modelos de simulación. A diferencia de los lenguajes de programación para propósitos generales, que otorgan

una flexibilidad completa sobre lo que se quiera realizar en la computadora, los lenguajes especializados fueron de gran relevancia pues permitían al investigador concentrarse en la formulación del modelo y programarlo con menor esfuerzo, ya que incluían generadores de números pseudoaleatorios y manejo de las relaciones entre las variables del modelo de manera simple, como la programación de ecuaciones, la posibilidad de elegir el tipo de registro de tiempo (discreto o continuo) en el que los eventos se simulaban, la generación de una serie de estadísticas resumen, además de contar con herramientas para reportar lo anterior por medio de tablas y gráficas.

Entre algunos de estos programas se pueden nombrar a: DYNAMO (Dynamic Models), desarrollado por Jay W. Forrester al final de la década de los 50 en el MIT, SIMSCRIPT, hecho por la corporación RAND en 1961, GPSS (General Purpose System Simulator), el cual fue muy popular, desde su introducción a principios de los años 60, SIMULA, generado para lenguaje Algol, pero que no alcanzó la popularidad de GPSS, pero sí mejores críticas. GASP (General Activity Simulation Program), desarrollado en 1962 por Philip J. Kiviat y Fortran (Formula Translating System) desarrollado en los años 40 para funcionar en computadoras IBM.

Los primeros autores que revisaban las técnicas de simulación desde sus conceptos y métodos como Hammersley y Handscomb (1964/1975), comienzan a establecer brevemente una convención de cómo se va a considerar el significado o definición de simulación, para posteriormente revisar ciertas técnicas y aplicaciones a campos específicos. En la literatura contemporánea como Law (2007) las técnicas Monte Carlo ahora son un apartado, ya que se revisa la simulación directamente asociada a un sistema desde una perspectiva más computacional, donde además se discuten diferentes tipos de simulación, así como aspectos importantes a considerar como la validez y credibilidad del modelo.

En particular para este trabajo la simulación por medio de la técnica Monte Carlo no se enfoca en modelar y estimar los parámetros de un sistema sino a aproximar su comportamiento, para producir ejemplos en los cuales sea clara la aplicación e interpretación de resultados en el uso de técnicas estadísticas. Si bien, el problema o ejemplo presentado puede ser posiblemente observado para posteriormente recabar información y ser estudiado, los datos obtenidos de éste pueden llegar a ser muy limitados, dependiendo de los objetivos y el tipo de estudio, en contraste, en la simulación por computadora cada variable utilizada puede ser registrada y manejada en una base de datos.

Actualmente la simulación como método de apoyo a la enseñanza se puede hallar en el ámbito militar con los llamados “juegos de guerra”, en los cuales dos equipos de un mismo ejercito pelean, en un escenario lo más verosímil posible.

En otro orden de ideas, en materias con relación al estudio económico se encuentran los “juegos gerenciales”, los cuales nacieron en 1956 por la American Managment Association. Dentro de los juegos gerenciales se pueden encontrar algunos con escenarios de diversos grados de complejidad, uno de los más desarrollados y complejos es el de la Universidad Carnegie denominado “Managment Game”, el cual tiene la duración de 8 semanas durante las cuales los alumnos de la Maestría en

Negocios son puestos a cargo de una empresa fabricante de relojes, con el objetivo de lograr desarrollar la empresa.

Históricamente, los juegos militares son más antiguos, pero ambos tienen la finalidad de exponer a los estudiantes a un ambiente controlado en el que aplicarían de manera práctica diversas técnicas que hayan aprendido.

Capítulo I:

Generación de valores pseudoaleatorios por medio de técnicas Monte Carlo

Métodos computacionales para generar números pseudoaleatorios

En la actualidad el software que se puede conseguir, para realizar diversas tareas automatizadas o resolver problemas, poseen generadores de números pseudoaleatorios, los cuales son puestos a prueba de distintas formas. Es necesario, entonces, mencionar su naturaleza, los algoritmos usados, y algunas pruebas usadas para poder evaluar su eficacia.

En general se puede establecer que una serie de números pseudoaleatorios es una sucesión de números enteros $Z_0, Z_1, \dots, Z_i \dots$ menores que un entero M $\forall i$ tales que la distribución del mapeo $U_i = \frac{Z_i}{M}$ se aproxime al comportamiento de una variable aleatoria uniforme en el intervalo (0,1).

Generadores Congruenciales

Este tipo de generadores determinan una sucesión de números obtenidos de la relación de congruencia en una regla recursiva de la forma

$$\boldsymbol{x_{n+1} = (ax_n + b) mod\ m}$$

$$\boldsymbol{U_n = x_n / m}$$

El parámetro a es llamado el multiplicador, b es el sesgo, m el modulo y el valor inicial del algoritmo x_0 es la semilla, cuando el parámetro $b = 0$ entonces el generador es llamado multiplicativo.

Las salidas de este tipo de generadores tienen las siguientes características:

- **Toda sucesión posee un ciclo o periodo.**
- **Si un cierto entero k, es el mínimo entero para el cual $x_k = x_0$ entonces $k - 1$ es la longitud del periodo y ésta, depende de la selección de los parámetros x_0, a y b.**

Algo que se puede notar es que $k \leq m$; en el caso en que $k = m$ el generador es de periodo completo, debido a esto m se puede elegir como el mayor entero posible, para una arquitectura de computadora de 32 bits el mayor número representable es $2^{32} - 1$.

Para mejorar también el comportamiento del generador se tienen además algunos resultados de la teoría de números con los que se deduce la siguiente proposición.

Un generador congruencial tiene periodo completo m si y sólo si

- $\boldsymbol{mcd\ (b, m) = 1}$
- $\boldsymbol{a = 1\ mod(p)\ para\ un\ factor\ primo\ p\ de\ m}$

- $\boldsymbol{a = 1\ mod(4)\ si\ 4\ divide\ a\ \ m}$

En el caso de los generadores multiplicativos ($b = 0$) el periodo máximo que se puede alcanzar es $m - 1$ sólo si m es primo. Un resultado útil en este caso es el siguiente:

El periodo es $\boldsymbol{m - 1}$ si y sólo si m es primo, $\boldsymbol{a}$ es una raíz primitiva de $\boldsymbol{m - 1}$, es decir, $\boldsymbol{a \neq 0}$, $\boldsymbol{a^{(m-1)/p} \neq 1\ mod\ m}$ para todos los factores primos $\boldsymbol{p}$ de $\boldsymbol{m - 1}$

Además para poder hallar las raíces positivas se pueden obtener mediante

Si $\boldsymbol{a}$ es raíz primitiva de $\boldsymbol{m}$, $\boldsymbol{a^k\ mod\ m}$ lo es, siempre que $\boldsymbol{mod(k, m - 1) = 1}$

Generadores congruenciales más generales.

La forma de generalizar a los generadores congruenciales es definiendo su fórmula recursiva de la siguiente forma:

$$\boldsymbol{Z_i = g(Z_{i-1}, Z_{i-2}, \ldots)(mod\ m)}$$

Donde g es una función función fija de los valores generados previamente que además deben cumplir la propiedad $Z_i < m\ \forall i$, como g es una función cualquiera se pueden tener funciones polinómicas, por ejemplo de la forma $g(Z_{i-1}, Z_{i-2}, \ldots) = a_1 Z_{i-1}^2 + a_2 Z_{i-1} + \cdots + c$.

Otra forma similar al primer tipo de generadores, es cuando se conserva la linealidad sobre los valores anteriores, cuando se recurre a esto se obtiene un generador recursivo múltiple se define por:

$$\boldsymbol{g(Z_{i-1}, Z_{i-2}, \ldots) = a_1 Z_{i-1} + a_2 Z_{i-2} + \cdots + a_q Z_{i-q}}$$

Donde $a_1, a_2, \ldots, a_q$ son constantes. El periodo máximo posible es $m^q - 1$ y es posible bajo una apropiada selección de los parámetros.

Un generador específico de esta clase se deriva cuando $q = 2$ y $a_1 = a_2 = 1$, llamado también generador de Fibonacci.

$$\boldsymbol{Z_i = (Z_{i-1} + Z_{i-2})\ (mod\ m)}$$

Pruebas estadísticas para los números pseudoaleatorios

Los números generados por un algoritmo de computadora que al ser determinado por sus valores iniciales y tener una precisión limitada a la capacidad de la maquina no se puede de ninguna forma afirmar que sean aleatorios, ese es el por qué se antepone el prefijo “pseudo”, aun así, el objetivo de la generación es el de reproducir la aleatoriedad, entonces surge la cuestión ¿cómo saber si la sucesión generada refleja un comportamiento aleatorio? En la medida que las sucesiones pseudoaleatorias pasan las pruebas estadísticas orientadas a probar la aleatoriedad ideal, se pueden utilizar como números aleatorios aunque no lo sean. Las siguientes menciones son algunas pruebas estadísticas usadas para comprobar las propiedades que debe un cumplir un generador.

Pruebas de rachas

Dada una sucesión de observaciones $X_1, X_2, \dots, X_n$, para realizar la prueba se define un criterio de tal manera que las observaciones se dividan en dos grupos distintos y se conservan en el mismo orden que fueron registradas. El criterio para probar la aleatoriedad de la muestra es mediante la relación binaria $1(0)\ si\ X_i < X_{i+1}$ $(si\ X_i > X_{i+1})$ donde se define como una racha $creciente$ $(decreciente)$ de tamaño l a un grupo consecutivo de l números $1(0)$. Posteriormente se cuenta el número de rachas (n) que se observen en la muestra, donde el número de rachas resultantes no debe ser tan pequeño o tan grande si se desea evidenciar la aleatoriedad de la muestra.

La forma de contrastar la hipótesis nula H_0: $la\ muestra\ es\ aleatoria$ contra la hipótesis alternativa H_1: $la\ muestra\ no\ es\ aleatoria$, es a partir de la distribución asintótica, que es posible obtener cuando se supone H_0 cierta. La distribución asintótica es

$$\boldsymbol{N\left(\frac{2n-1}{3}, \frac{16n-29}{90}\right)}$$

Otro criterio es contar las observaciones que se sitúen por encima de la mediana de aquellas que se localizan por debajo de la mediana, de esta manera el contraste se da por la distribución asintótica

$$\boldsymbol{N\left(1+\frac{n}{2}, \frac{n}{2}\right)}$$

Pruebas de bondad de ajuste

Esta clase de pruebas utiliza, como su nombre indica, cuando se desea identificar que tanto se ajusta la distribución de una serie de observaciones a una distribución de estudio $F_X(x)$. En este caso la distribución que se desea contrastar es la distribución Uniforme en el intervalo $(0,1)$. La función de distribución de la muestra se denota por $F_0(x)$, entonces, lo que se contrasta es la hipótesis nula H_0: $F_X(x) = F_0(x)$ para todo $x \in \mathbb{R}$ contra la hipótesis alternativa H_1: $F_X(x) \neq F_0(x)$ para algún x.

Entre las pruebas más usadas esta la χ^2 (Ji-cuadrada), para realizarla se toma el soporte de la variable aleatoria a contrastar y se divide en k subconjuntos mutuamente excluyentes y para cada una de las clases se cuenta el número de observaciones que caen en cada clase denotándose f_i $i = 1,2, \dots, k$. Además se calculan las frecuencias esperadas de cada clase de acuerdo a la distribución $F_X(x)$ que se denotan como $\boldsymbol{E_i}$[1]**.** La prueba se lleva a cabo por medio del estadístico

$$\boldsymbol{\chi^2 = \sum_{i=1}^{k} \frac{(f_i - E_i)^2}{E_i}}$$

El cual tiene una distribución asintótica χ^2_{k-1} y donde e_i es el valor esperado de observaciones para la i-ésima clase, que bajo la hipótesis de uniformidad es n/k.

Otra prueba de bondad de ajuste es la de Kolmogorov-Smirnov, que se restringe al caso en que F_0 es continua; se define la función de distribución empírica como

[1] Para una referencia completa sobre los cálculos de la prueba véase la parte I del apéndice.

$$\boldsymbol{F_n(x) = \frac{\#\{X_i \leq x\}}{n}}$$

La cual se espera sea similar a la función $F_X(x)$, que en este caso es la función de distribución de una variable Uniforme en (0,1). El estadístico para esta prueba es

$$\boldsymbol{D_n = \sup_{x \, \epsilon \, \mathbb{R}} |F_n(x) - F_X(x)|}$$

En el cálculo del estadístico se puede notar que la prueba mide la máxima diferencia entre la función de distribución empírica y la distribución del modelo de estudio. De forma exacta la distribución de $\boldsymbol{D_n}$ se puede hallar tabulada para valores menores que 40 en Gibbons (2003), para n mas grande se utiliza la distribución asintótica

$$\boldsymbol{\lim_{n\to\infty} P\left(\sqrt{n} D_n \leq z\right) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 z^2}}$$

Prueba del producto rezagado

A esta prueba se le considera también como una medida de independencia entre las observaciones X_i. Si se denomina a k como la longitud del rezago, entonces el coeficiente del producto rezagado para la muestra $X_1, X_2, \ldots, X_n$ es

$$\boldsymbol{C_k = \frac{1}{n-k} \sum_{i=1}^{n-k} X_i X_{i+k}}$$

Esta prueba identifica la existencia de correlación entre las mismas observaciones, en ausencia de ésta los valores de C_k se distribuyen normal con una esperanza de $1/4$ y una desviación estándar igual a $\frac{\sqrt{13n-19k}}{12(n-k)}$, por lo tanto debe de realizarse una prueba adicional de bondad de ajuste para los C_k.

Existen también otras pruebas estadísticas para probar la eficacia de los generadores de las cuales destacan

- **Pruebas de series**
- **Pruebas de corridas**
- **Pruebas de distancia**
- **Prueba de máximos**
- **Prueba de póker**

Métodos generales para simular distribuciones de probabilidad

Cuando se desea simular algún sistema, dentro de las relaciones descritas en un modelo, se puede establecer que el comportamiento de alguna de las variables asociadas a una entidad del sistema posee cierta incertidumbre. A esta clase de comportamientos es posible asociar una función de distribución, lo cual puede ser un supuesto del modelo que pudo haber estado basado en información previa, así que, es necesario el poder reproducir el comportamiento de alguna distribución y también

es importante señalar que siempre se puede conseguir un software que tenga integrado la generación de las distribuciones más conocidas.

Un supuesto importante de esta sección es el de poder generar valores pseudoaleatorios de una variable denotada como U, la cual se distribuye Uniforme en el intervalo (0,1), pues es la base de todos los algoritmos siguientes.

Método de la transformación inversa

Cuando se quieren simular valores de una variable aleatoria continua con función de distribución F_X, de la cual se pueda obtener su inversa F_X^{-1} , se puede igualar el resultado de la función de distribución a una variable uniforme en $(0,1)$ $U = F(x)$, de la cual si se obtiene el mapeo con la inversa $F^{-1}(U)$ se está generando un valor x con función de distribución F_X, lo cual se puede demostrar de la siguiente manera

$$\boldsymbol{Como\ P(U \leq u) = u\ \ por\ ser\ U \sim (0,1)}$$

$$\boldsymbol{ahora\ se\ obtiene\ la\ distribución\ de\ F^{-1}(U)}$$

$$\boldsymbol{P\left(F^{-1}(U) \leq x\right) = P(U \leq F(x)) = F(x)}$$

$$\boldsymbol{por\ lo\ tanto\ F^{-1}(U) \sim F(X)}$$

Supóngase que se puede generar un valor $U \sim U(0,1)$, entonces el algoritmo para generar una variable aleatoria X es

1. $\boldsymbol{Generar\ U \sim U(0,1)}$
2. $\boldsymbol{Hacer\ Y = F_X^{-1}(U)}$

También se debería comprobar que el método regresa una variable aleatoria con función de distribución F_X, es decir, que para la salida Y, $P(Y \leq x) = F_X(x)$ se tiene

$$\boldsymbol{P(Y \leq x) = P\left(F^{-1}(U) \leq x\right)}$$

$$\boldsymbol{= P\left(U \leq F_X(x)\right) = F_X(x)}$$ **Por la uniformidad de U.**

En el caso en que se quiera simular valores de una variable aleatoria discreta, pudiendo tomar los valores $x_1, x_2, \ldots$, tal que $x_1 < x_2 < \cdots$; el método se sigue basando en la misma idea de la función inversa F_X^{-1}, pero usando una forma generalizada de la función inversa. El algoritmo es como sigue

1. $\boldsymbol{Generar\ U \sim U(0,1)}$
2. $\boldsymbol{Hallar\ el\ entero\ positivo\ mas\ pequeño\ i\ que\ cumpla\ U \leq F_X(x_i)}$
3. $\boldsymbol{Hacer\ Y = x_i}$

Método de composición

Si la función de distribución $F_X(x)$, tiene una función de densidad f_X, que se desea simular y además, puede ser descrita como una mezcla de distribuciones, más fáciles de generar que la original, entonces un método que se puede seguir es el de la composición.

En el caso continuo la mezcla se puede ver en general como

$$\boldsymbol{f_X(x) = \int g(x|y)dH(y)}$$

Donde $g(x|y)$ es una función de densidad que depende de la variable y la cual tiene función de distribución $H(y)$.

Así que, el algoritmo es

1. $\boldsymbol{Generar\ Y \sim H}$
2. $\boldsymbol{Generar\ X \sim g(.|Y)}$

Para generar H y $g(.|Y)$ puede usarse el método de la transformación inversa u otro método que se verá más adelante. En el caso discreto la idea es la misma, a diferencia que, en este caso la distribución de Y es discreta, así que sus probabilidades son de la forma $p(Y = j) = P_j \ \ j = 1,2, \dots, n$, por consiguiente la distribución $f_X(x)$ es una combinación convexa de una serie de densidades f_j que se puede expresar de la siguiente forma

$$\boldsymbol{f_X(X) = \sum_{j=1}^{n} P_j f_j(x)}$$

Donde cada f_j es una densidad g de x, dado que $y = j$; el algoritmo se resume a: generar un entero aleatorio que siga las probabilidades P_j, y con este muestrear de la densidad indexada por dicho entero, más específicamente

1. $\boldsymbol{Generar\ Y\ tal\ que\ P(Y = j) = P_j}$
2. $\boldsymbol{Generar\ X \sim f_Y}$

Esta simulación también se puede realizar por distintos métodos de generación de variables aleatorias, además debe notarse que para el empleo del algoritmo, se deben generar por lo menos dos números aleatorios distintos.

Método de aceptación y rechazo

Con los algoritmos anteriores se puede encontrar una función g que pueda ser simulada fácilmente, en el caso que se desee simular otra función f, la cual conlleve a ciertas complicaciones en su simulación, es posible auxiliarse en la simulación de g para generar la función objetivo por medio del método de aceptación y rechazo, publicado en el año 1951 por Von Neumann.

La primera condición que debe de cumplir g, es tener el mismo soporte que la función a simular, la segunda condición es poder hallar una constante M tal que

$$\boldsymbol{f(x) \leq Mg(x) \qquad \forall x}$$

Si se define M como $Sup_x\{f(x)/g(x)\}$ se asegura que g sea una envolvente de f. La idea principal del algoritmo es generar aleatoriamente puntos por debajo de la curva $Mg(x)$, aceptando aquellos que se sitúen debajo de la curva f y rechazando los demás.

El algoritmo es el siguiente

1. $\boldsymbol{Generar\ x^* \sim g}$
2. $\boldsymbol{generar\ U \sim U(0,1)}$
3. $\boldsymbol{si\ \frac{f(x^*)}{Mg(x^*)} \geq U\ hacer\ X = x^*\ en\ caso\ contrario\ volver\ a}$ **1.**

Se debe probar que cuando el valor X es aceptado por el algoritmo entonces tal valor se distribuye $F(x)$

Demostración.

$$\boldsymbol{P(X \leq x | X\ es\ aceptado) = P(X \leq x\ | U \leq f(X)/Mg(X))}$$

$$= \frac{\boldsymbol{P\left(X \leq x\ , U \leq \frac{f(X)}{Mg(X)}\right)}}{\boldsymbol{P(U \leq f(X)/Mg(X))}}$$

Se realiza una doble integral tanto el numerador sobre la conjunta para encontrar la probabilidad deseada, mientras que en el denominador se obtiene la función marginal sobre g.

$$= \frac{\boldsymbol{\int_{-\infty}^{x} \int_{0}^{\frac{f(x)}{Mg(x)}} 1 du\, g(x) dx}}{\boldsymbol{\int_{-\infty}^{\infty} \int_{0}^{\frac{f(x)}{Mg(x)}} 1 du\, g(x) dx}}$$

$$= \frac{\boldsymbol{\int_{-\infty}^{x} \frac{f(x)}{Mg(x)} g(x) dx}}{\boldsymbol{\int_{-\infty}^{\infty} \frac{f(x)}{Mg(x)} g(x) dx}}$$

$$= \frac{\boldsymbol{\int_{-\infty}^{x} f(x) dx}}{\boldsymbol{\int_{-\infty}^{\infty} f(x) dx}} = \boldsymbol{F(x)}$$

■

Observando el denominador de la prueba anterior, se puede obtener la probabilidad de que un valor generado sea aceptado, esta probabilidad, también llamada tasa de aceptación, es un indicador de la eficiencia del método y su resultado es

$$\boldsymbol{\int_{-\infty}^{\infty} \frac{f(x)}{Mg(x)} g(x) dx = \frac{1}{M}}$$

Puesto que el valor de M depende de la elección de la densidad g se puede notar que cuanto más se aproxime g a la densidad objetivo, menor será el valor de M lo cual produce una mayor tasa de aceptación, que se traduce en una mayor eficiencia.

Método del cociente de uniformes

Este método, introducido por Kinderman y Monahan en 1977, retoma el hecho de que el cociente de dos valores, que se distribuyen uniformemente en el disco unitario, tiene

una distribución Cauchy. El algoritmo extiende el concepto a una cierta región C_h encontrando la forma de simular una cierta densidad.

Se define $C_h = \{(u,v)|0 \leq u \leq \sqrt{h(v/u)}\}$ donde h es una función no negativa, la cual la integral sobre todo su dominio, es finita.

Entonces si (u,v) se distribuyen uniforme sobre C_h entonces $X = v/u$ se distribuye $\frac{h}{\int h}$

Demostración:

Realizando el cambio de variable $x = \frac{v}{u} \quad y = u \quad xy = v \quad dv = udx$, además recordando el resultado de transformaciones sobre densidades conjuntas, se puede obtener la distribución de (u,x)

$$\boldsymbol{f_{y,x}(u,x) = f_{u,v}(y,yx)|J|}$$

Donde $|J|$ es el jacobiano de la transformación y queda determinado por

$$\boldsymbol{|J| = \begin{vmatrix} \frac{\partial u}{\partial y} & \frac{\partial u}{\partial x} \\ \frac{\partial v}{\partial y} & \frac{\partial v}{\partial x} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ x & y \end{vmatrix} = y}$$

Por lo tanto

$$\boldsymbol{f_{y,x}(y,x) = \frac{y}{A(C_h)}}$$

Donde $A(C_h)$ es el área de la región C_h. La densidad conjunta toma esta forma por la uniformidad de u y v sobre la región. Por otra parte, la condición $0 \leq u \leq \sqrt{h(v/u)}$ => $0 \leq u \leq \sqrt{h(x)}$. Para hallar el área de C_h con el cambio de variable se calcula

$$\boldsymbol{A(C_h) = \iint_{C_h} dudv = \int_{-\infty}^{\infty} \int_{0}^{\sqrt{h(x)}} y\, dy\, dx = \int_{-\infty}^{\infty} \frac{1}{2} h(x) dx}$$

Falta entonces obtener la distribución marginal de X a partir de la conjunta anterior, integrando sobre todos los valores y

$$\boldsymbol{\int_{0}^{\sqrt{h(x)}} \frac{y}{A(C_h)} dy = \frac{\frac{1}{2}h(x)}{\int_{-\infty}^{\infty} \frac{1}{2} h(x) dx} = \frac{h}{\int h}}$$

■

Para poder simular sobre el conjunto C_h, se debe encerrar la región en un rectángulo y luego usar una forma del método de aceptación y rechazo, la forma general de obtener el rectángulo es la siguiente.

Por medio del cambio de variable usado anteriormente se obtiene una cota para u, que cumple:

$0 \leq u \leq \sqrt{h(x)}$ Entonces sea $R = \sup_x\{\sqrt{h(x)}\}$

Como $v = ux$ => $0 \leq \frac{v}{x} \leq \sqrt{h(x)}$

Para $x \leq 0$ $v \geq x\sqrt{h(x)}$ => $D = Inf_{x\leq 0}\{x\sqrt{h(x)}\}$

Para $x \geq 0$ $v \leq x\sqrt{h(x)}$ => $U = \sup_{x\geq 0}\{x\sqrt{h(x)}\}$

Requiriendo que tanto $h(x)$ como $x^2h(x)$ estén acotadas en su dominio

Después de haber hallado los coeficientes anteriores, el algoritmo es el siguiente

1. $Generar\ u_1, v_1 \sim U(0,1)$
2. $hacer\ u = Ru_1\ y\ v = D + (U - D)v_1$
3. $Si\ u \leq \sqrt{h(v/u)}\ entonces\ X = \frac{v}{u}$ **caso contrario regresar a 1.**

Como medida de eficiencia del método, se puede obtener la probabilidad de aceptar una observación generada, debido a la uniformidad sobre C_h, sobre el rectángulo que acota tal región, el cálculo se remite a la probabilidad geométrica, obteniendo el cociente de las áreas, la región posible donde se obtendría un resultado favorable, entre la región total donde se lleva a cabo la simulación.

$$P(X\ sea\ aceptado) = \frac{A(C_h)}{R(U - D)}$$

Otro uso de esta probabilidad, es considerar que cada iteración del método va a realizar un cierto número de ensayos Bernoulli hasta que acepte una observación, con la probabilidad de éxito igual al cociente anterior, se sigue que tiene una distribución geométrica, así que tomando el inverso de la probabilidad, se obtiene el número promedio de veces que tomará aceptar un valor generado. Para mejorar la eficiencia del método se puede intentar ajustar una región envolvente más parecida a la región C_h , pero esto genera más operaciones a la hora de aceptar o rechazar un valor, además puede depender del tipo de distribución a simular; otros tipos de regiones que han sido propuestos son polígonos y elipses.

Simulación de distribuciones empíricas

Cuando se tiene información previa del fenómeno que se intenta modelar, los primeros intentos de abstraer su comportamiento son necesariamente basados en los registros obtenidos, tales datos son vistos como la muestra $x_i,\ i = 1,2,\dots,n$ de una variable aleatoria, con una función de distribución F desconocida, se procede a estimarse con una función F_n que se conoce comúnmente como función de distribución empírica. Hay diversas formas de estimación que difieren por ciertos detalles, aunque se basan en la misma idea.

En el caso que los valores que puede tomar la variable aleatoria sean discretos F_n sigue la ecuación

$$\boldsymbol{F_n(x)} = \begin{cases} \boldsymbol{0} & x < \boldsymbol{x_{(1)}} \\ \boldsymbol{\frac{i}{n}}, & \boldsymbol{x_{(i)} \le x < x_{(i+1)}} \\ \boldsymbol{1} & \boldsymbol{x \ge x_{(n)}} \end{cases}$$

Donde $x_{(i)}$ es la i-ésima observación de un ordenamiento ascendente de las $x_j's$, un problema a esta estimación es cuando una función de distribución empírica tenga el valor 1 para un valor discreto ($x = x_{(n)}$), por esta razón se usan funciones alternativas como por ejemplo restar 0.5 al numerador es decir

$$\boldsymbol{F_n(x)} = \frac{\boldsymbol{i - 0.5}}{\boldsymbol{n}}$$

Tal ajuste es menos notorio, cuando n es cada vez mayor, en particular esta alternativa es la usada para el tipo de gráfico *Q-Q plot;* otro tipo de ajuste sugerido es tomar

$$\boldsymbol{F_n(x)} = \frac{\boldsymbol{i}}{\boldsymbol{n+1}}$$

Para datos con posibles valores en un dominio continuo, una forma de estimar F es por medio de una interpolación lineal entre los puntos de una función del estilo de las anteriores explícitamente

$$\boldsymbol{F_n(x)} = \begin{cases} \boldsymbol{0} & x < \boldsymbol{x_{(1)}} \\ \boldsymbol{\frac{i-1}{n-1} + \frac{x - x_{(i)}}{(n-1)(x_{(i+1)} - x_{(i)})}}, & \boldsymbol{x_{(i)} \le x < x_{(i+1)}} \\ \boldsymbol{1} & \boldsymbol{x \ge x_{(n)}} \end{cases}$$

La simulación de una muestra en cualquiera de los casos, continuo o discreto, se remite al método de la transformación inversa donde el caso discreto la transformación considera un valor aleatorio de la variable Uniforme U en el intervalo (0,1) y devuelve a la $x_{(i)}$ que cumpla $U < F_n(x_{(i+1)})$, por lo que el algoritmo es

1. $\boldsymbol{Generar\ U \sim U(0,1)}$
2. $\boldsymbol{regresar\ la\ x_{(i)}\ que\ satisfaga\ \ F_n(x_{(i)}) \le\ U < F_n(x_{(i+1)})}$

Si el dominio de la variable aleatoria es continuo, entonces se debe invertir la interpolación, esto conlleva primero a encontrar, como en el algoritmo anterior, una cierta $x_{(i)}$ para establecer la ecuación de la recta entre $x_{(i)}$ y $x_{(i+1)}$ para luego obtener un valor aleatorio de salida en el dominio de la recta, es decir ya situados entre $x_{(i)}$ y $x_{(i+1)}$ se tiene

$$\boldsymbol{U = F_n(x)} = \frac{\boldsymbol{i-1}}{\boldsymbol{n-1}} + \frac{\boldsymbol{x - x_{(i)}}}{\boldsymbol{n - 1(x_{(i+1)} - x_{(i)})}}$$

$$\boldsymbol{x = (U(n-1) - i + 1)(x_{(i+1)} - x_{(i)})}$$

Un algoritmo que evita la búsqueda inicial es el siguiente

1. $\boldsymbol{Generar\ U \sim U(0,1), sea\ P = (n-1)U, y\ sea\ I = floor[P] + 1}$

2. $\boldsymbol{regresar\ X = x_{(I)} + (P - I + 1)(x_{(I+1)} - x_{(I)})}$

Donde $floor[P]$ es el máximo entero que no es mayor que P, también llamada función piso o parte entera; analizando el algoritmo, el valor P es uniforme entre 0 y $n-1$ por lo tanto I es un entero aleatorio entre 1 y $n-1$, entonces se toma un valor de las x_i de forma aleatoria (ya que no considera su ordenamiento) para luego a partir del factor $(P - I + 1)$, que es un equivalente a U, genera a partir de $x_{(I)}$ un numero aleatorio entre $x_{(I)}$ y $x_{(I+1)}$.

La desventaja de los valores por los algoritmos anteriores es que están limitados por $x_{(1)}$ en la parte inferior y por $x_{(n)}$en la parte superior, esto establece la nula posibilidad de valores extremos, lo cual para diversas aplicaciones cuando el tamaño de la muestra no es tan grande produce conflictos, una propuesta por parte de Bratley, Fox y Schrage (1987) es anexar una distribución exponencial al lado derecho de la distribución empírica, que funja como una cola derecha, para permitir generar valores mayores a $x_{(n)}$, con la condición de que, el nuevo ajuste conserve la misma media que los datos originales.

Generación de números pseudoaleatorios distribuidos acorde a variables aleatorias comunes dentro de la estadística.

Dentro de un modelo de simulación en muchos casos es necesario generar una o más variables aleatorias con ciertas propiedades específicas, ya sea porque las pruebas estadísticas sobre datos observados lo indiquen, o en otros casos quien estuviera realizando el modelo lo incluyera dentro de uno de los supuestos del comportamiento del algún componente del fenómeno de estudio.

Apuntando hacia el objetivo de este trabajo, también es necesario saber generar valores aleatorios que sigan las propiedades de ciertas distribuciones, con el objetivo de comprender conceptos fundamentales de la estadística, y hacerlos más evidentes; además de tener una gran gama de ejemplos posibles en los cuales se pueda enseñar cómo llevar a cabo una técnica estadística y estudiarla a detalle.

Dada la importancia de generar diversos tipos de comportamientos aleatorios se enlistan los algoritmos de generación de algunas de las variables aleatorias de gran importancia, en cuanto a temas de estadística.

Variables aleatorias discretas

Esta clase de variables son utilizadas cuando de antemano se sabe que el fenómeno tiene un numero finito (o numerable) de resultados posibles. Por ejemplo, al asignar las distintas opciones de una encuesta o tipos de productos, pues los datos obtenidos se pueden dividir por categorías finitas; también aparecen en variables numéricas que son medidas por unidades enteras, como el conteo de objetos que son puestos en un espacio definido, reproduciendo un sistema de inventarios.

En el contexto de la enseñanza, este tipo de variables fueron las primeras en ser estudiadas, desde inicios de la historia de la probabilidad, entonces su uso puede ser

acompañado por un marco histórico, en el que se desarrollan modelos realistas sobre problemas involucrados con juegos de azar o selección de elementos de una urna.

La simulación de estas variables puede servir como ayuda para ilustrar el resultado de los desarrollos teóricos, además realizando estadística descriptiva, se puede comprobar los valores de las características principales de estas variables. Todos los algoritmos de simulación presentados en el capítulo pueden ser consultados en la parte II del apéndice.

Uniforme discreta

Esta variable aleatoria, está definida en un dominio discreto finito de enteros consecutivos los cuales son equiprobables; una de las formas más común de hallar esta distribución es definida en el conjunto $\{1,2,3,\dots,n\}$ o en el conjunto$\{j, j+1,\dots,n\}$ $j,n \in \mathbb{Z}$ $j<n$.

Se le puede relacionar con una serie de objetos que se encuentran etiquetados de una manera consecutiva, por ejemplo en la segunda guerra mundial las fábricas de armamento alemán colocaban un número de serie en el chasis de los tanques de manera consecutiva, un problema que se utilizó para estimar el total de la producción pues la cifra exacta era información clasificada teniendo entonces que ser estimado; surgió entonces la idea de relacionarlo con la distribución uniforme, pues se puede suponer que seleccionar un tanque en particular es igualmente probable que elegir cualquier otro, de esta manera se pudo desarrollar una forma para estimar el número total de tanques fabricados.

En general esta distribución tiene importancia en relación con fenómenos en los cuales se supone un comportamiento de igual probabilidad para cada elemento del dominio. Otra forma de utilizar esta distribución es cuando se tiene una muestra aleatoria $x_1, x_2, \dots, x_n$ y se sospecha que son igualmente probables cada uno de los resultados.

Dado que al almacenar datos de manera automática quedan indexados, se podría simular valores de la uniforme discreta; para después tomar cada valor simulado como un índice y separar todos los datos que tengan tales índices, es decir se estaría obteniendo una muestra aleatoria con reemplazo, pues un cierto índice es posible que se repita.

Características principales

La función de masa de probabilidad para esta distribución cuando se define el dominio entre dos enteros i y n, es la siguiente

$$\boldsymbol{P(X=j)=\frac{1}{n-i+1}} \qquad \boldsymbol{i \le j \le n\ ;\ i,n \in N}$$

Al ser iguales todas sus probabilidades se puede expresar la función de distribución acumulativa de la siguiente forma

$$\boldsymbol{P(X \le j)=\sum_{k=i}^{j} P(X=j)=\frac{j-i+1}{n-i+1}} \qquad \boldsymbol{i \le j \le n}$$

La media de esta distribución está dada por la siguiente suma

$$\boldsymbol{E(X) = \sum_{k=i}^{n} kP(X = j) = \frac{1}{n - i + 1} \sum_{k=i}^{n} k}$$

La cual depende del signo de los enteros i, n, pero es importante señalar que la esperanza siempre toma el valor del punto medio entre tales enteros. En cuanto a su varianza su expresión es como sigue

$$\boldsymbol{VAR(X) = \sum_{k=i}^{n} k^2 P(X = j) - \left(\sum_{k=i}^{n} kP(X = j) \right)^2}$$

Simulación de valores

Para generar valores de una uniforme discreta se parte un valor U uniforme en $(0,1)$, luego se le multiplica por la distancia que separa los dos enteros que definen al dominio, que es $(n - i + 1)$, generando entonces valores entre 0 y $(n - i + 1)$, pero los cuales tienen decimales por lo cual se deben truncar, por medio de una función piso, y así generar valores entre 0 y $n - i$. Por último si a estos valores se les suma i, se obtienen valores entre i y n, que es lo que se desea. Entonces el algoritmo para generar enteros con igual probabilidad entre un entero i y otro entero n es el siguiente

1. $\boldsymbol{Generar\ U \sim U(0,1)}$
2. $\boldsymbol{hacer\ I = Entero[(n - i + 1)U] + i}$

La eficacia del algoritmo es mejor entendida de manera práctica por lo que se generaron 5000 valores de la distribución, además para ver que el algoritmo sirve para cualesquiera dos enteros se eligieron $i = -3$ y $n = 2$. La tabla 1.1 muestra de manera parcial los números pseudoaleatorios utilizados junto con la transformación que indica el algoritmo; además al final de la tabla se hallan un estimador de la media $\bar{X}$ junto con un estimador de la varianza S^2.

Tabla 1.1: Valores simulados de una Uniforme discreta en (-3,2)

j	$U_j \sim U(0,1)$	$x_j = [(n - i + 1)U_j] + i$	j	$U_j \sim U(0,1)$	$x_j = [(n - i + 1)U_j] + i$
1	0.42002	-1	6	0.44164	-1
2	0.90095	2	7	0.67309	1
3	0.07322	-3	8	0.70471	1
4	0.54471	0	9	0.04434	-3
5	0.19402	-2	10	0.00668	-3
			$\bar{X} = -0,496$		$S^2 = 2.92856$

Los valores de la esperanza y la varianza según los enteros n e i seleccionados son

$$\boldsymbol{E(X) = -0.5}$$

$$\boldsymbol{VAR(X) = 2.91\overline{6}}$$

Con lo que se puede notar la gran cercanía de los valores estimados con los teóricos, aunque esto no asegura nada sobre el comportamiento de los datos. Para tener una

visión completa de los datos en un formato compacto, se puede recurrir a una tabla de frecuencias.

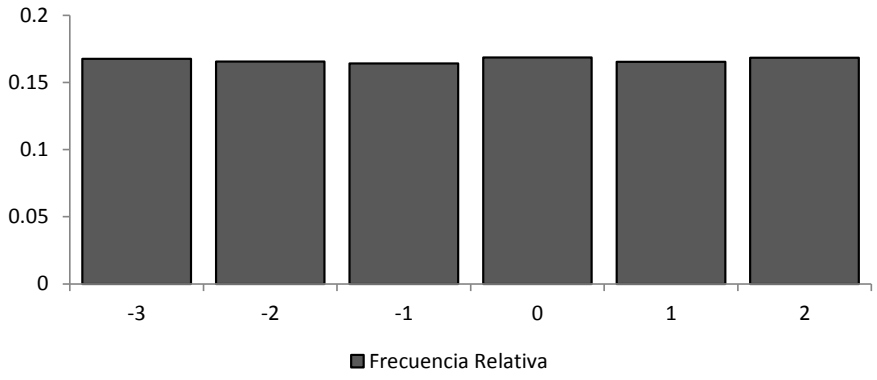
La forma de construirla es por medio de un conteo del número de observaciones con un valor determinado, y llevar a cabo el conteo por cada valor posible de la distribución; aunque es posible formar una serie de subconjuntos a partir del soporte de la distribución, sin embargo, al ser pocos los valores, no es de gran utilidad. Junto con las frecuencias también es útil incluir las frecuencias relativas, calculadas como la frecuencia de un grupo entre el número de observaciones, pues termina siendo el porcentaje de datos por cada grupo, pudiéndole interpretar como una probabilidad empírica.

Para realizar la tabla se puede emplear la función de Excel $FRECUENCIA()$ la cual recibe de parámetros una matriz de datos y un vector de grupos en los cuales se contarán los datos pertenecientes. La tabla 1.1 contiene el resultado del conteo antes descrito sobre los datos simulados, además para visualizar el comportamiento de los datos la gráfica 1.1 muestra las frecuencias relativas graficadas por cada grupo.

Tabla 1.2: Frecuencias de los valores simulados de una uniforme(-3,2)

valor	*Frecuencia*	*Frecuencia Relativa*
-3	838	0.1676
-2	828	0.1656
-1	821	0.1642
0	844	0.1688
1	827	0.1654
2	842	0.1684
Total	5000	1.00

Gráfica 1.1:Frecuencias relativas de los datos conseguidos vía simulación



La forma de evaluar estadísticamente que los valores simulados son uniformes es a través de la prueba de bondad de ajuste Ji-cuadrada, la cual parte de la idea comparar las frecuencias de una subdivisión específica en k conjuntos de la muestra, contra las frecuencias que se esperarían tuvieran los subintervalos de la subdivisión, dado que

los datos se distribuyen como una variable que se debe proponer, la cual debe estar totalmente especificada.

Como cualquier prueba de hipótesis, esta debe partir del contraste entre dos hipótesis, las cuales, para este caso, son:

$$\boldsymbol{H_0}: \boldsymbol{F(x) = F_0(x)} \ \forall \boldsymbol{x}$$

$$\boldsymbol{H_1}: \boldsymbol{F(x) \neq F_0(x)\ para\ alguna\ x}$$

Donde F es la función de distribución de los datos y F_0 es una distribución especificada. Puesto que para la distribución uniforme la subdivisión es natural, se puede tomar como referencia la tabla de frecuencias anterior, a las que ahora se denotarán como O_i, sin embargo se debe introducir el cálculo de las frecuencias esperadas E_i. Como cada valor tiene una probabilidad teórica de $1/6$ entonces el valor esperado de la frecuencia de cada intervalo es $E_i = 5000 * \left(\frac{1}{6}\right) = 833.33$; conociendo tal valor se compara la frecuencia empírica de cada valor con la formula $\frac{(O_i - E_i)^2}{E_i}$; después usándolos para calcular el siguiente estadístico de prueba:

$$\boldsymbol{\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}}$$

El cual asintóticamente se distribuye como una distribución Ji-cuadrada con $k-1$ grados de libertad, por lo que se puede obtener el $p - value = P\left(\chi^2 \geq \chi^2_{(k-1)}\right)$, por medio de la función$DISTR.CHI()$. La regla de decisión con un nivel de significancia α, después de haber calculado los valores anteriores es rechazar la hipótesis H_0 si $p - value \leq \alpha$. La tabla 1.3 resume los cálculos de diferencias entre las frecuencias de cada categoría y el valor esperado.

Tabla 1.3: Cálculos necesarios para realizar la prueba Ji-Cuadrada

i	$i-ésima\ categoría$	$\frac{(O_i - E_i)^2}{E_i}$
1	-3	0.02613
2	-2	0.03413
3	-1	0.18253
4	0	0.13653
5	1	0.04813
6	2	0.09013
$\chi^2 =$		**0.5176**
$p - value =$		**0.9914**

Dado que las diferencias entre las frecuencias empíricas y la esperada no son grandes, los valores individuales aportan poco al valor total del estadístico de prueba, implicando que la probabilidad de que sea al menos tan grande como su valor sea cercana a 1, así que aplicando el criterio o regla de decisión no se rechaza la hipótesis

nula $H_0: F(x)\ es\ Uniforma\ discreta\ en\ (-3,2)$, concluyendo que no existen diferencias significativas entre la distribución de los datos y la de una uniforme (-3,2).

Bernoulli

Introducción

La variable aleatoria Bernoulli está asociada a un único experimento en el que los resultados solamente pueden ser éxito o fracaso asignándoles el valor de 1 o 0 respectivamente, donde la probabilidad de obtener un éxito es p; al tipo de experimentos de este tipo se les llama ensayos Bernoulli, en honor al matemático suizo Jacob Bernoulli.

El campo de aplicaciones que se le han dado a esta variable son amplias pues es base de la construcción para algunas de las distribuciones, vistas más adelante como la distribución Binomial, también en el ámbito estadístico son el tipo de ensayos Bernoulli los que se asocian con la construcción de tablas de contingencia, partiendo del muestreo consecutivo de individuos en una población, donde la aparición de una característica o respuesta especifica es considera como un éxito y fracaso en caso contrario.

Características principales

La función de masa de probabilidad de esta variable es

$$\boldsymbol{P(X=x)=\begin{cases} p & si\ x=1 \\ 1-p & si\ x=0 \end{cases}} \qquad \mathbf{0}<p<1$$

La cual puede expresarse también como

$$\boldsymbol{P(X=x)=p^x(1-p)^{1-x}} \qquad \boldsymbol{x}\in\{\mathbf{0},\mathbf{1}\}$$

El valor esperado de experimentos cuyo resultado es éxito, es la media de la distribución también llamada la Esperanza de X denotada como $E(X)$, mientras que la variabilidad cuadrática entre la media y los resultados es explicada por su varianza $VAR(X)$. Las fórmulas para estos valores se resumen a continuación

$$\boldsymbol{E(X)=\sum_{x=0}^{1} xP(X=x)=p}$$

$$\boldsymbol{VAR(X)=\sum_{x=0}^{1} x^2P(X=x)-p^2=p\,(1-p)}$$

Simulación de valores

La motivación de generar valores de esta distribución es aproximar el comportamiento de un experimento con dos posibles resultados. El siguiente algoritmo de simulación para generar un valor Bernoulli, es de cierta manera intuitivo al recordar que la distribución toma los valores 1 con probabilidad p o 0 con probabilidad $1-p$.

1. $\boldsymbol{Generar\ U\sim U(0,1)}$

2. $\boldsymbol{si\ U \leq p\ hacer\ X = 1\ en\ otro\ caso\ hacer\ X = 0}$

Para comprobar este algoritmo se han generado 5000 números pseudoaleatorios para después asignarles vía una instrucción condicional, el valor 1 con probabilidad $p = 0.2$ y 0 con probabilidad $1 - p = 0.8$. En la tabla 1.4 se observa una parte de la muestra generada junto con la estimaciones de la esperanza y la varianza, $\bar{X}$ y S^2.

Tabla 1.4: Valores condicionados a ser 1 si $U_j \leq 0.2$ o 0 en otro caso

j	$U_j \sim U(0,1)$	$x_1 \sim Ber(0.2)$	j	$U_j \sim U(0,1)$	$x_1 \sim Ber(0.2)$
1	0.1099	1	6	0.1831	1
2	0.4703	0	7	0.4227	0
3	0.0319	1	8	0.4049	0
4	0.8484	0	9	0.8709	0
5	0.388	0	10	0.1044	1
			$\bar{X} = 0.201$		$S^2 = 0.16$

Se logra notar que el valor de la media es cercano al parámetro p el cual también es el valor de la esperanza, por otra parte, el estimador S^2 se aproxima al valor $p(1-p) = 0.16$. Debido a la naturaleza de la variable, la media, que es calculada como la suma de las observaciones entre el número total de observaciones, donde en realidad se realiza un conteo del número de veces que se obtuvo el valor 1, después se divide entre el numero de observaciones, con lo cual se obtiene el porcentaje de éxitos empíricos vía simulación. Por lo tanto el comportamiento de los datos en general se aproxima a una distribución del tipo Bernoulli ($p = 0.2$).

Para evaluar y concluir de manera objetiva que los datos cumplen con las proporciones antes mencionadas se puede usar la prueba de bondad de ajuste Ji-cuadrada la cual contrasta las siguientes hipótesis

$$\boldsymbol{H_0: F(x) = F_0(x)} \quad \forall \boldsymbol{x}$$

$$\boldsymbol{H_1: F(x) \neq F_0(x)\ para\ algunas\ x}$$

Para una función F_0 de distribución especificada, que en este caso es la función de distribución de una variable Bernoulli ($p = 0.2$), y F es la distribución de los datos. Entonces lo que se compara para la prueba son las frecuencias esperadas E_i por cada valor del soporte contra las frecuencias empíricasO_i con la operación $\frac{(O_i - E_i)^2}{E_i}$ para después calcular el estadístico de prueba

$$\boldsymbol{\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}}$$

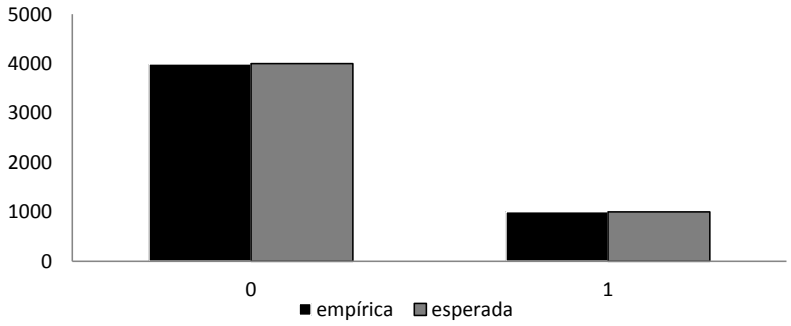
El cual se distribuye asintóticamente como una distribución Ji-cuadrada con $k-1$ grados de libertad conforme el tamaño de la muestra crece y donde k es el número de categorías en el que se divide la muestra que para este caso $k = 2$. Con el valor del estadístico se calcula el $p - value = P(\chi^2 \geq \chi^2_{(k-1)})$ para después concluir con un nivel de significancia α, el rechazar la hipótesis H_0 si $p - value \leq \alpha$.

Para la distribución Bernoulli se espera que el número de éxitos sea una proporción p de la muestra es decir $5000 * p = 5000 * .2 = 1000$. La tabla 1.5 plasma los cálculos hechos para llevar a cabo la prueba Ji-cuadrada. Por otra parte, la gráfica 1.2 muestra los valores de la tabla para poder compararlos de manera visual.

Tabla 1.5: Cálculos para realizar la prueba Chi-Cuadrada

$Categoría$	$Frecuencia\ O_i$	$Frecuencia\ Esperada\ E_i$	$\frac{(O_i - E_i)^2}{E_i}$
0	3,995	4,000	0.00625
1	1,005	1,000	0.025
$\chi^2 =$	**0.0312**	$p - value =$	**0.8596**

Gráfica 1.2: comparacion de los valores de la frecuencias esperadas y empiricas



Tomando un nivel de significancia $\alpha = 0.05$ se puede notar que lleva al resultado de no rechazar la hipótesis nula H_0 lo que implica que estadísticamente no existe una diferencia significativa entre la distribución Bernoulli(0.2) y la función de distribución de los datos simulados; lo cual también se puede entender porque la gráfica muestra que los valores de la tabla son prácticamente los mismos, considerando el tamaño de muestra que se eligió.

Binomial

Introducción

La distribución Binomial de parámetros n y p describe un fenómeno de la realización de n intentos, de tipo bernoulli independientes, sobre un experimento en el cual la probabilidad de que ocurra el evento estudiado es p, y tal probabilidad se mantiene constante para los n intentos.

Las aplicaciones que tiene la interpretación anterior es muy amplia, por ejemplo, dentro de la genética se han estudiado las características heredadas dentro de una población específica, también en ecología en el experimento de marcar un cierto número de animales, para dejarlos libres sobre ciertas zonas y después estimar a partir de un muestreo posterior, su sobrevivencia.

Dentro de las técnicas estadísticas se han encontrado aplicaciones de la distribución Binomial, pues los estadísticos de prueba se distribuyen de manera Binomial, en la pruebas de hipótesis de signos la cual contrasta si los datos de dos muestras provienen de la misma población y la prueba de McNemar que identifica el impacto que tiene un experimento sobre la respuesta dicotómica de una población.

Existen eventos a los cuales se les puede asociar con una distribución Binomial. Por ejemplo, dentro de los procesos industriales para un cierto tamaño de producción es de interés saber la proporción de productos que tienen algún defecto de fabricación. Otro ejemplo relacionado se encuentra dentro de las encuestas de opinión, pues se elige de manera aleatoria un número finito de personas de una población, para saber la proporción de personas que responderán de una manera específica a una pregunta.

Características principales

Esta variable aleatoria tiene dominio en el conjunto finito $\{0,1,\dots,n\}$, con la siguiente función de masa de probabilidad, la cual es similar a la función de masa de probabilidad de una variable Bernoulli con la diferencia de que si se valúa en un valor j, entonces los exponentes reflejan los j éxitos que ocurren con probabilidad p y los $n-j$ fracasos con probabilidad $1-p$, multiplicado por las combinaciones con repetición de los j elementos elegidos de los posibles n, como se muestra a continuación

$$\boldsymbol{P(X=j)=\binom{n}{j}p^j(1-p)^{n-j}}$$

$$\boldsymbol{0\leq j\leq n\,;\;\; n\in N; 0<p<1}$$

Su función de probabilidad acumulativa se define como

$$\boldsymbol{P(X\leq x)=\sum_{j=0}^{x}P(X=j)=\sum_{j=0}^{x}\binom{n}{j}p^j(1-p)^{n-j}}$$

Su cálculo es complicado cuando el valor de x aumenta por lo cual se usan aproximaciones de las probabilidades anteriores, aunque existen diversos métodos para realizar la aproximación, una de las más utilizadas por simplicidad es por medio de una aproximación a la variable aleatoria Normal, donde el uso de esta aproximación está justificada por el teorema de $De\ Moivre-Laplace$.

El teorema enuncia que a medida que n aumenta se pueden aproximar las probabilidades de una Binomial por medio de una variable con distribución Normal. Previo a desarrollar el método primero se debe definir la media y la varianza de la variable Binomial las cuales son

$$\boldsymbol{E(X)=\sum_{j=0}^{n}jP(X=j)=np}$$

$$\boldsymbol{VAR(X)=\sum_{j=0}^{x}j^2P(X=j)-(np)^2=np(1-p)}$$

El método de aproximación, requiere de estandarizar la variable de la siguiente forma

$$\frac{\boldsymbol{X - np + .5}}{\sqrt{\boldsymbol{np(1-p)}}}$$

De la fórmula anterior se puede notar que la estandarización para realizar la estimación de la probabilidad, es lograda restando la media y dividiendo entre la raíz de la varianza de la distribución y el término sumado de 0.5 es debido a una corrección por continuidad. El valor anterior es estimado por medio de una Normal(0,1) (estándar), y se lleva a cabo con la siguiente formula.

$$\boldsymbol{P(X \leq j) \approx (2\pi)^{-1/2} \int_{-\infty}^{j-np+.5/\sqrt{np(1-p)}} e^{-t^2}\, dt}$$

Los valores de la integral pueden hallarse en tablas especiales de valores de una Normal estándar[2] o por la función de Excel DISTR.NORM.ESTAND().

Simulación de valores

Simular un valor de la distribución Binomial es sencillo tomando como referencia inicial que lo que se está realizando es un ensayo del tipo Bernoulli de parámetro p del cual se obtienen como resultado 0 o 1, si se realiza n ensayos de manera consecutiva y cada uno generado de forma independiente y al final se suman los resultados, con lo cual se tendrá el número total de éxitos en n experimentos que es la descripción de una variable binomial. Por lo tanto, se pueden generar n valores Bernoulli y sumarlos para generar un valor de la variable Binomial. El método anterior se puede resumir con el siguiente algoritmo

1. $\boldsymbol{Generar\ y_1, y_2, \dots, y_n\ independientes\ distribuidas\ Bernoulli(p)}$
2. $\boldsymbol{Hacer\ \ X = \sum_{i=1}^{n} y_i}$

Puesto que a medida que n aumenta las operaciones a realizarse crecen manera lineal, para parámetros grandes una alternativa, dado que el dominio de la distribución binomial es finito, es utilizar el método de la transformación inversa, aplicando un método eficiente para hallar la función inversa generalizada.

Poniendo a prueba el algoritmo anterior se han simulado 5000 valores de la distribución Binomial con los parámetros $n = 20,\ p = 2/5$, de esta manera la esperanza teórica debe ser $E(X) = 8$, además de tener una varianza de $VAR(X) = \frac{24}{5} = 4.8$; una muestra parcial de la simulación, junto con los estimadores $\bar{X}, S^2$ de la esperanza y la varianza respectivamente, pueden observarse en la Tabla 1.6.

[2] La tabla de la Normal Estándar se puede hallar en el libro de Gibbons (2003).

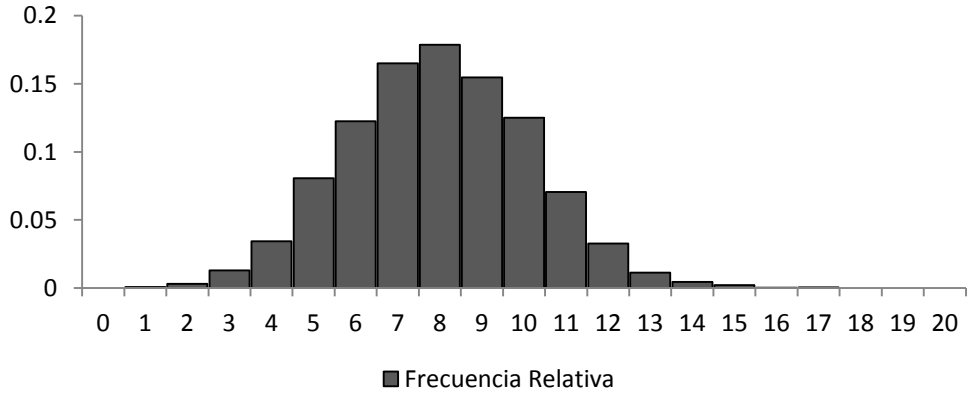
Tabla 1.6: Muestra parcial de valores simulados de una Binomial(20;0.4)			
i	$x_i \sim Bin(20;0.4)$	i	$x_i \sim Bin(20;0.4)$
1	6	6	5
2	12	7	9
3	8	8	9
4	9	9	11
5	12	10	10
$\bar{X} = 7.97$		$S^2 = 4.81$	

El estimador de la media es bastante cercano al valor de la esperanza calculada con los parámetros, la varianza por su parte, también toma un valor cercano al teórico. Para entender de mejor manera el comportamiento de los datos se puede recurrir, una vez más, a una tabla de frecuencias. Para construirla se debe dividir soporte de la distribución en grupos y contar el número de datos perteneciente a cada uno. También es útil incluir las frecuencias relativas, calculadas a partir de las frecuencias divididas entre el número total de datos, lo cual da una perspectiva de probabilidad empírica del grupo.

Como el soporte de la variable es de cardinalidad finita, una opción natural es tomar cada valor posible como un grupo; para llevar a cabo la tabla se puede emplear la función $FRECUENCIA()$ de Excel que recibe como parámetros un conjunto de valores, además de una lista de las distintas clases a agrupar, entonces se debe aplicar sobre los datos y a una lista de los valores posibles de la distribución. La Tabla 1.7 es el resultado de aplicar el proceso anterior a los datos simulados, mientras que para tener una visión del comportamiento de los datos la gráfica 1.3 muestra las frecuencias relativas de cada grupo.

Tabla 1.7: Frecuencias y frecuencias relativas de los 5000 valores simulados					
valor	*Frecuencia*	*Frecuencia Relativa*	*valor*	*Frecuencia*	*Frecuencia Relativa*
0	0	0	11	353	0.0706
1	4	0.0008	12	164	0.0328
2	15	0.003	13	57	0.0114
3	65	0.013	14	22	0.0044
4	172	0.0344	15	11	0.0022
5	403	0.0806	16	1	0.0002
6	613	0.1226	17	2	4
7	825	0.165	18	0	0
8	894	0.1788	19	0	0
9	773	0.1546	20	0	0
10	626	0.1252			

Gráfica 1.3: Frecuencias relativas de los datos simulados de una Binomial(20;0.4)



Sin embargo para probar que los datos provienen de una distribución Binomial(20;0.4) se realizará una prueba de hipótesis de bondad de ajuste Ji-cuadrada. Esta prueba contrasta las siguientes hipótesis

$$\boldsymbol{H_0}: \boldsymbol{F(x) = F_0(x)} \quad \forall \boldsymbol{x}$$

$$\boldsymbol{H_1}: \boldsymbol{F(x) \neq F_0(x)\ para\ algunas\ x}$$

Donde $\boldsymbol{F_0}$ es la función de distribución acumulativa de la variable Binomial(20;0.4). Para llevar a cabo la prueba primero se debe volver a subdividir el soporte de la distribución en k categorías denotadas ahora como C_i. Por la recomendación que se encuentra en el libro de Gibbons en la cual se menciona que debido a que la distribución JI-cuadrada es una aproximación a la verdadera distribución del estadístico que se desarrolla con el cociente verosimilitudes, por lo que, como criterio conservador, cada categoría debe tener una frecuencia de al menos 5. Lo anterior se resuelve mediante una agrupación en las colas de la distribución, además entre más valores contenga cada categoría menos son las operaciones que deben realizarse, lo que se traduce en mayor simplicidad.

Para este caso se agrupan los primeros tres valores del soporte en una categoría, para después tomar como valor individual a cada valor hasta el número 13, para después agrupar los últimos valores en una última categoría, así que se clasifican en 13 categorías en total. Después se obtiene la frecuencia de cada categoría ahora denotada por O_i; para luego calcular el valor esperado de valores que cada categoría debería tener.

Las frecuencias esperadas E_i son calculadas como el número de observaciones n, multiplicado por la probabilidad teórica de cada categoría siguiendo la distribución propuesta P_i, entonces para la primera categoría $P_1 = P(X \leq 2)$, para las demás categorías $P_i = P(X = i + 1)\ ; 2 \leq i \leq 13$ y para la última categoría $P_{13} = P(X > 13) = 1 - P(X \leq 13)$.

Posteriormente se procede a realizar una comparación de cada categoría por medio de la operación $\frac{(O_i - E_i)^2}{E_i}$ los cuales son usados para formar el estadístico de prueba

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

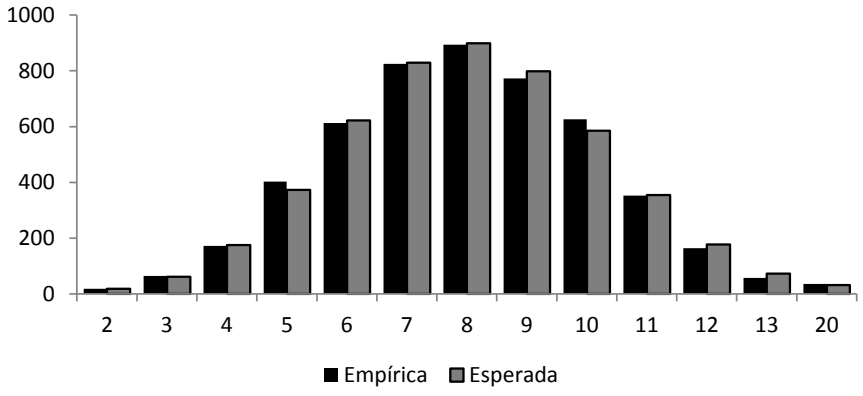
Una vez calculado el estadístico de prueba, se calcula el $p - value = P\left(\chi^2 \geq \chi^2_{(k-1)}\right)$, por medio de funciones de Excel, por último, la regla de decisión para esta prueba establece que con un nivel de significancia α rechazar la hipótesis nula H_0 si $p - value \leq \alpha$.

Para el caso de los valores simulados los cálculos anteriores se hallan resumidos en la Tabla 1.8, por otra parte la Gráfica 1.4 compara de manera visual los valores de las frecuencias esperadas y empíricas, con la finalidad de entender mejor el comportamiento de las diferencias, y el porqué de la futura conclusión.

Tabla 1.8: Cálculos necesarios para realizar la prueba Ji-Cuadrada

i	C_i	O_i	E_i	$\frac{(O_i - E_i)^2}{E_i}$
1	$\{0,1,2\}$	19	18.057	0.049
2	$\{3\}$	65	61.748	0.171
3	$\{4\}$	172	174.954	0.05
4	$\{5\}$	403	373.235	2.374
5	$\{6\}$	613	622.058	0.132
6	$\{7\}$	825	829.411	0.023
7	$\{8\}$	894	898.529	0.023
8	$\{9\}$	773	798.692	0.826
9	$\{10\}$	626	585.708	2.772
10	$\{11\}$	353	354.974	0.011
11	$\{12\}$	164	177.487	1.025
12	$\{13\}$	57	72.815	3.435
13	$\{14,15,16,17,18,19,20\}$	36	32.329	0.417
	$\chi^2 =$	11.308	$p - value =$	0.503

Grafíca 1.4: Comparación de las frecuencias empíricas contra esperadas.



Si se considera un nivel de significancia de $\alpha = 0.05$ se puede concluir dada la regla de decisión anterior, en no rechazar la hipótesis H_0, lo que también se puede interpretar en que no existen diferencias significativas entre la función de distribución de los datos y la función de distribución de la variable Binomial(20;0.4). Analizando la grafica se puede identificar de manera rápida cuales son las categorías más diferenciadas, las cuales si se toma como referencia a la tabla anterior son las que más peso aportan al valor del estadístico de prueba, sin embargo, en conjunto no se considera que sean significativas.

Geométrica

Introducción

La descripción de la variable discreta Geométrica es una secuencia de Y ensayos independientesde tipo Bernoulli en los cuales se puede obtener un éxito con probabilidad p o fracaso con probabilidad $(1-p)$, esta distribución se define entonces como el número $X = Y - 1$ de fracasos $X = 0,1,2,\ldots$ antes de obtener el primer el éxito.

Aplicaciones de esta distribución se pueden encontrar en estudios de temas Biométricos, como por ejemplo al estudiar el número consecutivo de árboles sanos dentro de un bosque que se considera en general infectado; en teoría de colas y modelos con Cadenas de Markov también se pueden encontrar aplicaciones, en estudios de datos de mortalidad y rentabilidad.

Características principales

Para caracterizar esta variable en primer lugar es posible enunciar la función de masa de probabilidad de la distribución Geométrica que está dada por a la siguiente expresión:

$$\boldsymbol{P(X = x) = p(1-p)^x} \quad \boldsymbol{x} \in \mathbb{N} \cup \{\mathbf{0}\}; \quad \mathbf{0} < p < 1$$

También es posible obtener la función de distribución acumulada por medio de la progresión geométrica

$$\boldsymbol{P(X \leq x) = \sum_{i=0}^{x} p(1-p)^i = \frac{p(1-(1-p)^x)}{1-(1-p)} = 1-(1-p)^x}$$

La esperanza y la varianza se pueden obtener de igual manera por medio de progresiones geométricas y otras técnicas de cálculo diferencial, pero, en resumen, su forma de cálculo es la siguiente

$$\boldsymbol{E(X) = \frac{1-p}{p} = \frac{1}{p} - 1}$$

$$\boldsymbol{VAR(X) = \frac{1-p}{p^2}}$$

Simulación de valores

Para la simulación de valores de esta variable se puede aplicar el método de la inversa, igualando a un valor $U \sim U(0,1)$ la función de probabilidad acumulada y despejando x, de la siguiente manera

$$\boldsymbol{U = 1 - (1 - p)^x}$$

$$\boldsymbol{x ln(1 - p) = \ln(U)}$$

$$\boldsymbol{x = \frac{\ln(U)}{\ln(1 - p)}}$$

Para conseguir un valor dentro del dominio de la variable, a este resultado debe aplicarse la función $Entero()$, es decir un truncamiento a valor entero. De acuerdo la fórmula para la simulación, el valor de cero se obtendrá en los casos que los valores generados de U sean cercanos a 1. El algoritmo de simulación es el siguiente

1. $\boldsymbol{Generar\ U \sim U(0,1)}$
2. $\boldsymbol{Regresar\ X = Entero[\ln(U)/\ln(1 - p)]}$

Pueden surgir dificultades numéricas en el caso en que el parámetro p sea cercano ya sea a 0 o a 1; una alternativa es volver a la idea inicial del método tomando los éxitos y fracasos como ensayos Bernoulli, iniciando una secuencia de valores con distribución Bernoulli de parámetro p, contando el número de ceros obtenidos (iteraciones) y parar hasta haber obtenido el primer uno.

La funcionalidad de este algoritmo se comprueba generando una muestra, en este caso de tamaño 5000 con parámetro $p = 0.4$, los resultados de manera parcial se pueden ver en la tabla 1.9, junto con las estimaciones de la media $\bar{X}$ y la varianza S^2.

Tabla 1.9: Valores simulados de una distribución Geométrica(p=0.7)

i	$U_i = U(0,1)$	$x_i = \left[\frac{\ln(U_i)}{\ln(1-p)}\right]$	i	$U_i = U(0,1)$	$x_i = \left[\frac{\ln(U_i)}{\ln(1-p)}\right]$
1	0.1602	3	6	0.7779	0
2	0.5766	1	7	0.0627	5
3	0.2432	2	8	0.6562	0
4	0.2807	2	9	0.3259	2
5	0.2482	2	10	0.9436	0
		$\bar{X} = 1.474$		$S^2 = 3.798$	

Puesto que el valor de la media y la varianza calculadas con los valores seleccionados para los parámetros de la distribución geométrica son $E(X) = \frac{0.6}{0.4} = 1.5$ y $VAR(X) = \frac{0.6}{0.4^2} = 3.75$, se nota que las estimaciones son bastante cercanas, pero para comprender mejor el comportamiento de los datos, también se revisarán por medio de una tabla de frecuencias.

Para realizarla se divide en grupos el soporte de la distribución, pero a pesar de que el soporte no es finito, en este caso los valores simulados presentaron un valor máximo de 17, por lo cual se toma como grupo a cada valor individual de 1 hasta 17. Luego

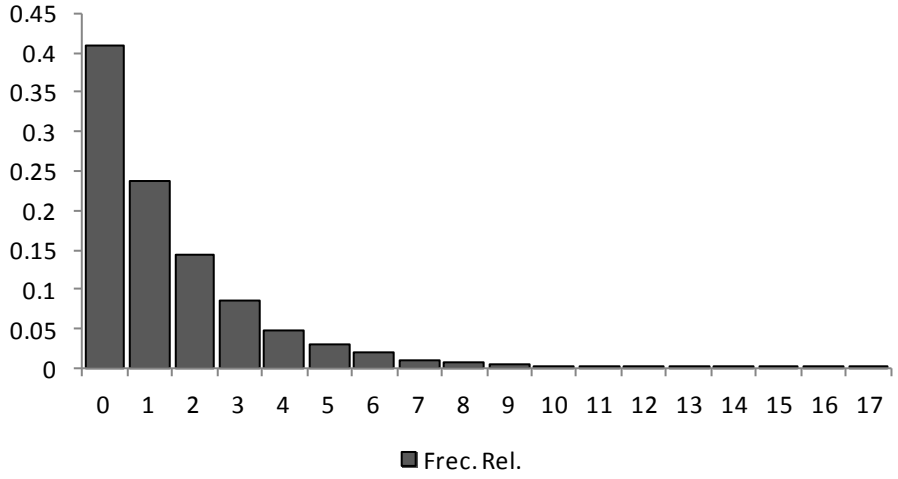
con la ayuda de la función $FRECUENCIA()$ de Excel, se realiza un conteo del número de datos pertenecientes a cada grupo; por otra parte, también se puede incluir en la tabla a las frecuencias relativas, tomadas como cada frecuencia entre el número de simulaciones, pues dan una perspectiva empírica de la probabilidad de cada grupo.

El resultado de llevar a cabo las operaciones anteriores se puede ver en la Tabla 1.10, mientras que, para tener una visión más clara del uso de la tabla, la gráfica 1.5 está hecha a partir de las frecuencias relativas de la tabla.

Tabla 1.10: Frecuencias y frecuencias relativas de los valores simulados de una Geométrica(p=0.4)

Valor	*Frecuencia*	*Frecuencia Relativa*	*Valor*	*Frecuencia*	*Frecuencia Relativa*
0	2048	0.4096	10	11	0.0022
1	1186	0.2372	11	5	0.001
2	716	0.1432	12	3	0.0006
3	429	0.0858	13	4	0.0008
4	237	0.0474	14	2	0.0004
5	149	0.0298	15	2	0.0004
6	99	0.0198	16	2	0.0004
7	50	0.01	17	1	0.0002
8	34	0.0068			
9	22	0.0044			

Gráfica 1.5: Frecuencias relativas de los valores simulados de una Geo(0.4)



A pesar de un comportamiento similar que se nota en el gráfico de los datos, a la distribución de la variable aleatoria objetivo, se debe evaluar de manera objetiva la simulación, lo cual se realiza con la prueba de bondad de ajuste Ji-cuadrada. Esta prueba parte de contrastar las hipótesis

$$\boldsymbol{H_0: F_0(x) = F(x) \quad \forall x}$$

$$\boldsymbol{H_1: F_0(x) \neq F(x) \; para \; alguna \; x}$$

Donde F es la función de distribución de los datos y F_0 es la función de distribución de la variable aleatoria Geométrica(0.4). Esta prueba inicia subdividiendo el soporte de la distribución en k categorías C_i, pero esta vez por recomendación antes mencionada se procura que cada categoría contenga al menos 5 valores, lo cual se remedia fácilmente agrupando en una sola categoría la cola de la distribución, obteniendo su frecuencia sumando las frecuencias de las categorías agrupadas; para este caso se decidió agrupar a partir de valor 12 en adelante.

En este caso puesto que X es una variable Geométrica(0.4) las probabilidades son $P(X = i),\ 0 \leq i \leq 11$ y para la última categoría la probabilidad es $P(X > 11) = 1 - P(X \leq 11)$. Los cálculos realizados para la prueba Ji-cuadrada en los datos simulados se resumen en la Tabla 1.11, además en la Gráfica 1.6 se observan las comparaciones de una manera más intuitiva de los valores de la tabla.

Gráfica 1.6: comparacion de frecuencias esperadas VS empíricas

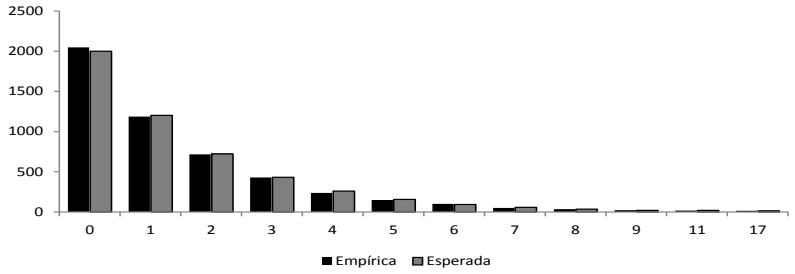


Tabla 1.11: Cálculos necesarios para la prueba Ji-Cuadrada

i	C_i	O_i	E_i	$\frac{(O_i - E_i)^2}{E_i}$
1	{0}	2048	2000	1.152
2	{1}	1186	1200	0.163
3	{2}	716	720	0.022
4	{3}	429	432	0.021
5	{4}	237	259.2	1.901
6	{5}	149	155.52	0.273
7	{6}	99	93.31	0.347
8	{7}	50	55.99	0.64
9	{8}	34	33.59	0.005
10	{9}	22	20.16	0.169
11	{11}	16	19.35	0.58
12	{12,13,14, 15,16,17}	14	10.88	0.892
	$\chi^2 =$	6.166	$p - value = 0.862$	

Dados los valores de la tabla se concluye, que no se debe rechazar la hipótesis H_0. La forma intuitiva de interpretar el resultado es a través de la Gráfica 1.6, donde las diferencias son casi imperceptibles, y en los casos en que las frecuencias de algunas categorías tienen ciertas diferencias con las frecuencias esperadas las cuales aportan más peso al estadístico de prueba, en conjunto son consideradas como no significativas.

Binomial Negativa

Introducción

Vista como una generalización de la distribución geométrica, una variable aleatoria Binomial Negativa, denotada por $BN(r,p)$ toma la misma idea de ensayos tipo Bernoulli sucesivos de probabilidad de éxito p, con la diferencia, que ahora el conteo se realiza sobre el número de fracasos hasta haberse obtenido el r-ésimo éxito. Entonces esta variable puede verse como la suma de r variables aleatorias independientes de distribución Geométrica.

En base a lo anterior, se puede relacionar a esta distribución con el modelaje del comportamiento de fenómenos que se observan con frecuencia en diversas ramas de la ciencia, es por eso que se le puede hallar en estudios de ajuste de la distribución de datos de índole psicológico, procesos de riesgo, teoría de colas, temas relacionados con el pasar de automóviles hasta que ocurra un cierto número de accidentes, o en temas de ecología.

Características principales

Para caracterizar esta distribución se define la función de masa de probabilidad, que se obtiene por medio de la siguiente fórmula

$$\boldsymbol{P(X=x)=\binom{r+x-1}{r-1}p^r(1-p)^x}$$

La función de probabilidad acumulada se expresa de la siguiente manera:

$$\boldsymbol{P(X\leq x)=\sum_{x=0}^{x}\binom{r+x-1}{r-1}p^r(1-p)^x}$$

Para obtener la función de probabilidad acumulada, es más sencillo realizar los cálculos de las probabilidades individuales de manera recursiva para después realizar la suma, pues a partir de la probabilidad $P(X=x)$ se puede obtener a $P(X=x+1)$ de la siguiente manera

$$\boldsymbol{P(X=x+1)=\binom{r+x}{r-1}p^r(1-p)^{x+1}=\frac{(r+x)!}{(x+1)!\,(r-1)!}p^r(1-p)^{x+1}}$$

$$\boldsymbol{=\frac{r+x}{x+1}(1-p)\left(\frac{(r+x-1)!}{x!\,(r-1)!}p^r(1-p)^x\right)}$$

$$\boldsymbol{=\frac{r+x}{x+1}(1-p)P(X=x)}$$

Para obtener la esperanza y la varianza basta con aplicar la idea que la Binomial Negativa es la suma de r Geométricas por lo que la esperanza es la suma de las esperanzas y la varianza es la suma de las varianzas suponiendo independencia entre las variables, entonces al tener esperanzas y varianzas iguales los momentos se calculan como

$$\boldsymbol{E(X) = \frac{r(1-p)}{p}}$$

$$\boldsymbol{VAR(X) = \frac{r(1-p)}{p^2}}$$

Simulación de valores

Construyendo el método de simulación de valores de una distribución Binomial Negativa, puede pensarse en un conteo de la distribución geométrica, cuando se obtiene el éxito en lugar de terminar se comienza otra nueva secuencia contando los fracasos, proceso que termina hasta haber obtenido el éxito numero r, como puede notarse se tendrán r variables geométricas que al sumarse se obtiene el total de fracasos a lo largo del experimento, por tal motivo el siguiente algoritmo se basa en una suma de variables aleatorias geométricas.

1. $\boldsymbol{Generar\ y_1, y_2, y_3, \dots, y_r\ distribuidas\ geométrica\ de\ parámetro\ p}$
2. $\boldsymbol{Regresar\ X = \sum_{i=1}^{r} y_i}$

Cuando el parámetro r aumenta, el número de operaciones de la maquina aumenta, por lo cual para un ejemplo dentro de una clase, puede elegirse un valor para el parámetro r razonable para que el ejemplo se desarrolle rápidamente.

Ejemplo, para probar este algoritmo se han generado 5000 simulaciones para aproximar el comportamiento de una Binomial Negativa de parámetros $r = 20$ y $p = 0.8$; para estos valores de los parámetros los valores de la esperanza y varianza deben ser

$$\boldsymbol{E(X) = 20 * \frac{0.2}{0.8} = 5}$$

$$\boldsymbol{VAR(X) = 20 * \frac{0.2}{0.8^2} = 6.25}$$

La Tabla 1.12 resume los primeros 10 valores generados; además se hallan las estimaciones de la esperanza $\bar{X}$ y de la varianza S^2.

Tabla 1.12: Muestra parcial de la simulación de una Binomial Negativa(20,0.8)			
i	x_i	i	x_i
1	4	6	5
2	1	7	12
3	8	8	6
4	3	9	3
5	4	10	4
$\bar{X} = 4.9884$		$S^2 = 6.227$	

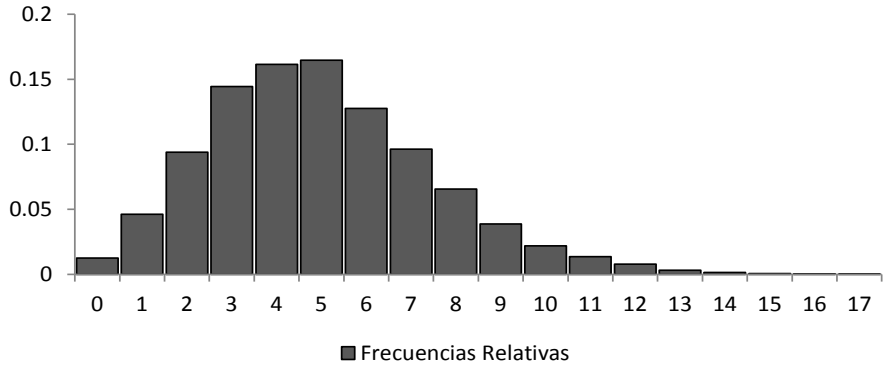
Lo primero que se puede notar es la aproximación de los estimadores de la media y la varianza a sus respectivos valores calculados con valores elegidos para los parámetros, aunque esto permite enunciar algo de forma concreta sobre el comportamiento de los datos, se tiene que recurrir a una tabla de frecuencias.

Primero se divide al soporte de la distribución en grupos para contar después el número de elementos de la muestra que pertenecen a cada grupo, aunque el soporte de esta distribución es no es finito, los valores obtenidos mediante simulación si lo son, pues el máximo de los valores generados es 17, lo que implica tener un intervalo relativamente corto de valores posibles, por lo que una opción es tomar a cada valor individual como grupo.

Otra serie de valores de interés, son las frecuencias relativas, que son las frecuencias observadas entre el número total de observaciones. Estas frecuencias relativas, aportan una información de la probabilidad empírica sobre cada grupo. La siguiente tabla contiene la información antes mencionada, además, para comprender de manera más intuitiva la funcionalidad de esta tabla, se puede observar en la gráfica 1.7 las frecuencias relativas que muestran la forma de la distribución de los datos y el peso que representa cada valor.

Tabla 1.13: Frecuencias de los valores generados de una distribución BinNeg(20;0.8)					
Valor	*Frecuencia*	*Frecuencia Relativa*	*Valor*	*Frecuencia*	*Frecuencia Relativa*
0	62	0.0124	11	68	0.0136
1	231	0.0462	12	39	0.0078
2	470	0.094	13	16	0.0032
3	722	0.1444	14	7	0.0014
4	807	0.1614	15	3	0.0006
5	823	0.1646	16	2	0.0004
6	638	0.1276	17	1	0.0002
7	481	0.0962			
8	328	0.0656			
9	193	0.0386			
10	109	0.0218			

Gráfica 1.7: Frecuencias relativas de los valores simulados de una BinNeg(20;0.8)



De manera visual el comportamiento de los datos no presenta algún tipo de perturbación clara. La forma de evaluar objetivamente que la simulación se distribuye como una Binomial Negativa, es a través de la prueba de bondad de ajuste Ji-cuadrada.

La prueba Ji-cuadrada se realiza para contrastar las siguientes 2 hipótesis

$$\boldsymbol{H_0}: \boldsymbol{F_0(x) = F(x)} \quad \forall \boldsymbol{x}$$

$$\boldsymbol{H_1}: \boldsymbol{F_0(x) \neq F(x)\ para\ alguna\ x}$$

Donde $F_0(x)$ es la función de distribución acumulativa de la variable Binomial Negativa de parámetros $r = 20, p = 0.8$.

Para realizar la prueba se continúa aplicando la recomendación de procurar definir las categorías con al menos cinco elementos en cada una, por lo cual se agrupan las categorías que forman la cola de la distribución, en una sola categoría. De a acuerdo a lo anterior, la última categoría consta de los valores mayores a 13.

Los cálculos realizados para obtener el estadístico de prueba se encuentran en la Tabla 1.14, los cuales se observan plasmados en la Gráfica 1.8 pues contiene los valores de la tabla, para poder compararlos de manera visual e intuitiva, y entender mejor que el aporte de las diferencias al estadístico que en este caso son bajas debido a la similitud de las frecuencias empíricas con las teóricas.

Gráfica 1.8: Frecuencias Esperadas VS Empíricas

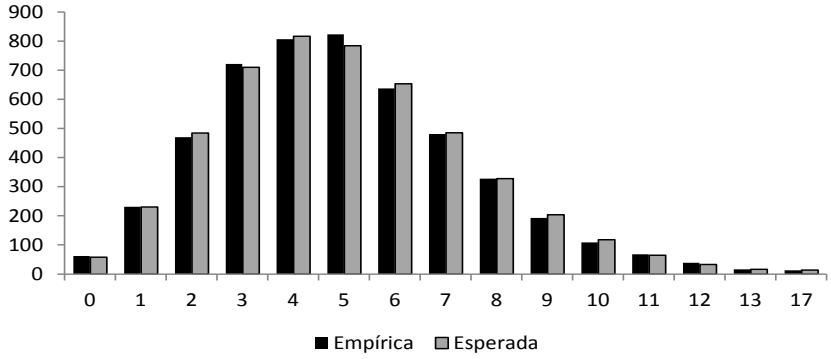


Tabla 1.14: cálculos para llevar a cabo la prueba Ji-Cuadrada				
i	$Categoria\ C_i$	$Frecuencia\ O_i$	$Frecuencia\ Esperada\ E_i$	$\frac{(O_i - E_i)^2}{E_i}$
1	0	62	57.65	0.33
2	1	231	230.58	0.00
3	2	470	484.23	0.42
4	3	722	710.20	0.20
5	4	807	816.73	0.12
6	5	823	784.06	1.93
7	6	638	653.38	0.36
8	7	481	485.37	0.04
9	8	328	327.63	0.00
10	9	193	203.86	0.58
11	10	109	118.24	0.72
12	11	68	64.49	0.19
13	12	39	33.32	0.97
14	13	16	16.40	0.01
15	19	13	13.86	0.05
	$\chi^2 = 5.917$		$p - value = 0.968$	

Si se considera un nivel de significancia de $\alpha = 0.05$, se observa en la tabla que $p - value > \alpha$, así que de acuerdo a la regla de decisión de la prueba, no se rechaza la hipótesis H_0 con lo que se concluye que, no existen diferencias significativas entre la función distribución de los datos simulados y la función de distribución de una Binomial Negativa(20;0.8). Tanto en la gráfica como en la tabla, se pueden localizar las categorías en las cuales las diferencias son mayores y por lo tanto son éstas las que más aportan al valor del estadístico de prueba, aunque en general las diferencias no se consideran como significativas.

Hipergeométrica

Introducción

Considérese un muestreo de n elementos sin remplazo sobre una población de tamaño N. La característica principal que se supone de esta población es que m elementos poseen una determinada característica de interés, y $N - m$ elementos no poseen tal característica, como por ejemplo individuos enfermos de una población o en un contexto de modelos de urnas elementos de un color específico. La distribución Hipergeométrica describe el número de elementos con la característica deseada en la muestra seleccionada de tamaño n.

Algunos ejemplos de las aplicaciones que se le han dado la distribución Hipergeométrica, se encuentran en el campo de la industria, en donde de una producción total de N productos se toma una muestra sin reemplazo para contar el numero de productos con defectos y se concluye en una mala producción si su número es mayor a una cierta cantidad c, donde el valor de c está relacionado con la

proporción máxima de productos defectuosos, de acuerdo al criterio que se tenga en el área de control de producción.

Otro caso de aplicación es en estudios biológicos donde se desea estimar el tamaño de una población, por ejemplo, de peces, donde se pueden marcar una cierta cantidad conocida de peces y mezclarlos en el ambiente de estudio, posteriormente se toma una muestra sin remplazo con la cual es posible estimar el total de la población original, considerando estos elementos como aquellos con la característica deseada.

Entre otros campos de aplicación también se le puede hallar en estudios de opinión política y en el estudio de las llamadas tablas de contingencia, donde se obtienen respuestas de tipo dicotómicas, pues si se deseara presentar los datos obtenidos posterior a un muestreo sin reemplazo, podría hacerse por medio de una tabla de $2\,X\,2$, por lo que las tablas de este tipo pueden ser estudiadas a través de la distribución hipergeométrica.

Características principales

La función de masa de probabilidad de esta distribución está dada por la siguiente expresión

$$\boldsymbol{P(X=x)=\frac{\binom{n}{x}\binom{N-n}{m-x}}{\binom{N}{m}}} \quad \mathbf{0} < n \leq N\,;\, 0 < m \leq N$$

Donde $\max(0, n+m-N) \leq x \leq \min(n, m)$, la función de probabilidad acumulativa de esta variable puede conllevar a costos computacionales altos cuando se pretende programar directamente como una función debido al cálculo de factoriales, por tal razón sólo es indicada como

$$\boldsymbol{P(X \leq x)=\sum_{j=0}^{x} P(X=j)=\sum_{j=0}^{x}\frac{\binom{n}{j}\binom{N-n}{m-j}}{\binom{N}{m}}}$$

Una forma más conveniente de obtener las probabilidades individuales y además la probabilidad acumulada es a través de la relación recursiva

$$\boldsymbol{P(X=x+1)=\frac{\binom{n}{x+1}\binom{N-n}{m-x-1}}{\binom{N}{m}}}$$

$$=\frac{\boldsymbol{n!\,(N-n)!}}{\boldsymbol{(n-x-1)!\,(x+1)!\,(N-n-m+x+1)!\,(m-x-1)!}}\Big/\binom{N}{m}$$

$$\frac{\boldsymbol{(n-x)(m-x)}}{\boldsymbol{(N-n-m+x+1)(x+1)}}\,\frac{\boldsymbol{n!\,(N-n)!}}{\boldsymbol{(n-x)!\,x!\,(N-n-m+x)!\,(m-x)!}}\Big/\binom{N}{m}$$

$$=\frac{\boldsymbol{(n-x)(m-x)}}{\boldsymbol{(N-n-m+x+1)(x+1)}}\boldsymbol{P(X=x)}$$

Además para caracterizar esta distribución es necesario indicar los valores de la esperanza y la varianza que son

$$\boldsymbol{E(X) = n * \frac{m}{N}}$$

$$\boldsymbol{VAR(X) = n\frac{m}{N}\left(1 - \frac{m}{N}\right)\frac{N-n}{N-1}}$$

Simulación de valores

La simulación de esta distribución es intuitiva, ya que reproduce paso a paso el proceso de muestreo antes explicado, considerando que m/N es la probabilidad de obtener el primer éxito y posteriormente en la nueva selección, se actualiza el valor m(los restantes con la característica de interés) y de N(el tamaño de donde se va a muestrear), de acuerdo a lo obtenido, entonces el algoritmo es:

1. $\boldsymbol{Sea\ C = N - m}$
2. $\boldsymbol{para\ i = 1\ hasta\ n}$
   - 2.1. $\boldsymbol{Generar\ U \sim U(0,1)}$
   - 2.2. $\boldsymbol{Si\ U \leq \frac{m}{N}\ hacer\ X = X + 1, m = m - 1\ en\ otro\ caso\ hacer\ C = C - 1}$
   - 2.3. $\boldsymbol{Hacer\ N = N - 1}$
3. $\boldsymbol{Regresar\ X}$

Aunque este algoritmo genera un programa de código más amplio[3], el cual generalmente no es mostrado al usuario final, es decir el estudiante, se debe mostrar su eficacia, para este propósito se generaron 5000 valores de la distribución Hipergeométrica con $N = 100,\ n = 50,\ m = 75$. En la Tabla 1.15 se pueden ver los primeros valores generados, junto con una estimación de la media $\bar{X}$ y la varianza S^2, que de acuerdo con valores elegidos para los parámetros la media y la varianza deben ser

$$\boldsymbol{E(X) = 50 * \frac{75}{100} = 37.5}$$

$$\boldsymbol{VAR(X) = 50 * \frac{75}{100}\left(1 - \frac{75}{100}\right)\frac{100-50}{99} = 4.734}$$

Tabla 1.15: muestra parcial de los valores generados de una HiperGe(N=100;n=50;m=75)

i	$HiperGe(100;50;75)$	i	$HiperGe(100;50;75)$
1	38	6	34
2	37	7	37
3	40	8	38
4	35	9	36
5	36	10	42
$\bar{X} = 37.562$		$S^2 = 4.7999$	

En la tabla se observa que el valor estimado de la media es cercano al valor teórico de la media con base en los parámetros, mientras que la estimación de la varianza es un

[3] Éste y todos los códigos expuestos se pueden consultar en la parte II del apéndice.

poco elevada, aunque esto no determina una conclusión sobre la eficacia de la simulación, pues se debe explorar el comportamiento de las observaciones, para lo que se empleará una tabla de frecuencias.

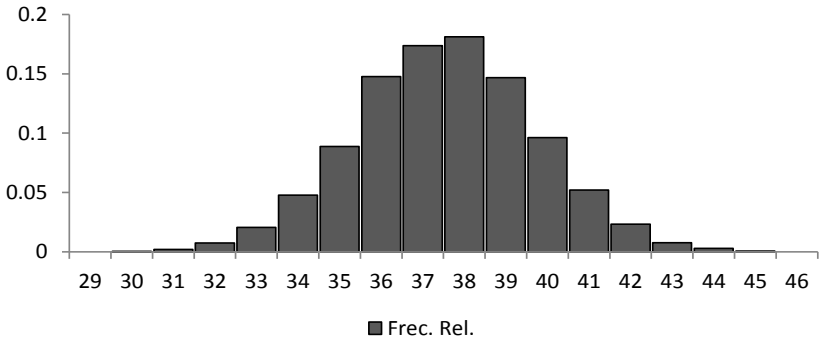
El rango de valores que puede tomar la variable de acuerdo a los valores seleccionados para los parámetros de la distribución son: $\max(0,25) = 25 \leq x \leq \min(50,75) = 50$, sin embargo la simulación mostró tener un mínimo de 29 y un máximo de 46, lo que se puede explicar si se calculan las probabilidades de los demás valores del soporte, pues son cercanas a 0.

En la Tabla 1.16 se muestran las frecuencias absolutas y relativas obtenidas a partir de un total de 5,000 simulaciones, además para visualizar el comportamiento de la distribución de los datos, la Gráfica 1.9 muestra las frecuencias relativas de la tabla.

Tabla 1.16: Frecuencias de los valores simulados de una HiperGeo(100;50;75)

Valor	*Frecuencia*	*Frecuencia Relativa*	*Valor*	*Frecuencia*	*Frecuencia Relativa*
29	1	0.0002	38	906	0.1812
30	2	0.0004	39	734	0.1468
31	10	0.002	40	481	0.0962
32	37	0.0074	41	260	0.052
33	103	0.0206	42	117	0.0234
34	239	0.0478	43	39	0.0078
35	444	0.0888	44	15	0.003
36	738	0.1476	45	4	0.0008
37	868	0.1736	46	2	0.0004

Gráfica 1.9:Frecuencias relativas de los valores imulados partir de una variable HiperGe(100;50;75)



Después de tener una idea intuitiva del comportamiento de los datos, es necesario pasar a una evaluación del comportamiento de los datos, por lo que se aplicará la prueba Ji-cuadrada a partir del contraste de las siguientes hipótesis:

$$\boldsymbol{H_0}: \boldsymbol{F_0(x) = F(x)} \quad \forall \boldsymbol{x}$$

$$\boldsymbol{H_1}: \boldsymbol{F_0(x) \neq F(x)\ para\ alguna\ x}$$

Donde F es la función de distribución de la muestra y F_0 es la distribución de una variable aleatoria HiperGeométrica(N=100; n=50; m=75).

Para la primera categoría se han considerado los valores menores o iguales a 31, a partir de ahí las categorías serán de la forma $\{i\}, 32 \leq i \leq 43$, y la última categoría estaría conformada por los valores mayores a 43, por lo cual se tienen un total de 14 categorias. Después para calcular la frecuencia esperada E_i, se obtienen las probabilidades $P_1 = P(X \leq 31)$, luego $P_i = P(X = i)$ $32 \leq i \leq 43$ y por ultimo $P_{14} = P(X > 43)$, donde X es una variable Hipergeométrica(100;50;75).

El resultado de llevar a cabo los cálculos necesarios para realizar la prueba sobre los valores simulados, se hallan en la Tabla 1.17, además se agrega la gráfica que compara las frecuencias esperadas contra las empíricas.

Gráfica 1.10: Frecuencias Esperadas VS Empíricas

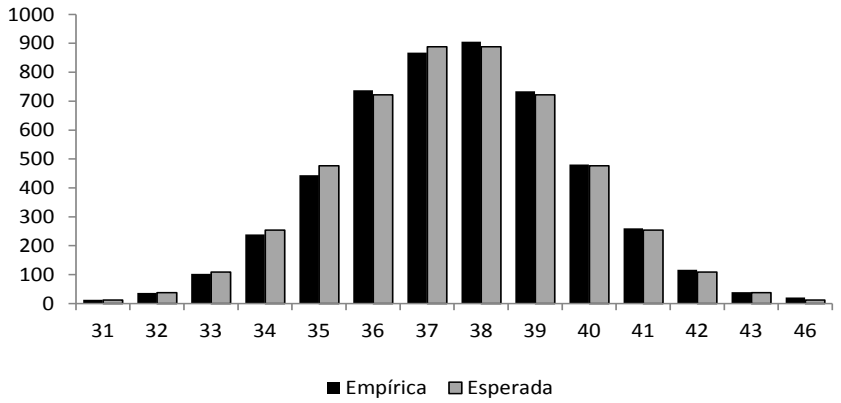


Tabla 1.17: Cálculos necesarios para realizar la prueba Ji-Cuadrada

i	$Categoría\ C_i$	$Frecuencia\ Empírica\ O_i$	$Frecuencia\ Esperada\ E_i$	$\frac{(O_i - E_i)^2}{E_i}$
1	{25,...,31}	13	12.379	0.031
2	{32}	37	37.178	0.001
3	{33}	103	108.999	0.330
4	{34}	239	254.330	0.924
5	{35}	444	476.687	2.241
6	{36}	738	722.253	0.343
7	{37}	868	888.176	0.458
8	{38}	906	888.176	0.358
9	{39}	734	722.253	0.191
10	{40}	481	476.687	0.039
11	{41}	260	254.330	0.126
12	{42}	117	108.999	0.587
13	{43}	39	37.178	0.089
14	{44,45,46, ...,50}	21	12.379	6.005
	$\chi^2 = 11.725$		$p - value = 0.55$	

Si se considera un nivel de significancia $\alpha = 0.05$, la regla de decisión del prueba es no rechazar la hipótesis nula H_0, de esta manera se puede concluir que no existen diferencias significativas entre la distribución de los datos y una distribución Hipergeométrica(100;50;75).

Poisson

El nombre de esta distribución es dedicado en honor al físico y matemático francés Simeon Denis Poisson, quien en 1837 publicó su derivación, a partir de tomar el límite de una distribución Binomial(n, p) tendiendo el parámetro n a infinito mientras el valor de la media de la distribución np permanece igual a una constante λ.

Otra derivación posterior surgió en el terreno de los procesos estocásticos, ya que describe una serie de eventos que ocurren de forma aleatoria y con independencia en el tiempo, los cuales se van contando en un intervalo de tamaño t. Una característica de este proceso es que el tiempo entre ocurrencias se distribuye de manera exponencial y para dos intervalos disjuntos, cumplen ser independientes las ocurrencias de los eventos, a este proceso de conteo se le llama proceso Poisson pues la variable aleatoria del conteo se distribuye de esa forma.

Las aplicaciones que se le han dado a esta variable son bastas desde las dos perspectivas de su generación, ha sido utilizada como aproximación de la distribución Binomial, en problemas de muestreo de poblaciones extensas donde la ocurrencia de un cierto evento atípico, pues el número de experimentos es grande mientras que la probabilidad es pequeña.

Dentro de su otra perspectiva de conteo, es de utilidad dentro de la teoría de colas, pues considera como eventos las llegadas de individuos a una cola, donde el tiempo de espera se supone se distribuye como una variable aleatoria Exponencial, con esto, se puede estimar el número de elementos en espera, además, se ha empleado dentro de la ecología, en estudios que muestran que al dividir un área específica en partes el número de individuos de una especie animal en cada parte, se distribuye como una variable Poisson. Existen muchas otras áreas de aplicación en las que se puede mencionar a la geografía, geología, ciencias sociales y la industria.

Características principales

Para caracterizar esta variable se tiene en primer lugar a su función de masa de probabilidad dada por

$$\boldsymbol{P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}} \quad \boldsymbol{x \in \{0, 1, 2, 3, \ldots\}} \; ; \; \boldsymbol{\lambda > 0}$$

Luego está su función de probabilidad acumulada

$$\boldsymbol{P(X \leq x) = \sum_{j=0}^{x} P(X = x) = \sum_{j=0}^{x} e^{-\lambda}\frac{\lambda^j}{j!}}$$

Su cálculo puede tornarse tedioso por el hecho de realizar las operaciones para obtener las probabilidades individuales, para aminorar esto es mejor apoyarse en la siguiente relación recursiva

$$\boldsymbol{P(X=x+1)=e^{-\lambda}\frac{\lambda^{x+1}}{(x+1)!}=\frac{\lambda}{x+1}e^{-\lambda}\frac{\lambda^{x}}{x!}=\frac{\lambda}{x+1}P(X=x)}$$

Una de las propiedades importantes de la distribución Poisson es que su media y su varianza son iguales a su parámetro λ, es decir

$$\boldsymbol{E(X)=\lambda=VAR(X)}$$

Simulación de valores

Para simular valores de la distribución Poisson es necesario recurrir a la siguiente propiedad

$$\boldsymbol{Sean\ U_1,U_2,\dots,U_n variables\ aleatorias\ Uniformes\ en\ (0,1)\ y\ N\ el\ mínimo\ entero\ \ tal\ que}$$

$$\boldsymbol{\prod_{j=1}^{N+1}U_j\le e^{-\lambda};entonces\ N\ es\ Poisson(\lambda)}$$

Así que los pasos para generar valores de la distribución Poisson son

1. $\boldsymbol{Sea\ i=0\ ,a=e^{-\lambda},b=1}$
2. $\boldsymbol{Generar\ U_{i+1}\sim U(0,1),hacer\ b=bU_{i+1},si\ b<}\ a\ regresar\ X=i\ en\ otro\ caso\ ir\ al\ paso\ 3$
3. $\boldsymbol{Hacer\ i=i+1\ e\ ir\ al\ paso\ 2}$

Este algoritmo se justifica por la propiedad del proceso Poisson y su relación con la distribución Exponencial que se pueden hallar con más detalle en Law (2007). Esta propiedad enuncia que si se consideran una secuencia IID de variables aleatorias no negativas $Y_1, Y_2, ...,$ y sea $X=\max\{i:\sum_{j=1}^{i}Y_j\le 1\}$. Entonces la distribución de las variables Y_i es exponencial de parámetro $1/\lambda$, si y sólo si, X ~ Poisson (λ).

El algoritmo define a la secuencia de variables aleatorias como $Y_i=-\ln(U_i)/\lambda$, donde U_i se distribuye Uiforme en el intervalo (0,1), que como se verá mas adelante es una aplicación del método de la transformación inversa a una variable Exponencial($1/\lambda$), ahora la suma que se encuentra en la definición de X se tranforma a una forma equivalente como

$$\sum_{j=1}^{i}Y_j=\sum_{j=1}^{i}-\ln(U_i)/\lambda\le 1$$

$$\sum_{j=1}^{i}\ln(U_i)\le-\lambda$$

$$\ln\left(\prod_{j=1}^{i}U_i\right)\le-\lambda$$

$$\boldsymbol{\prod_{j=1}^{i}U_j\le e^{-\lambda}}$$

Los principales problemas de este método son de índole numérica pues cuando λ toma valores altos puede tomar demasiado tiempo al programa poder realizar las operaciones del algoritmo, de hecho para el paso 2, el número de ejecuciones promedio es de $\lambda + 1$.

Por otra parte, el soporte de memoria para representar un número también es de gran importancia, puesto que normalmente para una representación de valores que salen del rango (como e^{λ} con λ de valor muy grande) ocurre un error, pues su inverso sería considerado como 0, es decir $e^{-\lambda}$ pudiera ocasionar que el paso 2. y 3. entraran en un ciclo infinito. Para el programa de Microsoft Excel en particular, la máxima potencia que puede soportar es $2^{2^{10}-1}$que igualándose a e^{λ} se obtiene que el máximo valor de λ es 709.

A manera de ejemplo sencillo considérese $\lambda = 50$ de la cual se han generado 5000 valores de la distribución Poisson, según las propiedades vistas anteriormente los valores de la media muestral y la varianza muestral deben ser iguales al parámetro λ. La Tabla 1.18 contiene parcialmente los valores generados, además de los valores $\bar{X}$ y S^2 estimadores de la media y de la varianza respectivamente.

Tabla 1.18: muestra parcial de valores generados de una Variable Poisson(50)

i	$Poisson(50)$	i	$Poisson(50)$
1	45	6	50
2	57	7	49
3	43	8	41
4	40	9	45
5	49	10	41
$\bar{X} = 50.322$		$S^2 = 52.501$	

El primer indicador de la eficacia del algoritmo se puede ver en que tanto el valor de la varianza como de la media son parecidos y además se acercan al valor del parámetros λ, aunque es un tanto subjetivo. Para mostrar de diferente manera que los datos tienen un comportamiento Poisson, se construirá una tabla de frecuencias.

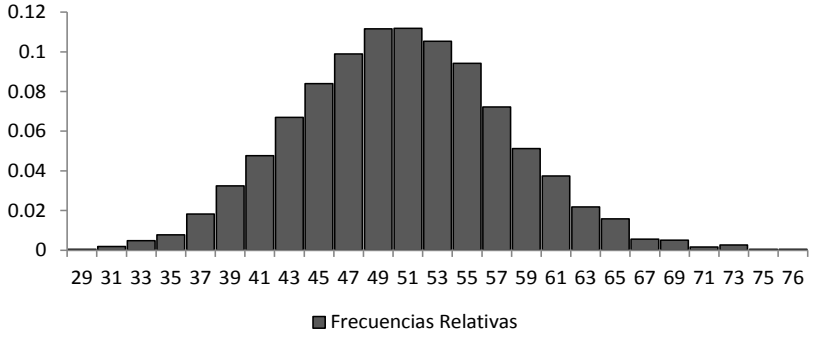
Lo primero en realizarse es una división del soporte de la muestra obtenida, dado que se obtuvo un mínimo de 28 y un máximo de 76, se tiene un rango de 49 posibles valores que se tomaron en la simulación. Tales valores se agrupan cada dos valores para formar un total de 25 grupos de la forma $\{28,29\}, \{30,31\}, \dots, \{74,75\}, \{76\}$.

Después de haber formado los grupos, se procede a contar el número de observaciones que pertenecen a cada grupo, por medio de la función de Excel $FRECUENCIA()$ que recibe como parámetros los datos, y un vector de grupos, este último vector debe contener el mayor valor de cada grupo. Definir el vector se puede hacer a partir de haber obtenido el mínimo de la muestra, con la fórmula $minimo - 1 + 2k$ para $1 \leq k \leq 25$(nótese que para grupos de cardinalidad mas grande el valor a cambiar es el 2).

La siguiente tabla muestra el resultado del proceso anterior, junto con las frecuencias relativas correspondientes, las cuales son plasmadas en la Gráfica 1.11.

Tabla 1.19: Frecuencias de los valores simulados a partir de una distribución Poisson(50)					
Categoría	*Frecuencia*	*Frecuencia Relativa*	*Categoría*	*Frecuencia*	*Frecuencia Relativa*
{28,29}	2	0.0004	{54,55}	471	0.0942
{30,31}	9	0.0018	{56,57}	361	0.0722
{32,33}	24	0.0048	{58,59}	256	0.0512
{34,35}	39	0.0078	{60,61}	187	0.0374
{36,37}	91	0.0182	{62,63}	109	0.0218
{38,39}	162	0.0324	{64,65}	79	0.0158
{40,41}	238	0.0476	{66,67}	28	0.0056
{42,43}	335	0.067	{68,69}	25	0.005
{44,45}	420	0.084	{70,71}	8	0.0016
{46,47}	495	0.099	{72,73}	13	0.0026
{48,49}	558	0.1116	{74,75}	2	0.0004
{50,51}	559	0.1118	{76}	2	0.0004
{52,53}	527	0.1054			
		$Mínimo = 28$		$Máximo = 76$	

Gráfica 1.11: Frecuencias relativas de los valores simulados de una Poisson(50)



Para dar certidumbre que la distribución de los valores generados, es el de una variable Poisson, se llevará a cabo la prueba Ji-cuadrada, la cual contrasta las siguientes hipótesis.

$$\boldsymbol{H_0}: \boldsymbol{F_0(x) = F(x)} \quad \forall \boldsymbol{x}$$

$$\boldsymbol{H_1}: \boldsymbol{F_0(x) \neq F(x)\ para\ alguna\ x}$$

Donde F es la función de distribución de los datos y F_0 es una función de distribución de una variable aleatoria Poisson(50). Para esta muestra se dividido el dominio en $k = 12$ categorias, por medio de hallar el valor más grande de la categoría usando la fórmula, $minimo - 1 + 4i$ para $1 \leq i \leq 11$, para introducirlo como el argumento que define los valores que definen los grupos en la función $FRECUENCIA()$ de Excel, pero esto no agrupa todos los datos pues deja 5 valores hasta llegar al $maximo$ los cuales se pueden considerar como una última categoría.

Luego se hallan las probabilidades de la distribución Poisson, de que un elemento pertenezca a cada categoría denotada por P_i, como $P_1 = P(X \leq minimo - 1 + 4 = 31)$,

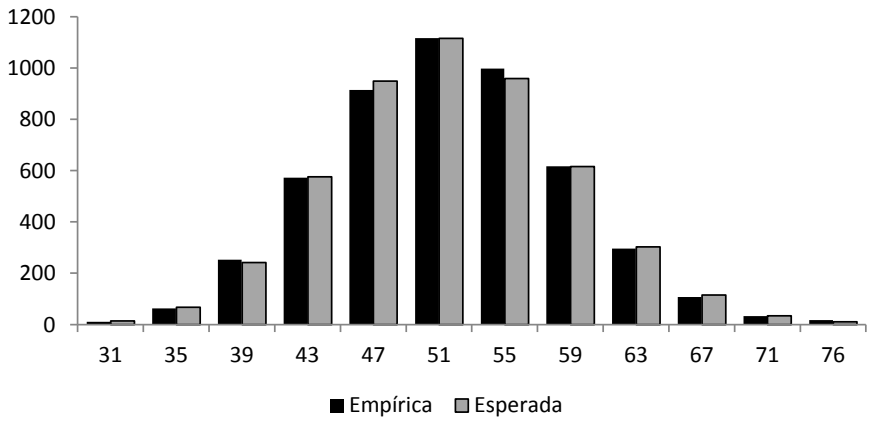
mientras que para la última categoría se tiene $P_{12} = P(X > 71)$. A partir de estos cálculos lo único que falta es multiplicar por el tamaño de la muestra para obtener las frecuencias esperadas, E_i.

Los cálculos intermedios para llevar a cabo la prueba como el estadístico de prueba y el $p - value$, se hallan plasmados en la Tabla 1.20, además en la Gráfica 1.12 se encuentra una comparación entre los valores de las frecuencias esperadas y las empíricas.

Tabla 1.20: Cálculos intermedios de la prueba Ji-Cuadrada

i	$Categoria\ C_i$	$Frecuencia\ O_i$	$Frecuencia\ Esperada\ E_i$	$\frac{(O_i - E_i)^2}{E_i}$
1	{28,29,30,31}	11	13.43	0.44
2	{32,33,34,35}	63	67.64	0.32
3	{36,37,38,39}	253	241.78	0.52
4	{40,41,42,43}	573	576.13	0.02
5	{44,45,46,47}	915	949.36	1.24
6	{48,49,50,51}	1117	1,115.35	0.002
7	{52,53,54,55}	998	958.67	1.61
8	{56,57,58,59}	617	616.32	1.61
9	{60,61,62,63}	296	302.11	0.001
10	{64,65,66,67}	107	114.82	0.12
11	{68,69,70,71}	33	34.34	0.53
12	{72,73,74,75,76}	17	10.05	0.05
	$p - value = 0.84$		$\chi^2 = 6.47$	

Gráfica 1.12: Frecuencias Esperadas VS Empíricas



Como se observa, si se elige un nivel de significancia $\alpha = 0.05$, entonces $p - value > \alpha$, por lo tanto no se debe rechazar la hipótesis nula H_0, concluyendo que no existen diferencias significativas entre la función de distribución de los datos y la función de distribución de una variable Poisson(50). Por otra parte analizando la gráfica permite ver que las frecuencias esperadas y las empíricas son de valores similares.

Variables aleatorias continuas

En el estudio de ciertos fenómenos, una o más variable que se registren son medidas en términos de una magnitud, la cual se pueda relacionar con un dominio continuo, como la temperatura. Debido a lo anterior es necesario modelar este tipo de variables por medio de las variables aleatorias catalogadas como de tipo continuo. La característica de dominio continuo hace referencia a que teóricamente se pueda obtener cualquier valor de un intervalo, aunque de manera estricta en la práctica esto no ocurre, pues por las limitantes de la tecnología, o por reglas establecidas de registro para los valores observados, pues son asentados con una representación decimal de longitud finita.

Aun así, su aproximación al comportamiento real es suficiente para modelar variables como el tiempo, el espacio, la temperatura, las cantidades monetarias o tasa de interés por medio de variables aleatorias con un soporte, ya sea en un fragmento o en su totalidad, continuo. Algunas variables de este tipo no poseen una función de distribución con expresión analítica como la distribución Normal y las variables conocidas como parte de la familia de variables Gamma transformadas, por lo que la única forma de usar el método de inversión es por medio de métodos numéricos lo que podría tener un costo computacional alto. Una alternativa es usar las diversas propiedades y relaciones que tienen con otras variables fáciles de simular.

Uniforme en (a,b)

Introducción

La aleatoriedad pura de un suceso que puede tomar cualquier valor dentro de un intervalo (a, b) está descrita por la distribución Uniforme. Esta distribución también es llamada rectangular por la forma de su densidad, pues es un valor constante, dentro de sus principales aplicaciones están: estudiar la ley de probabilidad del redondeo de decimales, la cual enuncia que si se registran observaciones hasta un cierto decimal de la forma $P.d_1d_2 \dots d_k$ con P, d_i, k enteros, entonces existen un error por haber truncado el verdadero valor, éste se supone se distribuye Uniforme, donde el valor registrado es considerado como el punto medio del intervalo, al cual se le suma y resta el número con $P = 0$ y sus decimales con las primeras k posiciones en 0 y en la posición $k + 1$ el valor cinco.

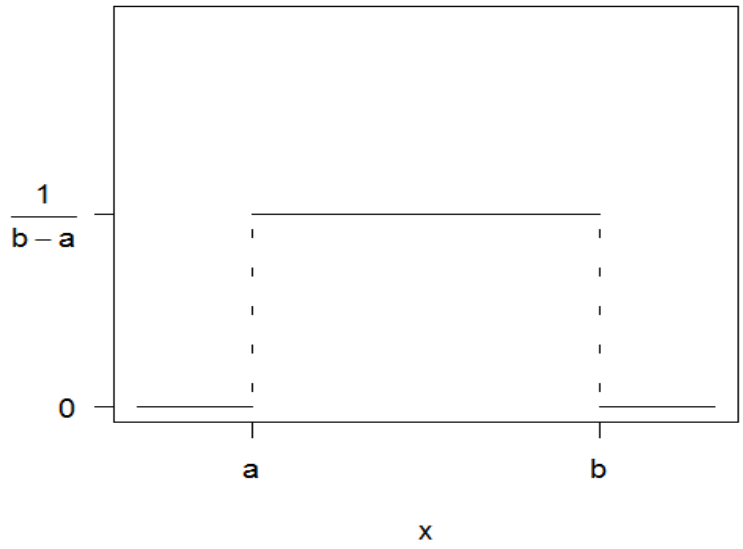
Características principales

La función de densidad de la distribución uniforme se define como:

$$\boldsymbol{f(x)} = \begin{cases} \dfrac{\mathbf{1}}{\boldsymbol{b-a}}, & a \leq x \leq b \\ \mathbf{0}, & en\ otro\ caso \end{cases}$$

La cual se puede observar en la Gráfica 1.13.

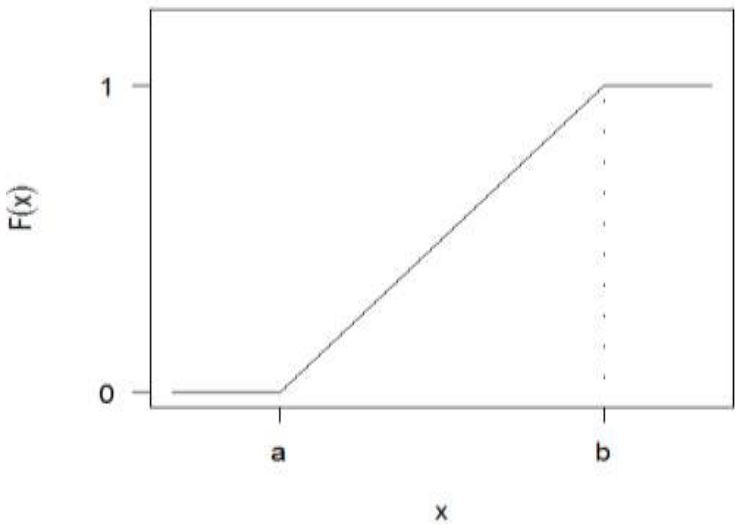
Grafica 1.13: función de densidad de una uniforme(a,b)



Su función de distribución es una línea recta que conecta los puntos $(a, 0)$ y a $(b, 1)$ justo como se ve en la Gráfica 1.14 y su expresión analítica es

$$\boldsymbol{F(x) = \int_{-\infty}^{x} f(x)dx = \frac{x-a}{b-a} \quad a \leq x \leq b}$$

Grafica 1.14: Función de distribución de una uniforme(a,b)



Su función de riesgo (o *hazard function*) está definida como

$$\boldsymbol{h(x) = \frac{f(x)}{1 - F(x)} = \frac{1}{b-x}}$$

$\boldsymbol{h(x)}$ es una función creciente de x, la interpretación de esta función de manera intuitiva, es una probabilidad instantánea de ocurrencia, aunque en realidad es una función que indica la fuerza o intensidad en la que los elementos ocurren en un punto

del dominio x, como por ejemplo la intensidad con la que puede ocurrir la muerte de un individuo a cierta edad.

Los valores de su esperanza $E(X)$ y su varianza $Var(X)$ son

$$\boldsymbol{E(X) = \frac{b+a}{2}} \qquad \boldsymbol{Var(X) = \frac{(b-a)^2}{12}}$$

Simulación de valores

Para poder simular valores de la distribución uniforme se puede usar el método de la distribución inversa, por lo cual se debe hallar la inversa igualando a un valor uniforme en $(0,1)$ U, a la función de distribución y despejar el valor de x.

$$\boldsymbol{F(x) = U = \frac{x-a}{b-a}}$$

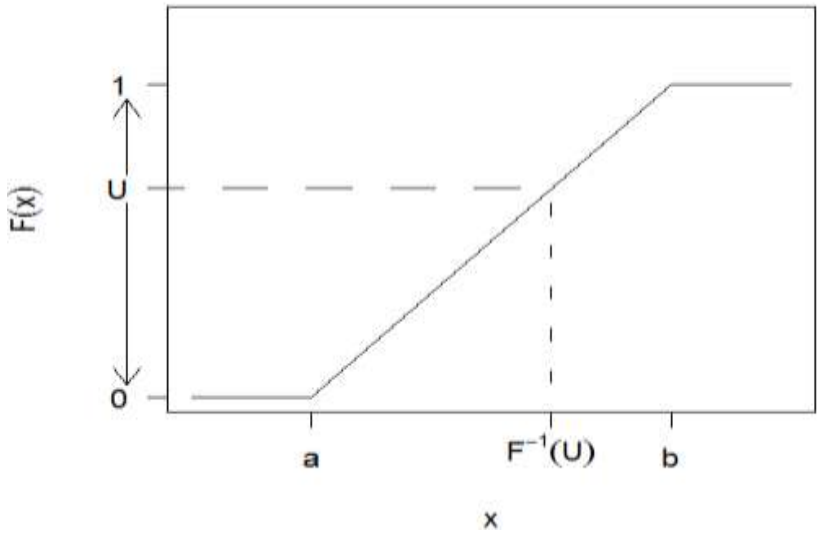
$$\boldsymbol{x = U(b-a) + a}$$

Por lo tanto, el pseudocódigo para generar un valor uniforme entre a y b es

1. $\boldsymbol{Generar\ U \sim U(0,1)}$
2. $\boldsymbol{Regresar\ X = a + (b-a)U}$

De manera gráfica lo que se está realizando es, dado un valor U en el dominio de $F(x)$, se refleja sobre la función, de tal manera que se halla el valor x proveniente de la distribución, como lo muestra la Gráfica 1.15.

Grafica 1.15: inversion de la función de distribución de una uniforme(a,b)



Ejemplo: Usando el generador de números aleatorios de Excel, se generó una muestra de tamaño 5000, a los cuales se les aplicó la inversa de la función de distribución antes deducida, con $a = -5$,$b = 5$, para obtener una variable aleatoria con distribución Uniforme sobre el intervalo $(-5,5)$. Una muestra parcial de los primeros 20 datos simulados se muestra en la Tabla 1.21, junto con las estimaciones de la media y la varianza.

Tabla 1.21: Muestra parcial de 5000 simulaciones de una uniforme en (-5,5)					
i	$U_i \sim U(0,1)$	$x_i = U_i(b-a)+a$	i	$U_i \sim U(0,1)$	$x_i = U_i(b-a)+a$
1	0.36	-1.401	11	0.732	2.33
2	0.451	-0.487	12	0.417	-0.83
3	0.28	-2.203	13	0.568	0.68
4	0.254	-2.459	14	0.634	1.34
5	0.145	-3.548	15	0.508	0.08
6	0.149	-3.509	16	0.213	-2.87
7	0.625	1.250	17	0.032	-4.68
8	0.679	1.787	18	0.435	-0.65
9	0.6	1.001	19	0.596	0.96
10	0.901	4.006	20	0.128	-3.72
		$\bar{X} = -0.006$		$S^2 = 8.36$	

Comparando los valores estimados de la media y varianza teóricas, que son

$$E(X) = \frac{5-5}{2} = 0 \qquad Var(X) = \frac{(5-(-5))^2}{12} = 8.\overline{33}$$

Se nota que las estimaciones son muy cercanas a los valores de la varianza y media calculados en base a sus parámetros, por otra parte, un análisis de los valores simulados puede dar una mejor perspectiva; primero, por medio de agrupar los datos en subintervalos, contando el número de datos (frecuencia) en cada subintervalo, esto otorga una visión completa de los datos, en un formato entendible; la Tabla 1.22 muestra las frecuencias de los valores anteriormente simulados. Se incluye también la perspectiva de la frecuencia relativa, que es la frecuencia entre el número de simulaciones totales, y puede ser interpretada como una probabilidad empírica del subintervalo.

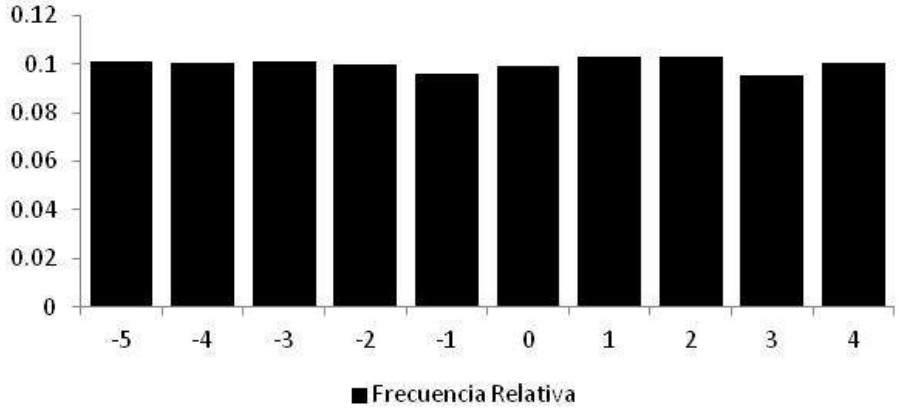
Tabla 1.22: frecuencias de la muestra de la simulación de una uniforme en (-5,5)		
Rango	*Frecuencia*	*Frecuencia relativa*
-5 ≤ x < -4	505	0.101
-4 ≤ x < -3	502	0.1004
-3 ≤ x < -2	505	0.101
-2 ≤ x < -1	501	0.1002
-1 ≤ x < 0	481	0.0962
0 ≤ x < 1	496	0.0992
1 ≤ x < 2	514	0.1028
2 ≤ x < 3	515	0.103
3 ≤ x < 4	478	0.0956
4 ≤ x	503	0.1006

Los subintervalos dividen de tal manera al intervalo$(-5,5)$ que cada uno tiene la misma longitud. La probabilidad de que la variable aleatoria uniforme en (a,b) quede

entre dos valores α y β , $a \leq \alpha < \beta \leq b$, es $P(\alpha \leq X \leq \beta) = P(X \leq \beta) - P(X \leq \alpha) = F(b) - F(b) = \frac{\beta - \alpha}{b-a}$; quiere decir que la probabilidad depende solo de la longitud del intervalo, por lo tanto la probabilidad de los subintervalos anteriores es la misma igual a $\frac{1}{10} = 0.1$.

Como cada subdivisión tiene probabilidad de 0.1, entonces se espera que el 10% del total de las observaciones quede agrupado en cada una de las subdivisiones, esto se puede apreciar de mejor manera en la Gráfica 1.16 pues permite comparar las frecuencias relativas de la tabla.

Gráfica 1.16: Frecuencias relativas de los valores simulados



Para dar validez estadística a la simulación, se procederá a realizar la prueba de bondad de ajuste de Kolmogorov-Smirnov[4] (K-S), pues compara la función de distribución empírica F_n, como estimación de la función de distribución de los datos(F), contra F_0 que es la función de distribución teórica de la variable aleatoria Uniforme en $(-5,5)$, en la cual se basa la hipótesis nula H_0, por lo cual el contraste de hipótesis es:

$$\boldsymbol{H_0: la\ funcion\ de\ distribicion\ de\ los\ datos}$$

$$\boldsymbol{se\ distribuyen\ como\ una\ uniforme(-5,5)}$$

$$\boldsymbol{(F(X) = F_0(X)\ \forall x).}$$

$$\boldsymbol{H_1: los\ datos\ provienen\ de\ una\ distribucion}$$

$$\boldsymbol{distinta\ a\ la\ uniforme(-5,5)}$$

$$\boldsymbol{(F(X) \neq F_o(X)\ para\ algún\ x).}$$

La comparación se realiza por medio de medir las distancias que separan las funciones de distribución, tomando como estadístico de prueba el supremo de estas distancias, es decir se calcula

[4] Detallada en el Apéndice

$$D_n = \sup\{|F_n(x) - F_0(x)|\}$$

Para poder realizar el cálculo anterior, lo primero que se debe hacer es ordenar los valores de la muestra simulada de manera ascendente, obteniendo los estadísticos de orden $x_{(i)}$, para después evaluarlos sobre F, obteniendo las siguientes cantidades

$$D_n^+ = \max_{1\le i\le n}\left\{\frac{i}{n} - F_0(x_{(i)})\right\}$$

$$D_n^- = \max_{1\le i\le n}\left\{F_0(x_{(i)}) - \frac{i-1}{n}\right\}$$

Luego se toma como estadístico de prueba

$$D_n = \max\{D_n^+, D_n^-\}$$

En la Tabla 1.23 se muestran los primeros y últimos datos ordenados, junto con la evaluación de la función de distribución uniforme, además de las diferencias antes mencionadas.

Tabla 1.23: Muestra parcial de la prueba de bondad de ajuste K-S para los datos simulados de una variable Uniforme(-5,5)

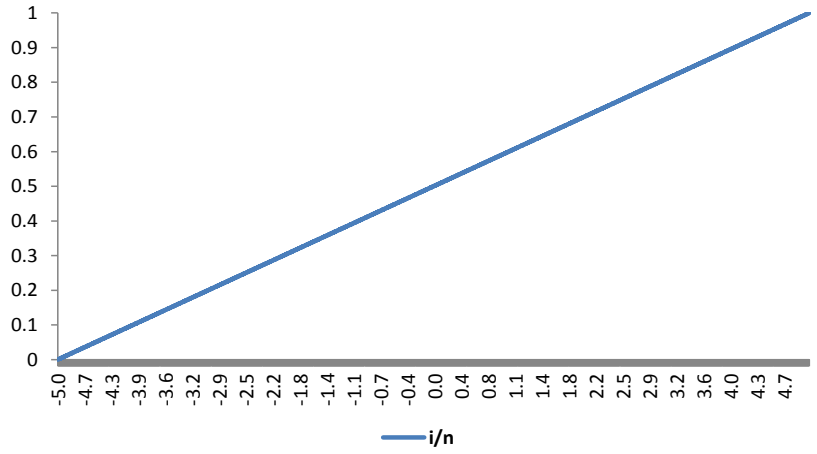
i	$x_{(i)}$	$F(x_{(i)})$	$\frac{i}{n} - F(x_{(i)})$	$F(x_{(i)}) - \frac{i-1}{n}$
1	-4.99964	0.00004	0.00016	0.00004
2	-4.99960	0.00004	0.00036	-0.00016
3	-4.99898	0.0001	0.0005	-0.0003
4	-4.99633	0.00037	0.00043	-0.00023
5	-4.99612	0.00039	0.00061	-0.00041
6	-4.99554	0.00045	0.00075	-0.00055
7	-4.99397	0.0006	0.0008	-0.0006
8	-4.99302	0.0007	0.0009	-0.0007
9	-4.98871	0.00113	0.00067	-0.00047
10	-4.98839	0.00116	0.00084	-0.00064
...	...	...	...	...
4990	4.97519	0.99752	0.00048	-0.00028
4991	4.97687	0.99769	0.00051	-0.00031
4992	4.97925	0.99793	0.00047	-0.00027
4993	4.97949	0.99795	0.00065	-0.00045
4994	4.97999	0.998	0.0008	-0.0006
4995	4.98056	0.99806	0.00094	-0.00074
4996	4.98074	0.99807	0.00113	-0.00093
4997	4.98793	0.99879	0.00061	-0.00041
4998	4.99032	0.99903	0.00057	-0.00037
4999	4.99713	0.99971	0.00009	0.00011
5000	4.99914	0.99991	0.00009	0.00011
	$d_{n,0.95} = 0.01923$		$D_n = 0.00615$	
	$D_n^- = 0.00615$		$D_n^+ = 0.00602$	

El valor del estadístico D_n debe ser comparado con el valor crítico de la prueba K-S, que en este caso es, al 5% de significancia, el mostrado en la tabla como $d_{n,0.95}$, obtenido por medio de tablas que muestran los valores críticos de la prueba K-S, para distintos niveles de significancia.

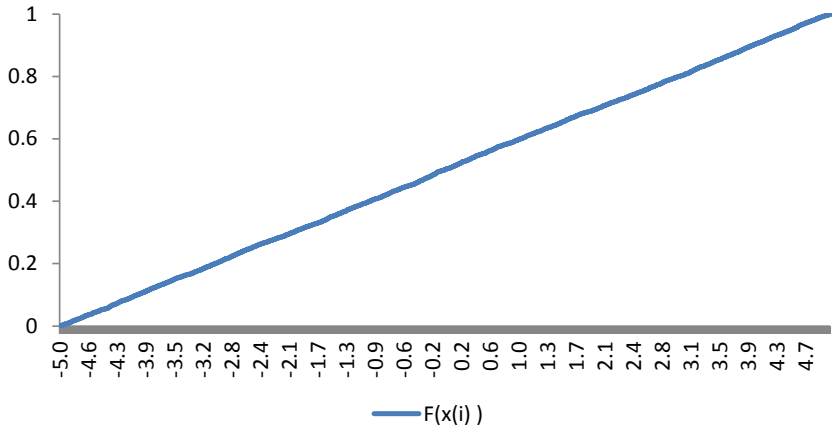
La regla para rechazar la hipótesis nula es si $D_n \geq d_{n,0.95}$, puesto que sucede el caso contrario, se concluye por medio de esta prueba que no existen razones estadísticas que indiquen que los datos simulados no se ajustan a una distribución uniforme $(-5,5)$.

Una forma gráfica de observar porqué no se rechazó la hipótesis nula es graficando $F(x_{(i)})$ y aparte $\frac{i}{n}$, sobre todos los valores simulados, logrando constatar su gran parecido comparando la gráfica 1-.17y la gráfica 1.18**,** aunque se podrían graficar yuxtapuestas, mostrando el mismo resultado; por otra parte se puede omitir en este caso los valores de $\frac{i-1}{n}$, debido a que el valor de n es lo suficientemente grande para que la diferencia sea imperceptible.

Gráfica 1.17**:** función de distribución empirica de los datos simulados



Grafica 1,.18: funcion de distribucion de una uniforme(-5,5) valuada sobre los estadisticos de orden de la muestra simulada



Exponencial

Introducción

La variable Exponencial también llamada Exponencial Negativa, está relacionada con el tiempo que tardan en ocurrir eventos independientes de un proceso de conteo conocido como Poisson de tasa λ constante, u homogéneo.

Aunque los fenómenos estudiados por esta distribución estén relacionados con la ocurrencia de eventos, en muchas ocasiones el ajuste de esta distribución puede no ser el óptimo, así que distribuciones alternativas son usadas, como la Weibull, la cual más adelante se verá, puede ser vista como una transformación de la distribución Exponencial.

Dentro de la teoría de colas, esta distribución es usada para describir el tiempo entre llegadas de elementos a un sistema, sin embargo, para sistemas más realistas, en los cuales la intensidad de llegada obedece un cierto patrón que aumenta a ciertas horas, el supuesto de independencia no se cumple.

En las telecomunicaciones y estudios relacionados con un ámbito tecnológico el supuesto de independencia es más realista, pues sin importar el tiempo de uso de algún aparato, puede tener la misma probabilidad en el tiempo de la ocurrencia de una falla que otro aparato nuevo.

Otra de las características por la cual esta variable es de gran importancia dentro de las aplicaciones es por su llamada pérdida de memoria, que consiste en la independencia de la variable, dado que se está empezando desde un cierto tiempo, es decir

$$\boldsymbol{P}(\boldsymbol{X} > t + s|\boldsymbol{X} > t) = \boldsymbol{P}(\boldsymbol{X} > s) \quad \boldsymbol{para\ s, t} > 0$$

Usando esta propiedad sobre la esperanza condicional se obtiene

$$\boldsymbol{E}(\boldsymbol{X}|\boldsymbol{X} > t) = \boldsymbol{E}(\boldsymbol{X})$$

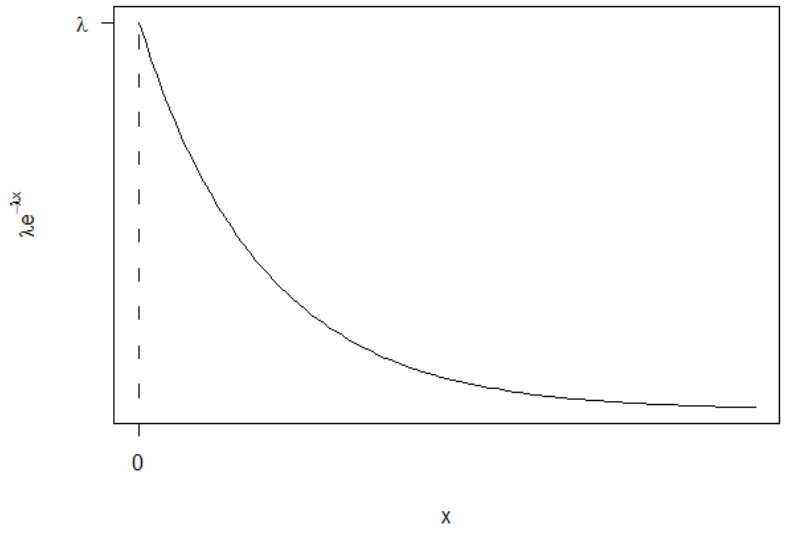
Lo anterior tiene importancia, cuando se trabaja con datos censurados al medir el tiempo de funcionamiento o vida, de ciertos elementos estudiados, pues a pesar de la pérdida de información, el tiempo esperado de vida es el mismo al que cuando comenzaron a ser observados.

Características principales

La primera característica de importancia es su función de densidad$f(x)$, mostrada en la Gráfica 1.19

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{cases} \boldsymbol{0}, & x < 0 \\ \boldsymbol{\lambda e^{-\lambda x}}, & x \geq 0 \end{cases}$$

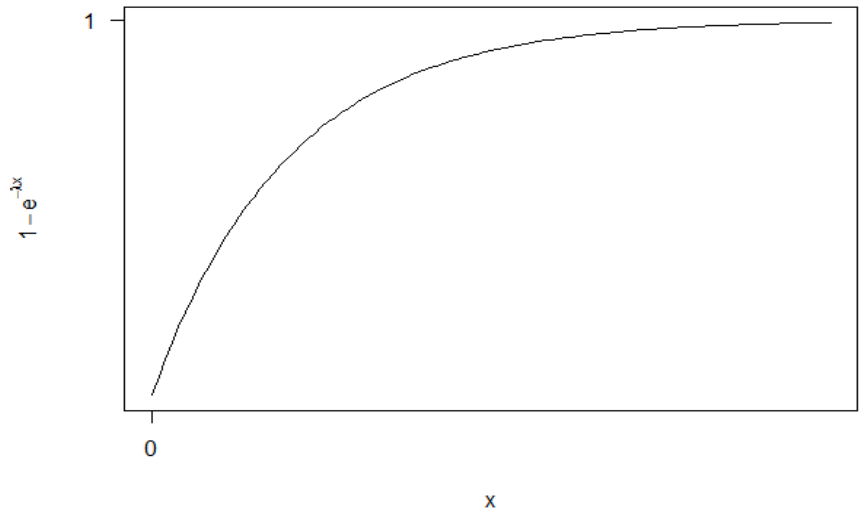
Gráfica 1.19: función de densidad de una distribución exponencial(λ)



La Gráfica 1.20 muestra la forma de la función de distribución la cual es

$$\boldsymbol{F(x) = \int_{-\infty}^{x} f(x) = 1 - e^{-\lambda x}}$$

Gráfica 1.20:Función de distribución de una exponencial (λ)



A continuación se resumen algunos valores de interés de la variable como su esperanza y sus momentos centrales

$$\boldsymbol{E(X) = \int_{-\infty}^{\infty} xf(x) = \frac{1}{\lambda}}$$

$$\boldsymbol{Var(X) = E(X^2) - E^2(X) = \frac{1}{\lambda^2}}$$

$$\boldsymbol{Asimetría} = \boldsymbol{E}\left(\left(\boldsymbol{X} - \boldsymbol{E}(\boldsymbol{X})\right)^{\boldsymbol{3}}\right) = \frac{\boldsymbol{2}}{\boldsymbol{\lambda^3}}$$

$$\boldsymbol{Curtosis} = \boldsymbol{E}\left(\left(\boldsymbol{X} - \boldsymbol{E}(\boldsymbol{X})\right)^{\boldsymbol{4}}\right) = \frac{\boldsymbol{9}}{\boldsymbol{\lambda^4}}$$

Una de las características que auxilian el estudio de la supervivencia es la fuerza de mortalidad, que en realidad es otra forma de llamar a la función de riesgo

$$\boldsymbol{h(x)} = \frac{\boldsymbol{\lambda e^{-\lambda x}}}{\boldsymbol{e^{-\lambda x}}} = \boldsymbol{\lambda}$$

Como se puede notar para esta distribución en particular, la intensidad de que ocurra el evento estudiado (muerte o falla) es constante en el tiempo, es decir no depende de su edad (momento de observación del sujeto), que es congruente con la mencionada pérdida de memoria.

Simulación de valores

Para la generación de esta variable aleatoria, la opción más simple es usar el método de inversión resolviendo

$$\boldsymbol{F(x)} = \boldsymbol{1 - e^{-\lambda x} = U}$$

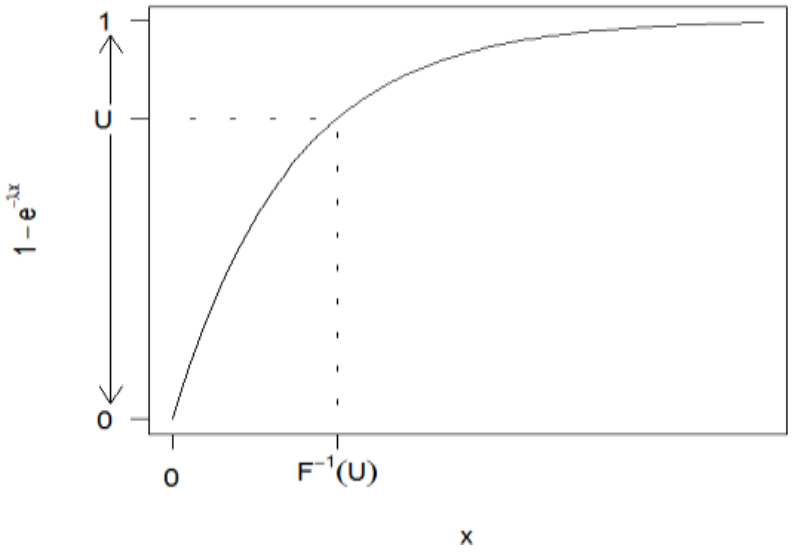
$$\boldsymbol{e^{-\lambda x} = 1 - U = U}$$

$$\boldsymbol{-\lambda x = \ln(U)}$$

$$\boldsymbol{x = -\frac{\ln(U)}{\lambda}}$$

La forma gráfica de entender cómo se aplica este método, en particular sobre la función de distribución de la distribución exponencial se encuentra en la Gráfica 1.21

Gráfica 1.21: inversion de la función de distribución de una exponencial (λ)



Con la sustitución de U por $1-U$, que como ya se ha mecionado poseen la misma distribución $U(0,1)$, el costo computacional solamente recae en la evaluación del logaritmo natural, por lo que la velocidad de generación es aceptable para su uso en diversas aplicaciones.

Para mostrar la funcionalidad de este algoritmo se generó una muestra de tamaño $n = 5000$ de datos aleatorios en (0,1) por medio de la función ALEATORIO() de Excel, para después aplicarle la transformación logarítmica con el parámetro $\lambda = 0.2$ por medio de la fórmula mostrada en la Tabla 1.24 donde se exhibe una muestra parcial de los datos y el resultado de su transformación logarítmica.

Tabla 1.24: muestra de 20 elementos de la simulación de una exponencial (λ=0.2)

i	$U_i \sim U(0,1)$	$x_i = -\ln(U_i)/\lambda$	i	$U_i \sim U(0,1)$	$x_i = -\ln(U_i)/\lambda$
1	0.73	1.59	11	0.38	4.84
2	0.48	3.72	12	0.87	0.69
3	0.64	2.2	13	0.02	20.41
4	0.84	0.88	14	0.75	1.46
5	0.96	0.19	15	0.59	2.66
6	0.94	0.31	16	0.35	5.22
7	0.2	8.03	17	0.27	6.53
8	0.13	10.38	18	0.02	18.86
9	0.96	0.21	19	0.32	5.77
10	0.44	4.15	20	0.02	19.47
		$\bar{X} = 5.0249$		$S^2 = 26.0620$	

Con el parámetro asignado, la media y varianza teóricas son

$$\boldsymbol{E(X) = \frac{1}{0.2} = 5}$$

$$\boldsymbol{Var(X) = \frac{1}{(0.2)^2} = 25}$$

Valores que son cercanos a los estimadores, que emplean el total de los datos simulados, de la media ($\bar{X}$) y de la varianza (S^2), mostrados en la tabla.

Para comprender de manera integral el comportamiento de los datos, deben ser presentados en un formato entendible, por medio de una tabla, que representa un agrupamiento de los datos en subintervalos y un conteo sobre cuantos datos pertenecen a cada subintervalo.

Para este caso las subdivisiones fueron hechas de tal manera que cada subdivisión tuviera la misma longitud, para esto fue necesario obtener el mínimo y máximo de la muestra para tomar $\max_i(x_i) - \min_i(x_i)$ como el rango en que se toman valores, luego se divide el rango entre el número k de subintervalos deseados; posteriormente los limites superiores de los subintervalos se determinan como $\min_i(x_i) + i\frac{\max_i(x_i)-\min_i(x_i)}{k}$; para valores de i entre 1 y k.

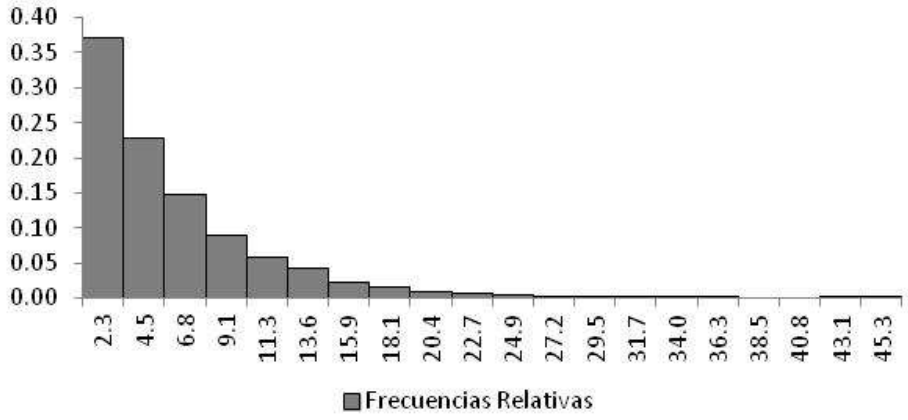
La tabla siguiente muestra los subintervalos obtenidos al elegir $k = 20$, junto con las frecuencias calculadas después de agrupar los datos, las frecuencias relativas, que se calculan como las frecuencias entre el número de simulaciones (5000) y se incluye tanto el máximo como el mínimo de los valores generados.

Tabla 1.25: Frecuencias de la muestra de una exponencial (λ=0.02)					
Rango	***Frecuencia***	***Frecuencia relativa***	***Rango***	***Frecuencia***	***Frecuencia relativa***
(0 , 2.27]	1851	0.3702	(20.40,24.94]	24	0.0048
(2.27,4.53]	1136	0.2272	(24.94,27.20]	15	0.003
(4.53,6.80]	734	0.1468	(27.20,29.47]	10	0.002
(6.80,9.07]	443	0.0886	(29.47,31.74]	6	0.0012
(9.07,11.34]	295	0.059	(31.74,34.00]	6	0.0012
(11.34,13.60]	208	0.0416	(34.00,36.27]	2	0.0004
(13.60,15.87]	111	0.0222	(36.27,38.54]	0	0
(15.87,18.14]	82	0.0164	(38.54,40.80]	0	0
(18.14,20.40]	46	0.0092	(40.80,43.07]	1	0.0002
(20.40,22.67]	29	0.0058	(43.07,45.34]	1	0.0002
		Mínimo = **0.0005**		***Máximo =*** **45.34**	

Si bien la Tabla 1.25 señala que a medida que se consideran valores más cerca del origen, el número de observaciones es cada vez mayor debido a que la función de densidad posee más área cerca del origen que lejos de él; es decir la probabilidad de un intervalo de una cierta longitud es mayor cerca de 0 y menor sobre valores cada vez mayores.

La mejor manera de entender lo anterior es de manera visual, por eso la Gráfica 1.22 muestra las frecuencias relativas, que en forma, es idéntica a la de las frecuencias absolutas pues son tan solo un reescalamiento; se puede además, se nota un gran parecido a la función de densidad de una exponencial.

Gráfica 1.22: Frecuencias relativas de los valores simulados de una Exp(λ=0.02)



Usando la prueba de bondad de ajuste de Kolmogorov-Smirnov[5], se da certeza de que los datos simulados de distribuyen como una variable exponencial, donde se compara la función de distribución F_0 de una exponencial ($\lambda = 0.2$) con la funcion de distribución empírica F_n de los datos por medio del contraste de las siguientes hipótesis

$$\boldsymbol{H_0}: \boldsymbol{Los\ datos\ se\ distribuyen\ exponecialmente}(\boldsymbol{\lambda = 0.2})\ (\boldsymbol{F(X) = F_0(X)}\ \forall \boldsymbol{x})$$

$$\boldsymbol{H_1}: \boldsymbol{Los\ datos\ provienen\ de\ una\ distribucion\ distinta\ a\ la\ exponencial}(\boldsymbol{\lambda = 0.2})\ (\boldsymbol{F(X) \neq F_0(X)\ \ para\ algun\ x})$$

Para realizar la comparación se emplea la aproximación de F, F_n. Primero se ordena la muestra simulada para obtener los estadísticos de orden $x_{(i)}$, para posteriormente calcular el estadístico de la prueba de bondad de ajuste como sigue

$$\boldsymbol{D_n = \max\{D_n^-, D_n^+\} = \max_{1 \leq i \leq n}\left\{F_0(x_{(i)}) - \frac{i-1}{n},\quad \frac{i}{n} - F_0(x_{(i)})\right\}}$$

Una muestra de los cálculos parciales, el estadístico D_n, además del valor crítico $d_{n,0.95}$[6], de la prueba al 5% de significancia (95% de confianza), obtenido de las tablas para la prueba K_S, se exhiben en la Tabla 1.26.

Tabla 1.26: Cálculos auxiliares y estadísticos para la prueba de bondad de ajuste K-S

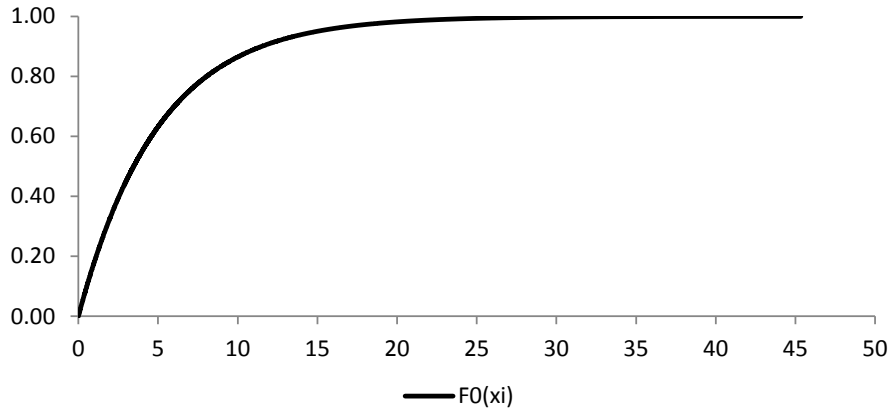
i	$x_{(i)}$	$F_0(x_{(i)}) = 1 - e^{-\lambda x_{(i)}}$	$\frac{i}{n} - F_0(x_{(i)})$	$F_0(x_{(i)}) - \frac{i-1}{n}$
1	0	0.000099	0.000101	0.000099
2	0.003	0.000564	-0.000164	0.000364
3	0.003	0.000591	0.000009	0.000191
4	0.004	0.00089	-0.00009	0.00029
5	0.005	0.000902	0.000098	0.000102
6	0.005	0.001087	0.000113	0.000087
7	0.008	0.001502	-0.000102	0.000302
8	0.009	0.001807	-0.000207	0.000407
9	0.009	0.001812	-0.000012	0.000212
10	0.009	0.001837	0.000163	0.000037
...	...	...	...	...
4990	31.692	0.998233	-0.000233	0.000433
4991	32.06	0.998358	-0.000158	0.000358
4992	32.456	0.998483	-0.000083	0.000283
4993	32.684	0.998551	0.000049	0.000151
4994	33.026	0.998647	0.000153	0.000047
4995	33.043	0.998651	0.000349	-0.000149
4996	33.552	0.998782	0.000418	-0.000218
4997	36.052	0.999261	0.000139	0.000061
4998	36.215	0.999285	0.000315	-0.000115
4999	40.829	0.999716	0.000084	0.000116
5000	45.339	0.999885	0.000115	0.000085
	$d_{n,0.95} = 0.01923$		$D_n = 0.011$	
	$D_n^- = 0.007$		$D_n^+ = 0.011$	

[5] Véase la parte I del apéndice para hallar una descripción más a detalle de la prueba.
[6] Véase la parte I del apéndice para hallar una descripción de cómo obtener este valor crítico.

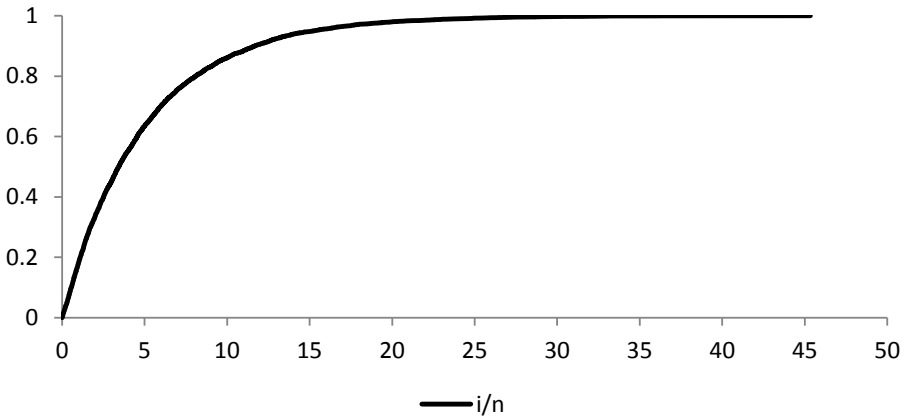
La regla de decisión para rechazar al hipótesis nula de ésta prueba, es si el valor crítico $d_{n,0.95} < D_n$, aunque como sucede lo contrario en este caso, se puede concluir que no existe justificación estadística que afirme que los datos simulados no provienen de una distribución exponencial ($\lambda = 0.2$).

Visualizando en las siguientes gráficas los puntos $\left(x_{(i)}, F_0\left(x_{(i)}\right)\right)$ y por otro lado los puntos $\left(x_{(i)}, \frac{i}{n}\right)$ se puede observar el porqué no se rechazó la hipótesis nula dándole validez a la conclusión de la prueba sobre la distribución de los valores simulados.

Gráfica 1.23: valores de $F_0(x_i)$



Gráfica 1.24: valores de i/n



Gamma

Introducción

La primera mención de esta variable aleatoria Gamma, fue cuando se intentó hallar la distribución del inverso de la varianza de una serie de datos distribuidos normal$(0, \sigma^2)$, es decir la ley de probabilidad de $h = \sigma^{-2}$, encontrando que $h \sim Gamma$.

Una de las propiedades importantes de la variable aleatoria Gamma es su relación con la distribución exponencial de parámetro λ, pues la suma de m variables aleatorias independientes distribuidas como una variable exponencial se distribuye como una variable aleatoria Gamma de parámetros m y λ, esta relación permite aplicarla en temas relacionados con el m-ésimo evento que ocurren en el tiempo de observación de un fenómeno. Diversos estudios en otras ramas de la ciencia han ajustado variables en distintos contextos a una variable Gamma, como por ejemplo en los procesos de precipitación meteorológica, estudios sobre el ingreso personal de una población, o la abundancia de una población cuando se desarrolla sobre un ambiente equilibrado.

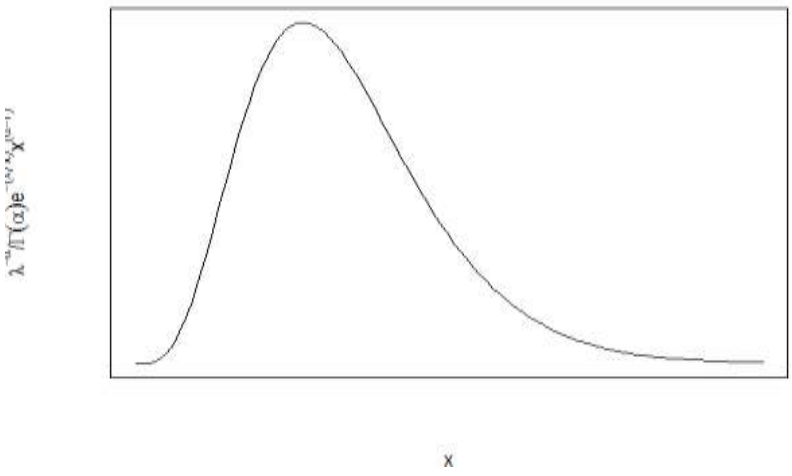
En teoría del riesgo, para un modelo agregado de siniestros cuya severidad es no negativa, un método de aproximación es ajustar los datos a una distribución Gamma debido a ser no negativa y al parecido de su función de densidad a la densidad de una variable Normal cuando el parámetro de forma α es lo suficientemente grande.

Características principales

La variable aleatoria Gamma tiene por función de densidad a la siguiente $f(x)$. La Gráfica 1.25 es la visualización de esta función

$$\boldsymbol{f(x) = \begin{cases} \mathbf{0}, & x < 0 \\ \dfrac{\boldsymbol{\lambda^{-\alpha} x^{\alpha-1} e^{-x/\lambda}}}{\boldsymbol{\Gamma(\alpha)}}, & x \geq 0 \end{cases}}$$

Grafica 1.30:densidad de una Gamma(α=5,λ=2)

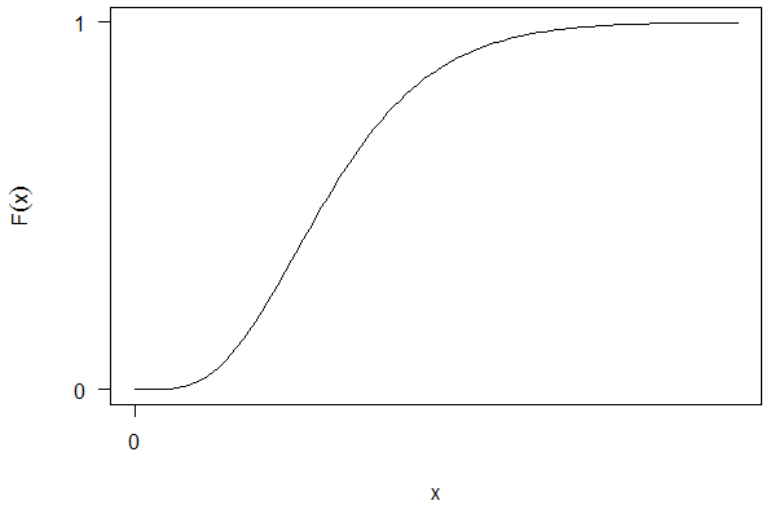


Los parámetros que definen a la distribución gamma son: α conocido como el parámetro de forma debido a que los cambios en su valor impactan directamente la forma de la función de densidad, y el parámetro λ llamado “tasa” aunque también se le

atribuye ser un parámetro de escala, ya que está relacionado directamente con la dispersión de la variable, y por lo tanto un cambio en su valor tendrá un cambio en la escala de la función de densidad, sin afectar su forma. La función de distribución de esta variable se define como la integral de la densidad desde $-\infty$ hasta un punto x, sin embargo no tiene una solución analítica, aunque se puede evaluar numéricamente por algún método de integración numérica como, el método de Simpson o el método de Gauss.

Uno de los programas que permiten la evaluación numérica de la función de distribución es R, por medio de su función *pgamma(),* la cual requiere especificaciones sobre los parámetros α y λ; usando esto, la forma de la función de distribución de una gamma se puede observar en la Gráfica 1.26.

Grafica 1.26:Función de distribución de una Gamma(α,λ)



Otras propiedades de interés, son su esperanza y sus momentos centrales de orden mayor e igual a 2, los cuales son

$$\boldsymbol{E(X) = \int_{-\infty}^{\infty} xf(x) = \lambda\alpha}$$

$$\boldsymbol{Var(X) = E\left(X^2\right) - E^2(X) = \alpha\lambda^2}$$

$$\boldsymbol{Asimetría = E\left(\left(X - E(X)\right)^3\right) = 2\alpha\lambda^3}$$

$$\boldsymbol{Curtosis = E\left(\left(X - E(X)\right)^4\right) = \lambda^4 3\alpha(1 + 2\alpha)}$$

Para cuestiones poblacionales una forma de analizar a qué tipo de poblaciones podría ajustarse esta distribución, es por medio de su función de riesgo (fuerza de mortalidad) $h(x)$, pero al no tener una forma analítica de la su función de distribución, se pueden aplicar alternativas numéricas por medio de *software,* para obtener h evaluada en una serie de puntos; aunque se pueden deducir ciertas propiedades de la función de riesgo por medio de técnicas de cálculo diferencial.

La primera de las propiedades es el límite de $h(x)$ cuando el valor de x crece indefinidamente es decir

$$\lim_{x\to\infty} \boldsymbol{h(x)} = \lim_{x\to\infty} \frac{\boldsymbol{f(x)}}{\boldsymbol{1 - F(x)}}$$

Para la densidad de una Gamma, una de las condiciones necesarias para que sea función de densidad es que $f(x) \to 0$ cuando $x \to \infty$, por la necesidad de la finitud de la integral de f, por otra parte $F(x) \to 1$ cuando $x \to \infty$, entonces se tiene una indeterminación de la forma $0/0$; para resolver tal problema se puede aplicar la regla de l'Hôpital

$$\lim_{x\to c} \boldsymbol{d(x)/g(x)} = \lim_{x\to c} \boldsymbol{d'(x)/\, g'(x)}$$

En este caso se considera como d a la densidad y a g como $1 - F(x)$ entonces estableciendo a $\alpha \neq 1$ y aplicando las reglas de derivación se obtiene

$$\boldsymbol{f'(x) = \frac{\lambda^{-\alpha}}{\Gamma(\alpha)}\left((\alpha - 1)x^{\alpha-2}e^{-\frac{x}{\lambda}} - \frac{x^{\alpha-1}e^{-\frac{x}{\lambda}}}{\lambda}\right)}$$

Y por otro lado usando el teorema fundamental del cálculo

$$\boldsymbol{\left(1 - F(x)\right)' = -f(x) = \frac{-\lambda^{-\alpha}}{\Gamma(\alpha)} x^{\alpha-1}e^{-x/\lambda}}$$

Con lo anterior el límite es

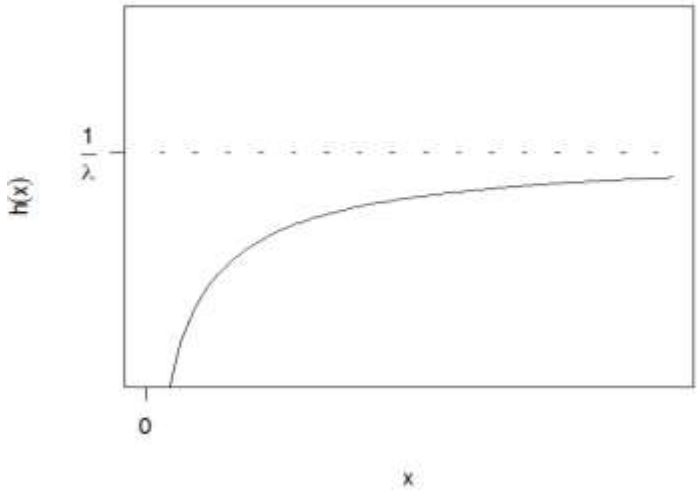
$$\lim_{x\to\infty} \boldsymbol{h(x)} = \lim_{x\to\infty} \frac{\boldsymbol{f(x)}}{\boldsymbol{1 - F(x)}} = \lim_{x\to\infty} \frac{\boldsymbol{f'(x)}}{\boldsymbol{-f(x)}} = \lim_{x\to\infty} \frac{\frac{\lambda^{-\alpha}}{\Gamma(\alpha)}\left((\alpha - 1)x^{\alpha-2}e^{-\frac{x}{\lambda}} - \frac{x^{\alpha-1}e^{-\frac{x}{\lambda}}}{\lambda}\right)}{\frac{-\lambda^{-\alpha}}{\Gamma(\alpha)} x^{\alpha-1}e^{-x/\lambda}}$$

$$= \lim_{x\to\infty} \frac{\boldsymbol{1}}{\boldsymbol{\lambda}} - \frac{\boldsymbol{\alpha - 1}}{\boldsymbol{x}} = \frac{\boldsymbol{1}}{\boldsymbol{\lambda}}$$

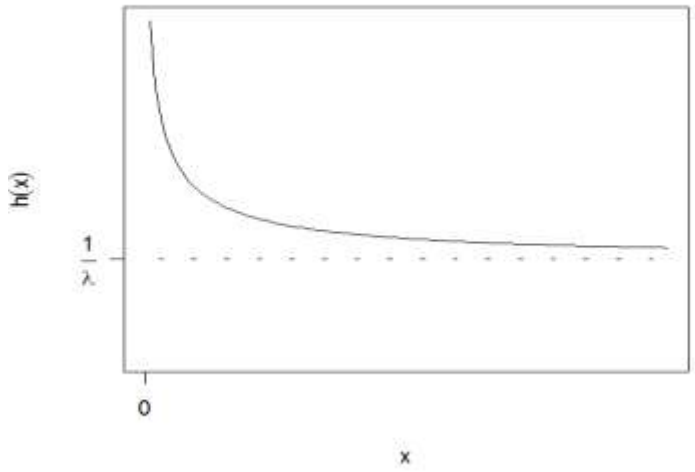
Concluyéndose que la función de riesgo de una Gamma con $\alpha \neq 1$, tiene un comportamiento asintótico a $1/\lambda$ conforme los valores de x aumentan. La función de riesgo también tiene diferentes formas de converger a la constante antes obtenida, por un lado si $\alpha < 1$ la función es decreciente y por el otro si $\alpha > 1$ la función es creciente.

La Gráfica 1.27 muestra a h, como función creciente de x para una α arbitraria con $\alpha > 1$, lo cual implica una mayor intensidad de la mortalidad conforme aumenta la edad, que es un supuesto más realista; mientras que la Gráfica 1.28 muestra a h, como función decreciente de x para una α menor que 1, donde ambas gráficas parten de densidades Gamma con el mismo parámetro λ.

Grafica 1.27: Función de riesgo de una Gamma($\alpha > 1, \lambda$)



Grafica 1.28: Función de riesgo de una Gamma($\alpha < 1, \lambda$)



Simulación de valores

Para la simulación de valores provenientes de una distribución Gamma, primero se debe retomar el hecho que dada una variable $X \sim Gamma(\alpha, 1)$ se puede obtener, para cualquier $\lambda > 0$ una $Gamma(\alpha, \lambda)$ por medio de la transformación $X' = \lambda X$; entonces, para su simulación basta con un mencionar algoritmos para generar una variable $Gamma(\alpha, 1)$.

Los valores que puede tomar el parámetro α se puede dividir en $0 < \alpha < 1$ y $\alpha > 0$; para el caso $\alpha = 1$ se obtendría una exponencial. Cuando α es menor a 1, se utiliza el método de aceptación y rechazo, con un algoritmo denominado GS, propuesto por Ahrens y Dieter en 1974, y se basa en las siguientes observaciones de una distribución $Gamma(\alpha, 1)$

$$\boldsymbol{f(x) = \frac{x^{\alpha-1}}{\Gamma(\alpha)} e^{-x}}$$

$$\boldsymbol{si\ 0 < x \leq 1\ \ \rightarrow\ \ e^{-x} < 1\ \rightarrow\ \frac{x^{\alpha-1}}{\Gamma(\alpha)}e^{-x} < \frac{x^{\alpha-1}}{\Gamma(\alpha)}}$$

$$\boldsymbol{si\ x > 1\ \ \rightarrow\ \ x^{\alpha-1} < 1\ \ \rightarrow\ \frac{x^{\alpha-1}}{\Gamma(\alpha)}e^{-x} < \frac{e^{-x}}{\Gamma(\alpha)}}$$

Luego se define una función r que sea la envolvente de f de la siguiente manera

$$\boldsymbol{r(x)} = \begin{cases} \boldsymbol{0}, & x \leq 0 \\ \boldsymbol{\frac{x^{\alpha-1}}{\Gamma(\alpha)}}, & \boldsymbol{0 < x \leq 1} \\ \boldsymbol{\frac{e^{-x}}{\Gamma(\alpha)}}, & x \geq 0 \end{cases}$$

Para que sea densidad debe de dividirse por una constante obtenida de la siguiente forma

$$\boldsymbol{c = \int_0^{\infty} r(x)dx = \int_0^1 \frac{x^{\alpha-1}}{\Gamma(\alpha)}dx + \int_1^{\infty} \frac{e^{-x}}{\Gamma(\alpha)}dx}$$

$$\boldsymbol{= \frac{1}{\alpha\Gamma(\alpha)} + \frac{1}{e\Gamma(\alpha)} = \frac{1}{\alpha\Gamma(\alpha)}\left(\frac{e+\alpha}{e}\right)}$$

Haciendo $b = \frac{e+\alpha}{e}$, la densidad denotada $t(x) = r(x)/c$

$$\boldsymbol{t(x)} = \begin{cases} \boldsymbol{0}, & x \leq 0 \\ \boldsymbol{\frac{\alpha x^{\alpha-1}}{b}}, & \boldsymbol{0 < x \leq 1} \\ \boldsymbol{\frac{\alpha e^{-x}}{b}}, & x \geq 0 \end{cases}$$

Después para generar valores de t se le puede aplicar el método de inversión a

$$\boldsymbol{T(x)} = \begin{cases} \boldsymbol{\frac{x^{\alpha}}{b}}, & \boldsymbol{0 < x \leq 1} \\ \boldsymbol{1 - \frac{\alpha e^{-x}}{b}}, & x > 1 \end{cases}$$

Esta función puede ser vista como una mezcla, pues para $x = 1\ \ T(x) = 1/b$, entonces se puede generar la inversa de la primera parte con probabilidad $1/b$ o de la segunda parte con probabilidad $1 - 1/b$, es decir para u uniforme en (0,1)

$$\boldsymbol{T^{-1}(u)} = \begin{cases} \boldsymbol{(bu)^{1/\alpha}}, & \boldsymbol{u \leq 1/b} \\ \boldsymbol{-\ln\left(\frac{b(1-u)}{\alpha}\right)}, & en\ otro\ caso \end{cases}$$

Realizando el cociente de densidades para un valor Y, que por la construcción de r ya no es necesario hallar una constante, queda seccionado en

$$\frac{\boldsymbol{f(Y)}}{\boldsymbol{r(Y)}} = \begin{cases} \boldsymbol{e^{-Y}}, & \boldsymbol{0 \leq Y \leq 1} \\ \boldsymbol{Y^{\alpha-1}}, & \boldsymbol{Y > 1} \end{cases}$$

El algoritmo completo es

1. $\boldsymbol{Generar\ U_1 \sim U, sea\ P = bU_1\ si\ P > 1}\ ir\ al\ paso\ 3\ en\ otro\ caso\ ir\ al\ paso\ 2$
2. $\boldsymbol{Sea\ Y = P^{1/\alpha}, generar\ U_2 \sim U(0,1), si\ U_2 \leq e^{-Y}},$ $\boldsymbol{regresar\ X = Y\ en\ otro\ caso\ volver\ a\ 1}$
3. $\boldsymbol{Sea\ Y = -\ln\left(\frac{b-P}{\alpha}\right)\ generar\ U_2 \sim U(0,1), si\ U_2 \leq Y^{\alpha-1},\ regresar\ X = Y\quad en\ otro\ caso\ \ ir\ al\ paso\ 1}$

Ahora se trata el caso en que $\alpha > 1$, usando de igual manera el método de aceptación y rechazo, se requiere una función que acote la densidad de la distribución Gamma, por lo que se define $L = \sqrt{2\alpha - 1}, \mu = \alpha^L, c = 4\alpha^\alpha e^{-\alpha}/(L\Gamma(\alpha))$

$$\boldsymbol{t(x)} = \begin{cases} \dfrac{\boldsymbol{L\mu x^{L-1}}}{\boldsymbol{(\mu + x^L)^2}}, & \boldsymbol{0 < x} \\ \boldsymbol{0}, & \boldsymbol{en\ otro\ caso} \end{cases}$$

Hallando la función de distribución e igualando a U

$$\boldsymbol{T(x) = \int_0^x \frac{L\mu y^{L-1}}{(\mu + y^L)^2} dy = \mu\left(\frac{1}{\mu} - \frac{1}{\mu + x^L}\right) = \frac{x^L}{\mu + x^L} = U}$$

$$\boldsymbol{x = \left(\frac{\mu U}{1 - U}\right)^{1/L}}$$

Para hacer más eficiente el algoritmo, un paso intermedio de aceptación o rechazo, se le puede adicionar de tal manera que se evite evaluar el logaritmo, suponiendo que las siguientes constantes son previamente definidas: $\boldsymbol{a = \frac{1}{\sqrt{2\alpha - 1}}, b = \alpha - \ln(4), q = \alpha + \sqrt{2\alpha - 1}, \theta = 4.5, d = 1 + \ln(\theta)}$ así que el algoritmo es

1. $\boldsymbol{Generar\ U_1, U_2\ independientes\ distribuidas\ U(0,1)}$
2. $\boldsymbol{Sea\ V = aln\left(\frac{U_1}{1-U_1}\right), Y = \alpha e^V,\ Z = U_1^2 U_2, y\ W = b + qV - Y}$
3. $\boldsymbol{Si\ W + d - \theta Z \geq 0\ regresar\ X = Y, en\ otro\ caso\ ir\ al\ paso\ 4}$
4. $\boldsymbol{W \geq \ln(Z), regresar\ X = Y\ en\ otro\ caso\ ir\ al\ paso\ 1}$

Debido a la definición de constantes, el método parece diferir un poco de la forma intuitiva de los algoritmos anteriores que se han mencionado, pero el paso 1., 2. y 4. son la aplicación del método de aceptación y rechazo.

Ejemplo: Para mostrar el parecido de la variable Gamma con la distribución Normal cuando el parámetro de forma toma valores relativamente grandes, se generó una muestra de tamaño $n = 5000$ con el algoritmo, para simular una variable Gamma($\alpha = 50, \lambda = 2$). La simulación se realizó por medio de un código realizado por procesos macros en Excel®, por ser los pasos para la simulación más complejos, en consecuencia, la generación de números aleatorios y el proceso son ocultos al

usuario, por lo cual la Tabla 1.27, muestra los primeros 20 valores generados con el algoritmo anterior, junto con las estimaciones de la media y la varianza, $\bar{X}$, S^2 respectivamente.

Tabla 1.27: primeros 20 valores de 5000 simulados de una Gamma (α=50 ; λ=2)

i	$Gamma(50;2)$	i	$Gamma(50;2)$
1	90.68	11	96.18
2	126.00	12	108.27
3	96.37	13	108.36
4	95.41	14	98.14
5	119.82	15	102.15
6	71.19	16	110.25
7	85.18	17	99.19
8	88.22	18	102.57
9	96.41	19	99.07
10	101.87	20	86.93
$\bar{X} = 100.079$		$S^2 = 197.472$	

Una evaluación intuitiva de la simulación se puede realizar a través de observar los valores de los momentos centrales de la distribución presentada los cuales son

$$E(X) = 50 * 2 = 100$$

$$Var(X) = 50 * 2^2 = 200$$

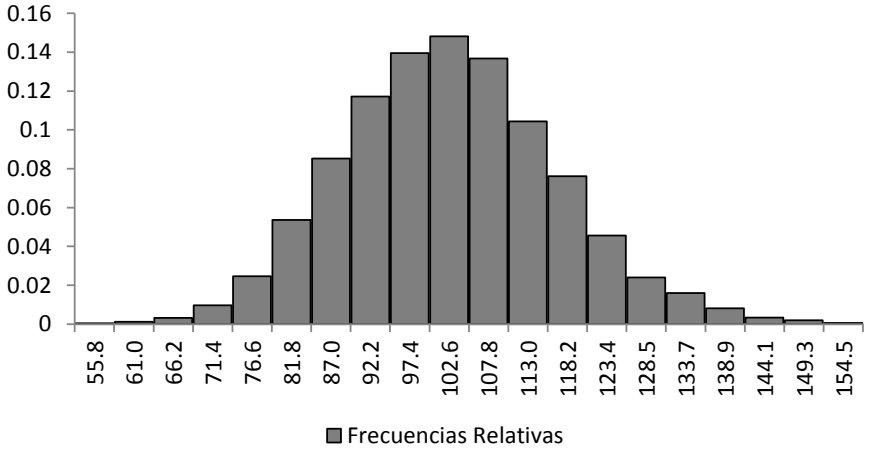
Donde se nota que los cálculos basados en los parámetros son cercanos a las estimaciones correspondientes. A continuación se realizará una tabla de frecuencias para observar la distribución de los valores generados.

Para obtener las frecuencias primero se divide el dominio de la variable, en subintervalos de igual longitud. Los límites superiores de los subintervalos serán de la forma: $mínimo + i\frac{maximo-mínimo}{k} 1 \leq i \leq k$, donde k es el numero de subintervalos a considerary además debe tomarse en cuenta que, debido al dominio de esta variable el ultimo límite superior se puede considerar como ∞.

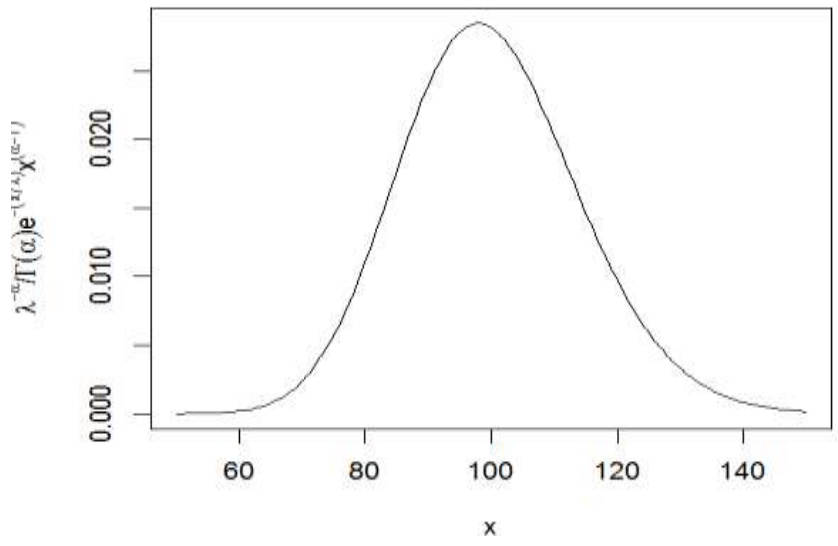
La Tabla 1.28 contiene las frecuencias obtenidas al elegir $k = 20$, junto con las frecuencias relativas, además de contener el máximo y el mínimo de los valores simulados. La Gráfica 1.29 muestra las frecuencias relativas en una gráfica de barras, la cual se puede comparar contra la función de densidad de una Gamma$(50; 2)$, mostrada en la Gráfica 1.30.

Tabla 1.28: Frecuencias de los valores simulados, agrupados en 20 subintervalos.					
Rango	*Frecuencia*	*Frecuencia Relativa*	*Rango*	*Frecuencia*	*Frecuencia Relativa*
(0,55.80]	2	0.0004	(102.56,107.76]	684	0.1368
(55.80,60.99]	6	0.0012	(107.76,112.96]	522	0.1044
(60.99,66.19]	16	0.0032	(112.96,118.15]	381	0.0762
(66.19,71.38]	48	0.0096	(118.15,123.35]	228	0.0456
(71.38,76.58]	123	0.0246	(123.35,128.55]	120	0.024
(76.58,81.78]	268	0.0536	(128.55,133.74]	80	0.016
(81.78,86.97]	426	0.0852	(133.74,138.94]	41	0.0082
(86.97,92.17]	586	0.1172	(138.94,144.14]	17	0.0034
(92.17,97.37]	698	0.1396	(144.14,149.33]	10	0.002
(97.37,102.56]	741	0.1482	(149.33,∞)	3	0.0006
		Mínimo = **50.60**		*Máximo* = **154.53**	

Gráfica 1.29: Frecuencias relativas de los valores simulados de una Gamma(50;2)



Grafica 1.30:Distribución de una Gamma(α=50,λ=2)



Para mostrar que la distribución los datos generados provienen de la distribución deseada, se aplicará la prueba de bondad de ajuste de Kolmogorov-Smirnov la cual contrasta las siguientes hipótesis

$$\boldsymbol{H_0}: \boldsymbol{F(X) = F_0(X)}$$

$$\boldsymbol{H_1}: \boldsymbol{F(X) \neq F_0(X)}$$

Donde F_0 es la función de distribución de una Gamma$(50; 2)$. Posteriormente se obtienen los estadísticos de orden $x_{(i)}$, para después calcular el siguiente estadístico de prueba

$$\boldsymbol{D_n = \max\{D_n^+, D_n^-\} = \max_{1 \leq i \leq n}\left\{\frac{i}{n} - F_0(x_{(i)}), F_0(x_{(i)}) - \frac{i-1}{n}\right\}}$$

La Tabla 1.29 muestra los cálculos individuales para realizar la prueba de K-S, el valor de los estadísticos obtenidos con funciones de Excel, además del valor crítico para comparar con el valor del estadístico y poder emitir una conclusión sobre la prueba, obtenido de tablas especificas para la prueba de K-S, donde para una prueba al 95% de confianza se obtiene como $1.36/\sqrt{n}$.

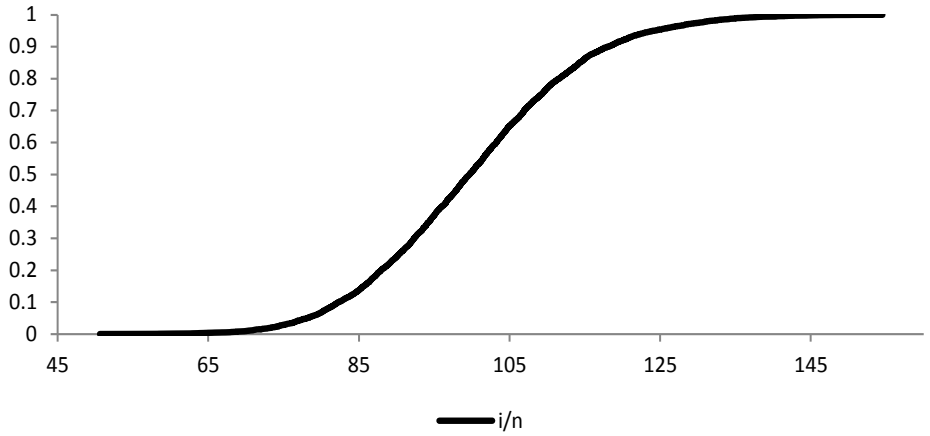
Tabla 1.29: Cálculos para realizar la prueba K-S sobre los valores simulados de una Gamma(α=50 ; λ=2)

i	$Gamma(50;2)$	$\frac{i}{n}$	$F_0(x_{(i)})$	$\frac{i}{n} - F_0(x_{(i)})$	$F_0(x_{(i)}) - \frac{i-1}{n}$
1	50.60	0.0002	0.000009	0.00019	0.000009
2	53.98	0.0004	0.000047	0.00035	-0.000153
3	56.61	0.0006	0.000144	0.00046	-0.000256
4	58.24	0.0008	0.000272	0.00053	-0.000328
5	58.39	0.001	0.000288	0.00071	-0.000512
6	59.08	0.0012	0.000373	0.00083	-0.000627
7	59.71	0.0014	0.000468	0.00093	-0.000733
8	60.50	0.0016	0.000619	0.00098	-0.000781
9	61.68	0.0018	0.000923	0.00088	-0.000677
10	62.61	0.002	0.001253	0.00075	-0.000547
...	...	...	...	...	...
4990	144.72	0.998	0.997678	0.00032	-0.000122
4991	145.12	0.9982	0.997833	0.00037	-0.000167
4992	145.37	0.9984	0.997926	0.00047	-0.000274
4993	145.89	0.9986	0.998108	0.00049	-0.000292
4994	146.09	0.9988	0.998172	0.00063	-0.000428
4995	146.86	0.999	0.998405	0.00060	-0.000395
4996	147.83	0.9992	0.998658	0.00054	-0.000342
4997	148.25	0.9994	0.998757	0.00064	-0.000443
4998	150.19	0.9996	0.999127	0.00047	-0.000273
4999	150.73	0.9998	0.999210	0.00059	-0.000390
5000	154.53	1	0.999615	0.0003853	-0.000185
	$1.36/\sqrt{n} =$ 0.0192			$D_n =$ 0.0118	
	$D_n^- =$ 0.0118			$D_n^+ =$ 0.00094	

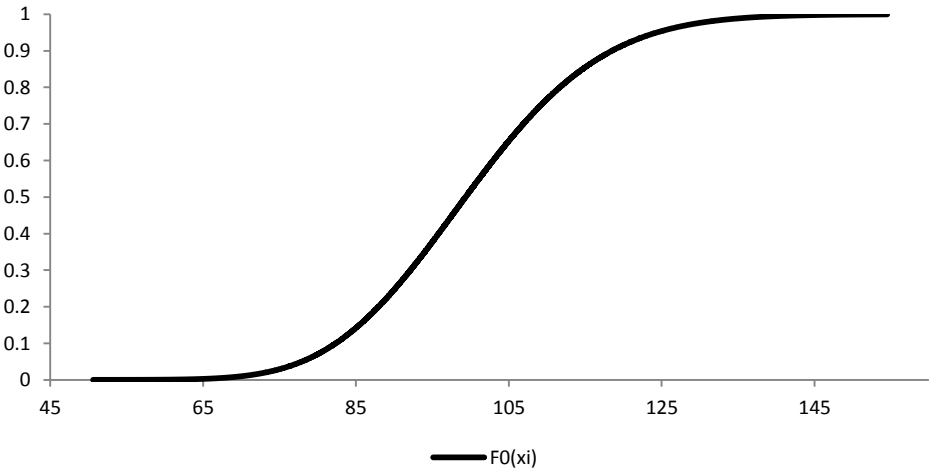
Al considerar la regla de decisión: rechazar la hipótesis si el estadístico D_n es mayor que un valor crítico al $(1-\alpha)$ nivel de confianza, entonces como sucede el escenario contrario, no se rechaza la hipótesis nula, por lo que se concluye que no existen diferencias importantes entre la función de distribución de los datos simulados y la función de distribución de una variable Gamma$(50;2)$.

La Gráfica 1.31 muestra porque la hipótesis nula no se rechazó, dando seguridad que los datos simulados provienen de la distribución propuesta, ya que es plasma los puntos $\left(x_{(i)}, \frac{i}{n}\right)$ mientras que la Gráfica 1.32 es visualiza los puntos $\left(x_{(i)}, F_0\left(x_{(i)}\right)\right)$, donde se puede observar su similitud.

Gráfica 1.31: función de distribución empírica de los datos simulados de una Gamma(50;2)



Grafica 1.32: Estadísticos de orden de la muestra simulada valuados sobre la función de distribución de una gamma(50;2)



Weibull

Introducción

El nombre de esta distribución es atribuido al físico sueco Waloddi Weibull, quien la utilizó para realizar diversos análisis sobre la resistencia de varios materiales de uso industrial; a lo largo de la historia ha sido de ayuda en estudios de control de calidad, comportamiento de la velocidad del viento, intensidad de la lluvia, estudios médicos sobre la distribución de la carcinogénesis, y diversos problemas que afectan al corazón o que provocan la aparición de tumores.

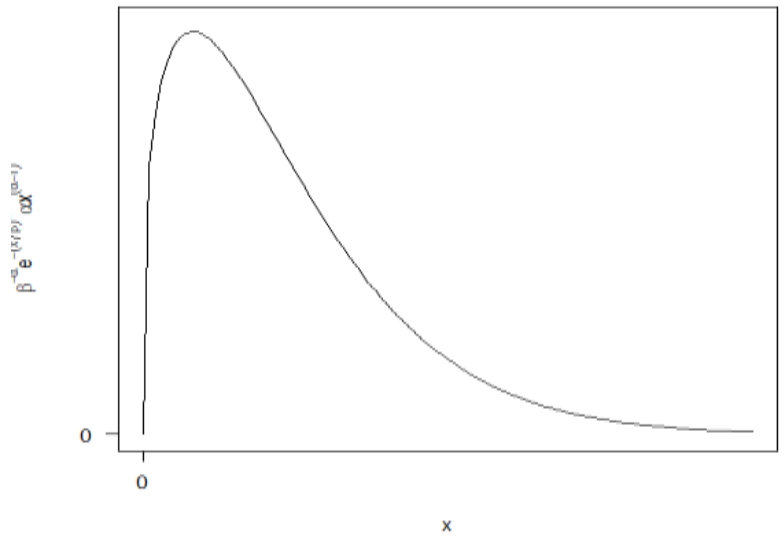
También se le puede encontrar mencionada en análisis de degradación de papel a nivel microscópico, estudios meteorológicos, de agricultura, o sobre algunos tipos de procesos, enfocando el análisis en el tiempo de reacción; además de análisis demográficos sobre la migración y el alcoholismo.

Características Principales

Esta distribución tiene como función de densidad$f(x)$, la cual depende de los parámetros α, β, ambos positivos, y su forma se puede ver en la Gráfica 1.33.

$$\boldsymbol{f(x) = \beta^{-\alpha}\alpha x^{\alpha-1}e^{-\left(\frac{x}{\beta}\right)^{\alpha}} \quad para\ x > 0,}$$

Gráfica 1.33:Función de densidad de una variable Weibull(α,β)

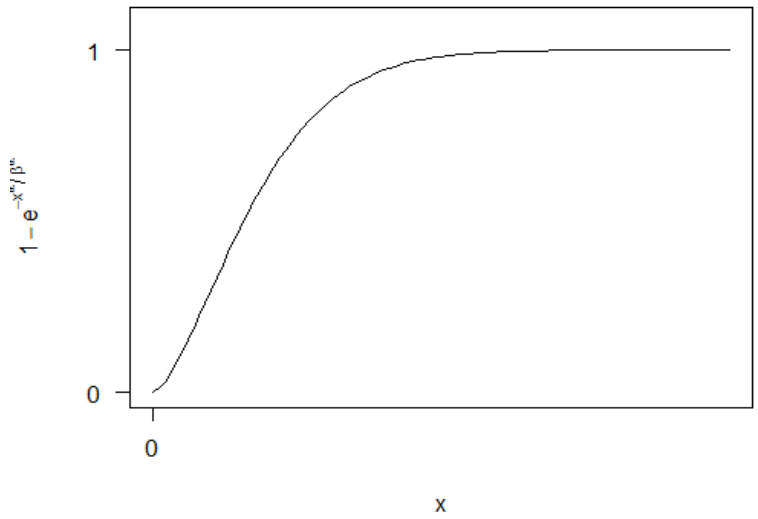


Una útil relación de la distribución Weibull con la distribución exponencial, es el hecho que si una variable Y tiene distribución exponencial con media β^{α} entonces $X = Y^{1/\alpha}$ se distribuye Weibull con parámetros α y β así que su distribución se puede obtener por medio de un despeje como se ve a continuación.

$$\boldsymbol{P(X \leq x) = P(Y \leq x^{\alpha}) = 1 - e^{-x^{\alpha}/\beta^{\alpha}} \quad \forall x > 0}$$

La forma que toma la función de distribución de la variable Weibull puede ser observada en la Gráfica 1.34.

Gráfica 1.34:función de distribución de una Weibull(α,β)



Con lo anterior, se obtiene la función de riesgo la cual explica el porqué de su gran uso sobre temas poblaciones.

$$\boldsymbol{h(x) = \frac{\beta^{-\alpha}\alpha x^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}}}{e^{-x^{\alpha}/\beta^{\alpha}}} = \frac{\alpha x^{\alpha-1}}{\beta^{\alpha}}}$$

De esta función se nota que cuando $\alpha < 1$ se tiene una función decreciente de x, mientras que si $\alpha > 1$ se tiene una función creciente. Un caso particular caso es cuando $\alpha = 1$, lo cual implica tener una distribución exponencial, aunque expresada de distinta forma, la cual tiene una función de riesgo constante.

Las siguientes propiedades de la distribución Weibull son la esperanza y los momentos centrales de orden mayor o igual a dos, listados hasta la kurtosis

$$\boldsymbol{E(X) = \beta\Gamma\left(1 + \frac{1}{\alpha}\right)}$$

$$\boldsymbol{Var(X) = \beta^2\left(\Gamma\left(1 + \frac{2}{\alpha}\right) - \Gamma^2\left(1 + \frac{1}{\alpha}\right)\right)}$$

$$\boldsymbol{Asimetría = As(X) = \beta^3\left(\Gamma\left(1 + \frac{3}{\alpha}\right) - 3\Gamma\left(1 + \frac{1}{\alpha}\right)\left(\Gamma\left(1 + \frac{2}{\alpha}\right) - \Gamma^2\left(1 + \frac{1}{\alpha}\right)\right) - \Gamma^3\left(1 + \frac{1}{\alpha}\right)\right)}$$

$$\boldsymbol{Curtosis = K(X) = \beta^4\Gamma\left(1 + \frac{4}{\alpha}\right) - 4E(X)As(X) - 6E^2(X)Var(X) - E(X)^4}$$

Simulación de valores

Para simular de esta distribución el método a usar, es el de inversión, por medio del siguiente despeje

$$\boldsymbol{1 - e^{-\frac{x^{\alpha}}{\beta^{\alpha}}} = U}$$

$$e^{-\frac{x^{\alpha}}{\beta^{\alpha}}} = U$$

$$\frac{x^{\alpha}}{\beta^{\alpha}} = -\ln(U)$$

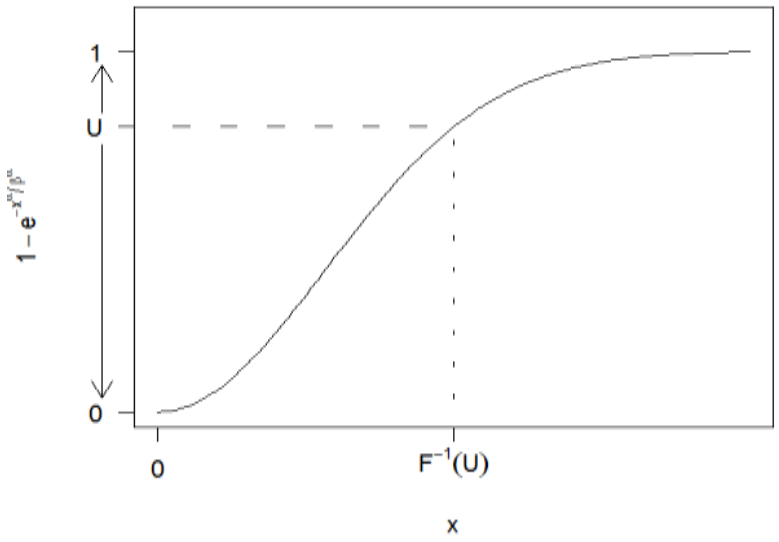
$$x^{\alpha} = \beta^{\alpha}(-\ln(U))$$

$$x = \beta(-\ln(U))^{\frac{1}{\alpha}}$$

Donde U es un número generado aleatoriamente Uniforme sobre el intervalo $(0,1)$. La forma en cómo trabaja este método sobre la distribución Weibull es vista en la Gráfica 1.35, por lo tanto el pseudocódigo para generar un valor distribuido Weibull(α, β), es el siguiente

1. $\boldsymbol{Generar\ U \sim U(0,1)}$
2. $\boldsymbol{Regresar\ X = \beta(-\ln(U))^{1/\alpha}}$

Grafica 1.35: Inversión de la función de distribución de una Weibull(α,β)



Para ejemplificar este algoritmo más detalladamente se ha generado una muestra de tamaño $n = 5000$,de números pseudoaleatorios Uniformes en $(0,1)$, generados en Excel® por medio de la función $ALEATORIO()$, para posteriormente aplicar la transformación anterior, con parámetros propuestos de $\alpha = 3$, $\beta = 2$; una muestra parcial de los datos se encuentra en la Tabla 1.30 donde además se hallan estimaciones para la media y la varianza de la distribución, las cuales se pueden comparar con sus respectivos valores basados en los parámetros, que son:

$$\boldsymbol{E(X) = 2\left(\Gamma\left(1+\frac{1}{3}\right)\right) = 2\Gamma\left(\frac{4}{3}\right) = 1.785959}$$

$$\boldsymbol{Var(X) = 4\left(\Gamma\left(1+\frac{2}{3}\right) - \Gamma^2\left(1+\frac{1}{3}\right)\right) = 4\left(\Gamma\left(\frac{5}{3}\right) - \Gamma^2\left(\frac{4}{3}\right)\right) = 0.421328}$$

Tabla 1.30: primeros 20 valores de la simulación de una variable Weibull(α=3 , β=2)					
i	$U_i \sim U(0,1)$	$x_i = 2*(-\ln(U_i))^{1/3}$	i	$U_i \sim U(0,1)$	$x_i = 2*(-\ln(U_i))^{1/3}$
1	0.242	2.246	11	0.087	2.692
2	0.98	0.542	12	0.02	3.148
3	0.739	1.342	13	0.221	2.294
4	0.611	1.579	14	0.326	2.077
5	0.92	0.874	15	0.206	2.33
6	0.549	1.686	16	0.111	2.601
7	0.039	2.957	17	0.527	1.724
8	0.781	1.256	18	0.972	0.607
9	0.415	1.916	19	0.607	1.587
10	0.988	0.457	20	0.389	1.962
		$\bar{X} = 1.7829$		$S^2 = 0.4246$	

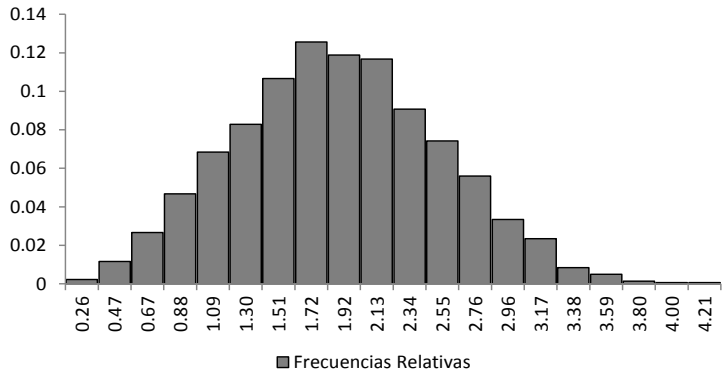
Como se ha comparado antes la media muestral $\bar{X}$ y la varianza muestral S^2 son valores cercanos a la esperanza y varianza teóricas. Para una visión integral sobre el comportamiento de los datos lo más sencillo es realizar una gráfica, a partir de una tabla de frecuencias, generada con la misma metodología con la que se definieron los subintervalos empleados en los valores anteriormente simulados de la distribución Gamma.

Una vez con los subintervalos definidos y con el apoyo de la función $FRECUENCIA()$ de Excel® se puede construir una tabla de frecuencias. La Tabla 1.31 contiene el resultado de agrupar los datos en 20 subdivisiones, se adjunta además la frecuencia relativa, así como los valores del mínimo y el máximo de la muestra.

Tabla 1.31: Frecuencias de valores simulados de una Weibull(3;2)					
Rango	*Frecuencia*	*Frecuencia Relativa*	*Rango*	*Frecuencia*	*Frecuencia Relativa*
(0,0.26]	11	0.002	(2.13,2.34]	454	0.091
(0.26,0.47]	58	0.012	(2.34,2.55]	371	0.074
(0.47,0.67]	133	0.027	(2.55,2.76]	280	0.056
(0.67,0.88]	234	0.047	(2.76,2.96]	167	0.033
(0.88,1.09]	342	0.068	(2.96,3.17]	117	0.023
(1.09,1.30]	414	0.083	(3.17,3.38]	42	0.008
(1.30,1.51]	533	0.107	(3.38,3.59]	25	0.005
(1.51,1.72]	628	0.126	(3.59,3.80]	7	0.001
(1.72,1.92]	594	0.119	(3.80,4.00]	3	0.001
(1.92,2.13]	584	0.117	(4.00,?)	3	0.001
		$Mínimo = 0.0505$		$Máximo = 4.213$	

En la Gráfica 1.36 están los valores de las frecuencias relativas, las cuales muestran la distribución de los datos.

Gráfica 1.36: Frecuencias relativas de los valores simulados de una variable Weibull(3;2)



La prueba estadística que se realizará para mostrar el ajuste de los valores simulados, realizada sobre la distribución de tales datos, es la prueba de bondad de ajuste de Kolmogorov-Smirnov que contrasta las siguientes hipótesis.

$$\boldsymbol{H_0: F_0(x) = F(x)} \ \ \forall \boldsymbol{x}$$

$$\boldsymbol{H_1: F_0(x) \neq F(x)} \ \ \boldsymbol{para\ alguna\ x}$$

Donde $F_0(x)$ la función de distribución Weibull de parámetros $\alpha = 3$, $\beta = 2$ y $F(x)$ es la función de distribución de los datos. Para realizar la prueba se debe calcular el estadístico de prueba $\boldsymbol{D_n}$ definido previamente.

Posteriormente se compara el valor $\boldsymbol{D_n}$ con el valor crítico $1.36/\sqrt{n}$, obtenido de tablas específicas de esta prueba, correspondiente a un nivel de significancia del 5%. La Tabla 1.32 resume los cálculos de la prueba junto con los estadísticos necesarios.

Tabla 1.32: Cálculos parciales de la prueba K-S

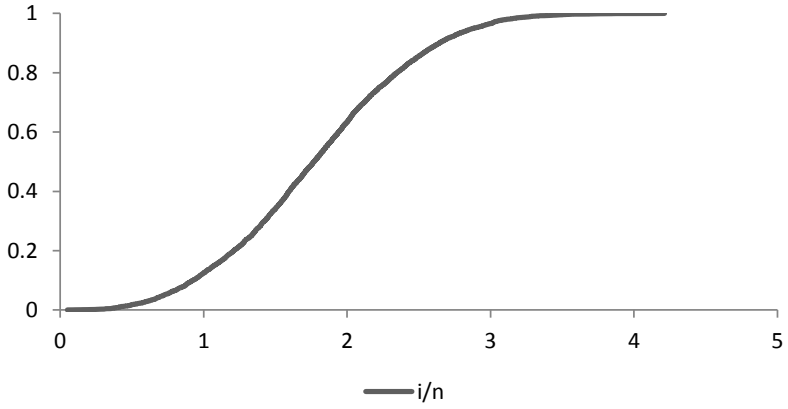
i	$Weibull(3;2)$	$\frac{i}{n}$	$F_0(x_{(i)}) = 1 - e^{\left(\frac{x_{(i)}}{\beta}\right)^{\alpha}}$	$i/n - F_0(x_{(i)})$	$F_0(x_{(i)}) - (i-1)/n$
1	0.05	0.0002	0.00002	0.0002	0.00002
2	0.09	0.0004	0.00009	0.0003	-0.00011
3	0.10	0.0006	0.00014	0.0005	-0.00026
4	0.14	0.0008	0.00036	0.0004	-0.00025
5	0.16	0.0010	0.00047	0.0005	-0.00033
...	...	...	...	...	...
4995	3.84	0.9990	0.99917	-0.00017	0.000367
4996	3.85	0.9992	0.99922	-0.00002	0.000216
4997	3.88	0.9994	0.99933	0.00007	0.00013
4998	4.11	0.9996	0.99983	-0.00023	0.000425
4999	4.20	0.9998	0.99991	-0.00011	0.000307
5000	4.21	1	0.99991	0.00009	0.000113
	$1.36/\sqrt{n} = 0.01923$			$D_n = 0.008555$	
	$D_n^- = 0.006807$			$D_n^+ = 0.008555$	

La regla de decisión que determina la conclusión de esta prueba bajo el nivel de significancia propuesto es: rechazar H_0 si $\frac{1.36}{\sqrt{n}} \leq D_n$ por lo tanto se concluye con un

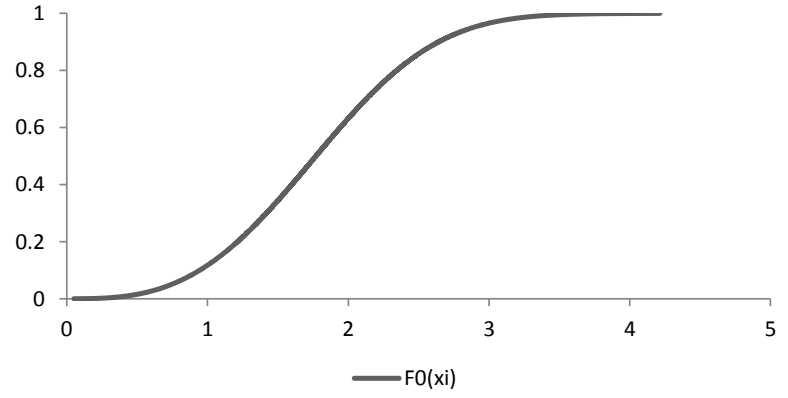
nivel de significancia del 5% que no existen evidencias estadísticas que muestren diferencias significativa entre la función de distribución de los valores simulados y la función de distribución de una variable Weibull($\alpha = 3, \beta = 2$).

La manera gráfica de justificar la conclusión anterior, es por medio de comparar las gráficas correspondientes a $F_n(x_{(i)})$ y $F_0(x_{(i)})$, las cuales se hallan en la Gráfica 1 37 y la Gráfica 1.38 respectivamente, en las cuales se observa la similitud entre ellas, lo cual explica porqué las diferencias sobre las cuales se realizó la prueba, sean no significativas.

Grafica 1.37: Función de distribución empírica de las simualciones de una Weibull(3;2)



Grafica 1.38: Función $F_0()$ de una Weibull(3;2) valuada sobre cada xi



Normal

La historia de la distribución Normal empieza como una aproximación de la variable Binomial por parte de De Moivre, quien publicó en 1733 un escrito en latín sobre la estrecha relación que tenían ambas distribuciones. Años más tarde, en 1774 Laplace uso la distribución Normal como una aproximación de la variable Hipergeométrica. La idea de ver la distribución Normal como la aproximación de una suma fue por parte de Gauss en 1816, quien estudió la relación entre la Normal y la suma de errores independientes.

La posibilidad de relacionarla con una suma de variables convierte al modelo de la Normal en una distribución importante dentro de la estadística, además que existen varios modelos que incluyen en alguno de sus supuestos un componente distribuido de manera Normal. La justificación de su uso sobre una amplia gama de problemas, es debido al Teorema de Central del Límite, el cual establece que la suma de variables aleatorias independientes e idénticamente distribuidas tiende a distribuirse como una variable aleatoria Normal.

En muchos casos el TCL es de gran utilidad cuando la cantidad de datos que se tienen sobre un fenómeno son grandes, es por eso que se emplean aproximaciones por medio de esta distribución.

Por otro lado en diversos estudios referentes a temas de la naturaleza, se ha mostrado que muchos fenómenos tienen una distribución Normal, o al menos un comportamiento simétrico en sus frecuencias, por ejemplo en variables antropométricas como el peso, la estatura o la longitud del pie, también en variables psicométricas como el coeficiente intelectual (I.Q. por sus siglas en ingles), e incluso en variables del ámbito educativo como la calificación de exámenes objetivos.

Otro ámbito importante en el que se puede encontrar a la distribución Normal es en la distribución de las diferencias entre las predicciones de un modelo matemático y las observaciones hechas, en el contexto de un fenómeno estudiado; a estos errores del modelo se le conoce como ruido blanco, término informalmente establecido dentro de las telecomunicaciones para describir lo que percibe un aparato en ausencia de una señal que pueda interpretar correctamente.

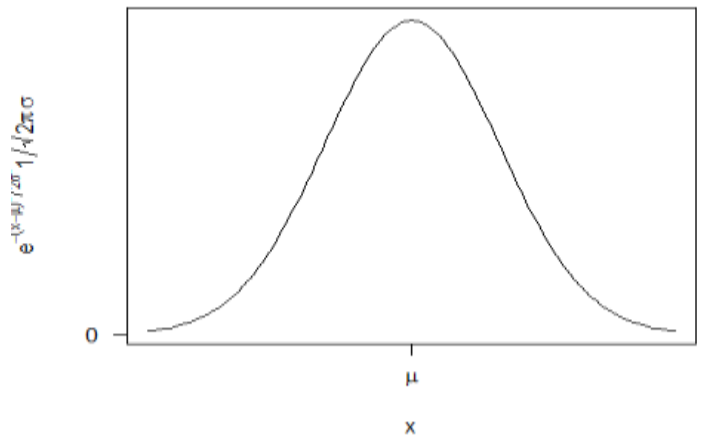
Características principales

La función de densidad de esta variable es:

$$\boldsymbol{f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad para \quad -\infty < x < \infty}$$

Esta función de densidad f depende de los parámetros μ, el cual es el centro de la función, y σ^2 que es su parámetro de escala, cumplen $-\infty < \mu < \infty$ y $\sigma^2 > 0$. La forma de f puede observarse en la Gráfica 1.39.

Gráfica 1.39: Densidad de una variable Normal(μ,σ^2)



La función de distribución para esta variable se define como

$$\boldsymbol{F(x) = \int_{-\infty}^{x} f(x)dx = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx}$$

Pero tal integral no tiene una solución analítica, por lo tanto es necesario recurrir a técnicas numéricas de integración, como el método de Simpson o la cuadratura de Gauss, para poder obtener un valor aproximado de F en un punto dado.

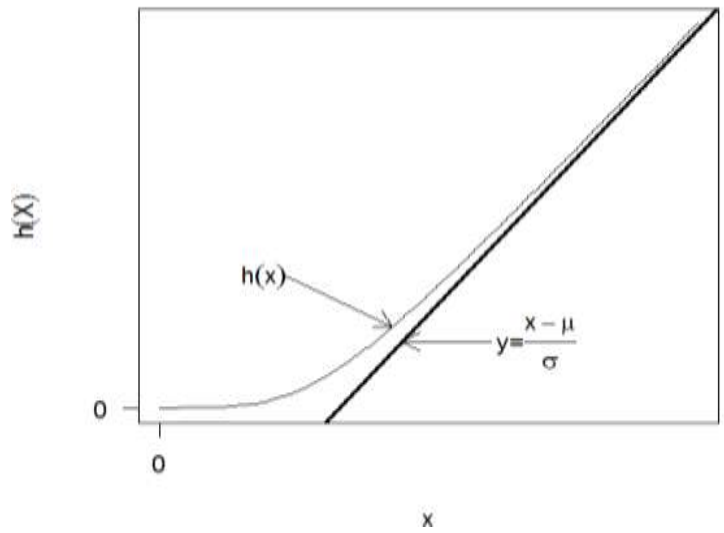
La función de riesgo definida como $h(x) = \frac{f(x)}{1-F(x)}$, la cual indica la intensidad con la que ocurren los eventos o el fenómeno, que describe la variable aleatoria, viéndola en un contexto de tiempo, también debe ser evaluada por medio de una alternativa numérica; aunque es posible realizar un ligero análisis haciendo tender a ∞ el valor de x, y con ayuda de la regla de l'Hôpital, conocer el comportamiento de $h(x)$

$$\boldsymbol{\lim_{x\to\infty} h(x) = \lim_{x\to\infty} \frac{f(x)}{1 - F(x)} = \lim_{x\to\infty} \frac{f'(x)}{-f(x)}}$$

$$\boldsymbol{\lim_{x\to\infty} \frac{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \left(-\frac{x-\mu}{\sigma}\right)}{-\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}} = \lim_{x\to\infty} \frac{x-\mu}{\sigma} = \infty}$$

Se observa que la función de riesgo se aproxima al comportamiento de una recta de pendiente $1/\sigma$ y ordenada al origen $-\frac{\mu}{\sigma}$ conforme el valor de x crece. La grafica 1.40 muestra la función de riesgo de una distribución Normal, junto con la recta antes mencionada para mostrar la aproximación.

Gráfica 1.40: Función de riesgo de una Normal(μ,σ^2)



Para determinar los momentos centrales de esta distribución es de utilidad la función generadora de momentos definida como

$$\boldsymbol{M_x(t) = E(e^{tX}) = \int_{-\infty}^{\infty} \frac{e^{tx}}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = e^{\mu t + \frac{t^2\sigma^2}{2}}}$$

La cual cumple la propiedad que su $n-$ésima derivada evaluada en cero es el $n-$ésimo momento de la variable, es decir:

$$\boldsymbol{\frac{d^n}{dt^n} M_X(t)|_{t=0} = E(X^n)}$$

Por medio de esta relación se determinan los momentos de la distribución Normal, por lo que se listan los primeros 4, asociados a la media ($E(X)$), la varianza ($VAR(X)$), asimetría ($E\left(\left(X-E(X)\right)^3\right)$) y kurtosis ($E\left(\left(X-E(X)\right)^4\right)$).

$$\boldsymbol{E(X) = \mu}$$

$$\boldsymbol{Var(X) = E\left(\left(X-E(X)\right)^2\right) = \sigma^2}$$

$$\boldsymbol{E\left(\left(X-E(X)\right)^3\right) = 0}$$

$$\boldsymbol{E\left(\left(X-E(X)\right)^4\right) = 3\sigma^4}$$

Simulación de valores

La distribución Normal aparece en estudios que describen el comportamiento de diversos fenómenos o en los supuestos de algunos modelos, como en el caso de la distribución de los residuos en el contexto del análisis de regresión. Generar valores una distribución Normal, es una de las primeras variables en intentar simularse, por lo que con el paso de los años se desarrollaron diversos métodos para generar valores de la distribución Normal.

Uno de los primeros algoritmos enfocados en la generación de valores Normales, es el llamado “de las 12 uniformes” y está basado en el Teorema Central del Límite. Surgió debido a que, en el comienzo del desarrollo de las computadoras era de suma importancia ahorrar lo más posible el número de operaciones que realizaba la computadora por cada iteración.

Primero se debe retomar algunas propiedades de la distribución uniforme: La media de una distribución uniforme en (0,1) es ½ y tiene una varianza de $1/12$. Por otro lado, para variables IID, $X_i, i = 1, \dots, n$ cada una con media μ y varianza σ^2, ambas finitas, al sumarse se obtiene una variable de media $n\mu$ y varianza $n\sigma^2$, entonces si cada una es uniforme en (0,1), el resultado de la suma es una variable de media $n/2$ y varianza $n/12$.

El Teorema Central del Límite enuncia que la suma de las variables aleatorias Independientes e Idénticamente Distribuidas (IID), en este caso uniformes ($\sum_{i=1}^{n} U_i = X, U_i \sim U(0,1)$) se aproximan al comportamiento de una Normal $(n/2, n/12)$. Esta variable Normal se puede estandarizar por medio de la transformación $Z = \frac{X - \frac{n}{2}}{\sqrt{n/12}}$ la cual se distribuye Normal $(0,1)$, entonces si se elige $n = 12$, implica que $Z = X - 6$, ahorra la operación de la división entre la varianza pues es 1. Por lo tanto el algoritmo de generación de una Normal estándar es:

1. $\boldsymbol{Generar\ U_1, \dots, U_{12}\ independientes\ Uniformes\ en\ el\ intervalo\ (0,1)}$
2. $\boldsymbol{Regresar\ X = \sum_{i=1}^{12} U_i - 6}$

Dado que el método anterior genera una variable X con distribución Normal(0,1) que también es llamada estándar, la forma de generalizar para cualquier media μ y varianza σ^2 es por medio de la transformación $X' = \mu + \sigma X$.

Generando una serie de valores de la distribución Normal, se muestra cómo se logra una aproximación a la distribución. El tamaño de la muestra simulada fue de 5000, los valores elegidos de los parámetros para este ejemplo son $\mu = 10$, $\sigma^2 = 5$. La Tabla 1.33 muestra de manera parcial, los primeros 20 valores obtenidos por medio del algoritmo, junto con estimaciones de la media y la varianza, $\bar{X}$ y S^2.

Tabla 1.33: Primeros 20 valores de la simulación de una v.a. Normal(10;5)			
i	$Normal(10;5)$	i	$Normal(10;5)$
1	11.288	11	8.145
2	11.497	12	8.300
3	6.621	13	9.008
4	10.773	14	8.491
5	11.285	15	4.691
6	8.649	16	10.152
7	7.457	17	6.361
8	12.317	18	15.392
9	9.158	19	9.659
10	7.848	20	10.421
$\bar{X} = 9.968$		$S^2 = 5.222$	

Como es de esperarse, el valor de la media muestral es cercano al valor del parámetro μ, la varianza muestral por su parte también muestra gran cercanía al valor del parametro σ^2.

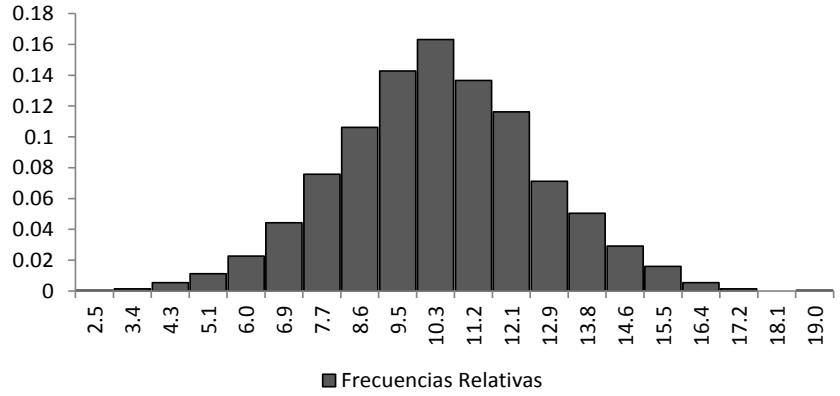
Acontinuación se muestra el comportamiento de los datos por medio de una tabla de frecuencias, en la cual se ha subdivido el dominio de la distribución Normal en $k = 20$, con la misma metodología antes empleada.

La Tabla 1.34 contiene la información anteriormente descrita para los valores simulados, junto con la frecuencia relativa. Los subintervalos extremos usan los conceptos $-\infty$ e ∞, sólo de manera simbólica, pues se limitan al mínimo y máximo de las simulaciones generadas.

Tabla 1.34: Frecuencias de los valores simulados de una Normal (μ=10;σ^2=5)					
Rango	*Frecuencia*	*Frecuencia Relativa*	*Rango*	*Frecuencia*	*Frecuencia Relativa*
(-∞,2.54]	2	0.0004	(10.32,11.19]	683	0.1366
(2.54,3.41]	7	0.0014	(11.19,12.05]	581	0.1162
(3.41,4.27]	27	0.0054	(12.05,12.92]	356	0.0712
(4.27,5.14]	56	0.0112	(12.92,13.78]	252	0.0504
(5.14,6.00]	113	0.0226	(13.78,14.64]	146	0.0292
(6.00,6.86]	221	0.0442	(14.64,15.51]	80	0.016
(6.86,7.73]	379	0.0758	(15.51,16.37]	27	0.0054
(7.73,8.59]	531	0.1062	(16.37,17.24]	7	0.0014
(8.59,9.46]	714	0.1428	(17.24,18.10]	0	0.0
(9.46,10.32]	816	0.1632	(18.10,∞)	2	0.0004
		Mínimo = 1.67		*Máximo =* 18.96	

La Gráfica 1.41 muestra el comportamiento de las frecuencias, donde se observa que su forma tiene cierta similitud con la curva de la Gráfica 1.39.

Gráfica 1.41: Frecuencias relativas de los valores simulados deuna distribución Normal(10,5)



Luego para corroborar que los datos generados se distribuyen como una Normal, es realizará la prueba de bondad de ajuste kolmogorov-smirnov. Partiendo de contrastar las siguientes hipótesis.

$$\boldsymbol{H_0}: \boldsymbol{F_0(x) = F(x)} \quad \forall \boldsymbol{x}$$

$$\boldsymbol{H_1}: \boldsymbol{F_0(x) \neq F(x)\ para\ alguna\ x}$$

Donde $F_0(x)$es la función de distribución de una Normal y F es aproximada por la función de distribución empírica de los datos. Los cálculos individuales necesarios para realizar la prueba, junto con el estadístico de prueba y el valor crítico para un nivel de confianza del 5%, se muestran en la Tabla 1.35.

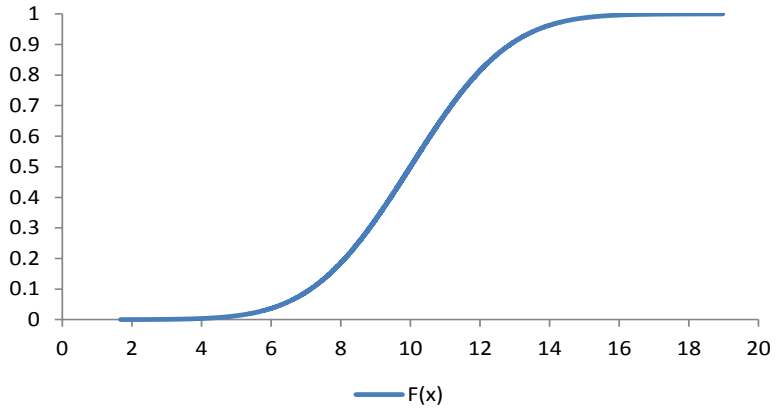
Tabla 1.35: Muestra parcial de los cálculos para realizar la prueba K-S

i	$x_{(i)}$	i/n	$F_0(x_{(i)})$	$i/n - F_0(x_{(i)})$	$F_0(x_{(i)}) - (i-1)/n$
1	1.678	0.0002	0.0001	0.0001	0.0001
2	2.301	0.0004	0.00029	0.00011	0.00009
3	2.567	0.0006	0.00044	0.00016	0.00004
4	2.905	0.0008	0.00075	0.00005	0.00015
5	3.013	0.001	0.00089	0.00011	0.00009
...	...	...	...	...	...
4996	16.667	0.9992	0.99857	0.09041	-0.09021
4997	17.068	0.9994	0.99921	0.07814	-0.07794
4998	17.11	0.9996	0.99926	0.07711	-0.07691
4999	18.177	0.9998	0.99987	0.05079	-0.05059
5000	18.967	1	0.99997	0.03646	-0.03626
		$1.36/\sqrt{n} = 0.00844$		$D_n = 0.01923$	
		$D_n^- = 0.01661$		$D_n^+ = 0.01661$	

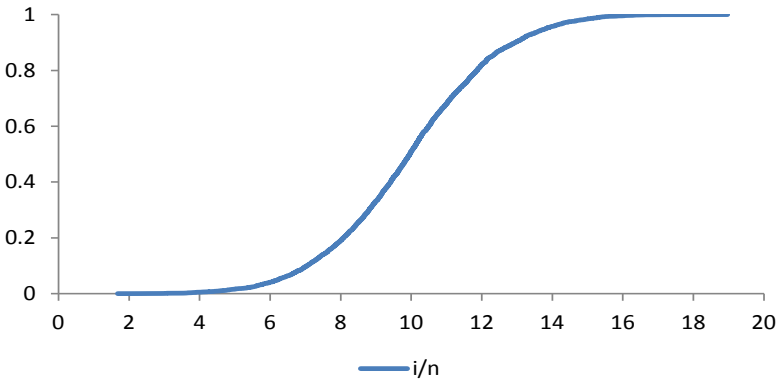
De acuerdo con la regla de decisión para la prueba de Kolmogorov-Smirnov , puesto que para los datos simulados $\frac{1.36}{\sqrt{n}} > D_n$ no se rechaza H_0, llegando a la conclusión que no existen diferencias estadísticamente significativas, entre la función de distribución de los datos simulados y la función de distribución de una Normal$(10; 5)$.

La Gráfica 1.42 muestra la función de distribución de la Normal$(10;5)$ valuada sobre los estadísticos de orden mientras que la Gráfica 1.43 muestra la relación de i/n contra los estadísticos de orden. Se observa en las gráficas que si se sobreponen una sobre otra, las diferencias serian prácticamente insignificantes.

Gráfica 1.42: Función de distribución Normal(10;5) valuada sobre las simulaciones



Gráfica 1.43: Función de distribución empírica delas simulaciones de unaNormal(10;5)



Log-Normal

La historia de la distribución Log-Normal se refiere a finales del siglo XIX y principios del XX, cuando Galton, McAlister, Kapteyn y Van Uven plantean esta distribución por la necesidad de estudiar modelos que involucraban efectos multiplicativos entre un conjunto de variables aleatorias $\{X_i\}_{i=1}^{n}$, es decir, cuando se desea conocer el comportamiento de

$$\boldsymbol{T_n = \prod_{i=1}^{n} X_i}$$

A esta nueva variable se le puede aplicar el logaritmo natural y usar sus propiedades para llegar a

$$\boldsymbol{Log(T_n) = \sum_{i=1}^{n} Log(X_i)}$$

Que por el Teorema Central del Límite se aproxima en distribución a una Normal, entonces si se define $T_n = Y$, ésta se distribuye Log-Normal si $X = \ln(Y)$ se distribuye Normal.

Otra aplicación importante se puede encontrar en la teoría económica, pues ha sido utilizada para ajustar funciones de producción en datos económicos, aunque en este contexto se le suele llamar distribución de Cobb-Douglas; por otra parte también se pueden encontrar referencias en un estudio de la edad al primer matrimonio realizado por Witsel en 1917; otros estudios destacables son: por parte de Gibrat en relación a variables económicas ; Gaddum sobre la distribución de las dosis críticas de diversas drogas; además de temas como geología, bioquímica, procesos de manufactura, seguros automovilísticos, y otros en los cuales su principal objetivo es estabilizar la varianza de una muestra por medio de una transformación logarítmica.

Características principales

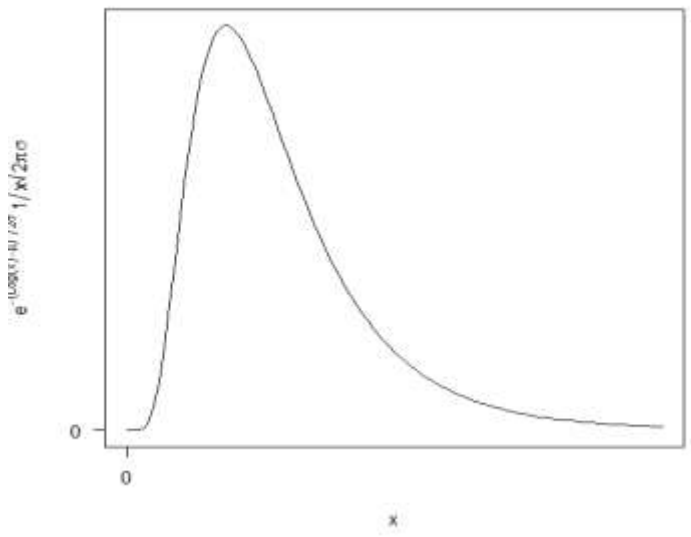
La función de densidad de esta variable es la siguiente $f(x)$; la cual se observa en la Gráfica 1.44

$$\boldsymbol{f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(log(x)-\mu)^2}{2\sigma^2}} \quad x > 0}$$

Donde los parámetros deben cumplir

$$\boldsymbol{\sigma^2 > 0}\,, -\infty < \mu < \infty$$

Gráfica 1.44: Función de densidad de una Log-Normal(μ,σ^2)



La función de distribución de esta variable es

$$\boldsymbol{F(x) = \int_{-\infty}^{x} f(x)dx = \int_{-\infty}^{x} \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\log(x)-\mu)^2}{2\sigma^2}} dx}$$

Esta integral no tiene una solución de forma analítica, aunque es posible aproximarla de manera muy exacta por métodos numéricos. Ahora se trata a la función de riesgo definida como:

$$\boldsymbol{h(x) = \frac{f(x)}{1 - F(x)}}$$

La función h aporta información sobre la intensidad en que el fenómeno ocurre, suponiendo que su comportamiento sigue la distribución $F(x)$; entonces para tener una primera idea de la función de riesgo se puede hacer tender los valores de x a infinito para ver cómo se comporta en el límite.

$$\boldsymbol{\lim_{x\to\infty} h(x) = \lim_{x\to\infty} \frac{f(x)}{1 - F(x)}}$$

Como $F(x) = \int_{-\infty}^{x} f(x)dx$ no tiene una solución entonces se recurre a la regla de l'Hôpital, ya que tanto el numerador como el denominador tienden a 0 conforme los valores de x aumentan

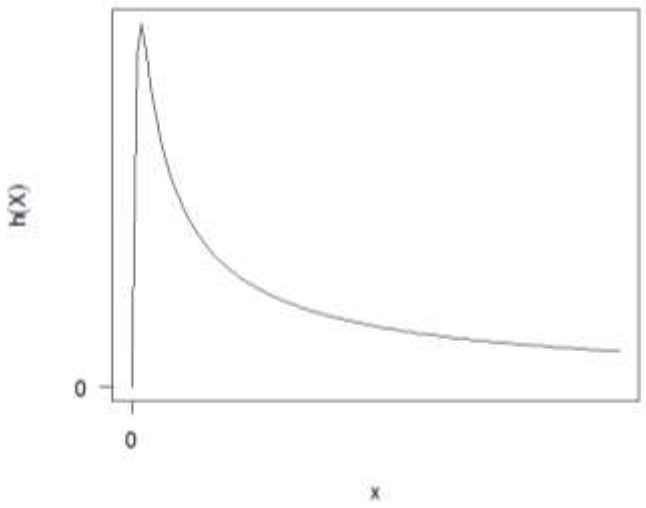
$$\boldsymbol{\lim_{x\to\infty} h(x) = \lim_{x\to\infty} \frac{f'(x)}{-f(x)}}$$

$$\boldsymbol{= \lim_{x\to\infty} \frac{\frac{-1}{x^2\sqrt{2\pi\sigma^2}} e^{-\frac{(log(x)-\mu)^2}{2\sigma^2}} - \frac{1}{x^2\sqrt{2\pi\sigma^2}} e^{-\frac{(log(x)-\mu)^2}{2\sigma^2}} \left(\frac{\log(x)-\mu}{\sigma^2}\right)}{\frac{-1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(log(x)-\mu)^2}{2\sigma^2}}}}$$

$$\boldsymbol{\lim_{x\to\infty} \frac{\left(\log(x) + \sigma^2 - \mu\right)}{\sigma^2 x} = 0}$$

Se llega al resultado dado que se tienen 2 límites de una constante entre x, entonces, tales limites van a cero; mientras que el límite de la forma $\log(x)/x$ va a cero puesto que $\log(x) < x$ para todo valor de x la Gráfica 1.45, muestra la forma que toma la función de riesgo de una distribución log-normal.

Gráfica 1.45: Función de riesgo de una Log-Normal(μ,σ^2)



Para determinar sus momentos centrales, primero se obtiene el $n-ésimo$ momento alrededor de 0 de la siguiente forma

$$\boldsymbol{E((X-0)^n) = E(X^n) = \int_{-\infty}^{\infty} x^n f(x)dx = \int_{0}^{\infty} \frac{x^n}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\log(x)-\mu)^2}{2\sigma^2}} dx}$$

Haciendo el cambio de variable $y = \log(x) \rightarrow e^y = x \ \rightarrow \ e^y dy = dx$; se tiene como resultado

$$\boldsymbol{\int_{-\infty}^{\infty} \frac{e^{ny}}{e^y\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} e^y dy = \int_{-\infty}^{\infty} \frac{e^{ny}}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy}$$

Ésta integral coincide con la función generadora de momentos de la normal de la cual se sabe que tiene como solución

$$\boldsymbol{e^{n\mu+\frac{n^2\sigma^2}{2}} = E(X^n)}$$

Con el cálculo anterior se puede entonces hallar los momentos centrales de la forma

$$\boldsymbol{E\Big(\big(X-E(X)\big)^n\Big)}$$

Desarrollando el producto dentro del operador esperanza y luego aplicando el operador, por ejemplo para obtener la asimetría de la distribución se considera $n = 3$ y para la kurtosis $n = 4$, entonces

$$\boldsymbol{asimetria = E\left(\big(x-E(x)\big)^3\right) = E\left(X^3 - 3X^2E(X) + 3XE^2(X) - E^3(X)\right)}$$

$$\boldsymbol{= E(X^3) - 3E(X^2)E(X) + 2E^3(X) = e^{3\mu+\frac{9\sigma^2}{2}} - 3e^{3\mu+\frac{5}{2}\sigma^2} + 2e^{3\mu+\frac{3\sigma^2}{2}}}$$

$$\boldsymbol{kurtosis = E\left(\big(x-E(x)\big)^4\right) = E\left(X^4 - 4X^3E(X) + 6E^2(X)X^2 - 4XE^3(X) + E^4(X)\right)}$$

$$\boldsymbol{= E(X^4) - 4E(X^3)E(X) + 6E^2(X)E(X^2) - 3E^4(X)}$$

$$= e^{4\mu+8\sigma^2} - 4e^{4\mu+5\sigma^2} + 6e^{4\mu+3\sigma^2} - 3e^{4\mu+2\sigma^2}$$

Simulación de valores

El método para su simulación se basa básicamente en la transformación de una variable Normal generada con cualquier método, pero es importante recalcar el hecho que aunque la variable Log-Normal está expresada en términos de los parámetros μ y σ^2, éstos no son ni la media ni la varianza, que con el método anteriormente visto para hallar momentos centrales se obtiene

$$E(Y) = e^{\mu+\sigma^2/2}$$

$$VAR(Y) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$$

En cambio los parámetros de la variable Normal derivada de aplicar logaritmo si son la media y la varianza, implicando la necesidad de despejar de una Normal($\mu', \sigma^{2\prime}$), sus parámetros para que, al introducir los valores al algoritmo genere una variable Log-Normal objetivo, en el caso que se conociera su media y su varianza, entonces se iguala la esperanza y varianza a los parámetros μ y σ^2respectivamente, por lo que se tiene

$$E(Y) = \mu = e^{\mu'+\sigma^{2'}/2}$$

$$VAR(Y) = \sigma^2 = e^{2\mu'+\sigma^{2'}}(e^{\sigma^{2'}} - 1)$$

$$como\ \mu^2 = e^{2\mu'+\sigma^{2'}}\ \ entonces\ \sigma^2 = \mu^2\left(e^{\sigma^{2'}} - 1\right)$$

$$\frac{\sigma^2}{\mu^2} + 1 = e^{\sigma^{2'}} \ \rightarrow\ \sigma^{2'} = \ln\left(\frac{\sigma^2}{\mu^2} + 1\right)$$

$$\mu = e^{\mu'+\ln\left(\frac{\sigma^2}{\mu^2}+1\right)/2} = e^{\mu'}e^{\ln\left(\sqrt{\frac{\sigma^2}{\mu^2}+1}\right)}$$

$$\mu = e^{\mu'}\sqrt{\frac{\sigma^2}{\mu^2} + 1} = \frac{e^{\mu'}\sqrt{\sigma^2+\mu^2}}{\mu}$$

$$\frac{\mu^2}{\sqrt{\mu^2+\sigma^2}} = e^{\mu'} \qquad entonces\ en\ resumen$$

$$\mu' = \ln\left(\frac{\mu^2}{\sqrt{\mu^2+\sigma^2}}\right)$$

$$\sigma^{2\prime} = \ln\left(1 + \frac{\sigma^2}{\mu^2}\right)$$

Conociendo los valores anteriores, el algoritmo es

1. $Generar\ Y \sim N\left(\mu', \sigma^{2'}\right)$

2. $\boldsymbol{Regresar\ X = e^{Y}}$

Para ejemplificar este algoritmo se ha generado una muestra de tamaño $n = 5000$ a partir de la generación de una muestra de una variable aleatoria de distribución Normal. La distribución que se desea obtener como resultado es una distribución Log-Normal de media 5 y varianza 2, entonces se deben obtener los parámetros de la Normal por medio de las formulas antes vistas

$$\boldsymbol{\mu' = \ln\left(\frac{5^2}{\sqrt{5^2+2}}\right) = \ln\left(\frac{25}{\sqrt{27}}\right) = 1.575312}$$

$$\boldsymbol{\sigma^{2'} = \ln\left(1+\frac{2}{5^2}\right) = \ln\left(\frac{27}{25}\right) = 0.076961}$$

Luego se generan valores x de una Normal(1.57,0.07) y después se calcula e^x. La Tabla 1.36 muestra parcialmente los valores generados de la Normal y su transformación exponencial, además de una estimación para la media $\bar{X}$, obtenido del promedio de las observaciones y un estimador de la varianza S^2 calculado como el promedio de las diferencias cuadráticas entre las observaciones y su media estimada, calculado tanto para la variable Normal($\bar{X}_{nor}, S^2_{nor}$) como para la transformación exponencial para obtener la variable Log-Normal($\bar{X}_{log}, S^2_{log}$).

Tabla 1.36: Muestra parcial de valores simulados de una variable Log-Normal

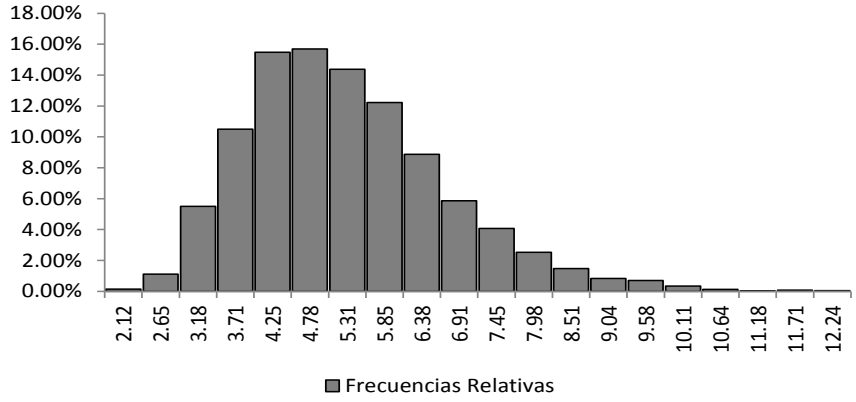
i	$x_i \sim N(1.57;0.07)$	e^{x_i}	i	$x_i \sim N(1.57;0.07)$	e^{x_i}
1	1.271	3.566	11	1.569	4.804
2	1.673	5.330	12	1.957	7.078
3	1.500	4.481	13	1.659	5.255
4	1.815	6.143	14	1.430	4.177
5	1.571	4.811	15	1.546	4.693
6	1.536	4.646	16	1.583	4.869
7	1.678	5.354	17	1.950	7.029
8	1.533	4.634	18	1.704	5.499
9	1.808	6.100	19	1.824	6.195
10	1.279	3.593	20	2.126	8.378
	$S^2_{nor} = 0.0775$			$S^2_{log} = 2.0284$	
	$\bar{X}_{nor} = 1.5753$			$\bar{X}_{log} = 5.0233$	

De manera análoga a las anteriores distribuciones, las estimaciones sobre la media y la varianza de la variable normal, como de su transformación exponencial, poseen valores cercanos a los cálculos realizados con los parámetros elegidos.

La tabla siguiente muestra el resultado de subdividir el dominio de los valores simulados en $k = 20$ subintervalos, análogamente a las distribuciones anteriores; además se muestra el máximo y el mínimo de la muestra. La Gráfica 1.46 muestra las frecuencias relativas de la muestra, con la que se puede realizar una comparación, al menos en cuanto a su forma, contra la curva de la Gráfica 1.46.

Tabla 1.37: Frecuencias de los datos simulados de una Log-Normal (μ=1.57;σ^2=0.07)					
Rango	***Frecuencia***	***Frecuencia Relativa***	***Rango***	***Frecuencia***	***Frecuencia Relativa***
(0,2.12]	7	0.0014	(6.91,7.45]	204	0.0408
(2.12,2.65]	56	0.0112	(7.45,7.98]	126	0.0252
(2.65,3.18]	275	0.055	(7.98,8.51]	74	0.0148
(3.18,3.71]	525	0.105	(8.51,9.04]	42	0.0084
(3.71,4.25]	774	0.1548	(9.04,9.58]	35	0.007
(4.25,4.78]	785	0.157	(9.58,10.11]	17	0.0034
(4.78,5.31]	719	0.1438	(10.11,10.64]	6	0.0012
(5.31,5.85]	611	0.1222	(10.64,11.18]	1	0.0002
(5.85,6.38]	444	0.0888	(11.18,11.71]	4	0.0
(6.38,6.91]	293	0.0586	(11.71,∞)	2	0.0004
		Mínimo = **1.58**		***Máximo =***	**12.24**

Gráfica 1.46: Frecuencias Relativas de los valores simulados de una distribución lognormal(1.57,0.07)



La verificación formal de que los valores simulados se comportan de manera acorde a una distribución Log-Normal se realizará por medio de la prueba de bondad de ajuste Kolmogorov-Smirnov, la cual contrasta las siguientes hipótesis

$$\boldsymbol{H_0: F_0(x) = F(x) \quad \forall x}$$

$$\boldsymbol{H_1: F_0(x) \neq F(x) \; para \; alguna \; x}$$

Donde $F_0(x)$ es la función de distribución de la variable Log-Normal(1.57,0.07), de la que se cree los datos provienen y $F(x)$ es la función de distribución de los datos observados.

La función de distribución F_0 se puede obtener por medio de la función de Excel® $DISTR.LOG.NORM()$, que recibe como parámetros los valores x a evaluar y los parámetros de la Normal utilizada, su media y desviación estándar, es decir, en este caso los argumentos de la función son $\left(x_{(i)}, 1.57, \sqrt{0.07}\right)$.

El estadístico de prueba D_n y el valor crítico necesarios para la realización de la prueba considerando un nivel significancia $\alpha = 0.05$, junto con una muestra parcial de los valores ordenados se hallan en la Tabla 1.38.

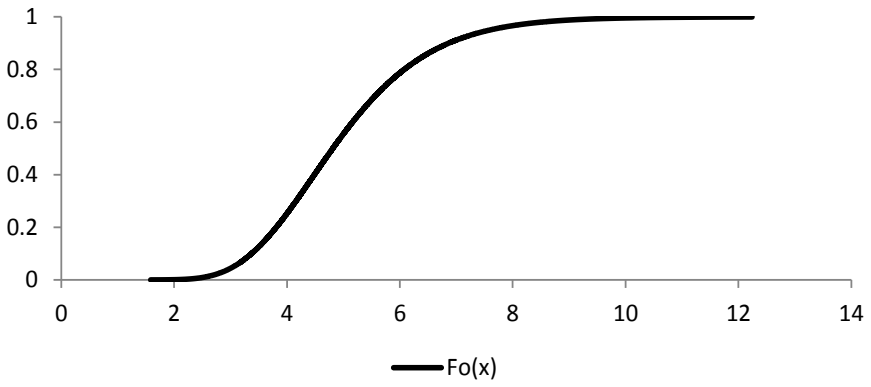
Tabla 1.38: Muestra parcial de la prueba K-S sobre los valores simulados de una Log-Normal(μ=1.57;σ²=0.07)

i	$x_{(i)}$	i/n	$F_0(x_{(i)})$	$i/n - F_0(x_{(i)})$	$F_0(x_{(i)}) - (i-1)/n$
1	1.58256	0.00003	0.0002	0.00017	0.00003
2	1.69478	0.00008	0.0004	0.00032	-0.00012
3	1.84245	0.00027	0.0006	0.00033	-0.00013
4	2.01238	0.00084	0.0008	-0.00004	0.00024
5	2.04862	0.00104	0.001	-0.00004	0.00024
...	...	...	...	...	...
4996	11.49977	0.99916	0.9992	0.00004	0.00016
4997	11.67974	0.99931	0.9994	0.00009	0.00011
4998	11.69174	0.99931	0.9996	0.00029	-0.00009
4999	12.03003	0.99952	0.9998	0.00028	-0.00008
5000	12.24261	0.99962	1	0.00038	-0.00018
		$1.36/\sqrt{n} = 0.01923$		$D_n = 0.01287$	
		$D_n^- = 0.01287$		$D_n^+ = 0.00425$	

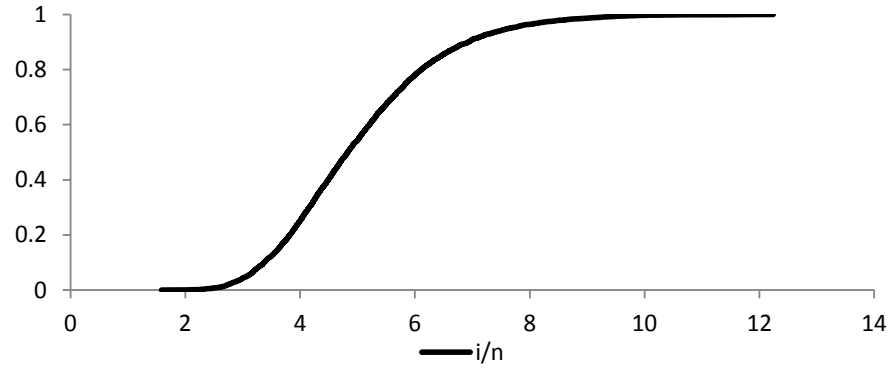
Para el caso de los valores anteriormente mostrados se puede notar que $D_n > 1.36/\sqrt{n}$ que de acuerdo a la regla de decisión de la prueba, implica en no rechazar la hipótesis nula. Entonces se concluye que no existen diferencias significativas entre la función distribución de los valores simulados y la función de distribución de una variable Log-Normal(1.57,0.07).

La forma visual de explicar porqué no se rechazó la hipótesis nula se puede realizar comparando las siguientes graficas que muestran los estadísticos de orden $x_{(i)}$ primero contra la función $F_0(x_{(i)})$ y luego contra los valores de la distribución empírica de los datos, donde no es posible notar diferencias, incluso al sobreponerlas una sobre la otra.

Gráfica 1.47: Función de distribución Log-Normal valuada sobre los estadísticos de orden



Gráfica 1.48 : Función de distribución empírica de los datos simulados de una lognormal(1.57,0.07)



Pareto

Introducción

Vilfredo Pareto (1848-1923) economista italiano utilizó una distribución desarrollada por él y la cual llevaría su nombre, para estudiar los ingresos de una población. Esta distribución estuvo basada en la siguiente ecuación:

$$\boldsymbol{N = Ax^{-a}}$$

Donde N es el número de personas con ganancias mayores o iguales a x, A y a parámetros; la distribución Pareto tiene la propiedad de tener una "cola pesada" en comparación con otras distribuciones, esto quiere decir que la convergencia de la densidad hacia 0 cuando los valores de x son cada vez mayores, es más lenta que para otras distribuciones. Dentro de las aplicaciones, tiene gran importancia la convergencia, pues implica asignar probabilidades considerables para rangos de valores grandes, por lo cual ha sido usada en estudios de la distribución de una población, sobre recursos naturales, fluctuaciones de precios, tamaño de empresas, y en astronomía para estudiar el brillo de los cometas.

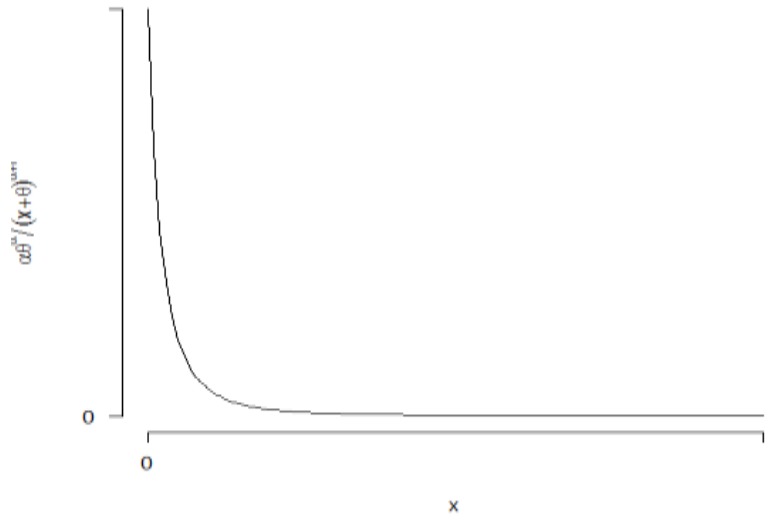
Características Principales

Esta distribución es parte de una familia de variables aleatorias conocida como la familia de la Beta transformada, de la cual existen diferentes distribuciones definidas de acuerdo a sus parámetros y función de densidad, clasificadas en "tipos". La forma en que se define en este trabajo es acorde al libro Loss Models para ser congruentes con aplicaciones actuariales, la cual coincide con ser de tipo II. La función de densidad de esta distribución es

$$\boldsymbol{f(x) = \frac{\alpha\theta^{\alpha}}{(x+\theta)^{\alpha+1}} \quad x > 0}$$

Donde α y θ son parámetros que deben ser positivos. La curva que forma esta función se puede ver en la Gráfica 1.49.

Gráfica 1.49: Densidad de una distribución Pareto(α,θ)

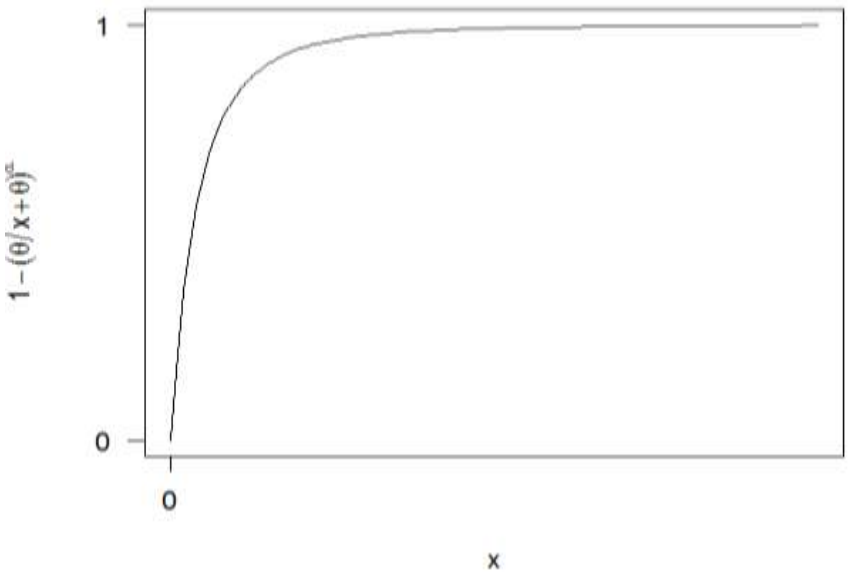


La función de distribución de esta variable es expresada de la siguiente manera

$$\boldsymbol{F(x) = 1 - \left(\frac{\theta}{x+\theta}\right)^{\alpha}}$$

Y para apreciar la forma que toma puede observarse a Gráfica 1.50

Gráfica 1.50: Función de distribución de una Pareto (α,θ)



Tomando estas dos funciones se puede obtener la función de riesgo, la cual indica otra propiedad de esta distribución que es tener una intensidad decreciente sobre la ocurrencia de valores en su dominio, es decir los valores de tamaño cada vez mayor, se consideran probables aunque suceden con menor intensidad.

$$\boldsymbol{h(x) = \frac{\alpha\theta^{\alpha}}{(x+\theta)^{\alpha+1}} \; / \; \left(\frac{\theta}{x+\theta}\right)^{\alpha} = \frac{\alpha}{x+\theta}}$$

El parámetro θ tiene como restricción $\theta > 0$. Para que la media de la distribución exista, se requiere que $\alpha > 1$. Al cumplir esta restricción la media se calcula como

$$\boldsymbol{E(X) = \frac{\theta}{\alpha - 1}}$$

En el caso de la varianza se necesita que $\alpha > 2$ para que exista. Si se cumple lo anterior la varianza se calcula de la siguiente forma

$$\boldsymbol{Var(X) = \frac{2\theta^2}{(\alpha - 2)(\alpha - 1)} - \left(\frac{\theta}{\alpha - 1}\right)^2}$$

Simulación de valores

Para la simulación de esta variable aleatoria la opción más sencilla es emplear el método de inversión, igualando $F(x)$ a un valor U uniforme y despejando el valor de x

$$\boldsymbol{1 - \left(\frac{\theta}{x + \theta}\right)^{\alpha} = U}$$

$$\boldsymbol{U = 1 - U = \left(\frac{\theta}{x + \theta}\right)^{\alpha}}$$

$$\boldsymbol{U^{1/\alpha} = \frac{\theta}{x + \theta}}$$

$$\boldsymbol{U^{\frac{1}{\alpha}}(x + \theta) = \theta}$$

$$\boldsymbol{U^{\frac{1}{\alpha}}x = \theta - \theta U^{\frac{1}{\alpha}}}$$

$$\boldsymbol{x = \theta U^{-\frac{1}{\alpha}} - \theta}$$

$$\boldsymbol{x = \theta\left(U^{-\frac{1}{\alpha}} - 1\right)}$$

El algoritmo resumido consta de los siguientes pasos

1. $\boldsymbol{Generar\ U \sim U(0, 1)}$
2. $\boldsymbol{Regresar\ X = \theta\left(U^{-\frac{1}{\alpha}} - 1\right)}$

Para demostrar este algoritmo se generaron por medio de un código en VBA, 5000 valores siguiendo el algoritmo anterior con los parámetros $\alpha = 3$ Y $\theta = 1000$, entonces su media y su varianza teóricas son 500 y $500{,}000$ respectivamente, la Tabla 1.39 muestra parcialmente los datos generados junto con estimaciones de la media $\bar{X}$ y la varianza S^2 calculados sobre toda la muestra.

Tabla 1.39: Muestra parcial de valores simulados de una Pareto (α=3;θ=1,000)					
i	$U_i \sim U(0,1)$	$x_i = \theta\left(U_i^{-\frac{1}{\alpha}} - 1\right)$	i	$U_i \sim U(0,1)$	$x_i = \theta\left(U_i^{-\frac{1}{\alpha}} - 1\right)$
1	0.594	189.426	11	0.773	89.796
2	0.418	337.778	12	0.388	371.061
3	0.826	65.763	13	0.848	56.514
4	0.424	331.237	14	0.069	1436.280
5	0.024	2451.491	15	0.506	255.294
6	0.809	73.368	16	0.561	212.818
7	0.981	6.279	17	0.319	463.422
8	0.542	226.307	18	0.840	59.826
9	0.198	717.070	19	0.232	627.651
10	0.417	338.969	20	0.204	699.108
	$\bar{X} = 499.004$			$S^2 = 536.892.4$	

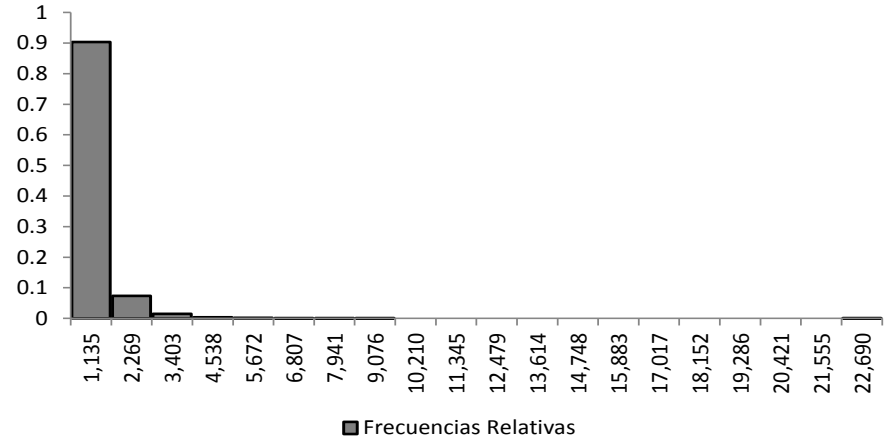
Debido a la propiedad de cola pesada de la distribución Pareto existe un sesgo marcado entre la estimación de la varianza y el valor computado en base a los parámetros, análogamente a la manera en que se ha trabajado, se realizará una tabla de frecuencias para observar la distribución de los datos.

La Tabla 1.40 contiene las frecuencias absolutas y relativas al considerar el número de subintervalos de igual longitud $k = 20$.

Tabla 1.40: frecuencias de los datos simulados de una Pareto(α=3;θ=1,000)					
Rango	*Frecuencia*	*Frecuencia Relativa*	*Rango*	*Frecuencia*	*Frecuencia Relativa*
(0,134.5]	4516	0.9032	(11344.8,12479.3]	0	0
(134.5 , 2268.9]	368	0.0736	(12479.3,13613.7]	0	0
(2268.9 , 3403.4]	76	0.0152	(13613.7,14748.2]	0	0
(3403.4,4537.9]	19	0.0038	(14748.2,15882.7]	0	0
(4537.9,5672.4]	11	0.0022	(15882.7,17017.2]	0	0
(5672.4,6806.9]	6	0.0012	(17017.2,18151.7]	0	0
(6806.9,7941.3]	2	0.0004	(18151.7,19286.1]	0	0
(7941.3,9075.8]	1	0.0002	(19286.1,20420.6]	0	0
(9075.8,10210.3]	0	0	(20420.6,21555.1]	0	0.0
(10210.3,11344.8]	0	0	(21555.1,∞)	1	0.0002
		Mínimo = 0.028		*Máximo* = 22689.620	

La siguiente gráfica muestra las frecuencias relativas de la muestra, en la cual se observa su similitud en forma con la densidad de la Pareto del Gráfico 1.49.

Gráfica 1.51: Frecuencias relativas de los valores simulados de una Pareto(3;1000)



Para concluir que los valores simulados provienen de una variable aleatoria Pareto, se aplicará la prueba de bondad de ajuste de Kolmogorov-Smirnov. Se parte entonces de realizar el contraste entre las siguientes hipótesis

$$\boldsymbol{H_0}: \boldsymbol{F_0(x) = F(x)} \quad \forall \boldsymbol{x}$$

$$\boldsymbol{H_1}: \boldsymbol{F_0(x) \neq F(x)\ para\ alguna\ x}$$

Donde F es la función de distribución de la muestra y F_0 es una funcion de distribución de una variable Pareto$(\alpha = 3, \theta = 1000)$.

El resultado de hacer los calculos necesarios para realizar la prueba de hipótesis, incluyendo el valor critico aproximado para un nivel de significancia del 5% el cual es $1.36/\sqrt{n}$, se muestra en la Tabla 1.41.

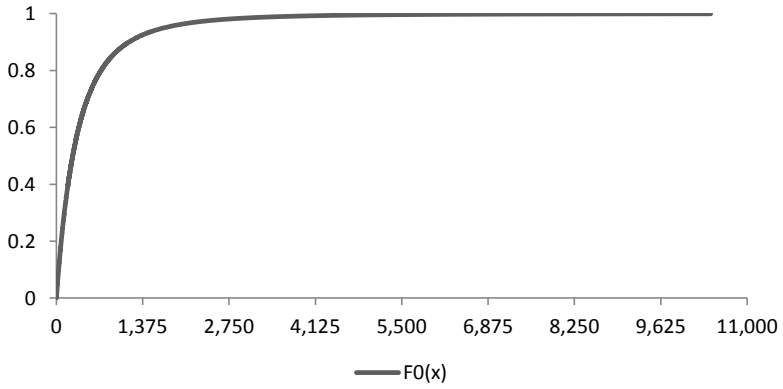
Tabla 1.41: muestra parcial de la prueba K-S para valores simulados de una Pareto (α=3;θ=1,000)

i	$Pareto(3;1000)$	$F_0(x_{(i)})$	i/n	$i/n - F_0(x_{(i)})$	$F_0(x_{(i)}) - (i-1)/n$
1	0.013	0.00004	0.0002	0.00016	0.00004
2	0.03863	0.00012	0.0004	0.00028	-0.00008
3	0.07335	0.00022	0.0006	0.00038	-0.00018
4	0.10717	0.00032	0.0008	0.00048	-0.00028
5	0.35179	0.00105	0.001	-0.00005	0.00025
...	...	...	...	...	...
4996	7541.07	0.9984	0.9992	0.0008	-0.0006
4997	9394.37	0.99911	0.9994	0.00029	-0.00009
4998	9605.53	0.99916	0.9996	0.00044	-0.00024
4999	10293.43	0.99931	0.9998	0.00049	-0.00029
5000	10419.12	0.99933	1	0.00067	-0.00047
		$1.36/\sqrt{n} = 0.01923$		$D_n = 0.00884$	
		$D_n^- = 0.00884$		$D_n^+ = 0.00641$	

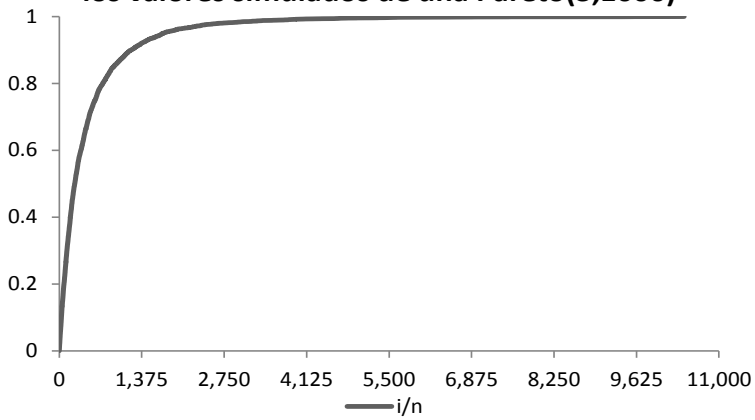
Los valores obtenidos demuestran que $D_n < 1.36/\sqrt{n}$, por lo tanto no se rechaza la hipótesis nula, por lo que se concluye con un nivel de significancia del 5%, que no existen evidencias estadísticas que muestren una diferencia significativa entre la función de distribución de los valores simulados y la función de distribución de una Pareto(3;1000).

Las siguientes gráficas muestran los estadísticos de orden contra los valores de $F_0(x_{(i)})$ y contra los valores i/n, donde se observa similitud entre ambas y explica, como en las distribuciones anteriores, porqué no se rechazó la hipotesis nula.

Gráfica 1.52: Función de distribución de una Pareto(3;1000) valuada sobre las simulaciones



Gráfica 1.53: Función de distribución empírica de los valores simulados de una Pareto(3,1000)



Beta

La distribución Beta en su forma estándar está definida para valores en el intervalo $(0,1)$, por lo cual, es de empleada para modelar fenómenos relacionados con proporciones de las cuales los datos demuestran, tener mayor (menor) recurrencia cerca de ciertos puntos del intervalo $(0,1)$.

La distribución Beta, en el contexto de modelar proporciones, se le puede hallar mencionada en estudios climáticos, sobre la proporción de luz de sol en el día que recibe una ciudad, o por el contrario la duración que las nubes obstruyen al sol, que es cuando una ciudad recibe menor luz solar.

A diferencia de la distribución Uniforme en (0,1), cuya densidad es un valor constante, la curva de la función de densidad de la distribución Beta definida en el intervalo (0,1), está determinada por dos parámetros α_1, α_2, cuyos valores determinan la forma de la curva de la función de densidad, la cual puede tener su valor máximo en alguna parte del intervalo, y por lo tanto modele eventos de los cuales se tenga un supuesto sobre la probabilidad de ocurrencia.

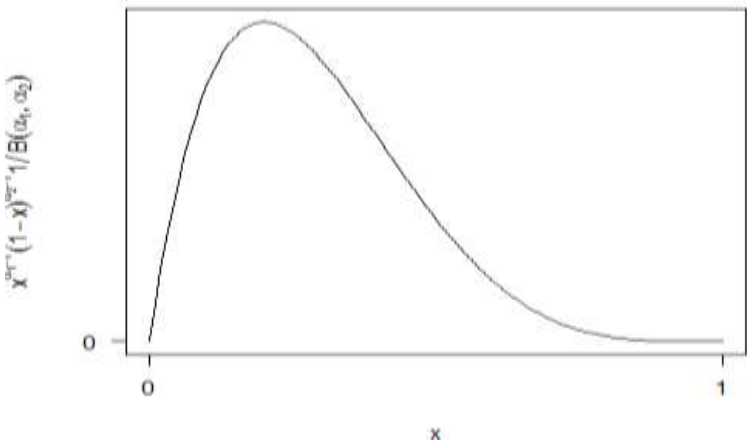
Características principales

La densidad de esta distribución está dada por la siguiente función

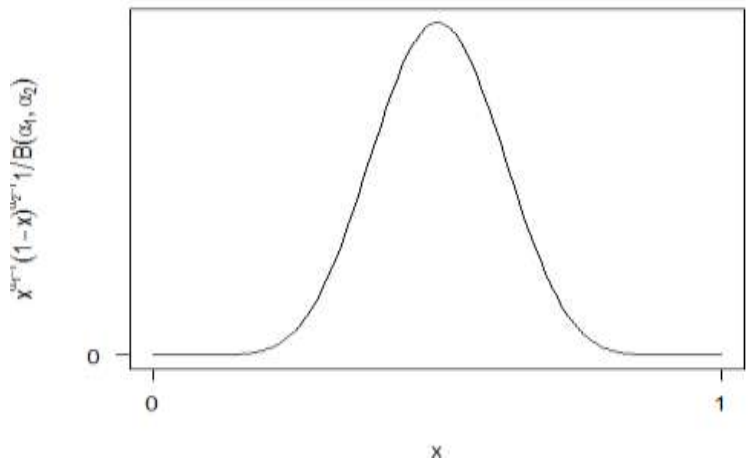
$$\boldsymbol{f(x) = \frac{x^{\alpha_1 - 1}(1-x)^{\alpha_2 - 1}}{B(\alpha_1, \alpha_2)} \qquad x \geq 0;\ \alpha_1, \alpha_2 > 0}$$

Donde $B(a,b)$ es la función Beta que se expresa como $\Gamma(a)\Gamma(b)/\Gamma(a+b)$. Los parámetros α_1 y α_2 inciden de manera muy sensible sobre la forma de la distribución por lo que en las siguientes gráficas se presentan tres casos que pueden ocurrir al elegir los valores de los parámetros.

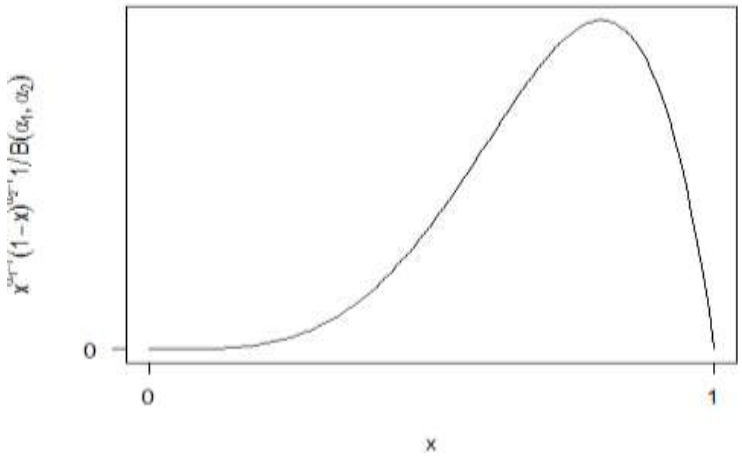
Gráfica 1.54: Densidad de una Beta($\alpha_1 < \alpha_2$)



Gráfica 1.55: Densidad de una Beta $\alpha_1 = \alpha_2$

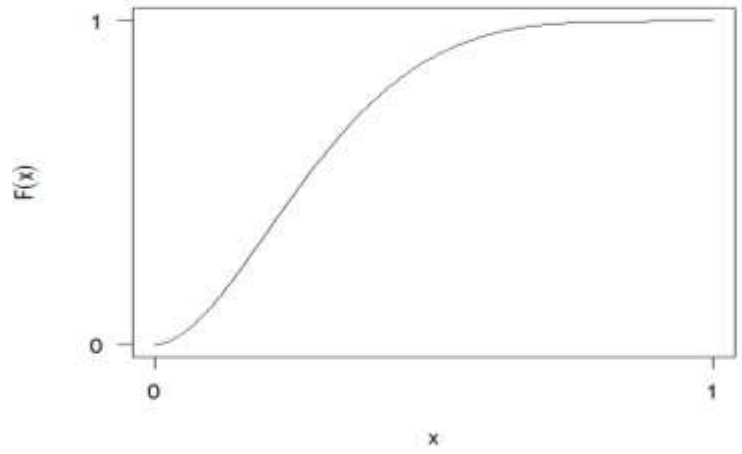


Gráfica 1.56: Densidad de una Beta $\alpha_1 > \alpha_2$

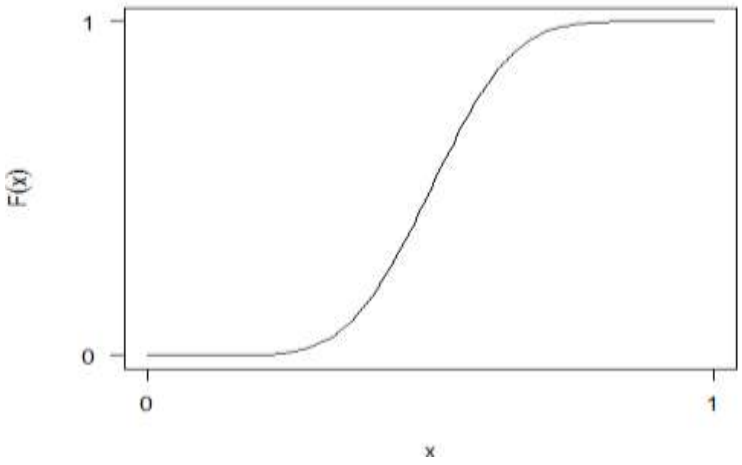


La función de distribución de esta variable aleatoria, se define como $F(x) = \int_{-\infty}^{x} f(u)du$, la cual no tiene una solución analítica, cuando sus parámetros son distintos de 1, por lo que es necesario recurrir a métodos numéricos para su evaluación, las gráficas siguientes muestran la forma de la función de distribución para los tres casos anteriores de acuerdo a la elección de los valores de los parámetros.

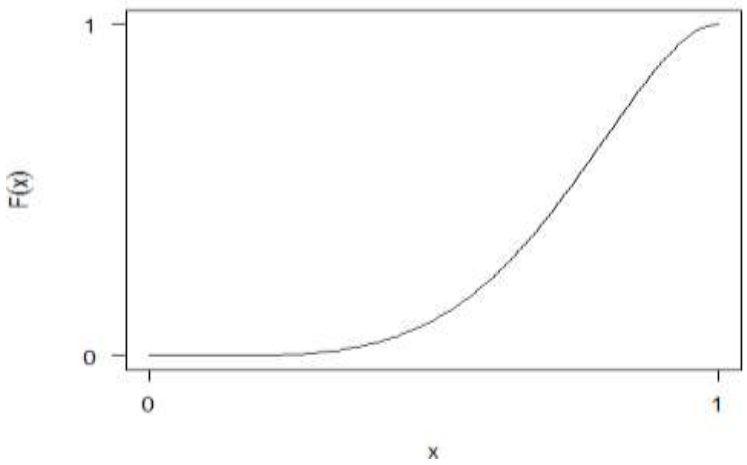
Gráfica 1.57: Distribución de una Beta $\alpha_1 < \alpha_2$



Gráfica 1.58: Distribución de una Beta $\alpha_1 = \alpha_2$



Gráfica 1.59: Distribución de una Beta $\alpha_1 > \alpha_2$



Ahora toca estudiar una función de interés en la modelación de un fenómeno, y es la función de riesgo, definida como $h(x) = \frac{f(x)}{1-F(x)}$. Para darse una idea del comportamiento de esta función, se puede recurrir a la regla de L'Hôpital ya que tanto el denominador y el numerador tienden a cero conforme x tiende a 1. Entonces se puede obtener el resultado con el siguiente límite:

$$\lim_{x\to 1} \boldsymbol{h(x) = \frac{f(x)}{1-F(x)} = \lim_{x\to 1} \frac{f'(x)}{-f(x)}}$$

$$\boldsymbol{si\ f'(x) = \frac{(\alpha_1 - 1)x^{\alpha_1 - 2}(1-x)^{\alpha_2 - 1} - (\alpha_2 - 1)x^{\alpha_1 - 1}(1-x)^{\alpha_2 - 2}}{B(\alpha_1, \alpha_2)}}$$

$$\boldsymbol{entonces\ \lim_{x\to 1} \frac{f'(x)}{-f(x)} = lim_{x\to 1} \frac{(\alpha_2 - 1)x^{\alpha_1 - 1}(1-x)^{\alpha_2 - 2} - (\alpha_1 - 1)x^{\alpha_1 - 2}(1-x)^{\alpha_2 - 1}}{-x^{\alpha_1 - 1}(1-x)^{\alpha_2 - 1}}}$$

$$\boldsymbol{= \lim_{x\to 1} (1-x)^{-1}(\alpha_2 - 1) - x^{-1}(\alpha_1 - 1) = \infty}$$

Se concluye entonces, que la función de riesgo tiende a infinito, pues el límite anterior no depende de los parámetros α_1, α_2.

Para caracterizar esta variable también es importante mostrar sus primeros momentos centrales los cuales son:

$$\boldsymbol{E(X) = \frac{\alpha_1}{\alpha_1 + \alpha_2}}$$

$$\boldsymbol{Var(X) = \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)}}$$

Simulación de valores

Para su generación, es necesario auxiliarse de su relación con otras distribuciones y otras propiedades. La primer propiedad importante dice que para una variable X, Beta(α_1, α_2), $\alpha_1, \alpha_2 > 0$ definida en (0,1) la transformación $Y = a + (b-a)X$ es Beta con los mismos parámetros pero definida en el intervalo (a, b), también conocida como Beta de cuatro parámetros, así que basta con enfocarse en generar variables definidas

en (0,1). Por otra parte $1-X$ se distribuye Beta(α_2, α_1), además, en el caso que algunos de sus parámetros sea 1, su función de densidad es

$$\boldsymbol{f(x) = \alpha x^{\alpha-1}}$$

La cual se puede generar por el método de inversión. Para el caso en que se quisiera simular una variable Beta de parámetros cualesquiera α_1, $\alpha_2 > 0$ un resultado útil es la siguiente:

Supóngase un proceso Poisson con una tasa de llegada de 1 por unidad de tiempo, además sea Y_1 el tiempo de espera hasta que el α_1-ésimio evento ocurra y Y_2 el tiempo hasta que el α_2-ésimo evento ocurra, se sabe que $Y_1 \sim Gamma(\alpha_1, 1)$, $Y_2 \sim Gamma(\alpha_2, 1)$, entonces $X = \frac{Y_1}{Y_1+Y_2}$ se distribuye Beta(α_1, α_2).

Basándose en el resultado anterior, el cual se puede extender para parámetros continuos, el algoritmo es

1. $\boldsymbol{Generar\ Y_1 \sim Gamma(\alpha_1, 1), Y_2 \sim Gamma(\alpha_2, 1)\ independientes}$
2. $\boldsymbol{Regresar\ X = \frac{Y_1}{Y_1+Y_2}}$

Para mostrar la eficacia del algoritmo, por medio de un código en VBA[7] se han generado 5000 valores simulados, con los parámetros $\alpha_1 = 2, \alpha_2 = 3$, eso implica que, según las formulas vistas anteriormente la media teóricamente debe ser $E(X) = \frac{2}{5} = 0.4$ y con una varianza $VAR(X) = \frac{1}{25} = 0.04$. La Tabla 1.41 contiene una muestra parcial de los valores simulados junto con la estimación de la media $\bar{X}$ y de la varianza S^2.

Tabla 1.42: Muestra parcial de valores simulados de una variable Beta(α_1=2;α_2=3)

i	$Beta(2;3)$	i	$Beta(2;3)$
1	0.48889	6	0.58486
2	0.73357	7	0.19975
3	0.12638	8	0.67641
4	0.65812	9	0.36585
5	0.11835	10	0.38561
$\bar{X} = 0.40011$		$S^2 = 0.03973$	

Como se puede esperar, las estimaciones son cercanas a los valores calculados con base en los parámetros. A continuación, como se ha estado trabajando, se observará el comportamiento general de los datos con una tabla de frecuencias absolutas y relativas, construida de igual forma que para las distribuciones anteriores.

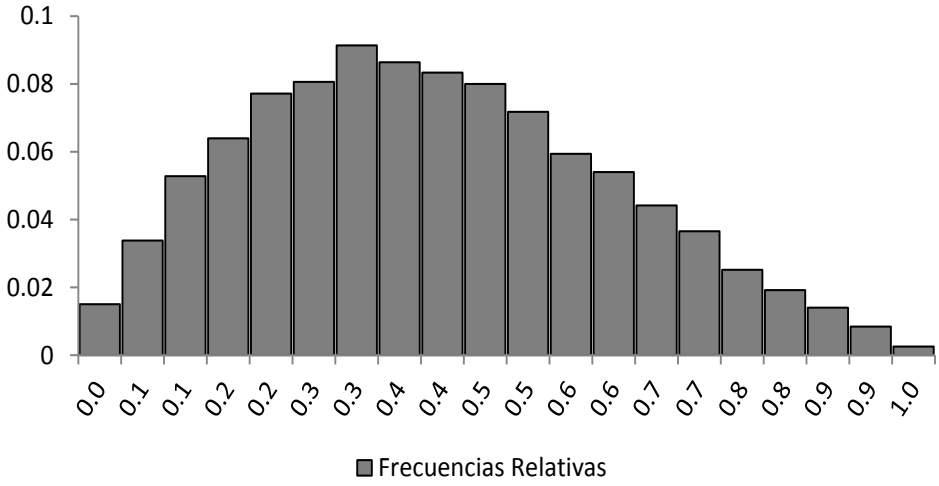
Los cálculos mencionados sobre los datos simulados de la variable Beta, al considerar una división de $k = 20$ subintervalos del dominio de la variable, se hallan en la Tabla 1.43.

[7] Consulte el código en la parte II del apéndice.

Tabla 1.43: Frecuencias de los valores simulados, divididos en 20 subintervalos					
Rango	*Frecuencia*	*Frecuencia Relativa*	*Rango*	*Frecuencia*	*Frecuencia Relativa*
(0,0.050]	0.015	75	(0.481,0.529]	0.07	291
(0.050,0.97]	0.034	169	(0.529,0.576]	0.06	356
(0.97,0.145]	0.053	264	(0.576,0.624]	0.05	240
(0.145,0.193]	0.064	320	(0.624,0.672]	0.04	274
(0.193,0.241]	0.077	386	(0.672,0.720]	0.04	136
(0.241,0.289]	0.081	403	(0.720,0.768]	0.03	183
(0.289,0.337]	0.091	457	(0.768,0.816]	0.02	78
(0.337,0.385]	0.086	432	(0.816,0.864]	0.01	79
(0.385,0.433]	0.083	417	(0.864,0.912]	0.01	26.0
(0.433,0.481]	0.08	400	(0.912,1)	0.003	9
		Mínimo = 0.0017		*Máximo* = 0.9595	

La gráfica 1.60 muestra las frecuencias relativas de la muestra, la cual es comparable a la función de densidad de la gráfica 1.54, de acuerdo a los parámetros elegidos en las simulaciones con $\alpha_1 < \alpha_2$.

Gráfica 1.60: Frecuencias relativas de los valores simulados de una Beta(2,3)



Análogamente al estudio de las distribuciones anteriores para comprobar que los valores simulados corresponden a los de una distribución Beta(2;3), se empleará la prueba de bondad de ajuste de Kolmogorov-Smirnov , la cual contrasta las siguientes hipótesis

$$\boldsymbol{H_0}: \boldsymbol{F_0(x) = F(x)} \quad \forall \boldsymbol{x}$$

$$\boldsymbol{H_1}: \boldsymbol{F_0(x) \neq F(x)\ para\ alguna\ x}$$

Donde $F_0(x)$ es una función de distribución de la variable Beta(2;3), y $F(x)$ es la función de distribución de los datos, la cual es aproximada por medio de su función de distribución empírica.

La función de apoyo de Excel® para realizar los cálculos de la distribución Beta(2;3) es $dist.beta()$. La Tabla 1.44 muestra de manera parcial los cálculos individuales realizados para esta prueba junto con los estadísticos de prueba D_n, D_n^+ y D_n^-, además al final de la tabla se halla el valor crítico aproximado de la prueba para un nivel de significancia del 5% dado por $1.36/\sqrt{n}$.

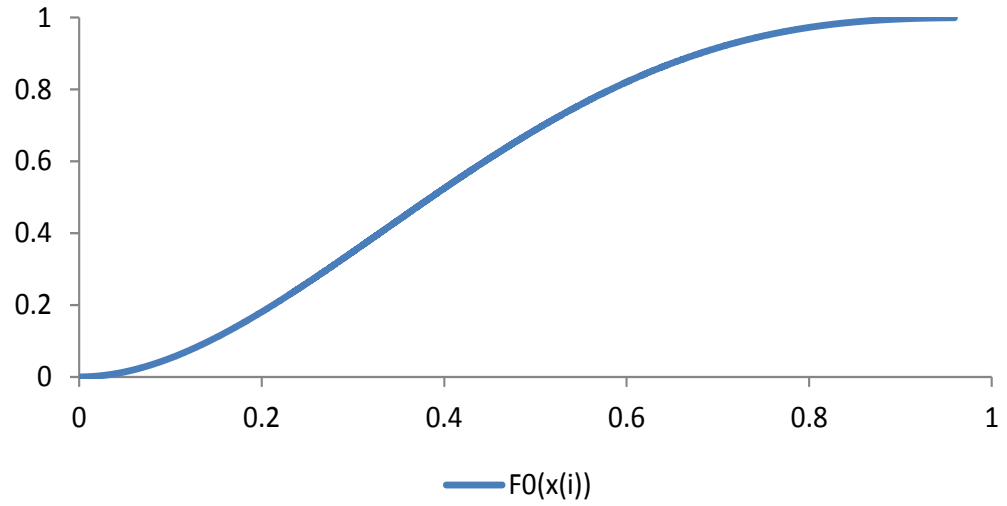
Tabla 1.44: Muestra parcial de los cálculos realizados para la prueba K-S

i	$x_{(i)}$	$F_0(x_{(i)})$	i/n	$i/n - F_0(x_{(i)})$	$F_0(x_{(i)}) - (i-1)/n$
1	0.01136	0.00076	0.0002	-0.00056	0.00076
2	0.01141	0.00077	0.0004	-0.00037	0.00057
3	0.0125	0.00092	0.0006	-0.00032	0.00052
4	0.01488	0.0013	0.0008	-0.0005	0.0007
5	0.01872	0.00205	0.001	-0.00105	0.00125
...	...	...	...	...	...
4996	0.92945	0.99867	0.9992	0.00053	-0.00033
4997	0.93035	0.99872	0.9994	0.00068	-0.00048
4998	0.93313	0.99886	0.9996	0.00074	-0.00054
4999	0.93586	0.999	0.9998	0.0008	-0.0006
5000	0.96676	0.99986	1	0.00014	0.00006
		$1.36/\sqrt{n} = 0.0192$		$D_n = 0.0098$	
		$D_n^- = 0.0098$		$D_n^+ = 0.0093$	

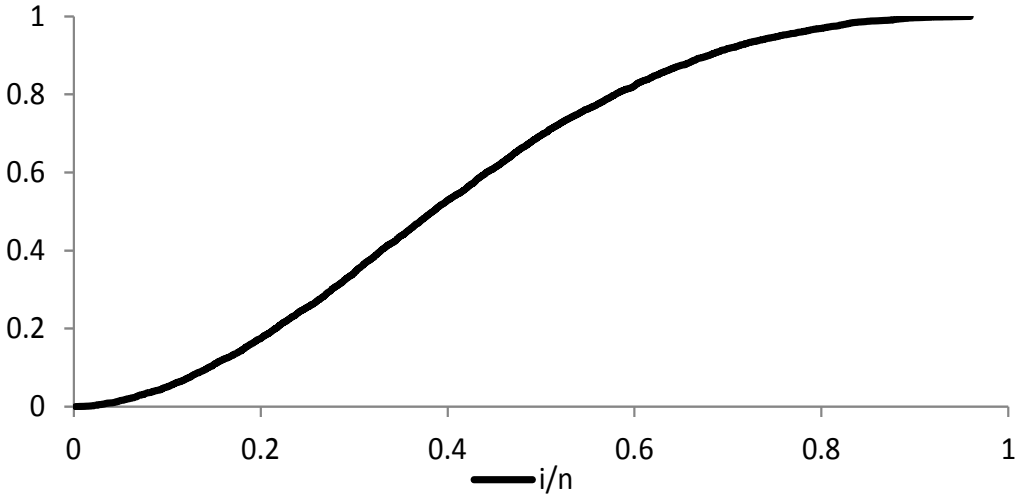
En la tabla se observa que $\frac{1.36}{\sqrt{n}} > D_n$, implicando no rechazar la hipótesis H_0, por lo tanto se concluye que no existe una diferencia significativa entre la función de distribución de los valores simulados y la función de distribución de una variable Beta(2;3).

Identificar que no existen diferencias significativas entre las funciones de distribución, se puede ver en las siguientes gráficas que muestran la función de distribución empírica de los datos y la función de distribución de una Beta(2;3), ambas valuadas en los estadísticos de orden de la muestra.

Gráfica 1.61: Función de distribución de una Beta(2;3) evaluada sobre los valores simulados



Gráfica 1.62: Función de distribución empírica de los valores simuladosdeuna distribución Beta(2,3)



Capítulo II:

Enseñanza de métodos descriptivos a través del enfoque de simulación Monte Carlo.

El desarrollo de algunos cursos académicos dedicados al aprendizaje de tópicos de economía, finanzas, ingeniería, mercadotecnia, física, ciencias políticas y sociales, entre otros, requieren revisar ciertos temas relacionados con la aplicación de técnicas estadísticas. Estás técnicas describen un conjunto de métodos matemáticos, los cuales se emplean para el análisis de información obtenida a través de la observación de un fenómeno que ocurre con cierta incertidumbre.

Con frecuencia, la compresión de las técnicas estadísticas, tanto en su aplicación como en la interpretación de resultados, es difícil para el estudiante, dependiendo de su formación matemática previa. La formación matemática previa a la realización de un curso estadístico, por lo regular difiere dentro de cada plan de estudios de diversas carreras universitarias, lo que lleva a la motivación de proponer un apoyo de fácil acceso al profesor, por medio de técnicas de simulación, para el desarrollo de algunos tópicos específicos de un curso estadístico introductorio.

Estadística descriptiva

Al iniciar un curso cada estudiante debe tener claro la diferencia entre la estadística descriptiva, la cual tiene como objetivo medir ciertas características de un fenómeno, a partir de la obtención de una muestra, y con base en esta muestra calcular indicadores de los cuales se obtenga una descripción general del fenómeno y su comportamiento. Mientras que la estadística inferencial se refiere al proceso de inducción lógica que supone un cierto comportamiento probabilístico a un fenómeno y que a partir de una muestra obtenida prueba la verosimilitud de los supuestos.

Para introducir al estudiante dentro de la aplicación de un análisis descriptivo, en muchos casos, no es costumbre tocar los temas relacionados a la obtención de la información, sin embargo, una lectura sobre el tema ayudará a entender la importancia que la calidad de la información tiene dentro de aplicaciones estadísticas. Aunque también es posible introducir como variable didáctica un proyecto de obtención de datos, ya sea con el desarrollo de encuestas o cuestionarios en su ambiente local.

Hoy en día es posible conseguir información oficial y de calidad de diversas fuentes como anuarios estadísticos, datos publicados de censos, encuestas o registros administrativos y en diversas páginas web de instituciones nacionales de estadística de diversos países, que como se verá más adelante, estas fuentes permiten obtener cierta información desagregada, es decir, cada uno de los datos individuales con sus diferentes variables , donde al agregado de datos finales recabados se le conoce como muestra aleatoria, y cuya estructura dependerá del tipo de fenómeno de estudio.

La alternativa aquí expuesta es tratar de generar datos con ciertas características similares a la información práctica para ejemplificar la forma de cálculo y además observar el impacto en el valor de los indicadores bajo diversos escenarios, empleando técnicas de simulación Monte Carlo. A continuación se revisarán conceptos de importancia dentro de un curso estadístico, además del uso de diversos ejemplos prácticos para su enseñanza en un aula de clases, donde se destaca el empleo de técnicas de simulación Monte Carlo.

Concepto de población

Para tener claridad en un análisis que se desea realizar, es necesario definir a los individuos que se considerarán a observar y las características que se desean registrar. A este conjunto de elementos se denomina población objetivo o simplemente población, donde el número total de individuos se representa comúnmente con la letra N.

Se puede tomar como contexto inicial el interés de las ciencias sociales y los gobiernos actuales, en estudiar el crecimiento de su población, para fundamentar en datos duros el desarrollo de políticas públicas. Por ejemplo, si el objetivo fuera realizar un análisis sobre el comportamiento del número de nacimientos de un país como Canadá, para una ventana de tiempo de un año, entonces la población objetivo serían todos aquellos individuos nacidos en Canadá en el año elegido.

La definición de la población objetivo es de gran importancia al inicio de cualquier estudio, debido a que las restricciones en esta definición determinarán las limitaciones al momento de interpretar y concluir, con base en los resultados de una muestra obtenida de la población.

Con el objetivo de mostrar los detalles del concepto anterior y las siguientes herramientas de un análisis descriptivo a través de ejemplos prácticos basados en datos reales, previo al uso de simulación, a continuación se retomará el contexto anterior sobre nacimientos en Canadá. Como fuente de información se obtuvo de la web el reporte de Milan (2013), el cual expone un análisis general de la fertilidad en Canadá, en una ventana de tiempo de 1981 a 2011.

Análisis tabular

El uso de tablas dentro de un análisis descriptivo es de gran importancia, ya que permite mostrar cálculos basados de los datos obtenidos de forma resumida, ordenada y lógica de acuerdo a las variables de interés. Por ejemplo, en el capitulo anterior el principal uso de tablas fue presentar el cálculo de frecuencias de los valores simulados, de forma ordenada.

Otro uso particular de tablas se da cuando se requiere observar el comportamiento de una variable a través del tiempo, el cual es el caso del reporte de *Statistics Canada* pues en un primer resumen de sus datos obtenidos, por medio de una tabla o cuadro estadístico mostrada en la Tabla 2.1, donde resume el total de nacimientos considerando a los renglones como los años en los que se realizó el censo y a las columnas como las 13 diferentes provincias censadas. Otro detalle a resaltar de la información obtenida es, que en este caso particular al tratarse de un censo, la muestra obtenida es la población completa.

Para mostrar la importancia de la definición de la población objetivo a través de un caso en que la definición puede cambiar por eventos externos, se puede considerar el caso de restringir la población objetivo como los niños canadienses nacidos en 1996 dentro de la provincia de los *Northwest Territories (N.W.T.),* entonces de acuerdo a la información de la Tabla 2.1 $N = 1{,}562$. En otro orden de ideas, se observa que en eso mismo año para la provincia *Nunavut (Nvt.)* se tiene un dato representado como vacío o en puntos, conocido comúnmente por el anglicismo *missing,* debido a que esta última provincia fue resultado de una división política de los N.W.T. avalada en 1999.

En consecuencia, para el año 2001 se inicia la diferenciación entre estos dos territorios, al censar el número de nacimientos, lo cual cambia significativamente la extensión que abarca la definición de las poblaciones y esto se refleja en la parte remarcada en rojo, ya que la suma de las poblaciones en 2001 de Nvt. y N.W.T. es $N_{Nvt} = 710$, $N_{NWT} = 613$, $N_{NWT} + N_{Nvt} = 1{,}323$; resultado similar al número de nacimientos en los N.W.T. previo al 2001.

Tabla 2.1 : Nacimientos Totales de las provincias y territorios de Canadá (1981 - 2011)

Parte del catálogo de Statistics Canada no. 91 - 209 - X

	Provincia													
	N.L.	P.E.I.	N.S.	N.B.	Que.	Ont.	Man.	Sask.	Alta.	B.C.	Y.T.	N.W.T.	Nvt.	Canadá
1981	9,120	1,897	12,079	10,503	95,322	122,183	16,073	17,209	42,638	41,474	536	1,302	...	370,336
1986	7,618	1,928	12,358	9,788	84,634	133,882	17,009	17,513	43,744	41,967	483	1,507	...	372,431
1991	7,166	1,885	12,016	9,497	97,311	151,480	17,282	15,304	42,777	45,613	568	1,634	...	402,533
1996	5,747	1,694	10,573	8,176	85,226	140,012	15,478	13,300	37,851	46,138	443	1,562	...	366,200
2001	4,716	1,380	8,909	7,195	73,695	131,710	14,002	12,275	37,619	40,575	344	613	710	333,744
2002	4,651	1,328	8,663	7,046	72,478	128,529	13,888	11,761	38,692	40,065	339	635	726	328,802
2003	4,629	1,417	8,650	7,117	73,905	130,928	13,940	12,038	40,287	40,496	335	701	758	335,202
2004	4,488	1,390	8,734	6,959	74,073	132,553	13,811	11,983	40,780	40,490	365	698	747	337,072
2005	4,501	1,340	8,557	6,892	76,346	133,760	14,145	11,967	42,110	40,827	320	712	699	342,176
2006	4,542	1,413	8,485	7,030	81,938	135,597	14,565	12,288	45,230	41,730	364	687	747	354,617
2007	4,553	1,389	8,868	7,146	84,387	138,436	15,285	13,248	49,028	43,649	355	725	794	367,864
2008	4,898	1,483	9,188	7,402	87,870	140,791	15,485	13,737	50,856	44,276	373	721	805	377,886
2009	4,915	1,457	8,989	7,391	88,868	140,372	15,940	14,243	51,722	44,993	383	711	877	380,863
2010	4,900	1,403	8,879	7,360	88,419	139,611	15,776	14,296	50,847	43,810	382	700	828	377,213
2011	4,478	1,436	8,862	7,124	88,583	140,135	15,620	14,271	51,040	44,129	431	690	837	377,636

Tabla 2.2: Catálogo de provincias canadienses

Abreviatura	Provincia	Abreviatura	Provincia
Alta.	Alberta	Nvt.	Nunavut
B.C.	British Columbia	Ont.	Ontario
Man.	Manitoba	P.E.I.	Prince Edward Islad
N.B.	New Brunswick	Que.	Québec
N.L.	New found land and Labrador	Sask.	Saskatchewan
N.S.	Nova Scotia	Y.T.	Yukon Territory
N.W.T.	Northwest Territories		

Debido al tamaño de la Tabla 2.1, puede ser difícil realizar comparaciones a primera vista entre los conteos observados. Para resolver lo anterior, dentro de un análisis descriptivo, se pueden emplear las facilidades de una hoja de cálculo para extraer una porción de la información y realizar los comparativos deseados.

En el contexto anterior, si lo que se desea comparar son los nacimientos ocurridos en el año 1981, contra los nacimientos al final de las lecturas en 2011, para todas las provincias, se debe calcular entonces la diferencia absoluta entre cada uno de los conteos correspondientes. Para aprovechar el formato de orden de la tabla, se puede realizar la operación como la diferencia entre los valores del renglón 2011 y los valores del renglón 1981.

Esta operación en ambiente de hoja de cálculo se simplifica al copiar la misma operación para las celdas continuas en dirección de los datos. A la comparación dentro de una tabla a través de diferencias, comúnmente se le denota por el símbolo griego Δ. Además otro tema que se debe retomar es la definición para los N.W.T., puesto que, para poder realizar la comparación, debe considerarse la misma definición para los territorios, por lo cual para el 2011 las provincias N.W.T. y Nvt. se agruparon en una sola región sumando sus poblaciones.

La siguiente tabla contiene los datos que se desean comparar y el resultado de las diferencias.

Tabla 2.3 : Diferencia en el crecimientos natural de las provincias canadienses 1981 vs. 2011

	Provincia												
	N.L.	P.E.I.	N.S.	N.B.	Que.	Ont.	Man.	Sask.	Alta.	B.C.	Y.T.	N.W.T.	Canadá
1981	9,120	1,897	12,079	10,503	95,322	122,183	16,073	17,209	42,638	41,474	536	1,302	370,336
2011	4,478	1,436	8,862	7,124	88,583	140,135	15,620	14,271	51,040	44,129	431	1,527	377,636
Δ	-4,642	-461	-3,217	-3,379	-6,739	17,952	-453	-2,938	8,402	2,655	-105	225	7,300

Al observar los resultados del total de Canadá, se puede observar que el crecimiento natural absoluto entre ambos años ha crecido; sin embargo entrando en detalle la tabla explica que este crecimiento se observa para las provincias de Ontario, British Columbia, Altavista y N.W.T. que compensan el decremento en el número de nacimientos de las otras nueve provincias.

En particular se ha revisado como métrica la diferencia absoluta entre los totales, pero existen otras métricas comunes dentro de las aplicaciones estadísticas, que se revisará a detalle más adelante. Antes se revisará una forma para clarificar la información de una tabla por medio de gráficas.

Análisis gráfico

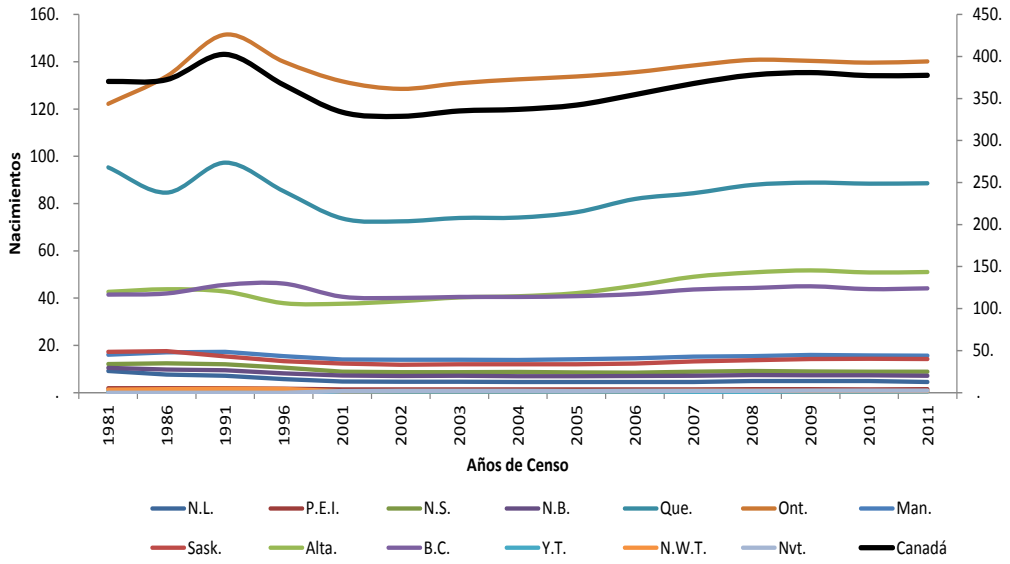
Un recurso más conveniente para mostrar la información de la Tabla 2.1 es a través de pictogramas que representen de una forma más clara el resumen de los datos. En ciertas ocasiones, una visión general del comportamiento de todos los datos, otorga un mayor aporte al análisis descriptivo, lo cual ocurre cuando una tabla es de tamaño considerable debido a la extensión del análisis.

Si se consideran los datos anteriores, para mostrar de manera gráfica el comportamiento de los nacimientos por provincia a través del tiempo, por medio de

una gráfica de línea se puede considerar a los años en el eje de las X, y los nacimientos totales en el eje Y. Otro detalle que surge al graficar los datos, ocurre debido a que el último renglón muestra el total de los nacimientos en Canadá, por lo que es de una magnitud distinta a los datos de cada provincia, cuando se selecciona la tabla para realizar la gráfica en Excel® se debe agregar esta última serie por separado e indicar que se plasmen sus valores en referencia a un eje secundario, que es el eje aledaño del lado derecho del gráfico siguiente.

En el Gráfico 2.1 se observa el comportamiento del crecimiento natural dentro de Canadá de forma más general, sin embargo no se entra en el detalle del valor de los datos.

Gráfico 2.1 Nacimientos totales por provincia de Canadá y total (1981-2011)
Miles de Nacimientos - Total referenciado al eje secundario



Como se puede ver, existe un aumento de nacimientos durante el año 1991 tendencia que se observa es principalmente alimentada por las provincias de Quebec y Ontario; aunque en total sólo representa un crecimiento con respecto a 1986 de 30,275 nacimientos. Posterior al 2005 se observa un alza generalizada, más notoria sobre la provincia de Altavista. Posteriormente durante los últimos años de observación, se ve una estabilización en el número de nacimientos. El análisis gráfico anterior se enriquece normalmente por medio de contextos históricos del país y las provincias, mismos que se extienden más allá del alcance de este trabajo.

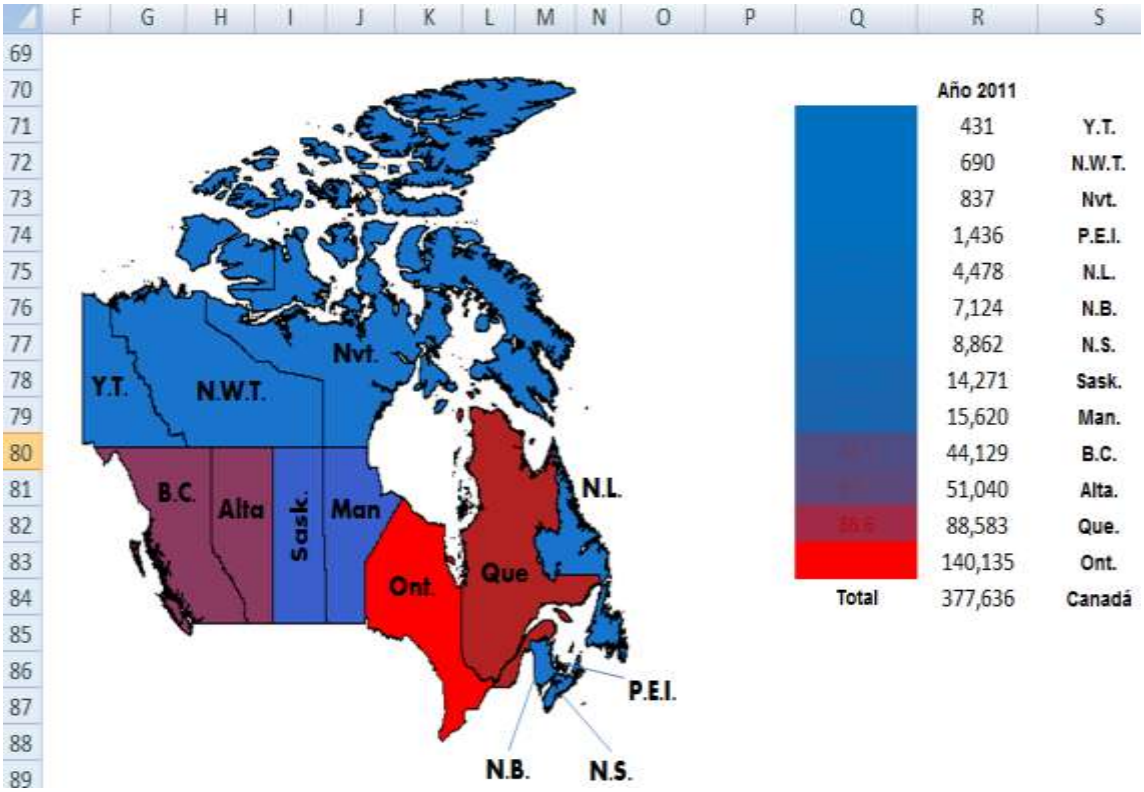
Otra forma de representar esta información es a través de relacionar el valor de la variable a la intensidad de un color en particular, o al cambio entre dos colores distintos , donde uno representa el mínimo del rango de valores y el segundo el máximo valor de los datos obtenidos o límite superior.

Si se aplicara esta técnica gráfica a la Tabla 2.1**,** la estabilidad de la población a través del tiempo, solo permitirá ver a la provincia de Ontario del color asignado para el mayor valor y los otros territorios del color contrario debido a la gran diferencia absoluta que se observa de nacimientos entre los territorios. Esta técnica grafica de análisis es de mayor utilidad en un plano geoespacial, ya que por medio de software

estadístico puede ser graficado en un mapa con los territorios de Canadá en color acorde a los valores de la Tabla 2.1.

La Grafica siguiente desarrollada en el software R[8], asigna una mayor intensidad de color rojo, a medida que el territorio tiene una mayor cantidad de nacimientos, y azul cuando los territorios tienen menores nacimientos totales, además si es copiada en imagen a una hoja de cálculo, se puede mejorar el formato de presentación incluyendo los datos de la tabla.

Gráfico 2.2 Nacimientos totales por provincia de Canadá para el año 2011

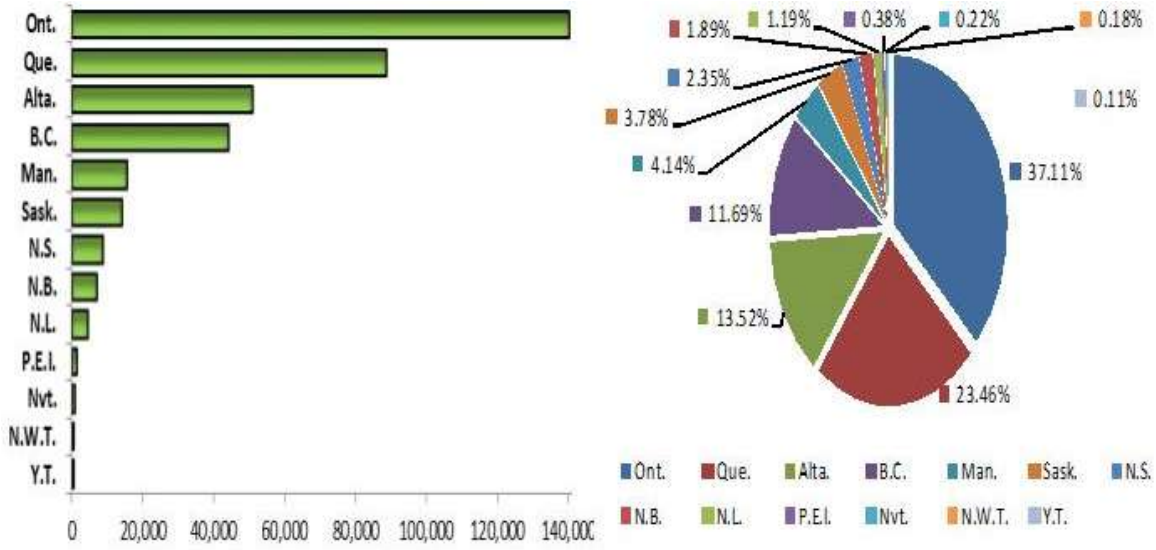


Usando la información correspondiente al año 2011, se nota una relación a un mayor número de nacimientos en la zona sur del territorio canadiense, y aunque esto arroja muchos cuestionamientos, es necesaria una mayor cantidad de información para poder relacionar temperatura, información económica, migratoria, o política de las diferentes provincias, y hallar una o varias explicaciones.

Existen diversas formas alternas a las presentadas hasta ahora, para la representación gráfica de esta u otra información agregada en categorías como las provincias canadienses, dos ejemplos más serían graficas de barra que muestra los valores absolutos y la gráfica de *pie* que muestra los porcentajes que representa cada categoría*;* ambas gráficas se concentran en realizar comparativos visuales y numéricos entre las provincias como se muestra a continuación.

[8] Consulte el código en el apéndice.

Gráfica 2.3: Nacimientos totales por provincias de Canadá para el año 2011



Función empírica de distribución acumulativa

Una vez definida la población objetivo, a la característica que se desea estudiar de la población se le asocia a una variable aleatoria X, la cual se define a través de su función de distribución acumulativa $F_X(x)$, por lo que para cada individuo de la población la probabilidad de que su respuesta o característica sea menor o igual a un valor x, está determinada por

$$\boldsymbol{F_X(x) = P(X \leq x)}$$

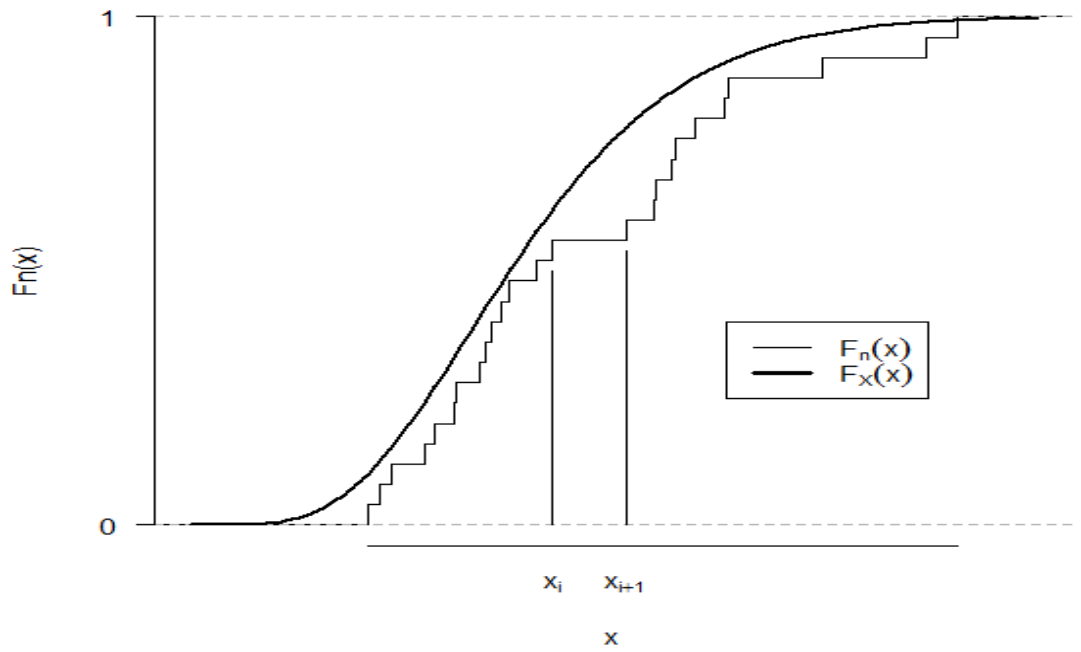
La cual se desconoce usualmente y al comenzar un análisis descriptivo no se realizan supuestos sobre esta función, sin embargo, para trabajar con simulaciones en el capítulo anterior se enunciaron las distribuciones más comunes en la teoría de la probabilidad y un método de simulación, donde se conocían de antemano sus distribuciones acumulativas.

Al no conocer la función de distribución F_X, se aproxima mediante una función de distribución calculada sobre una muestra obtenida de la población conocida como la función de distribución empírica, la cual se denota como $F_n(x)$ y se define como el conteo de los valores de la muestra menores al valor x, dividido entre el número de elementos n, es decir se obtiene mediante la siguiente expresión

$$\boldsymbol{F_n(x) = \frac{\#\{X_i \leq x\}}{n}}$$

Esta función estima la probabilidad calculando el porcentaje acumulado de la muestra en un punto x por lo cual tiene una forma escalonada debido a que sólo eleva su valor en los valores x_i y se define constante entre dos valores consecutivos x_i , x_{i+1}, similar al ejemplo ilustrativo en la Gráfica 2.4. Esta función en adelante será usada frecuentemente para realizar las simulaciones, la cual también puede ser objeto de un estudio más detallado dependiendo del nivel de curso.

Gráfica 2.4: Visualizacion de F(x) y el cálculo de la función de distribución empírica



Generación de muestras por simulación Monte Carlo

Una forma en que se puede acercar al estudiante a las técnicas descriptivas anteriores, es a través de ejercicios en clase o un *template*[9] que fuese guiando la generación de los gráficos, asignado como alguna tarea. Sin embargo, es mejor entregarles una visión desde el inicio real del proceso del análisis, es decir, entregar la información en datos individuales en forma de una base de datos.

Para generar un ejemplo esta clase de *template*, considérense la información de nacimientos de 2011 de Canadá. Para simular una muestra que derive en los resultados anteriores se deben generar en una hoja de cálculo dos columnas, que contenga las siguientes variables:

- **ID: Un identificador del individuo, puede ser una Índice numérico (1,2,…, n) ó una cadena de caracteres del estilo, "0001","0002",….,"0… n", para el año elegido $n = 377,636$.**
- **Provincia: La descripción de la provincia a la que pertenece cada individuo (N.L., P.E.I., N.S., N.B., Que., Ont., Man., Sask., Alta., B.C., Y.T., N.W.T., Nvt.)**

Para lograr que la muestra obtenida tenga la misma distribución que población original, se va a aplicar el método de la transformada inversa visto en el capítulo uno, que aplica el algoritmo:

1. $Generar\ U \sim U(0,1)$
2. $hallar\ el\ entero\ positivo\ mas\ pequeño\ i\ que\ cumpla\ U \leq F_n(x)$
3. $hacer\ Y = x_i$

[9] Término usado por los angloparlantes para designar un archivo en el que sólo se define una estructura o proceso que requiere el ingreso de parámetros o información por un usuario.

De acuerdo al algoritmo, en la hoja de cálculo se debe calcular la función de distribución empírica de los datos, por lo que se debe obtener las frecuencias relativas f_i como el porcentaje que cada una de las K provincias representa del total de los nacimientos Canadienses, para después acumular el porcentaje, a través de la provincias para obtener su función acumulativa empírica de frecuencias relativas F_n. El resultado de realizar los cálculos se encuentra en la Tabla 2.4:

Tabla 2.4: % por cada provincia (Año 2011)

K	Provincia	# Nacimientos	f_i	S_n
1	Y.T.	431	0.10%	0.10%
2	N.W.T.	690	0.20%	0.30%
3	Nvt.	837	0.20%	0.50%
4	P.E.I.	1,436	0.40%	0.90%
5	N.L.	4,478	1.20%	2.10%
6	N.B.	7,124	1.90%	4.00%
7	N.S.	8,862	2.30%	6.30%
8	Sask.	14,271	3.80%	10.10%
9	Man.	15,620	4.10%	14.20%
10	B.C.	44,129	11.70%	25.90%
11	Alta.	51,040	13.50%	39.40%
12	Que.	88,583	23.50%	62.90%
13	Ont.	140,135	37.10%	100%
	Canadá	**377,636**	**100%**	**-**

De esta manera el paso 2 del algoritmo se convierte en encontrar la provincia i de tal manera que $U \leq F_n(x)$ donde U es un número aleatorio Uniforme generado en la hoja de cálculo. Aprovechando el ambiente de hoja de cálculo, se puede simplificar el algoritmo generando un valor aleatorio $U_i \sim U(0,1)$ con la fórmula ALEATORIO(), a partir del cual se aplica la función $COINCIDIR()$[10] de Excel, que toma como parámetros el valor U_i y el rango de la tabla donde se encuentra F_n, para devolver el renglón K_i donde se halla la provincia cuyo porcentaje acumulado es el más cercano al valor de U_i.

A partir de este valor se emplea la función $Consultav()$* que localizará el valor de K_i sobre la Tabla 2.4, a la cual se le debe indicar también el número de columna a devolver, a partir del valor de referencia. La Tabla 2.5 es un ejemplo de cálculo para los primeros 10 valores de la simulación. Para conseguir un mejor entendimiento del proceso los números aleatorios generados están escritos en porcentaje.

[10] Estas fórmulas pueden encontrarse con un nombre distinto dependiendo de la versión de Excel.

Tabla 2.5: Ejemplo de cálculo para algunos valores aleatorios			
ID	Aleatorio	Provincia Correspondiente	Inversa
i	$U_i \sim U(0,1)$	$= \text{Consultar}(K_i, \text{TablaProvincias}, 2)$	$= \text{Coincidir}(U_i, S_n) = K_i$
1	42.75%	Que.	12
2	81.16%	Ont.	13
3	18.73%	B.C.	10
4	16.44%	B.C.	10
5	18.00%	B.C.	10
6	76.62%	Ont.	13
7	48.57%	Que.	12
8	95.83%	Ont.	13
9	10.91%	Man.	9
10	87.76%	Ont.	13

El tamaño del vector es el que determinara el tamaño de la muestra n simulada, en este caso se tomarán 1,500 valores para esta simulación. A continuación la Tabla 2.6 resume el conteo final obtenido de una muestra generada, referida en la Tabla como (Y), donde además se compara, en aras de demostrar su similitud, las frecuencias relativas $f_i(Y)$ contra $f_i$en la columna Δ. Se puede observar que la máxima diferencia es de alrededor de medio punto porcentual en la última provincia de la tabla, con lo que se puede concluir que el estudiante podrá reproducir todos los métodos antes vistos y otros impartidos en clase, a partir de interactuar con la muestra simulada.

Tabla 2.6: Resultado de la simulación y comparación contra la distribución de la población real				
Provincia	# Nacimientos Simulados	$f_i(Y)$	f_i	$\Delta = f_i - f_i(X)$
Y.T.	1	0.10%	0.10%	0.00%
N.W.T.	2	0.10%	0.20%	0.00%
Nvt.	5	0.30%	0.20%	0.10%
P.E.I.	8	0.50%	0.40%	0.10%
N.L.	22	1.50%	1.20%	0.30%
N.B.	25	1.70%	1.90%	-0.20%
N.S.	34	2.30%	2.30%	0.00%
Sask.	61	4.10%	3.80%	0.30%
Man.	63	4.20%	4.10%	0.10%
B.C.	170	11.30%	11.70%	-0.40%
Alta.	206	13.70%	13.50%	0.20%
Que.	356	23.70%	23.50%	0.20%
Ont.	547	36.50%	37.10%	-0.60%
Canadá	1,500	100%	100%	-

Histogramas

La muestra generada anteriormente, representa los nacimientos de las provincias de Canadá dentro de un rango de categorías finitas representadas dentro de la simulación Monte Carlo por números K_i.

Cuando la mediciones y observaciones se hacen clasificando los resultados de la muestra a través de un numero de categorías finitas, entonces se trata de una variable categórica, y el tipo de análisis dependerá si las categorías definidas para la muestra tienen o no un orden especificado, es decir, como en el caso de las provincias de Canadá, donde no existe una relación universal de "mayor o menor que" entre ellas. Otro enfoque puede ser por ejemplo, registrar el número de hijos de una familia, donde aunque no existe un límite previo sobre el máximo número de hijos de una población, se puede intuir que las categorías no serán mayores a un número finito K y seguirán siempre el orden: $sin\ hijos < 1\ hijo\ < 2\ hijos\ < 3\ hijos < \cdots < K\ hijos$.

Otro tipo de información se encuentra usualmente en datos numéricos, que se emplean para registrar mediciones de las cuales se sepa de antemano que los resultados pueden venir de un rango continuo de valores. Un detalle a resaltar para estas variables es que en las aplicaciones modernas cuando se requieren analizar millones de datos de información, la precisión con la que se pretende capturar cada dato observado, tiene un papel importante, puesto que una mayor precisión implicará la necesidad de un mayor espacio en la memoria de un computador debido al registro de decimales, lo cual impacta en algunas industrias, por ejemplo aquellas con sistemas transaccionales que pueden llegar a tener millones de clientes de quienes registran información de sus actividades por un tiempo de longitud indeterminada.

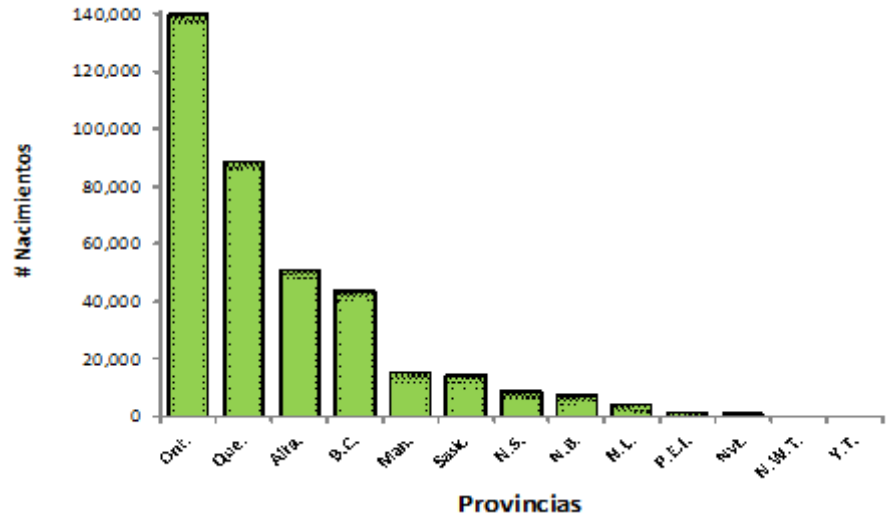
Una alternativa es mediante el truncamiento de los valores registrados a un menor número de decimales, sin embargo el simplificar demasiado derivará en desviaciones sobre los resultados de los cálculos estadísticos, que serán cada vez mayores conforme el tamaño de la muestra aumente.

Un gráfico universal dentro de las aplicaciones descriptivas es el histograma de frecuencias o simplemente histograma. Este gráfico considera el rango en el que se encuentra definida la variable de estudio para dividirlo en categorías o subintervalos, según el tipo dominio de la variable, con el fin de contar el número de datos (frecuencia) de cada categoría y mostrarlo en un gráfico de barras usualmente verticales.

Por ejemplo, considérense nuevamente los datos de las provincias de Canadá para el año 2011 vistos en un gráfico de barras vertical. La frecuencia de sus provincias, es decir su histograma, es similar a la Gráfica 2.3, aunque con orientación de las barras distinta.

De manera inicial no existe un orden predeterminado para las provincias de Canadá, por lo cual se opta por ordenarlos por el valor de su frecuencia, pues permite identificar en los extremos del gráfico las categorías con mayor y menor concentración de individuos, que es un cuestionamiento usual. Una forma de insertar los gráficos de barras predeterminados en la hoja de cálculo, usando software Excel® es a través de seleccionar el rango de valores que contenga la tabla de las categorías con sus respectivas frecuencias e insertando un grafico de barras, la cual se observa en la Gráfica 2.5.

Grafico 2.5: Histograma de frecuencias por provincias de Canadá (2011)



Si las categorías tienen un orden natural, entonces deben aparecer en el histograma en tal orden sobre el eje horizontal, y mostrar las frecuencias de cada una de sus categorías. De esta manera exhibe las categorías con mayores y menores concentraciones de individuos. Este gráfico es útil para identificar además, el cambio en la intensidad de la aparición de los individuos o su respuesta a medida que aumenta o disminuye el fenómeno de interés.

Para el caso en que las observaciones se registren de una variable continua, realizar el histograma requiere un paso previo, que es dividir el rango donde se encuentra definida la variable en subintervalos, para considerarlos como una serie de categorías ordenadas donde se calculan y grafican las frecuencias de cada uno.

La división del conjunto donde se define la muestra para las variables continuas, se realiza a discreción por parte del analista aunque existen prácticas recomendadas sobre el número de subintervalos a tomar, como la regla de Sturge que divide el rango de la muestra en m intervalos por medio de la fórmula

$$\boldsymbol{m = \log_2 n + 1}$$

Aplicación de técnicas Monte Carlo

Los detalles de los individuos de donde proviene el reporte utilizado anteriormente sobre fertilidad y en general sobre otros reportes publicados sobre distintos tópicos, es información privada que en muchos casos es considerada como delicada, por lo cual nunca se difunde a externos. Para poder practicar las técnicas estadísticas a nivel individual, se puede modelar algún contexto en particular y simular una muestra con el nivel de detalle deseado.

Con el propósito de tratar el tema de histogramas bajo los diferentes tipos de datos se propone la interacción desde un inicio con una serie de muestras de diversas características. Esta sección profundizará sobre los últimos detalles y tratamientos para ejecutar histogramas con datos categóricos ordenados y continuos por medio de simulaciones Monte Carlo.

Para ilustrar lo anterior supóngase la existencia de la empresa XYZ dedicada al giro de venta de diversos productos vía telefónica, la cual cuenta con 50 empleados, a los cuales les solicita una cuota diaria de ventas de 10 artículos. La empresa considera dentro de la medición del desempeño de sus empleados el número de ventas realizadas, el valor de los productos vendidos y el número de llamadas necesarias para concretar sus ventas, debido a los costos de oportunidad de intentos fallidos, más los costos telefónicos que genera cada llamada.

Una parte de su reporte de ventas del último mes mostrado en la Tabla 2.7, señala las ventas totales, V, suponiendo que todos los empleados cumplieron su cuota en 30 días y el total de llamadas generadas en el mismo periodo es, L, donde se muestra también un primer indicador como la probabilidad p que tiene un empleado elegido al azar de realizar una venta, la cual se estima como $p = V/L$. Como consultor de la firma la directiva le señala que desean analizar ¿Cuáles son los efectos de aumentar la cuota del número de ventas?, ¿Y cuáles los impactos, si la estrategia se dirigiera en lograr una mayor probabilidad de venta, a través de capacitación, nuevos modelos de conversación o mejores canales de venta?

Tabla 2.7: Reporte mensual Empresa XYZ	
Empleados	**15**
Ventas totales V=15*10*30	**4500**
Llamadas totales L	**8550**
Probabilidad de éxito p=V/L	**53%**

Para responder lo anterior se puede simular el modelo de negocio de tal manera que los parámetros definidos en la simulación puedan ser interpretados, para esto es posible considerar las llamadas de un día de trabajo como una serie de experimentos Bernoulli definidos como

$$\boldsymbol{I(x)} = \begin{cases} \mathbf{1} & \boldsymbol{x \in \{venta\ realizada\}} \\ \mathbf{0} & \boldsymbol{x \in \{Llamada\ sin\ venta\}} \end{cases}$$

Por lo cual de los cursos básicos de probabilidad se debe retomar el modelo de la variable aleatoria discreta Binomial Negativa, pues describe una serie de ensayos (llamadas) aleatorios de este tipo hasta lograr un número fijo de r éxitos (ventas). La variante para la definición de la variable es si se desea modelar el número Y de fracasos o el número X de ensayos realizados ($X = Y + 1$). En el escenario propuesto, se elige considerar la definición de Y para la simulación ya que representa las llamadas totales, lo cual está más alineado a los cuestionamientos. Los parámetros de la variable Binomial Negativa son el número de éxitos objetivo $r = 10$, y la probabilidad de éxito, que de acuerdo a la Tabla 2.7 $p = 53\%$, caracterizada por la función de probabilidad $f(y)$.

$$\boldsymbol{f(y) = P(Y = y) = \binom{r + y - 1}{r - 1} p^r (1 - p)^y}$$

La simulación de Y se consigue en ambiente de hoja de cálculo por medio de la suma de r variables geométricas como se vio en el capítulo 1. Cada variable geométrica será simulada a partir de un aleatorio U_i como $x_i = \lfloor \ln(U_i) / \ln(1 - p) \rfloor$, donde

$i := \{1,2,3,4, r = 10\}$ y los corchetes denotan la función piso, sin embargo para generar la variable aleatoria deseada se calcula $y_i = x_i + 1$ y se suman los r resultados como en los ejemplos de la Tabla 2.8

$$\boldsymbol{Y = \sum_{i=1}^{r} y_i} \qquad \boldsymbol{Y \sim BinNeg(r = 10, p = .53)}$$

Tabla 2.8: Simulación de un día de trabajo por medio de suma de variables geo(p=.53)

$\boldsymbol{i}$	$\boldsymbol{U_i}$	$\boldsymbol{y_i = \left\lfloor \frac{Ln(U_i)}{Ln(1-p)} \right\rfloor + 1}$
1	70.30%	1
2	40.70%	2
3	52.90%	1
4	10.40%	4
5	73.80%	1
6	48.30%	1
7	48.40%	1
8	25.10%	2
9	61.10%	1
10	37.70%	2
		$\boldsymbol{Y = \sum_i y_i = 16}$

La Tabla 2.8 obtenida para una simulación de Y, muestra que en 16 llamadas se alcanzó una cuota de ventas, además la construcción de la simulación permite observar el detalle de cada venta, pues cada resultado y_i representa el número de intentos para conseguir una venta. Como en el ambiente de hoja de cálculo el contenido de cada celda es en realidad una fórmula, el copiar y pegar un valor de Y generará automáticamente la actualización de la hoja de cálculo y generará una nueva observación de cada U_i.

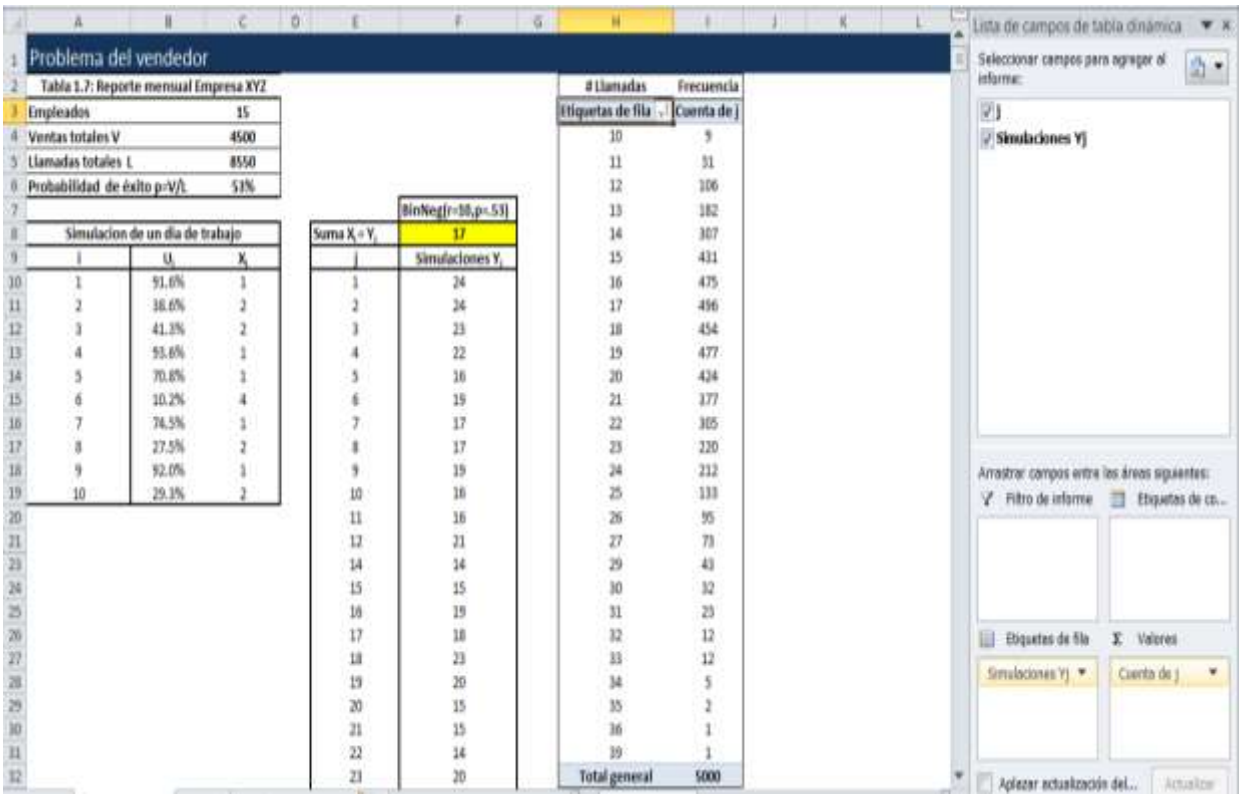
La simulación del total entonces se resume en automatizar en código VBA[11] , un proceso que copie el resultado de Y y lo pegue a valores, donde además se pueda definir el número de iteraciones, que determinará el tamaño de la simulación. Para este ejemplo se ha tomado $n = 5000$.

La muestra también debe contener una columna de enteros consecutivos j correspondiente a cada simulación de Y_j similar a la Tabla 1.5. Una vez realizado esto, una herramienta sencilla para realizar la tabla de frecuencias son las tablas dinámicas, pues basta con seleccionar un valor cualquiera de la tabla e indicar desde el menú la inserción de una tabla dinámica y reconocerá todo el rango de valores automáticamente. Por último, en el espacio de la tabla para colocar los valores de las columnas se coloca a la columna de las Y_j y en campo de valores se coloca la columna j, configurado para realizar la operación de cuenta de valores.

[11] Consulte el código en el apéndice.

La siguiente imagen muestra una estructura propuesta para la realización del ejercicio hasta la generación de la tabla dinámica como tabla de frecuencias.

Estructura en hoja de cálculo del problema de ventas en un *call center.*



Problema del vendedor

Tabla 1.7: Reporte mensual Empresa XYZ

Empleados	15
Ventas totales V	4500
Llamadas totales L	8550
Probabilidad de éxito p=V/L	53%

Simulacion de un dia de trabajo

i	U_i	X_i
1	91.6%	1
2	16.6%	2
3	41.3%	2
4	93.6%	1
5	70.8%	1
6	10.2%	4
7	74.5%	1
8	27.5%	2
9	92.0%	1
10	29.3%	2

BinNeg(r=10,p=.53)

Suma $X_i = Y_j$ 17

j	Simulaciones Y_j
1	24
2	24
3	23
4	22
5	16
6	19
7	17
8	17
9	19
10	16
11	16
12	21
14	14
15	15
16	19
17	18
18	23
19	20
20	15
21	15
22	14
23	20

# Llamadas	Frecuencia
Etiquetas de fila	Cuenta de j
10	9
11	31
12	106
13	182
14	307
15	431
16	475
17	496
18	454
19	477
20	424
21	377
22	305
23	220
24	212
25	133
26	95
27	73
29	43
30	32
31	23
32	12
33	12
34	5
35	2
36	1
39	1
Total general	5000

Lista de campos de tabla dinámica

Seleccionar campos para agregar al informe:

j

Simulaciones Yj

Arrastrar campos entre las áreas siguientes:

Filtro de informe | Etiquetas de co...

Etiquetas de fila | Valores

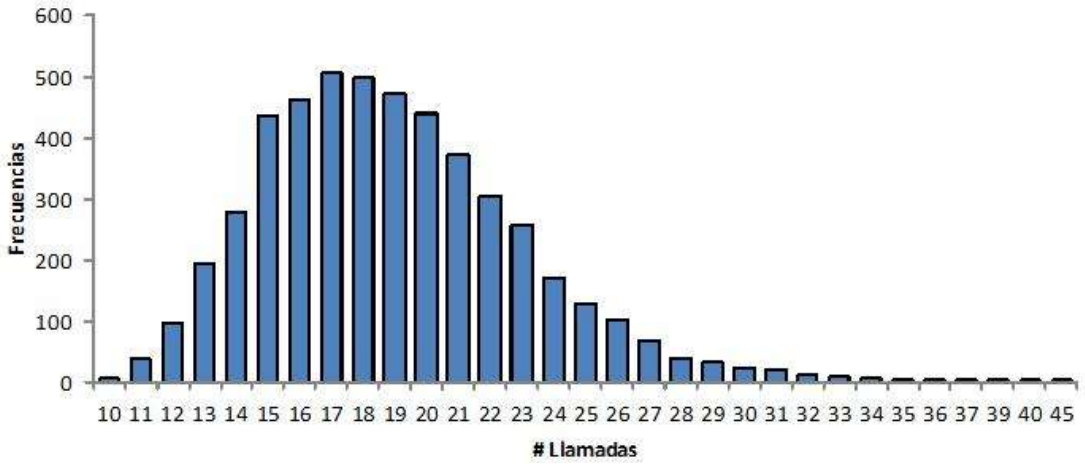
Simulaciones Yj | Cuenta de j

Aplazar actualización del... | Actualizar

El total de la tabla dinámica corroborará el total de las 5000 simulaciones generadas. Comúnmente al colocar el campo de las categorías la tabla dinámica ordena los valores automáticamente, en algunos casos es necesario indicarlo.

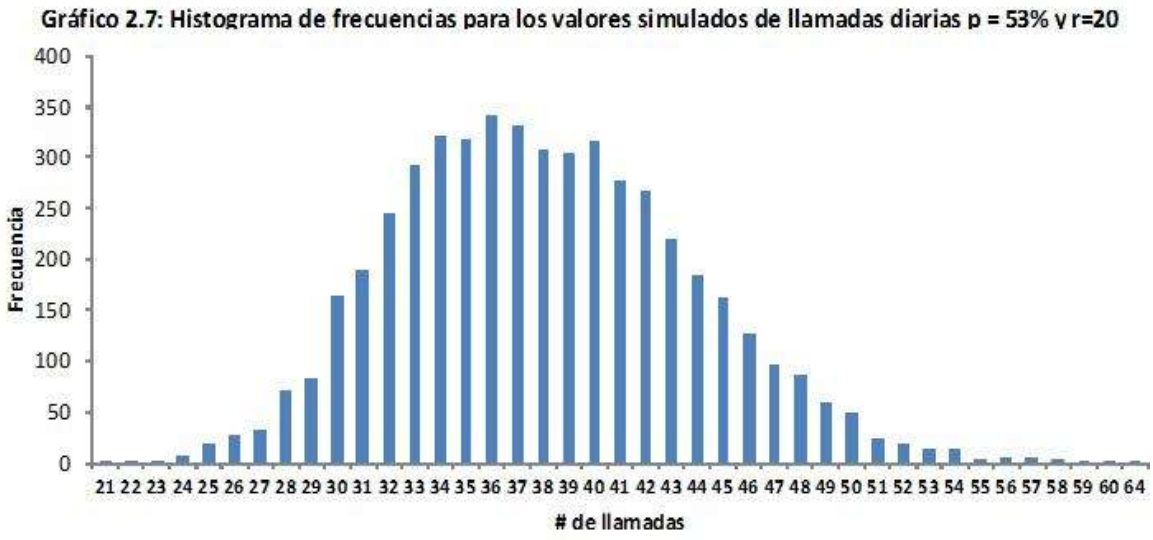
Para generar la vista gráfica a partir de esta información una vez seleccionada la tabla dinámica, se debe insertar un gráfico de columnas y el software Excel colocara las categorías en el eje horizontal y los conteos como barras, que mejorando el formato se obtiene como resultado la Gráfica 2.6.

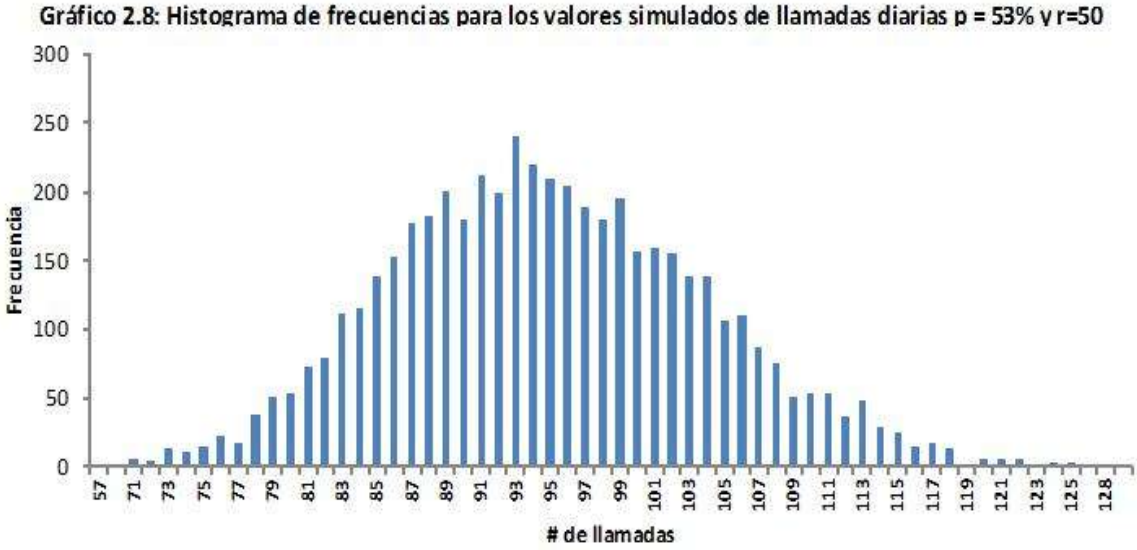
Gráfico 2.6: Histograma de frecuencias para los valores simulados de llamadas diarias p = 53% y r=10



El grafico muestra primeramente el rango en donde se define una segunda muestra después de la actualización de la hoja, donde indica que para lograr la cuota ahora se tuvieron que realizar entre 10 y 45 llamadas. Otro detalle que se destaca se observa en la curva que forman los valores de las frecuencias, la cual está cargada de forma asimétrica hacia la izquierda de la Gráfica 2.6 a partir del número de llamadas con mayor frecuencia, en este caso 17. A medida que se observan desde los extremos las frecuencias de cada categoría, son cada vez mayores, lo que quiere decir que se tiene una concentración de los resultados alrededor de este punto.

Explotando otra ventaja de la hoja de cálculo se pude incrementar el número de x_i generadas para experimentar con el cambio en la cuota de ventas. Las Gráficas 2.7 y 2.8 corresponden a repetir el ejercicio definiendo la cuota de ventas $r = 20$ y $r = 50$ cambiando el tamaño de la columna con las simulaciones de ventas individuales, pero conservando el valor de $p = .53$.



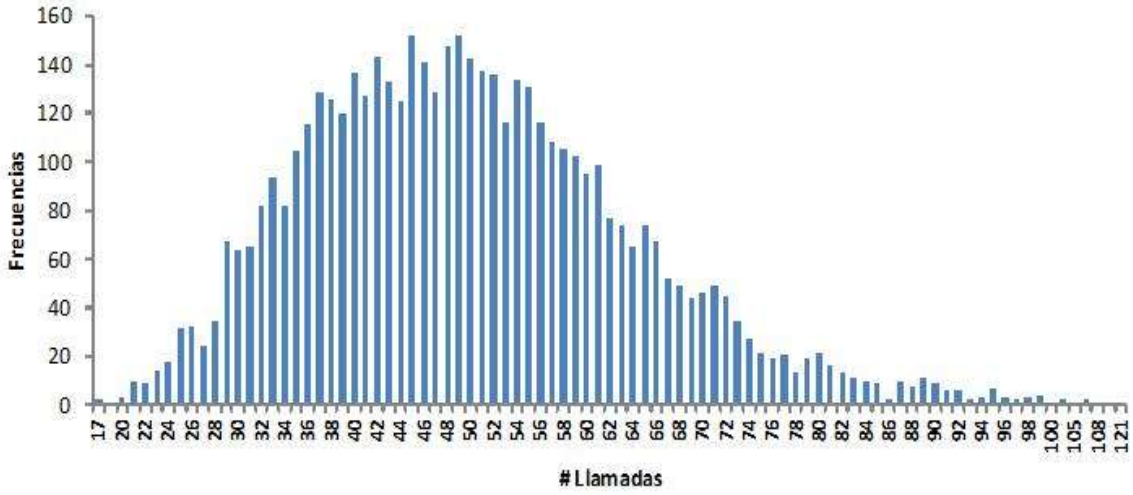


El haber aumentado la cuota, derivó en un incremento en el número máximo de llamadas y un aumento también en el número de llamadas alrededor del cual se concentran los resultados. Además el comportamiento de la concentración también ha

ido cambiando pues en el primer histograma se observan rasgos de asimetría en la forma en que se concentran las simulaciones, mientras que en el último histograma a pesar de la pérdida en la solidez del comportamiento, la forma que adopta el gráfico es más simétrico con respecto a las categorías donde se concentran las simulaciones.

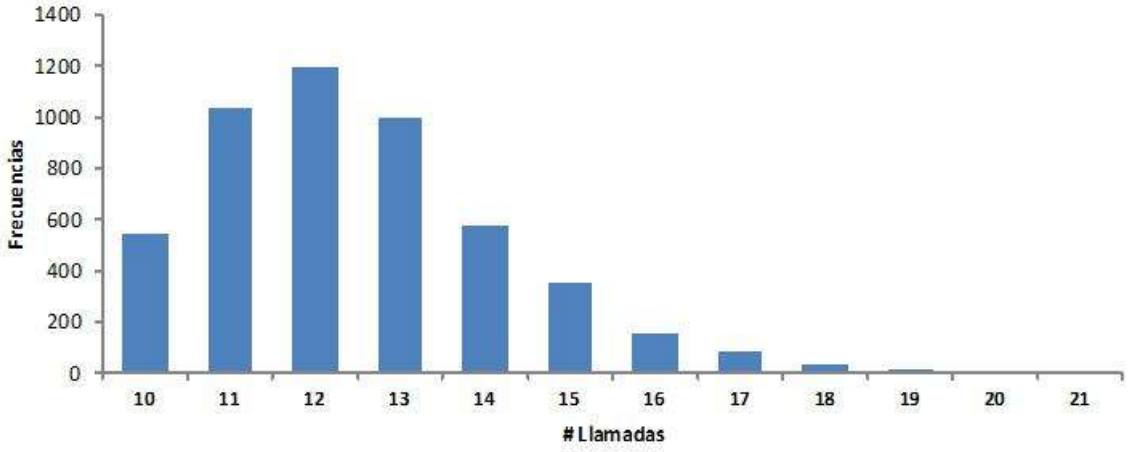
El otro parámetro que puede ser modificado para responder los temas iniciales en esta simulación es la probabilidad de éxito p, proponiendo directamente su valor o a partir del contexto, cambiando el número de llamadas totales realizadas bajo la restricción $L > V$. Ahora conservando la cuota de 10 ventas se repitió el ejercicio con los valores propuestos $p_1 = 20\%$ y $p_2 = 80\%$.

Gráfico 2.9: Histograma de frecuencias para los valores simulados de llamadas diarias p = 20% y r=10



Si se analiza el gráfico 2.9 se observa que debido a la elección de una baja probabilidad de éxito hubo días en que la simulación de ventas tomó hasta 121 llamadas en alcanzar la cuota de venta, además sus frecuencias se encuentran con una asimetría que se observa cargada al lado izquierdo del rango, lo que se considera una asimetría positiva. Hay una concentración alrededor de las 44 a 48 llamadas pero la frecuencia no es tan alta comparada cuando $p = 53\%$, pues en general la observaciones se encuentran más dispersas en todo el rango de la muestra.

Gráfico 2.10: Histograma de frecuencias para los valores simulados de llamadas diarias p = 80% y r=10



En el caso del gráfico 2.10 por el contrario se observa que el número de categorías es considerablemente menor debido a que una alta probabilidad de éxito genera un menor número de llamadas simuladas para cumplir la cuota, sin embargo las simulaciones tuvieron su mayor concentración en las 12 llamadas, y decae con rapidez conforme el número de llamada es mayor, lo que le da su forma asimétrica.

A través de estas simulaciones, se pudo ver el impacto que tienen en el modelo, tanto el cambio en mejorar o empeorar la calidad de los métodos de éxito como en la exigencia de un mayor número de ventas. Sin embargo la combinación de ambos efectos con la experimentación con otros valores de r y p se deja a la libre aplicación, con el fin de que genere discusiones sobre el tema.

Para practicar la generación de estos gráficos en clase con datos provenientes de rangos continuos de una forma práctica, se puede hacer por medio de contextos financieros, por ejemplo consultando series financieras publicadas, como precios e índices en diversas monedas, los cuales son almacenados con un mayor número de decimales debido al gran volumen de operaciones financieras que se llevan a cabo con base en ellos.

Para desarrollar un ejemplo en un contexto financiero, se continuará desarrollando el ejemplo anterior. Suponga que la empresa xyz tiene un especial interés en conocer en resumen cual es el comportamiento de las ganancias generadas bajo el esquema actual. Como consultor se le entrega un reporte adicional (Tabla 2.9) que contiene el anterior reporte de ventas desglosadas en unidades por los tres productos que comercializa junto con el costo unitario por llamada, el % de unidades vendidas, el total de ventas por producto y la ganancia resultado de sumar las ventas y restar los costos totales

Tabla 2.9: Ventas desglosadas por producto Empresa XYZ

Articulo	Precio/costo	Unidades	% Unidades	Total $
Llamada a cliente	$5	8,550	100%	$42,750
producto 1	$23	2,025	45%	$46,575
producto 2	$54	1,575	35%	$85,050
producto 3	$95	900	20%	$85,500
Ventas Totales		4,500	100%	**Venta - Costos=**
				$174,375

Los valores fijos propuestos en esta tabla son los porcentajes asignados a cada producto para calcular las unidades vendidas, multiplicándolas por el total de ventas V, y los precios unitarios de las llamadas (C) y productos (P_1, P_2, P_3). Esto permite establecer una distribución con tres productos (o más según se desee) o categorías, las cuales se basan en el supuesto simple de asignar un mayor porcentaje de venta a productos de menor precio y un costo por venta relativamente bajo a comparación de los precios de los productos; aunque en un contexto distinto el costo puede ser visto como porcentaje de los precios finales e incluso variar de acuerdo al volumen de venta. En la última columna se encuentra calculada la ganancia mensual como las ventas de todos los productos menos los costos generados.

Con la información anterior ahora es posible simular el comportamiento de las ganancias que genera cada empleado en diferentes escenarios de cuota de ventas o probabilidad de venta. Para esto se puede reutilizar la estructura en hoja de cálculo previamente generada para generar la Tabla 2.8 ya que los costos generados de realizar una venta es el número de llamadas necesarias Y_i multiplicada por el costo de cada llamada como $C_i = Y_i * C$.

Luego para simular el valor de la venta, se considera primero que la elección del producto vendido es independiente al número de llamadas fallidas previamente para lograr la venta, por lo que para simular los precios del producto seleccionado, se reproducirá el método previo al simular las provincias de Canadá, por lo que es necesario incluir en la hoja de cálculo el porcentaje acumulativo de unidades de forma análoga a la Tabla 2.4, para que a partir de un segundo numero aleatorio en $(0,1)$ W_i, se relacione al producto vendido por medio de la función COINCIDIR(). Posteriormente se emplea la función CONSULTAV() para devolver el valor correspondiente de la columna de precios, simulando el valor del producto elegido.

Las ganancias g_i derivadas de cada venta individual, se calculan restando los costos totales de las llamadas al valor de la venta. Para obtener la rentabilidad total diaria G se suma el total de los valores g_i. A continuación se muestra el ejemplo de la Tabla 2.8 adicionando las columnas necesarias para realizar la simulación de la ganancia final, a partir de las simulaciones previas y_i.

Tabla 2.10: Simulación de costos y ganancias generadas

i	U_i	y_i	$C_i = y_i * C$	$W_i \sim U(0,1)$	V_i	$g_i = V_i - C_i$
1	70.30%	1	$5	50.00%	$54	$49
2	40.70%	2	$10	1.70%	$23	$13
3	52.90%	1	$5	91.30%	$95	$90
4	10.40%	4	$20	17.50%	$23	$3
5	73.80%	1	$5	97.50%	$95	$90
6	48.30%	1	$5	36.20%	$23	$18
7	48.40%	1	$5	83.10%	$95	$90
8	25.10%	2	$10	78.60%	$54	$44
9	61.10%	1	$5	79.10%	$54	$49
10	37.70%	2	$10	44.90%	$23	$13
		$Y = \sum_i y_i = 16$	$\sum_i C_i = \$80$		$539	$G = \sum_i g_i = \$459$

Una vez generada una simulación de la variable G, se establece el tamaño de la muestra n, por ejemplo del mismo tamaño que las muestras anteriores, con $n = 5000$, y se utiliza un proceso que automatice el pegado y copiado consecutivo en la hoja de cálculo el valor de G[12], por último para completar la estructura de la muestra se agrega una columna de enteros consecutivos j como índice de la muestra.

En clase se puede generar una tabla dinámica con un conteo similar al del número de llamadas, con el objetivo de mostrar como una gran cantidad de valores generados tendrán frecuencia 1 implicando que son únicos en la muestra. Por esta razón se

[12] Consulte el código en el apéndice.

debe analizar de forma agrupada, entonces se requiere determinar el número de subintervalos m en los que se dividirá el rango de valores de la muestra.

En la hoja de trabajo se obtiene el cálculo de la regla de Sturge como $m = \lfloor \log_2(5000) + 1 \rfloor = 13$, por otra parte se debe calcular además, el mínimo y máximo de la muestra con las funciones MIN() y MAX() respectivamente para calcular los límites superiores $S_i \quad i \coloneqq \{0, \dots, 13\}$ de cada uno de los subintervalos por medio de la fórmula recursiva

$$\boldsymbol{S_i = S_{i-1} + \frac{maximo - mínimo}{m}}$$

Donde S_0 es el mínimo de la muestra. Posteriormente se utiliza la función FRECUENCIA() de Excel® para referenciar toda la muestra simulada, y como siguiente parámetro el valor de cada S_i, sin embargo esta función devolverá un conteo acumulado F_i, por lo que la frecuencia puntual se obtiene en una columna adicional con la fórmula

$$\boldsymbol{f_i = F_i - F_{i-1} \quad i \coloneqq \{2, \dots, 13\}}$$

$$\boldsymbol{f_1 = F_1}$$

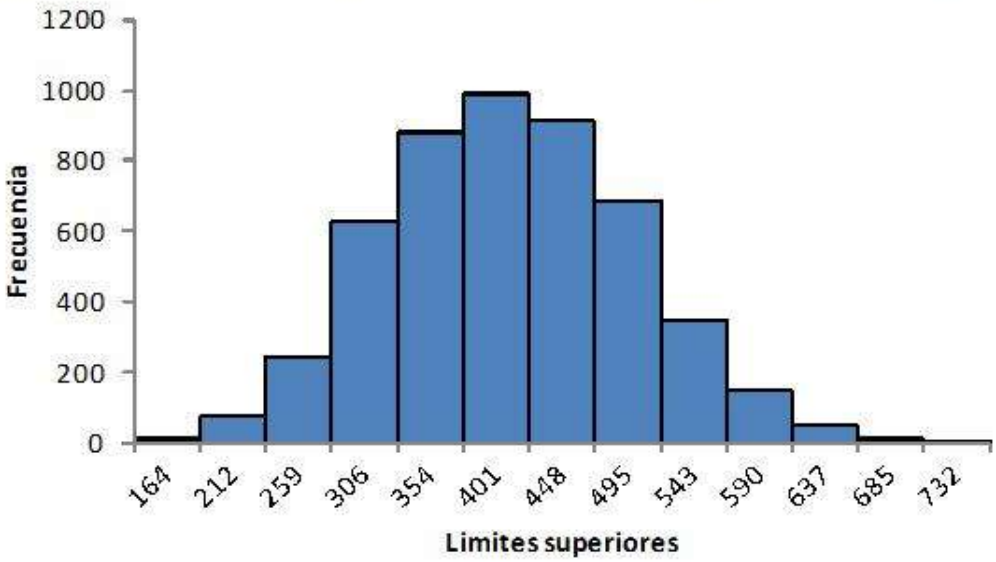
Tabla 2.11: Frecuencias acumuladas y puntuales de una realización de 5000 ganancias simuladas

i	S_i	F_i	f_i
1	164	10	10
2	212	85	75
3	259	330	245
4	306	955	625
5	354	1838	883
6	401	2831	993
7	448	3742	911
8	495	4431	689
9	543	4780	349
10	590	4931	151
11	637	4980	49
12	685	4996	16
13	732	5000	4
			Total =5000
n	**M**	**Máximo (S_m)**	**Mínimo (S_0)**
5000	13	732	117

En la Tabla 2.11 se encuentran los cálculos mencionados, los cuales muestran en el quinto renglón, referenciando al rango $(354, 401)$ obtuvieron la mayor frecuencia y dado que decaen a partir de este punto en ambas direcciones, hay una concentración alrededor de este intervalo. El histograma finalmente se genera insertando una gráfica, señalando la columna f_i como rango de valores y a S_i como etiquetas de datos,

aunque ahora que se trata del intervalo entero se elimina el espacio entre las barras, el resultado después de un arreglo en el formato, será similar a la Gráfica 2.11.

Gráfica 2.11: Histograma de frecuencias de las simulaciones de ventas diarias



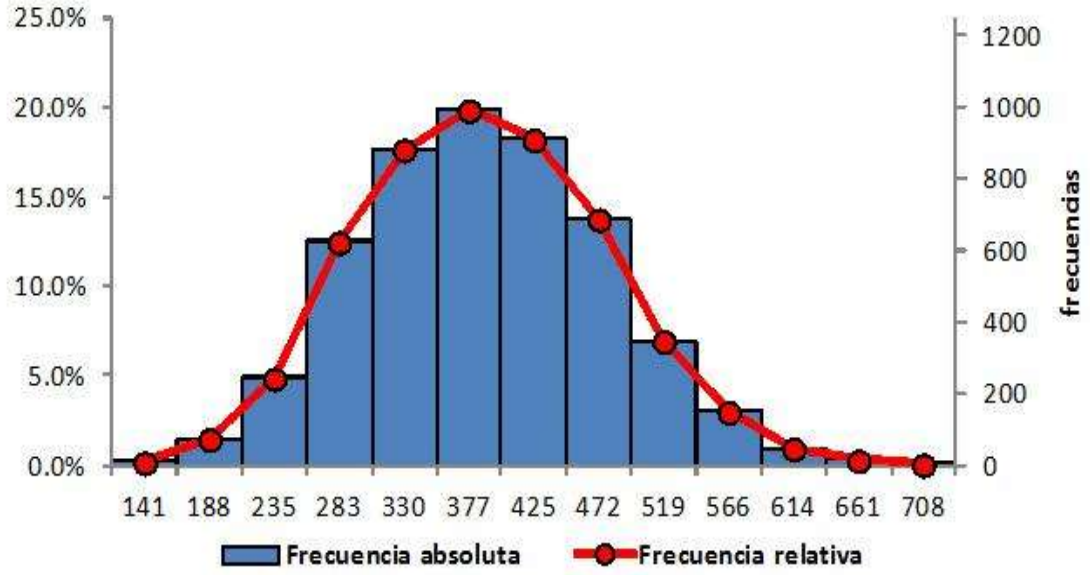
El grafico evidencia de forma clara la concentración de los resultados, además de la simetría en el comportamiento de las frecuencias. La variación de los parámetros del ejercicio permitirá revisar el comportamiento de las ganancias bajo diversos escenarios de capacitación y cuota. Al tratar el tema en un salón de clase la discusión se puede volver interesante cuando se elige la probabilidad de venta p baja, generando una gran cantidad de intentos o cuando los costos por cada intento aumentan y el margen de ganancia se reduce, ya que se puede revisar en qué escenarios se obtendrá una pérdida.

Curvas de frecuencias

Otra forma de representar la distribución de los datos en lugar de barras, es por medio de una gráfica línea conocida como el polígono de frecuencias, la cual se genera a través de trazar líneas entre los puntos (i, f_i). Cuando los datos son continuos, se considera como la primera coordenada al punto medio de cada uno de los m subintervalos del eje horizontal en los que se divide al dominio de la muestra.

Una variante es graficar las frecuencias relativas f_i/n, las cuales en lugar de las absolutas indican el porcentaje que representa el el i–ésimo valor o subintervalo en la muestra. Estas curvas se pueden sobreponer al histograma para mostrar como aproximan la forma de una función de densidad, lo cual deja más claro el concepto, como en la Gráfica 2.12 donde se la serie de las frecuencias relativas están referenciadas a un eje izquierdo o primario, mientras que las frecuencias absolutas están en referencia al eje secundario, además en el eje horizontal, se muestran los puntos medios calculados a partir de los límites superiores e inferiores de los subintervalos.

Gráfica 2.12 Histograma de frecuencias y poligono de frecuencias de las simulaciones de ventas diarias



Con el gráfico anterior se pretende identificar los comportamientos generales de la distribución de las ganancias simuladas, para posteriormente iniciar con el tema de los estadísticos que miden la tendencia central de los datos.

Medidas de tendencia central

El uso de graficas dentro de un reporte es de gran utilidad, aunque al presentar un reporte más completo resulta impráctico mostrar una gran cantidad gráficas, que explican algunas variaciones del mismo comportamiento, sin llegar a mostrar un número de interés o establecer una recomendación concreta con base en algún parámetro. Para resolver esto en una muestra aleatoria X con función de densidad teórica $f_X(x)$, la cual está compuesta de valores observables $x_i\,, i \leq n \;\; x_i \sim f_X(x)$, se usan estadísticos, que son definidos como toda función $T(X)$, cuyo cálculo incluye la muestra completa, y que como resultado sea a su vez una variable aleatoria sin parámetros desconocidos.

Estos estadísticos caracterizan diversos aspectos de una muestra, dependiendo de la forma en que se defina cada función. En esta parte se revisarán aquellos denominados de tendencia central, los cuales permiten determinar una posición central de la población, donde se verá que no siempre coinciden con los puntos de concentración identificados previamente en los histogramas.

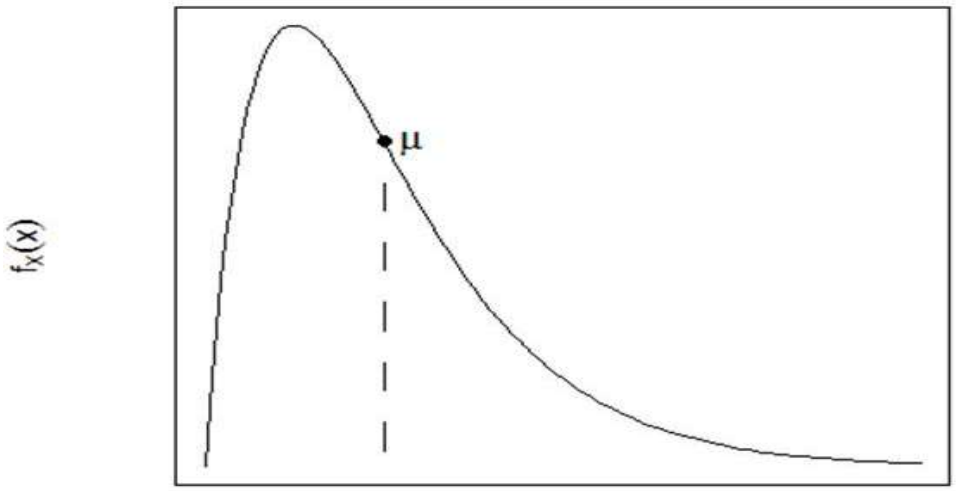
Media aritmética

Esta medida también conocida como promedio o simplemente media se calcula como la suma de todos los valores x_i de una muestra X, dividido entre el tamaño de la muestra n denotado comúnmente como $\bar{X}$

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

El valor del promedio de una muestra observada representa la localización de un punto en la curva de frecuencias, que en un contexto geométrico se conoce como centro de masa o de gravedad de la curva, es decir el punto donde un alambre rígido y delgado con la misma forma que la curva de frecuencias o la función de densidad de la población encontraría un equilibrio

Gráfica 2.13: Función de densidad f(x) y su media μ



Cuando se tiene la información de cada individuo de la muestra, como es el caso de las simulaciones, el computo de la media de una muestra en ambiente de hoja de cálculo, se realiza por medio de la función PROMEDIO(), sin embargo también se debe tocar como tema cuando la información publicada es un resumen de su distribución de frecuencias, reportes que son conocidos también como datos agrupados.

Cuando los datos se encuentran agrupados, y la naturaleza de los datos son similares a los tratados en el contexto del número de llamadas, es decir, en un número m razonable de categorías numéricas finitas C_i con sus respectivas frecuencias f_i, se calcula la suma de las observaciones en esa categoría como $C_i * f_i$, que en suma es el resultado de la suma total de la muestra por lo que el cálculo de la media es $\frac{\sum_{i=1}^{m} C_i * f_i}{n}$

En el caso que los datos provengan de un rango continuo, la distribución quedará tabulada en particiones de la muestra, por lo cual se toma como supuesto que los valores de la muestra dentro de cada subintervalo (S_i, S_{i+1}) se distribuyen uniformemente, con el objetivo de estimar la suma real de los elementos de cada subintervalo tomando el valor esperado de este subintervalo como mejor aproximación, el cual se sabe de las propiedades de la distribución uniforme que es el punto medio $m_i = (S_i + S_{i-1})/2$. La media de la población entonces queda estimada como $\frac{\sum_{i=1}^{m} m_i * f_i}{n}$

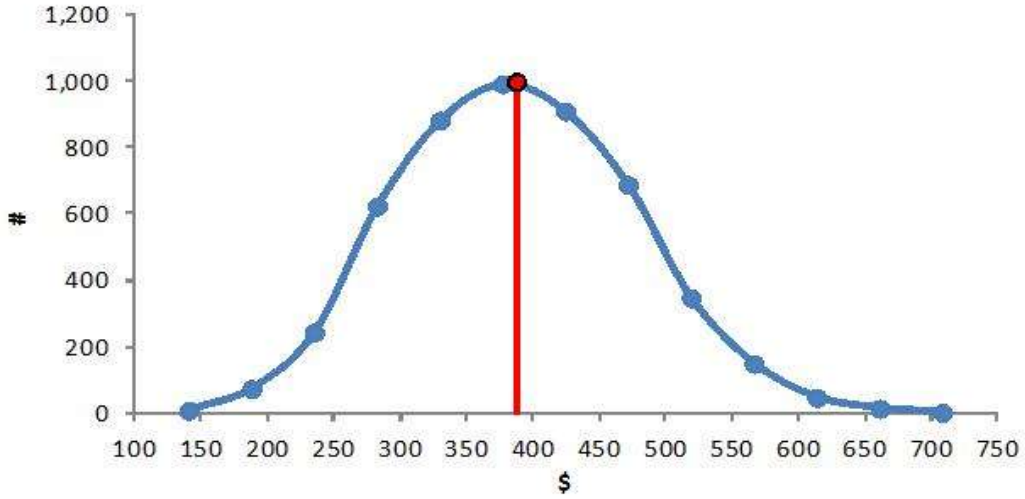
Como ejemplo se pueden retomar las ganancias simuladas en la Tabla 2.11 a la cual se le adicionan ahora las columnas m_i, $m_i * f_i$, para desglosar los cálculos y al final comparar los resultados entre la media con los datos individuales $\bar{X}$ y agrupados, mostrados al final de la Tabla 2.12

Tabla 2.12: Cálculo de la media con datos agrupados				
i	S_i	f_i	m_i	$m_i * f_i$
1	164	10	140.7	1,406.5
2	212	75	188	14,097.1
3	259	245	235.3	57,641.0
4	306	625	282.6	176,610.6
5	354	883	329.9	291,288.1
6	401	993	377.2	374,552.0
7	448	911	424.5	386,719.5
8	495	689	471.8	325,075.5
9	543	349	519.1	181,171.3
10	590	151	566.4	85,529.9
11	637	49	613.7	30,072.8
12	685	16	661	10,576.6
13	732	4	708.4	2,833.4
Total		5000		1,937,574.3
$N = 5,000$	$\bar{X} = 387.56$		$\frac{\sum_{i=1}^{m} m_i * f_i}{n} = 387.51$	

Se puede observar que la diferencia entre medias es del orden de los centésimos, lo cual indica una aproximación cercana a la verdadera media. Gráficamente la media se puede mostrar cambiando el tipo de gráfico de barras de los histogramas por un diagrama de dispersión, con los valores m_i como valores del eje X y las frecuencias f_i en el eje Y, obteniendo la curva de frecuencias.

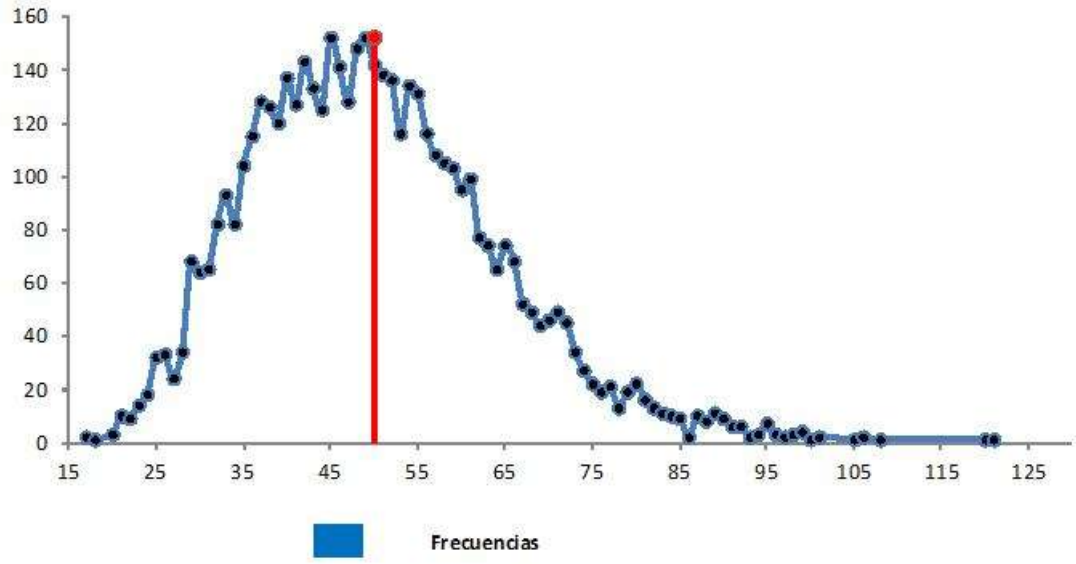
Una vez hecho esto se debe agregar al gráfico, la serie que contenga dos puntos $p_1 = (\bar{X}, 0)$, $p_2 = (\bar{X}, \max(f_i))$ el cual trazará una línea vertical localizando la media hasta la frecuencia más alta como referente. En la Gráfica 2.14 se muestra la curva de frecuencias de las ganancias simuladas anteriormente en donde se muestra la línea vertical en la posición de la media, la cual se encuentra muy cercana al punto de mayor frecuencia, donde también coincide la concentración de los resultados, lo cual sucede sólo cuando la distribución muestra simetría con respecto al valor de su media. Otra opción que se agregó al siguiente grafico fue suavizar las líneas que conforman el polígono de frecuencias lo que produce una curva que se asemeja aún más a una función de densidad.

Gráfica 2.14: polígono de frecuencias suavizado de las simulaciones de ventas diarias, donde se indica la posición de la media



Este ejemplo es uno de los que asemejan a aquellos que se pueden revisar en libros más teóricos, sin embargo cuando el número de puntos de la curva crece, usualmente se pierde cierta solidez en el comportamiento de las frecuencias, además puede ser acompañado por una asimetría en la curva, como es el caso del ejemplo del número de llamadas cuando $r = 10$ y $p = 20\%$, para mostrar esto la Gráfica 2.15 contiene la curva de frecuencias de este ejemplo particular, en la cual se ha localizado la media de la muestra con el valor de $\bar{X} = 50.01$. En la curva se puede observar que a pesar de seleccionar la opción de suavizar las líneas del polígono no genera una diferencia visual significativa debido al comportamiento de los puntos a unir.

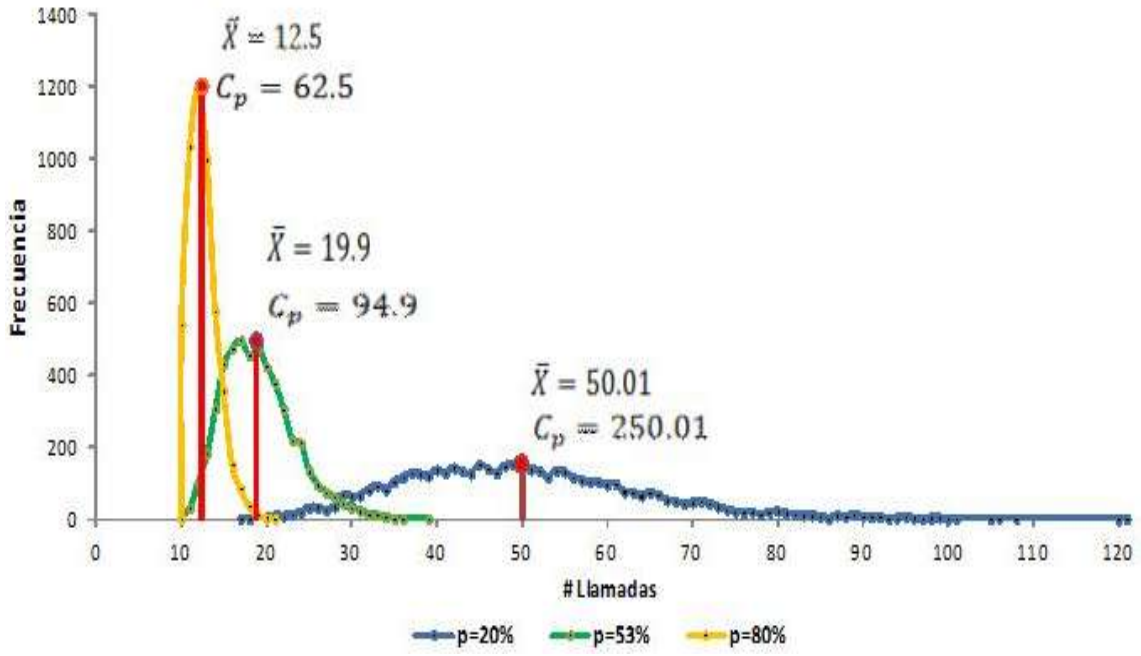
Gráfico 2.15: Polígono de frecuencias para los valores simulados de llamadas diarias p = 20% y r=10



Ahora se cuenta con un parámetro representativo que se puede comparar entre los diferentes casos que se plantearon anteriormente cuando se varió el parámetro de éxito p. En el gráfico 2.15 se encuentran estas distribuciones y sus medias señaladas, donde se observa que en cada caso la media permite hallar el centro de la distribución, además como se posee el costo por llamada, se multiplica por la media

para obtener el costo $C_p = \bar{X} * C$ asociado a las llamadas o costo promedio, señalado también en el gráfico, como referencia al contexto que se ha estado trabajando.

Gráfica 2.16: Polígono de frecuencias para los valores simulados de llamadas diarias p = 80%, p = 53% , p = 20%, con r=10



Cuando se requiera una forma resumida de reportar estos comportamientos, también se puede recurrir a una tabla que sólo contenga las medias y sus costos. El uso de la media, como se pudo analizar, es útil en representar a las distribuciones, sin embargo, es necesario usar otras medidas cuando el cálculo la media es afectada, por ejemplo, por valores en los extremos demasiado altos (o bajos) ya sea por situaciones extraordinarias o algún error inadvertido en el registro de la información, que comúnmente también es desconocido para el analista de la información.

Moda

La primera comparación que se hizo anteriormente al localizar la posición de la media fue contra aquel punto en el cual la distribución de frecuencias alcanza su máximo punto; a esta categoría se le conoce como la moda de la distribución. Las distribuciones tratadas anteriormente son ejemplos de distribuciones con una sola moda o unimodales, existen casos en los que se llegan a tener varias modas las cuales son necesarias de identificar pues usualmente son evidencia que sustenta la inferencia de que los datos provienen de un conjunto de varias poblaciones distintas; a este otro tipo de distribuciones se les conoce como multimodales.

Como ya se había identificado antes, cuando la distribución tiene la propiedad de ser simétrica entonces el centro coincide con el punto más alto de la curva de frecuencias lo que implica que la moda coincide con el valor de la media, por lo cual también se toma la proximidad de estos dos valores como indicador de la simetría de la distribución. Retomando el último comparativo de distribuciones en el contexto de llamadas de venta, se obtuvieron las modas correspondientes con la función MODA(), en los casos que se varió la probabilidad de éxito p a la cuales se calcularon la media y costos anteriormente.

Tabla 2.13 : Media y moda bajo variaciones en el parámetro p			
P	media	Moda	costo asociado a $\bar{X}$
0.2	50	45	250
0.53	19	17	95
0.8	12.5	12	62.5

Los valores de la Tabla 2.13 indican que la mayor asimetría se observa a medida que se disminuye la probabilidad de éxito. Adicionalmente la relación que guardan estos parámetros también indica la dirección de la asimetría, pues cuando $\bar{X} > Moda$ entonces se tiene una asimetría positiva es decir, se observara una concentración en el extremo izquierdo del rango de la muestra mientras que en el lado opuesto la curva del extremo (o cola) será más larga. Cuando la relación sea $\bar{X} < Moda$ entonces la asimetría es negativa y el comportamiento será inverso tanto en la concentración de la muestra como en sus colas.

Simulación de distribuciones multimodales

Para proponer ejemplos idóneos al tema, se simularan muestras con más de una moda, a partir de k distribuciones, las cuales la media μ es su parámetro de localización, para que al elegir los valores de estos parámetros, la tendencia central de las distribuciones se encuentre lo suficientemente separadas entre sí. A partir de estas distribuciones se debe simular el hecho de que un nuevo individuo puede pertenecer a cualquiera de las distribuciones propuestas con una probabilidad p_k utilizando la técnica antes vista para simular un número finito de categorías sin orden.

Para ejemplificar este método se eligieron como las distribuciones a simular a dos variables aleatorias normales $X_1 \sim N(\mu_1, {\sigma_1}^2)$, $X_2 \sim N(\mu_2, {\sigma_2}^2)$, donde se supone inicialmente que pueden provenir de cualquier distribución con igual probabilidad es decir $p_1 = p_2 = 1/2$. Cada variable Normal será simulada por el método de la transformada inversa, el cual toma un valor aleatorio U uniforme en $(0,1)$, y simula el valor de X por medio de aplicar un método numérico que aproxime la función $F_X^{-1}(U)$, inversa de la función de distribución $F_X(x)$.

La forma de hallar numéricamente la inversa de la función de distribución de una Normal, se encuentra ya definida en Excel® por medio de la función INV.NORM()[13], que recibe como parámetros: Un valor p que indique la probabilidad a la cual se cumple $F_X(x) = p$, el cual se considera como el valor simulado de la variable U.

Además como parámetros se introducen la media teórica μ de la distribución y la desviación estándar σ, este último relacionado con la amplitud que tendría teóricamente la función de densidad y que se reflejará en las simulaciones, por medio de la amplitud de en la curva de frecuencias

En una hoja de cálculo se debe colocar el valor de los parámetros en celdas individuales, para después generar una columna con naturales consecutivos i u otra

[13] El nombre de las funciones estadísticas, y las distribuciones más comunes conocidas en estadística que soporta el software Excel® llegan a cambiar entre versiones, por lo cual se recomienda revisar la documentación oficial, para cada función de este tipo.

especie de identificador de los individuos de la muestra, además se agrega una segunda columna con valores aleatorios U en $(0,1)$, con la función ALEATORIO() y una columna con la formula INV.NORM(), este proceso simulará valores de la variable X_1 por lo cual deben colocarse los títulos necesarios y copiar las columnas a un lado para generar valores de la variable X_2.

La simulación final se genera a través de decidir bajo la regla: si $U < p_1$ entonces el valor pertenece a X_1 , esto se logra usando la función SI() pues permite introducir la regla ALEATORIO() <"celda con p1" y arroja el valor X_1 en caso de que los valores generados cumplan la regla y en caso contrario arroja el valor X_2, regla similar con la que se simularon las provincias de Canadá sin embargo aquí se simplifica pues sólo se tienen 2 categorías. Los valores elegidos para los parámetros con los que se desarrolla la simulación, se encuentran en la Tabla 2.14, donde se tienen como valores iniciales propuestos, las medias lo suficientemente separadas y una desviación estándar baja, para poder identificar las distribuciones una vez mezcladas.

Tabla 2.14: Parámetros de las variables individuales X_1 y X_2

Distribución	X_1	X_2
Media	50	80
Desviación estándar	10	10
Peso	50%	50%

El tamaño de cada vector se tomó como $n = 5000$, generando así una muestra la cual se puede ver en la siguiente imagen, bajo la estructura una hoja cálculo propuesta. Otro detalle a considerar es que en este proceso todas las celdas son fórmulas por lo cual cualquier actualización a la hoja de cálculo genera una nueva muestra, y el cambio en cualquiera de los parámetros se verá reflejado inmediatamente.

Estructura en hoja de cálculo para el ejemplo de muestras multimodales.

	A	B	C	D	E	F	G	H	I
1	Distribuciones multimodales								
2		Parámetros							
3	Distribución	X_1	X_2						
4	Media	50	80						
5	Desv. estandar	10	10						
6	Peso	50%	50%						
7									
8	Simulacion de la variable X_1				Simulacion de la variable X_2				Simulación final
9	i	U_{1i}	X_{1i}		i	U_{2i}	X_{2i}		Yi
10	1	39.6%	47.4		1	36.1%	76.4		47.4
11	2	96.8%	68.5		2	53.8%	81.0		68.5
12	3	77.8%	57.7		3	8.0%	66.0		66.0
13	4	61.6%	52.9		4	74.3%	86.5		86.5
14	5	65.7%	54.0		5	42.7%	78.2		54.0
15	6	60.3%	52.6		6	15.3%	69.8		69.8
16	7	23.4%	42.8		7	69.3%	85.0		85.0
17	8	12.7%	38.6		8	86.5%	91.1		38.6
18	9	73.5%	56.3		9	82.4%	89.3		56.3
19	10	2.8%	30.8		10	63.5%	83.4		30.8
20	11	0.6%	24.9		11	22.8%	72.5		24.9
21	12	60.2%	52.6		12	59.2%	82.3		82.3
22	13	80.2%	58.5		13	69.2%	85.0		85.0
23	14	82.8%	59.5		14	57.1%	81.8		81.8
24	15	42.0%	48.0		15	86.8%	91.2		91.2
25	16	67.7%	54.6		16	26.3%	73.7		54.6
26	17	84.4%	60.1		17	25.5%	73.4		60.1
27	18	48.3%	49.6		18	15.3%	69.8		49.6

Dado que las simulaciones corresponden a variables aleatorias continuas, entonces se grafican las 3 curvas de frecuencias de las muestras X_1, X_2 y la simulación final denotada ahora como Z, que se realizan por medio de hallar los puntos medios m_i. Por la regla de Sturge de los ejemplos anteriores, con el tamaño de muestra $n = 5000$ el número de subintervalos es $m = 20$.

La Gráfica 2.17 contiene el histograma de frecuencias de Z donde además se señalan los valores de $\bar{X}$ de la muestra y las modas M_1 y M_2, las cuales se estimaron de los puntos m_i donde las frecuencias alcanzan valores que son máximos localmente. En las distribuciones continuas la moda se define como el máximo de la distribución, sin embargo Excel® no tiene una función específica para obtener una moda con datos continuos, pues depende de los límites que definen los subintervalos seleccionados. Por otra parte en la gráfica 2.18 se hallan las curvas de frecuencias de las variables X_1 y X_2 sobrepuestas, con sus medias y modas calculadas.

Gráfico 2.17: Curva de frecuencias para los valores simulados de la distribución Z mezcla de dos variables Normales , p1=50%

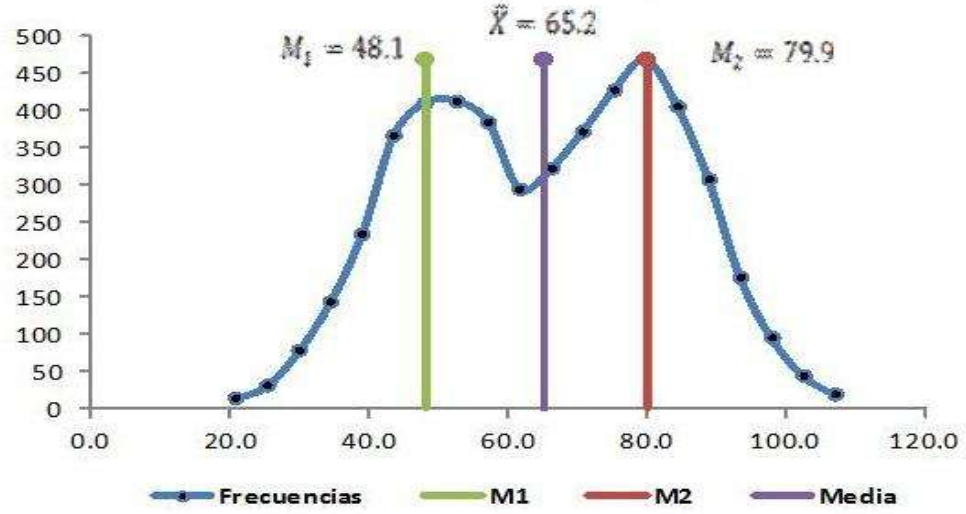
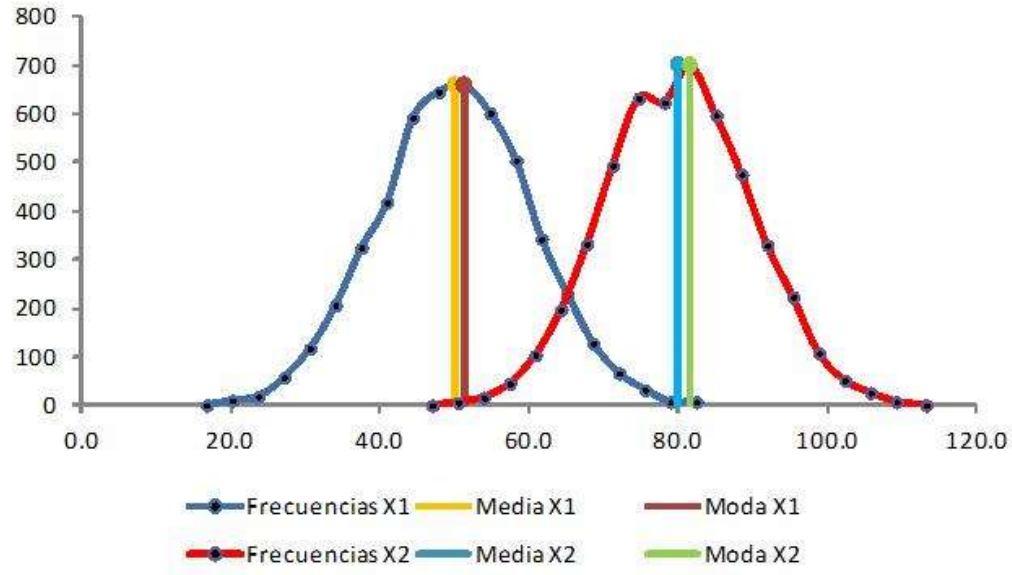


Gráfico 2.18: Curva de frecuencias para los valores simulados de distribuciones Normales X_1 y X_2



La segunda gráfica muestra cómo al ser estas distribuciones simétricas entonces la media coincide o es muy cercana al valor de la moda, las cuales son acordes al valor de los parámetros propuestos para el valor de las medias (50 y 80), de la Tabla 2.14. Otro detalle que se puede observar es el punto donde se cruzan ambas distribuciones, pues es exactamente en ese punto donde se posiciona la media de la distribución combinada, aunque esto no es una regla general pues también depende de varios factores, uno de ellos, el peso que se le atribuye a las distribuciones, por ejemplo a continuación se presentan las curvas de frecuencias de tres simulaciones de Z del mismo tamaño, en las cuales el parámetro de peso para la variable X_1 se definió como, $p_1 = 30\%,\ 60\%\ y\ 80\%$.

Gráfico 2.19: Curva de frecuencias para los valores simulados de la distribución Z mezcla de dos variables Normales , p1=30%

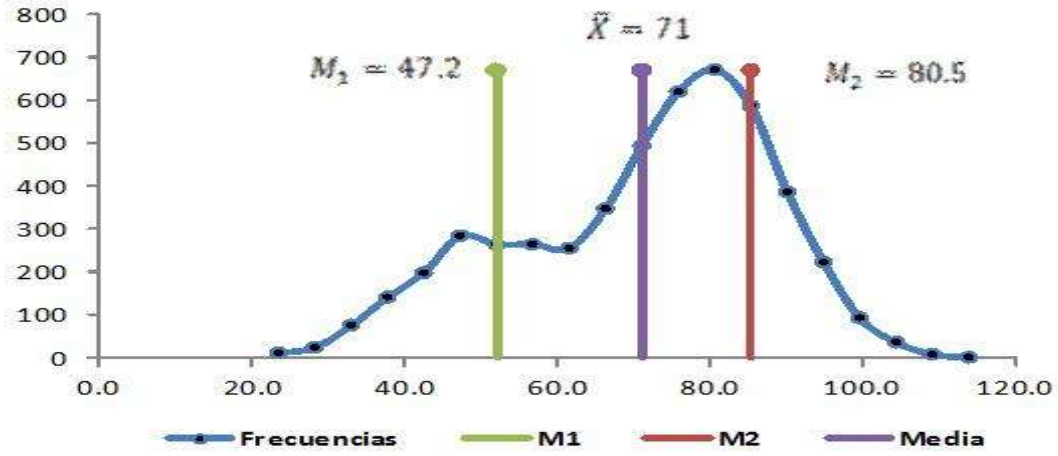


Gráfico 2.20: Curva de frecuencias para los valores simulados de la distribución Z mezcla de dos variables Normales , p1=60%

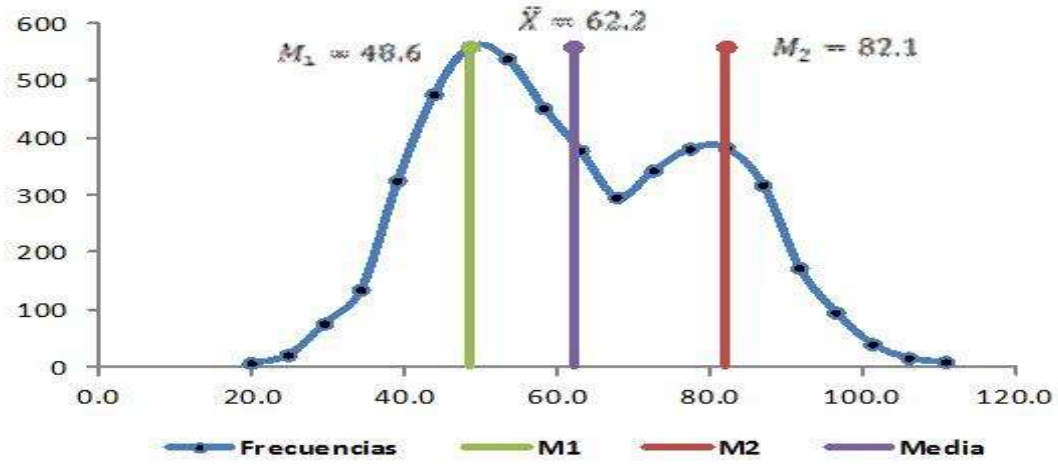
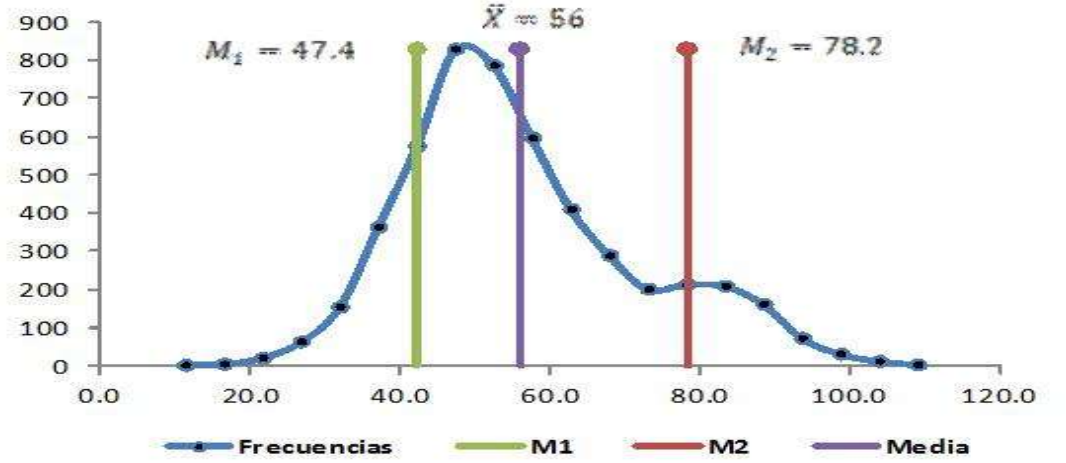


Gráfico 2.21: Curva de frecuencias para los valores simulados de la distribución Z mezcla de dos variables Normales , p1=80%



Como se puede observar en las gráficas, la proporción en la que aparece una distribución o la otra provoca que el valor de la media cambie drásticamente, sin embargo, no ha cambiado en esencia ninguna de las distribuciones origen. En los casos en que la proporción de una de las poblaciones es lo suficiente mente cercana a 0 o a 1 (por ejemplo $p_1 = 80\%$), entonces se vuelve cada vez más complicado diferenciar a las distribuciones e incluso se mezclan de tal manera que llegan a aparentar ser una distribución de una única población con un alto grado de asimetría.

En conclusión el estudio de la moda lleva un interés por su misma definición tanto en la teoría estadística como en cualquier contexto de aplicación al ser el valor de mayor aparición. La experimentación con el cambio en los demás parámetros se deja libre a la aplicación, y aunque este tema fue visto fuera de algún contexto la ventaja de usar la distribución Normal es que permite asociarla a una cantidad suficiente de ejemplos biométricos y financieros, por lo cual la mezcla de poblaciones se puede relacionar con mezclas de poblaciones de diversos orígenes geográficos, tipos de clientes, o portafolios financieros conformados por varios instrumentos que son valuados en base a diferentes *commodities*[14] o índices.

Mediana

La definición del cálculo de la mediana requiere una revisión previa en el tema de los estadísticos de orden, los cuales son de gran importancia para ahondar en temas de interpretación comparativa contra la media y la moda. Los estadísticos de orden se definen como los valores de una muestra x ordenados de menor a mayor, definidos como las variables $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$, es decir $X_{(1)}$ es el mínimo de la muestra y $X_{(n)}$ el máximo.

Ahora supóngase una muestra cualquiera de tamaño n, $X = \{x_i\}_{i=i}^{n}$ proveniente de una población con función de distribución $F_X(x)$, y función de densidad $f_X(x)$ entonces la mediana se define como el valor Me_X que cumple

[14] Termino en inglés utilizado para referenciar a materiales básicos y productos primarios que pueden ser vendidos o comprados.

$$\boldsymbol{F_X(x) = P(X \leq Me_X) = \int_{-\infty}^{Me_X} f_X(x)dx = \frac{1}{2}\ cuando\ f_X(x)\ es\ continua}$$

$$\boldsymbol{F_X(x) = P(X \leq Me_X) = \sum_{i=1}^{Me_X} P(X = x_i) = \frac{1}{2}\ cuando\ f_X(x)\ es\ discreta}$$

La Mediana de la muestra se encuentra dependiendo de si n es un número par o impar ya que es valor que se encuentra en la mitad de los estadísticos de orden, y se define en función de estos como sigue:

$$\boldsymbol{Mediana = Me_X} = \begin{cases} \boldsymbol{X_{\left(\frac{n+1}{2}\right)}}, & si\ n\ es\ impar \\ \dfrac{\left(\boldsymbol{X_{\left(\frac{n}{2}\right)}} + \boldsymbol{X_{\left(\frac{n}{2}+1\right)}}\right)}{\boldsymbol{2}}, & si\ n\ es\ par \end{cases}$$

Donde Me_X cumple $F_n(Me_X) = 1/2$, indicando entonces el punto donde la muestra acumula al 50% de los individuos. Otra propiedad importante es que el valor de la mediana tiene un menor impacto bajo los valores extremos de la distribución como el máximo o el mínimo por lo cual es usado, frecuentemente en la práctica cuando la distribución tiende a tener una forma asimétrica y en la cola alargada se encuentran frecuencias considerables cerca de los extremos, lo cual es conocido como una cola pesada

Por otro lado la aparición de un valor extremo demasiado alto, tendrá impacto drástico para el valor de la media, mientras que en el cálculo de la mediana impacta en un mismo peso que cualquiera de las observaciones.

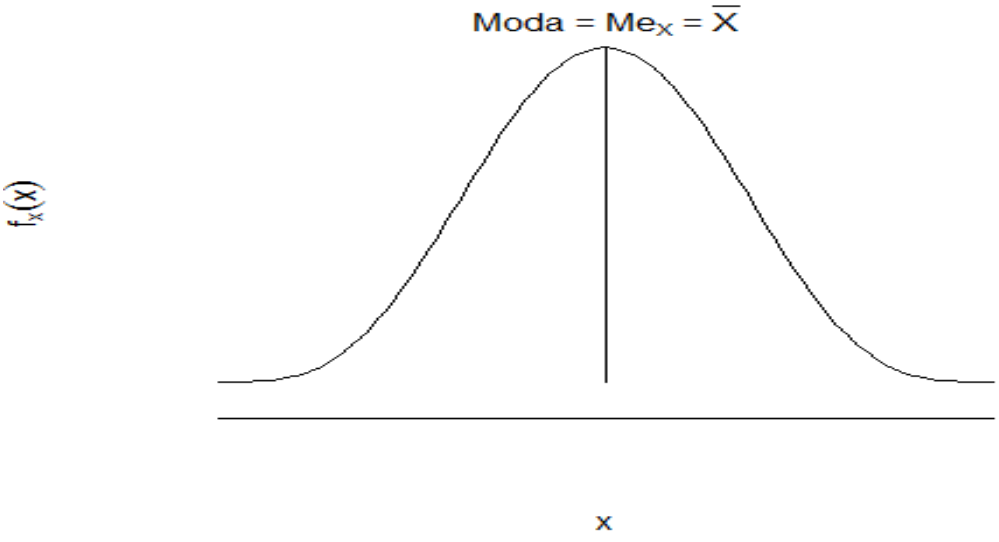
En un ambiente de hoja de cálculo como el que se ha venido utilizando, la mediana se puede hallar a través de una formula homónima, la función MEDIANA(), mientras que la comprobación a través de los estadísticos de orden se puede realizar de manera sencilla, seleccionando la columna que contenga la muestra para indicar ordenarlos de menor a mayor. Si la muestra está indexada con números de 1 a n , entonces si n es par el renglón $\frac{n}{2}$ contendrá el valor de la mediana, o en caso de n impar se realiza el promedio especificado en la fórmula.

La relación de la media y la mediana evalúa la forma de la distribución, puesto que en una población con distribución teórica perfectamente simétrica, las frecuencias a partir del punto máximo o moda, será la misma de ambos lados por lo cual la mediana siempre se posicionará en el centro de este tipo de distribuciones, por tanto la media y la moda coincidirán también con este valor.

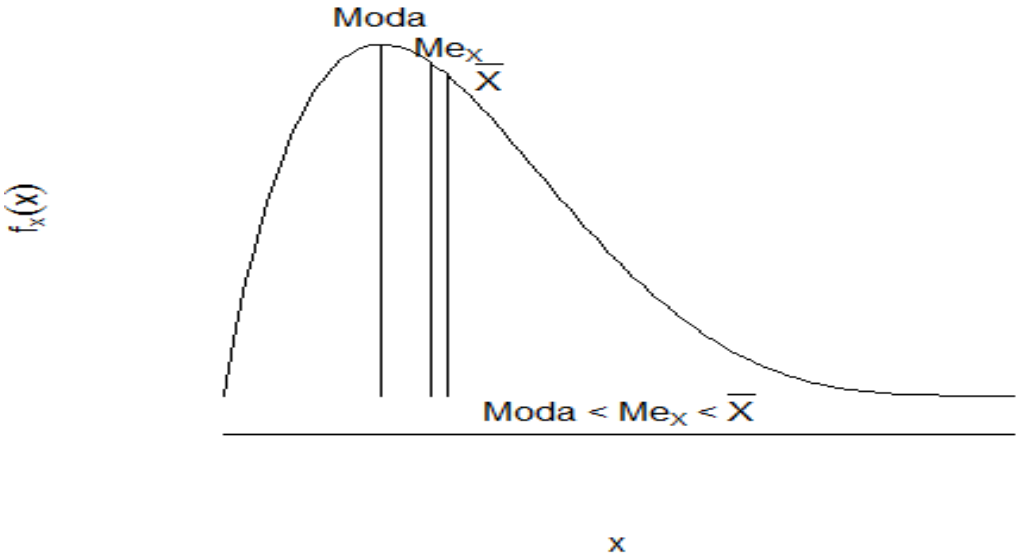
En el caso de las distribuciones asimétricas y unimodales se ha visto que la media es afectada por los valores altos de la distribución, lo que provoca que se separe del valor de la moda, en el caso de la mediana es similar pues la proporción de los individuos es arrastrada hacia donde se encuentre la moda y por consiguiente la mediana; aunque como se mencionó el impacto es menor por lo valores altos y la mediana se posiciona entonces siempre entre la moda y la media, guardando las relacionadas mostradas en las Gráficas 2.23 y 2.24, las cuales muestran las distribuciones teóricas generadas de

variables Beta asimétricas en el software R®, mientras que la gráfica 2.22 ejemplifica el caso de las distribuciones simétricas.

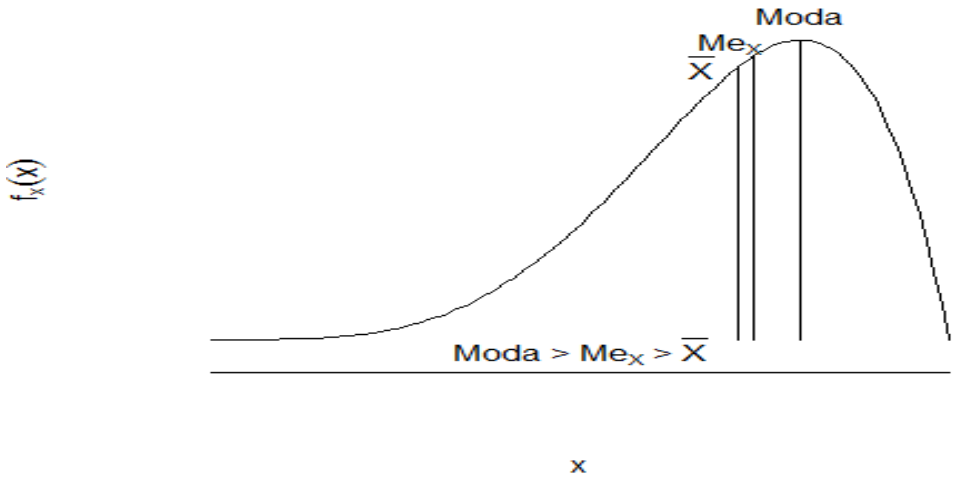
Gráfica 2.22:Densidad simétrica
media,mediana y moda coincidentes



Gráfica 2.23:Densidad asimétrica positiva
media,mediana y moda indicadas



Gráfica 2.24 :Densidad asimétrica negativa media,mediana y moda indicadas



Cuando la distribución de estudio provenga de una densidad multimodal como los casos simulados al revisar la moda, entonces el efecto en el valor de la media que tendrá será similar a los ilustrados, cumpliendo con las relaciones mencionadas. Este efecto sobre la mediana es debido a que la proporción de la variable más pesada acumulará un 50%.

Lo anterior se puede comprobar retomando la hoja de cálculo con las simulaciones que se hicieron para distribuciones con dos modas a partir de la generación de dos variables aleatorias normales X_1 y X_2, aunque ahora bajo nuevos escenarios en el parámetro p_1; la ventaja radica en la actualización inmediata de la hoja completa al introducir nuevos valores de p_1 generando nuevas muestras y las actualización de las gráficas; los valores propuestos para acercarse más al caso extremo, son $p_1 = 15\%, 35\% \; y \; 90\%$. Las gráficas siguientes contienen una realización de estas distribuciones mezcladas, donde además se hallan trazadas la media, la mediana y las modas correspondientes.

Gráfico 2.25: Curva de frecuencias para los valores simulados de la distribución Z mezcla de dos variables Normales , p1=15%

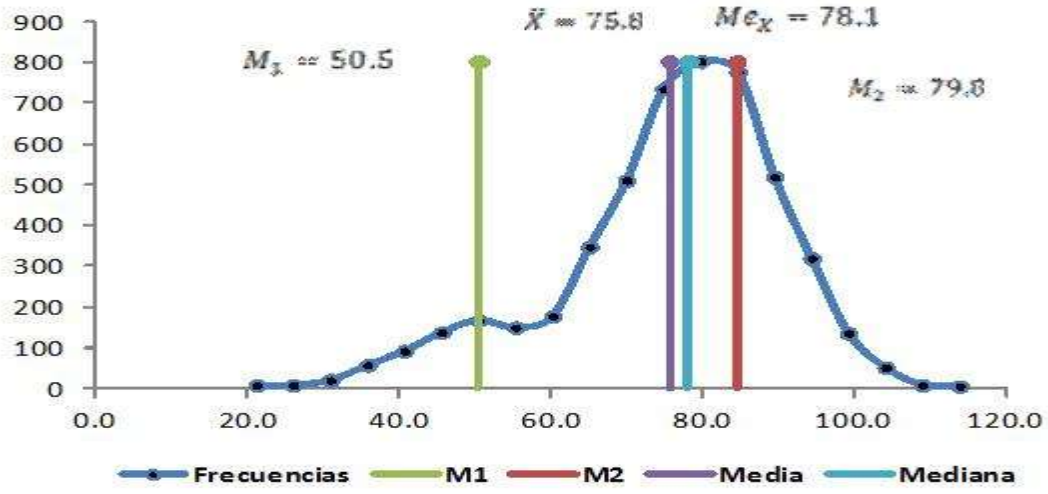


Gráfico 2.26: Curva de frecuencias para los valores simulados de la distribución Z mezcla de dos variables Normales , p1=35%

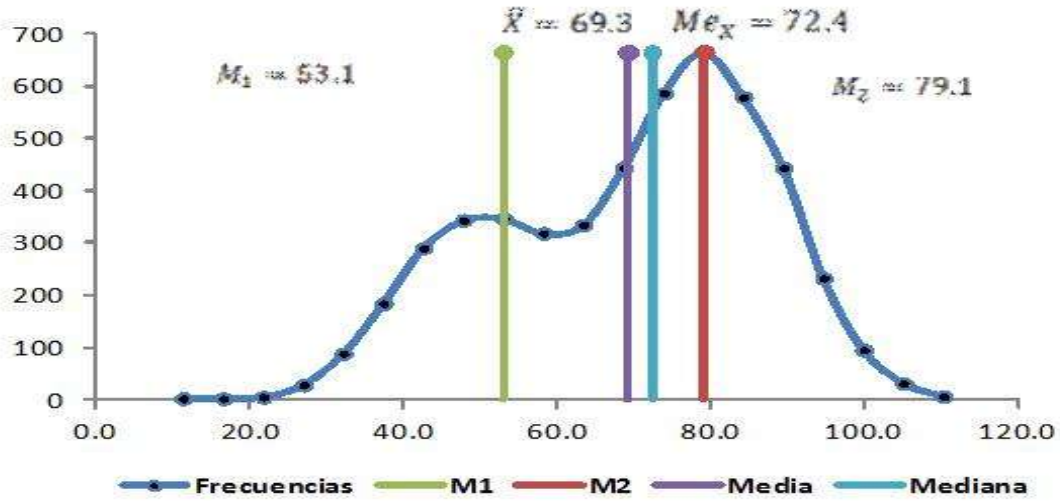
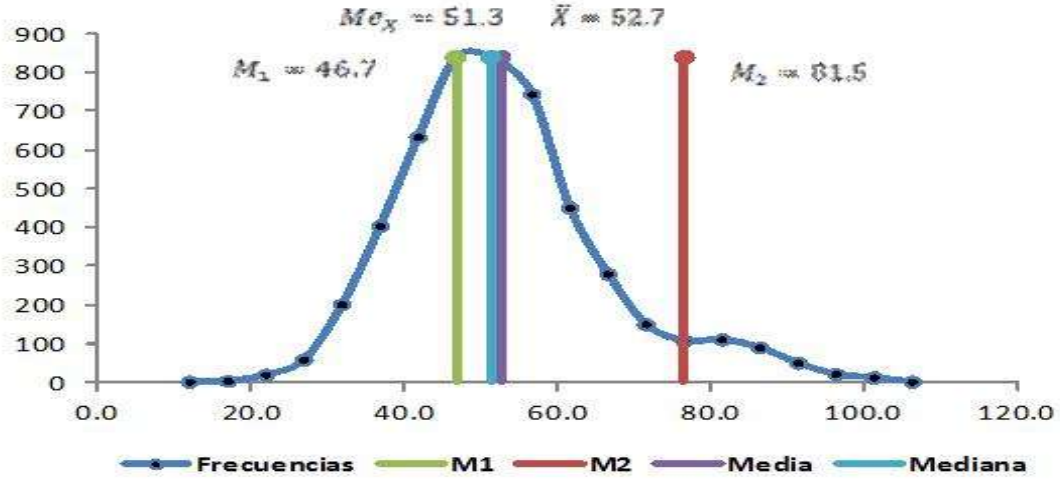


Gráfico 2.27: Curva de frecuencias para los valores simulados de la distribución Z mezcla de dos variables Normales , p1=90%



En las gráficas se puede observar el efecto en el movimiento de la mediana con respecto a la media que tiene la proporción en la que se encuentra cada una de las poblaciones, ya que se observa también en el caso $p_1 = 35\%$ que los valores de la mediana y la mediana tienen una mayor separación a comparación de los demás casos. Resumiendo todos los escenarios hasta ahora vistos y calculando todos sus parámetros, la Tabla 2.15 contiene un resumen en el cual se puede mostrar la información de todas las gráficas hasta ahora vistas y dar un panorama más general de las simulaciones al ver de forma holística el efecto que tiene las mezcla de dos poblaciones con distribución Normal, manteniendo constantes los parámetros μ_i y σ_i.

Tabla 2.15 : resumen de los escenarios vistos para la mezcla de dos poblaciones normales					
p_1	p_2	$\overline{X}$	Me_X	M_1	M_2
15%	85%	75.8	78.1	50.5	79.8
30%	70%	71	74.4	47.2	80.5
50%	50%	65.2	66.04	48.1	79.9
60%	40%	62.2	59.7	48.6	82.1
80%	20%	56	52.9	47.4	78.2
90%	10%	52.7	51.3	46.7	81.5

Puesto que la generación de más escenarios y su revisión se puede realizar con facilidad bajo la estructura propuesta, se puede concluir que al estudiar cada una de estas medidas de la tendencia central de una muestra, se podrá realizar con claridad, y generar ejemplos sencillos o elaborados sobre el tema. Ahora se revisará el estudio en otro aspecto de las distribuciones que es la forma de medir la forma en que se están dispersando los individuos en todo el rango de la muestra.

Medidas de dispersión

La forma en que los valores de una variable aleatoria se diseminan a través de todo su dominio, se le conoce como la dispersión de una variable aleatoria y se mide a través de observar cómo se desvían sus valores con respecto a un valor de referencia, siendo el más común la media de la distribución.

Los estudios sobre la variación de los valores obtenidos de una muestra, ha tomado una gran importancia puesto que para ciertos fenómenos, se tiene un interés cada vez mayor por la sensibilidad a los cambios externos. En ciertos casos es una prioridad la velocidad en que se modelan, diagnostican y proyectan estas variaciones. Algunos ejemplos de modelación de la dispersión y que además son de gran impacto en la sociedad son: los precios de los bienes debido a la introducción de sistemas financieros tecnológicamente robustos, que permiten la venta instantánea de bienes en enormes volúmenes, o el crecimiento de la poblaciones mundiales y la medición de la migración internacional, modelos que requieren enormes bases de datos con información a suficiente detalle demográfico político, económico, etc. y el impacto por las variaciones en el clima, las cuales pueden llegar a indicar una prosperidad para la agricultura local o un gran desastre para las naciones.

Existen una gran cantidad de estadísticos para medir la dispersión de una muestra, en esta sección se abordaran los más usuales dentro de la estadística descriptiva, es decir se considera que cada muestra proviene de una población con densidad $f(x)$, pero sin hacer algún supuesto sobre esta función, sin embargo para simplificar el estudio de los métodos descriptivos, se realizarán simulaciones a partir de distribuciones conocidas.

Amplitud total o Rango

La primera medida de diagnóstico de la dispersión se puede hallar mediante la identificación del mínimo $x_{(1)}$ y máximo $x_{(n)}$ de una muestra X, mediante conocer cuál es la distancia entre estos dos puntos como $x_{(n)} - x_{(1)}$, esta medida también es

conocida como el rango de la muestra, la cual se ha utilizado antes en este trabajo para la construccion de tablas.

Una manera de ilustrar el comportamiento del rango es retomando los ejemplos en los que se analizaba el número de llamadas necesarias para concretar una venta, donde ligeros cambios en la probabilidad de éxito p, genera un gran impacto en los puntos extremos de la muestra. Supóngase entonces que los modelos antes simulados $p = .2$,.53, .8 son muestras de empleados distintos e_i $i = 1,2,3$ los cuales se desean comparar, para resolverlo se genera una tabla similar a la Tabla 2.16 que identifique cada muestra y contenga el cálculo del rango, además como práctica se pueden agregar otros parámetros de interés en el contexto como la media de llamadas y costo medio asociado además de los máximos y mínimos correspondientes, para tener un diagnóstico más completo

Tabla 2.16 : Análisis de intentos de venta, bajo variaciones en la probabilidad de éxito p

$Empleado\ e_i$	P	$\overline{X}$	$costo\ medio\ asociado$	$x_{(1)}$	$x_{(n)}$	$x_{(n)} - x_{(1)}$
e_1	0.2	50	$250	17	121	104
e_2	0.53	19	$95	10	39	29
e_3	0.8	12.5	$62.50	10	21	11

Relacionando los datos de la tabla se interpreta que el empleado e_3, considerando a $x_{(1)}$ como el día que requirió el menor número de llamadas para cumplir su cuota de ventas, y su peor día como $x_{(n)}$, entonces en tal caso necesita hasta el doble de llamadas para terminar sus ventas. Mientras que, al enfocarse en el caso del empleado e_1, en su peor día puede llegar a necesitar más de 6 veces su número mínimo de llamadas (17) para alcanzar su cuota.

Por otro lado el mejor día ($x_{(1)}$) del empleado e_1 es de un valor menor que el de e_3 . Además de los puntos extremos del número de llamadas, se puede enriquecer el contexto al añadir también en una segunda tabla o columnas extra, los costos máximos, al multiplicar las últimas tres columnas por el costo unitario C. La Tabla 2.17 contiene los costos asociados a los estadísticos, de cada modelo de empleado.

Tabla 2.17 : Análisis de costos bajo variaciones en la probabilidad de éxito p

$Empleado\ e_i$	p	$Costo\ medio$	$Costo\ mínimo$	$Costo\ máximo$	$Rango\ de\ costos$
e_1	0.2	250	85	605	520
e_2	0.53	95	50	195	145
e_3	0.8	62.5	50	105	55

Los comentarios anteriores que comparan el comportamiento entre empleados, en términos del número de veces que representa el máximo, el mínimo y el rango en el contexto del número de llamadas, permite que se pueda concluir de la misma manera sobre los datos de la Tabla 2.17, que representan unidades monetarias, debido a que sólo se trata de un cambio en la escala, pues son multiplicados por la misma constante de costo.

Varianza muestral y poblacional

Para una variable aleatoria cualquiera X, con función de densidad $f(x)$, y exista la media teórica de la distribución $E(X) = \mu$, la varianza poblacional se define como

$$\boldsymbol{E\left((X-\mu)^2\right) = \sigma^2}$$

Supóngase una muestra $X = \{x_i\}_{i=1}^{n}$ de tamaño n, con la que se desea estimar la verdadera varianza de la población σ^2. La varianza muestral estima la varianza a través de las diferencias que existen de cada elemento contra la media aritmética $\bar{X}$, por medio de la formula

$$\boldsymbol{S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{X})^2}{n-1}}$$

En nuestra hoja de cálculo el cálculo de S^2 se realiza con la formula VAR.S()[15] tomando como parámetro la selección de donde se encuentra la muestra en la hoja de cálculo. Estos cálculos son las diferencias cuadráticas con respecto a la media como punto de referencia, para, a partir de ella, medir el comportamiento de la variable X, por ello para interpretar mejor con base en las unidades originales, se emplea la raíz cuadrada, $S = \sqrt{S^2}$ a lo que se conoce comúnmente como desviación estándar.

Cuando los datos de análisis se encuentren como datos agrupados o resumidos por una tabla similar a la Tabla 2.12, se puede utilizar una fórmula similar a la empleada para $\bar{X}$ con datos agrupados se consideran m divisiones de los datos. El cálculo se realiza como $\frac{\sum_{i=1}^{m} f_i*(m_i-\bar{X})^2}{n}$.

Una regla empírica en estadística sobre la desviación estándar conocida como la regla *68-95-99.7* establece que para variables distribuidas Normal(μ, σ^2), un intervalo a partir de la media entre $\bar{X} - S$ y $\bar{X} + S$,contendrá el 68.2% de los individuos. Cuando el intervalo es construido sobre 2 desviaciones estándar entonces el intervalo $(\bar{X} - 2S$, $\bar{X} + 2S)$ contendrá el 95.5% de la muestra, mientras que si se consideran hasta 3 desviaciones estándar entonces se encierra al 99.7% lo que simplifica la forma en que se secciona la distribución de una población Normal con base en su desviación estándar, por esta razón la regla ha sido adoptada en diversos campos como el control estadístico de procesos o en ciertos estándares como la metodología *six-sigma*.

Una forma de mostrar las medidas de dispersión gráficamente es iniciando con un ejemplo sencillo al que se irá integrando gradualmente complejidad y contexto, para esto la distribución normal ofrece una forma clásica de una distribución simétrica, por lo cual se estudiará una variable Z Normal con media $\mu = 0$ y con varianza $\sigma^2 = 1$, reutilizando la hoja de trabajo empleada en la simulación de distribuciones multimodales.

Para separar y ordenar los tópicos tratados se recomienda copiar la hoja de trabajo, y cambiar en las celdas de los parámetros de la variable x_1 los valores

[15] El nombre de las funciones llega a cambiar de acuerdo a la versión e idioma del software, por lo que se recomienda cofirmar en la documentación el nombre de la función que calcule la varianza con respecto a una muestra.

correspondientes, recordando que la fórmula INV.NORM() recibe como parámetro el valor de la desviación estándar en la lugar de la varianza, aunque en este caso particular coincide el valor.

Una vez generada la muestra, se genera la curva de frecuencias de x_1 ignorando, por el momento, las demás variables, esto develará el comportamiento de la muestra. El objetivo de emplear la curva de frecuencias en lugar del histograma en esta ocasión, tiene como función el poder señalar los parámetros calculados, sin que se sobrepongan con otros elementos como barras y le dé un formato más limpio al *template* de trabajo, aspecto importante para el entendimiento de los temas; puede verse así, si se quisiera entregar o publicar un reporte, el objetivo sería comunicar de la forma más clara el mensaje propuesto, en un gráfico esto refiere a un equilibrio entre los tipos de objetos insertados para señalar el valor de los elementos como barras, puntos, líneas, el color, tamaño de fuente y otros detalles, que si bien son temas engorrosos que tratar, también son indispensables en la práctica, por lo que se recomienda incluir un módulo que integre o discuta asuntos de esta índole[16].

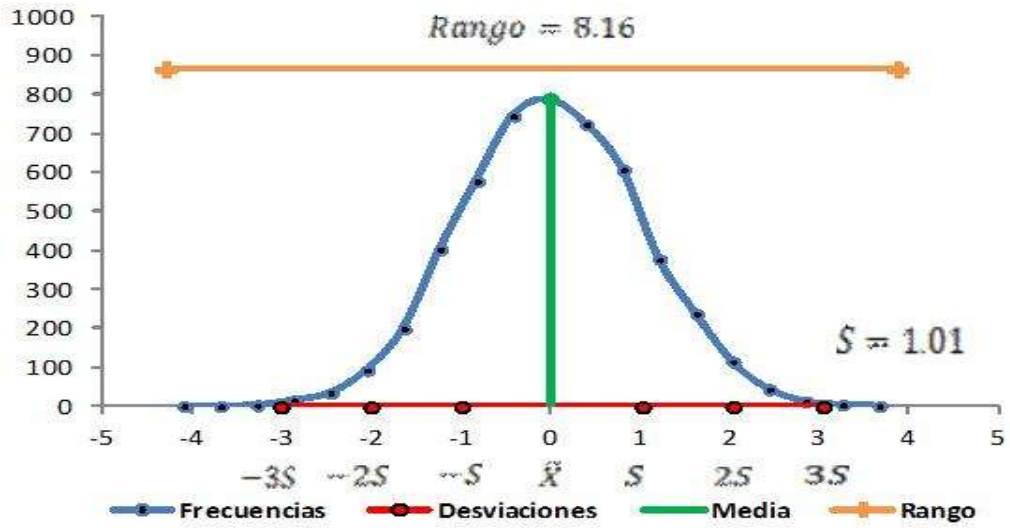
Para comprobar la regla *68-95-99.7* se pueden utilizar tanto métodos gráficos en la representación de estas distancias como cálculos en tablas para dejar más claro el concepto de lo que son y cómo se emplean las medidas de dispersión hasta ahora vistas. La forma en que se eligió representar gráficamente estas distancias es por medio de líneas horizontales que muestren, por una parte, el rango de la muestra por encima de la curva y la distancia S a partir de la media en ambos sentidos, en la base de la gráfica.

Para graficar el rango de la muestra, primero se deben colocar en alguna parte de la hoja de cálculo los puntos $\left(x_{(1)}, \max(f_i) + \varepsilon\right), \left(x_{(n)}, \max(f_i) + \varepsilon\right)$, donde ε es un valor, que permite elevar y ajustar la línea en el gráfico para diferenciarla correctamente de la curva. Al configurarlo en la gráfica de frecuencia como una serie se desplegara la línea horizontal por encima de la curva.

En el caso de la desviación estándar sólo es necesario calcular una columna con los puntos $x - 3S,\ X - 2S,\ X - S,\ X + S, x + 2S,\ X + 3S$, con una columna al lado con el valor 0 para colocarlos en la base de la gráfica. En la Gráfica 2.28 se muestra la curva de frecuencias de una realización de $n = 5000$ simulaciones de x_1 con los parámetros de media 0 y varianza 1, donde se señalan además los elementos antes mencionados.

[16] Para una referencia más extensa puede consultar el libro, de Gilbert, J. K.; Reiner M.; Nakhleh M. (2008). *Visualization: Theory and Practice in Science Education, Models and Modeling in Science Education*, Volumen 3, Editorial Springer.

Gráfica 2.28: Curva de frecuencias para 5000 valores simulados de la distribución Normal ($\mu=0, \sigma^2=1$)



Como se puede observar en el gráfico, la amplitud total da un diagnóstico de la variabilidad que puede presentar toda la muestra, mientras que las distancias en múltiplos S a partir de la media, permiten seccionar la muestra, de la misma forma que indica la regla *68-95-99.7*, que para comprobarla se calculan los porcentajes que encierran los intervalos, construidos en base a las desviaciones estándar. La siguiente Tabla muestra el número de casos que se encuentran en cada intervalo y el porcentaje que representa del total de la muestra

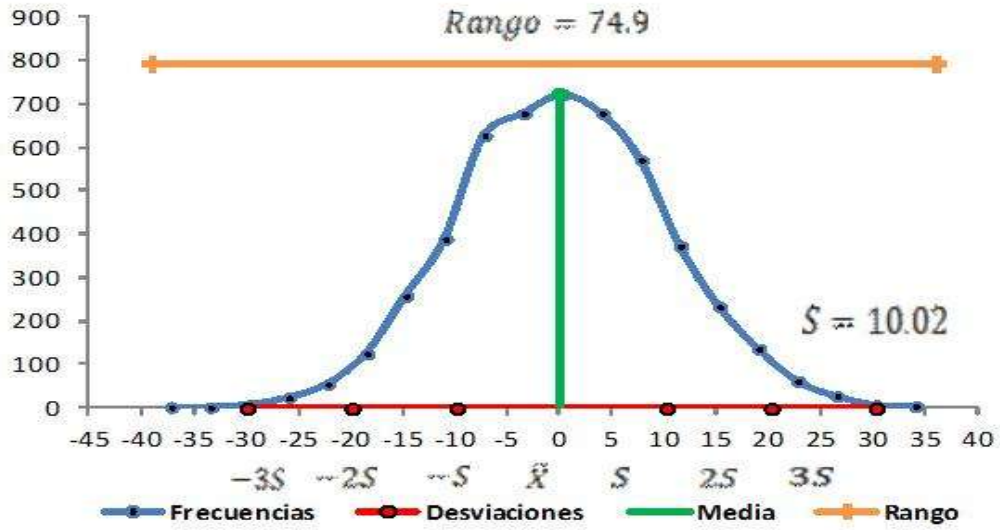
Tabla 2.18: Resumen de los intervalos basados en desviaciones estándar y sus frecuencias

Intervalo	casos	%
(μ-σ , μ+σ)	3,415	68.30%
(μ-2σ , μ+2σ)	4,783	95.70%
(μ-3σ , μ+3σ)	4,985	99.70%
Total de casos	**5,000**	**100%**

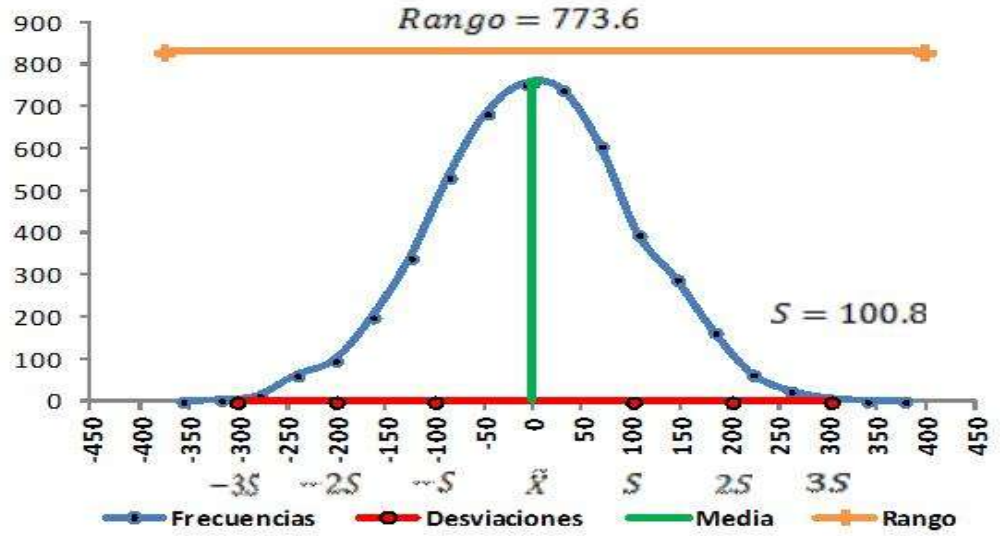
Una vez realizado esto, la modificación del valor de la desviación estándar permitirá generar suficientes muestras distintas, las cuales automáticamente actualizarán los gráficos para mostrar el efecto en el comportamiento de las muestras simuladas.

Los gráficos posteriores contienen las curvas de frecuencias de dos réplicas del ejercicio anterior bajo la misma media 0, pero asignando los parámetros $\sigma = 10$ y $\sigma = 100$, además el ajuste del gráfico permite apreciar que sólo hubo impactos en la escala de la distribución proporcional al aumento en el valor del parámetro σ, pues la relación que guardan el Rango con S es similar en los tres casos y en cada caso las distancias múltiplos de S a partir de la media se encuentran en los mismos puntos en cada curva, además los cálculos de frecuencia mostrarán que la regla *68-95-99.7* se sigue cumpliendo. Lo anterior permite mostrar de forma clara, porqué σ^2 es conocido como el parámetro de escala de la distribución Normal.

Gráfica 2.29: Curva de frecuencias para 5000 valores simulados de la distribución Normal ($\mu=0,\sigma^2=10$)



Gráfica 2.30: Curva de frecuencias para 5000 valores simulados de la distribución Normal ($\mu=0,\sigma^2=100$)



El estudio de la variabilidad también se debe llevar a cabo en distribuciones asimétricas y /o multimodales, para revisar como la regla anterior de porcentajes sólo se cumple para la distribución Normal. Para este trabajo se ha elegido trabajar con distribuciones multimodales, ya que, como se verá mas adelante con este tipo de distribuciones se pueden construir ejemplos para trabajar ambos escenarios, aunque también se puede hacer uso de las distribuciones asimétricas y con flexibilidad en cuanto a su forma del capítulo anterior, como la distribución Beta.

Para mostrar estos escenarios con una distribución multimodal, se pueden reusar los escenarios generados en la secciones anteriores, donde se introdujo una segunda variable Normal X_2 en el modelo para mezclarlas con pesos desiguales, por lo que se retomará el escenario cuando $p_1 = .15$ y $p_2 = .85$ asignando como parámetro de las

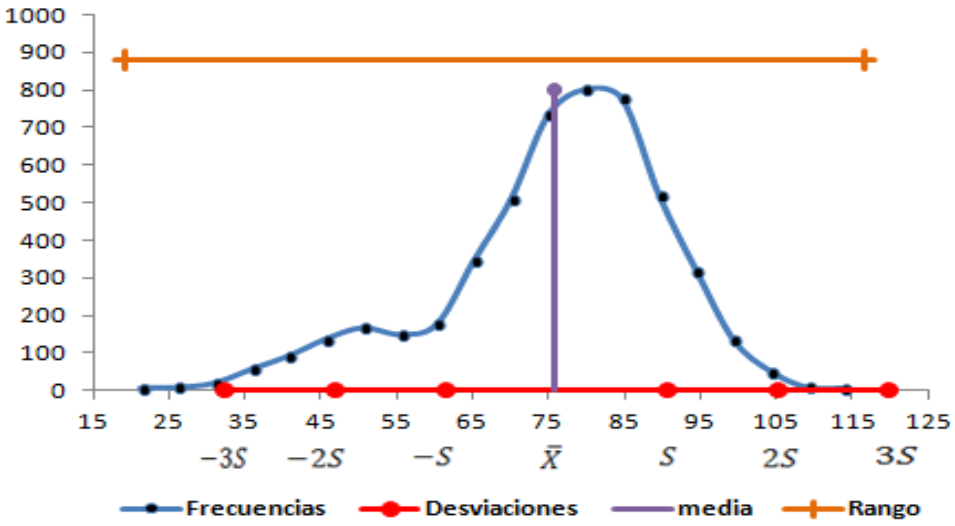
medias de las variables X_1 y X_2 como $\mu_1 = 50$ y $\mu_2 = 80$ respectivamente, que en la distribución mezcla se convierten en las modas M_1 y M_2 además se estableció una desviación estándar común $\sigma_1 = \sigma_2 = 10$.

La siguiente Tabla contiene el resumen de los estadísticos que se han revisado anteriormente junto con los intervalos definidos basados en múltiplos de la desviación estándar con respecto a la media, el número de casos que contienen en total y en porcentaje con el fin de revisar los impactos en los porcentajes para la regla antes vista, además en seguida se presenta la curva de frecuencias de la muestra en este escenario, con los elementos gráficos antes agregados.

Tabla 2.19: Resumen de una muestra de Z mezcla de dos normales, N_1 (μ_1=50,σ_1=10) , N_2 (μ_2=80,σ_2=10)

p_1	15%	S^2	212.6
p_2	85%	S	14.6
n	5,000	$\#(\bar{X}-S,\bar{X}+S)$	3,599
$\bar{X}$	75.8	$\#(\bar{X}-2S,\bar{X}+2S)$	4,712
Me_x	78.1	$\#(\bar{X}-3S,\bar{X}+3S)$	4,978
M_1	50.5	$\%(\bar{X}-S,\bar{X}+S)$	72.00%
M_2	79.8	$\%(\bar{X}-2S,\bar{X}+2S)$	94.20%
$x_{(n)}-x_{(1)}$	97.6	$\%(\bar{X}-3S,\bar{X}+3S)$	99.60%

Gráfica 2.31: Curva de frecuencias para 5000 valores simulados de la Mezcla de dos Normales N_1 (μ_1=50,σ_1=10) , N_2 (μ_2=80,σ_2=10)

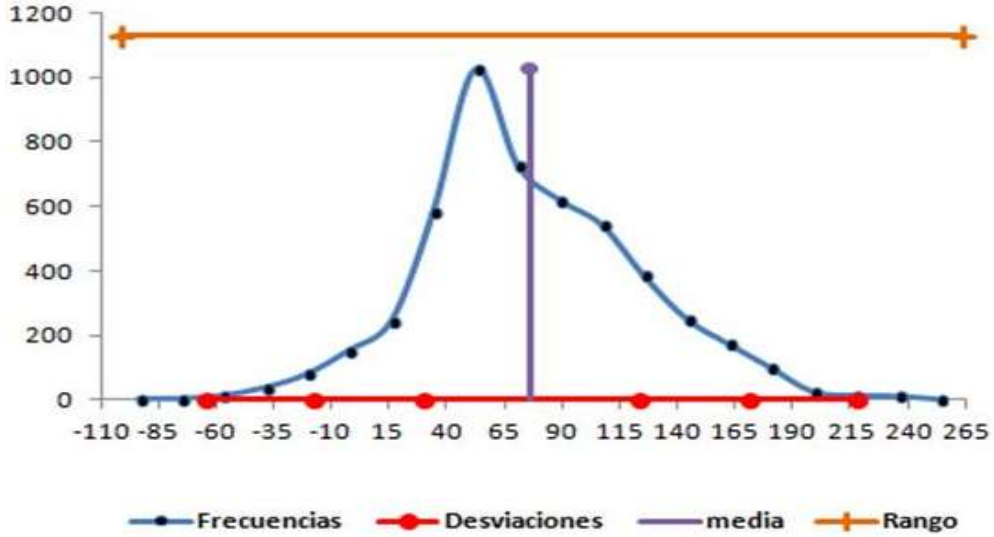


Analizando los elementos de la tabla se puede notar que los porcentajes de valores simulados que contienen los intervalos de las desviaciones son similares a los de la regla, sin embargo, con variaciones, por ejemplo una mayor concentración en el centro al estar el 72% de los valores, mientras que un porcentaje más cerrado comparado con el caso Normal con 94.2%. Se puede observar en el gráfico que el intervalo con tres desviaciones estándar sale del rango de la muestra por la derecha, por lo cual, a

pesar de ser el más cercano en cuanto a porcentajes con respecto a la regla, no se puede considerar como una intervalo que parta o seccione al rango de la muestra como en los casos anteriores.

Ahora lo que se puede realizar es ver el efecto que tiene el cambio en cualquiera de los parámetros, en este caso se decidió dejar ahora los pesos $p_1 = 15\%$, $p_2 = 85\%$ y las medias $\mu_1 = 50, \mu_2 = 80$ como constantes y primero variar los parámetros de las dispersiones individuales, por lo que en otro ejercicio se actualizó el modelo con los valores $\sigma_1 = 10$ y $\sigma_2 = 50$. Como se puede apreciar en la Gráfica 2.32 el resultado de las simulaciones derivó en la aparición de una sola moda $M = 52.5$, en la posición de la media de la distribución X_1 , a pesar de tener menor peso aunque esto le da su forma asimétrica positiva.

Gráfica 2.32: Curva de frecuencias para 5000 valores simulados de la Mezcla de dos Normales N_1 (μ_1=50,σ_1=10) , N_2 (μ_2=80,σ_2=50)



La explicación del cambio en el comportamiento se puede exponer mediante el uso de las curvas de frecuencias sobrepuestas las cuales deben estar referenciadas fijas a las series de X_1y X_2 con el objetivo de que se actualicen automáticamente junto con la generación de una nueva simulación. También es recomendable en este punto empezar a incluir una tabla con fórmulas, que resuma los estadísticos antes revisados para las tres muestras simuladas, para siempre incluir una visión evaluativa general de cada muestra. Aprovechando las características de una hoja de cálculo se pueden generar tablas similares a la Tabla 2.19, solo copiando las fórmulas de la primer columna, pegándola en algún espacio cercano al análisis de frecuencias y ajustando los parámetros de cada formula copiada, a cada muestra correspondiente.

La Gráfica 2.33 contiene los valores generados de las variables X_i que mezcladas fueron la base para la simulación anterior, ahora se puede ver que la amplitud de la segunda muestra x_2 abarca en su totalidad a la simulación de x_1, por consiguiente la aportación de x_1 son valores mayormente concentrados mientras que x_2 genera valores con una mayor lejanía alrededor de su media. Como resultado final sobre la variable Z se observa una desviación estándar promedio menor que si sólo se considerara x_2, pues los valores seleccionados a partir de x_1 aportan valores

cercanos a la media que reducen la suma de cuadrados de S_z^2 y en consecuencia el valor de S_z.

Gráfica 2.33: Curva de frecuencias para 5000 valores simulados de las distribuciones $X_1 \sim N_1$ (μ_1=50,σ_1=10) y $X_2 \sim N_2$ (μ_2=80,σ_2=50)

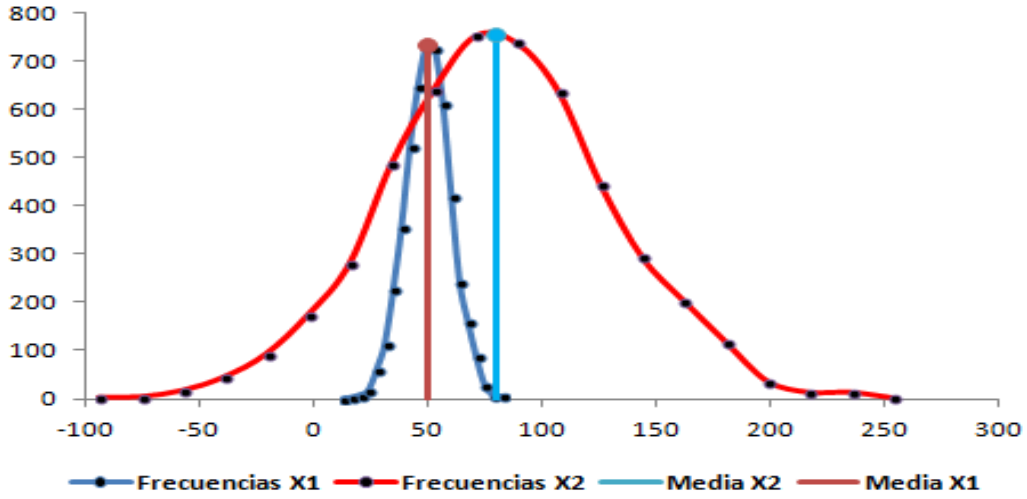


Tabla 2.20: Resumen de estadísticos, para los valores simulados de la distribución X_2, μ_2=80, σ_2=50, X_1 μ_1=50, σ_1=10, y Z p_1=15% p_2=85%

Estadístico	X_1	X_2	Z
n	5,000	5,000	5,000
$\bar{X}$	50.1	80.6	76
Me_x	50.2	80	69.9
M_1	49.4	70.8	52.5
$x_{(n)} - x_{(1)}$	72.7	365.7	365.7
S^2	96.6	2443.8	2215.4
S	9.8	49.4	47.1
$\#(\bar{X} - S, \bar{X} + S)$	3,439	3,466	3,563
$\#(\bar{X} - 2S, \bar{X} + 2S)$	4,756	4,759	4,742
$\#(\bar{X} - 3S, \bar{X} + 3S)$	4,982	4,987	4,974
$\%(\bar{X} - S, \bar{X} + S)$	68.80%	69.30%	71.30%
$\%(\bar{X} - 2S, \bar{X} + 2S)$	95.10%	95.20%	94.80%
$\%(\bar{X} - 3S, \bar{X} + 3S)$	99.60%	99.70%	99.50%

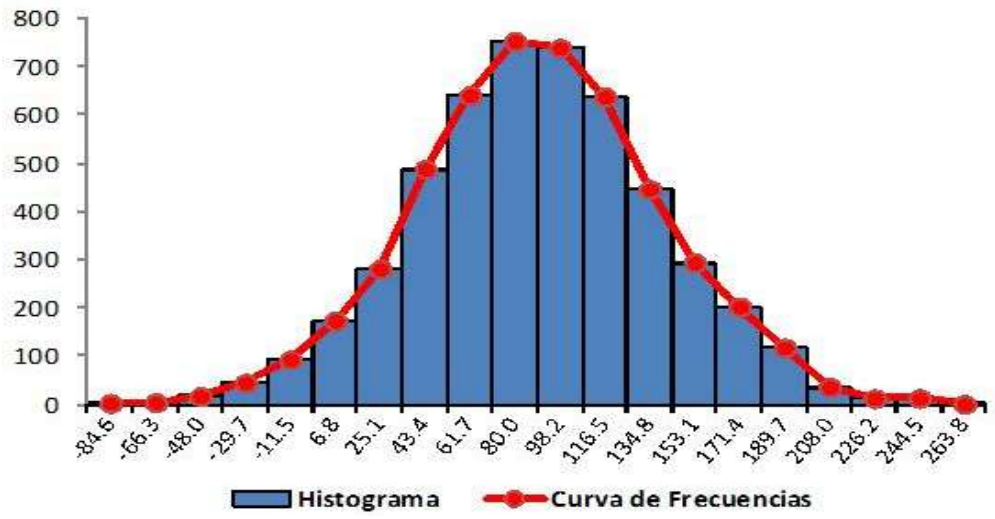
Los valores de la Tabla 2.20 muestran cómo en las variables Normales a pesar de tener distintos parámetros, los intervalos definidos con las desviaciones contienen los mismos porcentajes de individuos entre si y acorde a la regla. Por otra parte la distribución mezcla tiene variaciones importantes en los porcentajes considerando una desviación de distancia de la media, aunque cada vez más similares a los casos Normales conforme se toman más desviaciones lejos de la media.

Los cambios en los parámetros podrán ser observados en ambas gráficas y la tabla de manera automática, lo que permite evaluar diversos escenarios para las muestras x_1, x_2 y el impacto en su distribución de frecuencias, tanto de manera individual como una vez mezcladas para formar Z.

Un impacto que cabe resaltar en este punto, es que para la realización de x_2, en particular, el valor de su moda M_2 no está exactamente en el valor de $\bar{X}_2$ debido a la forma de cálculo de la moda pues el aumento en la desviación estándar de la muestra también ocasionó que el rango se incrementara, y a su vez la longitud de los subintervalos seleccionados para desarrollar la curva de frecuencias como $\frac{365.7}{20} = 19.3$, entonces para el intervalo con mayor frecuencias $(61.7 - 80)$ la estimación del punto medio se da en el valor 70.8.

Se recomienda que al trabajar este tipo de distribuciones se conserve de forma separada, la muestra generada de cada variable Normal, ya que si se desea mostrar de una forma más simple la identificación de la moda, se puede realizar mediante el desarrollo de un diagrama similar a la Gráfica 2.11, que en este caso se aplicó a la distribución de frecuencias de la muestra x_2.

Gráfico 2.34: Curvas de frecuencias para los valores simulados de la distribución $X_2 \sim N(\mu_2=80, \sigma_2=50)$



Ahora falta revisar un efecto en el cambio de los parámetros y es el cambio en las medias, puesto que este parámetro permite determinar el centro de las distribuciones, un cambio en su valor generará un desplazamiento a lo largo del eje X, del centro de la distribución de frecuencias. Cambiando sólo el parámetro $\mu_1 = 200$, se obtiene un efecto radical en la muestra de Z pues ahora la intención es separar x_1 de x_2. Una realización del ejercicio, la curva de frecuencias y el resumen de los modelos se pueden observar a continuación, donde ahora se pueden diferenciar claramente las dos distintas poblaciones.

Gráfica 2.35: Curva de frecuencias para 5000 valores simulados de la distribución mezcla de dos Normales $X_1 \sim N_1$ (μ_1=200,σ_1=10) y $X_2 \sim N_2$ (μ_2=80,σ_2=50) p_1=15%

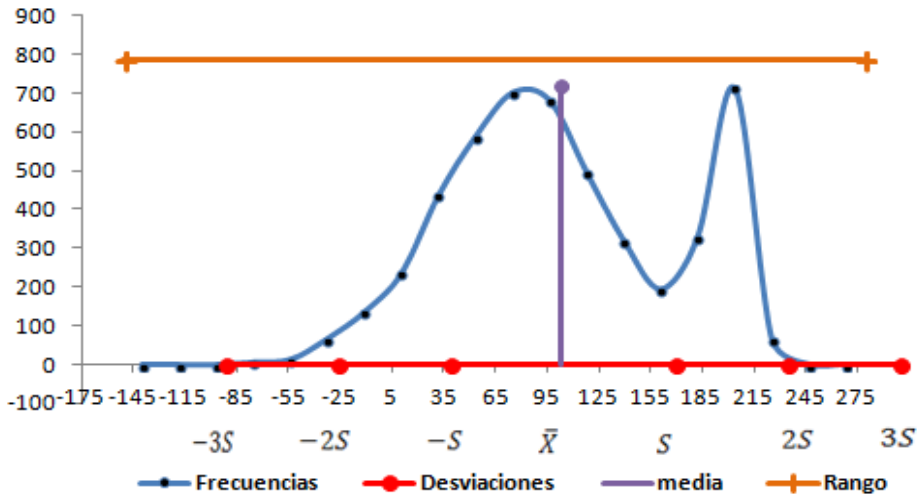


Tabla 2.21: Resumen de estadísticos, para los valores simulados de la distribución X_2, μ_2=80, σ_2=50, X_1 μ_1=200, σ_1=10, y Z p_1=15% p_2=85%

Estadístico	X_1	X_2	Z
n	5,000	5,000	5,000
$\bar{X}$	199.9	79.8	103.4
Me_x	199.9	79.9	95.1
M_1	200.1	74.7	203.7
$x_{(n)} - x_{(1)}$	72.5	429.9	429.9
S^2	98.7	2464.7	4250.6
S	9.9	49.6	65.2
$\#(\bar{X}-S, \bar{X}+S)$	3,433	3,445	3,065
$\#(\bar{X}-2S, \bar{X}+2S)$	4,766	4,750	4,932
$\#(\bar{X}-3S, \bar{X}+3S)$	4,987	4,984	4,999
$\%(\bar{X}-S, \bar{X}+S)$	68.70%	68.90%	61.30%
$\%(\bar{X}-2S, \bar{X}+2S)$	95.30%	95.00%	98.60%
$\%(\bar{X}-3S, \bar{X}+3S)$	99.70%	99.70%	99.90%

Los datos de la Tabla 2.21 indican cómo la media de z se sitúa en un punto entre las distribuciones origen, por lo cual las distancias a la media serán mayores, implicando que el valor de S_z sea mayor incluso que $\sigma_2 = 50$, mientras que la amplitud es prácticamente el mismo valor para z que para x_2, debido a que se desplazó la distribución de x_1 al mover la media μ_1 cerca del extremo de x_2, sin embargo, al tener

una desviación estándar baja sigue concentrando los simulaciones dentro del rango de X_2.

En cuanto a los porcentajes de cada intervalo basado en desviaciones se notan las diferencias en los porcentajes, pues ahora el primer intervalo contiene un menor porcentaje mientras que el segundo es mayor, por su parte en el último intervalo se observa desde el grafico que el intervalo sale del rango de valores al tomar 3 desviaciones por lo cual no se puede dividir el rango con base en este último intervalo.

Con lo anterior ahora se tiene tanto un entendimiento global de la simulación, además del impacto y funcionamiento de cada uno de sus componentes, tanto para generar nuevos ejemplos de mezcla de dos poblaciones normales, y con ello revisar una serie de ejemplos, brindando la posibilidad de generar fácilmente ejercicios a una población estudiantil, ya que incluso bajo los mismos parámetros cada actualización obtendrá distintas muestras.

Rango intercuartil

Ahora se tratarán brevemente otra medida usual en análisis descriptivo y la forma de señalarlo dentro de gráficas ilustrativas. Cuando se trató el tema de la mediana, también se retomó el tema de los estadísticos de orden, pues eran necesarios para determinar el punto donde la muestra acumulaba la mitad de las observaciones, o en otras palabras el 50% de los individuos. Al generalizar este concepto para el punto en el cual se acumulen los individuos de la muestra desde $x_{(1)}$ hasta acumular un $p\%$ del total de la muestra, se le conoce como Cuantil o percentil p denotado como X_p o q_p.

Cuando se calculan los percentiles de una muestra, hay casos en los cuales no existe un valor en la muestra para el cual se acumula un porcentaje en específico, por lo que se utiliza una interpolación lineal entre los dos valores que acumulen los porcentajes más cercanos. Como ejemplo de cálculo, si una muestra de datos sólo constaran de la serie de seis elementos {1,2,3,5,8,13} entonces no se podría hallar un valor que acumule un 25% de la muestra, sin embargo, se nota que el valor 1 acumula 16.6% de la muestra y el siguiente 2 acumula un 33.3% de la muestra, así que se utiliza la ecuación de la recta en los puntos $(X_{16.6\%}, 0.166)$ y $(X_{33.3\%}, 0.333)$ para aproximar el verdadero valor de $X_{25\%}$

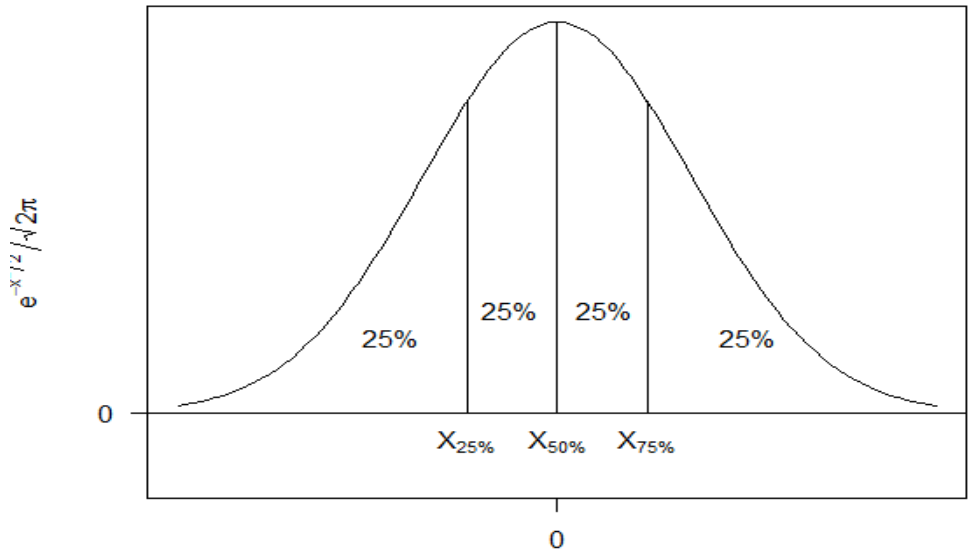
$$\frac{.25-.16.6}{X_{25\%}-1}=\frac{.33-.16.6}{2-1}$$

$$X_{25\%}=\frac{.25-.16.6}{.33-.16.6}+1=1.5$$

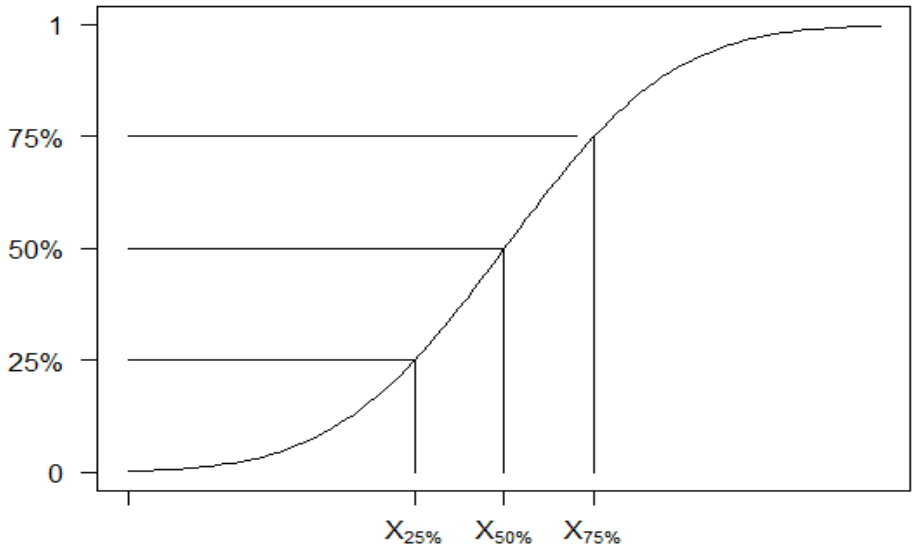
Cuando se habla de los percentiles $X_{25\%}$, $X_{50\%}$ y $X_{75\%}$ se les conoce también como cuartiles; de donde se puede notar que el segundo cuartil es la mediana. Estos valores deben su nombre a que parten el rango de la muestra en 4 puntos que acumulan cada uno ¼ o 25% de la población. La siguiente gráfica ilustra de forma teórica los cuartiles,

que son plasmados sobre las funciones de densidad y distribución acumulativa de una distribución Normal(μ, σ^2).

Gráfica 2.36: Posición de los cuartiles
en una función de densidad Normal(μ,σ^2)



Gráfica 2.37: Posición de los cuartiles
en distribución acumulativa de una Normal(μ,σ^2)



Con base en esto el rango intercuartil queda definido como $X_{75\%} - X_{25\%}$, interpretado como el rango alrededor de la mediana mismo que por su construcción contiene al 50% de los individuos más concentrados con respecto de la mediana en ambas direcciones, sin embargo, el intervalo alrededor de la mediana no siempre es simétrico, por esta razón también llega a considerarse otra medida de variabilidad, que es el rango semintercuartil, derivado de dividir el rango intercuartil entre 2.

Las ventajas de usar el rango intercuartil, es que tiene un impacto menor por los valores extremos a comparación de la desviación estándar S, pues en muchos contextos financieros, como la distribución de ingresos, pueden llegar a aparecer

valores extremos muy grandes que afecten el cálculo de la media y la desviación estándar.

Para calcular estos estadísticos dentro de la hoja de cálculo se utiliza la fórmula CUARTIL.INC(), la cual recibe como parámetros el rango de la muestra y un valor entero entre 0 y 4, donde 1 devuelve a $X_{25\%}$, 2 a $X_{50\%}$, 3 a $X_{75\%}$; 0 y 4 devuelven el mínimo y máximo de los valores de la muestra respectivamente. La ventaja de esta fórmula es que automáticamente realiza la interpolación si no encuentra el valor exacto para la probabilidad especificada.

Para mostrar gráficamente estos puntos, sin seguir encimando líneas a los gráficos, se sustituye la línea que se trazó para señalar el rango de la muestra, añadiendo los puntos $X_{25\%}$, $X_{50\%}$ y $X_{75\%}$, para lograr esto en la hoja de cálculo se pueden adicionar en un espacio libre, una tabla con tres columnas la primera con una serie consecutiva de 0 a 4 y en la segunda la fórmula CUARTIL.INC(), referenciando al valor de la primera columna, y por último se agrega una columna con la constante antes utilizada como $\max(f_i) + \varepsilon$ para darle altura a la recta.

Las dos últimas columnas sirven como referencia para sustituir la línea trazada para el rango dentro de la gráfica de la curva de frecuencias. Para mostrar las diferencia que tiene el rango intercuartil con el cálculo de la desviación estándar, se eligió cambiar el escenario antes visto con $\mu_1 = 50, \mu_2 = 80$, $\sigma_1 = 10$ y $\sigma_2 = 50$ pero con los siguientes pesos $p_1 = 30\%, p_2 = 70\%$. Este escenario con los cambios propuestos se observan en la gráfica 2.36 mientras que la Tabla 2.22 contiene el detalle del cálculo de los cuartiles y el rango intercuartil, donde además se agregó una columna extra para dar una referencia con los nombres de los estadísticos correspondientes.

Gráfica 2.38: Curva de frecuencias para 5000 valores simulados de la distribución mezcla de dos Normales $X_1 \sim N_1$ (μ_1=200,σ_1=10) y $X_2 \sim N_2$ (μ_2=80,σ_2=50) p_1=30%

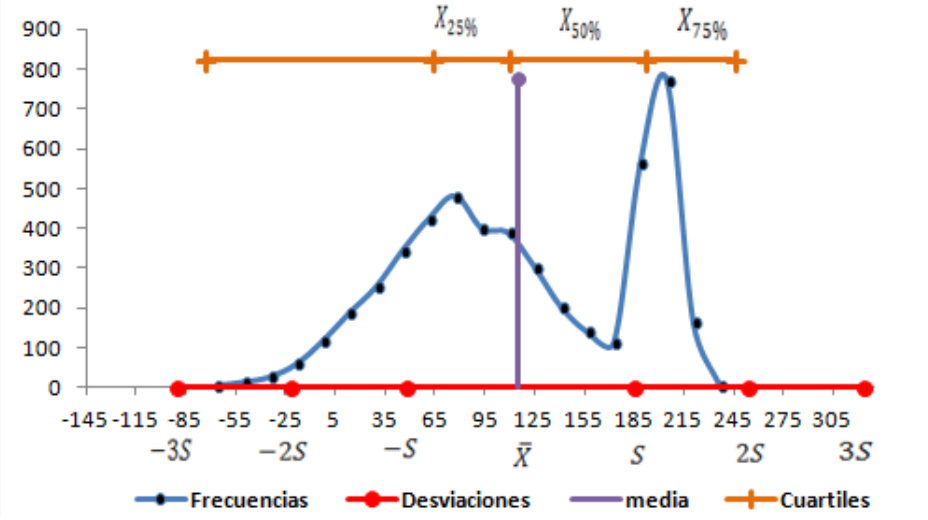


Tabla 2.22: Cálculo de los cuartiles $X_{(25\%)}$, $X_{(50\%)}$ y $X_{(75\%)}$ para la muestra de la distribución Z			
$Estadístico$	i	$Cuartil("rango\ de\ z", i)$	$MAX(f_i) + 50$
$MIN(x)$	0	-74.1	823
$X_{25\%}$	1	62.5	823
$X_{50\%}$	2	108	823
$X_{75\%}$	3	190.8	823
$MAX(x)$	4	244.8	823
$Rango\ Intercuartil$	128.4		

La primera característica que se identifica gráficamente del rango intercuartil es que, en este caso, es asimétrico alrededor de la mediana, a diferencia del intervalo que se encuentra rodeando la media por el valor S. Para este tipo de análisis es cuando se usa el rango semintercuartil, para otorgar una comparación entre ambas medidas de dispersión, el cual tiene una longitud de $128.4/2 = 64.2$, que es menor al valor de $S = 68.9$.

En la gráfica se aprecia que las características de las poblaciones individuales influyen en gran manera en la posición de los cuartiles, una de ellas el peso que se le asigna a cada distribución, ya que determina la acumulación de cada población. En este caso la población con mayor varianza, que también es a la que se le asignó mayor peso, así que acumula más peso cerca de la mediana por lo cual el cuartil $X_{25\%}$ queda más cerca de la mediana que el otro cuartil $X_{75\%}$.

Otro efecto que se puede notar es sobre los intervalos definidos con base en las desviaciones pues ahora el intervalo con 3 desviaciones sale por ambos lados del rango de la muestra, mientras que el intervalo con 2 desviaciones también sale del rango aunque sólo por la derecha.

El análisis anterior tiene el objetivo de mostrar una forma ordenada de realizar un análisis aislando y evaluando las características de interés con interpretaciones concisas, y al final obtener un resumen para lograr una visión integral de un fenómeno. Además el cambio en los parámetros de forma libre permitirá a un alumno explorar una gran cantidad de ejemplos teóricos, los cuales serán asimilables en la medida que se le pueda explicar el efecto que cada cambio tiene en el modelo final, y de formación práctica dependiendo de los análisis que deba producir.

Para lograr el entendimiento progresivo es recomendable cambiar un parámetro a la vez, ya que los efectos combinados al cambiar varios parámetros simultáneamente, aunque no son tan complejos hasta ahora, es mayor la dificultad de transmisión del mensaje al interpretar o explicar los resultados pues es más extenso.

Como ejemplo de resumir el comportamiento de la muestra anterior con los estadísticos hasta ahora revisados, la Tabla 2.23 contiene los datos de los parámetros de las distribuciones origen para identificar plenamente el modelo y los estadísticos calculados de la muestra generada.

Tabla 2.23: Resumen de una muestra de Z mezcla de dos normales, $N_1(\mu_1=50, \sigma_1=10)$, $N_2(\mu_2=200, \sigma_2=10)$ $p_1=30\%$, $p_2=70\%$			
p_1	15%	S^2	212.6
p_2	85%	S	14.6
n	5,000	$\#(\bar{X}-S, \bar{X}+S)$	3,599
$\bar{X}$	75.8	$\#(\bar{X}-2S, \bar{X}+2S)$	4,712
Me_x	78.1	$\#(\bar{X}-3S, \bar{X}+3S)$	4,978
M_1	50.5	$\%(\bar{X}-S, \bar{X}+S)$	72.00%
M_2	79.8	$\%(\bar{X}-2S, \bar{X}+2S)$	94.20%
$x_{(n)}-x_{(1)}$	97.6	$\%(\bar{X}-3S, \bar{X}+3S)$	99.60%
		Rango Intercuartil	128.4

Estadísticas Básicas

Después de realizar el cálculo de los estadísticos anteriores y resumirlos, en ciertos casos se requiere comparar los resultados contra las mediciones en otras poblaciones. Sin embargo los estadísticos hasta ahora vistos tienen una cualidad en común pues dentro de un contexto práctico, los resultados están expresados en términos de las unidades implícitas, usadas para plantear el análisis.

Por ejemplo, en un análisis del comportamiento financiero de una empresa se pueden hablar de clientes promedio, las desviaciones estándar de las ganancias en moneda nacional o extranjera, y los cuantiles que acumulan un porcentaje de individuos de una población objetivo. Para hacer comparables los comportamientos medidos entre poblaciones, se utilizan estadísticos que carezcan unidades implícitas en sus valores, los cuales se revisarán en las siguientes secciones.

Coeficiente de variación

El coeficiente de variación es un estadístico propuesto inicialmente por Fisher y Pearson, como una medida estandarizada de la variabilidad relativa de una muestra, independiente al contexto de las unidades en las que se expresa la información. Este coeficiente es empleado cuando se desean comparar distribuciones en unidades distintas como por ejemplo la distribución de pesos en recién nacidos expresada en gramos, y la distribución de su talla cuando son adultos la cual se registra en centímetros. Su cálculo se realiza como el cociente entre la desviación estándar y la media muestral $CV = \frac{S}{\bar{X}}$. Usualmente este coeficiente es expresado en términos porcentuales como $CV = \frac{S}{\bar{X}} * 100\,\%$, y es ampliamente usado en campos como al química, física y en campos actuariales en el ajuste de modelos de riesgo.

El problema con este coeficiente es que depende de la media el cual es un estadístico sensible a diferentes formas de las distribución de frecuencias, a casos individuales que en variables continuas toman valores muy elevados comparados con el resto de los individuos de la muestra y usualmente aparecen con muy poca frecuencia; además de no ser aplicable para aquellas distribuciones con valores negativos o para aquellas distribuciones en las que la media de la población sea cercana a cero.

Por esta razón algunos autores como Merce(2009) y Sharma(2011) proponen formas alternativas al cálculo de este coeficiente con el mismo objetivo de medir de forma comparable la variabilidad de la muestra y la relevancia de la media. E. Merce propone el cálculo del coeficiente de variación como

$$\boldsymbol{CV_M = \frac{S}{x_{(n)} - x_{(1)}} * 100\%}$$

Empleando la amplitud total en lugar de la media. Por su parte, la propuesta de Sharma es

$$\boldsymbol{CV_S = \frac{S}{\sqrt{(x_{(n)} - \bar{X})(\bar{X} - x_{(1)})}} * 100\%}$$

basado en la desigualdad de Muilwijk, para una muestra $\boldsymbol{X}$: $\boldsymbol{S^2 < (x_{(n)} - \bar{X})(\bar{X} - x_{(1)})}$ $\boldsymbol{\rightarrow \quad 0 \leq CV_S \leq 1.}$

La interpretación de cada coeficiente alternativo es, en el caso de CV_M la comparación relativa entre la desviación estándar, con respecto a la amplitud total de la muestra. Mientras que en el caso de la propuesta del coeficiente de variación de Sharma CV_S, mide la relación de la desviación de la media, con respecto a la posición de la media en el rango de la muestra, y puede ser afectado si existe asimetría en la muestra, ya que si la media se aproxima a algunos de los valores extremos entonces el denominador se reducirá.

Para poder visualizar y comprender los detalles de cada forma de cálculo se deben ejemplificar aplicándolos a una serie de muestras simuladas. Hasta el momento sólo se han trabajado como datos prácticos, la información de nacimientos totales del reporte de *Statistics Canada* de Julio 2013, los demás modelos aplicados al estar basados en contextos teóricos, parece que han dejado gradualmente de apegarse a la realidad, con el objetivo estresar el comportamiento de los estadísticos, sin embargo, ahora se analizara información real que demostrará, que los modelos expuestos anteriormente con distribuciones multimodales son comunes en realidad.

Aplicación del coeficiente de variación en datos reales

El Instituto Nacional de Estadística y Geografía (INEGI) realiza el cálculo del Índice Nacional de Precios al Consumidor (INPC), mismo que se publican en el Diario Oficial de la Federación de manera mensual, con el objetivo de medir la inflación de los precios de la canasta básica mexicana. Para llevar a cabo la generación del índice, se recaba una muestra de precios promedio calculados con las cotizaciones que se realizan en el mes de análisis, en 45 ciudades de la República Mexicana, para determinar los precios promedio en moneda nacional de cada producto componente del INPC.

Esta información se puede consultar en línea desde el banco de datos de la página web del INEGI donde se puede seleccionar en la consulta opciones como el periodo de análisis, los productos, y las ciudades donde se obtuvo la muestra, para que los datos sean descargados, en valores separados por comas (CSV, por sus siglas en

inglés), formato de fácil lectura para el software Excel®, por lo que es una gran fuente oficial de muestras.

Como objetivo particular de un análisis supóngase que se desea analizar el comportamiento de los precios de dos productos básicos: arroz y tortillas de maíz, los cuales son cotizados y registrados en una muestra sobre el precio de 1 Kilogramo de producto. La ventana de tiempo para el siguiente análisis general será sobre el comportamiento de los precios en el primer mes del año 2016.

Para iniciar el tratamiento de la información primero se debe establecer el tipo de variable a analizar, es decir si la variable es discreta o continua. Debido a que las unidades registradas en la información de precios son en moneda nacional, hasta con dos decimales, se le asociará con una variable aleatoria continua.

Sea entonces la variable A correspondiente a los precios de 1 Kilogramo de arroz y T la variable que corresponde al precio de 1 Kilogramo de tortilla de maíz. La muestra de precios consultada en línea fue sobre las 45 ciudades, para Enero 2016 el cual se puede abreviar como **Ene-2016**. En el caso de la variable A la muestra consultada fue de tamaño $n_A = 273$ cotizaciones; mientras que la muestra de T consta de $n_T = 910$ cotizaciones para el mes elegido.

El número de subintervalos para dividir el rango de las muestras de acuerdo la regla de Sturge son $m_A = 9$ y $m_T = 10$, sin embargo, a este nivel de agrupación puede simplificarse el análisis de la muestra, principalmente cuando se presentan en la muestra datos atípicos. Los efectos de estos valores dentro de la muestra son variados, por ejemplo, al realizar un histograma con el método visto se tienen subintervalos intermedios vacíos o también tienen un impacto en el valor de los estadísticos calculados.

En ambas muestras seleccionadas se hallan estos datos atípicos los cuales se analizarán más adelante, por lo cual se tomó una partición más fina del intervalo considerando $m = 20$ subdivisiones, aunque es recomendable usar como punto de partida la regla de Sturge y diferentes particiones del intervalo de acuerdo a lo que se observe en las curvas de frecuencia.

Para hacer un acercamiento al comportamiento de los datos similar al que se ha hecho en modelos teóricos, primero para la muestra de cotizaciones de T se generó la Gráfica 2.39, y además en la Tabla 2.24 se encuentra el resumen de los estadísticos calculados, donde se computaron los coeficientes de variación bajo las distintas definiciones.

Gráfica 2.39: Curvas de frecuencias y estadísticos principales para los precios promedio de 1Kg de Tortilla de maíz para el periodo Ene-2016.
Consulta en línea de los precios promedio del INPC. Publicados en el diario oficial de la Federación mensualmente ,Fecha de consulta: 05/05/2016

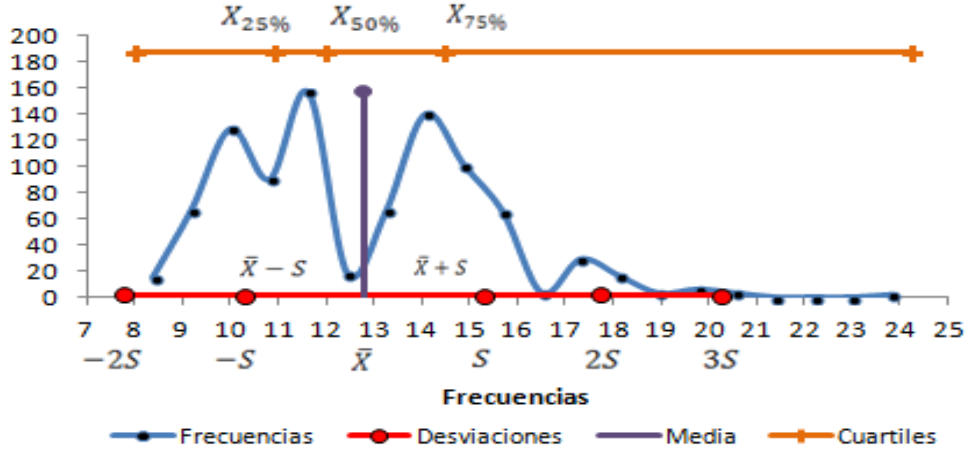


Tabla 2.24: Resumen de estadísticos principales para los precios promedio de 1 Kg de Tortilla de maíz para el periodo T1 2016.

Consulta en línea de los precios promedio del INPC, Publicados en el diario oficial de la Federación mensualmente ,Fecha de consulta: 05/05/2016

Estadístico	Valor	Estadístico	Valor
$\bar{X}$	12.8	$x_{(1)}$	8
Me_X	12	$X_{25\%}$	10.9
M_1	10	$X_{75\%}$	14.5
M_2	11.6	$x_{(n)}$	24.3
M_3	14.1	$x_{(n)} - x_{(1)}$	16.3
S^2	6.2	$Rango\ Intercuartil$	3.6
S	2.5	n	910
CV_P	19.50%	CV_S	33.60%
CV_M	15.30%		

Realizando un primer análisis general, de la gráfica se puede observar que la distribución de precios es multimodal, con 3 modas que se pueden identificar, como se había mencionado en la anterior sección, este efecto le resta interpretación a los valores de la media y la mediana pues se encuentran aproximadamente alrededor del punto medio entre 2 poblaciones. Lo que aporta el conocer el valor de éstas medidas de dispersión es que, de acuerdo al rango intercuartil y a la posición de los cuartiles es que cada posible grupo de la población contiene aproximadamente un 25% de la muestra. Otro aspecto que resalta en el gráfico es la cola derecha, pues la curva muestra una posible cuarta moda, pero de frecuencia baja, e incluso se encuentran precios de valores altos aunque en muy bajas frecuencias(es decir datos atípicos) que en este caso es el valor máximo de 24.3 MXN, lo que le da la forma plana a la cola de la distribución de precios.

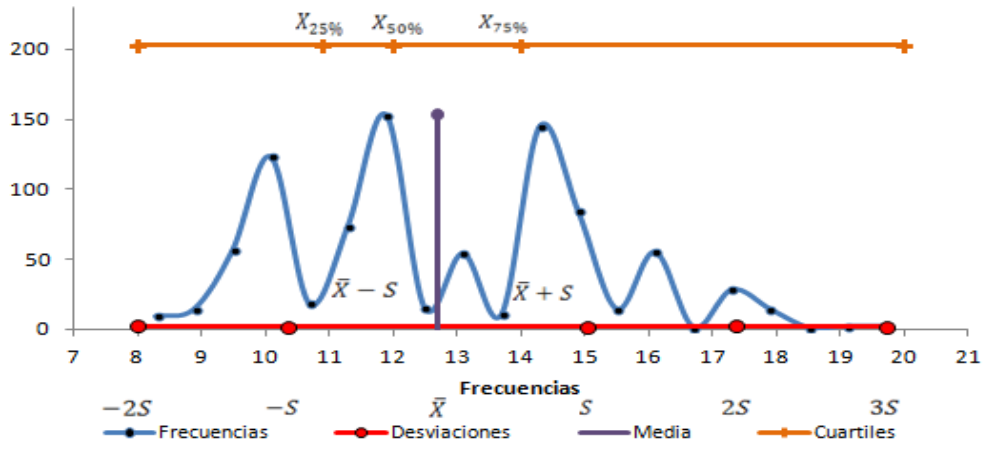
En cuanto a los valores de los coeficientes, en el caso de la propuesta de Merce, se puede interpretar que la desviación estándar de los precios equivale a un 15% de la amplitud total de la muestra. De acuerdo al cálculo de Pearson, la desviación estándar promedio representa un 19.4% del valor de la media; mientras que en la propuesta de Sharma el valor del coeficiente es de 33%, el cual aunque tenga la propiedad de siempre estar en el intervalo $(0,1)$, requiere una referencia conocida para poder comparar contra un caso en que el indicador arroje un valor cerca de algún punto extremo del intervalo. Aunque cada coeficiente tiene su propia interpretación, lo que se observa de este comparativo es que sus valores son muy distintos entre sí, por lo que el diagnóstico de variabilidad no está completo para esta muestra. A continuación se realizará primero un análisis más profundo desde los componentes de la muestra y después se comparará con el cálculo de los coeficientes de la distribución del precio promedio de 1 Kg de arroz, para poder explicar la variabilidad desde distintas perspectivas.

Una perspectiva, para poder explicar las variaciones en los datos, sin agregar supuestos, se puede hallar al revisar detalles en la forma de extracción de la muestra. En el caso de los precios cotizados la metodología publicada indica, que dentro de la muestra también se cotizan productos empaquetados los cuales pertenecen a una marca registrada, por lo que si un producto se vende en un paquete inferior a la unidad (por ejemplo 600 g) se realiza la conversión para obtener el precio unitario.

En el caso de la tortilla de maíz los datos obtenidos identifican al tipo de producto por medio de una columna con el título de “especificación” donde al tipo de tortilla la venta local a través de tortillerías usando maíz a granel y se le registra como “a granel”. En la muestra, 871 precios registrados son asociados a este tipo de producto y los 39 restantes son precios de productos empaquetados por alguna marca comercial, de los cuales se obtuvo la equivalencia del precio de 1 kg de producto.

Algunos de estos productos empaquetados se vendían en tamaños tales que el precio por kilo era más elevado que muchos precios de la muestra. Debido a que los productos de tortilla empaquetados representan el 4.3%, para poder observar un solo tipo de producto, se omitirán del análisis los productos empaquetados.

Gráfica 2.40: Curvas de frecuencias y estadísticos principales para los precios promedio de 1Kg de Tortilla de maíz exclusivo del tipo "A granel" para el periodo Ene-2016.
Consulta en línea de los precios promedio del INPC. Publicados en el diario oficial de la Federación mensualmente ,Fecha de consulta: 05/05/2016



La Gráfica 2.40 se obtuvo después de aplicar a la muestra del precio de tortillas un filtro sobre el tipo de producto para seleccionar únicamente productos a granel, donde se puede observar que, de igual manera a la información original, se tiene una distribución multimodal, en la cual se identifican tres modas alrededor de los valores 10, 12, y 14 aproximadamente, mientras que las demás frecuencias que se observan como modas se encuentran en proporciones menores alrededor de los valores 13, 16, y 17.5, las cuales no se descartan como poblaciones potenciales.

A continuación se muestra el cálculo de los coeficientes de variación y sus componentes, sobre los datos filtrados para identificar sólo productos de tipo a granel:

Tabla 2.25: Resumen de estadísticos principales para los precios promedio de 1 Kg de Tortilla de maíz de tipo "A granel" para el periodo T1 2016.

Consulta en línea de los precios promedio del INPC, Publicados en el diario oficial de la Federación mensualmente ,Fecha de consulta: 05/05/2016

Estadístico	Valor	Estadístico	Valor
$\bar{X}$	12.7	$x_{(1)}$	8
S	2.3	$x_{(n)}$	20
CV_P	18.50%	$x_{(n)} - x_{(1)}$	12
CV_M	19.50%	n	871
CV_S	40.00%		

Los datos de la tabla muestran que la media y la desviación estándar se mantuvieron cercanos a su valor previo, por otra parte el rango de la muestra fue menor debido al efecto de eliminar los productos empaquetados. Los impactos de este filtro se puede observar también en los valores de los coeficientes CV_M y CV_S pues aumentaron en 5 y 7 puntos porcentuales respectivamente, a diferencia del coeficiente de Pearson el cual se redujo en un punto porcentual, con lo que se puede observar que los coeficientes alternativos, son más sensibles a los valores extremos de una distribución.

Para poder explicar los comportamientos multimodales se pueden emplear algoritmos computacionales avanzados, que permiten separar de manera eficiente una muestra en diversos grupos lo más homogéneos posibles entre sí y que a su vez sean lo más heterogéneos entre los grupos. Esto se realiza por medio métodos que analizan las relaciones entre todas las características que se puedan obtener de los individuos de la poblaciones, como por ejemplo las redes neuronales, sin embargo para este tema de estadística descriptiva se mostrará una forma más intuitiva de poder explicar estos comportamientos.

Dado que la envergadura de la muestra obtenida fue a nivel nacional, en un contexto geográfico México es un país con territorios de diversos medios naturales y poblaciones que históricamente se han adaptado social y económicamente a su medio. Agrupar estos diferentes medios en regiones no es tarea sencilla pues involucran un gran número de consideraciones, aunque este comportamiento se puede observar en una gran cantidad de países multiculturales. Una primera aproximación para clasificar este tipo de información, son las agrupaciones que se asocian a las divisiones geopolíticas, como estados, provincias, municipios, etc. ya

que son divisiones fijas que perduran más en el tiempo, de los cuales usualmente se tienen información geográfica y estadística de manera pública. El nivel seleccionado en este caso para dividir a la República mexicana, es al nivel de los 31 estados y su capital la Ciudad de México.

La información sobre las ciudades, fecha de obtención de la muestra y detalles de los productos seleccionados, permite relacionar cada dato con el estado en el que fue tomado cada registro y colocarlo en una columna extra. Este paso se debe realizar con funciones del software, usando por ejemplo la función $Consultav()$ la cual es útil al usarla con un solo catalogo relacionando las ciudades de la muestra con los estados a los que pertenecen. Una vez que se tiene identificado el estado donde se obtuvo cada dato, en una tabla dinámica alimentada por toda la base de información, se puede seleccionar el estado como una variable para agrupar la información, luego se selecciona como variable de cálculo al precio promedio configurando el cálculo como un promedio de la variable. El resultado de calcular el promedio de los precios promedio por estado se encuentra en la Tabla 2.26.

Tabla 2.26: Precios promedio por Estado de Tortilla de maíz para el periodo Ene-2016.

Consulta en línea de los precios promedio del INPC, Publicados en el diario oficial de la Federación mensualmente ,Fecha de consulta: 05/05/2016

Estado	Precio Promedio	Estado	Precio Promedio	Estado	Precio Promedio
Aguascalientes	11.6	Guanajuato	11.3	Quintana Roo	13.7
Baja California	14.9	Guerrero	14.3	San Luis Potosí	11.7
Baja California Sur	14.3	Hidalgo	9.9	Sinaloa	14.3
Campeche	14.4	Jalisco	11.9	Sonora	15.5
Chiapas	11.7	Michoacán	12.3	Tabasco	13.5
Chihuahua	13.2	Morelos	12.8	Tamaulipas	13.4
Ciudad de México	11.1	Nayarit	14.3	Tlaxcala	10.2
Coahuila	14.3	Nuevo León	13.5	Veracruz	11.8
Colima	13.2	Oaxaca	11.7	Yucatán	14.3
Durango	10.9	Puebla	10.5	Zacatecas	11.5
Edo. de México	11.3	Querétaro	12.1		

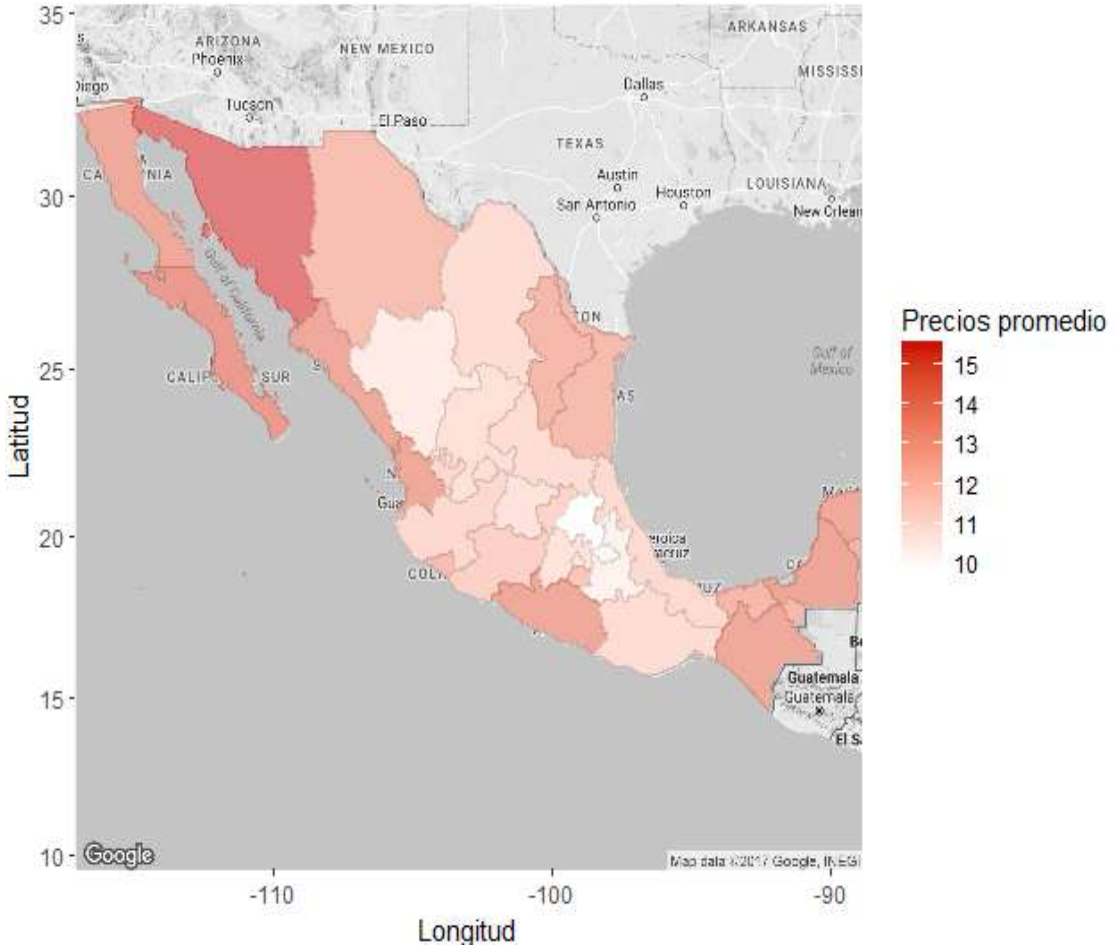
Los resultados se encuentran resumidos en un orden alfabético de acuerdo al nombre del Estado, por lo cual no es sencillo relacionar los distintos valores obtenidos, pero se puede observar que se tienen valores distintos desde 9.9 MXN en Hidalgo a 15.5 MXN en el Estado de Sonora, así que también refleja la variabilidad de la curva de frecuencias mostrada en la Gráfica 2.39.

A continuación se le dará un esquema visual de presentación a los cálculos anteriores. Puesto que el nivel elegido para resumir los datos, es una definición oficial de división geopolítica, es posible encontrar archivos en la web, conocidos como *Shapefile* los cuales describen detalles geográficos de alguna región y se pueden hallar para México. El tipo de archivo *Shapefile* es leído por software especializado en el trazo de mapas entre otros como R®, por lo que los resultados de la Tabla 2.26, se plasmarán en un mapa por medio del software R®, con base en un método publicado.

El código para generar el mapa, primero genera el mapa con fondo blanco y por medio de un vector con la información de la Tabla 2.26, se rellena cada estado con un color en diversas intensidades de acuerdo al valor del precio promedio[17]. Como resultado se obtiene la siguiente gráfica:

Gráfica 2.41:Precios promedio de 1 Kg de Tortilla por Estado, periodo Ene-2016.

Consulta en línea de los precios promedio del INPC



El mapa revela el origen de las múltiples modas de la curva de frecuencias de la muestra, a través de la relación que tienen los diversos precios promedio, además se observa la similitud que tienen diferentes estados en la intensidad del color por lo cual se pueden agrupar en zonas o regiones de una forma intuitiva. Por ejemplo localizando los lugares con los precios más bajos se encuentra la zona centro del país con: la Ciudad de México, Hidalgo, Tlaxcala, Puebla y Durango que coinciden con los promedios más bajos de la Tabla 2.26, mientras que en diversas zonas hacia los extremos norte y sureste del país resaltan los estados por ser los precios más elevados.

El uso de este tipo de mapas también puede servir como una mención introductoria, dependiendo de cada programa académico, hacia la estadística espacial, de la cual se pueden encontrar diversos ejemplos reales de aplicación, como el caso de

[17] Consulte el código completo y las referencias en la parte II del apéndice.

Barrera(1993) en el campo de la ecología analizando la dinámica y magnitud en la forma en que distribuyen manadas de elefantes en territorios de África, bajo los efectos de diferentes temporadas climatológicas.

Retomando el tema de los coeficientes de variación ahora se aplicará el análisis inicial, a la muestra de las cotizaciones de 1 kg de arroz para el mismo periodo de Enero 2016. Ahora es necesario revisar el detalle de la muestra previo a la realización del análisis pues es posible que la muestra contenga los mismos comportamientos.

El detalle de la muestra en cuanto al tipo de productos, es distinto al de la muestra con los precios de tortilla de maíz en el sentido que este producto es comercializado popularmente desde las tiendas de conveniencia hasta en los supermercados, de forma empaquetada por alguna marca productora o distribuidora, en el caso de esta muestra fue colectada de 125 diferentes marcas; este efecto en la muestra anterior provocaba valores extremos muy altos, debido a considerar precios unitarios, el tipo de producto y a que la competencia comercial provoca más variación en los precios cuando hay un mayor número de marcas involucradas.

Estos valores extremos también se encuentran en la muestra de precios de arroz al tener un valor mínimo de 9.9 MXN y un máximo de 75.2 MXN; así que es mejor mostrar en la tabla 2.27 los cálculos correspondientes en lugar de plasmar la curva de frecuencias. Además se dividió el rango de la muestra en 20 subdivisiones en lugar de considerar $m_A = 11$ subdivisiones de acuerdo a la regla de Sturge, con el objetivo de hacer un análisis más fino al tomar intervalos de menor longitud.

Tabla 2.27: Frecuencias acumuladas y puntuales de precios promedio de 1 Kg de arroz para el periodo Ene-2016.

Consulta en línea de los precios promedio del INPC, Publicados en el diario oficial de la Federación mensualmente ,Fecha de consulta: 05/05/2016

i	*Límite Superior*	*Punto medio* m_i	*Frecuencia Acumulada*	*Frecuencia de Subintervalo*
1	13.4	11.7	92	92
2	16.6	15	180	88
3	19.9	18.3	226	46
4	23.1	21.5	252	26
5	26.4	24.8	260	8
6	29.6	28	263	3
7	32.9	31.3	267	4
8	36.2	.4.5	269	2
9	39.4	37.8	269	0
10	42.7	41	270	1
11	45.9	44.3	270	0
12	49.2	47.6	270	0
13	52.4	50.8	270	0
14	55.7	54.1	271	1
15	59.0	57.3	271	0
16	62.2	60.6	271	0
17	65.5	63.8	271	0
18	68.7	67.1	271	0
19	72.0	70.3	271	0
20	75.2	73.6	272	1

Tabla 2.28: Resumen de estadísticos principales para los precios promedio de 1 Kg de Arroz para el periodo T1 2016.			
Consulta en línea de los precios promedio del INPC, Publicados en el diario oficial de la Federación mensualmente ,Fecha de consulta: 05/05/2016			
Estadístico	Valor	Estadístico	Valor
$\ddot{X}$	16.2	$x_{(1)}$	10.1
Me_X	14.5	$X_{25\%}$	12.7
M_1	11.7	$X_{75\%}$	18
S^2	39.9	$x_{(n)}$	75.2
S	6.3	$x_{(n)} - x_{(1)}$	65.1
CV_P	38.90%	$Rango\ Intercuartil$	5.3
CV_M	9.70%	n	272
CV_S	33.20%		

La tabla de frecuencias muestra una concentración de la muestra en los primeros subintervalos, sin embargo este efecto fue derivado de los valores extremos que se detectan al final de la tabla, donde el máximo de la muestra junto con otra observación intermedia se encuentran por valores muy por encima de los estadísticos de tendencia central y en general del resto de los valores de la muestra. A pesar de los efectos de los valores extremos, de acuerdo a la manera en que las frecuencias decaen exponencialmente, estos valores no se filtran de forma intuitiva de la muestra, pues aún con los metadatos no se pueden diferenciar tan claramente tipos de productos sino de marcas, a diferencia de la muestra anterior, por lo que estos valores atípicos son parte del comportamiento de los datos.

En este caso la regionalización de los datos no aporta gran información al análisis, esto es debido a que una marca puede establecer el precio de un producto, en varias zonas del país o a nivel nacional de acuerdo a su expansión comercial, lo anterior aunado a la gran cantidad de marcas que aparecen en la muestra, genera una variabilidad que para modelar se requiere de una clasificación distinta al agrupamiento por definiciones geopolíticas.

Enfocando el interés de esta sección, al valor de los coeficientes empezado por el método de Pearson, se observa un mayor valor la media de esta población, seguido por una desviación estándar aun mayor comparada contra los estadísticos de la muestra T, lo que refleja el 40% del valor de CV_P, indicando que A es la variable de mayor variabilidad. En cuanto a la forma de cálculo de Merce, fue impactada directamente por el tema de los valores extremos, pues el rango al ser amplio, deriva en que el porcentaje que representa la desviación estándar es de alrededor del 10%, que comparado contra el valor de CV_M de la distribución de T es menor, lo que indica un sentido contradictorio. Por último la alternativa CV_S se observa con un valor similar al correspondiente a CV_S de la variable T, así que del comparativo en el método de Sharma, se indica que ambas distribuciones tienen una variabilidad igual o al menos muy similar.

Los primeros análisis y comparativos anteriores arrojan que el Coeficiente de variación de Pearson es un indicador de la variabilidad de una muestra, que es más estable ante la aparición de datos atípicos, ya que por el contrario, los cálculos alternativos dependen directamente de los valores extremos de la muestra, los cuales en muestras como las revisadas u otras series de contexto financiero, son un dato atípico. El coeficiente de Merce sin embargo tiene una utilidad complementaria al relacionar la dispersión como porcentaje de la amplitud total de las observaciones.

Comparativo de las propuestas para el cálculo del coeficiente de variación empleando simulación Monte Carlo

El otro aspecto discutido por otros autores sobre el coeficiente de variación de Pearson, es su uso para variables con valores negativos, y cuando la media es cercana a cero, por tal motivo a continuación se utilizarán las simulaciones multimodales generadas anteriormente para revisar este tipo de escenarios. Para generar las condiciones anteriores, se puede retomar el primer ejemplo de distribuciones multimodales, empleándolo como *template* en Excel®, en el cual se suponen pesos iguales para ambas distribuciones, pero además se deben colocar sus medias de manera simétrica alrededor de 0, con una desviación estándar de tal manera que las distribuciones no estén completamente separadas. Por ejemplo si se define la distribución de X_2, como $X_2 \sim N(\mu_1 = 50, \sigma = 25)$ entonces se puede definir la otra distribución como $X_1 \sim N(\mu_1 = -50, \sigma = 25)$.

Al actualizar los parámetros sobre la hoja de cálculo y considerando el mismo tamaño de muestra $n = 5{,}000$, se genera una distribución con las condiciones deseadas, la curva de frecuencias como la tabla resumen de los coeficientes de variación y sus componentes se muestran a continuación:

Gráfica 2.42: Curva de frecuencias para 5000 valores simulados de la distribución mezcla de dos Normales X_1~N_1 (μ_1=-50,σ_1=25) y X_2~N_2 (μ_2=50,σ_2=25) p_1=p_2=50%

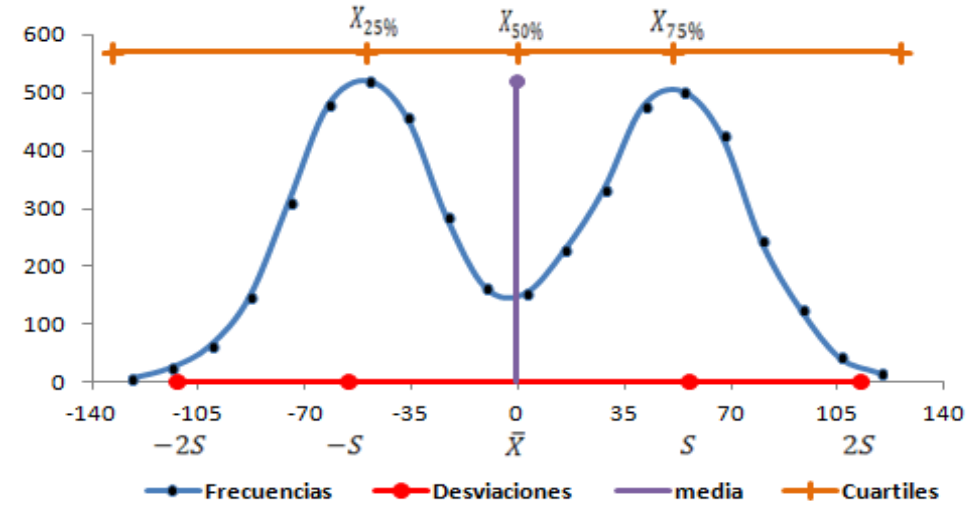


Tabla 2.29: Resumen de estadísticos principales para los valores simulados de la distribución Z mezcla de dos normales Donde μ_1=-50,μ_2=50, σ_1=σ_2=25 ,P_1=P_2=50%			
Estadístico	Valor	Estadístico	Valor
$\ddot{X}$	-0.3	$x_{(1)}$	-134
S	56.1	$x_{(n)}$	125
CV_P	-20048%	$x_{(n)} - x_{(1)}$	259
CV_M	21.60%	n	5,000
CV_S	43.30%		

Como se observa en la gráfica y en el valor de los estadísticos en la tabla, esta simulación cumple con las características deseadas, entonces comparando el valor de los coeficientes se nota que, al ser la media cercana a cero el coeficiente de Pearson tiende a valores muy elevados. Por otra parte CV_P es de valor negativo, sin embargo, en este caso es un tema de aleatoriedad de los valores simulados, aunque en algunos textos esta corrección se realiza al tomar el valor absoluto de la media. El coeficiente alternativo CV_M se observa estable, por lo que la desviación estándar promedio de la muestra representa un 21% de la amplitud total, y en cuanto al coeficiente CV_S se observa que al ser la media cero su denominador considerará el producto $-x_{(1)}$ con $x_{(n)}$ que al ser similares, aplicando la raíz cuadrada, el denominador toma un valor cercano en magnitud a alguno de los valores extremos(129.4 para esta muestra), por lo que se puede concluir que bajo la condición de una distribución simétrica alrededor de una media 0 entonces $CV_S \cong 2CV_M$.

Según la revisión anterior de estos coeficientes, se puede concluir, en general, que para valores no negativos el coeficiente de Pearson aun es el mejor indicador de variabilidad comparable entre distribuciones de diferentes tipos, mientras que el coeficiente de Merce, es un indicador alternativo que complementa el análisis, siendo útil también para muestras con valores negativos, aunque es directamente impactado por los valores extremos. Por último el coeficiente propuesta de Sharma CV_S carece de comparabilidad pues en muestras de diferentes comportamientos, no se observó una variación significativa que permitiera diferenciar a las poblaciones según el valor del estadístico.

Coeficiente de asimetría

La asimetría de una distribución, como se había comentado en secciones anteriores, permite identificar en qué sentido las observaciones se concentran, ya sea cerca de un punto medio del rango de la muestra o hacia alguno de sus extremos. Cuando la concentración de las observaciones está cargada hacia el lado izquierdo del rango entonces se tienen una asimetría positiva, en el caso que las observaciones se encuentren con mayor concentración del lado derecho, entonces la asimetría se considera negativa, mientras que cuando las observaciones se concentran en el medio y su distribución es muy similar en ambos sentidos entonces se considera como una distribución simétrica.

En las secciones anteriores se han revisado muestras en alguna de estas tres condiciones, sin embargo para poder medir la asimetría sin unidades de por medio

para poder comparar muestras de distintos contextos, es usado el coeficiente de asimetría de Pearson. Previo al cálculo de este coeficiente, se deben definir los momentos muestrales alrededor de la media o centrales de la distribución. Supóngase una muestra $X = \{x_i\}_{i=1}^{n}$, los momentos muestrales centrales μ_k se definen de la siguiente manera:

$$\boldsymbol{\mu_k = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^k}{n}}$$

Cuando $k = 1$ el resultado es 0, además es importante señalar la diferencia cuando $k = 2$, ya que la varianza muestral S^2 divide la suma entre $n - 1$ y no entre n. A partir del tercer momento es cuando se miden diferentes propiedades, puesto que al elevar un número al cubo no se pierde el signo original, entonces las diferencias $x_i - \bar{X}$ tendrán un mayor impacto a medida que los valores de la distribución x_i se encuentren más alejados hacia la derecha o a la izquierda de la media, de esta manera μ_3 tiene la siguiente regla:

$$\boldsymbol{Si\ \mu_3 > 0\ \rightarrow\ La\ distribución\ es\ asimétrica\ positiva}$$

$$\boldsymbol{Si\ \mu_3 < 0\ \rightarrow\ La\ distribución\ es\ asimétrica\ negativa}$$

$$\boldsymbol{Si\ \mu_3 = 0\ \rightarrow\ La\ distribución\ es\ simétrica}$$

La regla anterior sólo aporta el sentido de la asimetría, puesto que está en unidades cúbicas que no son posibles de comparar contra otras distribuciones; es necesario entonces eliminar las unidades implícitas de cualquier distribución, así que una opción es emplear el coeficiente de asimetría de Fisher-Pearson definido como:

$$\boldsymbol{\gamma_1 = \frac{\mu_3}{(\sqrt{S^2})^3} = \frac{\mu_3}{S^3} * 100\%} \qquad \boldsymbol{donde\ S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{X})^2}{n-1}}$$

Este estadístico ha recibido diferentes críticas en la literatura por las últimas décadas, sin embargo (Kim y White) resumen diferentes formas más robustas para medir la asimetría de las cuales la más sencilla es la definida a través de percentiles propuesto por Bowley (1920) como sigue

$$\boldsymbol{\gamma_2 = \frac{X_{75\%} + X_{25\%} - 2X_{50\%}}{X_{75\%} - X_{25\%}} * 100\%}$$

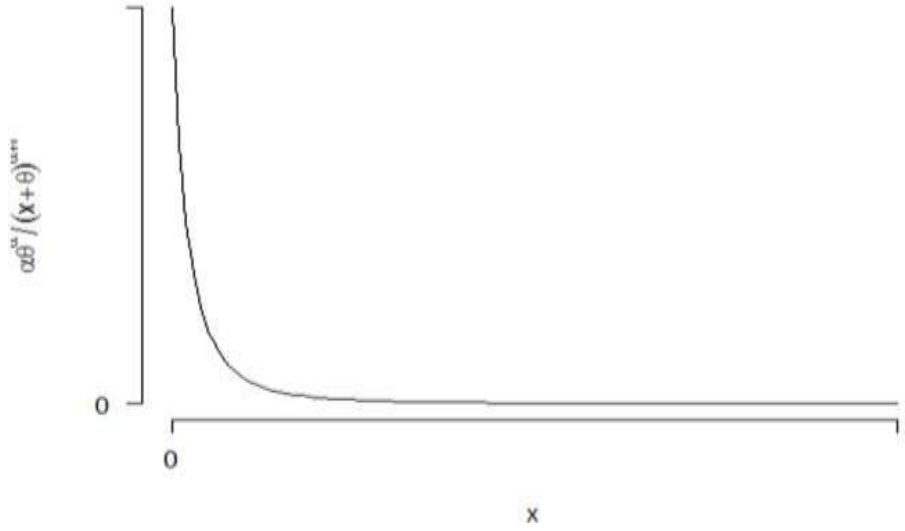
Analizando el numerador de γ_2 se puede reescribir como $\boldsymbol{(X_{75\%} - X_{50\%}) - (X_{50\%} - X_{25\%})}$, que se puede ver como la comparación de la distancia del primer cuartil a la media con la distancia de la mediana al tercer cuartil, así que en una distribución simétrica como las distribuciones Normales esta distancia es igual, implicando que el coeficiente sea 0. Por otra lado el denominador reescala los valores al dividir entre el rango intercuartil, limitando los valores del coeficiente en el intervalo $(-1,1)$.

Para ejemplificar el uso de estos estadísticos se puede emplear de inicio una distribución con una clara asimetría, en este caso se puede retomar la distribución del italiano Vilfredo Pareto conocida por la modelación de ingresos de una población y que

fue revisada en el capítulo anterior. Ésta distribución tiene su función de densidad definida como

$$\boldsymbol{f(x) = \frac{\alpha\theta^{\alpha}}{(x+\theta)^{\alpha+1}} \quad x > 0, \alpha > 1, \theta > 0}$$

Gráfica 2.43: Densidad de una distribución Pareto(α,θ)



Donde θ es un parámetro de escala y α es el parámetro de forma de la distribución. Ésta distribución totalmente cargada hacia la izquierda y además tiene una cola pesada, como se mostrará en la curva de frecuencia de los valores que serán simulados. Las simulaciones serán realizadas por el método de la transformada inversa visto también en el capítulo anterior donde se obtuvo la inversa de la distribución como

$$\boldsymbol{F^{-1}(U) = \theta\left(U^{-\frac{1}{\alpha}} - 1\right)}$$

Con esta fórmula es sencillo aplicar sobre la misma hoja de las distribuciones multimodales o en una hoja de cálculo nueva, una columna de valores aleatorios de tamaño n y una segunda columna adjunta, referenciando con la formula anterior de cada celda de esta segunda columna al valor del aleatorio como U y a los parámetros, los cuales deberán estar colocados en celdas individuales.

Hasta el momento el tamaño de muestra, en todos los ejemplos ha sido suficiente para mostrar los comportamientos deseados por lo cual se mantendrá en esta ocasión el tamaño de muestra $n = 5000$. Acerca de los parámetros, se eligió una escala relativamente reducida al tomar $\theta = 100$, por otra parte el parámetro α entre más cercano a 1 sea, la curva decae con mayor fuerza cerca del valor 0, por lo cual se eligió el parámetro $\alpha = 10$ para visualizar mejor la muestra.

Una vez generada la muestra se procede a realizar el conteo de frecuencias y la curva de frecuencias correspondiente, que se puede observar en la Gráfica 2.44, en la cual se nota que es similar a su densidad teórica. Para este conjunto de simulaciones se calcularon además los estadísticos principales, observados en la Tabla 2.30, aunque

para el coeficiente de variación se consideró la definición de Pearson al igual que para el coeficiente de asimetría.

Gráfica 2.44: Curva de frecuencias para 5000 valores simulados de la distribución Pareto (θ=100 , α=10)

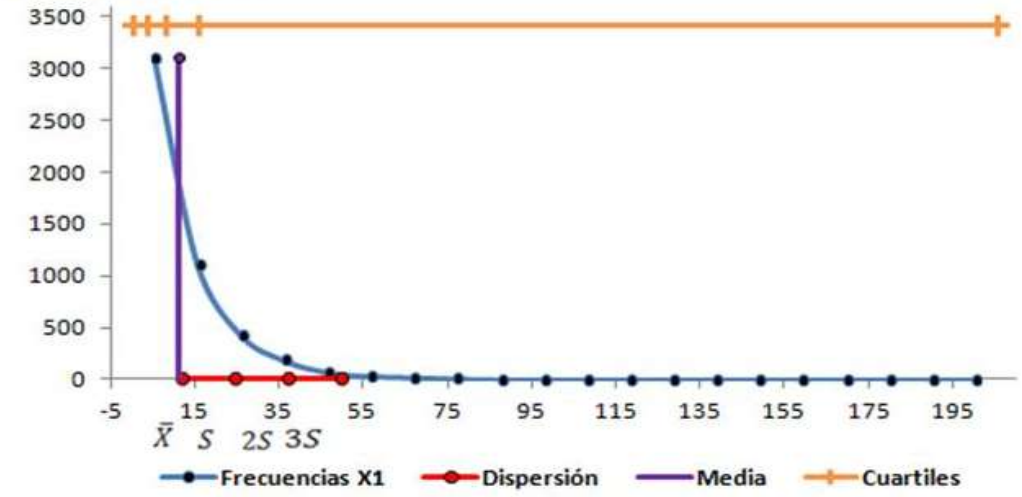


Tabla 2.30: Resumen de estadísticos principales de una muestra generada de una distribución Pareto(θ=100,α=10)			
Estadístico	Valor	Estadístico	Valor
$\bar{X}$	11.3	$x_{(1)}$	0.0
Me_X	7.3	$X_{25\%}$	3
M_1	5.1	$X_{75\%}$	15
S^2	159.9	$x_{(n)}$	205
S	12.6	$x_{(n)} - x_{(1)}$	205
CV_P	112.30%	$Rango\ Intercuartil$	12
μ_3	6,221.10	n	5,000
γ_1	307.60%	γ_2	29.90%

El valor de μ_3 al ser positivo corrobora la asimetría positiva de la muestra sin embargo observando el valor de γ_1 indica la medición relativa con respecto a la variabilidad, con la que las observaciones se concentran hacia el extremo izquierdo del rango. En el caso de γ_2 se interpreta de forma más sencilla pues indica que de los individuos concentrados alrededor de la mediana, hay un 30% más de individuos a su izquierda.

Una característica demostrable de la distribución Pareto es que el valor de la desviación estándar es mayor que la media. A continuación se obtiene ambos cálculos de la media y la desviación estándar, pues serán de utilidad para las simulaciones posteriores.

$$E(X) = \frac{\theta}{\alpha - 1} = \frac{100}{10 - 1} = 11.1$$

$$\boldsymbol{DesvE(X) = \sqrt{Var(X)} = \sqrt{\frac{2\theta^2}{(\alpha-2)(\alpha-1)} - \left(\frac{\theta}{\alpha-1}\right)^2} = \sqrt{154.3} = 12.4}$$

En la gráfica también se pude notar el efecto de una desviación estándar mayor a la media, ya que ahora no es necesario trazar el lado izquierdo de la línea de desviaciones, pues el cálculo de $\bar{X} - S$ es menor a 0 y sale del rango de valores. Es importante señalar que el objetivo de estas simulaciones, es que el estudiante comprenda la medición de la asimetría y logre diferenciar lo que cada estadístico aporta al análisis, sin embargo, hasta este punto se han revisado una serie de estadísticos que, aunque están relacionados en sus cálculos, analizan características diferentes e independientes en el comportamiento de la muestra, por lo que es importante diferenciarlos o aislarlos.

Por ejemplo, para esta distribución Pareto el coeficiente de variación es alto por la relación de la media con su desviación estándar, por este motivo la siguiente distribución a usar como ejemplo, será una distribución Normal, con un coeficiente de variación igual, al de la distribución Pareto pero con γ_1 y γ_2 cercanos a 0, debido a la simetría de la distribución Normal.

Aprovechando los métodos antes vistos para generar distribuciones multimodales, se puede reutilizar la estructura en hoja de cálculo de una simulación distribución Normal, ya sea X_1 o X_2, ajustando los valores de la media y desviación estándar, que en este caso se elegirán iguales a los de la distribución Pareto es decir ($\mu = 11.11, \sigma^2 = 154.3$) para mantener la variabilidad y conseguir una distribución con la menor asimetría posible.

A continuación se presentan la curva de frecuencias de la simulación de una variable Normal($\mu = 11.11, \sigma^2 = 154.3$) sobrepuesta a la curva de frecuencias de la distribución Pareto($\theta = 100, \alpha = 10$) antes generada. La Tabla 2.31 muestra un resumen de estadísticos de las simulaciones de la distribución Normal, para compararlos con los resultados de la Tabla 2.30.

Gráfica 2.45: Curvas de frecuencias para 5,000 valores simulados de la distribución Pareto(θ=100, α=10) y una Normal (μ=11.1, σ²=154.3)

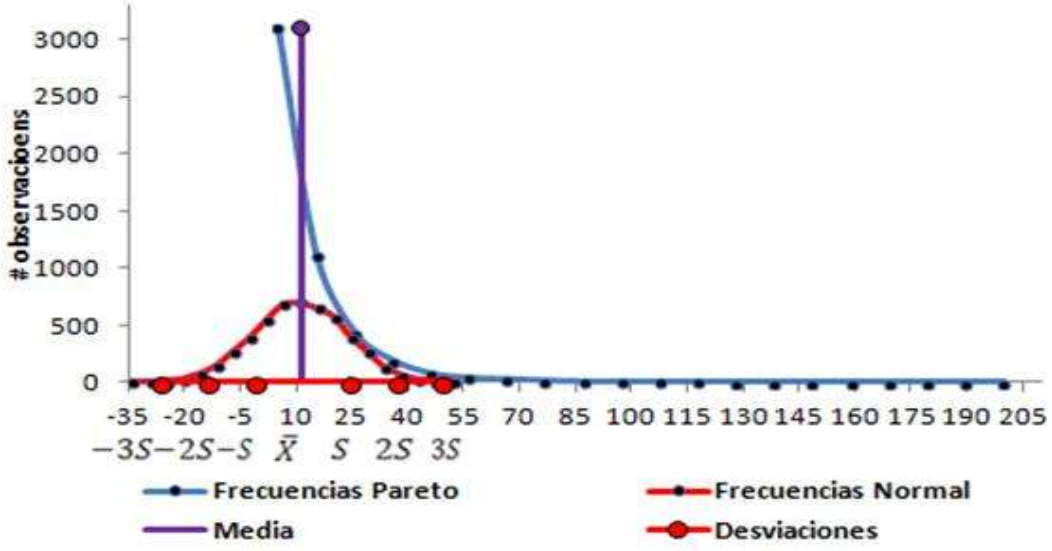


Tabla 2.31: Resumen de estadísticos principales de una muestra generada de una distribución Normal(μ=11.1,σ^2=154.3)			
Estadístico	Valor	Estadístico	Valor
$\bar{X}$	11.1	$x_{(1)}$	-36.2
Me_X	11.1	$X_{25\%}$	2.8
M_1	11.1	$X_{75\%}$	19.6
S^2	153.2	$x_{(n)}$	53.9
S	12.4	$x_{(n)} - x_{(1)}$	90.2
CV_P	111.60%	$Rango\ Intercuartil$	17
μ_3	-33.30	n	5,000
γ_1	-1.76%	γ_2	1.09%

Como era esperado los valores de la media muestral y desviación estándar, son prácticamente el valor de los parámetros, por consiguiente, el coeficiente de variación también es muy cercano al valor de la distribución Pareto simulada. El tercer momento de la distribución muestra un valor negativo lo que indicaría una asimetría negativa, es decir que los datos tienden en promedio a estar a la derecha de su media, sin embargo no indica en qué grado se encuentra esta tendencia al estar en unidades cúbicas, además de acuerdo al valor de γ_1 se puede observar que la asimetría de forma relativa con su desviación estándar sólo representa cerca de un -2% que es muy cercano al 0% teórico.

Por otra parte γ_2 toma un valor positivo debido a la diferencia entre los cuartiles, aunque su valor de igual manera es cercano a cero, este efecto ocurre debido a que es impactado igual por todas las observaciones mientras que $\boldsymbol{\gamma_1}$ es usualmente afectado por las simulaciones en los extremos de una muestra, aunque al generar muestras continuamente se podrá observar que el valor de los γ_i tomarán valores de diferentes combinaciones de signos aunque alrededor de 0.

Con los ejemplos anteriores se puede entender la necesidad de medir y comparar la asimetría de las distribuciones, como una medida complementaria, ya que en un análisis pueden coincidir en otros estadísticos de posición y dispersión. Por ejemplo en un contexto financiero a diferente escala, si ambas distribuciones representaran los rendimientos de un portafolio, en un sistema que compute sólo el rendimiento medio y su desviación estándar, no serían suficientes para diferenciar al mejor activo, mientras que con el conocimiento de la asimetría se puede interpretar que el activo de mayor asimetría positiva tendrá rendimientos más atractivos, pensando en una cola la derecha de la distribución cada vez más alargada. En estos tipos de contextos la siguiente pregunta sería que tan probable es que estos valores extremos de las colas largas ocurran, lo cual se conoce comúnmente como el peso de las colas, que se analizara a continuación.

Coeficiente de Kurtosis

En las diferentes simulaciones expuestas se ha revisado el efecto que los valores extremos pueden tener, cuando se desea medir el impacto del conjunto de individuos

que se localizan más lejos de centro de la distribución, es decir de las colas, entonces se usan las medidas de kurtosis. Pearson propuso medir este efecto a través de comparar el cuarto momento de la distribución de forma relativa contra su desviación estándar a la cuarta potencia con el siguiente coeficiente

$$\boldsymbol{K_1(X) = \frac{\mu_4}{(\sqrt{S^2})^4} = \frac{\mu_4}{S^4} * 100\%}$$

Un resultado conocido sobre la distribución Normal indica que cuando $\mu = 0$ y $\sigma^2 = 1$ entonces su kurtosis toma el valor de 3, como esta distribución es ampliamente usada en diversos campos, es común encontrar que se toma esta distribución como un estándar de kurtosis así que se calcula un coeficiente de exceso de kurtosis como

$$\boldsymbol{ExK_1(X) = K_1(X) - 3}$$

Este estadístico tiene las mismas desventajas que los otros coeficientes de Pearson al ser impactado por cualquiera de los valores extremos, es decir tiene un riesgo de sobreestimar sus mediciones cuando aparece un dato atípico. Una alternativa de cálculo fue propuesta por Moors(1988) empleando los percentiles de la distribución por medio de comparar los cuantiles que se encuentran alrededor de los cuartiles, dividiendo en 8 partes iguales el rango, profundizando más en el análisis de los pesos de las colas, minimizando el efecto de valores atípicos, con el coeficiente

$$\boldsymbol{K_2(X) = \frac{(X_{87.5\%} - X_{62.5\%}) + (X_{37.5\%} - X_{12.5\%})}{X_{75\%} - X_{25\%}} * 100\%}$$

Al igual que el cálculo alternativo en la asimetría, el estadístico se escala dividiendo entre el rango intercuartil. Para calcular el exceso de kurtosis se debe primero obtener el valor del estadístico aplicado a una distribución Normal$(0,1)$, por lo tanto con apoyo de tablas estadísticas se tiene que

$$\boldsymbol{K_2\big(N(0,1)\big) = \frac{(1.15 - 0.32) + (-0.32 - (-1.15))}{0.68 - (-0.68)} = 1.23}$$

$$\boldsymbol{ExK_2(X) = K_2(X) - 1.23}$$

Para ejemplificar la forma en que estos estadísticos miden la kurtosis, se pueden emplear las simulaciones generadas en el tema anterior ya que la distribución Pareto es famosa por considerarse una distribución con colas pesadas y además la distribución normal generada es un estándar para medir el exceso de kurtosis, sin embargo para ver el efecto aislado en la distribución Pareto se debe generar una distribución lo más similar posible en cuanto a su media , desviación estándar y asimetría pero que tenga un comportamiento diferente en las colas.

La distribución elegida para esta comparación es la distribución exponencial, ya que es una distribución con una función de densidad que es asimétrica de forma similar a la distribución Pareto y depende de un solo parámetro λ como se puede observar en la gráfica 2.42. Las características principales de esta distribución se resumen por las siguientes expresiones retomadas del capítulo anterior

$$f(x) = \begin{cases} 0, & x < 0 \\ \lambda e^{-\lambda x}, & x \geq 0 \end{cases}$$

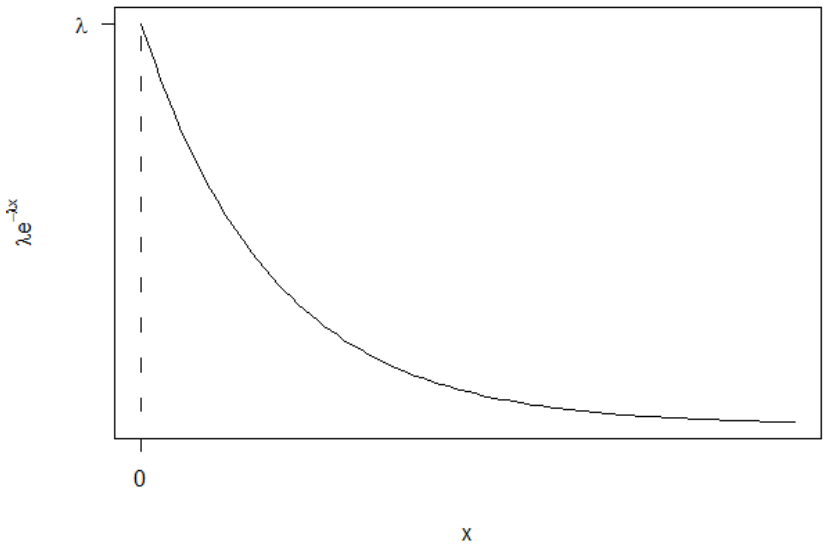
$$E(X) = \int_{-\infty}^{\infty} x f(x) = \frac{1}{\lambda}$$

$$Var(X) = E(X^2) - E^2(X) = \frac{1}{\lambda^2}$$

$$Asimetría = E\left(\left(X - E(X)\right)^3\right) = \frac{2}{\lambda^3}$$

$$Curtosis = E\left(\left(X - E(X)\right)^4\right) = \frac{9}{\lambda^4}$$

Gráfica 2.46: Función de densidad de una distribución exponencial(λ)



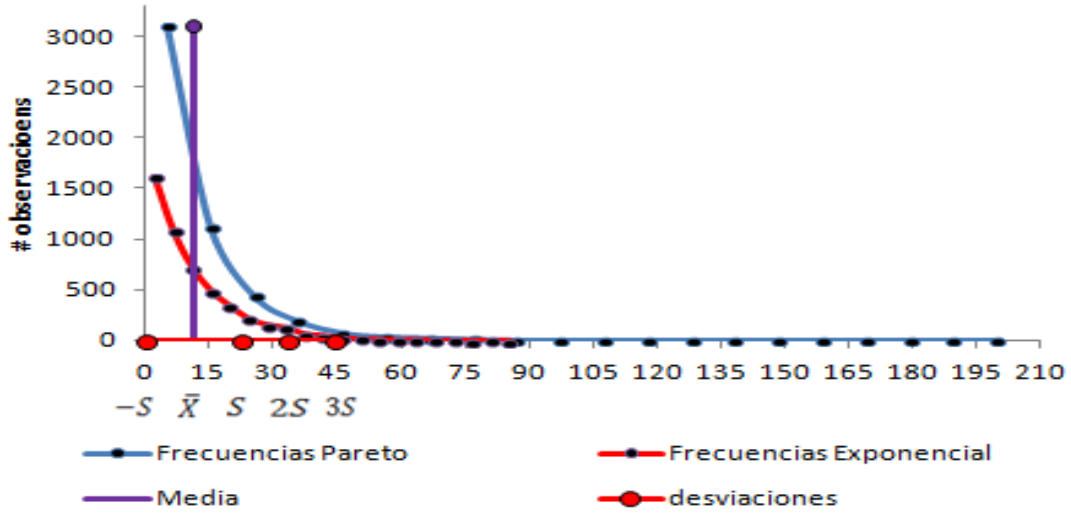
Para generar valores de esta distribución se empleará el método de la transformación inversa a partir de valores en la hoja de cálculo con la formula ALEATORIO() como $\boldsymbol{U}$, retomando la fórmula

$$F^{-1}(U) = -\frac{\ln(U)}{\lambda}$$

Para determinar el valor de λ, se debe recordar que el objetivo es igualar lo más posible los comportamientos de la distribución Pareto($\theta = 100, \alpha = 10$) en consecuencia se iguala el valor de las medias, $\frac{1}{\lambda} = 11.11 \rightarrow \lambda = 0.09$

Una forma de simplificar el proceso es copiar una de las distribuciones antes generadas y cambiar únicamente la fórmula de la última columna. Tomando el tamaño de muestra igual a las simulaciones anteriores $n = 5{,}000$, una vez generada la muestra se generan las curvas de frecuencia y se sobrepone a la de la distribución Pareto que se generó anteriormente lo cual se encuentra en la gráfica 2.47.

Gráfica 2.47: Curvas de frecuencias para 5,000 valores simulados de la distribución Pareto(θ=100, α=10) y una Exponencial (λ=0.09)



La gráfica muestra que las primeras características de posición, dispersión y asimetría son muy similares, aunque debido a la baja frecuencia de los valores extremos se dificulta la visibilidad de las colas, por esta razón se muestran a continuación las tablas de frecuencias que alimentan estos gráficos, para comparar el comportamiento en sus colas.

Tabla 2.32: Frecuencias acumuladas y puntuales de una simulación de la distribución Exponencial (λ=0.09)

i	*Límite Superior*	*Punto medio* m_i	*Frecuencia Acumulada*	*Frecuencia de Subintervalo*
1	4.4	2.2	1,620	1,620
2	8.8	6.6	2,707	1,087
3	13.1	10.9	3,436	729
4	17.5	15.3	3,933	497
5	21.9	19.7	4,280	347
6	26.3	24.1	4,500	220
7	30.6	28.4	4,659	159
8	35.0	32.8	4,790	131
9	39.4	37.2	4,859	69
10	43.8	41.6	4,911	52
11	48.1	45.9	4,935	24
12	52.5	50.3	4,954	19
13	56.9	54.7	4,970	16
14	61.3	59.1	4,978	8
15	65.6	63.4	4,989	11
16	70.0	67.8	4,991	2
17	74.4	72.2	4,996	5
18	78.8	76.6	4,996	0
19	83.1	80.9	4,999	3
20	87.5	85.3	5,000	1

Tabla 2.33: Frecuencias acumuladas y puntuales de una simulación de la distribución Pareto(θ=100,α=10)				
i	*Límite Superior*	*Punto medio* m_i	*Frecuencia Acumulada*	*Frecuencia de Subintervalo*
1	10.3	5.1	3,103	3,103
2	20.5	15.4	4,216	1,113
3	30.8	25.6	4,651	435
4	41.0	35.9	4,841	190
5	51.3	46.1	4,915	74
6	61.5	56.4	4,952	37
7	71.8	66.6	4,977	25
8	82.0	76.9	4,988	11
9	92.3	87.1	4,992	4
10	102.5	97.4	4,994	2
11	112.8	107.6	4,996	2
12	123.0	117.9	4,998	2
13	133.3	128.1	4,999	1
14	143.5	138.4	4,999	0
15	153.8	148.6	4,999	0
16	164.0	158.9	4,999	0
17	174.3	169.1	4,999	0
18	184.5	179.4	4,999	0
19	194.8	189.7	4,999	0
20	205.0	199.9	5,000	1

Como se observa en las tablas las frecuencias de la distribución Exponencial encuentran su máximo en 87.5, mientras que la distribución Pareto tiene observaciones en subintervalos de límites superiores mayores, aunque en bajas frecuencias, por lo tanto se puede inferir que la medición de la kurtosis será mayor en la distribución Pareto. Para corroborar que la simulación cumple con las características deseadas de posición, dispersión y asimetría la siguiente tabla contiene el resumen de estadísticos para la muestra de la distribución exponencial de tamaño $n = 5000$

Tabla 2.34: Resumen de estadísticos principales de una muestra generada de una distribución Exponencial (λ=0.09)			
Estadístico	Valor	Estadístico	Valor
$\bar{X}$	11.2	$x_{(1)}$	0.0
Me_X	7.8	$X_{25\%}$	3.2
M_1	2.2	$X_{75\%}$	15.8
S^2	123	$x_{(n)}$	87.5
S	11.1	$x_{(n)} - x_{(1)}$	87.5
CV_P	98.70%	*Rango Intercuartil*	12.6
μ_3	2,532.80	n	5,000
γ_1	185.70%	γ_2	27.90%

Los valores en cuanto a media y desviación estándar son los esperados aunque el coeficiente de variación de la exponencial es menor que el coeficiente de la distribución Pareto, pero muy cercano a su valor teórico de 1, en cuanto a los

indicadores de asimetría, los coeficientes de Pearson son afectados por los valores extremos, pues calculan diferencias mayores comparadas con los coeficientes basados en cuantiles, estos últimos demuestran que el nivel de asimetría de la muestra exponencial es muy similar a la de la muestra generada de la distribución Pareto.

Ahora que se ha comprobado que se tiene aislado el efecto de la kurtosis entre las distribuciones Exponencial y Pareto, se debe comparar las mediciones de la kurtosis aplicadas también a la muestra generada de la distribución Normal, con el objetivo de comparar los valores del exceso de kurtosis. La siguiente tabla contiene el cálculo de los coeficientes de kurtosis y sus componentes para las muestras generadas de las distribuciones Exponencial, Pareto y Normal.

Tabla 2.35: Cálculo de los Coeficientes de kurtosis y sus componentes de las muestras de las distribuciones simuladas

Estadístico	Distribución simulada		
	Pareto(θ=100,α=10)	Exp(λ=0.09)	Normal(μ=11.1,σ^2=154.3)
$X_{12.5\%}$	1.4	1.5	-3.2
$X_{25\%}$	3	3.2	2.8
$X_{37.5\%}$	4.9	5.2	7.1
$X_{50\%}$	7.3	7.8	11.1
$X_{62.5\%}$	10.4	11.2	15.2
$X_{75\%}$	15.2	15.8	19.6
$X_{87.5\%}$	23.7	23.5	25.6
μ_4	587,450.30	116,665.30	69,279.70
S	12.6	11.1	12.4
$K_1(X)$	2296.90%	771.50%	295.10%
$ExK_1(X)$	1996.90%	471.50%	-4.90%
$K_2(X)$	139.30%	127.50%	123.10%
$ExK_2(X)$	16.30%	4.50%	0.10%

En el comparativo de la tabla se puede identificar que la distribución Pareto en efecto, tiene una mayor kurtosis desde el valor del cuarto momento. Por otro lado, por construcción las distribuciones tienen dispersiones similares. Observado el valor de los cuantiles de la muestra Exponencial y la Pareto se nota que están dentro del mismo rango aunque en una separación mayor, incluso son similares a los cuantiles de la muestra de la distribución Normal, aunque con ciertas diferencias ya que por ejemplo el valor de $X_{12.5\%}$ es negativo.

En la tabla además se muestra cómo la distribución Normal conserva su propiedad en las colas y por lo tanto tiene el valor de $K_2(X)$ cercanos a 0, aunque $K_1(X)$ es negativo, lo cual se interpreta como que la muestras posee una cola más ligera a la de una Normal; sin embargo el valor del coeficiente es un valor muy cercano a 0. Al generar diversas muestras de estos mismos ejemplos se pueden llegar a variaciones en los valores, aunque con las mismas conclusiones en sus comparativos.

La distribución Pareto muestra los valores más altos de todos los indicadores de kurtosis como era esperado, aunque son los estadísticos basados en cuantiles los que permiten la comparación de los resultados, ya que el valor de $K_2(X)$ para la distribución Normal es muy similar al valor teórico. Con esto se puede concluir que la distribución exponencial tiene un 5%, más de kurtosis que la distribución Normal, aunque comparada con la distribución Pareto, bajo los valores particulares elegidos para los parámetros, es menor alrededor de un 12%.

Con los ejemplos anteriores queda más claro el concepto de la kurtosis y se detalla con base en ejemplos visuales y numéricos la forma en que mide una característica independiente, aunque se complica su visualización gráfica cuando las distribuciones son muy asimétricas, pero permite salir de los ejemplos típicos de la literatura estadística y financiera, que sólo comparan distribuciones similares la distribución Normal como la distribución T de Student.

Capítulo III:

Intervalos de confianza

En el tema anterior se revisaron el uso de técnicas de simulación Monte Carlo, como apoyo dentro de la enseñanza de temas de análisis de datos por medio de estadística descriptiva. La idea principal fue mostrar las distintas características generales que se pueden medir con base en una muestra.

Usualmente dentro de un programa educativo de temas estadísticos se incluye el tópico en el que a través de una muestra son construidos estadísticos con el objetivo de estimar el valor de los parámetros de una variable aleatoria conocida, como el parámetro λ de una distribución exponencial, o estimar el verdadero valor de una característica de la población como la media y la varianza.

Este tema es conocido como estimación puntual, el cual se desarrolla mayormente de forma teórica al comprobar diversas propiedades que deben tener los estadísticos construidos, como insesgamiento, consistencia y suficiencia, para identificar el mejor estimador.

En ciertas aplicaciones los posibles valores de parámetros de una distribución, se encuentran en un rango continuo, por lo que las distribuciones de los estimadores puntuales también serán variables aleatorias continuas. A pesar de que el estimador puntual cumpla con una serie de propiedades estadísticas y sea el más preciso, la probabilidad de que el valor del estimador sea igual al valor del parámetro es cero.

Considerando lo anterior, es preferible establecer con base en una muestra, un conjunto de valores como un estimador, de tal forma que contenga al verdadero valor del parámetro con una probabilidad conocida, es decir, con una cierta confianza de cubrirlo. Este conjunto de valores, cuando se trata de parámetros con posibles valores en los reales, se establece entonces como un intervalo, debido a lo cual se les conoce como estimador por intervalo y su definición es la siguiente:

Una estimación por intervalo de un parámetro θ se genera a partir cualquier par de estadísticos $L(x_1, \dots, x_n)$ y $U(x_1, \dots, x_n)$, que cumplan $L(X) \leq U(X)$ para una muestra X del universo de muestras posibles. Con una muestra observada x, la inferencia que se realiza es $L(X) \leq \theta \leq U(X)$ y el estimador por intervalo es el intervalo aleatorio $[L(X)\,,\ U(X)]$.

Una vez definido el intervalo, se puede obtener la probabilidad de que el estimador cubra al verdadero valor de θ, conocida como probabilidad de cobertura. Al ínfimo de las probabilidades de cobertura, se le conoce como el coeficiente de confianza o nivel de confianza del intervalo y se expresa como $\inf\big(P_\theta(\theta \in [L(X)\,,\ U(X)])\big) = 1 - \alpha$. Al estimador por intervalo con un coeficiente de confianza definido se le conoce como intervalo de confianza. El término α, como se verá en el siguiente capítulo, tiene una estrecha relación con las pruebas de hipótesis.

Para mejorar el entendimiento del nivel de confianza y la interpretación de los intervalos, lo que se propone en esta sección es clarificar las propiedades de algunos intervalos de confianza comúnmente tratados en la literatura estadística, simulando muestras de una población con el objetivo de generar ejemplos de cálculo, para posteriormente generar una cantidad suficiente de muestras comprobando la eficacia con la cual el intervalo cubre al verdadero valor del parámetro de interés.

Intervalo de confianza para la media de una población Normal con varianza conocida

El caso más sencillo en la estimación por intervalos es estimar la media μ de una población que se distribuye Normal(μ, σ^2) cuando σ^2 es conocida, a partir de una muestra $X = (X_1, \dots, X_n)$. Con base en el supuesto de que cada X_i proviene de la distribución Normal(μ, σ^2) se tiene que $\bar{X}$ se distribuye $N(\mu, \sigma^2/n)$. El estimador puntual del que parte la construcción del intervalo es $\bar{X}$, pues cumple $E(\bar{X}) = \mu$ y converge al verdadero valor de μ conforme $n \to \infty$.

Una técnica para construir los intervalos de confianza es por medio de una variable aleatoria que esté definida en términos del parámetro pero de tal forma que su distribución no dependa del parámetro, conocida como cantidad pivotal. En este caso se elimina la dependencia de los parámetros por medio de estandarizar la distribución de $\bar{X}$ con la transformación.

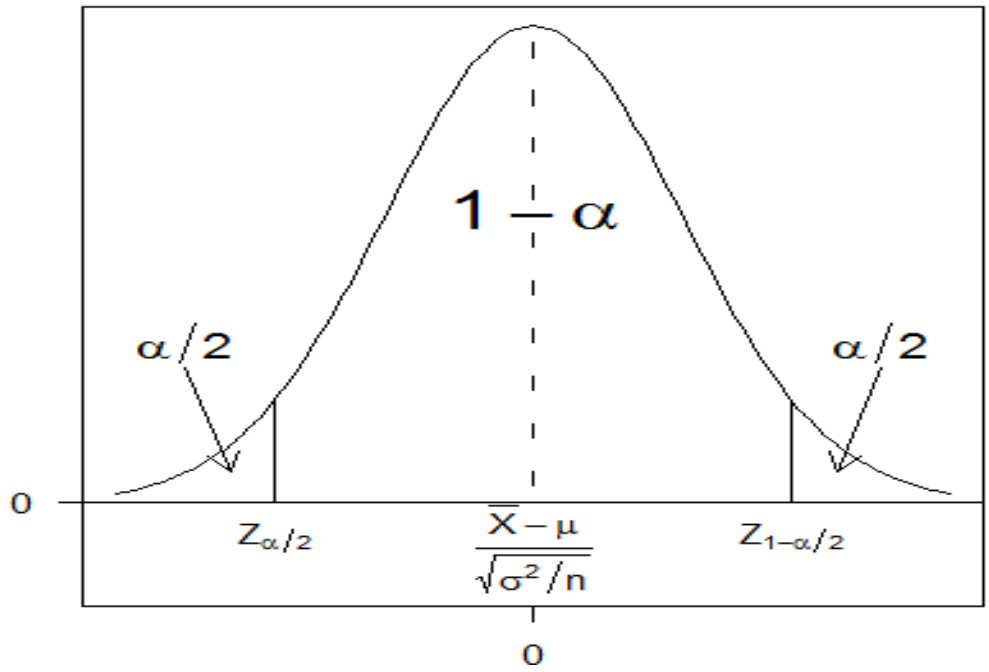
$$\boldsymbol{Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}}$$

La distribución de la variable Z es $Normal(0,1)$, así que no depende el parámetro μ. Ahora para conseguir un nivel de confianza $1 - \alpha$, se debe proponer un intervalo de tal forma que en la cola derecha de Z, haya $\alpha/2$ de probabilidad, y de igual manera debe haber $\alpha/2$ en la cola izquierda, esto se consigue por medio de los cuantiles de la Normal estándar $Z_{1-\alpha/2}$, $Z_{\alpha/2}$, para obtener la siguiente expresión

$$\boldsymbol{P\left(Z_{\frac{\alpha}{2}} < Z < Z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha}$$

Los cuantiles necesarios pueden ser hallados por medio de una tabla de la distribución Normal estándar incluida en un gran número de libros de temas estadísticos, en la Web, en algún software estadístico o de propósitos más generales, por ejemplo en Excel® la función $DISTR.NORM.ESTAND.INV()$ recibe como parámetro la probabilidad p y devuelve el valor x que cumpla $\int_{-\infty}^{x} \boldsymbol{f(Z)dZ = p}$. Debido a la simetría de la Normal se tiene que $\boldsymbol{Z_{\frac{\alpha}{2}} = -Z_{1-\frac{\alpha}{2}}}$ haciendo más fácil el establecimiento del intervalo. La siguiente gráfica muestra la distribución de la cantidad pivotal, donde se indican los elementos hasta el momento mencionados.

Gráfica 3.1: Distribución de $(\overline{X}-\mu)/\sqrt{\sigma^2/n}$



Con base en los cuantiles observados que encierran al parámetro se establece una desigualdad de los cuantiles y la media de la cual se despeja, al parámetro μ para encontrar un intervalo que refleje la confianza de que el verdadero valor del parámetro se encuentre en el intervalo

$$\boldsymbol{Z_{\frac{\alpha}{2}}} < Z < \boldsymbol{Z_{1-\frac{\alpha}{2}}} \quad = \quad \boldsymbol{Z_{\frac{\alpha}{2}}} < \frac{\boldsymbol{\overline{X}-\mu}}{\sqrt{\boldsymbol{\sigma^2/n}}} < \boldsymbol{Z_{1-\frac{\alpha}{2}}}$$

$$\sqrt{\frac{\boldsymbol{\sigma^2}}{\boldsymbol{n}}}\left(-\boldsymbol{Z_{1-\frac{\alpha}{2}}}\right) < \mu - \boldsymbol{\overline{X}} < \sqrt{\frac{\boldsymbol{\sigma^2}}{\boldsymbol{n}}}\boldsymbol{Z_{1-\frac{\alpha}{2}}}$$

$$\boldsymbol{\overline{X}} - \sqrt{\frac{\boldsymbol{\sigma^2}}{\boldsymbol{n}}}\boldsymbol{Z_{1-\frac{\alpha}{2}}} < \mu < \boldsymbol{\overline{X}} + \sqrt{\frac{\boldsymbol{\sigma^2}}{\boldsymbol{n}}}\boldsymbol{Z_{1-\frac{\alpha}{2}}}$$

Entonces el intervalo $\left(\boldsymbol{\overline{X}} - \sqrt{\frac{\boldsymbol{\sigma^2}}{\boldsymbol{n}}}\boldsymbol{Z_{1-\frac{\alpha}{2}}}\ ,\ \boldsymbol{\overline{X}} + \sqrt{\frac{\boldsymbol{\sigma^2}}{\boldsymbol{n}}}\boldsymbol{Z_{1-\frac{\alpha}{2}}}\right)$ contiene el valor del parámetro μ con nivel de confianza del $(1-\alpha)*100\%$. Para ejemplificar el cálculo supóngase por ejemplo que se tiene una muestra de tamaño $n=20$ proveniente de una población simulada por el método de la transformada inversa visto en el capítulo anterior de tal manera que esté distribuida de manera Normal($\mu=7, \sigma^2=10$) desplegada en la Tabla 3.1, con la cual se desea estimar la media de la población con un nivel de confianza del 90%.

Tabla 3.1: Muestra aleatoria de una Normal (μ=7,σ^2=10)					
i	x_i	i	x_i	i	x_i
1	4.01	8	7.6	15	3.25
2	9.59	9	10.28	16	8.31
3	6.54	10	11.5	17	6.24
4	4.67	11	10.85	18	1.59
5	4.06	12	7	19	5.46
6	5.14	13	6.48	20	10.51
7	4.83	14	10.61		
$\sigma^2 = 10$		$\bar{X} = 6.93$	$\alpha = 0.1$	$Q_{1-\alpha/2} = 1.64$	

Ahora lo único que falta hacer es aplicar las formulas anteriores. El límite inferior del intervalo es $6.93 - \sqrt{10/20} * 1.64 = 5.77$, mientras que el límite inferior del intervalo es $6.93 - \sqrt{10/20} * 1.64 = 8.09$, lo que constituye un intervalo símetrico alrededor de $\bar{X}$ cuya longitud depende del nivel de confianza y el tamaño de la muestra. Finalmente, se concluye con un 90% de confianza que el verdadero valor de la media poblacional se encuentra en el intervalo $(5.77,8.09)$, lo que corrobora con sólo 20 valores simulados como se cubre al verdadero valor de μ.

Perspectiva de los intervalos de confianza por medio de simulación Monte Carlo

Ahora se trabajará el concepto de intervalo de confianza por medio de un ejemplo más grande vía simulación, para calcular la proporción de los intervalos que contienen el verdadero valor del parámetro. Supóngase que se desea estimar la media de una distribución Normal$(\mu, \boldsymbol{\sigma^2} = 400)$ con un nivel de confianza del 95% con una muestra de tamaño $n = 100$.

En este caso se registró en una hoja de cálculo los valores verdaderos de los parámetros, considerando $\mu = 300$. Para simular muestras con distribución $N(300,400)$, se ha empleado el método de la transformación inversa con apoyo de la fórmula $INV.NORM()$[18], como en el capítulo anterior.

Puesto que $\alpha = 0.05$ se necesita obtener el cuantil $Z_{0.975}$, el cual tiene un valor de 1.96 obtenido usando la función inversa empleada en la simulación, pero con la probabilidad fija a 97.5%. Posteriormente se necesita se calcular $\bar{X}$ de la muestra.

Con los cálculos anteriores se puede obtener ahora los valores del límite inferior $(limI)$ y superior $(limS)$ aplicando las fórmulas para estimar el intervalo de confianza para la muestra generada. Se debe recordar que si se presiona la tecla $Supr$ toda la muestra se actualizará junto con los cálculos de los estadísticos, por lo que en cada repetición se puede verificar si el intervalo calculado cubre al verdadero valor de la media. Para automatizar la evaluación se usan condicionales a través de la función $SI()$, con la siguiente instrucción $SI(limI \leq \mu, SI(\mu \leq limS,$"SI","NO"$),"NO")$; con lo cual el

[18] Los nombres de las funciones pueden cambiar de acuerdo a la versión o idioma configurados.

resultado de la función devolverá SI en el caso que el intervalo calculado encierre el parámetro y NO en caso contrario.

La repetición consecutiva del cálculo del intervalo tendrá como resultado, que una gran cantidad de intervalos que produzcan el Valor SI y en menor cantidad resultados en NO. Para generar una muestra completa de los resultados binarios, en cada repetición se debe copiar y pegar a parte el resultado de la condicional[19], para después contar el número de muestras en las que se obtuvo el valor SI y dividir con respecto al número de muestras simuladas. Para este caso se generaron 5000 muestras.

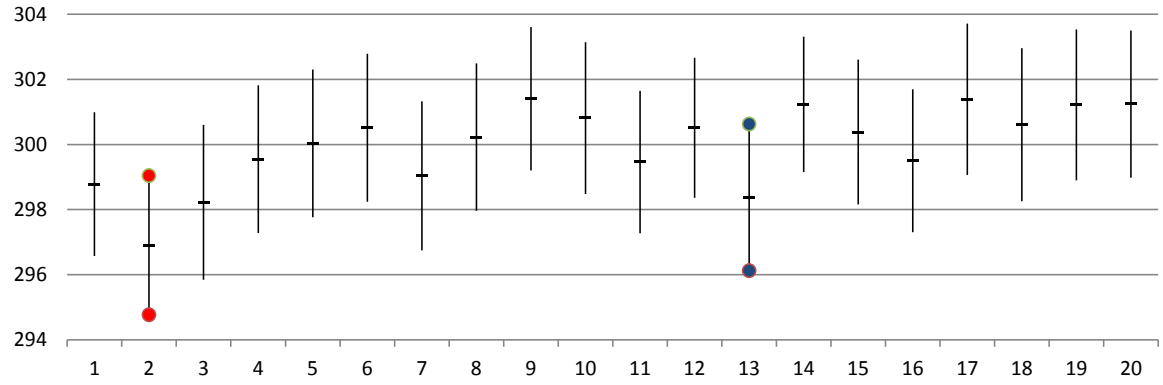
Después de contar el número de veces en las que se obtuvo SI en la condicional, se contaron $4{,}767$ resultados de SI lo que implica que en el 95.34% de las muestras el intervalo de confianza encerró al parámetro. A continuación se muestra una estructura propuesta en una hoja de cálculo para el ejemplo y los intervalos graficados como cotizaciones en Excel®.

Estructura propuesta en hoja de cálculo para el ejemplo

	F	G	H	I	J	K
1		μ verdadero	300	n=	100	
2		σ^2 verdadero	400	α	5%	
3		Confianza	95%	$Z_{1-\alpha/2}$	1.96	
4						
5		Límite Inferior	X barra	Límite Superior	¿Contiene a μ?	
6		300.08	304.00	307.92	NO	
7	Índice	Simulaciones: copias a valor por macro				% de SI
8	1	298.78	296.57	300.99	SI	95.34%
9	2	296.91	294.77	299.04	NO	
10	3	298.23	295.85	300.61	SI	
11	4	299.55	297.28	301.82	SI	
12	5	300.03	297.76	302.30	SI	
13	6	300.51	298.24	302.79	SI	
14	7	299.04	296.75	301.32	SI	
15	8	300.23	297.96	302.49	SI	
16	9	301.41	299.20	303.61	SI	
17	10	300.81	298.48	303.14	SI	
18	11	299.46	297.28	301.64	SI	
19	12	300.52	298.36	302.67	SI	
20	13	298.38	296.12	300.63	SI	
21	14	301.24	299.16	303.32	SI	
22	15	300.38	298.16	302.61	SI	
23	16	299.50	297.31	301.70	SI	
24	17	301.39	299.07	303.71	SI	

[19] La macro que automatiza este proceso se halla en el apéndice.

Gráfica 3.2: Primeros 20 Intervalos de confianza simulados de una Normal (μ=300,σ^2=400) bajo el supuesto de μ desconocida y σ^2 conocida



Analizando el gráfico lo que se observa es que cada uno de los intervalos tienen la misma amplitud, esto debido a que una vez elegido el nivel de confianza y al conocer la varianza el término que se suma y resta a la media es constante para todos los intervalos, sin embargo para un mayor nivel de confianza o un tamaño de muestra distinto la amplitud de los intervalos cambiarían; por esa razón lo único que influye en encerrar el parámetro es el valor de la media muestral, por ejemplo en el intervalo 2 con extremos en rojo, la posición de la media es tal que la amplitud del intervalo no logra encerrar el parámetro, por otra parte se nota que en el intervalo número 13 con los extremos señalados en azul, la media parece alejada del parámetro pero el intervalo apenas consigue encerrar la media verdadera.

El resultado parcial anterior es concordante con el resultado general, ya que se esperaría que 1 de cada 20 intervalos no encerrara al parámetro de interés $\left(\frac{1}{20} = 0.05\right)$. Ahora que el concepto es más intuitivo se pueden tratar de forma análoga los siguientes intervalos de estudio desde su definición, y revisar por medio de simulaciones controladas su comportamiento.

Intervalo de confianza para la media de una población Normal con varianza desconocida

En la sección anterior se consideró conocer dos características de la población, la primera es la definición explicita de su distribución y la segunda el valor su varianza. En la práctica el supuesto de conocer de manera previa alguna característica de una población en muchos casos no es fácil de justificar, pues incluso los supuestos de un analista experimentado pueden ser un error o las características del fenómeno de estudio puede que hayan cambiado en el tiempo.

El trabajar con esta clase de supuestos permite explorar modelos teóricos de la estadística, que son similares pero más simples que la realidad. Posteriormente, según sea necesario, se puede hacer más complejo y preciso el modelo, siempre pensando en que la parsimonia del mismo es también una característica valiosa.

En esta ocasión el supuesto que se desechará es el conocer la varianza de la población, y se conservará el supuesto de Normalidad de la población, con el objetivo de estimar el valor de la media por medio de intervalos de confianza.

El siguiente método se puede encontrar en el libro de Mood (1974) sección 3.1 el cual retoma la propiedad de la variable

$$\boldsymbol{Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1)}$$

Luego se considera el hecho de que si un conjunto variables aleatorias IID $X_i \sim N(0,1)$ se elevan al cuadrado, es decir, si se definen las variables $Y_i = (X_i)^2$, éstas se distribuyen Ji-cuadrada con un grado de libertad $\left(Y_i \sim X^2_{(1)}\right)$. Otra relación de importancia es que si $Y_i \sim X^2_{(1)}$ para $i = 1,2,\ldots,n$ entonces $\boldsymbol{\sum_{i=1}^{n} Y_i \sim X^2_{(n-1)}}$.

Con base en lo anterior otro resultado que se puede demostrar es que la variable $\boldsymbol{\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2}$ se distribuye Ji-cuadrada con $n-1$ grados de libertad.

Otra distribución importante para determinar el intervalo es la distribución t de Student, la cual se puede ver como el resultado de realizar el cociente entre una distribución $N(0,1)$ y la raíz cuadrada de una variable con distribución Ji-cuadrada dividida entre sus grados de libertad.

Determinación del intervalo de confianza

Para hallar el intervalo que con una confianza del $(1-\alpha)*100\%$ contenga al parámetro μ, calculado a partir de una muestra proveniente de la distribución $N(\mu, \sigma^2)$, donde σ^2 es un parámetro desconocido, se pueden emplear las relaciones anteriores entre las variables aleatorias distribuidas Normal, Ji-cuadrada y t de Student, dejando su demostración para una sesión de estudio sobre estos aspectos teóricos.

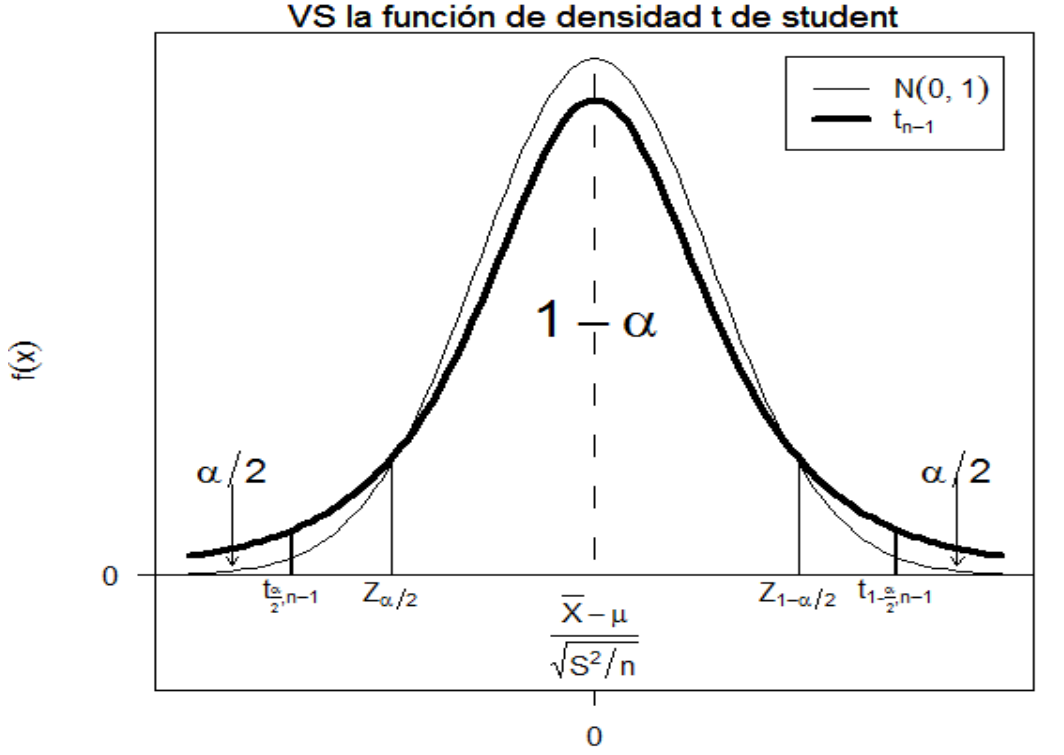
El procedimiento para conseguir una cantidad pivotal que no dependa de los parámetros, es realizado el cociente entre la variable $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ y la raíz de $\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2$ dividida por sus grados de libertad, obteniendo la distribución t de la siguiente manera,

$$\boldsymbol{\frac{\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 / (n-1)}} = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}}$$

Dada la distribución del cociente entonces se determina un intervalo con nivel de confianza de $(1-\alpha)100\%$ por medio de los cuantiles $\boldsymbol{t_{\frac{\alpha}{2},(n-1)}}$ **y** $\boldsymbol{t_{1-\frac{\alpha}{2},(n-1)}}$ los cuales pueden ser obtenidos en Excel® por medio de la función $DIST.T.2C()$ que considera ambas colas y recibe como parámetros una probabilidad p y los $n-1$ grados de libertad, para devolver el valor x que cumpla $\boldsymbol{\int_{-x}^{x} f_t(y)dy = 1-p}$ donde $f_t(y)$ es la función de densidad de la distribución $t_{(n-1)}$. Esta función de densidad se compara contra la densidad de una Normal en la Gráfica 3.3.

El efecto que tiene el desconocimiento de la varianza es tener una mayor incertidumbre en la estimación, lo que se ve reflejado en que la distribución t pues tiene colas más pesadas, esto quiere decir que los cuantiles que acumulan las probabilidades $\alpha/2$ y $1-\alpha/2$ se encuentran más alejados de la media. La gráfica siguiente visualiza el concepto de las colas más pesadas a través de comparar la función de densidad de una t de *Student* con la densidad de la variable construida la cual es Normal estándar.

Gráfica 3.3: Función de densidad Normal de $(\overline{X}-\mu)/\sqrt{S^2/n}$ VS la función de densidad t de student



Entendida gráficamente la idea del intervalo de confianza se puede determinar de forma analítica, no sin antes hacer notar el hecho que debido a la simetría de la distribución t $-\boldsymbol{t_{\frac{\alpha}{2},n-1}} = \boldsymbol{t_{1-\frac{\alpha}{2},n-1}}$ y se eligen de esa manera ya que determinan el intervalo de menor longitud para la probabilidad establecida. Luego, dada la probabilidad de $1-\alpha$ de cubrir a la variable construida con los cuantiles, lo que se debe hacer es trabajar únicamente con la desigualdad y despejar al parámetro de interés.

$$\boldsymbol{P\left(t_{\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n}}} < t_{1-\frac{\alpha}{2}}\right) = 1-\alpha}$$

$$\rightarrow \quad \boldsymbol{t_{\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n}}} < t_{1-\frac{\alpha}{2}} = t_{\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n}}} < t_{1-\frac{\alpha}{2}}}$$

$$\sqrt{\frac{S^2}{n}}\left(-t_{1-\frac{\alpha}{2}}\right) < \mu - \bar{X} < \sqrt{\frac{S^2}{n}}t_{1-\frac{\alpha}{2}}$$

$$\bar{X} - \sqrt{\frac{S^2}{n}}t_{1-\frac{\alpha}{2}} < \mu < \bar{X} + \sqrt{\frac{S^2}{n}}t_{1-\frac{\alpha}{2}}$$

Por lo tanto el intervalo $\left(\bar{X} - \sqrt{\frac{S^2}{n}}t_{1-\frac{\alpha}{2}}\ ,\ \bar{X} + \sqrt{\frac{S^2}{n}}t_{1-\frac{\alpha}{2}}\right)$ contiene el verdadero valor de μ con un $(1-\alpha)*100\%$ de confianza, para ejemplificar el cálculo se puede tratar un ejemplo similar al siguiente, que sea sencillo de aplicación.

Supóngase que se desea estimar la media de una población Normal (μ, σ^2) con un nivel de confianza del 95%, a partir de una muestra $Normal(\mu = 85, \sigma^2 = 90)$ simulada por el método de a transformada inversa de tamaño $n = 20$ que se muestra en la Tabla 3.2; para llevar a cabo el intervalo también se incluye los valores del cuantil al 97.5%, $\bar{X}$ y S^2, necesarios para realizar la determinación del intervalo de confianza.

Tabla 3.2: Muestra de tamaño n=20 de una Normal (μ,σ^2)

i	x_i	i	x_i	i	x_i
1	74.73	8	81.98	15	88.77
2	99.06	9	79	16	83.79
3	85	10	99.61	17	74.03
4	68.67	11	81	18	80.48
5	83.97	12	75.91	19	97.14
6	81.04	13	103.02	20	80.31
7	94.16	14	81.86		
$\bar{X} = 84.68$		$S^2 = 88.76$		$t_{.975,19} = 2.09$	$\alpha = 0.05$

Aplicando las formulas anteriores el límite superior del intervalo es $\mathbf{84.68} + \frac{\mathbf{2.09}}{\sqrt{\frac{\mathbf{88.76}}{\mathbf{19}}}} = \mathbf{89.08}$ mientras el límite inferior es $\mathbf{84.68} - \frac{\mathbf{2.09}}{\sqrt{\frac{\mathbf{88.76}}{\mathbf{19}}}} = \mathbf{80.26}$, con lo cual se concluye con un 95% de confianza que el verdadero valor de la media de la población se halla en el intervalo $(\mathbf{80.26}\ ,\ \mathbf{89.08})$**.**

Simulación de intervalos de confianza para poblaciones Normales con varianza desconocida

Ahora se modificará el esquema de simulación anterior utilizando las mismas técnicas de simulación, definiendo primero los verdaderos valores $\mu = 100$ y $\sigma^2 = 150$, para simular una muestra aleatoria de tamaño n. Aunque en este caso para ilustrar el proceso la elección del tamaño de la muestra Normal, tiene un fuerte impacto pues a medida que n crece el intervalo se reduce, por lo cual al tomar los parámetros de la Normal menores que el ejemplo anterior, se eligió un tamaño de muestra $n = 50$

Suponiendo un nivel de confianza del 95% para obtener el cuantil $t_{1-\frac{\alpha}{2},(n-1)}$, se calculan los limites inferior y superior, calculado con S^2 en lugar de σ^2, y luego se determina con las mismas condicionales anteriores, si el intervalo encerró el parámetro real. La siguiente imagen muestra una posible estructura para mostrar el ejemplo, calculado considerando a S^2 en el cálculo de los limites inferior y superior.

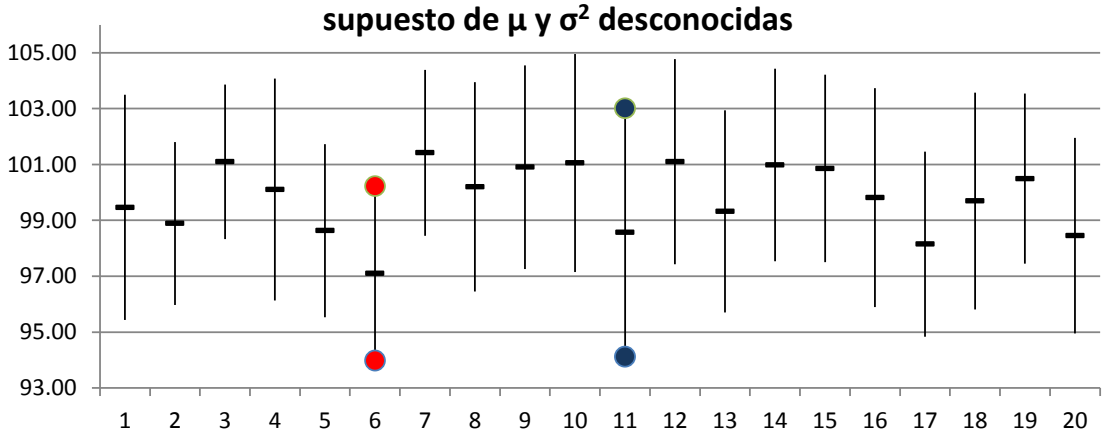
Estructura propuesta para realización en hoja de cálculo del ejemplo

	U	V	W	X	Y	Z	AA	AB
1			μ verdadero	100	n=	50		
2			σ^2 verdadero	150				
3			Confianza	95%	$t_{1-\alpha/2}$	2.01		
4								
5			S^2	Límite Inferior	X barra	Límite Superior	¿Contiene a μ?	
6			142.99	98.08	101.48	104.88	SI	
7								
8		Índice	Simulaciones de copias a valor por macro					% de SI
9		1	152.02	97.15	100.21	103.27	SI	94.8%
10		2	200.82	95.44	99.47	103.49	SI	
11		3	105.29	95.97	98.89	101.80	SI	
12		4	94.77	98.33	101.10	103.86	SI	
13		5	195.73	96.13	100.10	104.08	SI	
14		6	119.03	95.53	98.63	101.73	SI	
15		7	120.66	93.98	97.11	100.23	SI	
16		8	109.05	98.45	101.42	104.39	SI	
17		9	173.97	96.45	100.20	103.95	SI	
18		10	164.62	97.26	100.90	104.55	SI	
19		11	188.58	97.15	101.06	104.96	SI	
20		12	244.84	94.12	98.57	103.01	SI	
21		13	166.57	97.43	101.10	104.77	SI	
22		14	162.29	95.70	99.32	102.94	SI	
23		15	146.95	97.54	100.98	104.43	SI	
24		16	139.70	97.50	100.86	104.22	SI	

Una vez generada la muestra, se realiza el cálculo del porcentaje de veces en que el intervalo cubrió al parámetro. Para estas 5000 simulaciones 94.8% lograron cubrir el verdadero valor de la media, que es un porcentaje muy cercano al nivel de confianza que estableció en un principio.

Los intervalos simulados ahora tienen una longitud variable gracias a considerar a S^2, en el cálculo del intervalo, tal efecto también influye en si el intervalo encierra o no al parámetro de interés; la siguiente gráfica contiene a los primeros intervalos generados por simulación la cual permite visualizar el efecto de la incertidumbre en la estimación desconociendo el valor σ^2.

Gráfica 3.4: Primeros Intervalos de confianza de una muestra de valores simulados de una Normal (μ=100,σ²=150)bajo el supuesto de μ y σ² desconocidas



Como se puede notar en el gráfico se resaltan dos de los intervalos uno con extremos rojos el cual apenas logra encerrar al verdadero valor de la media pues se espera que 1 de estos 20 no cubra al parámetro, mientras que el intervalo señalado con extremos azules cubre al verdadero valor de la media.

Intervalo de confianza para la varianza de una población Normal con media desconocida

Ahora se verá el caso en que se necesita estimar un intervalo sobre la varianza, que en este caso será la σ^2 de una población Normal (μ, σ^2). Lo primero que se necesita es recordar que $\frac{(n-1)S^2}{\sigma^2}$ se distribuye $X^2_{(n-1)}$. Después se establece un nivel de confianza α, para trabajar con la siguiente desigualdad de la probabilidad

$$\boldsymbol{P\left(R_{\frac{\alpha}{2}} < \frac{(n-1)S^2}{\sigma^2} < R_{1-\frac{\alpha}{2}}\right) = 1-\alpha}$$

$$\boldsymbol{R_{\frac{\alpha}{2}} < \frac{(n-1)S^2}{\sigma^2} < R_{1-\frac{\alpha}{2}}}$$

Donde R_p es el cuantil de la distribución Ji-cuadrada que acumula un $p\%$ de probabilidad. Al ser la varianza siempre positiva entonces ambos cuantiles son positivos, por lo tanto

$$\boldsymbol{1/R_{1-\frac{\alpha}{2}} < \frac{\sigma^2}{(n-1)S^2} < 1/R_{\frac{\alpha}{2}}}$$

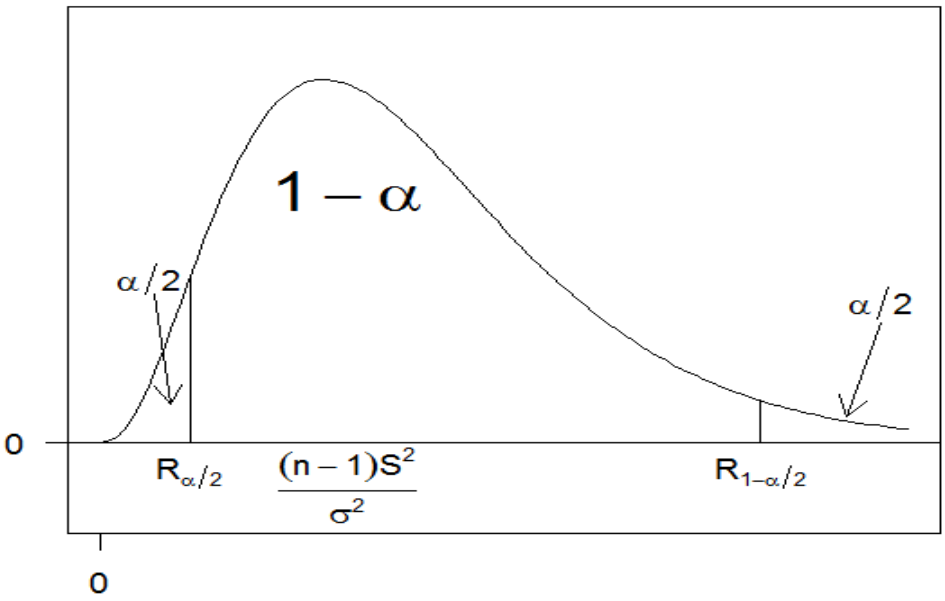
$$\boldsymbol{(n-1)S^2/R_{1-\frac{\alpha}{2}} < \sigma^2 < (n-1)S^2/R_{\frac{\alpha}{2}}}$$

Por lo tanto el intervalo $\left(\frac{(n-1)S^2}{R_{1-\frac{\alpha}{2}}}, \frac{(n-1)S^2}{R_{\frac{\alpha}{2}}}\right)$ contiene al parámetro σ^2 con nivel de confianza del $(1-\alpha)*100\%$. En la gráfica 3.5 se puede ver los cuantiles de la función de densidad de una variable Ji-cuadrada definida como

$$f(x) = \frac{1}{\left(2^{\frac{n}{2}}\ \Gamma\left(\frac{n}{2}\right)\right)}\ \ x^{\frac{n}{2}-1}\ e^{-\frac{x}{2}}$$

De la cual se parte para calcular el intervalo de confianza

Gráfica 3.5: Densidad de una X^2 y sus cuantiles



Para ejemplificar la forma de cálculo se puede revisar un ejemplo sencillo. Lo primero es generar una muestra de una distribución Normal($\mu = 100, \sigma^2 = 50$) que se encuentra totalmente especificada, de tamaño $n = 20$, que por facilidad se resume en cambiar los parámetros de las hojas de cálculo antes construidas. A partir de la muestra se desea establecer un intervalo en el cual se encuentre el parámetro σ^2 con nivel de confianza del 95%, pues se supone a σ^2 como desconocida.

La información de la muestra se halla desplegada en la Tabla 3.3 junto con los cuantiles de la distribución Ji-cuadrada y S^2 necesarios, para el cálculo del intervalo. Se puede notar que en esta estimación no es necesario conocer el valor de la media y tampoco está involucrada en los cálculos del intervalo, aunque se agrega para comprobar que el comportamiento de la muestra es el requerido.

Tabla 3.3: Ejemplo de una muestra Normal (μ,σ^2); n=20					
i	x_i	i	x_i	i	x_i
1	89.03	8	91.16	15	113.55
2	89.47	9	102.19	16	79.29
3	116.71	10	86.75	17	83.71
4	107.12	11	118.59	18	94.59
5	99.65	12	99.18	19	97.59
6	117.85	13	83.25	20	120.19
7	86.22	14	99.06		
$S^2 = 170.84$		$R_{0.025} = 8.91$		$R_{0.975} = 32.85$	

Lo que falta realizar es la aplicación de la fórmula obtenida para el intervalo, por lo que el cálculo del límite inferior es $\frac{(20-1)170.84}{32.85} = \mathbf{98.80}$ mientras que el imite superior es s $\frac{(20-1)170.84}{8.91} = \mathbf{364.45}$ con lo cual se puede concluir que el intervalo $(\mathbf{98.80}, \mathbf{364.45})$ contiene el verdadero valor de σ^2con un 95% de confianza.

Análisis de los intervalos de confianza para la varianza de una población Normal por medio de simulación Monte Carlo

Dado que se sigue trabajando con poblaciones de distribución Normal se puede seguir utilizando el mismo esquema de simulación sobre la misma hoja de cálculo que se ha estado manejando o se puede copiar para tener un mejor orden y separación en los temas. El objetivo es generar una cantidad suficiente de muestras simuladas para poder analizar el comportamiento de los intervalos de cada muestra, con respecto al valor de σ^2.

Basados en la estructura de la simulación anterior se puede generar diversas muestras, donde los elementos clave son: una tabla de los parámetros de la distribución Normal (μ, σ^2), y el nivel de confianza considerado que en este caso es del 90% ($\alpha = .1$), además de los respectivos cuantiles calculados empleando la fórmula $INV.CHICUAD()$, valuada en $\alpha/2$ y $1-\alpha/2$, con $n-1=99$ grados de libertad. Para referenciar los parámetros anteriores y cambiarlos a voluntad, se debe generar un par de columnas que contenga un índice junto con la simulación de la muestra Normal de tamaño n, finalmente el cálculo de S^2 con los limites inferior, superior y la condicional que permita saber al instante si el intervalo contiene el verdadero valor de la varianza.

Para llevar a cabo este análisis se generó una variable aleatoria Normal($\mu = 800, \sigma^2 = 500$) donde el tamaño de la columna que define a la muestra fue $n = 100$, Una vez con esto un programa escrito en VBA (con procesos Macro)[20] copió los estadísticos a valores para generar un total de 1000 realizaciones de intervalos sobre S^2 .

La siguiente imagen muestra una modificación a la estructura base de la simulación de intervalos de confianza, pues ahora el término S^2 debe estar en medio del límite

[20] Consulte el código en el apéndice.

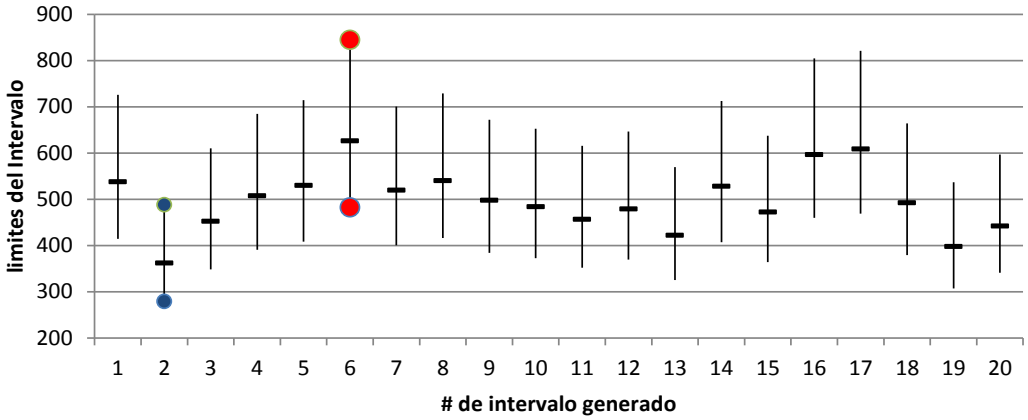
inferior y superior, para dejar en claro que el parámetro a estimar es la varianza, e incluso es opcional agregar la media muestral.

Estructura propuesta en hoja de cálculo para la generación de intervalos sobre la varianza

	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG
1	intervalo de confianza para sigma cuadrada										
2											
3							μ verdadero	800	n=	100	
4							σ^2 verdadero	500	$X^2_{\alpha/2}$	73.36	
5							α	5%	$X^2_{1-\alpha/2}$	128.42	
6											
7							Limite Inferior	S^2	Limite Superior	¿Contiene a σ^2?	
8		Muestra Normal					324.94	421.51	568.83	SI	
9		Indice	Aleatorio(0,1)	Normal(800, 500)		Indice	Simulaciones de copias a valor por macro				% de SI
10		1	0.693	811.31		1	414.61	537.83	725.80	SI	94.3%
11		2	0.585	804.78		2	278.93	361.82	488.28	NO	
12		3	0.839	822.16		3	348.50	452.07	610.07	SI	
13		4	0.337	790.57		4	390.97	507.16	684.41	SI	
14		5	0.196	780.85		5	408.23	529.56	714.63	SI	
15		6	0.111	772.68		6	482.63	626.07	844.87	SI	
16		7	0.360	791.98		7	400.28	519.24	700.71	SI	
17		8	0.232	783.61		8	416.31	540.04	728.78	SI	
18		9	0.207	781.75		9	383.77	497.83	671.81	SI	
19		10	0.142	776.05		10	372.76	483.54	652.54	SI	
20		11	0.181	779.61		11	351.77	456.31	615.79	SI	
21		12	0.994	856.21		12	369.20	478.93	646.31	SI	
22		13	0.009	747.02		13	325.23	421.89	569.34	SI	
23		14	0.983	847.55		14	406.92	527.86	712.34	SI	
24		15	0.227	783.25		15	364.06	472.26	637.31	SI	

Como resultado de la simulación, 943 de los intervalos contuvieron al verdadero valor de σ^2, lo que implica un porcentaje de cobertura del 94.3% que es muy cercano al nivel de confianza establecido. Para realizar un análisis gráfico del comportamiento de los intervalos, se puede recurrir a una gráfica de máximos y mínimos ya vista previamente para compararlos en tamaño y posición con respecto al verdadero valor de $\sigma^2 = 500$.

Gráfica 3.6: Primeros Intervalos de confianza para σ^2, a a partir de una muestra Normal (μ=800,σ^2=500)



En la gráfica anterior se nota además de la variabilidad de la posición, una variabilidad en el tamaño del intervalo, además de notar la asimetría del intervalo debido al estar basado en cuantiles de la distribución teórica Ji-cuadrada. Por ejemplo tomando como referencia el segundo intervalo con extremos en azul puesto que se esperaba que uno

de los 20 no contuviera al parámetro resultado tanto de su posición como su tamaño, mientras que el intervalo señalado con extremos en rojo logra contener al intervalo cerca de su límite inferior aunque en este caso comparando contra el verdadero valor de la varianza el límite superior del intervalo sobreestima la varianza cerca del valor 800 y se vuelve menos preciso.

Intervalo de confianza para la diferencia de medias de poblaciones normales con tamaños de muestra desigual

Dentro de las investigaciones científicas y comerciales es común al analizar problemas en los que se involucran más de una población, para medir que tan similares pueden ser entre sí. El primer comportamiento que se puede medir es la diferencia de las medias de las poblaciones para ver su cercanía o lejanía.

Considérese el caso sencillo en el que se desean estudiar la similitud de 2 poblaciones a partir de muestras de distintos tamaños $\boldsymbol{X = X_1, X_2, \dots, X_{n_1}}$ y $\boldsymbol{Y = Y_1, Y_2, \dots, Y_{n_2}}$ bajo el supuesto de que ambas se distribuyen Normal con $\boldsymbol{X \sim N(\mu_X, \sigma^2)}$, $\boldsymbol{Y \sim N(\mu_Y, \sigma^2)}$ donde se supone que la muestra X es independiente de Y y ambas varianzas se consideran desconocidas pero iguales. Como se verá más adelante, el supuesto de que sean de tamaños diferentes impacta en la determinación de la cantidad pivotal.

Se desea hallar un intervalo de confianza para la diferencia de las medias $\boldsymbol{\mu_X - \mu_Y}$ con un nivel de confianza del $(1-\alpha) * 100\%$, entonces se debe encontrar la distribución de la diferencia $X - Y$ como variable de apoyo para determinar el intervalo de $1 - \alpha$ de confianza.

Lo primero que se debe retomar de la teoría estadística es que $\boldsymbol{\overline{X} - \mu_X \sim N\left(0, \frac{\sigma^2}{n_1}\right)}$ **y** $\boldsymbol{\overline{Y} - \mu_y \sim N\left(0, \frac{\sigma^2}{n_2}\right)}$ por lo que $\boldsymbol{(\overline{X} - \mu_X) - (\overline{Y} - \mu_y) \sim N\left(0, \frac{\sigma^2}{n_2} + \frac{\sigma^2}{n_1}\right)}$. Por lo tanto el cociente $\boldsymbol{\frac{(\overline{X} - \mu_X) - (\overline{Y} - \mu_y)}{\sqrt{\frac{\sigma^2}{n_2} + \frac{\sigma^2}{n_1}}}}$ se distribuye $N(0,1)$

Por otra parte aunque las varianzas se suponen desconocidas, se debe utilizar las S_X^2 y S_Y^2 de las cuales se sabe además que $\boldsymbol{\frac{(n_1-1)S_X^2}{\sigma} \sim X^2_{n_1-1}}$ **y** $\boldsymbol{\frac{(n_2-1)S_Y^2}{\sigma} \sim X^2_{n_2-1}}$**,** las cuales son independientes, entonces la suma de ambas variables dará como resultado una Ji-cuadrada con grados de libertad igual a la suma de los grados de libertad individuales de cada variable, dando como resultado una variable Ji-cuadrada con $\boldsymbol{(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2}$ grados de libertad.

Realizando el cociente entre la Normal antes vista y la raíz de una Ji-cuadrada dividida entre sus grados de libertad, se obtiene una distribución t de *Student* con $n_1 + n_2 - 2$ grados de libertad, es decir:

$$\frac{\frac{(\bar{X}-\mu_X)-(\bar{Y}-\mu_y)}{\sqrt{\frac{\sigma^2}{n_2}+\frac{\sigma^2}{n_1}}}}{\frac{\frac{(n_1-1)S_X^2}{\sigma}+\frac{(n_2-1)S_Y^2}{\sigma}}{\sqrt{n_1+n_2-2}}} = \frac{(\bar{X}-\mu_X)-(\bar{Y}-\mu_y)}{\frac{\sqrt{(n_1-1)S_X^2+(n_2-1)S_Y^2}}{\sqrt{n_1+n_2-2}} * \sqrt{\frac{1}{n_2}+\frac{1}{n_1}}} \quad \sim t_{n_1+n_2-2}$$

Ahora si se define la varianza ponderada S_P^2 como $\frac{(n_1-1)S_X^2+(n_2-1)S_Y^2}{n_1+n_2-2}$ entonces la distribución t de Student toma la siguiente forma

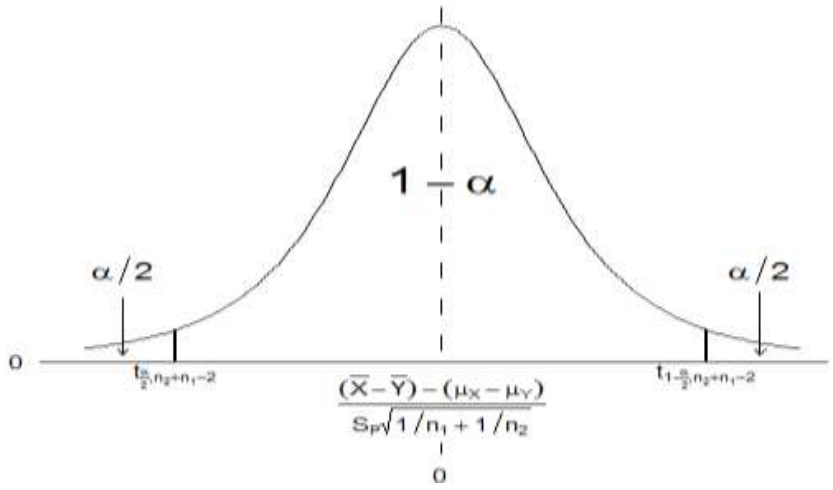
$$\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_y)}{S_P * \sqrt{\frac{1}{n_2}+\frac{1}{n_1}}} \sim t_{n_1+n_2-2}$$

Con la distribución anterior se establece el intervalo de confianza para la diferencia de medias $\mu_X - \mu_y$ despejando de la desigualdad interna de la siguiente probabilidad establecida

$$P\left(T_{\alpha/2} < \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_y)}{S_P * \sqrt{\frac{1}{n_2}+\frac{1}{n_1}}} < T_{1-\alpha/2}\right) = 1-\alpha$$

Donde $T_{\alpha/2}$ y $T_{1-\alpha/2}$ son los cuantiles de la distribución t que acumulen $\alpha/2$ y $1-\alpha/2$ respectivamente, aunque debido a la simetría de la distribución t se tiene $-T_{\alpha/2} = T_{1-\alpha/2}$

Gráfica 3.7: Densidad de la pivotal t de student con la cual se obtiene el intervalo de diferencias de medias



$$-T_{1-\alpha/2} < \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_y)}{S_P * \sqrt{\frac{1}{n_2}+\frac{1}{n_1}}} < T_{1-\alpha/2} \rightarrow$$

$$-T_{1-\alpha/2} * S_P * \sqrt{\frac{1}{n_2}+\frac{1}{n_1}} < (\bar{X}-\bar{Y}) - (\mu_X - \mu_y) < T_{1-\alpha/2} * S_P * \sqrt{\frac{1}{n_2}+\frac{1}{n_1}}$$

$$(\bar{X}-\bar{Y}) - T_{1-\alpha/2} * S_P * \sqrt{\frac{1}{n_2}+\frac{1}{n_1}} < (\mu_X - \mu_y) < (\bar{X}-\bar{Y}) + T_{1-\alpha/2} * S_P * \sqrt{\frac{1}{n_2}+\frac{1}{n_1}}$$

Entonces $\left((\bar{X}-\bar{Y}) - T_{1-\frac{\alpha}{2}} * S_P * \sqrt{\frac{1}{n_2}+\frac{1}{n_1}} \quad , \quad (\bar{X}-\bar{Y}) + T_{1-\alpha/2} * S_P * \sqrt{\frac{1}{n_2}+\frac{1}{n_1}} \right)$ es el intervalo para la diferencia de medias con un $(1-\alpha)$*100 % de confianza. Como ejemplo supóngase que se tienen dos poblaciones Normales X y Y de las cuales se extraen muestras de tamaño $n_1 = 15$ y $n_2 = 20$ respectivamente y se quiere obtener un intervalo con el 95% de confianza para la diferencia de medias, $\mu_X - \mu_y$. Las tablas siguientes muestran los valores de ambas muestras junto con los estadísticos resumen de importancia para la determinación del intervalo.

Tabla 3.4: Muestra de la variable X de tamaño 20

i	x_i	i	x_i	i	x_i
1	92.78	8	76.31	15	84.89
2	100.59	9	97.1	16	99.66
3	105.54	10	97.61	17	118.6
4	110.51	11	132.35	18	91.55
5	102.93	12	105.56	19	107.55
6	112.22	13	88.47	20	110
7	103.72	14	72.08		
$\bar{X} = 100.50$		$S_X^2 = 196.23$		$T_{1-\frac{\alpha}{2}} = 1.97$	$\alpha = 5\%$

Tabla 3.5: Muestra de la variable Y de tamaño 15

i	y_i	i	y_i
1	65.55	9	93.51
2	80.79	10	96.31
3	74.82	11	88.08
4	102.95	12	100.88
5	74.55	13	93.03
6	107.68	14	98.37
7	114.06	15	80.67
8	89.36		
$\bar{Y} = 90.71$		$S_Y^2 = 181.77$	

Con los datos de las tablas se calcula $\boldsymbol{S_P = \sqrt{\frac{(19)*196.23+9*181.77}{28}} = 13.84}$ por lo tanto el intervalo de confianza es $\boldsymbol{(100.50 - 90.71) \mp 1.97 * 13.84 * \sqrt{\frac{1}{20}+\frac{1}{10}} = (-0.77, 20.36)}$**,** así que se tiene un 95% de confianza de que la verdadera diferencia de medias se halla en el intervalo recién calculado.

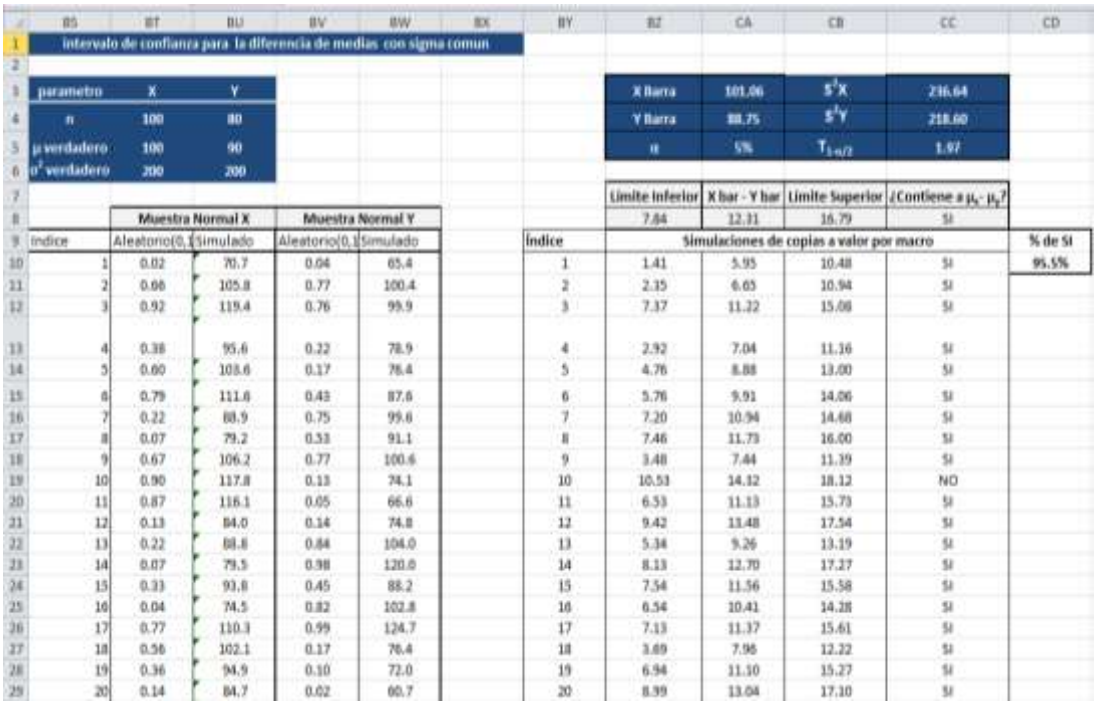
Simulación de intervalos para la diferencia de medias

Sin necesidad de cambiar la estructura de las simulaciones hechas en los anteriores subtemas de intervalos de confianza, se debe primero establecer los verdaderos parámetros de las muestras a simular, que deben ser las medias de ambas distribuciones y una varianza única. Luego se simularán dos poblaciones Normales de distintos tamaños para cada muestra, aunque de acuerdo a la construcción del intervalo al ser independientes las variables se puede tener el caso en que su tamaño de muestra sea el mismo.

Luego se establecen los parámetros del intervalo como el nivel de confianza α, a partir del cual se encuentra el cuantil de la distribución t, se calculan los límites inferior y superior del intervalo, así como la diferencia de medias $\bar{X} - \bar{Y}$ las S_X^2 , S_Y^2 y S_P^2. Por último se establecen las condicionales que permitan saber a simple vista si el intervalo contiene a la verdadera diferencia de medias.

Con esta estructura cualquier movimiento sobre la hoja de cálculo dará como resultado una nueva muestra e intervalo, por lo cual una macro de copiado y pegado a valores dará la posibilidad de analizar el porcentaje de veces que el intervalo ha encerrado a la verdadera diferencia de media. La imagen siguiente muestra un ejemplo diferente sobre la estructura de la hoja de cálculo.

Estructura propuesta para el cálculo de intervalos de confianza sobre la diferencia de medias



intervalo de confianza para la diferencia de medias con sigma comun

parametro	X	Y
n	100	80
μ verdadero	100	90
σ² verdadero	200	200

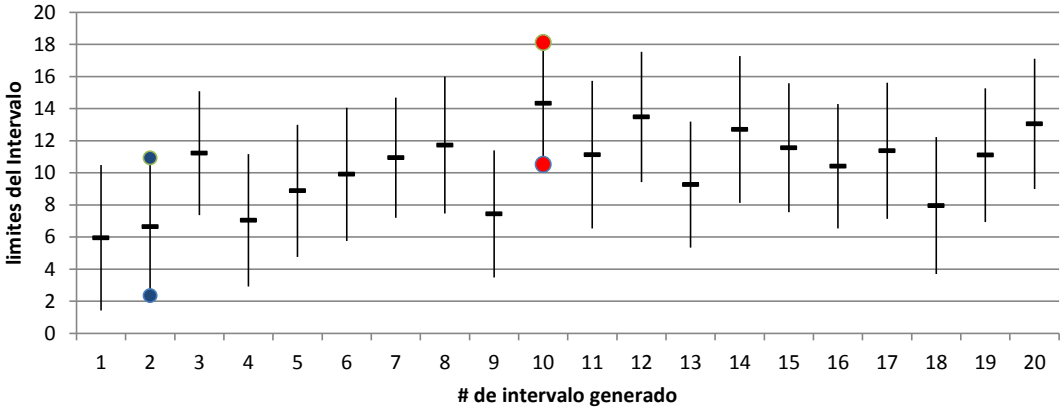
X Barra	101.06	S^2X	216.64
Y Barra	88.75	S^2Y	218.60
α	5%	$T_{1-\alpha/2}$	1.97

Limite Inferior	X bar - Y bar	Limite Superior	¿Contiene a $\mu_x - \mu_y$?
7.84	12.31	16.79	SI

	Muestra Normal X		Muestra Normal Y			Simulaciones de copias a valor por macro				
Indice	Aleatorio(0,1)	Simulado	Aleatorio(0,1)	Simulado	Indice					% de SI
1	0.02	70.7	0.04	65.4	1	1.41	5.95	10.48	SI	95.5%
2	0.66	105.8	0.77	100.4	2	2.35	6.65	10.94	SI	
3	0.92	119.4	0.76	99.9	3	7.37	11.22	15.08	SI	
4	0.38	95.6	0.22	78.9	4	2.92	7.04	11.16	SI	
5	0.60	103.6	0.17	76.4	5	4.76	8.88	13.00	SI	
6	0.79	111.6	0.43	87.6	6	5.76	9.91	14.06	SI	
7	0.22	88.9	0.75	99.6	7	7.20	10.94	14.68	SI	
8	0.07	79.2	0.53	91.1	8	7.46	11.73	16.00	SI	
9	0.67	106.2	0.77	100.6	9	3.48	7.44	11.39	SI	
10	0.90	117.8	0.13	74.1	10	10.53	14.12	18.12	NO	
11	0.87	116.1	0.05	66.6	11	6.53	11.13	15.73	SI	
12	0.13	84.0	0.14	74.8	12	9.42	13.48	17.54	SI	
13	0.22	88.8	0.84	104.0	13	5.34	9.26	13.19	SI	
14	0.07	79.5	0.98	120.0	14	8.13	12.70	17.27	SI	
15	0.33	93.8	0.45	88.2	15	7.54	11.56	15.58	SI	
16	0.04	74.5	0.82	102.8	16	6.54	10.41	14.28	SI	
17	0.77	110.3	0.99	124.7	17	7.13	11.37	15.61	SI	
18	0.56	102.1	0.17	76.4	18	3.89	7.96	12.22	SI	
19	0.36	94.9	0.10	72.0	19	6.94	11.10	15.27	SI	
20	0.14	84.7	0.02	60.7	20	8.99	13.04	17.10	SI	

En la estructura anterior, un cambio en alguno de los parámetros establecidos ocasionará una cambio en los estadísticos calculados y por lo tanto en el resultado de la condicional, la simulación anterior constó de un total de 1000 repeticiones de las cuales como se puede ver el 95.5% de las simulaciones contiene a la verdadera diferencia de medias, porcentaje cercano al nivel de confianza. Para la perspectiva visual la siguiente gráfica muestra el comportamiento de los primeros intervalos entre los cuales se puede apreciar su distinto tamaño debido al cálculo de la S_P^2 y se observa además la variabilidad de su posición. Se puede observar que el décimo intervalo es el intervalo que se espera, de los primeros 20 simulados, no contenga la verdadera diferencia de medias.

Gráfica 3.8: Primeros Intervalos de confianza para la diferencia de medias de muestras Normal (μ=100,σ^2=100) Vs Normal(μ=90,σ^2=200)



Intervalo de confianza para la diferencia de proporciones

En diversos estudios es común tener variables dicotómicas que determinan una característica de una población o una decisión que debe de tomar algún conjunto de personas por ejemplo la intención de voto en una elección popular, lo que conlleva a estudiar la proporción de cada respuesta. Una de las formas de comparar poblaciones distintas cuando se estudia una misma variable, es a través de tomar las diferencias de sus proporciones.

En esta sección se verá el método que se puede hallar en el libro de Hogg sección 5.4.2, el cual determina un intervalo de confianza aproximado a partir de la distribución asintótica de la diferencia de proporciones, de forma análoga al intervalo anterior para la diferencia de medias, y se estudiará las características del intervalo por medio de simulaciones Monte Carlo como se ha venido trabajando.

Supóngase que se tienen dos muestras aleatorias $X = X_1, X_2, \dots, X_{n_1}$ y $Y = Y_1, Y_2, \dots, Y_{n_2}$ independientes distribuidas Bernoulli es decir $X \sim Bernoulli(p_1)$, $Y \sim Bernoulli(p_2)$ para determinar un intervalo con una confianza del $(1-\alpha) * 100\%$ para la diferencia $p_1 - p_2$ lo primero que se retoma son las siguientes propiedades de la distribución Bernoulli

$$\boldsymbol{si\ X \sim Bernoulli(p)\ entonces}$$

$$\boldsymbol{E(X) = p}$$

$$\boldsymbol{Var(X) = p(1-p)}$$

Para esta variable en particular al tomar la media muestral $\bar{X}$ se computa la proporción de veces que la muestra contiene al valor 1 (Éxitos), por lo cual se denota como $\bar{X} = \bar{p}$. Además debe recordarse una propiedad de esta distribución, pues la suma de n variables Bernoulli dan como resultado una variable Binomial(n, p), conocida comúnmente por ser aproximada a medida que n aumenta, por medio de una distribución Normal$(np, np(1-p))$.

El razonamiento anterior aunado al Teorema Central del Límite justifica que el promedio de las variables $\bar{p}$ tienda a distribuirse como una $\boldsymbol{Normal\left(p, \frac{p(1-p)}{n}\right)}$ conforme el tamaño de la muestra aumente**.**

Si se tienen dos poblaciones de la cuales se desea comparar sus proporciones entonces se toma la diferencia $\overline{p_1} - \overline{p_2}$ que es posible aproximarla mediante la distribución $\boldsymbol{N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)}$. En el caso cuando se trató la diferencia de medias de distribuciones normales se tomó el supuesto que ambas varianza son iguales, sin embargo en este caso no es posible considerar ambas varianzas iguales pues se resume en los casos $p_1 = p_2$ o $p_1 = 1 - p_2$, por esta razón se considera la estimación de la variable Normal. Ahora si se toma $\frac{\overline{p_1}(1-\overline{p_1})}{n_1} + \frac{\overline{p_2}(1-\overline{p_2})}{n_2}$ como estimador de la varianza, un resultado estadístico indica que si $\bar{p}$ es un estimador que converge al valor de p cuando n aumenta entonces también la función $\bar{p}(1-\bar{p})$ convergerá al verdadero valor de $p(1-p)$, por lo cual la suma de los estimadores será la forma análoga a la varianza ponderada de la distribución Normal empleada en la sección anterior.

$$\frac{(\overline{p_1} - p_1) - (\overline{p_2} - p_2)}{\sqrt{\frac{\overline{p_1}(1-\overline{p_1})}{n_1} + \frac{\overline{p_2}(1-\overline{p_2})}{n_2}}} \rightarrow \boldsymbol{N(0,1)}$$

Puede notarse que en lugar de usar el símbolo de distribución "$\sim$" se usó una flecha para indicar que es una aproximación que converge a una Normal a medida que n crece. Bajo este esquema el intervalo de $1 - \alpha * 100\%$ de confianza se determina de la siguiente manera

$$\boldsymbol{P}\left(\boldsymbol{Z_{\alpha/2}} \leq \frac{(\overline{p_1} - p_1) - (\overline{p_2} - p_2)}{\sqrt{\frac{\overline{p_1}(1-\overline{p_1})}{n_1} + \frac{\overline{p_2}(1-\overline{p_2})}{n_2}}} \leq \boldsymbol{Z_{1-\alpha/2}} \right) = \boldsymbol{1-\alpha}$$

$$\boldsymbol{-Z_{1-\alpha/2}} \leq \frac{(\overline{p_1} - \overline{p_2}) - (p_1 - p_2)}{\sqrt{\frac{\overline{p_1}(1-\overline{p_1})}{n_1} + \frac{\overline{p_2}(1-\overline{p_2})}{n_2}}} \leq \boldsymbol{Z_{1-\alpha/2}}$$

$$-\boldsymbol{Z_{1-\alpha/2}}\sqrt{\frac{\overline{\boldsymbol{p_1}}(\boldsymbol{1}-\overline{\boldsymbol{p_1}})}{\boldsymbol{n_1}}+\frac{\overline{\boldsymbol{p_2}}(\boldsymbol{1}-\overline{\boldsymbol{p_2}})}{\boldsymbol{n_2}}} \leq (\overline{\boldsymbol{p_1}}-\overline{\boldsymbol{p_2}})-(\boldsymbol{p_1}-\boldsymbol{p_2}) \leq \boldsymbol{Z_{1-\alpha/2}}\sqrt{\frac{\overline{\boldsymbol{p_1}}(\boldsymbol{1}-\overline{\boldsymbol{p_1}})}{\boldsymbol{n_1}}+\frac{\overline{\boldsymbol{p_2}}(\boldsymbol{1}-\overline{\boldsymbol{p_2}})}{\boldsymbol{n_2}}}$$

$$(\overline{\boldsymbol{p_1}}-\overline{\boldsymbol{p_2}})-\boldsymbol{Z_{1-\frac{\alpha}{2}}}\sqrt{\frac{\overline{\boldsymbol{p_1}}(\boldsymbol{1}-\overline{\boldsymbol{p_1}})}{\boldsymbol{n_1}}+\frac{\overline{\boldsymbol{p_2}}(\boldsymbol{1}-\overline{\boldsymbol{p_2}})}{\boldsymbol{n_2}}} \leq (\boldsymbol{p_1}-\boldsymbol{p_2}) \leq (\overline{\boldsymbol{p_1}}-\overline{\boldsymbol{p_2}})+\boldsymbol{Z_{1-\alpha/2}}\sqrt{\frac{\overline{\boldsymbol{p_1}}(\boldsymbol{1}-\overline{\boldsymbol{p_1}})}{\boldsymbol{n_1}}+\frac{\overline{\boldsymbol{p_2}}(\boldsymbol{1}-\overline{\boldsymbol{p_2}})}{\boldsymbol{n_2}}}$$

Por lo tanto el intervalo aproximado que contiene a la diferencia de proporciones con nivel de confianza de $1-\alpha$ es $\left((\overline{\boldsymbol{p_1}}-\overline{\boldsymbol{p_2}})\pm\boldsymbol{Z_{1-\frac{\alpha}{2}}}\sqrt{\frac{\overline{\boldsymbol{p_1}}(\boldsymbol{1}-\overline{\boldsymbol{p_1}})}{\boldsymbol{n_1}}+\frac{\overline{\boldsymbol{p_2}}(\boldsymbol{1}-\overline{\boldsymbol{p_2}})}{\boldsymbol{n_2}}}\right)$.

Simulación de intervalos para la diferencia de proporciones

Para comprobar la efectividad del intervalo y su comportamiento bajo diferentes tamaño de muestra, se va a ejemplificar el cálculo sobre diversas simulaciones variando el tamaño de los vectores que contengan cada muestra. Para simular una variable Bernoulli basta con definir la probabilidad de éxito p en una celda, generar dos columnas de tamaño n_1y n_2 con valores U uniformes en $(0,1)$ generados por la fórmula ALEATORIO(). Por último la formula condicional SI($U \leq p_i, 1,0$) permitirá arrojar la simulación según cada parámetro p_i.

Con esto se pasa a realizar los cálculos de los parámetros para generar los intervalos, como el cuantil $Z_{1-\frac{\alpha}{2}}$, las medias y varianza muestrales, los límites inferior y superior del intervalo de confianza, además de una prueba lógica que permita reconocer si el intervalo contiene al verdadero valor de la diferencia de las proporciones. Una vez realizado esto un proceso automatizado que permita copiar el intervalo generado un cierto número de veces, generará la simulación necesaria para comprobar el porcentaje de intervalos que contienen a la verdadera diferencia de proporciones.

La siguiente imagen muestra una propuesta para la estructura en una hoja de cálculo para simular este tipo de intervalos de confianza.

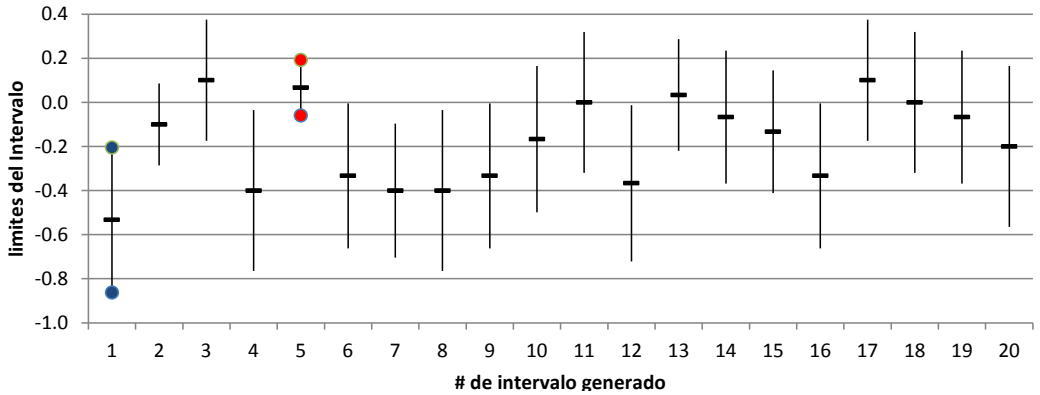
Estructura propuesta para el cálculo de intervalos de confianza sobre la diferencia de proporciones

	DR	DS	DT	DU	DV	DW	DX	DY	DZ	EA	EB	EC
1	intervalo de confianza para la diferencia de proporciones											
2												
3	parametro	X	Y					P1 Barra	0.90	P1(1-P1)	0.09	
4	P	70%	85%					P2 Barra	0.87	P2(1-P2)	0.12	
5	n	10	15					α	5%	$Z_{1-\alpha/2}$	1.96	
6												
7								Límite Inferior	P1 bar - P2 bar	Límite Superior	¿Contiene a p_1- p_2?	
8		Muestra Normal X		Muestra Normal Y				-0.22	0.03	0.29	SI	
9	Indice	Aleatorio(0,1	Simulado	Aleatorio(0,1	Simulado		Índice	Simulaciones de copias a valor por macro				% de SI
10	1	1.00	0.0	0.52	1.0		1	-0.86	-0.53	-0.20	NO	91.1%
11	2	0.06	1.0	0.28	1.0		2	-0.29	-0.10	0.09	SI	
12	3	0.20	1.0	0.65	1.0		3	-0.17	0.10	0.37	SI	
13	4	0.37	1.0	0.27	1.0		4	-0.76	-0.40	-0.04	SI	
14	5	0.06	1.0	0.65	1.0		5	-0.06	0.07	0.19	NO	
15	6	0.07	1.0	0.38	1.0		6	-0.66	-0.33	0.00	SI	
16	7	0.61	1.0	0.98	0.0		7	-0.70	-0.40	-0.10	SI	
17	8	0.19	1.0	0.78	1.0		8	-0.76	-0.40	-0.04	SI	
18	9	0.49	1.0	0.09	1.0		9	-0.66	-0.33	0.00	SI	
19	10	0.62	1.0	0.18	1.0		10	-0.50	-0.17	0.17	SI	
20	11			0.77	1.0		11	-0.32	0.00	0.32	SI	
21	12			0.30	1.0		12	-0.72	-0.37	-0.01	SI	
22	13			0.18	1.0		13	-0.22	0.03	0.29	SI	
23	14			0.99	0.0		14	-0.37	-0.07	0.24	SI	
24	15			0.22	1.0		15	-0.41	-0.13	0.14	SI	
25							16	-0.66	-0.33	0.00	SI	
26							17	-0.17	0.10	0.37	SI	
27							18	-0.32	0.00	0.32	SI	
28							19	-0.37	-0.07	0.24	SI	
29							20	-0.56	-0.20	0.16	SI	

En las tablas de la imagen se muestra un primer ejemplo de 1000 intervalos simulados, donde se tomaron tamaños de muestra pequeños $(n_1 = 10, n_2 = 15)$, como escenario inicial a la aproximación, para posteriormente variar los tamaños de muestra. Debido al tamaño de muestra, el porcentaje de los intervalos aproximados que cubren la diferencia es menor a la confianza esperada.

Para revisar el comportamiento de estos intervalos en la Gráfica 3.9 se encuentran los primeros 20 intervalos simulados los cuales muestran ahora que los dos resaltados en rojo y azul son aquellos que no contienen a la diferencia verdadera de las proporciones, afectados tanto por la posición como por la varianza de cada variable.

Gráfica 3.9: Primeros Intervalos de confianza para la diferencia de proporciones de 1000 muestras X(P_1) y Y (P_2)



Este ejercicio de 1000 simulaciones se replicó conservando las proporciones y nivel de confianza fijos, pero cambiando el tamaño de muestra para revisar los siguientes casos ($n_1 = 30 \quad n_2 = 50$), ($n_1 = 100 \quad n_2 = 15$) , $n_1 = 300 \quad n_2 = 500$) de los cuales un resumen de los resultados se puede hallar en la Tabla 3.6.

Tabla 3.6: Resumen de simulaciones de diferencias de proporciones p_1=70% p_2=85%

n_1	n_2	*Casos de éxito del intervalo*	*% de éxito de los intervalos*
10	15	911	91.10%
30	50	927	92.70%
100	15	897	89.70%
300	500	1,000	100.00%

De acuerdo a los casos observados el tamaño de cada muestra afecta la precisión de los intervalos en cubrir la verdadera diferencia de proporciones, pues en el tercer caso basta con que 1 sola muestra tenga un tamaño reducido, mientras que al ir aumentando el tamaño de ambas muestras aumenta la precisión de los intervalos. En el último caso al ser cercanas las probabilidades se observa que todos los intervalos contuvieron a la diferencia verdadera.

Capítulo IV:

Pruebas de hipótesis

El objetivo esencial de las ciencias es generar un conocimiento veraz sobre un fenómeno observado; para esto cada ciencia y sus ramas tienen métodos propios para observar el fenómeno de tal suerte que a partir de un proceso teórico-práctico se pueda llegar a una ley que gobierne los hechos observados, lo que implicaría que tales hechos son tan solo instancias de tal ley.

A principios del siglo pasado autores como Bertrand Russell (1931) exponía como otro de los fines de la ciencia, educar al sentido común, es decir, lograr que un estudiante tenga un pensamiento científico, basado en hechos, abstracciones y argumentos en lugar de una colección de técnicas finitas que sólo aplique de forma mecánica a todos los fenómenos que se le pongan en frente.

Contario a esto, hoy en día la educación básica y media superior actual sobrecarga al estudiante de los mayores conocimientos posibles sobre diversas ramas de una ciencia, provocando que sólo se revise de manera superficial los temas de un plan de estudios.

La estadística ha tenido desde su nacimiento un papel importante dentro del desarrollo de otras ciencias; tal importancia radica en que la estadística provee de técnicas que permiten analizar fenómenos de naturaleza aleatoria y realizar conclusiones con cierta confianza sobre estos, contribuyendo de esta manera a la generación de conocimientos. Aportar un conocimiento a una ciencia en específico, conlleva un proceso que ha ido evolucionando a través de los siglos llamado método científico.

El método científico consiste en un esquema en el cual debe basarse un investigador, para demostrar que sus afirmaciones son consistentes con lo que se ha observado sobre un fenómeno. El esquema fue concebido desde un inicio como una serie de pasos que iniciaran con una observación detallada de algún fenómeno y que finalizaría con una aportación de conocimiento, mismo que ha evolucionado a través de los siglos, sin embargo, se puede resumir en la siguiente lista:

1.- Observación y registro del fenómeno o de alguna de sus características de interés.

2.-Con base en los casos particulares intentar establecer premisas o enunciados sobre ellos (inducción).

3.-Con base en las premisas estructurar los supuestos formales acerca de las características observadas en el fenómeno, conocidas como hipótesis.

4.- Probar o contrastar la o las hipótesis por medio de experimentación.

5.-Demostrar la validez de las hipótesis o, según sea el caso, refutarlas.

6.-Establecer las conclusiones pertinentes de acuerdo a lo obtenido en el proceso.

Como se puede ver en este proceso se emplean procesos inductivos y deductivos de manera implícita, pues con base en observaciones particulares se infieren hipótesis, las cuales pueden ser verdaderas, sin embargo es de esperarse que conduzcan de una forma lógica a una explicación sobre el comportamiento del fenómeno.

Las hipótesis además también deben ser pensadas de acuerdo al objetivo de la investigación, es decir, si se desea confirmar, refutar, validar o, debido a la naturaleza aleatoria del fenómeno, sólo concluir con cierta certidumbre que la hipótesis es consistente con lo observado.

Hipótesis estadísticas

Las hipótesis que son usadas en temas relacionados con fenómenos de comportamiento aleatorio son denominadas hipótesis estadísticas, las cuales realizan suposiciones sobre alguna característica de la distribución del conjunto de variables aleatorias asociadas conceptualmente a las características del fenómeno de estudio, ya sea en cuanto a sus parámetros o hacia la forma de la distribución.

Una forma de enunciar las hipótesis estadísticas es por medio de especificar que la distribución sigue una función $F(X;\theta)$ o especificando la densidad $f(X;\theta)$, donde θ es un vector de parámetros. Cuando se establece un supuesto sobre algún parámetro de la distribución de la población, se divide en varios casos ya que las hipótesis pueden estar dirigidas a que el parámetro de interés tenga un valor en específico o se encuentre en un rango de valores. De acuerdo a esto las hipótesis estadísticas se les clasifican de la siguiente manera.

Hipótesis simples

Las hipótesis simples son aquellas que especifican un valor puntual para un parámetro, esto implica que bajo esta clase de hipótesis se logre definir completamente la distribución de la población, por ejemplo para una población de distribución Normal(μ, 10) el supuesto $\mu = 5$ es una hipótesis simple. Dentro del aprendizaje práctico de técnicas estadísticas, el uso de este tipo de hipótesis es muy común, pues después de aseverar una hipótesis simple su contraste se puede realizar intuitivamente viendo si el estimador del parámetro, al ser calculado con los datos observados, toma un valor cercano al de la hipótesis, entonces se tiene una cierta inclinación a pensar que tal hipótesis es cierta, mientras que si toma un valor alejado, es más factible que la hipótesis sea falsa.

Cuando se supone una forma para la distribución, y además los parámetros se proponen por medio de hipótesis simples, al confrontarlas se realiza un contraste entre creer que cada variable X_i de la muestra aleatoria se distribuye $f(X;\theta_0)$ que se puede enunciar como la hipótesis "$\theta = \theta_0$ o X se distribuye $f(x;\theta_0)$" que implica que la hipótesis contraria sea $\theta = \theta_1$ o X se distribuye $f(x;\theta_1)$.

Hipótesis compuestas

Cuando una hipótesis se traduce en que el parámetro de interés θ se localice dentro de un intervalo se le llama hipótesis compuesta, es decir, se supone que $\theta \subset \Theta$, donde Θ es un subconjunto del espacio paramétrico. El nombre de hipótesis compuesta se

debe a que el conjunto Θ esta compuesto un conjunto valores distintos de tal forma que la hipótesis no especifica completamente una distribución. Por ejemplo, para una población de distribución Gamma(α, β) la hipótesis $\beta > 100$ es una hipótesis compuesta de la forma $\Theta = \{\beta | \beta > 100\}$, de la cual se puede notar que cada uno de los valores del conjunto se les puede asociar a hipótesis simples.

Puesto que una hipótesis compuesta no define completamente una función de distribución para la población, se necesitan métodos más generales para hallar una regla que permita decidir de forma óptima, cual es la hipótesis a la cual se orienta la conclusión. Existen varios métodos para encontrar una regla que permita decidir de una manera óptima cuál hipótesis es la más verosímil, de los que se pueden destacar, la razón de verosimilitudes y el lema de Neyman-Pearson, aunque su tratamiento están en un campo más teórico al que se desea presentar, además debido a que el objetivo es presentar el tema de forma intuitiva, se revisarán contrastes de hipótesis de las cuales se pueda derivar sus estadísticos de prueba a partir de los supuestos de las poblaciones.

Proceso para realizar pruebas de hipótesis

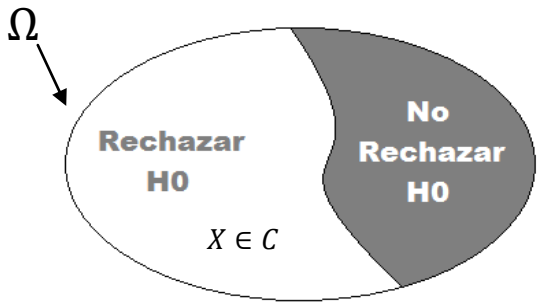
El procedimiento por el cual se lleva a cabo una comparación entre las hipótesis planteadas sobre un fenómeno se llama prueba de hipótesis. El proceso de realizar una prueba de hipótesis se puede resumir como la suposición de 2 hipótesis estadísticas, para que a partir de una regla objetiva, se pueda concluir si la hipótesis propuesta muestra evidencia de ser consistente con el comportamiento estadístico de las observaciones.

Según lo anterior lo primero que se debe hacer es proponer una hipótesis la cual se contrasta estableciéndola como una hipótesis nula denotada como H_0. Dentro de la literatura estadística, se relaciona el concepto de hipótesis nula, a aquella contraria a las creencias del investigador, mientras que sus creencias son comúnmente puestas en una hipótesis alternativa denotada H_1.

El siguiente paso es obtener una muestra aleatoria para que partir de ésta, se proceda a calcular un estadístico que determine la prueba. El valor del estadístico determinará cuál es la hipótesis más consistente con lo observado, es decir, conducirá a la decisión sobre si se debe rechazar o no una hipótesis nula. Normalmente sobre quien se trabaja esta perspectiva es la hipótesis nula.

A la región del espacio en la cual se hallan todas las muestras posibles que hagan rechazar la hipótesis nula, se llama región crítica y se denota como C. Para presentarlo de otra forma, dada una muestra aleatoria $X = X_1, \dots, X_n$ la regla para decidir si se debe rechazar o no la hipótesis nula se ilustra como:

$$\boldsymbol{Rechazar\ H_0\ si\ (X = X_1, \dots, X_n) \in C}$$



Como el valor de un estadístico depende de los valores observados de la muestra, entonces la regla de decisión sobre si rechazar H_0, puede ser traducida en términos del estadístico de prueba, que de hecho es la forma más común de encontrar tales reglas dentro de la literatura estadística.

Errores dentro de la conclusión sobre una hipótesis

Al realizar una prueba de hipótesis el valor del estadístico de prueba lleva a una conclusión, sin embargo, debido a la naturaleza aleatoria de los fenómenos de estudio es posible tener como resultado un valor del estadístico que apunte hacia la hipótesis alternativa pero los parámetros de la distribución correspondan en realidad a la hipótesis nula, es decir que se haya rechazado la hipótesis nula H_0, pero que esta hipótesis sea cierta, lo cual sería un error; de manera análoga otro error en el que se puede incurrir es el caso en que no se rechace la hipótesis nula, cuando en realidad esta hipótesis sea falsa.

El primer error en el que se puede incurrir rechazando la hipótesis nula H_0 cuando en realidad es cierta se llama error de tipo I; mientras que el error derivado de no se rechaza H_0, cuando H_0 es una hipótesis falsa, se conoce como el error de tipo II. Estos errores sirven como marco para saber qué es lo que se desea de una prueba de hipótesis, que sería minimizar la probabilidad de cometer esta clase de errores. La siguiente tabla resume los tipos de error que se pueden cometer, junto con las decisiones correctas que se pueden concluir al realizar una prueba de hipótesis.

	condición de la naturaleza	
Decisión	H_0 cierta	H_0 falsa
Rechazar H_0	Error Tipo I	decisión correcta
No Rechazar H_0	decisión correcta	Error Tipo II

Función potencia

Dada una regla o prueba para saber si una muestra se halla en la región de rechazo C, la función potencia se define como la probabilidad de rechazar la hipótesis H_0 cuando la muestra proviene una población definida por el parámetro θ, denotada como $\pi(\theta)$. Dicho de otra forma $\pi(\theta)$, es la probabilidad que la muestra obtenida se halle en la

región de rechazo C, bajo el supuesto de que θ sea el verdadero valor del parámetro, la cual se puede interpretar como la sensibilidad de la prueba, y se describe como

$$\boldsymbol{\pi(\theta) = P_{\theta}\left((X_1, \dots, X_n) \in C\right)}$$

Como existen posibilidades de cometer un error de la tabla anterior al ejecutar una prueba de hipótesis, para una muestra particular, la función potencia ayudará a medir la probabilidad de ocurrencia de los errores.

Denótese Θ_0 al conjunto de valores para el parámetro θ que cumplen con la definición de la hipótesis nula H_0. Si se valúa la función potencia sobre el conjunto de valores Θ_0, la función potencia calcula $\boldsymbol{P_{\theta}\left((X_1, \dots, X_n) \in C | \theta \in \Theta_0\right)} = P(Rechazar\ H_0 | H_0\ es\ cierta)$, con lo cual se mide la probabilidad del error de Tipo I.

Si se define ahora Θ_1 como el conjunto de valores para el parámetro que cumplen la definición de la hipótesis Alternativa H_1, al hallar la función potencia para los valores $\theta \in \Theta_1$ entonces se mide $P(Rechazar\ H_0 | H_1\ es\ cierta)$ que se interpreta como la probabilidad de ejecutar una decisión correcta. Esta probabilidad es conocida como la potencia de la prueba.

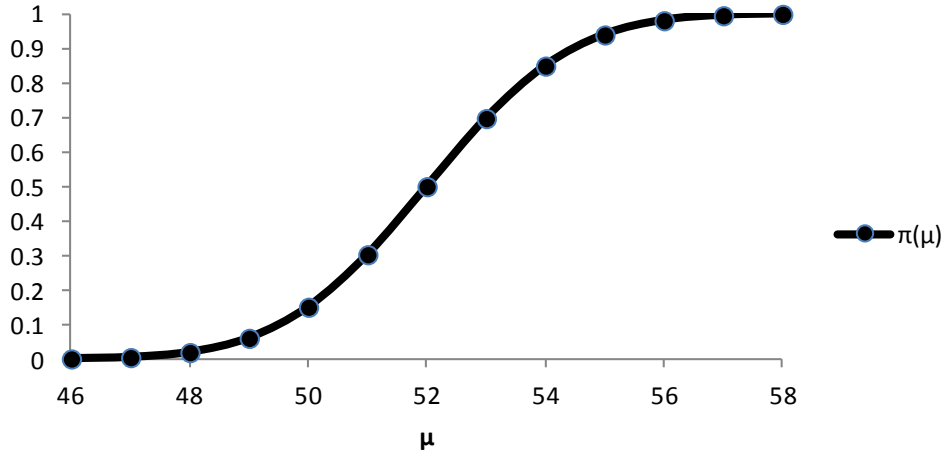
Para conseguir una visión más profunda sobre la función potencia se propone el siguiente ejemplo: supóngase que $X = x_1, x_2, \dots, x_n$ es una muestra aleatoria proveniente de una distribución Normal$(\mu, 75)$ de la cual se considera contrastar dos hipótesis,$H_0: \mu < 50$ contra $H_1: \mu \geq 50$. Para llevar a cabo la prueba se aplicará la regla de decisión $Rechazar\ H_0\ si\ \bar{X} \geq 50$ es decir que $C = \{X_1, X_2, \dots, X_n | \bar{X} \geq 50\}$. Considérese además de inicio un valor de $n = 20$ donde se aplica además la estandarización de la normal, con lo cual la función potencia toma la expresión

$$\boldsymbol{\pi(\mu) = P(\bar{X} \geq 50) = P\left(Z \geq \frac{50 - \mu}{\sqrt{75/20}}\right)}$$

Ahora, tabulando $\pi(\mu)$ para diversos valores μ se obtiene lo siguiente

μ	$\frac{50-\mu}{\sqrt{75/n}}$	$\pi(\mu)$	μ	$\frac{50-\mu}{\sqrt{75/n}}$	$\pi(\mu)$
46	3.1	0.001	53	-0.52	0.6972
47	2.58	0.0049	54	-1.03	0.8492
48	2.07	0.0194	55	-1.55	0.9393
49	1.55	0.0607	56	-2.07	0.9806
50	1.03	0.1508	57	-2.58	0.9951
51	0.52	0.3028	58	-3.1	0.999
52	0	0.5			

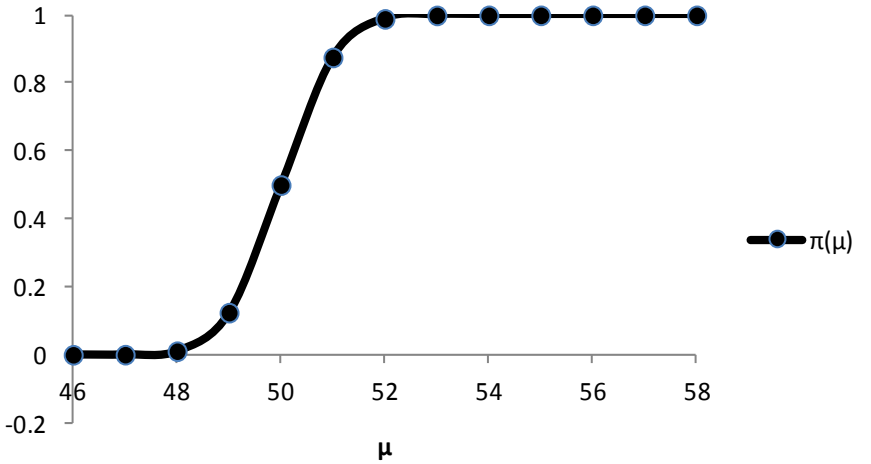
Grafica 4.1: Función potencia, n=20



De lo anterior se puede observar en la gráfica, que para valores menores a 50 la probabilidad de rechazar H_0 baja de manera gradual, mientras que para valores mayores a 50 la probabilidad de rechazo crece rápidamente. Ahora el valor de n se variará, para ver cómo se comporta la función potencia cuando el tamaño de la muestra aumenta, entonces para $n = 100$ bajo el mismo esquema, la tabla de valores y su gráfica que se obtienen son los siguientes

μ	$\frac{50-\mu}{\sqrt{75/n}}$	$\pi(\mu)$	μ	$\frac{50-\mu}{\sqrt{75/n}}$	$\pi(\mu)$
46	4.62	0.00	53	-3.46	0.9997
47	3.46	0.00	54	-4.62	1.0000
48	2.31	0.01	55	-5.77	1.0000
49	1.15	0.12	56	-6.93	1.0000
50	0.00	0.50	57	-8.08	1.0000
51	-1.15	0.88	58	-9.24	1.0000
52	-2.31	0.99			

Gráfica 4.2: Función potencia, n=100



Esta gráfica expone que la función potencia con un mayor tamaño de muestra consigue diferenciar mejor entre sí rechazar o no a la hipótesis nula, pues partiendo del valor 50 el valor menor inmediato de la tabla (49) oscila cerca del 10% para no rechazar a H_0, mientras que el inmediato mayor (51) está cerca del 90% para rechazar a H_0.

Otro papel de gran importancia de la función potencia es optimizar las decisiones tomadas por las pruebas de hipótesis; la forma en que lo consigue es por medio de establecer una máxima probabilidad α de cometer el error de Tipo I, conocida mejor como el tamaño de la prueba, o como el nivel de significancia de la prueba, el cual se expresa como

$$\boldsymbol{\alpha = \sup_{\theta \in \Theta_0} \pi(\theta)}$$

Una vez fijado el nivel de significancia α se opta por seguir la regla que busque maximizar la potencia de la prueba, lo cual es equivalente a una vez fija α minimizar la probabilidad de error Tipo II. Dentro de la teoría estadística existen formas óptimas para hallar estas reglas, pero para propósitos de este trabajo el enfoque estará en considerar una perspectiva más intuitiva sobre los conceptos explicados a través de ejemplos controlados vía simulación.

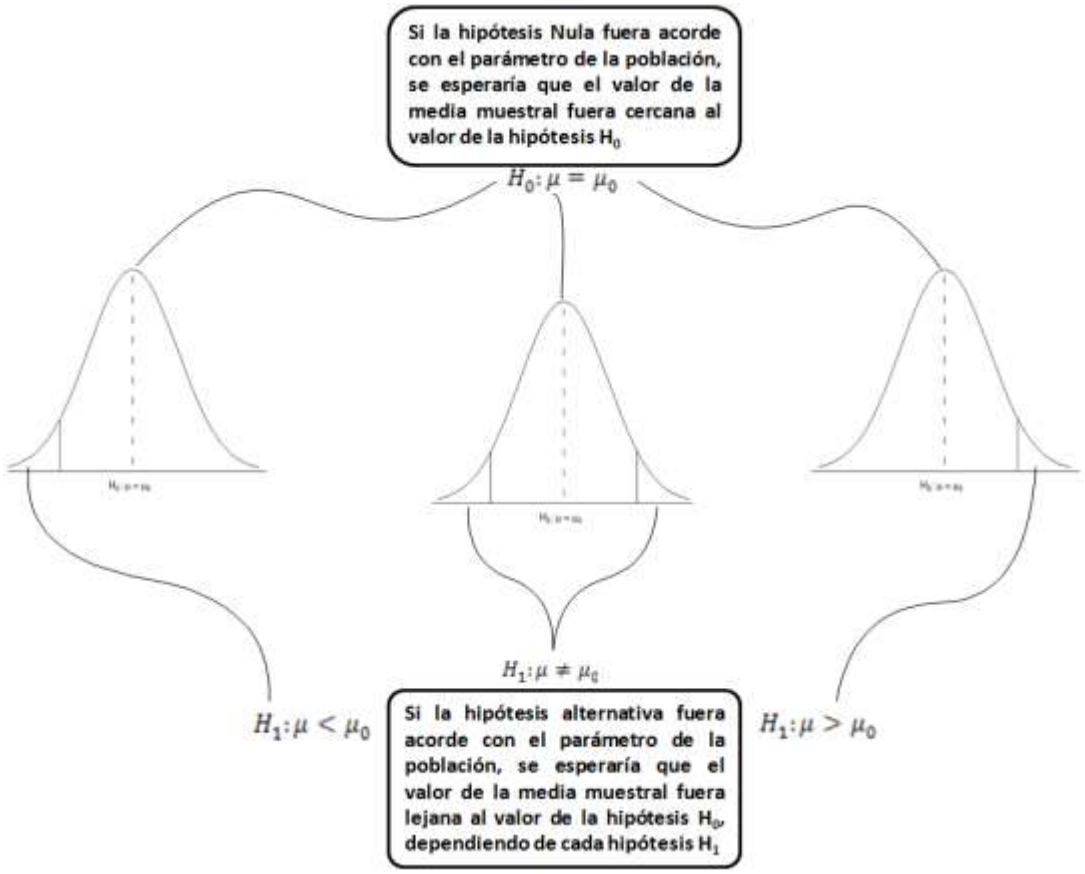
Prueba sobre la media de una población bajo el supuesto de normalidad

Para la siguiente serie de pruebas el supuesto que se manejará es que la distribución de la población es Normal y lo que se contrastará serán hipótesis sobre los parámetros de la distribución, primero desde una forma ortodoxa y luego con técnicas de la simulación Monte Carlo, con el fin de comprender mejor esta clase de pruebas de hipótesis. El objetivo es mostrar cómo los ejemplos simulados, pueden generar una pauta sobre cómo tratar el amplio rango de pruebas de hipótesis que existen en la teoría de la estadística.

Supóngase que se tiene una muestra aleatoria $X = x_1, x_2, \dots, x_n$ proveniente de una población Normal(μ, σ^2), el parámetro sobre el cual radicarán las hipótesis será **la media $\boldsymbol{\mu}$**, considérese **el caso en el que se conozca la varianza $\boldsymbol{\sigma^2}$** de la distribución. Primero se establece la hipótesis nula H_0 la cual afirma que el valor de la media poblacional es igual a un valor μ_0, es decir:

$$\boldsymbol{H_0: \mu = \mu_0}$$

A partir de esta hipótesis se pueden establecer 3 distintas hipótesis alternativas, la primera es que la media poblacional sea menor al valor de la hipótesis nula, $H_1: \mu < \mu_0$, otra hipótesis alternativa es aquella que afirma que el valor de la media poblacional es mayor que el establecido por H_0, $H_1: \mu > \mu_0$ y por último tomar como hipótesis alternativa que el valor de la media poblacional es otro distinto al establecido por H_0, $H_1: \mu \neq \mu_0$. Lo que se esperaría al contrastar la hipótesis nula contra las distintas hipótesis alternativas es que la región de rechazo cambie en cada caso, de acuerdo al siguiente diagrama:



Lo que muestra el diagrama anterior de forma más intuitiva es la forma en que se va a rechazar o no rechazar una hipótesis nula. De forma más precisa lo que se realizará es definir un estadístico que bajo el supuesto en que se cumpla H_0 se calcule sobre una muestra, para que de acuerdo a su valor y dependiendo de la hipótesis alternativa establecida, se decida si se cae en alguno de los casos anteriores.

Empleando el supuesto de Normalidad de la población y que se conoce el valor de $\boldsymbol{\sigma^2}$, se pueden emplear la media muestral $\bar{X}$, de la cual se conoce su distribución para construir una variable estandarizada Z. Si se supone inicialmente que se cumple la hipótesis H_0, es decir $\boldsymbol{\mu_0}$ se considera como la media verdadera, entonces la distribución de Z queda completamente especificada como $N(0,1)$. De acuerdo a lo anterior, el estadístico que se empleará para contrastar las hipótesis de esta prueba es el siguiente

$$\boldsymbol{Z = \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}}}$$

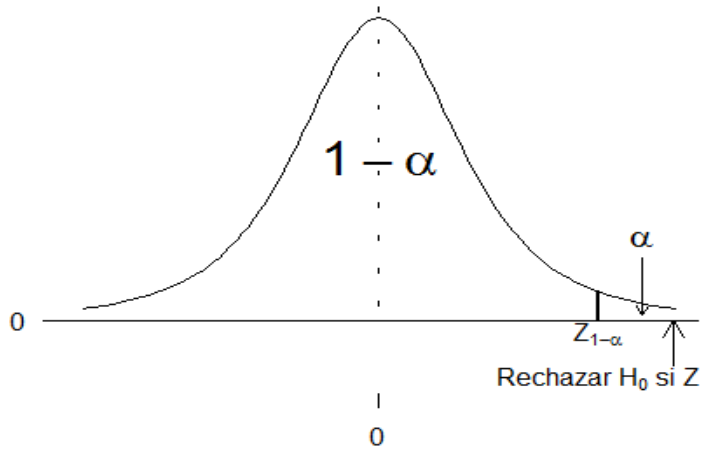
Para trabajar un ejemplo **considérese el caso en que la hipótesis alternativa es** $\boldsymbol{H_1: \mu > \mu_0}$**,** que es el caso de la derecha del diagrama anterior. Entonces la región de rechazo se define como $C = \{(x_1, x_2, \dots, x_n) | \, Z \geq k\}$, donde k es el límite de la región de rechazo. Lo siguiente en hacer es establecer la probabilidad máxima de cometer el error Tipo I igual a α y partir de ahí se debe encontrar el valor de la constante k que establezca la regla que establezca cuándo se debe rechazar H_0, es decir, se halla k tal que

$$\boldsymbol{P(X \in C) = P(Z \geq k) = \alpha}$$

$$\boldsymbol{P\left(\frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} \geq k\right) = \alpha}$$

El valor de la constante k es el cuantil $Z_{1-\alpha}$ de la distribución Normal(0,1). Esto implica que si la diferencia estandarizada entre $\bar{X}$ y el parámetro $\boldsymbol{\mu_0}$ es más grande que $Z_{1-\alpha}$ se debe rechazar H_0 con un nivel α de significancia. La gráfica siguiente muestra la región de rechazo de esta prueba bajo el supuesto de normalidad de la población y el área probabilidad que acumula.

Gráfica 4.3: Región de rechazo para H_0



Ahora se verá un ejemplo de cálculo considerando esta hipótesis alternativa $\boldsymbol{H_1: \mu > \mu_0}$ para ello se ha simulado una muestra de tamaño $n = 15$ de una población Normal($\mu, \sigma^2 = 25$), la cual se simuló deliberadamente con $\mu = 10$, y para la prueba se selecciona el mismo valor $\mu_0 = 10$ para revisar un caso en que se H_0 es cierta, por lo que las hipótesis a contrastar son

$$\boldsymbol{H_0: \mu = 10}$$

$$\boldsymbol{H_1: \mu > 10}$$

La prueba se realizará con un nivel de significancia del 5% ($\boldsymbol{\alpha = 0.05}$). La tabla siguiente despliega la muestra simulada por los métodos ya vistos para la distribución Normal considerando $\boldsymbol{\mu = 10, \sigma^2 = 25}$, además para la realización de la prueba, se incluyen los valores más importantes como $\boldsymbol{\bar{X}, \sigma^2, Z_{1-\alpha}, Z}$, y el valor de α.

Tabla 4.1: Muestra de una población Normal(μ,σ^2=25)

i	x_i	i	x_i	i	x_i
1	15.39	6	11.35	11	6.95
2	1.32	7	15.52	12	8.71
3	19.99	8	12.85	13	9.99
4	18.03	9	11.33	14	8.21
5	11.72	10	12.94	15	2.74

$\boldsymbol{\sigma^2 = 25}$ $\boldsymbol{Z_{1-\alpha} = 1.645}$ $\boldsymbol{\alpha = 0.05}$

$\boldsymbol{\bar{X} = 11.14}$ $\boldsymbol{Z = 0.440}$

Según las reglas que se habían establecido, se debe concluir el no rechazar H_0 ya que $\boldsymbol{Z < Z_{1-\alpha}}$, es decir, el valor de la media de los datos observados es significativamente cercano al valor de la hipótesis nula, por lo que no se debe rechazar la hipótesis que enuncia que son iguales. Ahora se trabajará esta misma prueba con un ejemplo detallado visto desde la perspectiva de la simulación Monte Carlo.

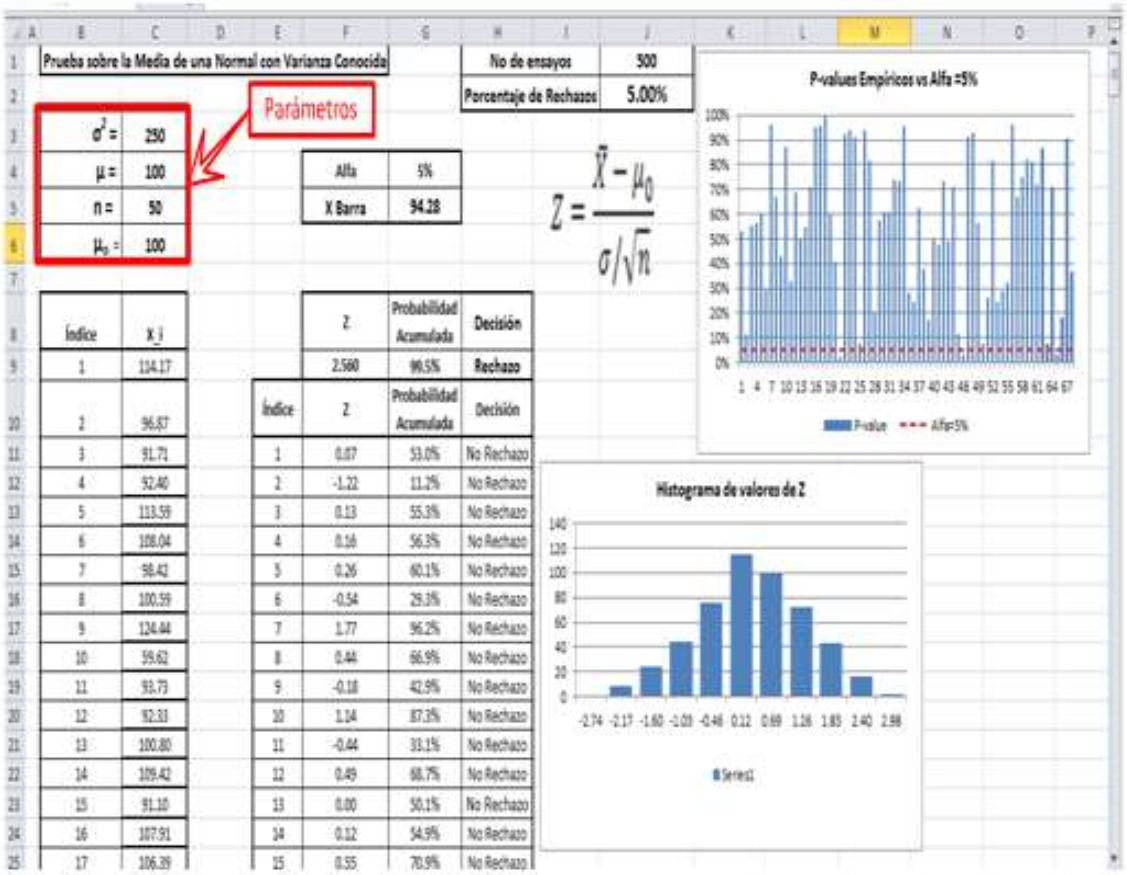
Para el siguiente ejemplo, se trabajará ahora una combinación distinta de hipótesis. Supóngase que se desea probar que la media μ de una población cuya distribución es Normal(μ, σ^2), donde σ^2 es conocida. Las hipótesis ahora son si μ es un valor igual a μ_0 o es un valor distinto a μ_0, es decir el contraste de hipótesis es el siguiente.

$$\boldsymbol{H_0: \mu = \mu_0}$$

$$\boldsymbol{H_1: \mu \neq \mu_0}$$

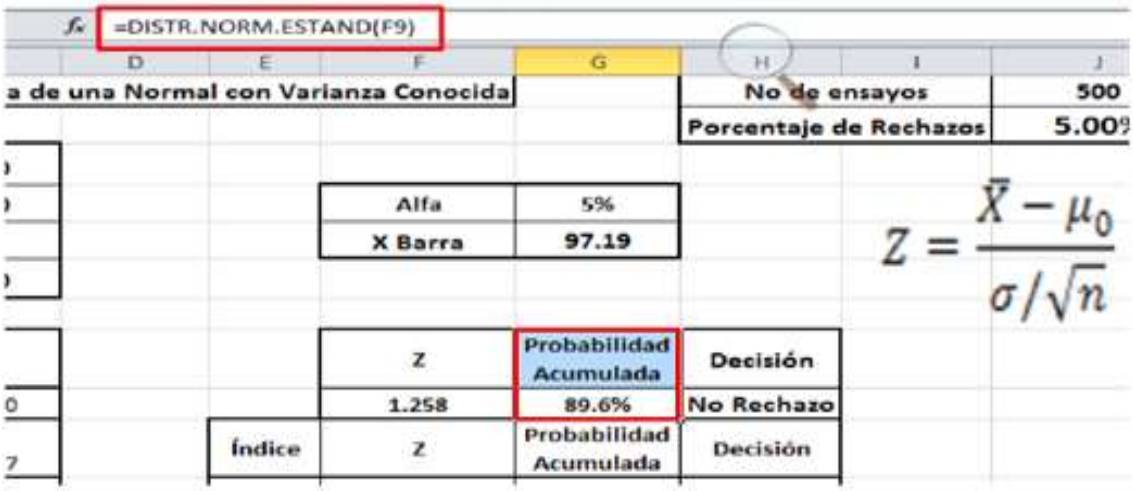
Este planteamiento conduce a realizar una prueba sobre las dos colas de la distribución de Z ya que la hipótesis alternativa se puede escribir de la forma $\boldsymbol{H_1: \mu > \mu_0 \ o \ \mu < \mu_0}$. La regla de decisión con esta hipótesis se dará explícitamente más adelante.
La siguiente estructura para realizar la simulación permitirá cambiar ciertos parámetros establecidos de inicio, para lograr una mejor visión de la prueba. Los parámetros iniciales son: el nivel de significancia $\alpha = 5\%$, $\mu = 100$, $\mu_0 = 100$ y $\sigma^2 = 250$.

En la siguiente imagen se presenta una propuesta en hoja de cálculo para realizar las simulaciones, donde se señala la tabla de parámetros con los valores de μ, σ^2 que se supone conocida, n y el valor de acuerdo a la hipótesis nula μ_0.



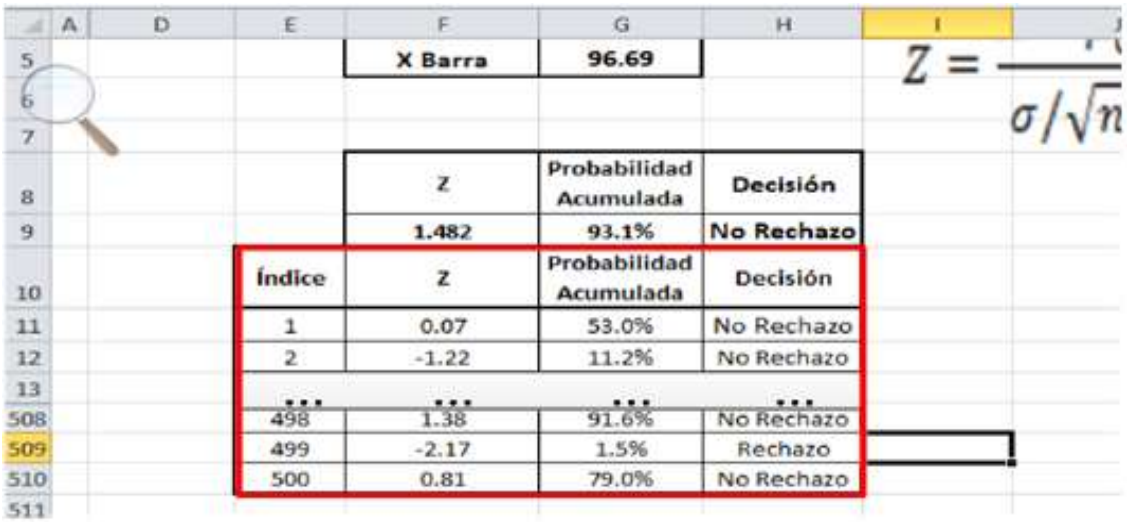
Primero para simular la muestra se puede utilizar la fórmula de la inversa de la distribución Normal introduciendo como argumentos los valores de la tabla de los parámetros a modificar, es decir se deben generar n celdas con la fórmula $= INV.DISTR.NORM(Aleatorio(), \mu, RAIZ(\sigma^2))$. A partir de esta muestra simulada se puede obtener los elementos clave para realizar la prueba de hipótesis como la media muestral $\bar{X}$, el estadístico de prueba Z , la probabilidad que acumula el valor del estadístico de prueba en la distribución Normal(0,1) y por último la decisión sobre si rechazar la hipótesis nula o no rechazarla.

Para determinar la Probabilidad (que se denotará como PA) que el estadístico Z acumula de la distribución Normal(0,1), se usará la fórmula $DISTR.NORM.EST(Z)$, la cual calcula $P(z \leq Z)$ para una distribución Normal estándar. Luego se formula la regla de decisión de una forma equivalente a la anterior mediante Pa, la cual es: *si* PA *es menor que* $\alpha/2$ *o* PA *es mayor a* $1-\alpha/2$ entonces se rechaza la hipótesis nula, en caso contrario no se rechaza.



La forma de automatizar la regla de decisión es por medio de la formula condicional $= SI(PA \geq \alpha/2, SI(PA \leq 1-\alpha/2, "No\ Rechazo", "Rechazo"), "Rechazo")$. A partir de este dictamen automático se tiene que repetir la generación de muestras y

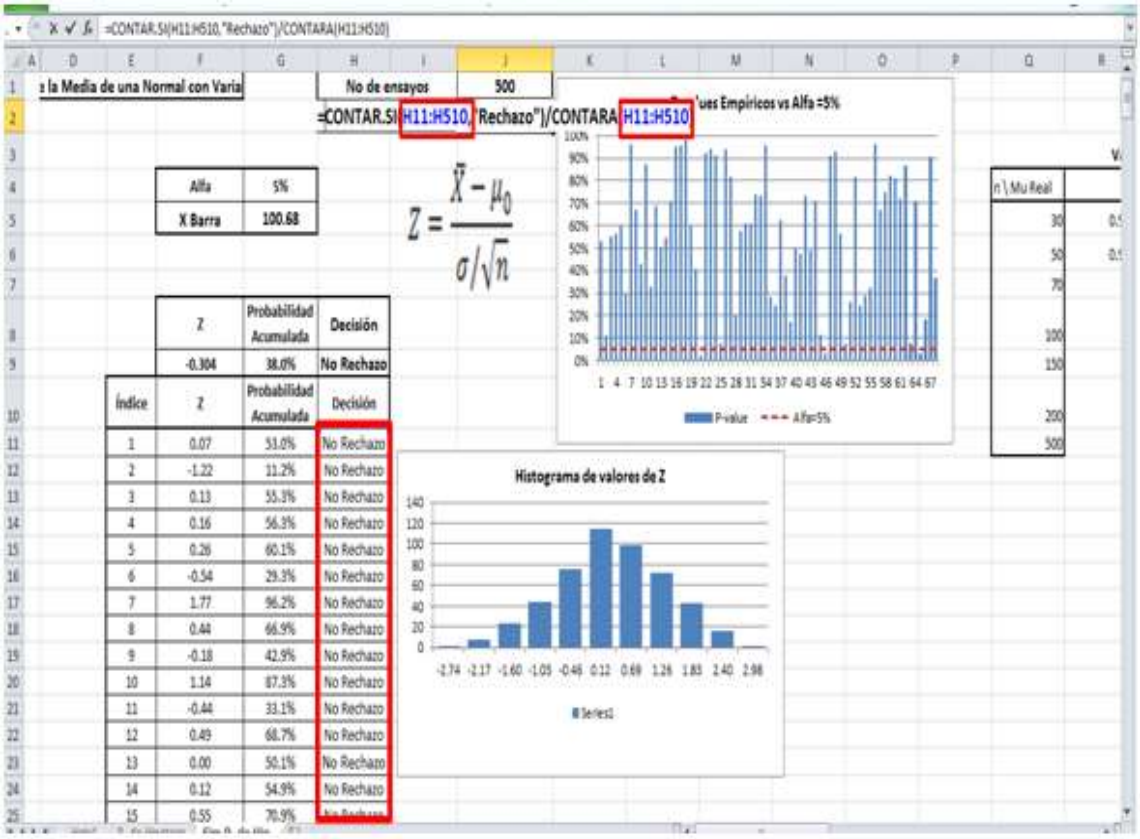
consecuentemente generar distintos valores de Z con lo cual se obtendrá una serie de decisiones, que para este caso práctico serán 500 repeticiones.



X Barra	96.69		
	Z	Probabilidad Acumulada	Decisión
	1.482	93.1%	No Rechazo

Índice	Z	Probabilidad Acumulada	Decisión
1	0.07	53.0%	No Rechazo
2	-1.22	11.2%	No Rechazo
...	...	...	...
498	1.38	91.6%	No Rechazo
499	-2.17	1.5%	Rechazo
500	0.81	79.0%	No Rechazo

Del total de las simulaciones generadas se espera que algunos de resultados aleatorios rechacen la hipótesis nula aunque por la selección de parámetros es verdadera. De acuerdo con la teoría antes explicada el porcentaje esperado debe ser cercano al nivel de significancia, pues es la probabilidad máxima de errores de Tipo I (Rechazar H_0 cuando H_0 es cierta) la cual se estableció del 5% .



En la imagen se puede observar un histograma de Z, el cual muestra la distribución de la población, que por construcción de la simulación es Normal. En otro orden de ideas, la gráfica de la parte superior compara la probabilidad acumulada de cada simulación

contra una línea punteada que indica el 5%, de esta manera no sólo se ve si se rechaza o no la hipótesis nula, sino que se está rechazando por probabilidades acumuladas por el estadístico Z lejanas a α, sin embargo se puede notar a simple vista que algunas barras llegan a estar por debajo de α, que son las ocasiones en los que se cometió el error de Tipo I.

Lo siguiente que se puede hacer, con la estructura de la hoja de cálculo y usando la misma prueba, es cambiar los valores de los parámetros, para observar cuál es el comportamiento de las decisiones tomadas bajo distintos escenarios. Primero se mantendrá fijo el valor de μ_0 mientras se cambia el valor de la media μ de la población simulada. En los casos que se elija una $\mu \neq \mu_0$, no rechazar la hipótesis nula sería ahora cometer un error de Tipo II.

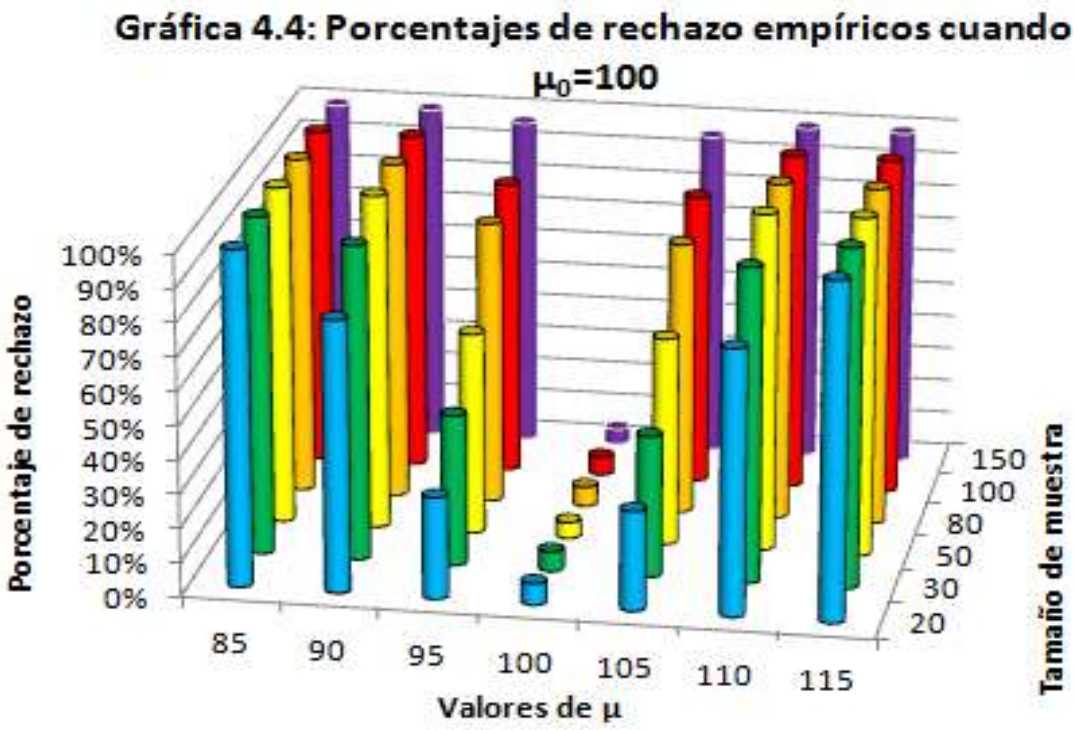
Teóricamente si se mide de igual manera la proporción de rechazos de H_0 se debería observar una disminución importante, sin embargo la sensibilidad con la que los rechazos aumenten conforme μ_0 se aleje de μ será la verdadera la potencia de la prueba. La siguiente tabla resume el resultado de realizar la prueba en 500 muestras simuladas de tamaño $n = 20$ bajo varios escenarios, modificando únicamente el valor de μ, en 5 unidades tanto en aumento como en disminución.

Tabla 4.2: Valores del % empírico de rechazo, cuando $\mu_0 = 100$; #Simulaciones = 500							
n \ μ	85	90	95	100	105	110	115
20	99.00%	80.20%	30.00%	6.40%	29.40%	78.40%	99.40%

Después de modificar el parámetro μ, se puede alterar ahora otra característica de la muestra como su tamaño n. En teoría al aumentar el tamaño de la muestra debería converger más rápido el porcentaje de rechazos de hipótesis falsas a 100%. La siguiente tabla muestra los resultados de haber realizado las simulaciones con un tamaño de muestra de 30, 50, 80, 100 y 150.

Tabla 4.3: Valores del Alfa empírico cuando $\mu_0 = 100$; # Simulaciones = 500							
n \ μ	85	90	95	100	105	110	115
20	99.00%	80.20%	30.00%	6.40%	29.40%	78.40%	99.40%
30	99.80%	93.00%	44.60%	5.80%	41.80%	92.80%	100%
50	100%	99.00%	59.60%	4.80%	61.40%	99.60%	100%
80	100%	100%	83.80%	5.80%	80.60%	100%	100%
100	100%	100%	87.00%	5.40%	86.20%	100%	100%
150	100%	100%	97.40%	4.60%	96.00%	100%	100%

Los resultados de la tabla demuestran que al aumentar el tamaño de la muestra, la sensibilidad de la prueba hacia rechazar hipótesis falsas, es decir su potencia, es cada vez mayor, lo que indica una buena potencia en la prueba. En la Gráfica 4.4 se pueden observar los valores de tabla anterior, en la cual se tiene una perspectiva general sobre la sensibilidad de la prueba en los distintos escenarios simulados.



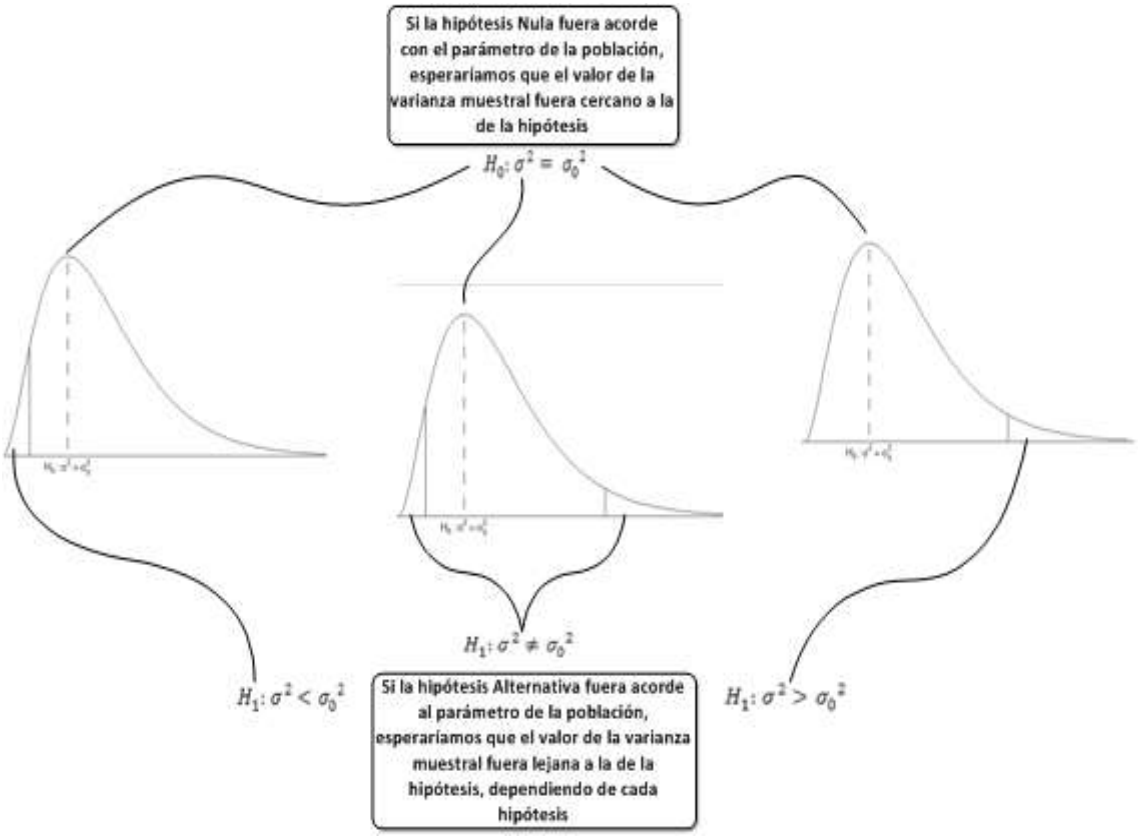
En conclusión para esta prueba se ha visto como por medio de la simulación se pueden explorar los conceptos prácticos y teóricos de la prueba de hipótesis sobre la media, pero ahora con una perspectiva distinta y más intuitiva.

Prueba sobre la varianza de una población bajo el supuesto de normalidad.

Supóngase que se tiene una muestra aleatoria $X_1, X_2, \dots, X_n$ de una población, donde cada variable X_i se distribuye Normal(μ, σ^2), de la que **se conoce la media poblacional $\boldsymbol{\mu}$** y en cambio se desconoce su varianza σ^2, por lo que se desea realizar una prueba de hipótesis, a partir de la siguiente hipótesis nula

$$\boldsymbol{H_0: \sigma^2 = \sigma_0^2}$$

Tomando esta hipótesis nula se tienen tres distintas hipótesis que pueden plantearse. Estas hipótesis alternativas y las regiones de rechazo para cada una, se resumen e Ilustran en el siguiente diagrama.



Como se puede ver en el diagrama cada hipótesis alternativa tiene un área de rechazo distinta donde se puede observar de manera intuitiva, que cada región de rechazo guarda una relación lógica con respecto a la descripción de la hipótesis alternativa. Otra característica que se puede notar es que cada zona tiene un cierto límite de acuerdo a cada hipótesis alternativa, el cual se determinará más delante.

Debido al supuesto inicial de conocer el valor de la media el estimador de la varianza S^2 se calcula de la siguiente forma

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n}$$

El cual es insesgado para el parámetro σ^2, es decir, $E(S^2) = \sigma^2$. Con base en lo anterior ahora el estadístico de prueba es

$$\boldsymbol{w = \frac{(n)S^2}{{\sigma_0}^2}}$$

Debido al supuesto de normalidad de la muestra y por conocer μ, la distribución del estadístico w es Ji-Cuadrada pero ahora con $\boldsymbol{n}$ **grados de libertad.** Como se mostró antes la región de rechazo depende de la definición de la hipótesis alternativa; para trabajar con hipótesis en particular, supóngase que se desea realizar el contraste de las siguientes hipótesis.

$$\boldsymbol{H_0: \sigma^2 = {\sigma_0}^2}$$

$$\boldsymbol{H_1: \sigma^2 < {\sigma_0}^2}$$

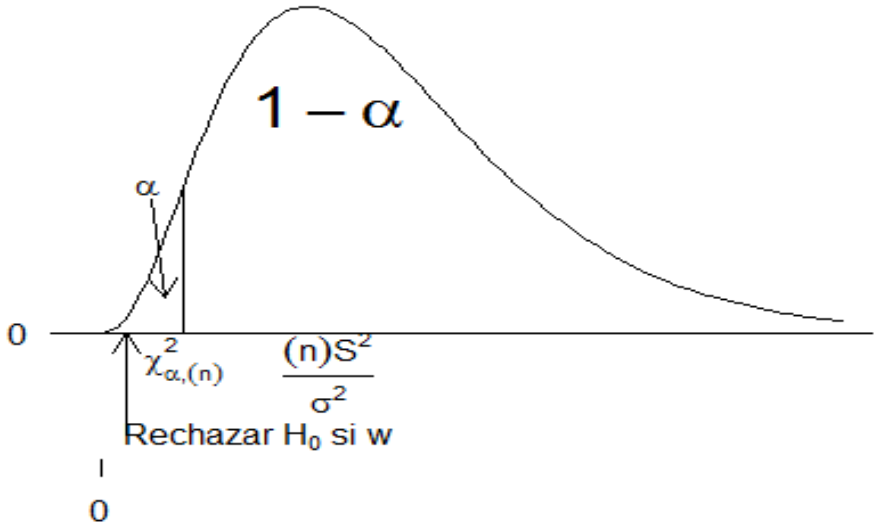
Con esta definición la región de rechazo es $C = \{(x_1, x_2, \dots, x_n) | w \leq k\}$, y es entonces que se puede definir el límite de la región de rechazo hallando k lo cual se hará estableciendo primero la máxima probabilidad de cometer el error de Tipo I, es decir el nivel de significancia α, para hallar una k que cumpla

$$\boldsymbol{P(X \in C) = P(w \leq k) = \alpha}$$

$$\boldsymbol{P\left(\frac{(n)S^2}{{\sigma_0}^2} \leq k\right) = \alpha}$$

Como la distribución del estadístico w es Ji-cuadrada con n grados de libertad entonces el valor de k es el cuantil de la distribución $\chi^2_{(n)}$ que acumula α, denotado como $\chi^2_{\alpha,(n)}$. Con base en el valor del cuantil, la regla de decisión es: si $w \leq \chi^2_{\alpha,(n)}$ entonces rechazar la hipótesis H_0 y en caso contrario no se debe rechazar H_0. Al usar las otras hipótesis alternativas se puede seguir un proceso análogo para hallar los cuantiles necesarios para determinar la regla de decisión correspondiente. De manera gráfica la región de rechazo de este caso se muestra a continuación.

Gráfica 4.5: Región de rechazo para H_0



Ahora se verá un ejemplo de la aplicación de esta prueba. Supóngase que se tiene una muestra de tamaño 20 que proviene de una distribución Normal $(15, \sigma^2)$, y se desea realizar, con un nivel de significancia de 5%, un contraste entre las siguientes hipótesis

$$\boldsymbol{H_0: \sigma^2 = 40}$$

$$\boldsymbol{H_1: \sigma^2} < 40$$

La siguiente tabla muestra los 20 valores de una muestra simulada, junto con valores auxiliares para realizar la prueba como son $S^2, w, \chi^2_{\alpha,(n)}$ y otros sólo como referencia como son $\alpha, {\sigma_0}^2$ y μ.

Tabla 4.4: Muestra de una población Normal (μ=15,σ^2)							
i	x_i	i	x_i	i	x_i	i	x_i
1	23.75	6	14.33	11	21.29	16	20.59
2	13.53	7	9.17	12	19.37	17	14.17
3	16.02	8	19.05	13	20.02	18	16.86
4	22.31	9	11.1	14	17.54	19	9.55
5	20.84	10	13.62	15	18.57	20	19.59
$\mu = 15$	$S^2 = 16.96$		$\sigma_0{}^2 = 40$	$w = 8.484$	$\chi^2_{\alpha,(20)} = 10.85$		$\alpha = 0.05$

El valor del estadístico de prueba se halla dentro de la zona de rechazo pues $\boldsymbol{w \leq \chi^2_{\alpha,(n)}}$ lo que quiere decir según la regla de decisión que se debe rechazar la hipótesis nula $H_0: \sigma^2 = 40$. Ahora se trabajará con un ejemplo parecido pero desde la perspectiva de la simulación Monte Carlo, el cual se basa en la estructura hecha para la prueba antes vista sobre la media de una población normal.

Simulación de Monte Carlo para pruebas sobre la varianza

Supóngase tener una muestra aleatoria $x_1, x_2, \dots, x_n$ de una población Normal(μ, σ^2) de la cual se conoce que $\mu = 100$. El tipo de prueba que se verá vía simulación será el mismo, sin embargo se usará en este caso parámetros distintos para contrastar ahora las hipótesis:

$$\boldsymbol{H_0: \sigma^2 = \sigma_0^2 = 250}$$

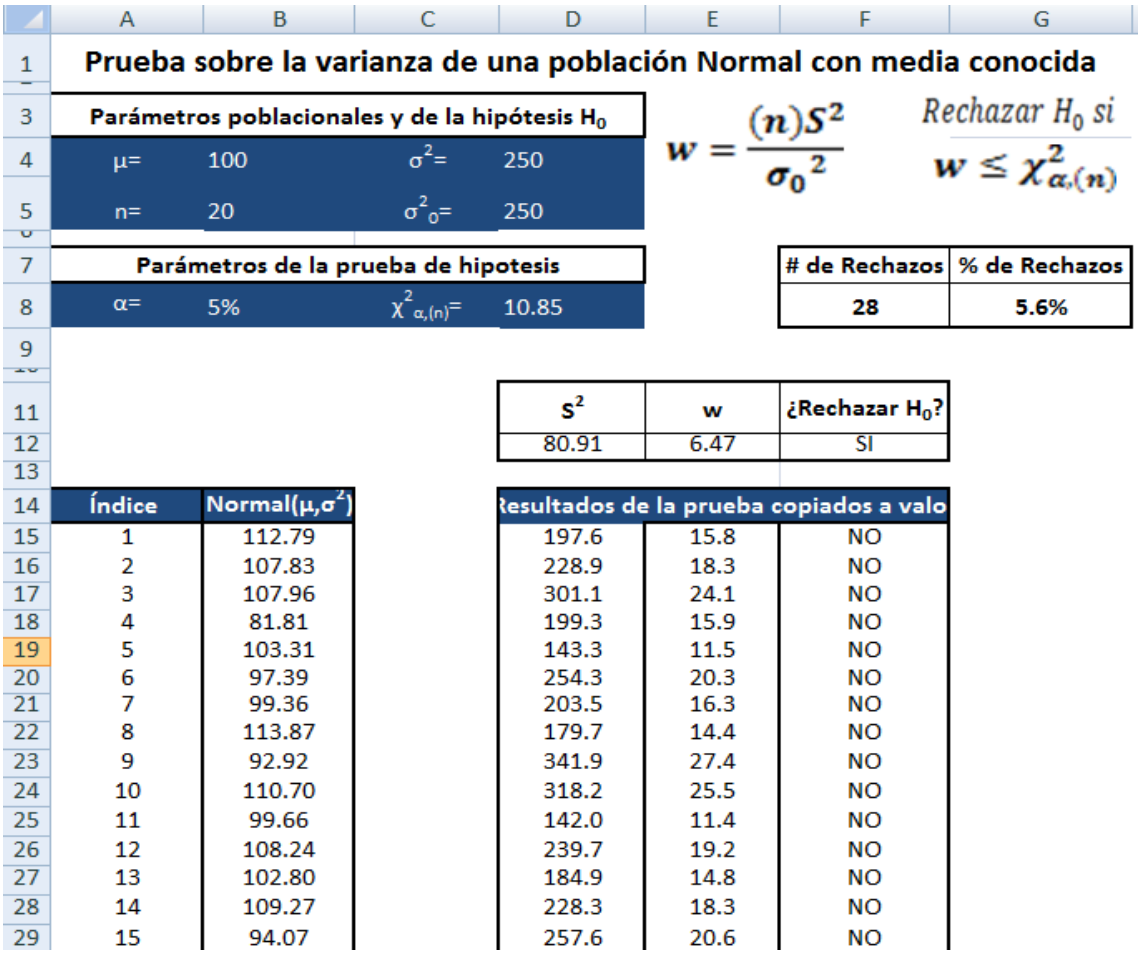
$$\boldsymbol{H_1: \sigma^2 < \sigma_0^2 = 250}$$

Para esta serie de pruebas se consideró en particular un nivel de significancia α del 5%, y un tamaño de muestra n inicial de 20 el cual se irá cambiando de manera dinámica conforme se requiera ver el tema de la potencia de la prueba.

En una de hoja de cálculo similar con la que se trabajó el tema anterior, se han establecidos los parámetros de la población y de la prueba, donde la varianza de la población se eligió de valor $\sigma^2 = 250$, para revisar el caso en que H_0 es cierta.

Luego se calcula el estadístico de prueba w, y los elementos antes vistos para realizar la prueba como S^2 y $\chi^2_{\alpha,(n)}$. Por último se evalúa por medio de una condicional si se rechaza la hipótesis nula con la regla $= \boldsymbol{SI}(\boldsymbol{w} \leq \chi^2_{\alpha,(n)}, \text{"}\boldsymbol{SI}\text{"}, \text{"}\boldsymbol{NO}\text{"})$. Una vez con los cálculos completos se puede correr una macro[21] que permita copiar y pegar cada evaluación de la prueba de manera automatizada. El siguiente diagrama es una muestra de cómo podría generarse la hoja de cálculo.

[21] Documentada en el Apéndice.



Prueba sobre la varianza de una población Normal con media conocida

Parámetros poblacionales y de la hipótesis H_0			
μ=	100	σ^2=	250
n=	20	σ^2_0=	250

$$w = \frac{(n)S^2}{{\sigma_0}^2} \qquad \text{Rechazar } H_0 \text{ si } w \leq \chi^2_{\alpha,(n)}$$

Parámetros de la prueba de hipotesis			
α=	5%	$\chi^2_{\alpha,(n)}$=	10.85

# de Rechazos	% de Rechazos
28	5.6%

S^2	w	¿Rechazar H_0?
80.91	6.47	SI

Índice	Normal(μ,σ^2)	Resultados de la prueba copiados a valo		
1	112.79	197.6	15.8	NO
2	107.83	228.9	18.3	NO
3	107.96	301.1	24.1	NO
4	81.81	199.3	15.9	NO
5	103.31	143.3	11.5	NO
6	97.39	254.3	20.3	NO
7	99.36	203.5	16.3	NO
8	113.87	179.7	14.4	NO
9	92.92	341.9	27.4	NO
10	110.70	318.2	25.5	NO
11	99.66	142.0	11.4	NO
12	108.24	239.7	19.2	NO
13	102.80	184.9	14.8	NO
14	109.27	228.3	18.3	NO
15	94.07	257.6	20.6	NO

Para el caso arriba observado la macro se corrió para realizar 500 simulaciones[22], y con los resultados se obtuvo el porcentaje de rechazos empíricos de 5.6% lo cual cercano al valor de la significancia $\alpha = 5\%$, interpretado como el máximo de la probabilidad de obtener un error de Tipo I (Rechazar H_0 cuando es cierta).

Con la estructura anterior ahora se pueden mover valores de la tabla de parámetros para poder ver cuál es el comportamiento de la prueba bajo estas modificaciones. La primer modificación que se puede hacer es sobre el valor de la varianza de la población lo que implicaría pasar de una condición en la que la hipótesis nula es cierta a otra condición en la que tal hipótesis es falsa, sin embargo conforme el valor de σ^2 se aleje de σ_0^2 se espera que el porcentaje de rechazos cambie de acuerdo a la dirección de la zona de rechazo.

La siguiente tabla de sensibilidad (potencia de la prueba) muestra los cambios sobre el porcentaje de rechazos obtenidos al cambiar el valor de σ^2. Además el número de simulaciones ahora se ha incrementado a 2,000.

Tabla 4.5: Valores del Alfa empírico cuando σ_0^2 = 250; # Simulaciones = 2000							
n \ μ	220	230	240	250	260	270	280
20	12.30%	10.15%	9.60%	6.25%	5.85%	4.40%	3.75%

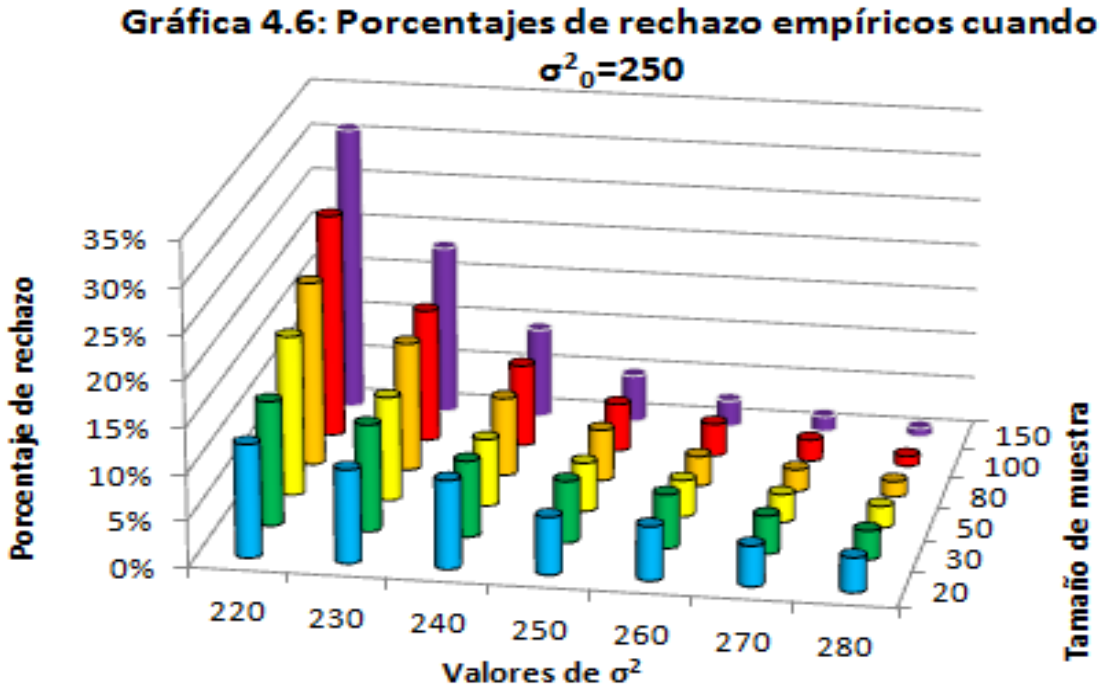
[22] Este número de simulaciones dentro de la hoja de cálculo tarda menos de medio minuto lo que permite visualizar lo que se está realizando, sin sacrificar tiempo dentro de una clase.

Lo que se nota en esta tabla es que conforme el valor de σ^2 se direcciona hacia la zona de rechazo el porcentaje de rechazos aumenta como se esperaría, mientras que en la dirección contraria los rechazos disminuyen pues la hipótesis alternativa está definida en una dirección. Los porcentajes de rechazo empíricos también se observan cercanos pero no iguales a lo esperado cuando se cumple H_0, esto se debe al tamaño de la muestra y al valor de los parámetros elegidos, ya que para realizaciones de un mismo escenario se obtienen porcentajes ligeramente distintos pero de valor similar.

Ahora lo que queda modificar es el tamaño de la muestra para ver la sensibilidad de la prueba en rechazar H_0 al ser falsa o no rechazarla al ser verdadera, lo que ésta directamente relacionado con la potencia de la prueba. La siguiente tabla muestra los resultados de aumentar el tamaño de la muestra y realizar el mismo análisis de sensibilidad para cada aumento de n.

Tabla 4.6: Valores del Alfa empírico cuando $\sigma_0^2 = 250$; # Simulaciones = 500							
$n \setminus \sigma^2$	220	230	240	250	260	270	280
20	9.60%	7.80%	6.20%	3.80%	2.80%	2.40%	2.20%
30	10.00%	7.00%	6.00%	5.20%	5.00%	1.80%	2.40%
50	13.80%	8.40%	6.60%	4.80%	3.20%	1.40%	0.60%
80	17.00%	13.20%	5.60%	4.80%	2.80%	2.20%	1.20%
100	21.60%	11.40%	6.00%	5.20%	2.80%	1.80%	0.80%
150	26.40%	17.40%	11.40%	4.90%	2.20%	1.00%	0.60%

Como se observa la potencia de la prueba concuerda con los resultados esperados, sin embargo se puede notar que el aumento de la potencia con respecto a n no es tan radical como en el caso de la prueba de la media muestral de la sección anterior. Por último en la Gráfica 4.6 se pueden ver las diferentes potencias obtenidas en la Tabla 4.6 donde se aprecia como la prueba discrimina mejor los cambios en σ^2 a medida que el tamaño de la muestra aumenta.



Prueba de contraste de medias de dos poblaciones, con observaciones por pares.

Cuando se tienen dos muestras aleatorias $X = X_1, X_2, \dots, X_n$y $Y = Y_1, Y_2, \dots, Y_n$, de tal forma que la i-ésima observación de X es asociada de forma natural con la i-ésima observación de Y, se tiene una observación en parejas (Xi, Yi). Un ejemplo de este tipo de observaciones es cuando se realizan mediciones sobre un mismo fenómeno, en dos periodos de tiempo distintos de tal manera que se pueda identificar en cada medición cuál se realizó antes y cuál después.

Este tipo de condiciones en las que se obtiene cada par de observaciones dan lugar a que $\boldsymbol{X}$ no sea independiente de $\boldsymbol{Y}$**.** Lo que se desea contrastar en este tipo de muestras es si existen diferencias entre las medias de ambas muestras. La forma de atacar este problema es definiendo la variable $D = X - Y$ la cual calcula las diferencias para las n parejas de la muestra como $d_i = x_i - y_i$.

El supuesto inicial es que ambas muestras se distribuyen Normal $X \sim N(\mu_X, \sigma_X^2)$ y $Y \sim N(\mu_Y, \sigma_Y^2)$, implicando que $D \sim N(\mu_D, \sigma_D^2)$; esta transformación permite hacer inferencias sobre una sola variable aleatoria D. El parámetro sobre el que se realizará una prueba de hipótesis será la media $\mu_D = \mu_X - \mu_Y$ de la cual se considera se desconoce el valor de la varianza σ_D^2.

Cuando se tiene una variable distribuida Normal(μ, σ^2) donde σ^2 es desconocida, no es factible el uso del cociente $\boldsymbol{Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}}$**,** pero se soluciona retomando de la teoría estadística que $\boldsymbol{w = \frac{(n-1)S^2}{\sigma_0{}^2}}$ se distribuye Ji-cuadrada con $n - 1$ grados de libertad. Con los elementos anteriores se puede usar la relación que guarda una distribución Normal con la Ji-cuadrada

$$\boldsymbol{T = \frac{Z}{\sqrt{\frac{w}{n-1}}} \sim t_{(n-1)}}$$

Donde $t_{(n-1)}$ es la distribución t de Student con $n - 1$ grados de libertad. T en términos de los estadísticos del problema inicial tendría la siguiente expresión:

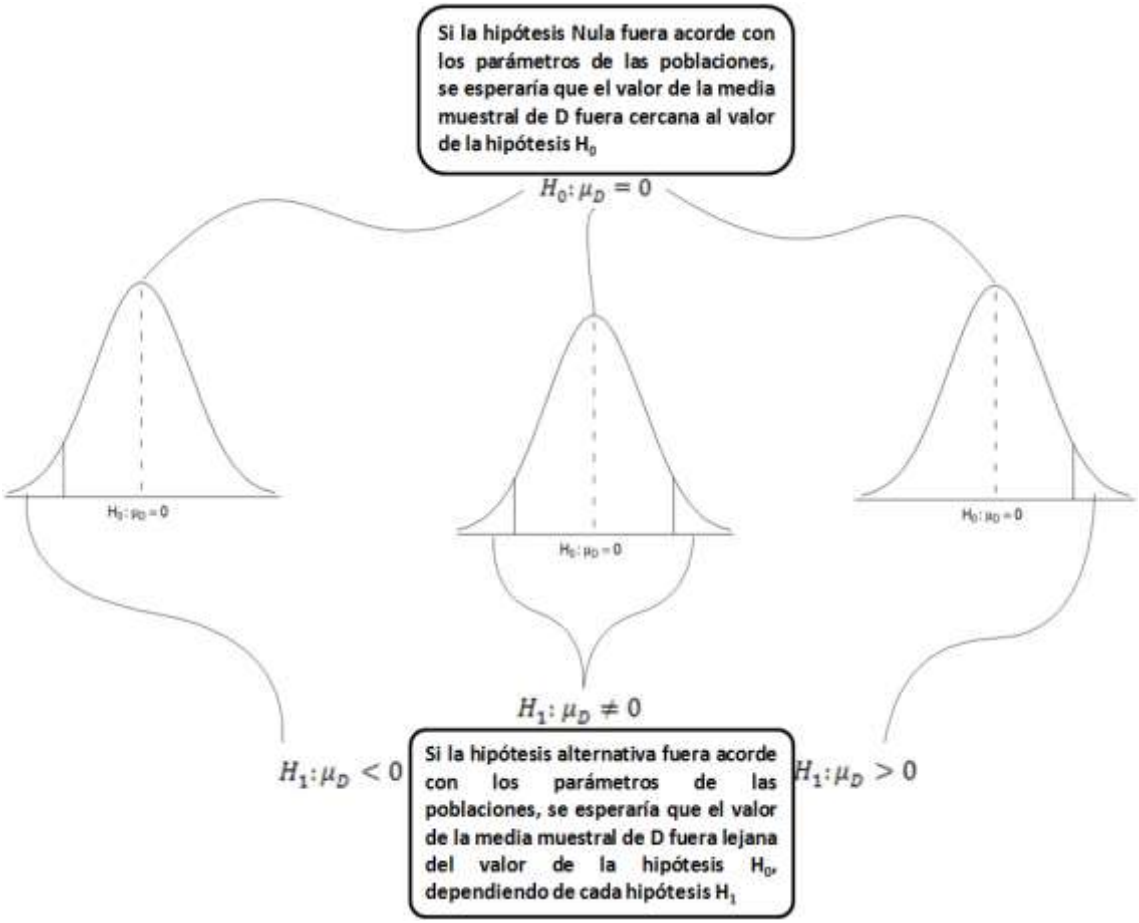
$$\boldsymbol{T = \frac{\bar{D} - \mu_0}{S_D/\sqrt{n}}} \quad , \quad \boldsymbol{S_D = \sqrt{S_D^2} = \sqrt{\sum_{i=1}^{n} \frac{(d_i - \bar{D})^2}{n-1}}}$$

El termino μ_0 dependerá de la hipótesis nula que se defina, en este caso considérese como motivación la necesidad de probar si las medias de las variables X y Y son iguales es decir $D = 0$, lo que define a H_0 como

$$\boldsymbol{H_0 : \mu_D = 0}$$

Con H_0 definida ahora se pueden contrastar contra 3 distintas hipótesis alternativas, la primera es que la media poblacional de X sea menor que el valor de la media de Y, es decir $H_1 : \mu_D < 0$, otra hipótesis alternativa es aquella que afirma que el valor de la media de X es mayor que la media poblacional de Y, $H_1 : \mu_D > 0$ y la tercer hipótesis

alternativa es usada cuando sólo se requiere probar que las medias no son iguales así que el contraste sería contra $H_1: \mu_D \neq 0$, las regiones de rechazo correspondientes a cada contraste se encuentran resumidas en el siguiente diagrama:



Como se puede ver en el diagrama la región de rechazo tiene una relación lógica con cada hipótesis alternativa pues si esta hipótesis estuviera más cerca del verdadero valor del parámetro entonces el estadístico T también reflejaría este comportamiento alejándose del valor propuesto por la hipótesis nula. Para trabajar un caso completo supóngase que se desea contrastar si las variables X y Y tienen o no la misma media, es decir se quiere realizar la siguiente prueba:

$$\boldsymbol{H_0: \mu_D = 0}$$

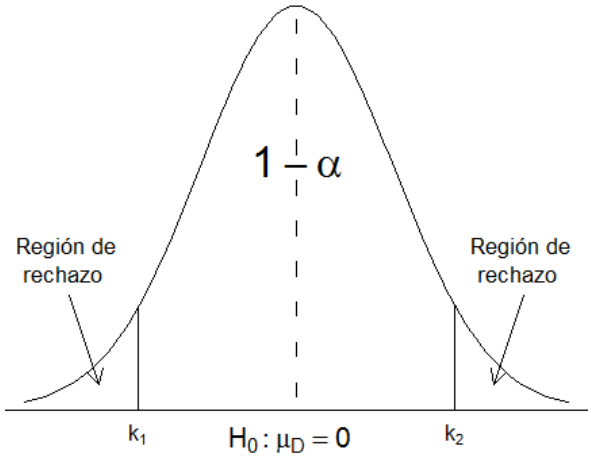
$$\boldsymbol{VS}$$

$$\boldsymbol{H_1: \mu_D \neq 0}$$

Seguido de la definición del contraste a realizar se puede establecer la región de rechazo, la cual toma en este caso la expresión $C = \{x_1, x_2, \dots, x_n; y_1, y_2, \dots, y_n | T \leq k_1 \; o \; T \geq k_2\}$. Para encontrar los valores de k_1 y k_2 se debe establecer primero el nivel de confianza α, con el cual se debe cumplir:

$$\boldsymbol{P(k_1 \leq T \leq k_2) = 1 - \alpha}$$

Por lo tanto, puesto que se conoce la distribución del estadístico T, k_1 es el cuantil que acumula $\frac{\alpha}{2}$ y k_2 es el cuantil que acumula $1-\frac{\alpha}{2}$ de la distribución $t_{(n-1)}$. La región de rechazo para esta prueba en particular que puede observar en el siguiente diagrama:



Ahora se verá un ejemplo de la aplicación de esta prueba; supóngase que se está probando un medicamento nuevo para reducir el colesterol “malo” en la sangre medido en mg/dL y se desea probar con un nivel de significancia del 5%, si el medicamento probado en una muestras de 15 pacientes produce algún cambio significativo en la concentración de colesterol malo. Considérese como X a la concentración de colesterol antes de tomar el medicamento y a Y como el colesterol medido después de haber tomado el medicamento.

La siguiente tabla muestra los resultados de las observaciones por pareja, las diferencias d_i $1 \leq i \leq 15$, junto con los valores de las medias, el estadístico de prueba y los valores de los cuantiles k_1 y k_2.

Tabla 4.7: Resultados para la prueba sobre la media de D

i	x_i	y_i	$d_i = x_i - y_i$	i	x_i	y_i	$d_i = x_i - y_i$
1	172.17	170.99	1.18	9	154.88	156.64	-1.759
2	191.67	181.49	10.179	10	128.48	130.56	-2.079
3	158.87	152.58	6.291	11	190.1	187.74	2.364
4	154.93	163.9	-8.964	12	152.42	151.24	1.174
5	157.03	151.76	5.272	13	186.01	183.64	2.367
6	196.36	193.4	2.96	14	140.18	134.55	5.632
7	145.36	140.35	5.011	15	159.3	152.65	6.652
8	171.28	166.67	4.602				
$\bar{X} = 163.94$	$\bar{Y} = 161.21$	$\mu_D = 2.725$	$S_D^2 = 20.87$	$T = 2.31$	$k_1 = -2.14$		$k_2 = 2.14$

Por la simetría de la distribución t se tiene que el valor de $k_1 = -k_2$; por otra parte contrastando el estadístico de prueba se puede ver que $T > k_2$ por lo que cae en la región de rechazo y por consecuente se debe rechazar la hipótesis nula, concluyendo así, con un nivel de significancia del 5% que existen evidencias estadísticas que demuestren que el uso del medicamento altera la concentración de colesterol malo en la sangre.

Simulación de Monte Carlo

El siguiente paso es darle la perspectiva de simulación Monte Carlo a esta prueba para estudiar su comportamiento bajo distintos escenarios, para así comprenderla mejor. Lo primero que se tiene que saber es cómo simular las dos muestras aleatorias. Considérese el caso en el que las variables X y Y están correlacionadas, planteando el siguiente modelo

$$\boldsymbol{Y = X + e}$$

Donde para conservar las hipótesis de trabajo sobre la normalidad de Y, se debe tomar a $e \sim N(\mu_e, {\sigma_e}^2)$ y $X \sim N(\mu_x, \sigma_x^2)$. La simulación entonces se puede hacer por medio de generar dos variables normales por medio de la fórmula $= INV.DISTR.NORM(Aleatorio(), \mu, raiz(\sigma^2))$, donde la primera es con $\mu = \mu_x, \sigma^2 = \sigma_x^2$ y la segunda con $\mu = \mu_e, \sigma^2 = \sigma_e^2$

Para trabajar la simulación Monte Carlo, el tipo de contraste para esta prueba será el mismo que el usado para el ejemplo anterior, es decir, la prueba T de dos colas. Se inicia con suponer que se poseen dos muestras aleatorias $x_1, x_2, \dots, x_n$ y $y_1, y_2, \dots, y_n$ provenientes cada una de una distribución Normal $N(\mu_x, \sigma_x^2)$, $N(\mu_y, \sigma_y^2)$ respectivamente, tomando una n inicial igual a 20, de las cuales se quiere realizar el contraste de las siguientes hipótesis, con un nivel de significancia del 5%:

$$\boldsymbol{H_0: \mu_D = \mu_x - \mu_y = 0 \ \rightarrow \ \mu_x = \mu_y}$$

$$\boldsymbol{H_1: \mu_D = \mu_x - \mu_y \neq 0 \ \rightarrow \ \mu_x \neq \mu_y}$$

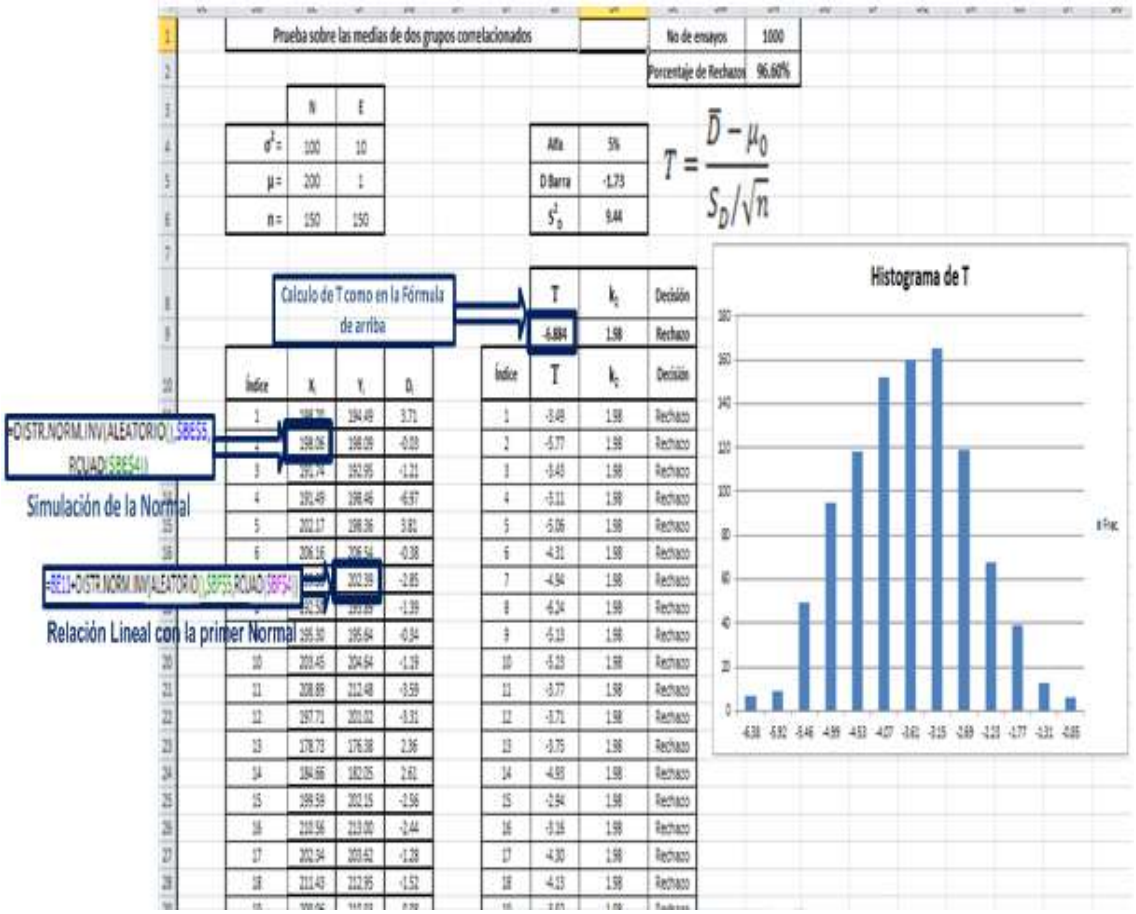
Trabajando de la misma manera en hojas de cálculo se tiene primero que simular las muestras aleatorias de tamaño $n = 20$ con la técnica anteriormente explicada, después hay que obtener las diferencias $d_i = x_i - y_i$, con las cuales se calculará el estadístico de prueba $T = \frac{\bar{D}}{S_D/\sqrt{n}}$. Posteriormente se debe hallar el valor del cuantil K_2 por medio de la fórmula $= DISTR.T.INV(T, n-1)$ y por último establecer la condicional de rechazo auxiliados de la relación $k_1 = -k_2$ con la fórmula:

$$= \boldsymbol{SI(T \leq k_2, SI(T \geq -k_2, \text{"No Rechazo"}, \text{"Rechazo"}), \text{"Rechazo"})}.$$

La simulación de N repeticiones se puede lograr copiando y pegando a valores a T, k_2 y a la decisión de la condicional, por medio de una macro[23] que repita esta operación N veces; además es recomendable agregar un histograma de los valores de T, para comprobar de forma gráfica, la teoría sobre su distribución, junto con una fórmula en imagen para incentivar la memoria del estudiante y tener siempre claro lo que se está calculando.

La siguiente imagen muestra una propuesta de hoja cálculo con los elementos anteriormente detallados.

[23] Para ver el código de esta macro revise el apéndice.



El número de simulaciones generadas fueron $N = 500$, con las cuales se obtuvo el porcentaje de veces que la prueba rechazó la hipótesis nula. Como se puede apreciar en la parte superior de la imagen, hubo un alto porcentaje de rechazos, debido al valor de la media de la variable e.

A partir de la definición anterior y dada la estructura de la hoja de cálculo se puede variar a voluntad los parámetros para revisar el comportamiento de la prueba. El primer parámetro a modificar es $\boldsymbol{\mu_e}$, pues impacta directamente a la media variable original X, afectando los resultados de la prueba, debido a que, conforme se aleje de 0 la media $\boldsymbol{\mu_e}$, la media de Y se distanciará de la media de X.

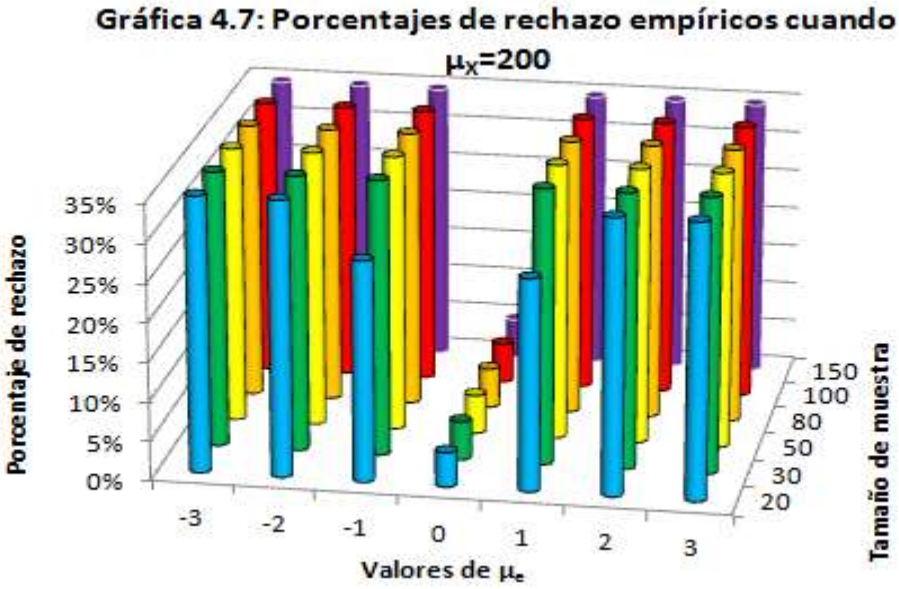
La Tabla 4.8 muestra los distintos escenarios en los que se varió únicamente el parámetro μ_e, manteniendo constante los parámetros de la distribución X y el tamaño de muestra $n = 20$.

Tabla 4.8: Valores del porcentaje de rechazos cuando μ_X=200, σ_X^2 = 100; σ_e^2=10 ; # Simulaciones = 1000							
n \ μ_e	-3	-2	-1	0	1	2	3
20	98.60%	76.20%	28.00%	4.40%	26.90%	77.40%	98.10%

Ahora lo siguiente que se puede variar es el tamaño de la muestra original n para estudiar cómo se comporta la potencia de la prueba en los distintos escenarios en los que se modifica la media de la variable e. La Tabla 4.9 muestra los resultados de haber generado las simulaciones correspondientes para cada uno de los escenarios.

Tabla 4.9: Valores del porcentaje de rechazos cuando $\sigma_0^2 = 100$; $\mu_X=200$; $\sigma_e^2=10$; # Simulaciones = 1000							
n \ μ_e	-3	-2	-1	0	1	2	3
20	98.60%	76.20%	28.00%	4.40%	26.90%	77.40%	98.10%
30	100.00%	91.30%	39.40%	4.90%	38.30%	92.40%	100.00%
50	100.00%	99.20%	61.00%	4.90%	64.00%	99.30%	100.00%
80	100.00%	100.00%	81.00%	5.00%	79.10%	100.00%	100.00%
100	100.00%	100.00%	87.00%	5.00%	88.50%	100.00%	100.00%
150	100.00%	100.00%	98.00%	5.00%	96.60%	100.00%	100.00%

Como se puede observar cuando el parámetro $\mu_e = 0$ el porcentaje de rechazos se aproxima al nivel de significancia señalando el porcentaje de errores sobre las simulaciones. A medida que el tamaño de la muestra aumenta se nota cómo la prueba se vuelve más potente, pues conforme la media $\boldsymbol{\mu_e}$ cambia de valor, entonces el porcentaje de rechazos aumenta pues la prueba detecta que la hipótesis nula cada vez es más probable que sea falsa. La Gráfica 4.7 muestra el comportamiento de los valores de la potencia, donde se puede observar la sensibilidad de la prueba ante los diversos escenarios.



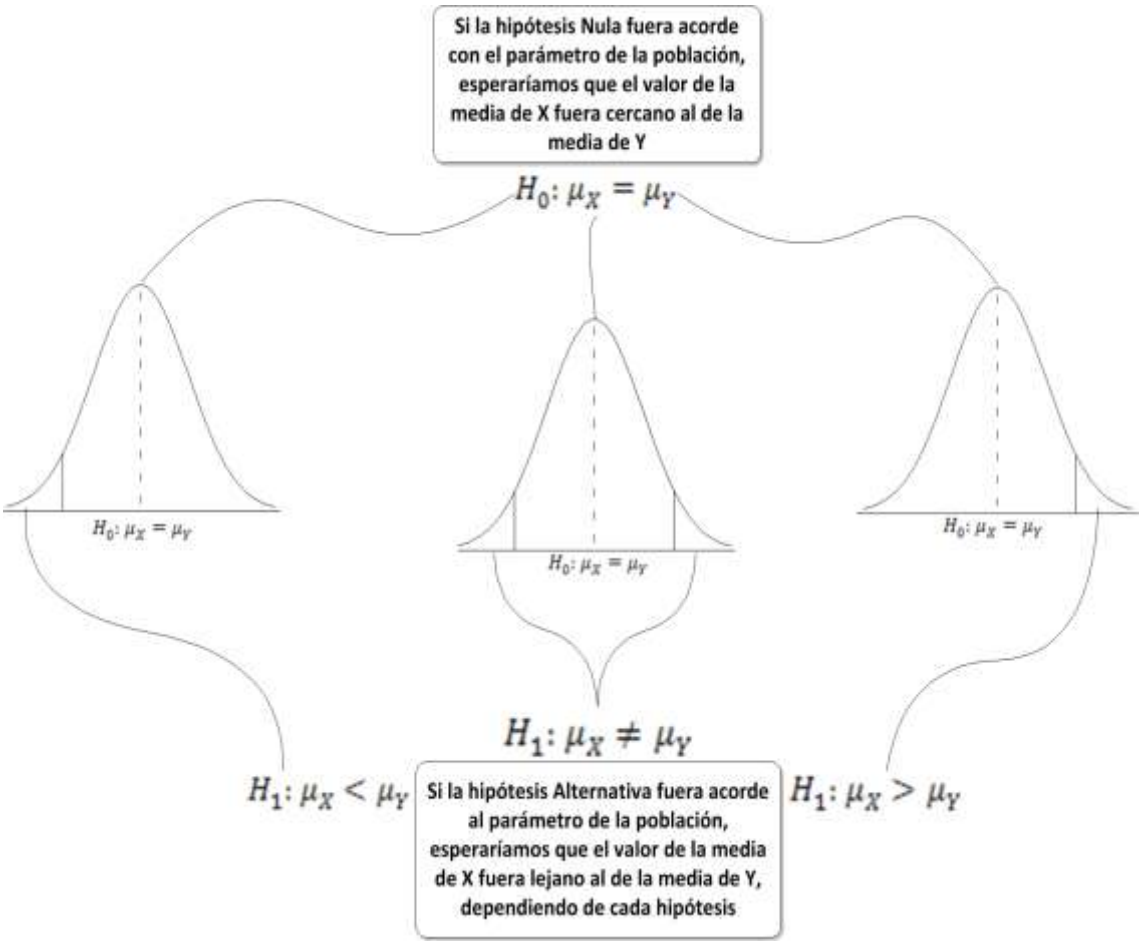
Prueba sobre la diferencia de medias de dos muestras independientes con varianzas conocidas

Otro uso común de las pruebas de hipótesis es el contraste entre dos poblaciones. El caso que se verá a continuación es el contraste entre las medias de dos poblaciones independientes. Supóngase que la muestra X de una población es de tamaño n mientras que para la segunda población, la muestra Y es de tamaño m, ambas distribuidas $\boldsymbol{N(\mu_X, \sigma_X^2), N(\mu_Y, \sigma_Y^2)}$ respectivamente, con el supuesto que **ambas varianzas son iguales** $\boldsymbol{\sigma_X^2 = \sigma_Y^2 = \sigma^2}$ **y** $\boldsymbol{\sigma^2}$ **es conocida.** Al ser ambas muestras Normales entonces $\boldsymbol{\overline{X} \sim N\left(\mu_X, \frac{\sigma^2}{n}\right)}$ **y** $\boldsymbol{\overline{Y} \sim N\left(\mu_Y, \frac{\sigma^2}{m}\right)}$**.**

La prueba consiste en comparar la diferencia de medias suponiéndola como hipótesis nula igual a una constante, es decir:

$$\boldsymbol{H_0}: \boldsymbol{\mu_X} - \boldsymbol{\mu_Y} = \Delta$$

Usualmente Δ se considera como 0 al definir las hipótesis para contrastar directamente la relación entre las medias de las poblaciones, es decir, la hipótesis nula se considera como $H_0: \boldsymbol{\mu_X = \mu_Y}$, la cual se puede contrastar contra tres distintas hipótesis alternativas: una posible hipótesis alternativa es que la media de X es menor a la media de Y expresada como $H_1: \mu_X < \mu_Y$; otra hipótesis alternativa es $H_1: \mu_X > \mu_Y$ que es el caso contrario al anterior; finalmente se tiene la hipótesis alternativa $H_1: \mu_X \neq \mu_Y$, que implica simplemente que las medias no son iguales. El siguiente diagrama muestra gráficamente las regiones de rechazo correspondientes a las hipótesis alternativas enunciadas.



Con base en los supuestos iniciales la diferencia de medias $\bar{X} - \bar{Y}$ tiene una distribución Normal $\left(\boldsymbol{\mu_X - \mu_Y}, \boldsymbol{\sigma^2}\left(\frac{\mathbf{1}}{\boldsymbol{n}} + \frac{\mathbf{1}}{\boldsymbol{m}}\right)\right)$, la cual se puede estandarizar para construir el estadístico de prueba

$$\boldsymbol{Z} = \frac{\boldsymbol{\bar{X} - \bar{Y}}}{\sqrt{\boldsymbol{\sigma^2}\left(\frac{\mathbf{1}}{\boldsymbol{n}} + \frac{\mathbf{1}}{\boldsymbol{m}}\right)}} \sim \boldsymbol{N(0,1)}$$

Considérese ahora el caso derecho del diagrama para contrastar las hipótesis

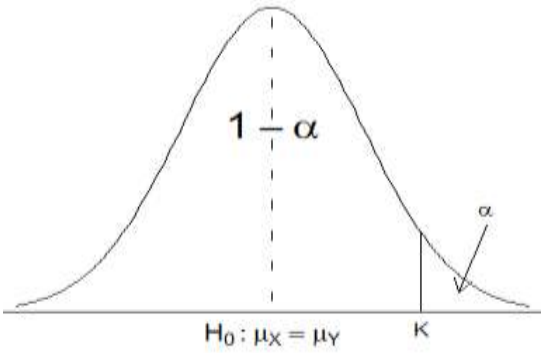
$$\boldsymbol{H_0}: \boldsymbol{H_0}: \boldsymbol{\mu_X = \mu_Y} \textbf{ Contra } \boldsymbol{H_1}: \boldsymbol{\mu_X > \mu_Y} \textbf{ .}$$

La región de rechazo para esta prueba es $C \coloneqq \{x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_m | \, Z > K\}$. El valor de K se obtiene como el cuantil que cumpla

$$\boldsymbol{P(Z > K) = \alpha}$$

$$\boldsymbol{P}\left(\frac{\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}}}{\sqrt{\boldsymbol{\sigma^2}\left(\frac{\boldsymbol{1}}{\boldsymbol{n}} + \frac{\boldsymbol{1}}{\boldsymbol{m}}\right)}} > K \right) = \boldsymbol{\alpha}$$

Donde α es el nivel de confianza considerado para realizar la prueba. La región de rechazo y su probabilidad asociada para este problema se ejemplifica en la siguiente gráfica.



Para ejemplificar el funcionamiento de la prueba con esta hipótesis alternativa en particular, se aprovechará la estructura generada para la prueba de diferencia de medias con observaciones por pares, en la cual se simularon dos muestras de una distribución Normal con cierta dependencia entre ellas, pero ahora se simularán muestras independientes, de tamaño distinto. Para llevar a cabo esto se debe generar una tabla con los parámetros elegidos para ambas distribuciones y se introduce la fórmula $INV.DISTR.NORM(Aleatorio(), \mu, raiz(\sigma^2))$ en ambas columnas con referencia a los parámetros propuestos para cada variable, generando una columna de tamaño n y otra de tamaño m.

Para mostrar un ejemplo, se decidió contrastar dos poblaciones bajo las hipótesis $\boldsymbol{H_0}: \boldsymbol{\mu_X = \mu_Y}$ contra $\boldsymbol{H_1}: \boldsymbol{\mu_X > \mu_Y}$ con un nivel de significancia elegido $\alpha = 5\%$. La primera distribución X se simulará como $\boldsymbol{N(\mu_x = 100, \sigma^2 = 20)}$, con $\boldsymbol{n = 200}$ y la segunda distribución Y como $\boldsymbol{N(\mu_y = 100, \sigma^2 = 20), m = 100,}$ para revisar el caso cuando se cumple $\boldsymbol{H_0}$.

Ahora, para evaluar el resultado de la prueba, se obtiene el cuantil de la distribución Normal estándar $Z_{1-\alpha}$, por medio de la fórmula $\boldsymbol{INV.DISTR.NORM.EST(1 - \alpha)}$ y por último se compara con una simple condicional para automatizar la decisión si se debe rechazar la hipótesis nula con la fórmula $\boldsymbol{SI(Z > Z_{1-\alpha}}$, "**SI**", "$\boldsymbol{NO}$").

Una vez con estos elementos al copiar a valor el resultado automáticamente se generará una nueva muestra, por lo cual la rutina macros empleada anteriormente que automatiza este proceso[24], que en esta ocasión se utilizó para realizar cada escenario 5,000 veces, así que, también ayudará a mostrar la potencia de la prueba.

En la siguiente imagen se muestra una propuesta en hoja de cálculo para llevar a cabo la prueba y sus simulaciones, en la que se puede observar del lado izquierdo la tabla de parámetros para las muestras y sus simulaciones en dos columnas, y en el lado derecho los parámetros de la prueba, la regla de decisión, además del resumen de las pruebas rechazadas en número y porcentaje. En este caso como la hipótesis nula se cumple se puede observar que el porcentaje de pruebas rechazadas es muy cercano a la significancia de la prueba.

Estructura propuesta para la simulación de la prueba de diferencia de medias

Pruebas sobre la diferencia de medias de poblaciones independientes con varianzas conocidas

Parámetros poblacionales y de la hipótesis H_0			
μ_X=	100	n=	200
μ_Y=	100	m=	100
σ^2=	20		

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

Rechazar H_0 *si* $Z > Z_{1-\alpha}$

Parámetros de la prueba de hipotesis				# de Pruebas	# de Rechazos	% de Rechazos
α=	5%	$Z_{1-\alpha}$=	1.645	5,000	245	4.90%

Índice	X	Y
1	91.7	94.7
2	101.8	106.3
3	97.1	101.4
4	99.0	100.2
5	109.5	100.1
6	100.8	100.8
7	101.5	99.0
8	103.7	106.1
9	102.5	98.2
10	104.4	96.7
11	103.2	91.2
12	96.2	104.4
13	98.1	105.7
14	104.5	97.0
15	98.4	102.6
16	91.3	103.6
17	104.2	100.0
18	111.8	108.8
19	101.4	102.4
20	97.1	107.1
21	98.4	102.5

Z	¿Rechazar H_0?
0.76	NO

# de muestra	Z	¿Rechazar H0?
1	0.4	NO
2	-0.6	NO
3	0.5	NO
4	2.1	SI
5	0.2	NO
6	0.1	NO
7	1.0	NO
8	0.7	NO
9	1.1	NO
10	0.8	NO
11	0.3	NO
12	0.7	NO
13	1.4	NO
14	0.2	NO
15	0.4	NO
16	-0.9	NO
17	-0.2	NO
18	0.7	NO
19	0.0	NO

Una vez construida esta estructura, se probaron diversos escenarios para los valores de las medias de ambas poblaciones cuyo resumen de los resultados se halla en la siguiente Tabla, donde se puede observar que el porcentaje de rechazo, es muy

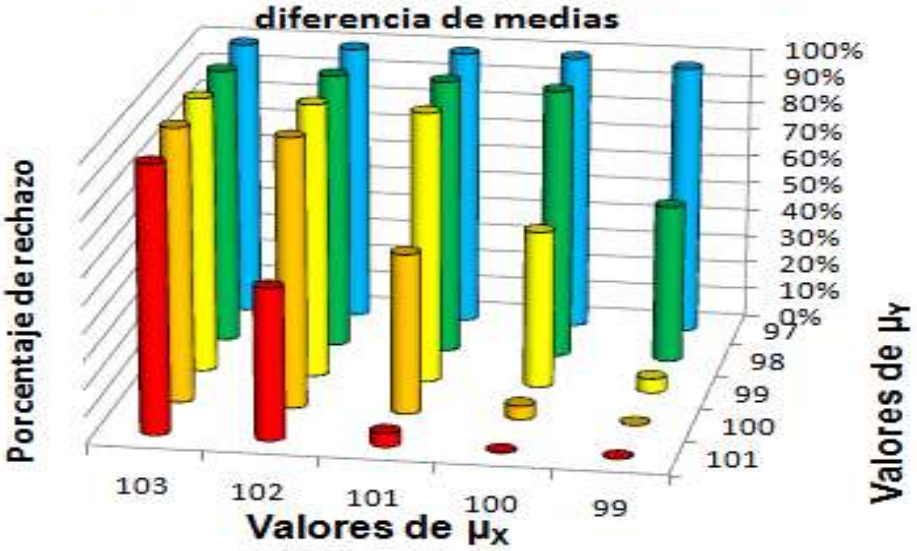
[24] Consulte el código en el apéndice.

similar en los casos cuando la diferencia real entre medias es la misma, y coincide con la significancia cuando las medias son iguales, mientras que a medida que la media de X aumenta, el porcentaje de rechazos es cada vez mayor pues la hipótesis alternativa se vuelve cada vez más evidente, mientras que en un sentido contrario cuando la media de Y es mayor el porcentaje de rechazos se aproxima a 0, pues en estos casos cuando la hipótesis nula se cumple de una forma mucho más fuerte.

Tabla 4.10: Valores del porcentaje de rechazos para le prueba de diferencia de medias, con σ^2=20, n=200 , m=100 # Simulaciones = 5,000					
$\mu_y \setminus \mu_x$	99	100	101	102	103
97	97.88%	99.98%	100.00%	100.00%	100.00%
98	57.26%	97.80%	99.54%	100.00%	100.00%
99	5.12%	56.92%	98.30%	99.58%	100.00%
100	0.00%	4.88%	57.90%	97.90%	99.60%
101	0.00%	0.20%	4.96%	55.00%	97.20%

En la Gráfica siguiente se muestra el comportamiento de los resultados de la tabla anterior, donde se puede observar de forma clara la potencia de la prueba.

Gráfica 4.8: Porcentajes de rechazo empíricos para la diferencia de medias



De esta manera se concluye esta sección, mostrando la idea principal en que las pruebas de hipótesis, tanto en su teoría como en su aplicación se pueden revisar de una forma más intuitiva, a través de las simulaciones Monte Carlo empleada como método didáctico en las explicaciones con diversas estructuras reproducibles por algún alumno o docente.

Capítulo V:

Enseñanza del método de análisis de regresión lineal simple mediante simulación Monte Carlo

En el siglo XIX los experimentos de Francis Galton sobre dimensiones de semillas de guisantes y posteriormente, la estatura de poblaciones humanas, publicados en 1886, lo llevaron a proponer el término de "regresión" el cual lo empleó en el estudio de una población, al describir la forma en que los padres que son más altos que la estatura media poblacional, tienden a tener hijos con una estatura menor y los padres de estaturas menores a la media tienden a tener hijos con una mayor altura. Entonces el concepto de regresión se entiende como un retroceso o regreso en el comportamiento progresivo de un conjunto de observaciones al comportamiento medio.

En su estudio, Galton estableció este comportamiento como una ley conocida como la ley de regresión universal, sin embargo, Galton notó también la simplicidad de sus explicaciones, pues también estudió las relaciones e influencias que tienen las estaturas de los padres, y los abuelos, sobre la estatura de la tercera generación. Además, en otra perspectiva, observó a la estatura total de un individuo, como la suma total de la altura de los huesos que conforman al ser humano, el cual clasificó en varias secciones. Por otro lado, al final de su estudio comenta que debido a falta de material, no podía aún describir otras relaciones interesantes, que compensó con experimentos posteriores.

El desarrollo de otras ciencias también ha llevado a estudiar este tipo de relaciones desde otra perspectiva. Por ejemplo, con respecto a la altura, en 2013 el *Tech museum of innovation,* apoyado por el Departamento de Genética de la Escuela de Medicina de la Universidad de Stanford, publicó un artículo en el que menciona más variables relacionadas a la altura final de los individuos, como el ambiente en el que se desarrollan los hijos, tomando el caso de la población japonesa, la cual bajo una ventana de observación de más generaciones (Galton consideró hasta la tercera generación) existe un aumento significativo de la altura debido a un mayor consumo de proteínas en la dieta diaria de las generaciones sucesoras.

Además, desde el punto de vista genético, menciona que se ha localizado a diversos genes que determinan la altura, los cuales son alrededor de 20, lo cual apoya la teoría de que la altura final está descrita por una serie de factores, en mayor o menor importancia, más un error aleatorio resultado de la incertidumbre de no conocer todas las variables que influyen en el crecimiento final de un individuo. El artículo menciona además como posibles factores la alimentación de la madre durante la gestación, sus cuidados médicos o el consumo de cigarro.

En general el modelo de regresión lineal es expresado matemáticamente como la descripción de una variable dependiente Y en función de una combinación lineal de una serie de variables independientes $X_1, X_2, \dots, X_k$ más un error aleatorio u, es decir

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u}$$

Donde $\beta_0, \beta_1, \ldots, \beta_k$ es una serie de parámetros desconocidos que establecen la importancia o peso de cada variable X_i para describir el comportamiento de la variable Y

Con el objetivo de instruir el tema de una forma más tangible para un alumno mediante y muestras simuladas por simulación Monte Carlo y gráficas, en este apartado se explorará el modelo de regresión simple, en el cual la variable Y depende de una única variable X. El modelo se puede emplear como introducción a modelos que consideren más variables y para clarificar los conceptos principales del análisis de regresión.

Modelo de regresión lineal simple

Este modelo se denomina simple ya que solo considera que los valores de la variable dependiente Y , están influenciados por los valores de la variable X, más un error aleatorio. Cuando los valores de las variables X y Y son observados, se registran como la muestra observada representada por una serie de n parejas de la forma (X_i, Y_i) .

El problema de hallar si existe una relación específica entre X y Y se reduce a un espacio de dos dimensiones, por lo que el modelo estará por una recta más un error aleatorio u. La relación entre las variables se expresa matemáticamente como:

$$\boldsymbol{Y = \alpha + \beta X + u}$$

Los parámetros que se desconocen son α la cual se puede asociar con el concepto de geometría como la ordenada al origen y β que se asocia con la pendiente de la recta.

El término de error aleatorio u, tiene varios supuestos, el primero es que su distribución sea Normal; esta característica hipotética es aceptada en muchos casos debido a que ciertos criterios matemáticos la identifican como una condición suficiente pero no necesaria para realizar pruebas de hipótesis sobre los parámetros, pero el supuesto es requerido para la construcción de intervalos de confianza sobre los estimadores de los parámetros $\boldsymbol{\alpha}$ **y** $\boldsymbol{\beta}$, los cuales se aplicarán más adelante.

Otra razón por la que se acepta este supuesto es que en ciertos estudios el objetivo es analizar el comportamiento medio de la variable dependiente Y y es posible trabajar con una muestra lo suficientemente grande para que la aproximación al comportamiento medio, sea razonable por medio de una Normal, como por ejemplo en investigaciones en tópicos de salud pública[25].

Otra interpretación más intuitiva sobre el error que incluye en el modelo, es que se puede considerar como una suma de todos los efectos aleatorios que se desconocen y que influyen en el valor final de la variable Y. Esta última idea puede servir como una ilustración intuitiva de los efectos conjuntos sobre un resultado, si se revisa como un breviario el comportamiento de la máquina de Galton.

Otro supuesto importante sobre los errores u_i, es que son independientes entre sí, por lo cual la correlación entre dos errores cualesquiera es nula. Además, se considera que los efectos de los errores se cancelan entre sí hacia la media, como en el caso del

[25] Para una referencia más extensa sobre el supuesto puede revisar el trabajo de Lumney T.; Diehr P. ; Emerson S.; Chen L. (2002) *The Importance of the Normality Assumption in large public Health data sets*

efecto de regresión en los experimentos de Galton, por lo cual se considera que la media de los errores es 0.

El tercer supuesto es que la varianza del error es un valor constante $Var(u) = \sigma^2$ sobre el rango en que están definidas las variables X y Y, debido a que un aspecto importante del análisis de regresión es también estimar la varianza de los estimadores de los parámetros, si no se cumple este supuesto entonces la medición de esta variabilidad se realizaría de forma incorrecta.

En resumen los supuestos son:

$$\boldsymbol{u_i \sim N(0, \sigma^2) \quad i = 1, 2, \dots, n}$$

$$\boldsymbol{E(u_i u_j) = \begin{cases} 0 \ para\ i \neq j \\ \sigma^2 \ para\ i = j \end{cases}}$$

Generación del modelo lineal por simulación Monte Carlo

Para comprender mejor el modelo de regresión a través de la visualización del comportamiento teórico bajo los supuestos anteriores, se construirá en una hoja de cálculo, una serie de simulaciones con una relación lineal entre dos variables, a las que se adicionará un término de error simulado, donde el objetivo es variar los parámetros de las variables y ver su impacto en el análisis.

Primero, en una tabla se deben agregar los valores propuestos de los parámetros α, β, n y σ^2. El tipo de valores de la variable X no está limitado a ser discreto o continuo, sin embargo en este apartado como tipo particular de variable se generará como una serie de k enteros aleatorios, los cuales se pueden acotar entre dos valores a y b $(a < b)$. La variable dependiente se simulará continua debido a la distribución del error, entonces se podría enunciar pero no limitar este modelo a algún contexto como por ejemplo el número de productos solicitados a un centro de distribución y el tiempo necesario para entregarlo a su destino.

La muestra de X se generará por medio de la fórmula empleada en el Capítulo 1 para generar una enteros uniformemente, la cual es ENTERO(ALEATORIO()*$(b - a)$+a). Para estas simulaciones se tomó $a = 10,\ b = 50$ .por otra parte, en otra columna adjunta a los valores de X, se coloca la fórmula de la recta $\alpha + \beta X_i$ sin sumar aún el término de error.

Luego se simulan los errores u_i como una variable $\boldsymbol{N(\mu = 0, \sigma^2)}$, empleando la técnica de la transformada inversa, en este caso calculada por la función INV.NORM(Probabilidad, Media, Desviación Estándar), la cual toma como parámetros un valor en (0,1) como Probabilidad al cual se desea acumular, introducida como un valor aleatorio en (0,1) por la función ALEATORIO(), además se introduce el valor de la media μ, que en este caso es igual a 0 y la desviación estándar σ, que se obtiene usando la función RAIZ(σ^2). Para completar la tabla, se agrega una columna con la fórmula de la simulación final como $\boldsymbol{Y_i = \alpha + \beta X_i + u_i}$.

Por último para visualizar los datos generados se inserta un diagrama de dispersión, agregando dos series: la primera compuesta de los puntos (X_i, Y_i) y la segunda serie con los puntos $\boldsymbol{(X_i, \alpha + \beta X_i)}$ Por ejemplo la siguiente imagen muestra una propuesta

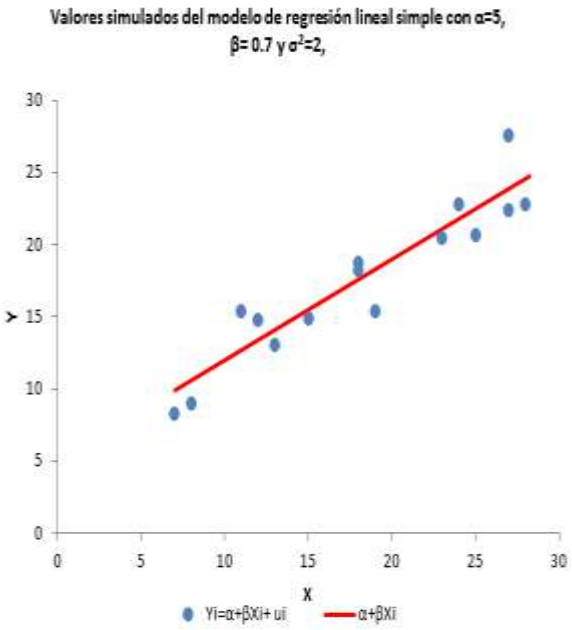
de estructura de hoja de cálculo para generar una muestra simulada de tamaño $n = 15$, seleccionando como parámetros iniciales $\boldsymbol{\alpha = 5, \beta = 0.7}$ y $\boldsymbol{\sigma^2 = 2}$

Propuesta en hoja de cálculo para generar el modelo de regresión lineal simple

Modelo de Regresión lineal simple

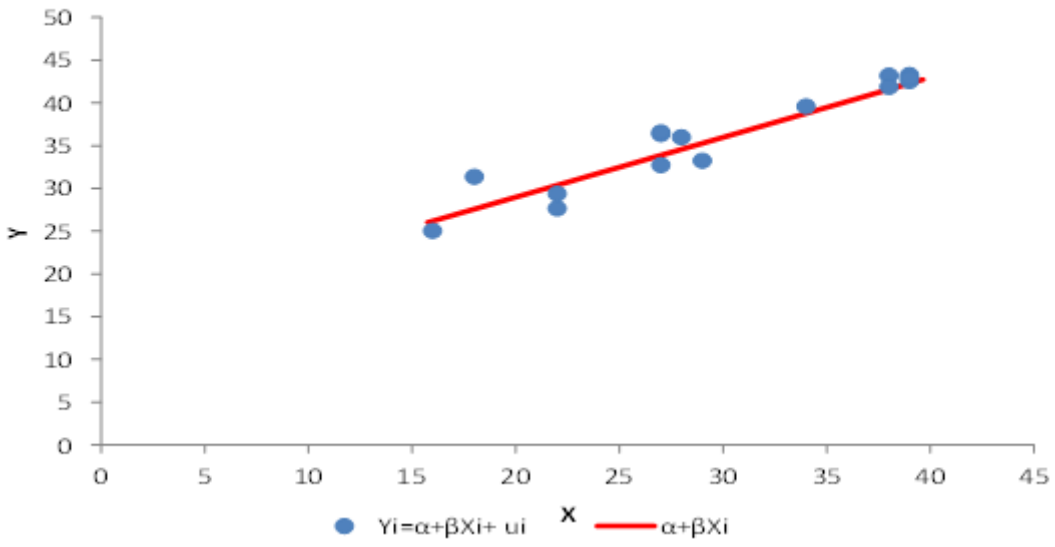
α=	5
β=	0.7
σ^2=	2

i	Xi	α+βXi	ui~N(0,σ^2)	Yi=α+βXi+ ui
1	15	15.5	-0.66	14.84
2	28	24.6	-1.84	22.76
3	23	21.1	-0.68	20.42
4	19	18.3	-2.92	15.38
5	12	13.4	1.29	14.69
6	8	10.6	-1.66	8.94
7	27	23.9	3.60	27.50
8	18	17.6	1.07	18.67
9	7	9.9	-1.64	8.26
10	18	17.6	0.63	18.23
11	13	14.1	-1.11	12.99
12	11	12.7	2.62	15.32
13	27	23.9	-1.54	22.36
14	24	21.8	0.93	22.73
15	25	22.5	-1.89	20.61



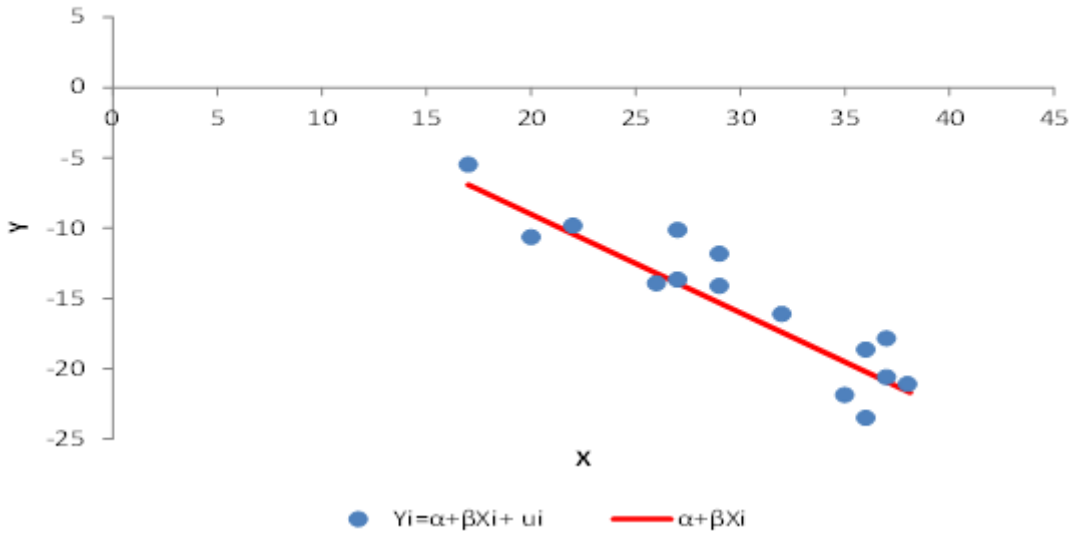
Una vez realizada esta estructura se pueden variar los parámetros α, β y σ^2 para comparar el impacto que tienen sobre el comportamiento de los datos simulados. Las gráficas siguientes muestran los escenarios en los cuales se cambiaron los valores de los parámetros de forma independiente, para poder compararlos contra el escenario generado inicialmente.

Gráfica 5.1: Valores simulados del modelo de regresión lineal simple con α=15, β= 0.7 y σ²=2,

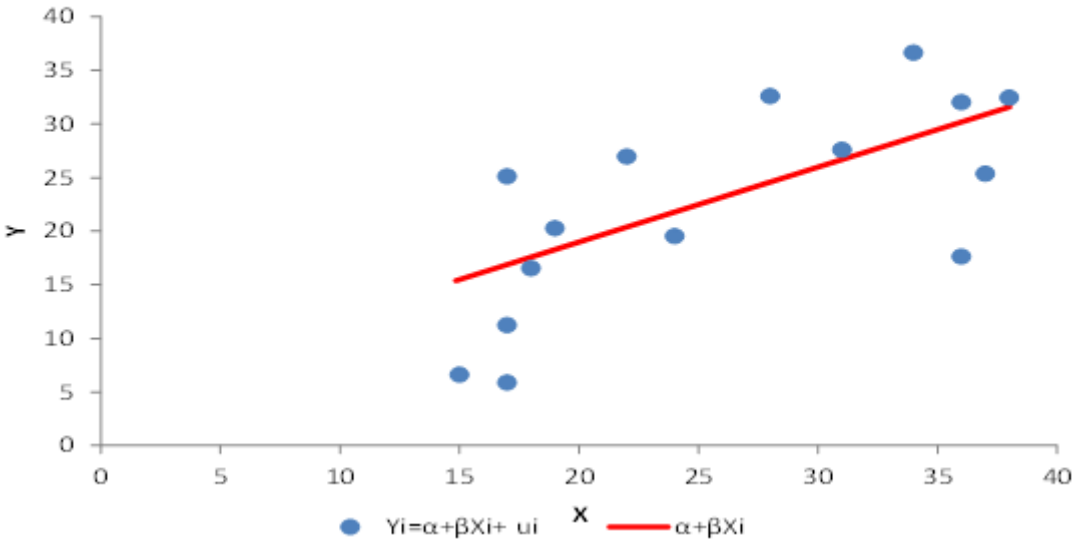


En este primer escenario se cambió el valor del parámetro α a 15, con lo cual el impacto observado es sobre la ordenada al origen, ya que ahora pasa por el punto $(0,15)$. En un segundo caso, se varió el signo de la pendiente β impactando en el sentido de la recta, sin embargo, se conservaron los parámetros $\alpha = 5$ y $\sigma^2 = 2$ por lo cual, la recta cruza en la ordenada en el mismo punto que el modelo inicial y también la separación de los puntos con respecto de la recta es similar. Por último, en la tercera gráfica se muestra el efecto de variar la varianza del error y conservar los parámetros $\alpha\ y\ \beta$ como en el escenario original, ya que la línea $\alpha X_i + \beta$ es la misma, sin embargo, la separación que muestran los valores simulados con respecto de la recta teórica es mucho mayor, lo cual se debe al incremento en el valor que se eligió para la varianza del error ($\sigma^2 = 25$) generando errores más dispersos.

Gráfica 5.2: Valores simulados del modelo de regresión lineal simple con α=5, β= -0.7 y σ²=2,



Gráfica 5.3: Valores simulados del modelo de regresión lineal simple con α=5, β= 0.7 y σ²=25



Desarrollo de los estimadores para α y β por mínimos cuadrados.

Con el modelo de regresión teórico construido con base en simulaciones, ahora lo pertinente es tener estimadores para los parámetros, que permitan tratar el modelo cuando se desconocen todos los parámetros. Como se vio en las simulaciones anteriores, el término de error causa en las observaciones una variación sobre su posición con respecto de la relación lineal original.

Cuando se tiene una muestra observada de n parejas (X_i, Y_i) y se desea ajustar los datos como una regresión lineal, es necesario estimar los valores de los parámetros desconocidos α y β con la información disponible. A los parámetros estimados se les denota como $\hat{\alpha}$ y $\hat{\beta}$ los cuales son sustituidos sobre la ecuación de la recta, para obtener las estimaciones de la variable dependiente denotadas como $\widehat{Y_i}$

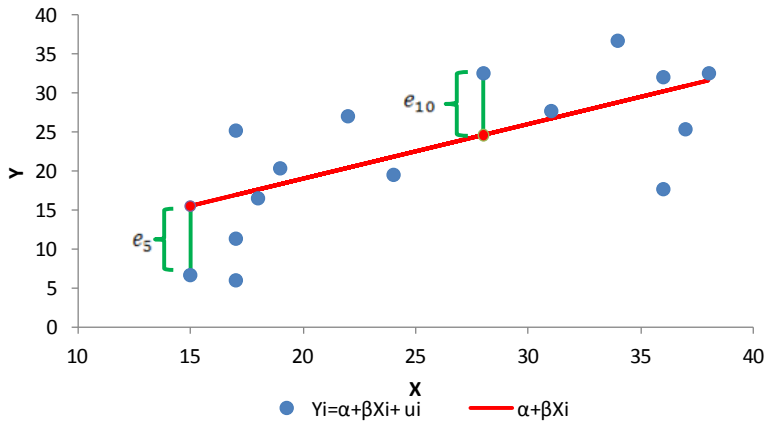
$$\widehat{Y_i} = \hat{\alpha} + \hat{\beta} X_i$$

Estos puntos generaran una recta estimada, la cual comparada contra los valores observados (X_i, Y_i) tendrán variaciones conocidas como los residuos del modelo, denotados como e_i. Estos residuos se calculan como la distancia vertical, del valor Y_i con respecto a la estimación $\widehat{Y_i}$

$$e_i = Y_i - \widehat{Y_i} = Y_i - (\hat{\alpha} + \hat{\beta} X_i)$$

Para mostrar visualmente las diferencias entre la relación lineal y los puntos generados, se puede utilizar la gráfica generada con las simulaciones generadas anteriormente; sobre la cual se pueden agregar una línea vertical que una al valor simulado (X_i, Y_i) con el valor de la línea $\alpha + \beta X_i$. Para lograr esto, se adicionan a la hoja de cálculo, una matriz auxiliar de 2X2 compuesta de una columna con el valor X_i repetido 2 veces, y otra columna con los valores $\alpha + \beta X_i$ y $\alpha + \beta X_i + U_i$ para alguna i correspondiente al residuo que se desee mostrar, y luego se inserta la matriz como una serie al gráfico de dispersión, lo que mostrará la línea deseada. En la Gráfica 5.4 se muestra el escenario simulado con mayor varianza, señalando dos diferencias, una que se encuentra por encima de la recta y otra que se posicionó por debajo.

Gráfica 5.4: Valores simulados del modelo de regresión lineal simple con α=5, β= 0.7 y σ²=25, con 2 residuales señalados



Una vez con el recurso gráfico se puede apelar, de cierta manera, a la imaginación para fijar los puntos azules, y pensar cuál sería el impacto sobre la distancia de los

puntos a la recta, al variar cualquiera de los parámetros del modelo, o puesto en otra perspectiva “mover” la recta, entonces los residuos también variarán en sus valores, unos disminuyendo por acortar su distancia hacia la recta, aunque en otros puntos esta distancia aumentará.

Entonces el objetivo para hallar los estimadores óptimos $\hat{\alpha}$ y $\hat{\beta}$ para los parámetros α y β será con respecto a minimizar la suma de cuadrados de los residuos, la cual se encuentra en términos de los estimadores de acuerdo a la definición del residuo, de donde también toma su nombre de mínimos cuadrados el método de estimación; por lo tanto el objetivo es

$$\boldsymbol{Min}_{\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}}\in\mathbb{R}}\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{e_i}^{\boldsymbol{2}}\right)$$

$$\boldsymbol{Min}_{\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}}\in\mathbb{R}}\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\left(\boldsymbol{Y_i}-\left(\widehat{\boldsymbol{\alpha}}+\widehat{\boldsymbol{\beta}}\boldsymbol{X_i}\right)\right)^{\boldsymbol{2}}\right)$$

Para resolverlo se utiliza la técnica de minimización de cálculo diferencial en varias variables, en la cual se hallan los puntos críticos de la función objetivo, mediante la derivación parcial con respecto de cada variable $\hat{\alpha}$ y $\hat{\beta}$ e igualando a 0 cada resultado. Para el caso de $\hat{\beta}$ se tiene

$$\frac{\boldsymbol{\partial}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{e_i}^{\boldsymbol{2}}}{\boldsymbol{\partial}\widehat{\boldsymbol{\beta}}}=\frac{\boldsymbol{\partial}}{\boldsymbol{\partial}\widehat{\boldsymbol{\beta}}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\left(\boldsymbol{Y_i}-\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\beta}}\boldsymbol{X_i}\right)^{\boldsymbol{2}}$$

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\frac{\boldsymbol{\partial}}{\boldsymbol{\partial}\widehat{\boldsymbol{\beta}}}\left(\boldsymbol{Y_i}-\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\beta}}\boldsymbol{X_i}\right)^{\boldsymbol{2}}=\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\frac{\boldsymbol{\partial}}{\boldsymbol{\partial}\widehat{\boldsymbol{\beta}}}\left(\boldsymbol{Y_i}-\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\beta}}\boldsymbol{X_i}\right)^{\boldsymbol{2}}$$

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}-\boldsymbol{2X_i}\left(\boldsymbol{Y_i}-\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\beta}}\boldsymbol{X_i}\right)=\boldsymbol{0}$$

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{X_iY_i}-\boldsymbol{X_i}\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\beta}}\boldsymbol{X_i}^{\boldsymbol{2}}=\boldsymbol{0}$$

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{X_iY_i}=\widehat{\boldsymbol{\alpha}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{X_i}+\widehat{\boldsymbol{\beta}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{X_i}^{\boldsymbol{2}}$$

Ahora tomando la derivada parcial con respecto a $\hat{\alpha}$ se tiene

$$\frac{\boldsymbol{\partial}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{e_i}^{\boldsymbol{2}}}{\boldsymbol{\partial}\widehat{\boldsymbol{\alpha}}}=\frac{\boldsymbol{\partial}}{\boldsymbol{\partial}\widehat{\boldsymbol{\alpha}}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\left(\boldsymbol{Y_i}-\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\beta}}\boldsymbol{X_i}\right)^{\boldsymbol{2}}$$

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\frac{\boldsymbol{\partial}}{\boldsymbol{\partial}\widehat{\boldsymbol{\beta}}}\left(\boldsymbol{Y_i}-\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\beta}}\boldsymbol{X_i}\right)^{\boldsymbol{2}}=\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}-\boldsymbol{2}\left(\boldsymbol{Y_i}-\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\beta}}\boldsymbol{X_i}\right)=\boldsymbol{0}$$

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{Y_i}-\boldsymbol{n}\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\beta}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{X_i}=\boldsymbol{0}$$

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{Y_i}=\boldsymbol{n}\widehat{\boldsymbol{\alpha}}+\widehat{\boldsymbol{\beta}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{X_i}$$

Como resultado de la derivación se obtiene un sistema de 2 ecuaciones con 2 incógnitas, conocido comúnmente como el sistema de ecuaciones normales

$$\sum_{i=1}^{n} \boldsymbol{X_i Y_i} = \widehat{\boldsymbol{\beta}} \sum_{i=1}^{n} \boldsymbol{X_i}^2 + \widehat{\boldsymbol{\alpha}} \sum_{i=1}^{n} \boldsymbol{X_i} \ldots\ldots (\mathbf{1})$$

$$\sum_{i=1}^{n} \boldsymbol{Y_i} = \widehat{\boldsymbol{\beta}} \sum_{i=1}^{n} \boldsymbol{X_i} + \widehat{\boldsymbol{\alpha}} \boldsymbol{n} \ldots\ldots (\mathbf{2})$$

Para resolverlo, se puede considerar el sistema en su representación matricial como

$$\begin{pmatrix} \sum_{i=1}^{n} \boldsymbol{X_i}^2 & \sum_{i=1}^{n} \boldsymbol{X_i} \\ \sum_{i=1}^{n} \boldsymbol{X_i} & \boldsymbol{n} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\alpha}} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} \boldsymbol{X_i Y_i} \\ \sum_{i=1}^{n} \boldsymbol{Y_i} \end{pmatrix}$$

Y posteriormente emplear el resultado de algebra lineal conocido como la regla de Cramer, la cual establece la solución al vector de variables $\begin{pmatrix} \hat{\beta} \\ \hat{\alpha} \end{pmatrix}$, como las siguientes fórmulas por determinantes

$$\widehat{\boldsymbol{\beta}} = \frac{\begin{vmatrix} \sum_{i=1}^{n} \boldsymbol{X_i Y_i} & \sum_{i=1}^{n} \boldsymbol{X_i} \\ \sum_{i=1}^{n} \boldsymbol{Y_i} & \boldsymbol{k} \end{vmatrix}}{\begin{vmatrix} \sum_{i=1}^{n} \boldsymbol{X_i}^2 & \sum_{i=1}^{n} \boldsymbol{X_i} \\ \sum_{i=1}^{n} \boldsymbol{X_i} & \boldsymbol{k} \end{vmatrix}} = \frac{\boldsymbol{n}(\sum_{i=1}^{n} \boldsymbol{X_i Y_i}) - (\sum_{i=1}^{n} \boldsymbol{Y_i})(\sum_{i=1}^{n} \boldsymbol{X_i})}{\boldsymbol{n} \sum_{i=1}^{n} \boldsymbol{X_i}^2 - (\sum_{i=1}^{n} \boldsymbol{X_i})^2}$$

$$\widehat{\boldsymbol{\alpha}} = \frac{\begin{vmatrix} \sum_{i=1}^{n} \boldsymbol{X_i}^2 & \sum_{i=1}^{n} \boldsymbol{X_i Y_i} \\ \sum_{i=1}^{n} \boldsymbol{X_i} & \boldsymbol{n} \sum_{i=1}^{n} \boldsymbol{Y_i} \end{vmatrix}}{\begin{vmatrix} \sum_{i=1}^{n} \boldsymbol{X_i}^2 & \sum_{i=1}^{n} \boldsymbol{X_i} \\ \sum_{i=1}^{n} \boldsymbol{X_i} & \boldsymbol{n} \end{vmatrix}} = \frac{(\sum_{i=1}^{n} \boldsymbol{X_i}^2)(\sum_{i=1}^{n} \boldsymbol{Y_i}) - (\sum_{i=1}^{n} \boldsymbol{X_i})(\sum_{i=1}^{n} \boldsymbol{X_i Y_i})}{\boldsymbol{n} \sum_{i=1}^{n} \boldsymbol{X_i}^2 - (\sum_{i=1}^{n} \boldsymbol{X_i})^2}$$

Sin embargo una expresión alternativa para $\hat{\alpha}$ se puede encontrar dividiendo la ecuación (2) entre n, de donde se obtiene

$$\overline{\boldsymbol{Y}} = \widehat{\boldsymbol{\beta}} \overline{\boldsymbol{X}} + \widehat{\boldsymbol{\alpha}}$$

$$\widehat{\boldsymbol{\alpha}} = \overline{\boldsymbol{Y}} - \widehat{\boldsymbol{\beta}} \overline{\boldsymbol{X}}$$

De esta forma además se encuentra que la recta estimada por mínimos cuadrados, pasa por el punto $(\bar{X}, \bar{Y})$

Ejemplo de cálculo y revisión de propiedades de los estimadores por mínimos cuadrados con muestras simuladas

Antes de proseguir con análisis del modelo de regresión lineal simple, los ejemplos de cálculo son muy útiles para reforzar la parte teórica del modelo. Para cumplir tal objetivo se utilizarán las simulaciones generadas anteriormente, mostrando la diferencia entre la recta inicial y la recta estimada con base en los estimadores por mínimos cuadrados.

El ejemplo en esta ocasión se realizó en una hoja de cálculo nueva, adjunta a la hoja de las simulaciones, para poder referenciar por fórmulas los parámetros del modelo simulado, junto con las columnas X_i y Y_i. Después se agregan otras dos columnas con los términos necesarios para emplear las fórmulas para los estimadores, con las operaciones $X_i * Y_i$ y ${X_i}^2$.

Con las columnas construidas, se obtiene después la suma total de cada una, para calcular los parámetros estimados con las formulas anteriormente obtenidas para $\hat{\alpha}$ y $\hat{\beta}$, las cuales se pueden pegar como imagen en la hoja de cálculo para generar un formulario de apoyo. Luego se computa la columna de los valores estimados como $\widehat{Y_i} = \hat{\alpha} + \hat{\beta} X_i$. Posteriormente se agrega a la gráfica de dispersión la serie con los puntos $(X_i, \widehat{Y_i})$ para comparar el resultado de trazar la línea con los parámetros estimados, contra la línea trazada en base a los parámetros propuestos.

A continuación se muestra una propuesta, tomando como valores para los parámetros $\alpha = 25$, $\beta = 1.2$ y $\sigma^2 = 15$ para una muestra de tamaño $n = 15$. Donde se puede observar que el valor del estimador $\hat{\beta} = 1.18$, mientras que el estimador $\hat{\alpha} = 24.35$, y la línea trazada en color verde representa los valores estimados $\widehat{Y_i}$.

Propuesta en hoja de cálculo para ilustrar el cálculo de los estimadores mínimos cuadrados del modelo de regresión lineal simple

Estimación de los parametros α y β

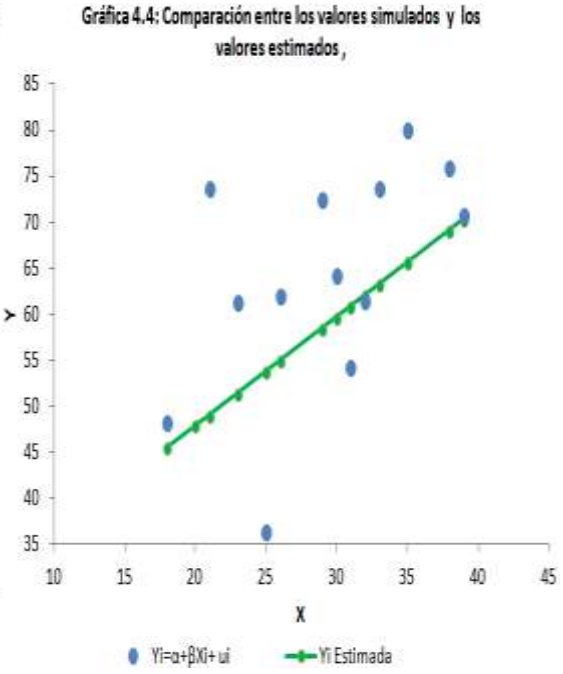
Parámetros Verdaderos			
α=	25	σ^2=	15
β=	1.2	n=	15

Parametros estimados	
$\hat{\beta}$ =	1.18
$\hat{\alpha}$ =	24.35

$$\hat{\beta} = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} Y_i \sum_{i=1}^{n} X_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

Xi	Yi	XiYi	Xi^2	Yi Estimada
38	75.91	2,884.6	1,444.0	69.16
39	70.86	2,763.4	1,521.0	70.34
23	61.33	1,410.5	529.0	51.47
31	54.30	1,683.4	961.0	60.91
26	62.16	1,616.1	676.0	55.01
32	61.49	1,967.7	1,024.0	62.09
18	48.26	868.6	324.0	45.58
25	36.40	910.0	625.0	53.83
30	64.26	1,927.8	900.0	59.73
35	80.12	2,804.0	1,225.0	65.63
29	72.51	2,102.8	841.0	58.55
21	73.80	1,549.8	441.0	49.11
20	16.19	323.8	400.0	47.93
33	73.68	2,431.3	1,089.0	63.27
35	26.98	944.4	1,225.0	65.63
Totales	**435**	**878.24**	**26,188.3**	**13,225.0**



Una vez hecha esta estructura se podrán generar muestras distintas, realizando algún cambio en los parámetros o actualizando la hoja de cálculo, lo que también actualizará el cálculo de los estimadores, y la gráfica. Como se podrá observar para cada muestra generada se tendrá un resultado distinto de los valores de los estimadores, ya que en sí cada estimador es una variable aleatoria, entonces como motivación para analizar a profundidad el comportamiento e impacto de los estimadores, se deben recopilar otras propiedades estadísticas como sus esperanzas y varianzas.

Estas propiedades se pueden hallar desarrolladas a detalle en diversas fuentes de literatura estadística, sin embargo, el desarrollo teórico vas más allá de este texto, con el objetivo de conservar la visión en el empleo de las simulaciones. Por lo tanto para el siguiente compendio, se consideraron como base tanto las notas de clase del profesor Francisco Sánchez (2015), como el libro de análisis de regresión de O. Rawlings (1998).

La primer propiedad de importancia de $\hat{\beta}$ y $\hat{\alpha}$, es sobre su comportamiento medio, expresado en su esperanza matemática, la cual es el valor de los parámetros en cuestión, por lo cual se consideran estimadores insesgados, es decir

$$\boldsymbol{E(\widehat{\beta}) = \beta}$$

$$\boldsymbol{E(\widehat{\alpha}) = \alpha}$$

En cuanto a la varianza de los estimadores, se calculan por medio de las siguientes expresiones, de las cuales se puede observar que la varianza del error tiene un impacto directo sobre la estimación de los parámetros.

$$\boldsymbol{Var(\widehat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

$$\boldsymbol{Var(\widehat{\alpha}) = \sigma^2\left(\frac{1}{n} + \left(\frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)\right)}$$

La covarianza entre los estimadores juega un papel importante en el análisis de regresión principalmente cuando se construyen modelos de más de dos dimensiones pues permite ver la relación entre variables independientes, ahora se presenta el cálculo en este caso con dos dimensione. La covarianza entre los estimadores, muestra cómo están correlacionados los estimadores de forma negativa, con una intensidad que depende de la varianza del error σ^2, además del comportamiento medio y dispersión de las variables X_i.

$$\boldsymbol{Cov(\widehat{\alpha}, \widehat{\beta}) = -\sigma^2 \frac{\bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

De acuerdo a las fórmulas anteriores, la varianza de los estimadores depende del valor de $\boldsymbol{\sigma^2}$, por lo cual es necesario estimar este parámetro y se calcula a partir de la suma de cuadrados de los residuos $\sum_{i=1}^{n} {e_i}^2$, pues es verificable que su esperanza matemática es $(n-2)\sigma^2$. Entonces el estimador de $\boldsymbol{\sigma^2}$ es

$$\widehat{\boldsymbol{\sigma^2}} = \frac{\sum_{i=1}^{n} \boldsymbol{e_i}^2}{\boldsymbol{n-2}}$$

Ahora, si se calcula la raíz cuadrada de $\widehat{\boldsymbol{\sigma^2}}$, se obtiene la estimación de la deviación estándar de $\boldsymbol{u}$ conocida como error estándar

$$\widehat{\boldsymbol{\sigma}} = \sqrt{\widehat{\boldsymbol{\sigma^2}}} = \sqrt{\frac{\sum_{i=1}^{n} \boldsymbol{e_i}^2}{\boldsymbol{n-2}}}$$

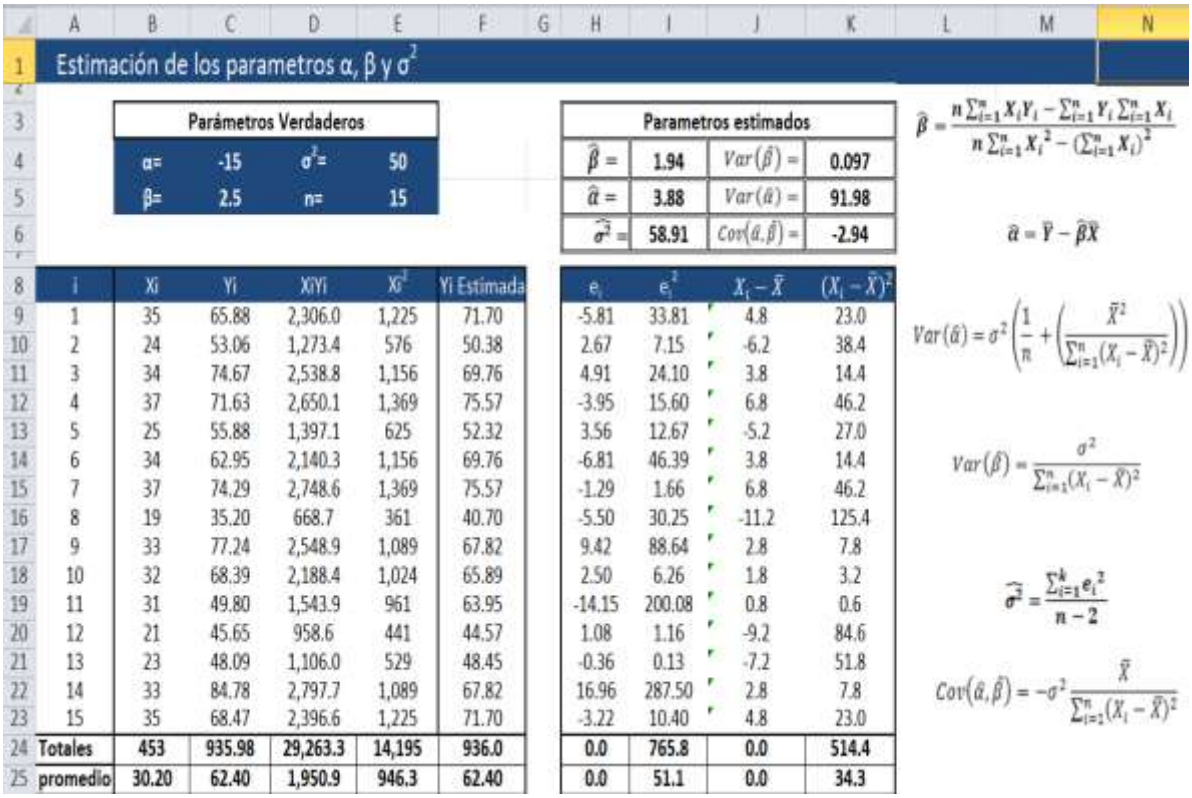
Para calcular las cantidades anteriores dentro de la hoja de cálculo de trabajo, primero se deben agregar, las columnas e_i, $e_i{}^2, X_i - \bar{X}, y\ (X_i - \bar{X})^2$ calculando al final las sumas totales de cada columna, junto con un renglón extra con los promedios de las columnas, los cuales son necesarios para computar las varianzas, covarianza entre los estimadores, y el estimador $\widehat{\boldsymbol{\sigma^2}}$. Lo anterior, con la intención de desglosar el cálculo y entrar en el detalle de identificar los puntos que tienen una mayor aportación a la variabilidad del parámetro estimado.

Se continúa con el ejemplo de cálculo que se ha simulado para mostrar el modelo de regresión simple, donde se agregó el cálculo de la varianza de los estimadores y del error. Ahora se considera el escenario donde los valores de los parámetros son $\alpha = -15$, $\beta = 2.5$, además se selecciona el valor del parámetro $\sigma^2 = 50$, conservando el tamaño de muestra $n = 15$.

En la siguiente imagen se muestra la estructura, anterior adicionando los cálculos mencionados donde se puede observar que el valor del estimador $\widehat{\boldsymbol{\alpha}} = \mathbf{3.88}$, que se encuentra distante del valor del parámetro definido por casi 20 unidades a comparación del escenario anterior. Lo anterior se justifica por el valor de $\boldsymbol{Var}(\widehat{\boldsymbol{\alpha}}) = \mathbf{91.98}$ pues implica una desviación estándar del estimador $\widehat{\boldsymbol{\alpha}}$ de 9.5 unidades, al reproducir el ejemplo se podrá observar que al generar muestras consecutivas el valor de $\widehat{\boldsymbol{\alpha}}$ tendrá variaciones importantes y cuando se reduzca el valor del parámetro σ^2 se reducirá esta variabilidad.

Por otro lado, el estimador $\widehat{\boldsymbol{\beta}} = \mathbf{1.94}$ el cual es más cercano al valor del parámetro debido a que el valor de su varianza es mucho menor al de $\boldsymbol{Var}(\widehat{\boldsymbol{\alpha}})$ con un valor de $\boldsymbol{Var}(\widehat{\boldsymbol{\beta}}) = \mathbf{0.09}$.

Estructura en Excel® como ejemplo de cálculo de los estimadores del modelo de regresión



Estimación de los parametros α, β y σ^2

Parámetros Verdaderos			
α=	-15	σ^2=	50
β=	2.5	n=	15

Parametros estimados			
$\hat{\beta}$ =	1.94	$Var(\hat{\beta})$ =	0.097
$\hat{\alpha}$ =	3.88	$Var(\hat{\alpha})$ =	91.98
$\hat{\sigma^2}$ =	58.91	$Cov(\hat{\alpha},\hat{\beta})$ =	-2.94

i	Xi	Yi	XiYi	Xi^2	Yi Estimada	e_i	e_i^2	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
1	35	65.88	2,306.0	1,225	71.70	-5.81	33.81	4.8	23.0
2	24	53.06	1,273.4	576	50.38	2.67	7.15	-6.2	38.4
3	34	74.67	2,538.8	1,156	69.76	4.91	24.10	3.8	14.4
4	37	71.63	2,650.1	1,369	75.57	-3.95	15.60	6.8	46.2
5	25	55.88	1,397.1	625	52.32	3.56	12.67	-5.2	27.0
6	34	62.95	2,140.3	1,156	69.76	-6.81	46.39	3.8	14.4
7	37	74.29	2,748.6	1,369	75.57	-1.29	1.66	6.8	46.2
8	19	35.20	668.7	361	40.70	-5.50	30.25	-11.2	125.4
9	33	77.24	2,548.9	1,089	67.82	9.42	88.64	2.8	7.8
10	32	68.39	2,188.4	1,024	65.89	2.50	6.26	1.8	3.2
11	31	49.80	1,543.9	961	63.95	-14.15	200.08	0.8	0.6
12	21	45.65	958.6	441	44.57	1.08	1.16	-9.2	84.6
13	23	48.09	1,106.0	529	48.45	-0.36	0.13	-7.2	51.8
14	33	84.78	2,797.7	1,089	67.82	16.96	287.50	2.8	7.8
15	35	68.47	2,396.6	1,225	71.70	-3.22	10.40	4.8	23.0
Totales	453	935.98	29,263.3	14,195	936.0	0.0	765.8	0.0	514.4
promedio	30.20	62.40	1,950.9	946.3	62.40	0.0	51.1	0.0	34.3

$$\hat{\beta} = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} Y_i \sum_{i=1}^{n} X_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

$$Var(\hat{\alpha}) = \sigma^2\left(\frac{1}{n} + \left(\frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)\right)$$

$$Var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

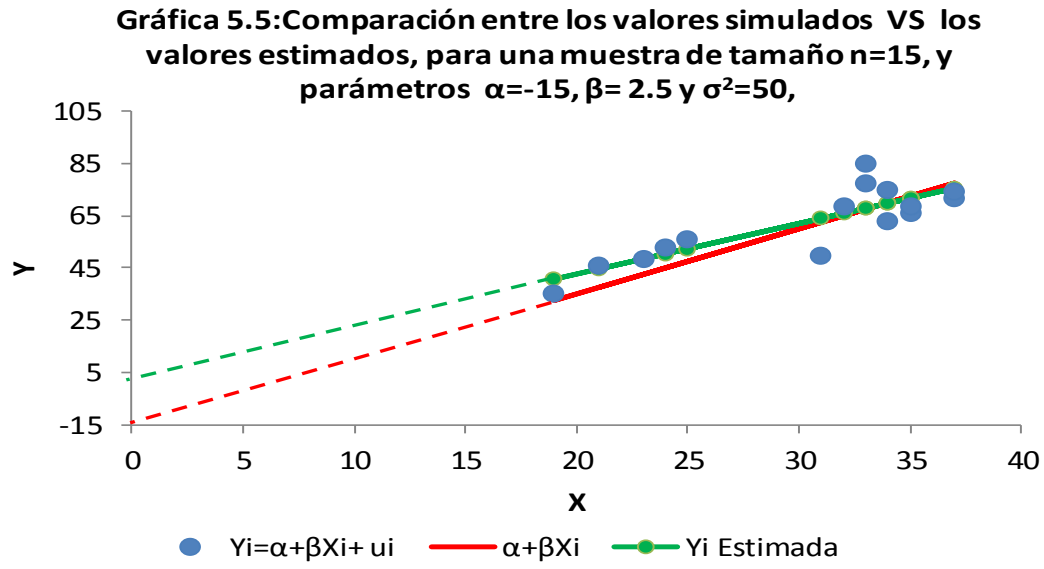
$$\hat{\sigma^2} = \frac{\sum_{i=1}^{k} e_i^2}{n-2}$$

$$Cov(\hat{\alpha},\hat{\beta}) = -\sigma^2 \frac{\bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Como se puede observar en la propuesta, se colocaron las fórmulas de apoyo a un costado para facilitar el seguimiento de los cálculos y reforzar su aprendizaje; además, se dividieron las operaciones en dos secciones para la estimación de la recta y otra para el cálculo de las columnas de apoyo para computar la varianza y covarianza de los estimadores, donde se puede observar el detalle de las desviaciones e identificar que el punto del renglón 14 es el que tuvo una desviación mayor con respecto a su estimación y por ende aportó más a la suma de cuadrados de los residuos.

En la Gráfica 5.5, se encuentra plasmado los resultados ejemplo anterior, donde se observa la similitud de las pendientes entre la relación lineal original y la recta con los valores estimados por mínimos cuadrados. De esta forma se puede apreciar visualmente cómo el estimador del parámetro de la ordenada al origen es más sensible, explicado por los valores de X_i simulados y que tan lejos se encuentran del punto $X = 0$, pues la fórmula de la varianza de $\hat{\alpha}$ esta directamente impacta por el valor de $\bar{X}$.

El valor del gráfico es mayor cuando se generan muestras sucesivas al actualizar la hoja de cálculo, lo que permitirá observar diversas muestras; además de mostrar cómo los estimadores variaran en sus valores y en consecuencia la recta estimada, por esta razón corroborar las propiedades de los estimadores basados en estas simulaciones, producirá una idea intuitiva sobre su impacto.



La primera propiedad a revisar será el insesgamiento de los estimadores. Lo anterior se puede comprobar por medio de simular una serie de muestras y calcular el valor de los estimadores. Luego al calcular la media aritmética de los estimadores se podrá identificar como se aproxima al valor del parámetro, conforme se incremente el número de estimaciones consideradas en el cálculo de la media.

Para construir en hoja de cálculo, la aproximación de la media de cada estimador hacia el valor del parámetro, se debe colocar en alguna parte libre de la hoja de cálculo o en una hoja nueva, las referencias a las celdas donde se hallan los cálculos de los estimadores. Posteriormente se copia el valor del estimador y se pega a valores debajo de la celda de referencia, con lo cual se fija el valor para la muestra actual y a su vez se genera una nueva muestra, de tal manera que al repetir el proceso se tendrá una serie de simulaciones para el valor de los estimadores. Para simplificar el proceso se puede aprovechar de una rutina en lenguaje macro que permita realizar el copiado a valores de una manera y automatizada[26].

Para este trabajo se consideró tomar un tamaño de muestra para los estimadores, repitiendo el proceso 1000 veces. El siguiente paso es calcular una columna con las medias calculadas con los valores de los estimadores, pero partiendo de considerar en el cálculo una observación y añadiendo una observación al cálculo de forma sucesiva, generando una columna de medias, por ejemplo para las realizaciones de $\hat{\beta}$, se toma como media la primera observación del estimador, en el siguiente renglón se calcula la media entre las dos primeras observaciones, en el tercer renglón la media de las primeras tres observaciones, y así de manera sucesiva hasta llegar al último renglón donde se calculará la media de todas las realizaciones.

Esta estructura tiene como objetivo graficar las columnas que contienen las medias sucesivas, la cuales se encuentran a continuación, donde se puede observar la convergencia de la media de $\hat{\beta}$ tiende a su verdadero valor de 2.5 al igual que la media de $\hat{\alpha}$ tiende su valor elegido $\alpha = -15$ y la media de $\widehat{\sigma^2}$ converge a 50.

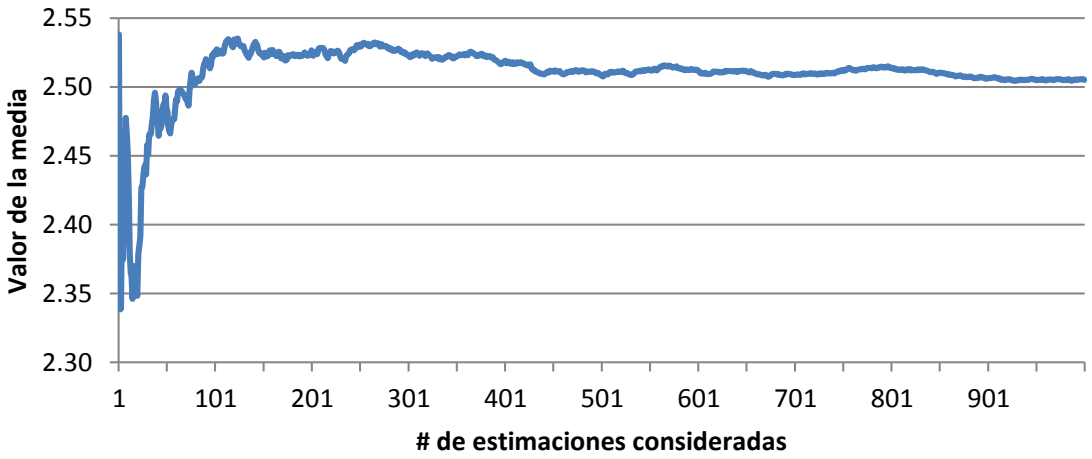
[26] Consulte el código completo en el apéndice.

La siguiente imagen de pantalla muestra, la propuesta para realizar los cálculos anteriores para los estimadores $\hat{\alpha}$, $\hat{\beta}$ y $\widehat{\sigma^2}$, donde se muestra el cálculo de una media intermedia con la función PROMEDIO(), considerando las observaciones correspondientes.

Estructura en Excel® para comprobar el insesgamiento de los estimadores

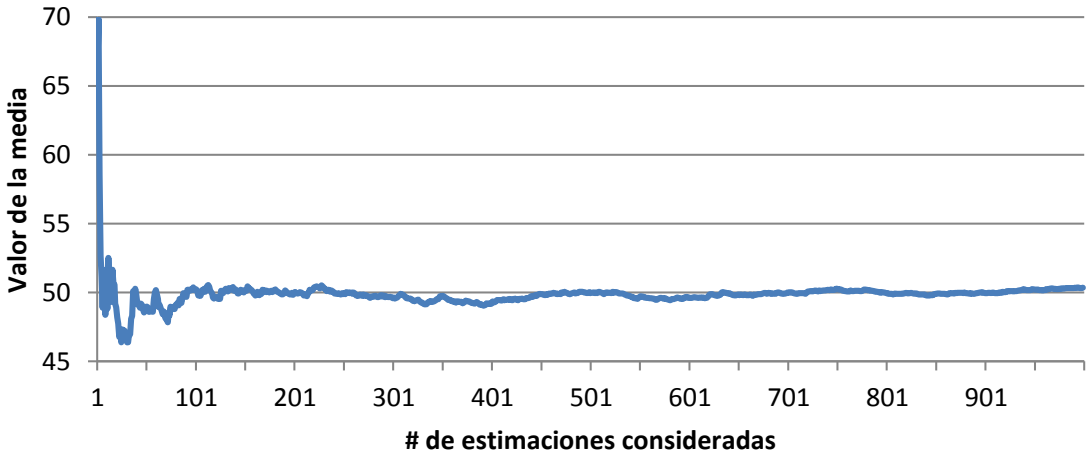
	O	P	Q	R	S	T	U	V
1		Insesgamiento de los estimadores						
3								
4		Estimadores						
5		$\hat{\beta}$ =	$\hat{\alpha}$ =	$\widehat{\sigma^2}$ =				
6		1.94	3.88	58.91				
8		j	Estimaciones copiadas a valor				Medias de los estimadores	
9		1	2.54	-16.08	49.59		2.54	-16.08	49.59
10		2	2.14	-6.83	83.41		2.34	-11.46	66.50
11		3	2.34	-10.27	76.65		2.34	-11.06	69.89
12		4	2.63	-18.54	24.66		2.41	-12.93	58.58
13		5	2.23	-8.14	29.05		2.37	-11.97	52.67
14		6	2.52	-17.52	43.68		2.40	-12.90	51.17
15		7	2.71	-17.69	35.30				48.91
16		8	2.72	-21.02	70.99		=PROMEDIO(P$9:P16)		51.67
17		9	2.37	-11.56	36.22		2.47	-14.18	49.95
18		10	2.32	-14.74	34.37		2.45	-14.24	48.39
19		11	2.12	-4.37	69.78		2.42	-13.34	50.34
20		12	1.90	1.55	32.14		2.38	-12.10	48.82
21		13	2.21	-6.70	96.79		2.36	-11.69	52.51
22		14	2.35	-8.58	24.15		2.36	-11.46	50.49
23		15	2.11	-5.04	37.98		2.35	-11.04	49.65
24		16	2.73	-18.52	44.44		2.37	-11.50	49.33
25		17	2.11	-2.99	88.97		2.35	-11.00	51.66

Gráfica 5.6: Convergencia del valor medio de las realizaciones del estimador de β



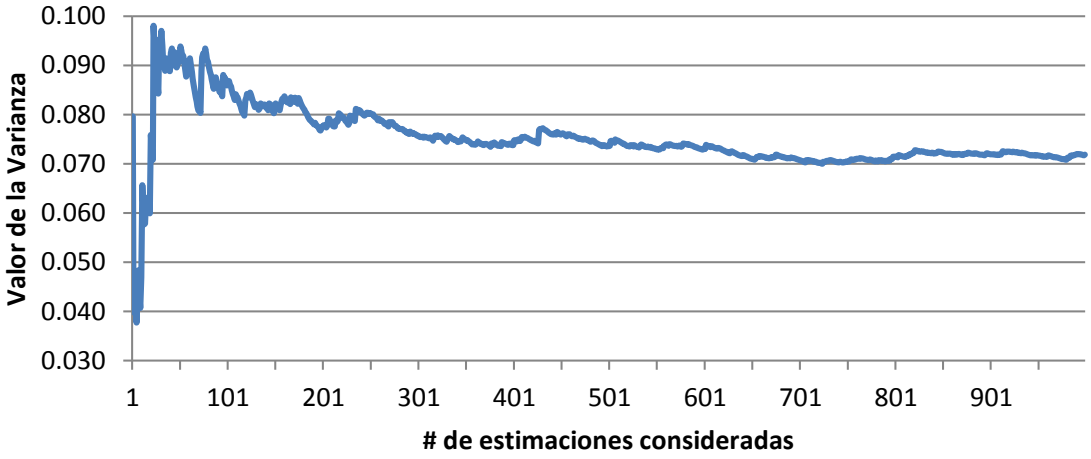
Gráfica 5.7: Convergencia del valor medio de las realizaciones del estimador de α

Gráfica 5.8: Convergencia del valor medio de las realizaciones del estimador de σ^2

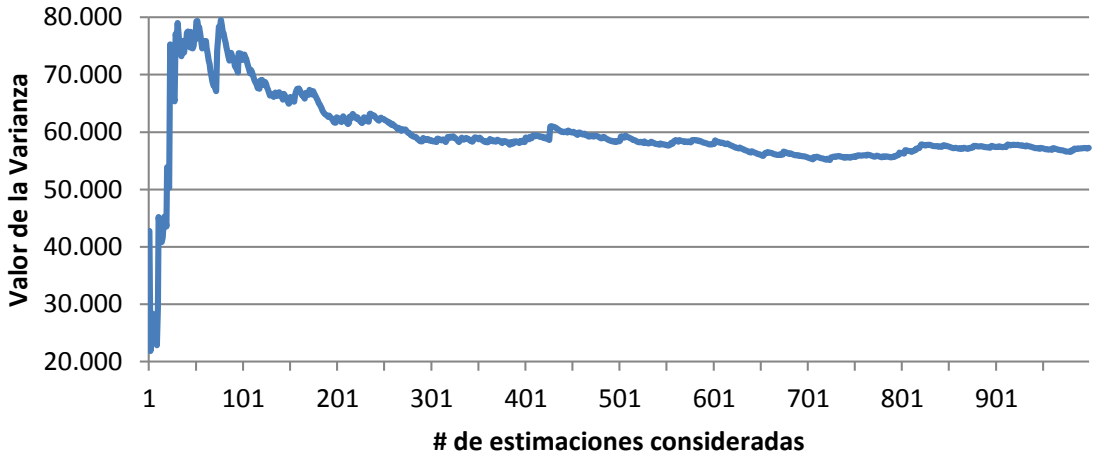


Una vez comprobado el insesgamiento de los estimadores se puede comprobar ahora el cálculo hecho para la varianza de los estimadores. Esto se puede realizar con la misma idea que la última estructura generada, ya que se debe calcular la varianza de los valores generados de los estimadores considerando análogamente un mayor número de observaciones de los estimadores, las cuales teóricamente apuntarían o estarían cercanos a las estimaciones $Var(\hat{\alpha}) = 91.98$ y $Var(\hat{\beta}) = 0.097$. La forma de realizarlo en sobre las simulaciones es con la función VAR(), que computa la varianza muestral del rango que se introduzca. Las siguientes graficas muestran la convergencia de las varianzas para cada estimador:

Gráfica 5.9: Convergencia de la varianza de las realizaciones del estimador de β



Gráfica 5.10: Convergencia de la varianza de las realizaciones del estimador de α



Con las comprobaciones anteriores, además de corroborar los desarrollos teóricos con datos prácticos, se puede comprender la naturaleza aleatoria de los parámetros, además de su sensibilidad, por lo cual sus valores también tienen una variación de acuerdo a la naturaleza de los datos. Bajo esta motivación de hallar los límites para la variabilidad en los estimadores, se verá en la sección siguiente la estimación por medio de intervalos de confianza.

Pruebas de hipótesis e intervalos de confianza para los estimadores $\widehat{\alpha}$,$\widehat{\beta}$ y $\widehat{\sigma^2}$

A pesar de construir los estimadores óptimos para α, β y para σ^2, como se ha visto la carga de incertidumbre del error, influye en que los estimadores varíen en su precisión, además al poder tomar cualquier valor en los números reales, los parámetros estimados se consideran variables aleatorias continuas, por lo cual la probabilidad que el estimador sea un valor en específico es 0; esto genera la motivación para construir intervalos de confianza para los parámetros, los cuales además sirven como previo en la construcción de pruebas de hipótesis sobre los valores de α, β y σ^2 .

Una propiedad verificable de los estimadores $\hat{\beta}$ y $\hat{\alpha}$, es que se pueden expresar como una combinación lineal de los errores u_i, entonces por el supuesto de normalidad de los errores y a propiedades de la distribución Normal ambos estimadores poseen una distribución Normal. Como se tienen las expresiones para la esperanza y varianza de los estimadores entonces las distribuciones de los estimadores se encuentran especificadas como

$$\widehat{\boldsymbol{\beta}} \sim \boldsymbol{N}\left(\boldsymbol{\beta}, \frac{\boldsymbol{\sigma^2}}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{X_i} - \overline{\boldsymbol{X}})^{\boldsymbol{2}}}\right)$$

$$\widehat{\boldsymbol{\alpha}} \sim \boldsymbol{N}\left(\boldsymbol{\alpha}\ ,\ \boldsymbol{\sigma^2}\left(\frac{\boldsymbol{1}}{\boldsymbol{n}} + \left(\frac{\overline{\boldsymbol{X}}^{\boldsymbol{2}}}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{X_i} - \overline{\boldsymbol{X}})^{\boldsymbol{2}}}\right)\right)\right)$$

Para eliminar la dependencia del parámetro $\boldsymbol{\sigma^2}$ se emplea el siguiente estadístico basado en el estimador $\widehat{\boldsymbol{\sigma^2}}$, del cual se puede verificar que su distribución es Ji-cuadrada con $n-2$ grados de libertad.

$$\frac{(\boldsymbol{n-2})\widehat{\boldsymbol{\sigma^2}}}{\boldsymbol{\sigma^2}} \sim \boldsymbol{\chi^2_{n-2}}$$

Luego para construir los intervalos de confianza, se estandarizará las distribuciones de los estimadores para obtener Normal, $N(0,1)$ y calcular el cociente entre la Ji-Cuadrada anterior, para obtener una variable t de *Student* con $n-2$ grados de libertad, pues estas variables son independientes. Con base en lo anterior para construir un intervalo de confianza sobre el parámetro α, se estandarizará la variable para conseguir una $N(0,1)$ de la siguiente manera

$$\frac{\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}}{\sqrt{\boldsymbol{\sigma^2}\left(\frac{\boldsymbol{1}}{\boldsymbol{n}} + \left(\frac{\overline{\boldsymbol{X}}^{\boldsymbol{2}}}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{X_i} - \overline{\boldsymbol{X}})^{\boldsymbol{2}}}\right)\right)}} \sim \boldsymbol{N(0,1)}$$

Entonces la variable t de *Student* se calcula por medio del siguiente cociente

$$\frac{\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}}{\sqrt{\boldsymbol{\sigma^2}\left(\frac{\boldsymbol{1}}{\boldsymbol{n}} + \left(\frac{\overline{\boldsymbol{X}}^{\boldsymbol{2}}}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{X_i} - \overline{\boldsymbol{X}})^{\boldsymbol{2}}}\right)\right)}} \Big/ \sqrt{\frac{(\boldsymbol{n-2})\widehat{\boldsymbol{\sigma^2}}}{\boldsymbol{\sigma^2(n-2)}}}$$

$$= \frac{\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}}{\sqrt{\widehat{\boldsymbol{\sigma^2}}\left(\frac{\boldsymbol{1}}{\boldsymbol{n}} + \left(\frac{\overline{\boldsymbol{X}}^{\boldsymbol{2}}}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{X_i} - \overline{\boldsymbol{X}})^{\boldsymbol{2}}}\right)\right)}} \sim \boldsymbol{t_{n-2}}$$

Con esta variable se define la probabilidad de contener al estadístico en un intervalo como $1-\delta$

$$\boldsymbol{P}\left(\boldsymbol{t}_{\boldsymbol{n-2},\frac{\boldsymbol{\delta}}{\boldsymbol{2}}} < \frac{\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}}{\sqrt{\widehat{\boldsymbol{\sigma^2}}\left(\frac{\boldsymbol{1}}{\boldsymbol{n}}+\left(\frac{\bar{\boldsymbol{X}}^{\boldsymbol{2}}}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{X_i}-\bar{\boldsymbol{X}})^{\boldsymbol{2}}}\right)\right)}} < \boldsymbol{t}_{\boldsymbol{n-2},\boldsymbol{1-\frac{\delta}{2}}}\right) = \boldsymbol{1-\delta}$$

Donde $t_{n-2,\frac{\delta}{2}}$ es el cuantil de la distribución t con $n-2$ grados de libertad, que acumula una probabilidad de $\delta/2$. Por simetría la distribución t se cumple $t_{n-2,\frac{\delta}{2}} = -t_{n-2,1-\frac{\delta}{2}}$ [27], así que substituyendo se puede despejar de la desigualdad el parámetro α con lo cual se obtiene el intervalo de confianza que contiene el verdadero valor de $\boldsymbol{\alpha}$ con un nivel de confianza del $(1-\delta)*100\%$ como:

$$\boldsymbol{\alpha} \in \left(\widehat{\boldsymbol{\alpha}} \pm \sqrt{\widehat{\boldsymbol{\sigma^2}}\left(\frac{\boldsymbol{1}}{\boldsymbol{n}}+\left(\frac{\bar{\boldsymbol{X}}^{\boldsymbol{2}}}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{X_i}-\bar{\boldsymbol{X}})^{\boldsymbol{2}}}\right)\right)}\,\boldsymbol{t}_{\boldsymbol{n-2},\boldsymbol{1-\frac{\delta}{2}}}\right)$$

Intervalos de confianza para α vía Simulación Monte Carlo

Con el objetivo de comprobar el comportamiento y dar ejemplos de cálculo para el intervalo de confianza sobre el parámetro α, se va a aprovechar la estructura de simulación generada anteriormente, en la cual se tiene el cálculo de los estimadores.

Se reproducirá el mismo método visto en el Capítulo III que calcula el intervalo de confianza y copia los resultados de una condicional por medio de la fórmula **SI(LINF< α,SI(α<LSUP,"SI","NO"))** la cual permite saber de forma inmediata si el intervalo calculado contiene al verdadero valor del parámetro α. Donde LINF es la celda donde se computó el límite inferior y LSUP la celda donde se encuentra el límite superior del intervalo y α se refiere a la celda donde se halla la celda con el valor real del parámetro α.

Los cálculos anteriores, al ser fórmulas que referencian otras celdas, se verán impactados por la generación de diversas muestras, pues la fórmulas se actualizarán a la par de las nuevas generaciones de números aleatorios por parte de Excel®. A partir de este punto al copiar a valor los límites inferiores y superiores, junto con la estimación y la fórmula que determina si el intervalo cubrió al verdadero valor del parámetro, se obtendrán realizaciones simuladas del intervalo de confianza.

Con el empleo de una rutina en procesos macro se puede automatizar el proceso de generación de muestras[28]. Para este caso se consideró cambiar el escenario a los valores de los parámetros $\alpha = 10, \beta = -0.5$ y $\sigma^2 = 100$ de donde se generaron 1000 simulaciones con el proceso. Una vez completada la generación de la muestra de intervalos, se calculó el conteo de las veces que el intervalo cubrió al valor real del parámetro y el porcentaje que representa de la muestra como una tasa de cobertura.

[27] Para la referencia visual vea la Gráfica 3.3 del capítulo III.
[28] Consulte el código de la rutina en el apéndice.

La siguiente imagen muestra la estructura propuesta para generar las realizaciones de los intervalos de confianza en hoja de cálculo, donde se puede apreciar que los primeros 3 intervalos contienen al parámetro, mientras que en el cuarto intervalo generado el límite inferior es 10.2, siendo mayor que el valor del parámetro $\alpha = 10$, lo cual se refleja en el resultado de la condicional. Del total de las 1000 simulaciones se cubrió a α en 948 ocasiones por lo intervalos generados, por lo cual la tasa de cobertura es del 94.8%, el cual en efecto, es casi el valor de la confianza definida para estos intervalos

Estructura en Excel® para generar intervalos de confianza sobre α

Intervalos de confianza para los parametros α y β

Parámetros Verdaderos			
α=	10	σ^2=	100
β=	-0.5	n=	15

Parámetros del intervalo	
δ	5%
$t_{n-2,1-\delta}$	2.16

$$\hat{\alpha} \pm \sqrt{\widehat{\sigma^2}\left(\frac{1}{n} + \left(\frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2}\right)\right)} t_{n-2,1-\frac{\delta}{2}}$$

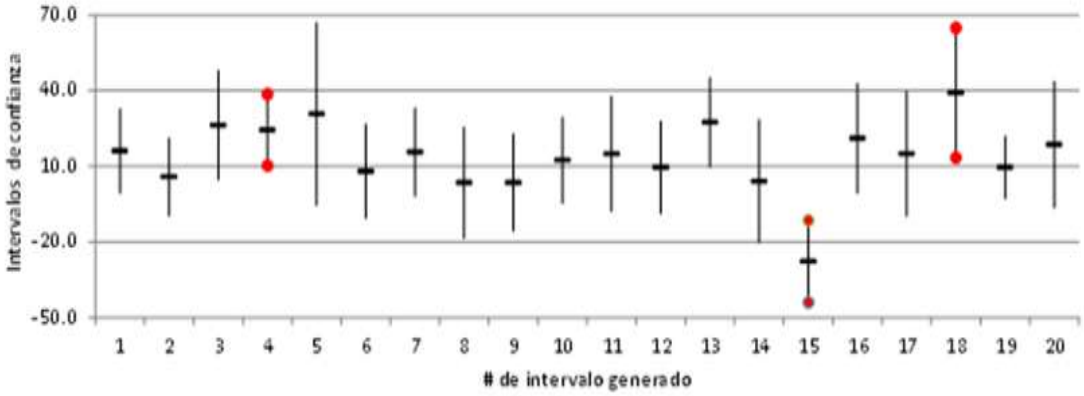
limite inferior	$\hat{\alpha}$ =	limite superior	¿contiene al parametro?
-16.64	9.45	35.55	SI

# de SI	% de SI
948.00	95%

# muestra simulada	Intervalos copiados a valor			
1	-0.6	16.0	32.7	SI
2	-9.7	5.8	21.2	SI
3	4.7	26.3	47.9	SI
4	10.2	24.4	38.6	NO
5	-5.4	30.8	67.0	SI
6	-10.6	8.0	26.7	SI
7	-1.8	15.6	33.0	SI
8	-18.6	3.5	25.5	SI
9	-15.7	3.5	22.7	SI
10	-4.7	12.3	29.4	SI
11	-7.9	14.9	37.7	SI
12	-8.7	9.6	27.9	SI

Para visualizar el impacto en cuanto a la posición del intervalo por los valores de $\boldsymbol{\alpha}$ como en su longitud debido a los diferentes valores de la estimación de la varianza del error $\widehat{\sigma^2}$, se agrega un gráfico de máximos y mínimos incluido en las opciones de Excel®, que coloca los intervalos de forma vertical como lo muestra la gráfica 5.12, donde se pueden observar señalados en color rojo los intervalos que no cubrieron al parámetro, aunque de acuerdo a los resultados obtenidos en estos 20 intervalos se esperarían que uno de ellos (1/20=0.05) no contuviera al parámetro aunque en esta ventana de observación en particular se observan 3.

Gráfica 5.12: Primeros Intervalos de confianza de 1000 muestras para el parametro α



Ahora, para una mejor comprensión del nivel de confianza se variaron el número de intervalos simulados para calcular el número de intervalos que contienen al parámetro y el porcentaje que representa del total de simulaciones. Los escenarios resumidos se hallan en la Tabla 5.1 con el número de intervalos generados, su total de intervalos efectivos y la tasa en porcentaje que representa de la muestra, donde se puede ver reflejada la confianza de los intervalos.

Tabla 5.1: Resumen de simulaciones de intervalos de confianza para α

Intervalos Generados	# de intervalos que encerraron α	% de intervalos que encerraron α
500	469	93.80%
1,000	948	94.89%
3,000	2,850	95.00%
5,500	5,184	94.25%
10,000	9,418	94.18%

Intervalos de confianza para $\boldsymbol{\beta}$

En el caso del intervalo de confianza para $\boldsymbol{\beta}$, los procedimientos para determinar el intervalo y comprobar sus comportamientos se harán de forma análoga que para el parámetro α. Primero se debe estandarizar la variable que define el estimador por medio de la transformación

$$\frac{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}} \sim \boldsymbol{N(0,1)}$$

Después al dividir entre el estadístico basado en el estimador de la varianza $\widehat{\sigma^2}$, que se distribuye Ji-cuadrada, con el cual se obtiene la siguiente variable t de *Student*

$$\frac{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}} \Big/ \sqrt{\frac{(\boldsymbol{n-2})\widehat{\boldsymbol{\sigma^2}}}{\boldsymbol{\sigma^2(n-2)}}}$$

$$= \frac{\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}}{\sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}} \sim \boldsymbol{t_{n-2}}$$

De forma análoga al intervalo para α, se define el valor δ con el cual se define la probabilidad que cubra a la variable t y acumule $(1-\delta)*100\%$ de confianza, a partir de la cual se despeja el parámetro $\boldsymbol{\beta}$, obteniéndose el siguiente intervalo de confianza

$$\boldsymbol{\beta} \in \left(\widehat{\boldsymbol{\beta}} \pm \boldsymbol{t}_{\boldsymbol{n-2,1-\frac{\delta}{2}}} \left(\sqrt{\frac{\widehat{\boldsymbol{\sigma^2}}}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \left(\boldsymbol{X_i} - \bar{\boldsymbol{X}} \right)^2}} \right) \right)$$

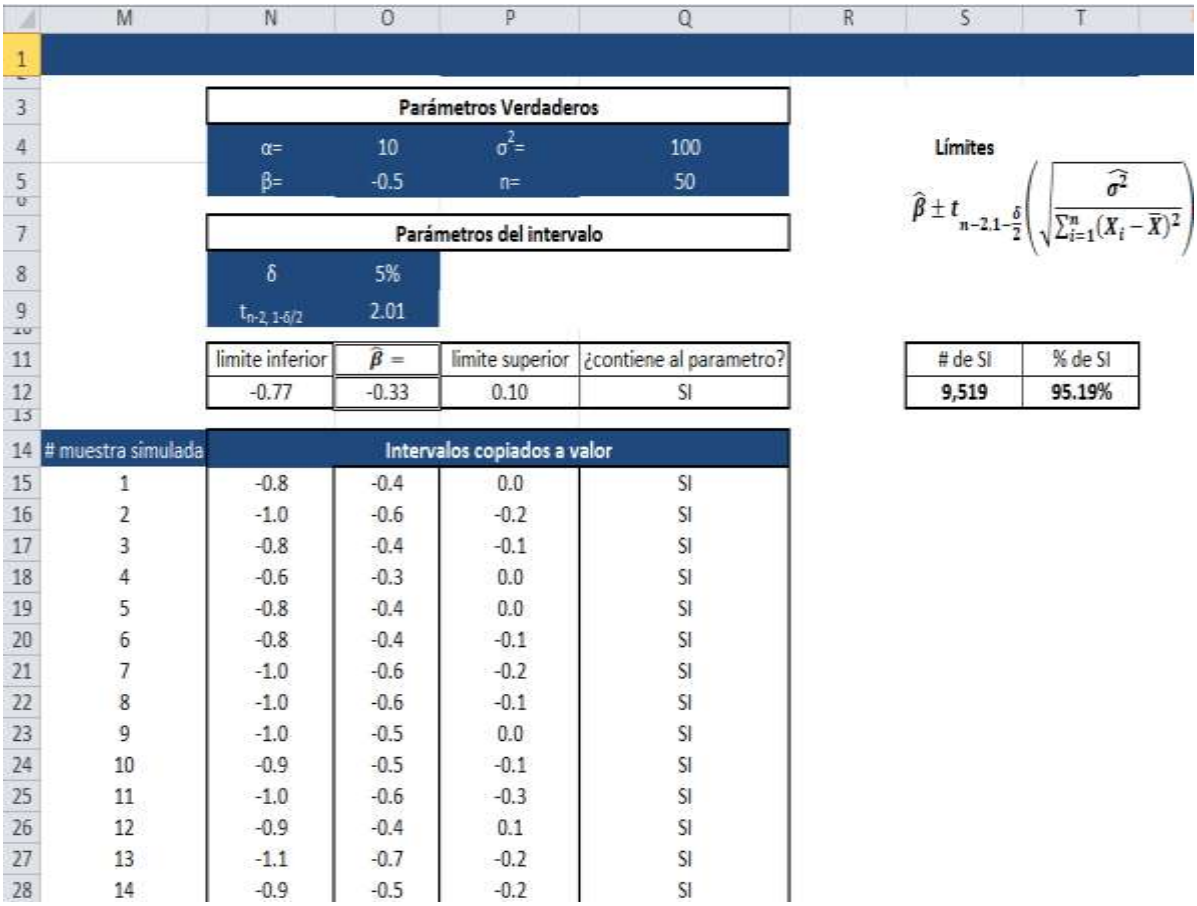
Ahora partir de estas fórmulas de los límites del intervalo, se pueden comprobar sus propiedades vía simulación Monte Carlo, empleando la misma estructura que se creó para las simulaciones de los intervalos de confianza de α. Para simplificar, la hoja completa se puede copiar y emplearla como plantilla de trabajo.

Los cambios que deben realizarse en la hoja para adaptarla son, la referencia al cálculo del estimador $\hat{\beta}$, las fórmulas de los límites inferior y superior, de forma opcional se puede colocar en imagen las fórmulas como apoyo, además de cambiar la referencia del condicionante, que ahora apuntará al verdadero valor del parámetro β.

Otro aspecto de la muestra que se ha mantenido constante es el tamaño de la muestra, por lo que para las siguientes simulaciones, se incrementó el número de parejas simuladas a $n = 50$. De igual manera que para el parámetro anterior, las simulaciones se generan por la rutina automatizada para copiar los valores del renglón con los cálculos de interés.

La siguiente imagen muestra la estructura en hoja de cálculo adaptada para simular intervalos sobre β al 95% de nivel de confianza, donde se mantuvieron como parámetros verdaderos los valores $\alpha = 10, \beta = -0.5$ y $\sigma^2 = 100$. De igual manera que en el caso anterior, para mostrar la confianza de los intervalos se pueden considerar diferentes cantidades de muestras simuladas, los cuales se resumen en la tabla 5.2. La imagen muestra el resultado después de generar 10,000 muestras, donde se puede observar que el número de intervalos que encerraron al parámetro fueron 9,519 representando el 95.19% del total.

Estructura en Excel® para generar intervalos de confianza sobre β



Parámetros Verdaderos			
α=	10	σ^2=	100
β=	-0.5	n=	50

Límites

$$\widehat{\beta} \pm t_{n-2,1-\frac{\delta}{2}} \left(\sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^{n}(X_i - \overline{X})^2}} \right)$$

Parámetros del intervalo	
δ	5%
$t_{n-2,\,1-\delta/2}$	2.01

limite inferior	$\widehat{\beta}$ =	limite superior	¿contiene al parametro?
-0.77	-0.33	0.10	SI

# de SI	% de SI
9,519	95.19%

# muestra simulada	Intervalos copiados a valor			
1	-0.8	-0.4	0.0	SI
2	-1.0	-0.6	-0.2	SI
3	-0.8	-0.4	-0.1	SI
4	-0.6	-0.3	0.0	SI
5	-0.8	-0.4	0.0	SI
6	-0.8	-0.4	-0.1	SI
7	-1.0	-0.6	-0.2	SI
8	-1.0	-0.6	-0.1	SI
9	-1.0	-0.5	0.0	SI
10	-0.9	-0.5	-0.1	SI
11	-1.0	-0.6	-0.3	SI
12	-0.9	-0.4	0.1	SI
13	-1.1	-0.7	-0.2	SI
14	-0.9	-0.5	-0.2	SI

Tabla 5.2: Resumen de simulaciones de intervalos de confianza para β

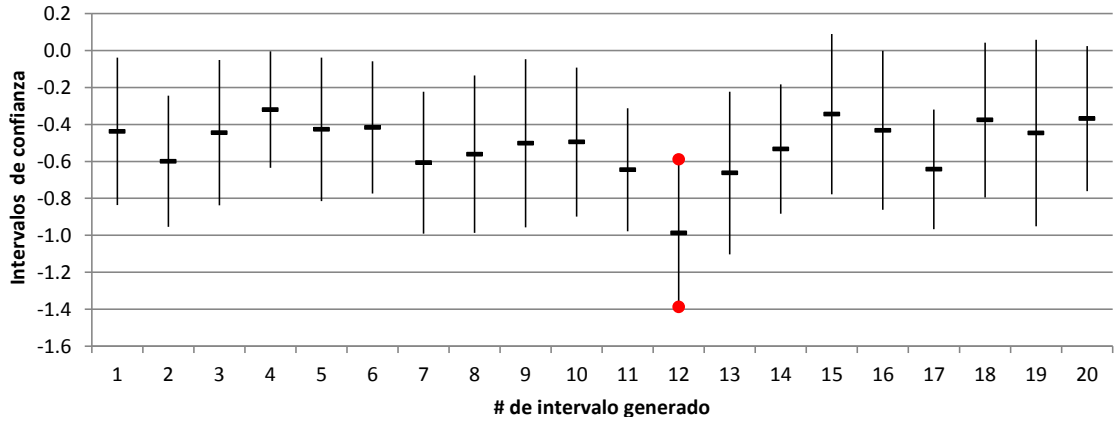
Intervalos Generados	# de intervalos que encerraron β	% de intervalos que encerraron β
100	97	97.00%
500	487	97.40%
1,000	950	95.00%
3,000	2,829	94.30%
5,500	5,204	94.62%
10,000	9,519	95.19%

En la tabla se puede observar como los intervalos construidos bajo diferentes números de simulaciones, mantienen constante el porcentaje de intervalos efectivos en cubrir al parámetro β, lo que comprueba, de la misma forma, lo que es la confianza para estos intervalos. Para visualizar el comportamiento de los intervalos generados para la pendiente se inserta un gráfico de máximo y mínimos, similar al grafico 5.13 en el cual se señala un intervalo de los primeros 20, que no encerró al parámetro $\beta = -0.5$.

Es destacable señalar que la longitud del intervalo también fue impactada además por el tamaño de muestra, en el sentido que el cuantil que determina el intervalo ahora toma el valor de $t_{50-2,1-\frac{\delta}{2}} = 2.01$ a comparación de cuando se tomaba $n = 15$ donde el

valor del cuantil era $t_{15-2,1-\frac{\delta}{2}} = 2.16$ reduciendo de forma sistemática la longitud de cada intervalo, lo cual también, visto desde otra perspectiva, se puede interpretar que con un tamaño de muestra mayor se espera que los intervalos sean más precisos. Sin embargo este intervalo que no cubre el verdadero valor del parámetro está dentro de lo esperado para estas 20 simulaciones ya que los 19 intervalos que cubren a β representan el 95% de los casos.

Gráfica 5.14: Primeros Intervalos de confianza de 1000 muestras para el parametro β



Intervalos de confianza para σ^2

Para construir los intervalos de confianza sobre la varianza del error, basta con retomar el estadístico basado en el estimador $\widehat{\sigma^2}$, empleado para generar la distribución t en los intervalos anteriores, del cual se sabe que su distribución es Ji-cuadrada con $n-2$ grados de libertad.

$$\frac{(n-2)\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-2}$$

Con esta variable se desea definir entonces un intervalo de confianza alrededor de σ^2, el cual acumule la probabilidad $1-\delta$ en encerrarlo; análogamente como los intervalos anteriores se despeja el parámetro σ^2 a partir del intervalo que lo contenga con la confianza deseada, de la siguiente forma:

$$P\left(\chi^2_{n-2,\delta/2} \leq \frac{(n-2)\widehat{\sigma^2}}{\sigma^2} \leq \chi^2_{n-2,1-\delta/2}\right) = 1-\delta$$

$$P\left(\frac{1}{\chi^2_{n-2,1-\frac{\delta}{2}}} \leq \frac{\sigma^2}{(n-2)\widehat{\sigma^2}} \leq \frac{1}{\chi^2_{n-2,\frac{\delta}{2}}}\right) = 1-\delta$$

$$P\left(\frac{(n-2)\widehat{\sigma^2}}{\chi^2_{n-2,1-\frac{\delta}{2}}} \leq \sigma^2 \leq \frac{(n-2)\widehat{\sigma^2}}{\chi^2_{n-2,\frac{\delta}{2}}}\right) = 1-\delta$$

Por lo tanto el intervalo $\left(\frac{(n-2)\widehat{\sigma^2}}{\chi^2_{n-2,1-\frac{\delta}{2}}} \ , \ \frac{(n-2)\widehat{\sigma^2}}{\chi^2_{n-2,\frac{\delta}{2}}}\right)$ contiene al parámetro con una confianza de $(1-\delta)*100\%$. De este intervalo es importante señalar que cada uno de los cuantiles del intervalo debe ser computado en celdas distintas, debido a que la distribución Ji-cuadrada es asimétrica por lo cual ambos cuantiles se encuentran a distancias diferentes del valor medio.

Los cuantiles de ésta distribución se obtienen a través de la función INV.CHICUAD(p,n) la cual toma como parámetros la probabilidad p acumulada desde la cola izquierda de la distribución hasta el cuantil $\chi^2_{n,p}$ que se desea ubicar y n como los grados de libertad de la distribución. La ventaja de estas fórmulas radica en que es sencillo calcular los límites del intervalo, pues se resume en dividir la suma de cuadrados de los residuos entre el cuantil correspondiente.

Ahora, para mostrar los intervalos por medio de simulaciones, reutilizando lo construido en hoja de cálculo, se puede copiar y usar como plantilla a la cual se requiere adaptar los cálculos necesarios cuidando las referencias para generar el intervalo y se actualice con cada muestra generada. En esta ocasión se decidió generar una serie de intervalos simulados con un nivel de confianza del 90%, conservando los parámetros anteriores y el tamaño de muestra $n = 50$ iguales a la simulación anterior. El resultado de variar el número de intervalos generados se encuentra en la Tabla 5.3 donde se puede observar que el porcentaje que representan los intervalos que contuvieron al valor de σ^2, como en los casos de los parámetros anteriores, se aproxima a la confianza teórica que se definió.

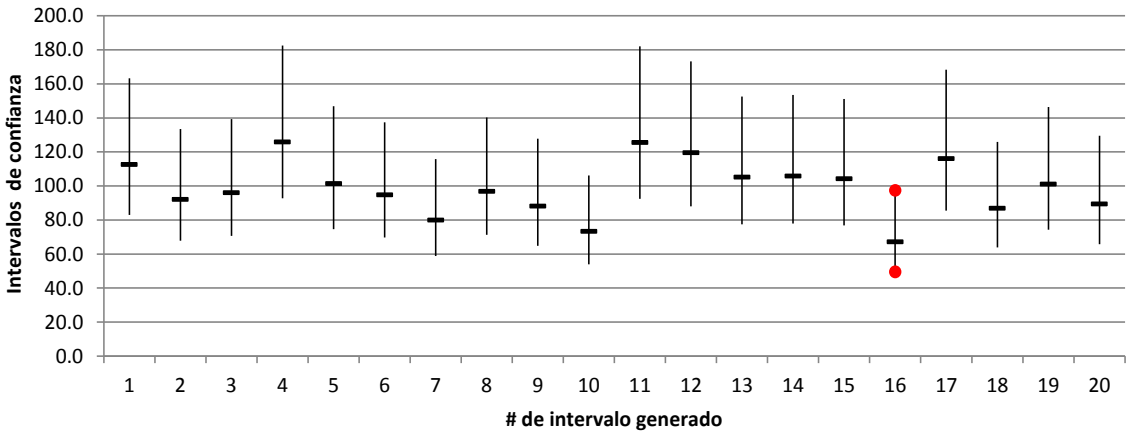
Tabla 5.3: Resumen de simulaciones de intervalos de confianza para σ^2

Intervalos Generados	# de intervalos que encerraron σ^2	% de intervalos que encerraron σ^2
100	81	81.00%
500	446	89.20%
1,000	899	89.90%
3,000	2,691	89.70%
5,500	4,953	90.05%
10,000	9,011	90.11%

Observar los intervalos generados en un gráfico es de utilidad para comprender cómo afecta la asimetría de la distribución Ji-cuadrada a los intervalos, derivado de la asimetría de su densidad[29] por lo cual se puede ver en la Gráfica 5.15, que el límite superior del intervalo se aleja más del estimador, lo que le da una forma asimétrica alrededor de $\widehat{\sigma^2}$. De acuerdo a los resultados de la tabla anterior sobre significancia de la prueba, en esta ventana de observación de 20 intervalos se esperaría que un intervalo no encerrara al parámetro, que en este caso fue el intervalo señalado en rojo.

[29] Para una referencia gráfica de la distribución, consulte el capítulo III Gráfica 3.5.

Gráfica 5.15: Primeros Intervalos de confianza para el parametro σ²



Pruebas de hipótesis

Después de tratar los intervalos de confianza para los parámetros, el siguiente tema de importancia son las pruebas de hipótesis sobre los parámetros, pues a diferencia de las simulaciones, con datos prácticos no se puede conocer previamente el valor de los parámetros.

Sin embargo en ocasiones ya sea al repetir experimentos que arrojen el registro de una muestra en el tiempo o por la necesidad de comparar el comportamiento de un fenómeno contra poblaciones similares, donde se hayan realizado ajustes de regresión o simplemente por un juicio personal de valor de un investigador experimentado, permite proponer como hipótesis nula que los parámetros de la regresión sean iguales a ciertos valores.

A continuación se verán las pruebas de hipótesis sobre los parámetros del modelo de regresión y la forma de entender sus componentes por medio de simulación Monte Carlo.

En el caso del parámetro α se propone como hipótesis nula que el parámetro α sea igual a un α_0, es decir de la forma $\boldsymbol{H_0}: \boldsymbol{\alpha} = \boldsymbol{\alpha_0}$, la cual se puede contrastar contra la hipótesis alternativa $\boldsymbol{H_1}: \boldsymbol{\alpha} \neq \boldsymbol{\alpha_0}$. El siguiente estadístico T_α es el que se emplea como base para determinar el criterio de rechazo, pues bajo la hipótesis nula, se sabe que T_α se distribuye t de *Student*

$$\boldsymbol{T_\alpha} = \frac{\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha_0}}{\sqrt{\widehat{\boldsymbol{\sigma^2}}\left(\frac{\mathbf{1}}{\boldsymbol{n}} + \left(\frac{\bar{\boldsymbol{X}}^{\mathbf{2}}}{\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}(\boldsymbol{X_i}-\bar{\boldsymbol{X}})^{\mathbf{2}}}\right)\right)}}$$

Para saber si se debe rechazar la hipótesis nula con base en el valor de T_α, supóngase que se realiza la prueba en una muestra particular, entonces si la H_0 fuera cierta se esperaría que el estadístico T_α tome valores cercanos a 0, pues el valor del estimador $\hat{\alpha}$ sería cercano al valor α_0, pero si la hipótesis H_1 fuera cierta el valor de T_α se alejaría de 0 a medida que sea menor o mayor que α_0 . Con base en lo anterior, la regla de decisión para esta prueba de hipótesis con un nivel de significancia δ es

$$\boldsymbol{Rechazar\ H_0: \alpha = \alpha_0\ si}$$

$$\boldsymbol{T_\alpha > t_{n-2,1-\frac{\delta}{2}} \quad o \quad T_\alpha < -t_{n-2,1-\frac{\delta}{2}}}$$

$$\boldsymbol{considerando\ que\ t_{n-2,\frac{\delta}{2}} = -t_{n-2,1-\frac{\delta}{2}}}$$

Cuando no se rechace la prueba para realizar la conclusión, se debe recordar que éste rechazo de H_0 no implica aceptar la hipótesis alternativa, ya que en ese caso se enuncia que no hay evidencias estadísticas, que $\alpha \neq \alpha_0$.

Para mostrar el comportamiento de las pruebas sobre este parámetro, se puede copiar o construir desde un inicio una hoja de cálculo con las mismas características vistas para simular pruebas de hipótesis del Capítulo IV. Entonces se deben introducir los valores de los parámetros de la prueba y de la regresión para poder cambiar sus valores a voluntad. Luego se calculan el estadístico de prueba donde se debe recordar que al computar $\boldsymbol{T_\alpha}$ se deben referenciar las celdas con los cálculos anteriores para los estimadores y sus varianzas, aprovechando que estos elementos se encuentran calculados en hojas de cálculo de las secciones anteriores.

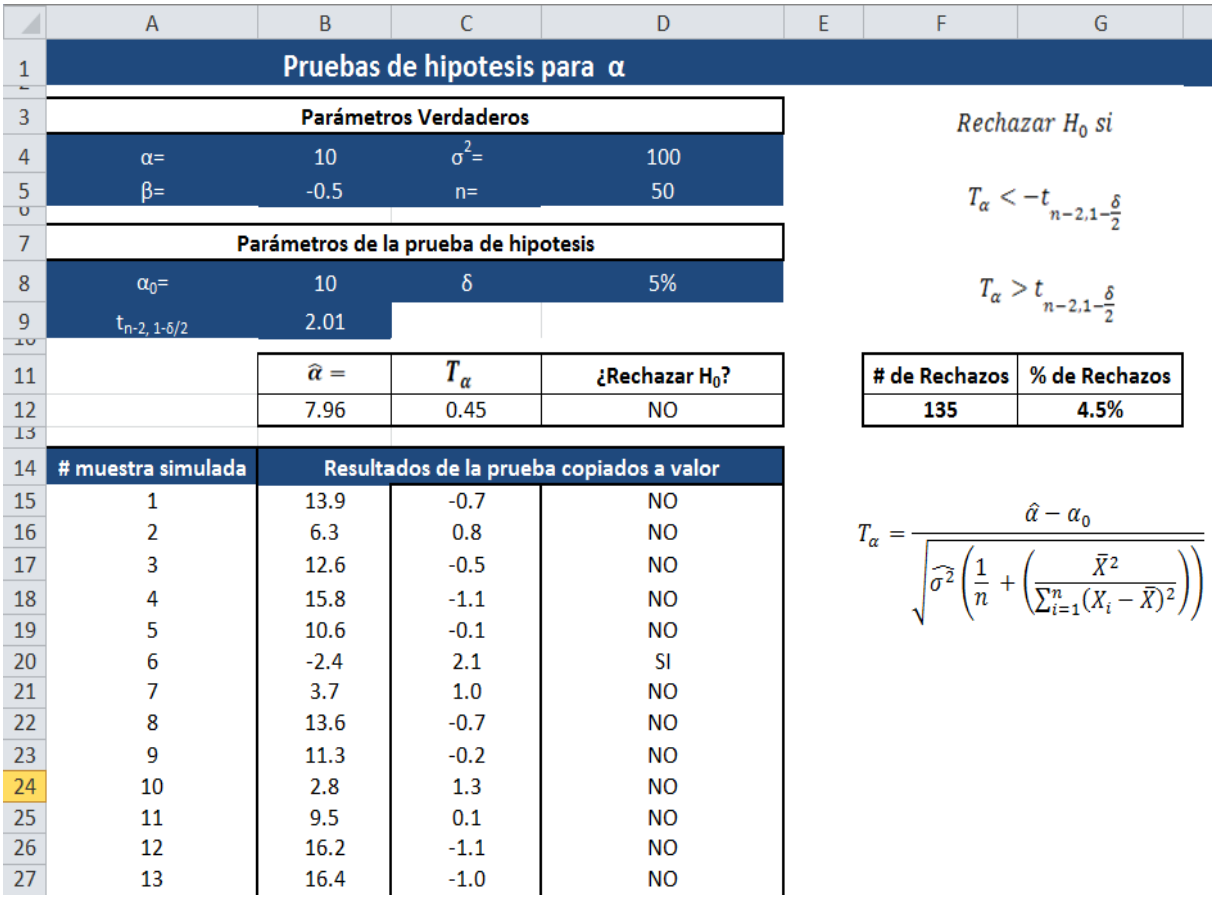
La regla de decisión se puede introducir a la hoja de cálculo por medio de la siguiente fórmula **SI(-$t_{n-2,1-\frac{\delta}{2}}$ <= T_α, SI(T_α<= $t_{n-2,1-\frac{\delta}{2}}$, "NO" , "SI") , "SI")** la cual comprueba si el estadístico está en el rango de los cuantiles en cuyo caso no rechaza la prueba y devuelve “SI” en el caso de salir del rango de los cuantiles.

Puesto que por cada actualización de la hoja se generará una nueva realización de la prueba, la rutina Macro empleada anteriormente será de utilidad para generar y registrar las simulaciones. En la siguiente imagen se observa la adaptación de la hoja de cálculo propuesta para las pruebas de hipótesis, donde se consideró conservar los parámetros $\alpha = 10, \beta = -0.5,\ \sigma^2 = 100$, con un tamaño de muestra $n = 50$.

Para realizar los contrastes simulados se consideraron diversos escenarios de la hipótesis nula tomando $\alpha_0 = -15, 10, 20, 35,$ considerando además, para cada escenario diferentes cantidades de muestras simuladas. Del lado derecho de la imagen se puede observar el conteo de veces que la prueba rechazó la hipótesis nula y el porcentaje que representa del total de simulaciones de un total de 3,000 generadas, que en el caso particular mostrado, es de alrededor del 5%, porcentaje esperado ya que $\alpha = 10$ y $\alpha_0 = 10$, pues se obtiene la significancia definida.

Para resumir los escenarios y simulaciones computadas la Tabla 5.4 contiene los resultados de realizar los contrastes de las diversas hipótesis con diferentes cantidades de muestras simuladas.

Estructura en Excel® para generar pruebas de hipótesis sobre α



Pruebas de hipotesis para α			
Parámetros Verdaderos			
α=	10	σ^2=	100
β=	-0.5	n=	50
Parámetros de la prueba de hipotesis			
α_0=	10	δ	5%
$t_{n-2,\ 1-\delta/2}$	2.01		

Rechazar H_0 si

$T_\alpha < -t_{n-2,1-\frac{\delta}{2}}$

$T_\alpha > t_{n-2,1-\frac{\delta}{2}}$

$\hat{\alpha}$ =	T_α	¿Rechazar H_0?
7.96	0.45	NO

# de Rechazos	% de Rechazos
135	4.5%

# muestra simulada	Resultados de la prueba copiados a valor		
1	13.9	-0.7	NO
2	6.3	0.8	NO
3	12.6	-0.5	NO
4	15.8	-1.1	NO
5	10.6	-0.1	NO
6	-2.4	2.1	SI
7	3.7	1.0	NO
8	13.6	-0.7	NO
9	11.3	-0.2	NO
10	2.8	1.3	NO
11	9.5	0.1	NO
12	16.2	-1.1	NO
13	16.4	-1.0	NO

$$T_\alpha = \frac{\hat{\alpha} - \alpha_0}{\sqrt{\widehat{\sigma^2}\left(\frac{1}{n} + \left(\frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)\right)}}$$

Tabla 5.4: Resumen de simulaciones de pruebas de hipótesis para α

% de pruebas rechazadas				
# de Pruebas	Valores de α_0			
	-15	10	20	35
500	99.40%	5.40%	44.00%	98.80%
1,000	99.00%	4.90%	44.20%	99.00%
3,000	99.20%	4.50%	41.60%	99.20%

En la tabla se puede identificar que bajo un mismo escenario el porcentaje de pruebas rechazadas es muy similar además cuando $\alpha_0 = 10$ el porcentaje de rechazos es igual a la significancia de la prueba, sin embargo cuando se revisa de forma horizontal la tabla a través de los escenarios se puede observar que de manera sistemática la prueba rechazará cada vez a un mayor número de resultados, conforme se vaya distanciando las hipótesis nula del valor real de α, ya que en los casos extremos cuando se tomó una diferencia de 25 unidades del valor de α la hipótesis nula se

rechazó en un 99% de los casos, donde también se puede retomar y reforzar el concepto de la potencia de la prueba[30].

La prueba de hipótesis en el caso del parámetro β es muy similar a la del parámetro anterior, ya que para los intervalos de confianza también se trabajó con un estadístico t de *Student* para despejar β, por lo que retomando del tema anterior el estadístico de prueba es

$$\boldsymbol{T_{\beta} = \frac{\widehat{\beta} - \beta_0}{\sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}} \sim t_{(n-2)}}$$

Se contrasta entonces una hipótesis nula acerca del valor β como $H_0: \beta = \beta_0,\ \ \beta_0 \in \mathbb{R}$ y la hipótesis alternativa como $H_1: \beta \neq \beta_0$. Eligiendo un nivel de significancia δ, de forma análoga al parámetro anterior, la regla de rechazo se encuentra en términos de los cuantiles de la distribución de la siguiente manera

$$\boldsymbol{Rechazar\ H_0: \beta = \beta_0\ si}$$

$$\boldsymbol{T_{\beta} > t_{n-2,1-\frac{\delta}{2}} \quad ó}$$

$$\boldsymbol{T_{\beta} < -t_{n-2,1-\frac{\delta}{2}}}$$

Para mostrar el comportamiento de esta prueba se realizó el mismo ejercicio de simulación, sobre diversos escenarios de β_0. Por la gran similitud con la prueba anterior, construirla en este caso se puede conseguir de forma simple copiando la estructura anterior y modificar las formulas correspondientes, cuidando hacer referencia a los cálculos previos hechos del estimador al adaptar los cálculos del estadístico de prueba y la condicionante que permita evaluar si se debe o no rechazar la hipótesis nula, colocando además las fórmulas como imagen de los cálculos como apoyo didáctico.

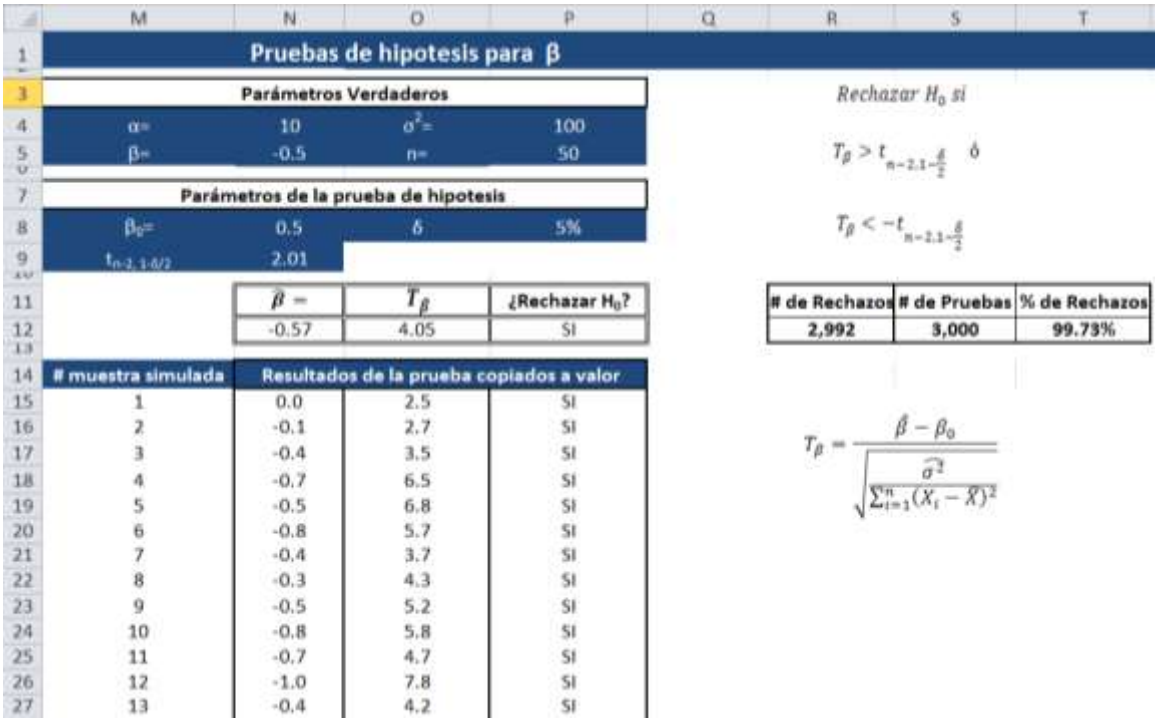
Se consideró conservar los parámetros y tamaños de muestra anteriores, con $\boldsymbol{\beta = -0.5}$, partir del cual se consideró tomar los siguientes escenarios para las pruebas $\boldsymbol{\beta_0 = -1.5, -1, -0.5, 0, 0.5}$, con diversas cantidades de muestras simuladas.

En la siguiente imagen se muestra la estructura de hoja de cálculo adaptada para la prueba sobre los valores de β, mostrando el escenario cuando $\boldsymbol{\beta_0 = 0.5}$, donde se observa que la hipótesis nula se aleja del verdadero valor del parámetro y por esta razón fue rechazada ésta hipótesis nula en el 99.7% de las simulaciones.

En la Tabla 5.5 se encuentran resumidos los escenarios considerando un **nivel de significancia del 5**% para realizar cada prueba, donde se muestra cómo al desviarse la hipótesis del valor real, el número de rechazos y en consecuencia el porcentaje que representan de la muestra, aumentan de forma muy sensible, incluso mayor que en el caso de α, explicado en este caso porque el estimador $\hat{\beta}$ tiene menor varianza.

[30] Para una referencia más a detalle consulte el capítulo IV, donde se trata el tema de la función potencia en las pruebas de hipótesis.

Estructura en Excel® para generar pruebas de hipótesis sobre β



Pruebas de hipotesis para β

Parámetros Verdaderos			
α=	10	σ^2=	100
β=	-0.5	n=	50

Parámetros de la prueba de hipotesis			
β_0=	0.5	δ	5%
$t_{n-2,\,1-\delta/2}$	2.01		

$\hat{\beta}$ =	$T_{\hat{\beta}}$	¿Rechazar H_0?
-0.57	4.05	SI

# muestra simulada	Resultados de la prueba copiados a valor		
1	0.0	2.5	SI
2	-0.1	2.7	SI
3	-0.4	3.5	SI
4	-0.7	6.5	SI
5	-0.5	6.8	SI
6	-0.8	5.7	SI
7	-0.4	3.7	SI
8	-0.3	4.3	SI
9	-0.5	5.2	SI
10	-0.8	5.8	SI
11	-0.7	4.7	SI
12	-1.0	7.8	SI
13	-0.4	4.2	SI

Rechazar H_0 si

$T_\beta > t_{n-2,1-\frac{\delta}{2}}$ ó

$T_\beta < -t_{n-2,1-\frac{\delta}{2}}$

# de Rechazos	# de Pruebas	% de Rechazos
2,992	3,000	99.73%

$$T_\beta = \frac{\hat{\beta} - \beta_0}{\sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}}$$

Tabla 5.5: Resumen de simulaciones de pruebas de hipótesis para β

% de pruebas rechazadas					
# de Pruebas	Valores de β_0				
	-1.5	-1	-0.5	0	0.5
500	99.60%	69.20%	4.80%	69.20%	99.80%
1,000	99.60%	68.60%	5.50%	70.00%	99.90%
3,000	99.70%	68.90%	5.20%	68.10%	99.70%

Otro detalle que se puede observar de la tabla es la simetría del comportamiento de los rechazos alrededor del verdadero valor de β, debido a la simetría de la distribución t de *Student* del estadístico de prueba, lo que asegura rechazar la prueba cuando la hipótesis nula subestime o sobreestime el verdadero valor del parámetro.

El siguiente parámetro para el cual se realizan pruebas de hipótesis es sobre el valor de la varianza del error σ^2. Para construir esta prueba se retoma como estadístico de prueba, el mismo que se empleó en los intervalos de confianza, distribuido como una Ji-cuadrada con $n-2$ grados de libertad.

$$\boldsymbol{T_{\sigma^2} = \frac{(n-2)\widehat{\sigma^2}}{\sigma_0^2} \sim \chi^2_{(n-2)}}$$

La hipótesis nula, como en los casos anteriores propone un valor σ_0^2, del cual se considera es igual al parámetro de la población, por lo que la hipótesis nula se enuncia como $\boldsymbol{H_0: \sigma^2 = \sigma_0^2}$. Acerca del estadístico, se puede observar que si el valor del

estimador $\widehat{\sigma^2}$ es cercano a σ_0^2 entonces su cociente se aproximará a 1, y en consecuencia T_{σ^2} se aproximará a $n-2$.

Otro aspecto de las pruebas de hipótesis, es que hasta el momento en los dos casos anteriores se consideró como hipótesis alternativa, que el parámetro no fuera igual al valor de la hipótesis nula, por esta razón se rechazaba la prueba cuando el estadístico de prueba se desviaba en cualquier sentido del verdadero valor de parámetro.

Sin embargo la hipótesis alternativa puede tener varias formas, la que se considerará en este caso es cuando se desea contrastar que σ^2 es mayor a un valor σ_0^2, por lo que la hipótesis alternativa se enuncia de la forma $\boldsymbol{H_1: \sigma^2 > \sigma_0^2}$.

Si se considera un valor δ para la significancia de la prueba, el contraste de hipótesis y la regla de decisión son:

$$\boldsymbol{H_0: \sigma^2 = \sigma_0^2 \; contra \;\; H_1: \sigma^2 > \sigma_0^2}$$

$$\boldsymbol{Rechazar\, H_0: \sigma^2 = \sigma_0^2 \;\; si \;\; T_{\sigma^2} > \chi^2_{n-2,1-\delta}}$$

Para mostrar el comportamiento de esta prueba se reutilizó la estructura de trabajo anterior para simular las pruebas de hipótesis. Después de adaptar los parámetros y cálculos para esta prueba, se modifica también la fórmula de la condicional construida para evaluar el resultado de la prueba, la cual se simplifica al sólo realizar la evaluación de la siguiente forma **SI(** T_{σ^2}**>** $\chi^2_{n-2,1-\delta}$**, "SI" , "NO")**.

Para este ejemplo se conservaron también los parámetros antes definidos, con $\sigma^2 = 100$, así que considerando un **nivel de significancia del 10%** se generaron los siguientes escenarios sobre el valor de $\boldsymbol{\sigma_0^2 = 95, 100, 110, 130}$, donde para cada uno de los escenarios de la hipótesis nula, se generaron $\boldsymbol{500, 1000, y\, 3000}$ muestras por la rutina macro mencionada anteriormente, con el objetivo de ver el efecto combinado al variar estos aspectos de la prueba.

En la siguiente imagen se muestra la hoja actualizada para realizar las pruebas sobre σ^2, donde se encuentra el cálculo del cuantil $\boldsymbol{\chi^2_{n-2,90\%}}$, las fórmulas de apoyo y el resultado de realizar 3,000 simulaciones. En el escenario mostrado se eligió $\sigma_0^2 = 130$, el cual muestra un porcentaje de rechazos de la prueba de 0.47%, debido a que la hipótesis alternativa se prueba en una dirección y considerando que el parámetro de la varianza del error es $\boldsymbol{\sigma^2 = 100}$ entonces la prueba muestra con una mayor potencia como H_1 es falsa.

Además en la Tabla 5.6 se halla el resumen de los diversos escenarios y número de muestras generadas, donde se observa cómo a medida que se incrementa el valor de la hipótesis es menor el número de veces que fue rechazada la hipótesis nula, sin embargo al mantener la hipótesis nula y variar el número de simulaciones el porcentaje de rechazos se mantuvo estable, resaltando además que en el escenario $\sigma_0^2 = 100$ el porcentaje fue muy similar a la significancia de la prueba.

Tabla 5.6: Resumen de simulaciones de pruebas de hipótesis para σ^2				
% de pruebas rechazadas				
# de Pruebas	Valores de σ^2_0			
	95	100	110	130
500	14.20%	10.20%	3.80%	0.60%
1,000	13.50%	9.30%	2.90%	0.40%
3,000	15.50%	9.80%	4.30%	0.50%

Estructura en Excel® para generar pruebas de hipótesis sobre σ^2

	X	Y	Z	AA	AB	AC	AD	AE
1	Pruebas de hipótesis para σ^2							
3	Parámetros Verdaderos					$Tomando\ H_1{:}\sigma^2 > \sigma_0^2$		
4	α=	10	σ^2=	100				
5	β=	-0.5	n=	50		$Rechazar\ H_0{:}\sigma^2 = \sigma_0^2\ si$		
7	Parámetros del intervalo							
8	σ^2_0=	130	$\chi^2_{n-2,1-\delta}$	60.91		$T_{\sigma^2} > \chi^2_{n-2,1-\delta}$		
9	δ	10%						
11		$\widehat{\sigma^2}$ =	T_{σ^2}	¿Rechazar H_0?		# de Rechazos	# de Pruebas	% de Rechazos
12		93.19	34.41	NO		14	3,000	0.47%
14	# muestra simulada	Resultados de la prueba copiados a valor						
15	1	95.5	35.3	NO		$T_{\sigma^2} = \frac{(n-2)\widehat{\sigma^2}}{\sigma_0^2}$		
16	2	90.4	33.4	NO				
17	3	108.7	40.1	NO				
18	4	132.2	48.8	NO				
19	5	84.3	31.1	NO				
20	6	133.7	49.4	NO				
21	7	121.8	45.0	NO				
22	8	121.6	44.9	NO				
23	9	88.3	32.6	NO				
24	10	109.3	40.3	NO				
25	11	87.7	32.4	NO				
26	12	102.7	37.9	NO				
27	13	123.8	45.7	NO				
28	14	69.9	25.8	NO				
29	15	106.1	39.2	NO				
30	16	124.6	46.0	NO				
31	17	69.0	25.5	NO				
32	18	103.5	38.2	NO				
33	19	115.4	42.6	NO				
34	20	127.5	47.1	NO				

La concusión de las pruebas cuando se elige la hipótesis alternativa como en estas simulaciones, tiene un detalle, debido a que sólo se está contrastando que σ^2 sea mayor a σ_0^2, por lo tanto cuando no se rechaza la hipótesis nula, sólo se puede afirmar que no hay evidencias estadísticas significativas que $\sigma^2 > \sigma_0^2$.

Coeficiente de correlación y coeficiente de determinación

Cuando se tiene un conjunto de n parejas (X,Y) cada una con datos numéricos y se grafican en un diagrama de dispersión, se puede hallar una idea general de la forma en que se relacionan sus valores, por ejemplo en particular para las simulaciones generadas, se construyó una relación lineal más un factor de error. Sin embargo los tipos de relación que se pueden hallar en la naturaleza o en experimentos entre dos

variables de interés son muy diversos, pues puede que los datos muestren una relación con alguna curvatura, infiriéndose que sea una relación cuadrática o polinómica o por otra parte los datos posean una alta dispersión de tal forma que no se pueda intuir de forma clara su asociación.

El concepto de correlación es la medición, en cuanto a su magnitud, de la intensidad con la que la concentración de los valores de las variables (X, Y) tienen cierta similitud con una línea. En cuanto al sentido de la correlación se expresa como positiva cuando existe una asociación similar a proporcionalidad directa, mientras que se expresa como una correlación negativa si la relación de los datos tiende a ser inversamente proporcional.

Karl Pearson, basado en los trabajos de F. Galton, desarrolló un estadístico para poder medir la magnitud y sentido de la correlación, conocido como el coeficiente de correlación producto-momento de Pearson entre dos variable (X,Y), denotado comúnmente por la letra griega ρ, definido como

$$\boldsymbol{\rho = \frac{Cov(X,Y)}{\sqrt{Var(Y)}\sqrt{Var(X)}}}$$

Cuando se desea obtener el valor del coeficiente de forma empírica en una muestra de parejas (X_i, Y_i), el coeficiente es denotado como r y se calcula de la siguiente manera

$$\boldsymbol{r = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})}{\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}}}$$

r es utilizado ampliamente para medir la correlación, ya que posee la propiedad de estar limitado en el intervalo $(-1,1)$ donde los extremos del intervalo indican una correlación perfecta entre los datos, mientras que en caso de ser $r = 0$ refleja la ausencia de asociación entre las parejas, y además el signo de r expresará el sentido de la correlación. Estas características se pueden comprender mejor si se expresa al coeficiente de la siguiente forma equivalente

$$\boldsymbol{r = \sum_{i=1}^{n}\left(\frac{Y_i - \overline{Y}}{\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}\right)\left(\frac{X_i - \overline{X}}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}}\right)}$$

De esta manera se puede identificar que lo que se está computando es la suma total del producto de una forma similar o vinculada a la estandarización de las variables X y Y, realizando una traslación de ejes y un cambio de escala. Para poder visualizar estos comportamientos y realizar ejemplos de cálculo para r, se reutilizaron las simulaciones generadas anteriormente donde se tiene el cálculo de los estimadores, junto con sus columnas de cálculo desglosadas, a la cuales se pueden agregar las columnas $\frac{Y_i - \overline{Y}}{\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$, $\frac{X_i - \overline{X}}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$ y colocar el cálculo de r en otra celda.

Una vez realizado esto, después se introducen en un diagrama de dispersión, las nuevas columnas computadas, además de agregar un gráfico con los puntos

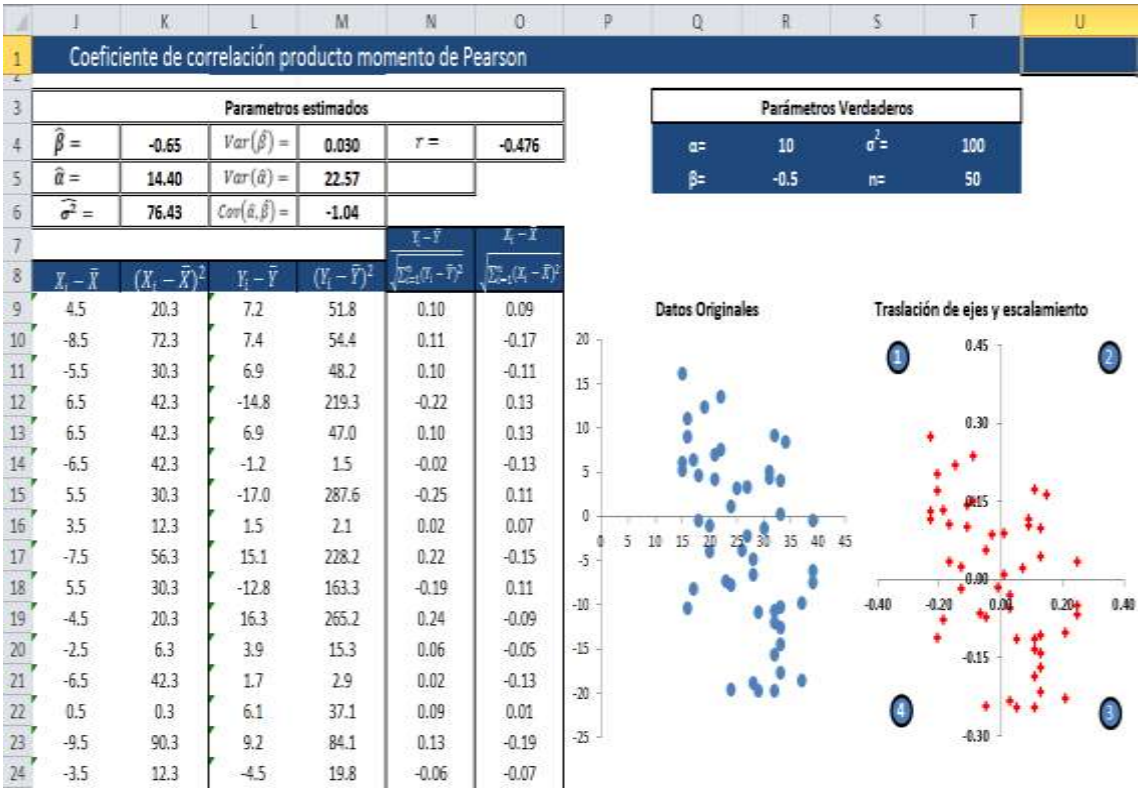
simulados (X_i, Y_i) para poder observar la relación original entre las variables. En la siguiente imagen se encuentra la propuesta en hoja de cálculo para ejemplificar el comportamiento de la correlación, donde los parámetros elegidos para tomar un escenario base, fueron los mismos que los de la sección anterior con $\alpha = 10, \beta = -0.5$ y $\sigma^2 = 100$.

Se observa en la imagen que la relación inversamente proporcional, derivada de una pendiente negativa, se encuentra en ambas gráficas, sin embargo la gráfica derivada de las nuevas columnas con los puntos transformados se encuentra centrada en el origen con los cuadrantes señalados, donde se podrá cotejar con los cálculos de las columnas, que en los puntos localizados en los cuadrantes 2 y 4 el producto de tales parejas son siempre positivos, mientras que los productos de los puntos localizados en los cuadrantes 1 y 3 serán negativos.

En general para esta realización en particular la forma de la relación original hace que los puntos estandarizados con mayor aporte sean los negativos, sin embargo, la dispersión de los datos impacta en los puntos localizados en los cuadrantes 2 y 4, los cuales, aunque son pocos, compensan hacia el sentido positivo, lo cual impacta en la medición de la magnitud de la correlación obteniéndose el valor de $r = -0.476$.

Además se agregaron en la hoja de cálculo los parámetros verdaderos y estimados de la regresión, ya que a partir de estas tablas se pueden cambiar los parámetros para poder observar los resultados de la medición de la correlación bajo diversos escenarios.

Estructura en Excel® para mostrar el cálculo del coeficiente de correlación r



Coeficiente de correlación producto momento de Pearson

Parametros estimados					
$\hat{\beta}$ =	-0.65	$Var(\hat{\beta})$ =	0.030	r =	-0.476
$\hat{\alpha}$ =	14.40	$Var(\hat{\alpha})$ =	22.57		
$\hat{\sigma^2}$ =	76.43	$Cov(\hat{\alpha}, \hat{\beta})$ =	-1.04		

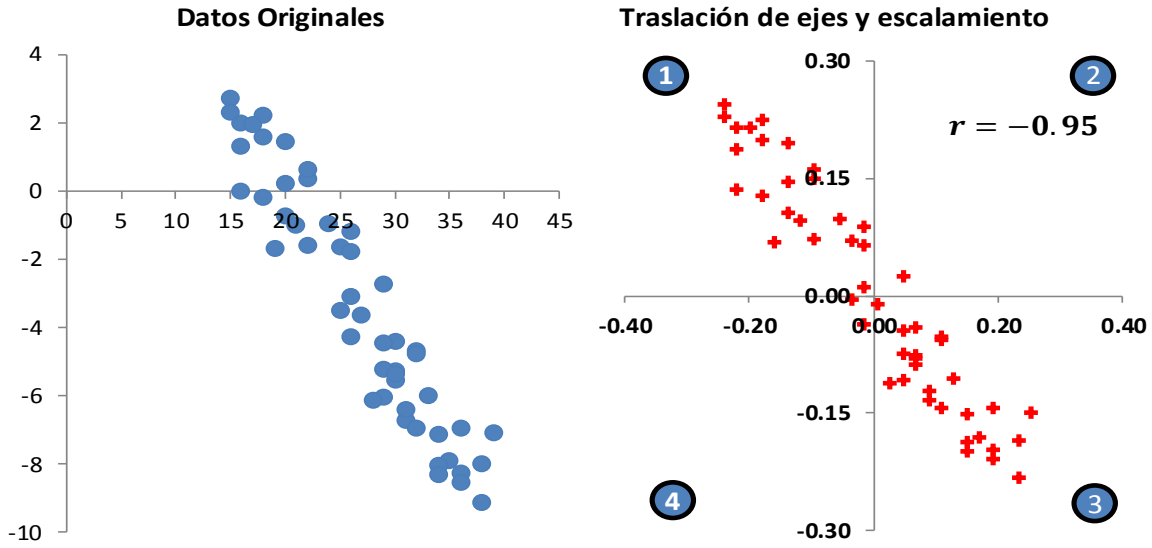
Parámetros Verdaderos			
α=	10	σ^2=	100
β=	-0.5	n=	50

$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$	$\frac{X_i - \bar{X}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$	$\frac{Y_i - \bar{Y}}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$
4.5	20.3	7.2	51.8	0.10	0.09
-8.5	72.3	7.4	54.4	0.11	-0.17
-5.5	30.3	6.9	48.2	0.10	-0.11
6.5	42.3	-14.8	219.3	-0.22	0.13
6.5	42.3	6.9	47.0	0.10	0.13
-6.5	42.3	-1.2	1.5	-0.02	-0.13
5.5	30.3	-17.0	287.6	-0.25	0.11
3.5	12.3	1.5	2.1	0.02	0.07
-7.5	56.3	15.1	228.2	0.22	-0.15
5.5	30.3	-12.8	163.3	-0.19	0.11
-4.5	20.3	16.3	265.2	0.24	-0.09
-2.5	6.3	3.9	15.3	0.06	-0.05
-6.5	42.3	1.7	2.9	0.02	-0.13
0.5	0.3	6.1	37.1	0.09	0.01
-9.5	90.3	9.2	84.1	0.13	-0.19
-3.5	12.3	-4.5	19.8	-0.06	-0.07

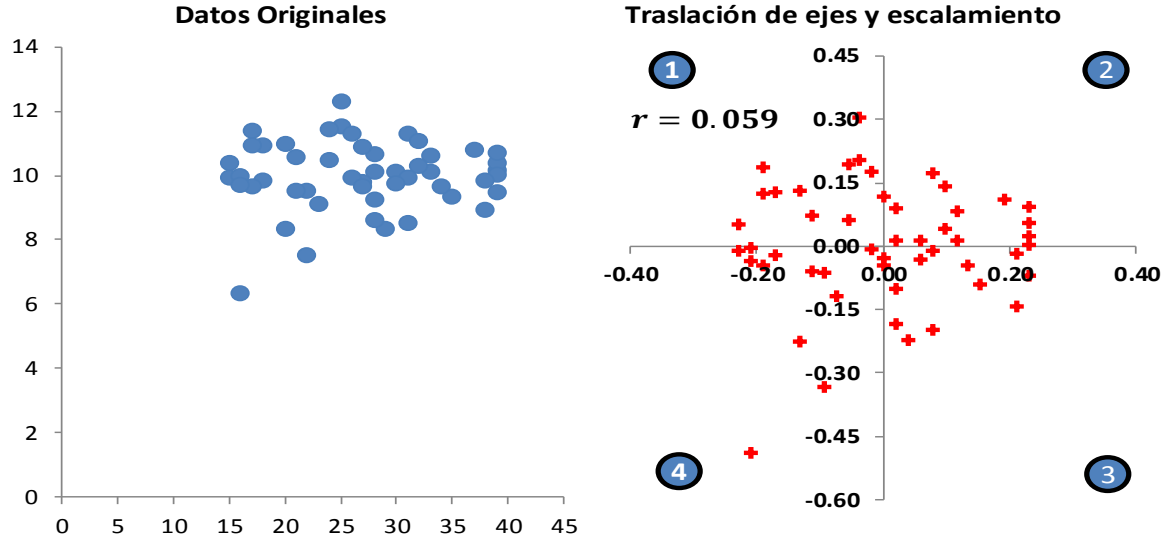
En un siguiente escenario se redujo la varianza a $\sigma^2 = 1$, para mostrar el efecto de generar datos más ajustados a una línea; en la gráfica 5.16 se muestran el par de gráficas con la relación original y la estandarizada de una realización. Para la muestra graficada, el coeficiente de correlación fue de $r = -0.95$ mostrando el sentido al ser un valor negativo y reflejando en su magnitud la similitud de los datos a una recta, al ser cercana a -1.

En otro escenario a partir de este último se conservó la varianza unitaria, para reducir después la magnitud de la pendiente al valor $\beta = 0.001$, cuyo resultado se puede observar en la gráfica 5.17, donde la relación original no parece tener una relación lineal muy clara, además en la forma estandarizada se puede ver cómo la distribución de puntos en cada cuadrante son similares, por lo que en suma se compensan los productos negativos y positivos, arrojando una correlación de $r = 0.059$.

Gráfica 5.16: Muestra generada del modelo lineal y su forma vinculada a la estandarizada, con los parámetros α=10, β=-0.5, σ²=1



Gráfica 5.17: Muestra generada del modelo lineal y su forma vinculada a la estandarizada, con los parámetros α=10, β=0.001, σ²=1



Al realizar diversos cambios en los parámetros se podrá mostrar su impacto en la medición de la correlación, principalmente la pendiente de la recta y la dispersión del error, ya que al cambiar el parámetro de la ordenada al origen sólo se observará un cambio de posición, lo cual no afectará la medición de la correlación. La forma teórica con la que se puede observar la relación anteriormente vista vía simulación entre el coeficiente de correlación y el estimador de la pendiente $\hat{\beta}$, es por medio de la expresión equivalente para como

$$\widehat{\boldsymbol{\beta}} = \frac{\boldsymbol{r}\sqrt{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{Y_i} - \overline{\boldsymbol{Y}})^{\boldsymbol{2}}}}{\sqrt{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{X_i} - \overline{\boldsymbol{X}})^{\boldsymbol{2}}}}$$

A continuación se trabajará ahora con el coeficiente de determinación el cual se puede calcular de forma sencilla como el cuadrado del coeficiente de correlación, por lo que toma valores entre 0 y 1, y representa la proporción de la varianza de la respuesta de la variable dependiente Y que es explicada por la variable independiente X. Otra relación de utilidad sobre el coeficiente de determinación que se empleará más adelante en el análisis del modelo de regresión es

$$\boldsymbol{r^2 = 1 - \frac{\sum_{i=1}^{n} e_i^{\,2}}{\sum_{i=1}^{n} y_i^{\,2}}}$$

$$\boldsymbol{Donde\ y_i = Y_i - \overline{Y}}$$

Donde se puede identificar al segundo término de la suma como el cuadrado del coeficiente de correlación, el cual se sustituye en la ecuación y se despeja para obtener la siguiente expresión

$$\boldsymbol{r^2 = 1 - \frac{\sum_{i=1}^{n} e_i^{\,2}}{\sum_{i=1}^{n} y_i^{\,2}}}$$

El valor del coeficiente de determinación se aproximará a 1 a medida que la suma de cuadrados de los residuos se vuelva cero, lo cual sucede cuando los datos se ajustan cada vez más a una línea recta.

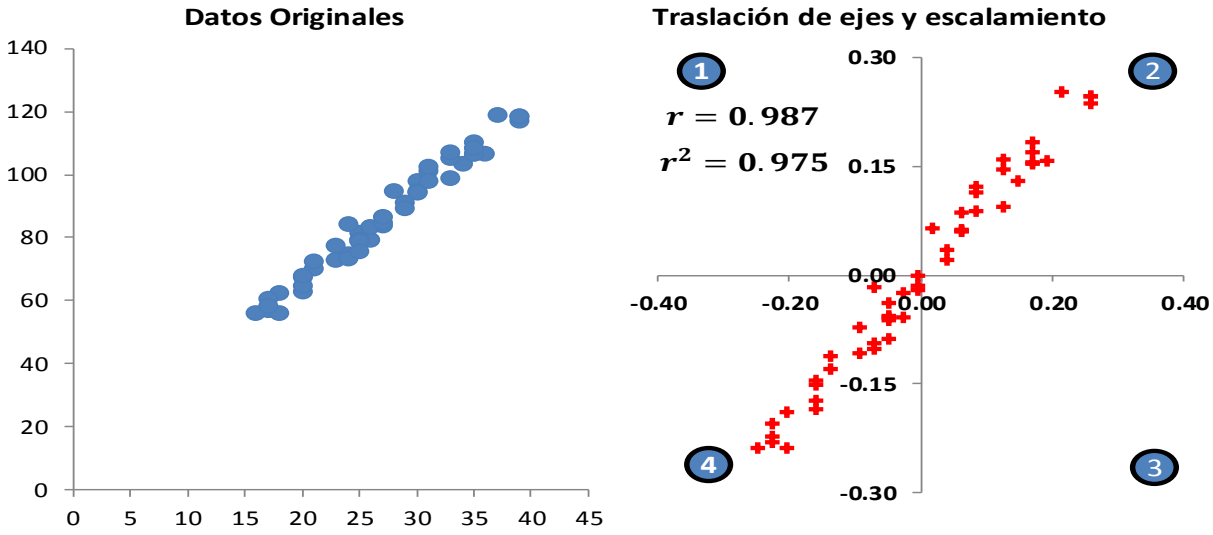
Cuando el valor de r^2 se aproxima a 0 pueden ocurrir diversos casos sobre la relación de los datos, el primero es que no haya una relación lineal entre las variables, por ejemplo en las simulaciones anteriores cuando la pendiente la recta se aproximó a 0, el valor de la correlación fue muy cercano a 0 y por ende también el coeficiente de determinación, al ser el cuadrado de un valor casi nulo. La forma de los datos en la Gráfica 5.17 que muestra este caso se asemeja a una nube de datos sin relación aparente.

Otros posibles casos son cuando la relación no es lineal sin embargo al estandarizar los datos en cuadrantes existe un número similar de datos de un cuadrante que el cuadrante opuesto (el cuadrante 1 es el opuesto del 3 y el cuadrante 2 es el opuesto del 4) de tal forma que se cancelen al calcular r^2, casos como estos pueden darse cuando los datos forman una especie de curva cuadrática o cuando los datos se asemejan a un círculo alrededor del centro los cuadrantes.

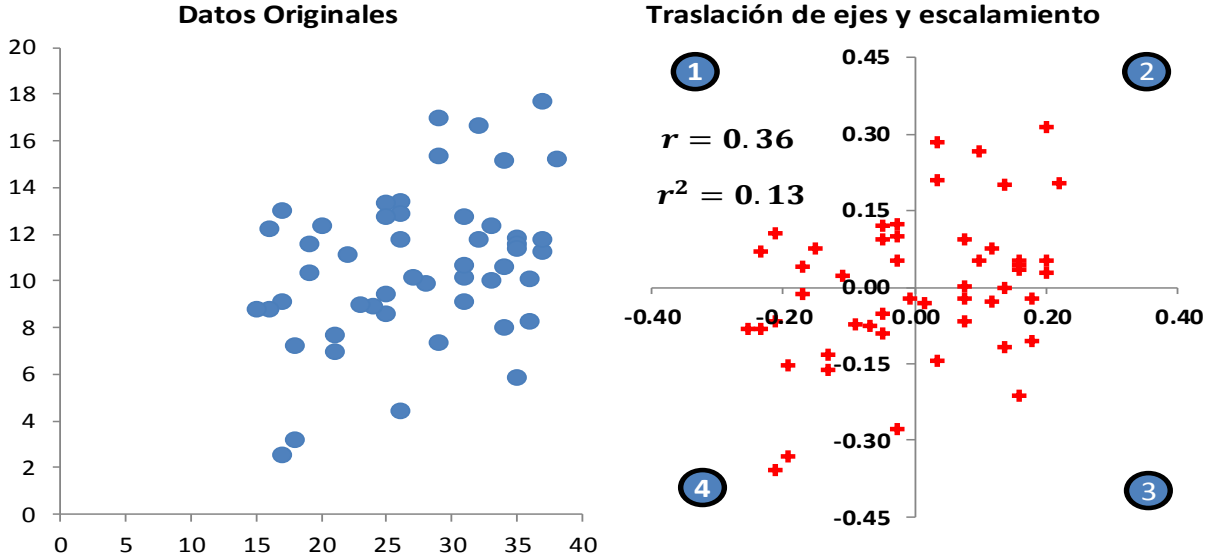
Para ejemplificar estos casos del coeficiente de determinación se tomaron dos escenarios más para la hoja de trabajo de esta sección, en la cual en un primer escenario se consideró tomar una varianza $\sigma^2 = 10$ con una pendiente $\beta = 3$ y ordenada al origen $\alpha = 5$, plasmado en la Gráfica 5.18 donde el coeficiente de correlación fue $r = 0.987$, mientras que el coeficiente de determinación fue $r^2 = 0.975$.

En otro escenario mostrado en la gráfica 5.19, se conservaron los valore de $\sigma^2 = 10$, y $\alpha = 5$, tomando ahora el valor de una pendiente $\beta = 0.2$, con la cual se puede observar que con la dispersión definida, la relación lineal no es tan clara, por lo cual el coeficiente correlación es $r = 0.36$ y el coeficiente de determinación es de $\boldsymbol{r^2 = 0.13}$, del cual se interpreta que el modelo ajustado logra explicar sólo el 13% de la varianza de la respuesta de los valores de la variable dependiente.

Gráfica 5.18: Muestra generada del modelo lineal y su forma vinculada a la estandarizada, con los parámetros α=5, β=3, σ²=10



Gráfica 5.19: Muestra generada del modelo lineal y su forma vinculada a la estandarizada, con los parámetros α=5, β=0.2, σ²=10



Para revisar el caso cuando se toma como nula la relación lineal, se puede referenciar al ejemplo de la Gráfica 5.17, en la que de acuerdo al valor de $\boldsymbol{r = 0.057}$, entonces $\boldsymbol{r^2 = 0.0032}$, esto acorde también al bajo valor de su pendiente $\boldsymbol{\beta = 0.001}$.

Con estos análisis y con un mayor juego sobre los valores de parámetros se puede mostrar el comportamiento de ambos coeficientes y su manejo en la interpretación de los resultados al momento de emplearlos.

Pruebas de hipótesis sobre el coeficiente de correlación ρ

Para realizar pruebas de hipótesis sobre el coeficiente ρ, se debe destacar que la prueba mas usual sobre el coeficiente es su nulidad, es decir se toma como hipótesis nula $\boldsymbol{H_0}$: $\boldsymbol{\rho = 0}$ contra la hipótesis alternativa $\boldsymbol{H_1}$: $\boldsymbol{\rho \neq 0}$. Basados en los ejemplos anteriores considerar que la correlación es nula también es equivalente a decir que el parámetro de la pendiente de regresión $\boldsymbol{\beta}$ es igual a 0.

Por lo cual se parte del mismo estadístico de prueba T_β para probar el caso cuando $\beta = 0$ como

$$\frac{\widehat{\boldsymbol{\beta}}}{\sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^{n}(X_i-\overline{X})^2}}} = \frac{\widehat{\boldsymbol{\beta}}\sqrt{\sum_{i=1}^{n}(\boldsymbol{X_i}-\overline{\boldsymbol{X}})^2}}{\sqrt{\frac{\sum_{i=1}^{n}\boldsymbol{e_i}^2}{\boldsymbol{n-2}}}}$$

El cual se puede transformar en términos del coeficiente de correlación empleando la relación conocida

$$\widehat{\boldsymbol{\beta}} = \frac{\boldsymbol{r}\sqrt{\sum_{i=1}^{n}(\boldsymbol{Y_i}-\overline{\boldsymbol{Y}})^2}}{\sqrt{\sum_{i=1}^{n}(\boldsymbol{X_i}-\overline{\boldsymbol{X}})^2}}$$

Expresando el estadístico como sigue

$$\frac{\boldsymbol{r}\sqrt{\sum_{i=1}^{n}(\boldsymbol{Y_i}-\overline{\boldsymbol{Y}})^2}}{\sqrt{\sum_{i=1}^{n}(\boldsymbol{X_i}-\overline{\boldsymbol{X}})^2}} * \frac{\sqrt{\sum_{i=1}^{n}(\boldsymbol{X_i}-\overline{\boldsymbol{X}})^2}}{\sqrt{\frac{\sum_{i=1}^{n}\boldsymbol{e_i}^2}{\boldsymbol{n-2}}}} = \frac{\boldsymbol{r}\sqrt{\sum_{i=1}^{n}(\boldsymbol{Y_i}-\overline{\boldsymbol{Y}})^2}}{\sqrt{\frac{\sum_{i=1}^{n}\boldsymbol{e_i}^2}{\boldsymbol{n-2}}}}$$

$$= \frac{\boldsymbol{r}\sqrt{\boldsymbol{n-2}}}{\sqrt{\frac{\sum_{i=1}^{n}\boldsymbol{e_i}^2}{\sum_{i=1}^{n}(\boldsymbol{Y_i}-\overline{\boldsymbol{Y}})^2}}}$$

Luego de la derivación del coeficiente de determinación se sabe que $\boldsymbol{r^2 = 1 - \frac{\sum_{i=1}^{n}\boldsymbol{e_i}^2}{\sum_{i=1}^{n}(\boldsymbol{Y_i}-\overline{\boldsymbol{Y}})^2}}$, por lo tanto la expresión final para el estadístico de prueba es

$$\boldsymbol{T_r = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}}$$

El cual se distribuye como una variable t de Student con $n-2$ grados de libertad a partir del cual se define la regla de decisión sobre la hipótesis nula, dado un nivel de significancia δ como

$$\boldsymbol{Rechazar\ H_0}:\ \rho = \mathbf{0}\ \boldsymbol{si}$$

$$\boldsymbol{T_r > t_{n-2,1-\frac{\delta}{2}}} \quad \boldsymbol{ó}$$

$$\boldsymbol{T_r < -t_{n-2,1-\frac{\delta}{2}}}$$

Para mostrar el comportamiento de la prueba se decidió tomar diversos escenarios sobre el verdadero valor de la pendiente como $\beta = -0.2\,, -0.1\,, 0\,, 0.1\,, 0.2$ variando además en cada caso el otro parámetro que impacta la medición de la correlación, que es la varianza del error, tomando los escenarios $\sigma^2 = 1\,, 10\,, 20$, realizando 5,000 pruebas simuladas para cada caso.

Para llevar a cabo las simulaciones se reutilizó la estructura en hoja de cálculo usada para la prueba de hipótesis sobre β , ya que al estar basado en los cuantiles de la distribución t, sólo se tuvo que cambiar la referencia de $\hat{\beta}$ a r y modificar el cálculo del estadístico de prueba.

Tomando una significancia del $\delta = 1\%$, se generaron los escenarios propuestos, los cuales se pueden hallar resumidos en la tabla 5.7 donde se puede observar que en el escenario $\beta = 0$ el porcentaje de rechazos fue el mismo definido en la significancia de la prueba y no tuvo mayores variaciones con diferentes valores de la varianza del error. En los demás escenarios a medida que se alejó el valor de la pendiente de 0, los rechazos fueron cada vez mayores.

Por otra parte en los escenarios cuando la pendiente es distinta de cero, el impacto de la varianza del error, se observa en la Tabla con un menor porcentaje de rechazos, pues al incrementar la dispersión de los datos, se generan nubes de puntos cada vez más dispersas, asemejándose más a una nube de puntos aleatoria, por lo que la medición de la correlación en esos casos es menor, lo que llevó a que un mayor número de casos a no rechazar la hipótesis nula.

Tabla 5.7: Resumen de simulaciones de pruebas de hipótesis para r

% de pruebas rechazadas					
Valores de σ_0^2	Valores de β				
	-0.2	-0.1	0	0.1	0.2
1	100.00%	98.60%	0.96%	98.90%	100.00%
10	68.10%	14.90%	0.86%	14.70%	68.60%
20	35.60%	6.80%	1.04%	7.10%	35.30%

Análisis de varianza en regresión

Con el análisis del coeficiente de correlación se muestra como una porción de la varianza de la variable dependiente se encuentra explicada por el ajuste de regresión,

mientras que existe otra porción de la varianza no se encuentra explicada y se refleja en la suma de cuadrados de los residuos.

Con base en esto la motivación ahora es analizar las fuentes de la varianza y con ellas realizar pruebas de hipótesis en la que se contrasta la hipótesis sobre la nulidad de la pendiente, es decir, que no haya una asociación lineal entre la variables, por lo que se toma como hipótesis nula $\boldsymbol{H_0}: \boldsymbol{\beta} = \mathbf{0}$ contra la hipótesis alternativa $\boldsymbol{H_1}: \boldsymbol{\beta} \neq \mathbf{0}$. Basados en estas hipótesis se necesita partir del estadístico T_β, justo como en la prueba del coeficiente de correlación.

Entonces para realizar la prueba y llevar a cabo un análisis de la varianza, se requiere primero expresar de una forma equivalente a T_β, por lo que se escribe la fórmula del estimador de la varianza del error de la siguiente manera

$$\frac{\widehat{\boldsymbol{\beta}}}{\sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^{n}(X_i-\overline{X})^2}}} = \frac{\widehat{\boldsymbol{\beta}}\sqrt{\sum_{i=1}^{n}(\boldsymbol{X_i}-\overline{\boldsymbol{X}})^2}}{\sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}}$$

Ahora con esta expresión se aplica un teorema de la teoría de probabilidad, el cual enuncia que al elevar al cuadrado una variable T de Student con n grados de libertad, entonces se obtiene una variable con una distribución F con 1 y n grados de libertad, aunque en este caso como la estadística T_β posee $n-2$ grados de libertad entonces

$$\boldsymbol{F} = \frac{\widehat{\boldsymbol{\beta}}^2 \sum_{i=1}^{n}(\boldsymbol{X_i}-\overline{\boldsymbol{X}})^2}{\frac{\sum_{i=1}^{n} e_i^2}{n-2}} \sim \boldsymbol{F_{1,n-2}}$$

Para simplificar las fórmulas, de la sección anterior se retomó la siguiente ecuación

$$\widehat{\boldsymbol{y_i}} = \widehat{\boldsymbol{Y_i}} - \overline{\boldsymbol{Y}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{X_i}-\overline{\boldsymbol{X}})$$

Con la cual el estadístico de prueba se puede expresar como

$$\boldsymbol{F} = \frac{\sum_{i=1}^{n}\left(\widehat{\boldsymbol{Y_i}}-\overline{\boldsymbol{Y}}\right)^2}{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$$

De este estadístico se puede identificar al denominador como la parte de la varianza explicada por la regresión, mientras que el denominador es la parte que no queda explicada por el ajuste de regresión. Usualmente la forma de hacer un análisis sobre esta estadística es desagregándola por cada uno de sus componentes tabulados, generando una tabla de análisis de varianza.

En un formato usual se incluye sólo el cálculo de la suma de cuadrados , sus grados de libertad, los cuadrados medios y el resultado de estadístico F, sin embargo en la tabla que se encuentra a continuación se agregó el porcentaje que representa cada componente de la varianza, relacionándolos con el coeficiente de determinación.

Tabla de varianza					
Prueba de hipótesis H_0:β=0 VS. H_1: β ≠0					
Fuente de Variación	**Suma de cuadrados**	**Grados de libertad**	**% de explicación de varianza**	**Media de los cuadrados**	**Estadístico F**
Regresión	$\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2$	**1**	$r^2=\frac{\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}$	$\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2$	$\frac{\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2}{\frac{\sum_{i=1}^{n}{e_i}^2}{n-2}}$
Residuales	$\sum_{i=1}^{n}{e_i}^2$	**n-2**	$1-r^2=\frac{\sum_{i=1}^{n}{e_i}^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}$	$\frac{\sum_{i=1}^{n}{e_i}^2}{n-2}$	
Total	$\sum_{i=1}^{n}(Y_i-\bar{Y})^2$	**n-1**			

Realizando los cálculos de la Tabla y una vez elegido un nivel de significancia δ, se realiza el contraste de las hipótesis empleando el estadístico F, bajo la siguiente regla de decisión, basada en el cuantil de la distribución F que acumule el $(1-\delta)\%$ de probabilidad $F_{1,n-2,1-\delta}$:

$$\boldsymbol{Rechazar\ H_0:\ \beta=0\ si}$$

$$\boldsymbol{F>F_{1,n-2,1-\delta}}$$

De igual manera que con las pruebas anteriores, se mostrará su comportamiento bajo diversos escenarios por medio de muestras simuladas, que para simplificar el proceso de simulación se puede copiar alguna de las estructuras pasadas en la que se simularon pruebas de hipótesis para los parámetros; en este caso se eligió adaptar una copia de la prueba de hipótesis para la pendiente β.

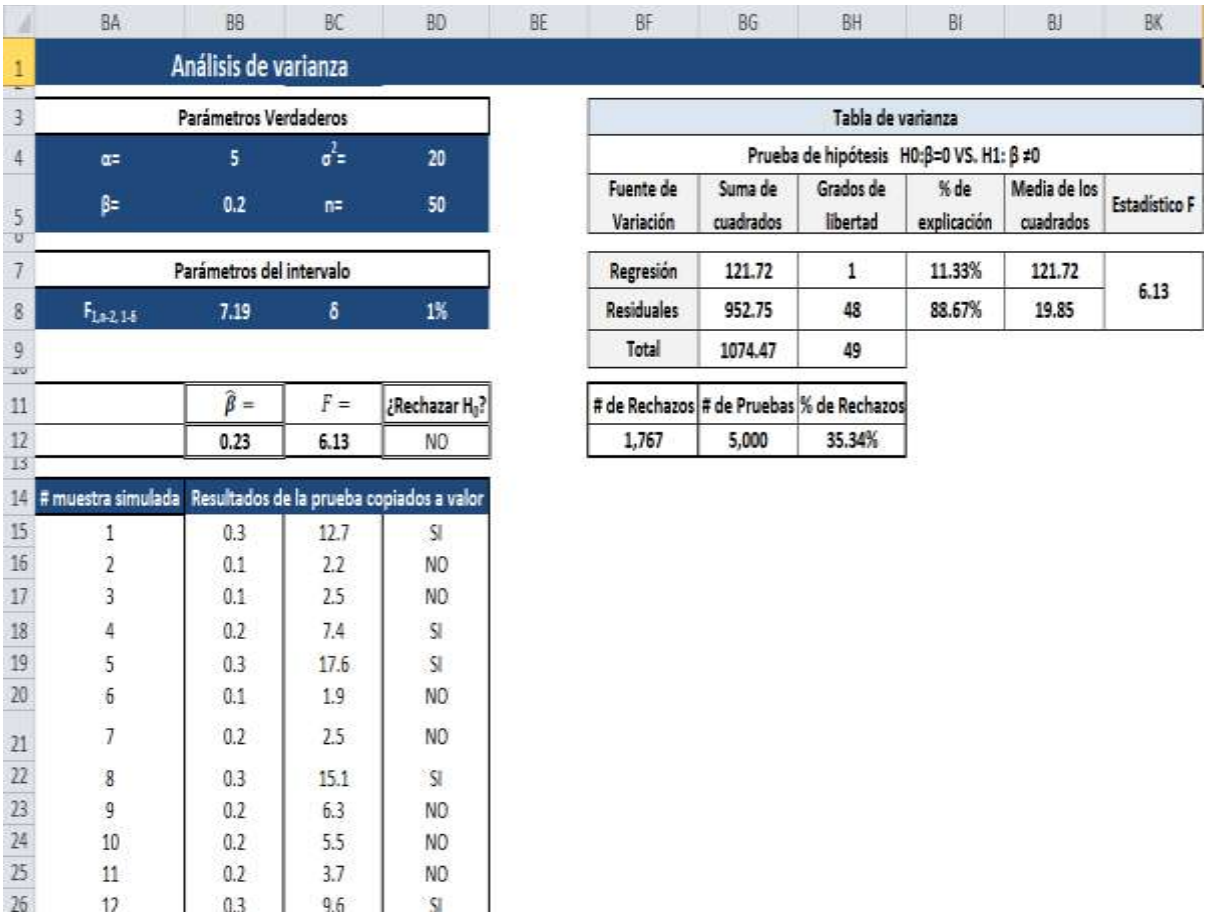
A la hoja de cálculo se adaptaron los cálculos de estadístico de prueba, además de obtener el cuantil de la distribución F por medio de la función INV.F(P, N_1, N_2) , donde el parámetro P es la probabilidad acumulada a la cual se desea ubicar el cuantil , N_1 y N_2 son los grados de libertad definidos para la variable F. En la hoja de cálculo se agregó también la tabla de varianza con los cálculos correspondientes, con lo que en cada simulación individual se podrán observar los resultados del análisis de varianza y el valor del estadístico F, empleado para construir una condicional que determine si se rechaza o no la hipótesis nula, basándose en la regla de decisión anterior.

Por medio de las simulaciones generadas, se decidió comparar el comportamiento de esta prueba contra la prueba de hipótesis sobre la correlación para mostrar su equivalencia, aplicando los mismos escenarios con respecto al valor de la pendiente tomando $\beta=-0.2\,,-0.1\,,0\,,0.1\,,\ 0.2$ con valores para la varianza $\sigma^2=1\,,10\,,20$, realizando 5,000 pruebas simuladas para cada uno de los escenarios propuestos.

En la siguiente imagen se muestra la estructura en Excel® con las modificaciones pertinentes donde se puede observar el resultado de 5000 simulaciones bajo el escenario $\beta=0.2$ y $\sigma^2=20$, en la cual se rechazó la hipótesis nula en el 35% de los casos. Además se puede ver la tabla de varianza, donde se puede observar que para una de las muestras simuladas el coeficiente de determinación fue de 11%, y el valor del estadístico $F=6.13$ cuyo valor no fue suficientemente grande como para rechazar la hipótesis nula.

En la Tabla 5.8 se hallan resumidos los resultados de las simulaciones para cada escenario en los cuales se puede observar que el porcentaje de rechazo en cada caso es muy similar con el de las pruebas realizadas para el coeficiente de correlación, realizadas bajo los mismos parámetros y significancia, de donde se concluye su equivalencia.

Estructura en Excel® para mostrar la tabla de varianza y su prueba de hipótesis asociada



Análisis de varianza

Parámetros Verdaderos			
α=	5	σ^2=	20
β=	0.2	n=	50

Parámetros del intervalo			
$F_{1-\delta/2,\,1-\delta}$	7.19	δ	1%

	$\hat{\beta}$ =	F =	¿Rechazar H_0?
	0.23	6.13	NO

# muestra simulada	Resultados de la prueba copiados a valor		
1	0.3	12.7	SI
2	0.1	2.2	NO
3	0.1	2.5	NO
4	0.2	7.4	SI
5	0.3	17.6	SI
6	0.1	1.9	NO
7	0.2	2.5	NO
8	0.3	15.1	SI
9	0.2	6.3	NO
10	0.2	5.5	NO
11	0.2	3.7	NO
12	0.3	9.6	SI

Tabla de varianza

Prueba de hipótesis H0:β=0 VS. H1: β ≠0

Fuente de Variación	Suma de cuadrados	Grados de libertad	% de explicación	Media de los cuadrados	Estadístico F
Regresión	121.72	1	11.33%	121.72	6.13
Residuales	952.75	48	88.67%	19.85	
Total	1074.47	49			

# de Rechazos	# de Pruebas	% de Rechazos
1,767	5,000	35.34%

Tabla 5.8: Resumen de simulaciones de la prueba H_0:β=0 VS. H_1: β ≠0 , empleando el estadístico F

% de pruebas rechazadas					
Valores de σ_0^2	Valores de β				
	-0.2	-0.1	0	0.1	0.2
1	100.00%	98.60%	0.92%	98.30%	100.00%
10	69.00%	15.40%	0.88%	14.80%	67.70%
20	34.50%	6.80%	0.96%	7.10%	35.30%

Con la estructura anterior se podrá comprender el impacto de los parámetros en la prueba de hipótesis sobre la nulidad de la pendiente, pues a través de generar distintos escenarios se observará cuál es la contribución de la variable independiente, en la explicación de la varianza de la variable dependiente. La compresión de la tabla de varianza es crucial en el análisis del modelo, ya que en diversos programas

estadísticos al realizar un ajuste de regresión, suele devolverse también una tabla de análisis de varianza, la cual contendrá elementos para evaluar si la pendiente de la recta de regresión es nula a través de la prueba de hipótesis anterior, y realizar análisis de las fuentes varianza.

Estimación del valor medio y puntual de Y a partir del modelo ajustado

Luego de ajustar un modelo de regresión lineal simple a un conjunto de datos, es de interés ahora estimar, a partir de un valor particular en el rango donde se encuentra definida la variable independiente, ya sea que se encuentre o no en los datos observados, una predicción sobre el valor medio y puntual de la variable dependiente.

Cuando se toma un punto en particular de la variable dependiente X_0, se puede demostrar que la verdadera media de Y dado X_0, denotado de la misma que en la teoría de probabilidad como $E(Y|X_0)$, está determinada por

$$\boldsymbol{E(Y|X_0) = E(\alpha + \beta X_0 + u)}$$

$$\boldsymbol{E(Y|X_0) = \alpha + \beta X_0 + E(u)}$$

$$\boldsymbol{E(Y|X_0) = \alpha + \beta X_0}$$

Entonces para estimar este valor medio se sustituyen los parámetros α y β por sus estimadores $\hat{\alpha}$ y $\hat{\beta}$ ya que como se ha mostrado anteriormente ambos estimadores son insesgados, por lo cual al tomar la esperanza de $\hat{\alpha} + \hat{\beta}X_0$ se obtendrá la expresión deseada. A este valor estimado de la media se le denota como $\widehat{Y_0}|X_0$ y se define entonces como

$$\boldsymbol{\widehat{Y_0}|X_0 = \hat{\alpha} + \hat{\beta}X_0}$$

Luego a partir de este estimador se obtiene su varianza, para la cual se retoma la relación obtenida de restar a un valor estimado $\hat{Y}$, la media de Y

$$\boldsymbol{\widehat{Y_0}|X_0 - \overline{Y} = \hat{\alpha} + \hat{\beta}X_0 - \overline{Y}}$$

$$\boldsymbol{puesto\ que\ \hat{\alpha} = \overline{Y} - \beta\overline{X}\ \rightarrow}$$

$$\boldsymbol{\widehat{Y_0}|X_0 - \overline{Y} = \hat{\beta}(X_0 - \overline{X})}$$

$$\boldsymbol{\widehat{Y_0}|X_0 = \overline{Y} + \hat{\beta}(X_0 - \overline{X})}$$

Ahora con esta nueva expresión se obtiene la varianza del estimador como

$$\boldsymbol{Var\left(\widehat{Y_0}|X_0\right) = Var(\overline{Y}) + (X_0 - \overline{X})^2 Var\left(\hat{\beta}\right) + 2(X_0 - \overline{X})Cov\,(\overline{Y}, \hat{\beta})}$$

Por una parte es verificable que $Cov\left(\overline{Y}, \hat{\beta}\right) = 0$, además los otros términos con ambas varianzas son conocidas, $\boldsymbol{Var(\overline{Y}) = \frac{\sigma^2}{n}}$ y $\boldsymbol{Var\left(\hat{\beta}\right) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$ entonces al sustituir sobre la expresión anterior se obtiene

$$\boldsymbol{Var(\widehat{Y_0}|X_0) = \left(\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)\sigma^2}$$

Del cálculo de la varianza del valor medio, se puede observar que el punto donde se minimiza la varianza, es cuando se toma el valor $X_0 = \bar{X}$ donde se sabe que el valor estimado $\widehat{Y_0}|\bar{X}$ es $\bar{Y}$, mientras que a medida que se estime $\widehat{Y_0}|X_0$ a partir de un punto más alejado de la media de X, la varianza será cada vez mayor.

Cuando lo que se desea es estimar un valor puntual de la variable independiente Y_0 a partir de un valor X_0, entonces se supone que Y_0 es resultado del modelo lineal $\alpha + \beta X_0$, más el termino de error asociado al modelo u_0, el cual se considera, como los demás términos de error, se distribuye como una Normal con media 0 y varianza σ^2 , por lo tanto el valor a estimar es

$$\boldsymbol{Y_0 = \alpha + \beta X_0 + u_0}$$

Puesto que la media de Y_0 es $\alpha + \beta X_0$, el estimador insesgado para la nueva predicción es el mismo que para el valor medio $\widehat{Y_0}|X_0 = \hat{\alpha} + \hat{\beta} X_0$, por lo que al tomarse la esperanza de la diferencia entre Y_0 y $\widehat{Y_0}|X_0$ se tiene

$$\boldsymbol{E\left(Y_0 - \widehat{Y_0}|X_0\right) = 0}$$

A esta diferencia $Y_0 - \widehat{Y_0}|X_0$ también se le conoce como el error de la predicción. Como la precisión del estimador depende de la magnitud del error de la predicción, entonces calcular la varianza del error será equivalente a calcular la varianza asociada a la nueva predicción, por lo que la varianza de la predicción se calcula como

$$\boldsymbol{Var\left(Y_0 - \widehat{Y_0}|X_0\right) = Var\,(Y_0) + Var\left(\widehat{Y_0}|X_0\right) - 2Cov(Y_0, \widehat{Y_0}|X_0)}$$

Donde es verificable que $Cov\left(Y_0, \widehat{Y_0}|X_0\right) = 0$, debido a Y_0 depende sólo del error u_0 y el valor estimado $\boldsymbol{\widehat{Y_0}|X_0}$ puede expresarse en términos de una suma ponderada de las Y_i.[31] por lo tanto

$$\boldsymbol{Var\left(Y_0 - \widehat{Y_0}|X_0\right) = \sigma^2 + \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)\sigma^2}$$

$$\boldsymbol{Var\left(Y_0 - \widehat{Y_0}|X_0\right) = \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)\sigma^2}$$

De la expresión para la varianza de la predicción, se puede observar que de igual manera que la varianza del valor medio, $Var\left(Y_0 - \widehat{Y_0}|X_0\right)$ se minimiza en el punto $X_0 = \bar{X}$, a partir del cual a medida que se tome el punto X_0 más lejos de $\bar{X}$, la varianza será cada vez mayor.

Intervalos de confianza para la predicción y el valor medio

En cada caso los estimadores de la observación puntual y del valor medio son una combinación lineal de los estimadores $\hat{\alpha}$, $\hat{\beta}$ los cuales como se había visto antes, poseen una distribución Normal, por esta razón la estimación de la predicción y el valor medio también se distribuyen de manera Normal. Ahora si, en aras de diferenciar estas dos estimaciones, se denota a la predicción como $Y_{pred}|X_0$ se tiene entonces

[31] Para una justificación completa de este hecho revise el libro de Rawlings sección 1.5

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{pred}}|\boldsymbol{X_0}\sim\boldsymbol{N}\left(\boldsymbol{\alpha}+\boldsymbol{\beta X_0}\,,\left(\mathbf{1}+\frac{\mathbf{1}}{\boldsymbol{n}}+\frac{(\boldsymbol{X_0}-\overline{\boldsymbol{X}})^{\mathbf{2}}}{\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}(\boldsymbol{X_i}-\overline{\boldsymbol{X}})^{\mathbf{2}}}\right)\boldsymbol{\sigma}^{\mathbf{2}}\right)$$

Mientras que en el caso del valor $\widehat{Y_0}|X_0$ su distribución es

$$\widehat{\boldsymbol{Y_0}}|\boldsymbol{X_0}\sim\boldsymbol{N}\left(\boldsymbol{\alpha}+\boldsymbol{\beta X_0}\,,\left(\frac{\mathbf{1}}{\boldsymbol{n}}+\frac{(\boldsymbol{X_0}-\overline{\boldsymbol{X}})^{\mathbf{2}}}{\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}(\boldsymbol{X_i}-\overline{\boldsymbol{X}})^{\mathbf{2}}}\right)\boldsymbol{\sigma}^{\mathbf{2}}\right)$$

Para determinar los intervalos de confianza se requiere una vez más el empleo del estadístico $\frac{(n-2)\widehat{\sigma^2}}{\sigma^2}$, pues al distribuirse Ji-cuadrada, fue usado en los intervalos para los parámetros $\alpha,\ \beta$, para generar una variable t de Student a partir de una Normal. Aplicando el mismo procedimiento para el cálculo del intervalo se puede observar que análogamente a los casos anteriores se estandariza la variable Normal restando la estimación $\hat{\alpha}+\hat{\beta}X_0$ y dividiendo entre la raíz cuadrada de su varianza, para después dividir entre la estadística distribuida Ji-cuadrada, lo que al final cancela el parámetro σ sustituyéndolo por $\hat{\sigma}$.

En el caso de realizar este procedimiento sobre la distribución del valor medio, se obtiene la siguiente variable t de *Student* con $n-2$ grados de libertad.

$$\frac{\widehat{\boldsymbol{\alpha}}+\widehat{\boldsymbol{\beta}}\boldsymbol{X_0}-(\boldsymbol{\alpha}+\boldsymbol{\beta X_0})}{\sqrt{\frac{\mathbf{1}}{\boldsymbol{n}}+\frac{(\boldsymbol{X_0}-\overline{\boldsymbol{X}})^{\mathbf{2}}}{\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}(\boldsymbol{X_i}-\overline{\boldsymbol{X}})^{\mathbf{2}}}\,\widehat{\boldsymbol{\sigma}}^{\mathbf{2}}}}\sim\boldsymbol{t_{n-2}}$$

Después al definir un nivel de confianza $(1-\delta)\%$ el intervalo queda definido de forma simétrica alrededor de $\hat{\alpha}+\hat{\beta}X_0$, $\pm$ la raíz cuadrada de la varianza estimada en base a $\hat{\sigma}$, por el cuantil de la distribución t que acumula $\left(1-\frac{\delta}{2}\right)\%$ de probabilidad, por lo que en el caso del intervalo para el valor medio $\widehat{Y_0}|X_0$ queda definido como

$$\left(\widehat{\boldsymbol{\alpha}}+\widehat{\boldsymbol{\beta}}\boldsymbol{X_0}\ \pm\ \boldsymbol{t}_{\boldsymbol{n-2},\mathbf{1}-\frac{\boldsymbol{\delta}}{\mathbf{2}}}\sqrt{\frac{\mathbf{1}}{\boldsymbol{n}}+\frac{(\boldsymbol{X_0}-\overline{\boldsymbol{X}})^{\mathbf{2}}}{\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}(\boldsymbol{X_i}-\overline{\boldsymbol{X}})^{\mathbf{2}}}\widehat{\boldsymbol{\sigma}}^{\mathbf{2}}}\ \right)$$

Mientras que en el caso de la estimación de la predicción puntual $\widehat{Y}_{pred}|X_0$ el intervalo cambia ligeramente por la forma de su varianza, siendo el intervalo de la siguiente forma

$$\left(\widehat{\boldsymbol{\alpha}}+\widehat{\boldsymbol{\beta}}\boldsymbol{X_0}\ \pm\ \boldsymbol{t}_{\boldsymbol{n-2},\mathbf{1}-\frac{\boldsymbol{\delta}}{\mathbf{2}}}\sqrt{\mathbf{1}+\frac{\mathbf{1}}{\boldsymbol{n}}+\frac{(\boldsymbol{X_0}-\overline{\boldsymbol{X}})^{\mathbf{2}}}{\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}(\boldsymbol{X_i}-\overline{\boldsymbol{X}})^{\mathbf{2}}}\,\widehat{\boldsymbol{\sigma}}^{\mathbf{2}}}\ \right)$$

Ahora para mostrar el cálculo y comportamiento de estos intervalos, se necesita desarrollar una estructura en la hoja de cálculo diferente a las generadas anteriormente, ya que ahora es posible establecer intervalos para una serie de puntos en el soporte de X, por lo que es necesario tener los cálculos desglosados por cada una de los valores elegidos, a partir de los cuales se hacen las estimaciones de los límites de cada intervalo de confianza, que como se verá más adelante, al unirlos en

un gráfico generarán una banda de confianza, donde la banda que se dibuja con los límites de las predicciones será para las parejas (X_0, Y_0), mientras que la banda de confianza del valor medio será sobre la recta de regresión.

Lo primero que se requiere para calcular los intervalos, es generar una tabla en la que se hagan referencias a los cálculos de los estimadores que se encuentran en hojas de cálculo previamente construida, además de otra tabla para los parámetros verdaderos, luego se define el valor δ partir del cual se obtiene el cuantil $t_{n-2,1-\frac{\delta}{2}}$.

Un detalle dentro de las gráficas de Excel®, es que para poder trazar correctamente las bandas de confianza en un diagrama de dispersión es necesario que los valores se encuentren ordenados. Para lograr lo anterior se realiza es una partición del soporte de la variable X, pues como se había mencionado antes, los valores X_0 pueden no ser necesariamente parte de la muestra.

Para conseguir la partición primero se debe calcular el mínimo y el máximo de los valores de las X_i, por medio de la función MIN(X_i) y MAX(X_i) de Excel®. Posteriormente se generan k nuevos valores de la variable X, a los cuales se les denotará como ${X_0}^{(i)}$, por medio de la fórmula recursiva

$$\boldsymbol{{X_0}^{(i)} = {X_0}^{(i-1)} + \frac{\mathbf{MAX}(X_i) - \mathbf{MIN}(X_i)}{k} \qquad i := 2, 3, \dots, k}$$

$$\boldsymbol{{X_0}^{(1)} = \mathbf{MIN}(X_i)}$$

Donde lo que se está realizando es una partición uniforme sobre el soporte de los valores de la variable independiente. Después, a partir de cada uno de los valores computados, se calcula la estimación de la variable dependiente para cada uno de los puntos como $\widehat{Y_0}|{X_0}^{(i)} = \hat{\alpha} + \hat{\beta}{X_0}^{(i)}$, los cuales son la base para generar los intervalos de confianza pues se debe sumar y restar el término de variabilidad de las formulas anteriores, según sea el caso del límite superior o inferior para la predicción o para el valor medio.

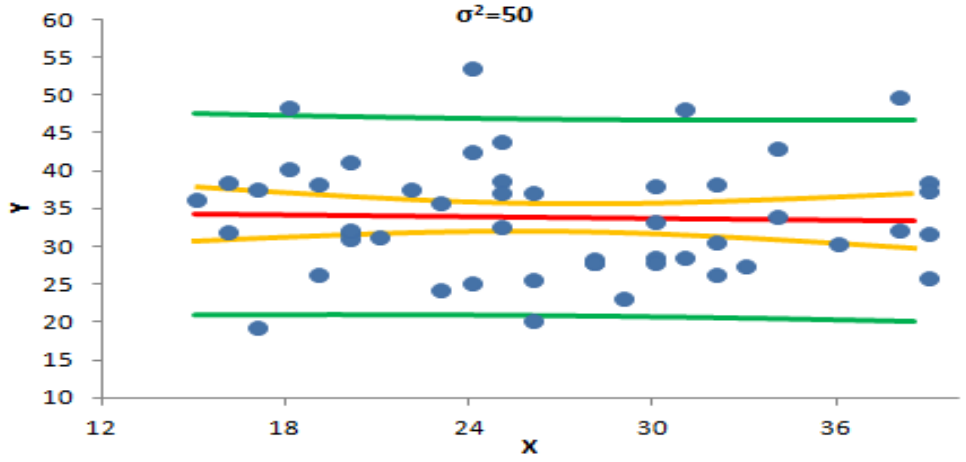
Una vez computadas estas columnas se grafican las series correspondientes a las estimaciones $\widehat{Y_0}|{X_0}^{(i)}$ junto con los límites de los intervalos de confianza, a los cuales se les agrega además la serie de datos originalmente simulados (X_i, Y_i), obteniéndose una gráfica de los datos con las bandas de confianza correspondientes y la recta de regresión.

En la siguiente imagen se encuentra una propuesta para la estructura de los cálculos anteriores, en la que se tomó un nivel de confianza del 90% ($\delta = 10\%$) para ambas bandas, y el valor de los parámetros de regresión fueron $\alpha = 35$, $\beta = 1.5$ y $\sigma^2 = 300$, donde se puede observar que se agregó la gráfica con los límites y la estimación adjunta a los datos. En el gráfico de la imagen se nota que la banda de confianza para el valor medio trazada en color amarillo toma la forma de una hipérbola, mientras que la banda que se genera para los valores puntuales trazada en color verde se asemeja a un par de líneas rectas.

Las bandas sobre los datos puntuales al estar calculados sobre el 90% de confianza se espera que para una muestra de tamaño $n = 50$, como la simulada en la imagen, la banda cubra en cada muestra aproximadamente 45 de los datos simulados (50*90% =45).

Diversos escenarios para los valores de los parámetros fueron generados, tomando el mismo nivel de confianza del 90%, los cuales se hallan en las gráficas 5.20 a 5.22, donde es destacable que al variar los parámetros lo datos en ciertos casos se vuelven más dispersos, sin embargo, las bandas de confianza en cada caso se recalculan para que en su espectro contengan aproximadamente el 90% de los datos.

Gráfica 5.20: Bandas de confianza generadas sobre el valor medio y puntuales de los datos simulados, α= 35, β=0 , σ²=50



Estructura en Excel® para mostrar las bandas de confianza en el modelo de regresión.

Intervalos de confianza para la predicción

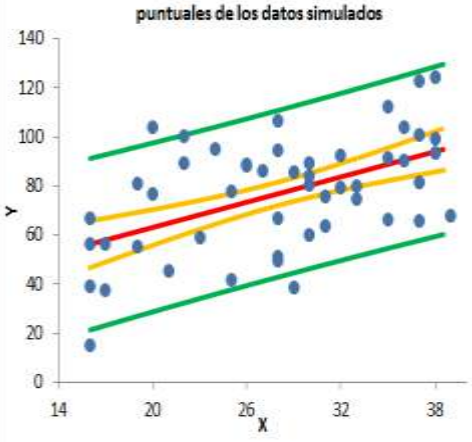
Parámetros Verdaderos			
α=	35	σ^2=	300
β=	1.5	n=	50

Parametros estimados			
$\hat{\beta}$ =	1.72	$Var(\hat{\beta})$ =	0.165
$\hat{\alpha}$ =	28.78	$Var(\hat{\alpha})$ =	138.43
$\widehat{\sigma^2}$ =	402.00	$Cov(\hat{\alpha},\hat{\beta})$ =	-3.46

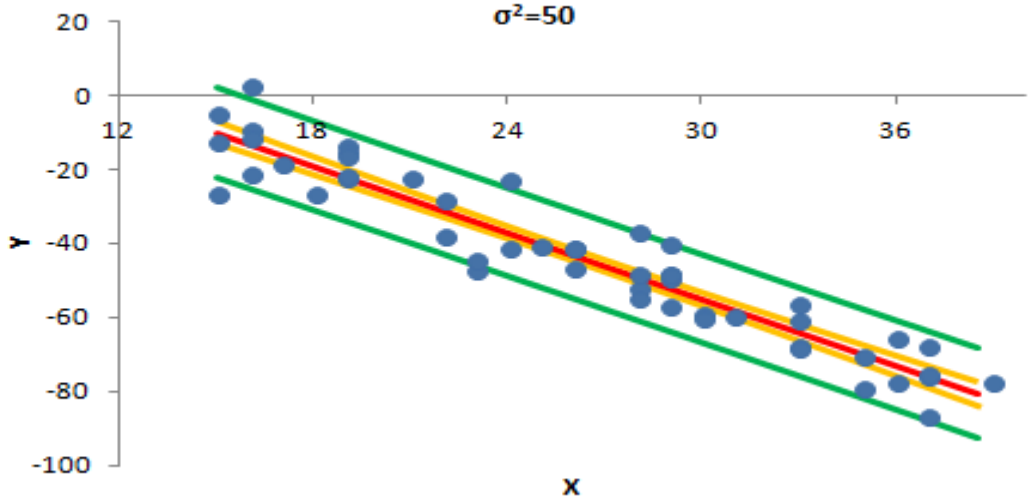
Parámetros de los intervalos de confianza			
$t_{n-2,\,1-\delta/2}$	1.68	δ	10%

		Estimación valor medio		Estimación puntual	
		Limite Inferior	Limite Superior	Limite Inferior	Limite Superior
$X_0^{(i)}$	$\hat{Y}_0\|X_0^{(i)}$	$\hat{Y}_0\|X_0^{(i)}$	$\hat{Y}_0\|X_0^{(i)}$	$\hat{Y}_{pred}\|X_0^{(i)}$	$\hat{Y}_{pred}\|X_0^{(i)}$
16.00	56.25	46.7	65.8	21.3	91.2
16.46	57.04	47.8	66.3	22.2	91.9
16.92	57.83	48.8	66.8	23.0	92.6
17.38	58.62	49.9	67.3	23.9	93.4
17.84	59.41	50.9	67.9	24.7	94.1
18.30	60.20	52.0	68.4	25.6	94.8
18.76	60.99	53.0	68.9	26.4	95.5
19.22	61.77	54.1	69.5	27.3	96.3
19.68	62.56	55.1	70.0	28.1	97.0
20.14	63.35	56.1	70.6	29.0	97.7
20.60	64.14	57.2	71.1	29.8	98.5
21.06	64.93	58.2	71.7	30.6	99.2
21.52	65.72	59.2	72.3	31.5	100.0
21.98	66.51	60.2	72.8	32.3	100.7
22.44	67.30	61.2	73.4	33.1	101.5
22.90	68.09	62.2	74.0	33.9	102.2

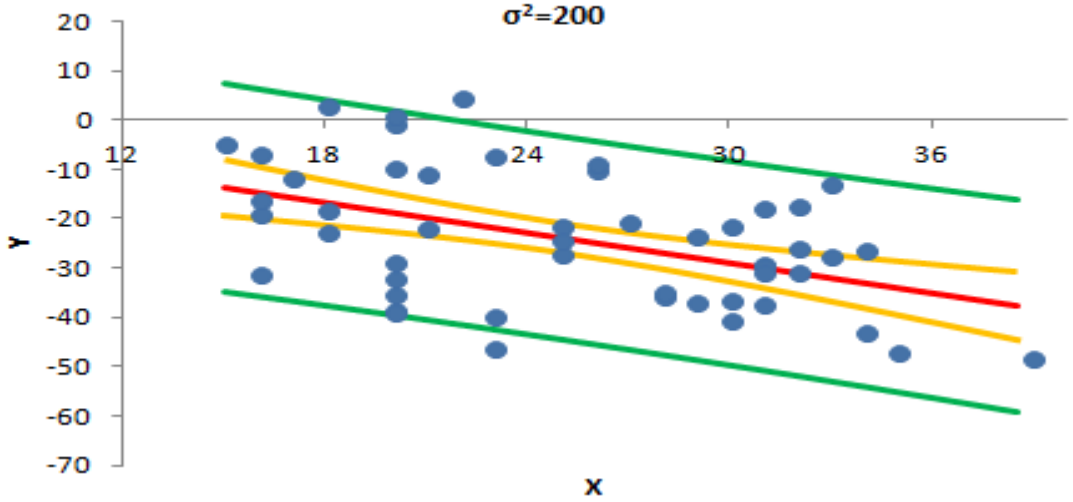
Bandas de confianza generadas sobre le valor medio y puntuales de los datos simulados



Gráfica 5.21: Bandas de confianza generadas sobre el valor medio y puntuales de los datos simulados, α= 35, β=-3 , σ^2=50



Gráfica 5.22: Bandas de confianza generadas sobre el valor medio y puntuales de los datos simulados, α= 0, β=-1 , σ^2=200



Análisis de residuos

Como se enunció al inicio del capítulo el modelo de regresión lineal simple tiene varios supuestos principales sobre las características del término de error asociado a la variable independiente, como son: distribuirse Normal, tener una varianza homogénea y que no hubiera correlación entre los errores.

Sin embargo, existe la posibilidad de que en los datos recabados, alguno o más de estos supuestos no se cumplan en ciertos fenómenos observados, por lo que en esta sección se revisará la forma de identificar y simular datos que no cumplan alguno de los supuestos, generando diversos ejemplos prácticos.

Uno de los supuestos que se pueden evaluar de forma aislada sobre los errores es su varianza, la cual se supone como una constante σ^2 en el soporte de la variable X. Este comportamiento es también conocido como homoscedasticidad, sin embargo en ciertas ocasiones los datos muestran una dispersión desigual en alguna o varias

secciones sobre todo el soporte de X, comportamiento conocido como heteroscedasticidad.

Simulación Monte Carlo de datos con heteroscedasticidad

Para la siguiente simulación de datos con un cierto grado de heteroscedasticidad, se debe recordar que las simulaciones de los errores aleatorios u_i anteriormente realizadas, se consideró a un valor σ^2 fijo para cada valor generado, sin embargo, para romper el supuesto ahora la varianza se propone como una función lineal de la variable independiente X, que dependa de un parámetro Q , que funcione como el grado de sensibilidad de la heteroscedasticidad de la varianza de acuerdo al valor de la variable independiente.

En la definición de la función lineal de la varianza, lo que se desea es que la simulación del error en el punto mínimo del soporte de X la varianza del error sea σ^2, y en el punto máximo del soporte sea una proporción de la varianza inicial $\sigma^2(1+Q\%)$ donde Q es el parámetro de sensibilidad que define un incremento lineal de la varianza cuando $Q>0$ y genera un decremento lineal cuando $Q<0$.

El parámetro Q tiene la restricción $Q>-100$ ya que en caso contrario se llega a definir valores negativos o cero para la varianza lo cual arroja un error en el modelo y en la hoja de cálculo. Para definir la fórmula de la varianza heterogénea se requiere obtener tanto el mínimo y el máximo de los valores X_i, puesto que estos valores fueron construidos como enteros aleatorios, el punto inicial del soporte de X se debe obtener como $Min(X_i)$ y el punto final de las X_i se calcula como $Max(X_i)$. Con estos elementos, para establecer la ecuación de la varianza, se debe retomar que los puntos esenciales que se desean tocar en la función lineal de la varianza son

$$\boldsymbol{P_1 = (Min(X_i), \sigma^2)}$$

$$\boldsymbol{P_2 = (Max(X_i), \sigma^2(1+Q\%))}$$

Ahora empleando la ecuación de una recta que pase por estos puntos valuada para alguna X_i, se obtiene el valor de la varianza heterogénea, que se denotará como $\sigma^2_{(X_i)}$.

$$\boldsymbol{\sigma^2_{(X_i)} - \sigma^2 = \frac{\sigma^2(1+Q\%) - \sigma^2}{Max(X_i) - Min(X_i)}\left(X_i - Min(X_i)\right)}$$

$$\boldsymbol{\sigma^2_{(X_i)} = \frac{\left(X_i - Min(X_i)\right)}{Max(X_i) - Min(X_i)}\sigma^2(Q\%) + \sigma^2}$$

Con esta relación se puede sustituir el valor de la varianza en la simulación de los errores, obteniéndose datos con diversos grados de heteroscedasticidad, de acuerdo al valor de Q. Para construir el uso de la relación anterior en una hoja de cálculo, se puede copiar la estructura generada inicialmente para la simulación de los datos, donde se tenían los parámetros α, β definidos, los valores de la variable independiente, el error aleatorio simulado en una columna independiente como una Normal$(0, \sigma^2)$, y los resultados de simular la variable dependiente.

Una vez con esto, se debe colocar el parámetro Q en una celda y modificar la fórmula de la simulación del error realizada anteriormente, para la cual ahora se deben elegir

los parámetros para tener media 0 y varianza definida con la fórmula de $\sigma^2_{(X_i)}$, donde el mínimo y máximo de los valores de X, que se hallan en la fórmula se recomienda fijar el argumento del rango de celdas de las X_i, con los símbolos $ $ o por medio de la tecla $F4$.

Una vez realizado esto, se debe graficar después la serie de datos simulados (X_i, Y_i), donde se podrá observar una dispersión creciente o decreciente de acuerdo al valor del parámetro Q. Un ejemplo de una simulación generada, se muestra en la siguiente imagen, donde se cambiaron el rango de valores en el que se simulan los valores de la variable X_i como enteros aleatorios entre los valores $(100{,}200)$; mientras que en el caso de los parámetros de la relación lineal se tomaron $\alpha = 5$, $\beta = 3$ y $\sigma^2 = 200$.

Para generar una varianza notoriamente creciente se tomó el valor de $Q = 2{,}000\%$, lo cual implica que los datos en el punto mínimo del soporte de X el error se simula como una $N(0{,}200)$ mientras que al extremo derecho las X_i se simula una distribución $N(0, (1+20) * 200 = 4{,}200)$, reflejado en la gráfica de la imagen adjunta a las columnas computadas, donde se observa cómo la nube datos tiene una mayor dispersión aproximada de $\sqrt{4200} = 64.8$ al final del soporte. Conforme se experimente con diversos valores de Q podrá darse cuenta que se necesitan valores altos para que el efecto de heteroscedasticidad sea lo suficientemente evidente.

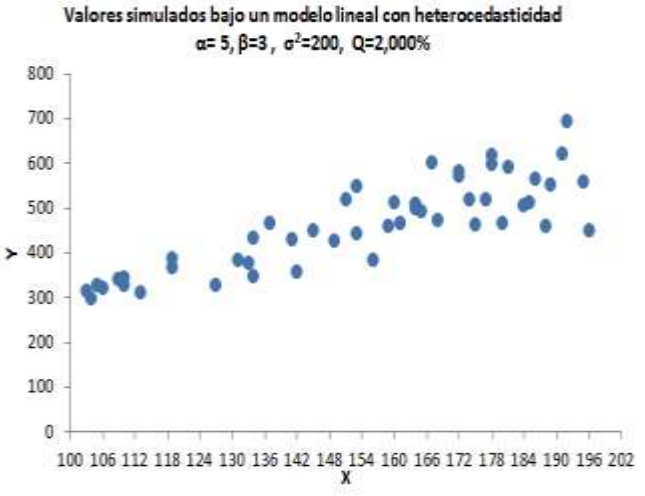
Estructura en Excel® para mostrar el modelo de regresión con heteroscedasticidad

Simulación de datos con heterocedasticidad

Parámetros Verdaderos			
α=	5	Q =	2000%
β=	3		
σ^2=	200		

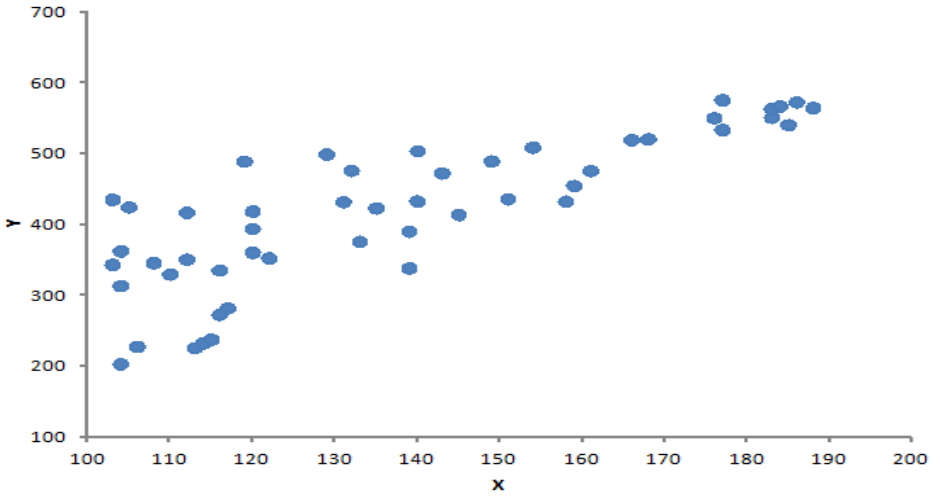
$$\sigma^2_{(X_i)} = \frac{(X_i - Min(X_i))}{Max(X_i) - Min(X_i)} \sigma^2(Q\%) + \sigma^2$$

i	Xi	ui~N(0, σ^2_{Xi})	Yi=α+βXi+ ui
1	180	-79.39	465.61
2	189	-20.77	551.23
3	103	-0.65	313.35
4	145	8.77	448.77
5	175	-67.11	462.89
6	188	-108.28	460.72
7	178	57.64	596.64
8	137	51.37	467.37
9	185	-48.81	511.19
10	131	-13.31	384.69
11	196	-143.48	449.52
12	161	-21.72	466.28
13	110	-6.96	328.04
14	106	-0.36	322.64
15	149	-25.01	426.99
16	156	-89.53	383.47
17	153	85.51	549.51
18	142	-72.03	358.97



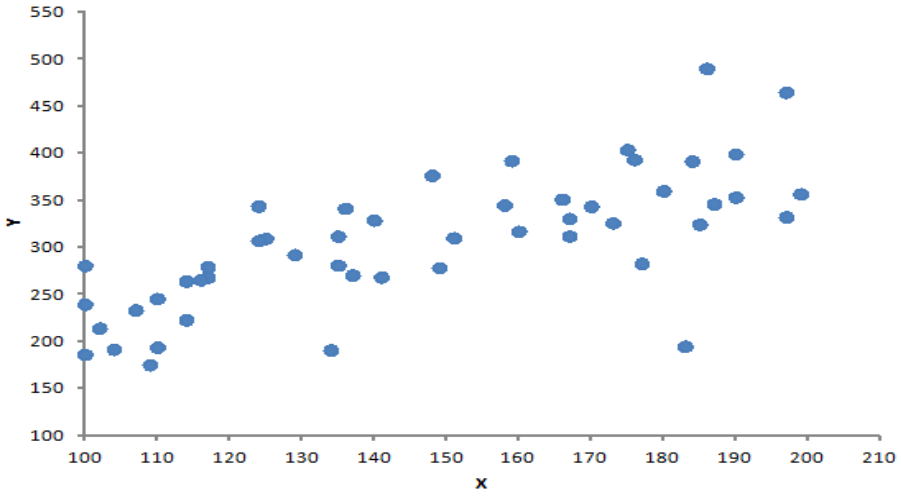
Ahora en otro escenario en la Gráfica 5.23 se muestra la nube de datos tras considerar, el caso extremo de un valor negativo del parámetro $Q = -99\%$ para generar en los datos una varianza linealmente decreciente, donde se observa cómo la varianza al extremo derecho del soporte de X es del 1% de la varianza original, considerando un valor inicial para $\sigma^2 = 4{,}000$ para poder mostrar de forma más evidente el decrecimiento, que se encuentra reflejado en como la nube de datos tiene un comportamiento de dispersión en un sentido inverso al primer escenario.

Gráfica 5.23: Muestra generada del modelo lineal con heteroscedasticidad, con α=5, β=3, σ^2=4,000, Q=-99%



Por otra parte en un tercer caso, mostrado en la Gráfica 5.24 se tomó el parámetro $Q = 100\%$ con un parámetro de varianza inicial $\sigma^2 = 2{,}000$, cambiando además el valor del parámetro de la pendiente $\beta = 2$; como resultado se puede observar, como se puede detectar de manera visual una varianza creciente, sin embargo no tan notoria, comparándolo contra el primer escenario, mostrando como este método proporciona el control que se puede ejercer en el grado de heteroscedasticidad de los datos.

Gráfica 5.24: Muestra generada del modelo lineal con heteroscedasticidad, con α=5, β=2, σ^2=2,000, Q=100%



Una vez con los datos simulados con las características deseadas, se debe realizar la comprobación por medio de los pasos necesarios para realizar el ajuste de regresión vistos anteriormente, como el cálculo de la estimación de los parámetros, las estimaciones $\widehat{Y}_i$, hasta el cálculo de los residuos, con lo cual se puede generar todas las muestras que se deseen, con diversos grados de heteroscedasticidad.

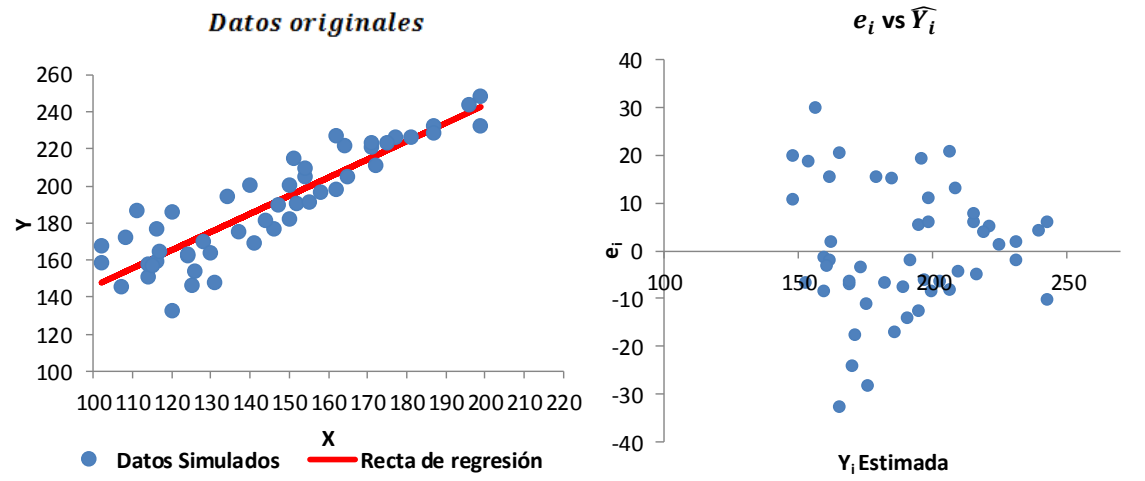
Método gráfico para detectar heteroscedasticidad.

En las simulaciones anteriores, se revisaron escenarios con varianza heterogénea, en las cuales se eligieron valores muy altos o muy bajos para el parámetro Q, para que se pudiera notar a simple vista la creciente dispersión en los datos, sin embargo, cuando el grado de heteroscedasticidad es menor, se puede emplear un gráfico para detectar con mayor sensibilidad una varianza heterogénea.

El siguiente gráfico se construye a partir de realizar un diagrama de dispersión colocando los residuos e_i en el eje Y y en el eje X los valores ajustados por la regresión $\widehat{Y_i}$, lo cual visualiza el comportamiento de los residuos a través de la línea de regresión. Para generar ejemplos basta con remitirse a la hoja de cálculo anteriormente generada, e insertar el diagrama de dispersión seleccionando como insumo las columnas correspondientes.

En las gráficas 5.26 a 5.28 se encuentran tanto la nube de datos original como el diagrama de dispersión de e_i vs. $\widehat{Y}_i$, donde para el modelo lineal se consideraron los parámetros $\alpha = 50$, $\beta = 1$, $\sigma^2 = 250$, variando en cada grafico el valor del parámetro de heterocedasticidad $Q = -70\%, 0\%, 400\%$, con el objetivo de mostrar escenarios en los cuales los datos originales no permitan la detección de una varianza heterogénea, no obstante, en el gráfico de residuos se detecte la variación de la dispersión de los datos, como en el caso del construido en la hoja de datos anterior.

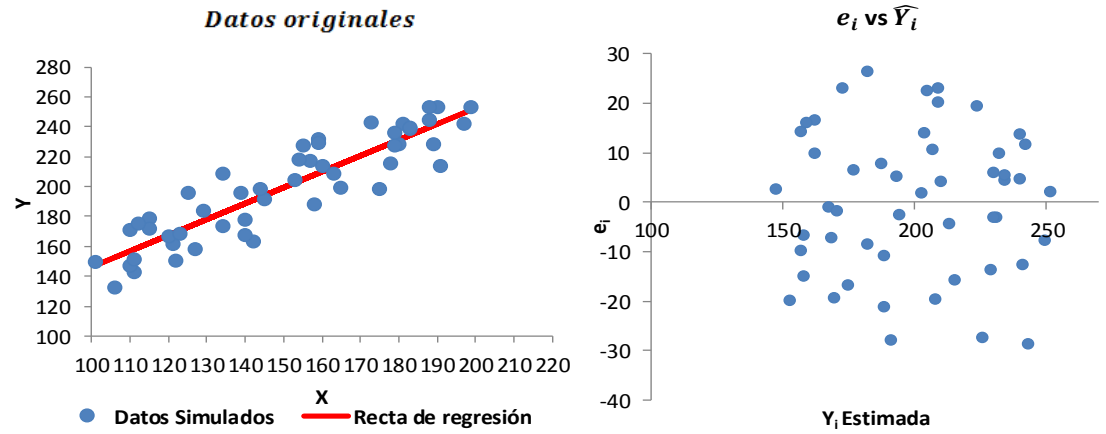
Gráfica 5.26: Datos originales tomando, α= 50, β= 1 , σ^2=250 , Q=-70%, junto con la dispersión de e_i VS la estimación



En la Gráfica 5.26 se puede observar como la relación lineal de la varianza, se nota con menor claridad en el gráfico izquierdo mientras que en el gráfico de residuos VS $\widehat{Y}_i$ se nota de forma más evidente cómo la varianza de los residuos decrece conforme aumenta el valor de la estimación.

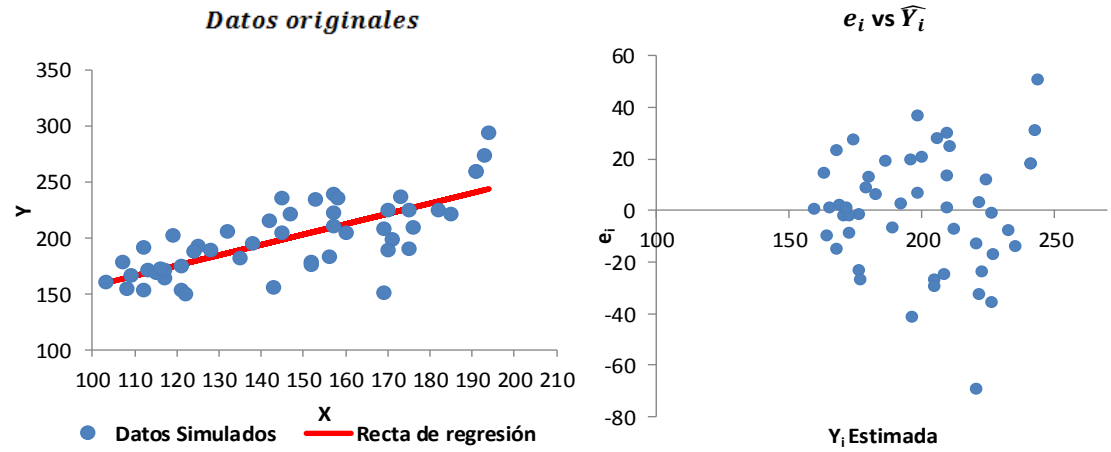
En el siguiente escenario se consideró una varianza constante al elegir $Q = 0\%$, lo cual se refleja en la dispersión homogénea de la nube de datos y cómo la variación entre los residuos parece ser la misma a lo largo del soporte de las $\widehat{Y}_i$. Este caso se considera como el comportamiento esperado de la gráfica cuando se cumple el supuesto de una varianza homogénea. Como se puede observar en la Gráfica 5.27 en la realización particular mostrada del lado derecho los datos parecen poder cubrirse si se rodean por un círculo, aunque, en generaciones posteriores de muestras con $Q = 0$ esta gráfica mostrará patrones de una nube aleatoria.

Gráfica 5.27: Datos originales tomando, α= 50, β= 1 , σ^2=250 , Q=0%, junto con la dispersión de e_i VS la estimación



Ahora en el tercer escenario tomando el parámetro $Q = 400\%$, se tiene una varianza creciente linealmente, que análogamente al primer caso se logra observar de una forma más clara estos comportamientos en el gráfico de residuos VS $\widehat{Y}_i$.

Gráfica 5.28: Datos originales tomando, α= 50, β= 1 , σ^2=250 , Q=400%, junto con la dispersión de e_i VS la estimación



Prueba de Breusch-Pagan para la detección de Heteroscedasticidad

T. S. Breusch y A. R. Pagan desarrollaron una prueba de hipótesis a finales de los años 70, la cual es utilizada ampliamente en temas econométricos, para la detección de un comportamiento con heteroscedasticidad en un conjunto de datos, es decir, que su varianza no sea constante. En esta prueba los autores consideran como principal

supuesto que la varianza del modelo de regresión no es constante y se puede describir por medio de una función h que dependa del tiempo t

$$\boldsymbol{\sigma_t^2 = h(Z_t \delta_i) \quad i =: 1, 2, \dots, n}$$

Donde la variable Z_t, se considera como la unidad cuando $i = 1$ y para otros valores de i, como una selección de las variables independientes del modelo, que en este caso al tratarse del modelo de regresión simple, sólo se considera una variable independiente X, por lo tanto en la prueba de Breusch-Pagan el principal supuesto es que las variaciones de σ_t^2 deben estar relacionadas a la variable independiente.

Lo anterior traducido al modelo de regresión simple, quiere decir que mientras el modelo original se define como

$$\boldsymbol{Y_i = \alpha + \beta X_i + u_i}$$

La varianza se describe mediante el modelo lineal

$$\boldsymbol{\sigma_i^2 = \delta_0 + \delta_1 X_i + \epsilon_i}$$

Donde ϵ_i es el término de error de este segundo modelo.

Se puede notar que, cuando el coeficientes δ_1 es igual a 0, entonces se tendría sólo el término constante para el valor de la varianza más un error aleatorio y por consiguiente se cumpliría la homoscedasticidad en el primer modelo, debido a esto la hipótesis nula se enuncia como $\boldsymbol{H_0: \delta_1 = 0}$. Para llevar a cabo la segunda regresión y contrastar la hipótesis se debe recordar que el contenido de la información de la varianza en el modelo inicial, se encuentra en los errores cuadráticos ${u_i}^2$, por lo tanto es equivalente comprobar la hipótesis con el modelo

$$\boldsymbol{{u_i}^2 = \delta_0 + \delta_1 X_i + \epsilon_i}$$

Para aplicarlo ahora de manera práctica con datos observados, se debe sustituir el término de error cuadrático con el cuadrado del residuo observado ${e_i}^2$, además de tomar los valores observados de X como x_i, con lo cual el modelo final para la prueba es

$$\boldsymbol{{e_i}^2 = \delta_0 + \delta_1 x_i + \epsilon_i}$$

Los autores de la prueba demuestran que al obtenerse el coeficiente de determinación de este último modelo, denotado como $r^2_{{e_i}^2}$ multiplicado por el tamaño de la muestra n, se distribuye Ji-cuadrada con los grados de libertad igual al número de variables independientes consideradas para la segunda regresión (En este caso 1), por lo tanto el estadístico de prueba es el siguiente

$$\boldsymbol{LM = n * r^2_{{e_i}^2} \sim {\chi_1}^2}$$

Con base en el estadístico LM y el cuantil de la distribución Ji-cuadrada que acumula $1 - \alpha$ de probabilidad, ${\chi^2}_{1,1-\alpha}$, se define la siguiente regla de decisión: Si $LM >$

$\chi^2{}_{1,1-\alpha}$ entonces se debe rechazar la hipótesis nula H_0 y por lo tanto existe un grado estadísticamente significativo de heteroscedasticidad en los datos.

Para aplicar la prueba de forma sencilla en los datos simulados sobre la hoja de cálculo que se ha estado trabajando, se debe recordar que el coeficiente $r^2_{e_i{}^2}$ tiene la propiedad de ser el cuadrado del coeficiente de correlación entre las variables ${e_i}^2$ y X_i que se puede obtener por la fórmula de Excel® COEF.DE.CORREL(V1,V2) donde V1 y V2 son rangos de las columnas que contienen a las variables correspondientes para medir su correlación, que en esta ocasión son las columnas previamente computadas ${e_i}^2$ y X_i.

Una vez con el valor coeficiente de correlación, éste se eleva al cuadrado y se multiplica por n para obtener el estadístico LM y comparar contra el cuantil de la distribución Ji-cuadrada obtenido por la formula INV.CHICUAD(P,K) con P la probabilidad acumulada que es $1-\alpha$ y K los grados de libertad que para éste modelo lineal simple es $\mathrm{K}=1$.

Con esta construcción se puede comparar el resultado de la prueba para cada muestra generada por la actualización de la hoja de cálculo, agregando una condicional simple que contenga la regla decisión que en Excel® la fórmula es =SI($LM > \chi^2{}_{1,1-\alpha}$,"SI","NO").

Para ver el resultado de aplicar la prueba sobre una de las muestras anteriores simuladas con cierto grado de heteroscedasticidad, se retomó la muestra observada en la Gráfica 5.26 donde el parámetro $Q=-70\%$ y $n=50$. En la tabla 5.9 se encuentran los resultados de calcular el coeficiente de correlación $r_{e_i{}^2}$, su cuadrado para obtener el coeficiente de determinación, el estadístico LM, además, considerando un $\alpha=5\%$ se encuentra el cuantil de la distribución Ji- cuadrada $\chi^2{}_{1,95\%}$. De acuerdo a los valores de la tabla se puede ver que debido a la correlación entre X y los residuos cuadráticos $LM > \chi^2{}_{1,95\%}$, por lo tanto se rechaza la hipótesis nula H_0 y se comprueba entonces que existe un grado significativo de heteroscedasticidad.

Tabla 5.9: Prueba de Breusch-Pagan para la detección de heteroscedasticidad

α	5%	$\chi^2{}_{1,95\%}$	3.84	n	50
$r_{e_i{}^2}$	-0.37	$r^2_{e_i{}^2}$	0.14	LM	6.81

Ahora para comprobar el comportamiento de la prueba bajo diferentes grados de heteroscedasticidad, se registrarán los resultados de repetir esta prueba un cierto número de veces para observar el % de pruebas rechazadas tras la generación de una serie de muestras aleatorias. Primero sobre las características de la muestra, se consideró aumentar el tamaño de la muestra simulada a $n=100$, con el objetivo de lograr resultados más consistentes. Para llevar a cabo la repetición y registro de resultados, se utilizó la rutina en leguaje Macro que permite copiar un resultado y

copiarlo a valores un determinado número de veces[32] actualizando la hoja y generando una nueva muestra en cada repetición; esto se aplica sobre una condicional que devuelva el valor lógico si se debe o no rechazar la prueba de acuerdo al valor de LM, con la cual se copiaron y registraron 5,000 resultados de la prueba aplicados a las muestras simuladas, para cada uno de los diversos valores del parámetro Q propuestos a continuación.

El resumen de los resultados se halla en la Tabla 5.10 donde se consideró la misma significancia $\alpha = 5\%$ para todas las pruebas realizadas, por lo que se agregó el valor del cuantil contra el cual se realizaron los contrastes. Los valores de los parámetros del modelo lineal original son $\alpha = 50$, $\beta = 2$, $\sigma^2 = 250$ como base para todos los escenarios con muestras de tamaño $n = 100$.

Una la forma para llevar a cabo el análisis de la tabla es a partir del escenario $Q = 0\%$ ya que representa el escenario donde se cumple la homoscedasticidad del modelo y por lo tanto el 5% de rechazos que se observa es el reflejo de la probabilidad del error de Tipo I para esta prueba. Mientras los valores de Q aumentan junto con el grado de heteroscedasticidad se puede ver que son rechazadas un porcentaje cada vez mayor de las realizaciones de la prueba, hasta los escenarios extremos donde en el caso positivo se muestra que se limita a la probabilidad $1 - \alpha$ pues es en este casos, donde es evidente que la varianza no es constante.

Tabla 5.10: Resultados de 5,000 simulaciones de pruebas de Breusch-Pagan para diversos grados de heteroscedasticidad

Valores de Q	# de pruebas rechazadas	% de pruebas rechazadas
-99%	4,887	97.70%
-75%	3,287	65.70%
-50%	1,276	25.50%
0%	254	5.10%
100%	1,274	25.50%
500%	4,027	80.50%
1000%	4,565	91.30%
1500%	4,711	94.20%
$\boldsymbol{Muestras\ por\ caso = 5{,}000}$		$\boldsymbol{\chi^2_{1,95\%} = 3.84}$

De esta forma se puede clarificar la eficacia y el comportamiento de la prueba, mientras se muestra además su potencia, aunque es bueno notar la discrepancia en el caso negativo, donde en el escenario extremo se rechaza un porcentaje ligeramente mayor del esperado. Ahora los siguientes métodos a revisar son los correspondientes para detectar el incumplimiento del supuesto de la falta de correlación entre los errores del modelo lineal.

[32] Consulte el código completo en el apéndice.

Prueba de Durbin-Watson para detectar residuos correlacionados

El siguiente de los supuestos principales en el modelo de regresión a revisar es que los errores u_i no estén correlacionados entre sí, sin embargo, en diversos contextos comúnmente los económicos, puesto que el término de error en el modelo de regresión, debe representar todas las demás variable desconocidas que explican el valor final de la variable independiente, también debe incluir el caso en que dependa del resultado en el periodo anterior. Ejemplos de este tipo de comportamientos se hallan en los datos de ventas de una empresa o su valor de mercado, aunque en otras ocasiones, este tipo de datos se encuentran en un contexto distinto como al observar la evolución de un fenómeno como la temperatura.

Para detectar correlación entre los errores Durbin y Watson desarrollaron una prueba de hipótesis, la cual parte del hecho que si un error cualquiera e_i no estuviera correlacionado con el error anterior e_{i-1} entonces se tendría que $Cov(e_i, e_{i-1}) = 0$. Desde otra perspectiva se puede considerar en su lugar al coeficiente de correlación entre estos dos errores que se denotara como $\rho_{i,i-1}$ el cual se encuentra expresado como

$$\boldsymbol{\rho_{i,i-1} = \frac{Cov(e_i, e_{i-1})}{\sqrt{Var(e_i)Var(e_{i-1})}}}$$

$\rho_{i,i-1}$ Tomará diferentes valores de acuerdo a la correlación que tengan cada error con su valor previo, pues en el caso de no estar correlacionados se sabe que $\rho_{i,i-1} = 0$, mientras que si existe una correlación positiva tomara valores positivos con limite en 1, límite que representa la correlación positiva perfecta, mientras que será negativo con limite en -1, donde el valor límite representa cuando la correlación sea negativa y perfecta entre los residuos.

Después el razonamiento de la prueba, se emplea la esperanza de la diferencia cuadrática entre e_i ye_{i-1} como $E(e_i - e_{i-1})^2$ dividido entre la raíz de cada una de las varianzas, cuyo resultado se denota como $\sigma_{i,i-1}$ que se desglosa de la siguiente manera

$$\boldsymbol{\sigma_{i,i-1} = \frac{E(e_i - e_{i-1})^2}{\sqrt{Var(e_i)Var(e_{i-1})}}}$$

$$\boldsymbol{= \frac{E(e_i)^2 + E(e_i)^2 + 2E(e_i * e_{i-1})}{\sqrt{Var(e_i)Var(e_{i-1})}}}$$

Ahora puesto que los residuos tienen media 0 y considerando que se cumpla el supuesto de una varianza constante entonces $E(e_i)^2 = Var(e_i) = \sigma^2$, con lo cual el cociente se expresa finalmente como

$$\boldsymbol{\sigma_{i,i-1} = 2 - 2\frac{E(e_i * e_{i-1})}{\sqrt{Var(e_i)Var(e_{i-1})}}}$$

El segundo término de la suma es igual a dos veces el coeficiente de correlación entre los residuos, por lo que analizando $\sigma_{i,i-1}$ su valor tiene tres casos importantes

$$\sigma_{i,i-1} = \begin{cases} \mathbf{0} & \boldsymbol{Cuando\ la\ correlacion\ de\ los\ residuales\ sea\ positiva\ y\ perfecta} \\ \mathbf{4} & \boldsymbol{Cuando\ la\ correlacion\ de\ los\ residuales\ sea\ negativa\ y\ perfecta} \\ \mathbf{2} & \boldsymbol{Cuando\ los\ residuales\ no\ se\ encuentren\ correlacionados} \end{cases}$$

Basados en esto Durbin y Watson proponen la siguiente estadística para medir la correlación de los residuos

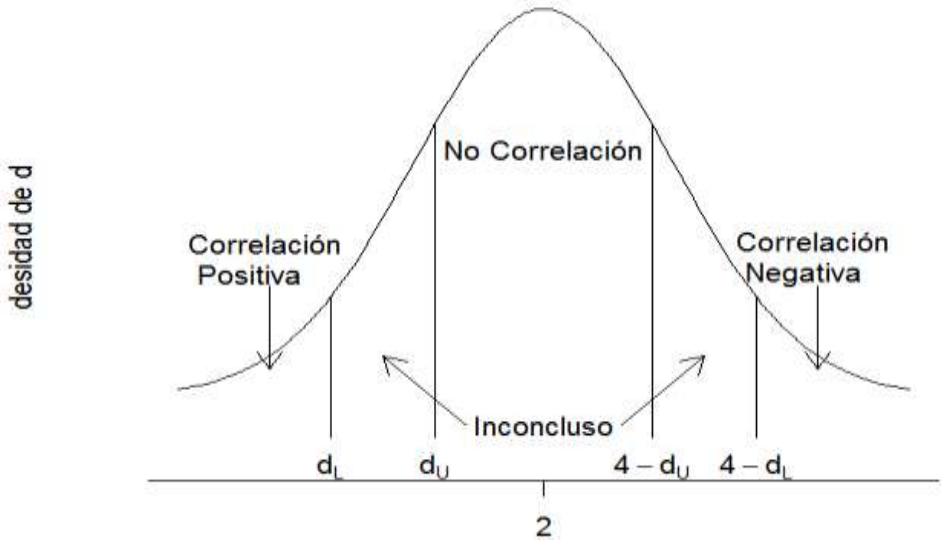
$$\boldsymbol{d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}}$$

Luego para plantear la prueba, se establece como hipótesis nula H_0 que los errores no están correlacionados, y como hipótesis alternativa H_1 que los errores tienen algún tipo de correlación, ya sea positiva o negativa, aunque se puede realizar la prueba enfocada a un sentido de la correlación en particular.

Tomando el caso de la hipótesis alternativa para probar si existe algún tipo de correlación en ambos sentidos, las reglas decisión para realizar la prueba en base a d, están determinadas por su densidad y una serie de límites denotados como d_U como límite superior y d_L para el límite inferior, los cuales se obtienen de una tabla desarrollada por Durbin y Watson la cual se halla en diversos libros de la literatura estadística, para diversos niveles de significancia o en diversos recursos en la Web.

Estas reglas decisión también contienen zonas para las cuales la prueba de Durbin-Watson queda inconclusa, tanto para el sentido positivo como para sentido negativo de la correlación, por lo que, para comprender mejor las reglas de esta prueba se puede emplear como referencia el siguiente gráfico.

Gráfica 5.29: Reglas de decisión para la prueba de Durbin-Watson



En el gráfico se pueden observar como los límites inferiores y superiores dividen los diferentes casos donde se señala el centro donde no se rechaza la hipótesis nula,

además de las zonas inconclusas y de rechazo. Para ejemplificar la realización de la prueba y su comportamiento con errores correlacionados, a continuación se revisará un método para poder generar los errores u_i con un alto grado de correlación positiva y negativa.

Simulación Monte Carlo de datos con errores correlacionados

Una serie de valores que compongan una muestra, de tal manera que para generar una nueva observación se tenga una cierta dependencia con los valores previamente generados, se conoce como un proceso autocorrelacionado, el cual dependiendo del número de valores previos de los que dependa se define el orden del proceso. Este tipo de procesos son el caso opuesto al supuesto de no correlación de los errores, por lo que generarlos será el objetivo de esta sección.

Partiendo de un error u_i, considerando que depende sólo del valor previo entonces se trata de un proceso de primer orden, relación que usualmente se describe de la siguiente forma

$$\boldsymbol{u_i = Cu_{i-1} + \varepsilon_i \quad donde\ |C| < 1}$$

Donde C es un parámetro menor a 1 en valor absoluto ya que tomando el caso $C > 1$ se tiene una caminata aleatoria y ε_i es una variación aleatoria, sin embargo, ya que se está trabajando bajo el supuesto de normalidad en los errores, se puede evitar la simulación de variables aleatorias adicionales y simplificar el razonamiento de dependencia, tomando como la media la distribución de u_i, en la posición de u_{i-1}, y luego la variación será determinada por la simulación generada por la técnica de la transformada inversa. Este método requiere de un valor inicial como semilla para el proceso, el cual se puede obtener como un error con el supuesto de media 0 y varianza σ^2, con lo cual la generación de los errores se realiza con el método de la transformada inversa con las siguientes distribuciones

$$\boldsymbol{u_i \sim N(u_{i-1}, \sigma^2) \qquad i := 2, 3, \dots, n}$$

$$\boldsymbol{u_1 \sim N(0, \sigma^2)}$$

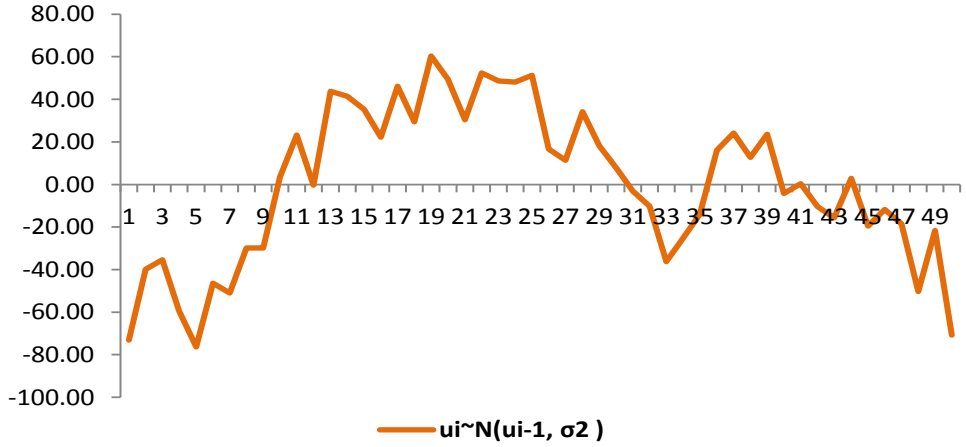
Con esta relación entre los residuos se genera un proceso autocorrelacionado positivamente. Por otro lado, si lo que se deseara fuera una alta correlación negativa, entonces basta con tomar la media del nuevo error como el valor previo pero con signo negativo, ya que considerando el valor 0 como eje de referencia entonces se generará el nuevo valor a partir del valor reflejado en el lado contrario al anterior, simulando entonces un comportamiento de correlación negativa.

En ambiente de hoja de cálculo es sencillo simular este proceso al copiar la hoja con el modelo ajustado hasta el cálculo de los residuos, ya que basta con referenciar la media desde u_2 a u_n como la celda anterior, donde se halla el valor simulado anterior, sin cambiar la fórmula del primer error u_1.

A continuación se encuentra el resultado de graficar los errores simulados después de realizar el cambio de la fórmula para considerar el primer caso de correlación positiva, donde se observa que los errores ya no se hayan alrededor de 0, sino que prosiguen una cierta tendencia de acuerdo al comportamiento anterior de la serie. Al realizar

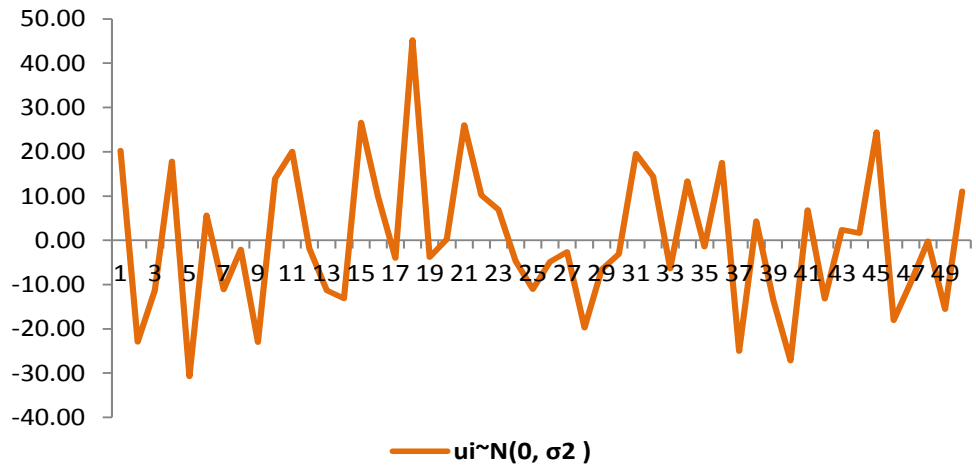
actualizaciones a la hoja de cálculo y generar nuevas muestras, se podrá observar como en ciertas ocasiones se generan procesos que toman una tendencia creciente, otros mayormente decreciente y otros, casos como el de la Gráfica 5.30, en el que la tendencia inicia creciente y cambia a una decreciente al final.

Gráfica 5.30: Simulación de errores u_i autocorrelacionados positivamente

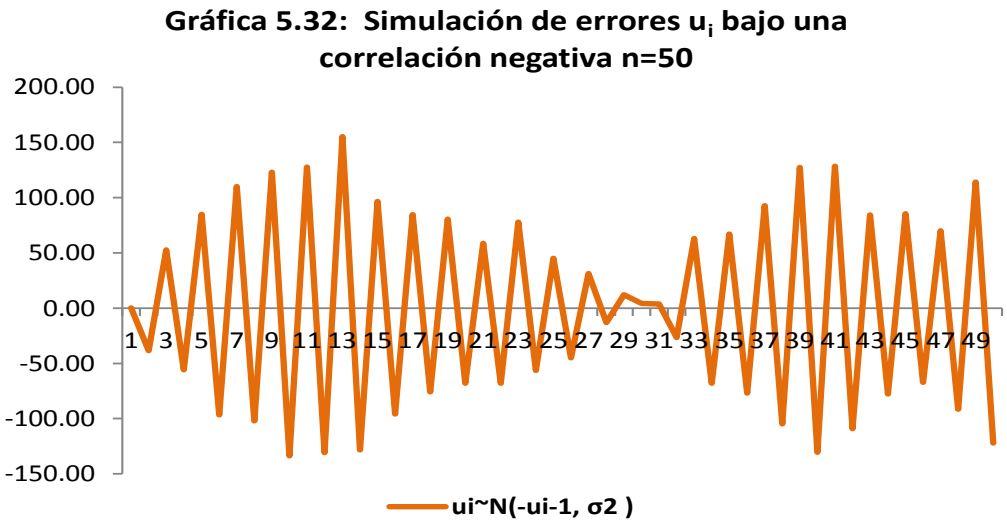


En la Gráfica 5.31, se encuentran el resultado de una simulación de los errores sin el cambio de la fórmula, es decir los errores no correlacionados generados al inicio del capítulo, en los cuales se puede observar que al tener un media constante 0, las simulaciones no presentan una tendencia, el cual es el comportamiento esperado de los residuos cumpliendo el supuesto de no correlación.

Gráfica 5.31: Simulación de errores u_i no correlaciónados n=50



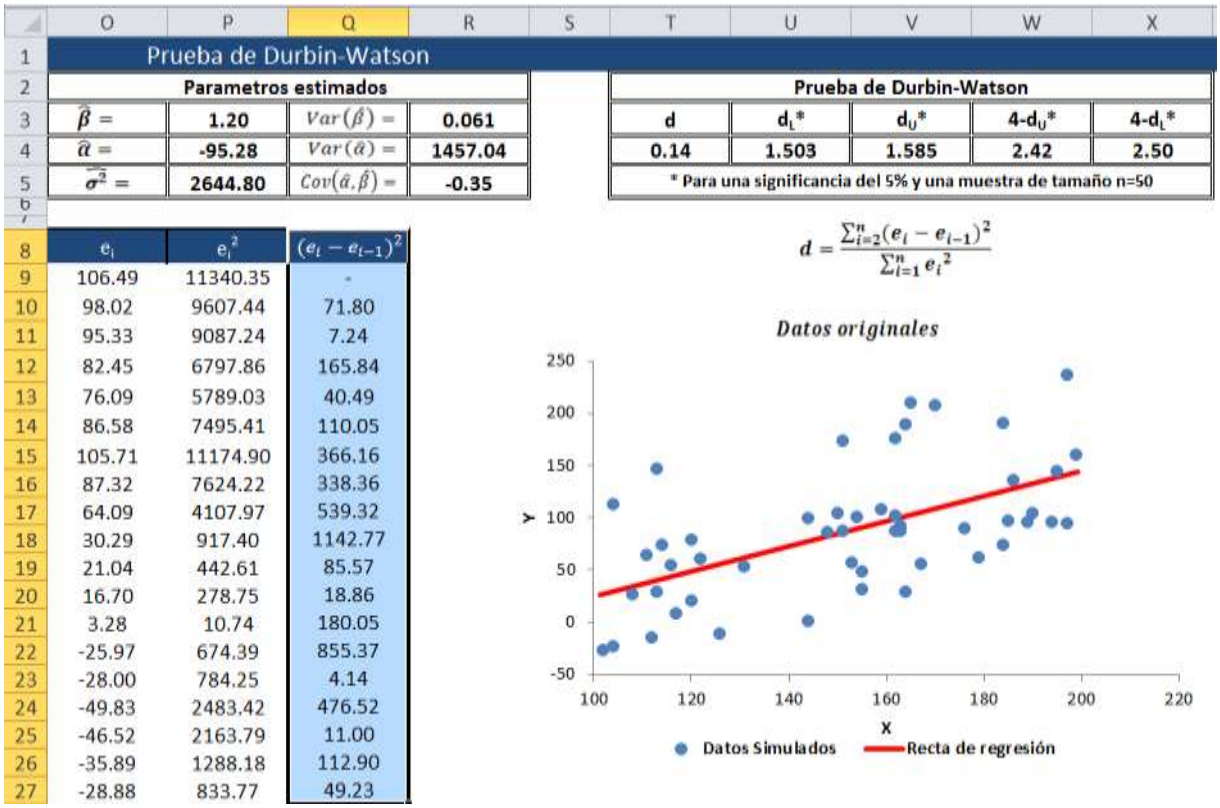
En el siguiente escenario se simularon datos cuya definición de la media se consideró como el valor generado anteriormente, pero con signo contrario, es decir la distribución de cada error es $u_i \sim N(-u_{i-1}, \sigma^2)$ con $u_1 \sim N(0, \sigma^2)$. En la gráfica se observa una serie en la que, para un valor positivo el siguiente es un valor negativo, lo cual describe una correlación negativa entre los errores simulados.



Para comprobar que el tipo de correlación es correcto , se adicionaron los tres tipos de errores al modelo de regresión, en la hoja en la cual se tiene computado hasta los residuos, donde además se adicionó una columna con los valores $(e_i - e_{i-1})^2$ y su suma total al final, para luego adicionar el cálculo del estadístico d y los limites d_U , d_L para realizar la prueba de Durbin-Watson, consultándolos en una tabla de los valores críticos de esta prueba, a partir de un recurso hallado en la Web.

La siguiente imagen muestra la adición de una tabla con los elementos necesarios para la prueba de Durbin-Watson, donde se simularon datos con los parámetros $\alpha = 50$, $\beta = 1$, $\sigma^2 = 250$, y se realizó el cambio de fórmula en la simulación de los errores u_i para conseguir errores correlacionados positivamente. Los límites de la prueba de Durbin-Watson d_U y d_L mostrados en la parte superior derecha de la hoja fueron consultados al 5% de significancia para una muestra de tamaño $n = 50$, además se puede observar el valor del estadístico d que al ser 0.14 se posiciona en la zona de rechazo de la hipótesis nula y determina que en efecto hay un alto grado de correlación positiva en los errores ya que d es muy cercano a 0.

Estructura en Excel® donde se agrega la prueba de Durbin-Watson



Prueba de Durbin-Watson

Parametros estimados			
$\hat{\beta} =$	1.20	$Var(\hat{\beta}) =$	0.061
$\hat{\alpha} =$	-95.28	$Var(\hat{\alpha}) =$	1457.04
$\hat{\sigma^2} =$	2644.80	$Cov(\hat{\alpha}, \hat{\beta}) =$	-0.35

Prueba de Durbin-Watson				
d	d_L*	d_U*	4-d_U*	4-d_L*
0.14	1.503	1.585	2.42	2.50
* Para una significancia del 5% y una muestra de tamaño n=50				

e_i	e_i^2	$(e_i - e_{i-1})^2$
106.49	11340.35	-
98.02	9607.44	71.80
95.33	9087.24	7.24
82.45	6797.86	165.84
76.09	5789.03	40.49
86.58	7495.41	110.05
105.71	11174.90	366.16
87.32	7624.22	338.36
64.09	4107.97	539.32
30.29	917.40	1142.77
21.04	442.61	85.57
16.70	278.75	18.86
3.28	10.74	180.05
-25.97	674.39	855.37
-28.00	784.25	4.14
-49.83	2483.42	476.52
-46.52	2163.79	11.00
-35.89	1288.18	112.90
-28.88	833.77	49.23

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Sobre esta misma estructura se pueden realizar los cambios sobre la fórmula de la simulación errores u_i, lo que producirá que se actualicen los cálculos. En la Tabla 5.11 se encuentra el resultado de la prueba de Durbin-Watson para una muestra simulada con los errores simulados originalmente con media constante 0, en la que se puede ver que d toma un valor cercano a 2, por lo que de acuerdo con el diagrama con la reglas de decisión este tipo de errores no se hallan correlacionados, mientras que en la Tabla 5.12 se encuentra el resultado de simular los errores bajo una correlación negativa definiendo la media de cada u_i como $-u_{i-1}$, lo que produjo que le estadístico de prueba sea de un valor cercano a 4, lo que comprueba el tipo de correlación negativa en los residuos.

Tabla 5.11: Prueba sobre los errores simulados con media constante 0

Prueba de Durbin-Watson				
d	d_L*	d_U*	4-d_U*	4-d_L*
1.98	1.503	1.585	2.42	2.5
* Para una significancia del 5% y una muestra de tamaño n=50				

Tabla 5.12: Prueba sobre los errores simulados con $u_i \sim N(-u_{i-1}, \sigma^2)$				
Prueba de Durbin-Watson				
d	**d_L***	**d_U***	**4-d_U***	**4-d_L***
3.934	1.503	1.585	2.42	2.5
* Para una significancia del 5% y una muestra de tamaño n=50				

Hasta ahora se ha visto como evaluar la correlación de los errores, aunque ahora para hacer un análisis más profundo se verán un par de métodos gráficos para detectar correlación en los errores.

Métodos gráficos para detectar correlación entre los errores

Para detectar ahora la correlación de los errores desde evaluaciones de los residuos del modelo, se introducirán dos gráficos que complementarán la conclusión de la prueba de Durbin-Watson y que se pueden agregar fácilmente a la hoja de trabajo.

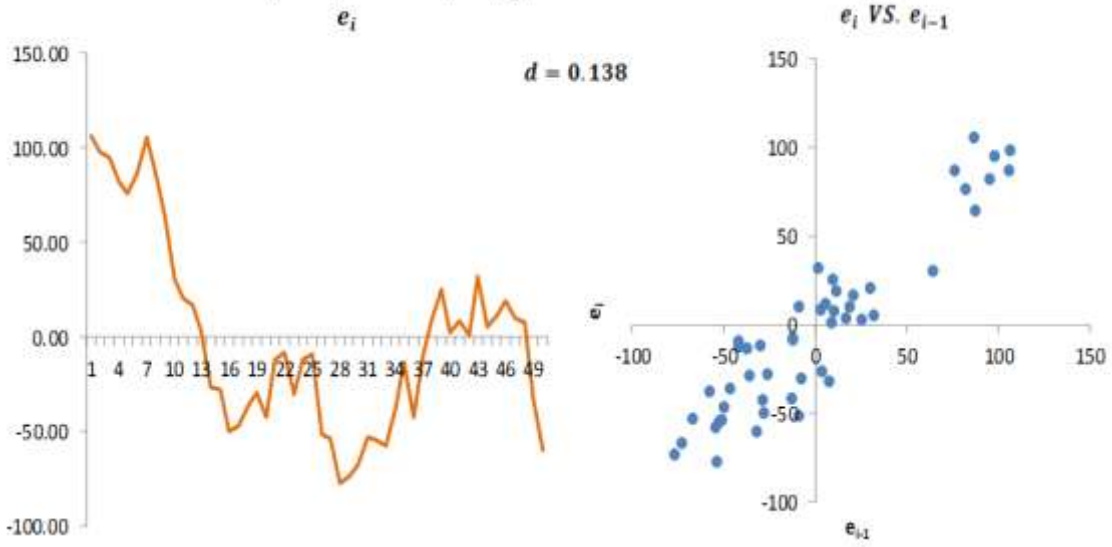
La primera de estas gráficas, consiste en mostrar los residuos como una secuencia de valores, en un diagrama de línea similar a las gráficas 5.30 a 5.32, en la que se podrá observar si existe alguna dependencia de los residuos con sus valores previos y el tipo de serie que forman en conjunto.

El otro gráfico que permite detectar correlación se desarrolla por medio de un diagrama de dispersión colocando en el eje X los residuos e_i y en el eje Y a los residuos de tal manera que e_i coincida como pareja con el valor e_{i-1} , de esta manera se desea visualizar la relación que cada residuo tiene con su valor previo.

Para agregar el gráfico de dispersión en la hoja de cálculo, se puede seleccionar la columna de los residuos y para la otra serie sólo se debe omitir el primer residuo, con lo que se generará la coincidencia de las parejas. Una vez agregados ambos gráficos, se contrastan junto con la estadística d para relacionar el comportamiento de los residuos con el tipo de correlación simulada. Por ejemplo en la siguiente gráfica se encuentra tanto los residuos graficados en línea, así como el gráfico de e_i vs. e_{i-1}, para el caso de la simulación de datos con errores correlacionados positivamente, es decir con $\boldsymbol{u_i \sim N(u_{i-1}, \sigma^2)}$**.**

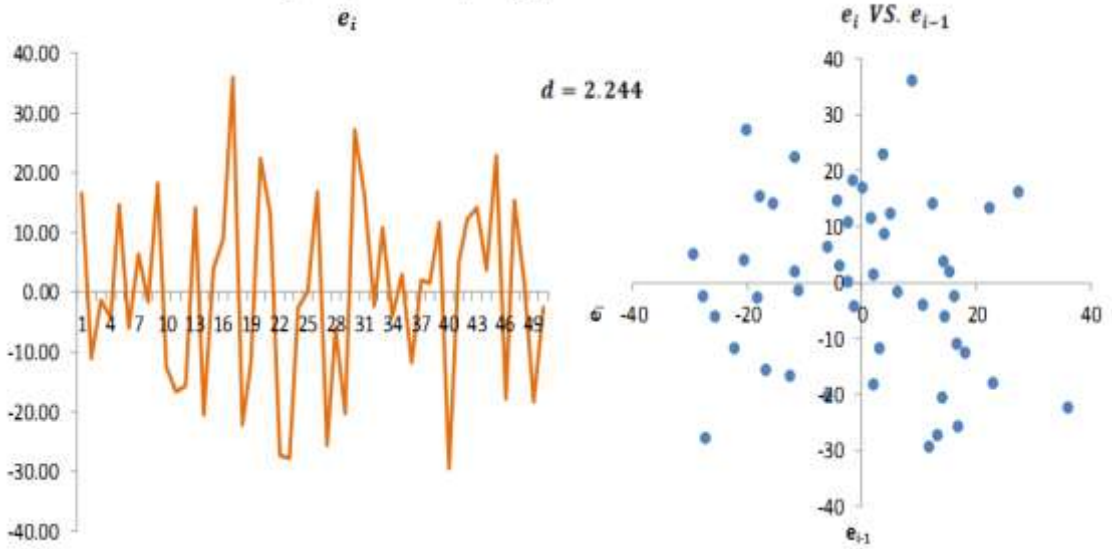
En la serie de residuos se observa como los errores tienen cierta tendencia a partir del valor anterior, mientras que en el diagrama de dispersión se puede observar la formación de una nube de puntos que asemeja una línea recta de pendiente positiva debido a que en la serie un valor positivo por lo regular es seguido de otro valor positivo y viceversa, lo que ejemplifica la correlación positiva de cada residuo con respecto del residuo anterior y que se puede complementar además mostrando, como en la gráfica, el valor del estadístico $d = 0.138$ que por ser cercano a cero también diagnostica la correlación positiva en los residuos.

Gráfica 5.33: Residuos y comparativo de e_i vs. e_{i-1}, para datos simulados con correlación positiva en los errores



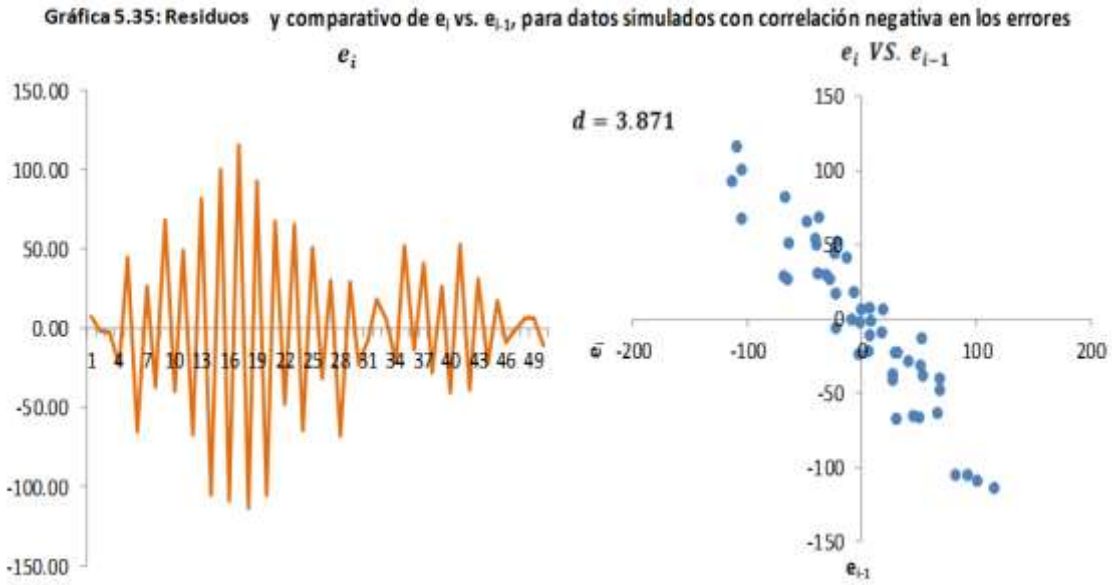
Ahora en otra actualización de las fórmulas de la hoja de cálculo se generó el escenario cuando los errores son simulados con media 0 y varianza constante σ^2 , comportamiento deseado de los datos, que ocurre cuando se cumple el supuesto de falta de correlación entre los errores. En el gráfico también se puede notar que el comportamiento de la serie de residuos se encuentra alrededor de 0 sin una tendencia observable, además en el diagrama de dispersión, se puede ver que no existe una relación entre los residuos con su valor previo, lo que ocasiona que en el diagrama de dispersión se hallen los puntos dispersos en todos los cuadrantes y que el estadístico d tengaun valor cercano a 2 como lo supondría el razonamiento detrás de la prueba de Durbin-Watson cuando los errores no están correlacionados.

Gráfica 5.34: Residuos y comparativo de e_i vs. e_{i-1}, para datos simulados con los errores no correlacionados



En un tercer escenario se realizaron ambas gráficas de los residuos, con los datos simulados bajo una correlación negativa considerando $\boldsymbol{u_i \sim N(-u_{i-1}, \sigma^2)}$. Para una simulación en particular las siguientes gráficas contienen los diagramas antes revisados, en los que se puede observar que para este tipo de correlación el comportamiento de la serie de residuos se asemeja a la Gráfica 5.32, mientras que el

diagrama de dispersión se asemeja ahora con una recta de pendiente negativa, pues de acuerdo al construcción de los errores, para un valor positivo el siguiente tomaría como media el valor anterior pero un signo contrario. Para esta simulación de acuerdo al valor del estadístico $d = 3.871$ y la gráfica con las reglas de decisión, la correlación entre los residuos es negativa, además de ser cercana al valor de la correlación negativa perfecta.



Prueba de Jarque-Bera para evaluar la normalidad de los errores

Esta prueba fue desarrolla por Carlos M. Jarque y Anil K. Bera en el año 1980, para probar la Normalidad de los residuos en un modelo de regresión, la cual toma como hipótesis nula H_0, que la asimetría y kustoris de los residuos son iguales a los calculados a partir de una población distribuida de manera Normal N(0,σ^2) y concluye tras aplicar una regla de decisión con cierto grado de significancia, si poseen la misma distribución. Con base en lo anterior la hipótesis alternativa H_1 establece que los residuos poseen una distribución distinta a una Normal.

El computo del estadístico de prueba, se basa en el coeficiente de asimetría de Pearson y el coeficiente de Kurtosis de Pearson, revisados en el capítulo 2 sobre estadística descriptiva denotados como γ_1 y K_1 respectivamente los cuales se calculan como sigue

$$\boldsymbol{K_1 = \frac{\mu_4}{(\sqrt{S^2})^4} = \frac{\mu_4}{S^4} * 100\%}$$

$$\boldsymbol{\gamma_1 = \frac{\mu_3}{(\sqrt{S^2})^3} = \frac{\mu_3}{S^3} * 100\%}$$

Donde $\boldsymbol{\mu_k = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^k}{n}}$ **y** $\boldsymbol{S^2 = \sum_{i=1}^{n}\frac{(x_i - \bar{X})^2}{n-1}}$

Entonces el estadístico de prueba denotada como JB se calcula de la siguiente manera:

$$\boldsymbol{JB = \frac{n}{6}\left(\gamma_1{}^2 + \frac{(K_1 - 3)^2}{4}\right)}$$

En la forma de cálculo del estadístico JB se puede ver que bajo el cumplimiento de H_0 el estadístico de prueba es teóricamente 0, por lo tanto conforme la distribución de los residuos tengan algún grado de asimetría, o sus colas sean más o menos pesadas con respecto a la de una Normal (teóricamente K_1 =3) produciendo que la medición de la kurtosis sea diferente, entonces el estadístico de prueba incrementará su valor.

Además bajo el supuesto de cumplirse la hipótesis nula, JB se distribuye asintóticamente como una variable Ji-cuadrada con 2 grados de libertad, con esto la regla de decisión establece que se debe comparar el cuantil de la distribución Ji-cuadrada $\chi^2_{2\,1-\alpha}$, para un nivel de significancia α, se debe aplicar la siguiente regla de decisión: en caso de obtenerse $JB > \chi^2_{2\,1-\alpha}$ entonces se debe rechazar la hipótesis nula.

Una de las ventajas de esta prueba es que al calcular JB, se puede mantener también a la vista el cálculo de los coeficientes de asimetría y kurtosis de los residuos, pues en caso de rechazar la hipótesis nula, se puede detectar cuál es la fuente de incremento del estadístico, ya sea en la asimetría, lo que daría como referencia a pensar que los errores podrían distribuirse como una variable gamma o alguna otra con otro grado de asimetría, o en el caso de ser la kurtosis la fuente de discrepancia, entonces se tomaría como una mejor referencia para la distribución de los residuos un variables de colas más pesadas como la *t de Student*, y en caso de ser ambos factores los que difieran de los valores de una Normal se puede pensar como referencia una distribución Pareto para los residuos.

Para poder revisar el comportamiento de la prueba bajo diferentes escenarios, se deben simular los errores u_i con diversas distribuciones, diferentes de la distribución Normal, originalmente simulada para los errores. En este caso las distribuciones elegidas para llevar a cabo la comparación son la distribución exponencial, para revisar el caso asimétrico, y por otra parte la distribución conocida como Laplace o doble exponencial, que más adelante se detallará, donde se verá una forma sencilla de simularla, con el objetivo de mostrarla como escenario de una distribución de colas pesadas.

En el primer escenario se realizó el cálculo del estadístico JB con los errores u_i originalmente simulados distribuidos $N(0, \sigma^2)$, cuyo resultado de obtener los cálculos necesarios, para aplicar la prueba sobre una muestra en particular se puede observar en la Tabla 5.13 en la cual también se agregaron los coeficientes de asimetría y Kurtosis. Considerando un nivel de confianza de $\alpha = 5\%$ se obtuvo el cuantil $\chi^2_{2\,1-\alpha}$ que acumula $(1-\alpha)\%$ por medio de la función INV.CHICUAD($1-\alpha$,2) que se compara contra el estadístico de prueba, que este caso se tiene $JB < \chi^2_{2\,1-\alpha}$ y por lo tanto no se debe rechazar la hipótesis nula, concluyendo que no existen diferencias estadísticas significativas entre la distribución de los residuos y una distribución Normal.

Tabla 5.13:Prueba de Jarque-Bera aplicada a los residuales, basados en errores simulados como una N(0,σ^2)				
γ_1	K_1	JB	α	$\chi^2_{2\,1-\alpha}$
0.39	2.75	1.397	0.05	5.99

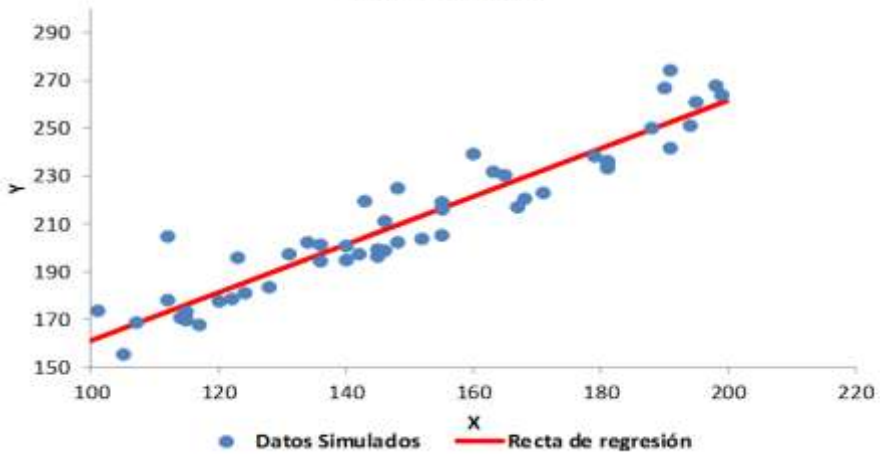
En el siguiente escenario se simularon los errores u_i como valores distribuidos Exponencial($\lambda = 0.1$), sustituyendo la columna u_i de la hoja con las simulaciones de una Normal, por la fórmula LN(ALEATORIO())/λ, que aplica el mismo método de la transformada inversa para generar valores de la distribución Exponencial(λ).

Por medio de las fórmulas referenciadas para el ajuste de regresión en la hoja de cálculo, tras la sustitución estos nuevos errores se incorporan automáticamente al modelo lineal. Los resultados de realizar los cálculos necesarios para aplicar la prueba de Jarque-bera se pueden observar en la Tabla 5.14, donde se consideró ahora una significancia del 1%, además se puede observar que la hipótesis nula es rechazada pues $\chi^2_{2\,1-\alpha} < JB$. Analizando el resultado de la prueba se nota que la fuente de discrepancia con respecto a la Normal fue la asimetría de la distribución exponencial, reflejado en el valor del coeficiente de asimetría, mientras que la Kurtosis para esta muestra en particular es cercana a 3 que es el valor de la kurtosis de una Normal.

Tabla 5.14:Prueba de Jarque-Bera aplicada a los residuales, basados en errores simulados como una Exponencial (λ=0.1)				
γ_1	K_1	JB	α	$\chi^2_{2\,1-\alpha}$
1.1	3.633	10.92	0.01	9.21

En la Gráfica 5.36 se pueden observar el modelo lineal simulado con los errores distribuidos exponencialmente, donde se puede notar como los puntos simulados se sitúan en parte concentrados por debajo de la recta, debido a que la distribución exponencial acumula una mayor probabilidad cerca del valor 0[33].

Gráfica 5.36: Datos Simulados del modelo lineal con errores simulados deuna Exponencial(λ=.1)



[33] Para una referencia más a detalle sobre estas características la distribución Exponencial(λ) revise el Capítulo I.

Ahora para mostrar un escenario en el cual la kurtosis sea la fuente de discrepancia con respecto de la Normal, se tomará a la distribución Laplace(μ, b) para simular los errores u_i. Esta distribución nombrada por el matemático Pierre Simón Laplace, también conocida como la distribución doble exponencial debido a la forma de su función de densidad la cual se muestra en la Gráfica 5.37 y se encuentra definida como

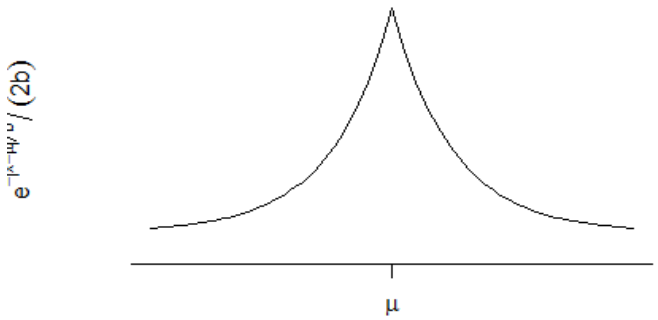
$$\boldsymbol{f_X(x) = \frac{1}{2b} e^{\left(-\frac{|x-\mu|}{b}\right)}}$$

La esperanza y varianza de esta variable aleatoria están calculadas de la siguiente manera

$$\boldsymbol{E(X) = \mu}$$

$$\boldsymbol{Var(X) = 2b^2}$$

Gráfica 5.37: densidad de la distribución Laplace(μ , b)

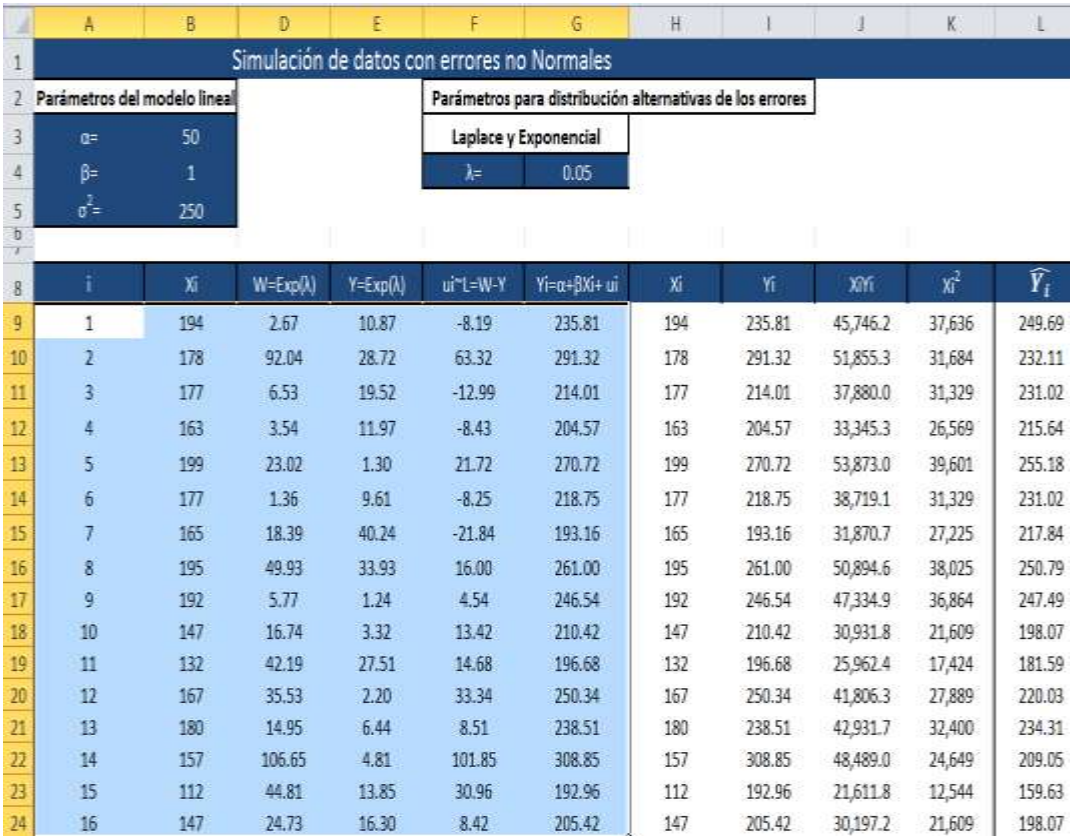


Para generar valores de esta distribución de una forma sencilla, se retomará el teorema de la teoría de la probabilidad, el cual enuncia que si se tienen dos variables aleatorias $W \sim Exponencial(\lambda_1)$ y $Y \sim Exponencial(\lambda_2)$, y además $\lambda_1 = \lambda_2 = \lambda$, entonces la variable obtenida de la diferencia $L = W - Y$ se distribuye Laplace$(0,1/\lambda)$, la cual es una distribución centrada en 0, pero de colas pesadas similares a las de una exponencial por ambos lados de su soporte.

En la hoja de cálculo se puede emplear la columna de la simulación de los errores exponenciales simulada anteriormente como W y copiar la columna en una columna adjunta, para así obtener la segunda variable exponencial Y con el mismo parámetro λ, una vez con esto, sólo se necesitaría calcular en una tercera columna adjunta la diferencia entre los valores de las columnas para generar errores $u_i \sim Laplace(0,1/\lambda)$.

En la siguiente imagen se muestra una propuesta para la estructura de cálculo muy similar a las anteriores, en donde en la parte sombreada se señalan los valores generados para dos variables exponenciales W y Y con el mismo método anterior, con un parámetro único $\lambda = 0.05$, además junto a esta columnas se halla la columna con la diferencia entre las variables para simular los errores con una distribución $Laplace\left(0, \frac{1}{0.05}\right)$, los cuales son incluidos en el modelo al referenciarlos como la columna de valores u_i.

Estructura en Excel® donde se simulan los errores u_i como una variable Laplace$(0, 1/\lambda)$.



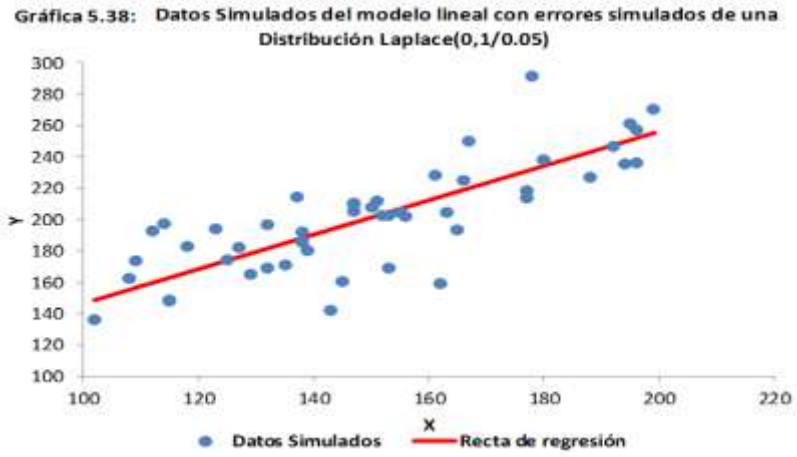
	A	B	D	E	F	G	H	I	J	K	L
1	Simulación de datos con errores no Normales										
2	Parámetros del modelo lineal				Parámetros para distribución alternativas de los errores						
3	α=	50			Laplace y Exponencial						
4	β=	1			λ=	0.05					
5	σ^2=	250									
8	i	Xi	W=Exp(λ)	Y=Exp(λ)	ui~L=W-Y	Yi=α+βXi+ ui	Xi	Yi	XiYi	Xi^2	$\widehat{Y_i}$
9	1	194	2.67	10.87	-8.19	235.81	194	235.81	45,746.2	37,636	249.69
10	2	178	92.04	28.72	63.32	291.32	178	291.32	51,855.3	31,684	232.11
11	3	177	6.53	19.52	-12.99	214.01	177	214.01	37,880.0	31,329	231.02
12	4	163	3.54	11.97	-8.43	204.57	163	204.57	33,345.3	26,569	215.64
13	5	199	23.02	1.30	21.72	270.72	199	270.72	53,873.0	39,601	255.18
14	6	177	1.36	9.61	-8.25	218.75	177	218.75	38,719.1	31,329	231.02
15	7	165	18.39	40.24	-21.84	193.16	165	193.16	31,870.7	27,225	217.84
16	8	195	49.93	33.93	16.00	261.00	195	261.00	50,894.6	38,025	250.79
17	9	192	5.77	1.24	4.54	246.54	192	246.54	47,334.9	36,864	247.49
18	10	147	16.74	3.32	13.42	210.42	147	210.42	30,931.8	21,609	198.07
19	11	132	42.19	27.51	14.68	196.68	132	196.68	25,962.4	17,424	181.59
20	12	167	35.53	2.20	33.34	250.34	167	250.34	41,806.3	27,889	220.03
21	13	180	14.95	6.44	8.51	238.51	180	238.51	42,931.7	32,400	234.31
22	14	157	106.65	4.81	101.85	308.85	157	308.85	48,489.0	24,649	209.05
23	15	112	44.81	13.85	30.96	192.96	112	192.96	21,611.8	12,544	159.63
24	16	147	24.73	16.30	8.42	205.42	147	205.42	30,197.2	21,609	198.07

La visualización de los datos simulados en la imagen se halla en la Gráfica 5.38 en la cual se puede observar que los puntos se concentran sobre la recta de regresión, no obstante, las colas pesadas de los errores generaron algunos puntos que se alejan lo suficiente como para identificarlos como datos atípicos, resultado de las colas pesadas de la distribución Laplace, y que se encuentran, tanto por debajo como por encima de la recta de regresión, debido a la simetría de la distribución.

Aplicando la prueba de Jarque-Bera a los datos recién simulados, se espera que ahora la fuente de discrepancia con respecto de la distribución Normal sea la medición de la kurtosis. El resultado de aplicar la prueba a los datos anteriores de la imagen anterior se encuentran en la Tabla 5.15 donde se observar que el valor del coeficiente de asimetría fue cercano a 0 por lo que el aporte al valor del estadístico JB fue el coeficiente de kurtosis que toma una valor de $K_1 = 6.09$. Considerando una significancia del 1%, la prueba concluye de acuerdo al valor de JB que se debe rechazar la hipótesis nula y por lo tanto existen diferencias significativas entre la distribución de los residuos y una Normal.

Tabla 5.15: Prueba de Jarque-Bera aplicada a los residuales, basados en errores simulados como una distribución Laplace(0,1/0.05=20)

γ_1	K_1	JB	α	$\chi^2_{2\,1-\alpha}$
0.28	6.095	20.625	0.01	9.21



Puesto que para todos los casos se pueden generar nuevas simulaciones actualizando la hoja de cálculo se podrá observar el comportamiento de la prueba, donde para unas muestras se observarán las variaciones del coeficiente de Asimetría alrededor de 0 en el caso de la distribución Laplace y los valores de la Kurtosis muy por encima de 3 lo que provocará nuevamente la conclusión de rechazar la hipótesis nula, que de acuerdo a la significancia de la prueba sería en el 99% de los casos.

Gráfico de probabilidad Normal de residuos

El siguiente gráfico está diseñado para detectar discrepancias sobre la distribución de los residuos en el modelo de regresión con respecto a una distribución Normal, para ello compara los estadísticos de orden de los residuos observados denotados como $e_{(i)}$, contra los estadísticos de orden teóricos bajo una distribución Normal de media 0 y varianza unitaria los cuales son denotados como $z_{(i)} \quad i := 1,2,\dots,n$.

Para generar el gráfico primero se debe agregar el cálculo de los estadísticos de orden de los residuos a la hoja de cálculo, generando una columna adjunta a los residuos con la formula K.ESIMO.MENOR(R,i) donde R es el rango que se debe seleccionar donde se encuentran los residuos y el valor de i puede ser la referencia a la una columna donde se encuentra el índice de las observaciones, con esta fórmula al tomar el valor 1 se obtendrá el mínimo del rango, mientras que en el valor $i = n$ se obtendrá el máximo de los residuos**.**

Para obtener los estadísticos de orden normales primero se debe obtener una aproximación del valor teórico de la probabilidad asociada que cada $z_{(i)}$ acumula para el estadístico de orden correspondiente, denotada como p_i la cual se calcula como

$$\boldsymbol{p_i = \frac{R_i - \frac{3}{8}}{n + \frac{1}{4}}}$$

Donde R_i es el rango del residuo ordenado, que en la hoja de cálculo es el valor de i. Con las p_i se obtiene el valor del cuantil de la distribución Normal estándar, por medio de su función de distribución inversa, que en el software Excel® se halla con la formula INV.NORM.ESTAND(p_i). Puesto que la función de distribución de una variable normal es conocida comúnmente por la letra griega Φ entonces como título a la nueva columna computada se coloca comúnmente $\boldsymbol{\Phi^{-1}(p_i)}$**.**

Ahora para hacer el comparativo del gráfico con su comportamiento esperado primero se debe trazar un diagrama de dispersión de los valores $e_{(i)}$ versus los valores de $e_{(i)}$ divididos entre su desviación estándar $\sqrt{\mathbf{Var}(\boldsymbol{e_i})}$, ya que teóricamente si los residuos se distribuyeran de manera normal entonces su estandarización se distribuiría como una Normal(0,1).

El comportamiento esperado de un gráfico de dispersión entre los valores de $e_{(i)}$ versus $\Phi^{-1}(p_i)$ bajo el supuesto de Normalidad de los residuos, es el de una línea recta que pasa por el origen y cuya pendiente está determinada por la dispersión de los residuos. Cuando existen discrepancias entre las distribuciones entonces el comportamiento de la dispersión puede tomar diferentes formas, las cuales se revisarán sobre diversos escenarios de datos, simulados con las diversas distribuciones anteriores para observar el resultado final del gráfico.

En la siguiente imagen se muestra una propuesta para la generación del gráfico de probabilidad Normal en la hoja de cálculo que se ha venido trabajando, la cual muestra los resultados aplicados a los datos simulados en la Tabla 5.11 donde se pueden apreciar los mismos resultados de la tabla en la que no se rechazó la hipótesis Nula la cual menciona que la distribución de los errores es Normal. En el gráfico se puede observar como el patrón de los puntos de la dispersión coinciden con el de la recta esperada.

Estructura en Excel® donde se muestra el gráfico de probabilidad Normal en datos normales.

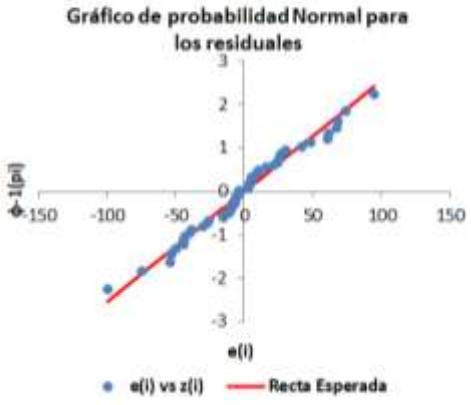
Gráfico de probabilidad Normal

Prueba de Jarque-Bera aplicada a los residuales, basados en errores simulados como una distribución Laplace(0,1/0.05=20)

γ_1	K_1	JB	α	$\chi^2_{2,1-\alpha}$
0.11	3.027	0.102	0.01	9.21

$$p_i = \frac{R_i - \frac{3}{8}}{n + \frac{1}{4}}$$

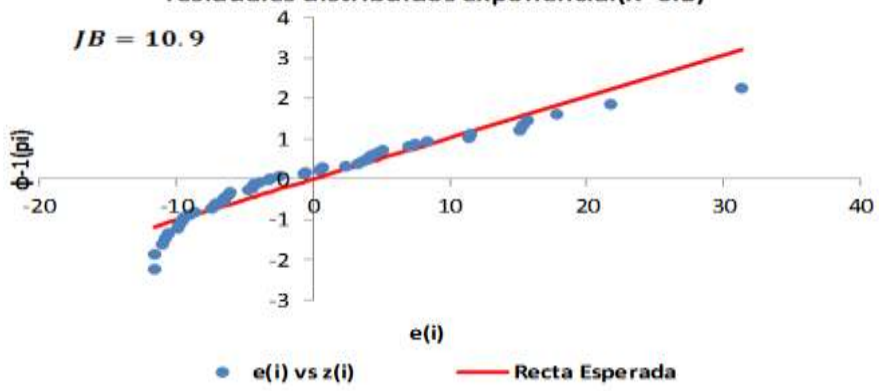
e_i	$e_{(i)}$	p_i	$\Phi^{-1}(p_i)$	$e_{(i)}/\sqrt{\mathrm{Var}(e_i)}$
29.71	-99.90	0.01	-2.243	-2.54
74.17	-74.62	0.03	-1.847	-1.90
24.50	-54.28	0.05	-1.624	-1.38
-7.38	-53.41	0.07	-1.460	-1.36
66.90	-50.32	0.09	-1.328	-1.28
5.17	-44.33	0.11	-1.216	-1.13
4.95	-44.10	0.13	-1.118	-1.12
61.40	-43.26	0.15	-1.029	-1.10
60.60	-39.89	0.17	-0.948	-1.01
-6.08	-38.37	0.19	-0.872	-0.98
2.62	-30.20	0.21	-0.801	-0.77
48.14	-27.10	0.23	-0.734	-0.69
25.26	-26.04	0.25	-0.671	-0.66
94.62	-15.76	0.27	-0.609	-0.40
-10.86	-15.06	0.29	-0.550	-0.38
15.91	-10.94	0.31	-0.493	-0.28



En el siguiente escenario se generó el gráfico de probabilidad normal para los datos de la Tabla 5.12 los cuales se generaron a través de errores distribuidos exponencialmente con $\lambda = 0.1$ y que son mostrados en la gráfica 5.36; para estos datos en particular se puede ver en el grafico 5.39 como la figura que forma la

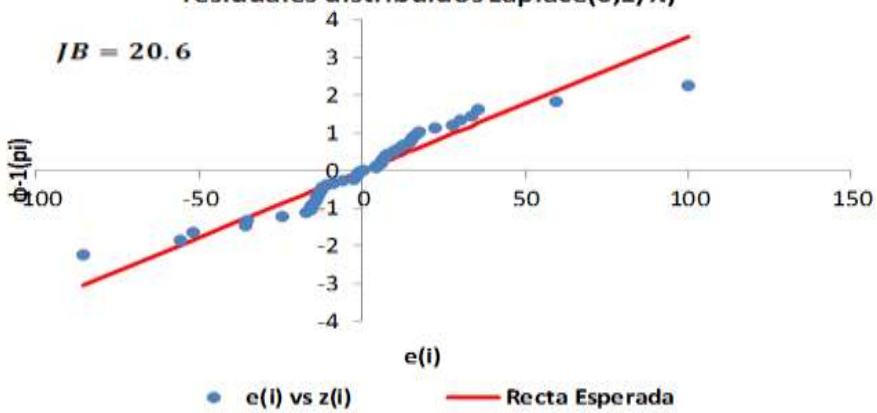
dispersión de los residuos es una curva sesgada hacia el origen, comportamiento provocado por la asimetría de la distribución exponencial. En este gráfico cuando los datos se distribuyen con una asimetría negativa entonces el sesgo de la curva se observará orientada hacia el otro extremo del soporte de los residuos.

Gráfica 5.39: Gráfico de probabilidad Normal para los residuales distribuidos exponencial(λ=0.1)



Por otra parte en un tercer escenario proveniente de los datos de la Tabla 5.13, los cuales fueron generados de la distribución Laplace(0,1/λ) y se visualizan en la Gráfica 5.38, se generó el gráfico de probabilidad Normal, que se puede observar en la gráfica 5.40, donde se puede notar que existen discrepancias entre los estadísticos de orden graficados con la recta esperada. La nube de residuos toma ahora una forma sigmoidea lo cual significa que las colas de la distribución son la fuente de discrepancia con respecto a una Normal, pues entre más alejados estén los puntos en los extremos del soporte con respecto de la recta esperada, entonces será un reflejo de colas pesadas en la distribución de los residuos.

Gráfica 5.40: Gráfico de probabilidad Normal para los residuales distribuidos Laplace(0,1/λ)



En conclusión, con los métodos es posible observar diferentes escenarios sobre los datos modificado desde la muestra, algunos de los supuestos más comunes dentro del análisis de regresión lineal simple. Lo cual sirve también como apoyo previo de análisis más complejos, como el análisis de regresión múltiple o la selección óptima de variables dependientes.

Conclusiones y comentarios finales

Este trabajo tuvo como objetivo proponer un enfoque para la enseñanza de tópicos dentro del currículo básico de un profesionista especializado en la aplicación de técnicas estadísticas con base en métodos de simulación de Monte Carlo.

La estructura y los enfoques mostrados, se han pensado para profesores que deseen generar una serie de ejemplos prácticos al ver los tópicos básicos de estadística, bajo diversos contextos, ya sea basado y acompañado con datos reales o en un contexto hipotético, para enseñar de forma clara las características de las técnicas estadísticas empleadas. Además, se pueden emplear los métodos expuestos, para la generación de muestras con distintas características para que sirvan como ejercicios de tarea y que produzcan distintos resultados.

Otra recomendación para introducir otro tipo de herramientas con el uso de Excel junto con software estadístico, es por medio de la inclusión de complementos de software (Add-in en inglés) que permiten la conexión con el ambiente de hoja de cálculo. Algunos de los complementos que se pueden hallar documentados en la red son por ejemplo, para software comercial como SAS®, Mathematica® o Spreadsheet Link™ para el uso de Matlab® y para software libre como RExcel para la conexión con R®. Estos complementos, por lo general, incluyen la transmisión de datos entre las plataformas utilizadas, la posibilidad de emplear códigos y métodos de forma indistinta entre ellas, además de una interfaz gráfica más familiar con el usuario.

Asimismo, este escrito puede ser consultado por alumnos que deseen iniciar en el aprendizaje de técnicas de simulación Monte Carlo, a quienes se recomienda adicionalmente consultar la bibliografía mas general en temas de simulación como Law(2007), Ríos(2009) y Fishman(2001).

En el apéndice, se encuentran los códigos de programación de las diversas Macro instrucciones en lenguaje *Visual Basic for Applications* (VBA) para Excel® y los códigos en R® que apoyaron el desarrollo tanto de las muestras simuladas como sus gráficos. Se publican para su libre reproducción y modificación, para este propósito, se recomienda consultar la versión digital de este trabajo, para que reproducir el código se resuma en copiar y pegar el texto en un script o en la línea de comandos. Finalmente, ante cualquier error se recomienda revisar la documentación correspondiente de cada software.

Apéndice
Parte I

Pruebas de bondad de ajuste Ji-cuadrada, y Kolmogorov-Smirnov

Cuando se realiza un experimento y se recaban datos, los cuales en conjunto forman una muestra aleatoria, es una cuestión de interés inferir estadísticamente como se distribuyen los datos.

Si la función de distribución es desconocida entonces se puede emplear una prueba bondad de ajuste, pues este tipo de pruebas supone desde un principio que se conoce la distribución de los datos, es decir se toman como hipótesis que los datos provienen de una variable aleatoria con función de distribución F_0, que idealmente, estaría totalmente especificada. Entonces una prueba de este tipo proporcionará los criterios necesarios para medir que tanto se ajusta la distribución de los datos a la distribución propuesta.

Prueba Ji-cuadrada

Esta prueba fue introducida por Karl Pearson en 1900, como una variante de las pruebas que involucraban a las tablas también llamadas de contingencia, que son de utilidad para probar características y relaciones de interés sobre los datos, cuando éstos recogidos dentro de la observación de un fenómeno y puedan presentarse como un arreglo de valores en forma tabular.

La forma general de realizar la prueba es clasificar una muestra de tamaño N, en clases o categorías C_i, propuestas previamente por quien realiza la prueba, para posteriormente agrupar los datos de tal manera que se pueda contar el número de observaciones pertenecientes a cada clase, lo cual se denomina frecuencia observada (O_i); en el caso que se tuvieran k clases predeterminadas, debería obtenerse en primer lugar una tabla como la siguiente

$\boldsymbol{Clase}$	$\boldsymbol{C_1}$	$\boldsymbol{C_2}$	…	$\boldsymbol{C_k}$	$\boldsymbol{Total}$
$\boldsymbol{Freciencia\ Observada}$	$\boldsymbol{O_1}$	$\boldsymbol{O_2}$	…	$\boldsymbol{O_k}$	$\boldsymbol{N}$

Las clases C_k varían de acuerdo al tipo de distribución que se desee ajustar, si es discreta entonces las clases pueden verse como subconjuntos del dominio con cierta cardinalidad cada uno; seccionando el dominio de la forma $\{\{i, i+1, \dots\}, \{m, m+1, \dots\}, \dots\}$.

En el caso de tener una variable categórica los conjuntos pueden verse como los valores de cada categoría o conjuntos que consten de varias categorías.

Cuando se tienen variables que sean definidas en un dominio continuo, entonces las clases son subintervalos del dominio, los cuales se definen de la forma $(a, b]$, puesto que posteriormente se requerirá hacer el cálculo $\boldsymbol{P}(\boldsymbol{a} < X \leq b)$, además los extremos de los intervalos teóricamente no tienen gran importancia debido a la continuidad de la variable.

La forma en que se dividen los intervalos a lo largo de esta tesis, están realizados con la ayuda de la función de Excel, FRECUENCIA ().

Habiendo calculado las frecuencias observadas O_i, se procede a contrastar las siguientes hipótesis

$$\boldsymbol{H_0}: \boldsymbol{la\ funcion\ de\ distribución\ de\ la\ muestra\ observada\ es\ F_0(X)}$$

$$\boldsymbol{H_1}: \boldsymbol{la\ funcionde\ distribucion\ de\ la\ muestra\ es\ distinta\ a\ F_0(X)}$$

Para calcular el estadístico de prueba primero se deben obtener las probabilidades que, suponiendo cierta la hipótesis nula H_0, corresponderían a cada clase, denotándolas como P_{C_k}. Entonces se define E_k como

$$\boldsymbol{E_k = P_{C_k} * N}$$

Que representa el valor esperado que debería tomar la frecuencia de la clase C_k. Dado lo anterior el estadístico de prueba es dado por

$$\boldsymbol{\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}}$$

La distribución del estadístico χ^2 converge asintóticamente a una variable Ji-cuadrada con $k-1$ grados de libertad conforme N crece, tal aproximación es inapropiada si algunos de los valores E_k es pequeño, según recomendaciones del libro de J. D. Gibbons(2003), se debe tomar las clases con un criterio, tal que cada agrupación contenga al menos 5 individuos.

La regla de decisión que implica rechazar la hipótesis nula, dado un nivel de significancia para la prueba de α, es si $\boldsymbol{\chi^2 > X^2_{1-\alpha,(k-1)}}$; donde $\boldsymbol{X^2_{1-\alpha,(k-1)}}$ es el cuantil $(1-\alpha)$ de una variable distribuida Ji-cuadrada con $(k-1)$ grados de libertad.

El cuantil, también llamado valor crítico, se puede obtener por medio de la función DISTR.CHI() de Excel[34] o por medio de una tabla estadística que puede hallarse en diversos libros de Estadística, como por ejemplo en el libro de Gibbons. Por último, es importante aclarar que en el caso que $\chi^2 \leq X^2_{1-\alpha,(k-1)}$ no se concluye aceptar H_0, sino simplemente a no rechazarla.

Prueba de Kolmogorov-Smirnov

La prueba de bondad de ajuste Ji-cuadrada tiene algunas desventajas, en primer lugar al ser pensada originalmente para datos puestos de manera tabulada, migrando la prueba para las distribuciones continuas llega a ser arbitrario el número de clases o subintervalos en los que se divide al dominio de la variable.

Un criterio expuesto en diversos libros es la recomendación de construir los subintervalos de manera que el valor esperado de cada subintervalos sea mayor o igual a 5, con el objetivo de asegurar una prueba insesgada.

[34] El nombre de la función pueden cambiar de acuerdo a la versión o idioma configurados.

Otro tema importante es que la distribución del estadístico se alcanza de manera asintótica, por lo cual la validez de la prueba depende de cierta forma en el tamaño de la muestra, requiriendo una muestra lo suficientemente grande; lo cual en las aplicaciones no siempre es posible.

La prueba de Kolmogorov-Smirnov es más potente que la prueba Ji-cuadrada debido a los detalles anteriormente mencionados, puesto que para realizar la prueba no existe una especificación arbitraria por parte del analista, además el estadístico de prueba cuenta con una distribución exacta para muestras de cierto tamaño, ya que se emplea una distribución asintótica cuando el tamaño de la muestra es mayor a 40.

Para poder realizar esta prueba deben llevarse a cabo cierto manejo de los datos, junto con otros cálculos previos que se enuncian y justifican a continuación.

El primer cálculo previo es la función de distribución empírica de la muestra, que quedó definida en el capítulo II y además cumple ser un estimador insesgado de la distribución de la muestra F, es decir

$$\boldsymbol{E[F_n(X)] = F(X)}$$

Por otro lado su varianza puede ser calculada como

$$\boldsymbol{Var\big(F_n(X)\big) = \frac{\big[F(X)\big(1 - F(X)\big)\big]}{n}}$$

De donde se puede ver que

$$\boldsymbol{\lim_{n\to\infty} Var\big(F_n(X)\big) \Rightarrow 0}$$

Esta propiedad es conocida como consistencia, la cual es deseable cuando se estima uno o más parámetros, pues quiere decir que la aproximación es cada vez mejor cuando el tamaño de la muestra aumenta. Otro resultado importante sobre la distribución empírica es el teorema de Gilvenko-Canteli el cual se enuncia como

$$\boldsymbol{P\left[\lim_{n\to\infty} \sup_{-\infty<x<\infty} |F_n(x) - F(x)| = 0\right] = 1}$$

Otro cálculo importante para realizar la prueba de bondad de ajuste son los estadísticos de orden, los cuales se pueden establecer como, dada una muestra aleatoria $x_1, x_2, \dots, x_n$, se ordena de manera ascendente, el valor que queda en la posición k a partir del menor se le conoce como el k-ésimo estadístico de orden y se denota como $x_{(k)}$.

Los estadísticos de orden tienen importancia dentro de la inferencia estadística pues se puede notar que $X_{(i)} = \min_i x_i$ y $x_{(n)} = \max_i x_i$, lo que nos indica los valores extremos que tomó la muestra; por otra parte calculando $\boldsymbol{x_{(n)} - x_{(1)}}$ se conoce el rango o tamaño de intervalo en el que la muestra toma valores.

Hacer uso de una prueba de bondad de ajuste implica el contrastar dos hipótesis que se infieren sobre la distribución F de una muestra obtenida; las hipótesis de las cuales se parten para la prueba de K-S son

$$\boldsymbol{H_0: F(X) = F_0(X)\ \forall x}$$

$$\boldsymbol{H_1: F(X) \neq F_0(X)\ \ para\ algún\ x}$$

Como se puede notar las hipótesis son idénticas a las de la prueba Ji-cuadrada, aunque en este caso la prueba realiza cálculos donde se nota el contraste sobre las hipótesis de manera más trasparente.

La forma de comparar el ajuste entre la distribución de los datos y la distribución propuesta es por medio de medir para cada valor de la muestra que tanto se separan; para poder hacer la comparación se emplea F_n, pues como ya se enunció, es un buen estimador de F; la separación se mide calculando las distancias verticales entre F_n y F_0.

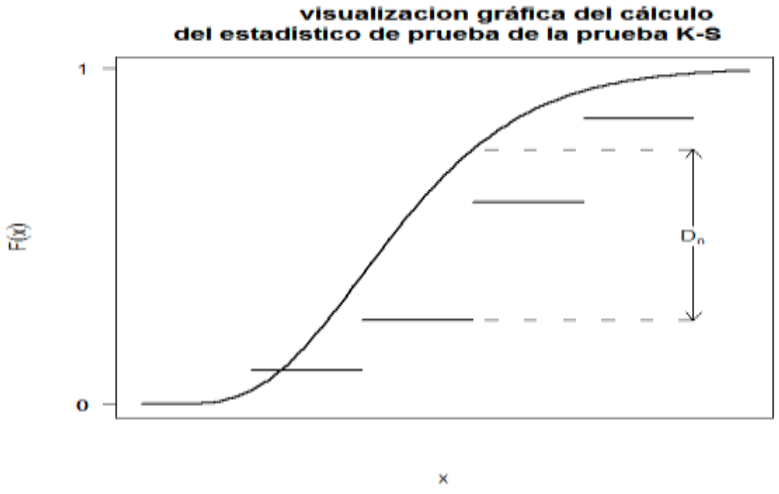
Ya habiendo calculado previamente los estadísticos de orden de la muestra $x_1, x_2, \dots, x_n$; se procede a evaluarlos sobre la función F_0 para calcular las cantidades auxiliares

$$\boldsymbol{D_n^+ = \max_{1 \le i \le n}\left\{\frac{i}{n} - F_0(x_{(i)})\right\}} \quad \boldsymbol{y} \quad \boldsymbol{D_n^- = \max_{1 \le i \le n}\left\{F_0(x_{(i)}) - \frac{i-1}{n}\right\}}$$

Luego para realizar una prueba de dos colas con la distancia $D_n = \sup\{|F_n(x) - F_0(x)|\}$ se toma como estadístico de prueba

$$\boldsymbol{D_n = \max\{D_n^+, D_n^-\}}$$

La grafica de la función de distribución empírica es escalonada, y como se mencionó anteriormente se desea comparar por medio de las distancias verticales las distancias entre la función empírica y la función de distribución F_0; la visualización del proceso de comparación por medio del cálculo de D_n se muestra a continuación.



En el caso de la gráfica el estadístico D_n coincide con D_n^- , esto es debido a que la función de distribución F_0 queda por encima de la empírica; además, de las ecuaciones del cálculo de D_n^- se puede notar que, toma las distancias del extremo derecho del "escalón" con respecto a F_0; inversamente la distancia máxima por el lado

izquierdo del "escalón" a F_0 vendrá dado por D_n^+, en el caso que la función de distribución F_0 quede lo suficientemente por debajo de la empírica.

La Regla establecida para enunciar una conclusión, con el valor previamente obtenido del valor del estadístico de prueba y estableciendo un nivel de significancia α, puede ser de dos formas.

La primera es por medio del cálculo del $p-value$,es decir tomando a D_n como el cuantil de una cierta distribución D, $p-value = P(D_n > D)$, el cual se puede obtener en alguna tabla estadística.

Entonces para el caso en que se conoce el valor del $p-value$, la regla de decisión es: Si $p-value \leq \alpha$ entonces se debe rechazar la hipótesis nula, en caso contrario, la hipótesis nula no debe ser rechazada.

Otra manera de establecer una conclusión es por medio de valores críticos obtenidos de una tabla especifica que se usa para realizar la prueba K-S.

Cuando la muestra es mayor a 40 se usan aproximaciones para el valor del valor crítico, las cuales son muy útiles para experimentos de simulación pues en general el tamaño de muestra es establecido por el analista. Los valores críticos que se pueden hallar en las tablas para muestras grandes son

Valores críticos para la prueba K-S si n>50								
$\boldsymbol{nivel\ de\ significancia\ \alpha}$	$\mathbf{0.2}$	$\mathbf{0.1}$	$\mathbf{0.05}$	$\mathbf{0.02}$	$\mathbf{0.01}$	$\mathbf{0.005}$	$\mathbf{0.002}$	$\mathbf{0.001}$
$\boldsymbol{valor\ crítico\ d_{1-\alpha,n}}$	$\frac{\mathbf{1.07}}{\sqrt{\boldsymbol{n}}}$	$\frac{\mathbf{1.22}}{\sqrt{\boldsymbol{n}}}$	$\frac{\mathbf{1.36}}{\sqrt{\boldsymbol{n}}}$	$\frac{\mathbf{1.52}}{\sqrt{\boldsymbol{n}}}$	$\frac{\mathbf{1.63}}{\sqrt{\boldsymbol{n}}}$	$\frac{\mathbf{1.73}}{\sqrt{\boldsymbol{n}}}$	$\frac{\mathbf{1.85}}{\sqrt{\boldsymbol{n}}}$	$\frac{\mathbf{1.95}}{\sqrt{\boldsymbol{n}}}$

Entonces tomado un valor crítico $d_{1-\alpha,n}$ a un nivel de significancia α, para el tamaño de muestra n, la regla de decisión es: Si $d_{1-\alpha,n} \leq D_n$ entonces se debe rechazar la hipótesis nula; en caso contrario, no debe rechazarse la hipótesis nula.

Parte II

Códigos en lenguaje macros para realizar las simulaciones Monte Carlo

Como herramienta de cálculo, a lo largo de este fue empleado el lenguaje VBA para trabajar en ambiente de hoja de cálculo, desarrollado por la Microsoft Corporation®, aunque no es el software más avanzado y especializado en el análisis estadístico, ni en el manejo eficiente bases de datos, en la actualidad es ampliamente usada, por lo que se eligió, debido a su accesibilidad y fácil manejo.

A pesar de su manejo relativamente sencillo para cualquier iniciado en su uso, es importante señalar que existe un lenguaje intrínseco en el software que permite la automatización de las acciones de Excel, conocido como proceso Macro el cual se basa en el lenguaje Visual Basic. A continuación, se presentan los diferentes códigos empleados en este trabajo, que, en una primera parte, son aquellos que se utilizaron para la generación de valores simulados de las diferentes distribuciones revisadas en el Capítulo I.

Se tuvo como objetivo simplificar lo más posible la programación, con instrucciones básicas de bucles y condicionales, automatizando los mismos pasos que alguien realizaría en la hoja de cálculo para simular estas variables, por lo que todas las fórmulas y detalles quedan plasmados.

Variables discretas

Uniforme Discreta

```
Sub Unif_Disc()
'Los siguientes códigos son trabajo del autor, y se publican para su libre reproducción y/o modificación
'Guillermo Cuauhtemoctzin Granados García, Facultad de ciencias, UNAM(2017)
'El programa genera una variable discreta uniforme entre dos enteros (a,b), e incluye una gráfica de barras opcional
'Definición de variables
Dim ene As Long
Dim ini As Double
Dim N As Double
Dim R As Range
Dim rand As Double
Dim X(1) As Double
'los parámetros a elegir se insertan en un cuadro de diálogo
ene = InputBox("Indique el número total de simulaciones", "N", 1000)
ini = InputBox("Indique el entero inicial i", "i", 0)
N = InputBox("Indique el entero final n" & vbCrLf & _
"da la forma: entero, decimales", "n", 10)
'coloca titulos a las columnas por generar
If Selection.Column > 1 Then
ActiveCell.Offset(0, -1).Value = "Indice"
```

```
ActiveCell.Offset(1, -1).Value = 1
ActiveCell.Offset(2, -1).Value = 2
Set r1 = Range(Selection.Offset(1, -1), Selection.Offset(2, -1))
Set R = Range(Selection.Offset(1, -1), Selection.Offset(ene, -1))
r1.AutoFill Destination:=R, Type:=xlFillDefault
End If
Selection.Value = "UDiscreta(" & ini & ";" & N & ")"
Selection.Columns(1).EntireColumn.AutoFit
'Se generan los valores aleatorios uniformes
For i = 1 To ene
Randomize
rand = Rnd()
'se trunca el valor generado
ActiveCell.Offset(i, 0).Value = Int(((N - ini + 1) * rand)) + ini
Next i
'se calcula la media y varianza muestral de los valores generados
ActiveCell.Offset(0, 1).Value = "Media Est."
ActiveCell.Offset(0, 2).FormulaR1C1 = "=AVERAGE(R[1]C[-2]:R[" & ene & "]C[-2])"
ActiveCell.Offset(1, 1).Value = "Var. Est."
ActiveCell.Offset(1, 2).FormulaR1C1 = "=VAR(R[0]C[-2]:R[" & (ene - 1) & "]C[-2])"
'Pregunta en un cuadro si se desea agregar el histograma de frecuencias de la muestra
Dim resp As Integer
resp = MsgBox("¿Desea agregar una gráfica de frecuencias?", vbQuestion + vbYesNo,
"Histograma")
If resp = 6 Then
' Se calcula el rango, una partición y las frecuencias de cada segmento.
Dim Z1 As Range
ActiveCell.Offset(0, 3).Value = "Rango"
ActiveCell.Offset(0, 4).Value = "Frecuencia"
Selection.Columns(5).EntireColumn.AutoFit
ActiveCell.End(xlToRight).Select
For i = 1 To (N - ini + 1)
ActiveCell.Offset(i, -1).Value = ini + i - 1
Next i
Set R = Range(Selection.Offset(1, 0), Selection.Offset((N - ini + 1), 0))
R.Select
Selection.FormulaArray = "=FREQUENCY(RC[-4]:R[" & (ene - 1) & "]C[-4],RC[-1]:R[" & (N - ini +
1) & "]C[-1])"
ActiveCell.Offset(-1, 0).Select
' se selecciona la muestra y se genera el gráfico de barras
Set R = Range(Selection, Selection.Offset((N - ini + 1), 0))
Set Z1 = Range(Selection.Offset(1, -1), Selection.Offset((N - ini + 1), -1))
   ActiveSheet.Shapes.AddChart.Select
   ActiveChart.SetSourceData Source:=R
   ActiveChart.ChartType = xlColumnClustered
```

```
   ActiveChart.SeriesCollection(1).XValues = Z1
      ActiveChart.SeriesCollection(1).Select
      ActiveChart.ChartGroups(1).GapWidth = 0
   ActiveChart.SetElement (msoElementChartTitleAboveChart)
   ActiveChart.ChartTitle.Text = "Frecuencias Empíricas"
Else
Exit Sub
End If
'fin del programa
End Sub
```

Distribución Bernoulli

```
Sub Bernoulli()
'Genera una variable discreta con distribución Bernoullli entre dos enteros (a,b)
'Además incluye el cálculo de la media, varianza y genera el histograma de forma opcional
'Definición de variables
Dim ene As Long
Dim pe As Double
Dim R As Range
Dim rand As Double
Dim X(1) As Double
'se introducen los parámetros en cuadros de dialogo
ene = InputBox("Indique el número total de simulaciones", "N", 1000)
pe = InputBox("Indique la probabilidad de éxito", "P", 0.5)
'se generan los títulos de las columnas a generar
If Selection.Column > 1 Then
ActiveCell.Offset(0, -1).Value = "Indice"
ActiveCell.Offset(1, -1).Value = 1
ActiveCell.Offset(2, -1).Value = 2
Set r1 = Range(Selection.Offset(1, -1), Selection.Offset(2, -1))
Set R = Range(Selection.Offset(1, -1), Selection.Offset(ene, -1))
r1.AutoFill Destination:=R, Type:=xlFillDefault
End If
Selection.Value = "Bernoulli(" & pe & ")"
Selection.Columns(1).EntireColumn.AutoFit
'se generan los resultados de los ensayos Bernoulli con probabilidad p
For i = 1 To ene
Randomize
rand = Rnd()
If rand <= pe Then
ActiveCell.Offset(i, 0).Value = 1
Else
ActiveCell.Offset(i, 0).Value = 0
End If
Next i
```

```
'se calculan la media y la varianza de los valores generados
ActiveCell.Offset(0, 1).Value = "Media Est."
ActiveCell.Offset(0, 2).FormulaR1C1 = "=AVERAGE(R[1]C[-2]:R[" & ene & "]C[-2])"
ActiveCell.Offset(1, 1).Value = "Var. Est."
ActiveCell.Offset(1, 2).FormulaR1C1 = "=VAR(R[0]C[-2]:R[" & (ene - 1) & "]C[-2])"
'Se pregunta si se desea el histograma
Dim resp As Integer
resp = MsgBox("¿Desea agregar una Gráfica de Frecuencias?", vbQuestion + vbYesNo,
"Histograma")
If resp = 6 Then
Dim ent As Long
Dim Z1 As Range
'En caso afirmativo se identifica el rango de la muestra y se divide para calcular las 'frecuencias
ActiveCell.Offset(0, 3).Value = "Rango"
ActiveCell.Offset(0, 4).Value = "Frecuencia"
Selection.Columns(5).EntireColumn.AutoFit
ActiveCell.End(xlToRight).Select
ActiveCell.Offset(1, -1).Value = 0
ActiveCell.Offset(2, -1).Value = 1
Set R = Range(Selection.Offset(1, 0), Selection.Offset(2, 0))
R.Select
Selection.FormulaArray = "=FREQUENCY(RC[-4]:R[" & (ene - 1) & "]C[-4],RC[-1]:R[1]C[-1])"
ActiveCell.Offset(-1, 0).Select
Set R = Range(Selection, Selection.Offset(2, 0))
Set Z1 = Range(Selection.Offset(1, -1), Selection.Offset(2, -1))
   ' se genera el gráfico de barras
   ActiveSheet.Shapes.AddChart.Select
   ActiveChart.SetSourceData Source:=R
   ActiveChart.ChartType = xlColumnClustered
   ActiveChart.SeriesCollection(1).XValues = Z1
      ActiveChart.SeriesCollection(1).Select
      ActiveChart.ChartGroups(1).GapWidth = 2
   ActiveChart.SetElement (msoElementChartTitleAboveChart)
   ActiveChart.ChartTitle.Text = "Frecuencias Empíricas"
Else
Exit Sub
End If
End Sub
```

Distribución Binomial

```
Sub Binomial()
' genera una variable discreta con distribución Binomial con los parámetros (p,n)
'además incluye el cálculo de la media, varianza y genera el histograma de forma opcional
'Definición de variables
```

```
Dim ene As Long
Dim pe As Double
Dim N As Double
Dim R As Range
Dim rand As Double
Dim X(1) As Double
'se introducen los parámetros en cuadros de dialogo
ene = InputBox("Indique el número total de simulaciones", "N", 1000)
pe = InputBox("Indique la probabilidad de éxito" & vbCrLf & _
"da la forma: entero,decimales" & vbCrLf & _
"Recuerde que E(X)= nP" & vbCrLf & _
"VAR(X)=nP(1-P)", "P", 0.5)
N = InputBox("Indique el numero de intentos n" & vbCrLf & _
"Recuerde que E(X)=nP" & vbCrLf & _
"VAR(X)=nP(1-P)", "n", 10)
'se generan los títulos de las columnas a simular
If Selection.Column > 1 Then
ActiveCell.Offset(0, -1).Value = "Indice"
ActiveCell.Offset(1, -1).Value = 1
ActiveCell.Offset(2, -1).Value = 2
Set r1 = Range(Selection.Offset(1, -1), Selection.Offset(2, -1))
Set R = Range(Selection.Offset(1, -1), Selection.Offset(ene, -1))
r1.AutoFill Destination:=R, Type:=xlFillDefault
End If
Selection.Value = "Binomial(" & N & ";" & pe & ")"
Selection.Columns(1).EntireColumn.AutoFit
Dim sum As Long
'se generan los valores como la suma de resultados Bernoulli
For i = 1 To ene
sum = 0
For j = 1 To N
rand = Rnd()
If rand < pe Then
sum = sum + 1
End If
Next j
ActiveCell.Offset(i, 0).Value = sum
Next i

'Se calcula la media y la varianza de la muestra generada
ActiveCell.Offset(0, 1).Value = "Media Est."
ActiveCell.Offset(0, 2).FormulaR1C1 = "=AVERAGE(R[1]C[-2]:R[" & ene & "]C[-2])"
ActiveCell.Offset(1, 1).Value = "Var. Est."
ActiveCell.Offset(1, 2).FormulaR1C1 = "=VAR(R[0]C[-2]:R[" & (ene - 1) & "]C[-2])"
'Pregunta si se debe incluir el histograma
```

```
Dim resp As Integer
resp = MsgBox("¿Desea agregar una gráfica de frecuencias?", vbQuestion + vbYesNo,
"Histograma")
If resp = 6 Then
'En caso afirmativo identifica el rango y lo divide para obtener las frecuencias de cada
'segmento
Dim Z1 As Range
ActiveCell.Offset(0, 3).Value = "Rango"
ActiveCell.Offset(0, 4).Value = "Frecuencia"
Selection.Columns(5).EntireColumn.AutoFit
ActiveCell.End(xlToRight).Select
For i = 0 To N
ActiveCell.Offset((i + 1), -1).Value = i
Next i
Set R = Range(Selection.Offset(1, 0), Selection.Offset((N + 1), 0))
R.Select
Selection.FormulaArray = "=FREQUENCY(RC[-4]:R[" & (ene - 1) & "]C[-4],RC[-1]:R[" & (N + 1) &
"]C[-1])"
ActiveCell.Offset(-1, 0).Select
Set R = Range(Selection, Selection.Offset((N + 2), 0))
Set Z1 = Range(Selection.Offset(1, -1), Selection.Offset((N + 2), -1))
'se genera el gráfico de barras
   ActiveSheet.Shapes.AddChart.Select
   ActiveChart.SetSourceData Source:=R
   ActiveChart.ChartType = xlColumnClustered
   ActiveChart.SeriesCollection(1).XValues = Z1
      ActiveChart.SeriesCollection(1).Select
      ActiveChart.ChartGroups(1).GapWidth = 0
   ActiveChart.SetElement (msoElementChartTitleAboveChart)
   ActiveChart.ChartTitle.Text = "Frecuencias Empíricas"

Else
Exit Sub
End If
'fin del programa
End Sub
```

Distribución Geométrica

```
Sub Geometrica()
'Genera una variable discreta con distribución Geométrica con parámetro (p)
'Además incluye el cálculo de la media, varianza, máximo, mínimo y genera el histograma de
forma opcional
```

```
'Definición de variables
Dim ene As Long
Dim pe As Double
Dim R As Range
Dim rand As Double
'los parámetros se introducen en un cuadro de diálogo
ene = InputBox("Indique el número total de simulaciones", "N", 1000)
pe = InputBox("Indique la probabilidad de exito", "P", 0.5)
'genera una columna de números índice
If Selection.Column > 1 Then
ActiveCell.Offset(0, -1).Value = "Indice"
ActiveCell.Offset(1, -1).Value = 1
ActiveCell.Offset(2, -1).Value = 2
Set r1 = Range(Selection.Offset(1, -1), Selection.Offset(2, -1))
Set R = Range(Selection.Offset(1, -1), Selection.Offset(ene, -1))
r1.AutoFill Destination:=R, Type:=xlFillDefault
End If
'genera el título de la variable a simular
Selection.Value = "Geométrica(" & pe & ")"
Selection.Columns(1).EntireColumn.AutoFit
'se emplea la técnica de la distribución inversa generalizada para esta variable discreta
For i = 1 To ene
Randomize
rand = Rnd()
ActiveCell.Offset(i, 0).Value = Int((Log(rand) / Log((1 - pe))))

Next i
'se calcula la media, varianza , el mínimo y máximo de la  muestra
ActiveCell.Offset(0, 1).Value = "Media Est."
ActiveCell.Offset(0, 2).FormulaR1C1 = "=AVERAGE(R[1]C[-2]:R[" & ene & "]C[-2])"
ActiveCell.Offset(1, 1).Value = "Var. Est."
ActiveCell.Offset(1, 2).FormulaR1C1 = "=VAR(R[0]C[-2]:R[" & (ene - 1) & "]C[-2])"
ActiveCell.Offset(0, 3).Value = "Máximo"
ActiveCell.Offset(0, 4).FormulaR1C1 = "=max(R[1]C[-4]:R[" & ene & "]C[-4])"
ActiveCell.Offset(1, 3).Value = "Mínimo"
ActiveCell.Offset(1, 4).FormulaR1C1 = "=min(R[0]C[-4]:R[" & (ene - 1) & "]C[-4])"
'pregunta si se desea el histograma
Dim resp As Integer
resp = MsgBox("¿Desea agregar una Gráfica de Frecuencias?", vbQuestion + vbYesNo,
"Histograma")
'en caso afirmativo obtiene la frecuencia de cada uno de los valores, en el rango de la muestra
If resp = 6 Then
Dim ent As Long
ent = ActiveCell.Offset(0, 4).Value
Dim Z1 As Range
```

```
ActiveCell.Offset(0, 5).Value = "Rango"
ActiveCell.Offset(0, 6).Value = "Frecuencia"
Selection.Columns(7).EntireColumn.AutoFit
ActiveCell.End(xlToRight).Select
'se grafican los valores del 0 al máximo obtenido
 For i = 0 To ent
ActiveCell.Offset((i + 1), -1).Value = i
Next i
Set R = Range(Selection.Offset(1, 0), Selection.Offset((ent + 1), 0))
R.Select
Selection.FormulaArray = "=FREQUENCY(RC[-6]:R[" & (ene - 1) & "]C[-6],RC[-1]:R[" & ent &
"]C[-1])"
ActiveCell.Offset(-1, 0).Select
Set R = Range(Selection, Selection.Offset((ent + 1), 0))
Set Z1 = Range(Selection.Offset(1, -1), Selection.Offset((ent + 1), -1))
   ' se genera la gráfica de barras
   ActiveSheet.Shapes.AddChart.Select
   ActiveChart.SetSourceData Source:=R
   ActiveChart.ChartType = xlColumnClustered
   ActiveChart.SeriesCollection(1).XValues = Z1
      ActiveChart.SeriesCollection(1).Select
      ActiveChart.ChartGroups(1).GapWidth = 0
   ActiveChart.SetElement (msoElementChartTitleAboveChart)
   ActiveChart.ChartTitle.Text = "Frecuencias Empíricas"

Else
Exit Sub
End If
'fin del programa
End Sub
```

Distribución Binomial Negativa

```
Sub BinomialNeg()
' este código genera una variable discreta con distribución Binomial Negativa de parámetros
(r,p)
'además incluye el cálculo de la media, varianza, máximo, mínimo y genera el histograma de
forma opcional
' Aunque tenga un dominio infinito la simulación tendrá un máximo que puede variar, el cual
se calcula y toma la gráfica, el ingreso de decimales puede depender de la versión de Excel
'Definición de variables
Dim ene As Long
Dim pe As Double
```

```
Dim R As Range
Dim rand As Double
'se introducen los parámetros en un cuadro de dialogo
ene = InputBox("Indique el número total de simulaciones", "N", 1000)
pe = InputBox("Indique la probabilidad de éxito", "P", 0.5)
N = InputBox("Indique el numero de éxitos a contar: r", "r", 10)
'genera una columna de números índice
If Selection.Column > 1 Then
ActiveCell.Offset(0, -1).Value = "Indice"
ActiveCell.Offset(1, -1).Value = 1
ActiveCell.Offset(2, -1).Value = 2
Set r1 = Range(Selection.Offset(1, -1), Selection.Offset(2, -1))
Set R = Range(Selection.Offset(1, -1), Selection.Offset(ene, -1))
r1.AutoFill Destination:=R, Type:=xlFillDefault
End If
'Título de la columna de valores simulados
Selection.Value = "BinNeg(" & N & ";" & pe & ")"
Selection.Columns(1).EntireColumn.AutoFit
Dim sum As Long
'se generar los valores de la distribución BN como la suma de valores de una distribución
'geométrica
For i = 1 To ene
sum = 0
For j = 1 To N
rand = Rnd()
sum = sum + Int((Log(rand) / Log((1 - pe))))
Next j
ActiveCell.Offset(i, 0).Value = sum
Next i
'Se calcula, la media, varianza, mínimo y máximo de la muestra
ActiveCell.Offset(0, 1).Value = "Media Est."
ActiveCell.Offset(0, 2).FormulaR1C1 = "=AVERAGE(R[1]C[-2]:R[" & ene & "]C[-2])"
ActiveCell.Offset(1, 1).Value = "Var. Est."
ActiveCell.Offset(1, 2).FormulaR1C1 = "=VAR(R[0]C[-2]:R[" & (ene - 1) & "]C[-2])"
ActiveCell.Offset(0, 3).Value = "Máximo"
ActiveCell.Offset(0, 4).FormulaR1C1 = "=max(R[1]C[-4]:R[" & ene & "]C[-4])"
ActiveCell.Offset(1, 3).Value = "Mínimo"
ActiveCell.Offset(1, 4).FormulaR1C1 = "=min(R[0]C[-4]:R[" & (ene - 1) & "]C[-4])"
'Pregunta si se desea el histograma
Dim resp As Integer
resp = MsgBox("¿Desea agregar una Gráfica de Frecuencias?", vbQuestion + vbYesNo,
"Histograma")
If resp = 6 Then
Dim ent As Long
ent = ActiveCell.Offset(0, 4).Value
```

```
ini = ActiveCell.Offset(1, 4).Value
Dim Z1 As Range
'en caso afirmativo, calcula la frecuencia para cada valor en el rango de la muestra
ActiveCell.Offset(0, 5).Value = "Rango"
ActiveCell.Offset(0, 6).Value = "Frecuencia"
Selection.Columns(7).EntireColumn.AutoFit
ActiveCell.End(xlToRight).Select
'valores 0 al maximo generado y graficar
For i = 1 To (ent - ini + 1)
ActiveCell.Offset(i, -1).Value = ini + i - 1
Next i
Set R = Range(Selection.Offset(1, 0), Selection.Offset((ent - ini + 1), 0))
R.Select
Selection.FormulaArray = "=FREQUENCY(RC[-6]:R[" & (ene - 1) & "]C[-6],RC[-1]:R[" & (ent - ini +
1) & "]C[-1])"
ActiveCell.Offset(-1, 0).Select
Set R = Range(Selection, Selection.Offset((ent - ini + 1), 0))
Set Z1 = Range(Selection.Offset(1, -1), Selection.Offset((ent - ini + 1), -1))
' aquí genera la gráfica de barras
   ActiveSheet.Shapes.AddChart.Select
   ActiveChart.SetSourceData Source:=R
   ActiveChart.ChartType = xlColumnClustered
   ActiveChart.SeriesCollection(1).XValues = Z1
      ActiveChart.SeriesCollection(1).Select
      ActiveChart.ChartGroups(1).GapWidth = 0
   ActiveChart.SetElement (msoElementChartTitleAboveChart)
   ActiveChart.ChartTitle.Text = "Frecuencias Empíricas"
Else
Exit Sub
End If
End Sub 'fin de programa
```

Distribución Hipergeométrica

```
Sub Hipergeometrica()
'este código genera valores de una variable discreta con distribución Hipergeométrica de
parámetros (m,n,N)
'además incluye el cálculo de la media, varianza, máximo, mínimo y genera el histograma de
forma opcional
'Definición de variables
Dim ene As Long
Dim m As Double
Dim N As Long
Dim rand As Double
```

```
Dim ntent As Long
'se introducen los parámetros en un cuadro de dialogo
ene = InputBox("Indique el número total de simulaciones", "Simulaciones", 1000)
N = InputBox("Indique el total de elementos N", "N", 100)
ntent = InputBox("indique el número de intentos n <= N", "n", 50)
m = InputBox("Indique el número de elementos" & vbCrLf & _
"con la caracteristica objetivo m < N", "m", 10)
'Genera una columna de números índice
If Selection.Column > 1 Then
ActiveCell.Offset(0, -1).Value = "Indice"
ActiveCell.Offset(1, -1).Value = 1
ActiveCell.Offset(2, -1).Value = 2
Set r1 = Range(Selection.Offset(1, -1), Selection.Offset(2, -1))
Set R = Range(Selection.Offset(1, -1), Selection.Offset(ene, -1))
r1.AutoFill Destination:=R, Type:=xlFillDefault
End If
'se genera el título de la columna de valores simulados
Selection.Value = "HiperGe(" & N & ";" & ntent & ";" & m & ")"
Selection.Columns(1).EntireColumn.AutoFit
Dim N1 As Long
Dim m1 As Long
'aquí reproduce el pseudocódigo revisado en el capítulo I, reproduciendo el proceso del
muestro de la hipergeométrica
For i = 1 To ene
N1 = N
m1 = m
sum = 0
For j = 1 To ntent
rand = Rnd()
If rand <= (m1 / N1) Then
sum = sum + 1
m1 = m1 - 1
End If
N1 = N1 - 1
Next j
ActiveCell.Offset(i, 0).Value = sum
Next i
'se calcula la media, varianza, máximo y mínimo de la muestra
ActiveCell.Offset(0, 1).Value = "Media Est."
ActiveCell.Offset(0, 2).FormulaR1C1 = "=AVERAGE(R[1]C[-2]:R[" & ene & "]C[-2])"
ActiveCell.Offset(1, 1).Value = "Var. Est."
ActiveCell.Offset(1, 2).FormulaR1C1 = "=VAR(R[0]C[-2]:R[" & (ene - 1) & "]C[-2])"
ActiveCell.Offset(0, 3).Value = "Máximo"
ActiveCell.Offset(0, 4).FormulaR1C1 = "=max(R[1]C[-4]:R[" & ene & "]C[-4])"
ActiveCell.Offset(1, 3).Value = "Mínimo"
```

```
ActiveCell.Offset(1, 4).FormulaR1C1 = "=min(R[0]C[-4]:R[" & (ene - 1) & "]C[-4])"
'pregunta si se desea incluir el histograma
Dim resp As Integer
resp = MsgBox("¿Desea agregar una Gráfica de Frecuencias?", vbQuestion + vbYesNo,
"Histograma")
'en caso afirmativo calcula la frecuencia
If resp = 6 Then
Dim ent As Long
ent = ActiveCell.Offset(0, 4).Value
ini = ActiveCell.Offset(1, 4).Value
Dim Z1 As Range
ActiveCell.Offset(0, 5).Value = "Rango"
ActiveCell.Offset(0, 6).Value = "Frecuencia"
Selection.Columns(7).EntireColumn.AutoFit
ActiveCell.End(xlToRight).Select
'los cálculos los realiza para os valores entre el mínimo y el máximo obtenidos
For i = 1 To (ent - ini + 1)
ActiveCell.Offset(i, -1).Value = ini + i - 1
Next i
Set R = Range(Selection.Offset(1, 0), Selection.Offset((ent - ini + 1), 0))
R.Select
Selection.FormulaArray = "=FREQUENCY(RC[-6]:R[" & (ene - 1) & "]C[-6],RC[-1]:R[" & (ent - ini +
1) & "]C[-1])"
ActiveCell.Offset(-1, 0).Select
Set R = Range(Selection, Selection.Offset((ent - ini + 1), 0))
Set Z1 = Range(Selection.Offset(1, -1), Selection.Offset((ent - ini + 1), -1))
   ' se genera el grafico de barras
   ActiveSheet.Shapes.AddChart.Select
   ActiveChart.SetSourceData Source:=R
   ActiveChart.ChartType = xlColumnClustered
   ActiveChart.SeriesCollection(1).XValues = Z1
      ActiveChart.SeriesCollection(1).Select
      ActiveChart.ChartGroups(1).GapWidth = 0
   ActiveChart.SetElement (msoElementChartTitleAboveChart)
   ActiveChart.ChartTitle.Text = "Frecuencias Empíricas"
Else
Exit Sub
End If
'fin del programa
End Sub
```

Distribución Poisson

```
Sub Poisson()
' este código genera valores de una variable discreta con distribución Poisson de parámetro
(Lambda)
```

```
'además incluye el cálculo de la media, varianza, máximo, mínimo y genera el histograma de forma opcional
'Definición de variables
Dim ene As Long
Dim lam As Double
Dim rand As Double
'se introduce el parámetro y tamaño de la muestra en cuadros de dialogo
ene = InputBox("Indique el número total de simulaciones", "Simulaciones", 1000)
lam = InputBox("Indique el parámetro Lambda" & vbCrLf & _
"Recuerde que E(X)=Var(X)=Lambda", "Lambda", 5)
'genera una columna con números índice
If Selection.Column > 1 Then
ActiveCell.Offset(0, -1).Value = "Indice"
ActiveCell.Offset(1, -1).Value = 1
ActiveCell.Offset(2, -1).Value = 2
Set r1 = Range(Selection.Offset(1, -1), Selection.Offset(2, -1))
Set R = Range(Selection.Offset(1, -1), Selection.Offset(ene, -1))
r1.AutoFill Destination:=R, Type:=xlFillDefault
End If
'Coloca el título a la columna con los valores a simular
Selection.Value = "Poisson(" & lam & ")"
Selection.Columns(1).EntireColumn.AutoFit
Dim a As Double
Dim b As Double
'aquí reproduce el pseudocódigo que se halla en el capítulo I, para esta distribución empleando una propiedad
a = Exp(-lam)
Dim j As Long
For i = 1 To ene
b = 1
j = -1
While a <= b
b = b * Rnd()
j = j + 1
Wend
ActiveCell.Offset(i, 0).Value = j
Next i
'se calcula la media, varianza, máximo y mínimo de la muestra
ActiveCell.Offset(0, 1).Value = "Media Est."
ActiveCell.Offset(0, 2).FormulaR1C1 = "=AVERAGE(R[1]C[-2]:R[" & ene & "]C[-2])"
ActiveCell.Offset(1, 1).Value = "Var. Est."
ActiveCell.Offset(1, 2).FormulaR1C1 = "=VAR(R[0]C[-2]:R[" & (ene - 1) & "]C[-2])"
ActiveCell.Offset(0, 3).Value = "Máximo"
ActiveCell.Offset(0, 4).FormulaR1C1 = "=max(R[1]C[-4]:R[" & ene & "]C[-4])"
ActiveCell.Offset(1, 3).Value = "Mínimo"
```

```
ActiveCell.Offset(1, 4).FormulaR1C1 = "=min(R[0]C[-4]:R[" & (ene - 1) & "]C[-4])"
'pregunta si se desea incluir el histograma
Dim resp As Integer
resp = MsgBox("¿Desea agregar una Gráfica de Frecuencias?", vbQuestion + vbYesNo,
"Histograma")
'en caso afirmativo identifica el rango de 0 al máximo y obtiene la frecuencia para cada valor
If resp = 6 Then
Dim ent As Long
ent = ActiveCell.Offset(0, 4).Value
ini = ActiveCell.Offset(1, 4).Value
Dim Z1 As Range
ActiveCell.Offset(0, 5).Value = "Rango"
ActiveCell.Offset(0, 6).Value = "Frecuencia"
Selection.Columns(7).EntireColumn.AutoFit
ActiveCell.End(xlToRight).Select
'aquí coloca la fórmulas de frecuencia en la hoja
For i = 1 To (ent - ini + 1)
ActiveCell.Offset(i, -1).Value = ini + i - 1
Next i
Set R = Range(Selection.Offset(1, 0), Selection.Offset((ent - ini + 1), 0))
R.Select
Selection.FormulaArray = "=FREQUENCY(RC[-6]:R[" & (ene - 1) & "]C[-6],RC[-1]:R[" & (ent - ini +
1) & "]C[-1])"
ActiveCell.Offset(-1, 0).Select
Set R = Range(Selection, Selection.Offset((ent - ini + 1), 0))
Set Z1 = Range(Selection.Offset(1, -1), Selection.Offset((ent - ini + 1), -1))
 'aqui se genera la gráfica de barras
   ActiveSheet.Shapes.AddChart.Select
   ActiveChart.SetSourceData Source:=R
   ActiveChart.ChartType = xlColumnClustered
   ActiveChart.SeriesCollection(1).XValues = Z1
      ActiveChart.SeriesCollection(1).Select
      ActiveChart.ChartGroups(1).GapWidth = 0
   ActiveChart.SetElement (msoElementChartTitleAboveChart)
   ActiveChart.ChartTitle.Text = "Frecuencias Empíricas"
Else
Exit Sub
End If
'fin del programa
End Sub
```

Variables continuas

Distribución Uniforme

```
Sub Unif_cont()
'Este código genera valores de una variable continua con distribución uniforme en el intervalo (a,b)
'además incluye el cálculo de la media, varianza, máximo, mínimo y genera el histograma de forma opcional
'Definición de variables
Dim a As Double
Dim b As Double
Dim ene As Long
Dim R As Range
Dim r1 As Range
'se introducen los parámetros en cuadros de dialogo
ene = InputBox("Ingrese el número de simulaciones", "N", 1000)
a = InputBox("ingrese el inicio del intervalo: a", "Parámetro a", 0)
b = InputBox("ingrese el final del intervalo: b", "Parámetro b", 1)
'genera una columna con los números índice de la muestra
If Selection.Column > 1 Then
ActiveCell.Offset(0, -1).Value = "Indice"
ActiveCell.Offset(1, -1).Value = 1
ActiveCell.Offset(2, -1).Value = 2
Set r1 = Range(Selection.Offset(1, -1), Selection.Offset(2, -1))
Set R = Range(Selection.Offset(1, -1), Selection.Offset(ene, -1))
r1.AutoFill Destination:=R, Type:=xlFillDefault
End If
'título de la columna de valores simulados
Selection.Value = "Uniforme(" & a & ";" & b & ")"
'introduce la fórmula de simulación para el primer valor simulado
ActiveCell.Offset(1, 0).FormulaR1C1 = "=rand()*(" & (b - a) & ") + " & a & ""
'luego copia la fórmula para generar todos los valores simulados deseados
Set R = Range(Selection.Offset(1, 0), Selection.Offset(ene, 0))
ActiveCell.Offset(1, 0).AutoFill Destination:=R, Type:=xlFillDefault
'se calcula la media, varianza, mínimo y máximo de la muestra
ActiveCell.Offset(0, 1).Value = "Media Est."
ActiveCell.Offset(0, 2).FormulaR1C1 = "=AVERAGE(R[1]C[-2]:R[" & ene & "]C[-2])"
ActiveCell.Offset(1, 1).Value = "Var. Est."
ActiveCell.Offset(1, 2).FormulaR1C1 = "=VAR(R[0]C[-2]:R[" & (ene - 1) & "]C[-2])"
ActiveCell.Offset(0, 3).Value = "Máximo"
ActiveCell.Offset(0, 4).FormulaR1C1 = "=max(R[1]C[-4]:R[" & ene & "]C[-4])"
ActiveCell.Offset(1, 3).Value = "Mínimo"
ActiveCell.Offset(1, 4).FormulaR1C1 = "=min(R[0]C[-4]:R[" & (ene - 1) & "]C[-4])"
' ventana interactiva que da la opcion de agregar un histograma
resp = MsgBox("¿Desea agregar un histograma?", vbQuestion + vbYesNo, "Histograma")
If resp = 6 Then
'este paso fija la muestra a valores, es necesario si la simulación
'es una trasformación de la función ALEATORIO() pues pueden ocurrir errores
```

```
Set R = Range(Selection.Offset(1, 0), Selection.Offset(ene, 0))
R.Select
   Selection.Copy
   Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
      :=False, Transpose:=False
  Application.CutCopyMode = False
'Ventana interactiva que permite indicar el número de intervalos del histograma
ent = InputBox("Indique el número de clases(Intervalos)", "Clases", 20)
ActiveCell.End(xlToRight).Select ' posicionamiento para empezar a calcular las frecuencias
ActiveCell.Offset(-1, 0).Select
If ent > 2 Then ' aquí se evita el error después de haber indicados 2 o menos intervalos para el
histograma
ActiveCell.Offset(2, -1).Value = "Clases" 'encabezado de las clases
'generación de las clases a través de la formula recursiva (máximo -mínimo)/(total de
intervalos) + celda anterior
ActiveCell.Offset(2, 0).FormulaR1C1 = "=(R" & Selection.Row & "C" & Selection.Column & "-R"
& (Selection.Row + 1) & "C" & Selection.Column & ")/" & ent & "+R[-1]C"
'autorella el rango de intervalos con la formula anterior
Set Z1 = Range(Selection.Offset(2, 0), Selection.Offset((1 + ent), 0))
Selection.Offset(2, 0).AutoFill Destination:=Z1, Type:=xlFillDefault
' se selecciona el rango para calcular frecuencias
Set Z = Range(Selection.Offset(2, 1), Selection.Offset((1 + ent), 1))
Z.Select
' se introduce la formula de FRECUENCIA con los rangos anteriores sobre toda la muestra
Selection.FormulaArray = "=FREQUENCY(R[-1]C[-5]:R[" & (ene - 2) & "]C[-5],RC[-1]:R[" & (ent -
1) & "]C[-1])"
ActiveCell.Select
ActiveCell.Offset(-1, 0).Value = "Frecuencias" ' encabezado
Selection.Columns(1).EntireColumn.AutoFit

'se seleccionan los rangos para realizar las gráfica
Set Z = Range(Selection.Offset(-1, 0), Selection.Offset((ent - 1), 0)) ' frecuencias (eje y)
Set R = Range(Selection.Offset(0, -1), Selection.Offset((ent - 1), -1)) ' rangos (límites)
 ' Aquí se genera el gráfico de barras
  ActiveSheet.Shapes.AddChart.Select
   ActiveChart.SetSourceData Source:=Z ' se indica aqui el eje y para las barras
   ActiveChart.ChartType = xlColumnClustered ' tipo de gráficos: de barras
   ActiveChart.SeriesCollection(1).XValues = R ' rango de valores del eje X
'se selecciona de nuevo la grafica para agregar características para una mejor presentación
      ActiveChart.SeriesCollection(1).Select
      ActiveChart.ChartGroups(1).GapWidth = 5 ' espacio entre barras
   ActiveChart.SetElement (msoElementChartTitleAboveChart) ' poner nuevo titulo
   ActiveChart.ChartTitle.Text = "Histograma" ' valor del nuevo titulo

Else 'caso intervalos menores que 2
```

```
MsgBox ("Error: deben ser más de dos intervalos") ' mensaje de error
Exit Sub
End If
Else
Exit Sub
End If
'fin del programa
End Sub
```

Distribución Exponencial

```
Sub ExponencialNeg()
'Este código genera valores de una variable continua con distribución exponencial negativa con parámetros lambda
'Además incluye el cálculo de la media, varianza, máximo, mínimo y genera el histograma de forma opcional
'Definición de variables
Dim med As Double
Dim ene As Long
Dim r1 As Range
Dim R As Range
'se ingresan los parámetros en cuadros de diálogo
ene = InputBox("Indique el número total de simulaciones", "N", 1000)
med = InputBox("Indique la media de la distribución" & vbCrLf & _
"Recuerde que E(X)=1/lambda" & vbCrLf & _
"VAR(X)=1/lambda^2", "Media", 1)
'genera una columna de números índice de 1 hasta el tamaño de muestra
If Selection.Column > 1 Then
ActiveCell.Offset(0, -1).Value = "Indice"
ActiveCell.Offset(1, -1).Value = 1
ActiveCell.Offset(2, -1).Value = 2
Set r1 = Range(Selection.Offset(1, -1), Selection.Offset(2, -1))
Set R = Range(Selection.Offset(1, -1), Selection.Offset(ene, -1))
r1.AutoFill Destination:=R, Type:=xlFillDefault
End If
'genera un vector de números aleatorios para aplicar la transformada inversa
Selection.Value = "Aleatorio(0,1)"
Selection.Columns(1).EntireColumn.AutoFit
ActiveCell.Offset(1, 0).Formula = "=rand()"
Set R = Range(Selection.Offset(1, 0), Selection.Offset(ene, 0))
Selection.Offset(1, 0).AutoFill Destination:=R, Type:=xlFillDefault
'aquí aplica la fórmula de la transformada inversa y copia la formula en todo el vector
Selection.Offset(0, 1).Value = "Exponencial"
Selection.Columns(2).EntireColumn.AutoFit
Selection.Offset(1, 1).FormulaR1C1 = "=-ln(RC[-1])*" & med
```

```
Set R = Range(Selection.Offset(1, 1), Selection.Offset(ene, 1))
Selection.Offset(1, 1).AutoFill Destination:=R, Type:=xlFillDefault
Selection.Offset(0, 1).Select
'se calcula la media, varianza, mínimo y máximo de la muestra
ActiveCell.Offset(0, 1).Value = "Media Est."
ActiveCell.Offset(0, 2).FormulaR1C1 = "=AVERAGE(R[1]C[-2]:R[" & ene & "]C[-2])"
ActiveCell.Offset(1, 1).Value = "Var. Est."
ActiveCell.Offset(1, 2).FormulaR1C1 = "=VAR(R[0]C[-2]:R[" & (ene - 1) & "]C[-2])"
ActiveCell.Offset(0, 3).Value = "Máximo"
ActiveCell.Offset(0, 4).FormulaR1C1 = "=max(R[1]C[-4]:R[" & ene & "]C[-4])"
ActiveCell.Offset(1, 3).Value = "Mínimo"
ActiveCell.Offset(1, 4).FormulaR1C1 = "=min(R[0]C[-4]:R[" & (ene - 1) & "]C[-4])"
' ventana interactiva que da la opcion de agregar un histograma
resp = MsgBox("¿Desea agregar un histograma?", vbQuestion + vbYesNo, "Histograma")
If resp = 6 Then
'este paso fija la muestra a valores, es necesario si la simulación
'es una trasformacion de la funcion ALEATORIO() pues pueden ocurrir errores
Set R = Range(Selection.Offset(1, 0), Selection.Offset(ene, 0))
R.Select
   Selection.Copy
   Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
      :=False, Transpose:=False
  Application.CutCopyMode = False
' ventana interactiva que permite indicar el número de intervalos del histograma
ent = InputBox("Indique el número de clases(Intervalos)", "Clases", 20)
ActiveCell.End(xlToRight).Select ' posicionamiento para empezar a calcular las frecuencias
ActiveCell.Offset(-1, 0).Select
If ent > 2 Then ' aqui se evita el error despues de haber indicados 2 o menos intervalos para el
histograma
ActiveCell.Offset(2, -1).Value = "Clases" 'encabezado de las clases
'generación de las clases a través de la formula recursiva (máximo -mínimo)/(total de
intervalos) + celda anterior
ActiveCell.Offset(2, 0).FormulaR1C1 = "=(R" & Selection.Row & "C" & Selection.Column & "-R"
& (Selection.Row + 1) & "C" & Selection.Column & ")/" & ent & "+R[-1]C"
'autorella el rango de intervalos con la formula anterior
Set Z1 = Range(Selection.Offset(2, 0), Selection.Offset((1 + ent), 0))
Selection.Offset(2, 0).AutoFill Destination:=Z1, Type:=xlFillDefault
' se selecciona el rango para calcular frecuencias
Set Z = Range(Selection.Offset(2, 1), Selection.Offset((1 + ent), 1))
Z.Select
' se introduce la formula de FRECUENCIA con los rangos anteriores sobre toda la muestra
Selection.FormulaArray = "=FREQUENCY(R[-1]C[-5]:R[" & (ene - 2) & "]C[-5],RC[-1]:R[" & (ent -
1) & "]C[-1])"
ActiveCell.Select
ActiveCell.Offset(-1, 0).Value = "Frecuencias" ' encabezado
```

```
Selection.Columns(1).EntireColumn.AutoFit

'se seleccionan los rangos para realizar las grafica
Set Z = Range(Selection.Offset(-1, 0), Selection.Offset((ent - 1), 0)) ' frecuencias (eje y)
Set R = Range(Selection.Offset(0, -1), Selection.Offset((ent - 1), -1)) ' rangos (límites)
 ' Aquí se genera el gráfico de barras
  ActiveSheet.Shapes.AddChart.Select
   ActiveChart.SetSourceData Source:=Z ' se indica aqui el eje y para las barras
   ActiveChart.ChartType = xlColumnClustered ' tipo de graficos: de barras
   ActiveChart.SeriesCollection(1).XValues = R ' rango de valores del eje X
'se selecciona de nuevo la grafica para agregar caracteristicas para una mejor presentación
      ActiveChart.SeriesCollection(1).Select
      ActiveChart.ChartGroups(1).GapWidth = 5 ' espacio entre barras
   ActiveChart.SetElement (msoElementChartTitleAboveChart) ' poner nuevo titulo
   ActiveChart.ChartTitle.Text = "Histograma" ' valor del nuevo titulo
Else 'caso intervalos menores que 2
MsgBox ("Error: deben ser más de dos intervalos") ' mensaje de error
Exit Sub
End If
Else
Exit Sub
End If
'fin del programa
End Sub
```

Distribución Gamma

En el caso de la distribución Gamma(α, β) se requirió codificar una función auxiliar, para generar el algoritmo de simulación del Capítulo I, ya que la generación de valores depende de los casos del valor del parámetro α, además que el método fue reutilizado en el código para la generación de valores con distribución Beta.

```
'En esta función está depositado el algoritmo de generación
'De valores de una variable Gamma(alfa, beta)
Function Gam(al As Double, la As Double)

'Definición de variables
Dim b As Double
Dim P As Double
```

```
Dim ran As Double
Dim Y As Double
Dim flag As Integer
Dim ranex As Double
Dim a As Double
Dim b1 As Double
Dim q As Double
Dim t As Double
Dim d As Double
Dim v As Double
Dim Y1 As Double
Dim Z As Double
Dim W As Double
Dim ran1 As Double
Dim ran2 As Double
Dim flag1 As Integer

If al = 1 Then 'en el caso alfa=1 se tiene una exponencial se puede generar por inversión
Gam = -Log(rnd()) * la
Else ' en otro caso se procede con un algoritmo más elaborado
' De acuerdo al Capítulo I, estos valores se generan según varios casos una Gamma(alfa,1) y
'después se devuelve el valor multiplicado por lambda, resultando en una Gamma(alfa,
'lambda)
If al < 1 Then ' primer caso del parámetro alfa
flag = 0 ' variable auxiliar de control(bandera)
While flag = 0

ran = rnd()
b = (Exp(1) + al) / Exp(1)
P = b * ran
If P > 1 Then    ' luego se divide a su vez en dos casos de acuerdo al valor de la cantidad P
Y = -Log(((b - P) / al))
ranex = rnd()
If ranex <= (Y ^ (al - 1)) Then
Gam = Y * la   ' Aquí devuelve el valor de la función
flag = 1
End If  'fin del primer caso de P

End If ' caso P>1
If P <= 1 Then
Y = P ^ (1 / al)
ranex = rnd()
If ranex <= Exp(-Y) Then
Gam = Y * la ' Aquí devuelve el valor de la función
flag = 1
```

```
End If
End If   ' caso P<1
Wend

Else 'ahora se trata el caso alfa mayor a 1
a = (1 / (2 * al - 1)) ^ 0.5
b1 = al - Log(4)
q = al + (2 * al - 1) ^ 0.5
t = 4.5
d = 1 + Log(t)
flag1 = 0 ' Bandera
While flag1 = 0
ran1 = rnd()
ran2 = rnd()
v = a * Log((ran1 / (1 - ran1)))
Y1 = al * Exp(v)
Z = (ran1 ^ 2) * ran2
W = b1 + q * v - Y1
If (W + d - t * Z) >= 0 Then
Gam = Y1 * la  ' Aquí devuelve el valor de la función
flag1 = 1
End If
If W >= Log(Z) Then
Gam = Y1 * la ' Aquí devuelve el valor de la función
flag1 = 1
End If
Wend

End If  ' fin de los casos de alfa menor a 1 y mayor a 1

End If
End Function

 Ahora se

Sub Gamma()
'El siguiente Código genera valores de una distribución Gamma de parámetros alfa y lambda
'Los cuales se piden al usuario junto con el número total de simulaciones;
'Cuenta también con la opción de agregar un histograma de frecuencias.
'Nota: Depende de la función Gam()
'Definición de variables a utilizar
Dim ene As Long ' total de simulaciones
Dim alf As Double ' parámetro de forma
Dim lam As Double ' parámetro de escala
Dim R As Range ' rango para realizar diversas operaciones
```

```
Dim r1 As Range ' rango auxiliar para llevar a cabo diversas operaciones
Dim Z As Range ' rango auxiliar para realizar el histograma
Dim Z1 As Range ' rango auxiliar para realizar el histograma
Dim resp As Integer ' entero que indica un valor para una ventana de mensaje
Dim ent As Long ' numero de intervalos del histograma (pedido al usuario)
' se piden valores de los parámetros al usuario a través de ventanas interactivas
ene = InputBox("Indique el número total de simulaciones", "N", 1000)
alf = InputBox("Indique el parametro alfa >0" & vbCrLf & _
"da la forma: entero,decimales" & vbCrLf & _
"Recuerde que E(X)=alfa*lambda" & vbCrLf & _
"VAR(X)=alfa* lambda^2", "Alfa", 1)
lam = InputBox("Indique el parámetro lambda >0" & vbCrLf & _
"da la forma: entero, decimales" & vbCrLf & _
"Recuerde que E(X)=alfa*lambda" & vbCrLf & _
"VAR(X)=alfa* lambda^2", "Lambda", 1)
' se genera una lista de índices para enumerar los valores generados
If Selection.Column > 1 Then ' este paso evita un error si se inicia la macro en el primer renglón
ActiveCell.Offset(0, -1).Value = "Indice"
ActiveCell.Offset(1, -1).Value = 1
ActiveCell.Offset(2, -1).Value = 2
Set r1 = Range(Selection.Offset(1, -1), Selection.Offset(2, -1))
Set R = Range(Selection.Offset(1, -1), Selection.Offset(ene, -1))
r1.AutoFill Destination:=R, Type:=xlFillDefault
End If ' si se inicia en el primer renglón el índice no se genera
Selection.Value = "Gamma(" & alf & ";" & lam & ")"  'encabezado del vector de la muestra
Selection.Columns(1).EntireColumn.AutoFit
' se generan variables de distribución Gamma el algoritmo recae en la función Gam()
For i = 1 To ene
ActiveCell.Offset(i, 0).Value = Gam(alf, lam)
ActiveCell.Select
Next i
'se calcula la media, varianza, mínimo y máximo de la muestra
ActiveCell.Offset(0, 1).Value = "Media Est."
ActiveCell.Offset(0, 2).FormulaR1C1 = "=AVERAGE(R[1]C[-2]:R[" & ene & "]C[-2])"
ActiveCell.Offset(1, 1).Value = "Var. Est."
ActiveCell.Offset(1, 2).FormulaR1C1 = "=VAR(R[0]C[-2]:R[" & (ene - 1) & "]C[-2])"
ActiveCell.Offset(0, 3).Value = "Máximo"
ActiveCell.Offset(0, 4).FormulaR1C1 = "=max(R[1]C[-4]:R[" & ene & "]C[-4])"
ActiveCell.Offset(1, 3).Value = "Mínimo"
ActiveCell.Offset(1, 4).FormulaR1C1 = "=min(R[0]C[-4]:R[" & (ene - 1) & "]C[-4])"
'Ventana interactiva que da la opción de agregar un histograma
resp = MsgBox("¿Desea agregar un histograma?", vbQuestion + vbYesNo, "Histograma")
If resp = 6 Then
'este paso fija la muestra a valores, es necesario si la simulación
'es una trasformación de la función ALEATORIO() pues pueden ocurrir errores
```

```
Set R = Range(Selection.Offset(1, 0), Selection.Offset(ene, 0))
R.Select
   Selection.Copy
   Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
      :=False, Transpose:=False
  Application.CutCopyMode = False
'Ventana interactiva que permite indicar el número de intervalos del histograma
ent = InputBox("Indique el número de clases(Intervalos)", "Clases", 20)
ActiveCell.End(xlToRight).Select ' posicionamiento para empezar a calcular las frecuencias
ActiveCell.Offset(-1, 0).Select
If ent > 2 Then ' aquí se evita el error después de haber indicados 2 o menos intervalos para el
histograma
ActiveCell.Offset(2, -1).Value = "Clases" 'encabezado de las clases
'generación de las clases a través de la formula recursiva (máximo -mínimo)/(total de
intervalos) + celda anterior
ActiveCell.Offset(2, 0).FormulaR1C1 = "=(R" & Selection.Row & "C" & Selection.Column & "-R"
& (Selection.Row + 1) & "C" & Selection.Column & ")/" & ent & "+R[-1]C"
'autorella el rango de intervalos con la formula anterior
Set Z1 = Range(Selection.Offset(2, 0), Selection.Offset((1 + ent), 0))
Selection.Offset(2, 0).AutoFill Destination:=Z1, Type:=xlFillDefault
' se selecciona el rango para calcular frecuencias
Set Z = Range(Selection.Offset(2, 1), Selection.Offset((1 + ent), 1))
Z.Select
'Se introduce la fórmula de FRECUENCIA con los rangos anteriores sobre toda la muestra
Selection.FormulaArray = "=FREQUENCY(R[-1]C[-5]:R[" & (ene - 2) & "]C[-5],RC[-1]:R[" & (ent -
1) & "]C[-1])"
ActiveCell.Select
ActiveCell.Offset(-1, 0).Value = "Frecuencias" ' encabezado
Selection.Columns(1).EntireColumn.AutoFit

'se seleccionan los rangos para realizar las gráfica
Set Z = Range(Selection.Offset(-1, 0), Selection.Offset((ent - 1), 0)) ' frecuencias (eje y)
Set R = Range(Selection.Offset(0, -1), Selection.Offset((ent - 1), -1)) ' rangos (límites)
 'Aquí se genera el gráfico de barras
  ActiveSheet.Shapes.AddChart.Select
   ActiveChart.SetSourceData Source:=Z ' se indica aquí el eje y para las barras
   ActiveChart.ChartType = xlColumnClustered ' tipo de gráficos: de barras
   ActiveChart.SeriesCollection(1).XValues = R ' rango de valores del eje X
'se selecciona de nuevo la grafica para agregar características para una mejor presentación
      ActiveChart.SeriesCollection(1).Select
      ActiveChart.ChartGroups(1).GapWidth = 5 ' espacio entre barras
   ActiveChart.SetElement (msoElementChartTitleAboveChart) ' poner nuevo titulo
   ActiveChart.ChartTitle.Text = "Histograma" ' valor del nuevo título

Else 'caso intervalos menores que 2
```

```
MsgBox ("Error: deben ser más de dos intervalos") ' mensaje de error
Exit Sub
End If
Else
Exit Sub
End If
'fin del programa
End Sub
```

Distribución Beta

Esta distribución se muestra en un orden distinto al que aparecen las distribuciones en el Capítulo I, puesto que el siguiente algoritmo se basa en la propiedad de poder generar una variable beta a partir del cociente de 2 variables Gamma, y por lo tanto depende de la función anterior para la simulación de las variables distribuidas Gamma, como auxiliares para la generación de valores distribuidos como una variable Beta.

```
Sub Beta()
'El siguiente Código genera valores de una distribución Gamma de parámetros alfa y lambda
'Los cuales se piden al usuario junto con el número total de simulaciones;
'Cuenta también con la opción de agregar un histograma de frecuencias.
'Nota1: Depende de la función Gam()
'Nota2: Se pueden especificar 4 parámetros, referentes al intervalo de soporte y la forma de la distribución
'Definición de variables a utilizar
Dim ene As Long
Dim alf1 As Double
Dim a As Double
Dim b As Double
Dim alf2 As Double
Dim R As Range
Dim gam1 As Double
Dim gam2 As Double
' se piden valores de los parámetros al usuario a través de ventanas interactivas
ene = InputBox("Indique el número total de simulaciones", "N", 1000)
a = InputBox("Límite inferior del intervalo (soporte)", "Límite Inferior", 0)
b = InputBox("Límite superior del intervalo (soporte)", "Límite Superior", 1)
alf1 = InputBox("Indique el parámetro Alfa1 >0" & vbCrLf & _
"da la forma: entero, decimales" & vbCrLf & _
"Recuerde que E(X)=alfa1/(alfa1+alfa2)" & vbCrLf & _
"VAR(X)=alfa1*alfa2/[(alfa1 +alfa2)^2(alfa1 +alfa2 +1)]", "Alfa1", 1)
alf2 = InputBox("Indique el parámetro Alfa2 >0" & vbCrLf & _
"da la forma: entero, decimales" & vbCrLf & _
"Recuerde que E(X)=alfa1/(alfa1+alfa2)" & vbCrLf & _
"VAR(X)=alfa1*alfa2/[(alfa1 +alfa2)^2(alfa1 +alfa2 +1)]", "Alfa2", 1)
```

```
'se genera una lista  índice para enumerar los valores generados
If Selection.Column > 1 Then ' este paso evita un error si se inicia la macro en el primer renglón
ActiveCell.Offset(0, -1).Value = "Indice"
ActiveCell.Offset(1, -1).Value = 1
ActiveCell.Offset(2, -1).Value = 2
Set r1 = Range(Selection.Offset(1, -1), Selection.Offset(2, -1))
Set R = Range(Selection.Offset(1, -1), Selection.Offset(ene, -1))
r1.AutoFill Destination:=R, Type:=xlFillDefault
End If ' si se inicia en el primer renglón el índice no se genera
Selection.Value = "Beta(" & alf1 & ";" & alf2 & ")"'encabezado del vector de la muestra
Selection.Columns(1).EntireColumn.AutoFit
' se genera los valores de la distribución Beta como el cociente de dos variables Gamma
For i = 1 To ene
gam1 = Gam(alf1, 1)
gam2 = Gam(alf2, 1)
ActiveCell.Offset(i, 0).Value = a + (b - a) * (gam1 / (gam1 + gam2))
Next i
'se calcula la media, varianza, mínimo y máximo de la muestra
ActiveCell.Offset(0, 1).Value = "Media Est."
ActiveCell.Offset(0, 2).FormulaR1C1 = "=AVERAGE(R[1]C[-2]:R[" & ene & "]C[-2])"
ActiveCell.Offset(1, 1).Value = "Var. Est."
ActiveCell.Offset(1, 2).FormulaR1C1 = "=VAR(R[0]C[-2]:R[" & (ene - 1) & "]C[-2])"
ActiveCell.Offset(0, 3).Value = "Máximo"
ActiveCell.Offset(0, 4).FormulaR1C1 = "=max(R[1]C[-4]:R[" & ene & "]C[-4])"
ActiveCell.Offset(1, 3).Value = "Mínimo"
ActiveCell.Offset(1, 4).FormulaR1C1 = "=min(R[0]C[-4]:R[" & (ene - 1) & "]C[-4])"
'Ventana interactiva que da la opción de agregar un histograma
resp = MsgBox("¿Desea agregar un histograma?", vbQuestion + vbYesNo, "Histograma")
If resp = 6 Then
'este paso fija la muestra a valores, es necesario si la simulación
'es una trasformación de la función ALEATORIO() pues pueden ocurrir errores
Set R = Range(Selection.Offset(1, 0), Selection.Offset(ene, 0))
R.Select
   Selection.Copy
   Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
      :=False, Transpose:=False
 Application.CutCopyMode = False
'Ventana interactiva que permite indicar el número de intervalos del histograma
ent = InputBox("Indique el número de clases(Intervalos)", "Clases", 20)
ActiveCell.End(xlToRight).Select ' posicionamiento para empezar a calcular las frecuencias
ActiveCell.Offset(-1, 0).Select
If ent > 2 Then ' aquí se evita el error después de haber indicados 2 o menos intervalos para el
histograma
ActiveCell.Offset(2, -1).Value = "Clases" 'encabezado de las clases
```

```
'generación de las clases a través de la formula recursiva (máximo -mínimo)/(total de
intervalos) + celda anterior
ActiveCell.Offset(2, 0).FormulaR1C1 = "=(R" & Selection.Row & "C" & Selection.Column & "-R"
& (Selection.Row + 1) & "C" & Selection.Column & ")/" & ent & "+R[-1]C"
'autorella el rango de intervalos con la formula anterior
Set Z1 = Range(Selection.Offset(2, 0), Selection.Offset((1 + ent), 0))
Selection.Offset(2, 0).AutoFill Destination:=Z1, Type:=xlFillDefault
' se selecciona el rango para calcular frecuencias
Set Z = Range(Selection.Offset(2, 1), Selection.Offset((1 + ent), 1))
Z.Select
'Se introduce la formula de FRECUENCIA con los rangos anteriores sobre toda la muestra
Selection.FormulaArray = "=FREQUENCY(R[-1]C[-5]:R[" & (ene - 2) & "]C[-5],RC[-1]:R[" & (ent -
1) & "]C[-1])"
ActiveCell.Select
ActiveCell.Offset(-1, 0).Value = "Frecuencias" ' encabezado
Selection.Columns(1).EntireColumn.AutoFit

'Se seleccionan los rangos para realizar las gráfica
Set Z = Range(Selection.Offset(-1, 0), Selection.Offset((ent - 1), 0)) ' frecuencias (eje y)
Set R = Range(Selection.Offset(0, -1), Selection.Offset((ent - 1), -1)) ' rangos (límites)
 'Aquí se genera el gráfico de barras
  ActiveSheet.Shapes.AddChart.Select
   ActiveChart.SetSourceData Source:=Z ' se indica aquí el eje y para las barras
   ActiveChart.ChartType = xlColumnClustered ' tipo de gráficos: de barras
   ActiveChart.SeriesCollection(1).XValues = R ' rango de valores del eje X
'se selecciona de nuevo la grafica para agregar características para una mejor presentación
      ActiveChart.SeriesCollection(1).Select
      ActiveChart.ChartGroups(1).GapWidth = 5 ' espacio entre barras
   ActiveChart.SetElement (msoElementChartTitleAboveChart) ' poner nuevo titulo
   ActiveChart.ChartTitle.Text = "Histograma" ' valor del nuevo titulo
Else 'caso intervalos menores que 2
MsgBox ("Error: deben ser más de dos intervalos") ' mensaje de error
Exit Sub
End If
Else
Exit Sub
End If
'fin del programa
End Sub
```

Distribución Weibull

```
Sub Weibull()
'El siguiente Código genera valores de una distribución Weibull de parámetros alfa y beta
'Los cuales se piden al usuario junto con el número total de simulaciones;
```

```
' cuenta también con la opción de agregar un histograma de frecuencias.
'Definición de variables a utilizar
Dim ene As Long
Dim bet As Double
Dim alf As Double
Dim R As Range
Dim rand As Double
' se piden valores de los parámetros al usuario a través de ventanas interactivas
ene = InputBox("Indique el número total de simulaciones", "N", 1000)
alf = InputBox("Indique el parámetro Alfa >0" & vbCrLf & _
"da la forma: entero,decimales" & vbCrLf & _
"Recuerde que E(X)=Beta*Gamma(1+1/Alfa)" & vbCrLf & _
"VAR(X)=" & vbCrLf & "Beta^2[Gamma(1+2/Alfa)-Gamma^2(1+1/Alfa)]", "Alfa", 1)

bet = InputBox("Indique el parámetro Beta >0" & vbCrLf & _
"da la forma: entero,decimales" & vbCrLf & _
"Recuerde que E(X)=Beta*Gamma(1+1/Alfa)" & vbCrLf & _
"VAR(X)=" & vbCrLf & "Beta^2[Gamma(1+2/Alfa)-Gamma^2(1+1/Alfa)]", "Beta", 1)


'Se genera una lista de índices para enumerar los valores generados
If Selection.Column > 1 Then ' este paso evita un error si se inicia la macro en el primer renglón
ActiveCell.Offset(0, -1).Value = "Indice"
ActiveCell.Offset(1, -1).Value = 1
ActiveCell.Offset(2, -1).Value = 2
Set r1 = Range(Selection.Offset(1, -1), Selection.Offset(2, -1))
Set R = Range(Selection.Offset(1, -1), Selection.Offset(ene, -1))
r1.AutoFill Destination:=R, Type:=xlFillDefault
End If ' si se inicia en el primer renglón el índice no se genera
Selection.Value = "Weibull(" & alf & ";" & bet & ")" 'encabezado del vector de la muestra
Selection.Columns(1).EntireColumn.AutoFit
'Se generan los valores de la distribución con el método de la transformada inversa
For i = 1 To ene
Randomize
rand = rnd()
ActiveCell.Offset(i, 0).Value = bet * ((-Log(rand)) ^ (1 / alf))
Next i
'se calcula la media, varianza, mínimo y máximo de la muestra
ActiveCell.Offset(0, 1).Value = "Media Est."
ActiveCell.Offset(0, 2).FormulaR1C1 = "=AVERAGE(R[1]C[-2]:R[" & ene & "]C[-2])"
ActiveCell.Offset(1, 1).Value = "Var. Est."
ActiveCell.Offset(1, 2).FormulaR1C1 = "=VAR(R[0]C[-2]:R[" & (ene - 1) & "]C[-2])"
ActiveCell.Offset(0, 3).Value = "Máximo"
ActiveCell.Offset(0, 4).FormulaR1C1 = "=max(R[1]C[-4]:R[" & ene & "]C[-4])"
ActiveCell.Offset(1, 3).Value = "Mínimo"
```

```
ActiveCell.Offset(1, 4).FormulaR1C1 = "=min(R[0]C[-4]:R[" & (ene - 1) & "]C[-4])"
'Ventana interactiva que da la opción de agregar un histograma
resp = MsgBox("¿Desea agregar un histograma?", vbQuestion + vbYesNo, "Histograma")
If resp = 6 Then
'Este paso fija la muestra a valores, es necesario si la simulación
'Es una trasformación de la función ALEATORIO() pues pueden ocurrir errores
Set R = Range(Selection.Offset(1, 0), Selection.Offset(ene, 0))
R.Select
   Selection.Copy
   Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
      :=False, Transpose:=False
  Application.CutCopyMode = False
'Ventana interactiva que permite indicar el número de intervalos del histograma
ent = InputBox("Indique el número de clases(Intervalos)", "Clases", 20)
ActiveCell.End(xlToRight).Select ' posicionamiento para empezar a calcular las frecuencias
ActiveCell.Offset(-1, 0).Select
If ent > 2 Then ' aquí se evita el error después de haber indicados 2 o menos intervalos para el
histograma
ActiveCell.Offset(2, -1).Value = "Clases" 'encabezado de las clases
'generación de las clases a través de la formula recursiva (máximo -mínimo)/(total de
intervalos) + celda anterior
ActiveCell.Offset(2, 0).FormulaR1C1 = "=(R" & Selection.Row & "C" & Selection.Column & "-R"
& (Selection.Row + 1) & "C" & Selection.Column & ")/" & ent & "+R[-1]C"
'autorella el rango de intervalos con la formula anterior
Set Z1 = Range(Selection.Offset(2, 0), Selection.Offset((1 + ent), 0))
Selection.Offset(2, 0).AutoFill Destination:=Z1, Type:=xlFillDefault
' se selecciona el rango para calcular frecuencias
Set Z = Range(Selection.Offset(2, 1), Selection.Offset((1 + ent), 1))
Z.Select
'Se introduce la formula de FRECUENCIA con los rangos anteriores sobre toda la muestra
Selection.FormulaArray = "=FREQUENCY(R[-1]C[-5]:R[" & (ene - 2) & "]C[-5],RC[-1]:R[" & (ent -
1) & "]C[-1])"
ActiveCell.Select
ActiveCell.Offset(-1, 0).Value = "Frecuencias" ' encabezado
Selection.Columns(1).EntireColumn.AutoFit

'se seleccionan los rangos para realizar las gráfica
Set Z = Range(Selection.Offset(-1, 0), Selection.Offset((ent - 1), 0)) ' frecuencias (eje y)
Set R = Range(Selection.Offset(0, -1), Selection.Offset((ent - 1), -1)) ' rangos (límites)
 ' Aquí se genera el gráfico de barras
  ActiveSheet.Shapes.AddChart.Select
   ActiveChart.SetSourceData Source:=Z ' se indica aquí el eje Y para las barras
   ActiveChart.ChartType = xlColumnClustered ' tipo de gráficos: de barras
   ActiveChart.SeriesCollection(1).XValues = R ' rango de valores del eje X
'se selecciona de nuevo la gráfica para agregar características para una mejor presentación
```

```
    ActiveChart.SeriesCollection(1).Select
    ActiveChart.ChartGroups(1).GapWidth = 5 ' espacio entre barras
  ActiveChart.SetElement (msoElementChartTitleAboveChart) ' pone nuevo titulo
  ActiveChart.ChartTitle.Text = "Histograma" ' valor del nuevo titulo
Else 'caso intervalos menores que 2
MsgBox ("Error: deben ser más de dos intervalos") ' mensaje de error
Exit Sub
End If
Else
Exit Sub
End If
'fin del programa
End Sub
```

Distribución Normal

```
Sub Normal_DoceU()
'El siguiente Código genera valores de una distribución Normal de parámetros mu y sigma
'Los cuales se piden al usuario junto con el número total de simulaciones;
'Cuenta también con la opción de agregar un histograma de frecuencias.
'Definición de variables a utilizar
Dim med As Double
Dim vari As Double
Dim ene As Long
Dim R As Range
Dim r1 As Range
'Se piden valores de los parámetros al usuario a través de ventanas interactivas
ene = InputBox("Ingrese el número de simulaciones", "N", 1000)
med = InputBox("ingrese la media de la Normal", "Media", 0)
vari = InputBox("ingrese la varianza de la Normal", "Varianza", 1)
'primero genera 12 valores aleatorios uniformes en (0,1)
For i = 1 To 12
ActiveCell.Offset(0, (i - 1)).Value = i
Next i
ActiveCell.Offset(0, 12).Value = "Normal"
For i = 1 To 12
ActiveCell.Offset(1, (i - 1)).Formula = "=rand()"
Next i
' luego genera los números índice para los valores de la muestra
Set r1 = Range(Selection.Offset(1, 0), Selection.Offset(1, 11))
Set R = Range(Selection.Offset(1, 0), Selection.Offset(ene, 11))
r1.AutoFill Destination:=R, Type:=xlFillDefault
If Selection.Column > 1 Then
ActiveCell.Offset(0, -1).Value = "Indice"
ActiveCell.Offset(1, -1).Value = 1
```

```
ActiveCell.Offset(2, -1).Value = 2
Set r1 = Range(Selection.Offset(1, -1), Selection.Offset(2, -1))
Set R = Range(Selection.Offset(1, -1), Selection.Offset(ene, -1))
r1.AutoFill Destination:=R, Type:=xlFillDefault
End If
'Después coloca la fórmula para hacer la suma de las 12 uniformes menos 6 que es su media
'y aprovecha la linealidad de la normal para generar la distribución con los parámetros
'deseados
ActiveCell.End(xlToRight).Select
ActiveCell.Offset(1, 0).FormulaR1C1 = "=(sum(RC[-12]:RC[-1])-6)*SQRT(" & vari & ") +" & med
'se calcula la media, varianza, mínimo y máximo de la muestra
ActiveCell.Offset(0, 1).Value = "Media Est."
ActiveCell.Offset(0, 2).FormulaR1C1 = "=AVERAGE(R[1]C[-2]:R[" & ene & "]C[-2])"
ActiveCell.Offset(1, 1).Value = "Var. Est."
ActiveCell.Offset(1, 2).FormulaR1C1 = "=VAR(R[0]C[-2]:R[" & (ene - 1) & "]C[-2])"
ActiveCell.Offset(0, 3).Value = "Máximo"
ActiveCell.Offset(0, 4).FormulaR1C1 = "=max(R[1]C[-4]:R[" & ene & "]C[-4])"
ActiveCell.Offset(1, 3).Value = "Mínimo"
ActiveCell.Offset(1, 4).FormulaR1C1 = "=min(R[0]C[-4]:R[" & (ene - 1) & "]C[-4])"
' ventana interactiva que da la opción de agregar un histograma
resp = MsgBox("¿Desea agregar un histograma?", vbQuestion + vbYesNo, "Histograma")
If resp = 6 Then
'este paso fija la muestra a valores, es necesario si la simulación
'es una trasformación de la función ALEATORIO() pues pueden ocurrir errores
Set R = Range(Selection.Offset(1, 0), Selection.Offset(ene, 0))
R.Select
   Selection.Copy
   Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
      :=False, Transpose:=False
 Application.CutCopyMode = False
'Ventana interactiva que permite indicar el número de intervalos del histograma
ent = InputBox("Indique el número de clases(Intervalos)", "Clases", 20)
ActiveCell.End(xlToRight).Select ' posicionamiento para empezar a calcular las frecuencias
ActiveCell.Offset(-1, 0).Select
If ent > 2 Then ' aquí se evita el error después de haber indicados 2 o menos intervalos para el
histograma
ActiveCell.Offset(2, -1).Value = "Clases" 'encabezado de las clases
'generación de las clases a través de la formula recursiva (máximo -mínimo)/(total de
intervalos) + celda anterior
ActiveCell.Offset(2, 0).FormulaR1C1 = "=(R" & Selection.Row & "C" & Selection.Column & "-R"
& (Selection.Row + 1) & "C" & Selection.Column & ")/" & ent & "+R[-1]C"
'auto rellena el rango de intervalos con la formula anterior
Set Z1 = Range(Selection.Offset(2, 0), Selection.Offset((1 + ent), 0))
Selection.Offset(2, 0).AutoFill Destination:=Z1, Type:=xlFillDefault
' se selecciona el rango para calcular frecuencias
```

```
Set Z = Range(Selection.Offset(2, 1), Selection.Offset((1 + ent), 1))
Z.Select
'Se introduce la formula de FRECUENCIA con los rangos anteriores sobre toda la muestra
Selection.FormulaArray = "=FREQUENCY(R[-1]C[-5]:R[" & (ene - 2) & "]C[-5],RC[-1]:R[" & (ent -
1) & "]C[-1])"
ActiveCell.Select
ActiveCell.Offset(-1, 0).Value = "Frecuencias" ' encabezado
Selection.Columns(1).EntireColumn.AutoFit
'Se seleccionan los rangos para realizar las gráfica
Set Z = Range(Selection.Offset(-1, 0), Selection.Offset((ent - 1), 0)) ' frecuencias (eje y)
Set R = Range(Selection.Offset(0, -1), Selection.Offset((ent - 1), -1)) ' rangos (límites)
 ' Aquí se genera el gráfico de barras
  ActiveSheet.Shapes.AddChart.Select
   ActiveChart.SetSourceData Source:=Z ' se indica aquí el eje Y para las barras
   ActiveChart.ChartType = xlColumnClustered ' tipo de gráficos: de barras
   ActiveChart.SeriesCollection(1).XValues = R ' rango de valores del eje X
'Se selecciona de nuevo la grafica para agregar características para una mejor presentación
      ActiveChart.SeriesCollection(1).Select
      ActiveChart.ChartGroups(1).GapWidth = 5 ' espacio entre barras
   ActiveChart.SetElement (msoElementChartTitleAboveChart) ' poner nuevo titulo
   ActiveChart.ChartTitle.Text = "Histograma" ' valor del nuevo titulo
Else 'caso intervalos menores que 2
MsgBox ("Error: deben ser más de dos intervalos") ' mensaje de error
Exit Sub
End If
Else
Exit Sub
End If
'fin del programa
End Sub
```

Distribución Log-Normal

El siguiente código genera valores de una variable Log-normal a través de calcular la exponencial de una Normal(μ, σ^2), donde los valores de la Normal son generados a través del método de Marsaglia, el cual es una forma similar a la forma de simulación de Box-Muller, que genera un par de variables que se distribuyen uniforme en un círculo de radio 1 y por medio de una transformación en coordenadas polares obtiene dos variables independientes con distribución Normal.

Esto se debe a que el método de las 12 uniformes es muy ilustrativo sobre el comportamiento de la suma de variables aleatorias y como introducción al estudio del teorema central del límite, sin embargo el método de Marsaglia es más eficiente cuando se pretende estudiar muestras de gran tamaño.

Marsaglia se basa en el mismo proceso de Box-Muller, pero lo combina con el método de aceptación y rechazo, por lo que su método se resume en:

1. $\boldsymbol{Generar\ U_1, U_2\ independientes\ distribuidas\ U(0,1)}$
2. $\boldsymbol{Sea\ V_1 = 2U_1 - 1,\ V_2 = 2U_2 - 1 \quad y\ W = {V_1}^2 + {V_2}^2}$
3. $\boldsymbol{Si\ W < 1\ regresar\ X = \sqrt{-2Log(W)/W} * V_1, en\ otro\ caso\ ir\ al\ paso\ 1}$

Lo anterior se incluye en el siguiente código.

```
Sub Lognormal()
'El siguiente Código genera valores de una distribución Log-Normal de parámetros mu y sigma
'Los cuales se piden al usuario junto con el número total de simulaciones
'Cuenta también con la opción de agregar un histograma de frecuencias.
'NOTA: los parámetros mu y sigma son de la normal, esto se aclara en los cuadros de dialogo
'Definición de variables a utilizar
Dim ene As Long
Dim mu As Double
Dim sig As Double
Dim R As Range
Dim rand As Double
Dim X(1) As Double
'Se piden valores de los parámetros al usuario a través de ventanas interactivas
ene = InputBox("Indique el número total de simulaciones", "N", 1000)
mu = InputBox("Indique el parámetro Mu " & vbCrLf & _
"da la forma: entero,decimales" & vbCrLf & _
"Recuerde que E(X)=Exp(mu+Sigma^2/2)" & vbCrLf & _
"VAR(X)=Exp(2Mu+Sigma^2)*[Exp(Sigma^2)-1]", "Mu", 0)

sig = InputBox("Indique el parámetro Sigma^2 >0" & vbCrLf & _
"da la forma: entero,decimales" & vbCrLf & _
"Recuerde que E(X)=Exp(mu+Sigma^2/2)" & vbCrLf & _
"VAR(X)=Exp(2Mu+Sigma^2)*[Exp(Sigma^2)-1]", "Sigma^2", 1)
'Se generan los números índice para los valores de la muestra
If Selection.Column > 1 Then
ActiveCell.Offset(0, -1).Value = "Indice"
ActiveCell.Offset(1, -1).Value = 1
ActiveCell.Offset(2, -1).Value = 2
Set r1 = Range(Selection.Offset(1, -1), Selection.Offset(2, -1))
Set R = Range(Selection.Offset(1, -1), Selection.Offset(ene, -1))
r1.AutoFill Destination:=R, Type:=xlFillDefault
End If
Selection.Value = "Lognormal(" & mu & ";" & sig & ")"
Selection.Columns(1).EntireColumn.AutoFit
'Simulación
For i = 1 To ene
'primero se generan los valores de la normal por el método de Marsaglia
f = 0
```

```
While f = 0
Randomize
v1 = 2 * rnd() - 1
v2 = 2 * rnd() - 1
W = v1 ^ 2 + v2 ^ 2
If W <= 1 Then
X(1) = Sqr(((-2 * Log(W)) / W)) * v1
f = 1
End If
Wend
'Se calcula la exponencial de la Normal generada para obtener la variable deseada
ActiveCell.Offset(i, 0).Value = Exp(X(1) * Sqr(sig) + mu)
Next i
'Se calcula la media, varianza, mínimo y máximo de la muestra
ActiveCell.Offset(0, 1).Value = "Media Est."
ActiveCell.Offset(0, 2).FormulaR1C1 = "=AVERAGE(R[1]C[-2]:R[" & ene & "]C[-2])"
ActiveCell.Offset(1, 1).Value = "Var. Est."
ActiveCell.Offset(1, 2).FormulaR1C1 = "=VAR(R[0]C[-2]:R[" & (ene - 1) & "]C[-2])"
ActiveCell.Offset(0, 3).Value = "Máximo"
ActiveCell.Offset(0, 4).FormulaR1C1 = "=max(R[1]C[-4]:R[" & ene & "]C[-4])"
ActiveCell.Offset(1, 3).Value = "Mínimo"
ActiveCell.Offset(1, 4).FormulaR1C1 = "=min(R[0]C[-4]:R[" & (ene - 1) & "]C[-4])"
'Ventana interactiva que da la opción de agregar un histograma
resp = MsgBox("¿Desea agregar un histograma?", vbQuestion + vbYesNo, "Histograma")
If resp = 6 Then
'este paso fija la muestra a valores, es necesario si la simulación
'es una trasformación de la función ALEATORIO() pues pueden ocurrir errores
Set R = Range(Selection.Offset(1, 0), Selection.Offset(ene, 0))
R.Select
   Selection.Copy
   Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
      :=False, Transpose:=False
  Application.CutCopyMode = False
' ventana interactiva que permite indicar el número de intervalos del histograma
ent = InputBox("Indique el número de clases(Intervalos)", "Clases", 20)
ActiveCell.End(xlToRight).Select ' posicionamiento para empezar a calcular las frecuencias
ActiveCell.Offset(-1, 0).Select
If ent > 2 Then ' aquí se evita el error después de haber indicados 2 o menos intervalos para el
histograma
ActiveCell.Offset(2, -1).Value = "Clases" 'encabezado de las clases
'generación de las clases a través de la formula recursiva (máximo -mínimo)/(total de
intervalos) + celda anterior
ActiveCell.Offset(2, 0).FormulaR1C1 = "=(R" & Selection.Row & "C" & Selection.Column & "-R"
& (Selection.Row + 1) & "C" & Selection.Column & ")/" & ent & "+R[-1]C"
'autorella el rango de intervalos con la formula anterior
```

```
Set Z1 = Range(Selection.Offset(2, 0), Selection.Offset((1 + ent), 0))
Selection.Offset(2, 0).AutoFill Destination:=Z1, Type:=xlFillDefault
' se selecciona el rango para calcular frecuencias
Set Z = Range(Selection.Offset(2, 1), Selection.Offset((1 + ent), 1))
Z.Select
'Se introduce la formula de FRECUENCIA con los rangos anteriores sobre toda la muestra
Selection.FormulaArray = "=FREQUENCY(R[-1]C[-5]:R[" & (ene - 2) & "]C[-5],RC[-1]:R[" & (ent -
1) & "]C[-1])"
ActiveCell.Select
ActiveCell.Offset(-1, 0).Value = "Frecuencias" ' encabezado
Selection.Columns(1).EntireColumn.AutoFit
'se seleccionan los rangos para realizar las grafica
Set Z = Range(Selection.Offset(-1, 0), Selection.Offset((ent - 1), 0)) ' frecuencias (eje y)
Set R = Range(Selection.Offset(0, -1), Selection.Offset((ent - 1), -1)) ' rangos (límites)
 ' Aquí se genera el gráfico de barras
  ActiveSheet.Shapes.AddChart.Select
   ActiveChart.SetSourceData Source:=Z ' se indica aquí el eje Y para las barras
   ActiveChart.ChartType = xlColumnClustered ' tipo de gráficos: de barras
   ActiveChart.SeriesCollection(1).XValues = R ' rango de valores del eje X
'se selecciona de nuevo la grafica para agregar características para una mejor presentación
      ActiveChart.SeriesCollection(1).Select
      ActiveChart.ChartGroups(1).GapWidth = 5 ' espacio entre barras
   ActiveChart.SetElement (msoElementChartTitleAboveChart) ' poner nuevo titulo
   ActiveChart.ChartTitle.Text = "Histograma" ' valor del nuevo titulo
Else 'caso intervalos menores que 2
MsgBox ("Error: deben ser más de dos intervalos") ' mensaje de error
Exit Sub
End If
Else
Exit Sub
End If
'fin del programa
End Sub
```

Distribución Pareto

```
Sub Pareto()
'El siguiente Código genera valores de una distribución Log-Normal de parámetros mu y sigma
'Los cuales se piden al usuario junto con el número total de simulaciones
'Cuenta también con la opción de agregar un histograma de frecuencias.
'NOTA: La parametrización sea elegido para coincidir con la del libro Loss Models from data
'decisions
'Definición de variables a utilizar
Dim ene As Long
Dim thet As Double
```

```
Dim alf As Double
Dim R As Range
Dim rand As Double
' se piden valores de los parámetros al usuario a través de ventanas interactivas
ene = InputBox("Indique el número total de simulaciones", "N", 1000)
alf = InputBox("Indique el parámetro Alfa >0" & vbCrLf & _
"da la forma: entero,decimales" & vbCrLf & _
"Recuerde que E(X)=Theta/(Alfa-1)" & vbCrLf & _
"VAR(X)=" & vbCrLf & "2Theta^2/(Alfa-1)(Alfa-2)-Media^2", "Alfa", 1)

thet = InputBox("Indique el parámetro Theta >0" & vbCrLf & _
"da la forma: entero,decimales" & vbCrLf & _
"Recuerde que E(X)=Theta/(Alfa-1)" & vbCrLf & _
"VAR(X)=" & vbCrLf & "2Theta^2/(Alfa-1)(Alfa-2)-Media^2", "Theta", 1)
' se generan los números índice para los valores de la muestra
If Selection.Column > 1 Then
ActiveCell.Offset(0, -1).Value = "Indice"
ActiveCell.Offset(1, -1).Value = 1
ActiveCell.Offset(2, -1).Value = 2
Set r1 = Range(Selection.Offset(1, -1), Selection.Offset(2, -1))
Set R = Range(Selection.Offset(1, -1), Selection.Offset(ene, -1))
r1.AutoFill Destination:=R, Type:=xlFillDefault
End If
Selection.Value = "Pareto(" & alf & ";" & thet & ")"
Selection.Columns(1).EntireColumn.AutoFit
'Se generan los valores de la distribución por el método de la función de distribución inversa
For i = 1 To ene
Randomize
rand = rnd()
ActiveCell.Offset(i, 0).Value = thet * ((rand ^ (-1 / alf)) - 1)
Next i
'Se calcula la media, varianza, mínimo y máximo de la muestra
ActiveCell.Offset(0, 1).Value = "Media Est."
ActiveCell.Offset(0, 2).FormulaR1C1 = "=AVERAGE(R[1]C[-2]:R[" & ene & "]C[-2])"
ActiveCell.Offset(1, 1).Value = "Var. Est."
ActiveCell.Offset(1, 2).FormulaR1C1 = "=VAR(R[0]C[-2]:R[" & (ene - 1) & "]C[-2])"
ActiveCell.Offset(0, 3).Value = "Máximo"
ActiveCell.Offset(0, 4).FormulaR1C1 = "=max(R[1]C[-4]:R[" & ene & "]C[-4])"
ActiveCell.Offset(1, 3).Value = "Mínimo"
ActiveCell.Offset(1, 4).FormulaR1C1 = "=min(R[0]C[-4]:R[" & (ene - 1) & "]C[-4])"
'Ventana interactiva que da la opcion de agregar un histograma
resp = MsgBox("¿Desea agregar un histograma?", vbQuestion + vbYesNo, "Histograma")
If resp = 6 Then
'este paso fija la muestra a valores, es necesario si la simulación
'es una trasformación de la función ALEATORIO() pues pueden ocurrir errores
```

```
Set R = Range(Selection.Offset(1, 0), Selection.Offset(ene, 0))
R.Select
   Selection.Copy
   Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
      :=False, Transpose:=False
  Application.CutCopyMode = False
'Ventana interactiva que permite indicar el número de intervalos del histograma
ent = InputBox("Indique el número de clases(Intervalos)", "Clases", 20)
ActiveCell.End(xlToRight).Select ' posicionamiento para empezar a calcular las frecuencias
ActiveCell.Offset(-1, 0).Select
If ent > 2 Then ' aquí se evita el error después de haber indicados 2 o menos intervalos para el
histograma
ActiveCell.Offset(2, -1).Value = "Clases" 'encabezado de las clases
'generación de las clases a través de la formula recursiva (máximo -mínimo)/(total de
intervalos) + celda anterior
ActiveCell.Offset(2, 0).FormulaR1C1 = "=(R" & Selection.Row & "C" & Selection.Column & "-R"
& (Selection.Row + 1) & "C" & Selection.Column & ")/" & ent & "+R[-1]C"
'auto rellena el rango de intervalos con la fórmula anterior
Set Z1 = Range(Selection.Offset(2, 0), Selection.Offset((1 + ent), 0))
Selection.Offset(2, 0).AutoFill Destination:=Z1, Type:=xlFillDefault
'Se selecciona el rango para calcular frecuencias
Set Z = Range(Selection.Offset(2, 1), Selection.Offset((1 + ent), 1))
Z.Select
' se introduce la formula de FRECUENCIA con los rangos anteriores sobre toda la muestra
Selection.FormulaArray = "=FREQUENCY(R[-1]C[-5]:R[" & (ene - 2) & "]C[-5],RC[-1]:R[" & (ent -
1) & "]C[-1])"
ActiveCell.Select
ActiveCell.Offset(-1, 0).Value = "Frecuencias" ' encabezado
Selection.Columns(1).EntireColumn.AutoFit
'Se seleccionan los rangos para realizar las gráfica
Set Z = Range(Selection.Offset(-1, 0), Selection.Offset((ent - 1), 0)) ' frecuencias (eje y)
Set R = Range(Selection.Offset(0, -1), Selection.Offset((ent - 1), -1)) ' rangos (límites)
 ' Aquí se genera el gráfico de barras
  ActiveSheet.Shapes.AddChart.Select
   ActiveChart.SetSourceData Source:=Z ' se indica aquí el eje Y para las barras
   ActiveChart.ChartType = xlColumnClustered ' tipo de gráficos: de barras
   ActiveChart.SeriesCollection(1).XValues = R ' rango de valores del eje X
'se selecciona de nuevo la gráfica para agregar características para una mejor presentación
      ActiveChart.SeriesCollection(1).Select
      ActiveChart.ChartGroups(1).GapWidth = 5 ' espacio entre barras
   ActiveChart.SetElement (msoElementChartTitleAboveChart) ' poner nuevo titulo
   ActiveChart.ChartTitle.Text = "Histograma" ' valor del nuevo titulo
Else 'caso intervalos menores que 2
MsgBox ("Error: deben ser más de dos intervalos") ' mensaje de error
Exit Sub
```

```
End If
Else
Exit Sub
End If
'fin del programa
End Sub
```

Automatización del proceso de copiado y pegado

El siguiente código es ampliamente mencionado y utilizado a lo largo de este trabajo, pues dada una fórmula dinámica que cambie de valor a cada actualización de la hoja de cálculo, copia su valor y lo deposita otra celda, además permite introducir como parámetros el número de veces a copiar y el rango de celdas como el objetivo a copiar.
Este código tiene la ventaja de ser bastante simple, lo que permite su fácil comprensión y aplicación.

```
Sub copiaypega()
'El siguiente código permite registrar el valor de algún rango en especial y copiarlo a valores un
 'Cierto número de veces, el cual es empleado para registrar la simulación de diversos
'estadísticos
'Definición de variables
Dim myNum As Long
   Dim myCell As Range
'Introduce como parámetros el número de veces a copiar y la selección del rango sobre la hoja
'de calculo
     myNum = Application.InputBox("número de veces a copiar")
   Set myCell = Application.InputBox( _
   prompt:="seleccione el rango de celdas a copiar", Type:=8)

  myCell.Select
'Realiza el copiado a valores hacia debajo del rango de referencia
   Selection.Copy
   ActiveCell.Offset(2, 0).Select
'pega a valores el registro
   For i = 1 To myNum
   Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
      :=False, Transpose:=False
   ActiveCell.Offset(1, 0).Select
   Next i
   'fin del programa
End Sub
```

Códigos de programación en R®

Los siguientes códigos fueron desarrollados en el software libre especializado en temas estadísticos R® Basado en el lenguaje S, y comprenden el conjunto de graficas mostradas a lo largo de los 5 capítulos del presente trabajo, se publican con la intención de ser reproducidos como auxiliares en el salón de clase, como un método didáctica en la enseñanza de la estadística que incluso puede ser explicada a detalle a partir del propio código.

Gráficos del Capítulo I

Distribución Uniforme

```
#Gráficas de las funciones densidad, distribución y otras de importancia para las variables
#aleatorias continuas
#Función de densidad
#Se traza como una línea recta horizontal
plot(c(3,6),c(.5,.5),type="l",xaxt="n",yaxt="n",xlim=c(2,7),ylim=c(0,1),xlab="x",ylab="")
#se colocan títulos
title("Grafica 1.13: función de densidad\n de una uniforme(a,b)",cex.main=1)
# límites de la distribución
points(c(2,3),c(0,0),type="l",xaxt="n",yaxt="n")
points(c(6,7),c(0,0),type="l",xaxt="n",yaxt="n")
yax=seq(0,.5,by=.1)
yaxa=rep(3,length(yax))
yaxb=rep(6,length(yax))
points(yaxa,yax,type="c",xaxt="n",yaxt="n")
points(yaxb,yax,type="c",xaxt="n",yaxt="n")
#Se personalizan las caracteristicas de los ejes
axis(2,at=c(0,.5),labels=c(0,expression(frac(1,b-a))),las=2,cex=1) #eje y
axis(1,at=c(3,6),labels=c("a","b"),las=0) #eje x

# Función de distribución
#Se traza como una línea de pendiente positiva
plot(c(2,3,6,7),c(0,0,1,1),type="l",xaxt="n",yaxt="n",xlim=c(2,7),ylim=c(0,1.2),xlab="x",ylab=ex
pression(F(x)))
#se colocan titulos
title("Grafica 1.14: Función de distribución\n de una uniforme(a,b)",cex.main=1)
yax=seq(0,1,by=.1)
yaxb=rep(6,length(yax))
points(yaxb,yax,type="c",xaxt="n",yaxt="n")
axis(2,at=c(0,1),labels=c(0,1),las=2)#eje y
axis(1,at=c(3,6),labels=c("a","b"),las=0)#eje x

#Método de simulación de la transformación inversa vista sobre la función de distribución
#Se coloca la misma gráfica de distribución anterior
```

```
plot(c(2,3,6,7),c(0,0,1,1),type="l",xaxt="n",yaxt="n",xlim=c(2,7),ylim=c(0,1.2),xlab="x",ylab=ex
pression(F(x)))
title("Grafica 1.15: inversión de la función de \ndistribución de una uniforme(a,b)",cex.main=1)
# se traza un valor aleatorio y su trayectoria
yax=seq(0,5,by=.7)
yax1=rep(2/3,length(yax))
points(yax,yax1,type="c",xaxt="n",yaxt="n")
xax=seq(0,2/3,2/15)
xax1=rep(5,length(xax))
points(xax1,xax,type="c",xaxt="n",yaxt="n")
#se personalizan los ejes
axis(2,at=c(0,2/3,1),labels=c(0,"U",1),las=2)#eje y
axis(1,at=c(3,5,6),labels=c("a",expression(paste(F^-1,(U))),"b"),las=0)#eje x
#Se apoya la señalización con flechas
arrows(1.46, .711, 1.46,.95, xpd = TRUE,length=.13)
arrows(1.46, .62, 1.46,.04, xpd = TRUE,length=.13)
```

Distribución Exponencial

```
#Función de densidad
#Se traza a partir de definir una función que define la densidad
f<-function(x){0.2*exp(-0.2*x)}
curve(f,0,25,ylab=expression(paste(lambda, e^-{lambda*x})),xaxt="n",yaxt="n")
yax=seq(0,.2,by=.02)
yax1=rep(0,length(yax))
points(yax1,yax,type="c",xaxt="n",yaxt="n")
#se personalizan los ejes
axis(1,at=c(0),las=0)#eje x
axis(2,at=c(.2),labels=c(expression(lambda)),las=2)#eje y
#se colocan los títulos
title(expression(paste("Gráfica 2.44:densidad de una distribución
exponencial",(lambda))),font.main=3 )

#Función de distribución
#De igual forma se define una función con la forma de la distribución
f<-function(x){1-exp(-0.2*x)}
curve(f,0,25,ylab=expression(paste( 1-e^-{lambda*x})),xaxt="n",yaxt="n")
yax=seq(0,.2,by=.02)
yax1=rep(0,length(yax))
points(yax1,yax,type="c",xaxt="n",yaxt="n")
#se personalizan los ejes
axis(1,at=c(0),las=0)#eje x
axis(2,at=c(1),las=2)#eje y
#se colocan los títulos
```

```
title(expression(paste("Gráfica 1.20:función de distribución\n                de una
exponencial",(lambda))),font.main=3 )

# Método de simulación de la transformación inversa vista sobre la función de distribución
# se define en una función la transformada inversa
f<-function(x){1-exp(-0.2*x)}
curve(f,0,25,ylab=expression(paste( 1-e^-{lambda*x})),xaxt="n",yaxt="n")
#se colocan los títulos
title(expression(paste("Gráfica 1.21: inversión de la función de \n    distribución de una
exponencial",(lambda))),cex.main=1)
xax=seq(0,f(7),by=(f(7)/9))
xax1=rep(7,length(xax))
yax=seq(0,7,by=1.7)
yax1=rep(f(7),length(yax))
#Se traza un valor aleatorio y su trayectoria
points(yax,yax1,type="c",xaxt="n",yaxt="n")
points(xax1,xax,type="c",xaxt="n",yaxt="n")
#se personalizan los ejes
axis(2,at=c(0,f(7),1),labels=c(0,"U",1),las=2)#eje y
axis(1,at=c(0,7),labels=c(0,expression(paste(F^-1,(U)))),las=0)#eje x
# se apoya la señalización con flechas
arrows(-2.6, .79, -2.6,.96, xpd = TRUE,length=.13)
arrows(-2.6, .71, -2.6,.04, xpd = TRUE,length=.13)
```

Distribución Gamma

```
##Función de densidad, generada a partir de funciones internas de R
curve(dgamma(x,shape=50,scale=2),50,150,ylab=expression(paste(lambda^-
alpha,"/",Gamma(alpha), e^-(x/lambda),x^(alpha-1))),xaxt="n",yaxt="n")

#Se colocan los títulos
title(expression(paste("Grafica 1.30:densidad de una
Gamma(",alpha,'=50',",",lambda,"=2)")),font.main=3 )

##Función de distribución, generada a partir de funciones internas  de R
curve(dgamma(x,shape=50,scale=2),50,150,ylab=expression(paste(lambda^-
alpha,"/",Gamma(alpha), e^-(x/lambda),x^(alpha-1))))
# se colocan los títulos
title(expression(paste("Grafica 1.31:Distribución de una
Gamma(",alpha,'=50',",",lambda,"=2)")),font.main=3 )

#Función de riesgo de la distribución Gamma
#se define a partir de una función con la fórmula de la función de riesgo
f<-function(x){dgamma(x,shape=.9,scale=2)/pgamma(x,shape=.9,scale=2,lower.tail=FALSE)}
curve(f,0,16,ylab=expression(h(x)),ylim=c(0.45,.62),xaxt="n",yaxt="n")
```

```
yax=seq(0,16,by=1)
yax1=rep(.5,length(yax))
points(yax,yax1,type="c",xaxt="n",yaxt="n")
#se personalizan los ejes
axis(2,at=c(0,.5),labels=c(0,expression(frac(1,lambda))),las=2)#eje y
axis(1,at=c(0),labels=c(0),las=0)#eje x
#se colocan el título
title(expression(paste("Grafica 1.28: Función de riesgo de una
Gamma(",alpha<1,",",lambda,")")),font.main=3 )
```

Distribución Weibull

```
# Función de densidad, generada a partir de funciones internas de R
curve(dweibull(x,shape=3,scale=2),0,4,ylab=expression(paste(beta^-alpha, e^-
(x/beta)^alpha,alpha, x^(alpha-1))),ylim=c(0,.6),xaxt="n",yaxt="n")
axis(1,at=c(0),las=0)#eje x
axis(2,at=c(0,1),las=2)#eje y
title(expression(paste("Gráfica 1.33:Densidad de una Weibull(",alpha,",",beta,")")),font.main=3
)

# Función de distribución, generada a partir de funciones internas de R
curve(pweibull(x,shape=2,scale=2),0,5,ylab=expression(1-e^{-
{x^alpha}/{beta^alpha}}),ylim=c(0,1.08),xaxt="n",yaxt="n")
axis(1,at=c(0),las=0)#eje x
axis(2,at=c(0,1),las=2)#eje y
title(expression(paste("Grafica 1.34:función de distribución de una
Weibull(",alpha,",",beta,")")),font.main=3 )



# Método de simulación de la transformación inversa vista sobre la función de distribución
#Se define en una curva con la función de la transformación inversa
curve(pweibull(x,shape=2,scale=2),0,5,ylab=expression(1-e^{-
{x^alpha}/{beta^alpha}}),ylim=c(0,1.08),xaxt="n",yaxt="n")
#se colocan títulos
title(expression(paste("Grafica 1.35: Inversión de la función de \n                distribución de
una Weibull(",alpha,",",beta,")")),cex.main=1.2)
xax=seq(0,pweibull(2.5,shape=2,scale=2),by=(pweibull(2.5,shape=2,scale=2)/9))
xax1=rep(2.5,length(xax))
yax=seq(0,2.5,by=.4)
yax1=rep(pweibull(2.5,shape=2,scale=2),length(yax))
# se indica la trayectoria del valor aleatorio
points(yax,yax1,type="c",xaxt="n",yaxt="n")
points(xax1,xax,type="c",xaxt="n",yaxt="n")
# se personalizan los ejes
```

```
axis(2,at=c(0,pweibull(2.5,shape=2,scale=2),1),labels=c(0,"U",1),las=2)#eje y
axis(1,at=c(0,2.5),labels=c(0,expression(paste(F^-1,(U)))),las=0)#eje x
#se apoya la señalización con flechas
arrows(-.47, .83, -.47,.96, xpd = TRUE,length=.13)
arrows(-.47, .74, -.47,.04, xpd = TRUE,length=.13)
```

Distribución Normal

```
# Función de densidad, generada a partir de funciones internas de R
curve(dnorm(x,10,sqrt(7)),2,18,ylab=expression(paste( e^-{(x-
mu)^2/{2*sigma^2}},1/sqrt(2*pi)*sigma  )), xaxt="n",yaxt="n")
#se coloca el titulo
title(expression(paste("Gráfica 1.39: Densidad de una variable
Normal(",mu,",",sigma^2,")")),font.main=3 )
# se personalizan los ejes
axis(1,at=c(0,10),labels=c(0,expression(mu)),las=0)#eje x
axis(2,at=c(0),las=2)#eje y

#Función de Riego de la distribución normal
# Se define una función con la formula de la distribución acumulativa
f<-function(x){dnorm(x,10,sqrt(7))/pnorm(x,10,sqrt(7),lower.tail=FALSE)}
curve(f,0,30,ylab=expression(paste(h(X))),xaxt="n",yaxt="n")
abline(a=-10/7,b=1/7,lwd=2.5)
# se coloca el titulo
title(expression(paste("Gráfica 1.40: Función de riesgo de una
Normal(",mu,",",sigma^2,")")),font.main=3 )
#se personalizan los ejes
axis(1,at=c(0),labels=c(0),las=0)#eje x
axis(2,at=c(0),las=2)#eje y

# se apoya la señalización con flechas
arrows(7, 1, 13,f(13),length=.13)
arrows(18.5, .5, 13.5,.5,length=.13)
text(5.8,1,expression(h(x)))
text(21,0.5,expression(paste("y=",frac(x-mu,sigma))))
```

Distribución Lognormal

```
# Función de densidad, generada a partir de funciones internas de R
curve(dlnorm(x,0,sqrt(.3)),0,4,ylab=expression(paste( e^-{(Log(x)-
mu)^2/{2*sigma^2}},1/x*sqrt(2*pi)*sigma  )),xaxt="n",yaxt="n")
#Se coloca el titulo
title(expression(paste("Gráfica 1.44: Densidad de una Log-Normal(",mu,
",",sigma^2,")")),font.main=3 )
# se personalizan los ejes
```

```
axis(1,at=c(0,10),labels=c(0,expression(mu)),las=0)#eje x
axis(2,at=c(0),las=2)#eje y

# Función de riesgo de la distribución Lognormal
#Se define a partir de una función con la fórmula de la función de riesgo
curve(x^1 ,1,100)
curve(log(x),1,100,add=T)
f<-function(x){dlnorm(x,0,sqrt(.3))/plnorm(x,0,sqrt(.3),lower.tail=FALSE)}
curve(f,0,100,ylab=expression(paste(h(X))),xaxt="n",yaxt="n")
title(expression(paste("Gráfica 1.45: Función de riesgo de una Log-
Normal(",mu,",",sigma^2,")")),font.main=3 )
axis(1,at=c(0),labels=c(0),las=0)#eje x
axis(2,at=c(0),las=2)#eje y
log(25/sqrt(29))
```

Distribución Pareto

```
#Función de densidad
# Se traza a partir de definir una función que define la densidad
p<-function(x,t,a){(a*t^a)/((x+t)^a)}
curve(p(x,1000,3),0,15000,ylab=expression(paste( alpha*theta^alpha/{(x+theta)}^{alpha+1}
)),bty="n",xaxt="n",yaxt="n")
#Se coloca el titulo
title(expression(paste("Grafica 2.41: Densidad de una distribución
Pareto(",alpha,",",theta,")")),font.main=3 )
#Se generan los ejes
axis(1,at=c(0,15000),labels=c(0,""),las=0)#eje x
axis(2,at=c(0,3),labels=c(0,""),las=2)#eje y
#Función de distribución
# Se traza a partir de definir una función que define la distribución acumulativa
f<-function(x,t,a){1-(t/(x+t))^a}
t<-1000
a<-3
curve(f(x,1000,3),0,8000,ylab=expression(1-(theta/{x+theta})^{alpha}),xaxt="n",yaxt="n")
#Se coloca el título
title(expression(paste("Gráfica 1.50: Función de distribución de una Pareto
(",alpha,",",theta,")")),font.main=3 )
# se generan los ejes
axis(1,at=c(0),labels=c(0),las=0)#eje x
axis(2,at=c(0,1),las=2)#eje y
```

Distribución Beta

```
# Función de densidad, generada a partir de funciones internas de R
#Se divide la ventana para poder mostrar al mismo tiempo los diferentes casos de los
parámetros
```

```
win.graph()
par(mfrow=c(1,3))
#caso1
curve(dbeta(x,2,5),0,1,ylab=expression(paste( x^{alpha[1]-1}*(1-x)^{alpha[2]-
1},1/B(alpha[1],alpha[2])  )),xaxt="n",yaxt="n")
title(expression(paste("Grafica 1.54: Densidad de una
Beta(",alpha[1],'<',alpha[2],')')),font.main=3 )
#se generan los ejes
axis(1,at=c(0,1),labels=c(0,1),las=0)#eje x
axis(2,at=c(0),las=2)#eje y
#caso 2
curve(dbeta(x,10,10),0,1,ylab=expression(paste( x^{alpha[1]-1}*(1-x)^{alpha[2]-
1},1/B(alpha[1],alpha[2])  )),xaxt='n',yaxt='n')
title(expression(paste("Gráfica 1.55: Densidad de una Beta ",alpha[1]==alpha[2])),font.main=3
)
#se generan los ejes
axis(1,at=c(0,1),labels=c(0,1),las=0)#eje x
axis(2,at=c(0),las=2)#eje y
#caso 3
curve(dbeta(x,5,2),0,1,ylab=expression(paste( x^{alpha[1]-1}*(1-x)^{alpha[2]-
1},1/B(alpha[1],alpha[2])  )),xaxt='n',yaxt='n')
title(expression(paste("Gráfica 1.56: Densidad de una Beta ",alpha[1]>alpha[2])),font.main=3 )
#se generan los ejes
axis(1,at=c(0,1),labels=c(0,1),las=0)#eje x
axis(2,at=c(0),las=2)#eje y


#se colocan las graficas de las funciones de distribuciones para los diversos casos de los
#parámetros de la distribución Beta
win.graph()
par(mfrow=c(1,3))
#caso 1
curve(pbeta(x,2,5),0,1,ylab=expression(F(x)),xaxt='n',yaxt='n')
title(expression(paste("Gráfica 1.57: Distribución de una Beta
",alpha[1]<alpha[2])),font.main=3 )
axis(1,at=c(0,1),labels=c(0,1),las=0)#eje x
axis(2,at=c(0,1),las=2)#eje y
#caso 2
curve(pbeta(x,10,10),0,1,ylab=expression(F(x)),xaxt='n',yaxt='n')
title(expression(paste("Grafica 1.58: Distribución de una Beta
",alpha[1]==alpha[2])),font.main=3 )
axis(1,at=c(0,1),labels=c(0,1),las=0)#eje x
axis(2,at=c(0,1),las=2)#eje y
#caso 3
curve(pbeta(x,5,2),0,1,ylab=expression(F(x)),xaxt='n',yaxt='n')
```

```
title(expression(paste("Grafica 1.59: Distribución de una Beta
",alpha[1]>alpha[2])),font.main=3 )
axis(1,at=c(0,1),labels=c(0,1),las=0)#eje x
axis(2,at=c(0,1),las=2)#eje y

#se colocan las funciones de riego al mismo tiempo de acuerdo a los casos de los parámetros,
y partiendo de la definición de una función
win.graph()
par(mfrow=c(1,3))
#Caso 1
b<-function(x,a,b){dbeta(x,a,b)/pbeta(x,a,b,lower.tail=FALSE)}
curve(b(x,2,5),0,1,ylab=expression(paste( h[1](x)  )))
title(expression(paste("Grafica:funcion de riesgo de una
Beta(",alpha[1],"=2",",",alpha[2],"=5)")),font.main=3 )
#Caso 2
curve(b(x,10,10),0,1,ylab=expression(paste( h(x)  )))
title(expression(paste("Grafica:densidad de una
Beta(",alpha[1],"=10",",",alpha[2],"=10)")),font.main=3 )
#Caso 3
curve(b(x,5,2),0,1,ylab=expression(paste( h[3](x)  )))
title(expression(paste("Grafica:densidad de una
Beta(",alpha[1],"=5",",",alpha[2],"=2)")),font.main=3 )
```

Gráficos del Capítulo II

```
## Guillermo Cuauhtemoctin Granados García
## Reporte de actividad Docente: Técnicas Monte Carlo aplicadas a la enseñanza de la
estadística.
```

Código para la generación de un mapa de las provincias de Cánada

```
##Con asignación de color de acuerdo al número de nacimientos del año 2011 de acuerdo al
reporte de Statistics Canada

##debe instalarse la siguiente librería
library(raster)

## aquí se obtienen los datos de la base precargada en la librería
canada <- getData("GADM",country="CAN",level=1)
ca.provinces <- canada[canada$NAME_1 %in% provinces,]

## Se emplea el metodo bbox para interpretar los datos como dimensiones graficas
ca.bbox <- bbox(ca.provinces)
xlim <- c(min(us.bbox[1,1],ca.bbox[1,1]),max(us.bbox[1,2],ca.bbox[1,2]))
ylim <- c(min(us.bbox[2,1],ca.bbox[2,1]),max(us.bbox[2,2],ca.bbox[2,2]))
```

```
## Aquí se asigna los colores del grafico en orden de las provincias
## Los números son parte de un catalogo interno de R
y<-colors()[c(371,371,565,131,131,131,131,131,131,131,553,131,133,565,131)]

## Se genera un vector de etiquetas para las provincias
z<-c('Alta.',  'B.C.',  'Man.',  'N.B.',  'N.L.',  'N.W.T.',  'N.S.',  'Nvt.',  'Ont.',  'P.E.I.',
     'Que.',  'Sask.',  'Y.T.')

## Se grafican los datos
plot(ca.provinces, col=y)
## las etiquetas son colocadas
text(ca.provinces, labels=z, cex=.8)
```

Gráfica para mostrar el comportamiento de la función de distribución empírica

```
## Se realiza la simulación de valores distribuidos como una variable gamma,
## Y se obtiene la función de distribución empírica para comparar contra la teórica
u<-rgamma(25,6,2.5)
F10 <- ecdf(u)
#distribución empírica
plot(F10,xlim=c(0,6), verticals = TRUE, do.points =
FALSE,xaxt="n",yaxt="n",main=NULL,axes="FALSE",ylab=expression(S[n](x)))
#Distribución teórica
curve(pgamma(x,6,2.5),0,6,lwd=2,xaxt="n",yaxt="n",add=T)
axis(2,at=c(0,1),labels=c(0,1),las=2)#eje y
#se coloca el título
title("visualización gráfica de F(x) y el cálculo\n de la función de distribución empírica")
legend(3.8,.4,c(expression(S[n](x)),expression(F[X](x))), lwd=c(1,3))
#se señalan dos puntos para mostrar el salto en la distribución empírica
points(c(sort(u)[14],sort(u)[14]),c(0,13/26),type="l")
points(c(sort(u)[15],sort(u)[15]),c(0,14/26),type="l")
axis(1,at=c(min(u),sort(u)[14],sort(u)[15],max(u)),labels=c("",expression(x[i]),expression(x[i+1])
,""),las=0,tck=0)#eje x
```

Gráfica para mostrar el tema de media aritmética, sobre una densidad Gamma

```
##densidad Gamma, con parámetros (2,1) para asegurar una media de valor 2
curve(dgamma(x,shape=2,scale=1),0,8,xlab="", ylab=expression(f[X](x)),xaxt="n",yaxt="n")
title(expression(paste("Grafica 2.12:densidad f(x) y su media " ,mu )),cex.main=1 )
#se genera una línea vertical sobre el valor teórico de la media
yay=seq(0,dgamma(2,shape=2,scale=1),by=.05)
yax=rep(2,length(yay))
points(yax,yay,type="c",xaxt="n",yaxt="n")
```

```
points(2,dgamma(2,shape=2,scale=1),pch = 19,xaxt="n",yaxt="n" ,cex=1)
#se apoya la señalización con flechas y texto
text(2.3,dgamma(2,shape=2,scale=1),expression(mu),cex=1.3)
arrows(16.15,-.016,16.15,-.00008,length=.13)

## Grafica para mostrar las tres medidas de posición, sobre una serie de densidades beta
#la distribución beta se eligió por su versatilidad en generar densidad con asimetrías positivas
o negativas

#funciones de densidad
win.graph()
par(mfrow=c(1,1))
##  Densidad asimétrica positiva
curve(dbeta(x,2,5),0,1,ylab=expression(f[x](x)) ,axes="FALSE",ylim=c(-.15,(dbeta(.2,2,5) +.3)))
title(expression(paste("Grafica 2.21:Densidad Asimétrica positiva\n media,mediana y moda
indicadas")),font.main=2 )
axis(1,at=c(0,1),labels=c("",""),las=0,tck=0)#eje x
#moda
points(c(.2,.2),c(0,dbeta(.2,2,5)),type="l",xaxt="n",yaxt="n")
text(1/5,(dbeta(.2,2,5) +.2),expression(Moda),cex=1)
#mediana
points(c(5/19,5/19),c(0,dbeta(5/19,2,5)),type="l",xaxt="n",yaxt="n")
text(2/7,(dbeta(2/7,2,5) +.2),expression(Me[X]),cex=1)
#media
points(c(2/7,2/7),c(0,dbeta(2/7,2,5)),type="l",xaxt="n",yaxt="n")
text((2/7+.05),(dbeta(2/7,2,5) ),expression(bar(X)),cex=1)
text(.5,-.1,expression(paste(Moda, " < " ,Me[X] ," < ", bar(X))),cex=1)

##  medidas de posición sobre una densidad simétrica
curve(dbeta(x,5,5),0,1,ylab=expression(f[x](x)) ,axes="FALSE",ylim=c(-.15,(dbeta(.5,5,5) +.3)))
title(expression(paste("Grafica 2.20:Densidad simétrica\n media, mediana y moda
coincidentes")),font.main=2 )
axis(1,at=c(0,1),labels=c("",""),las=0,tck=0)#eje x
#moda mediana y media
points(c(.5,.5),c(0,dbeta(.5,5,5)),type="l",xaxt="n",yaxt="n")
text(.5,(dbeta(.5,5,5)+.2),expression(paste(Moda, " = " ,Me[X] ," = ", bar(X))),cex=1)

##  medidas de posición sobre una densidad asimétrica negativa
curve(dbeta(x,5,2),0,1,ylab=expression(f[x](x)) ,axes="FALSE",ylim=c(-.15,(dbeta(4/5,5,2) +.3)))
title(expression(paste("Grafica 2.22 :Densidad Asimétrica negativa\n media,mediana y moda
indicadas")),font.main=2 )
axis(1,at=c(0,1),labels=c("",""),las=0,tck=0)#eje x
#moda
points(c(4/5,4/5),c(0,dbeta(4/5,5,2)),type="l",xaxt="n",yaxt="n")
text(4/5,(dbeta(4/5,5,2) +.2),expression(Moda),cex=1)
```

```
#mediana
points(c(14/19,14/19),c(0,dbeta(14/19,5,2)),type="l",xaxt="n",yaxt="n")
text((5/7-.01),(dbeta(5/7,5,2) +.2),expression(Me[X]),cex=1)
#media
points(c(5/7,5/7),c(0,dbeta(5/7,5,2)),type="l",xaxt="n",yaxt="n")
text((5/7-.05),(dbeta(5/7,5,2) ),expression(bar(X)),cex=1)
text(.5,-.1,expression(paste(Moda, " > " ,Me[X] ," > ", bar(X))),cex=1)
```

Gráficas para mostrar el tema de rango intercuartil sobre una densidad Normal

```
#se genera una curva con la densidad Normal
curve(dnorm(x,10,sqrt(8)),2,18,ylab=expression( e^-{x^2/{2}}/sqrt(2*pi)  ),
xaxt="n",yaxt="n",ylim=c(-.024,.14),xlab="")
title(expression(paste("Gráfica 2.34: Posición de los cuartiles\n en una función de densidad
Normal(", mu,",",sigma^2,")")),font.main=3 )
axis(1,at=c(0,10),labels=c(0,expression(0)),las=0)#eje x
axis(2,at=c(0),las=2)#eje y
#puntos para hacer el dash sobre la densidad
ye=seq(0,dnorm(10,10,sqrt(8)),by=(dnorm(10,10,sqrt(8))/10))
eq=rep(10,length(ye))
points(eq,ye,type="l")
points(c(-1,30),c(0,0),type="l") # linea y=0
points(c(qnorm(.25,10,sqrt(8)),qnorm(.25,10,sqrt(8))),c(0,dnorm(qnorm(.25,10,sqrt(8)),10,sqrt
(8))),type="l",lwd=1.5) #cuantil izquierdo
points(c(qnorm(.75,10,sqrt(8)),qnorm(.75,10,sqrt(8))),c(0,dnorm(qnorm(.75,10,sqrt(8)),10,sqrt
(8))),type="l",lwd=1.5) #cuantil derecho
#Etiquetas de % acumulado
text(9,.037,expression("25%"),cex=1)
text(11,.037,expression("25%"),cex=1)
text(6.5,.027,expression("25%"),cex=1)
text(14,.027,expression("25%"),cex=1)
# etiquetas de los cuantiles
text(qnorm(.25,10,sqrt(8)),-.01,expression(X['25%']))
text(qnorm(.50,10,sqrt(8)),-.01,expression(X['50%']))
text(qnorm(.75,10,sqrt(8)),-.01,expression(X['75%']))

# Rango intercuartil sobre la distribución acumulativa de una Normal
curve(pnorm(x,10,sqrt(8)),2,18,ylab="", xaxt="n",yaxt="n",ylim=c(0,1),xlab="")
title(expression(paste("Gráfica 2.35: Posición de los cuartiles \n en distribución acumulativa de
una Normal(", mu,",",sigma^2,")")),cex.main=1)
xax=seq(0,.25,by=(.25/9))
yax=seq(0,.50,by=(.5/9))
zax=seq(0,.75,by=(.75/9))
#se generan los vectores que conforman las líneas para señalar los cuantiles
xax1=rep(qnorm(.25,10,sqrt(8)),length(xax))
yax1=rep(qnorm(.50,10,sqrt(8)),length(xax))
```

```
zax1=rep(qnorm(.75,10,sqrt(8)),length(xax))
#se trazan líneas sobre los cuantiles
points(xax1,xax,type="l",xaxt="n",yaxt="n")
points(yax1,yax,type="l",xaxt="n",yaxt="n")
points(zax1,zax,type="l",xaxt="n",yaxt="n")
ax1=seq(2,qnorm(.25,10,sqrt(8)),by=(qnorm(.25,10,sqrt(8))/12))
ay1=seq(2,qnorm(.50,10,sqrt(8)),by=(qnorm(.50,10,sqrt(8))/10))
az1=seq(2,qnorm(.75,10,sqrt(8)),by=(qnorm(.75,10,sqrt(8))/10))
ax=rep(.25,length(ax1))
ay=rep(.5,length(ay1))
az=rep(.75,length(az1))
#también se traza el mapeo sobre el eje y
points(ax1,ax,type="l",xaxt="n",yaxt="n")
points(ay1,ay,type="l",xaxt="n",yaxt="n")
points(az1,az,type="l",xaxt="n",yaxt="n")
axis(2,at=c(0,.25,.5,.75,1),labels=c(0,'25%','50%','75%',1),las=2)#eje y
axis(1,at=c(2,qnorm(.25,10,sqrt(8)),qnorm(.5,10,sqrt(8)),qnorm(.75,10,sqrt(8))),labels=c("",exp
ression(X['25%']) ,expression(X['50%']),expression(X['75%']) ),las=0)#eje x
```

#Mapa de México para el tema de coeficiente de variación
 # El siguiente código es una modificación, al publicado en la página Web
#https://chanchullopolitico.wordpress.com/

```
#en su publicación:
#"MAPAS EN R… LOS PARTIDOS QUE GOBIERNAN LOS ESTADOS EN MÉXICO "
# Los archivos del shapefile se pueden descargar de la misma publicación
# El cambio sustancial se produce en el uso de un gradiente de color para colocar los precios
promedio
#El working directory donde se encuentra el dichoso shapefile.
setwd("C:/Mexico/MEX_adm")

#Los paquetes a utilizar.
require(ggmap)
require(ggplot2)
require(maptools)
require(car)

#Se descarga el shapefile en R. La función readShapeSpatial viene en el paquete maptools
estados.shape = readShapeSpatial("MEX_adm1.shp")

#Se convierte este shapefile a un data.frame de R para poder utilizarlo con ggmap, ggplot2 o
para generar más variables.
#La función para convertir este shapefile en un data.frame se llama fortify y viene en el
paquete ggplot2.
estados.poly = fortify(estados.shape)
```

```
#se revisa la información que tiene el data.frame que acabamos de generar.
# tiene dos variables numéricas para longitud y latitud (long y lat respectivamente), order
(variable de enteros que une los puntos para generar los polígonos), etc.
str(estados.poly)
# Cada argumento lo explico en el párrafo de aquí arriba.
imagen = get_map(location=c(right=-85, left=-121, bottom=13, top=33), source="google",
color="bw")
#se observa el mapa con la función ggmap.
ggmap(imagen)
#se cambia el valor de cada una de las etiquetas del estado por el valor de su precio promedio
de tortillas
estados.poly$id2 = as.numeric(estados.poly$id) + 1
y<-seq(1,32,by=1)
#promedio por estado del precio de tortillas
preciosT<-
c(11.6,14.9,14.3,14.4,14.3,13.2,11.7,13.2,11.1,10.9,11.3,14.3,9.9,11.9,11.3,12.3,12.8,14.3,13.5,
11.7,10.5,12.1,13.7,11.7,14.3,15.5,13.5,13.4,10.2,11.8,14.3,11.5)

#aquí se sustituyen los valores promedio al dataframe
for(i in y){
estados.poly$id2[estados.poly$id2==i] <-preciosT[i]
}

mapa = ggmap(imagen) + #Esto genera la imagen de México que ya se vio anteriormente.
geom_polygon(data=estados.poly, aes(x=long, y=lat, group=group), colour="grey",
fill="#ffffff") + #Genera el relieve de los estados con un contorno gris y un fondo blanco.
geom_polygon(data=estados.poly, aes(x=long, y=lat, group=group, fill=estados.poly$id2),
alpha=0.5) +
#se utiliza el siguiente método para relacionar los precios promedio con una escala de color
 scale_fill_gradient("Precios promedio", low="white", high="#cc0000", space="Lab",
breaks=c(10,11,12,13,14,15),labels=c(10,11,12,13,14,15))+
labs(x="Longitud", y="Latitud") + #El nombre de los ejes.
ggtitle("Grafica 2.37:Precios promedio de 1 Kg de Tortilla por Estado Ene-2016. \n
Consulta en línea de los precios promedio del INPC \n") #El título
#se ejecuta la observación  del mapa:
mapa
```

Gráficos del Capítulo III

###Gráficas para mostrar la densidad de una Normal estandarizada

```
#se emplea una grafica para el intervalo de confianza de la media por medio de una densidad
Normal,
curve(dnorm(x,10,sqrt(8)),2,18,ylab=expression( e^-{x^2/{2}}/sqrt(2*pi)  ),
xaxt="n",yaxt="n",ylim=c(-.024,.14),xlab="")
```

```
#se coloca el título
title(expression(paste("Gráfica 3.1: Intervalo para ", (bar(X)-mu)/sqrt(sigma^2/n))),font.main=3
)
#se adicionan los ejes
axis(1,at=c(0,10),labels=c(0,expression(0)),las=0)#eje x
axis(2,at=c(0),las=2)#eje y

#puntos para hacer el dash
ye=seq(0,dnorm(10,10,sqrt(8)),by=(dnorm(10,10,sqrt(8))/10))
eq=rep(10,length(ye))
points(eq,ye,type="c")
points(c(-1,30),c(0,0),type="l") # linea y=0
points(c(qnorm(.04,10,sqrt(8)),qnorm(.04,10,sqrt(8))),c(0,dnorm(qnorm(.04,10,sqrt(8)),10,sqrt
(8))),type="l",lwd=1.5) #cuantil izquierdo
points(c(qnorm(.96,10,sqrt(8)),qnorm(.96,10,sqrt(8))),c(0,dnorm(qnorm(.96,10,sqrt(8)),10,sqrt
(8))),type="l",lwd=1.5) #cuantil derecho
#se etiqueta el área de probabilidad 1- alfa
text(10,.087,expression(1-alpha),cex=2)
# se etiqueta las áreas de probabilidad
arrows(qnorm(.009,10,sqrt(8)),.04,qnorm(.02,10,sqrt(8)),.009,length=.13)
arrows(qnorm(1-.009,10,sqrt(8)),.04,qnorm(1-.02,10,sqrt(8)),.009,length=.13)
 #se colocan las etiquetas en el eje X
text(qnorm(.009,10,sqrt(8)),.047,expression(alpha/2),cex=1.3)
text(qnorm(1-.009,10,sqrt(8)),.047,expression(alpha/2),cex=1.3)
text(qnorm(.04,10,sqrt(8)),-.01,expression(Z[alpha/2]))
text(qnorm(1-.04,10,sqrt(8)),-.01,expression(Z[1-alpha/2]))
text(10,-.014,expression(frac(bar(X)-mu,sqrt(sigma^2/n))))
```

####Gráfico para comparar la densidad de una Normal VS la densidad de una variable t de Student

```
#densidad Normal
curve(dnorm(x,0,1),-3.5,3.5,ylab=expression( f(x)  ), xaxt="n",yaxt="n",ylim=c(-.07,.4),xlab="")
#se coloca el titulo
title(expression(paste("Gráfica 3.3: Función de densidad Normal de ",(bar(X)-
mu)/sqrt(S^2/n))),font.main=3 )
mtext("VS la función de densidad t de Student",3, ,cex=1.2)
#se colocan los ejes
axis(1,at=c(0,10),labels=c(0,expression(0)),las=0)#eje x
axis(2,at=c(0),las=2)#eje y
#se adiciona sobre la gráfica anterior la densidad de una t de Student
curve(dt(x,3),-3.5,3.5, xaxt="n",yaxt="n",xlab="",add=TRUE,lwd=4)
#puntos para hecer el dash
ye=seq(0,dnorm(0,0,sqrt(1)),by=(dnorm(0,0,sqrt(1))/10))
eq=rep(0,length(ye))
points(eq,ye,type="c")
```

```
points(c(-4,4),c(0,0),type="l") # línea y=0
# líneas de  cuantiles
points(c(qnorm(.04,0,sqrt(1)),qnorm(.04,0,sqrt(1))),c(0,dnorm(qnorm(.04,0,sqrt(1)),0,sqrt(1))),
type="l",lwd=1.5) #cuantil izquierdo
points(c(qnorm(.96,0,sqrt(1)),qnorm(.96,0,sqrt(1))),c(0,dnorm(qnorm(.96,0,sqrt(1)),0,sqrt(1))),
type="l",lwd=1.5) #cuantil derecho
points(c(qt(.04,3),qt(.04,3)),c(0,dt(qt(.04,3),3)),type="l",lwd=2.8) #cuantil izquierdo
points(c(qt(.96,3),qt(.96,3)),c(0,dt(qt(.96,3),3)),type="l",lwd=2.8) #cuantil derecho
#se etiqueta el área de probabilidad
text(0,.2,expression(1-alpha),cex=2)
#se apoya el grafico con flechas
arrows(qnorm(.0009,0,sqrt(1)),.07,qnorm(.0009,0,sqrt(1)),.007,length=.07)
arrows(qnorm(1-.0009,0,sqrt(1)),.075,qnorm(1-.0009,0,sqrt(1)),.007,length=.07)
 #se etiquetan las áreas laterales de probabilidad
text(qnorm(.0009,0,sqrt(1)),.082,expression(alpha/2),cex=1.3)
text(qnorm(1-.0009,0,sqrt(1)),.082,expression(alpha/2),cex=1.3)
text(qnorm(.04,0,sqrt(1)),-.015,expression(Z[alpha/2]))
text(qnorm(1-.04,0,sqrt(1)),-.015,expression(Z[1-alpha/2]))
text(qt(.04,3),-.015,expression(t[paste(frac(alpha,2),",",n-1)]))
text(qt(1-.04,3),-.015,expression(t[paste('1-',frac(alpha,2),",",n-1)]))
text(0,-.048,expression(frac(bar(X)-mu,sqrt(S^2/n))))
# se coloca la diferenciación de los gráficos con una leyenda
legend(1.65,.4,c(expression(N(0,1)),expression(t[n-1])), lwd=c(1,4))
```

Curva y cuantiles de la distribución Ji-cuadrada

```
# se genera la curva con la ayuda de funciones internas
curve(dchisq(x,df=7),0,18,ylab=expression( 1 / {2^{n/2}*Gamma(n/2)}*x^{n/2-1}*e^{-x/2}),
xaxt="n",yaxt="n",ylim=c(-.024,.14),xlab="")
title(expression(paste("Gráfica 3.5: Densidad de una ",Ji^2, " y sus cuantiles" ),font.main=3 ))
axis(1,at=c(0),labels=c(0),las=0)#eje x
axis(2,at=c(0),las=2)#eje y
#puntos para hecer el dash
points(c(-1,30),c(0,0),type="l") # linea y=0
points(c(qchisq(.04,df=7),qchisq(.04,df=7)),c(0,dchisq(qchisq(.04,df=7),df=7)),type="l",lwd=1.5)
#cuantil izquierdo
points(c(qchisq(.96,df=7),qchisq(.96,df=7)),c(0,dchisq(qchisq(.96,df=7),df=7)),type="l",lwd=1.5)
#cuantil derecho
text(5.5,.085,expression(1-alpha),cex=2)
# se colocan flechas para apoyar el gráfico
arrows(qchisq(.009,df=7),.05,qchisq(.02,df=7),.013,length=.13)
arrows(qchisq(1-.015,df=7),.04,qchisq(1-.02,df=7),.009,length=.13)
 # se etiquetan las áreas de probabilidad y los cuantiles
text(qchisq(.007,df=7),.055,expression(alpha/2),cex=1)
```

```
text(qchisq(1-.015,df=7),.047,expression(alpha/2),cex=1)
text(qchisq(.04,df=7),-.01,expression(R[alpha/2]))
text(qchisq(1-.04,df=7),-.01,expression(R[1-alpha/2]))
text(5.5,-.014,expression(frac((n-1)*S^2,sigma^2)))
```

####Intervalos de confianza t, para la diferencia de medias

```
# se genera una densidad de la distribución t de Student
curve(dt(x,3),-3.5,3.5,ylab=expression( f(x)  ), xaxt="n",yaxt="n",ylim=c(-
.07,.4),xlab="",axes="FALSE")
# se coloca el titulo
title(expression(paste("Gráfica 3.7: Distribución t para el intervalo \n de diferencias de
medias")),font.main=3 )
#se colocan los ejes
axis(1,at=c(0),labels=c(0),las=0)#eje x
axis(2,at=c(0),las=2)#eje y

#puntos para hacer el dash
ye=seq(0,dnorm(0,0,sqrt(1)),by=(dnorm(0,0,sqrt(1))/10))
eq=rep(0,length(ye))
points(eq,ye,type="c")
points(c(-4,4),c(0,0),type="l") # linea y=0
# líneas de  cuantiles
points(c(qt(.04,3),qt(.04,3)),c(0,dt(qt(.04,3),3)),type="l",lwd=2.8) #cuantil izquierdo
points(c(qt(.96,3),qt(.96,3)),c(0,dt(qt(.96,3),3)),type="l",lwd=2.8) #cuantil derecho
#Se apoya el gráfico con flechas
arrows(qnorm(.0009,0,sqrt(1)),.07,qnorm(.0009,0,sqrt(1)),.007,length=.07)
arrows(qnorm(1-.0009,0,sqrt(1)),.075,qnorm(1-.0009,0,sqrt(1)),.007,length=.07)
#se colocan las etiquetas de las áreas de probabilidad y los cuantiles
text(0,.2,expression(1-alpha),cex=2)
text(qnorm(.0009,0,sqrt(1)),.1,expression(alpha/2),cex=1.3)
text(qnorm(1-.0009,0,sqrt(1)),.1,expression(alpha/2),cex=1.3)
text(qt(.04,3),-.015,expression(t[paste(frac(alpha,2),",",n[2]+n[1]-2)]))
text(qt(1-.04,3),-.015,expression(t[paste('1-',frac(alpha,2),",",n[2]+n[1]-2)]))
text(0,-.048,expression(frac((bar(X)-bar(Y))-(mu[X]-mu[Y]),S[P]*sqrt(1/n[1]+1/n[2]))))
```

Gráficos del Capítulo IV

Gráfica básica para el tema de pruebas de hipótesis

```
# Se utiliza la densidad normal
curve(dnorm(x,10,sqrt(8)),2,18,ylab=expression( e^-{x^2/{2}}/sqrt(2*pi)  ),
xaxt="n",yaxt="n",ylim=c(-.024,.14),xlab="")
title(expression(paste("Gráfica_: Región de rechazo para ", H[0])),font.main=3 )
axis(1,at=c(0,10),labels=c(0,expression(0)),las=0)#eje x
axis(2,at=c(0),las=2)#eje y
```

```
#puntos para hacer el dash
ye=seq(0,dnorm(10,10,sqrt(8)),by=(dnorm(10,10,sqrt(8))/10))
eq=rep(10,length(ye))
points(eq,ye,type="c")
points(c(-1,30),c(0,0),type="l") # línea y=0
points(c(qnorm(.95,10,sqrt(8)),qnorm(.95,10,sqrt(8))),c(0,dnorm(qnorm(.95,10,sqrt(8)),10,sqrt
(8))),type="l",lwd=1.5) #cuantil derecho
text(10,.087,expression(1-alpha),cex=2)
# las siguientes etiquetas y flechas se pueden omitir de acuerdo a la hipótesis presentada
arrows(qnorm(1-.009,10,sqrt(8)),.04,qnorm(1-.02,10,sqrt(8)),.009,length=.13)
text(qnorm(1-.009,10,sqrt(8)),.047,expression(alpha),cex=1.3)
text(qnorm(1-.05,10,sqrt(8)),-.008,expression(Z[1-alpha]))
text(14,-.02,expression(paste("Rechazar ",H[0]," si ",Z)))
arrows(16.15,-.016,16.15,-.00008,length=.13)
```

Gráficos del Capítulo V

```
# Gráficos de apoyo al tema: regresión lineal, densidad del libro de estadístico Durbin Watson
tomando como referencia el libro Statistics de Yamane (1969)
#densidad Normal
curve(dnorm(x,10,sqrt(8)),2,18,ylab=expression("densidad de d" ), xaxt="n",yaxt="n",ylim=c(-
.024,.14),xlab="",axes="FALSE")
# se coloca el título
title(expression(paste("Gráfica 5.29: Reglas de decisión\n para la prueba de Durbin-
Watson")),font.main=3 )
# se coloca el eje x
axis(1,at=c(0,10,20),labels=c(0,expression(2),""),las=0)#eje x

#puntos para hacer el dash
points(c(qnorm(.20,10,sqrt(8)),qnorm(.2,10,sqrt(8))),c(-
.015,dnorm(qnorm(.2,10,sqrt(8)),10,sqrt(8))),type="l",lwd=1.5) #cuantil izquierdo
points(c(qnorm(.8,10,sqrt(8)),qnorm(.8,10,sqrt(8))),c(-
.015,dnorm(qnorm(.8,10,sqrt(8)),10,sqrt(8))),type="l",lwd=1.5) #cuantil derecho
points(c(qnorm(.05,10,sqrt(8)),qnorm(.05,10,sqrt(8))),c(-
.015,dnorm(qnorm(.05,10,sqrt(8)),10,sqrt(8))),type="l",lwd=1.5) #cuantil izquierdo
points(c(qnorm(.95,10,sqrt(8)),qnorm(.95,10,sqrt(8))),c(-
.015,dnorm(qnorm(.95,10,sqrt(8)),10,sqrt(8))),type="l",lwd=1.5) #cuantil derecho
#etiquetas de áreas de rechazo
text(10,-.01,expression("Inconcluso"),cex=1)
text(10,.09,expression("No Correlación"),cex=1)
text(4,.045,expression("Correlación \n Positiva"),cex=1)
text(16,.045,expression("Correlación \n Negativa"),cex=1)

#flechas para señalar las zonas de rechazo e inconclusas
```

```
arrows(8.3, -.01, 6.5,0.015, xpd = TRUE,length=.13)
arrows(11.7, -.01, 13.5,0.015, xpd = TRUE,length=.13)
arrows(4, .04, 4,0.01, xpd = TRUE,length=.13)
arrows(16, .04, 16,0.01, xpd = TRUE,length=.13)
#etiquetas de los limites superiores e inferiores
text(qnorm(.05,10,sqrt(8)),-.025,expression(d[L]))
text(qnorm(.20,10,sqrt(8)),-.025,expression(d[U]))
text(qnorm(.8,10,sqrt(8)),-.025,expression(4-d[U]))
text(qnorm(.95,10,sqrt(8)),-.025,expression(4-d[L]))

# Gráfico de la densidad de la distribución Laplace empleada para simular valores de una
distribución de cola pesada
# se requiere un paquete extra
require(smoothmest)

#Se emplea una función del paquete para graficar la densidad Laplace
curve(ddoublex(x,10,2),2,18,ylab=expression( e^-{group("|",x-mu,"|")/{b}}/(2*b)     ),
xaxt="n",yaxt="n",ylim=c(-.024,.25),xlab="",axes="FALSE")
# se coloca el título
title(expression(paste("Gráfica 5.36: densidad de la\n          distribución Laplace(" ,mu," , b)"
)),font.main=3 )
#se coloca el eje X
axis(1,at=c(0,10,20),labels=c(0,expression(mu),""),las=0)#eje
```

Bibliografía

BARRERA S. sustentante, (1993) *Algunas aplicaciones de modelos estadísticos y actuariales en ecología*, Facultad de Ciencias, UNAM.

BOWLEY, A. L. (1920) *Elements of Statistics*, Charles Scribner's Sons, New York.

BRATLEY, P.; FOX, B. L.; SCHRAGE, L. E.; (1987) *A Guide to Simulation*, Springer Science & Business Media, New York.

CASELLA, G.; BERGER, R. L. (2002) *Statistical Inference*, Segunda Edición, Duxbury Advanced Series, Universidad de Florida, U.S.A.

DEPARTMENT OF GENETICS, STANFORD SCHOOL OF MEDICINE (2013), *Height Hopes: A new study identifies potential height genes*, The Tech Museum of Innovation ©, disponible en: *http://genetics.thetech.org/original_news/news60.*

FISHMAN, G. S.; (2001) *Discrete-Event Simulation Modeling, Programming and Analysis*, Editorial Springer-Verlag, New York.

GALTON, F. F.R.S. &C; (1886) *regression towards mediocrity in hereditary stature*, Anthropological Miscellanea, disponible en: *http://galton.org/essays/1880-1889/galton-1886-jaigi-regression-stature.pdf.*

GIBBONS, J. D.; SUBHABRATA C.; (2003) *Nonparametric statistical inference*, 4a. Edición, New York.

GILBERT, J. K.; REINER, M.; NAKHLEH, M.; (2008) *Visualization: Theory and Practice in Science Education, Models and Modeling in Science Education*, Volumen 3, Editorial Springer.

HAMMERSLEY, J. M.; (1979) *Monte Carlo Methods*, Chapman and Hall, London. *http://www.cs.fsu.edu/~mascagni/Hammersley-Handscomb.pdf.*

HEIBERGER, R. M.; NEUWIRTH, E.; (2009) *R Through Excel A Spreadsheet Interface for Statistics Data Analysis, and Graphics,* Springer Science+Business Media LLC.

HOGG, R. V.; MCKEAN, J. W.; CRAIG, A. T.; (2005) *Introduction to Mathematical Statistics*, Sexta Edición, Pearson Education Inc., U.S.A.

JOHNSON, N. L., KEMP, A. W.; (2005) *Univariate Discrete Distributions*, Tercera Edición, John Wiley & Sons, Inc.

JOHNSON, N. L.; KOTZ, S.; BALAKRISHNAN; (1970) *Continuous Univariate Distributions*, Volumen 1, Segunda edición, John Wiley & Sons, Inc.

JOHNSON, N. L.; KOTZ, S.; BALAKRISHNAN; (1970) *Continuous Univariate Distributions*, Volumen 2, Segunda edición, John Wiley & Sons, Inc.

KIM, T.; WHITE, H.; (2003) *On More Robust Estimation of Skewness and Kurtosis: Simulation and Application to the S&P500 Index*, University of California, San Diego. Disponible en: *http://www.cirano.qc.ca/realisations/grandes_conferences/methodes_econometriques/white.pdf.*

KLUGMAN, S. A.; PANJER, H., H.; WILLMOT, G. E.; (2004) *Loss Models: from data decisions*, Segunda Edición, Wiley &Sons, New Jersey.

LAW, A. M. (2007) *Simulation Modeling & Analysis*, 4ª. Edición, Editorial McGraw Hill, U.S.A.

LUMNEY, T.; DIEHR, P. ; EMERSON, S.; CHEN, L. (2002) *The Importance of the Normality Assumption in large public Health data sets* Copyright® 2002 por Annual Reviews,Washington. Disponible en: http://rctdesign.org/techreports/arphnonnormality.pdf

MERCE, E., (2009) *Pearson's Coefficient of Variation, an Erroneous History in Assessing the Degree of Significance of the Mean Value*, University of Agricultural Sciences and Veterinary Medicine, Boletín UASVM sobre Horticultura, 66(2)/2009 *http://journals.usamvcluj.ro/index.php/horticulture/article/download/4355/4047.*

MILAN, A. (2013) *Fertility: Overview, 2009 to 2011*, Parte del catálogo de Statistics Canada No. 91-209-X, Canadá, Disponible en: http://www.statcan.gc.ca/pub/91-209-x/2013001/article/11784-eng.pdf

MOOD, A. M. (1974) *Introduction to the theory of statistics*, Editorial McGraw-Hill, U.S.A.

NANCE, R. E.; (1995) *Simulation Programming Languages: an abridged history*, Systems Research Center and Department of Computer Science, Virginia Polytechnic Institute and State, University Blacksburg, Virginia, U.S.A. disponible en: *https://repository.lib.ncsu.edu/bitstream/handle/1840.4/6177/1995_0199.pdf?sequence=1&isAllowed=y.*

NAYLOR, T. H.; (1971) *Técnicas de simulación en computadoras*, Editorial Limusa-Wiley.

RAND CORPORATION, (2001) *A million random digits with 100,000 normal deviates*, Santa Monica California, U.S.A.

RAWLINGS, J. O.; PANTULA S. G.; DICKEY D. A.; (1998) *Applied Regression Analysis: A Research Tool*, Segunda Edición Springer – Verlag, U.S.A., New York. Disponible en: *http://web.nchu.edu.tw/~numerical/course1012/ra/Applied_Regression_Analysis_A_Research_Tool.pdf.*

RÍOS, D., RÍOS, S., JIMÉNEZ, J. M., JIMÉNEZ, A; (2009) *Simulación Métodos y Aplicaciones*, 2ª Edición, Alfaomega Ra-Ma, España.

RUSSELL, B. (1949) *The scientific outlook*, Segunda Edición, U.K.

SÁNCHEZ, V. F.; (2013) *Notas de clase, Tema: Pruebas de hipótesis,* Facultad de Ciencias, UNAM.

SÁNCHEZ, V. F.; (2015) *Notas de clase, Tema: Análisis de regresión simple*, Facultad de Ciencias, UNAM.

SHARMA, R.; SHANDIL, R., G.; KAPOOR, G.; (2011) *A Note on Karl Pearson's Coefficient of Dispersion, Himachal Pradesh University Journal*, India. *http://www.hpuniv.nic.in/Journal/Jul_2011_R%20Sharma%20Shandil%20and%20Kapoor.pdf.*

TODHUNTER, I.; (1965) History of the theory of probability, Chelsea.

YAMANE, T.; (1979) *Estadística*, Traducción por la Dra. Nuria Cortado de Kohan y Nicolas Civetta Arzayús Tercera Edición, Editorial Harla, México.

Recursos WEB

https://es.wikipedia.org/
http://genetics.thetech.org/
https://chanchullopolitico.wordpress.com/
http://www.inegi.org.mx/
http://www3.inegi.org.mx/sistemas/inp/preciospromedio/
http://www.jstor.org/
http://www.rand.org/
http://www.statcan.gc.ca/eng/start
https://www.mathworks.com/help/exlink/index.html
https://www.microsoft.com/es-mx/
http://support.sas.com/software/products/addin/
https://www.r-project.org/
https://www.wolfram.com/products/applications/excel_link/
Juegos Gerenciales: http://tepper.cmu.edu/prospective-students/course-page/45990/management-game