35

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

" MATERIAL DE APOYO PARA LA MATERIA ESTADISTICA II DEL BACHILLERATO DEL CCH DE LA UNAM "

T E S I S

QUE PARA OBTENER EL TITULO DE

MATEMATICO

PRESENTA

TIMOTEO PEREZ DOMINGUEZ

FACULTAD DE CIENCIAS
UNAM

Director de Tesis:
Mat. José Luis Castrejón Caballero

2002

DIVISION DE ESTUDIOS PROFESIONALES

FACULTAD DE CIENCIAS
SECCION ESCOLAR

M. EN C. ELENA DE OTEYZA DE OTEYZA
Jefa de la División de Estudios Profesionales de la
Facultad de Ciencias
Presente

Comunicamos a usted que hemos revisado el trabajo escrito:

"Material de apoyo para la materia ESTADISTICA II del Bachillerato del CCH de la UNAM"

realizado por Timoteo Pérez Domínguez

con número de cuenta 8II3I57-6 , quién cubrió los créditos de la carrera de Matemático.

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis Propietario Mat. José Luis Castrejón Caballero

Propietario Dr. Luis Antonio Rincón Solís

Propietario Mat. Laura Pastrana Ramírez Laura Pastrana R.

Suplente Mat. Marco Antonio Esquivel Pichardo

Suplente Mat. José Federico Olvera Rincón

Consejo Departamental de Matemáticas

M en C Alejandro Bravo Mojica

A los profesores de México
Por esa tan generosa lucha diaria contra la ignorancia

A mis padres y hermanos
Por su ininterrumpido y desinteresado apoyo

A la memoria de Pin

CONTENIDO

PREFACIO

Este trabajo se inspiró en la falta de material para impartir la materia "ESTADÍSTICA II" en el bachillerato del CCH de la UNAM.

Se intentó amenidad y sencillez en la exposición de los temas para que el lector disfrute la obra.

De la experiencia se sabe que el profesor necesitará ejercicios para exponer, ejercicios para dejar de tarea, ejercicios para exámenes parciales, finales, extraordinarios, etc. A ello se debe la abundancia de ejercicios para que el lector resuelva, aunque casi todos tienen respuesta en la obra misma.

Agradezco al director de tesis y a los sinodales el tiempo, la paciencia, dedicados para mejorar este trabajo.

Timoteo Pérez Domínguez

INTRODUCCIÓN

Cuando nos disponemos a aprender algo nuevo debemos estar conscientes de que habremos de realizar un esfuerzo intelectual. Con esto en mente y sabiendo que las matemáticas que los alumnos necesitan como son la aritmética y el álgebra, para asimilar este curso ya las tienen, seguramente se dedicarán a su estudio con más ánimo.

Aquí presento los temas de la materia ESTADÍSTICA II del bachillerato del CCH de la UNAM en la forma que se pensó más sencilla posible y que la materia permite. En el primer capítulo se revisan conceptos que aunque no forman parte del programa, sí son necesarios para la inferencia estadística, tema de éste segundo semestre de PROBABILIDAD Y ESTADÍSTICA.

En el capítulo 2 se presentan las distribuciones maestrales. Este capítulo prepara al lector para la inferencia estadística.

El capítulo 3 se refiere ya a la estadística inferencial; concretamente, se refiere a la estimación.

El capítulo 4 se refiere a pruebas de hipótesis y el quinto y último capítulo trata de la predicción.

Espero que el lector disfrute de ésta no solo bonita sino también muy útil materia.

Timoteo Pérez Domínguez

CAPITULO 1. CONCEPTOS PRELIMINARES.

SONRÍE:

1) Consejo paterno:
   - Hijo mío, si quieres que los demás te quieran, que te estimen mucho, házles creer que son capaces de pensar. Pero si quieres que te odien, pónlos a pensar.

2) Un sabio entomólogo, de ésos especializados en insectos, está investigando en su laboratorio con una pulga, obligándola a brincar por medio de unas voces de mando :

   - ¡Brinca, pulga, brinca!

El sabio registra en su libreta que con todas las patas la pulga brinca un metro. Le arranca una pata, le da la orden y el insecto brinca cincuenta centímetros. Lo anota. Le arranca otra pata, y ahora solo brinca veinte centímetros. Todo lo va anotando. Por fin le arranca la última pata y le ordena:

- ¡Brinca, pulga, brinca!

La pulga no se mueve. El investigador insiste:

- ¡Brinca, brinca!

Después de varios intentos inútiles, el sabio anota en su libreta:

" Pulga sin ninguna pata se vuelve sorda"

CAPITULO 1. CONCEPTOS PRELIMINARES.

En este capítulo vas a repasar algunos conceptos que seguramente ya dominaste en tu primer curso de PROBABILIDAD Y ESTADISTICA (ESTADISTICA I) para facilitarte la asimilación del contenido temático de éste segundo curso.

Qué mejor que empezar con una exposición de lo que es, en general, la estadística. Es lo que haremos en la sección ...

1.1 Panorama general de la estadística.

Cierto día, un profesor del CONALEP me preguntó ¿ qué es la estadística? y " que se me lengua la traba" porque no supe si decirle que era una rama de las matemáticas (un montón de teoremas deducidos a partir de otros teoremas, o tal vez de axiomas, mediante las leyes de la lógica) o decirle que era " un proceso lógicamente inductivo en el que se tomaba una muestra de una población para con aquella hacer inferencias acerca de la segunda", o mejor decirle los pasos que sigue un profesional de la estadística en su trabajo : " la selección de una muestra de los datos –aleatoriamente- de una población, su organización en tablas y gráficos para analizarlos y finalmente hacer conclusiones acerca de la población". No recuerdo mi respuesta de aquella ocasión pero al respecto te digo aquí que la estadística es todo lo anterior. A lo largo de tus dos cursos de ésta materia se te presentan algunos de los aspectos que mencioné excepto la deducción de los teoremas por razones del nivel (medio superior) y por razones didácticas. Veamos lo que , referente a la pregunta que nos ocupa y a otros conceptos relacionados, nos dicen los trabajadores intelectuales de la UPN :

" ... Veremos que la estadística consta de tres partes y discutiremos con algún detalle el funcionamiento de cada una. También daremos algunas definiciones fundamentales.

Recordemos que en el primer tema vimos que la estadística es una herramienta que nos ayuda en las siguientes etapas :

1) *La planeación de la búsqueda y la obtención de la información.*
2) *La sistematización y la organización de la información, para describirla y analizarla con facilidad.*
3) *La inferencia sobre la realidad a partir de la información obtenida, mediante estimaciones y contrastación de hipótesis.*

Para cada una de estas etapas existe una parte de la estadística que proporciona los métodos que ayudan a resolver los problemas correspondientes. Con el fin de aclarar en qué consisten estas tres partes, introduciremos dos conceptos fundamentales en la estadística : el de población y el de muestra.

En el lenguaje cotidiano la palabra "población" designa a un conjunto de personas. El concepto estadístico de población está ligado históricamente a este sentido de la palabra, pero

se ha pasado por varias etapas de abstracción para llegar a la definición actual de población estadística.

Por una parte, un mismo conjunto de personas, objetos o entidades puede ser estudiado desde varios puntos de vista, y es posible medir en ellos características distintas. Así, para evaluar un conjunto de libros de texto nos pueden interesar las características " grado de relación entre los ejemplos y el contenido de cada libro", "grado de amenidad de cada libro", "adecuación del formato al uso de cada libro", etc. Es necesario, pues, aclarar a cuál o cuáles de las posibles características nos estamos refiriendo cada vez. Como dijimos anteriormente, expresamos la medición de cada característica a través de una variable. Es entonces la variable asociada a cada característica lo que nos permitirá definir el concepto de población estadística.

Si una variable está asociada a una característica, entonces llamamos ***población*** *a la colección de los* ***valores*** *que toma la variable, es decir a la colección de los resultados de las mediciones en todos los individuos, objetos o entidades en los que interese medir la característica.*

Veamos un ejemplo: supongamos que nos interesa conocer la característica "número de hijos de los alumnos de la UPN". La población asociada a la variable ***número de hijos*** *es la colección de valores que se obtiene al ver cuántos hijos tiene cada uno de los alumnos de la UPN. Así, si el alumno José Domínguez tiene ocho hijos (o sea que el resultado de la medición en José Domínguez es ocho), y el alumno Arturo López tiene también ocho hijos, aparecerá el número ocho cuando menos dos veces en la población.*

Es pertinente hacer algunas observaciones sobre el concepto de población :

1) *Una población comprende tantas repeticiones de un mismo valor como individuos, objetos o entidades lo tengan asociado en la medición.*
2) *Cuando se habla de una población, es necesario aclarar cuál es el conjunto de personas, objetos o entidades en los que interesa medir la característica. Por ejemplo, consideremos la característica "estatura": Las poblaciones asociadas a esta característica serán colecciones de valores como 1.58, 1.61, 1.53, etc. Y son muchas las poblaciones asociadas a la estatura: El conjunto de todos los mexicanos da origen a una población estadística, mientras que el conjunto de los mexicanos hombres da origen a otra y el conjunto de los mexicanos hombres mayores de treinta años da origen a otra población estadística más, distinta de las anteriores.*
3) *Cuando es claro cuál es la característica de la que se está tratando, suele hablarse de población como del conjunto de individuos, objetos o entidades en los que interesa medir dicha característica. Así, la expresión "Población Económicamente Activa" en el artículo citado en el tema uno se refiere a un conjunto de 23 millones de personas en los que interesa medir la característica "nivel de empleo".*
4) *Es común que en el estudio de un fenómeno interese medir más de una característica en un mismo conjunto de personas, objetos o entidades, o sea que se tengan los valores de más de una variable. En éstos casos, la población es la colección de todas las combinaciones que se pueden formar con un valor de cada variable. Por ejemplo, si en el estudio sobre el fenómeno del desempleo en Francia se hubiera considerado también el sexo de los integrantes de la " Población Económicamente Activa", la*

población estadística estaría formada por una sucesión de las siguientes combinaciones de valores : "Hombre empleado", "mujer empleada", "Hombre desempleado", "Mujer desempleada"

La información que nos interesa conocer para el manejo estadístico de un problema está contenida en la población, puesto que la población incluye todos los valores que toma la variable en el conjunto de individuos, objetos o entidades sobre los que queremos obtener conclusiones. Este conjunto puede ser tan grande como el formado por todos los habitantes de un País (o aún más grande), y tan pequeño como el formado por los alumnos de un grupo escolar (o aún más pequeño). Por ello, la obtención de los datos correspondientes a todos los individuos, objetos o entidades puede ser un proceso más o menos difícil y costoso. Un censo, por ejemplo, significa un esfuerzo que ningún País puede hacer con mucha frecuencia. Cuando éste proceso de recopilación de la información es difícil y costoso, o cuando es imposible, resulta conveniente seleccionar algunos de los individuos, objetos o entidades y medir en ellos la característica de interés. Obtenemos así solo algunos de los datos que conforman la población, es decir, una muestra de la población, que nos puede dar una idea de los datos de toda la población.

Una muestra es una subcolección de una población, constituida por los valores que toma la variable en algunos de los individuos, objetos o entidades en los que interesa medir la característica.

Por ejemplo, supongamos que nos interesa conocer la ***distancia*** *que recorre cada niño del país para ir de su casa a la escuela. La población asociada a esta variable es muy difícil de obtener. Tendríamos que recorrer todo el país para preguntar a cada niño cuántos kilómetros hay entre su casa y la escuela a la que asiste. Podemos entonces seleccionar a algunos niños y preguntarles solo a ellos. La muestra obtenida es una colección de valores como 10.5, 1, 3.7, etc. Si los niños son seleccionados adecuadamente, la muestra nos puede dar una idea de los datos de toda la población.*

El concepto de ***muestra*** *es tan importante en la estadística como el de población. Las siguientes observaciones son importantes :*

1) *Una muestra incluye, como una población, tantas repeticiones de un mismo valor de la variable como individuos, objetos o entidades seleccionados lo tengan asociado en la medición.*
2) *Cuando se habla de una muestra, es necesario aclarar de qué población es subcolección, es decir, de qué población fue extraída.*
3) *Es frecuente que se hable de una muestra como del subconjunto de individuos, objetos o entidades en los que se mide la característica. En el ejemplo sobre la* ***distancia*** *que recorre cada niño para ir de su casa a la escuela, la muestra obtenida es una colección de números, pero podemos también referirnos a los niños a quienes preguntamos qué distancia hay entre su casa y la escuela diciendo que es la muestra de niños.*

---

* En este caso se dice que la población es bivariada. Si están involucradas más de dos variables se habla de una población multivariada

4) *Cuando en el estudio de un fenómeno se consideran dos o más características, una muestra es, como la población de la que proviene, una colección de combinaciones de los valores que toman las variables.*
5) *El número de datos que conforman una muestra se llama tamaño de la muestra y se simboliza en general por la letra* ***n****.*
6) *Aunque, por definición, cualquier subcolección de la población es una muestra, no cualquier muestra es adecuada para fines estadísticos. Es necesario que la muestra con la que se trabaja reproduzca, en la medida de lo posible, los rasgos generales de la población. Supongamos por ejemplo que en el estudio de la característica "distancia entre la casa y la escuela" consideramos una muestra que incluya únicamente niños de las colonias proletarias de las ciudades del país. Esta muestra nos conduciría sin duda a grandes errores en la apreciación de los valores de la población, puesto que no quedarían representados otros grupos como los pertenecientes a niños de regiones desérticas. Llegaríamos a menos errores con una muestra elegida de tal manera que sea muy probable que ésta represente a toda la población. Nos referiremos a estas muestras con el nombre de muestras representativas ...*

... Los conceptos de población y de muestra nos permiten considerar nuevamente las tres partes de la estadística que hemos mencionado :

La primera es la que consiste en la planeación de la búsqueda y la obtención de la información; en particular, esta parte incluye métodos que permiten diseñar un esquema para la selección de una muestra representativa de la población.

Otra parte de la estadística es la que se encarga de sistematizar y organizar la información contenida en una muestra o en una población, es decir los valores de la variable. Esta parte incluye métodos que permiten describir y analizar la información.

La tercera parte que hemos mencionado comprende métodos estadísticos que permiten inferir, a partir de la información contenida en una muestra, cuáles pueden ser los rasgos principales de los valores de la población de la que proviene la muestra.

La primera de éstas tres partes está estrechamente ligada a las otras dos, ya que el buen éxito de éstas depende de una planeación adecuada. El esquema de búsqueda y obtención de la información debe permitir que se realicen, en los términos deseados, tanto la sistematización como las inferencias planteadas. Para aclarar esto hablaremos primero de las dos últimas partes, y finalizaremos el tema con algunas consideraciones sobre la primera.

La parte de la estadística que proporciona los métodos para sistematizar y describir la información contenida en una muestra o en una población recibe el nombre de ***estadística descriptiva****. Los métodos de la estadística descriptiva permiten resumir los aspectos principales de los valores de una muestra o de una población, tanto gráfica como numéricamente. Cuando el fenómeno bajo estudio involucra dos o más variables, algunos métodos de la estadística descriptiva permiten saber de qué tipo y de qué magnitud es la relación entre ellas.*

La parte de la estadística que permite inferir los rasgos principales de los valores de poblaciones, a partir de los valores de muestras extraídas de ellas, se llama ***estadística inferencial****. Ciertos métodos de la estadística inferencial permiten obtener estimaciones de algunos valores de interés de una población (por ejemplo, con los datos de una muestra se*

algunos valores de interés de una población (por ejemplo, con los datos de una muestra se puede obtener una aproximación o estimación del promedio de todos los valores de la población). Otros métodos de la estadística inferencial permiten contrastar hipótesis; es decir, dada una conjetura que se plantea sobre una o más poblaciones, permiten ver si la información contenida en la o las muestras es compatible o no con la conjetura. Por ejemplo, se puede plantear la hipótesis de que los promedios de las estaturas de dos poblaciones (digamos una de hombres y otra de mujeres) son distintos; entonces, la estadística inferencial permite decir, mediante la comparación de los promedios obtenidos con los datos de las muestras, si la información obtenida corrobora o contradice la hipótesis.

Vale la pena que nos detengamos a hacer algunas consideraciones sobre la estadística inferencial.

Cuando hacemos inferencias sobre la población a partir de la información contenida en una muestra extraída de la población, estamos obteniendo ciertas conclusiones sobre la realidad a pesar de que solo conocemos una porción de ella. Como dijimos anteriormente, el proceso de selección de los individuos, objetos o entidades que se estudiarán debe estar bien diseñado, para que la muestra sea representativa de la población. Pero por muy representativa de la población que sea la muestra, no deja de ser una subcolección de la población, una de las muchas muestras posibles que son igualmente representativas. Es probable que si se hubiera obtenido otra muestra, los valores de ésta habrían sido distintos. Si ambas muestras son representativas de la población, los valores de cada una nos dan una idea acerca de los valores de la población, pero no nos informan totalmente acerca de ellos.

Nuestro punto de partida es, en general, una incertidumbre casi total sobre la población que nos interesa. Una muestra de ella nos permite obtener un cierto grado de certeza sobre algunos de sus aspectos, pero la certeza no podrá ser nunca total (a menos, desde luego, que la muestra fuera toda la población). Esto es, la confianza que tengamos al hacer afirmaciones sobre la población no podrá ser nunca total, puesto que habrá siempre un margen de incertidumbre sobre la población. Esto implica que al hacer inferencias sobre la población hay un riesgo de cometer algún error.

Cuando se tiene una muestra representativa de una población, los métodos de la estadística inferencial permiten hacer ciertas afirmaciones acerca de la población con la máxima confianza posible, es decir, con el mínimo riesgo posible de cometer error. La probabilidad ocupa un lugar importante en éste proceso, ya que aporta un lenguaje que permite expresar en forma cuantificable tanto la confianza que podemos tener en una inferencia estadística como el riesgo de cometer un error en esa inferencia.

Supongamos que la inferencia que se desea hacer es una estimación del valor del promedio de la población formada por los tiempos que requieren para resolver cierto examen los alumnos de secundaria, a partir de una muestra representativa de la población. Entonces, los métodos estadísticos de inferencia proporcionarán la estimación deseada en términos parecidos a los siguientes : " con 95 % de confianza, se puede afirmar que el promedio de la población está entre 80 y 100 minutos". Observe que en la vida cotidiana hacemos con frecuencia inferencias parecidas. En éste caso diríamos tal vez algo como " creo que el promedio es de 90 minutos aproximadamente". Lo que permiten los métodos estadísticos es precisar las expresiones "creo que" y "alrrededor de". Observe también que ambas estimaciones están expresadas con la cautela con la que por fuerza nos movemos cuando

estamos en un ámbito de incertidumbre. Sin embargo, la estimación estadística proporciona mucha más información : nos dice con qué confianza podemos hacer la afirmación y entre cuáles dos valores puede estar el promedio de la población.

Hemos señalado que la estadística inferencial hace posible no solo hacer estimaciones sino también probar hipótesis. Esto significa poner a prueba conjeturas que se hacen acerca del fenómeno bajo estudio, es decir contrastar los hechos que se deducen de ellas con los hechos observados. Los métodos estadísticos de prueba de hipótesis permiten decidir si la información contenida en una muestra contradice o corrobora una hipótesis planteada sobre la población.

Para aclarar este punto veamos un ejemplo: Supongamos que se plantea la hipótesis de que cierto método nuevo de enseñanza de la notación musical es más efectivo en cierta población de alumnos que el utilizado tradicionalmente, y que se argumenta esto diciendo que el tiempo promedio que tardan los alumnos en aprender los elementos básicos del uso del pentagrama y las notas es menor con el método nuevo que el que tardaban con el tradicional. Para probar la hipótesis, se toman dos muestras representativas; a una se le enseña la notación musical con el método nuevo (muestra N), y a la otra con el método tradicional (muestra T). Si el tiempo promedio requerido por los alumnos de la muestra N para aprender el tema es mayor o igual que el requerido por los alumnos de la muestra T, resulta obvio que la información contenida en las muestras contradice la hipótesis planteada, y por lo tanto no se puede afirmar que el nuevo método sea más efectivo que el tradicional. Por otra parte, si el tiempo promedio es mucho menor en la muestra N que en la muestra T, la información contenida en las muestras corrobora la hipótesis planteada. Los métodos estadísticos de inferencia permiten decidir, en términos probabilísticos si la información contenida en las muestras corrobora la hipótesis en grado suficiente para considerar, con cierta confianza, que los alumnos requieren en promedio un tiempo menor con el método nuevo que con el tradicional.

Así, la estadística inferencial posibilita el hacer inferencias sobre la población a partir de la información contenida en la muestra, mediante estimaciones de los valores de la población o mediante la prueba de hipótesis planteadas acerca de la población.

Recapitulemos brevemente : los métodos de la estadística descriptiva permiten ordenar la información contenida en una muestra (o una población) y resumir sus aspectos principales, y los de la estadística inferencial permiten inferir ciertas características de la población a partir de la información contenida en una muestra extraída de ella.

Por último, haremos algunas consideraciones sobre la primera parte de la estadística, que consiste, como ya lo hemos señalado, en la planeación de la búsqueda y la obtención de la información. La exposición general que hemos hecho sobre la estadística descriptiva y la estadística inferencial hace patente la necesidad de que la muestra sobre la que se hace el estudio sea representativa de la población sobre la que interesa obtener conclusiones. Esto solo puede garantizarse mediante una adecuada planeación.

La estadística contiene dos ramas que auxilian en ésta planeación : el muestreo y el diseño experimental.

El muestreo aporta métodos que permiten diseñar un esquema de la recolección de la información, es decir que permiten diseñar un proceso adecuado para obtener una muestra de la población de interés que sea representativa de ella y de la que se obtenga la máxima

información con el mínimo tamaño de muestra o con el mínimo costo. El muestreo adquiere particular importancia cuando los objetivos del estudio incluyen inferencias sobre la población. Si una muestra no es representativa de la población de la que fué extraída, se corre un riesgo muy grande de cometer errores en las inferencias. El muestreo también indica qué tamaño debe tener la muestra para poder hacer las inferencias con la confiabilidad deseada : una muestra demasiado pequeña no nos daría mucha información acerca de la población, ya que tendría pocas oportunidades de reproducir toda la variabilidad de la población.

El diseño experimental, por otra parte, permite planear experimentos en forma óptima y tomando en cuenta las condiciones reales en que se efectuarán los experimentos, condiciones que suelen imponer restricciones de tiempo, de espacio, de recursos, de ética, etc. Por ejemplo, supongamos que tres profesores de español de una secundaria desean hacer una investigación para determinar qué libro de texto de español es el más adecuado para el tipo de alumnos de la escuela. Para ello, deciden que en una primera etapa analizarán tres textos distintos en los seis grupos de primer grado que les toca atender. Entonces, el diseño experimental indicará, dadas las condiciones reales, cómo se puede efectuar la investigación, de tal modo que se eviten errores. Un error consistiría, por ejemplo, en que cada texto solo fuera utilizado por un profesor. Un diseño así no permitiría ver si las eventuales diferencias detectadas se deben a una diferencia entre los textos o a una diferencia entre los profesores.

Es claro que cuando se desea hacer una investigación que involucrará tratamiento estadístico de la información, debe planearse con cuidado cómo se va a recolectar ésta. Es frecuente que al hablar de estadística se enfatice la importancia de los niveles descriptivo e inferencial de la disciplina, pero es un error pasar por alto la importancia de la etapa de planeación ¡ cuántas fallas podrían remediarse si se recordara siempre que la estadística empieza ***antes*** *de la obtención de datos!"*

1.2 Variables aleatorias.

Existen en el universo, tanto en el macromundo – el mundo de lo grande, incluído astronomía - como en el micromundo – el mundo de lo pequeño, incluído el átomo - infinidad de magnitudes variables (o simplemente variables) y de ellas te han hablado tus profesores de otras materias: que si la velocidad de un móvil, o su posición, que si el ingreso per cápita, etc. De todas las variables, las que interesan en probabilidad y estadística son las que dependen del azar, como sería, por ejemplo, lanzar volados con una moneda legal; el resultado es variable, unas veces aguila, otras veces sol, dependiendo del azar. Pero en vez de hablar de águila o sol para los resultados posibles, podríamos hablar de 1,2, conviniendo en que águila es 1 y sol es 2 o viceversa, y así iniciaríamos la matematización del sencillo experimento aleatorio "lanzamiento de un volado".

---

* Universidad Pedagógica Nacional. (1983). ¿Cómo funciona la estadística? En Introducción a los métodos estadísticos. Vol. 1 Sistema de educación a distancia. México : SEP p.p. 17-23.

El hombre está rodeado de fenómenos aleatorios, pero es inteligente y sabe que si los matematiza podrá comprenderlos mejor y resolverá con más eficacia los problemas que le interese resolver en dichos fenómenos.

En estadística hay varios tipos de variables, para los propósitos de éste trabajo se hablará solo de las cuantitativas.

Las variables aleatorias, ya traducidas a las matemáticas, se han clasificado en **discretas** y **continuas.**

En efecto, una variable aleatoria se dice que es discreta si solo toma valores numerables y es continua cuando toma un conjunto contínuo de valores; o en otras palabras, es discreta si se cuenta y es continua si se mide.

De lo dicho hasta el momento en éste párrafo, debe entenderse que el nombre de variable aleatoria continua, se aplica a variables tales como: longitud, peso, temperatura y tiempo, que se pueden considerar capaces de tomar cualquier valor dentro de un intervalo de valores. Así pues, el peso de un estudiante en la gama 70-75 kg puede considerarse capaz de tomar cualquier valor dentro de ésta gama. Variables tales como el número de accidentes automovilísticos en un día, número de insectos que mueren al ser rociados con insecticida o número de niños en una familia son ejemplos de lo que se conoce como una variable aleatoria discreta. Consideremos las variables discretas como variables cuyos valores posibles son enteros: luego, son contables más que mensurables.

Puesto que cualquier aparato de medición es de exactitud limitada, las mediciones en la vida real son en realidad discretas en su naturaleza más bien que continuas; sin embargo, esto no debe ser obstáculo para pensar en éstas variables como contínuas. Aún cuando, por ejemplo, los pesos de quienes duermen en algún dormitorio se hayan registrado al kilogramo más cercano, deben considerarse como valores de una variable continua que se ha redondeado al entero más próximo. Cuando un peso se registra como, por ejemplo 51 kg se supone que el peso real puede encontrarse entre 50.5 y 51.5 kg.

1.2.1 Variables aleatorias discretas.

Una variable aleatoria X es una función cuyo dominio son los elementos del espacio muestral, Ω , y cuyo contradominio es el conjunto de números reales $\mathbf{R}$.

$$X : \Omega \longrightarrow R$$
$$w \mapsto X(w)$$

Veamos un ejemplo de variable aleatoria discreta, desde el fenómeno aleatorio hasta su traducción a matemáticas.

EJEMPLO 1. En algún punto del tiempo, la selección mexicana de fútbol llega a la gran final de la copa del mundo, pero por desgracia debido a empate llegaron hasta penalties. Hugo, Luis, Carlos, David y "Matador" son los encargados de tirar la primera tanda de cinco tiros ¿cuántos goles podrían anotar entre los cinco? ¿con qué probabilidades?

SOLUCION:

Un resultado posible de éste fenómeno aleatorio es que ninguno anotara : denotemos este resultado con (n,n,n,n,n) conviniendo en que n = no anota y a = anota. También convengamos en que la primer componente de la quíntupla ordenada corresponde a Hugo, la segunda a Luis, La tercera a Carlos, la cuarta a David y la quinta a "Matador".

Con las mismas convenciones anteriores, otro posible resultado es (n,n,n,n,a), que indica que todos fallaron su tiro excepto Matador.

Y ya llevamos dos elementos (dos resultados posibles) del espacio muestra .

$$\Omega = \{(n,n,n,n,n),(n,n,n,n,a), \dots, \}$$

Introduzcamos aquí la variable aleatoria X = Número de penalties anotados entre los cinco tiradores.

Entonces X vale 0 en (n,n,n,n,n); es decir X[(n,n,n,n,n)] = 0.

Análogamente, X[(n,n,n,n,a)] = 1.

Pero, ¿cuántos elementos tiene el espacio muestra Ω de éste fenómeno aleatorio?. Si recuerdas tu cálculo combinatorio, hay $2^5 = 32$ resultados posibles. Estos aparecen a continuación, en columna, junto con los correspondientes valores de X

$$X : \Omega \_\_\_\_\_ R$$

$$w \quad \mapsto \quad X(w)$$

H	L	C	D	M		X
n	n	n	n	n	-------	0
n	n	n	n	a	------	1
n	n	n	a	n	------	1
n	n	n	a	a	------	2
n	n	a	n	n	------	1
n	n	a	n	a	------	2
n	n	a	a	n	------	2
n	n	a	a	a	------	3

H	L	C	D	M		X
a	n	n	n	n	-------	1
a	n	n	n	a	-------	2
a	n	n	a	n	-------	2
a	n	n	a	a	-------	3
a	n	a	n	n	-------	2
a	n	a	n	a	-------	3
a	n	a	a	n	-------	3
a	n	a	a	a	-------	4

n	a	n	n	n	------	1		a	a	n	n	n	--------	2
n	a	n	n	a	------	2		a	a	n	n	a	--------	3
n	a	n	a	n	------	2		a	a	n	a	n	--------	3
n	a	n	a	a	------	3		a	a	n	a	a	--------	4
n	a	a	n	n	------	2		a	a	a	n	n	--------	3
n	a	a	n	a	------	3		a	a	a	n	a	--------	4
n	a	a	a	n	------	3		a	a	a	a	n	--------	4
n	a	a	a	a	------	4		a	a	a	a	a	--------	5

A la imagen de ésta función se le llama " Recorrido de la variable aleatoria". En éste ejemplo, el recorrido de la variable aleatoria X es el conjunto $R_x = \{0,1,2,3,4,5\}$. Se llama "Distribución de probabilidad" a la función $f_x(x)$ cuyo dominio es el recorrido de la variable aleatoria y cuya imagen son las probabilidades asociadas al recorrido :

$$f_x : R_x \longrightarrow [0,1]$$
$$x \mapsto f_x(x) = P(X = x)$$

O sea, tenemos la composición de funciones $f_x \circ X$ siguiente :

$$f_x \circ X : \Omega \longrightarrow [0,1]$$
$$w \mapsto (f_x \circ X)(w) = f(X(w))$$

En nuestro ejemplo, el espacio muestra Ω tiene 32 elementos y asignamos de modo subjetivo* las probabilidades siguientes :

$$f : \{0,1,2,3,4,5\} \longrightarrow [0,1]$$
$$0 \mapsto f(0) = P[X = 0] = 1/32$$
$$1 \mapsto f(1) = P[X = 1] = 5/32$$
$$2 \mapsto f(2) = P[X = 2] = 10/32$$
$$3 \mapsto f(3) = P[X = 3] = 10/32$$
$$4 \mapsto f(4) = P[X = 4] = 5/32$$
$$5 \mapsto f(5) = P[X = 5] = 1/32$$

Podemos compactar lo último anterior en la siguiente tabla

X	0	1	2	3	4	5
f (x)=P(X=x)	1/32	5/32	10/32	10/32	5/32	1/32

* Recuérdese que la probabilidad subjetiva es la creencia personal acerca de la ocurrencia de un evento.

Debe cumplirse que

$$\sum_x f(x) = 1$$

y en efecto : 1/32 + 5/32 + 10/32 + 10/32 + 5/32 + 1/32 = 1

Con lo que hemos hecho hasta aquí, contestamos las preguntas del ejemplo : entre los cinco jugadores podrían anotar entre 0 y 5 goles y las probabilidades están dadas en la última tabla escrita arriba.

Pero sigamos adelante, repasando :

La gráfica de la función f(x), de la "distribución de probabilidades", es :

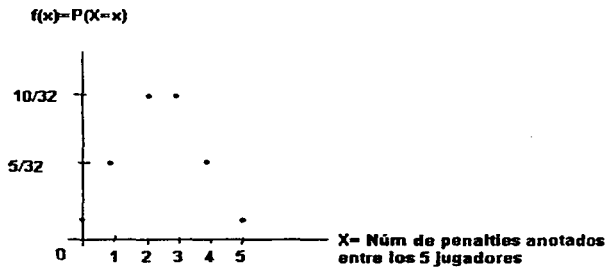


Fig. 1.1 función de densidad correspondiente al ejemplo1.

Por cierto, a f(x) también le llaman "función de densidad".

Otra función importante es la "función de distribución acumulada" o simplemente "función de distribución", $F_X(x)$, que se refiere a las probabilidades acumuladas; es decir

$F_X(x) = P(X \leq x)$

Así, en nuestro ejemplo y suponiendo independencia entre los tiros de penalty :

$F_X(0) = P(X \leq 0) = P(X=0) = 1/32$

$F_X(1) = P(X \leq 1) = P(X=0) + P(X=1) = 1/32 + 5/32 = 6/32$

$F_X(2) = P(X \leq 2) = P(X=0) + P(X=1) + P(X=2) = 1/32 + 5/32 + 10/32 = 1/2$

$F_X(3) = P(X \leq 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3) =$
$= 1/32 + 5/32 + 10/32 + 10/32 = 26/3$

$F_X(4) = P(X \leq 4) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) =$
$= 1/32 + 5/32 + 10/32 + 10/32 + 5/32 = 31/32$

$F_X(5) = P(X \leq 5) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) + P(X=5) =$
$= 1/32 + 5/32 + 10/32 + 10/32 + 5/32 + 1/32 = 32/32 = 1$

La gráfica de $F_X(x)$

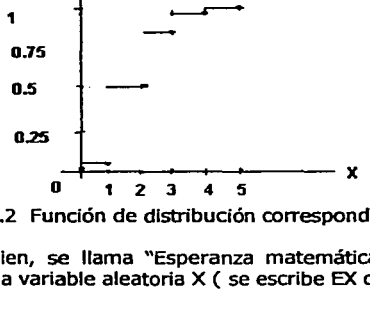


Fig. 1.2 Función de distribución correspondiente al ejemplo 1.

Ahora bien, se llama "Esperanza matemática" o simplemente "Esperanza" o "Valor esperado" de una variable aleatoria X (se escribe EX o E(X)) al resultado de la expresión

$$E(X) = \sum_x x \cdot f(x)$$

La esperanza es el promedio "a la larga" de la variable aleatoria.
En nuestro ejemplo :

E(X) = 0 (1/32) + 1 (5/32) + 2 (10/32) + 3 (10/32) + 4 (5/32) + 5 (1/32) = 2.5

Se define "Varianza" de una variable aleatoria X (se escribe V(X)) al resultado de V(X) $= E(X-EX)^2$. Pero como $E(X-EX)^2 = E\{X^2 - 2XEX + E^2X\} =$
$E(X^2) - E^2(X)$, tenemos el

TEOREMA : V(X) = $E(X^2) - E^2(X)$.
En la práctica no se usa la definición sino el teorema ya que éste último facilita más los cálculos.

En nuestro ejemplo, tenemos que, como

$$Ex^2 = \sum_x x^2 f(x)$$

X	0	1	2	3	4	5
X^2	0	1	4	9	16	25
$f(x) = P(X=x)$	1/32	5/32	10/32	10/32	5/32	1/32

$EX^2 = 0(1/32) + 1(5/32) + 4(10/32) + 9(10/32) + 16(5/32) + 25(1/32) = 240/32 = 7.5$
Y entonces

$$V(X) = E(X^2) - E^2(X) = 7.5 - 6.25 = 1.25$$

EJERCICIO 1 : Se lanzan dos dados distinguibles normales,es decir no cargados. Hallar la distribución de probabilidad de la suma de las caras superiores, la esperanza de dicha suma, la varianza, $f_X(x), F_X(x)$ con sus gráficas. ¿Cuál es la probabilidad P(X=10)? Y ¿Cuál es la probabilidad de que $P(X \leq 8)$? (ver resultado al final de la obra).

Ya se han desarrollado modelos matemáticos de variables aleatorias discretas determinando sus medias y sus varianzas para que nosotros muy cómodamente solo las utilicemos. Así, se habla de la variable aleatoria binomial, Poisson, etc. Para propósitos de éste curso ocupémonos solo de la binomial:

Distribución binomial.

Si X = número de éxitos en n ensayos independientes de Bernoulli donde la probabilidad de éxito es p y la probabilidad de fracaso es q=1-p, (P(E)=p, P(F)= q= 1-p) escribiremos

$X \equiv B(n, p)$

que se leerá " la variable aleatoria X tiene distribución binomial con parámetros n y p". De ésta distribución se sabe que la E(X) = np y la V(X) = npq. También se sabe que

$$P(X = k) = C_n^k p^k q^{n-k} \quad \text{y que} \quad P(r \leq X < s) = \sum_{i=r}^{s-1} C_n^i p^i q^{n-i}$$

EJEMPLO 2: Si el 20 % de fusibles son defectuosos y se compra una caja de 10 fusibles

a) ¿Cuál es la probabilidad de que exactamente 6 sean buenos?
b) ¿Cuál es la probabilidad de que por lo menos 7 sean buenos?
c) ¿Cuál es la probabilidad de que a lo más, 4 sean buenos?
d) ¿cuál es la probabilidad de que haya entre 6 y 9 buenos?

SOLUCIÓN :

Sea Éxito = fusible bueno, Fracaso = fusible defectuoso. P(F) = 0.2 porque el 20 por ciento de fusibles son defectuosos. Luego, la P(E) = 1 – 0.2 = 0.8. Sea X = número de éxitos en n = 10, entonces $X \cong B(10,0.8)$.

a) Con la fórmula de la binomial tenemos
$P(X = k) = C_n^k p^k q^{n-k}$. En nuestro caso:
$P(X = 6) = C_{10}^6 (0.8)^6 (0.2)^4 = (210)(0.262144)(0.0016) = 0.088080384$

Con el uso de las tablas procedemos como sigue:
Las tablas denotan nuestra k con r. El cuerpo de la tabla es la probabilidad de r o más éxitos, es decir, $P(X \geq r)$.
Si queremos saber cuál es P(X=6) tenemos que restar $P(X \geq 7)$ de $P(X \geq 6)$
Así, para n=10 y p=0.8 :
$P(X = 6) = P(X \geq 6) - P(X \geq 7) = 0.967 - 0.879 = 0.088$ que es la probabilidad pedida en este inciso.

b) $P(X \geq 7)$. Con la fórmula de la binomial, tenemos :
$P(X \geq 7) = P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10) =$
$\sum_{k=7}^{10} C_n^k p^k q^{n-k} = C_{10}^7 (0.8)^7 (0.2)^3 + C_{10}^8 (0.8)^8 (0.2)^2 + C_{10}^9 (0.8)^9 (0.2) +$
$C_{10}^{10} (0.8)^{10} (0.2)^0 = 0.879126118.$

Con el uso de la tabla, vemos que el renglón para r=7, n=10 y la columna p=0.8 obtenemos 0.879.

c) Con la fórmula es muy laborioso y como ya se describió el procedimiento con ella, pasamos al uso de la tabla :
$P(X \leq 4) = 1 - P(X \geq 5) = 1 - 0.994 = 0.006$

d) Con las tablas vemos que
$P(6 \leq X \leq 9) = P(X \geq 6) - P(X \geq 10) = 0.967 - 0.107 = 0.86$

EJERCICIO 2: Supóngase que se sabe que la probabilidad de recuperación de cierta enfermedad es de 0.4. Si 15 personas contraen la enfermedad ¿cuál es la probabilidad de que
a) 3 o más se recuperen?
b) 4 o más?
c) 5 o más?
d) Menos de 3? (ver resultado al final de ésta obra)

1.2.2 Distribución normal.

A continuación te presento una curva muy distinguida de la estadística, la "campana de Gauss" o "curva normal". Es la siguiente :

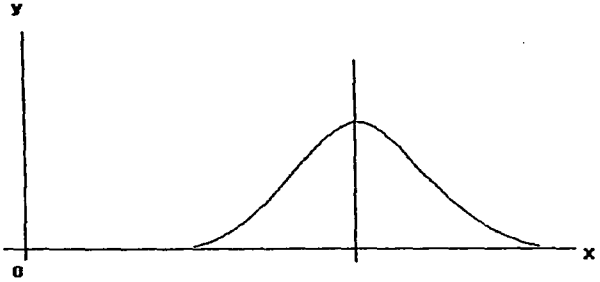


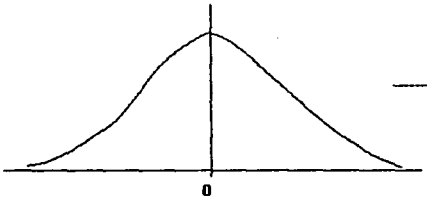
Fig. 1.3 Distribución normal. μ

con la siguiente fórmula

$$y = f(x) = \frac{1}{\sqrt{2\Pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\textit{donde} \quad x, \mu \in R; \sigma > 0$$

Cuando $\mu = 0$ y $\sigma = 1$, tenemos la "curva normal estándar" :



$$\text{----------(1)}$$

Fig. 1.4 Distribución normal estándar.

No te asustes, no necesitas memorizar las fórmulas, aunque tampoco se te prohibe que lo hagas.

Algunas propiedades de ésta curva son (sea estándar o no) :

1) Es simétrica respecto a la recta $x = \mu$.
2) Es cóncava hacia abajo en el intervalo $[\mu - \sigma, \mu + \sigma]$ y cóncava hacia arriba en el resto de la recta real.
3) Es asíntótica con respecto al eje x en ambos sentidos.

Además, para el caso de la normal estándar (I), el valor máximo es aproximadamente 0.4

Puedes percibir experimentalmente la presencia de esta curva "acampanada" en la naturaleza y en el trabajo experimental realizando las siguientes actividades :

1) Dobla un trozo de papel por la mitad, fijando un lado a una superficie plana y sosteniendo el otro lado con un libro formando un ángulo de aproximadamente 60 grados con la horizontal. Echa después sal con un embudo sobre la parte inclinada del papel hacia el doblez. El contorno de la pila de sal que toque el papel formará una clara y bien formada curva normal.

2) Toma un trozo de papel de aproximadamente 25 cm de largo y numera columnas de 1 a 10; coloca un punto grande en el medio del papel, aproximadamente en la columna seis. Alguien que te ayude, que deje caer arroz, un grano por vez y sobre el punto desde una altura de aproximadamente 30 cm. Cuenta, después, el número de granos de arroz caídos en cada columna y construye el histograma de frecuencias correspondiente. Aproxima el histograma por una curva suave y obtendrás una curva normal.

Ahora bien, ya que sabemos qué es la curva normal (función, gráfica, algunas propiedades) déjame decirte que cuando sabemos que una variable aleatoria X tiene distribución normal con media μ y varianza σ^2, lo escribiré así : $X \equiv N(\mu, \sigma^2)$. Y si se sabe que X tiene distribución aproximadamente normal, escribiré $X \approx N(\mu, \sigma^2)$. En cualquiera de los dos casos, la probabilidad de que X esté entre a y b es el área entre la curva, el eje x y las rectas x=a, x=b; es decir :

Cuando $X \equiv N(\mu, \sigma^2) \Rightarrow P(a < X \leq b) = A$

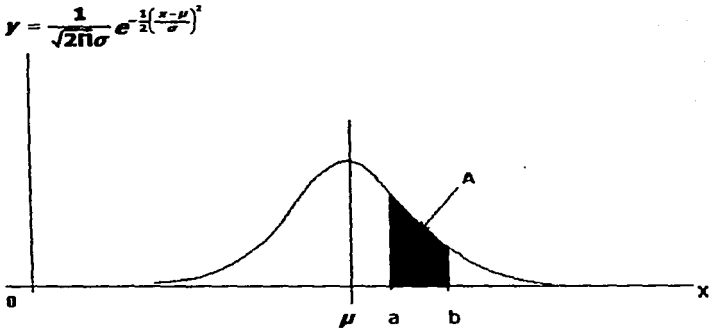


Fig. 1.5 El área sombreada bajo la curva, es una probabilidad.

Pero, ¿cuál es el valor numérico de dicha área? La respuesta involucra un concepto matemático que seguramente aún no dominas: "la integral definida". Debe bastarte saber que hay tablas con los valores de dicha área.

No encontrarás tablas del área bajo la curva para cualquier distribución normal sino solo de la "normal estándar" (la que tiene $\mu = 0$ y $\sigma = 1$). Así que si tu distribución normal no es N(0,1), si no es estándar, tendrías que realizar un proceso conocido como "estandarización" de tu variable aleatoria X del siguiente modo :

Si $X \cong N(\mu, \sigma^2)$, entonces la variable $Z = (x - \mu)/\sigma \cong N(0,1)$ se distribuye como normal estándar y trabajarías con ésta en las tablas.

En efecto, se sabe que

$$P(x_1 \leq X \leq x_2) = P(z_1 \leq Z \leq z_2), \quad \textit{donde} \quad z_i = \frac{x_i - \mu}{\sigma}$$

En la realidad ningún fenómeno se rige exactamente por una distribución normal ni siquiera en los casos en que decimos que hay distribución normal. Pero en dichos casos se puede tomar como modelo matemático a la distribución normal y trabajar sin problemas. La línea recta es un modelo que se adopta en infinidad de casos y se ajusta bien a nuestras necesidades a pesar de que tampoco existe en el mundo real.

EJEMPLO 3: Supuesto que se sabe que los CI (coeficientes intelectuales) se distribuyen normalmente con media $\mu = 100$ y desviación estándar $\sigma = 10$. Hállese la probabilidad de que un individuo elegido aleatoriamente tenga un CI comprendido entre 90 y 120.

SOLUCIÓN:

Sea la población X= CI de los individuos. Entonces la probabilidad buscada es el valor numérico del área bajo la curva dibujada en la siguiente gráfica denotada con A

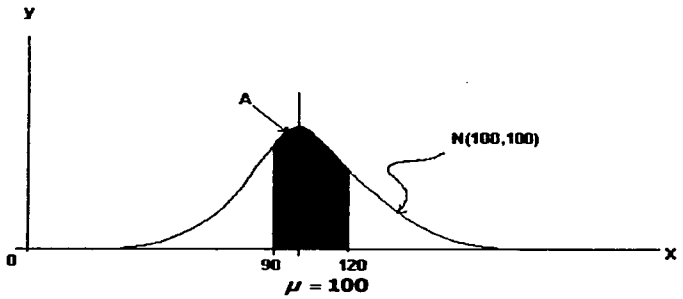


Fig. 1.6 Figura correspondiente al ejemplo 3.

No hay tablas para conocer el valor de ésta área pero podemos estandarizar.

$Z = (X - \mu) / \sigma$. Aquí $\mu = 100$ y $\sigma = 10$:

$P(90 < X < 120) = P[(90 - 100)/10 < (X - 100)/10 < (120 - 100)/10] = P(-1 < Z < 2)$

Como las tablas que existen son de la N(0,1) y $Z \cong N(0,1)$, el problema de encontrar el valor del área fue simplificado :

$P(-1 < Z < 2) = 0.8186$ según tabla de la normal estándar que viene en el apéndice de ésta obra.

La probabilidad pedida es 0.8186.

Al estandarizar ocurrió una traslación. El área equivalente es :

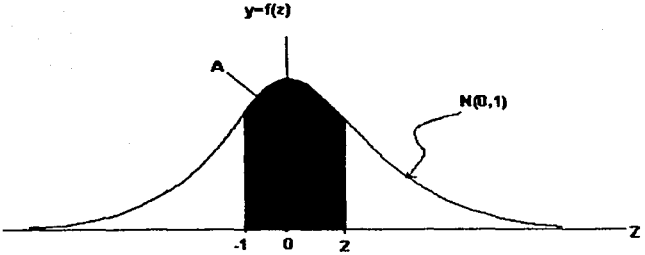


Fig. 1.7 Figura correspondiente al ejemplo 3 con una traslación.

Hay varias importantes razones por las cuales la distribución normal ocupa un lugar prominente en la estadística; lo que puedo decir es que la normal llega a encajar muy bien en las distribuciones observadas de frecuencia de multitud de fenómenos, entre ellos las características humanas (peso, talla y coeficiente intelectual), la producción de procesos físicos (dimensiones y rendimientos) y otras medidas de interés para diversos profesionistas.

Aunque muchas distribuciones apropiadas para el campo sanitario, por ejemplo, no pueden describirse de manera adecuada por una distribución normal, esta curva acampanada es muy útil en estadística.

EJERCICIO 3: Si los valores de colesterol total para cierta población están distribuídos aproximadamente en forma normal, con una media de 200 mg/100 ml y una desviación estándar de 20 mg/100 ml, hallar la probabilidad de que un individuo elegido al azar de esta población tenga un valor de colesterol:

a) entre 180 y 200 mg/100 ml
b) mayor que 225 mg/100 ml
c) menor que 150 mg/100 ml
d) entre 190 y 210 mg/100 ml

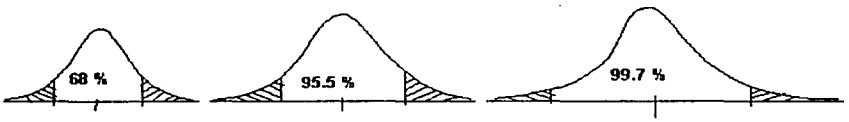
Ya para terminar ésta sección, veamos otras propiedades de la curva normal:

Aclaro aquí que no importa cuáles sean los valores de μ y σ^2 para una distribución normal de probabilidad, el área total bajo la curva normal será de 1.0, por lo cual podemos considerar que las áreas bajo la curva son probabilidades.

En términos matemáticos, es verdad que :

1) Aproximadamente 68 % de todos los valores en una población distribuída normalmente se encuentran dentro de una desviación estándar (con signo positivo y negativo) respecto de la media.
2) Aproximadamente 95.5 % de todos los valores en una población con distribución normal se hallan dentro de dos desviaciones estándar (con signo positivo y negativo) respecto de la media.
3) Aproximadamente 99.7 % de todos los valores en una población con distribución normal se encuentran dentro de tres desviaciones estándar (con signo positivo y negativo) respecto de la media.

Los tres enunciados anteriores se muestran gráficamente en la siguiente figura :



$\sigma\ \mu\ \sigma$ $\qquad 2\sigma\ \mu\ 2\sigma$ $\qquad \Leftarrow 3\sigma \Rightarrow\ \mu \Leftarrow 3\sigma$

Fig. 1.8 Relación existente entre el área bajo la curva de una distribución normal de probabilidad y la distancia respecto de la media medida en desviación estándar.

1.3 Aproximación normal a la binomial.

Recordemos que si $X \cong B(n,p)$:

$$P(X = k) = C_n^k p^k q^{n-k}, \quad k = 0,\ldots,n \quad y$$

$$P(r \le X \le s) = \sum_{k=r}^{s} C_n^k p^k q^{n-k}$$

Pues resulta que bajo ciertas condiciones hacer cálculos con las fórmulas anteriores es muy laborioso o pudiera ser que las tablas no contengan los datos que buscas. Por suerte, hay ocasiones en que la binomial se distribuye aproximadamente normal. Concretamente $X \approx N(np,npq)$. Es decir, $\mu = np$ y $\sigma^2 = npq$.

La experiencia ha demostrado que la aproximación es bastante buena mientras np y nq son 5 por lo menos. Incluso es buena cuando esto no se cumple, como en el siguiente

EJEMPLO 4: Si la probabilidad de que un tirador acierte a un blanco es de 1/3 y si dispara 12 tiros ¿cuál es la probabilidad de que acierte 6 tiros?

SOLUCIÓN: Sea X= número de tiros acertados. $X \approx B(12,1/3)$; np= 12(1/3) = 4 y nq = 12(2/3) = 8 . Puedes constatar que la probabilidad **real** pedida es P(X=6) = 0.111.

El resultado, aproximando a la normal es :

$$X \approx N\left(4, \frac{8}{3}\right) \Rightarrow \frac{X-4}{\sqrt{8/3}} \approx N(0,1)$$

Luego: $P(X=6) \approx P(6 - \frac{1}{2} < X < 6 + \frac{1}{2})$ =
$= P((5.5 - 4)/1.63 < (X-4)/1.63 < (6.5-4)/1.63) = P(0.92 < Z < 1.53) =$
$= 0.4370 - 0.3212 = 0.1158.$

½ es llamado "factor de corrección por continuidad" y se usa por el hecho de estar aproximando una distribución discreta con una contínua.

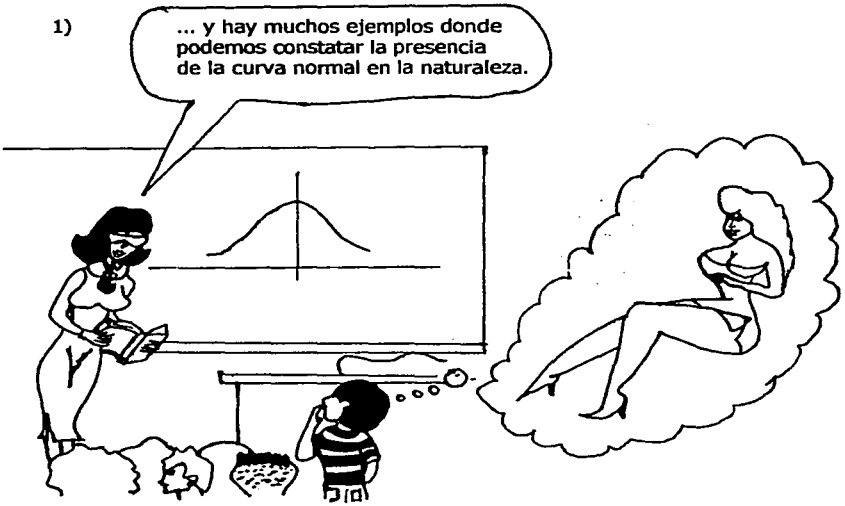
EJERCICIO 4: Resolver el siguiente problema utilizando la aproximación de la curva normal :

Si el 30 % de los estudiantes tienen visión defectuosa, ¿cuál es la probabilidad de que por lo menos la mitad de los miembros de una clase de 20 estudiantes posean visión defectuosa? (ver solución al final de esta obra)

Hasta aquí termina el repaso de los conceptos que te facilitarán el aprendizaje de los temas de tu segundo curso de estadística.

SONRIE ...

1)



2) Sabido es que con la ayuda de los números, casi cualquier cosa se puede demostrar, y que por eso las estadísticas son a veces tan engañosas. Por ejemplo, en el pequeño y rico estado de Luxemburgo, donde el nivel de vida es uno de los más altos del mundo, sucedió que en 1980 había dos hombres sin empleo, mientras que en 1981 la cifra se elevó a tres. Esto sirvió para que un periódico de la oposición gubernamental publicara el siguiente encabezado truculento:

"Caos económico en el país. La situación es catastrófica. El desempleo ha aumentado en un 50 %"

CAPITULO 2. DISTRIBUCIONES MUESTRALES.

SONRIE ...

1) Un experto es aquel que sabe más y más acerca de menos y menos, hasta que lo sabe absolutamente todo de nada.

2) A propósito de inferencias a partir de una muestra:

En un examen de zoología, el profesor le entrega a un alumno una pata de pájaro y le dice :

- A la vista de ésta extremidad, ha de decirme la familia, el género y la especie del animal, así como sus costumbres migratorias y el número de crías por nidada.

Y el alumno responde :
- Pero, ¿cómo le voy a decir todo eso con una pata solamente?

Y el profesor le dice:
- Está usted suspendido. A ver, dígame su nombre y apellidos.

Y el alumno se quita un zapato, le enseña el pie desnudo al profesor y le dice:
- Adivine ...

CAPITULO 2. DISTRIBUCIONES MUESTRALES.

2.1 Generalidades.

El razonamiento es una característica del ser humano; le permite avanzar en el orden del conocer, en él se basa todo pensamiento científico.
Hay varios tipos de razonamiento: deductivo, inductivo y analógico.

Razonamiento deductivo.- Es el razonamiento que parte de lo universal para llegar a conclusiones de menor grado de universalidad.

Ejemplo A) Todos los seres humanos sienten amor alguna vez en su vida.
Roxana fué un ser humano.
CONCLUSIÓN: Roxana sintió amor alguna vez en su vida.

Ejemplo B) Todos los metales sometidos al calor, se dilatan.
El oro es un metal.
CONCLUSIÓN: El oro se dilata.

Razonamiento inductivo.- Se organiza a la inversa del razonamiento deductivo: parte de conocimientos de menor grado de universalidad, para concluír con un conocimiento nuevo de mayor grado de universalidad.

Ejemplo C) El cobre es maleable.
El plomo es maleable.
CONCLUSIÓN: Todos los metales son maleables.

Ejemplo D) Gladys engaña a su novio.
Rodrigo engaña a su novia.
CONCLUSIÓN: Todos los jóvenes engañan a sus novios(as).

Razonamiento analógico.-Es el razonamiento que, después de haber reconocido algunas semejanzas comunes a dos o más objetos, concluye que convienen a otros objetos que también son semejantes a los ya conocidos.
Fue un razonamiento analógico el que llevó a los sabios a la conclusión de que la tierra gira alrrededor del sol y no como se pensó durante siglos, a la inversa. El hecho sucedió así: al descubrirse el planeta júpiter, se vió que era una gran masa alrrededor de la cual giraban doce satélites. Se encontró semejanza entre este hecho y el sistema solar, en el que el sol es

también una gran masa rodeada de otras más pequeñas –los planetas- y se infirió por ANALOGÍA que en ambos casos el funcionamiento era semejante.

Los razonamientos analógicos se emplean con frecuencia en ciencia y ayudan especialmente a sacar conclusiones en aquellos casos en los que los estudios directos se dificultan. Sin embargo, el razonamiento que comprueba con certeza la validez de la conclusión obtenida es el razonamiento deductivo.

¿A qué viene todo lo anterior? ¿qué tiene que ver con la estadística? Pues bien, resulta que la estadística es la disciplina que nos permite sacar conclusiones respecto a una población partiendo de tan solo una muestra de ella mediante un proceso inductivo (de lo particular, la muestra, a lo general, la población).

En el quehacer científico del mundo real en general, se realizan procesos inductivos. Es bien sabido que la inferencia inductiva constituye un proceso arriesgado. En efecto, los estudiosos de la lógica han encontrado que toda inferencia inductiva exacta es imposible. ¿Acaso estás de acuerdo con la conclusión del ejemplo D) de razonamiento inductivo ("todos los jóvenes engañan a sus novios(as)")?. Una generalización perfectamente válida no puede hacerse – en razonamiento inductivo- pero sí cabe hacer inferencias inseguras y el grado de incertidumbre es susceptible de medición si el experimento en cuestión se ha realizado de acuerdo con determinados principios. Una de las misiones de la estadística consiste en conseguir técnicas para efectuar inferencias inductivas y para medir el grado de incertidumbre de tales inferencias. Uno de los principios es que la muestra represente aceptablemente a la población y la medida de la incertidumbre viene expresada en probabilidad.

Por cierto, te preguntarás ¿por qué querríamos trabajar con los pocos datos de una muestra y no con todos los de la población?. La respuesta es: porque no siempre se tienen recursos humanos, materiales (dinero) o tiempo, suficientes para un censo. Además, pudiera ser que al estar recolectando los datos muestrales de la población, ésta última se destruyera, como sucedería si estuviéramos registrando el tiempo de vida útil de una producción de focos.

2.2 Población y muestra.

Ahora bien,¿cuáles son las muestras **que representan aceptablemente a la población**? Y una vez que lo sepamos, ¿cómo obtenemos dichas muestras?

En estadística, lo que se quiere es que la muestra represente en miniatura a la población. ¿Cómo lograr que nuestra muestra represente aceptablemente a la población de donde proviene?. Como ya dijimos, lo que debe buscarse en una muestra es que miniaturice a la población, que la distribución de frecuencias de ella represente de modo satisfactorio a la población muestreada. Si no fuera así, la muestra no serviría de base para las inferencias estadísticas respecto a la población que se muestrea. Las muestras que satisfacen las características deseables son las llamadas muestras aleatorias, que son aquellas que se

obtienen dándole a cada dato de la población la misma oportunidad de ser elegida para la muestra. Además, es en las muestras aleatorias donde, como seguramente ya habrás intuído, puede aplicarse la teoría de probabilidades.

Las muestras **que representan aceptablemente a la población** son las que **se obtienen al azar .**

Muy bien, pero ¿cómo obtener muestras aleatorias?. Una forma es proceder de modo análogo a como se hace en las rifas domésticas: numerando papelitos, haciéndolos bolita, metiéndolos en un recipiente, revolviendo bien y sacando alguno con los ojos cerrados. Otra forma es utilizando tablas de números aleatorios elaborados por especialistas. Una forma más es generar números aleatorios con las computadoras programadas para mezclar números, o con las calculadoras con la tecla RAN. Para propósitos de este curso, nos serviremos del siguiente ejemplo.

EJEMPLO 5: En éste ejemplo seleccionaremos una muestra aleatoria simple usando la tabla de números aleatorios que viene en el apéndice de ésta obra. Supóngase que la población de interés consiste de las 150 concentraciones de azúcar en la sangre extraída en ayunas que se muestra en la siguiente tabla (Núm significa número de individuo).

CONCENTRACIONES DE AZÚCAR EN LA SANGRE EXTRAÍDA EN AYUNAS DE 150 INDIVIDUOS APARENTEMENTE NORMALES

NUM	Concen	num	Concen	num	Concen	num	Concen	num	Concen	num	Concen	num	Concen
1	91	23	98	45	119	67	88	89	107	111	96	133	102
2	94	24	89	46	90	68	107	90	97	112	104	134	101
3	115	25	105	47	82	69	113	91	91	113	85	135	111
4	85	26	101	48	90	70	95	92	104	114	108	136	91
5	89	27	81	49	113	71	102	93	109	115	103	137	92
6	107	28	108	50	104	72	94	94	92	116	90	138	98
7	94	29	94	51	97	73	99	95	85	117	105	139	81
8	105	30	104	52	101	74	87	96	108	118	99	140	117
9	94	31	107	53	90	75	102	97	99	119	88	141	103
10	103	32	94	54	88	76	105	98	103	120	103	142	96
11	104	33	101	55	108	77	80	99	81	121	90	143	101
12	105	34	95	56	95	78	90	100	96	122	105	144	88
13	88	35	80	57	100	79	108	101	105	123	100	145	100
14	107	36	104	58	103	80	105	102	91	124	89	146	100
15	90	37	94	59	10[illegible]	81	90	103	115	125	90	147	95
16	95	38	102	60	85	82	115	104	108	126	106	148	103
17	104	39	89	61	87	83	82	105	102	127	94	149	101
18	93	40	98	62	104	84	90	106	101	128	100	150	90
19	109	41	106	63	109	85	102	107	94	129	92		
20	87	42	85	64	93	86	91	108	93	130	91		
21	92	43	93	65	95	87	103	109	102	131	87		
22	117	44	103	66	107	88	107	110	119	132	105		

TABLA I

Extraigamos de esta población una muestra aleatoria simple de tamaño 10, usando la tabla de números aleatorios que viene en el apéndice de ésta obra. Como primer paso, localícese un punto de partida aleatorio en la tabla. Esto puede hacerse en varias formas, una de las cuales es quitar la vista de la página, mientras se le toca con la punta de un lápiz. El punto de partida aleatorio es el dígito más próximo a donde el lápiz tocó la página. Supongamos que, siguiendo éste procedimiento, se llegó a un punto de partida aleatorio en la primera página de la tabla (de números aleatorios), en la intersección del renglón 21 y la columna 28. El dígito en éste punto es 5. Como se tienen 150 valores de los cuales elegir, solo pueden usarse los números aleatorios 1 al 150. Resultará conveniente elegir números de tres dígitos de modo que solo sean elegibles los números del 001 al 150. El primer número de tres dígitos empezando en el punto de partida hallado, es el 532, un número que no puede usarse. Recorramos la tabla hacia abajo y pasemos el 196,372, 654 y el 928 hasta llegar al 137, un número que puede usarse. El 137vo valor de la tabla I es el 92, el primer valor de la muestra. En la tabla II se han registrado el número aleatorio y la concentración de azúcar en la sangre correspondiente.

MUESTRA DE 10 CONCENTRACIONES DE AZÚCAR EN LA SANGRE EXTRAÍDA EN AYUNAS OBTENIDAS DE LOS QUE SE DAN EN LA TABLA I		
Número aleatorio	Número del sujeto en la muestra	Concentración
137	1	92
114	2	108
028	3	108
085	4	102
018	5	93
042	6	85
053	7	90
108	8	93
144	9	88
126	10	106

TABLA II

Se han registrado los números aleatorios para que se vea cuáles fueron los seleccionados. Como se desea una muestra sin reemplazo, no se desea incluír dos veces la misma concentración de un individuo. Procediendo en la forma que acaba de describirse se llega a los nueve números aleatorios restantes y las concentraciones de azúcar en la sangre que se muestran en la tabla II. Nótese que cuando se llega hasta abajo, simplemente se avanza tres dígitos hacia la derecha hasta el 028 y se sigue hacia arriba de la columna. También pudo haberse empezado por la parte superior de esta última columna con el número 369.

Así se ha extraído una muestra aleatoria simple de tamaño 10 de una población de tamaño 150. En toda discusión futura, siempre que se use el término muestra aleatoria simple, se entenderá que la muestra se extrajo en ésta forma o en otra equivalente.

Ejercicio 2.2.1 Usando la tabla de números aleatorios, seleccionar un nuevo punto de partida aleatorio y extraer otra muestra aleatoria simple de tamaño 10 de los datos de la población de la tabla I.

Aquí conviene un comentario : si el resultado de nuestro trabajo estadístico fuera erróneo ¿a qué puede deberse? Tal vez se debió a que el muestreo no fue al azar. O sí lo fue pero el tamaño de la muestra fue muy pequeño o, si se trató de una encuesta, los encuestados pudieron haber mentido para impresionar, etc.

2.3 Estadísticos y parámetros.

Desde el punto de vista matemático, podemos describir las muestras y poblaciones mediante medidas como la media, la moda y la desviación estándar. Cuando éstos términos describen las características de una muestra, se les llama **estadísticos**. Cuando describen las características de una población, reciben el nombre de **parámetros**. El estadístico es una característica de la muestra; el parámetro es una característica de la población.

Supóngase que la altura media en pulgadas de todos los niños del décimo grado es de sesenta pulgadas en México. En éste caso, 60 pulgadas es una característica de la población "todos los alumnos del décimo grado" y podemos llamarla un parámetro de la población. Por otra parte, si decimos que la altura media del grupo escolar del décimo grado de la señora Jones es de 60 pulgadas estaremos usando ésta cifra para describir una característica de la muestra "alumnos del décimo grado de la señora Jones". En tal caso, 60 pulgadas será un estadístico muestral. Si estamos convencidos de que la altura media de los alumnos del décimo grado de la señora Jones es una estimación precisa (un estimador es un numerito y no uno de tus amigos que te quiere mucho) de la talla media de todos los alumnos del décimo grado en México, podríamos emplear el estadístico muestral "altura media de los alumnos del décimo grado de la señora Jones" para estimar el parámetro de la población "altura media de todos los alumnos del décimo grado en México", sin tener que contar todos los millones de ese grupo escolar en el país.

Para ser congruentes, los expertos en estadística utilizan letras minúsculas cuando quieren denotar los estadísticos muestrales y letras griegas o mayúsculas para indicar los parámetros de la población. Así tenemos la siguiente tabla :

DEFINICION	GRUPO DE ELEMENTOS QUE VAN A SER CONSIDERADOS	PARTE O PORCIÓN DE LA POBLACIÓN SELECCIONADA PARA EL ESTUDIO
CARACTERISTICAS	"Parámetros"	"Estadísticos"
SIMBOLOS	Tamaño de la población: N Media de la población: μ Desv. Estándar de la pob. : σ	Tamaño de la muestra : n Media muestral: $\bar{x}$ Desv est de la muestra: s

TABLA 2.3.1

2.4 Los estadísticos como variables aleatorias. Interpretación del Teorema Central Del Límite.

Formalmente un estadístico **E** se define como una función de la muestra

$$E : M \longrightarrow R$$
$$(x_1, x_2, \ldots, x_n) \mapsto E[(x_1, x_2, \ldots, x_n)]$$

Donde M = todas las muestras posibles de tamaño n de una población.

Por otro lado, recuérdese que una variable aleatoria es un resultado numérico de un fenómeno aleatorio; más precisamente, recuérdese que es una función que a cada elemento de un espacio muestra (ver el CAPITULO 1) le asigna un número real. Puesto que nuestra muestra es aleatoria ya que cada elemento de la población tuvo la misma oportunidad de aparecer en nuestra muestra, estarás de acuerdo en que nuestra muestra particular es un elemento de algún espacio muestra Ω = todos los resultados posibles de un fenómeno aleatorio.

Simplemente hagamos M= Ω y resulta entonces que un estadístico es también una variable aleatoria.

$$E: \Omega \longrightarrow R$$

Ejemplos de estadísticos son

i) $\overline{x} : M \longrightarrow R$ (media aritmética)
ii) $\overline{p} : M \longrightarrow R$ (proporción muestral)

No son los únicos que existen desde luego, pero sí son los únicos que estudiaremos en este trabajo.

2.4.1 Distribuciones de la media de la muestra.

Empecemos con un ejemplo:

EJEMPLO 6: Sea P una población finita de tamaño N=5 que consiste de las edades de 5 niños {6, 8, 10, 12, 14} pacientes externos en un centro de enfermedades mentales.

Puedes verificar fácilmente que la media de ésta población es el parámetro $\mu = 10$; es decir, la edad promedio de la población es 10, y la varianza es el parámetro

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N} = 8$$

Extraeremos todas las muestras tamaño n sin reemplazo y sin importar el orden, de diferente tamaño y sin llegar al censo. Calcularemos la media y la varianza de dichas muestras para ver qué encontramos. Los resultados están en la siguiente tabla

Muestras tamaño n = 1	Muestras tamaño n = 2	Muestras tamaño n = 3	Muestras tamaño N = 4
{ 6 } $\bar{x} = 6$	{6,8} $\bar{x} = 7$	{6,8,10} $\bar{x} = 8$	{6,8,10,12} $\bar{x} = 9$
{ 8 } $\bar{x} = 8$	{6,10} $\bar{x} = 8$	{6,8,12} $\bar{x} = 26/3$	{6,8,10,14} $\bar{x} = 38/4$
{ 10 } $\bar{x} = 10$	{6,12} $\bar{x} = 9$	{6,8,14} $\bar{x} = 28/3$	{8,10,12,14} $\bar{x} = 11$
{ 12 } $\bar{x} = 12$	{6,14} $\bar{x} = 10$	{8,10,12} $\bar{x} = 10$	{6,10,12,14} $\bar{x} = 42/4$
{ 14 } $\bar{x} = 14$	{8,10} $\bar{x} = 9$	{10,12,14} $\bar{x} = 12$	{6,8,12,14} $\bar{x} = 10$
	{8,12} $\bar{x} = 10$	{6,12,14} $\bar{x} = 32/3$	
	{8,14} $\bar{x} = 11$	{8,12,14} $\bar{x} = 34/3$	
	{10,12} $\bar{x} = 11$	{6,10,12} $\bar{x} = 28/3$	
	{10,14} $\bar{x} = 12$	{6,10,14} $\bar{x} = 10$	
	{12,14} $\bar{x} = 13$	{8,10,14} $\bar{x} = 32/3$	
$\mu_{\bar{x}} = 10$ $\sigma^2_{\bar{x}} = \frac{\sum(x_i - \bar{x})^2}{5} = 8$	$\mu_{\bar{x}} = 10$ $\sigma^2_{\bar{x}} = \frac{\sum(x_i - \bar{x})^2}{10} = 3$	$\mu_{\bar{x}} = 10$ $\sigma^2_{\bar{x}} = 4/3 \approx 1.333$	$\mu_{\bar{x}} = 10$ $\sigma^2_{\bar{x}} = 1/2 = 0.5$

La "distribución de probabilidad" de la media para muestras tamaño n=1 (primera columna) está dada en la siguiente tabla

X	6	8	10	12	14
$f(x)$	1/5	1/5	1/5	1/5	1/5

Y para n=2

X	7	8	9	10	11	12	13
$f(x)$	1/10	1/10	2/10	2/10	2/10	1/10	1/10

Para las muestras de tamaño n=3 (tercera columna), la "distribución de probabilidad" es

X	8	26/3	28/3	10	32/3	12	34/3
$f(x)$	1/10	1/10	2/10	2/10	2/10	1/10	1/10

Y por último, para n=4 (cuarta columna) :

X	9	38/4	11	42/4	10
$f(x)$	1/5	1/5	1/5	1/5	1/5

Hagamos unas observaciones interesantes respecto al cuadro anterior a las 4 tablas de "distribuciones de probabilidad" :

1) Para cualquier tamaño de muestra n, la media de todas las medias es igual a la media de la población.
2) La varianza disminuye a medida que n, el tamaño de la muestra, crece.

Estos resultados no son coincidencias; vienen expresados en el llamado TEOREMA CENTRAL DEL LIMITE : **Dada una población de cualquier forma funcional con una media μ y una varianza finita σ^2, la distribución muestral de $\overline{x}$, calculada a partir de muestras de tamaño n de ésta población, estará distribuída aproximadamente en forma normal con media $\mu_{\overline{x}} = \mu$ y varianza $\sigma^2_{\overline{x}} = \frac{\sigma^2}{n}$, cuando el tamaño de la muestra es grande.**

Formalmente, el teorema central del límite dice :

TEOREMA CENTRAL DEL LÍMITE : Sean $x_1, x_2, \cdots, x_n$ una muestra de variables aleatorias independientes e idénticamente distribuídas muestreada con reemplazo o bien de una población infinita con alguna distribución no importa si discreta o contínua, normal o no normal de media μ y varianza σ^2 (observación : una población normal es por definición una población infinita). Entonces $\overline{x}$ se distribuye mas o menos normal donde $\mu_{\overline{x}} = \mu$ y $\sigma^2_{\overline{x}} = \frac{\sigma^2}{n}$.

Esto último lo escribiremos así: $\overline{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$ y se lee "$\overline{X}$ se distribuye aproximadamente en forma normal con media μ y varianza $\frac{\sigma^2}{n}$ ".

Así es que si la población se distribuye en forma normal, la media se distribuye normalmente para cualquier tamaño de muestra. Pero si no sabemos cómo se distribuye la población o sabemos que no se distribuye normalmente, entonces $\overline{X}$ se distribuye mas o

menos normal para tamaños de muestra n > 30 y la distribución (de $\bar{X}$) tiende a tener una distribución normal a medida que el tamaño de la muestra crece.

Sin embargo, hay ocasiones en que es necesario muestrear sin reemplazo de una población finita como es el caso que nos acaba de ocupar, el del ejemplo 6. En este caso, el TCL nos dice que $\bar{X}$ se distribuye aproximadamente en forma normal con media $\mu_{\bar{x}} = \mu$ y varianza

$$\sigma^2_{\bar{x}} = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$$

Es decir

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)\right)$$

El factor (N-n)/(N-1) se llama corrección por población finita y se usa si n/N >0.05. Si no se cumple esto último, se toma como varianza $\sigma^2_{\bar{x}}$ a $\frac{\sigma^2}{n}$; o sea

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

En la tabla anterior, correspondiente al ejemplo 6, puedes constatar que
$\sigma^2_{\bar{x}} = \sigma^2$(N-n) / n(N-1) ya que el tamaño de la muestra es mayor que el 5 % del tamaño de la población (n > 0.05N =(0.05)(5) = 0.25).

El conocimiento y comprensión de las distribuciones muestrales será un requisito indispensable para entender los conceptos de la inferencia estadística. La aplicación más sencilla del conocimiento de la distribución muestral de la media de la muestra es al calcular la probabilidad de obtener una muestra con una media de alguna magnitud especificada. Veamos algunos ejemplos :

EJEMPLO 7: Supóngase que se sabe que los salarios por hora de cierto tipo de empleados de hospital están distribuídos aproximadamente en forma normal con una media y una desviación estándar de $ 45.00 y $ 5.00 respectivamente. Si se selecciona una muestra aleatoria de tamaño 16 de ésta población, encontrar la probabilidad de que la media del salario por hora para la muestra sea :

a) Mayor que $ 42.50
b) Entre $ 42.50 y $ 47.50
c) Mayor que $ 48.00
d) Menor que $ 42.00

SOLUCION: Sea la población X= Salario por hora de los empleados del hospital. Aquí, $\mu = 45$, $\sigma = 5$, n=16. Como $\bar{X} \approx N(45,25)$, el TCL dice que

$$\bar{X} \approx N\left(45, \frac{5^2}{16}\right)$$

a) Mayor que $ 42.50 . $P(\bar{X} > 42.5)$. Estandarizamos para poder usar las tablas de la normal (0,1):

$$P\left(\frac{\bar{x}-45}{5/4} > \frac{42.5-45}{5/4}\right) = P(Z > -2) = 0.9772$$

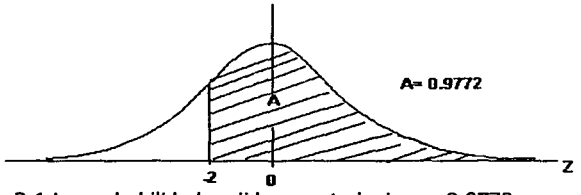


Fig. 2.1 La probabilidad pedida en este inciso es 0.9772.

b) Entre $ 42.50 y $ 47.50. $P(42.5 < \bar{X} < 47.5)$. Estandarizamos para poder usar las tablas de la Normal que viene en el apéndice de ésta obra.

$$P\left(\frac{42.5-45}{5/4} < \frac{\bar{x}-45}{5/4} < \frac{47.5-45}{5/4}\right) = P(-2 < Z < 2) =$$
$$2(0.4772) = 0.9544$$

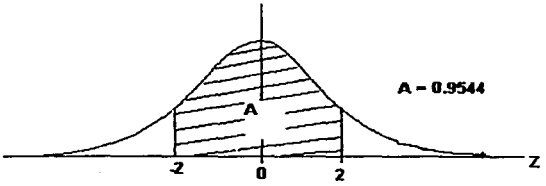


Fig. 2.2 La probabilidad pedida en éste inciso es 0.9544

c) Mayor que $ 48.00. $P(\overline{X} > 48)$. Estandarizamos para poder usar las tablas de la normal estándar (la N(0,1))

$$P(\overline{X} > 48) = P\left(\frac{\overline{X} - 45}{5/4} > \frac{48 - 45}{5/4}\right) = P(Z > 2.4) = 0.0082$$

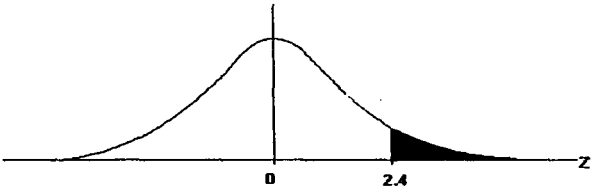


Fig. 2.3 La probabilidad pedida en éste inciso es 0.0082, representada por el área sombreada.

d) Menor que $ 42.00. $P(\overline{X} < 42)$. Estandarizamos para poder usar las tablas de la normal estándar, N(0,1) :

$$P(\overline{X} < 42) = P\left(\frac{\overline{X} - 45}{5/4} < \frac{42 - 45}{5/4}\right) = P(Z < -2.4) = 0.0082$$

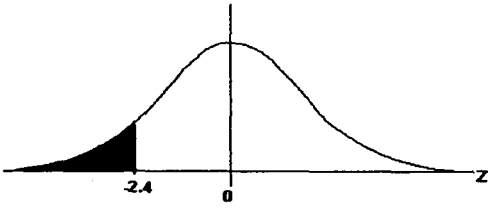


Fig. 2.4 La probabilidad pedida en éste inciso es 0.0082, representada por el área sombreada.

EJEMPLO 8: Si las concentraciones de ácido úrico en los adultos del sexo masculino normales están distribuídos aproximadamente en forma normal con una media y una desviación estándar de 5.7 y 1 mgs por ciento respectivamente, encontrar la probabilidad de que una muestra de tamaño 9 tenga una media :

a) Mayor que 6
b) Entre 5 y 6
c) Menor que 5.2

SOLUCION : Sea la población X = Concentraciones del ácido úrico (en mgs por ciento). Aquí n=9, μ =5.7 y σ =1. Entonces por el Teorema Central del Límite (TCL) : $\bar{X} \approx N(5.7, 1/9)$.

a) Mayor que 6. $P(\bar{X} > 6)$. Estandarizando para poder usar las tablas de la normal (0,1):

$$P(\bar{X} > 6) = P\left(\frac{\bar{X} - 5.7}{1/3} > \frac{6 - 5.7}{1/3}\right) = P(Z > 0.9) = 0.1841$$

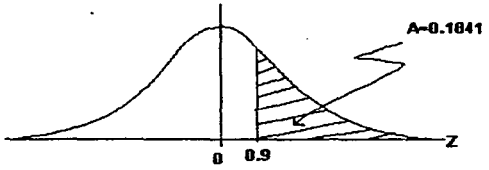


Fig. 2.5 La probabilidad pedida en éste inciso es 0.1841

b) Entre 5 y 6. $P(5 \leq \bar{X} \leq 6)$. Estandarizando : $P(5 \leq \bar{X} \leq 6) =$

$$P\left(\frac{5 - 5.7}{1/3} \leq \frac{\bar{X} - 5.7}{1/3} \leq \frac{6 - 5.7}{1/3}\right) = P(-2.1 \leq Z \leq 0.9) = 0.798$$

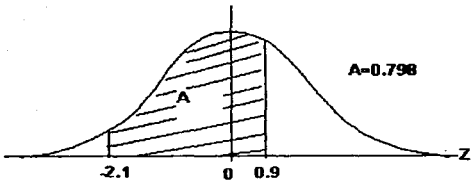


Fig. 2.6 La probabilidad pedida en este inciso es 0.798.

c) Menor que 5.2 . P($\bar{X}$ <5.2) . Estandaricemos : P($\bar{X}$ <5.2) =

$$P\left(\frac{\bar{X}-5.7}{1/3} < \frac{5.2-5.7}{1/3}\right) = P(Z < -1.5) = 0.0668$$

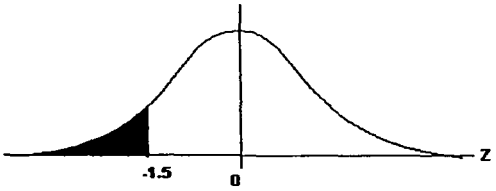


Fig. 2.7 La probabilidad pedida en este inciso es 0.0668 representada por el área sombreada.

EJERCICIOS PROPUESTOS DE LA SECCION 2.4.1

1) Se ha encontrado que , siguiendo un período de entrenamiento, el tiempo medio requerido por ciertas personas impedidas para realizar una tarea particular es de 25 segundos, con una desviación estándar de 5 segundos. Suponiendo una distribución normal de los tiempos, encontrar la probabilidad de que una muestra de 25 individuos proporcionen una media

   a) de 26 segundos o más.
   b) Entre 24 y 27 segundos.
   c) 26 segundos o menos.
   d) Mayor que 22 segundos.

2) Para cierta porción grande de la población, para un año particular, supóngase que el número medio de días de incapacidad es de 5.4, con una desviación estándar de 2.8 días. Hallar la probabilidad de que una muestra aleatoria de tamaño 49 de esta población tenga una media :

   a) mayor que 6 días
   b) Entre 4 y 6 días
   c) Entre 4.5 y 5.5 días

3) Dada una población normalmente distribuída con una media de 100 y una desviación estándar de 20, encontrar las probabilidades siguientes, basadas en una muestra de tamaño 16 :

   a) $P(\bar{X} \geq 100)$
   b) $P(96 \leq \bar{X} \leq 108)$
   c) $P(\bar{X} \leq 110)$

4) Dado $\mu = 50$, $\sigma = 16$, n=64, encontrar :

   a) $P(45 \leq \bar{X} \leq 55)$
   b) $P(\bar{X} \geq 53)$
   c) $P(\bar{X} \leq 47)$
   d) $P(49 \leq \bar{X} \leq 56)$

2.4.2 Distribución de la proporción de la muestra.

Si hiciéramos un procedimiento análogo al que hicimos para la media al principio de la sección anterior, encontraríamos que el TEOREMA CENTRAL DEL LÍMITE (TCL) nos dice lo siguiente referente a las proporciones:

Considérese una población finita y sea p la proporción de éxitos en ella. Para una muestra de tamaño n, sea $\overline{p}$ la proporción de éxitos en ella (en la muestra). Si el tamaño de la muestra es grande, la proporción de éxitos en ella, $\overline{p}$, se distribuye aproximadamente en forma normal con media $\mu_{\overline{p}} = p$ y varianza $\sigma^2_{\overline{p}} = \frac{p(1-p)}{n}$.

Esto lo escribiremos así : $\overline{p} \approx N(\mu_{\overline{p}}, \sigma^2_{\overline{p}})$. O sea, $\overline{p} \approx N\left(p, \frac{p(1-p)}{n}\right)$

¿Qué tan grande debe ser n para que sea válido el uso de la aproximación normal? Un criterio ampliamente usado es que tanto *np* como *n(1-p)* sean mayores que 5.

Puede mejorarse la aproximación normal mediante la corrección por continuidad, un artificio que hace un ajuste por el hecho de que se está aproximando una distribución discreta por una distribución contínua. Para emplear la corrección por continuidad, se resta 1/2n del valor absoluto de $\overline{p} - p$ al calcular Z, para dar

$$Z = \frac{|\overline{p} - p| - \frac{1}{2n}}{\sqrt{\frac{pq}{n}}}$$

Donde q=1-p.

La corrección por continuidad no producirá una gran diferencia cuando n es grande. Por lo tanto en éstos casos (cuando n es grande) es indistinto usar o no la corrección por continuidad.

EJEMPLO 9: Si en una población de mujeres, 15% están sometidas a cierta dieta, ¿cuál es la probabilidad de que una muestra aleatoria de tamaño 100 dé una proporción de aquellas que se encuentran a dieta :

a) Mayor que o igual a 0.20?
b) Entre 0.10 y 0.20?
c) No mayor que 0.12?

SOLUCIÓN: Como np= 100(0.15) = 15 y nq= 100(0.85) son ambos mayores que 5 entonces es válido el uso de la aproximación normal:

$$\bar{p} \approx N\left(0.15, \frac{(0.15)(0.85)}{100}\right) = N(0.15, 0.001275)$$

a) $P(\bar{p} \geq 0.2)$. Estandarizando, para poder usar las tablas de la normal que existen:

$$P(\bar{p} \geq 0.2) = P\left(\frac{\bar{p} - 0.15}{0.03570} \geq \frac{0.2 - 0.15}{0.03570}\right) \approx P(Z \geq 1.4) \approx 0.0808$$

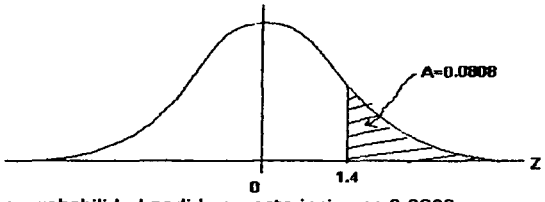


Fig 2.8 La probabilidad pedida en este inciso es 0.0808

Utilizando la corrección por continuidad,tendríamos:

$$Z_c = \frac{|0.2 - 0.15| - \frac{1}{2(200)}}{0.035707142} = 1.26$$

Y entonces $P(\bar{p} \geq 0.2) = P(Z \geq Z_c) = P(Z \geq 1.26) = 0.1038$

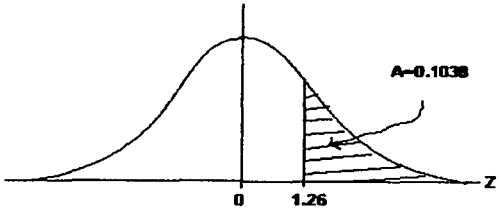


Fig. 2.9 Figura correspondiente al inciso a) pero usando corrección por continuidad.

b) P(0.10 ≤ $\overline{p}$ ≤ 0.20). Estandarizando:

$$P(0.10 \le \overline{p} \le 0.20) = P\left(\frac{0.10-0.15}{0.0357} \le \frac{\overline{p}-0.15}{0.0357} \le \frac{0.20-0.15}{0.0357}\right) =$$

$$= P(-1.4 \le Z \le 1.4) = 0.8384$$

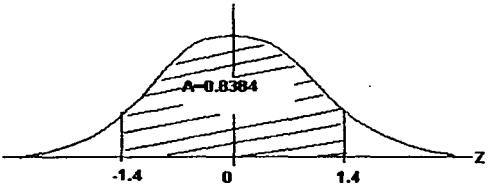


Fig. 2.10 La probabilidad pedida en este inciso es 0.8384

c) P($\overline{p}$ ≤ 0.12). Estandaricemos:

$$P(\overline{p} \le 0.12) = P\left(\frac{\overline{p}-0.15}{0.0357} \le \frac{0.12-0.15}{0.0357}\right) = P(Z \le -0.84) = 0.2005$$

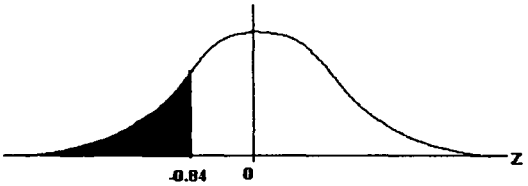


Fig. 2.11 La probabilidad pedida en este inciso es 0.2005, representada por el área sombreada.

Ejercicios propuestos de la sección 2.4.2

1) En cierta ciudad, se observa que el 20 % de las familias tienen al menos un miembro que está sufriendo de algún malestar debido a la contaminación del aire. Una muestra aleatoria de 150 familias dio $\overline{p} = 0.27$. Si el valor del 20 % es correcto, ¿cuál es la probabilidad de obtener una proporción de la muestra así o mayor?

2) En una muestra aleatoria de 75 adultos, 35 dijeron que consideran que el cáncer de la mama es curable. Si en la población de la cual se extrajo la muestra, la proporción verdadera de quienes creen que el cáncer de la mama es curable es 0.55, ¿cuál es la probabilidad de obtener una proporción de la muestra tan pequeña o menor que la que se obtuvo en esta muestra?

3) Se sabe que el medicamento estándar usado para tratar a cierta enfermedad ha resultado ser efectivo en un lapso de 3 días en el 75% de los casos en los que se usó. Al evaluar la efectividad de un nuevo medicamento para tratar a la misma enfermedad, se le dio a 150 personas que la sufrían. Al término de los 3 días, se habían recuperado 97 personas. Si la nueva medicina es tan efectiva como la estándar, ¿cuál es la probabilidad de observar esta pequeña proporción de recuperación?

4) Dada una población en la que p= 0.6 y una muestra aleatoria extraída de esta población, de tamaño 100, encontrar

   a) $P(\overline{p} \geq 0.65)$
   b) $P(\overline{p} \leq 0.58)$
   c) $P(0.56 \leq \overline{p} \leq 0.66)$

2.5 MISCELÁNEA DE EJERCICIOS DEL CAPÍTULO 2.

1) Sugerir una forma para tomar una muestra al azar de 100 estudiantes del estudiantado de una Universidad.
2) Las líneas aéreas dejan generalmente cuestionarios en las bolsas de los asientos de sus aviones para obtener información de sus clientes con respecto a sus servicios. Hacer una crítica de éste método para obtener información.
3) Una firma comercial envió cuestionarios para una muestra al azar de 1000 amas de casa en una determinada ciudad, concerniendo sus puntos de vista respecto a servilletas de papel. De ellas, respondieron 400. ¿ Serían satisfactorias estas 400 respuestas para juzgar los puntos de vista generales de las amas de casa con respecto a servilletas?
4) Una agencia desea tomar una muestra de 200 adultos en una cierta sección residencial de una ciudad. Se proponen hacerlo tomando una muestra al azar de 200 casas obtenidas de una lista de casas en ese distrito y seleccionando entonces al azar un adulto de cada casa. ¿Por qué no puede considerarse que éste procedimiento arroje muestras al azar?
5) Tomando X normalmente distribuída con una media de 22 y desviación estándar de 4, calcular la probabilidad de que la media de la muestra $\overline{x}$, basada en una muestra de tamaño 64

   a) Exceda a 23
   b) Exceda 21.5
   c) Se encuentre entre 21 y 23
   d) Exceda a 24
   e) Exceda a 18

6) Bosquejar en la misma hoja de papel la gráfica de una curva normal con media 10 y desviación estándar 2 y la correspondiente curva media para una muestra de tamaño 9.
7) Si la desviación estándar del peso de niños de primer grado es de 3 kg, ¿Cuál es la probabilidad de que el peso medio de una muestra al azar de 100 de éstos niños difiera en más de medio kilogramo, con respecto al peso medio para todos los niños?
8) El peso medio de alumnos entrantes a una Universidad, tomado en los últimos 5 años, es de 70 kg y la desviación estándar es 8 kg. Si el peso medio de los 100 primeros alumnos inscritos es de 72.5 kg, ¿estaría usted en lo correcto al pensar que la nueva generación de estudiantes es más pesada que las anteriores? Dé algunas explicaciones posibles.
9) Con un calendario muestree sistemáticamente cada día decimoctavo del año, comenzando con el 6 de enero.

10) Una organización no lucrativa está efectuando una encuesta domiciliaria de opinión sobre los centros municipales de atención diurna. La organización ha ideado un esquema para realizar el muestreo aleatorio de las casas y planea efectuar la encuesta los días laborables de las 12 del día a las 5 de la tarde. ¿Producirá este esquema una muestra aleatoria?
11) Consulte la tabla de números aleatorios del apéndice de ésta obra. ¿cuál es la probabilidad de que un 4 aparezca en el dígito del extremo izquierdo en cada conjunto de 10 dígitos?¿y de que aparezca un 7 o un 2?¿cuántas veces espera usted ver esos dígitos en la posición extrema izquierda?¿cuántas veces se encuentra cada uno en ella?¿puede explicar las diferencias entre el número encontrado y el número esperado?
12) ¿A qué tipo de error se refiere el término *error* en el error estándar de la media?
13) Una empresa gasera ha determinado que el costo medio de 100 pies cuadrados para el servicio eléctrico de la población residencial es de $ 0.314, con un error estándar de $ 0.07. Se seleccionaron dos muestras al azar, y las medias son $ 0.30 y $ 0.35, respectivamente. El ayudante encargado de la obtención de datos llega a la conclusión de que la segunda muestra es mejor porque conviene más sobreestimar que subestimar la media verdadera. Comente esta afirmación. ¿Es mejor en algún aspecto una de las medias, si se tiene la verdadera media de la población?
14) El presidente de la American Dental Association quiere determinar el número promedio de veces que cada paciente de un dentista se enjuaga la boca por día. Para lograrlo pide a 100 dentistas seleccionados aleatoriamente encuestar a 50 de sus pacientes en forma aleatoria y entregar a la American Dental Association el número medio de enjuagues diarios. Esos números se calculan y se envían al presidente de la asociación. ¿Ha recibido el presidente una muestra extraída de la población de pacientes o de alguna otra distribución?.
15) En una muestra de 25 observaciones de una distribución normal con una media de 98.6 y una desviación estándar de 17.2, encuentre:

    a) P(92< $\bar{X}$ <102)
    b) La probabilidad correspondiente si tenemos una muestra de 36.

16) En una distribución normal con una media de 375 y una desviación estándar de 48, ¿de qué tamaño debe ser una muestra para que haya por lo menos 0.95 probabilidades de que la muestra se encuentre entre 370 y 380?
17) El costo promedio de un condominio con estudio en un desarrollo urbano es de $ 62,000 con una desviación estándar de $ 4 200.

    a) ¿Cuál es la probabilidad de que un condominio de éste desarrollo cueste por lo menos $ 65 000?
    b) ¿Es la probabilidad de que el costo promedio de una muestra de dos condominios sea al menos $ 65 000 mayor o menor que la probabilidad de que un condominio cueste esa cantidad? ¿en qué cantidad?

18) Una refinería de petróleo tiene monitores de reserva para llevar un control constante del flujo y prevenir que las fallas de la máquina desorganicen el proceso. Un monitor tiene un promedio de vida de 4 300 horas, con una desviación estándar de 730 horas. Además del monitor primario, la refinería ha instalado dos unidades de emergencia, que son un duplicado de la unidad primaria. En caso de avería de uno de los monitores, el otro se activa en forma automática. La vida de operación de los dos es independiente de los otros.

   a) ¿Cuál es la probabilidad de que determinado conjunto de monitores dure por lo menos 13 000 horas?
   b) ¿y un máximo de 12 630 horas?

19) El presidente de una empresa telefónica está molesto con el número de teléfonos producidos por la empresa que tienen aparatos defectuosos. En promedio, 120 teléfonos son devueltos diariamente a causa de ese problema, con una desviación estándar de 81. El presidente ha decidido que, a menos que logre una seguridad promedio de 85 % de que al día no serán devueltos más de 135 teléfonos en los próximos 40 días, ordenará revisar el producto. ¿Tomará esa medida?

20) Un agricultor, que vende trigo a Alemania Occidental, posee 60 acres de campos de trigo. Basándose en su experiencia, sabe que el rendimiento de cada acre tiene una distribución normal con una media de 120 bushels y una desviación estándar de 12 bushels. Ayúdele a planear la cosecha del próximo año, calculando :

   a) La media esperada de los rendimientos de 60 acres de campo, destinados al cultivo de trigo.
   b) La desviación estándar de la media muestral de los rendimientos de los 60 acres.
   c) La probabilidad de que el rendimiento medio por acre rebase los 123.8 bushels.
   d) La probabilidad de que el rendimiento medio por acre oscile entre 117 y 122 bushels.

21) Un técnico de rayos X está haciendo lecturas con su máquina para cerciorarse de que se ajusta a las pautas de seguridad del gobierno. Sabe que la desviación estándar de la radiación emitida por la máquina es de 150 milirems, pero quiere tomar lecturas hasta que el error estándar de la distribución muestral no supere los 25 milirems. ¿Cuántas lecturas deberá tomar?

22) Se escogieron 64 elementos en una población de 125 elementos, con una media de 105 y una desviación estándar de 17.

   a) ¿Cuál es el error estándar de la media?
   b) ¿Cuál es la $P(107.5 < \bar{X} < 109)$?

23) Un equipo de salvamento submarino está preparándose para explorar un sitio mar adentro en Florida, en el cual se hundió una flotilla de 45 galeones españoles. Según documentos históricos, el equipo espera que éstos naufragios generen un promedio de 225,000 de ingresos cuando se exploren, con una desviación estándar de $ 39,000. Sin embargo, el financiero del equipo se muestra escéptico y ha manifestado que, si los gastos de exploración de $ 2.1 millones no se recuperan en los 9 primeros naufragios, cancelará el resto de la exploración. ¿Cuál es la probabilidad de que la exploración continúe después de los primeros 9 naufragios?
24) Sara Torres encabeza una campaña de recaudación de fondos para un colegio. Quiere concentrarse en la décima reunión de la clase actual, confiando lograr contribuciones de 36 % de los 250 miembros de dicha clase. Los datos anteriores revelan que los que cooperan para el regalo de la reunión donarán 4 % de sus ingresos anuales. Sara piensa que los miembros de la clase tienen un sueldo anual de $ 32 000, con una desviación estándar de $ 9600. Si sus expectativas se cumplen (36 % de la clase dona 4 % de su sueldo anual), ¿ qué probabilidades hay de que el regalo de la décima reunión fluctúe entre $ 110 000 y $ 12 000?
25) Una compañía de alimentos, que tiene 122 supermercados, ha sido adquirida por una gran cadena de supermercados en todo el país. Antes de cerrar el trato, ésta última quiere estar segura de que hará una buena inversión. Ha decidido analizar los registros financieros de 40 de las tiendas de comestibles de la empresa que va a adquirir. La gerencia de ésta afirma que las utilidades de cada establecimiento tienen una distribución aproximadamente normal con la misma media y una desviación estándar de $ 1 000. Si la gerencia tiene razón, ¿qué probabilidades hay de que la media muestral de las 40 tiendas caiga dentro de $ 150 de la media real?

SONRIE ...

1) A propósito de razonamiento deductivo :

JUSTIFICACIÓN MATEMÁTICA DE LA POBREZA MATERIAL

El "Teorema del salario" de Dilbert establece que : << Los científicos nunca pueden ganar tanto como los ejecutivos o los comerciantes >>.

Este teorema es posible demostrarlo matemáticamente a partir de los dos siguientes y evidentes postulados :

Postulado 1 : "El conocimiento es poder".
Postulado 2 : "El tiempo es dinero"

También usamos el axioma : Poder (o Potencia) = Trabajo / Tiempo

DEMOSTRACIÓN :

Como Conocimiento= Poder, entonces Conocimiento= Trabajo / Tiempo
Si tiempo = Dinero, entonces Conocimiento = Trabajo / Dinero.
Resolviendo para "Dinero", obtenemos :

Dinero = Trabajo / conocimiento

Así, si "conocimiento" se aproxima a cero, entonces "Dinero" tiende a infinito independientemente de la cantidad de trabajo hecho.

¡Demostrado! : Cuanto menos sepas, ¡ más ganarás !

2) La generación de números aleatorios es una cuestión demasiado importante como para dejarla al azar (Donald Knuth)

3) En una escuela de un país imaginario, Paolo enseñaba estadística :

" ... Un problema evidente de afirmaciones tales como <<el 67 % (o el 75 %) de los encuestados prefirieron la pastilla X >> es que fácilmente podrían estar basadas en muestras pequeñas de 3 o 4 individuos. Más descarado aún es el caso en que una celebridad avala una dieta, un medicamento o lo que sea; en tal caso tenemos una muestra tamaño 1, que generalmente, además, ha cobrado por ello ... "

4) - ¿ Qué sucede cuando n tiende a infinito?
- que infinito se seca.

CAPITULO 3. ESTIMACIÓN.

1) A propósito de proporciones :

En cierta ocasión le preguntaron a un vendedor que cómo podía vender tan baratos sus sandwiches de conejo, a lo que respondió :

- "Bueno, tengo que admitir que hay un poco de carne de caballo. Pero la mezcla es solo 50:50; uso el mismo número de conejos que de caballos".

2) ¿Cómo se calcula el volumen de una vaca?

INGENIERO : Metemos la vaca dentro de una gran cuba de agua y la diferencia de volumen es el de la vaca.

MATEMÁTICO: Parametrizamos la superficie de la vaca y se calcula el volumen mediante una integral triple.

FÍSICO : Supongamos que la vaca es esférica ...

ESTADÍSTICO : Tomamos como muestra las tripas de la vaca, les medimos longitud y diámetro y de ahí inferimos el volumen del animal.

CAPITULO 3. ESTIMACION.

3.1 Estimación puntual. Características de un buen estimador.

Llegamos a la parte de inferencia estadística, que es donde obtendremos conclusiones acerca de la población analizando las características de una muestra de ella.

Por ejemplo, si quisiéramos saber cuál es el promedio del número de calzado que usan los mexicanos (sea μ dicho promedio), μ es un parámetro porque es característica de la población y es desconocida porque si conociéramos el valor de μ, nada tendríamos que hacer. Tomaríamos una muestra aleatoria de, tal vez, 100 mexicanos, anotaríamos el número de calzado de cada uno y calcularíamos la media aritmética de nuestros datos

$$\overline{x} = \frac{\sum x_i}{100}$$

Este valor particular $\overline{x}$, por ejemplo 7 del estadístico $\overline{X}$, es decir este numerito $\overline{x} = 7$ que obtenemos de nuestra muestra aleatoria particular en la función

$$\overline{X} : \Omega_N \dashrightarrow R$$
$$(x_1, x_2, \ldots, x_n) \mapsto \overline{X}[(x_1, x_2, \ldots, x_n)]$$

podemos tomarlo como estimador de μ; o sea, concluír sin más ni más que la media de la población es $\mu = 7$. Concluír que los millones de mexicanos calzan en promedio del número 7.

Hacer lo descrito anteriormente es hacer lo que en estadística se conoce como "estimación puntual" (un punto, el 7, para estimar el parámetro μ).

¿Qué es entonces la "estimación puntual"? Es la estimación habitual; esto es, el número que se obtiene mediante cálculos a partir de los valores de la muestra y que usamos como aproximación al parámetro que se está estimando, asignando solo un punto a dicho parámetro. Otro ejemplo : la proporción de muestra x/n de votantes que favorecen a un cierto candidato sería una estimación por punto de la proporción p de la población.

La desventaja de la estimación puntual es que acierta o se equivoca. Si se equivoca, ignoraremos el grado de error y no podremos estar seguros de la confiabilidad de la estimación.

Un estadístico, desde el momento en que lo usas para estimar a un parámetro se convierte en un **estimador**; un valor particular (el de una muestra particular) del estadístico sería una **estimación** para el parámetro.

Es muy importante elegir un buen estadístico para estimar a determinado parámetro, ¿tomarías por ejemplo al estadístico

$$s^2 = \frac{\Sigma(x_i - \overline{x})^2}{n}$$

para estimar siempre a la media μ de los mexicanos del ejemplo con que se inició este capítulo? Claramente no, ¿verdad? La intuición nos dice (aunque a veces nos engaña) que $\overline{X}$ es mejor estimador y además, y esto es importante, en el campo lógico-deductivo resulta ser lo que se llama un estimador insesgado para μ, por fortuna.

¿Qué quiere decir que un estimador, por ejemplo $\overline{X}$ sea insesgado? Quiere decir que $\overline{X}$ en promedio, para todas las muestras posibles, le "pega" al parámetro en cuestión, en éste caso μ, como puede verificarse en la tabla de la página 36 donde $\mu_{\overline{x}} = \mu$ (en promedio $\overline{X}$ le "pega" a μ).

Lo mismo puede decirse de $\overline{p}$; éste también es un estimador insesgado (para p), pues $\mu_{\overline{p}}$ = p ($\overline{p}$ en promedio le pega al parámetro p).

Para una población infinita, E(X) se define en términos del cálculo.

Veamos características deseables en un estimador:

INSESGAMIENTO: Es una propiedad conveniente de un buen estimador. El término insesgado se refiere a que en promedio, el estimador le "pega" al parámetro. Por ejemplo, una media muestral es un estimador insesgado de la media de la población, pues la media de la distribución de muestreo de las medias maestrales tomadas de una misma población es igual a la media de esta última. Podemos decir que un estadístico es un estimador insesgado si, en promedio, tiende a asumir valores que se hallan arriba del parámetro de la población que está siendo estimado en el mismo grado y con la misma frecuencia con que tiende a asumir valores que se hallan por debajo del parámetro que está siendo estimado.

EFICIENCIA: Una propiedad muy conveniente de un buen estimador es que sea **eficiente**. La eficiencia designa el tamaño del error estándar del estadístico. Si comparamos dos estadísticos de una muestra del mismo tamaño y tratamos de decidir cuál es el estimador más eficiente, seleccionaremos el estadístico que tenga el error estándar o la desviación estándar más pequeños de la distribución muestral. Supóngase que escogemos una muestra de determinado tamaño y debemos decidir si utilizamos la media muestral o la mediana muestral para estimar la media de la población. Si calculamos el error estándar de la media muestral y descubrimos que es 1.05 y luego obtenemos el error estándar de la mediana muestral y resulta ser 1.6, diremos que la media de la muestra es un estimador más eficiente de la media de la población porque su error estándar es más pequeño. Es natural que un estimador con un error estándar menor (con menos variación) tenga mayores probabilidades de producir una estimación más cercana al parámetro de la población en cuestión.

CONSISTENCIA: Un estadístico es un estimador **consistente** del parámetro de una población si, al aumentar el tamaño de la muestra se logra una seguridad casi absoluta de que el valor del estadístico se acerca mucho al valor del parámetro de la población. Si un estimador es consistente, se torna más confiable en las muestras grandes. En consecuencia, en caso de

dudarse si debemos incrementar el tamaño de la muestra para conseguir más información sobre un parámetro de la población, primero se averigua si el estadístico es un estimador consistente. Si no lo es, perderemos tiempo y dinero al escoger muestras más grandes.

SUFICIENCIA: Un estimador es **suficiente** si utiliza la información contenida en la muestra, al punto que ningún otro estimador podría extraer de ésta última más información referente al parámetro de la población que va a ser estimado.

Un estadístico muestral no siempre es el mejor estimador de su análogo parámetro de población.

Se presentan aquí estos criterios para que conozcas el cuidado con que los estadísticos proceden al seleccionar un estimador.

Cabe mencionar aquí dos cosas :

1) El mejor estimador de la media μ de una población es la media de la muestra $\overline{X}$. $\overline{X}$ es insesgada, consistente, el estimador más eficiente y, mientras la muestra sea lo bastante amplia, su distribución muestral puede ser aproximada por la distribución normal.

2) El mejor estimador de la proporción p de la población es la proporción $\overline{p}$ de la muestra. Es insesgada, consistente, eficiente y suficiente.

3.2 Ejemplos de estimaciones puntuales (de μ, σ^2, σ y p)

Como el mejor estimador puntual de μ ***es*** $\overline{X}$, la utilizaremos en el siguiente ejemplo:

EJEMPLO 10-A) ESTIMACIÓN PUNTUAL DE LA MEDIA POBLACIONAL: Examinemos una compañía de artículos médicos que produce jeringas hipodérmicas desechables. Las jeringas se envuelven en un paquete estéril y luego se meten en una gran caja corrugada. El empaque hace que las cajas contengan distintos números de jeringas. Puesto que las jeringas se venden por unidades, la compañía necesita una estimación del número de las que se incluyen en cada caja para poder hacer la facturación. Hemos tomado una muestra de 35 cajas al azar y hemos registrado el número de jeringas en cada una. La tabla siguiente contiene nuestros resultados, donde la media es

$$\overline{X} = \frac{\sum x_i}{n} = \frac{3570}{35} = 102 \quad \textit{jeringas}$$

101	103	112	102	98	97	93
105	100	97	107	93	94	97
97	100	110	106	110	103	99
33	98	106	100	112	105	100
114	97	110	102	98	112	99

Resultados de una muestra de 35 cajas de jeringas hipodérmicas (jeringas por caja)

Así pues, al emplear como estimador la media de la muestra $\overline{x}$, la estimación puntual de la media de la población μ es 102 jeringas por caja. El precio de fabricación de una jeringa hipodérmica desechable es muy pequeño (unos 25 centavos), por lo cual tanto el vendedor como el comprador aceptarán utilizar esta estimación puntual como base de la facturación y el fabricante se ahorrará el tiempo y dinero que supone contar cada jeringa que se mete a la caja.

Como estimadores de la varianza tenemos a los estadísticos

$$s_1^2 = \frac{\sum(x_i - \overline{x})^2}{n-1} = \frac{\sum x_i^2}{n-1} - \frac{n\overline{x}^2}{n-1}$$

$$s_2^2 = \frac{\sum(x_i - \overline{x})^2}{n}$$

OJO: Si al estimar la varianza de la población se considera s_2^2 en lugar de s_1^2, el resultado sería algún sesgo como un estimador de la varianza de la población; en concreto, tendería a ser demasiado bajo. Si se usa un divisor de n-1 se obtiene un estimador insesgado de μ.

ESTIMACION PUNTUAL DE LA PROPORCIÓN DE LA POBLACIÓN (Con la proporción de la muestra) : La proporción de unidades que poseen una característica particular en determinada población se representa con p. Si conocemos la proporción de unidades de una muestra que tiene esas mismas características (denotada con $\overline{p}$), podemos utilizar ésta última como un estimador de p. Puede demostrarse que $\overline{p}$ tiene todas las propiedades deseables que ya se mencionaron antes : es insesgada, consistente, eficiente y suficiente.

Continuando con el ejemplo del fabricante de artículos médicos, trataremos de estimar la proporción de la población a partir de la proporción de la muestra.

Supóngase que la gerencia desea estimar el número de cajas que llegarán dañadas por un deficiente manejo en el embarque después que las cajas salen de la fábrica. Podemos vigilar una muestra de 50 cajas desde el punto de embarque hasta el arribo a su destino y luego registrar la presencia o ausencia de daño. Si en este caso encontramos que la proporción de cajas dañadas de la muestra es 0.08, podremos afirmar que

$\overline{p} = 0.08$ ---------- proporción dañada de la muestra

Y como la proporción de la muestra $\overline{p}$ es un estimador adecuado de la proporción p de la población, podemos afirmar que la proporción de cajas dañadas en la población será 0.08

EJERCICIOS DE LA SECCIÓN 3.2

1) Un auditorio está considerando la posibilidad de ampliar su capacidad de asientos y necesita conocer el número promedio de personas que asisten a los eventos y la variabilidad de ese número. Se transcribe la asistencia (en miles) a 9 eventos deportivos seleccionados aleatoriamente. Calcule las estimaciones puntuales de la media y la varianza de la población de donde se extrajo la muestra

2) En una muestra de 400 trabajadores de la industria textil, 184 expresaron una insatisfacción extrema ante el plan de modificar las condiciones de trabajo. Dado que la insatisfacción fue lo suficientemente vehemente como para permitir a la gerencia interpretar que la reacción frente al plan era muy desfavorable, quiso conocer la proporción del total de empleados que albergaban ese sentimiento. Dé una estimación puntual de dicha proporción.

3.3 Estimación por intervalo.

La estimación por intervalo es una gama de valores que sirven para estimar el parámetro de una población. Indica el error en dos formas : por el grado de su intervalo y por la probabilidad de que el verdadero parámetro de la población se encuentre dentro de él.

3.3.1 Estimación por intervalo de la media de una población. Muestras grandes.

Ilustraré los conceptos a la luz de un ejemplo:

EJEMPLO 11: Supongamos que el director de investigación de mercados necesita una estimación en meses de la vida promedio de las baterías para automóvil fabricadas por su

compañía. Podemos seleccionar una muestra aleatoria de 200 baterías, registrar los nombres de los dueños de automóviles y sus domicilios que se conservan en los archivos de la tienda y entrevistarlos respecto a la vida de la batería que han usado. La muestra de 200 usuarios tiene una vida media de 36 meses y se sabe que la desviación estándar de la población de las baterías es de 10 meses.

a) Darle al director la estimación que pide.

b) El director también pide una afirmación sobre la incertidumbre que seguramente acompaña a la estimación; es decir, una declaración sobre el intervalo en que posiblemente se halle la media de la población desconocida.

c) Con los requerimientos de los incisos anteriores a) y b) ¿quedaría satisfecho el director?

SOLUCIÓN:

a) Podemos emplear la estimación puntual $\bar{x} = 36$ meses como mejor estimador de la media de la población μ.

b) En este inciso necesitamos encontrar el error estándar* de la media

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{200}} = 0.707 \quad meses$$

Y podemos decirle al director que nuestra estimación de la vida de las baterías de la compañía es de 36 meses, y el error estándar que acompaña a ésta estimación es de 0.707. Es decir, μ puede encontrarse en alguna parte de la estimación por intervalo entre 35.293 y 36.707 meses

c) La información anterior es útil pero insuficiente para el director. A continuación debe calcularse la probabilidad de que la vida real esté en ese intervalo (o en otro de diferente ancho) : 2σ, (2x0.707); 3σ,(3x0.707); etc.

Para calcular dicha probabilidad se procede del siguiente modo:

Sea la población X= Tiempo de vida útil de las baterías (en meses).

Por el TCL

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(\mu, \frac{10^2}{200}\right)$$

El problema se reduce a calcular el área sombreada bajo la curva en la siguiente figura

* No olvidar que se llama "error estándar" a la desviación estándar al referirnos a un estimador.

$$N\left(\mu, \frac{1}{2}\right) = f(\bar{x})$$

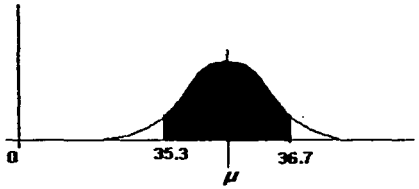


Fig. 3.1 Probabilidad de que la vida real de las baterías esté en el intervalo (35.3, 36.7)

Estandaricemos para poder usar las tablas de la normal estándar N(0,1)

$P(35.293 \leq \mu \leq 36.707) = P(-36.707 \leq -\mu \leq -35.293) =$

$$P\left(\frac{\bar{x} - 36.707}{0.707} \leq \frac{\bar{x} - \mu}{0.707} \leq \frac{\bar{x} - 35.293}{0.707}\right) =$$

$$P\left(\frac{36 - 36.707}{0.707} \leq Z \leq \frac{36 - 35.293}{0.707}\right) =$$

$$P(-1 \leq Z \leq 1) = 0.6826$$

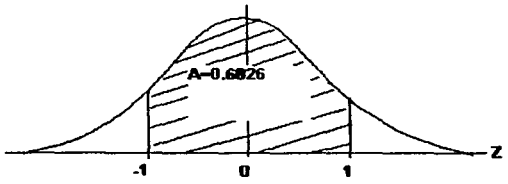


Fig. 3.2 Esta es la figura con traslación correspondiente a la fig. 3.1

Y presentaríamos nuestro informe final de las baterías al director de mercadotecnia del siguiente modo:

"Nuestra mejor estimación de la vida de las baterías de la compañía es de 36 meses, y tenemos una confianza de 68.3 % de que la vida de ellas se encuentra en el intervalo comprendido entre 35.293 y 36.707 meses $(36 \pm 1\sigma_{\bar{x}})$"

Se llama nivel de confianza a la probabilidad que asociamos a una estimación de intervalo. Esta probabilidad indica la confianza que tenemos de que la estimación por intervalo comprenda el parámetro de la población. Una probabilidad mayor significa más confianza. Los niveles más utilizados son 90, 95 y 99 % pero se puede aplicar cualquier otro.

En la práctica, los altos niveles de confianza producen grandes intervalos de confianza y éstos no son precisos, dan estimaciones muy confusas.

Ahora atendamos un asunto que podría confundirnos:

Si en el ejemplo de las baterías que nos ocupamos, un estadístico dijera:

" Tenemos una confianza de 95 % de que la vida media de las baterías de la población fluctúa entre 30 y 42 meses " ¿en qué error podrían caer los oyentes?

Hay que advertirle que ésta afirmación **no significa que haya 0.95 de probabilidades de que la vida media de todas las baterías caiga dentro del intervalo establecido a partir de esta muestra. Como μ es un parámetro y está fijo, o bien está en el intervalo que construimos o bien no está ahí. Es decir, la probabilidad de que nuestro intervalo contenga a μ es 0 o 1.**

Lo que sí significa es que si seleccionamos muchas muestras aleatorias de ese tamaño y si calculamos el intervalo de confianza de todas ellas, en 95 % de estos casos, la media de la población μ, se halla dentro de dicho intervalo. Y esta interpretación es la que se da a cualquier estimación por intervalo.

EJEMPLO 12: En un experimento para determinar el número promedio de latidos del corazón por minuto para cierta población, bajo las condiciones del experimento, se encontró que el número promedio de latidos por minuto para 49 sujetos era de 130. Si resulta razonable suponer que estos 49 pacientes constituyen una muestra aleatoria y que la población está distribuída normalmente, con una desviación estándar de 10, encontrar:

a) El intervalo de confianza del 90 % para μ.
b) El intervalo de confianza del 95 % para μ.
c) El intervalo de confianza del 99 % para μ.

SOLUCIÓN: El procedimiento de solución es el siguiente :

Sea la población X= Número de latidos del corazón por minuto.
Como $X \cong N(\mu, 10)$, por el TCL

$$\bar{X} \cong N\left(\mu, \frac{10^2}{49}\right) \quad porque \quad \bar{X} \cong N\left(\mu, \frac{\sigma^2}{n}\right)$$

Estandaricemos para poder usar las tablas de la distribución normal estándar:

$$\frac{\bar{X}-\mu}{\sigma_{\bar{x}}} \cong N(0,1). \; Es\ decir,\ Z = \frac{\bar{X}-\mu}{10/7} \cong N(0,1)$$

Viendo en tablas de la normal estándar:

a) $P(-z^* \le z \le z^*) = 0.9 =$

$$= P\left(-z^* \le \frac{\bar{X}-\mu}{10/7} \le z^*\right); \quad z^* \approx 1.645. \quad Así\ que$$

$$-z^*\left(\frac{10}{7}\right) \le \bar{X}-\mu \le z^*\left(\frac{10}{7}\right).$$

$$-1.645\left(\frac{10}{7}\right) \le \bar{X}-\mu \le 1.645\left(\frac{10}{7}\right)$$

$$-1.645\left(\frac{10}{7}\right) \le \mu-\bar{X} \le 1.645\left(\frac{10}{7}\right)$$

$$\bar{X}-1.645\left(\frac{10}{7}\right) \le \mu \le \bar{X}+1.645\left(\frac{10}{7}\right)$$

Es decir, $\mu \in \bar{X} \pm 1.645(10/7)$. El intervalo pedido en éste inciso es $(\bar{x} \pm z^* \sigma_{\bar{x}})$:
130 ± 1.645(10/7).
Dicho de otro modo :

$$P(130-1.645(\frac{10}{7}) < \mu < 130+1.645(\frac{10}{7})) = 0.9$$

Pero esto no quiere decir que nuestro intervalo particular, de una muestra específica, contiene a μ con probabilidad 0.9, sino que al muestrear muchas veces y construír el intervalo $(\bar{x} \pm z^* \sigma_{\bar{x}})$, el 90 % de las veces, dicho intervalo contendrá a μ.
Pues

$$P(\mu \in 130+1.645(\frac{10}{7})) = 0 \ \ o \ \ 1.$$

Así que el intervalo del 90 % de confianza para μ es 130 ± 2.35, es decir :

(127.65, 132.35). Los valores que delimitan al intervalo se **llaman límites de confianza.** 127.65 es el **límite inferior de confianza** y 132.35 es **el límite superior de confianza.**

Resumiendo lo anterior podemos decir entonces, en general, que una estimación de intervalo puede expresarse como sigue:

(Estimador) ± (factor de confiabilidad)x(error estándar del estimador).

$$\theta \pm Z_{(1-\alpha/2)}\sigma_\theta$$

En los ejemplos que nos ocupan $\theta = \bar{X}$ y el intervalo es del (1- α) % de confianza Con $0<\alpha<1$. (1- α) es llamado **coeficiente de confianza.**

La gráfica correspondiente al inciso que acabamos de resolver es

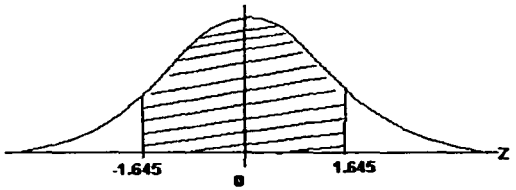


Fig. 3.3 gráfica del inciso a) ejemplo 12)

b) Se quiere un intervalo del 100(1-α)% = 100(1-0.05)% para μ.
$\alpha = 0.05$, $\alpha/2 = 0.025$; 1- $\alpha/2 = 0.975$. El intervalo buscado es

$$\bar{X} \pm Z_{(1-\alpha/2)}\sigma_{\bar{x}}; \quad O\ sea \quad \bar{X} \pm Z_{.975}\sigma_{\bar{x}}$$

Consultando tablas de la normal estándar y sustituyendo valores, tenemos:

130 ± 1.96(10/7). O sea 130 ± 2.8

Es decir, el intervalo pedido es (127.2, 132.8)
c) En este caso se quiere un intervalo del 100 (1-0.01)% para μ
Es decir $\alpha = 0.01$, $\alpha/2 = 0.005$, 1- $\alpha/2 = 0.995$

Como sabemos, el intervalo es

$$\bar{X} \pm Z_{.995}\sigma_{\bar{x}}$$

O sea, consultando en tablas: 130 ± 2.575(10/7). Es decir, 130 ± 3.679
El intervalo pedido es (126.321, 133.679)

EJERCICIOS DE LA SECCIÓN 3.3.1

1) Se encontró que el nivel indirecto medio de bilirrubina en el suero de 16 niños de 4 días de nacidos era de 5.98 mg/100 cc. Suponiendo que los niveles de bilirrubina en los niños de 4 días de nacidos están distribuídos aproximadamente en forma normal con una desviación estándar de 3.5 mg/100 cc. Encontrar

   a) El intervalo de confianza del 90 % para μ.
   b) El intervalo de confianza del 95 % para μ.
   c) El intervalo de confianza del 99 % para μ.

2) En un estudio de duración de la hospitalización realizado por varios hospitales en cooperación, se extrajo una muestra aleatoria de 64 pacientes de úlcera péptica, de una lista de todos los pacientes de úlcera péptica admitidos alguna vez en los hospitales participantes y se determinó, para cada uno, la duración de la hospitalización. Se encontró que la duración media de hospitalización fue de 8.25 días. Si se sabe que la desviación estándar de la población es de tres días, hallar:

   a) El intervalo de confianza del 90 % para μ.
   b) El intervalo de confianza del 95 % para μ.
   c) El intervalo de confianza del 99 % para μ.

3) Una muestra de 100 adultos del sexo masculino aparentemente normales, de 25 años de edad, tuvo una presión sistólica sanguínea media de 125. Si se tiene la sensación de que la desviación estándar de la población es de 15, encontrar:

   a) El intervalo de confianza del 90 % para μ.
   b) El intervalo de confianza del 95 % para μ.

3.3.2 Estimación por intervalo de la proporción de una población. Muestras grandes.

La binomial es la distribución correcta que ha de utilizarse al construír los intervalos de confianza para estimar una proporción de la población. Pero como es tedioso, si tanto np como nq son al menos 5 puede usarse la normal como sustituto de la binomial.

Como en la binomial μ = np = número medio de éxitos, al dividir entre n obtenemos la proporción de éxitos:

$$\frac{\mu}{n} = p. \quad \textit{Es decir} \quad \mu_{\bar{p}} = p \quad \textit{donde } p = \textit{probabilidad de éxito}.$$

Igualmente podemos modificar la desviación estándar de la binomial, $\sqrt{npq}$

Para pasar del tratamiento del número de éxitos al tratamiento de la proporción de éxitos: Dividimos entre n

$$\frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}}. \quad \textit{Es decir, } \sigma_{\bar{p}} = \sqrt{\frac{pq}{n}} = \textit{error estándar de la proporción. Así que } \bar{p} \approx N\left(p, \frac{pq}{n}\right)$$

Y se obtiene un intervalo de confianza mediante la fórmula general:

Estimador ± (Coeficiente de confiabilidad)x(error estándar)

En este caso, de las proporciones, el intervalo del 100(1- α)% de confianza para p es

$$\bar{p} \pm Z_{(1-\alpha/2)}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \quad \text{pues} \quad \sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$$

$$\text{se estima por medio de} \quad \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Ilustremos lo anterior con un ejemplo

EJEMPLO 13: Un encargado del archivo de expedientes médicos extrajo una muestra aleatoria de 100 expedientes de pacientes y encontró que en el 8 % de ellos, la carátula tenía, al menos, un detalle de información que contradecía a la demás información que aparecía en el

expediente. Construír los intervalos de confianza del 90, 95 y 99 por ciento para la proporción verdadera de los expedientes que contienen tales discrepancias.

SOLUCIÓN:

$$\textit{Como} \quad \bar{p} \approx N\left(p, \frac{pq}{n}\right), \quad \bar{p} \approx N\left(p, \frac{pq}{100}\right)$$

$$\bar{p} = 0.08$$

Para el intervalo del 90 %, $\alpha = 0.10$, $\alpha/2 = 0.05$, $1 - \alpha/2 = 0.95$.
Estandaricemos para poder usar las tablas de la normal

$$Z = \frac{\bar{p} - p}{\sqrt{\frac{pq}{100}}} \approx N(0,1), \quad \textit{Entonces}$$

$$P\left(-Z_{.95} \leq \frac{\bar{p} - p}{\sqrt{\frac{pq}{100}}} \leq Z_{.95}\right) = 0.9$$

Procediendo de modo análogo a como se hizo en el caso de la media se llega a

$$P\left(\bar{p} - Z_{.95}\sqrt{\frac{pq}{100}} \leq p \leq \bar{p} + Z_{.95}\sqrt{\frac{pq}{100}}\right) = 0.9$$

Es decir

$$P\left(p \in \bar{p} \pm Z_{.95}\sqrt{\frac{pq}{100}}\right) = 0.9$$

El intervalo de confianza que nos ocupa ya aparece en la última expresión anterior. Solo hay que sustituírle los valores correspondientes.
El intervalo del 90 % se obtiene sustituyendo valores en

$$\bar{p} \pm Z_{(1-\alpha/2)}\sqrt{\frac{pq}{n}}. \quad \bar{p} \pm Z_{.95}\sqrt{\frac{\bar{p}(1-\bar{p})}{100}}$$

y consultando las tablas de la normal estándar. Observamos que $Z_{.95} = 1.645$, por lo tanto

$$0.08 \pm 1.645\left(\frac{\sqrt{(0.08)(0.92)}}{10}\right)$$

Haciendo la aritmética correspondiente llegamos a que el intervalo pedido en este inciso es (0.0354, 0.1246).

Para el intervalo del 95 % : 0.95 = (1-0.05), así que $\alpha = 0.05$, $\alpha/2 = 0.025$ y $1-\alpha/2 = 0.975$.

Sabiendo cuál es el intervalo, y consultando la tabla de la normal estándar, tenemos:

$$0.08 \pm Z_{.975}\left(\frac{\sqrt{(0.08)(0.92)}}{10}\right) = 0.08 \pm 1.96(0.027129)$$

$$o\ sea \quad (0.02683, 0.13317)$$

Por último, para el intervalo del 99% para p, $\alpha = 0.01$, $\alpha/2 = 0.005$, $1-\alpha/2 = 0.995$.

Sustituyendo valores en la fórmula correspondiente y consultando la tabla de la normal estándar:

$$0.08 \pm 2.575\sqrt{\frac{(0.08)(0.92)}{100}} = 0.08 \pm 0.0699$$

O sea que el intervalo del 99 % para p es (0.0101, 0.1499)

Estos intervalos se interpretan del mismo modo que en el caso de las medias.

EJERCICIOS DE LA SECCIÓN 3.3.2

1) Una encuesta, que condujo a una muestra aleatoria de 150 familias en cierta comunidad urbana, reveló que, en el 87 % de los casos, al menos uno de los miembros de la familia tenía alguna forma de seguro relacionada con la salud. Construír los intervalos de confianza del 90, 95 y 99 % para p, la proporción verdadera de familias en la comunidad con la característica de interés.
2) En un estudio diseñado para averiguar la relación entre cierto medicamento y cierta anomalía en los embriones de polluelos, se inyectaron con el medicamento 50 huevos fertilizados, en el cuarto día de incubación. En el vigésimo día de incubación, se examinaron los embriones y se observó la presencia de la anormalidad en 12 de ellos. Encontrar el intervalo de confianza del 90, 95 y 99 por ciento para p.

3.3.3 La distribución t de Student.

Hay ocasiones en que se desconoce la varianza de la población y el tamaño de la muestra n es menor que 30. En tales casos, dicha varianza es estimada por la de la muestra y dicha estimación es usada en nuestro trabajo estadístico. Se utiliza en éstos casos la cantidad

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

La cual tiene una distribución llamada t de Student con las siguientes propiedades:

1) Tiene una media de cero.
2) Es simétrica respecto a la media.
3) En general, tiene una varianza mayor que 1, pero la varianza tiende hacia 1 a medida que el tamaño de la muestra se hace grande.
4) La variable t toma valores desde $-\infty$ hasta $+\infty$.
5) En realidad la distribución t es una familia de distribuciones ya que se tiene una distribución diferente para cada valor de n-1, el divisor usado al calcular s, correspondiente a la muestra
6) Comparada con la distribución normal, la distribución t es menos alta en el centro y tiene colas más altas

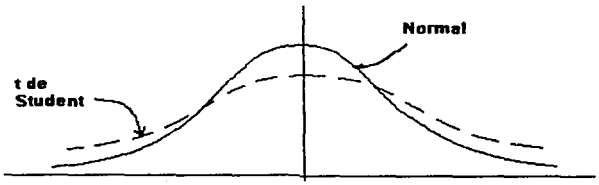


Fig.3.4 Comparación de la distribución normal y la distribución t

Al usar la distribución t suponemos que la distribución de la población es normal o aproximadamente normal.

En la práctica se usan tablas de la t-Student, ya que existen.

La cantidad n-1, usada para calcular la varianza, se conoce como grados de libertad; existe una distribución t diferente para cada valor de los grados de libertad y éstos deben tomarse en cuenta cuando se usa la tabla de la distribución t (la tabla viene al final de este trabajo, en el apéndice).

El intervalo de confianza sigue siendo (estimador)± (coeficiente de confiabilidad) (error estándar).

Más precisamente

$$\bar{X} \pm t_{(1-\alpha/2)}\left(\frac{s}{\sqrt{n}}\right)$$

EJEMPLO 14: Una muestra de 25 niños de 10 años proporcionaron un promedio y una desviación estándar de 36.5 y 5 kg respectivamente. Suponiendo una población normalmente distribuída, encontrar los intervalos de confianza del 90, 95 y 99 por ciento para la media de la población de la cual provino la muestra.

SOLUCIÓN: Sea X= Peso de los niños (en kilogramos)

$$X \equiv N(\mu, \sigma^2), \quad n = 25$$

Como se desconoce la varianza de la población y el tamaño de la muestra es pequeño (menor que 30) la distribución que debemos usar es la t-Student. Como sabemos de lo expuesto más arriba, el intervalo de confianza es:

$$\bar{X} \pm t_{(1-\alpha/2)}\left(\frac{s}{\sqrt{n}}\right)$$

a) Para 90 %: $100(1-\alpha)$ %, $\alpha = 0.1$, $\alpha/2 = 0.95$, $n-1 = 25-1 = 24$.

Sustituyendo valores en la última fórmula anterior y consultando la tabla de la t-Student que viene en el apéndice de ésta obra, el intervalo es:

$$\bar{X} \pm t_{.95}\left(\frac{5}{\sqrt{25}}\right) = 36.5 \pm 1.7109(1)$$

Es decir, el intervalo es (34.7891, 38.2109)

b) Para 95 % : $100(1-\alpha)$%, $\alpha = 0.05$, $\alpha/2 = 0.025$, $1-\alpha/2 = 0.975$, $n-1 = 24$
El intervalo es, consultando tablas de la t :

$$\bar{X} \pm t_{.975}\left(\frac{5}{5}\right) = 36.5 \pm 2.0639(1)$$

Es decir, el intervalo del 95 % de confianza es (34.44, 38.56)

c) Para 99%: $100(1-\alpha)$%, $\alpha = 0.01$, $\alpha/2 = 0.005$, $1-\alpha/2 = 0.995$, $n-1 = 24$.
Procediendo como en los dos incisos anteriores, el intervalo es:

$$36.5 \pm t_{.995}\left(\frac{5}{\sqrt{25}}\right) = 36.5 \pm 2.7969(1)$$

El intervalo del 99% de confianza para p es (33.7031, 39.2969)

A continuación se resume qué distribución usar, según tamaño de muestra, σ conocida o desconocida y distribución normal o no normal.

Población normal, $n \geq 30$, σ^2 conocida : Usar Z.
Población normal, $n \geq 30$, σ^2 desconocida : Usar t o Z.
Población normal, $n < 30$, σ^2 conocida : Usar Z.

Población normal, $n < 30$, σ^2 desconocida : Usar t.
Población no normal, $n \geq 30$, σ^2 conocida : Usar Z (por el TCL).
Población no normal, $n \geq 30$, σ^2 desconocida : Usar Z (por el TCL).
Población no normal, $n < 30$, σ^2 conocida : Nada puede hacerse; o mejor dicho, lo que puede hacerse queda fuera del alcance de esta obra.
Población no normal, $n < 30$, σ^2 desconocida : Nada puede hacerse; o mejor dicho, lo que puede hacerse queda fuera del alcance de esta obra.

Ejercicios de la sección 3.3.3

1) A nueve pacientes que sufren de la misma incapacidad física, pero de lo contrario comparables, se les pidió que llevaran a cabo cierta tarea como parte de un experimento. El tiempo promedio requerido para realizar la tarea fue de 7 minutos, con una desviación estándar de dos minutos. Suponiendo normalidad, construír los intervalos de confianza del 90, 95 y 99 por ciento para el tiempo medio verdadero requerido para realizar la tarea por este tipo de paciente.
2) Un administrador de un hospital tomó una muestra de 25 cuentas vencidas, de las cuales calculó una media de $ 2 500 y una desviación estándar de $ 750. Suponiendo que las cantidades de todas las cuentas vencidas están distribuídas normalmente, encontrar los intervalos de confianza del 90, 95 y 99 por ciento para μ.

3.4 MISCELÁNEA DE EJERCICIOS DEL CAPÍTULO 3.

1) Una televisora selecciona una muestra de 900 casas y calcula la proporción de éstas que tienen un aparato de televisión a color. Si la proporción real p es 0.6, ¿cuál es el error estándar de $\bar{p}$?

2) Un investigador médico calcula la presión sistólica sanguínea promedio de 100 varones adultos. Supongamos que $\sigma = 3$ mmHg ¿cuál es el error estándar de la media muestral?
3) Una oficina local de escuelas debe estimar el número de niños por familia. De una muestra de 100 familias, se calcula la media y la desviación estándar. Si s=1.2,¿ cuál es el error estándar de la media muestral?¿cuáles son los límites de confianza?
4) Para estimar el tiempo de servicio en el Burger Boy, el encargado del negocio anotó los tiempos que requirieron 35 personas, de una muestra aleatoria, para tomar una orden estándar (2 hamburguesas, 2 paquetes de papas y 2 refrescos). En promedio éstas personas requirieron 72.2 segundos con una desviación estándar de 12.8 seg para tomar las órdenes, ¿qué puede afirmar el encargado con una confianza del 95% acerca del error máximo, si usa la media muestral de 72.2 segundos como una estimación del tiempo promedio requerido para tomar esta orden?
5) Supóngase que los tiempos que los estudiantes tardan en llegar a la escuela utilizando su bicicleta se distribuyen normalmente. Una muestra aleatoria de 35 estudiantes proporcionó una media de 32 minutos y una desviación estándar de 9 minutos. Determina los límites de confianza del 98% para la media poblacional.
6) Supóngase que el peso de los bebés al nacer se distribuye normalmente. En un hospital se desea determinar el peso medio de la población de bebés con un intervalo de confianza del 95 %, con base en la evidencia de su muestra. Su media muestral fue de 3 kg con una desviación estándar de 1 kg, en una muestra de 40 bebés.
7) Las longitudes de 200 peces capturados en una laguna tuvieron una media de 35.7 cm. La desviación estándar poblacional es de 6.3 cm.

   a) Encuentra un intervalo de confianza del 90 % para la longitud media poblacional
   b) Encuentra un intervalo de confianza del 98 % para la longitud media poblacional.

8) En un estudio sobre jugadores de fútbol un cardiólogo estima que el promedio real de los latidos cardíacos, en estado de reposo es de 63 a 70 latidos por minuto; con una confianza del 95 % ¿cuál es el valor de t si nos basamos en 22 grados de libertad?
9) Se preguntó a 10 niños menores de 12 años cuántas horas semanales dedicaban a ver televisión, los resultados fueron: 28, 34, 42, 35, 16, 32, 38, 40, 39, 31. Determina el intervalo de confianza del 90 % para estimar el número medio de horas empleadas para ver televisión.
10) La compañía Marinela desea estimar la proporción de consumidores que prefieren su marca. Los agentes de la compañía observan a 450 compradores. Del número total observado, 300 compraron Marinela. Calcula un intervalo de confianza del 99 % para la verdadera proporción de compradores que prefieren dicha marca.
11) Una auditoría aleatoria sobre ciertas operaciones en una industria, arrojó para una muestra de 50 de ellas un error promedio de $ 150.00 con un error estándar de $ 60.00. Calcular los límites de confianza al 95 % para μ (error promedio verdadero del total de operaciones) suponiendo que X se distribuye normalmente.
12) Los inversionistas que adquieren acciones que producen ingresos se interesan únicamente en las que producen réditos anuales por su tipo de acciones ordinarias.

Para ayudar a inversionistas de éste tipo, la más importante firma de corredores de bolsa selecciona aleatoriamente 50 acciones ordinarias. El promedio anual y la desviación estándar de los réditos por obligaciones en los 5 años anteriores ha mostrado que $\overline{x}$ = 8.71 % y σ =2.1 %. Estime el verdadero promedio anual de réditos para ésta clase de acciones, usando un intervalo de confianza del 90 %

13) Si se utiliza una variable aleatoria X para representar el peso de una persona que hace un viaje en avión y se desea conocer μ, esto es, la media del peso de todos los pasajeros del avión hacemos lo siguiente: Puesto que hay limitaciones de tiempo y dinero para pesar a todos, se toma una muestra aleatoria de 36 pasajeros obteniéndose una media muestral $\overline{x}$ = 160 libras. Supóngase además que la distribución de los pesos de los pasajeros del avión tenga una desviación estándar de 30 ¿cuál sería entonces el valor probable de μ si la probabilidad de estimación correcta es 0.95?

14) Una industria de aparatos eléctricos ha inventado una lámpara incandescente nueva y de mayor duración. Los especialistas de la compañía piensan que el período de vida de este nuevo modelo presenta la misma varianza que el anterior σ^2 =16 pero no saben cuál será el período de vida media μ (μ representa la media de la población que consiste de todas las lámparas incandescentes nuevas). Para calcular μ, toman una muestra aleatoria de tamaño n= 100 y dejan funcionando las lámparas de esa muestra hasta que se funden. Si la media de la muestra resulta ser $\overline{x}$ = 150 horas, encuentra el intervalo de confianza del 95 % para μ.

15) Para investigar el aprovechamiento potencial de un producto en servicio, generalmente se incluye la posibilidad de estudiar medidas de la demanda. En el estudio de las posibilidades de extender la programación de la televisión comercial, un investigador ha encontrado que 76 de 180 familias propietarias de televisores seleccionadas al azar, miran por lo menos dos horas a la semana programas de televisión. Encontrar un intervalo del 90 % de confianza de familias en la población que miran programas de televisión.

16) Una industria desea conocer la proporción de amas de casa que en el DF preferirían una licuadora moderna. Se toma al azar una muestra de 100 amas de casa y las entrevista un investigador de mercados de la compañía. De las 100 amas de casa, 20 dicen que les gustaría la licuadora y 80 dicen que no, ¿Cuál sería entonces la verdadera proporción p si se desea un intervalo de confianza del 95 %?

17) Una investigación reciente reporta que el 82 % de 1200 residentes de Sao Paulo, Brasil consideran que la contaminación del aire de su ciudad es muy seria. Encontrar un intervalo del 95 % de confianza para estimar qué proporción de la población de Sao Paulo considera que la contaminación del aire de su ciudad es muy seria.

18) Los artículos defectuosos resultan ser costosos a un fabricante en función del costo de reemplazo y de la pérdida de la confianza en el producto entre el público consumidor. Un fabricante de radios portátiles cree que no más del 10 % de los productos manufacturados por la firma son defectuosos. Si el fabricante desea estimar la proporción actual de radios defectuosos con una probabilidad del 95.45 %

y que el error de estimación sea de 0.03, ¿ qué tan grande debe seleccionar la muestra de radios terminados?

19) Un fabricante de refacciones de autos cree que aproximadamente el 5 % de su producto tiene fallas. Si desea estimar el porcentaje verdadero con margen de 0.5 % y tomar la certeza de estar correcto con una probabilidad de de 0.99, ¿qué tan grande debe ser la muestra para satisfacer tales cifras?

20) Mantener los precios fijos puede hacerse costoso si el promedio de cuentas por cobrar cae por debajo de cierto nivel. El administrador de una tienda de departamentos desea estimar el monto promedio de cuentas por cobrar por mes de sus clientes con cuentas por cobrar dentro de $ 2.50, con una probabilidad de aproximadamente 0.95, ¿ cuántas cuentas debe seleccionar de los registros de la tienda, si la desviación estándar mensual se sabe por balances contables es de $ 7.50?

21) Una muestra aleatoria de n= 24 productos en una tienda de abarrotes enseña una diferencia en contra de valores registrados de los productos. La media y la desviación estándar de las diferencias en contra de valores registrados para los 24 productos son $ 37.40 y $ 6.42 respectivamente. Encontrar un intervalo del 95 % de confianza para la diferencia promedio en contra de los actuales valores registrados por producto en la tienda de abarrotes.

22) Un vendedor industrial obtuvo los siguientes ingresos por ventas diarias en nueve días diferentes: $ 58.00, $ 42.00, $ 71.00, $ 59.00, $ 66.00, $ 49.00, $ 68.00, $ 63.00 y $ 55.00. Suponer que estos ingresos representan una muestra aleatoria de una población normal. Construír un intervalo de confianza al 95 % para la ganancia diaria del vendedor si para los datos de la muestra anterior $\overline{x}$ = $ 59.00 y una desviación estándar de s = $ 9.33

23) Se desea obtener un intervalo de confianza al 99% para el tiempo medio requerido para realizar una tarea. Se sabe que el tiempo requerido sigue una distribución normal. Una muestra aleatoria de 16 mediciones de tiempo produce los resultados siguientes: $\overline{x}$ = 13 mns y s= 5.6 mns.

SONRIE ...

1) 2 + 2 = ?
Ingeniero: 3.9968743
Físico: 4.000000004 ± 0.00000006
Matemático: Espere,solo unos minutos más, ya he probado que la solución existe y es única, ahora la estoy acotando ...
Filósofo: ¿ Qué quiere decir 2 + 2?
Lógico: Defina mejor 2 + 2 y le responderé
Estadístico: Con cierta confianza estimo que 2 + 2 = 4

2) En un examen se les pide a los estudiantes que demuestren que todos los números impares son primos
Matemático: Se da cuenta que el enunciado es falso, pero tiene que demostrarlo, así que escribe " 3 es primo, 5 es primo, 7 es primo y por inducción, todos los números impares son primos"
Físico: También "se da cuenta" de que es falso ...
" 3 es primo, 5 es primo, 7 es primo y por inducción, todos los números impares son primos. Nota: al llegar al 9 se obtiene un error experimental"
Ingeniero: " 3 es primo, 5 es primo, 7 es primo, 9 es primo y por inducción todos los números impares son primos"
Programador de computadoras: " 3 es primo, 5 es primo, 7 es primo, 7 es primo, 7 es primo, 7 es primo, 7 es primo, 7 es primo, ..."
Político: " 3 es primo, 7 es primo, y por lo tanto todos los números impares son primos, de acuerdo con la doctrina del partido. Esta verdad ha sido revelada al gran líder y campeón de la paz. Aquél que no esté de acuerdo es un conspirador contrarrevolucionario".
Médico: " 3 es primo, 5 es primo, 7 es primo y a los demás se les aplica el mismo tratamiento hasta que se curen"
Estadístico: Se da cuenta que el enunciado es falso, pero debe demostrarlo:
"... De toda la población de números primos obtuve aleatoriamente la muestra suficientemente grande siguiente :
{3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 67, 71, 73, 79, 83, 89, 97, 101, 103, 107, 109, 113, 127, 131, 137, 139, 149, 151, 157, 163, 167, 173, 179,181, 191, 193, 197, 199} de donde podemos asegurar con confianza que todos los números impares son primos".

CAPITULO 4. PRUEBAS DE HIPÓTESIS.

SONRIE ...

1) Guía de bolsillo de la ciencia moderna

i) Si es verde o repta, es biología.
ii) Si huele mal, es química.
iii) Si no funciona, es física.
iv) Si no se entiende, es matemáticas.
v) Si no tiene sentido, es económicas o psicología.
vi) Si hay que empezar adivinando con riesgo de equivocarse, es pruebas de hipótesis estadísticas.

2) Dos matemáticos ingleses en vacaciones visitan los Estados Unidos y alquilan un coche. Mientras van por la carretera, oyen esta advertencia por radio : " maneje con precaución. Piense que, según las estadísticas, este fin de semana habrá 400 norteamericanos muertos en los caminos por accidentes de tránsito".
Entonces uno de ellos visiblemente nervioso comenta
- Estamos en riesgo de muerte
A lo que el otro le contesta
- Ni te preocupes, estaríamos en riesgo de muerte bajo la fuerte hipótesis de que fuéramos norteamericanos.

CAPITULO 4. PRUEBAS DE HIPÓTESIS.

4.1 Conceptos básicos del procedimiento de la prueba de hipótesis.

Una hipótesis estadística es una suposición acerca de la expresión matemática del modelo; en otras palabras: es una hipótesis sobre la distribución de una variable aleatoria. Generalmente, en problemas estadísticos, una hipótesis especifica uno o más parámetros asociados con la población.

A veces es difícil formular una hipótesis para un problema dado. Sin embargo, suponiendo que se ha dado este paso, nuestro objetivo inmediato es probar la hipótesis. Para realizar esto, observamos un estadístico cuya distribución muestral se conoce si la hipótesis es verdadera. Algunos valores del estadístico pueden llevarnos a sospechar que la hipótesis no es razonable y debe ser rechazada. Otros valores pueden considerarse como justificación de la hipótesis. Sin embargo, la obtención de un valor razonable del estadístico no demuestra que la hipótesis sea verdadera; simplemente no la contradice.

Lo anterior y otros conceptos básicos serán expuestos con un ejemplo :

EJEMPLO 15: El contrato del techado de un nuevo complejo deportivo del Distrito Federal ha sido concedido a PA-construcciones, una gran compañía constructora. Las especificaciones del edificio exigen un techo móvil cubierto aproximadamente por 10,000 láminas de aluminio de 0.04 pulgadas de espesor. Las láminas no deben tener un espesor mucho mayor que 0.04 pulgadas, debido a que entonces la resistencia del techo sería insuficiente. A causa de ésta restricción del espesor, la compañía constructora examina cuidadosamente las hojas de aluminio recibidas de su proveedor. Desde luego, la empresa no quiere medirlas todas, por lo cual aleatoriamente muestrea 100. Las láminas de la muestra presentan un espesor medio de 0.0408 pulgadas. Por la experiencia pasada con su proveedor, PA piensa que provienen de una población de espesor con una desviación estándar de 0.004 pulgadas. Basándose en estos datos, debe decidir si las 10,000 láminas cumplen con las especificaciones. ¿debe PA aceptar el embarque de láminas?

SOLUCIÓN: Las láminas de aluminio para el techo deben tener un espesor promedio de 0.04 pulgadas y resultarían insatisfactorias si fueran demasiado gruesas o muy delgadas. El contratista toma una muestra de 100 hojas y determina que el espesor medio de la muestra es de 0.0408 pulgadas. Basándose en su experiencia, sabe que la desviación estándar de la población es de 0.004 pulgadas. ¿ Indica ésta evidencia de la muestra que el lote de 10,000 láminas de aluminio es idónea para construír el techo del nuevo complejo deportivo?

Formulación de la hipótesis: Si suponemos que el verdadero espesor de la media es 0.04 pulgadas y si sabemos que la desviación estándar de la población es de 0.004 pulgadas, ¿qué probabilidades habrá de que obtengamos de la población una media muestral de 0.0408

o más? En otras palabras, si la verdadera media es de 0.04 pulgadas y la desviación estándar es de 0.004 pulgadas ¿qué probabilidades habrá de obtener una media muestral que difiera de 0.04 pulgadas por 0.0008 pulgadas o más?

Las preguntas anteriores muestran que, para determinar si la media de la población es realmente 0.04 pulgadas, debemos calcular la probabilidad de que una muestra aleatoria con una media de 0.0408 sea seleccionada de una población con una μ de 0.04 pulgadas y una σ de 0.004 pulgadas. Esta probabilidad indicará si es razonable observar una muestra como ésta si la media de la población es en realidad de 0.04 pulgadas. Si dicha probabilidad es demasiado baja, hemos de concluir que la afirmación de la compañía sobre el aluminio es falsa y que el espesor medio de las láminas no es de 0.04 pulgadas.

Plantearé una pregunta con la siguiente figura:

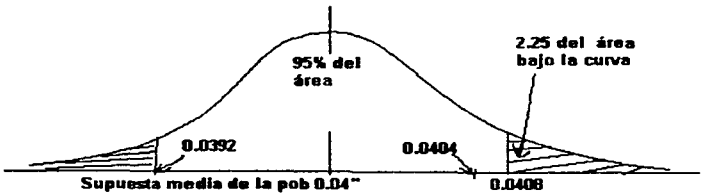


Fig. 4.1

Si la media supuesta de la población es de 0.04 pulgadas y la desviación estándar de la población es de 0.004 pulgadas ¿Cuáles son las probabilidades de obtener una media muestral (0.0408 pulgadas) que difiera de 0.04 pulgadas en 0.0008 pulgadas o más?

O sea

$$Si \quad \bar{X} \approx N\left(0.04, \left(\frac{0.004}{\sqrt{100}}\right)^2\right) = N(0.04, (0.0004)^2)$$

$$¿Cuál\ es \quad P(|\bar{X} - 0.04| \geq 0.0008)?$$

$$¿Cuál\ es \quad P(\bar{X} \geq 0.0408)?$$

Para el cálculo, estandarizamos

$$Z = \frac{\bar{X} - 0.04}{0.0004} = \frac{0.0408 - 0.04}{0.0004} = 2 \textit{ errores estándar de la media}$$

Interpretación de la probabilidad asociada a esa diferencia.- Usando tablas de la normal, descubrimos que 4.5 % es la probabilidad total de que la media muestral difiera de la media de la población por dos o más errores estándar; es decir, las probabilidades de que la media de la muestra sea 0.0408 pulgadas o más, o bien 0.392 pulgadas o menos son apenas 4.5 % (P(Z>2 o Z<-2) = 2(0.5-0.4722)=0.0456, o sea aproximadamente 4.5 %). Con esta probabilidad tan baja, PA llegaría a la conclusión de que una población con una media verdadera de 0.04 pulgadas posiblemente no produzca una muestra como ésta. El supervisor de proyectos rechazaría la afirmación del proveedor relativa al espesor medio de las láminas.

Función del que toma la decisión en la formulación de la hipótesis.- En este caso, la diferencia entre la media de la muestra y la supuesta media de la población resulta demasiado grande, y la probabilidad de que la población produzca esa muestra aleatoria es demasiado pequeña. La razón por la cual ésta probabilidad de 4.5 % resulta demasiado baja o no, es un juicio que incumbe a los que toman la decisión. Ciertas situaciones exigen que ellos estén muy seguros respecto a las características de lo que estén probando; tratemos de averiguar los costos resultantes de una decisión incorrecta y el grado exacto de riesgo que estamos dispuestos a correr.

Riesgo del rechazo.- En el ejemplo anterior, rechazamos la afirmación de la compañía fabricante de aluminio de que la media de la población es de 0.04 pulgadas. Pero supongamos, por un momento, que la media de la población fuera realmente de esa magnitud. Si observamos nuestra regla de rechazo de 2 errores estándar o más (la probabilidad de 4.5 % o menos en los extremos de la figura 4.1), 4.5 % de las veces rechazamos un lote perfectamente bueno de hojas de aluminio. Por tanto, el requisito mínimo de una probabilidad aceptable, 4.5 % es también el riesgo que corremos de rechazar una hipótesis verdadera. Esto sucede en todas las pruebas de hipótesis.

La hipótesis de que $\mu = 0.04$ de la que tanto hemos hablado hasta el momento se llama hipótesis nula y se representa con H_0 ; cada vez que rechazamos la hipótesis nula, la conclusión que aceptamos se llama hipótesis alternativa y se representa con H_1 o H_a.

La finalidad de la prueba de hipótesis no es poner en tela de juicio el valor calculado del estadístico muestral, sino emitir un juicio sobre la diferencia existente entre él y un supuesto parámetro de la población. El siguiente paso, luego de formular la hipótesis nula y la hipótesis alternativa, será decidir qué criterio aplicar para decidir si se acepta o rechaza la primera.

En el ejemplo del complejo deportivo, hemos determinado que la diferencia observada entre la media muestral $\bar{x}$ y la supuesta media de la población μ tenía apenas 4.5 %, o sea, 0.045 probabilidades de ocurrir. Por tanto, rechazamos la hipótesis nula de que la media de la

población era de 0.04 pulgadas ($H_0 : \mu = 0.04$ pulgadas). En términos estadísticos, el valor 0.045 recibe el nombre de nivel de significancia.

Función del nivel de significancia : ¿qué sucederá si probamos la hipótesis en un nivel de significancia de 5 %? Ello significa que rechazaremos la hipótesis nula si en promedio la diferencia entre el estadístico muestral y el supuesto parámetro de la población es tan grande que ella o una diferencia mayor podría ocurrir, en promedio, apenas cinco o menos veces en cada 100 muestras, cuando sea correcto el parámetro de la población. Así pues, suponiendo que la hipótesis es correcta, el nivel de significancia indica el porcentaje de medias muestrales que se encuentran fuera de ciertos límites (al hacer la estimación, no olvidar que el nivel de confianza indica el porcentaje de las medias muestrales que caían dentro de los límites definidos de confianza).

Área donde no existe una diferencia significativa.- La figura 4.2 ilustra cómo interpretar un nivel de significancia de 5 %. Adviértase que 2.5 % del área bajo la curva está situado en cada extremo. Si consultamos la tabla de la normal estándar, podremos determinar que 95 % del área bajo la curva queda incluída en un intervalo que se extiende $1.96\sigma_{\bar{x}}$ a ambos lados de la supuesta media. En 95 % del área, no existe diferencia de significancia entre el estadístico muestral y el supuesto parámetro de la población. En el restante 5 % (las regiones sombreadas de la figura 4.2) sí hay una diferencia significativa

$$\mu_{H_0} - 1.96\sigma_{\bar{x}} \qquad \mu_{H_0} = 0.04 \qquad \mu_{H_0} + 1.96\sigma_{\bar{x}}$$

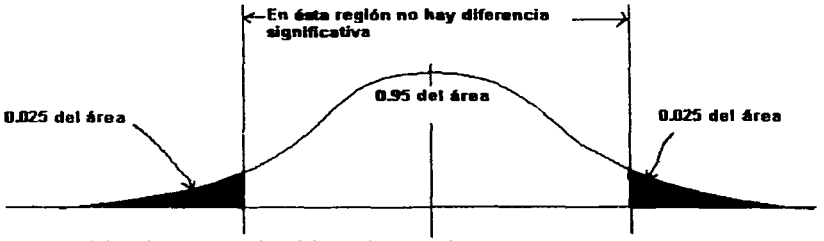


Fig 4.2 Regiones de diferencia significativa y no significativa en un nivel de significancia de 5 %. En las dos regiones sombreadas sí existe una diferencia significativa entre el estadístico de la muestra y el supuesto parámetro de la población.

En la figura 4.3 se examina esta misma muestra en una forma diferente. En ella 95% del área bajo la curva se halla donde aceptaríamos la hipótesis nula. Las dos partes sombreadas bajo la curva que representan un total de 5 % del área, se encuentran donde rechazaríamos la hipótesis nula

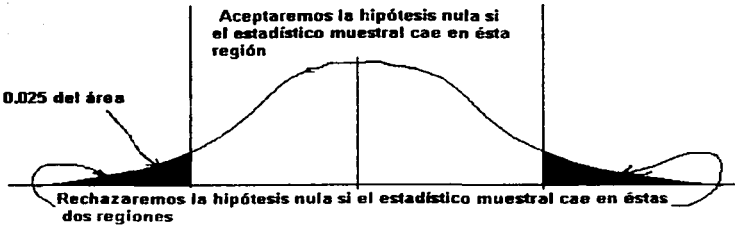


Figura 4.3 Un nivel de significancia de 5 % , con las regiones designadas de aceptación y rechazo.

Aquí conviene hacer una advertencia. Aún cuando los estadísticos muestrales en la figura 4.3 caigan en la región no sombreada (que abarca 95 % del área bajo la curva) ello no prueba que nuestra hipótesis nula H_0 sea verdadera; simplemente no ofrece evidencia estadística para rechazarla. ¿Por qué? Porque la única manera de aceptarla con certidumbre consiste en que conozcamos el parámetro de la población y, por desgracia, ello no es posible. Por tanto, cada vez que decimos que la aceptamos, en realidad queremos decir que no se cuenta con suficiente evidencia estadística para rechazarla. Se ha generalizado el uso del verbo aceptar, en lugar de no rechazar. Significa sencillamente que, cuando los datos de la muestra no nos llevan a rechazar una hipótesis nula, nos portamos como si ésta fuera verdadera.

Compromisos que deben hacerse al seleccionar un nivel de significancia : No hay un nivel oficial o universal de significancia con el cual probar las hipótesis. En algunos casos se recurre a un nivel de 5 %. A menudo los resultados publicados de las investigaciones prueban las hipótesis en un nivel de 1 % de significancia. Es posible probar una hipótesis en cualquier nivel de significancia. Pero tómese en cuenta que nuestra elección del criterio mínimo de una probabilidad aceptable, o nivel de significancia, es asimismo el riesgo que corremos de rechazar una hipótesis nula aunque sea verdadera. Cuanto más alto sea el nivel de significancia que utilizamos al probar una hipótesis, mayores probabilidades habrá de rechazar una hipótesis nula que sea verdadera.

Definición del error de tipo I y del error de tipo II : El rechazar una hipótesis nula que sea verdadera recibe el nombre de error de tipo I, y su probabilidad (que, como hemos visto,

es también el nivel de significancia de la prueba) se representa con α (alfa). En cambio, el aceptar una hipótesis nula que sea falsa se llama error de tipo II, y su probabilidad se representa con β (beta). Hay una especie de compromiso entre ambos tipos de error: la probabilidad de cometer uno de ellos se reduce solo si estamos dispuestos a aumentar la probabilidad de incurrir en el otro tipo de error. Por ejemplo, a fin de conseguir una β baja, habremos de conformarnos con una α alta. Para sortear este compromiso en situaciones personales y profesionales, los encargados de tomar decisiones eligen el nivel apropiado de significancia examinando los costos o castigos que conllevan a ambos tipos de error.

Preferencia por un error de tipo I : Supóngase que el cometer un error de tipo I (rechazar una hipótesis nula cuando es verdadera) implica el tiempo y el trabajo de reelaborar un lote de sustancias químicas que debería de haber sido aceptado. En cambio, el incurrir en un error de tipo II (aceptar una hipótesis nula que es falsa) significa correr el riesgo de que se envenene un grupo entero de usuarios de la sustancia. Sin duda, la gerencia de ésta compañía preferiría el error de tipo I al de tipo II y, en consecuencia, establecería niveles muy elevados de significancia en sus pruebas para conseguir "betas" bajas.

Preferencia por un error de tipo II: Supongamos ahora que el cometer un error de tipo I exige desarmar totalmente un motor en la fábrica y que, en cambio, incurrir en un error de tipo II requiere reparaciones garantizadas y relativamente baratas por parte del distribuidor. En tal caso, el fabricante probablemente preferirá un error de tipo II y fijará bajos niveles de significancia a sus pruebas.

Observación : No olvidar otra regla más al probar el supuesto valor de una media. Como en la estimación, se utiliza el multiplicador de población finita cuando ésta es de tamaño finito, el muestreo se realiza sin reemplazamiento y la muestra constituye más del 5 % de la población.

Ya en la práctica, por conveniencia presentaremos la prueba de hipótesis como un procedimiento de 8 pasos aunque nada hay de sagrado o mágico en éste formato:

1) Datos.- Debe comprenderse la naturaleza de los datos que forman la base de los procedimientos de prueba, ya que ésta determina la prueba particular que debe emplearse. Debe determinarse, por ejemplo, si los datos consisten de conteos o medidas.
2) Suposiciones.- Un procedimiento general se modifica, dependiendo de las suposiciones. De hecho, las mismas suposiciones que tienen importancia en la estimación, también son importantes en la prueba de hipótesis. Se ha visto que éstas incluyen, entre otras, suposiciones acerca de la normalidad de la distribución de la población e independencia de las muestras.
3) Hipótesis.- En la prueba de hipótesis se trabaja con dos hipótesis que deben enunciarse explícitamente: la hipótesis nula (la que debe probarse) y la hipótesis alternativa. En general, la hipótesis nula se establece con el propósito expreso de ser

desacreditada. Como consecuencia, la opuesta a la conclusión que el investigador desea alcanzar se convierte en el enunciado de la hipótesis nula.

4) Estadística de prueba.- La estadística de prueba es alguna estadística que puede calcularse a partir de los datos de la muestra. Como regla, existen muchos valores posibles que puede tener la estadística de prueba, dependiendo del valor particular observado de la muestra particular extraída. Como se verá, la estadística de prueba sirve como un productor de decisiones, ya que la decisión de rechazar o no rechazar la hipótesis nula depende de la magnitud de la estadística de prueba. Un ejemplo de estadística de prueba es la cantidad

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

con la que ya estamos familiarizados. Otro ejemplo es

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

Los valores de la estadística de prueba son puntos sobre una recta que sirve como eje horizontal para la distribución de tal estadística.

5) Distribución de la estadística de prueba.- La clave para la inferencia estadística es la distribución muestral. Se hace necesario especificar la distribución de probabilidad de la estadística de prueba. Por ejemplo, la distribución de la estadística de prueba Z escrita anteriormente sigue la distribución normal unitaria si la hipótesis nula es verdadera y se satisfacen las suposiciones; mientras que la t anterior está distribuída como t-Student.
6) Regla de decisión.- Los valores de la estadística de prueba que comprenden la región de rechazo son aquellos que tienen la menor probabilidad de ocurrir si la hipótesis nula es verdadera, mientras que los valores que forman la región de aceptación son los que tienen mayor probabilidad de ocurrir si la hipótesis nula es verdadera. La regla de decisión nos dice que se rechace la hipótesis nula, si el valor de la estadística de prueba que se calcule a partir de la muestra es uno de los valores en la región de rechazo y que no se rechace (o que se "acepte") la hipótesis nula, si el valor calculado de la estadística de prueba es uno de los valores en la región de aceptación.

   La decisión por lo que respecta a cuáles valores van hacia la región de rechazo y cuáles a la región de aceptación se toma con base en el nivel de significación deseado (α).
7) Estadística de prueba calculada.- A partir de los datos contenidos en la muestra, se calcula un valor de la estadística de prueba y se le compara con las regiones de aceptación y de rechazo que ya se han especificado.

8) Decisión estadística.- La decisión estadística consiste en el rechazo o no rechazo de la hipótesis nula. Se rechaza, si el valor calculado de la estadística de prueba cae en la región de rechazo y no se rechaza, si el valor calculado de la estadística de prueba cae en la región de aceptación.

Existe aún un noveno paso, correspondiente a la decisión técnica pero lo omitimos porque corresponde ya al científico del trozo de realidad donde se esté aplicando la estadística; o sea, al químico, al administrador, al físico, al econometrista, etc.

Las reglas para elegir la distribución apropiada (la estadística de prueba) en el punto 4 del formato, se parecen a las descritas en el capítulo dedicado a la estimación. En la siguiente tabla se resume cuándo usar la distribución normal y la distribución t al efectuar pruebas de las medias

	CUANDO SE CONOCE LA DESVIACIÓN ESTÁNDAR DE LA POBLACIÓN	CUANDO NO SE CONOCE LA DESVIACIÓN ESTÁNDAR DE LA POBLACIÓN
EL TAMAÑO n DE LA MUESTRA ES MAYOR QUE 30	Distribución Normal Tabla Z	Distribución normal Tabla Z
EL TAMAÑO n DE LA MUESTRA ES 30 O MENOS Y SUPONEMOS QUE LA POBLACIÓN ES NORMAL O APROXIMADAMENTE NORMAL	Distribución normal Tabla Z	Distribución t-Student Tabla t

4.2 **Prueba de hipótesis para una media poblacional.**

Resolvamos el siguiente ejemplo con el formato de los 8 pasos:

EJEMPLO 16 : Una compañía que vende camarón congelado imprime sobre el envase "Contiene 12 onzas". Una muestra de 35 envases da una media de 11.83 onzas. De las experiencias anteriores se conoce que la población de los pesos de los envases tiene una desviación estándar de 0.5 onzas. Utilizando α =0.05, ¿qué conclusión se sacaría acerca de la producción que la compañía intenta obtener?

SOLUCIÓN :

1) Datos.- Con la variable aleatoria X designamos a la población X= número de onzas en los envases. Los datos de que se dispone son las determinaciones de los pesos hechas en una muestra de 25 envases de la población de interés. A partir de ésta muestra, se ha calculado una media de $\overline{x}$ =11.83 onzas.
2) Suposiciones.-La muestra proviene de una población de valores de los pesos de distribución desconocida con una varianza conocida $\sigma^2 = (0.5)^2$ onzas.
3) Hipótesis.- La hipótesis que debe probarse, o hipótesis nula es que el peso medio de los envases en la población es igual a 12 onzas. La hipótesis alternativa es que el peso

medio de los envases en la población no es igual a 12 onzas. Nótese que se está identificando a la hipótesis alternativa con la conclusión que la compañía intenta obtener de que la media de la población es de 12 onzas, de modo que si los datos permiten el rechazo de la hipótesis nula, su conclusión tendrá más peso, puesto que la probabilidad acompañante de rechazar una hipótesis nula verdadera es pequeña. La hipótesis en cuestión se presenta en forma compacta como sigue:

$$H_0 : \mu = 12 \quad \text{vs} \quad H_a : \mu \neq 12 \quad (\alpha = 0.05)$$

4) Estadística de prueba.-

Como X tiene distribución desconocida con media μ y varianza σ^2, por el TCL $\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$

La estadística de prueba es

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

5) Distribución de la estadística de prueba.- Con base en nuestro conocimiento de las distribuciones muestrales y la distribución normal, se sabe que la última Z de arriba tiene distribución N(0,1) si H_0 es verdadera. Existen muchos valores posibles de la estadística de prueba que la situación presente puede generar; uno para cada muestra posible de tamaño 25 que pueda extraerse de la población. Puesto que solo se extrajo una muestra solo se tiene uno de estos valores posibles sobre el cual basar una decisión.

6) Regla de decisión.- La regla de decisión nos dice que se rechace H_0, si el valor calculado de la estadística de prueba cae en la región de rechazo y se acepte H_0, si cae en la región de aceptación. Ahora deben especificarse las regiones de rechazo y aceptación. Podemos empezar por preguntarnos qué magnitud de los valores de la estadística de prueba provocarán el rechazo de H_0. Si la hipótesis nula es falsa puede ser porque la media verdadera es menor que 12, o bien, porque la media verdadera es mayor que 12. Por lo tanto, ya sea los valores extremadamente pequeños o los valores extremadamente grandes de la estadística de prueba provocarán el rechazo de la hipótesis nula. Se desea que éstos valores extremos constituyan la región de rechazo. ¿ Qué tan extremo debe ser un valor posible de la estadística de prueba para poder formar parte de la región de rechazo? La respuesta depende del nivel de significancia (o significación) que se elija, es decir, la magnitud de la probabilidad de cometer un error del tipo I. Puesto que la probabilidad de rechazar una hipótesis nula

verdadera es $\alpha = 0.05$, la región de rechazo va a consistir de dos partes, parte de α tendrá que asociarse con los valores grandes y parte con los valores pequeños. Parece razonable que deba dividirse α en partes iguales y considerar a $\alpha/2 = 0.025$ asociada con los valores extremadamente pequeños y a $\alpha/2 = 0.025$ asociada con los valores extremadamente grandes.

¿Qué valor de la estadística de prueba es tan grande que, cuando la hipótesis nula es verdadera, la probabilidad de obtener un valor así de grande o mayor es 0.025? En otras palabras ¿cuál es el valor de Z hacia la derecha del cual esta 0.025 del área bajo la distribución normal unitaria? El valor de Z hacia la derecha del cual está 0.025 es el mismo valor entre cero y hasta el cual está 0.475 del área. Se busca en el cuerpo de la tabla normal estándar hasta que se encuentra 0.475 o su valor más próximo y se leen las anotaciones correspondientes en el margen, para obtener el valor de Z. En el presente ejemplo, el valor de Z es 1.96. Razonando de manera semejante, se llegará a encontrar -1.96 como el valor de la estadística de prueba tan pequeño que, cuando la hipótesis nula es verdadera, la probabilidad de obtener un valor así de pequeño o menor es 0.025. Entonces la región de rechazo consiste de todos los valores de la estadística de prueba iguales o mayores que 1.96 o menores que o iguales a −1.96. La región de aceptación consiste de todos los valores entre éstos. En la siguiente figura se muestran las regiones de aceptación y rechazo

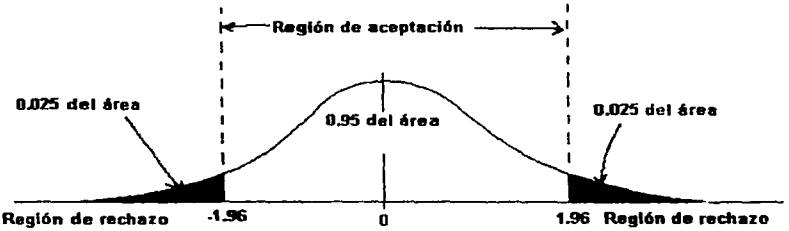


Fig 4.4 Regiones de aceptación y rechazo para el ejemplo 16.

En ocasiones se da el nombre de valores críticos de la estadística de prueba a los valores de ésta que separan a las regiones de aceptación y rechazo y a la región de rechazo se le conoce como región crítica.

La regla de decisión nos dice que se calcule un valor de la estadística de prueba a partir de los datos de la muestra y que se rechace H_0 si se obtiene un valor igual o mayor que 1.96 o igual o menor que −1.96; y que se "acepte" H_0 si se obtiene cualquier otro valor. El valor de α y, de aquí, la regla de decisión debe decidirse antes de reunir los datos. Esto evita que se nos acuse de permitir que los resultados de la muestra influyan en nuestra decisión. Esta condición de objetividad es importantísima y debe conservarse en todas las pruebas.

7) Estadística de prueba calculada.- De la muestra se calcula

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{11.83-12}{0.084515425} = -2.011467138$$

8) Decisión estadística.- Ateniéndose a la regla de decisión se puede rechazar la hipótesis nula ya que −2 está en la región de rechazo. Puede decirse que el valor calculado de la estadística de prueba es significativo en el nivel 0.05

Es decir, por la evidencia en la muestra, la media del contenido de los envases no es 12 onzas con un nivel de significancia de 0.05.

Este que acabamos de resolver, es un ejemplo de los llamados "de dos colas" porque las regiones de rechazo están en las dos colas de la distribución normal.

Un ejemplo de "cola inferior" o "cola izquierda" es:

EJEMPLO 17: Una empresa industrial supone que la vida de su prensa rotativa más grande es de 14,500 horas con una desviación estándar de 2,100 horas. De una muestra de 25 prensas la compañía obtiene una media muestral de 13,000 horas. En un nivel de significancia de 0.01, ¿debe la compañía concluír que la vida media de las prensas es menor que las 14,500 supuestas?

SOLUCIÓN: Podríamos resolver este ejemplo desglosando detalladamente los 8 pasos. Pero como eso ya lo hicimos en el ejemplo anterior, éste lo resolveremos de modo más compacto:

Sea la población X= número de horas de vida de las prensas rotativas. La compañía debe concluír que la vida media de las prensas es menor que las 14,500 supuestas si puede rechazar la hipótesis nula de que dicha media es mayor o igual que las 14,500 supuestas:

Hay que probar $H_0 : \mu \geq 14{,}500$ vs $H_a : \mu < 14{,}500$ ($\alpha = 0.01$)

Como bajo H_0 :

$$X \cong ?\left(14500, (2100)^2\right), \quad \textit{por el TCL}$$

$$X \approx N\left(14500, \frac{(2100)^2}{25}\right). \quad \textit{La estadística de prueba}$$

$$\textit{es} \quad Z = \frac{\bar{X}-\mu}{2100/5} \quad \textit{que tiene distribución } N(0,1)$$

Según la hipótesis alternativa, la región de rechazo son los valores muy pequeños. Como α =0.01, ¿Cuál es la Z* tal que P(Z < Z*)= 0.01 ?

Viendo las tablas de la normal, encontramos Z* = -2.33; y se tiene la figura:

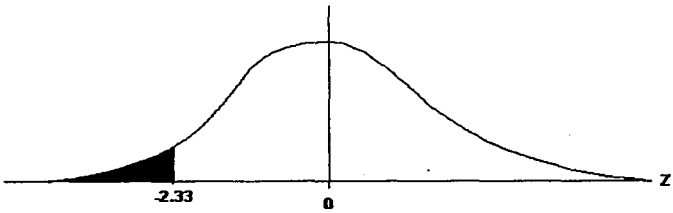


Fig. 4.5 Para el ejemplo 17

Como la estadística calculada

$$E_c = Z_c = \frac{1300 - 14,500}{420} = -3.57 \quad cae \ en \ la \ región \ de \ rechazo$$

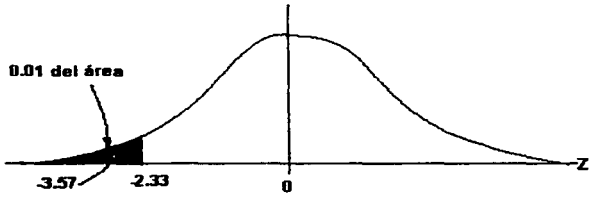


Fig. 4.6 Para el ejemplo 17

Rechazamos H_0. La compañía sí debe concluír que la vida media de las prensas es menor que las 14,500 supuestas.

Como ejemplo de "cola superior" o "cola derecha" presentamos el siguiente

EJEMPLO 18: La comisión promedio que cobran las empresas norteamericanas de servicios completos de corretaje en la venta de acciones comunes es de $ 144.00, con una desviación estándar de $ 52.00. Un corredor ha extraído una muestra aleatoria de 121 transacciones de sus clientes y determinó que pagaron una comisión promedio de $ 151.00. En

un nivel de significancia de 0.10 ¿podemos afirmar que las comisiones de sus clientes son superiores al promedio de la industria?

SOLUCIÓN: Podemos afirmar que las comisiones de los clientes son superiores a 144, el promedio de la industria, si podemos rechazar la hipótesis nula de que las comisiones de los clientes son menores o iguales a 144 :

Hay que probar $H_0 : \mu \leq 144$ vs $H_a : \mu > 144$ ($\alpha = 0.10$)

Sea la población X= comisiones
Como bajo H_0, por el TCL

$$\bar{X} \approx N\left(144, \frac{(52)^2}{121}\right). \quad \textbf{Entonces, estandarizando}$$

$$Z = \frac{\bar{X} - 144}{52/11} \approx N(0,1)$$

Ésta Z es la estadística de prueba.
Y según la hipótesis alternativa, la región de rechazo de H_0 son los valores muy altos.
Como α =0.10, ¿cuál es la z* tal que P(Z > z*) = 0.1?
Viendo tablas de la normal, z* = 1.28

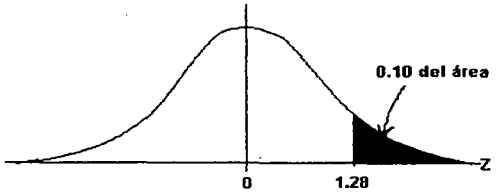


Fig. 4.7 Para el ejemplo 18

La estadística calculada es E_c = (151–144) / (52/11) = 1.48 que cae en la región de rechazo. Así que se rechaza H_0.

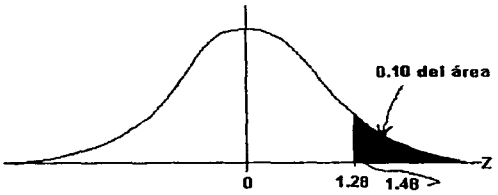


Fig. 4.8 Para el ejemplo 18

Puede afirmarse que las comisiones del cliente son superiores al promedio de la industria; es decir, las comisiones son significativamente más altas.

Ejercicios de la sección 4.2

1) Una tienda de artículos deportivos ha iniciado una promoción especial para su horno de propano y piensa que la promoción deberá culminar en un cambio de precio. Sabe que, antes de comenzar la promoción, el precio promedio al menudeo del horno era de $ 41.95, con una desviación estándar de $ 5.36. La tienda muestrea 16 de sus detallistas una vez iniciada la promoción y descubre que el precio medio de los hornos ahora es de $ 38.95. En un nivel de significancia de 0.02, ¿tiene motivos para pensar que el precio promedio al menudeo ha disminuído?
2) De 1980 hasta 1985, la razón media de precio / ganancias de aproximadamente 1800 acciones cotizadas en la bolsa de valores de México fue de 14.35 con una desviación estándar de 9.73. En una muestra de 30 acciones elegidas al azar, la razón media de precio/ganancias fue de 11.77 en 1986. ¿Esta muestra ofrece suficiente evidencia para afirmar (en un nivel de significancia de 0.05) que en 1986 esa razón media de la bolsa de valores de México ha cambiado su valor anterior?
3) American Theaters sabe que una película de gran éxito se exhibirá un promedio de 84 días en cada ciudad, y la desviación estándar correspondiente ha sido de 10 días. El gerente del distrito del sureste quería comparar la popularidad de la película en su región con la que alcanzó en otros cinemas del país. Seleccionó aleatoriamente 75 cinemas de su región y descubrió que exhibieron la película un promedio de 81.5 días.
   a) Formule las hipótesis correspondientes para probar si hubo una diferencia significativa en las semanas de exhibición de la película entre los cinemas del distrito del sureste y en todas las demás salas del país.
   b) En un nivel de significancia de 1 %, pruebe éstas hipótesis.

4) Todos los días, La oficina de aduanas descubre siempre unos 28 millones de dólares en artículos introducidos ilegalmente en el país, con una desviación estándar de $ 16 millones al día. En 64 días elegidos aleatoriamente en 1986, esa oficina interceptó un promedio de 30.3 millones de dólares en artículos de contrabando. ¿Indica esa muestra (en un nivel de significancia de 5 %) que al jefe de aduanas debe preocuparle el hecho de que el contrabando haya rebasado su nivel histórico?
5) Antes del embargo petrolero de 1973 y de los incrementos posteriores en los precios del petróleo crudo, el consumo de gasolina en Estados Unidos había crecido a una tasa mensual de 0.57 % ajustado a la estación del año, con una desviación estándar de 0.10 % al mes. En 15 meses escogidos aleatoriamente entre 1975 y 1985, el consumo de gasolina aumentó a un porcentaje promedio de apenas 0.33 % por mes. En un nivel de significancia de 0.01, ¿puede afirmar usted que el crecimiento en el consumo de gasolina disminuyó a raíz del embargo y sus consecuencias?
6) Un equipo semiprofesional de béisbol tiene un jugador que encabezó la liga en el promedio de bateo durante muchos años. En los últimos años, ese jugador consiguió un promedio de bateo de 0.343, con una desviación estándar de 0.018. Sin embargo, éste año su promedio es apenas de 0.306. El jugador está renegociando su contrato de la próxima temporada, y el sueldo que obtendrá depende mucho de su capacidad para convencer al dueño del equipo de que su promedio de bateo en éste año no difiere muy significativamente del logrado en años anteriores. Si el dueño está dispuesto a usar un nivel de significancia de 0.02, ¿reducirá el sueldo del jugador en el próximo año?

4.3 Prueba de hipótesis para una proporción poblacional.

Ante todo, recuérdese que la distribución binomial es la distribución teóricamente correcta que debe usarse en el caso de las proporciones, pues los datos son discretos, no contínuos. A medida que aumenta el tamaño de la muestra, la distribución binomial se acerca a la distribución normal en sus características y podemos aplicar ésta última para aproximar la distribución de muestreo. En concreto np y nq necesitan cada uno ser por lo menos 5 para que pueda utilizarse la distribución normal en sustitución de la binomial.

Pasemos ahora a los problemas.

EJEMPLO 19: De acuerdo con la teoría de la herencia de Mendel, ciertas cruzas de chícharos arrojan chícharos amarillos y verdes en la relación 3:1. En un experimento se han obtenido 176 chícharos amarillos y 48 verdes. Con $\alpha = 0.05$ ¿Son éstos números compatibles con la teoría mendeliana?

SOLUCIÓN : Resolvamos este problema desglosando el formato de los 8 pasos.

1) Datos.- Se tiene una muestra de 224 chícharos, de los cuales 176 son amarillos, es decir

$$\bar{p} = \frac{176}{224} = 0.79, \quad n = 224$$

2) Suposiciones.- Como elegimos que p sea la proporción de chícharos amarillos y np y nq son ambos mayores que 5, por el TCL:

$$p \approx N\left(p, \frac{pq}{n}\right). \quad \textit{Es decir} \quad \bar{p} \approx N(p, 0.1659)$$

3) Hipótesis.- H_0 : p= ¾ vs H_a : p≠ ¾ ($\alpha = 0.05$)

4)Estadística de prueba

$$Z = \frac{\bar{p} - p}{\sqrt{pq/n}}$$

5) Distribución de la estadística de prueba.- Si la hipótesis nula es verdadera, la estadística de prueba está distribuída aproximadamente en forma normal.

$$Z = \frac{\bar{p} - p}{\sqrt{pq/n}} \approx N(0,1)$$

6) Regla de decisión.- Siendo α =0.05, los valores críticos de Z son ± 1.96. Se rechaza H_0 a menos que $-1.96 < E_c < 1.96$

7) Estadística de prueba calculada (E_c)

$$E_c = Z_c = \frac{0.79 - 0.75}{\sqrt{\frac{(0.79)(0.21)}{224}}} \approx 1.47$$

8) Decisión estadística.- Como $-1.96 < 1.47 < 1.96$ se "acepta" H_0 .

Fig. 4.9 Para el ejemplo 19

Se acepta H_0. Los números del experimento sí son compatibles con la teoría mendeliana.

EJEMPLO 20: Un fabricante de una patente médica sostiene que la misma tiene un 90 % de efectividad en el alivio de una alergia con un período de 8 horas. En una muestra de 200 individuos que tenían la alergia la medicina suministrada alivió a 160 personas. Determinar si la aseveración del fabricante es cierta, a un nivel de significancia del 1 %

SOLUCIÓN: Como el ejemplo anterior ya fue resuelto detalladamente con el formato de los 8 pasos, éste ejemplo lo resolveremos de modo compacto, aunque en realidad sí daremos los 8 pasos.

Establecemos $H_0 : p = 0.9$ vs $H_a : p < 0.9$ ($\alpha = 0.01$)

Bajo H_0 :

Como np y nq son ambos mayores que 5, podemos aproximar con la normal y usar el TCL, es decir:

$$\bar{p} \approx N\left(p, \frac{pq}{n}\right). \quad \text{Estandarizando:}$$

$$Z = \frac{\bar{p} - p}{\sqrt{pq/n}} \approx N(0,1) \quad \text{donde} \quad \bar{p} = \frac{160}{200} = 0.8$$

Puesto que los valores que rechazan la hipótesis nula son los muy pequeños, esta es una prueba de cola izquierda (o cola inferior). Y como $\alpha = 0.01$, de las tablas de la normal, el valor crítico es –2.33 y la estadística de prueba calculada es

$$E_c = Z_c = \frac{0.8 - 0.9}{\sqrt{\frac{(0.8)(0.2)}{200}}} \approx -3.54$$

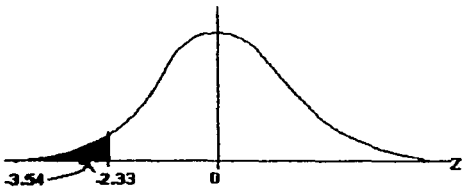


Fig. 4.10 Para el ejemplo 20

Por lo tanto se rechaza H_0 ; no hay evidencia estadística para aceptarla. La aseveración del fabricante es falsa al nivel de significancia del 1 %.

EJEMPLO 21: Un fabricante de salsa de tomate está a punto de decidir si producir una marca nueva de mucho condimento. El departamento de investigación de la compañía aplicó una encuesta telefónica a nivel nacional en 6,000 familias y averiguó que la salsa sería comprada por 335 de ellas. Un estudio mucho más exhaustivo hecho dos años antes reveló que 5 % de las familias compraría la marca entonces. En un nivel de significancia de 2 %, ¿deberá la compañía concluír que hay un mayor interés en el sabor tan condimentado?

SOLUCIÓN: Deberá concluirlo si puede rechazar H_0 :"hay menor o igual interés".
Así que establecemos

$H_0 : p \leq 0.05$ vs $H_a : p > 0.05$ (α =0.02)

Bajo H_0 :

Como np y nq son ambos mayores que 5 (¡Verifícalo!), podemos aproximar con la normal y usar el TCL. Entonces

$$p \approx N\left(p, \frac{pq}{n}\right) \quad donde \quad p = \frac{335}{6000} \approx 0.056$$

Si estandarizamos obtenemos

$$Z = \frac{\bar{p} - p}{\sqrt{pq/n}} \approx N(0,1)$$

Los valores que nos harán rechazar H_0 son los muy grandes, así que esta prueba es de cola superior (o cola derecha). Sabiendo esto último, con los datos del problema y con la tabla de la normal construímos la siguiente figura:

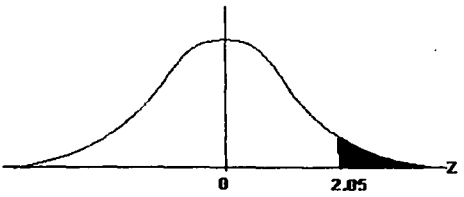


Fig. 4.11 Para el ejemplo 21

La estadística de prueba calculada es $Z_c = 1.97$ (¡Verifícalo!)

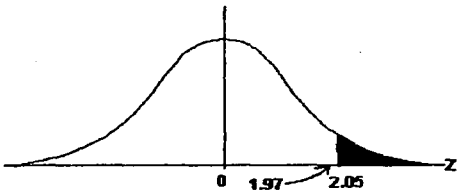


Fig. 4.12 Para el ejemplo 21

la cual cae en la región de aceptación de H_0 : la compañía no puede concluír que hay un mayor interés en el sabor tan condimentado.

Ejercicios de la sección 4.3

1) Un fabricante de blusas para dama, sabe que su marca se expende en 19 % de las tiendas de ropa de mujer al este del río Mississippi. Hace poco, muestreó 85 de ésas tiendas en la costa occidental y descubrió que 14.12 % de las tiendas venden la marca. En un nivel de significancia de 0.04, ¿hay pruebas de que el fabricante tiene una distribución menos satisfactoria en la costa occidental que la región oriental del Mississippi?
2) De un total de 9,700 préstamos, otorgados por la asociación cooperativa de crédito de los empleados, en el último período de 5 años, 350 fueron muestreados para determinar qué proporción fue concedida a las mujeres. La muestra reveló que 41 % de los préstamos se concedían a las empleadas. Un estudio similar, efectuado hace 5 años, mostró que 35 % de los prestatarios eran mujeres. En un nivel de significancia de 0.02, ¿puede usted afirmar que la proporción de préstamos otorgados a las mujeres ha cambiado significativamente durante los últimos 5 años?
3) Una compañía se especializa en la aplicación de técnicas de ingeniería genética para producir nuevos productos farmacéuticos. Hace poco inventó un pulverizador que contiene interferón, el cual reducirá la transmisión del resfriado común entre parientes. En la población general, 15.1 % de todos contraerán el resfriado causado

por rhinovirus cuando otro miembro de la familia sufre la enfermedad. El pulverizador de interferón fue probado en 180 personas, y uno de los miembros de sus familias contrajo resfriado por rhinovirus. Solo 17 de los sujetos sufrieron un resfriado semejante.

a) En un nivel de significancia de 0.05, ¿deberá la empresa farmacéutica concluír que el nuevo pulverizador logra reducir la transmisión de los resfriados?
b) ¿A qué conclusión deberá llegar cuando α=0.02?
c) Basándose en los resultados anteriores,¿cree usted que la empresa farmacéutica debería vender el nuevo pulverizador? Explique su repuesta.

4) Algunos teóricos de las finanzas piensan que los precios diarios del mercado de valores, constituyen una "fluctuación aleatoria con tendencia positiva". Si tienen razón, entonces el promedio industrial Dow Jones habrá de mostrar una ganancia en más del 50 % los días de actividad bursátil. Si el promedio aumentará en 101 de 175 días seleccionados aleatoriamente, ¿qué pensaría usted de la teoría anterior? Use un nivel de significancia de 0.01.
5) Un comerciante vende cortadoras de pasto en su ferretería y quiere comparar la confiabilidad de los aparatos que él vende con la de otra marca que se vende a nivel nacional. Sabe que apenas 15 % de éstas cortadoras requieren reparación durante el primer año de uso. Una muestra de 120 clientes del comerciante reveló que, exactamente 22 de ellos, necesitaban reparaciones el primer año de uso. En el nivel de significancia de 0.02, ¿hay evidencia de que las cortadoras del comerciante difieren en su confiabilidad de las que se venden a nivel nacional?

4.4 Prueba de hipótesis de las medias. Muestras pequeñas.

Cuando estimamos los intervalos de confianza, en el capítulo correspondiente, aprendimos que la diferencia de tamaño entre muestras grandes y pequeñas es importante cuando la desviación estándar σ de la población se desconoce y debe ser estimada a partir de la desviación estándar de la muestra. Si el tamaño de la muestra n es 30 o menos y σ^2 no se conoce, se usa la distribución t con n-1 grados de libertad. Estas reglas se aplican asimismo a la prueba de hipótesis.

EJEMPLO 22: Un documental de televisión dedicado a la glotonería afirmó que, en promedio, los norteamericanos tienen un exceso de peso de 10 libras aproximadamente. Para probar tal aserción, se examinó a 18 individuos seleccionados aleatoriamente, y se descubrió que su exceso promedio de peso era de 12.4 libras, con una desviación estándar de la muestra de 2.7 libras. En un nivel de significancia de 0.01, ¿hay razones para dudar de la validez de ese valor de 10 libras?

SOLUCIÓN: Sea la variable aleatoria X= número de libras de exceso de peso en los individuos norteamericanos.

Probar $H_0 : \mu = 10$ vs $H_a : \mu \neq 10$ ($\sigma = 0.01$)

Trabajando bajo H_0 :

Como se desconoce σ, tomamos como dato una estimación a partir de la muestra s= 2.7 y porque la muestra es pequeña (n= 18), debemos usar la t-Student con n-1 = 17 grados de libertad:

$t_{(1-\alpha/2)} = t_{.995}$. **Consultando tablas de la t – Student, vemos que el valor de $t_{.995}$ para 17 grados de libertad, es $t_{.995} = 2.8982$**

Y el valor calculado de la estadística de prueba es

$$E_c = t_c = \frac{12.4 - 10}{2.7/\sqrt{18}} \approx 3.77$$

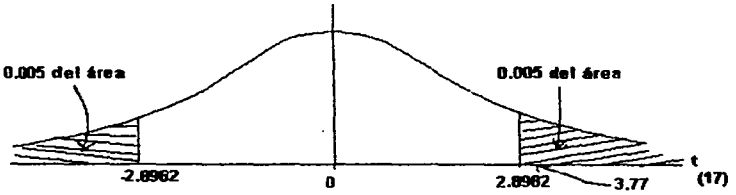


Fig. 4.13 Para el ejemplo 22

el cual cae en la región de rechazo de H_0.

Así que, hay razones para dudar de la validez del valor de 10 libras que dio la televisión, al nivel de 0.01.

4.5 **Potencia de una prueba.**

Un concepto importante es el de potencia. Me limitaré aquí a hacer algunos comentarios generales. La **potencia de una prueba** se define como la probabilidad de rechazar H_0.

EJEMPLO 23: Supongamos que tenemos $x_1, x_2, \ldots, x_{25}$ una muestra aleatoria de una población con función de densidad de probabilidad N(μ, 100). Se desea probar

$H_0 : \mu \leq 75$ vs $H_a : \mu > 75$

La región crítica o región de rechazo es

$$C = \{(x_1, x_2, \ldots, x_{25}) / \frac{\sum x_i}{25} > 75\}$$

La función potencia de esta prueba es

P(rechazar H_0 / μ) = P($\bar{X} > 75 / \mu$) =?
Estandarizando : P($\bar{X} > 75 / \mu$)=

$$P\left(\frac{\bar{X}-\mu}{2} > \frac{75-\mu}{2}\right) \quad porque \quad \bar{X} \cong N\left(\mu, \frac{100}{25}\right) = N(\mu, 4)$$

$$Así, \quad Z = \frac{\bar{X}-\mu}{2} \cong N(0,1)$$

Entonces la función potencia de la prueba es P($\bar{X} > 75 / \mu$) =

$$P(Z > \frac{75-\mu}{2} / \mu) \ para \ cada \ \mu$$

$$Si \ \mu = 75, \ P(Z > 0) = 0.5$$
$$Si \ \mu = 77, P(Z > -1) = 0.8413$$
$$Si \ \mu = 79, P(Z > -2) = 0.9772$$
$$Si \ \mu = 73, P(Z > 1) = 0.1587$$
$$Si \ \mu = 71, P(Z > 2) = 0.0228$$

La gráfica de ésta función es

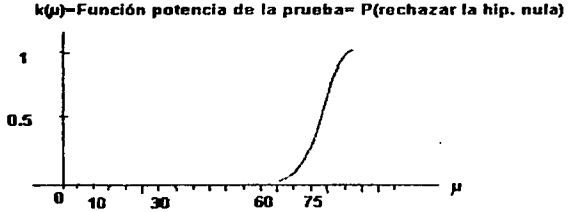


Figura 4.14. Función potencia de la prueba del ejemplo 23

Este es solo un trozo de la función potencia de la prueba.

Por supuesto que la potencia podría obtenerse para muchos más valores de μ. Marcando los puntos correspondientes a los valores de la potencia sobre un gráfico, se obtiene la figura 4.14. Una curva de potencia ideal sería de altura 1 para todos los valores del parámetro descrito por H_a, y de altura 0 para los valores del descrito por H_0 . Esto es, si la hipótesis es verdadera, sería lógico aceptarla siempre y si la hipótesis es falsa, sería lógico rechazarla. Mucha literatura de la estadística matemática tiene por objeto encontrar pruebas para una H_0 dada, con la potencia máxima, cuando H_a es verdadera.

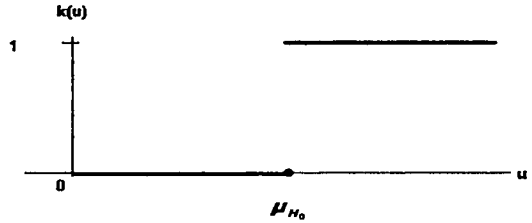


Figura 4.15 Función potencia ideal

Pues bien, con todo lo anterior resulta la siguiente definición:

Sea H_0 una hipótesis que va a ser probada contra la hipótesis alternativa H_a de acuerdo a una prueba preescrita. El nivel de significancia de la prueba (o el tamaño de la región crítica) es el valor numérico (o supremo) de la función potencia de la prueba cuando H_0 es cierta.

4.6 MISCELÁNEA DE EJERCICIOS DEL CAPÍTULO 4.

1) ACTIVIDAD 1

i) Los frascos de café soluble dicen en su etiqueta: "CONTENIDO NETO 200 gr" ¿Qué significado tiene esta frase en las etiquetas?
ii) El encargado de revisar el llenado de los frascos percibe a través de la observación visual que el nivel de los frascos ha subido. La observación del encargado conduce a una hipótesis de trabajo. Escribe esta hipótesis de trabajo.
iii) Tradúcela al lenguaje estadístico.
iv) Redacta lo contrario a la hipótesis de trabajo
v) Tradúcelo al lenguaje estadístico
vi) Para el inspector de la Secretaría de comercio, ¿Cuál sería una hipótesis de trabajo en relación con lo ofrecido por el fabricante en las etiquetas "CONTENIDO NETO 200gr"?
vii) Traduce la hipótesis de trabajo al esquema de pruebas de hipótesis
viii) ¿cuál sería una hipótesis de trabajo para la frase "CONTENIDO NETO 200 gr" si el interesado es el gerente de producción de la fábrica?
ix) Traduce esta hipótesis de trabajo en el esquema de pruebas de hipótesis.

2) ACTIVIDAD 2. Para contrastar la hipótesis de trabajo que ha sido observada por el encargado de revisar el llenado de los frascos, se ha tomado una muestra de 16 frascos y se ha pesado en contenido neto de cada uno de ellos, obteniéndose los siguientes resultados, en gramos:
199, 208, 202, 203, 205, 207, 201, 198, 202, 206, 205, 204, 203, 198, 202, 204
Tomando como base el siguiente esquema, realiza la prueba de hipótesis sabiendo que la varianza de la población es $\sigma^2 = 9$ y que la población está normalmente distribuída.

i) Planteamiento estadístico de las hipótesis
H_0 : H_a :
ii) Estadístico de prueba y su respectiva distribución de probabilidades.

iii) Nivel de significancia y regiones de aceptación y de rechazo de la hipótesis nula.
iv) Regla de decisión
v) Decisión estadística.
vi) Decisión técnica.

3) Si el inspector selecciona al azar 8 envases y la media de la muestra resultó ser $\bar{x}$ = 199.6 gr ¿a qué decisión se llega con α=0.10 y los mismos supuestos que en la actividad número 2?

i) Planteamiento estadístico de las hipótesis
H_0 : H_a :
ii) Estadístico de prueba y su respectiva distribución de probabilidades
iii) Nivel de significancia y regiones de aceptación y de rechazo de la hipótesis nula.
iv) Regla de decisión.
v) Decisión estadística.
vi) Decisión técnica.

4) El gerente de la fábrica necesita saber si la máquina envasadora está o no ajustada, para lo cual seleccionó 20 envases al azar de la línea de producción, obteniendo una media muestral $\bar{x}$ = 196.4, ¿a qué decisión se llega con un nivel de significancia α=0.01, suponiendo que la población está normalmente distribuida con varianza σ^2 =9?

i) Planteamiento estadístico de las hipótesis.
H_0 : H_a :
ii) Estadístico de prueba y su respectiva distribución de probabilidades.
iii) Nivel de significancia y regiones de aceptación y de rechazo de la hipótesis nula.
iv) Regla de decisión.
v) Decisión estadística.
vi) Decisión técnica.

5) Un fisicoterapeuta estudia la fuerza máxima media de un músculo particular en cierto grupo de individuos. El se inclina a suponer que las calificaciones de esa fuerza están distribuídas aproximadamente en forma normal, con una varianza de 144. Una muestra de 15 sujetos, quienes participaron en el experimento proporcionaron una media de 84.3, ¿debe concluír el fisicoterapeuta que la fuerza media para la población es menor que 95? Sea α=0.01.

6) En un estudio del flujo de pacientes a través de las oficinas de médicos generales, se encontró que, en promedio, una muestra de 35 pacientes llegaban 17.2 minutos tarde a las citas. Una investigación previa había demostrado que la desviación estándar era de 15 minutos aproximadamente. Se tuvo la sensación que la distribución de la población no era normal. ¿Puede concluírse que el tiempo medio de retraso verdadero es mayor que 12 minutos? Sea α=0.05.

7) En un experimento para determinar el número promedio de latidos del corazón por minuto para cierta población, bajo las condiciones del experimento, se encontró que el número

promedio de latidos por minuto para 49 sujetos era de 130. Si resulta razonable suponer que estos 49 pacientes constituyen una muestra aleatoria y que la población está distribuída normalmente, con una desviación estándar de 10, recórrase el procedimiento de 8 pasos para la prueba de hipótesis con el fin de probar la hipótesis nula de que $\mu = 125$. Sea $\alpha = 0.05$.

8) Se encontró que el nivel indirecto medio de bilirrubina en el suero de 16 niños de 4 días de nacidos era de 5.98 mg/100 cc. Suponiendo que los niveles de bilirrubina en los niños de 4 días de nacidos están distribuídos aproximadamente en forma normal con una desviación estándar de 3.5 mg/100 cc. ¿Proporcionan estos datos evidencia suficiente como para indicar que la media verdadera es menor que 6.5? Sea $\alpha = 0.05$ y recórrase el procedimiento de 8 pasos.

9) En un estudio de duración de la hospitalización realizado por varios hospitales en cooperación, se extrajo una muestra aleatoria de 64 pacientes de úlcera péptica admitidos alguna vez en los hospitales participantes, y se determinó, para cada uno, la duración de la hospitalización. Se encontró que la duración media de hospitalización fue de 8.25 días. Si se sabe que la desviación estándar de la población es de tres días, ¿debe concluirse que $\mu > 7.5$ días? Sea $\alpha = 0.01$ y complete los 8 pasos.

10) Una muestra de 100 adultos del sexo masculino aparentemente normales, de 25 años de edad, tuvo una presión sistólica sanguínea media de 125. Si se tiene la sensación de que la desviación media de la población es de 15, ¿indican éstos datos que $\mu < 130$? Sea $\alpha = 0.05$ y complétense los 8 pasos.

11) A 9 pacientes que sufren de la misma incapacidad física, pero de lo contrario comparables, se les pidió que llevaran a cabo cierta tarea como parte de un experimento. El tiempo promedio requerido para realizar la tarea fue de 7 minutos, con una desviación estándar de 2 minutos. Supóngase normalidad y pruébese la hipótesis nula de que $\mu = 10$ minutos. Sea $\alpha = 0.10$ y complétense los 8 pasos.

12) Un administrador de un hospital tomó una muestra de 25 cuentas vencidas, de las cuales calculó una media de $ 2,500 y una desviación estándar de $ 750. Suponiendo que las cantidades de todas las cuentas vencidas están distribuídas normalmente, ¿debe concluír el administrador que $\mu > \$ 2{,}250$? Sea $\alpha = 0.05$.

13) Una muestra de 25 niños de 10 años proporcionaron un peso medio y una desviación estándar de 36.5 y 5 kg respectivamente. Suponiendo una población normalmente distribuída, ¿proporcionan los datos evidencia suficiente como para indicar que μ es diferente de 37.5? Sea $\alpha = 0.05$

14) Una muestra de 16 niñas de 10 años dieron un peso medio de 35.8 kg y una desviación estándar de 6 kg respectivamente; suponiendo normalidad, ¿proporcionan los datos evidencia suficiente como para indicar que $\mu < 37.5$? Sea $\alpha = 0.05$

15) Para hacer frente a la disminución de recursos energéticos, la NASA está trabajando con instrumentos a través de la nación para encontrar lugares para la instalación de grandes máquinas de viento para generar energía eléctrica. D.J. Vargo, el funcionario a cargo de éste proyecto, dice que la velocidad promedio del viento debe de ser de al menos 15 millas por hora para que el lugar sea considerado aceptable. Se toman 36 registros de las velocidades del viento en intervalos al azar de un lugar considerado para instalar la máquina de viento; la velocidad promedio registrada fue de 14.2 mph con una desviación estándar de 3 mph ¿indican

estos datos que el lugar no satisface los requerimientos de la NASA para la instalación de una planta de energía eléctrica generada por el viento? Úsese σ =0.10

16) La calificación promedio de un grupo de 49 estudiantes resultó ser de 76 puntos con desviación estándar de 8 puntos. Si μ es la calificación promedio de todos los estudiantes de esa escuela, pruebe la hipótesis μ = 72 puntos contra la hipótesis μ > 72 puntos para un nivel de significancia de σ =0.05

17) Una compañía fabrica una pluma que presenta una media y una desviación estándar medidas en horas de escritura contínua de μ = 100 y σ =9 respectivamente. Para incrementar las ventas, modifica ligeramente el proceso de fabricación para producir una pluma que dure más que las plumas anteriores. Sin embargo, no se tiene razón para pensar que con el nuevo proceso se altera la variabilidad de estas plumas en términos de horas de escritura contínua. Para probar eso, elige en forma aleatoria n= 400 plumas fabricadas con el nuevo proceso y las utiliza hasta que se agota su tinta. Si la media de la muestra, en términos de horas de escritura contínua, resulta ser $\bar{x}$ = 101.5 horas y el nivel de significancia es σ =0.05, ¿qué conclusión puede deducir la compañía respecto a esta nueva pluma?

18) En el pasado una planta química ha fabricado un promedio de 1,100 libras de productos químicos por día. Una muestra aleatoria de 260 días de operación en los pasados años muestra que la media muestral es $\bar{x}$ = 1,060 libras y que σ =340 libras por día. Si se desea probar que la producción promedio por día ha disminuído en forma significante sobre los años pasados:

a) Escriba la hipótesis nula y la alternativa.
b) Si se usa el estadístico de prueba Z, determine la región de rechazo correspondiente a un nivel de significancia del 0.05.
c) ¿los datos proporcionan suficiente evidencia que indique una disminución en el promedio de la producción diaria?

19) La especificación para una nueva aleación resistente al calor requiere que la cantidad de cobre en la aleación debe ser menor al 23.2 %. Una muestra de 10 análisis de una hornada común del producto muestra que el contenido promedio de cobre es de 23 % con una desviación estándar de 0.24 % ¿éstos análisis proporcionan evidencia suficiente para indicar que la cantidad promedio de cobre en la hornada es menor que el límite especificado? Use σ =0.10.

20) Un contratista constructor ha construído un gran número de casas del mismo tamaño y valor. El contratista afirma que el valor promedio de esas casas (o casas similares) no excede de $ 35,000. Un tasador de la categoría de los inmuebles selecciona al azar 5 de las nuevas casas construídas por el contratista y aprecia su valor en $ 34,500, $ 37,000, $ 36,000, $ 35,000, $ 35,000 con una desviación estándar de $ 801.46 ¿Estas 5 valuaciones contradicen la afirmación del contratista acerca del valor promedio de las casas? Pruebe ésta hipótesis al nivel de significancia de 0.05.

21) Compañía de servicio de alimentos instantáneos se esfuerzan por idear proyectos que proporcionen incentivos y produzcan salarios para sus administradores que sean competitivos con las correspondientes compañías competidoras. Una muestra aleatoria simple de 12 administradores de cada una de las compañías muestra que gana un salario promedio de $ 16,750 con una desviación estándar de $ 3,100 ¿sugieren estos datos que los salarios

promedio ganados por los administradores difieren de $ 18,500 que es el salario anual pagado por las compañías competidoras a sus administradores? Pruebe la hipótesis nula de que μ = $ 18,500 contra la alternativa de que $\mu \neq$ $ 18,500 al nivel de significancia del 5 %

22) Un fabricante afirma pretensiosamente que al menos el 95 % del equipo que suministra a las factorías satisface las especificaciones requeridas. Un examen de 700 piezas de equipo revela que 53 son defectuosas, ¿el resultado obtenido proporciona evidencia suficiente para rechazar la afirmación del fabricante? Usar α =0.05

23) Se ha detectado que el 90 % de infracciones de tránsito son levantadas incorrectamente. Se cree que con las medidas administrativas para realizar la renovación moral de la sociedad aumentará el número de infracciones legales. En un experimento con 100 automovilistas infraccionados, se detectaron 5 infracciones levantadas incorrectamente. ¿Esta evidencia es suficiente para afirmar que el procedimiento modificado es mayor? Resolver para un nivel de significancia α =0.05.

24) Un fabricante de focos afirma que éstos tienen una vida promedio de 100 horas, ¿se justifica esto si una muestra de tamaño 50 arroja como resultado $\bar{x}$ = 95, con σ = 15? Emplear α =0.01

SONRÍE...

1) Un lunes llegan un químico, un ingeniero, un estadístico y un contador a una empresa a solicitar empleo. Como sencillo examen les preparan la pregunta : 1+1=? Y les programan dicho examen. El martes lo presenta el ingeniero, el miércoles el químico, el jueves el estadístico y el viernes el contador. Además, cada uno puede auxiliarse en los resultados que dieron los aspirantes que ya presentaron el examen:

Martes.- El ingeniero se presenta con un ábaco, una regla de cálculo, su calculadora, una computadora portátil y además se auxiliaría con los dedos si fuera necesario. Como con los 5 instrumentos de cálculo obtiene que el resultado es o dos o aproximadamente dos, responde ya muy fatigado : "1+1=2"

Miércoles.- El químico responde: " 1+1=1 y voy a demostrarlo experimentalmente". Saca sus tubos de ensayo, matraces, etc. " ... si unimos una gota de agua con otra gota de agua, obtenemos otra gota de agua... por lo tanto – responde ya cansado y todo mojado - 1+1=1"

Jueves.- El estadístico solicita los resultados del ingeniero y del químico y escribe: "por el trabajo del ingeniero y del químico tengo 6 observaciones del resultado 1+1. Hay fuerte sospecha de que 1+1=2.

Sea la variable aleatoria X=1+1 y probaremos

$$H_0 : 1+1=2 \quad \text{vs} \quad H_a : 1+1 \neq 2 \quad (\sigma = \sigma_0)$$

Como n=6 es pequeña y se desconoce σ, deberemos usar la t-Student como estadística de prueba ... calcularé s...

... Se acepta H_0 (o mejor dicho, no se rechaza H_0)

Total que, ya muy cansado, terminó exponiendo su resultado:

" ... Hay fuerte evidencia estadística para concluír que 1+1 = 2"

Viernes.- El contador se presenta muy fresco, muy seguro de sí mismo y al leer en el examen: 1+1=?, les pregunta muy tranquilamente a los de la empresa: "¿cuánto quieren que sea el resultado?"

SONRÍE ...

2) En la escuela de un lugar imaginario Paolo enseñaba estadística:

" ... La apuesta de Pascal acerca de la existencia de Dios, puede presentarse como una elección entre las probabilidades de los errores de tipo I y II, y sus posibles consecuencias. Deberíamos aceptar la existencia de Dios y actuar en consecuencia, arriesgándonos a cometer un error de tipo II (que Dios no exista), o deberíamos negar su existencia y actuar también en consecuencia, corriendo el riesgo de cometer un error de tipo I (que exista)...

... Cuando se distribuye dinero, el conservador típico prefiere como sea los errores del tipo II (que el que ha hecho méritos no reciba su parte), mientras que el neoliberal típico aceptaría los errores del tipo I (que el que no lo merece reciba más de lo que le toca). Cuando se reparten castigos, el conservador típico se interesa más por aceptar los errores del tipo II (que el culpable no reciba el castigo que se merece), mientras que el neoliberal típico se preocupa más por aceptar los errores del tipo I (que el inocente reciba un castigo inmerecido)...

... Y en la administración federal de drogas, algunos directivos aceptan una alta probabilidad de cometer un error del tipo II (dando el visto bueno a un mal medicamento), mientras que otros prefieren un error del tipo I (negando la autorización a un buen fármaco) ... "

CAPITULO 5. PREDICCIÓN ESTADÍSTICA.

SONRÍE...

1) En cierta ocasión, el hombre que anuncia las predicciones del tiempo en un canal local informó que la probabilidad de que lloviera el sábado era del 50 % y la de que lloviera el domingo, el 50 % también, y por lo tanto, concluyó :

" ... Predecimos: la probabilidad de que llueva este fin de semana es del 100 %"

2) Otro hombre del tiempo informó un sábado que al día siguiente, domingo, haría el doble de calor (que el sábado) porque la temperatura pasaría de 5 grados a 10 grados.

3) - ¿Podrías predecir qué sería un niño complejo?
   - Pues uno con la madre real y el padre imaginario.
   - Y qué un oso polar?
   - Pues un oso rectangular después de un cambio de coordenadas.
   - Y qué le diría la curva a la tangente si se encontraran?
   - ¡ No me toques!
   - Y si dos vectores se encontraran, ¿qué le diría uno al otro?
   - ¿tienes un momento?

CAPITULO 5. PREDICCIÓN ESTADISTICA.

Los métodos estadísticos presentados hasta ahora han tratado, todos, con una sola variable x y su distribución de frecuencia. En particular, se ha tratado del cálculo y prueba de hipótesis sobre los parámetros de distribuciones de frecuencia de variables binomiales y normales; recuérdese que para el caso de una proporción, la distribución que está implícita es la binomial. Muchos de los problemas de trabajo estadístico, sin embargo, tratan con un número múltiple de variables. Este capítulo se dedicará a la explicación de dos de las técnicas para la elaboración de datos asociados con dos o más variables. El método se encuentra aplicado al caso de dos variables; pero puede aplicarse a más de dos, aunque esto último no lo veremos aquí.

5.1 Relación entre dos variables (Diagrama de dispersión)

Hay ocasiones en que necesitamos saber si dos variables están o no relacionadas y si lo están, de qué manera lo hacen. Por ejemplo, ¿están relacionados la estatura y el peso de las personas?, ¿el promedio en el bachillerato y el promedio en la licenciatura?, ¿la concentración de un medicamento inyectado y la rapidez de los latidos del corazón?, etc. Si aceptamos que las dos variables de interés, llamémoslas "x" e "y", están relacionadas, nos gustaría saber cómo están relacionadas. Conociendo la relación podremos predecir el valor de una de las variables conociendo el valor de la otra.

No estudiaremos en esta obra el tipo de relación llamado de "causa-efecto" como serían:

I = E / R (ley de Ohm) donde I = Intensidad de corriente eléctrica
E = Fuerza electromotríz
R = Resistencia del conductor

O

F = m a (segunda ley de Newton)

Etc.

El tipo de relación que veremos aquí, es el llamado de "naturaleza estadística".

Para empezar, debemos ser cuidadosos en nuestra pretensión de que las dos variables de interés están relacionadas porque, por ejemplo, ¿están relacionados de algún modo el número de choques de autos en la Ciudad de México y el número de helados vendidos en China?

Después de la observación anterior debemos tomar una muestra de los datos, que serán parejas ordenadas (x, y) tomadas de la misma entidad llamada "unidad de asociación".

Si se tiene interés en la relación entre la estatura y el peso, por ejemplo, éstas dos medidas se toman sobre el mismo individuo. Por lo general, no tiene sentido hablar de la relación, digamos, entre las estaturas de un grupo de individuos y los pesos de otro grupo.

Enseguida, con los datos recopilados (x , y), graficamos en el plano cartesiano. Dicha gráfica, que consta de puros puntos, recibe el nombre de "diagrama de dispersión" y es ella la que nos sugiere el tipo de relación que hay entre las dos variables. Ejemplos

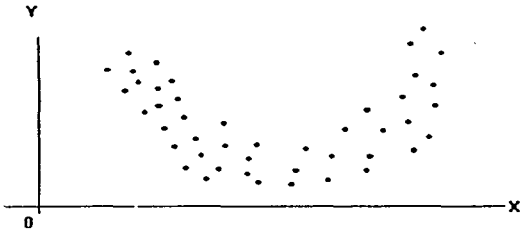


Fig. 5.1 Este diagrama de dispersión nos sugiere que X e Y se relacionan según
$y = ax^2 + bx + c$ (parábola)

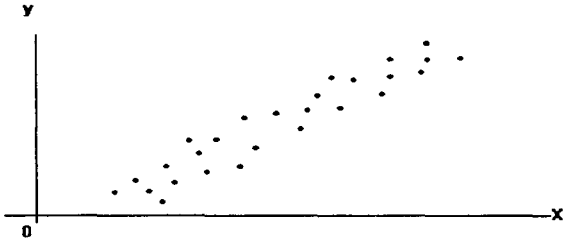


Fig. 5.2 Este diagrama de dispersión nos sugiere que X e Y se relacionan según
Y = a + bx (línea recta)

Cuando la relación sugerida sea una línea recta, diremos simplemente que hay "relación lineal" entre X e Y.

EJEMPLO 24. El vicepresidente de International Motors está trabajando en la relación que hay entre el sueldo de un empleado y el ausentismo. Dividió el intervalo de sueldos de la empresa en 12 grados o niveles (1 es el grado más bajo y 12 el máximo grado) y luego muestreó aleatoriamente un grupo de empleados. Determinó el grado del sueldo de cada empleado y el número de días que ese trabajador faltó en los tres últimos años.

RANGO DE SUELDOS	11	10	8	5	9	9	7	3	11	8	7	2	9	8	6	3
INASISTENCIAS	18	17	29	36	11	26	28	35	14	20	32	39	16	26	31	40

Construír un diagrama de dispersión para los datos anteriores e indicar el tipo de relación

SOLUCIÓN:

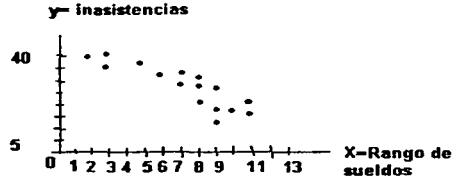


Fig. 5.3 Diagrama de dispersión para el ejemplo 24

El diagrama de dispersión nos sugiere que hay relación lineal (según y= a+bx. Una recta con pendiente negativa) entre las variables bajo estudio.

EJERCICIO A : El National Institute of Environmental Health Sciences (NIEHS) ha estado estudiando las relaciones estadísticas entre muchas variables diferentes y el resfriado común. Una de las variables es el empleo de toallas faciales (x) y el número de días en que aparecieron los síntomas de resfriado (y) en 7 personas durante un período de 12 meses, ¿qué relación , si la hay, parece existir entre las dos variables?, ¿indica esto un efecto causal?

X	2000	1500	500	750	600	900	1000
y	60	40	10	15	5	25	30

5.2 Correlación lineal simple.

Como ya decía, en algunos problemas las variables se estudian simultáneamente, para ver la forma en que se encuentran interrelacionadas; en otros, se tiene una variable de interés particular y las restantes se estudian por la posibilidad de que ayuden a arrojar luz sobre la primera. Estas dos clases de problemas se conocen generalmente con los nombres de correlación y regresión, respectivamente. Los métodos de correlación se discuten en primer término.

Un problema de correlación se presenta cuando un individuo se pregunta si existe alguna relación entre un par de variables que le interesan. Por ejemplo, ¿existe alguna relación entre el uso del tabaco y las afecciones cardíacas?, ¿entre la capacidad para aprender música y

la aptitud científica?, ¿entre la recepción de ondas de radio y la actividad de las manchas solares?, ¿entre la belleza y la inteligencia?

Para ilustrar la forma en que se procede a estudiar la relación entre dos variables, considérense los datos de la tabla I, que consisten en promedios de calificaciones correspondientes a escuela preparatoria y a primer año universitario. El promedio en la escuela preparatoria se denota con "x" y con "y" el promedio universitario.

Como ya vimos en la sección anterior, la investigación de la relación entre las dos variables comienza, generalmente, con un intento de descubrir la forma aproximada de la relación, trazando los datos como puntos en el plano x-y. Ya sabemos que ésta gráfica recibe el nombre de "diagrama de dispersión". Por este medio, puede decirse fácilmente si existe o no una relación acentuada y, en este caso, si puede tratarse como aproximadamente lineal.

EJEMPLO 25: El diagrama de dispersión para las 30 calificaciones obtenidas por los estudiantes, según la tabla I, se muestra en la figura 5.4

x	3	2.4	3.7	3.6	3.8	2.9	3.5	3	2.3	3	2.9	2.7	3.7	2.7	3.3	2.8
y	2.4	2.6	3	3.9	3.6	3	3.1	2.8	2.2	2.9	1.9	2.2	3.1	2.6	2.8	2.7
x	3.1	2.8	3	2.2	3.1	3.3	2.7	3.5	2.9	2.7	2.9	3.2	3.4	2.5		
y	2.4	3	3.3	1.8	2.8	3.2	1.8	2.7	2.1	1.7	1.7	2.3	2.6	2.7		

Tabla I

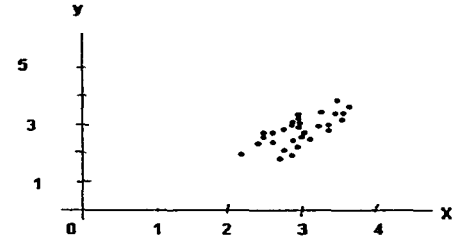


Fig. 5.4 Diagrama de dispersión para promedios de calificaciones.

La observación de este diagrama de dispersión muestra que existe una tendencia a que los valores bajos de x, estén asociados a los valores bajos de y, así como de que también estén asociados entre sí los valores altos de ambas variables. Por otro lado, y en términos aproximados, el aspecto general del diagrama de dispersión, es el de una línea recta. Para determinar la naturaleza de una tendencia, se busca cualquier tendencia de los puntos de agruparse sobre ambos lados de alguna curva simple, quizás con unas cuantas ondulaciones, o bien a ambos lados de una línea recta. Para el ejemplo que nos ocupa, sería conveniente poder medir en alguna forma el grado en que ambas variables se encuentran relacionadas

linealmente (en forma de línea recta). La medida deseada para dicha relación está dada por r en la fórmula

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Y se sabe que $-1 \leq r \leq 1$. Luego : $0 \leq r^2 \leq 1$

r se llama "coeficiente de correlación" y su cuadrado r^2 se llama "coeficiente de determinación".

Si se calcula el coeficiente de correlación a partir de los datos de la muestra de la tabla I, obtenemos r = 0.63. Para poder interpretar este valor de r y descubrir cuáles valores de r son los más probables para obtenerse en los diversos tipos de relaciones entre X y Y se han trazado en la figura 5.5 varios diagramas de dispersión acompañados con los valores calculados correspondientes de r. Los primeros cuatro diagramas corresponden a dispersiones con relación lineal cada vez más acentuada. El quinto diagrama ilustra una dispersión en la que X y Y se encuentran estrechamente relacionadas, pero en la que la relación no es lineal. Este ejemplo ilustra el hecho de que r es una medida útil del grado de relación entre dos variables, solo cuando las variables están relacionadas linealmente

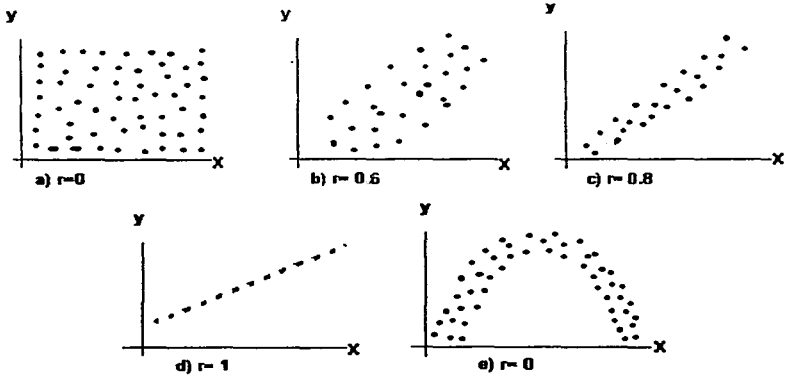


Fig. 5.5 Varios diagramas de dispersión con su correspondiente r.

Así pues, el grado de la relación está dado por la magnitud de r, mientras que el signo de r solo indica si el valor de Y tiende a crecer o a decrecer con X; el signo positivo indica que Y tiende a crecer con X, y el signo negativo, que Y tiende a decrecer cuando X crece

De las 5 figuritas anteriores (fig 5.5) concluímos que el valor de r=1 si y solo si todos los puntos del diagrama se encuentran sobre una línea recta.

Interpretación de r.- La interpretación de un coeficiente de correlación como medida del grado de relación lineal entre dos variables es una interpretación matemática pura y está completamente desprovista de implicaciones de causa y efecto. El hecho de que dos variables tiendan a aumentar o disminuír no implica que una tenga algún efecto directo o indirecto sobre la otra. Ambas pueden estar sujetas a la influencia de otras variables, de modo que resulten con una estrecha relación matemática. Por ejemplo, en un período de años, los coeficientes de correlación entre los sueldos de maestros y el consumo de licor ha resultado ser de 0.98. Durante éste lapso, se ha presentado una tendencia ascendente en sueldos y salarios de todos los tipos y una tendencia general a mayores comodidades de vida. Bajo tales condiciones, los salarios de los maestros también habrían de aumentar. Además, la tendencia general de aumento de salarios y poder adquisitivo, así como el aumento de población, se vería reflejada en un aumento en el consumo de licor. Así pues, la alta correlación refleja solamente el efecto común de una tendencia ascendente de dos variables. Los coeficientes de correlación deben manejarse con cuidado si se va a dar una información sensata respecto a la relación entre pares de variables. El utilizarlas correctamente requiere familiaridad con el campo de aplicación así como con sus propiedades matemáticas.

Los coeficientes de correlación han probado ser muy útiles; por ejemplo, para pruebas psicológicas y en otros campos en que es importante determinar la interrelación de algunas variables que se estudian simultáneamente. Así, las correlaciones entre promedios universitarios, de escuela preparatoria, puntuaciones en pruebas de aptitud, o de vocabulario y otras variables, han permitido evaluar la importancia relativa de éstos factores respecto al éxito en estudios universitarios. Aunque la correlación entre los promedios de escuela y los promedios universitarios de estudiantes presentados en la tabla I no parece muy alta, es de hecho un valor de correlación típico de esa situación. La magnitud de r depende mucho de la calidad y de la severidad del método de calificación de las escuelas de proveniencia.

Ejercicios de la sección 5.2

1) Para los siguientes datos de estaturas (x) y pesos (y) de 12 estudiantes

   a) Trazar el diagrama de dispersión
   b) Estimar el valor de r
   c) Calcular el valor de r

x	65	73	70	68	66	69	75	70	64	72	65	71
y	124	184	161	164	140	154	210	164	126	172	133	150

2) Calcular el valor de r para los siguientes datos sobre pruebas de inteligencia y promedios de calificación, después de trazar el diagrama y estimar el valor de r

P I	295	152	214	171	131	178	225	141	116	173
P C	3.4	1.6	1.2	1.0	2.0	1.6	2.0	1.4	1.0	3.6
P I	230	195	174	236	198	217	143	135	146	227
P C	3.6	1.0	2.8	2.8	1.8	2.0	1.2	2.4	2.2	2.4

3) ¿Por qué no se ve afectado el valor del coeficiente de correlación si las variables X y Y se intercambian?
4) En cuanto estimaría usted el valor de r para los siguientes pares de variables:
   a) Número de horas-hombre de trabajo y número de unidades de un producto elaborado en una industria dada.
   b) Tamaño de una ciudad, e índice de criminalidad.
   c) Costo unitario de producción de un artículo y número de unidades producidas?

5) Estimar el valor de r para los siguientes pares de variables:

   a) Calificaciones en matemáticas y en idiomas extranjeros.
   b) Consumo de mantequilla y precio de mantequilla
   c) Cantidad de lluvia en la primavera y temperatura media.

6) ¿Qué interpretación darías a la información de que la correlación entre el número de accidentes automovilísticos por año y la edad del conductor es de r= -0.60 si solo se han considerado conductores con un accidente por lo menos?

5.3 **Regresión lineal simple.**

5.3.1 La regresión lineal simple y la predicción.

Supongamos que ya determinamos que sí hay una relación lineal entre X y Y, con el diagrama de dispersión y el valor de r; ahora nos gustaría ajustar la línea recta con las observaciones, con los datos, con el diagrama de dispersión. ¿Para qué? Para hacer predicción. Para saber qué valor tomaría una variable si sabemos el valor que tomó la otra.

En este trabajo estudiaremos lo que se conoce como "modelo clásico de regresión". Dicho modelo está basado en los siguientes supuestos:

i) Se dice que los valores de la variable independiente X son "fijos". Esto significa que los valores de X son preseleccionados por el investigador, de modo que en la recolección de los datos no se permite que varíen de estos valores

preseleccionados. En este modelo, algunos autores le dan a X el nombre de variable no aleatoria y otros el de variable "matemática". Debe señalarse en este momento que es este supuesto el que clasifica al modelo que vamos a estudiar como "modelo de regresión clásico". También puede llevarse a cabo el análisis de regresión sobre datos en los que X es una variable aleatoria.

ii) La variable X se mide sin error. Debido a que ningún procedimiento de medición es perfecto, esto significa que se desprecia la magnitud del error de medición en X.

iii) Para cada valor de X existe una subpoblación de valores Y. Para que fueran válidos los procedimientos inferenciales usuales de estimación y pruebas de hipótesis, éstas subpoblaciones deben estar normalmente distribuídas

iv) Las varianzas de las subpoblaciones de Y son todas iguales.

v) Todas las medias de las subpoblaciones de Y están sobre la misma recta. Esto se conoce como suposición de linealidad. Esta suposición se puede expresar simbólicamente como

$$\mu_{y|x} = \alpha + \beta x$$

Donde $\mu_{y|x}$ es la media de la subpoblación de valores Y para un valor particular de X y α y β se llaman "coeficientes de regresión" de la población. Geométricamente, α y β representan la ordenada al origen y la pendiente respectivamente, de la recta sobre la cual se supone que están las medias.

vi) Los valores Y son estadísticamente independientes. En otras palabras, al extraer la muestra, se supone que los valores de Y elegidos en un valor de X en ninguna forma dependen de los valores de Y elegidos en otro valor de X.

Las anteriores suposiciones se pueden resumir por medio de la siguiente ecuación que se conoce como modelo de regresión

$$y = \alpha + \beta x + e$$

Donde e es llamado "término de error"

Ahora bien, si ya tenemos el diagrama de dispersión de un conjunto de datos, arbitrariamente podrías trazar ahí la recta que consideres que es la que mejor describe la tendencia de dichos datos. Pero otra persona podría trazar otra recta diferente (que no coincide con la tuya); una tercera persona podría dar otra recta diferente, etc. Se ve que necesitamos un método que nos dé la recta de mejor ajuste a los datos de la muestra; la recta que objetivamente describa mejor la tendencia lineal de los datos. Uno de ésos métodos es el que se conoce como "método de mínimos cuadrados" y es el que nos proporciona los valores a y b que son las letras con las que denotaremos a los estimadores de α y β. Los valores de a y b se obtienen resolviendo el sistema de ecuaciones lineales de 2 x 2.

$$\sum y_i = na + b\sum x_i$$
$$\sum x_i y_i = a\sum x_i + b\sum x_i^2$$

Este sistema es conocido como "ecuaciones normales" y lo puedes resolver por cualquiera de los métodos que ya conoces: suma-resta, sustitución, igualación o por la regla de Cramer.

Así es como obtenemos y= a+bx la recta de mejor ajuste a los datos, a los puntos del diagrama de dispersión. En éste caso se dice que hacemos regresión de Y sobre X.

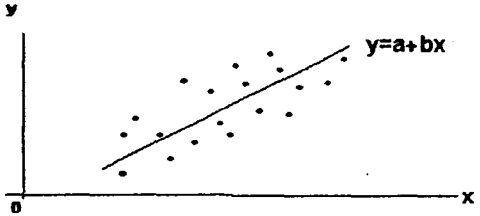


Fig. 5.6 Recta de regresión ajustada a los datos

Este método es llamado de mínimos cuadrados porque minimiza la suma de los cuadrados de las distancias

$$\sum e_i^2$$

de cada punto del diagrama de dispersión a la recta ajustada y_c.

En efecto, cada ordenada de cada punto del diagrama de dispersión satisface y=a+bx+e, y entonces e=y-(a+bx) o más precisamente

$$e_i = y_i - (a + bx_i)$$

En el trabajo matemático para minimizar

$$\sum e^2 = \sum (y-(a+bx))^2$$

llegamos a las ecuaciones normales y es así como encontramos los valores de a y de b; es decir, a y b son los valores que minimizan

$$\sum e_i^2$$

Cuando se dice que la recta de mínimos cuadrados es la mejor, quiere decir que es la mejor en el sentido antes descrito, en el sentido de que minimiza la última suma de cuadrados indicada arriba.

Ya que resolvimos las ecuaciones normales y tenemos la recta de mejor ajuste en el sentido de mínimos cuadrados, y_c =a+bx, podemos predecir el valor de "y" correspondiente a una "x" dada.

EJEMPLO 26: Una gran compañía dedicada a la venta por correo utiliza el peso del correo que le llega para determinar cuántos de sus empleados debe asignar para empaquetar pedidos en un determinado día. Supóngase que se toman los datos de la tabla 5.3.1 de los archivos de la compañía en diez días diferentes. Suponiendo una relación lineal, calcúlese la línea de regresión estimada. Utilícese para predecir la cantidad de hombres-hora necesarios para empaquetar los pedidos medios si el correo a las 7 de la mañana daba un peso de 673 kilogramos. Supóngase que cada hombre trabaja unas 8 horas diarias y estímese el número de hombres necesarios.

PESO DEL CORREO A LAS 7 HORAS (En cientos de kilogramos)	HOMBRES-HORA NECESARIOS PARA EMPAQUETAR LOS PEDIDOS (En miles de horas)
5.21	12.6
7.16	17.3
6.34	15.2
8.41	18.5
6.94	15.8
6.52	15.0
7.33	16.8
5.87	13.8
6.61	14.9
8.03	18.0

Tabla 5.3.1

SOLUCIÓN: Claramente se siente que el peso del correo es la variable independiente y no los hombres-hora necesarios; así que haremos regresión de los hombres-hora sobre el peso del correo, y no al revés:

Sean X= peso del correo a las 7 horas y Y= hombres-hora necesarios para empaquetar los pedidos

$$Como \sum y_i = 157.9, \quad n = 10, \quad \sum x_i = 68.42$$

$$\sum x_i^2 = 476.3662, \quad \sum x_i y_i = 1096.098$$

las ecuaciones normales se convierten en

$$10a + (68.42)b = 157.9$$

$$(68.42)a + (476.3662)b = 1096.098$$

Resolviendo con la regla de Cramer:

$$a = \frac{\Delta a}{\Delta} = \frac{223.19782}{82.3656} = 2.709$$

$$y \quad b = \frac{\Delta b}{\Delta} = \frac{157.462}{82.3656} = 1.912$$

La recta ajustada es $y_c = 2.7 + (1.9)x$

Dicha recta y el diagrama de dispersión se muestran en la siguiente figura 5.3.1

Fig. 5.7 Diagrama de dispersión para el ejemplo 26

Para x= 6.73 tenemos y_c =2.7+(1.9)(6.73)= 15.487

Es decir, cuando el correo pese 673 kilogramos debemos asignar 15.487 hombres-hora; o sea 15,487 horas. Si cada hombre trabaja 8 horas diarias, se necesitan 1936 hombres para acabar en un día (pues el resultado teórico es 15,487 / 8 = 1935.875)

EJERCICIO B: Los siguientes datos son para dureza y resistencia a la tensión del aluminio vaciado en troqueles. Encontrar la ecuación de la línea de regresión para estimar la resistencia a la tensión, partiendo de la dureza y trácela sobre el diagrama de dispersión. ¿Cuál sería la resistencia a la tensión para una dureza con valor de 50?

RT (y)	293	349	368	301	340	308	354	313	322	334	377	247
D (x)	53	70	84	55	78	64	71	53	82	67	70	56

Por cierto, en general la predicción de "y" solo puede hacerse para valores de x entre el valor más pequeño x_m y el valor más grande x_M de los datos. No puedes extrapolar- a menos que conozcas lo suficientemente bien el trozo de realidad en estudio y concluyas que sí puedes hacerlo- porque pudiera ser que la situación real fuera como el de la figura siguiente

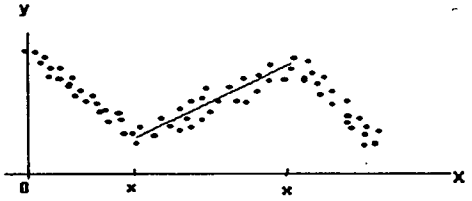


Fig. 5.8

Al predecir "y" para x=0 resulta $y_c(0) < 0$, cuando en realidad es $y_c(0) > 0$. Algo similar ocurre en el otro extremo de la recta de regresión ajustada.

5.3.2 Regresión y correlación.

En el modelo clásico de regresión que es el que hemos discutido hasta este punto, solo "Y" es aleatoria, a la cual se le ha dado el nombre de variable dependiente. La variable "X" se define como una variable fija (no aleatoria o matemática) y se conoce como variable independiente. Recuérdese también que se han descrito las observaciones como que se obtienen mediante la preselección de los valores de X y determinando los valores correspondientes de Y.

Cuando tanto Y como X son variables aleatorias, se tiene lo que se conoce como *modelo de correlación*. Típicamente, bajo el modelo de correlación, se obtienen observaciones muestra, seleccionando una muestra aleatoria de *las unidades de asociación* (que pueden ser personas, lugares, animales, puntos en el tiempo, o cualquier otro elemento sobre el cual se toman las dos medidas) y tomando una medida de X y una medida de Y sobre cada una. En éste procedimiento, no se preseleccionan los valores de X, sino que ocurren al azar, dependiendo de la unidad de asociación seleccionada en la muestra.

Aunque no puede llevarse a cabo con sentido el análisis de correlación bajo el modelo clásico de regresión, puede realizarse el análisis de regresión bajo el modelo de correlación. La correlación, que comprende dos variables, implica una correlación entre las variables que pone a ambas sobre un mismo terreno y no las distingue, refiriéndose a una como la dependiente y a la otra como la variable independiente. En efecto, en los procedimientos básicos de cálculo, que son los mismos que para el modelo de regresión, puede ajustarse una recta a los datos, ya sea minimizando

$$\Sigma(y_i - y_c)^2$$

o bien, minimizando

$$\Sigma(x_i - x_c)^2$$

En otras palabras, puede hacerse una regresión de X sobre Y. En general, las rectas ajustadas en los dos casos serán diferentes y surge una pregunta lógica : ¿cuál recta ajustar?

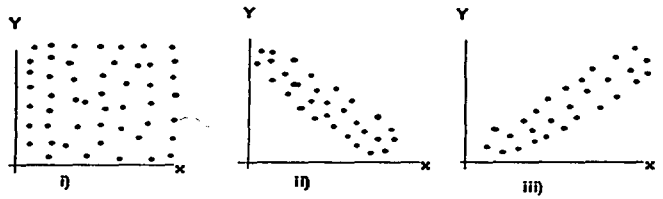
Si el objetivo es únicamente obtener una medida de la relación entre las dos variables, no importa qué recta se ajuste, ya que la medida que por lo común se calcula será la misma en cualquier caso. Pero si se desea usar la ecuación que describe la relación entre las dos variables para estimación, prueba de hipótesis o predicción, sí importa cuál recta se ajuste. La variable para la que se desean estimar las medias o hacer predicciones debe tratarse como la variable dependiente, es decir, debe regresarse esta variable sobre la otra variable.

Por cierto, la estimación por intervalo y la prueba de hipótesis en regresión son temas que podrías empezar a investigar.

Buena suerte en tus cursos superiores de estadística.

5.4 Miscelánea de ejercicios del capítulo 5

1) Actividad I.
   a) En las siguientes gráficas indica cuándo sí y cuándo no existe correlación y si ésta es positiva o negativa.



   b) Indica la relación que con respecto a cero guardan los coeficientes de correlación de las gráficas anteriores
   c) Traza una recta que tú consideres de "mejor ajuste" en las gráficas anteriores e intenta asignarles una expresión matemática.

2) Actividad II. Los siguientes valores representan la relación que existe entre el peso en kilogramos (x) y la presión sanguínea diastólica (y) de 16 adultos varones cuyas edades están entre 21 y 28 años:

X	86	89	72	73	78	87	86	67	100	66	76	85	81	85	67	79
y	76	76	74	70	80	68	90	70	96	80	90	72	76	80	68	72

   a) Construye el diagrama de dispersión correspondiente a éstos pares de valores
   b) Estima en base a la gráfica por ti trazada, el signo del coeficiente de correlación r.
   c) Calcula el valor de r, la correlación entre el peso y la presión sanguínea.
   d) Obtener la ecuación de la recta de mejor ajuste.
   e) ¿Parece existir una relación entre las dos variables?
   f) ¿Significa esto que un aumento en el peso, tendrá por resultado un incremento en la presión sanguínea?
   g) Estime la presión sanguínea de alguien que pesa 120 kg
   h) Estime la presión sanguínea de alguien que pesa 84 kg.
   i) ¿Cuál de estas dos estimaciones será más confiable y por qué? Explica tu respuesta.

SONRÍE ...

Si un inventor fracasa, es un chiflado. Si triunfa, es un genio.

APÉNDICE A (TABLAS)

Binomial acumulada

n	r	p 0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99	r
2	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	020	098	190	360	510	640	750	840	910	960	990	998	1 —	1
	2	0 +	002	010	040	090	160	250	360	490	640	810	902	980	2
3	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	030	143	271	488	657	784	875	936	973	992	999	1 —	1 —	1
	2	0 +	007	028	104	216	352	500	648	784	896	972	993	1 —	2
	3	0 +	0 +	001	008	027	064	125	216	343	512	729	857	970	3
4	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	039	185	344	590	760	870	938	974	992	998	1 —	1 —	1 —	1
	2	001	014	052	181	348	525	688	821	916	973	996	1 —	1 —	2
	3	0 +	0 +	004	027	084	179	312	475	652	819	948	986	999	3
	4	0 +	0 +	0 +	002	008	026	062	130	240	410	656	815	961	4
5	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	049	226	410	672	832	922	969	990	998	1 —	1 —	1 —	1 —	1
	2	001	023	081	263	472	663	812	913	969	993	1 —	1 —	1 —	2
	3	0 +	001	009	058	163	317	500	683	837	942	991	999	1 —	3
	4	0 +	0 +	0 +	007	031	087	188	337	528	737	919	977	999	4
	5	0 +	0 +	0 +	0 +	002	010	031	078	168	328	590	774	951	5
6	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	059	265	469	738	882	953	984	996	999	1 —	1 —	1 —	1 —	1
	2	001	033	114	345	580	767	891	959	989	998	1 —	1 —	1 —	2
	3	0 +	002	016	099	256	456	656	821	930	983	999	1 —	1 —	3
	4	0 +	0 +	001	017	070	179	344	544	744	901	984	998	1 —	4
	5	0 +	0 +	0 +	002	011	041	109	233	420	655	886	967	999	5
	6	0 +	0 +	0 +	0 +	001	004	016	047	118	262	531	735	941	6
7	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	068	302	522	790	918	972	992	998	1 —	1 —	1 —	1 —	1 —	1
	2	002	044	150	423	671	841	938	981	996	1 —	1 —	1 —	1 —	2
	3	0 +	004	026	148	353	580	773	904	971	995	1 —	1 —	1 —	3
	4	0 +	0 +	003	033	126	290	500	710	874	967	997	1 —	1 —	4
	5	0 +	0 +	0 +	005	029	096	227	420	647	852	974	996	1 —	5
	6	0 +	0 +	0 +	0 +	004	019	062	159	329	577	850	956	998	6
	7	0 +	0 +	0 +	0 +	0 +	002	008	028	082	210	478	698	932	7

Binomial acumulada

n	r	p = 0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99	r
8	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	077	337	570	832	942	983	996	999	1 —	1 —	1 —	1 —	1 —	1
	2	003	057	187	497	745	894	965	991	999	1 —	1 —	1 —	1 —	2
	3	0 +	006	038	203	448	685	855	950	989	999	1 —	1 —	1 —	3
	4	0 +	0 +	005	056	194	406	637	826	942	990	1 —	1 —	1 —	4
	5	0 +	0 +	0 +	010	058	174	363	594	806	944	995	1 —	1 —	5
	6	0 +	0 +	0 +	001	011	050	145	315	552	797	962	994	1 —	6
	7	0 +	0 +	0 +	0 +	001	009	035	106	255	503	813	943	997	7
	8	0 +	0 +	0 +	0 +	0 +	001	004	017	058	168	430	663	923	8
9	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	086	370	613	866	960	990	998	1 —	1 —	1 —	1 —	1 —	1 —	1
	2	003	071	225	564	804	929	980	996	1 —	1 —	1 —	1 —	1 —	2
	3	0 +	008	053	262	537	768	910	975	996	1 —	1 —	1 —	1 —	3
	4	0 +	001	008	086	270	517	746	901	975	997	1 —	1 —	1 —	4
	5	0 +	0 +	001	020	099	267	500	733	901	980	999	1 —	1 —	5
	6	0 +	0 +	0 +	003	025	099	254	483	730	914	992	999	1 —	6
	7	0 +	0 +	0 +	0 +	004	025	090	232	463	738	947	992	1 —	7
	8	0 +	0 +	0 +	0 +	0 +	004	020	071	196	436	775	929	997	8
	9	0 +	0 +	0 +	0 +	0 +	0 +	002	010	040	134	387	630	914	9
10	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	096	401	651	893	972	994	999	1 —	1 —	1 —	1 —	1 —	1 —	1
	2	004	086	264	624	851	954	989	998	1 —	1 —	1 —	1 —	1 —	2
	3	0 +	012	070	322	617	833	945	988	998	1 —	1 —	1 —	1 —	3
	4	0 +	001	013	121	350	618	828	945	986	998	1 —	1 —	1 —	4
	5	0 +	0 +	002	033	150	367	623	834	953	994	1 —	1 —	1 —	5
	6	0 +	0 +	0 +	006	047	166	377	633	850	967	998	1 —	1 —	6
	7	0 +	0 +	0 +	001	011	055	172	382	650	879	987	999	1 —	7
	8	0 +	0 +	0 +	0 +	002	012	055	167	383	678	930	988	1 —	8
	9	0 +	0 +	0 +	0 +	0 +	002	011	046	149	376	736	914	996	9
	10	0 +	0 +	0 +	0 +	0 +	0 +	001	006	028	107	349	599	904	10
11	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	105	431	686	914	980	996	1 —	1 —	1 —	1 —	1 —	1 —	1 —	1
	2	005	102	303	678	887	970	994	999	1 —	1 —	1 —	1 —	1 —	2
	3	0 +	015	090	383	687	881	967	994	999	1 —	1 —	1 —	1 —	3

Binomial acumulada

n	r	p 0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99	r
11	4	0+	002	019	161	430	704	887	971	996	1−	1−	1−	1−	4
	5	0+	0+	003	050	210	467	726	901	978	998	1−	1−	1−	5
	6	0+	0+	0+	012	078	247	500	753	922	988	1−	1−	1−	6
	7	0+	0+	0+	002	022	099	274	533	790	950	997	1−	1−	7
	8	0+	0+	0+	0+	004	029	113	296	570	839	981	998	1−	8
	9	0+	0+	0+	0+	001	006	033	119	313	617	910	985	1−	9
	10	0+	0+	0+	0+	0+	001	006	030	113	322	697	898	995	10
	11	0+	0+	0+	0+	0+	0+	0+	004	020	086	314	569	895	11
12	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	114	460	718	931	986	998	1−	1−	1−	1−	1−	1−	1−	1
	2	006	118	341	725	915	980	997	1−	1−	1−	1−	1−	1−	2
	3	0+	020	111	442	747	917	981	997	1−	1−	1−	1−	1−	3
	4	0+	002	026	205	507	775	927	985	998	1−	1−	1−	1−	4
	5	0+	0+	004	073	276	562	806	943	991	999	1−	1−	1−	5
	6	0+	0+	001	019	118	335	613	842	961	996	1−	1−	1−	6
	7	0+	0+	0+	004	039	158	387	665	882	981	999	1−	1−	7
	8	0+	0+	0+	001	009	057	194	438	724	927	996	1−	1−	8
	9	0+	0+	0+	0+	002	015	073	225	493	795	974	998	1−	9
	10	0+	0+	0+	0+	0+	003	019	083	253	558	889	980	1−	10
	11	0+	0+	0+	0+	0+	0+	003	020	085	275	659	882	994	11
	12	0+	0+	0+	0+	0+	0+	0+	002	014	069	282	540	886	12
13	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	122	487	746	945	990	999	1−	1−	1−	1−	1−	1−	1−	1
	2	007	135	379	766	936	987	998	1−	1−	1−	1−	1−	1−	2
	3	0+	025	134	498	798	942	989	999	1−	1−	1−	1−	1−	3
	4	0+	003	034	253	579	831	954	992	999	1−	1−	1−	1−	4
	5	0+	0+	006	099	346	647	867	968	996	1−	1−	1−	1−	5
	6	0+	0+	001	030	165	426	709	902	982	999	1−	1−	1−	6
	7	0+	0+	0+	007	062	229	500	771	938	993	1−	1−	1−	7
	8	0+	0+	0+	001	018	098	291	574	835	970	999	1−	1−	8
	9	0+	0+	0+	0+	004	032	133	353	654	901	994	1−	1−	9
	10	0+	0+	0+	0+	001	008	046	169	421	747	966	997	1−	10
	11	0+	0+	0+	0+	0+	001	011	058	202	502	866	975	1−	11
	12	0+	0+	0+	0+	0+	0+	002	013	064	234	621	865	993	12
	13	0+	0+	0+	0+	0+	0+	0+	001	010	055	254	513	878	13

Binomial acumulada

n	r	p 0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99	r
14	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	131	512	771	956	993	999	1 —	1 —	1 —	1 —	1 —	1 —	1 —	1
	2	008	153	415	802	953	992	999	1 —	1 —	1 —	1 —	1 —	1 —	2
	3	0 +	030	158	552	839	960	994	999	1 —	1 —	1 —	1 —	1 —	3
	4	0 +	004	044	302	645	876	971	996	1 —	1 —	1 —	1 —	1 —	4
	5	0 +	0 +	009	130	416	721	910	982	998	1 —	1 —	1 —	1 —	5
	6	0 +	0 +	001	044	219	514	788	942	992	1 —	1 —	1 —	1 —	6
	7	0 +	0 +	0 +	012	093	308	605	850	969	998	1 —	1 —	1 —	7
	8	0 +	0 +	0 +	002	031	150	395	692	907	988	1 —	1 —	1 —	8
	9	0 +	0 +	0 +	0 +	008	058	212	486	781	956	999	1 —	1 —	9
	10	0 +	0 +	0 +	0 +	002	018	090	279	584	870	991	1 —	1 —	10
	11	0 +	0 +	0 +	0 +	0 +	004	029	124	355	698	956	996	1 —	11
	12	0 +	0 +	0 +	0 +	0 +	001	006	040	161	448	842	970	1 —	12
	13	0 +	0 +	0 +	0 +	0 +	0 +	001	008	047	198	585	847	992	13
	14	0 +	0 +	0 +	0 +	0 +	0 +	0 +	001	007	044	229	488	869	14
15	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	140	537	794	965	995	1 —	1 —	1 —	1 —	1 —	1 —	1 —	1 —	1
	2	010	171	451	833	965	995	1 —	1 —	1 —	1 —	1 —	1 —	1 —	2
	3	0 +	036	184	602	873	973	996	1 —	1 —	1 —	1 —	1 —	1 —	3
	4	0 +	005	056	352	703	909	982	998	1 —	1 —	1 —	1 —	1 —	4
	5	0 +	001	013	164	485	783	941	991	999	1 —	1 —	1 —	1 —	5
	6	0 +	0 +	002	061	278	597	849	966	996	1 —	1 —	1 —	1 —	6
	7	0 +	0 +	0 +	018	131	390	696	905	985	999	1 —	1 —	1 —	7
	8	0 +	0 +	0 +	004	050	213	500	787	950	996	1 —	1 —	1 —	8
	9	0 +	0 +	0 +	001	015	095	304	610	869	982	1 —	1 —	1 —	9
	10	0 +	0 +	0 +	0 +	004	034	151	403	722	939	998	1 —	1 —	10
	11	0 +	0 +	0 +	0 +	001	009	059	217	515	836	987	999	1 —	11
	12	0 +	0 +	0 +	0 +	0 +	002	018	091	297	648	944	995	1 —	12
	13	0 +	0 +	0 +	0 +	0 +	0 +	004	027	127	398	816	964	1 —	13
	14	0 +	0 +	0 +	0 +	0 +	0 +	0 +	005	035	167	549	829	990	14
	15	0 +	0 +	0 +	0 +	0 +	0 +	0 +	0 +	005	035	206	463	860	15
16	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	149	560	815	972	997	1 —	1 —	1 —	1 —	1 —	1 —	1 —	1 —	1
	2	011	189	485	859	974	997	1 —	1 —	1 —	1 —	1 —	1 —	1 —	2
	3	001	043	211	648	901	982	998	1 —	1 —	1 —	1 —	1 —	1 —	3

Binomial acumulada

n	r	p 0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99	r
16	4	0+	007	068	402	754	935	989	999	1—	1—	1—	1—	1—	4
	5	0+	001	017	202	550	833	962	995	1—	1—	1—	1—	1—	5
	6	0+	0+	003	082	340	671	895	981	998	1—	1—	1—	1—	6
	7	0+	0+	001	027	175	473	773	942	993	1—	1—	1—	1—	7
	8	0+	0+	0+	007	074	284	598	858	974	999	1—	1—	1—	8
	9	0+	0+	0+	001	026	142	402	716	926	993	1—	1—	1—	9
	10	0+	0+	0+	0+	007	058	227	527	825	973	999	1—	1—	10
	11	0+	0+	0+	0+	002	019	105	329	660	918	997	1—	1—	11
	12	0+	0+	0+	0+	0+	005	038	167	450	798	983	999	1—	12
	13	0+	0+	0+	0+	0+	001	011	065	246	598	932	993	1—	13
	14	0+	0+	0+	0+	0+	0+	002	018	099	352	789	957	999	14
	15	0+	0+	0+	0+	0+	0+	0+	003	026	141	515	811	989	15
	16	0+	0+	0+	0+	0+	0+	0+	0+	003	028	185	440	851	16
17	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	157	582	833	977	998	1—	1—	1—	1—	1—	1—	1—	1—	1
	2	012	208	518	882	981	998	1—	1—	1—	1—	1—	1—	1—	2
	3	001	050	238	690	923	988	999	1—	1—	1—	1—	1—	1—	3
	4	0+	009	083	451	798	954	994	1—	1—	1—	1—	1—	1—	4
	5	0+	001	022	242	611	874	975	997	1—	1—	1—	1—	1—	5
	6	0+	0+	005	106	403	736	928	989	999	1—	1—	1—	1—	6
	7	0+	0+	001	038	225	552	834	965	997	1—	1—	1—	1—	7
	8	0+	0+	0+	011	105	359	685	908	987	1—	1—	1—	1—	8
	9	0+	0+	0+	003	040	199	500	801	960	997	1—	1—	1—	9
	10	0+	0+	0+	0+	013	092	315	641	895	989	1—	1—	1—	10
	11	0+	0+	0+	0+	003	035	166	448	775	962	999	1—	1—	11
	12	0+	0+	0+	0+	001	011	072	264	597	894	995	1—	1—	12
	13	0+	0+	0+	0+	0+	003	025	126	389	758	978	999	1—	13
	14	0+	0+	0+	0+	0+	0+	006	046	202	549	917	991	1—	14
	15	0+	0+	0+	0+	0+	0+	001	012	077	310	762	950	999	15
	16	0+	0+	0+	0+	0+	0+	0+	002	019	118	482	792	988	16
	17	0+	0+	0+	0+	0+	0+	0+	0+	002	023	167	418	843	17
18	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	165	603	850	982	998	1—	1—	1—	1—	1—	1—	1—	1—	1
	2	014	226	550	901	986	999	1—	1—	1—	1—	1—	1—	1—	2
	3	001	058	266	729	940	992	999	1—	1—	1—	1—	1	1	3

Binomial acumulada

n	r	p 0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99	r
	4	0+	011	098	499	835	967	996	1−	1−	1−	1−	1−	1−	4
	5	0+	002	028	284	667	906	985	999	1−	1−	1−	1−	1−	5
	6	0+	0+	006	133	466	791	952	994	1−	1−	1−	1−	1−	6
	7	0+	0+	001	051	278	626	881	980	999	1−	1−	1−	1−	7
	8	0+	0+	0+	016	141	437	760	942	994	1−	1−	1−	1−	8
	9	0+	0+	0+	004	060	263	593	865	979	999	1−	1−	1−	9
	10	0+	0+	0+	001	021	135	407	737	940	996	1−	1−	1−	10
	11	0+	0+	0+	0+	006	058	240	563	859	984	1−	1−	1−	11
	12	0+	0+	0+	0+	001	020	119	374	722	949	999	1−	1−	12
	13	0+	0+	0+	0+	0+	006	048	209	534	867	994	1−	1−	13
	14	0+	0+	0+	0+	0+	001	015	094	333	716	972	998	1−	14
	15	0+	0+	0+	0+	0+	0+	004	033	165	501	902	989	1−	15
	16	0+	0+	0+	0+	0+	0+	001	008	060	271	734	942	999	16
	17	0+	0+	0+	0+	0+	0+	0+	001	014	099	450	774	986	17
	18	0+	0+	0+	0+	0+	0+	0+	0+	002	018	150	397	835	18
19	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	174	623	865	986	999	1−	1−	1−	1−	1−	1−	1−	1−	1
	2	015	245	580	917	990	999	1−	1−	1−	1−	1−	1−	1−	2
	3	001	067	295	763	954	995	1−	1−	1−	1−	1−	1−	1−	3
	4	0+	013	115	545	867	977	998	1−	1−	1−	1−	1−	1−	4
	5	0+	002	035	327	718	930	990	999	1−	1−	1−	1−	1−	5
	6	0+	0+	009	163	526	837	968	997	1−	1−	1−	1−	1−	6
	7	0+	0+	002	068	334	692	916	988	999	1−	1−	1−	1−	7
	8	0+	0+	0+	023	182	512	820	965	997	1−	1−	1−	1−	8
	9	0+	0+	0+	007	084	333	676	912	989	1−	1−	1−	1−	9
	10	0+	0+	0+	002	033	186	500	814	967	998	1−	1−	1−	10
	11	0+	0+	0+	0+	011	088	324	667	916	993	1−	1−	1−	11
	12	0+	0+	0+	0+	003	035	180	488	818	977	1−	1−	1−	12
	13	0+	0+	0+	0+	001	012	084	308	666	932	998	1−	1−	13
	14	0+	0+	0+	0+	0+	003	032	163	474	837	991	1−	1−	14
	15	0+	0+	0+	0+	0+	001	010	070	282	673	965	998	1−	15
	16	0+	0+	0+	0+	0+	0+	002	023	133	455	885	987	1−	16
	17	0+	0+	0+	0+	0+	0+	0+	005	046	237	705	933	999	17
	18	0+	0+	0+	0+	0+	0+	0+	001	010	083	420	755	985	18
	19	0+	0+	0+	0+	0+	0+	0+	0+	001	014	135	377	826	19

Binomial acumulado

n	r	p 0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99	r
20	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	182	642	878	988	999	1—	1—	1—	1—	1—	1—	1—	1—	1
	2	017	264	608	931	992	999	1—	1—	1—	1—	1—	1—	1—	2
	3	001	075	323	794	965	996	1—	1—	1—	1—	1—	1—	1—	3
	4	0+	016	133	589	893	984	999	1—	1—	1—	1—	1—	1—	4
	5	0+	003	043	370	762	949	994	1—	1—	1—	1—	1—	1—	5
	6	0+	0+	011	196	584	874	979	998	1—	1—	1—	1—	1—	6
	7	0+	0+	002	087	392	750	942	994	1—	1—	1—	1—	1—	7
	8	0+	0+	0+	032	228	584	868	979	999	1—	1—	1—	1—	8
	9	0+	0+	0+	010	113	404	748	943	995	1—	1—	1—	1—	9
	10	0+	0+	0+	003	048	245	588	872	983	999	1—	1—	1—	10
	11	0+	0+	0+	001	017	128	412	755	952	997	1—	1—	1—	11
	12	0+	0+	0+	0+	005	057	252	596	887	990	1—	1—	1—	12
	13	0+	0+	0+	0+	001	021	132	416	772	968	1—	1—	1—	13
	14	0+	0+	0+	0+	0+	006	058	250	608	913	998	1—	1—	14
	15	0+	0+	0+	0+	0+	002	021	126	416	804	989	1—	1—	15
	16	0+	0+	0+	0+	0+	0+	006	051	238	630	957	997	1—	16
	17	0+	0+	0+	0+	0+	0+	001	016	107	411	867	984	1—	17
	18	0+	0+	0+	0+	0+	0+	0+	004	035	206	677	925	999	18
	19	0+	0+	0+	0+	0+	0+	0+	001	008	069	392	736	983	19
	20	0+	0+	0+	0+	0+	0+	0+	0+	001	012	122	358	818	20
21	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	190	659	891	991	999	1—	1—	1—	1—	1—	1—	1—	1—	1
	2	019	283	635	942	994	1—	1—	1—	1—	1—	1—	1—	1—	2
	3	001	085	352	821	973	998	1—	1—	1—	1—	1—	1—	1—	3
	4	0+	019	152	630	914	989	999	1—	1—	1—	1—	1—	1—	4
	5	0+	003	052	414	802	963	996	1—	1—	1—	1—	1—	1—	5
	6	0+	0+	014	231	637	904	987	999	1—	1—	1—	1—	1—	6
	7	0+	0+	003	109	449	800	961	996	1—	1—	1—	1—	1—	7
	8	0+	0+	001	043	277	650	905	988	999	1—	1—	1—	1—	8
	9	0+	0+	0+	014	148	476	808	965	998	1—	1—	1—	1—	9
	10	0+	0+	0+	004	068	309	668	915	991	1—	1—	1—	1—	10
	11	0+	0+	0+	001	026	174	500	826	974	999	1—	1—	1—	11
	12	0+	0+	0+	0+	009	085	332	691	932	996	1—	1—	1—	12
	13	0+	0+	0+	0+	002	035	192	524	852	986	1—	1—	1—	13

n	r	p 0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99	r
21	14	0+	0+	0+	0+	001	012	095	350	723	957	999	1—	1—	14
	15	0+	0+	0+	0+	0+	004	039	200	551	891	997	1—	1—	15
	16	0+	0+	0+	0+	0+	001	013	096	363	769	986	1—	1—	16
	17	0+	0+	0+	0+	0+	0+	004	037	198	586	948	997	1—	17
	18	0+	0+	0+	0+	0+	0+	001	011	086	370	848	981	1—	18
	19	0+	0+	0+	0+	0+	0+	0+	002	027	179	648	915	999	19
	20	0+	0+	0+	0+	0+	0+	0+	0+	006	058	365	717	981	20
	21	0+	0+	0+	0+	0+	0+	0+	0+	001	·009	109	341	810	21
22	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	198	676	902	993	1—	1—	1—	1—	1—	1—	1—	1—	1—	1
	2	020	302	661	952	996	1—	1—	1—	1—	1—	1—	1—	1—	2
	3	001	095	380	846	979	998	1—	1—	1—	1—	1—	1—	1—	3
	4	0+	022	172	668	932	992	1—	1—	1—	1—	1—	1—	1—	4
	5	0+	004	062	457	835	973	998	1—	1—	1—	1—	1—	1—	5
	6	0+	001	018	267	687	928	992	1—	1—	1—	1—	1—	1—	6
	7	0+	0+	004	133	506	842	974	998	1—	1—	1—	1—	1—	7
	8	0+	0+	001	056	329	710	933	993	1—	1—	1—	1—	1—	8
	9	0+	0+	0+	020	186	546	857	979	999	1—	1—	1—	1—	9
	10	0+	0+	0+	006	092	376	738	945	996	1—	1—	1—	1—	10
	11	0+	0+	0+	002	039	228	584	879	986	1—	1—	1—	1—	11
	12	0+	0+	0+	0+	014	121	416	772	961	998	1—	1—	1—	12
	13	0+	0+	0+	0+	004	055	262	624	908	994	1—	1—	1—	13
	14	0+	0+	0+	0+	001	021	143	454	814	980	1—	1—	1—	14
	15	0+	0+	0+	0+	0+	007	067	290	671	944	999	1—	1—	15
	16	0+	0+	0+	0+	0+	002	026	158	494	867	996	1—	1—	16
	17	0+	0+	0+	0+	0+	0+	008	072	313	733	982	999	1—	17
	18	0+	0+	0+	0+	0+	0+	002	027	165	543	938	996	1—	18
	19	0+	0+	0+	0+	0+	0+	0+	008	068	332	828	978	1—	19
	20	0+	0+	0+	0+	0+	0+	0+	002	021	154	620	905	999	20
	21	0+	0+	0+	0+	0+	0+	0+	0+	004	048	339	698	980	21
	22	0+	0+	0+	0+	0+	0+	0+	0+	0+	007	098	324	802	22
23	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	206	693	911	994	1—	1—	1—	1—	1—	1—	1—	1—	1—	1
	2	022	321	685	960	997	1—	1—	1—	1—	1—	1—	1—	1—	2
	3	002	105	408	867	984	999	1—	1—	1—	1—	1—	1—	1—	3

Binomial acumulada

n	r	p 0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99	r
23	4	0+	026	193	703	946	995	1−	1−	1−	1−	1−	1−	1−	
	5	0+	005	073	499	864	981	999	1−	1−	1−	1−	1−	1−	
	6	0+	001	023	305	731	946	995	1−	1−	1−	1−	1−	1−	
	7	0+	0+	006	160	560	876	983	999	1−	1−	1−	1−	1−	
	8	0+	0+	001	072	382	763	953	996	1−	1−	1−	1−	1−	
	9	0+	0+	0+	027	229	612	895	987	999	1−	1−	1−	1−	
	10	0+	0+	0+	009	120	444	798	965	998	1−	1−	1−	1−	1
	11	0+	0+	0+	003	055	287	661	919	998	1−	1−	1−	1−	1
	12	0+	0+	0+	001	021	164	500	836	979	999	1−	1−	1−	1
	13	0+	0+	0+	0+	007	081	339	713	945	997	1−	1−	1−	1
	14	0+	0+	0+	0+	002	035	202	556	880	991	1−	1−	1−	1
	15	0+	0+	0+	0+	001	013	105	388	771	973	1−	1−	1−	1
	16	0+	0+	0+	0+	0+	004	047	237	618	928	999	1−	1−	1
	17	0+	0+	0+	0+	0+	001	017	124	440	840	994	1−	1−	1
	18	0+	0+	0+	0+	0+	0+	005	054	269	695	977	999	1−	1
	19	0+	0+	0+	0+	0+	0+	001	019	136	501	927	995	1−	1
	20	0+	0+	0+	0+	0+	0+	0+	005	054	297	807	974	1−	2
	21	0+	0+	0+	0+	0+	0+	0+	001	016	133	592	895	998	2
	22	0+	0+	0+	0+	0+	0+	0+	0+	003	040	315	679	978	2
	23	0+	0+	0+	0+	0+	0+	0+	0+	0+	006	089	307	794	2
24	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	214	708	920	995	1−	1−	1−	1−	1−	1−	1−	1−	1−	1
	2	024	339	708	967	998	1−	1−	1−	1−	1−	1−	1−	1−	2
	3	002	116	436	885	988	999	1−	1−	1−	1−	1−	1−	1−	3
	4	0+	030	214	736	958	996	1−	1−	1−	1−	1−	1−	1−	4
	5	0+	006	085	540	889	987	999	1−	1−	1−	1−	1−	1−	5
	6	0+	001	028	344	771	960	997	1−	1−	1−	1−	1−	1−	6
	7	0+	0+	007	189	611	904	989	999	1−	1−	1−	1−	1−	7
	8	0+	0+	002	089	435	808	968	998	1−	1−	1−	1−	1−	8
	9	0+	0+	0+	036	276	672	924	992	1−	1−	1−	1−	1−	9
	10	0+	0+	0+	013	153	511	846	978	999	1−	1−	1−	1−	10
	11	0+	0+	0+	004	074	350	729	947	996	1−	1−	1−	1−	11
	12	0+	0+	0+	001	031	213	581	886	988	1−	1−	1−	1−	12
	13	0+	0+	0+	0+	012	114	419	787	969	999	1−	1−	1−	13
	14	0+	0+	0+	0+	004	053	271	650	926	996	1−	1−	1−	14

Binomial acumulada

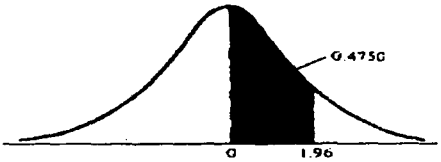
n	r	p 0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99	r
24	15	0+	0+	0+	0+	001	022	154	489	847	987	1−	1−	1−	15
	16	0+	0+	0+	0+	0+	008	076	328	725	964	1−	1−	1−	16
	17	0+	0+	0+	0+	0+	002	032	192	565	911	998	1−	1−	17
	18	0+	0+	0+	0+	0+	001	011	096	389	811	993	1−	1−	18
	19	0+	0+	0+	0+	0+	0+	003	040	229	656	972	999	1−	19
	20	0+	0+	0+	0+	0+	0+	001	013	111	460	915	994	1−	20
	21	0+	0+	0+	0+	0+	0+	0+	004	042	264	786	970	1−	21
	22	0+	0+	0+	0+	0+	0+	0+	001	012	115	564	884	998	22
	23	0+	0+	0+	0+	0+	0+	0+	0+	002	033	292	661	976	23
	24	0+	0+	0+	0+	0+	0+	0+	0+	0+	005	080	292	786	24
25	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	222	723	928	996	1−	1−	1−	1−	1−	1−	1−	1−	1−	1
	2	026	358	729	973	998	1−	1−	1−	1−	1−	1−	1−	1−	2
	3	002	127	463	902	991	1−	1−	1−	1−	1−	1−	1−	1−	3
	4	0+	034	236	766	967	998	1−	1−	1−	1−	1−	1−	1−	4
	5	0+	007	098	579	910	991	1−	1−	1−	1−	1−	1−	1−	5
	6	0+	001	033	383	807	971	998	1−	1−	1−	1−	1−	1−	6
	7	0+	0+	009	220	659	926	993	1−	1−	1−	1−	1−	1−	7
	8	0+	0+	002	109	488	846	978	999	1−	1−	1−	1−	1−	8
	9	0+	0+	0+	047	323	726	946	996	1−	1−	1−	1−	1−	9
	10	0+	0+	0+	017	189	575	885	987	1−	1−	1−	1−	1−	10
	11	0+	0+	0+	006	098	414	788	966	998	1−	1−	1−	1−	11
	12	0+	0+	0+	002	044	268	655	922	994	1−	1−	1−	1−	12
	13	0+	0+	0+	0+	017	154	500	846	983	1−	1−	1−	1−	13
	14	0+	0+	0+	0+	006	078	345	732	956	998	1−	1−	1−	14
	15	0+	0+	0+	0+	002	034	212	586	902	994	1−	1−	1−	15
	16	0+	0+	0+	0+	0+	013	115	425	811	983	1−	1−	1−	16
	17	0+	0+	0+	0+	0+	004	054	274	677	953	1−	1−	1−	17
	18	0+	0+	0+	0+	0+	001	022	154	512	891	998	1−	1−	18
	19	0+	0+	0+	0+	0+	0+	007	074	341	780	991	1−	1−	19
	20	0+	0+	0+	0+	0+	0+	002	029	193	617	967	999	1−	20
	21	0+	0+	0+	0+	0+	0+	0+	009	090	421	902	993	1−	21
	22	0+	0+	0+	0+	0+	0+	0+	002	033	234	764	966	1−	22
	23	0+	0+	0+	0+	0+	0+	0+	0+	009	098	537	873	998	23
	24	0+	0+	0+	0+	0+	0+	0+	0+	002	027	271	642	974	24
	25	0+	0+	0+	0+	0+	0+	0+	0+	0+	004	072	277	778	25

Areas bajo la distribución normal estandarizada

Ejemplo

$\Pr(0 \leq z \leq 1.96) = 0.4750$

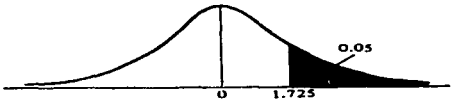
$\Pr(z \geq 1.96) = 0.5 - 0.4750 = 0.025$



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Puntos porcentuales de la distribución *t*

Ejemplo
Pr $(t > 2.086) = 0.025$
Pr $(t > 1.725) = 0.05$ para df $= 20$
Pr $(|t| > 1.725) = 0.10$



g de l \ Pr	0.25 0.50	0.10 0.20	0.05 0.10	0.025 0.05	0.01 0.02	0.005 0.010	0.001 0.002
1	1.000	3.078	6.314	12.706	31.821	63.657	318.31
2	0.816	1.886	2.920	4.303	6.965	9.925	22.327
3	0.765	1.638	2.353	3.182	4.541	5.841	10.214
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	0.718	1.440	1.943	2.447	3.143	3.707	5.208
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785
8	0.706	1.397	1.860	2.306	2.896	3.355	4.501
9	0.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.700	1.372	1.812	2.228	2.764	3.169	4.144
11	0.697	1.363	1.796	2.201	2.718	3.106	4.025
12	0.695	1.356	1.782	2.179	2.681	3.055	3.930
13	0.694	1.350	1.771	2.160	2.650	3.012	3.852
14	0.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	0.690	1.337	1.746	2.120	2.583	2.921	3.686
17	0.689	1.333	1.740	2.110	2.567	2.898	3.646
18	0.688	1.330	1.734	2.101	2.552	2.878	3.610
19	0.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	0.686	1.323	1.721	2.080	2.518	2.831	3.527
22	0.686	1.321	1.717	2.074	2.508	2.819	3.505
23	0.685	1.319	1.714	2.069	2.500	2.807	3.485
24	0.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.684	1.316	1.708	2.060	2.485	2.787	3.450
26	0.684	1.315	1.706	2.056	2.479	2.779	3.435
27	0.684	1.314	1.703	2.052	2.473	2.771	3.421
28	0.683	1.313	1.701	2.048	2.467	2.763	3.408
29	0.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	0.681	1.303	1.684	2.021	2.423	2.704	3.307
60	0.679	1.296	1.671	2.000	2.390	2.660	3.232
120	0.677	1.289	1.658	1.980	2.358	2.167	3.160
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.090

Nota: El valor más pequeño en el encabezamiento de cada columna es el área en una cola; el mayor valor es el área en ambas colas.

Fuente: Tomado de E. S. Pearson y H. O. Hartley, eds., *Biometrika Tables for Statisticians*, vol. 1, 3a. ed., tabla 12, Cambridge University Press, New York, 1966. Reproducido con permiso de los editores y depositarios de *Biometrika*.

Tabla G
Dígitos aleatorios

85967	73152	14511	85285	36009	95892	36962	67835	63314	50162
07483	51453	11649	86348	76431	81594	95848	36738	25014	15460
96283	01898	61414	83525	04231	13604	75339	11730	85423	60698
49174	12074	98551	37895	93547	24769	09404	76548	05393	96770
97366	39941	21225	93629	19574	71565	33413	56087	40875	13351
90474	41469	16812	81542	81652	45554	27931	93994	22375	00953
28599	64109	09497	76235	41383	31555	12639	00619	22909	29563
25254	16210	89717	65997	82667	74624	36348	44018	64732	93589
28785	02760	24359	99410	77319	73408	58993	61098	04393	48245
84725	86576	86944	93296	10081	82454	76810	52975	10324	15457
41059	66456	47679	66810	15941	84602	14493	65515	19251	41642
67434	41045	82830	47617	36932	46728	71183	36345	41404	81110
72766	68816	37643	19959	57550	49620	98480	25640	67257	18671
92079	46784	66125	94932	64451	29275	57669	66658	30818	58353
29187	40350	62533	73603	34075	16451	42885	03448	37390	96328
74220	17612	65522	80607	19184	64164	66962	82310	18163	63495
03786	02407	06098	92917	40434	60602	82175	04470	78754	90775
75085	55558	15520	27038	25471	76107	90832	10819	56797	33751
09161	33015	19155	11715	00551	24909	31894	37774	37953	78837
75707	48992	64998	87080	39333	00767	45637	12538	67439	94914
21333	48660	31288	00086	79889	75532	28704	62844	92337	99695
65626	50061	42539	14812	48895	11196	34335	60492	70650	51108
84380	07389	87891	76255	89604	41372	10837	66992	93183	56920
46479	32072	80083	63868	70930	89654	05359	47196	12452	38234
59847	97197	55147	76639	76971	55928	36441	95141	42333	67483
31416	11231	27904	57383	31852	69137	96667	14315	01007	31929
82066	83436	67914	21465	99605	83114	97885	74440	99622	87912
01850	42782	39202	18582	46214	99228	79541	78298	75404	63648
32315	89276	89582	87138	16165	15984	21466	63830	30475	74729
59388	42703	55198	80380	67067	97155	34160	85019	03527	78140
58089	27632	50987	91373	07736	20436	96130	73483	85332	24384
61705	57285	30392	23660	75841	21931	04295	00875	09114	32101
18914	98982	60199	99275	41967	35208	30357	76772	92656	62318
11965	94089	34803	48941	69709	16784	44642	89761	66864	62803
85251	48111	80936	81781	93248	67877	16498	31924	51315	79921
66121	96986	84844	93873	46352	92183	51152	85878	30490	15974
53972	96642	24199	58080	35450	03482	66953	49521	63719	57615
14509	16594	78883	43222	23093	58645	60257	89250	63266	90858
37700	07688	65533	72126	23611	93993	01848	03910	38552	17472
85466	59392	72722	15473	73295	49759	56157	60477	83284	56367

Tabla G (Continuación)

52969	55863	42312	67842	05673	91878	82738	36563	79540	61935
42744	68315	17514	02878	97291	74851	42725	57894	81434	62041
26140	13336	67726	61876	29971	99294	96664	52817	90039	53211
95589	56319	14563	24071	06916	59555	18195	32280	79357	04224
39113	13217	59999	49952	83021	47709	53105	19295	88318	41626
41392	17622	18994	98283	07249	52289	24209	91139	30715	06604
54684	53645	79246	70183	87731	19185	08541	33519	07223	97413
89442	61001	36658	57444	95388	36682	38052	46719	09428	94012
36751	16778	54888	15357	68003	43564	90976	58904	40512	07725
98159	02564	21416	74944	53049	88749	02865	25772	89853	88714

APÉNDICE B (PROGRAMA DE ESTUDIO DE ESTADÍSTICA II)

TEMÁTICA DE ESTADÍSTICA Y PROBABILIDAD II

PRIMERA UNIDAD. DISTRIBUCIONES MUESTRALES.

1) Población y muestra.
2) Selección de muestras pequeñas de poblaciones finitas.
3) Parámetros y estadísticos.
4) Los estadísticos como variables aleatorias (en particular la media muestral y la proporción muestral)
5) Distribución de la media muestral y de la proporción muestral (la media y la varianza de ambas distribuciones).
6) Interpretación del Teorema Central del Límite.
7) Relaciones entre parámetros y estadísticos.
8) Cálculo de probabilidades en la distribución de medias y de proporciones.

SEGUNDA UNIDAD. ESTIMACIÓN.

1) Introducción a la inferencia estadística.
2) Estimación puntual y por intervalos.
3) Características de los buenos estimadores (estimadores insesgados, eficientes y consistentes).
4) Estimación de intervalos de confianza para las medias poblacionales con varianza desconocida.
5) Estimación de intervalos de confianza para las proporciones poblacionales.

TERCERA UNIDAD. PRUEBAS DE HIPÓTESIS.

1) Modelo general de una prueba de hipótesis.
2) Hipótesis estadística (hipótesis nula e hipótesis alterna).
3) Significación de una prueba estadística (nivel de significancia)
4) Tipos de error en una prueba de hipótesis.
5) Prueba de hipótesis para una media poblacional.
6) Prueba de hipótesis para una proporción poblacional.

CUARTA UNIDAD. PREDICCIÓN ESTADÍSTICA.

1) Relación entre dos variables (diagramas de dispersión).
2) Correlación lineal simple.
3) Regresión lineal simple.

APÉNDICE C (RESPUESTAS A LOS EJERCICIOS)

CAPÍTULO 1.

Ejercicio 1. Suponiendo dados distinguibles, por ejemplo uno rojo y otro blanco (el espacio muestra constaría de 6 x 6 = 36 parejas ordenadas).

Sea X= Suma de las caras superiores. La distribución de probabilidad aparece compactada en la siguiente tabla :

X	2	3	4	5	6	7	8	9	10	11	12
$f_x(x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

E(X) = 7; V(X)= 54.83 – 7 = 5.83

Para F(x) tenemos la tabla siguiente :

X	2	3	4	5	6	7	8	9	10	11	12
F(x)	1/36	3/36	6/36	10/36	15/36	21/36	26/36	30/36	33/36	35/36	36/36

Según la tabla para f(x), P(X=10)= 3/36

Y según la tabla para F(x), P(X ≤ 8) = 26/36

Ejercicio 2. P(E) = 0.4, P(F) = 0.6, X = # de éxitos en 15. X≅B(15,0.4)

a) P(X ≥ 3) = 0.973
b) P(X ≥ 4) = 0.909
c) P(X ≥ 5) = 0.783
d) P(X < 3) = 0.027

Ejercicio 3. Siendo la variable aleatoria X= valores de colesterol total en los individuos de la población.

Como $X \approx N(200,(20)^2)$

a) P(180 ≤ X ≤ 200) = P(-1 ≤ Z ≤ 0) = 0.3413
b) P(X > 225) = P(Z > 1.25) =0.1056
c) P(X < 150) = P(Z < -2.5) = 0.0062
d) P(190 ≤ X ≤ 210) = P(-0.5 ≤ Z ≤ 0.5) = 0.383

Ejercicio 4. Sea E = Tener visión defectuosa
F = No tener visión defectuosa

P(E) = 0.3, P(F) = 0.7

Sea la variable aleatoria X = # de éxitos en 20 ensayos
X≅B(20,0.3). Como np=20(0.3)=6 y nq=14 son ambos mayores que 5, podemos usar la normal
X≈N(np, npq); X ≈ N(6, 4.2)
P(X ≥ 10) = P(10 - ½ ≤ $\mathcal{X}$ ≤ 20 + ½) = P(1.707≤ $\mathcal{Z}$ ≤ 7.073) = 0.0384

CAPÍTULO 2.

Ejercicio 2.2.1 .- Es aleatorio.

Ejercicios propuestos de la sección 2.4.1

1) X≅N(25,25)
   a) 0.4207
   b) 0.2347
   c) 0.5793
   d) 0.7257

2) X~N(5.4, $(2.8)^2$)
   a) 0.4168
   b) 0.2747
   c) 0.1415

3) X≅N(100, $(20)^2$), $\overline{\mathcal{X}}$ ≅N(100,$(20)^2$ /n), $\overline{\mathcal{X}}$ ≅N(100, $(20/4)^2$)
   a) 0.5
   b) 0.7333
   c) 0.9772

4)X≅?(50, $(16)^2$), $\overline{\mathcal{X}}$ ≈N(50, $(16)^2$/n).
   a) 0.9876
   b) 0.0668
   c) 0.0668
   d) 0.6902

Ejercicios propuestos de la sección 2.4.2

1) 0.017
2) 0.1131
3) 0.0021
4) a) 0.1539
   b)0.3409
   c) 0.5949

MISCELÁNEA DE EJERCICIOS DEL CAPÍTULO 2.

1) Emplear los archivos de registro y números al azar, o bien seleccionar 20 grupos de las horas semanales al azar, seleccionando 5 estudiantes también al azar, de cada grupo.
2) Los individuos que contestan son generalmente de buenos sentimientos o difíciles de complacer. La mayor parte de las personas no se molestan en responder; por lo tanto, la muestra tiene elementos de éstas dos opiniones extremas.
3) Sería necesario descubrir si los 600 que no contestaron tendrían las mismas opiniones que los que sí lo hicieron, antes de poderse fiar de los datos obtenidos de éstos 400.
4) Las casas difieren respecto al número de adultos ahí alojados; por lo tanto, las casas con un número crecido de adultos no estarían representadas adecuadamente.
5) A) 0.02, b) 0.84, c) 0.95, d) 0.00, e) 1
6) –
7) 0.10
8) P(Z > 3.0) = 0.001; parece ser más pesado éste grupo, o los 100 primeros estudiantes registrados no serían típicos (usuales)
9) Suponiendo que se trata de un año no bisiesto, con febrero de 28 días :
   1/6, 1/24, 2/11, 3/1, 3/19, 4/6, 4/24, 5/12, 5/30, 6/17, 7/5, 7/23, 8/10, 8/28, 9/15, 10/3, 10/21, 11/8, 11/26, 12/14. (1/6 = 6 de enero)
10) No. Si trabajan ambos cónyuges, nadie estará en casa entre el mediodía y las 5 de la tarde, y entonces algunos de los principales usuarios de la atención diurna quedarán excluídos de la encuesta.
11) ---
12) Error por muestreo
13) En general, la sobreestimación de la media no es ni mejor ni peor que la subestimación. En éste caso, la subestimación (30 ctvs) se acerca más a la verdadera media (31.4 ctvs) que la sobreestimación (35 ctvs).
14) Es una muestra de la distribución de muestreo de la media de las muestras de tamaño 50, extraídas de la población.

15) a) 0.8116
b)0.8703
16) Por lo menos 355
17) a) 0.2389
b)0.1562 . Ha disminuído en 0.0827
18) a) 0.4681 b) 0.4168
19) P(Z < 1.17) = 0.8790 > 0.85 . No se ordenará la compostura general.
20) a) 120 bu; b) 1.549 bu; c) 0.0071; d) 0.8753
21) Por lo menos 36
22) a) 1.490; b)0.0428
23) 0.2389
24) 0.9120
25) 0.7498

CAPÍTULO 3. ESTIMACIÓN.

Ejercicios de la sección 3.2

1) $\overline{x}$ = 14.278 mil personas, s^2 = 21.119 (miles de personas)
2) 0.46

Ejercicios de la sección 3.3.1

1) a) (4.54, 7.42); b) (4.265, 7.695); c) (3.727, 8.233)
2) a) (7.63, 8.87); b) (7.51, 8.99); c) (7.28, 9.22)
3) a) (122.533, 127.4675); b) (122.06, 127.94)

Ejercicios de la sección 3.3.2

1) 90 % : (0.824, 0.916); 95 %: (0.815, 0.925); 99 %: (0.798, 0.942)
2) (0.14, 0.34); (0.12, 0.36); (0.09, 0.39)

Ejercicios de la sección 3.3.3

1) (5.75, 8.25); (5.45, 8.55); (4.45, 9.55)
2)90%: (2243.35, 2756.65); 95 %: (2190.4, 2809.6); 99 %: (2080.45, 2919.55)

MISCELÁNEA DE EJERCICIOS DEL CAPÍTULO 3.

1) $\sigma_{\bar{p}} = 0.016$
2) $\sigma_{\bar{X}} = 3/10$
3) $\sigma_{\bar{X}}$ se estima mediante 0.12. Los límites de confianza son $\bar{X} \pm Z_{(1-\alpha/2)}(0.12)$
4) El error máximo es 4.238 aproximadamente con el 95% de confianza.
5) 32 ± 3.544
6) (2.69, 3.31)
7) a) (34.966, 36.434); b) (34.661, 36.739)
8) $t_{.975} = 2.074$
9) (29.125, 37.875)
10) (0.61, 0.71)
11) (133.37, 166.63)
12) (8.22, 9.20)
13) $150.2 \leq \mu \leq 169.8$
14) (149.21, 150.78)
15) (0.362, 0.483)
16) (0.1216, 0.2784)
17) (0.798, 0.842)
18) n = 400
19) 12549
20) n = 35
21) (40.11, 34.69)
22) (51.83, 66.17)
23) (8.87, 17.13)

CAPITULO 4

En estas respuestas, pondremos un subíndice c a la estadística de prueba calculada y un superíndice derecho * a los límites de confianza.

Ejercicios de la sección 4.2

1) $H_0 : \mu \geq 41.95$ vs $H_a : \mu < 41.95$ ($\alpha = 0.02$)
$Z_c = -0.56$ y $Z^* = -2.05$. Se "acepta" H_0.

2) $H_0 : \mu = 14.35$ vs $H_a : \mu \neq 14.35$ ($\alpha = 0.05$)
$Z_c = -1.452$ y $Z^* = \pm 1.96$. Se acepta H_0.

3) a) $H_0 : \mu = 84$ vs $H_a : \mu \neq 84$
b) $\alpha = 0.01$. $Z_c = -2.165$ y $Z^* = \pm 2.575$. Se acepta H_0.

4) $H_0 : \mu \leq 28$ vs $H_a : \mu > 28$ ($\alpha = 0.05$)
$Z_c = 1.15$ y $Z^* = 1.645$. Se acepta H_0.

5) $H_0 : \mu \geq 0.57$ vs $H_a : \mu < 0.57$ ($\alpha = 0.01$)
$Z_c = -9.3$ y $Z^* = -2.33$. Se rechaza H_0.

6)-------

Ejercicios de la sección 4.3

1) $H_0 : p \geq 0.19$ vs $H_a : p < 0.19$ ($\alpha = 0.04$)
$Z_c = -1.1455$ y $Z^* = -1.75$. Se acepta H_0.

2) $H_0 : p = 0.35$ vs $H_a : p \neq 0.35$ ($\alpha = 0.02$)
$Z_c = 2.353$ y $Z^* = \pm 2.33$. Se rechaza H_0.

3) a) $H_0 : p \geq 0.151$ vs $H_a : p < 0.151$ ($\alpha = 0.05$)
$Z_c = -2.135$ y $Z^* = -1.645$. Se rechaza H_0 y se concluye que la transmisión ha disminuído
b) $Z_c = -2.135$ y $Z^* = -2.055$ y también se rechaza H_0.
c) No necesariamente. No se tiene información sobre efectos secundarios.

4) $H_0 : p \leq 0.5$ vs $H_a : p > 0.5$ ($\alpha = 0.01$)
$Z_c = 2.037$ y $Z^* = 2.33$. Se acepta H_0 . La tendencia no es positiva al nivel del 0.01

5) $H_0 : p = 0.15$ vs $H_a : p \neq 0.15$ ($\alpha = 0.02$)
$Z_c = 1.0123$ y $Z^* = \pm 2.33$. Se acepta H_0.

MISCELÁNEA DE EJERCICIOS DEL CAPÍTULO 4.

1) i) Que el contenido es poco más o poco menos de 200 gramos (que en promedio, los frascos contienen 200 gramos)
ii) "El contenido ha aumentado"

iii) $\bar{X} > 200$
iv) " El contenido no ha aumentado"
v) $\bar{X} \leq 200$
vi) "El contenido de los frascos tal vez sea menor de 200 gramos"
vii) $\bar{X} < 200$
viii) " Tal vez la media del contenido de los frascos es mayor que 200 gramos"
ix) $\bar{X} > 200$

2) i) $H_0 : \mu \leq 200$ vs $H_a : \mu > 200$

ii) $$Z = \frac{\bar{X} - 200}{0.75} \cong N(0,1)$$

iii) Podemos usar cualquier nivel de significancia. Sea $\alpha = 0.05$. Como los valores que llevarán a rechazar la hipótesis nula son los muy altos, esta es una prueba de cola superior o cola derecha. El valor crítico es 1.645.
iv) Se rechaza H_0 si $Z_c > 1.645$
v) Como $Z_c = 3.92$, se rechaza H_0.
vii) Revisar la maquinaria. Reajustarla.

3) i) $H_0 : \mu \geq 200$ vs $H_a : \mu < 200$
ii)
$$Z = \frac{\bar{X} - 200}{1.06} \cong N(0,1)$$
iii) $\alpha = 0.10$. El valor crítico es -1.28.
iv) Rechazamos H_0 si $Z_c < -1.28$
v) Como $Z_c = -0.38$, aceptamos H_0. Los envases de café contienen al menos 200 gramos.
vi) Ni se les clausura ni se les multa a los de la empresa cafetera. Nada se ejerce contra ellos.

4) i) $H_0 : \mu \leq 200$ vs $H_a : \mu > 200$
ii)
$$Z = \frac{\bar{X} - 200}{0.67} \cong N(0,1)$$
iii) $\alpha = 0.01$. El valor crítico es 2.33
iv) Rechazamos H_0 si $Z_c > 2.33$
v) Como $Z_c = -5.37$, aceptamos H_0.
vi) Que se siga envasando en las mismas condiciones. La empresa no se ve afectada.

5) $H_0 : \mu \geq 95$ vs $H_a : \mu < 95$ ($\alpha = 0.01$)
$Z_c = -3.45$ y $Z^* = -2.33$. Rechazamos H_0. El fisicoterapeuta debe concluír que la fuerza media para la población es menor que 95.
6) $H_0 : \mu \leq 12$ vs $H_a : \mu > 12$ ($\alpha = 0.05$)
$Z_c = 2.05$ y $Z^* = 1.645$. Se rechaza H_0. Sí puede concluírse que el tiempo medio de retraso verdadero es mayor que 12 minutos.
7) Se rechaza H_0, $Z_c = 3.5$
8) $H_0 : \mu \geq 6.5$ vs $H_a : \mu < 6.5$ ($\alpha = 0.05$)
$Z_c = -0.594$ y $Z^* = -1.645$. Aceptamos H_0.
9) No, $Z_c = 2.0$
10) $H_0 : \mu \geq 130$ vs $H_a : \mu < 130$ ($\alpha = 0.05$)
$Z_c = -1.67$ y $Z^* = -1.645$. Se rechaza H_0.
11) Se rechaza H_0, $t_c = -4.5$
12) $H_0 : \mu \leq 2{,}250$ vs $H_a : \mu > 2{,}250$ ($\alpha = 0.05$)
$t_c = 1.67$ y $t^* = 1.725$. Aceptamos H_0.
13) No se rechaza H_0, $t_c = -1$.
14) $H_0 : \mu \geq 37.5$ vs $H_a : \mu < 37.5$ ($\alpha = 0.05$)
$t_c = -1.13$ y $t^* = -1.725$. Aceptamos H_0.
15) $H_0 : \mu = 15$ vs $H_a : \mu < 15$ ($\alpha = 0.10$)
$Z_c = -1.6$ y $Z^* = -1.285$. Se rechaza H_0.
16) $H_0 : \mu = 72$ vs $H_a : \mu > 72$ ($\alpha = 0.05$)
$Z_c = 3.5$ y $Z^* = 1.645$. Se rechaza H_0.
17) $H_0 = 100$ vs $H_a : \mu > 100$ ($\alpha = 0.05$)
$Z_c = 3.3$ y $Z^* = 1.645$. Se rechaza H_0.
18) $H_0 : \mu = 1100$ vs $H_a : \mu < 1100$ ($\alpha = 0.05$)
$Z_c = -1.9$ y $Z^* = -1.645$. Se rechaza H_0. Sí hay evidencia suficiente que indica disminución en el promedio de la producción diaria.
19) $H_0 : \mu = 23.2\ \%$ vs $H_a : \mu < 23.2\%$ ($\alpha = 0.10$)
$t_c = -2.64$ y $t^* = -1.833$. Se rechaza H_0. Los datos indican que la cantidad promedio de cobre es menor.
20) $H_0 : \mu = 35{,}000$ vs $H_a : \mu > 35{,}000$ ($\alpha = 0.05$)
$t_c = 1.395$ y $t^* = 2.132$. Se acepta H_0. La evidencia estadística no contradice la afirmación del contratista.
21) $t_c = -1.956$ y $t^* = \pm 2.201$. Se acepta H_0. Los salarios promedio no difieren de los salarios anuales pagados por las compañías competidoras.
22) $H_0 : p = 0.95$ vs $H_a : p < 0.95$ ($\alpha = 0.05$)
$Z_c = -3.12$ y $Z^* = -1.645$. Se rechaza H_0.

23) H_0 : p = 0.9 vs H_a : p > 0.9 (α = 0.05)
Z_c = 1.67 y Z* = 1.645 . Se rechaza H_0.

24) $H_0 : \mu = 100$ vs $H_a : \mu < 100$ (α =0.01)
Z_c = -2.35 y Z*= -2.579. No se rechaza H_0.

CAPITULO 5

Ejercicio A) : Un diagrama de dispersión sugiere una relación lineal directa (que a valores altos de "Y" corresponden valores altos de "x"). Desde luego, el uso de pañuelos faciales de papel no provoca resfriados. Además seamos cuidadosos en nuestra pretensión de que las variables están relacionadas.

Ejercicios de la sección 5.2

1) r = 0.93
2) r = 0.480
3) La fórmula muestra que el intercambio de "x" y "y" no afecta el valor de r.
4) Es subjetivo pues no hay datos:
   a) r tal vez ande por 0.9
   b) r tal vez ande por 0.6
   c) r tal vez ande por –0.7
5) a) 0.4, b) –0.6, c) –0.5
6) Que entre los accidentados en automóvil hay una regular dependencia lineal. Mientras más edad tenga el conductor menor número de accidentes tendrá.

Ejercicio B): y_c (50) = 174.69 + 2.25 (50) = 287.19

MISCELÁNEA DE EJERCICIOS DEL CAPÍTULO 5.

1) a) i) No existe correlación.
   ii) Parece que sí existe correlación y es negativa.
   iii) Parece que sí existe correlación y es positiva.

   b) i) r= 0
   ii) r < 0
   iii) r > 0

2)

c) r= 0.473
d) y = 43.76 + (0.42)x
e) Parece que sí.
f)Ligeramente, en relación lineal.
g)94.16
h)79.04
i)Como x_M = 100 y x_m = 66, la extrapolación para 120 kg no es confiable. La interpolación para x= 84 kg sí es confiable. ¿Por qué? Por lo mencionado en la figura al final de la sección 5.3.1

BIBLIOGRAFÍA

1) **Wayne W. Daniel. " Bioestadística. Base para el análisis de las ciencias de la salud". Editorial Limusa. Segunda reimpresión : 1980.**
2) **G. Hoel Paul. "Estadística elemental". C.E.C.S.A. Reimpresión: Noviembre de 1974.**
3) **Damodar gujarati. "Econometría Básica" . Mc Graw Hill. Abril de 1984**
4) **Stephen S. Willoughby. "Probabilidad y estadística". Publicaciones cultural, S.A. Novena reimpresión, 1981.**
5) **William C. Guenther. "Introducción a la inferencia estadística". Mc Graw Hill. Julio de 1977**