Universidad Nacional Autónoma de México

Facultad de Ciencias

Uso de `R` en estadística no paramétrica y análisis de regresión lineal simple

REPORTE DE ACTIVIDAD DOCENTE

PARA OBTENER EL TÍTULO DE:
ACTUARIO

PRESENTA:
MISRAIM GUTIÉRREZ MESTAS

ASESOR:
DR. LUIS ANTONIO RINCÓN SOLÍS

2016

Hoja de Datos del Jurado

1. **Datos del Alumno**
Nombre: Gutiérrez Mestas Misraim
Teléfono: 55 36 63 84 63
Universidad Nacional Autónoma de México
Facultad de Ciencias
Carrera: Actuaría
Número de cuenta: 30307758-9

2. **Datos del primer sinodal**
Dra. Ruth Selene Fuentes García

3. **Datos del segundo sinodal**
Dr. Ricardo Ramírez Aldana

4. **Datos del tercer sinodal**
Dr. Luis Antonio Rincón Solís
Asesor

5. **Datos del cuarto sinodal**
Mat. Margarita Elvira Chávez Cano

6. **Datos del quinto sinodal**
Dra. Lizbeth Naranjo Albarrán

7. **Datos del trabajo escrito**
Título: Uso de `R` en estadística no paramétrica
y análisis de regresión lineal simple.
No. de páginas: 230
Año: 2016

A mis padres por haberme brindado todo su apoyo y forjarme como la persona que soy; gracias a ellos he podido realizar parte de mis logros incluyendo este. Me orientaron con reglas y con libertad, al final del día siempre estuvieron conmigo y me motivaron continuamente para alcanzar mis anhelos.

Contenido

Prólogo

El presente texto contiene material básico sobre temas de estadística no paramétrica, análisis de regresión lineal simple, una breve introducción al uso del paquete estadístico R, incluyendo su manejo en estas áreas de la estadística a nivel licenciatura y al final se presenta un breve apartado de pruebas de hipótesis. Es un trabajo realizado para obtener el título de Actuario y dicho material, es el producto de la recopilación de notas de clase que he utilizado para impartir el curso de Estadística II en la Facultad de Ciencias de la UNAM como ayudante de profesor. Está dirigido a estudiantes de las carreras de Actuaría, Matemáticas, Ciencias de la Computación y otras carreras a fines, para contribuir, apoyar y complementar el trabajo docente de los profesores. Una breve revisión al índice temático le dará al lector una idea de los temas expuestos y el orden en el que se presentan. Los capítulos 1 y 2 corresponden al cuerpo de este trabajo, en ellos se estudia estadística no paramétrica y análisis de regresión lineal, respectivamente. En el apéndice A se da una breve introducción a R mientras que en el apéndice B se estudian conceptos de pruebas de hipótesis desde un punto de vista paramétrico y muchos de los conceptos que se presentan aquí se usarán a lo largo del texto. Finalmente, en el último apéndice se presentan tablas estadísticas útiles para un mejor entendimiento de los temas expuestos. El texto fue escrito en el editor de textos científicos de alta calidad LaTeX junto con el paquete *Sweave* que permite la conección de LaTeX y R generando texto y código automáticamente. Un archivo fuente simple contiene el texto de documentación y el código R, los cuales son entrelazados dentro de un documento final que contiene el texto de documentación junto con el código R y/o la salida del código (texto, gráficos). Además, las ilustraciones que se presentan en este trabajo fueron elaboradas con la ayuda del paquete *Pstricks*. El material que se presenta se basa completamente de las fuentes bibliográficas que se muestran al final del texto.

Agradezco todos los comentarios, sugerencias, ayuda y correcciones que he recibido por parte de mi asesor Luis Rincón para una mejor presentación de este trabajo y sobre todo, por su gran apoyo académico. Doy gracias a los sinodales por todas sus aportaciones y retroalimentación que me brindaron y finalmente, muchas gracias a los profesores y colegas Jésica Rojano y Ricardo Ramírez por haberme impulsado en el ámbito académico y por todo su apoyo para la preparación de este trabajo.

Misraim Gutiérrez
Marzo del 2016
mizra@ciencias.unam.mx

Capítulo 1

Estadística no paramétrica

1.1. Introducción

En los problemas de pruebas de hipótesis se supone que las observaciones disponibles para el estadístico provienen de distribuciones cuya forma exacta es conocida, aún cuando los valores de algunos parámetros sean desconocidos. En otras palabras, se ha supuesto que las observaciones provienen de una cierta familia paramétrica de distribuciones y que se debe hacer una inferencia estadística acerca de los valores de los parámetros que definen dicha familia, normalmente la media μ, la varianza σ^2, o la proporción p.

En este capítulo, no se supondrá que las observaciones disponibles provienen de una familia paramétrica de distribuciones, en su lugar, se estudiarán inferencias que se pueden realizar sobre la distribución de donde provienen los datos, sin hacer suposiciones especiales acerca de la forma de esa distribución. Cuando se realizan inferencias que no se aplican a los valores que toman los parámetros se denominan *inferencias no paramétricas*. También existen otro tipo de inferencias que no necesariamente se basan en la distribución de la población, las cuales reciben el nombre de pruebas de distribución libre. Aunque ambos tipos de métodos inferenciales no son necesariamente idénticos, ambos forman parte de lo que se denominan *métodos no paramétricos*. Por ejemplo, se puede suponer que las observaciones constituyen una muestra aleatoria de una distribución continua, sin especificar la forma de esta distribución y entonces investigar la posibilidad de que esta distribución sea una distribución normal.

Escalas de medición

En este apartado se definen los tipos de escalas que se le asigna a una variable. Se entenderá por *escalación* al procedimiento que asigna una observación a una categoría. Este procedimiento utiliza diversas escalas. Hay cuatro escalas básicas las cuales son: nominal, ordinal, de intervalo y de razón. A las variables que se consideran con escala ordinal y nominal también se les conocen como

variables categóricas, mientras que a las variables que se consideran con escala de intervalo y razón se conocen como *variables cuantitativas*.

Variables categóricas

Escala nominal. La escala nominal sólo permite asignar un nombre o categoría a cada observación convirtiendola en la escala que brinda menor información. Se trata de agrupar objetos en clases o categorías, de modo que todos los que pertenezcan a la misma sean equivalentes respecto del atributo o propiedad en estudio, en otras palabras, una escala nominal es un esquema de etiquetado figurado. Por ejemplo, cuando en una encuesta se pregunta cuál es el estado civil de una persona, las posibles clases, etiquetas o categorías son "casado", "soltero", "divorciado" o "viudo" y no hay distinción alguna entre todas las personas que sean solteros. El hecho de que a veces, se le atribuyan números a cada categoría, puede ser una de las razones por las cuales se le conoce como *medidas nominales*.

Escala ordinal. La escala ordinal, además de las propiedades de la escala nominal, permite establecer un orden entre las categorías, o bien, cada categoría indica una posición. Por ejemplo, si en un estudio de mercado se desea conocer la calidad del servicio telefónico de cierta compañía, entonces la calidad del servicio puede ser medida como "excelente", "buena", "regular" o "mala". De esta forma, es común que a este tipo de escala se les asignen valores numéricos, por ejemplo, puede medirse la calidad del servicio telefónico seleccionando un número del 1 al 5, con la característica de que se selecciona 1 si el servicio es malo y 5 si el servicio es excelente.

Variables cuantitativas

Escala de intervalo. Una escala de intervalo corresponde aquellas variables que necesitan de un punto arbitrario que corresponde al cero. Por ejemplo, la temperatura, aunque hay diferencia en el cero asignado si se utilizan grados centígrados en vez de grados Fahnrenheit.

Escala de razón. En una escala de razón hay una medida que de manera natural se vuelve el cero. Por ejemplo, los salarios, peso, estatura, etc.

Los métodos no paramétricos son utilizados preferentemente cuando la distribución de los datos no es normal o el tamaño de las muestras es pequeño, ya que es difícil establecer la distribución de la población y la escala de medida de la variable puede ser de tipo nominal u ordinal.

1.2. Prueba para proporciones

Supongamos que un investigador desea saber si la tasa de desempleo va en aumento, si la tasa de pobreza está cambiando o si la tasa de algún grupo cívico está a favor de una política en particular. En muchos casos existe una proporción hipotética p de una población en estudio y una proporción específica p^* y se pretende llevar a cabo una comparación para saber si la proporción hipotética es igual, mayor o menor que la proporción específica p^*. Una prueba de proporciones puede ser útil para ayudar a responder este tipo de preguntas. En esta sección se presenta la *prueba de proporciones* o también conocida como *prueba binomial*, la cual utiliza técnicas no paramétricas para contrastar hipótesis relacionadas con la proporción de la población.

Datos. Los datos consisten en una muestra $X_1, X_2, \ldots, X_n$ de tamaño n y cada elemento de la muestra pertenece a una de dos distintas clases, a la *clase 1* ó a la *clase 2*. Se definen

T: El número de elementos de la *clase 1* en la muestra.
T^c: El número de elementos de la *clase 2* en la muestra,

donde $T^c = n - T$.

Supuestos.

a) Las n observaciones de la muestra son mutuamente independientes.

b) Cada observación tiene probabilidad p de pertenecer a la *clase 1*.

Esta situación se puede modelar con una distribución Bernoulli donde el éxito corresponde a que cada observación pertenece a la *clase 1* con probabilidad p y el fracaso corresponde a pertenecer a la *clase 2* con probabilidad $1 - p$.

Hipótesis a probar. A continuación se describen los contrastes de hipótesis para la prueba de proporciones.

Caso A (Dos colas)

Sea p^* un valor fijo conocido o la proporción que se especifica en la hipótesis nula. Nuestro interés es contrastar las hipótesis

$$H_0 : p = p^* \quad vs \quad H_1 : p \neq p^*.$$

Como la variable T representa el número de elementos de la *clase 1* en la muestra, entonces T sigue una distribución binomial con parámetros n y probabilidad p desconocida. Se desea encontrar una regla de decisión para rechazar la hipótesis nula H_0 a favor de la hipótesis alternativa H_1. De manera natural,

se rechaza H_0 cuando la distancia entre la proporción estimada de la muestra $\widehat{p}$ y la proporción poblacional p sea muy grande, es decir, cuando

$$|\widehat{p} - p| \geq t,$$

o bien,

$$\left|\frac{1}{n}T - p\right| \geq t,$$

para algún $t > 0$. Esta desigualdad se puede escribir como:

$$|T - np| \geq nt$$

entonces

$$np - T \geq nt \quad \text{ó} \quad T - np \geq nt$$

si y sólo si

$$T \leq n(p - t) \quad \text{ó} \quad T \geq n(p + t).$$

El tamaño de la región de rechazo es la máxima probabilidad de rechazar H_0 cuando p es igual a p^*, bajo la hipótesis nula, entonces

$$\begin{aligned}
\alpha &= \sup_{p=p^*} \mathbb{P}(\text{“Rechazar } H_0\text{”}|\text{“}H_0 \text{ es cierta”}) \\
&= \sup_{p=p^*} \mathbb{P}(T \leq n(p-t) \text{ ó } T \geq n(p+t)|p = p^*) \\
&= \mathbb{P}(T \leq n(p^* - t) \text{ ó } T \geq n(p^* + t)).
\end{aligned}$$

Por lo anterior, la distribución nula de T es binomial con parámetros n y probabilidad p^*. La región de rechazo de tamaño α corresponde a dos colas de la distribución de T, donde el tamaño de la cola inferior denotado por α_1 es cercano a $\alpha/2$ y el tamaño de la cola superior denotado por α_2 es cercano a $\alpha/2$, así, el verdadero valor de α será $\alpha_1 + \alpha_2$.
De esta manera, se quiere encontrar t_1 y t_2 tales que

$$\mathbb{P}(T \leq t_1) \approx \alpha/2$$

y

$$\mathbb{P}(T \leq t_2) \approx 1 - \alpha/2.$$

De esta manera tenemos la siguiente estadística de prueba y regla de decisión.

Estadística de prueba. La estadística de prueba propuesta es

$$T = \text{el número de elementos de la } \textit{clase 1} \text{ en la muestra,}$$

donde $T \sim binom(n, p^*)$.

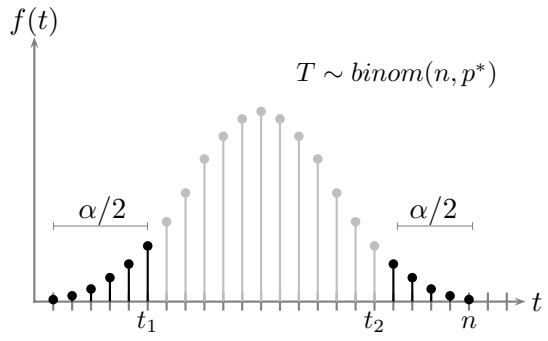


Figura 1.1: Región de rechazo para el contraste de dos colas.

Regla de decisión. Se rechaza la hipótesis nula H_0 al nivel de significancia α si $T \leq t_1$ ó $T > t_2$, donde t_1 es el cuantil que acumula aproximadamente $\alpha/2$ de probabilidad en la cola inferior y t_2 es el cuantil que acumula aproximadamente $\alpha/2$ de probabilidad en la cola superior de la distribución de T (véase la Figura 1.1).

$$\mathbb{P}(T \leq t_1) \approx \alpha/2,$$
$$\mathbb{P}(T > t_2) \approx \alpha/2.$$

En la tabla C.1 se muestran los valores de la función de densidad $f(t)$ para una variable T que se distribuye binomial cuando $n \leq 15$ y distintos valores de p^*. Para valores donde n es mayor a 20 se pueden calcular los cuantiles utilizando la aproximación normal de la siguiente manera. Si la variable $T \sim binom(n, p^*)$ entonces $\mathbb{E}(T) = np^*$ y $Var(T) = np^*(1-p^*)$. En consecuencia,

$$Z \approx \frac{T - np^*}{\sqrt{np^*(1-p^*)}} \tag{1.1}$$

entonces

$$T \approx np^* + Z\sqrt{np^*(1-p^*)}. \tag{1.2}$$

De la ecuación (1.2) el cuantil aproximado t_q esta dado por

$$t_q \approx np^* + z_q\sqrt{np^*(1-p^*)} \tag{1.3}$$

donde z_q es el cuantil de orden q de una variable aleatoria normal estándar. En particular, si $\alpha = 0.05$, $z_{\alpha/2} = -1.96$ y $z_{1-\alpha/2} = 1.96$, entonces t_1 y t_2 estan dados por

$$t_1 \approx np^* - 1.96\sqrt{np^*(1-p^*)}$$
$$t_2 \approx np^* + 1.96\sqrt{np^*(1-p^*)}.$$

Valor p. Si definimos t_{obs} como el valor observado de la estadística de prueba T, el valor p para el contraste de dos colas, es dos veces el mínimo entre la probabilidad de que T sea menor o igual al valor observado t_{obs} y la probabilidad de que T sea mayor o igual que t_{obs}, o bien

$$v_p = 2\,\text{mín}\{\mathbb{P}(T \leq t_{obs}), \mathbb{P}(T \geq t_{obs})\}$$

Para valores de $n > 20$ y usando p^*, las probabilidades $\mathbb{P}(T \leq t_{obs})$ y $\mathbb{P}(T \geq t_{obs})$ se calculan de la siguiente manera

$$\mathbb{P}(T \leq t_{obs}) \approx \mathbb{P}\left(Z \leq \frac{t_{obs} - np^* + 0.5}{\sqrt{np^*(1-p^*)}}\right) \tag{1.4}$$

$$\mathbb{P}(T \geq t_{obs}) \approx 1 - \mathbb{P}\left(Z \leq \frac{t_{obs} - np^* - 0.5}{\sqrt{np^*(1-p^*)}}\right) \tag{1.5}$$

En las expresiones anteriores se incorpora el valor 0.5 como una corrección de continuidad que mejora la aproximación normal a la binomial.

Corrección de continuidad

Dada una variable aleatoria discreta X, se cumple que $\mathbb{P}(X \leq x) = \mathbb{P}(X < x+1)$ y $\mathbb{P}(X \geq x) = \mathbb{P}(X > x-1)$. Cuando se busca hacer una aproximación normal, *la corrección de continuidad* establece que

a) $\mathbb{P}(X \leq x) = \mathbb{P}(X < x + 0.5)$,

b) $\mathbb{P}(X \geq x) = \mathbb{P}(X > x - 0.5)$,

c) $\mathbb{P}(x_1 \leq X \leq x_2) = \mathbb{P}(x_1 - 0.5 < X < x_2 + 0.5)$.

Ejemplo 1.1. *Sea* $X \sim binom(100, 1/3)$ *y se desea calcular* $\mathbb{P}(X < 45)$.

Calculando la esperanza, la varianza y la desviación estándar se tiene que $\mathbb{E}(X) \approx 3.3\bar{3}$, la $Var(X) \approx 22.2\bar{2}$ y $\sigma \approx 4.714$. Utilizando la aproximación normal tenemos

$$\begin{aligned}\mathbb{P}(X < 45) &= \mathbb{P}\left(\frac{X - 33.33}{4.714} < \frac{45 - 33.33}{4.714}\right) \\ &\approx \mathbb{P}(Z < 2.476)\end{aligned}$$

donde $Z \sim N(0,1)$. Calculando esta probabilidad se tiene

```
> pnorm(q=2.476,mean=0,sd=1)
```

```
[1] 0.9933568
```

mientras que la probabilidad $\mathbb{P}(X < 45) = \mathbb{P}(X \leq 44)$ original de la binomial es

```
> pbinom(q=44,size=100,prob=1/3)
```

```
[1] 0.9899949
```

Aplicando la corrección de continuidad para mejorar la aproximación vemos que

$$\begin{aligned}\mathbb{P}(X < 45) &= \mathbb{P}(X \leq 44) \\ &= \mathbb{P}(X < 44.5) \\ &\approx \mathbb{P}\left(Z < \frac{44.5 - 33.33}{4.714}\right) \\ &= \mathbb{P}(Z < 2.3695)\end{aligned}$$

donde nuevamente $Z \sim N(0,1)$. Calculando esta probabilidad se obtiene

```
> pnorm(q=2.3695,mean=0,sd=1)
```

```
[1] 0.9910939
```

Por lo tanto, comparando las probabilidades, podemos ver que la aproximación es mejor aplicando la corrección de continuidad de 0.5. ▪

Caso B (Cola inferior)

Nuestro interés es contrastar las hipótesis

$$H_0 : p \geq p^* \quad vs \quad H_1 : p < p^*.$$

Si se observan valores pequeños de la estadística de prueba T, esto indica que la proporción hipotética p es pequeña, lo que favorece a la hipótesis alternativa H_1. De manera natural, se rechaza H_0 a favor de H_1 si la proporción estimada de la muestra $\widehat{p}$ más un valor t es menor o igual que p^*, para algún $t > 0$. Lo anterior se puede escribir como:

$$\widehat{p} + t \leq p^*$$

entonces

$$\frac{1}{n}T \leq p^* - t$$

si y sólo si

$$T \leq n(p^* - t).$$

La región de rechazo de tamaño α es la máxima probabilidad de rechazar H_0 cuando p es igual a p^*, bajo la hipótesis nula, es decir

$$\begin{aligned}\alpha &= \sup_{p \geq p^*} \mathbb{P}(\text{“Rechazar } H_0\text{”}|\text{“}H_0 \text{ es cierta”}) \\ &= \sup_{p \geq p^*} \mathbb{P}(T \leq n(p^* - t)|p \geq p^*) \\ &= \mathbb{P}(T \leq n(p^* - t)).\end{aligned}$$

Por lo tanto, la región de rechazo consiste de todos los valores de T menores o iguales a un número t_3 el cual debe acumular aproximadamente α de probabilidad. De esta manera se tiene la siguiente regla de decisión.

Regla de decisión. Se rechaza la hipótesis nula H_0 al nivel de significancia α si $T \leq t_3$, donde t_3 es el cuantil que acumula aproximadamente α de probabilidad de la distribución de T (véase la Figura (1.2)).

$$\mathbb{P}(T \leq t_3) \approx \alpha.$$

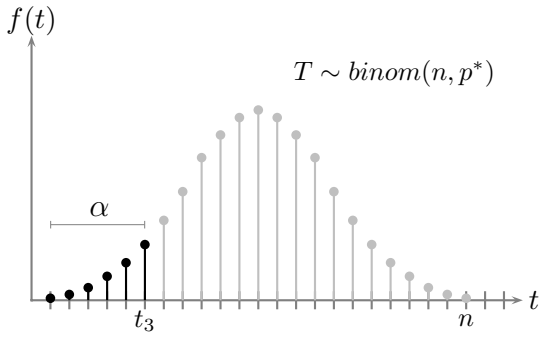


Figura 1.2: Región de rechazo de la prueba de cola inferior.

Si el tamaño de la muestra n es mayor que 20 se puede calcular el cuantil t_3 utilizando la aproximación normal de la ecuación (1.3), obteniendo

$$t_3 \approx np^* + z_\alpha \sqrt{np^*(1-p^*)},$$

donde z_α es el cuantil que acumula α de probabilidad de una distribución normal estándar.

Valor p. El valor p para la prueba de cola inferior es la probabilidad de que T sea menor o igual que el valor observado de la estadística t_{obs}

$$v_p = \mathbb{P}(T \leq t_{obs}).$$

Para valores grandes de $n > 20$ la probabilidad $\mathbb{P}(T \leq t_{obs})$ se puede calcular aproximadamente con la ecuación (1.4)

$$\mathbb{P}(T \leq t_{obs}) \approx \mathbb{P}\left(Z \leq \frac{t_{obs} - np^* + 0.5}{\sqrt{np^*(1-p^*)}}\right).$$

Caso C (Cola superior)

Nuestro interés es contrastar las hipótesis

$$H_0 : p \leq p^* \quad vs \quad H_1 : p > p^*.$$

Si se observan valores grandes de la estadística de prueba T, esto indica que la proporción hipotética p es grande, lo que favorece a la hipótesis alternativa H_1. Se rechaza H_0 a favor de H_1 si la proporción estimada $\widehat{p}$ menos t es mayor o igual que p^*, para algún $t > 0$. Lo anterior se puede escribir como:

$$\widehat{p} - t \geq p^*$$

entonces

$$\frac{1}{n}T - t \geq p^*$$

si y sólo si

$$T \geq n(p^* + t).$$

La región de rechazo de tamaño α es la máxima probabilidad cuando $p = p^*$, bajo la hipótesis nula, es decir,

$$\begin{aligned}\alpha &= \sup_{p \leq p^*} \mathbb{P}(\text{“Rechazar } H_0\text{”}|\text{“}H_0 \text{ es cierta”}) \\ &= \sup_{p \leq p^*} \mathbb{P}(T \geq n(p^* + t)|p \leq p^*) \\ &= \mathbb{P}(T \geq n(p^* + t)).\end{aligned}$$

Por lo tanto, la región de rechazo corresponde a todos los valores de T mayores a un número t_4 tal que acumule aproximadamente α de probabilidad. De esta forma, se tiene la siguiente regla de decisión.

Regla de decisión. Se rechaza la hipótesis nula H_0 al nivel de significancia α si $T > t_4$, donde t_4 es el cuantil de cola superior que acumula aproximadamente α de probabilidad de la distribución de T (véase la Figura (1.3)).

$$\mathbb{P}(T \leq t_4) \approx 1 - \alpha \quad \text{ó} \quad \mathbb{P}(T > t_4) \approx \alpha.$$

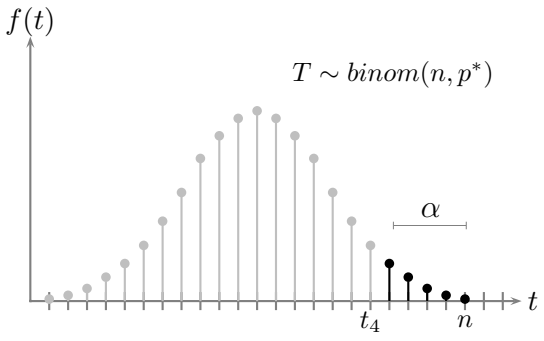


Figura 1.3: Región de rechazo de la prueba de cola superior.

Si el tamaño de la muestra n es mayor que 20 se puede calcular el cuantil t_4 utilizando la aproximación normal de la ecuación (1.3), obteniendo

$$t_4 \approx np^* + z_{1-\alpha}\sqrt{np^*(1-p^*)},$$

donde $z_{1-\alpha}$ es el cuantil que acumula $1-\alpha$ de probabilidad de una distribución normal estándar.

Valor p. El valor p para la prueba de cola superior es la probabilidad de que T sea mayor o igual que el valor observado de la estadística t_{obs}.

$$v_p = \mathbb{P}(T \geq t_{obs}).$$

Para valores de $n > 20$ la probabilidad $\mathbb{P}(T \geq t_{obs})$ se puede calcular aproximadamente con la ecuación (1.5)

$$\mathbb{P}(T \geq t_{obs}) \approx 1 - \mathbb{P}\left(Z \leq \frac{t_{obs} - np^* - 0.5}{\sqrt{np^*(1-p^*)}}\right).$$

Ejemplo 1.2. *En cada uno de los siguientes incisos, una observación aleatoria T se toma de una distribución binom(n, p) y se probarán las hipótesis dadas con el nivel de significancia establecido.*

a) Cuando $T = 6$ y $n = 8$. Se contrastan las hipótesis

$$H_0 : p \leq 0.45 \quad vs \quad H_1 : p > 0.45$$

con un nivel de significancia de $\alpha = 5\,\%$.

b) Cuando $T = 1$ y $n = 10$. Se contrastan las hipótesis

$$H_0 : p \geq 0.45 \quad vs \quad H_1 : p < 0.45$$

con un nivel de significancia de $\alpha = 5\,\%$.

Solución: Para el inciso a) se contrastan las hipótesis

$$H_0 : p \leq 0.45 \quad vs \quad H_1 : p > 0.45$$

y utilizando la prueba de cola superior (Caso C), la estadística de prueba es $T = 6$ con $n = 8$, $p^* = 0.45$ y $\alpha = 0.05$, tenemos que $T \sim binom(8, 0.45)$. Queremos buscar el valor del cuantil t_4 de tal manera que $\mathbb{P}(T > t_4) \approx 0.05$, o en su defecto $\mathbb{P}(T \leq t_4) \approx 0.95$. Calculando las probabilidades

```
# Sucesión de valores del 0 al 8
> t <- 0:8
# P(T > t) para valores de t
> pbinom(q=t, size=8, prob=0.45, lower.tail=F)
```

```
[1] 0.991627 0.936819 0.779870 0.523044 0.260381 0.088456 0.018123
[8] 0.001682 0.000000
```

vemos que los posibles valores de $\mathbb{P}(T > t_4)$ que se aproximan a $\alpha = 0.05$ son:

$$\mathbb{P}(T > 5) = 0.08845 \quad \text{ó} \quad \mathbb{P}(T > 6) = \mathbf{0.018123}.$$

Se toma la $\mathbb{P}(T > 6) = 0.018123$ porque no excede el nivel de significancia α, en consecuencia, como sabemos que la región de rechazo está dada por

$T > t_4$ y esta condición no se cumple ya que $6 \not> 6$, entonces no se rechaza la hipótesis nula H_0. Si tomamos la probabilidad $\mathbb{P}(T > 5) = 0.0885$, la región de rechazo se vuelve más grande que el nivel de significancia 0.05 lo que implica que se incrementa la probabilidad de cometer un error. El valor p, o bien la probabilidad $\mathbb{P}(T \geq t_{obs})$ es

```
# Valor p ℙ(T ≥ 6)
> pbinom(q=5, size=8, prob=0.45, lower.tail=F)
```

```
[1] 0.08846
```

y como 0.08846 es mayor que 0.05, entonces no se rechaza la hipótesis nula H_0. Por lo tanto, no existe evidencia suficiente para contradecir que $p \leq 0.45$.

Análogamente, para el inciso b) se tiene las hipótesis a contrastar

$$H_0 : p \geq 0.45 \quad vs \quad H_1 : p < 0.45$$

por lo que utilizaremos el contraste de cola inferior (Caso B). La estadística de prueba es $T = 1$ con $n = 10$, $p^* = 0.45$ y $\alpha = 0.05$ entonces $T \sim binom(10, 0.45)$. Se quiere encontrar el valor t_3 tal que $\mathbb{P}(T \leq t) \approx 0.05$. Calculando las probabilidades

```
# Sucesión de valores del 0 al 10
> t <- 0:10
# ℙ(T ≤ t) para valores de t
> pbinom(q=t, size=10, prob=0.45)
```

```
 [1] 0.002533 0.023257 0.099560 0.266038 0.504405 0.738437 0.898005
 [8] 0.972608 0.995498 0.999659 1.000000
```

vemos que los posibles valores de $\mathbb{P}(T \leq t_3)$ que se aproximan a $\alpha = 0.05$ son:

$$\mathbb{P}(T \leq 1) = \mathbf{0.023257} \quad \text{ó} \quad \mathbb{P}(T \leq 2) = 0.099560$$

entonces, se toma $t_3 = 1$, ya que este cuantil acumula una probabilidad que no excede al nivel de significancia 0.05 y sabemos que se rechaza la hipótesis nula H_0 si $T \leq t_4$ y tenemos que $1 \leq 1$, por lo que rechazamos la hipótesis nula H_0. El valor p dado por $\mathbb{P}(T \leq t_{obs})$ es

```
# Valor p ℙ(T ≤ 1)
> pbinom(q=1, size=10, prob=0.45)
```

```
[1] 0.02326
```

como el valor p es menor que el nivel de significancia 0.05, entonces se rechaza H_0. Por lo tanto, se dice que la proporción p es menor que 0.45. ▪

En `R` existe el comando `binom.test()` que realiza la prueba exacta binomial la cual sirve para realizar este tipo de contrastes. Este comando cuenta con los siguientes argumentos:

```
binom.test(e, n, p, alternative=, conf.level=)
```

donde `e` corresponde al número de éxitos de `n` ensayos Bernoulli con probabilidad de éxito `p`. El argumento `alternative=` puede tomar los siguientes valores:

a) `"two.sided"` realiza un contraste de dos colas.

b) `"greater"` realiza un contraste de cola superior.

c) `"less"` realiza un contraste de cola inferior.

Por último, el argumento `conf.level=` especifica el nivel de confianza del intervalo para la proporción p, si no se especifica ningún valor, la salida muestra por defecto el intervalo al 95 % de confianza.

De esta manera, utilizando el comando `binom.test()` podemos resolver el inciso a) de la siguiente manera:

```
# Prueba binomial exacta
> binom.test(x=6, n=8, p=0.45, alternative="greater")

        Exact binomial test

data:  6 and 8
number of successes = 6, number of trials = 8, p-value = 0.08846
alternative hypothesis: true probability of success is greater than 0.45
95 percent confidence interval:
 0.4003 1.0000
sample estimates:
probability of success
                 0.75
```

Igualmente y con los argumentos adecuados, se puede resolver el inciso b) como sigue:

```
# Prueba binomial exacta
> binom.test(x=1, n=10, p=0.45, alternative="less")

        Exact binomial test

data:  1 and 10
number of successes = 1, number of trials = 10, p-value = 0.02326
alternative hypothesis: true probability of success is less than 0.45
95 percent confidence interval:
 0.0000 0.3942
sample estimates:
probability of success
                   0.1
```

Observación 1.1. *Si el número de éxitos es grande, el comando* `prop.test()`*, el cual cuenta con los mismos argumentos que el comando* `binom.test()`*, realiza un contraste de hipótesis utilizando una distribución* $\chi^2(1)$ *con una corrección de continuidad de 0.5 para aproximar a la binomial. De esta manera la estadística de prueba está dada por*

$$\chi^2 = \left[\frac{t_{obs} - np^* \pm 0.5}{\sqrt{np^*(1-p^*)}}\right]^2.$$

La corrección de continuidad del 0.5 la acompaña el signo "+" si se lleva a cabo un contraste de cola inferior y el signo "−" si se lleva a cabo un contraste de cola superior.

■

1.3. Prueba de signos

La prueba de signos merece una consideración especial debido a su versatilidad, su gran utilidad y simplicidad. Esta prueba es una prueba de proporciones cuando el valor específico $p^* = 1/2$ y para los contrastes de dos colas, cola inferior y cola superior la máxima probabilidad para rechazar la hipótesis nula H_0 se da cuando $p = 1/2$. Frecuentemente la prueba de signos también es apropiada para analizar datos de un vector aleatorio (X, Y) y ver si alguna de sus entradas tiene valores más grandes que la otra. De esta manera, si una variable tiende a tener valores mayores que la otra, se puede utilizar la prueba de signos para determinar si las medias de estas variables son diferentes, por lo cual, se podría plantear la hipótesis nula $H_0 : \mu_X = \mu_Y$ o equivalentemente $H_0 : \mu_X - \mu_Y = 0$. Cada par de datos en la muestra se reemplazará por un signo "+" cuando $X_i < Y_i$, por un signo "−" cuando $X_i > Y_i$ y se omitirán las parejas cuando $X_i = Y_i$. Si cada signo "+" se identifica como *éxito* y cada signo "−" como *fracaso*, entonces se tiene la hipótesis de que $\mathbb{P}(X_i < Y_i) = 1/2 = \mathbb{P}(X_i > Y_i)$ para cada i y la estadística de prueba definida como el número de signos "+" que se obtengan de la muestra sigue una distribución $binom(n, 1/2)$ donde n es el número total de parejas asignadas con un signo "+" o "−".

Datos. Sea $(X_1, Y_1), (X_2, Y_2), \ldots, (X_m, Y_m)$ una muestra aleatoria de tamaño m de una variable aleatoria bivariada. Supondremos que los vectores aleatorios (X_i, Y_i) para $i = 1, \ldots, m$ son independientes y las $X's$ y $Y's$ *no son necesariamente independientes.*

Por ejemplo, si se está estudiando la pérdida de peso en 50 pacientes sometidos a un cambio de dieta para bajar de peso, los valores de las $X's$ podrían ser los pesos de dichos pacientes antes de someterse a la dieta y los valores de las $Y's$ los pesos de los mismos pacientes después de la dieta. Por otro lado, la prueba de signos también tiene la ventaja de ser utilizada para datos dicotómicos, como son bueno-malo, si-no, alto-bajo, etc. Por ejemplo, si se estudia la preferencia

de los consumidores de café con respecto a dos marcas nuevas en el mercado, A y B, la prueba de signos puede ser utilizada para determinar si la marca B ha tenido mayor preferencia que la marca A.

Para cada observación de la muestra (X_i, Y_i) con $i = 1, 2, \ldots, m$

a) Se asigna un signo "+" si $X_i < Y_i$.

b) Se asigna un signo "−" si $X_i > Y_i$.

c) Se asigna un signo "0" o se omite la observación si $X_i = Y_i$.

A la probabilidad $\mathbb{P}(Xi < Y_i)$ se le denotará por $\mathbb{P}(+)$ y a la probabilidad $\mathbb{P}(X_i > Y_i)$ por $\mathbb{P}(-)$.

Supuestos.

1. La observación (X_i, Y_i) es independiente de la observación (X_j, Y_j) para $i \neq j$.

2. La escala de medida es al menos ordinal en cada par, es decir, a cada pareja (X_i, Y_i), puede ser asignada con un "+", "−", o "0".

3. Los vectores $(X_1, Y_1), (X_2, Y_2), \ldots, (X_m, Y_m)$ son independientes e idénticamente distribuidos y para cada uno de ellos ocurrirá uno y sólo uno de los siguientes tres casos:

   a) $\mathbb{P}(+) > \mathbb{P}(-)$,

   b) $\mathbb{P}(+) < \mathbb{P}(-)$,

   c) $\mathbb{P}(+) = \mathbb{P}(-)$.

Estadística de prueba. La estadística de prueba propuesta es

$$T = \text{El número de parejas a las cuales se les asigna el signo "+"}$$

ya que se considera a la *clase 1* como el conjunto de todos los signos "+" obtenidos.

Observación 1.2. *Si n es el número total de parejas que son asignadas con un signo "+" o signo "−", sin tomar en cuenta las parejas a las que se les asigna el signo "0", y como se cosidera el valor específico de $p^* = 1/2$, entonces la variable aleatoria $T \sim binom(n, 1/2)$.*

Hipótesis a probar. A continuación se describen los contrastes de hipótesis para la prueba del signo.

Caso A (Dos colas)

Nuestro interés es contrastar las hipótesis

$$H_0 : \mathbb{P}(+) = \mathbb{P}(-) \quad vs \quad H_1 : \mathbb{P}(+) \neq \mathbb{P}(-)$$

para el nivel de significancia $\alpha \in (0, 1)$.

Regla de decisión. Se rechaza la hipótesis H_0 al nivel de significancia α si $T \leq t_1$ ó $T \geq t_2$, donde t_1 es el cuantil que acumula aproximadamente $\alpha/2$ de probabilidad en la cola inferior, t_2 es el cuantil que acumula aproximadamente $\alpha/2$ de probabilidad en la cola superior y $t_2 = n - t_1$ (véase la Figura 1.4).

$$\mathbb{P}(T \leq t_1) \approx \alpha/2,$$
$$\mathbb{P}(T \geq t_2) \approx \alpha/2.$$

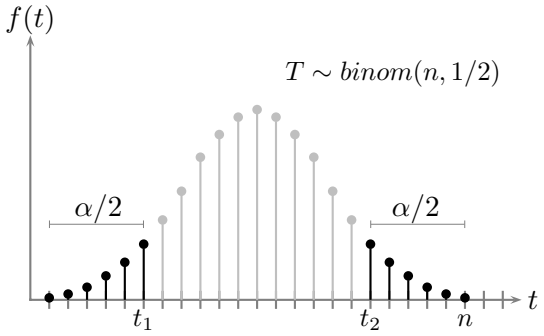


Figura 1.4: Región de rechazo para el contraste de dos colas.

En la tabla C.1 se muestran los valores de la distribución binomial para $n \leq 15$. Si n es mayor que 20 se calcula el cuantil con la aproximación normal de la siguiente manera. Si la variable aleatoria $T \sim binom(n, 1/2)$ entonces $\mathbb{E}(T) = n/2$ y $Var(T) = n/4$. Entonces aplicando el teorema central del límite se tiene que

$$Z \approx \frac{T - n/2}{\sqrt{n/4}} \sim N(0, 1) \tag{1.6}$$

entonces

$$T \approx \frac{1}{2}\big(n + Z\sqrt{n}\big). \tag{1.7}$$

De la ecuación (1.7) el cuantil t_1 está dado por

$$t_1 \approx \frac{1}{2}\big(n + z_{\alpha/2}\sqrt{n}\big) \tag{1.8}$$

donde $z_{\alpha/2}$ es el cuantil que acumula $\alpha/2$ de probabilidad de una distribución normal estándar. Si $\alpha = 0.05$, $z_{\alpha/2} = -1.96$ por lo que la ecuación (1.8) tiene la forma

$$t_1 \approx \frac{1}{2}\big(n - 1.96\sqrt{n}\big)$$

entonces, t_1 y t_2 pueden ser aproximadamente

$$t_1 \approx \frac{n}{2} - \sqrt{n} \qquad\qquad t_2 \approx \frac{n}{2} + \sqrt{n} \tag{1.9}$$

y estas dos expresiones pueden ser recordadas con facilidad.

Valor p. Si definimos a t_{obs} como el valor observado de la estadística T, el valor p, para un contraste de dos colas, es dos veces el mínimo entre la probabilidad de que T sea menor o igual al valor observado de T y la probabilidad de que T sea mayor o igual que el valor observado de T, es decir

$$v_p = 2\,\text{mín}\{\mathbb{P}(T \leq t_{obs}), \mathbb{P}(T \geq t_{obs})\}$$

Para valores de $n > 20$, las probabilidades $\mathbb{P}(T \leq t_{obs})$ y $\mathbb{P}(T \geq t_{obs})$ se calculan de la siguiente manera

$$\mathbb{P}(T \leq t_{obs}) \approx \mathbb{P}\left(Z \leq \frac{2 \cdot t_{obs} - n + 1}{\sqrt{n}}\right), \tag{1.10}$$

$$\mathbb{P}(T \geq t_{obs}) \approx 1 - \mathbb{P}\left(Z \leq \frac{2 \cdot t_{obs} - n - 1}{\sqrt{n}}\right). \tag{1.11}$$

En las expresiones anteriores se incorpora el valor 1.0 como una corrección de continuidad que mejora la aproximación normal a la binomial.

Caso B (Cola inferior)

Se desea contrastar

$$H_0 : \mathbb{P}(+) \geq \mathbb{P}(-) \quad vs \quad H_1 : \mathbb{P}(+) < \mathbb{P}(-)$$

Si se observan valores pequeños de la estadística T, esto indica que la ocurrencia de un signo "$-$" es más probable que la ocurrencia de un signo "$+$", lo cual corresponde a no rechazar la hipótesis alternativa H_1.

Regla de decisión. Se rechaza la hipótesis H_0 al nivel de significancia α si $T \leq t_3$, donde t_3 es el cuantil que acumula aproximadamente α de probabilidad en la cola inferior de la distribución de T (véase la Figura 1.5),

$$\mathbb{P}(T \leq t_3) = \alpha$$

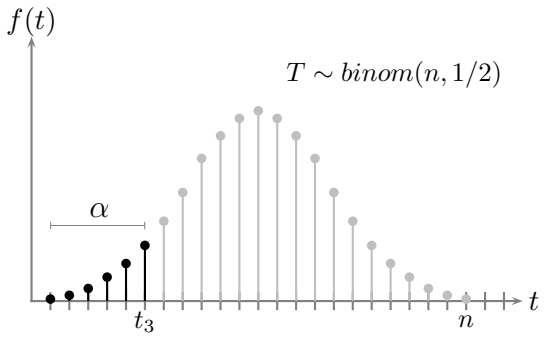


Figura 1.5: Región de rechazo de la prueba de cola inferior.

Si n es mayor que 20 se calcula el cuantil t_3 utilizando la aproximación normal de la ecuación (1.7) obteniendo

$$t_3 \approx \frac{1}{2}(n + z_\alpha \sqrt{n})$$

donde z_α es el cuantil que acumula α de probabilidad de una distribución normal estándar.

Valor p. El valor p es la probabilidad de que T sea menor o igual que el valor observado de la estadística de prueba t_{obs}.

$$v_p = \mathbb{P}(T \leq t_{obs}).$$

Para valores de $n > 20$ la probabilidad $\mathbb{P}(T \leq t_{obs})$ se aproxima como en la ecuación (1.10)

$$\mathbb{P}(T \leq t_{obs}) = \mathbb{P}\left(Z \leq \frac{2 \cdot t_{obs} - n + 1}{\sqrt{n}}\right).$$

Caso C (Cola superior)

Nuestro interés es contrastar las hipótesis

$$H_0 : \mathbb{P}(+) \leq \mathbb{P}(-) \quad vs \quad H_1 : \mathbb{P}(+) > \mathbb{P}(-)$$

Si se observan valores grandes de la estadística T, esto indica que es más probable la ocurrencia de un signo "+" que la ocurrencia de un signo "−", lo cual corresponde a no rechazar la hipótesis alternativa H_1.

Regla de decisión. Se rechaza la hipótesis H_0 al nivel de significancia α si $T \geq t_4$, donde t_4 es el cuantil que acumula α de probabilidad en la cola superior de la distribución de T. El valor t_4 puede ser encontrado a partir del cuantil t que acumula la probabilidad $\mathbb{P}(T \leq t) = \alpha$ como $t_4 = n - t$ (véase la Figura 1.6).

$$\mathbb{P}(T \geq t_4) = \alpha$$

Si n es mayor que 20 se calcula el cuantil con la aproximación normal como en la ecuación (1.7) obteniendo

$$t_4 \approx \frac{1}{2}(n + z_{1-\alpha} \sqrt{n})$$

donde $z_{1-\alpha}$ es el cuantil que acumula $1 - \alpha$ de probabilidad de la distribución normal estándar.

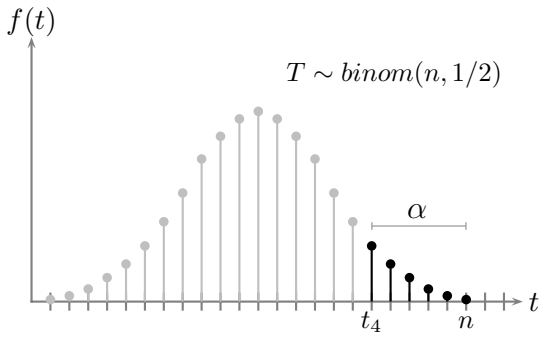


Figura 1.6: Región de rechazo de la prueba de cola superior.

Valor p. El valor p es la probabilidad de que T sea mayor o igual que el valor observado de la estadística t_{obs}

$$v_p = \mathbb{P}(T \geq t_{obs}).$$

Para valores de $n > 20$ la probabilidad $\mathbb{P}(T \geq t_{obs})$ se aproxima como en la ecuación (1.11)

$$\mathbb{P}(T \geq t_{obs}) = 1 - \mathbb{P}\left(Z \leq \frac{2 \cdot t_{obs} - n + 1}{\sqrt{n}}\right).$$

Ejemplo 1.3. *En la Facultad de Ciencias, 8 estudiantes se someten a una dieta para bajar de peso. Se pesó a los estudiantes antes y después de la dieta, obteniendo los siguientes resultados:*

Estudiante	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Antes	*74*	*91*	*88*	*82*	*101*	*88*	*77*	*82*
Después	*65*	*89*	*83*	*73*	*103*	*81*	*80*	*82*

¿Es efectiva la dieta para bajar de peso? Se llevará a cabo el contraste de hipótesis correspondiente utilizando un nivel de significancia de $\alpha = 1\,\%$.

Solución: Como se quiere probar la afirmación de que la dieta no es efectiva para bajar de peso, se utilizará el Caso B (cola inferior) para contrastar las hipótesis

$$H_0 : \mathbb{P}(+) \geq \mathbb{P}(-) \quad vs \quad H_1 : \mathbb{P}(+) < \mathbb{P}(-)$$

donde la hipótesis nula H_0 plantea que la dieta no es efectiva para bajar de peso. Asignando los signos correspondientes a cada pareja de observaciones por alumno, se obteniene la siguiente tabla de resultados

Estudiante	1	2	3	4	5	6	7	8
X_i: Antes	74	91	88	82	101	88	77	82
Y_i: Después	65	89	83	73	103	81	80	82
Signo	$-$	$-$	$-$	$-$	$+$	$-$	$+$	0

Bajo los resultados obtenidos, podemos observar que los datos muestran la ocurrencia de más signos "$-$" que signos "$+$". El tamaño de la muestra bivariada

es $m = 8$, el número de parejas a las cuales se les asigna el signo "+" es $T = 2$ y el total de signos "+" y "−" asignados es $n = 7$, entonces $T \sim binom(7, 1/2)$. Estamos interesados en encontrar el cuantil t_1 tal que $\mathbb{P}(T \leq t_1)$ sea aproximada a 0.01, calculando las probabilidades, tenemos los valores

```
# Genera la secuencia del 0 al 7
> t <- 0:7
# Calcula las probabilidades de t
> pbinom(t, size=7, prob=1/2)
```

```
[1] 0.007813 0.062500 0.226563 0.500000 0.773437 0.937500
[7] 0.992188 1.000000
```

vemos que los posibles valores de $\mathbb{P}(T \leq t)$ que se aproximan a $\alpha = 0.01$ son:

$$\mathbb{P}(T = 0) = \mathbf{0.0078} \quad \text{ó} \quad \mathbb{P}(T \leq 1) = 0.0625$$

El cuantil adecuado es $t = 0$ ya que es el valor que acumula una probabilidad más cercana y que no excede el nivel de significancia $\alpha = 0.01$. Sabemos que se rechaza H_0 si $T \leq t$, esto es, $2 \nleq 0$, entonces no se rechaza la hipótesis H_0, es decir, la probabilidad de ocurrencia de signos "+" es mayor que la de signos "−". El valor p está dado por $\mathbb{P}(T \leq t_{obs})$, usando `R` tenemos que

```
# Calcula el valor p
> pbinom(q=2, size=7,prob=1/2)
```

```
[1] 0.2266
```

Como el valor p excede el nivel de significancia α, no se rechaza la hipótesis H_0, entonces podemos concluir que no existe evidencia suficiente para contradecir que la dieta no ayuda a bajar de peso con un nivel de significancia $\alpha = 1\,\%$.

Observación 1.3. *Nótese que si se tiene un nivel de significancia $\alpha = 23\,\%$, la dieta sí es efectiva.*

Usando la prueba binomial exacta `binom.test()` podemos resolver el ejercicio rapidamente de la siguiente manera:

```
# Prueba binomial exacta
> binom.test(x=2, n=7, p=1/2, alternative="less")
```

```
        Exact binomial test

data:  2 and 7
number of successes = 2, number of trials = 7, p-value = 0.2266
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000 0.6587
sample estimates:
probability of success
                0.2857
```

▪

Observación 1.4. *Si el número de éxitos es grande, se puede utilizar el comando* `prop.test()`*, el cual utiliza una distribución* $\chi^2(1)$ *con una corrección de continuidad de 1.0 para aproximar a la binomial, entonces cuando* $p = 1/2$ *el estadístico de prueba está dado por*

$$\chi^2 = \left[\frac{2 \cdot t_{obs} - n \pm 1}{\sqrt{n}}\right]^2.$$

La corrección de continuidad de 1.0 la acompaña el signo "+" si se lleva a cabo el contraste de cola inferior y un signo "−" si se lleva a cabo el contraste de cola superior.

1.4. Pruebas de aleatoriedad

En esta sección estudiaremos las principales pruebas que se utilizan para detectar la existencia de aleatoriedad en un conjunto de datos, o bien, para determinar si los datos muestrales en una secuencia son aleatorios o muestran alguna tendencia. Las pruebas de aleatoriedad se basan en datos muestrales que cuentan con dos características y se analizan las rachas de esas características para determinar si las rachas parecen ser el resultado de algún proceso aleatorio, o si las rachas sugieren que el orden de los datos no es aleatorio.

1.4.1. Pruebas basadas en el número total de corridas

Supongamos que se tienen n observaciones $X_1, X_2, \ldots, X_n$, y se quiere averiguar si el proceso que las generó es aleatorio. Supongamos además que cada observación puede clasificarse en una de dos categorías, la forma en que se definan las dos categorías va a depender del problema específico que se tenga y de la hipótesis que se desee probar. Si la información está dada en forma numérica, puede tomarse un punto a partir del cual todas las observaciones muestrales mayores o iguales a ese punto sean clasificadas como tipo 1, y las menores a él del tipo 2. Este punto podría ser la mediana de las observaciones, un cuantil conveniente de acuerdo al problema, o un punto cualquiera.

Definición 1.1. *Dada una sucesión ordenada de dos o más tipos de símbolos, una corrida o racha, es una sucesión de uno o más tipos de símbolos que tienen antes y después un símbolo diferente o ninguno.*

Por ejemplo, en la sucesión

$$\underline{\text{M}}\ \underline{\text{H H}}\ \underline{\text{M}}\ \underline{\text{H}}\ \underline{\text{M M}}\ \underline{\text{H}}$$

donde se puede pensar que la letra M representa una mujer y la letra H un hombre, hay seis corridas o rachas. Otro ejemplo sería el siguiente: supóngase que todos los números menores o iguales a 20 son semejantes entre sí y los mayores a 20 son semejantes entre sí, se tendrá que en la siguiente sucesión

13 21 9 12 20 26 24 15 29

existen seis rachas o corridas. El siguiente código en R crea la función `nruns()` la cual calcula el número total de corridas o rachas de un vector con datos numéricos, donde los números mayores a 0 son semejantes entre sí o del tipo 1 y los números menores a cero son semejantes entre sí o del tipo 2.

```
# Calcula el número de corridas en un vector de datos
> nruns <- function(x) {
signs <- sign(x)
runs <- rle(signs)
r <- length(runs$lengths)
return(r)
}
```

Por ejemplo, considere el siguiente conjunto de datos:

14, 11, 110, 9, 5, 2, -4, 2, -1, 5, 2, -8, -9, -7, -10, -6, 4, 8, 11, 5, 1, 2, 4, 7, -1, -6, -5, -3, -3, -1

entonces el número total de corridas se puede calcular utilizando la función `nruns()` de la siguiente forma.

```
# Vector de datos
> datos <- c(14, 11, 10, 9, 5, 2, -4, 2, -1, 5, 2, -8, -9,
-7, -10, -6, 4, 8, 11, 5, 1, 2, 4, 7, -1, -6, -5, -3, -3,
-1)
# Uso de la función nruns()
> nruns(datos)
```

```
[1] 8
```

Muchas rachas, pocas rachas, rachas muy grandes, o muy pequeñas se pueden utilizar para saber si hay aleatoriedad, ya que son cosas que en una secuencia aleatoria no ocurrirán. Las rachas se pueden utilizar tanto para variables cuantitativas o cualitativas. Considérese una muestra de n elementos, en donde

$$n_1 : \text{Son del tipo 1,}$$
$$n_2 : \text{Son del tipo 2,}$$

y $n = n_1 + n_2$. Además supóngase que existen

$$r_1 : \text{Número de corridas del tipo 1,}$$
$$r_2 : \text{Número de corridas del tipo 2,}$$

por lo que el *total de corridas* está dado por $r = r_1 + r_2$. Debido a que el número de corridas es el que ayudará a decidir si la muestra es aleatoria o no, la prueba se basa en la distribución de la variable aleatoria número total de corridas, denotada como R. La función de densidad de probabilidad de la

variable aleatoria R, el número total de corridas de $n = n_1 + n_2$ objetos, n_1 del tipo 1 y n_2 del tipo 2 de una muestra aleatoria está dada por

$$f_R(r) = \begin{cases} 2\binom{n_1-1}{r/2-1}\binom{n_2-1}{r/2-1} \Big/ \binom{n_1+n_2}{n_1} & \text{si } r \text{ es par} \\ \left[\binom{n_1-1}{(r-1)/2}\binom{n_2-1}{(r-3)/2} + \binom{n_1-1}{(r-3)/2}\binom{n_2-1}{(r-1)/2}\right] \Big/ \binom{n_1+n_2}{n_1} & \text{si } r \text{ es impar} \end{cases} \tag{1.12}$$

para $r = 2, 3, \ldots, n_1 + n_2$. Para esta variable aleatoria se tiene que

$$\mathbb{E}(R) = 1 + \frac{2n_1n_2}{n_1 + n_2} \tag{1.13}$$

y

$$Var(R) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)} \tag{1.14}$$

Estas características de la variable aleatoria R se pueden encontrar en [11] donde se explica detalladamente su desarrollo. Existen tablas para la variable aleatoria R donde $n_1 \leq n_2 \leq 12$, lo cual indica que el tipo de elementos que ocurren con menos frecuencia debe considerarse del tipo 1.

Las siguientes líneas de código en `R` generan la función `fd.runs()` la cual representa la función de densidad de la variable aleatoria R y facilita el cálculo de probabilidades cuando se conoce n_1, n_2 y para algún valor que toma la variable aleatoria R en el conjunto $\{1, 2, \ldots\}$.

```
# Función de densidad de R
> fd.runs <- function(n1,n2,r){
i=r%%2
if(i==0){
fd.runs = (2 * choose(n1-1,(r/2)-1) * choose(n2-1,(r/2)-
1))/choose(n1+n2,n1)
}
else
fd.runs = ((choose(n1-1,(r-1)/2) * choose(n2-1,(r-3)/2))
+ (choose(n1-1,(r-3)/2) * choose(n2-1,(r-1)/2))) /
choose(n1+n2,n1)
}
```

Por ejemplo, si $n_1 = 5$, $n_2 = 4$ y cuando el número total de corridas es $R = 9$, $R = 8$, $R = 2$ y $R = 3$ se tienen los siguientes resultados

```
> j <- fd.runs(5,4,9)
> j
```

```
[1] 0.007937
```

```
> k <- fd.runs(5,4,8)
> k
```

```
[1] 0.06349
```

```
> l <- fd.runs(5,4,2)
> l
```

```
[1] 0.01587
```

```
> m <- fd.runs(5,4,3)
> m
```

```
[1] 0.05556
```

Para calcular $P(R \leq r)$, o bien, la función de distribución se puede emplear el siguiente código con ayuda de la función `fd.runs()` generada anteriormente.

```
# Función de distribución de R
> F.runs <- function(n1,n2,r){
S <- rep(1,r)
S[1] <- fd.runs(n1,n2,1)
# Ciclo for
for(i in 2:r){
S[i] <- fd.runs(n1,n2,i) + S[i-1]
}
print(S[i])
}
```

Por ejemplo, si deseamos calcular $P(R \leq 8)$ cuando $n_1 = 6$ y $n_2 = 14$ tenemos

```
> F.runs(6,14,8)
```

```
[1] 0.299
```

Estadística de prueba. La estadística de prueba es el valor de la variable aleatoria R, o bien, el número total de corridas.

Hipótesis a probar. Los tipos de contrastes de interés se presentan en los siguientes tres casos.

Caso A (Cola inferior)

$$H_0: \text{Los elementos de la muestra presentan aleatoriedad} \quad vs \quad H_1: \text{Los elemetos de la muestra presentan tendencia a agruparse}$$

Regla de decisión. Rechazar H_0 al nivel de significancia α si $R \leq r_\alpha$, donde r_α es el cuantil que acumula α de probabilidad. Por esta razón, este contraste se realiza cuando hay pocas corridas o rachas.

Valor p. El valor p viene dado por $\mathbb{P}(R \leq r)$.

Caso B (Cola superior)

$$H_0: \text{Los elementos de la muestra presentan aleatoriedad} \quad vs \quad H_1: \text{Los elementos de la muestra presentan tendencia a revolverse}$$

Regla de decisión. Rechazar H_0 al nivel de sigificancia α si $R \geq r_{1-\alpha}$, donde $r_{1-\alpha}$ es el cuantil que acumula $1-\alpha$ de probabilidad. Por esta razón este contraste se utiliza cuando hay muchas corridas o rachas.

Valor p. El valor p viene dado por $\mathbb{P}(R \geq r)$.

Caso C (Dos colas)

$$H_0: \text{Los elementos de la muestra presentan aleatoriedad} \quad vs \quad H_1: \text{Los elementos de la muestra no presentan aleatoriedad}$$

Para poder contrastar este caso se tendrá que encontrar $r_{\alpha/2}$ y $r_{1-\alpha/2}$ tales que

$$\mathbb{P}(R \leq r_{\alpha/2}) \leq \frac{\alpha}{2} \quad \text{y} \quad \mathbb{P}(R \geq r_{1-\alpha/2}) \leq \frac{\alpha}{2}.$$

Así el nivel de significancia exacto está dado por

$$\alpha_{exacto} = \mathbb{P}(R \leq r_{\alpha/2}) + \mathbb{P}(R \geq r_{1-\alpha/2})$$

Regla de decisión. Se rechaza H_0 para un nivel de significancia α_{real} si $R \leq r_{\alpha/2}$ o $R \geq r_{1-\alpha/2}$.

Valor p. El valor p está dado por dos veces el mínimo entre $\mathbb{P}(R \leq r)$ y $\mathbb{P}(R \geq r)$, es decir

$$v_p = 2\,\text{mín}\{\mathbb{P}(R \leq r), \mathbb{P}(R \geq r)\}$$

donde r es el valor observado de la variable aleatoria R de la muestra aleatoria.

Aproximación normal

Para $n_1 > 12$ y $n_2 > 12$, los valores críticos r_α y $r_{1-\alpha}$ se pueden encontrar a partir de la aproximación normal de la distribución nula del número total de corridas. Tomando la esperanza y la varianza de R tenemos que

$$Z \approx \frac{R - 1 - 2n_1n_2/n}{\sqrt{2n_1n_2(2n_1n_2 - n)/[n^2(n-1)]}}.$$

Utilizando una corrección de continuidad de 0.5, las regiones críticas para el caso de cola inferior, cola superior y dos colas se presentan a continuación.

Caso A (Cola inferior)

$$\frac{R + 0.5 - 1 - 2n_1n_2/n}{\sqrt{2n_1n_2(2n_1n_2 - n)/[n^2(n-1)]}} \leq z_\alpha$$

El valor p está dado por

$$\Phi\left(\frac{R + 0.5 - 1 - 2n_1n_2/n}{\sqrt{2n_1n_2(2n_1n_2 - n)/[n^2(n-1)]}}\right)$$

Caso B (Cola superior)

$$\frac{R - 0.5 - 1 - 2n_1n_2/n}{\sqrt{2n_1n_2(2n_1n_2 - n)/[n^2(n-1)]}} \geq z_{1-\alpha}$$

El valor p está dado por

$$1 - \Phi\left(\frac{R - 0.5 - 1 - 2n_1n_2/n}{\sqrt{2n_1n_2(2n_1n_2 - n)/[n^2(n-1)]}}\right)$$

Caso C (Dos colas)

La región crítica se calcula utilizando ambas expresiones con $z_{\alpha/2}$ en lugar de z_α y el valor p está dado por dos veces el mínimo de los dos casos anteriores.

Ejemplo 1.4. *Un equipo profesional de futbol tiene la siguiente sucesión de triunfos y derrotas en la última temporada*

g p g g p g p p p g p p g g p g g p g p g

donde la letra "g" indica que el juego fue ganado y la letra "p" que el juego fue perdido. ¿Puede decirse que el récord de sus triunfos y derrotas es aleatorio?. En este ejemplo, se quiere probar si el récord de triunfos y derrotas es aleatorio o no, utilizando un nivel de significancia $\alpha = 5\,\%$.

Solución: Tenemos que el número de observaciones del tipo 1 (juegos perdidos) es $n_1 = 10$ y el número de observaciones del tipo 2 (juegos ganados) es $n_2 = 11$, por lo que el número de observaciones totales es $n = 21$ (total de juegos). Por otro lado vemos que el número de corridas del tipo 1 es de $r_1 = 7$ y el número de corridas del tipo 2 es de $r_2 = 8$ por lo que el número total de corridas es $r = 15$. Como sólo se quiere comprobar si los triunfos y derrotas son o no son aleatorios, se utiliza el Caso C (Dos colas), contrastando

$$H_0: \text{Los triunfos y derrotas del equipo son aleatorios} \quad vs \quad H_1: \text{Los triunfos y derrotas del equipo no son aleatorios}$$

Para la cola inferior queremos encontrar el valor $r_{0.025}$ para el cual $\mathbb{P}(R \leq r_{0.025}) = 0.025$, buscando en la Tabla C.6 o utilizando la función `F.runs()` tenemos

```
# P(R ≤ 6) cuando n1 = 10 y n2 = 11
> F.runs(10,11,6)
```

```
[1] 0.01192
```

```
# P(R ≤ 7) cuando n1 = 10 y n2 = 11
> F.runs(10,11,7)
```

```
[1] 0.03489
```

entonces, las probabilidades aproximadas a $\mathbb{P}(R \leq r_{0.025}) = 0.025$ son:

$$\mathbb{P}(R \leq 6) = \mathbf{0.01192} \quad \text{y} \quad \mathbb{P}(R \leq 7) = 0.03489$$

La probabilidad que se debe tomar es 0.01192 ya que este valor no excede a 0.025, porque de lo contrario estaríamos haciendo el error α más grande al tomar el valor 0.03488926, por lo tanto, el valor $r_{0.025}$ para la cola inferior es 6.

Análogamente, para la cola superior queremos encontrar el valor $r_{0.975}$ para el cual $\mathbb{P}(R \geq r_{0.975}) = 0.025$, nuevamente buscando en la Tabla C.6 o usando la función `F.runs()` tenemos

```
# P(R ≤ 15) y P(R > 15) = P(R ≥ 16) cuando n1 = 10 y n2 = 11
> 1 - F.runs(10,11,15)
```

```
[1] 0.9651
[1] 0.03489
```

```
# P(R ≤ 16) y P(R > 16) = P(R ≥ 17) cuando n1 = 10 y n2 = 11
> 1 - F.runs(10,11,16)
```

```
[1] 0.9896
[1] 0.01039
```

entonces, las probabilidades aproximadas a $\mathbb{P}(R \geq r_{0.975}) = 0.025$ son:

$$\mathbb{P}(R \geq 16) = 0.03489 \quad \text{y} \quad \mathbb{P}(R \geq 17) = \mathbf{0.01039}$$

En consecuencia, se toma el valor $r_{0.975} = 17$. Sumando estas dos probabilidades tenemos que $\alpha_{real} = 0.01192 + 0.01039 = 0.02231$. Entonces, se rechaza la hipótesis H_0 si $r \leq r_{0.025}$ ó $r \geq r_{0.975}$, pero $15 \nleq 6$ ó $15 \ngeq 17$ con un nivel de significancia $\alpha = 0.02231$, por lo tanto no se rechaza la hipótesis H_0. Podemos concluir que no hay evidencia suficiente para decir que el récord de triunfos y derrotas del equipo no es aleatorio. Para conocer el valor p, basta con calcular $\mathbb{P}(R \geq 15)$ y compararla contra la probabilidad $\mathbb{P}(R \leq 15)$ antes encontrada.

```
# P(R ≤ 14) y P(R > 14) = P(R ≥ 15)
> 1 - F.runs(10, 11, 14)
```

```
[1] 0.9151
[1] 0.0849
```

Entonces, como el valor p es dos veces el mínimo entre $\mathbb{P}(R \leq 15) = 0.9651$ y $\mathbb{P}(R \geq 15) = 0.0849$, el valor p es 0.1698 y como este valor es mayor que el nivel de significancia $\alpha = 0.05$, se concluye que no existe aleatoriedad en la muestra,

por lo tanto, el récord de triunfos y derrotas del equipo de futbol es aleatorio.

En `R` existe la instrucción `runs.test()` que permite obtener un resultado inmediato de la prueba del número total de corridas. Esta instrucción utiliza la aproximación normal y proporciona el valor de la estadística de prueba y el valor p. Para utilizar la instrucción `runs.test()` es necesario utilizar la biblioteca `tseries` y cuenta con los siguientes argumentos

```
runs.test(datos, alternative)
```

donde `datos` es un vector con los resultados de la muestra aleatoria y `alternative` especifica qué tipo de prueba se quiere llevar a cabo.

a) `alternative="less"` prueba de cola inferior.

b) `alternative="greater"` prueba de cola superior.

c) `alternative="two.sided"` prueba de dos colas.

Las siguientes líneas de código proporcionan los resultados que se obtienen usando la instrucción `runs.test()` para resolver el ejemplo anterior.

```
# Carga la biblioteca tseries
> library(tseries)
# Vector de datos
> x <- c( 1 , 0 , 1 , 1 , 0 , 1 , 0 , 0 , 0 , 1 , 0 , 0 , 1
, 1 , 0 , 1 , 1 , 0 , 1 , 0 , 1 )
# Factores o categorías el vector x
> rdata <- factor(x, labels=c("p","g"))
> rdata
```

```
 [1] g p g g p g p p p g p p g g p g g p g p g
Levels: p g
```

```
# Lleva a cabo la prueba
> runs.test(rdata, alternative="two.sided")
```

```
        Runs Test

data:  rdata
Standard Normal = 1.582, p-value = 0.1137
alternative hypothesis: two.sided
```

▪

1.4.2. Corridas hacia arriba y hacia abajo

La prueba de corridas hacia arriba y hacia abajo es útil cuando se tienen observaciones numéricas y se analiza si una sucesión de observaciones es aleatoria de acuerdo con el número de corridas que se obtengan. En esta prueba, en lugar de comparar contra un punto fijo, se compara cada observación con la observación

siguiente asignando un signo "+" si ésta última *es mayor*, o un signo "−" *si es menor*. De esta forma, el número de corridas R estará determinado por las sucesiones de signos "+" y signos "−" que se obtengan. Por ejemplo, si tenemos las observaciones 9, 14, 2, 4, 8, 7, entonces, la sucesión de signos es +,−,+,+,− la cual indica que el número de *corridas hacia arriba* (tipo 1) fue de $r_1 = 2$ y el número de *corridas hacia abajo* (tipo 2) también fue de $r_2 = 2$ teniendo un total de corridas $r = 4$. Cuando se analiza aleatoriedad en una sucesión numérica de acuerdo al número de corridas hacia arriba y hacia abajo de la *mediana*, se pierde información para identificar un patrón en las observaciones a través del tiempo.

Datos. Sea $X_1, X_2, \ldots, X_n$ una muestra de tamaño n. Se construye la sucesión de signos de la siguiente manera:

a) Si $X_i < X_{i+1}$ se asigna un signo "+",

b) Si $X_i > X_{i+1}$ se asigna un signo "−",

c) Si $X_i = X_{i+1}$ no hay cambio de signo,

donde el tamaño máximo de la sucesión de signos será $n - 1$.

Estadística de prueba. La estadística de prueba es el valor de la variable aleatoria R, a partir del número total de corridas de signos "+" y signos "−".

Hipótesis a probar. Para la prueba de aleatoriedad basada en el número total de corridas hacia arriba y hacia abajo, en una secuencia ordenada de n observaciones numéricas, o equivalentemente, una secuencia de $n - 1$ signos "+" ó "−" las regiones de rechazo para cada tipo de contraste son las mismas que en la sección 1.4.1, la cual se basa en el número total de corridas de dos tipos de elementos. Específicamente, si la hipótesis alternativa en que los signos muestran tendencia a agruparse, la región apropiada es para valores de R pequeños. Si la hipótesis alternativa es que los signos tienden a revolverse la región apropiada es para valores de R grandes. De esta manera, los tipos de contrastes son los siguientes:

Caso A (Cola inferior)

$$H_0 : \text{Los signos muestran aleatoriedad} \quad vs \quad H_1 : \text{Los signos muestran tendencia a agruparse}$$

En este contraste, si existe tendencia de los signos a agruparse se tiene:

a) Tendencia creciente si la mayoría de signos son "+".

b) Tendencia decreciente si la mayoría de signos es "−".

Sin embargo, se rechaza H_1 cuando hay pocas corridas.

Regla de decisión. Rechazar la hipótesis H_0 si $R \leq r_\alpha$, donde r_α cumple con $\mathbb{P}(R \leq r_\alpha) = \alpha$.

Valor p. El valor p está dado por $\mathbb{P}(R \leq r)$.

Caso B (Cola superior)

$$H_0 : \text{Los signos muestran aleatoriedad} \quad vs \quad H_1 : \text{Los signos muestran tendencia a revolverse}$$

Regla de decisión. Rechazar H_0 si $R \geq r'_\alpha$ donde r_α cumple con $\mathbb{P}(R \geq r_\alpha) = \alpha$. En este contraste se rechaza la hipótesis alternativa H_1 si se observan muchas corridas.

Valor p. El valor p está dado por $\mathbb{P}(R \geq r)$.

Caso C (Dos colas)

$$H_0 : \text{Los signos muestran aleatoriedad} \quad vs \quad H_1 : \text{Los signos no muestran aleatoriedad}$$

La hipótesis nula puede ser interpretada como: "los datos son independientes y tienen la misma distribución".

Regla de decisión. Rechazar H_0 si $R \leq r_{\alpha/2}$ o $R \geq r'_{\alpha/2}$.

Valor p. El valor p está dado por dos veces el mínimo entre $\mathbb{P}(R \leq r)$ y $\mathbb{P}(R \geq r)$, es decir

$$v_p = 2\,\text{mín}\{\mathbb{P}(R \leq r), \mathbb{P}(R \geq r)\}$$

donde r es el valor observado de la variable aleatoria R de la muestra aleatoria.

Aproximación normal

Para valores de $n > 25$ los valores críticos se pueden encontrar utilizando la aproximación normal.

Caso A (Cola inferior)

$$\frac{R + 0.5 - (2n-1)/3}{\sqrt{(16n-29)/90}} \leq -z_\alpha.$$

El valor p está dado por

$$\Phi\left(\frac{R+0.5-(2n-1)/3}{\sqrt{(16n-29)/90}}\right)$$

Caso B (Cola superior)

$$\frac{R-0.5-(2n-1)/3}{\sqrt{(16n-29)/90}} \geq z_\alpha$$

El valor p está dado por

$$1-\Phi\left(\frac{R-0.5-(2n-1)/3}{\sqrt{(16n-29)/90}}\right)$$

Caso C (Dos colas)

La región crítica se calcula utilizando las expresiones anteriores con $\alpha/2$ y el valor p está dado por dos veces el mínimo de los dos casos anteriores.

En la práctica este tipo de prueba es muy utilizada en el análisis de series de tiempo y cuando los signos presentan tendencia a revolverse se conoce como Variaciones Cíclicas. También se puede usar esta prueba si los datos de la serie se comparan con cierto periodo, con la media o mediana *(corridas hacia arriba y hacia abajo de la mediana)* asignando signo "+" a los valores mayores a la media o mediana, y signo "−" a los valores menores a la media o mediana sin olvidar la definición de corridas o rachas.

Ejemplo 1.5. *Cierto analista anotó el número de bonos vendidos cada mes para un periodo de 12 meses:*

Enero	*19*	*Julio*	*22*
Febrero	*23*	*Agosto*	*24*
Marzo	*20*	*Septiembre*	*25*
Abril	*17*	*Octubre*	*28*
Mayo	*18*	*Noviembre*	*30*
Junio	*20*	*Diciembre*	*21*

Utilizando la prueba de corridas hacia arriba y hacia abajo probaremos si los datos muestran una tendencia hacia alguna dirección utilizando un nivel de significancia α =0.05.

Solución: Tenemos que el número de observaciones es de $n = 12$, por lo que tenemos una secuencia de $n - 1 = 11$ signos.

19	23	+	22	24	+
23	20	−	24	25	+
20	17	−	25	28	+
17	18	+	28	30	+
18	20	+	30	21	−
20	22	+			

En caso de que un dato se repita, se asigna el mismo signo. Sabemos que r_1 es el número de corridas del tipo 1, es decir, $r_1 = 2$ (para el signo "+") y r_2 es el número de corridas del tipo 2, es decir, $r_2 = 2$ (para el signo "−"), por lo que el número total de corridas es $r = 4$.
Para el Caso C (Dos colas) para ver si existe aleatoriedad o no, contrastamos

$$H_0: \text{El número de bonos vendidos muestran un comportamiento aleatorio} \quad vs \quad H_1: \text{El número de bonos vendidos no muestran un comportamiento aleatorio}$$

Para la cola inferior queremos encontrar el valor $r_{\alpha/2}$ de tal forma que $\mathbb{P}(R \leq r_{\alpha/2}) = 0.025$. Buscando en tablas tenemos que las probabilidades más aproximadas a 0.025 son:

$$\mathbb{P}(R \leq 4) = \mathbf{0.0082} \quad \text{ó} \quad \mathbb{P}(R \leq 5) = 0.0539.$$

Para la cola superior queremos encontrar el valor $r_{\alpha/2}$ de tal suerte que $\mathbb{P}(R > r'_{\alpha}) = 0.025$. Buscando en tablas tenemos que las probabilidades aproximadas a 0.025 son:

$$\mathbb{P}(R > 10) = 0.0821 \quad \text{ó} \quad \mathbb{P}(R > 11) = \mathbf{0.0113}.$$

Sumando estas dos probabilidades tenemos que $\alpha_{real} = 0.0082 + 0.0113 = 0.0195$. Sabemos que se rechaza H_0 si $r \leq r_{\alpha/2}$ ó $r \geq r'_{\alpha/2}$, es decir, $4 \leq 4$ ó $4 \not\geq 11$, por lo que se rechaza la hipótesis H_0, por lo tanto podemos concluir que no hay aleatoriedad, es decir hay tendencia.

Ahora, utilizando el Caso A de cola izquierda probaremos las hipótesis

$$H_0: \text{El número de bonos vendidos muestran un comportamiento aleatorio} \quad vs \quad H_1: \text{El número de bonos vendidos muestran una tendencia a agruparse}$$

Queremos encontrar el valor r_α de tal suerte que $\mathbb{P}(R \leq r_\alpha) = 0.05$, buscando en tablas para $n = 12$, tenemos que las probabilidades aproximadas son:

$$\mathbb{P}(R \leq 4) = \mathbf{0.0082} \quad \text{ó} \quad \mathbb{P}(R \leq 5) = 0.0529.$$

Sabemos que se rechaza la hipótesis H_0 si $r \leq r_\alpha$, es decir, $4 \leq 4$, por lo que se rechaza H_0, se puede concluir que, existe tendencia de los signos a agruparse.

▪

1.5. Pruebas basadas en rangos

En esta sección se presentan dos de las pruebas de rangos más utilizadas en estadística no paramétrica, la prueba de Kruskal-Wallis y la prueba de Friedman.

1.5.1. Prueba de Kruskal-Wallis

La prueba de Kruskal-Wallis es el equivalente no paramétrico a la prueba paramétrica del análisis de varianza ANOVA de un factor. Es un método no paramétrico útil para poner a prueba una hipótesis nula relativa a que k muestras independientes provienen de poblaciones continuas idénticas. Es una prueba de homogeneidad para más de dos muestras, es decir, la prueba de suma de rangos para dos muestras independientes se extiende al problema de analizar k muestras independientes, para $k \geq 3$. El siguiente material fue tomado de [3].

Datos. Sean k muestras aleatorias independientes, posiblemente de distintos tamaños, que se obtienen de k poblaciones distintas. Si denotamos a la i-ésima muestra aleatoria de tamaño n_i como $X_{i1}, X_{i2}, \ldots, X_{in_i}$ entonces los datos pueden ser odenados de la siguiente manera:

Muestra 1	Muestra 2	$\cdots$	Muestra i	$\cdots$	Muestra k
X_{11}	X_{21}		X_{i1}		X_{k1}
X_{12}	X_{22}		X_{i2}		X_{k2}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
X_{1n_1}	X_{2n_2}	$\cdots$	X_{in_i}	$\cdots$	X_{kn_k}

Sea N el número total de observaciones, es decir,

$$N = \sum_{i=1}^{k} n_i$$

Para llevar a cabo esta prueba se ordenan las observaciones de menor a mayor, incluyendo repeticiones, y se les asignan los números $1, 2, \ldots, N$ respectivamente. Si dos o más observaciones son idénticas, el orden en que se consideren éstas será irrelevante. Al número asignado a cada observación se le llamará *rango*. De esta forma, se asigna el rango 1 a la observación más pequeña de todas, el rango 2 a la segunda observación más pequeña y así sucesivamente hasta la última observación más grande. Al rango de la observación X_{ij} se le denotará por $R(X_{ij})$. Sea R_i la suma de los rangos asignados a la i-ésima muestra (i-ésima columna):

$$R_i = \sum_{j=1}^{n_i} R(X_{ij}), \quad \text{para } i = 1, 2, \ldots, k.$$

Gráficamente se calcula R_i para cada muestra como se indica en la siguiente tabla.

Rangos Muestra 1	Rangos Muestra 2	$\cdots$	Rangos Muestra k
$R(X_{11})$	$R(X_{21})$	$\cdots$	$R(X_{k1})$
$R(X_{12})$	$R(X_{22})$	$\cdots$	$R(X_{k2})$
$\vdots$	$\vdots$		$\vdots$
$R(X_{1n_1})$	$R(X_{2n_2})$	$\cdots$	$R(X_{kn_k})$
R_1	R_2	$\cdots$	R_k

Si dos o más observaciones son iguales, se asigna nuevamente el promedio de los rangos correspondientes. Por ejemplo, si $X_{11} = X_{26} = X_{35}$ y les corresponden los rangos 5, 6 y 7 respectivamente, el rango asignado será igual a $(5+6+7)/3 = 6$.

Supuestos.

1. Todas las muestras son muestras aleatorias de sus respectivas poblaciones.

2. Además de la independencia dentro de cada muestra, se supone que las muestras son independientes.

3. La escala de medida es al menos ordinal.

4. Supondremos que las k poblaciones tienen función de distribución idéntica, o que al menos una de las poblaciones tiende a tener valores más grandes que las otras poblaciones, ya que la prueba está diseñada para detectar alguna de estas dos situaciones.

Hipótesis a probar. La hipótesis nula y la alternativa que se contrastan son:

H_0 : Todas las k poblaciones tienen funciones de distribución idénticas.

vs

H_1 : Al menos una de las poblaciones tiende a producir observaciones más grandes que las otras poblaciones.

En particular, la prueba de Kruskal-Wallis puede detectar diferencia entre las medias de las k poblaciones de modo que la hipótesis alternativa H_1 se escribe en algunas ocasiones de la siguiente manera:

H_1 : Al menos una de las medias es más grande que las otras

aunque no existe equivalencia entre esta hipótesis alternativa y la anterior.

Estadística de prueba. En 1952, Kruskal y Wallis publican la siguiente estadística de prueba

$$H = \frac{1}{S^2}\left(\sum_{i=1}^{k}\frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4}\right), \tag{1.15}$$

donde N y R_i son como los definimos anteriormente y

$$S^2 = \frac{1}{N-1}\left(\sum_{\text{todos los rangos}} R(X_{ij})^2 - N\frac{(N+1)^2}{4}\right).$$

En 1976, Iman y Devenport publican una nueva aproximación de la distribución exacta de la prueba estadística de Kruskal-Wallis. Para profundizar sobre este resultado puede consultar [12] y [15].

En el caso en que las k muestras son idénticas y del mismo tamaño $n = n_1 = \cdots = n_k$ se espera que la hipótesis nula H_0 no sea rechazada, ya que en esta situación se puede mostrar que $\sum_{i=1}^{k} R_i^2/n_i$ es igual a $N(N+1)^2/4$. Esto indica que para valores pequeños de esta diferencia, la hipótesis nula H_0 no se rechaza y para valores grandes se rechaza.

Cuando *no hay empates* se puede verificar que $S^2 = N(N+1)/12$, por lo que la estadística de prueba es

$$H = \left(\frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i}\right) - 3(N+1) \tag{1.16}$$

Regla de decisión. Rechazar H_0 al nivel de significancia α si $H > w_{1-\alpha}$. Donde el cuantil $w_{1-\alpha}$ se obtiene de la tabla de cuantiles C.7 de la distribución de H. En el caso de no encontrarse este valor en tablas, se utiliza la aproximación Ji-cuadrada con $k-1$ grados de libertad, ya que la estadística se distribuye asintóticamente como una $\chi^2_{(k-1)}$ cuando el mín$\{n_1, \ldots, n_k\} \to \infty$ siempre que la hipótesis nula sea cierta, aunque supondremos válida dicha aproximación cuando $n_i > 5$ para todo i.

La aproximación de la estadística H para una muestra grande se basa en el hecho de que R_i es la suma de n_i variables aleatorias, y para n_i grande se puede utilizar el teorema central del límite. Así

$$\frac{R_i - \mathbb{E}(R_i)}{\sqrt{Var(R_i)}}$$

se distribuye aproximadamente normal estándar cuando la hipótesis nula H_0 es verdadera. Como R_i se puede definir como la suma de n_i enteros seleccionados al azar, sin reemplazo, del conjunto $\{1, 2, \ldots, N\}$ se puede demostrar que

$$\mathbb{E}(R_i) = \frac{n_i(N+1)}{2},$$

$$Var(R_i) = \frac{n_i(N+1)(N-n_i)}{12}.$$

Por consiguiente,

$$\left[\frac{R_i - \mathbb{E}(R_i)}{\sqrt{Var(R_i)}}\right]^2 = \frac{\left(R_i - \frac{n_i(N+1)}{2}\right)^2}{\frac{n_i(N+1)(N-n_i)}{12}}$$

se distribuye aproximadamente $\chi^2_{(1)}$. Si las R_i's fueran independientes

$$H' = \sum_{i=1}^{k} \frac{\left(R_i - \frac{n_i(N+1)}{2}\right)^2}{\frac{n_i(N+1)(N-n_i)}{12}}$$

se distribuye aproximadamente $\chi^2_{(k)}$. Sin embargo

$$\sum_{i=1}^{k} R_i = \frac{N(N+1)}{2}$$

por lo que hay una dependencia entre las R_i's. Kruskal en 1952 mostró que si el i-ésimo término en H' es multiplicado por $\frac{N-n_i}{N}$ para $i = 1, 2, \ldots, k$ entonces se tiene

$$\begin{aligned} H &= \sum_{i=1}^{n} \frac{\left(R_i - \frac{n_i(N+1)}{2}\right)^2}{\frac{n_i(N+1)N}{12}} \\ &= \frac{12}{N(N+1)} \sum_{i=1}^{k} \left(\frac{R_i^2}{n_i} - R_i(N+1) + \frac{n_i(N+1)^2}{4}\right) \\ &= \frac{12}{N(N+1)} \sum_{i=1}^{k} \left(\frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4}\right) \\ &= \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) \end{aligned}$$

se distribuye asintóticamente como una variable aleatoria $\chi^2_{(k-1)}$. La cual corresponde a la estadística H de la ecuación (1.16). Para ver más acerca de este resultado se puede consultar [14].

Valor p. El valor p en este caso se calcula de manera aproximada como la probabilidad de que una variable aleatoria que se distribuye Ji-cuadrada con $k-1$ grados de libertad sea mayor o igual al valor observado h_{obs} de la estadística H. En otras palabras, el valor p se calcula como la probabilidad acumulada en la cola derecha de la distribución Ji-cuadrada con $k-1$ grados de libertad

$$\text{valor } p = \mathbb{P}(H \geq h_{obs}).$$

Comparaciones múltiples. En caso de que la hipótesis nula se rechace y alguna de las poblaciones tenga observaciones mayores que alguna de las otras, nuestro problema ahora es saber cúal es, por lo que se comparan una a una. Se dice que la población i es diferente de la población j si se satisface la desigualdad

$$\left|\frac{R_i}{n_i} - \frac{R_j}{n_j}\right| > t_{1-\alpha/2}^{N-k}\left(S^2\frac{N-1-H}{N-k}\right)^{1/2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)^{1/2}$$

donde R_i y R_j son la sumas de los rangos de las muestras i y j, $t_{1-\alpha/2}^{N-k}$ es el cuantil $1-\alpha/2$ de la distribución t de Student con $N-k$ grados de libertad, S^2 como la definimos anteriormente y H el valor de la estadística de prueba.

Ejemplo 1.6. *Se hizo un estudio para comparar la efectividad de cuatro tipos de dietas diferentes para bajar de peso en 20 personas de la misma edad. Cada dieta la llevaron a cabo 5 personas elegidas aleatoriamente, registrando la pérdida de peso en kilogramos despues de 4 semanas. Los resultados obtenidos son*

Dieta			
A	*B*	*C*	*D*
6.1	*5.0*	*7.6*	*6.2*
4.3	*5.6*	*6.8*	*8.0*
4.5	*7.3*	*3.9*	*7.4*
2.4	*5.7*	*7.9*	*4.6*
9.1	*2.1*	*5.9*	*7.0*

Con $\alpha = 5\,\%$, determine si hay diferencia significativa entre las dietas en términos de la pérdida de peso promedio.

Solución: En este ejemplo las muestras son independientes por lo que se puede utilizar la prueba Kruskal-Wallis y además se tienen más de 2 muestras. El número de muestras es $k = 4$ y son de igual tamaño, es decir, $n_i = 5$ para $i = 1, 2, 3, 4$. Asignando los rangos a las observaciones obtenidas y calculando la sumas R_i tenemos

A	$R(X_{1i})$	B	$R(X_{2i})$	C	$R(X_{3i})$	D	$R(X_{4i})$
6.1	11	5.0	7	7.6	17	6.2	12
4.3	4	5.6	8	6.8	13	8.0	19
4.5	5	7.3	15	3.9	3	7.4	16
2.4	2	5.7	9	7.9	18	4.6	6
9.1	20	2.1	1	5.9	10	7.0	14
	$R_1 = 42$		$R_2 = 40$		$R_3 = 61$		$R_4 = 67$

Como no hay empates entre las observaciones, usamos la ecuación (1.16) para calcular el valor observado h_{obs} de la estadística H. La estadística de prueba es

$$H = \frac{12}{(20)(21)}\left(\frac{(42)^2+(40)^2+(61)^2+(67)^2}{5}\right) - 3(12) = 3.1371$$

Rechazamos la hipótesis al nivel de significancia del 5 % si $H > w_{0.95}$. Haciendo uso de R, calculando este valor para una variable aleatoria que se distribuye Ji-cuadrada con 3 grados de libertad escribimos

```
# Cuantil w_0.95
> qchisq(p=0.95, 3, lower.tail=T)
```

```
[1] 7.815
```

Como el valor de la estadística $H = 3.1371$ no es mayor que el valor del cuantil $w_{0.95} = 7.814728$, no rechazamos la hipótesis nula H_0. Calculando el valor p con el valor observado de la estadística $h_{obs} = 3.1371$ obtenemos

```
# Valor p
> pchisq(3.1371, df=3, lower.tail=F)
```

```
[1] 0.371
```

Como el valor $p = 0.3709663$ es mayor que el nivel de significancia $\alpha = 0.05$, entonces no rechazamos la hipótesis nula H_0 de igualdad de medias entre la pérdida de peso para las 4 dietas. Por lo tanto podemos concluir que la pérdida de peso promedio es significativamente la misma para las 4 diferentes dietas.
En R existe la instrucción `kruskal.test()` que nos permite obtener un resultado inmediato de la prueba Kruskal-Wallis. Esta instrucción nos proporciona el valor que toma la estadística H, sus grados de libertad y el valor p. La instrucción `kruskal.test()` cuenta con los siguientes argumentos

```
kruskal.test(datos, grupos)
```

donde `datos` representan todas las observaciones de las k muestras y `grupos` especifica los diferentes grupos para los que se quiere contrastar de que todos ellos proceden de la misma distribución. El siguiente código proporciona los mismos resultados obtenidos para la solución del ejemplo anterior.

```
# Introduce el total de observaciones N
> datos <- c(6.1, 4.3, 4.5, 2.4, 9.1, 5.0, 5.6, 7.3,
5.7, 2.1, 7.6, 6.8, 3.9, 7.9, 5.9, 6.2, 8.0, 7.4, 4.6,
7.0)
# Especifica las observaciones de cada dieta
> dietas <- c(rep(1,5),rep(2,5), rep(3,5), rep(4,5))
# Realiza la prueba Kruskal-Wallis
> kruskal.test(datos,dietas)
```

```
        Kruskal-Wallis rank sum test

data:  datos and dietas
Kruskal-Wallis chi-squared = 3.137, df = 3, p-value = 0.371
```

▪

1.5.2. Prueba de Friedman

Esta prueba no paramétrica ha sido muy conocida y ampliamente utilizada, no únicamente porque los cálculos son fáciles, sino que la prueba tiene buen desempeño sobre un amplio rango de condiciones. La prueba de Friedman es muy similar a la prueba de Kruskal-Wallis; la diferencia básica es la forma en que se asignan los rangos. Para la prueba de Kruskal-Wallis, los k grupos de tratamientos se combinan en un grupo grande donde se asignan los rangos en orden ascendente, mientras que en la prueba de Friedman, los rangos son asignados en orden ascendente a las k medidas para cada bloque.

Trataremos con problemas donde tendremos varias muestras relacionadas de igual tamaño. Se trata de un experimento diseñado para detectar si existe diferencia significativa en k tratamientos distintos, para $k \geq 2$. Se cuenta con *observaciones en bloques* (a veces los elementos muestrales reciben el nombre de bloques) que son grupos de k unidades experimentales similares entre ellas en algún aspecto importante. Se comparan los efectos de los tratamientos. Se trabajará con b bloques donde $b > 1$.

Datos. Se tendrán b vectores aleatorios de k entradas mutuamente independientes, es decir, $(X_{i1}, X_{i2}, \ldots, X_{ik})$ llamadas b-bloques donde $i = 1, 2, \ldots, b$, donde la entrada X_{ij} pertenece al i-ésimo bloque y j-ésimo tratamiento.

	Tratamientos			
Bloque	1	2	$\ldots$	k
1	X_{11}	X_{12}	$\cdots$	X_{1k}
2	X_{21}	X_{22}	$\cdots$	X_{2k}
$\vdots$	$\vdots$			$\vdots$
b	X_{b1}	X_{b2}	$\cdots$	X_{bk}

Asignaremos rangos a todas las observaciones de tal forma que $R(X_{ij})$ sea el rango del i-ésimo bloque y j-ésimo tratamiento, sólo que en esta prueba se asignan los rangos por *bloques* (renglones), es decir, los rangos son asignados en orden ascendente a las k observaciones para cada bloque. Cada bloque de rangos será una permutación de los enteros $1, 2, \ldots, k$.

Si hay empates se asigna nuevamente a cada valor repetido el promedio de los rangos correspondientes. Por ejemplo, si en el i-ésimo bloque se tienen que $X_{i,1} = X_{i,5} = X_{i,8} = X_{i,10}$ y les corresponden los rangos 8, 9, 10 y 11 respectivamente, el verdadero valor del rango asignado será $(8 + 9 + 10 + 11)/4$ que es igual a 9.5. Sea R_j la suma de los rangos asignados para el j-ésimo tratamiento (j-ésima columna) esto es

$$R_j = \sum_{i=1}^{b} R(X_{ij}) \quad \text{para} \quad j = 1, 2, \ldots, k$$

Rangos Tratamiento 1	Rangos Tratamiento 2	$\cdots$	Rangos Tratamiento k
$R(X_{11})$	$R(X_{12})$	$\cdots$	$R(X_{1k})$
$R(X_{21})$	$R(X_{22})$	$\cdots$	$R(X_{2k})$
$\vdots$	$\vdots$		$\vdots$
$R(X_{b1})$	$R(X_{b2})$	$\cdots$	$R(X_{bk})$
R_1	R_2	$\cdots$	R_j

Supuestos.

1. Los vectores aleatorios de k entradas son mutuamente independientes (resultados dentro de un bloque no influyen en los resultados dentro de los demás bloques)

2. Dentro de cada bloque las observaciones pueden ser clasificadas de acuerdo a cierto criterio de interés.

Hipótesis a probar. Las hipótesis que se contrastan son:

H_0: Los valores de las variables aleatorias dentro de un bloque son igualmente probables (los tratamientos son igualmente efectivos)

$$vs$$

H_1: Al menos un tratamiento tiende a mostrar observaciones más grandes que al menos uno de los otros tratamientos

Estadística de prueba. En ausencia de empates la estadística de prueba es:

$$T_1 = \frac{12}{bk(k+1)} \sum_{j=1}^{k} \left(R_j - \frac{b(k+1)}{2} \right)^2$$

Si hay empates se necesita realizar un ajuste. Sea A_1 dada por

$$A_1 = \sum_{i=1}^{b} \sum_{j=1}^{k} [R(X_{ij})]^2$$

Además, se calcula el "factor de corrección" C_1 dado por

$$C_1 = \frac{bk(k+1)^2}{4}$$

entonces la estadística T_1, ajustada en la presencia de empates, queda como

$$T_1 = \frac{(k-1)\sum_{j=1}^{k}(R_j^2 - bC_1)}{A_1 - C_1} = \frac{(k-1)\sum_{j=1}^{k}\left(R_j - \frac{b(k+1)}{2}\right)^2}{A_1 - C_1}$$

Una estadística más exacta es

$$T_2 = \frac{(b-1)T_1}{b(k-1) - T_1}$$

La distribución de T_1 (o T_2) es difícil de calcular por lo que es común utilizar su aproximación asintótica cuando el número total de valores para cada muestra $b \geq 10$ y el de tratamientos $k \geq 4$. La aproximación de T_1 es Ji-cuadrada con $k-1$ grados de libertad. Sin embargo, se recomienda utilizar la estadística T_2 en lugar de T_1, la cual se aproxima a una distribución F con $k_1 = k-1$, $k_2 = (b-1)(k-1)$ grados de libertad.

La aproximación de la distribución de la estadística T_1 por una Ji-cuadrada y la distribución de T_2 que se aproxima por una F se justifican usando el teorema central del límite. Estos resultados fueron presentados por Iman y Davenport en 1976 para T_2 como se muestra en [12] y por Friedman en 1937 para T_1 como se puede ver en [10].

Regla de decisión.

a) Rechazar H_0 al nivel de significancia α si $T_1 > \chi^2_{(k-1),1-\alpha}$ y el cuantil se busca en tablas de la distribución Ji-cuadrada.

b) Rechazar H_0 al nivel de significancia α si $T_2 > f^{1-\alpha}_{(k_1,k_2)}$ donde se busca el cuantil en tablas de la distribución F para $k_1 = k-1$ y $k_2 = (b-1)(k-1)$ ya que es la mejor aproximación conforme el valor de b crece.

Valor p. El valor p es aproximadamente la probabilidad de que una variable aleatoria Ji-cuadrada con $k-1$ grados de libertad sea mayor o igual al valor observado de la estadística T_1. En otra palabras, el valor p es la probabilidad acumulada en la cola derecha de la distribución Ji-cuadrada con $k-1$ grados de libertad

$$\text{valor } p = \mathbb{P}(T_1 \geq t_{obs})$$

Comparaciones múltiples. En caso de que se rechace H_0, es decir, que algún tratamiento tenga observaciones mayores que alguna de las otras, ahora nos interesa saber cúal de ellas es la que muestra esta situación, por lo que se compara una a una. Se consideran diferentes los tratamientos si satisfacen la siguiente desigualdad

$$|R_j - R_i| > t^{1-\alpha/2}_{k_2} \left[\frac{2(bA_1 - \sum_{j=i}^{k} R_j^2)}{(b-1)(k-1)} \right]^{1/2}$$

En términos de T_1

$$|R_j - R_i| > t^{1-\alpha/2}_{k_2} \left[\frac{(A_1 - C_1)2b}{(b-1)(k-1)} \left(1 - \frac{T_1}{b(k-1)} \right) \right]^{1/2}$$

Si no hay empates

$$A_1 = \frac{bk(k+1)(2k+1)}{6}$$

y

$$A_1 - C_1 = \frac{bk(k+1)(k-1)}{12}$$

Ejemplo 1.7. *12 estudiantes seleccionados aleatoriamente participan en un experimento de aprendizaje. El investigador construye cuatro listas de palabras. Cada una contiene 20 pares de palabras, pero se utilizaron diferentes métodos en cada lista para aparearlas. A cada estudiante se le dio una lista y se le dieron 5 minutos para estudiarla y luego se examinó su habilidad para recordar las palabras. Este procedimiento se repite para las cuatro listas para cada estudiante, el orden de las listas se rotó de un estudiante al siguiente. Los resultados del examen son como sigue (20 es perfecto):*

Estudiante:	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Lista 1:	*18*	*7*	*13*	*15*	*12*	*11*	*15*	*10*	*14*	*9*	*8*	*10*
Lista 2:	*14*	*6*	*14*	*10*	*11*	*9*	*16*	*8*	*12*	*9*	*6*	*11*
Lista 3:	*16*	*5*	*16*	*12*	*12*	*9*	*10*	*11*	*13*	*9*	*9*	*13*
Lista 4:	*20*	*10*	*17*	*14*	*18*	*16*	*14*	*16*	*15*	*10*	*14*	*16*

¿Algunas listas son más fáciles de aprender que otras? Utilice un nivel de significancia de $\alpha = 5\,\%$.

Solución: Se tiene que $b = 12$ estudiantes (bloques) y $k = 4$ listas (tratamientos) y asignando rangos suponiendo que 1 se le asigna a la lista que mejor se aprendieron y 4 a la que peor se aprendieron, tenemos la siguiente tabla de rangos:

Estudiante:	**R**	**R**	**R**	**R**	**R**	**R**	**R**	**R**	**R**	**R**	**R**	**R**	R_i
Lista 1:	2	2	4	1	2.5	2	2	3	2	3	3	4	30.5
Lista 2:	4	3	3	4	4	3.5	1	4	4	3	4	3	40.5
Lista 3:	3	4	2	3	2.5	3.5	4	2	3	3	2	2	34
Lista 4:	1	1	1	2	1	1	3	1	1	1	1	1	15
$\sum_{j=1}^{4}[R_{ij}]^2$	30	30	30	30	29.5	29.5	30	30	30	28	30	30	$A_1 = 357$

Como hay empates, tenemos que

$$C_1 = \frac{bk(k+1)^2}{4} = \frac{12(4)(5)^2}{4} = 300$$

y

$$\begin{aligned} T_1 &= \frac{(k-1)\sum_{j=1}^{k}\left(R_j - \frac{b(k+1)}{2}\right)^2}{A_1 - C_1} \\ &= \frac{3[(30.5-30)^2 + (40.5-30)^2 + (34-30)^2 + (15-30)^2]}{57} \\ &= 18.5 \end{aligned}$$

Calculando el cuantil $\chi^2_{(3),0.95}$

```
> qchisq(p=0.95, df=3, lower.tail=T)
```

```
[1] 7.815
```

Rechazamos la hipótesis nula H_0 si $18.5 > 7.814728$. En consecuencia, rechazamos H_0. Obteniendo el valor p con $t_{obs} = 18.5$

```
> pchisq(q=18.5, df=3, lower.tail=F)
```

```
[1] 0.0003468
```

el valor $p = 0.0003468294$ es menor que el nivel de significancia 0.05, entonces rechazamos la hipótesis nula H_0.
Calculando T_2 tenemos que

$$\begin{aligned} T_2 &= \frac{(b-1)T_1}{b(k-1)-T_1} \\ &= \frac{11(18.5)}{12(3)-18.5} \\ &= 11.629 \end{aligned}$$

Calculando el cuantil de la distribución F con $k_1 = 3$ y $k_2 = 33$ grados de libertad y para $1 - \alpha = 0.95$

```
# Cuantil f^0.95_(3,33)
> qf(p=0.95, df1=3, df2=33, lower.tail=T)
```

```
[1] 2.892
```

Sabemos que se rechaza H_0 si $T_2 > f^{1-\alpha}_{k_1,k_2}$, es decir, $11.629 > 2.891564$, por lo que se rechaza H_0, por lo tanto hay una lista más fácil de aprender. Sólo falta ver cúal de las listas es la que resulta más fácil aprender. Entonces, realizando las comparaciones múltiples, obteniedo el cuantil $t^{0.975}_{(33)}$

```
# Cuantil t^0.975_(33)
> qt(p=0.975, df=33, lower.tail=T)
```

```
[1] 2.035
```

Sustituyendo los valores obtenidos en

$$t_{1-\alpha/2}\left[\frac{(A_1 - C_1)2b}{(b-1)(k-1)}\left(1 - \frac{T_1}{b(k-1)}\right)\right]^{1/2}$$

tenemos que

$$\begin{aligned} &= 2.034515\left[\frac{(57)(2)(12)}{(11)(3)}\left(1 - \frac{18.5}{12(3)}\right)\right]^{1/2} \\ &= 9.133. \end{aligned}$$

La diferencia de $R_4 = 15$ con cualquier otra R_i es mayor que 9.133, entonces la lista 4 es significativamente diferente a las demás, por lo que es más fácil de

aprender que las demás listas.

En R existe la instrucción `friedman.test()` que permite obtener el resultado inmediato de la prueba de Friedman. Esta instrucción nos proporciona el valor que toma la estadística de prueba T_1, los $k-1$ grados de libertad y el valor p. La instrucción `friedman.test()` cuenta con los siguientes argumentos

```
friedman.test(datos, tratamientos, bloques)
```

donde `datos` representa todas las observaciones de las k muestras, `tratamientos` son los diferentes tratamientos y `bloques` son los diferentes bloques referentes al problema. El siguiente código proporciona la solución del ejemplo anterior.

```
# Datos ordenados por listas (tratamientos)
> resultados <- c(18,14,16,20,7,6,5,10,13,14,16,17,15,10,12,
+ 14,12,11,12,18,11,9,9,16,15,16,10,14,10,8,11,16,14,12,13,
+ 15,9,9,9,10,8,6,9,14,10,11,13,16)
# Estudiantes b = 12
estudiantes <- c(rep(1,4), rep(2,4), rep(3,4), rep(4,4),
+ rep(5,4), rep(6,4), rep(7,4), rep(8,4), rep(9,4),
+ rep(10,4), rep(11,4),rep(12,4))
# Tratamientos k = 4
> listas <- c(rep(seq(1,4,1),12))
# Realiza la prueba de Friedman
> friedman.test(resultados, listas, estudiantes)

        Friedman rank sum test

data:  resultados, listas and estudiantes
Friedman chi-squared = 18.5, df = 3, p-value = 0.0003468
```

▪

1.6. Pruebas de bondad de ajuste

En esta sección se muestran unas de las principales *pruebas de bondad de ajuste*. Por bondad de ajuste, entenderemos los métodos que examinan qué tan bien ajustan las observaciones $X_1, X_2, \ldots, X_n$ de una muestra aleatoria con una distribución dada para su población. Una prueba de bondad de ajuste usualmente requiere estudiar una muestra aleatoria de alguna distribución desconocida para probar la hipótesis nula de que la función de distribución desconocida $F_X(x)$ es en realidad una función específica $F^*(x)$, es decir, la hipótesis nula H_0 especifica completamente alguna función de distribución hipotética $F^*(x)$. Si $X_1, X_2, \ldots, X_n$ es una muestra aleatoria de una población con función de distribución $F_X(x)$ desconocida, una manera intuitiva para saber si la función hipotética $F^*(x)$ es la verdadera función de distribución de la muestra, es comparar la función de distribución empírica $S_n(x)$ con $F^*(x)$ ya que $S_n(x)$ sirve como estimador para $F_X(x)$.

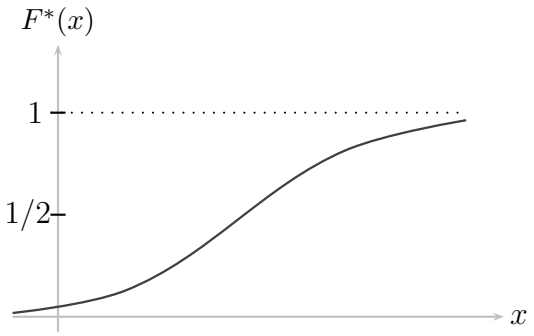


Figura 1.7: Función de distribución hipotética $F^*(x)$.

Función de distribución empírica

En la mayoría de los casos la función de distribución de una variable aleatoria X es desconocida. Una forma para estimar la función de distribución $F_X(x)$ es construyendo la *función de distribución empírica o muestral*, denotada por $S_n(x)$, la cual se basa en las estadísticas de orden para una muestra aleatoria.

Definición 1.2. *Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria. La función de distribución empírica $S_n(x)$ (f.d.e.) es una función de x definida como la proporción de valores muestrales menores o iguales que el valor específico x para cada $x \in (-\infty, \infty)$, es decir,*

$$S_n(x) = \frac{\textit{número de valores muestrales} \leq x}{n}$$

La función de distribución empírica $S_n(x)$ está dada por

$$S_n(x) = \begin{cases} 0 & \text{si} \quad x < X_{(1)}, \\ \frac{i}{n} & \text{si} \quad X_{(i)} \leq x < X_{(i+1)} \quad i = 1, \ldots, n-1, \\ 1 & \text{si} \quad x \geq X_{(n)} \end{cases} \tag{1.17}$$

donde $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ son las estadísticas de orden. Si se definen adicionalmente las estadísticas de orden $X_{(0)} = -\infty$ y $X_{(n+1)} = \infty$, la función de distribución empírica se puede escribir como

$$S_n(x) = \frac{i}{n} \quad \text{para} \quad X_{(i)} \leq x \leq X_{(i+1)} \quad i = 0, 1, \ldots, n. \tag{1.18}$$

Observación 1.5.

a) $S_n(x)$ es una función escalonada con saltos en los distintos valores muestrales ordenados.

b) Cuando más de una observación tiene el mismo valor, se dice que están empatadas o repetidas.

c) Los saltos ocurren solamente en los valores diferentes de los valores muestrales ordenados X_j y la altura del salto es igual a m/n donde m es el número de valores de los datos repetidos en X_j.

La gráfica de la distribución empírica $S_n(x)$ se muestra en la Figura (1.8).

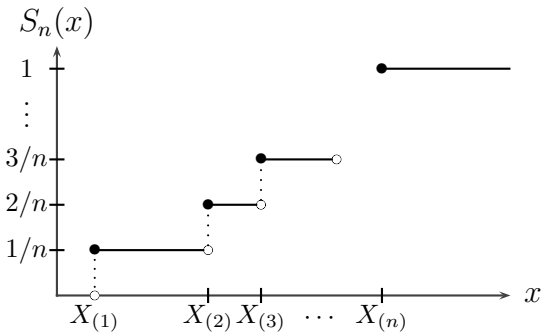


Figura 1.8: Gráfica de la función de distribución empírica $S_n(x)$.

Ejemplo 1.8. *Suponga que una muestra aleatoria de tamaño $n = 5$ está dada por 8.7, 10.4, 10.6, 11 y 11.5. La función de distribución empírica de esta muestra es:*

$$S_5(x) = \begin{cases} 0 & \text{si} \quad x < 8.7, \\ 1/5 & \text{si} \quad 8.7 \leq x < 10.4, \\ 2/5 & \text{si} \quad 10.4 \leq x < 10.6, \\ 3/5 & \text{si} \quad 10.6 \leq x < 11, \\ 4/5 & \text{si} \quad 11 \leq x < 11.5, \\ 1 & \text{si} \quad x \geq 11.5. \end{cases}$$

La gráfica de la función $S_5(x)$ se muestra en la Figura (1.9).

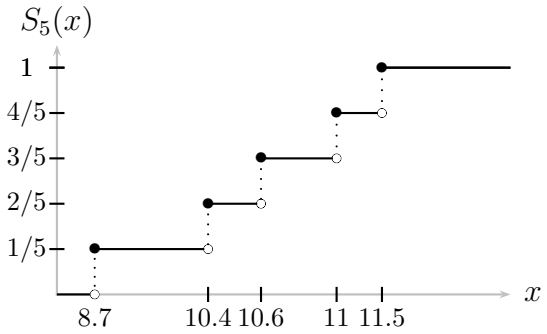
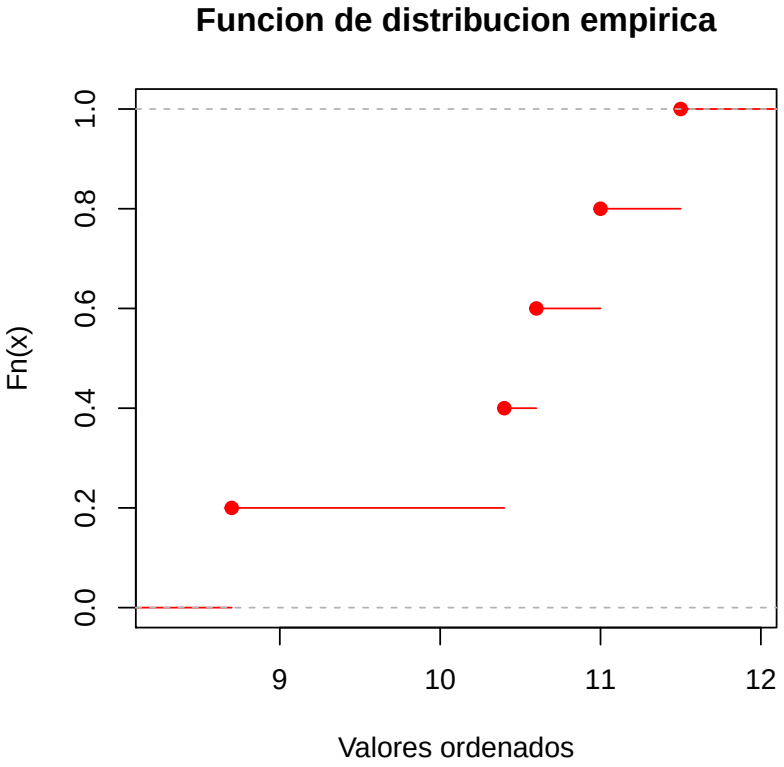


Figura 1.9: Gráfica de la función de distribución empírica $S_5(x)$.

En este caso, la altura de cada salto es igual al recíproco del tamaño de la muestra, es decir 1/5 o 0.2 ya que no se cuenta con valores empatados o repetidos.

En R existe la instrucción `ecdf()` que permite conocer la función de distribución empírica para un conjunto de datos. El siguiente código muestra cómo obtener la función y su gráfica de $S_5(x)$ del ejemplo anterior con ayuda de la función `plot()`.

```
# Vector de datos
> datos <- c(8.7, 10.4, 10.6, 11, 11.5)
# Función de distribución empírica.
> S <- ecdf(datos)
# Gráfica de la función de distribución empírica
> plot(S, col=red", main="Funcion de distribucion empiri-
ca", xlab="Valores ordenados")
```



1.6.1. Prueba Ji-cuadrada

Considere una muestra aleatoria de tamaño n de una v.a. X con función de distribución $F_X(x)$ desconocida. Para realizar la prueba Ji-cuadrada de bondad de ajuste, las observaciones de la muestra se clasifican en k categorías o clases, en el número de resultados posibles obtenidos al realizar cierto experimento o en algún esquema con la finalidad de formar una distribución de frecuencias. De esta forma, los datos se presentan organizados en una tabla de distribuciones de frecuencias de la siguiente manera. Sea

$$f_j : \text{el número de observaciones de la } j\text{-ésima categoría}$$

	Categorías					
	1	2	3	$\cdots\cdots$	k	
Frecuencias observadas	f_1	f_2	f_3	$\cdots\cdots$	f_k	Total N

En el caso de *datos cualitativos*, donde la distribución hipotética debería ser discreta, se toman las frecuencias de cada categoría o clasificaciones numéricas. Por ejemplo, en el lanzamiento de un dado, las categorías estarán dadas por el número de puntos que se obtengan; en el lanzamiento de una moneda, las posibles categorías estarán definidas como águila o sol; en las encuestas de calidad, las categorías podrían estar dadas por excelente, bueno, regular y malo. Si las observaciones son *datos cuantitativos*, las categorías estarán definidas por el investigador, por ejemplo, si la muestra consta de las estaturas en centímetros de un grupo de personas, las posibles categorías podrían ser definidas por intervalos [1.70, 1.75), [1.75, 1.80), [1.80, 1.85] y [1.85, 1.90). En este caso, la distribución de frecuencias no será única y por ello se pierde información, además, se espera que la distribución hipotética sea continua y los datos deben ser clasificados para ser analizados por la prueba Ji-cuadrada para bondad de ajuste.

Hipótesis a probar. La hipótesis nula y alternativa que se contrastan son:

$$H_0 : F_X(x) = F^*(x) \quad \forall\, x \quad vs \quad \begin{array}{l} H_1 \,:\, F_X(x) \neq F^*(x) \\ \text{(para al menos una } x\text{)} \end{array}$$

donde $F^*(x)$ es la distribución completamente específica propuesta por el investigador. Estas hipótesis también pueden escribirse de la siguiente manera:

$H_0 : \mathbb{P}(X$ pertenece a la j-ésima categoría$) = p_j$ para $j = 1, \ldots, k$

$H_1 : \mathbb{P}(X$ pertenece a la j-ésima categoría$) \neq p_j$ para al menos una clase

Procedimiento.

a) Se calcula p_j la probabilidad de que una observación de la variable aleatoria X esté en la clase j, bajo el supuesto de que la hipótesis nula H_0 es verdadera usando la distribución que se quiere probar $F^*(x)$.

b) Se calculan los valores esperados $e_j = np_j$ para $j = 1, 2, \ldots, k$ que representan las frecuencias esperadas en un total de n observaciones del experimento bajo la hipótesis nula H_0.

La prueba Ji-cuadrada compara los valores observados f_j contra los valores esperados e_j y mide las distancias para ver qué tan cercanas o qué tan lejanas se encuentran entre sí, es decir, la decisión respecto al ajuste se basa en las desviaciones $f_j - e_j$. Las correspondientes frecuencias observadas y esperadas pueden ser comparadas visualmente usando un histograma, un polígono de frecuencias o una gráfica circular.

Estadística de prueba. La estadística de prueba está dada por:

$$Q = \sum_{j=1}^{k} \frac{(f_j - e_j)^2}{e_j} \quad \text{ó} \quad Q = \sum_{j=1}^{k} \frac{f_j^2}{e_j} - n.$$

Esta estadística fue propuesta por Karl Pearson en 1900 al tomar la suma de los cuadrados de las desviaciones $f_j - e_j$ normalizadas por la frecuencia esperada e_j, la cual se aproxima asintóticamente a una distribución Ji-cuadrada con $k-1$ grados de libertad cuando el tamaño de la muestra n tiende a infinito. Un desarrollo más profundo de esta prueba Ji-cuadrada para bondad de ajuste se puede consultar en [16] o en [11].

Observación 1.6.

a) Un valor grande de Q significa que existe gran discrepacia entre las frecuencias observadas f_j y las frecuencias esperadas e_j y en consecuencia la hipótesis nula H_0 debe ser rechazada, en otras palabras, para valores grandes de Q las distribuciones $F_X(x)$ y $F^*(x)$ no son parecidas.

b) La Ji-cuadrada es una aproximación que se puede usar con confianza cuando todos los valores observados $e_j > 5$. Si esto no ocurre, se sugiere unir categorías o clases adyacentes hasta que $e_j > 5$. Si se unen clases el cuantil $\chi^2_{(k-1,1-\alpha)}$ va perdiendo grados de libertad así como cuando se estiman parámetros.

Regla de decisión. La estadística Q se compara contra el cuantil $\chi^2_{(k-1,1-\alpha)}$. Se rechaza H_0 al nivel de significancia α si:

$$Q > \chi^2_{(k-1,1-\alpha)}.$$

Si se desconocen los parámetros se estiman con la muestra aleatoria y se utiliza la regla

$$Q > \chi^2_{(k-1-r,1-\alpha)}$$

donde r es el número de parámetros desconocidos.

Valor p. El valor p es aproximadamente la probabilidad de que una variable aleatoria Ji-cuadrada con $k-1$ grados de libertad sea mayor que el valor observado de la estadística Q, o bien, si q_{obs} representa el valor observado de la estadística Q y si $X \sim \chi^2_{(k-1)}$

$$\text{Valor } p = \mathbb{P}(X > q_{obs})$$

Ejemplo 1.9. *Un dado se lanza 600 veces obteniendo los siguientes resultados*

	1	2	3	4	5	6
Frecuencias observadas	87	96	108	89	122	98

$$N = 600$$

¿El dado está balanceado? Probar con un nivel de significancia $\alpha = 10\,\%$.

Solución: Sabemos que el experimento aleatorio de lanzar un dado se traduce en una variable aleatoria $X \sim unif\{1, \ldots, 6\}$ por lo que $p_j = \dfrac{1}{6}$ y $e_j = 600(\frac{1}{6}) = 100$ para $j = 1, \ldots, 6$, se quiere probar si el dado está balanceado por lo que se tiene

$$H_0 : \mathbb{P}(X\text{caiga en la clase } j) = \frac{1}{6}, \quad j = 1, \ldots, 6.$$

Haciendo los cálculos tenemos que la estadística de prueba es:

$$Q = \frac{\sum_{j=1}^{6} f_j^2}{100} - 600 = \frac{60858}{100} - 600 = 8.58.$$

Por otro lado, tenemos que el cuantil $\chi^2_{(5,0.90)}$ es

```
> qchisq(p=0.9, df=5, lower.tail=TRUE)
```

```
[1] 9.236357
```

en consecuencia, $8.58 \ngeq 9.236357$. Por lo tanto, no se rechaza la hipótesis nula H_0 al nivel de significancia 0.1, y se puede concluir que no existe evidencia suficiente para contradecir que el dado está balanceado. Calculando el valor p se obtiene

```
# Valor p
> pchisq(8.58, df=5, lower.tail=FALSE)
```

```
[1] 0.1270355
```

En el paquete estadístico `R` existe la instrucción `chisq.test()` la cual lleva a cabo una prueba de la Ji-cuadrada de bondad de ajuste. Sin embargo, existen varias modalidades de esta instrucción para analizar distintas pruebas basadas en la distribución Ji-cuadrada, por ejemplo, la prueba de independencia. Las siguientes instrucciones proporcionan la solución del ejemplo anterior utilizando `chisq.test()`.

```
# Vector de datos (frecuencias observadas)
> frecuencias <- c(87, 96, 108, 89, 122, 98)
# Matrix de frecuencias observadas
> tabla<-matrix(frecuencias, nrow=1)
# Vector de categorias
> categorias <- c("1", "2", "3", "4", "5", "6")
# Asigna nombres a la matrix de cada categoria
> dimnames(tabla) <- list(NULL, categorias)
# Muestra la matrix con cada categoria
> tabla
```

```
     1  2   3  4   5  6
[1,] 87 96 108 89 122 98
```

```
# Prueba Ji-cuadrada
> chisq.test(tabla)
```

```
        Chi-squared test for given probabilities

data:  tabla
X-squared = 8.58, df = 5, p-value = 0.127
```

Ejemplo 1.10. *Una computadora genera 300 números durante diez horas, donde el valor esperado de números que se generan por hora es de 30. La siguiente tabla resume cómo se obtienen los números:*

Hora	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*Total*
Números observados	*22*	*28*	*41*	*35*	*19*	*25*	*25*	*40*	*30*	*35*	*300*
Esperados	*30*	*30*	*30*	*30*	*30*	*30*	*30*	*30*	*30*	*30*	*300*

Se desea probar con un nivel de significancia $\alpha = 5\,\%$ *si la cantidad de números que se generan en cada hora se comportan uniformemente.*

Solución: Se desea probar el siguiente contraste de hipótesis

H_0: Los números provienen de una distribución uniforme discreta	*vs*	H_1: Los números no provienen de una distribución uniforme discreta.

Nótese que en este caso se tiene que $p_j = \frac{1}{10}$, donde $e_j = 30$ para $j = 1, \ldots, 10$. Calculando la estadística de prueba

$$\begin{aligned} Q &= \sum_{j=1}^{10} \frac{f_j^2}{e_j} - n \\ &= \left(\frac{22^2}{30} + \frac{28^2}{30} + \cdots + \frac{30^2}{30} + \frac{35^2}{30} \right) - 300 \\ &= 317 - 300 \\ &= 17 \end{aligned}$$

entonces $Q = 17$ y para $k = 10$, calculando el cuantil $\chi^2_{(9,0.95)}$

```
> qchisq(p=0.95, df=9, lower.tail=TRUE)
```

```
[1] 16.91898
```

Se rechaza H_0 si $Q > \chi^2_{(9,0.95)}$, es decir, $17 > 16.91898$, en consecuencia se rechaza la hipótesis H_0. Por lo tanto, podemos decir que los números no provienen de una distribución uniforme discreta.

▪

Ejemplo 1.11. *Se ha tomado una muestra aleatoria de 50 baterías y se ha registrado su duración en años. Estos resultados se han agrupado en 6 clases como se muestra en la siguiente tabla:*

i	*Duración*	*Frecuencia* (f_i)
1	$[1, 2)$	*10*
2	$[2, 3)$	*11*
3	$[3, 4)$	*11*
4	$[4, 5)$	*13*
5	$[5, 6)$	*3*
6	$[6, 7)$	*2*

Se va a verificar con un nivel de significancia del α =5 % que la duración en años de las baterías producidas sigue una distribución normal con los parámetros estimados $\mu = 3.38$ y $\sigma = 1.366$.

Solución: Sea X la variable aleatoria continua definida como la duración en años de las baterías. Se desea probar el siguiente contraste de hipótesis

H_0: La duración en años de las baterías $X \sim N(3.38, 1.366^2)$ *vs* H_1: La duración de las baterías en años X siguen una distribución distinta a la normal.

Calculando las probabilidades p_i para $i = 1, \ldots, 6$, correspondientes a cada clase:

```
# p_1 = P(X < 2)
> pnorm(2,mean=3.38,sd=1.366)
```

```
[1] 0.1562
```

```
# p_2 = P(2 ≤ X < 3)
> pnorm(3,3.38,1.366)-pnorm(2,3.38,1.366)
```

```
[1] 0.2342
```

```
# p_3 = P(3 ≤ X < 4)
> pnorm(4,3.38,1.366)-pnorm(3,3.38,1.366)
```

```
[1] 0.2846
```

```
# p_4 = P(4 ≤ X < 5)
> pnorm(5,3.38,1.366)-pnorm(4,3.38,1.366)
```

```
[1] 0.2071
```

```
# p_5 = P(5 ≤ X < 6)
> pnorm(6,3.38,1.366)-pnorm(5,3.38,1.366)
```

```
[1] 0.09027
```

```
# p_6 = P(6 ≤ X < 7)
> pnorm(7,3.38,1.366)-pnorm(6,3.38,1.366)
```

```
[1] 0.02353
```

Calculando las frecuencias esperadas e_j para $j = 1, \ldots, 6$ tenemos:

$e_1 = 0.1562 \times 50 = 7.81$
$e_2 = 0.2342 \times 50 = 11.71$
$e_3 = 0.2846 \times 50 = 14.23$
$e_4 = 0.2071 \times 50 = 10.355$
$e_5 = 0.09027 \times 50 = 4.5135$
$e_6 = 0.02353 \times 50 = 1.1765$

Nótese que $\sum_{j=1}^{6} e_i \approx 50$ debido a que se pierden cifras decimales. Es necesario que se cumpla la condición $e_j > 5$ para toda j, por lo que se deben unir clases adyacentes, es decir, se unen las clases 5 y 6, y como resultado tenemos $k = 5$ clases como se muestra a continuación:

i	frecuencias observadas f_i	frecuencias esperadas e_i
1	10	7.81
2	11	11.71
3	11	14.23
4	13	10.355
5	5	5.69

Haciendo los cálculos utilizando esta nueva tabla, tenemos que la estadística de prueba es:

$$Q = \sum_{j=1}^{5} \frac{(f_j - e_j)^2}{e_j} = 2.15$$

entonces Q =2.15, así la estadística de prueba se compara con el cuantil $\chi^2_{(k-1-r,1-\alpha)}$, o bien, para $k = 5$ y los parámetros estimados $r = 2$, entonces, el cuantil $\chi^2_{(2,0.95)}$ es

```
> qchisq(p=0.95,df=2,lower.tail=TRUE)
```

```
[1] 5.991
```

Se rechaza H_0 si $Q > \chi^2_{(2,0.95)}$. Como 2.15$\not>$5.991, entonces, no se rechaza H_0, en consecuencia, se dice que no hay evidencia suficiente para rechazar H_0 al nivel de significancia del 5 %, por lo tanto la distribución normal con media 3.38 y desviación estándar 1.366 da un ajuste razonable a los datos. Calculando el valor p con $k = 4$ grados de libertad se obtiene

```
> pchisq(2.15,df=4,lower.tail=FALSE)
```

```
[1] 0.7082
```

Como el valor p es mayor que el nivel de significancia, entonces, no se rechaza H_0. ▪

1.6.2. Prueba de Kolmogorov-Smirnov

La prueba de Kolmogorov-Smirnov de una muestra consiste en estimar la función de distribución $F_X(x)$ basándose en la diferencia entre la función de distribución hipotética $F^*(x)$ y la función de distribución empírica $S_n(x)$ para todo valor de x. La estadística que proporciona Kolmogorov en 1933 consta de tomar la distancia vertical más grande entre la gráfica de $S_n(x)$ y $F_X(x)$ dada por

$$D = \sup_x |S_n(x) - F_X(x)| \tag{1.19}$$

la cual es, para cualquier n, una medida razonable para la estimación. La estadística D es conocida como *estadística de una muestra de Kolmogorov-Smirnov*, y es particularmente muy usada en inferencia estadística no paramétrica porque la distribución de probabilidad de D no depende de $F_X(x)$ siempre y cuando $F_X(x)$ sea continua. Por consiguiente, a D se le conoce como estadística de distribución libre.

Las desviaciones definidas como

$$D^+ = \sup_x [S_n(x) - F_X(x)] \quad \text{y} \quad D^- = \sup_x [F_X(x) - S_n(x)] \tag{1.20}$$

se conocen como *estadísticas de una cola de Kolmogorov-Smirnov.*

Teorema 1.1. *Las estadísticas D, D^+ y D^- son completamente de distribución libre para cualquier $F_X(x)$ continua.*

Demostración.

$$D = \sup_x |S_n(x) - F_X(x)| = \max_x \{D^+, D^-\}.$$

De la definición (1.18) de la función de distribución empírica se tiene que

$$\begin{aligned} D^+ &= \sup_x [S_n(x) - F_X(x)] \\ &= \max_{0\le i\le n} \quad \sup_{X_{(i)}\le x\le X_{(i+1)}} [S_n(x) - F_X(x)] \\ &= \max_{0\le i\le n} \quad \sup_{X_{(i)}\le x\le X_{(i+1)}} \left[\frac{i}{n} - F_X(x)\right] \\ &= \max_{0\le i\le n} \left[\frac{i}{n} - \inf_{X_{(i)}\le x\le X_{(i+1)}} F_X(x)\right] \\ &= \max_{0\le i\le n} \left[\frac{i}{n} - F_X(X_{(i)})\right] \end{aligned}$$

por lo tanto

$$D^+ = \max\left\{\max_{1\le i\le n}\left[\frac{i}{n} - F_X(X_{(i)})\right], 0\right\} \tag{1.21}$$

Similarmente

$$D^- = \max\left\{\max_{1\le i\le n}\left[F_X(X_{(i)}) - \frac{i-1}{n}\right], 0\right\} \tag{1.22}$$

así que

$$D = \text{máx}\left\{ \underset{1\leq i\leq n}{\text{máx}} \left[\frac{i}{n} - F_X(X_{(i)}) \right], \underset{1\leq i\leq n}{\text{máx}} \left[F_X(X_{(i)}) - \frac{i-1}{n} \right], 0 \right\}. \tag{1.23}$$

La distribución de probabilidad de D, D^+ y D^- dependen sólo de las variables aleatorias $F_X(X_{(i)})$, para $i = 1, \ldots, n$. Éstas son las estadísticas de orden de la distribución uniforme en $(0, 1)$, independientemente de la distribución original $F_X(x)$ siempre que sea continua, debido a la transformación integral de probabilidad[1]. Así la distribución de D, D^+ y D^- es independiente de $F_X(x)$. ▪

Datos. Los datos consisten de una muestra aleatoria $X_1, X_2, \ldots, X_n$ que proviene de una función de distribución $F_X(x)$ desconocida.

Supuestos.

1. La muestra es una muestra aleatoria.

Sea $F^*(x)$ la función de distribución completamente especificada con la que se propondrán las hipótesis H_0. La estadística de prueba y la regla de decisión se define para cada uno de los siguientes casos:

Caso A (Dos colas)

$$H_0 : F(x) = F^*(x) \quad \forall x \in (-\infty, \infty)$$

$$vs$$

$$H_1 : F(x) \neq F^*(x) \quad \text{para alguna } x$$

La estadística de prueba D es la distancia vertical más grande entre $S_n(x)$ y $F^*(x)$, dada por

$$D = \sup_x |S_n(x) - F^*(x)| \tag{1.24}$$

la cual se lee como, "D es el supremo, sobre todos los valores de x, del valor absoluto de la diferencia entre la función de distribución empírica $S_n(x)$ y la distribución hipotética $F^*(x)$". El valor de la estadística D puede calcularse usando (1.23) si todas las n observaciones no presentan empates. Sin embargo, la siguiente expresión se considera más fácil para el cálculo y se aplica cuando existen empates en las observaciones.

$$D = \sup_x |S_n(x) - F^*(x)| = \underset{x}{\text{máx}} \{|S_n(x) - F^*(x)|, |S_n(x - \epsilon) - F^*(x)|\}$$

donde $\epsilon > 0$.

Regla de decisión. Se rechaza H_0 al nivel de significancia α si $D > w^n_{1-\alpha}$, donde $w^n_{1-\alpha}$ es el cuantil que acumula $(1-\alpha)\,\%$ de probabilidad y dicho cuantil se busca en la tabla C.8 para valores de $n \leq 40$.

[1]Para encontrar más información sobre la transformación integral de probabilidad puede consultar [16].

Caso B (Cola inferior)

$$H_0 : F(x) \geq F^*(x) \quad \forall x \in (-\infty, \infty)$$

$$vs$$

$$H_1 : F(x) < F^*(x) \quad \text{para alguna } x$$

La estadística de prueba, denotada por D^-, es igual a la distancia vertical más grande que se alcanza cuando la función de distribución hipotética $F^*(x)$ se encuentra por encima de la función de distribución empírica $S_n(x)$ y se puede calcular a partir de (1.22) como

$$D^- = \sup_x [F^*(x) - S_n(x)] = \max_x \{F^*(x) - S_n(x - \epsilon)\} \tag{1.25}$$

Regla de decisión. Se rechaza H_0 al nivel de siginificancia α si $D^- > w^n_{1-\alpha}$, donde nuevamente $w^n_{1-\alpha}$ es el cuantil que acumula $(1-\alpha)\,\%$ de probabilidad y se busca en tablas del estadístico de Kolmogorov para valores de $n \leq 40$.

Caso C (Cola superior)

$$H_0 : F(x) \leq F^*(x) \quad \forall x \in (-\infty, \infty)$$

$$vs$$

$$H_1 : F(x) > F^*(x) \quad \text{para algun } x$$

La estadística de prueba, denotada por D^+, está definida como la distancia vertical más grande que se alcanza cuando la función de distribución empírica $S_n(x)$ está por encima de la función de distribución hipotética $F^*(x)$ y se puede calcular a partir de (1.21) como

$$D^+ = \sup_x [S_n(x) - F^*(x)] = \max_x \{S_n(x) - F^*(x)\} \tag{1.26}$$

Regla de decisión. Se rechaza H_0 al nivel de significancia α si $D^+ > w^n_{1-\alpha}$, donde nuevamente $w^n_{1-\alpha}$ es el cuantil que acumula $(1-\alpha)\,\%$ de probabilidad y se busca en tablas del estadístico de Kolmogorov para valores de $n \leq 40$. En la Figura (1.10) se muestran gráficamente las estadísticas de prueba D^- y D^+.

Valor p. El valor p de esta prueba para el caso de una cola, se puede calcular de la siguiente manera

$$\text{valor } p = d \sum_{j=0}^{[n(1-d)]} \binom{n}{j} \left(1 - d - \frac{j}{n}\right)^{n-j} \left(d + \frac{j}{n}\right)^{j-1} \tag{1.27}$$

donde $[n(1-d)]$ es el entero más grande, menor o igual a $n(1-d)$ y d es el valor de la estadística D, D^+ ó D^-. Para el caso de una prueba de dos colas, el valor p es igual a dos veces el valor p de una cola.

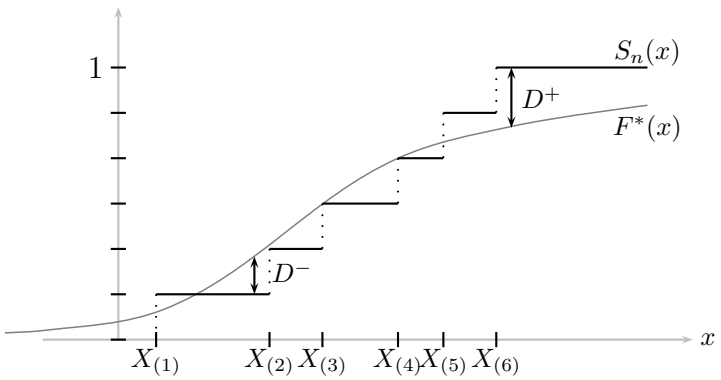


Figura 1.10: Estadísticas de prueba D^- y D^+.

Cuando $F(x)$ es continua y la hipótesis H_0 es verdadera la distribución exacta de D^+ y D^- está dada por

$$G(x) = 1 - x \sum_{j=0}^{[n(1-x)]} \binom{n}{j} \left(1 - x - \frac{j}{n}\right)^{n-j} \left(x + \frac{j}{n}\right)^{j-1}.$$

La distribución aproximada de D es $\mathbb{P}(T \leq x) = [G(x)]^2$ ya que D es menor a x sólo cuando D^+ y D^- son menores a x.
La función de distribución asintótica ($n \to \infty$) de $\sqrt{n}D^+$ y $\sqrt{n}D^-$ está dada por

$$H(x) = \lim_{n\to\infty} G\left(\frac{x}{\sqrt{n}}\right) = 1 - e^{-2x^2}$$

y es con esta distribución asintótica que se estiman los cuantiles para las estadísticas de la prueba Kolmogorov para valores de $n > 40$. Para mayor información sobre estos resultados, puede consultalse [3].

La derivación de la distribución de la estadística de Kolmogorov está fuera del alcance de este texto. La distribución de la estadística D, en el caso de dos colas, fue encontrada por Kolmogorov en 1933 y este resultado puede consultarse en [13]. Además, esta estadística fue tabulada por Smirnov en 1948 y puede consultarse en [24]. La distribución asintótica de las estadísticas D^+ y D^- para una cola fueron obtenidas por Smirnov en 1939 y se puede consultar en [23]. Por esta razón, en muchos textos esta prueba es conocida como *Prueba de bondad de ajuste de Kolmogorov-Smirnov.*

Ejemplo 1.12. *La emisión de óxido nitroso de automóviles a partir del modelo del año pasado se ha medido para mil automóviles y se encontró que se aproxima a una distribución normal con media 5.6 y desviación estándar 1.2. Se tomó una muestra de 12 automóviles modelo de este año que han sido probados obteniendo los siguientes resultados*

4.8, 6.6, 5.7, 5.5, 6.0, 5.9, 5.8, 6.2, 6.1, 6.3, 6.5, 5.0

¿Tiene el modelo de este año la misma distribución que el modelo del año pasado?

Solución: Tenemos que la función de distribución hipotética $F^*(x)$ se distribuye normal con media $\mu = 5.6$ y desviación estándar $\sigma = 1.2$, y nuestro interés es saber si la emisión de óxido nitroso de los automóviles modelo de este año tienen la misma distribución que los del modelo del año pasado, por lo que contrastaremos las hipótesis correspondiente al caso de dos colas de la siguiente manera:

$$H_0\text{: La distribución de } F(x) \text{ es normal con } \mu = 5.6 \text{ y } \sigma = 1.2 \quad \textit{vs} \quad H_1\text{: } F(x) \text{ no se distribuye normal para alguna } x.$$

Sabemos que para este caso, la estadística de prueba está dada por

$$D = \text{máx}\{|S_n(x) - F^*(x)|, |S_n(x-\epsilon) - F^*(x)|\}.$$

Usando `R` podemos calcular los valores de la función de distribución $F^*(x)$ normal con media $\mu = 5.6$ y desviación estándar $\sigma = 1.2^2$ en cada valor x de la siguiente forma

```
# Vector de datos
> x <- c(4.8, 6.6, 5.7, 5.5, 6.0, 5.9, 5.8, 6.1, 6.2, 6.3,
6.5, 5.0)
# Vector de datos ordenado en forma ascendente
> xi <- sort(x)
# Valores de la distribución N(5.6,1.2^2)
> pnorm(q = xi, mean = 5.6, sd = 1.2)
```

```
[1] 0.2525 0.3085 0.4668 0.5332 0.5662 0.5987 0.6306 0.6615
[9] 0.6915 0.7202 0.7734 0.7977
```

Usando los valores anteriores podemos obtener la siguiente tabla:

$X_{(i)}$	$S_n(x)$	$F^*(x)$	$\lvert S_n(x) - F^*(x)\rvert$	$\lvert S_n(x-\epsilon) - F^*(x)\rvert$
4.8	1/12	0.2525	0.1692	0.2525
5.0	2/12	0.3085	0.1418	0.2252
5.5	3/12	0.4668	**0.2168**	**0.3001**
5.7	4/12	0.5332	0.1999	0.2832
5.8	5/12	0.5662	0.1495	0.2329
5.9	6/12	0.5987	0.0987	0.1820
6.0	7/12	0.6306	0.0473	0.1306
6.1	8/12	0.6615	0.0052	0.0782
6.2	9/12	0.6915	0.0585	0.0248
6.3	10/12	0.7202	0.1131	0.0298
6.5	11/12	0.7734	0.1433	0.0599
6.6	1	0.7977	0.2023	0.1190

La estadística de prueba es

$$D = \text{máx}\{0.2168, 0.3001\}$$

Buscando en tablas de cuantiles de la distribución de Kolmogorov para $n = 12$ y considerando un nivel de significancia $\alpha = 5\,\%$ tenemos $w^{12}_{0.95} = 0.375$. Sabemos que se rechaza H_0 si $D > w^n_{1-\alpha}$, o bien, $0.3001 \not> 0.375$ entonces no se rechaza la hipótesis H_0, por lo tanto, se dice que no hay evidencia suficiente para contradecir que la distribución de las emisiones de óxido nitroso de los automoviles de modelo de este año es igual a la distribución de los automóviles modelo del año pasado.
En `R` existe el comando `ks.test()` el cual permite obtener un resultado inmediato de la prueba de Kolmogorov Smirnov. Este comando proporciona el valor que toma la estadística de prueba junto con el valor p, el cual no es tan sencillo de calcular a partir de la ecuación (1.27). El comando `ks.test()` cuenta con los siguientes argumentos

```
ks.test(x, y, mean=, sd=, alternative=)
```

donde `x` representa todas las observaciones de la muestra (vector de datos), `y` especifica la función de distribución hipotética $F^*(x)$, `mean` representa el valor de la media, `sd` el valor de la desviación estándar de $F^*(x)$ (si es el caso), y `alternative=` indica si la prueba es de dos colas, cola inferior o superior, es decir, el argumento `alternative` puede ser

a) `alternative="two.sided"`,

b) `alternative="less"`,

c) `alternative="greater"`.

Si el argumento `alternative=` no se especifica con ninguna de las tres formas anteriores `R` por default realizará una prueba de dos colas. En el siguiente código se realiza la prueba de Kolmogorov para los datos del ejercicio anterior.

```
# Vector de datos
> datos <- c(4.8, 6.6 ,5.7 ,5.5, 6.0, 5.9, 5.8, 6.1, 6.2,
6.3, 6.5, 5.0)
# Prueba Kolmogorov-Smirnov
> ks.test(datos ,y ="pnorm", 5.6, 1.2, alternati-
ve="two.sided")
```

```
        One-sample Kolmogorov-Smirnov test

data:  datos
D = 0.3001, p-value = 0.187
alternative hypothesis: two-sided
```

■

1.6.3. Prueba Lilliefors para normalidad

En esta sección consideraremos el problema de una prueba de bondad de ajuste para la distribución normal cuando la media y varianza no son especificadas. Este problema es muy importante en la práctica porque el supuesto de una distribución normal general con media μ y desviación estándar σ desconocidas es muy común e implica la estimación de estos parámetros a partir de la muestra aleatoria. Esta prueba fue presentada por Lilliefors en 1967 la cual es una modificación de la prueba de Kolmogorov-Smirnov donde la hipótesis nula especifica que la población corresponde a una distribución normal con media y desviación estándar desconocidas. Como antes, la estadística de dos colas de Kolmogorov-Smirnov es definida como

$$D_1 = \sup_x |S_n(x) - F^*(x)|$$

donde $F(x)$ se calcula como la distribución normal estándar acumulada $\Phi(Z)$, con $Z = (X - \bar{X})/\widehat{S}$ para cada observación X de la muestra. La variable aleatoria $\bar{X}$ es la media de la muestra de n observaciones, y $\widehat{S}$ en el denominador es el estimador insesgado de σ calculado con $n - 1$.

Datos. Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de tamaño n de una población con función de distibución desconocida $F(x)$. Se calculan la media y la desviación estándar muestrales, o bien, los estimadores insesgados de μ y σ respectivamente.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{y} \qquad \widehat{S} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Calculamos los valores estandarizados de la muestra definidos como

$$Z_i = \frac{X_i - \bar{X}}{\widehat{S}} \qquad \text{para } i = 1, 2, \ldots, n.$$

Como se ha mencionado, esta prueba es como la prueba de Kolmogorov-Smirnov sólo que la función de distribución empírica $S_n(x)$ se calcula a partir de los valores estandarizados Z_i y no de los originales X_i. Así, la estadística de prueba se calcula sobre los valores estandarizados de la muestra.

Supuestos.

1. La muestra $X_1, X_2, \ldots, X_n$ es una muestra aleatoria.

Hipótesis a probar. La hipótesis nula y la alternativa que se contrastan son:

H_0: La muestra aleatoria proviene de una población con distribución normal, con media y varianza desconocidas.

vs

H_1: La muestra aleatoria proviene de una población con distribución distinta a la de una normal.

Estadística de prueba. La estadística de prueba está dada por la máxima distancia vertical

$$D_1 = \sup_z |S_n(z) - F^*(z)|$$

donde $S_n(z)$ es la función de distribución empírica de la muestra estandarizada (de los valores Z_i, $i = 1, \ldots, n$) y $F^*(z)$ es la función de distribución normal estándar $\Phi(Z)$.

Regla de decisión. Rechazar H_0 al nivel de significancia α si:

$$D_1 > w^n_{1-\alpha}$$

donde $w^n_{1-\alpha}$ es el cuantil que acumula $(1-\alpha)\,\%$ de probabilidad y se busca en tablas de cuantiles de la estadística de Lilliefors. En otras palabras, se rechaza la hipótesis nula H_0 si la estadística de prueba D_1 excede el cuantil $1-\alpha$ de la distribución de D_1 que se presentan el la tabla C.9.

Valor p. El valor p se calcula a partir de la ecuación (1.27) como en el caso de la prueba de Kolmogorov-Smirnov.

Ejemplo 1.13. *Quince estudiantes de primer año tenían puntuaciones de rendimiento de la siguiente manera*

481	*620*	*642*	*515*	*740*
562	*395*	*615*	*596*	*618*
525	*584*	*540*	*580*	*598*

Pruebe la normalidad usando la prueba de Lilliefors usando un nivel de significancia $\alpha = 5\,\%$.

Solución: Tenemos que el tamaño de la muestra es de $n = 15$. Usando `R` se calcula $\bar{X}$ y $\widehat{S}$ de la siguiente manera:

```
# Vector de datos
> a <- c(481, 562, 525, 620, 395, 584, 642, 615, 540, 515,
596, 580, 740, 618, 598)
# Valor de X̄
> mean(a)
```

```
[1] 574.1
```

```
# Valor de Ŝ
> sd(x)
```

```
[1] 78.82
```

Ordenando la muestra, obteniendo los valores estandarizados, sus respectivos valores de la distribución normal estándar y los valores de la distribución empírica para cada uno de ellos, tenemos la siguiente tabla:

$X_{(i)}$	$Z_{(i)} = \frac{X_{(i)} - \bar{X}}{\widehat{S}}$	$\Phi(Z_{(i)})$	$S_n(x)$	$\lvert S_n(x) - \Phi(Z_{(i)})\rvert$	$\lvert S_n(x-\epsilon) - \Phi(Z_{(i)})\rvert$
395	-2.2718	0.0115	1/15	0.0552	0.0115
481	-1.1807	0.1189	2/15	0.0144	0.0522
515	-0.7494	0.2268	3/15	0.0268	0.0935
525	-0.6225	0.2668	4/15	0.0001	0.0668
540	-0.4322	0.3328	5/15	0.0005	0.0661
562	-0.1531	0.4392	6/15	0.0392	0.1059
580	0.0753	0.5300	7/15	0.0633	**0.1300**
584	0.1260	0.5501	8/15	0.0168	0.0834
596	0.2783	0.6096	9/15	0.0096	0.0763
598	0.3036	0.6193	10/15	0.0474	0.0193
615	0.5193	0.6982	11/15	0.0351	0.0315
618	0.5574	0.7114	12/15	0.0886	0.0219
620	0.5827	0.7200	13/15	**0.1467**	0.0800
642	0.8619	0.8056	14/15	0.1277	0.0611
740	2.1052	0.9824	1	0.0176	0.0491

Sabemos que la estadística $D_1 = \text{máx}\{0{,}1467, 0{,}1300\}$, por lo que $D_1 = 0.1467$. Buscando en tablas de la estadística de Lilliefors tenemos que $w^{15}_{0.95} = 0.219$ y la regla de decisión está dada por $D_1 > w^n_{1-\alpha}$, es decir, $0.1467 \not> 0.219$. Por lo que no se rechaza la hipótesis H_0, por lo tanto la muestra de puntuaciones de los estudiantes proviene de una distribución normal con media y varianza desconocida.

La instrucción que ayuda a solucionar este tipo de problemas usando el paquete estadístico `R` es `lillie.test()`. Este comando permite obtener un resultado inmediato de la prueba de Lillefors para normalidad, arrojando el valor de la estadística D_1 y el valor p de la prueba. El comando `lillie.test()` solo cuanta con un argumento, el cual consta de un vector `x` de valores, además es necesario instalar y cargar la biblioteca `nortest` para poder realizar esta prueba en `R`.

Aunque la estadística de prueba obtenida del comando `lillie.test(x)` es la misma que la obtenida a partir de la instrucción `ks.test(x, "pnorm", mean(x), sd(x))`, no es correcta para utilizar el valor p de este último para la hipótesis compuesta de normalidad con media y varianza desconocidas, ya que la distribución de la estadística de prueba es diferente cuando se estiman los parámetros. El siguiente código proporciona la solución del ejemplo anterior.

```
# Vector de datos
> a <- c(481, 562, 525, 620, 395, 584, 642, 615, 540, 515,
596, 580, 740, 618, 598)
# Carga la biblioteca nortest
> library(nortest)
# Realiza la prueba de Lilliefors para normalidad
> lillie.test(a)

        Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data:  a
D = 0.1467, p-value = 0.5201
```

El siguiente código genera una función llamada `lilliefors.normal.test` la cual calcula el valor de la estadística D_1.

```
# Función para calcular la estadística D1
> lilliefors.normal.test <- function(x){
+ n <- length(x)
+ i <- 1:n
+ xi <- sort(x)
+ zi <- (xi-mean(x))/sd(x)
+ fi <- pnorm(zi,0,1)
+ t1 <- (i / n) - fi
+ t2 <- fi - ((i - 1)/n)
+ D <- max(abs(t1),abs(t2))
+ return(D)
}
# Estadística D usando el vector de datos a
> lilliefors.normal.test(a)
```

```
[1] 0.1467
```

▪

1.6.4. Prueba Lilliefors para la distribución exponencial

En la práctica, otro problema importante de bondad de ajuste es la prueba para la distribución exponencial con media no especificada. Este problema es importante porque el supuesto de una distribución exponencial con media desconocida tiene muchas aplicaciones, particularmente donde las variables aleatorias bajo estudio representan el tiempo de espera o el tiempo en que ocurre cierto evento. Lilliefors en 1969 desarrolló una prueba análoga a la prueba de Kolmogorov-Smirnov y dió una tabla de valores críticos basados en simulaciones Monte Carlo. Como en el caso de normalidad con parámetros desconocidos, la estadística de dos colas de Kolmogorov-Smirnov se define como

$$D_2 = \sup_x |S_n(x) - F^*(x)|.$$

En este caso, $F^*(x)$ se calcula como $1-e^{-x/\bar{x}}$, donde $\bar{x}$ es la media muestral. De esta forma, se calcula los valores $z = x/\bar{x}$ para cada observación de la muestra y podemos escribir la función como $F^*(z) = 1 - e^{-z}$.

Datos. Los datos consisten en una muestra aleatoria $X_1, X_2, \ldots, X_n$ de una función de distribución desconocida $F(x)$. Se calcula la media muestral

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

y los valores

$$Z_i = \frac{X_i}{\bar{X}} \qquad \text{para } i = 1, 2, \ldots, n.$$

La estadística de prueba se calcula con los valores Z_i de la muestra.

Supuestos.

1. La muestra $X_1, X_2, \ldots, X_n$ es una muestra aleatoria.

Hipótesis a probar. Las hipótesis que se contrastan son:

H_0: La muestra aleatoria proviene de una población con función de distribución exponencial

$$F(x) = \begin{cases} 1 - e^{-x/\theta} & \text{para } x > 0 \\ 0 & \text{e.o.c} \end{cases}$$

donde θ es un parámetro desconocido.

$$vs$$

H_1: La muestra aleatoria proviene de una población con función de distribución distinta a la exponencial.

Estadística de prueba. La estadística de prueba está dada por la máxima distancia vertical

$$D_2 = \sup_x |S_n(z) - F^*(z)|$$

donde $S_n(z)$ es la función de distribución empírica y $F^*(z) = 1 - e^{-z}$, las cuales se basan en los valores z_i para $i = 1, \ldots, n$.

Regla de decisión. Se rechaza H_0 al nivel de significancia α si:

$$D_2 > w^n_{1-\alpha}$$

donde $w^n_{1-\alpha}$ es el cuantil que acumula el $(1-\alpha)\,\%$ de probabilidad y se busca en la tabla C.10 de la estadística D_2 de prueba de Lilliefors para la distribución exponencial.

Ejemplo 1.14. *Un conductor de autos de carreras está probando un nuevo modelo y se miden los tiempos que tarda en llegar a cada indicación de la pista, los cuales son los siguientes:*

1.087	*0.788*	*0.192*	*0.503*	*0.752*
0.775	*0.431*	*0.967*	*1.757*	*0.688*

¿ Los datos obtenidos, provienen de una distribución exponencial? Probar con un nivel de significancia de $\alpha = 5\,\%$.

Solución: Queremos probar las hipótesis

H_0 : La muestra proviene de una distribución exponencial.	*vs*	H_1 : La muestra proviene de una distribución diferente a una exponencial.

El tamaño de la muestra es $n = 10$. Usando `R` calculamos el valor de $\bar{x}$ de la siguiente manera:

```
# Vector de datos
> b <- c(1.087, 0.788, 0.192, 0.503, 0.752, 0.775,
0.431, 0.967, 1.757, 0.688)
> mean(b)
```

```
[1] 0.794
```

Ordenando, calculando los valores z_i para $i = 1, \ldots, 10$ y los valores de $F^*(z)$ tenemos la siguiente tabla:

$x_{(i)}$	$z_{(i)}$	$S_n(z)$	$F^*(z) = 1 - e^{-z_i}$	$\lvert S_n(z) - F^*(z)\rvert$	$\lvert S_n(z-\epsilon) - F^*(z)\rvert$
0.192	0.242	0.1	0.215	0.115	0.215
0.431	0.543	0.2	0.419	0.219	**0.319**
0.503	0.634	0.3	0.469	0.169	0.269
0.688	0.866	0.4	0.579	0.179	0.279
0.752	0.947	0.5	0.612	0.112	0.212
0.775	0.976	0.6	0.623	0.023	0.123
0.788	0.992	0.7	0.629	0.071	0.029
0.967	1.218	0.8	0.704	0.096	0.004
1.087	1.369	0.9	0.746	0.154	0.054
1.757	2.213	1	0.891	0.109	0.009

Tenemos que la estadística de prueba está dada por $D_2 = \sup_x |F^*(z) - S_n(z)|$, entonces $D_2 = 0.319$. Buscando en la tabla C.10 de cuantiles para la estadística D_2 de Lilliefors para la distribución exponencial tenemos que $w^{10}_{0.95}$ =0.3244. Sabemos que se rechaza la hipótesis H_0 al nivel de significancia α si $D_2 > w^{10}_{0.95}$, pero como $0.319 \not> 0.3244$, entonces no rechazamos H_0. Por lo tanto, no existe evidencia suficiente para contradecir que los datos provienen de una distribución exponencial.

Aunque en `R` no está implementada la prueba de Lilliefors para la distribución exponencial, el siguiente código genera una función llamada `lillie-fors.exp.test`, la cual calcula el valor de la estadística D_2.

```
# Función para calcular la estadística D2
> lilliefors.exp.test <- function(x){
+ n <- length(x)
+ i <- 1:n
+ xi <- sort(x)
+ zi <- xi / mean(x)
+ fi <- 1 - exp(-zi)
+ t1 <- (i / n) - fi
+ t2 <- fi - ((i - 1)/n)
+ D <- max(abs(t1),abs(t2))
+ return(D)
}
# Estadística D2 usando el vector de datos b
> lilliefors.exp.test(b)
```

```
[1] 0.3189
```

▪

1.6.5. Prueba Shapiro-Wilk para normalidad

La prueba de Shapiro-Wilk se considera también como una prueba de bondad de ajuste y es usada para contrastar la normalidad para un conjunto de datos. Se plantea como hipótesis nula que una muestra $X_1, X_2 \ldots, X_n$ proviene de una distribución normal. Ésta prueba fue publicada en 1965 por Samuel Shapiro y Martin Wilk y se considera como una de las pruebas más potentes para el contraste de normalidad, sobre todo para muestras pequeñas cuando el número total de datos es $n < 30$.

Datos. Sea $X_1, X_2, \ldots, X_n$ muestra aleatoria de una distribución $F(x)$ desconocida.
Se calcula el valor

$$D = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$$

Se calculan las estadísticas de orden de la muestra

$$X_{(1)} < X_{(2)} < \cdots < X_{(n)}$$

Hipótesis a probar. Las hipótesis que se contrastan son:

H_0 : $F(x)$ es una función de distribución normal con parámetros desconocidos	*vs*	H_1 : $F(x)$ no es una función de distribución normal.

Estadística de prueba. La estadística de prueba está dada por

$$W = \frac{1}{D}\left[\sum_{j=1}^{k} a_j(X_{(n-j+1)} - X_{(j)})\right]^2.$$

En la tabla C.12, para la muestra de tamaño n, se obtienen los coeficientes $a_{j,n}$ para $j = 1, 2, \ldots, [n/2]$ y $n = 2, 3, \ldots$

La estadística W es básicamente el cuadrado del coeficiente de correlación de Pearson y se calcula entre las estadísticas de orden $X_{(i)}$ en la muestra y los coeficientes a_i, lo cual representa que las estadísticas de orden deben verse como una población normal. Así, si W es cercana a 1 la muestra se comporta como una muestra normal. Si W es demasiado pequeña, es decir, si W está muy por debajo de 1, la muestra no es normal. De esta forma, tenemos las siguientes observaciones:

Observación 1.7.

a) W es básicamente el cuadrado del coeficiente de correlación de Pearson.

b) W cercano a 1 significa que la muestra se comporta como una muestra de una distribución normal.

c) W muy pequeño, lejos de 1, significa que la muestra no pertenece a una distribución normal.

La teoría detrás de la prueba Shapiro-Wilk está fuera del alcance de este texto, pero si es de inteŕes del lector profundizar en el tema, puede consultar las publicaciones originales por Shapiro y Wilk [21] y [22].

Regla de decisión. Rechazar H_0 a un nivel de significancia α si: $W < w_\alpha$, donde w_α es el cuantil que acumula $\alpha\,\%$ de probabilidad y se busca en tablas de cuantiles de la estadística de Shapiro-Wilk.

Ejemplo 1.15. *Un programa generó 20 observaciones enteras del intervalo* $(0, 100)$ *y se obtuvo*

77	*57*	*40*	*23*	*56*
68	*87*	*93*	*61*	*73*
31	*58*	*66*	*74*	*81*
42	*89*	*70*	*33*	*75*

Se desea saber si la muestra proviene de una distribución normal. Probar con un nivel de significancia del 5 % y 10 %.

Solución: Ordenando la muestra se tiene

23	31	33	40	42
56	57	58	61	66
68	70	73	74	75
77	81	87	89	93

Como $n = 20$ y buscando los coeficientes $a_{j,n}$ para $j = 1, 2, \ldots, [n/2]$ y $n = 2, 3, \ldots$ en la tabla C.12 y haciendo las diferencias $X_{(n-j+1)} - X_{(j)}$, obtenemos

j	a_j	$X_{(n-j+1)} - X_{(j)}$
1	0.4734	$93 - 23 = 70$
2	0.3211	$89 - 31 = 58$
3	0.2565	$87 - 33 = 54$
4	0.2085	$81 - 40 = 41$
5	0.1686	$77 - 42 = 35$
6	0.1334	$75 - 56 = 19$
7	0.1013	$74 - 57 = 17$
8	0.0711	$73 - 58 = 15$
9	0.0422	$70 - 61 = 9$
10	0.0140	$68 - 66 = 2$

Calculando $\bar{X}$ y D tenemos

$$\bar{X} = 62.7 \quad \text{y} \quad D = 7726.2$$

calculando el valor de la estadística de prueba tenemos

$$W = \frac{1}{D}\left[\sum_{j=1}^{k} a_j (X_{(n-j+1)} - X_{(j)})\right]^2 = \frac{1}{7726.2}(85.7933)^2 = 0.9527$$

por lo que $W = 0.9527$. Buscando en la tabla C.11, para el nivel de significancia $\alpha = 5\,\%$ se tiene $w_{0.05} = 0.905$ y para $\alpha = 10\,\%$ tenemos $w_{0.1} = 0.92$, sabemos que se rechaza la hipótesis H_0 si, $W < w_\alpha$, por lo que no se rechaza H_0 para estos dos niveles de significancia, por lo tanto, podemos concluir que los números generados por el programa si provienen de una distribución normal.

En `R` existe la función `shapiro.test()` que permite obtener el resultado inmediato de la prueba Shapiro-Wilk para normalidad. Esta instrucción consta sólo de un vector numérico y proporciona la estadística de prueba W y el valor p de la prueba. El ejercicio anterior puede ser resuelto de la siguiente manera:

```
# Vector de datos
> z <- c(77, 57, 40, 23, 56, 68, 87, 93, 61, 73, 31, 58,
66, 74, 81, 42, 89, 70, 33, 75)
# Prueba Shapiro-Wilks
> shapiro.test(z)

        Shapiro-Wilk normality test

data:  z
W = 0.9525, p-value = 0.407
```

▪

Capítulo 2

Modelo de regresión lineal simple

2.1. Introducción

El contenido de este capítulo fue tomado de [17] y en éste se presenta el modelo de regresión lineal simple. Dicho modelo es el más sencillo de los modelos lineales e involucra una variable de interés y llamada *dependiente o respuesta* y su relación con la variable *predictoria o independiente* x, estableciendo que la media de la variable dependiente y cambia a razón constante cuando el valor de la variable independiente x crece o decrece. El modelo de regresión lineal simple es

$$y = \beta_0 + \beta_1 x + \epsilon, \tag{2.1}$$

donde

a) x es la variable regresora,

b) y es la variable de respuesta,

c) β_0 ordenada al origen,

d) β_1 pendiente del modelo,

e) ϵ es un error aleatorio.

Conviene considerar a la variable regresora x como una variable determinista, o bien, una variable controlada por el investigador la cual puede ser medida, mientras que la variable respuesta y es una variable aleatoria. Por ejemplo, en una compañía que vende y repara computadoras, los tiempos de reparación de las computadoras dependen del número de componentes electrónicos que deben ser reparados o reemplazados en cada una de las computadoras que se tienen para reparar.

Supongamos que se tiene un conjunto de datos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, obtenidos de un experimento en estudio. Con base en la ecuación (2.1) el modelo de regresión lineal simple se puede escribir como

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{para} \quad i = 1, 2, \ldots, n. \tag{2.2}$$

De esta forma, a la ecuación (2.1) se le conoce como ***modelo poblacional de regresión***, mientras que a la ecuación (2.2) se le llama ***modelo muestral de regresión***. Desde el punto de vista práctico, ϵ_i es una variable aleatoria y se considera como un error estadístico el cual describe la incapacidad de tener un modelo exacto de la realidad, este error explica por qué el modelo no ajusta con exactitud a los datos.

Suponiendo que la media de los errores $\mathbb{E}(\epsilon_i) = 0$ y que la varianza de los errores es constante, común y desconocida $Var(\epsilon_i) = \sigma^2$, y que el modelo es lineal respecto a los parámetros, es decir, los parámetros entran al modelo como coeficientes simples sobre las variables independientes, así se tiene que la media para las observaciones y_i es

$$\mathbb{E}(y_i|x_i) = \beta_0 + \beta_1 x_i$$

y la varianza es

$$\begin{aligned} Var(y_i|x_i) &= Var(\beta_0 + \beta_1 x_i + \epsilon_i) \\ &= Var(\epsilon_i) \\ &= \sigma^2 \quad \text{(constante)} \end{aligned}$$

para $i = 1, 2, \ldots, n$. Así, para el modelo poblacional de regresión, la media de y es una función lineal de x, aunque la varianza de y no depende del valor de x

$$\mathbb{E}(y|x) = \beta_0 + \beta_1 x \tag{2.3}$$

$$Var(y|x) = \sigma^2. \tag{2.4}$$

En la ecuación (2.3), β_0 es la intersección de la recta con el eje ordenado, mientras que β_1 es la pendiente de dicha recta y se interpreta como la razón de cambio en $\mathbb{E}(y)$ por unidad de cambio en x, o bien, la pendiente β_1 es el cambio en la media de la distribución de y producida por un cambio unitario en la variable x. Usualmente β_0 y β_1 son llamados *coeficientes de la regresión* y son desconocidos.

Nótese que si el rango de valores de la variable regresora x incluye el valor $x = 0$, entonces β_0 es la media de la distribución de la variable de respuesta y cuando $x = 0$, es decir $\mathbb{E}(y|x = 0) = \beta_0$. En caso de que $x = 0$ no esté en el dominio, entonces β_0 no tiene interpretación. El valor aleatorio ϵ es un error no observable y se suele suponer que los errores ϵ_i para cada i, no están correlacionados.

Al supuesto de que $\mathbb{E}(\epsilon_i) = 0$ y $Var(\epsilon_i) = \sigma^2$ se le conoce como *hipótesis distribucional* y a la suposición de que los parámetros entran al modelo como coeficientes simples se le llama *hipótesis estructural*. Como los errores ϵ_i son elementos aleatorios en el modelo, y es una variable aleatoria y existe una distribución de probabilidad para y en cada posible valor de x, en consecuencia las variables y_i son variables aleatorias y por lo tanto tienen la misma varianza y son mutuamente independientes. Además, se introduce la hipótesis de que los errores ϵ_i se distribuyen normal con media cero y varianza σ^2, son independientes para $i = 1, 2, \ldots, n$, o bien, $Cov(\epsilon_i, \epsilon_j) = 0$ para $i \neq j$, esta hipótesis

establece que las variables y_i también sigan una distribución normal con la finalidad de realizar inferencias sobre los parámetros del modelo construyendo intervalos de confianza y pruebas significativas.

En general, podemos decir que para el modelo de regresión lineal simple se tendrán los siguientes supuestos:

a) Linealidad (relación entre dichas variables).

b) $\mathbb{E}(\epsilon) = 0$ (la media del error es igual a cero).

c) $Var(\epsilon) = \sigma^2$ (varianza constante u homoscedasticidad).

d) $\epsilon \sim N(0, \sigma^2)$ (normalidad).

e) $Cov(\epsilon_i, \epsilon_j) = 0$ para $i \neq j$ (los errores son no correlacionados).

2.2. Estimación por mínimos cuadrados

El modelo de regresión lineal simple

$$y = \beta_0 + \beta_1 x + \epsilon$$

cuenta con dos parámetros desconocidos, β_0 y β_1, los cuales deben ser estimados a partir de los datos de la muestra. Con la hipótesis de varianza constante sobre los errores, aparece otro parámetro σ^2 desconocido, aunque no está incluido en el modelo también debe ser estimado. Un procedimiento para estimar los parámetros de un modelo lineal simple es el *método de mínimos cuadrados* que se puede ilustrar sencillamente aplicándolo para ajustar una línea recta a (x_1, y_1),(x_2, y_2),...,(x_n, y_n). También conocido como *principio de mínimos cuadrados*, se estiman β_0 y β_1 tales que la suma de los cuadrados de las diferencias entre las observaciones y_i y la línea recta sea mínima.
Sean $\widehat{\beta}_0$ y $\widehat{\beta}_1$ estimadores de los parámetros β_0 y β_1 respectivamente, entonces

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \quad \text{para} \quad i = 1, 2, \ldots, n \tag{2.5}$$

es un estimador de y_i para cada x_i y para $i = 1, 2, \ldots, n$. A la ecuación (2.5) se le conoce como *recta estimada o ajustada*. El objetivo es minimizar las distancias $y_i - \widehat{y}_i$ para $i = 1, 2, \ldots, n$ y poder encontrar los valores de $\widehat{\beta}_0$ y $\widehat{\beta}_1$ (véase Figura 2.1). A las desviaciones $y_i - \widehat{y}_i$ se les conoce como *residuales o errores*, y son denotadas como e_i, es decir,

$$e_i = y_i - \widehat{y}_i \quad \text{para} \quad i = 1, 2, \ldots, n.$$

Se define la suma de los cuadrados de los residuales, denotada por **SCE**, como

$$\mathbf{SCE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2.$$

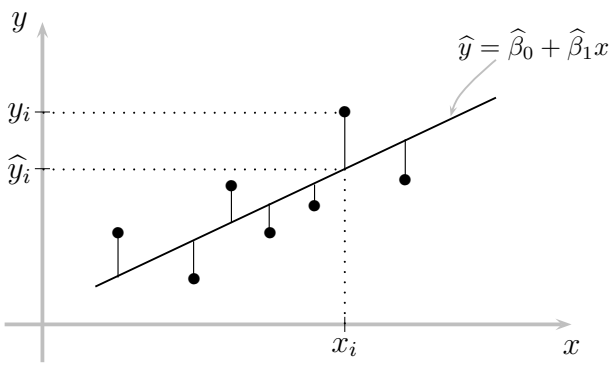


Figura 2.1: Ajuste de una línea recta a través de un conjunto de puntos.

El principio de mínimos cuadrados elige a $\widehat{\beta}_0$ y a $\widehat{\beta}_1$ de tal forma que minimicen la suma de los cuadrados de los residuales **SCE**. Estos estimadores, son variables aleatorias y se encuentran utilizando técnicas de cálculo diferencial.

Proposición 2.1. *Para el modelo de regresión lineal simple ajustado*

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \quad \textit{para} \quad i = 1, 2, \ldots, n$$

los estimadores por mínimos cuadrados $\widehat{\beta}_0$ *y* $\widehat{\beta}_1$ *están dados por*

a) $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$

b) $\widehat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$

donde

$$S_{xy} = \sum_{i=1}^{n} y_i (x_i - \bar{x}) \quad y \quad S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Demostración. Se tiene que

$$\begin{aligned} \textbf{SCE} &= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2, \end{aligned}$$

derivando esta ecuación con respecto a $\widehat{\beta}_0$ y $\widehat{\beta}_1$, e igualando a cero

$$\frac{\partial\, \textbf{SCE}}{\partial \widehat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0$$

$$\frac{\partial\, \textbf{SCE}}{\partial \widehat{\beta}_1} = -2 \sum_{i=1}^{n} x_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0,$$

entonces

$$\sum_{i=1}^{n} y_i - n\widehat{\beta}_0 - \widehat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \tag{2.6}$$

$$\sum_{i=1}^{n} x_i y_i - \widehat{\beta}_0 \sum_{i=1}^{n} x_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0. \tag{2.7}$$

Resolviendo estas ecuaciones para $\widehat{\beta}_0$ y $\widehat{\beta}_1$ se obtienen los estimadores de β_0 y β_1. De (2.6) tenemos

$$\sum_{i=1}^{n} y_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i = n\widehat{\beta}_0$$

entonces

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}.$$

Sustituyendo este valor en la ecuación (2.7)

$$\sum_{i=1}^{n} x_i y_i - (\bar{y} - \widehat{\beta}_1 \bar{x}) \sum_{i=1}^{n} x_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0,$$

$$\sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i + \widehat{\beta}_1 \bar{x} \sum_{i=1}^{n} x_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0,$$

y despejando $\widehat{\beta}_1$

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\bar{y}\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i}{\bar{x}\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i^2} \\
&= \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i\right)}{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i} \\
&= \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \bar{x}}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \\
&= \frac{\sum_{i=1}^{n} y_i (x_i - \bar{x})}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\
&= \frac{\sum_{i=1}^{n} y_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\
&= \frac{S_{xy}}{S_{xx}}.
\end{aligned}$$

▪

Comúnmente, a las ecuaciones (2.6) y (2.7) se les conocen como *ecuaciones normales* del modelo de regresión lineal simple.

Ejemplo 2.1. *En este ejemplo se utilizará el método de mínimos cuadrados para ajustar una línea recta para el siguiente conjunto de 5 datos.*

x	y
-2	*3*
-1	*2*
0	*1*
1	*1*
2	*0.5*

Solución: Obteniendo los siguientes valores

x_i	y_i	x_iy_i	x_i^2
-2	3	-6	4
-1	2	-2	1
0	1	0	0
1	1	1	1
2	0.5	1	4
$\sum_{i=1}^{5} x_i = 0$	$\sum_{i=1}^{5} y_i = 7.5$	$\sum_{i=1}^{5} x_iy_i = -6$	$\sum_{i=1}^{5} x_i^2 = 10$

y calculando

$$\begin{aligned}\widehat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \\ &= \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= -0.6,\end{aligned}$$

tenemos que

$$\begin{aligned}\widehat{\beta}_0 &= \bar{y} - \widehat{\beta}_1\bar{x} \\ &= \frac{7.5}{5} + (0.6)(0) \\ &= 1.5,\end{aligned}$$

por lo tanto, la recta ajustada viene dada por

$$\widehat{y} = 1.5 - 0.6x.$$

La interpretación de $\widehat{\beta}_1 = -0.6$ es que en promedio los valores de la variable respuesta y disminuirán aproximadamente 0.6 por cada unidad que aumente la variable regresora x, mientras que la intersección con el eje ordenado es $\widehat{\beta}_0 = 1.5$. Para encontrar los valores ajustados $\widehat{y}_i$ para $i = 1, \ldots, 5$, sustituimos los valores de la variable regresora x en el modelo ajustado obteniendo la siguiente tabla:

x	y	$\widehat{y}_i$
-2	3	2.7
-1	2	2.1
0	1	1.5
1	1	0.9
2	0.5	0.3

En la Figura 2.2 se grafica la recta ajustada incluyendo el conjunto de datos del ejemplo. Nótese que en la gráfica, la ordenada al origen es el valor promedio de las observaciones de y.

▪

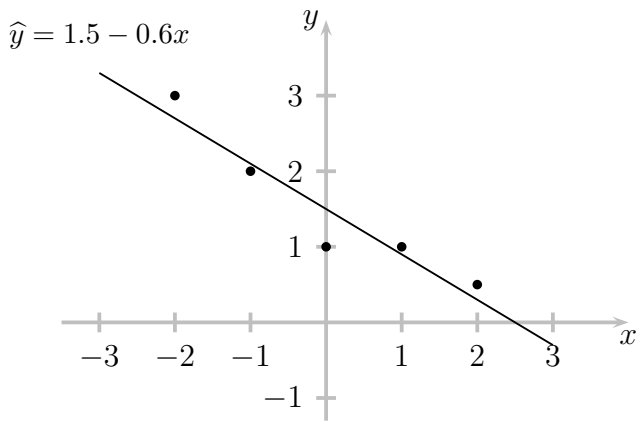


Figura 2.2: Recta ajustada de un conjunto de 5 datos.

2.3. Propiedades de los estimadores por mínimos cuadrados

En esta sección mostraremos que los estimadores por mínimos cuadrados, $\widehat{\beta}_0$ y $\widehat{\beta}_1$, cuentan con varias propiedades estadísticas importantes.

Proposición 2.2. *Los estimadores por mínimos cuadrados $\widehat{\beta}_0$ y $\widehat{\beta}_1$ son combinaciones lineales de las observaciones y_i.*

Demostración. Tenemos que

$$\begin{aligned}\widehat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \\ &= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) y_i \\ &= \sum_{i=1}^{n} \frac{(x_i - \bar{x}) y_i}{S_{xx}} \\ &= \sum_{i=1}^{n} c_i y_i\end{aligned}$$

donde $c_i = (x_i - \bar{x})/S_{xx}$ para $i = 1, 2, \ldots, n$. Por lo tanto, los estimadores $\widehat{\beta}_0$ y $\widehat{\beta}_1$ son combinaciones lineales de las observaciones y_i. ▪

Al ser $\widehat{\beta}_0$ y $\widehat{\beta}_1$ combinaciones lineales de las observaciones y_i, las cuales tienen distribución normal, entonces $\widehat{\beta}_0$ y $\widehat{\beta}_1$ también se distribuyen normal.

Proposición 2.3. *Las constantes c_i definidas anteriormente cumplen las siguientes características:*

a) $\sum_{i=1}^{n} c_i = 0$.

b) $\sum_{i=1}^{n} c_i x_i = 1$.

Demostración. Calculando

$$\begin{aligned}\sum_{i=1}^{n} c_i &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{S_{xx}} \\ &= \frac{\sum_{i=1}^{n} x_i - n\bar{x}}{S_{xx}} \\ &= 0.\end{aligned}$$

De manera similar tenemos que

$$\begin{aligned}\sum_{i=1}^{n} c_i x_i &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})x_i}{S_{xx}} \\ &= \frac{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i}{S_{xx}} \\ &= \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{S_{xx}} \\ &= \frac{S_{xx}}{S_{xx}} \\ &= 1.\end{aligned}$$

∎

Proposición 2.4. *Los estimadores* $\widehat{\beta}_0$ *y* $\widehat{\beta}_1$ *son* estimadores insesgados *para los parámetros* β_0 *y* β_1 *respectivamente.*

Demostración. Obteniendo la esperanza de $\widehat{\beta}_1$ y utilizando los resultados de la Proposición 2.3 se tiene que

$$\begin{aligned}\mathbb{E}(\widehat{\beta}_1) &= \mathbb{E}\left(\sum_{i=1}^{n} c_i y_i\right) \\ &= \sum_{i=1}^{n} \mathbb{E}(c_i y_i) \\ &= \sum_{i=1}^{n} c_i \mathbb{E}(y_i) \\ &= \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i) \\ &= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i \\ &= \beta_1.\end{aligned}$$

Análogamente, para $\widehat{\beta}_0$ tenemos

$$\begin{aligned}
\mathbb{E}(\widehat{\beta}_0) &= \mathbb{E}(\bar{y} - \widehat{\beta}_1\bar{x}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(y_i) - \mathbb{E}(\widehat{\beta}_1)\bar{x} \\
&= \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) - \beta_1\bar{x} \\
&= \frac{1}{n}[n\beta_0 + \beta_1\sum_{i=1}^{n}x_i] - \beta_1\bar{x} \\
&= \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} \\
&= \beta_0,
\end{aligned}$$

por lo tanto, $\widehat{\beta}_0$ y $\widehat{\beta}_1$ son estimadores insesgados para los parámetros β_0 y β_1, respectivamente. ▪

Proposición 2.5. *Los estimadores $\widehat{\beta}_0$ y $\widehat{\beta}_1$ tienen varianza finita y están dadas por*

$$Var(\widehat{\beta}_1) = \frac{\sigma^2}{S_{xx}} \qquad y \qquad Var(\widehat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right),$$

respectivamente.

Demostración. Calculando

$$\begin{aligned}
Var(\widehat{\beta}_1) &= Var\left(\sum_{i=1}^{n}c_i y_i\right) \\
&= \sum_{i=1}^{n}c_i^2 Var(y_i) \\
&= \sigma^2\sum_{i=1}^{n}c_i^2. \\
&= \sigma^2\sum_{i=1}^{n}\left[\frac{(x_i - \bar{x})}{S_{xx}}\right]^2 \\
&= \sigma^2\cdot\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{S_{xx}^2} \\
&= \sigma^2\cdot\frac{S_{xx}}{S_{xx}^2} \\
&= \frac{\sigma^2}{S_{xx}}
\end{aligned}$$

Para calcular la $Var(\widehat{\beta}_0)$, demostraremos primero que $Cov(\bar{y}, \beta_1) = 0$.

$$\begin{aligned} Cov(\bar{y}, \widehat{\beta}_1) &= Cov\left(\sum_{i=1}^{n} \frac{y_i}{n}, \sum_{i=1}^{n} c_i y_i\right) \\ &= \sum_{i=1}^{n} \left(\frac{c_i}{n}\right) Var(y_i) + \sum\sum_{i \neq j} \left(\frac{c_j}{n}\right) Cov(y_i, y_j). \end{aligned}$$

Como y_i y y_j, donde $i \neq j$, son independientes, $Cov(y_i, y_j) = 0$. Además, $Var(y_i) = \sigma^2$, por lo tanto

$$Cov(\bar{y}, \widehat{\beta}_1) = \frac{\sigma^2}{n} \sum_{i=1}^{n} c_i = \frac{\sigma^2}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_{xx}}\right) = 0.$$

Recordando que para dos variables aleatorias $Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X, Y)$, se tiene que

$$\begin{aligned} Var(\widehat{\beta}_0) &= Var(\bar{y} - \widehat{\beta}_1 \bar{x}) \\ &= Var(\bar{y}) + Var(\widehat{\beta}_1 \bar{x}) - 2Cov(\bar{y}, \widehat{\beta}_1 \bar{x}) \\ &= Var(\bar{y}) + \bar{x}^2 Var(\widehat{\beta}_1) - 2\bar{x} Cov(\bar{y}, \widehat{\beta}_1) \\ &= \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{xx}} \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right). \end{aligned}$$

Además, se puede demostrar que $Cov(\widehat{\beta}_0, \widehat{\beta}_1) = -\bar{x}\sigma^2 / S_{xx}$. Observe que $\widehat{\beta}_0$ y $\widehat{\beta}_1$ están correlacionadas y por tanto son dependientes, a menos que $\bar{x} = 0$.

Por último se enuncia el *Teorema de Gauss-Markov*, el cual resume las propiedades de los estimadores.

Teorema 2.1. *(Teorema de Gauss-Markov). Para el modelo de regresión lineal simple*

$$y = \beta_0 + \beta_1 x + \epsilon$$

con el supuesto de que $\mathbb{E}(\epsilon) = 0$ *y* $Var(\epsilon) = \sigma^2$ *y que dichos errores no están correlacionados, entonces los estimadores* $\widehat{\beta}_0$ *y* $\widehat{\beta}_1$ *por el método de mínimos cuadrados son los mejores estimadores lineales insesgados con varianza mínima.*

La demostración de este teorema es algo extensa y puede consultarse en [17]. Con frecuencia se dice que los estimadores por mínimos cuadrados son los *estimadores lineales insesgados óptimos*, donde "óptimos" implica que son de varianza mínima.

Otras propiedades

A continuación se enuncian otras propiedades de suma importancia para los estimadores por mínimos cuadrados.

a) Para el modelo de regresión lineal que contiene al estimador $\widehat{\beta}_0$, la suma de los residuales es igual a cero

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(y_i - \widehat{y}_i) = 0. \tag{2.8}$$

Esta propiedad es consecuencia directa de la ecuación normal (2.6). Es fácil ver que esta propiedad se cumple. Tenemos que

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \widehat{y}_i) &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \widehat{y}_i \\
&= n\bar{y} - \sum_{i=1}^{n}(\widehat{\beta}_0 + \widehat{\beta}_1 x_i) \\
&= n\bar{y} - \sum_{i=1}^{n} \widehat{\beta}_0 - \sum_{i=1}^{n} \widehat{\beta}_1 x_i \\
&= n\bar{y} - \sum_{i=1}^{n}(\bar{y} - \widehat{\beta}_1 \bar{x}) - \widehat{\beta}_1 \sum_{i=1}^{n} x_i \\
&= n\bar{y} - \sum_{i=1}^{n} \bar{y} + \widehat{\beta}_1 \sum_{i=1}^{n} \bar{x} - \widehat{\beta}_1 \sum_{i=1}^{n} x_i \\
&= n\bar{y} - n\bar{y} - \widehat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x}) \\
&= -\widehat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x}) \\
&= 0,
\end{aligned}$$

ya que $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

b) La suma de los valores observados y_i es igual a la suma de los valores ajustados $\widehat{y}_i$, es decir,

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y}_i. \tag{2.9}$$

Esta propiedad es consecuencia inmediata de la propiedad anterior.

c) La recta de la regresión por mínimos cuadrados pasa por el *centroide* de los datos el cual corresponde al punto $(\bar{x}, \bar{y})$. En efecto, evaluando $\bar{x}$ en la recta de la regresión y usando el hecho de que $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$ se obtiene

$$\begin{aligned}
\widehat{y} &= \widehat{\beta}_0 + \widehat{\beta}_1 \bar{x} \\
&= (\bar{y} - \widehat{\beta}_1 \bar{x}) + \widehat{\beta}_1 \bar{x} \\
&= \bar{y}.
\end{aligned}$$

d) La suma de los residuales, ponderados por el valor correspondiente de la variable regresora siempre es igual a cero, es decir,

$$\sum_{i=1}^{n} x_i e_i = 0. \tag{2.10}$$

En efecto,

$$\begin{aligned}
\sum_{i=1}^{n} x_i e_i &= \sum_{i=1}^{n} x_i (y_i - \widehat{y}_i) \\
&= \sum_{i=1}^{n} x_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) \\
&= \sum_{i=1}^{n} x_i y_i - \widehat{\beta}_0 \sum_{i=1}^{n} x_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\
&= \sum_{i=1}^{n} x_i y_i - (\bar{y} - \widehat{\beta}_1 \bar{x}) \sum_{i=1}^{n} x_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\
&= \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i + \widehat{\beta}_1 \bar{x} \sum_{i=1}^{n} x_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\
&= \sum_{i=1}^{n} x_i y_i - \bar{x} \sum_{i=1}^{n} y_i - \widehat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i \right) \\
&= \sum_{i=1}^{n} x_i y_i - \bar{x} \sum_{i=1}^{n} y_i - \frac{S_{xy}}{S_{xx}} \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) \\
&= \sum_{i=1}^{n} (x_i y_i - \bar{x} y_i) - \frac{S_{xy}}{S_{xx}} S_{xx} \\
&= \sum_{i=1}^{n} y_i (x_i - \bar{x}) - \sum_{i=1}^{n} y_i (x_i - \bar{x}) \\
&= 0.
\end{aligned}$$

e) La suma de los residuales, ponderados por el valor ajustado correspondiente siempre es igual a cero. o bien,

$$\sum_{i=1}^{n} \widehat{y}_i e_i = 0. \tag{2.11}$$

Por las propiedades a) y d) vemos que

$$\begin{aligned}
\sum_{i=1}^{n} \widehat{y}_i e_i &= \sum_{i=1}^{n} (\widehat{\beta}_0 + \widehat{\beta}_1 x_i) e_i \\
&= \widehat{\beta}_0 \sum_{i=1}^{n} e_i + \widehat{\beta}_1 \sum_{i=1}^{n} x_i e_i \\
&= 0.
\end{aligned}$$

2.4. Estimación de σ^2

Como ya hemos visto, se hallaron los estimadores $\widehat{\beta}_0$ y $\widehat{\beta}_1$ para los parámetros β_0 y β_1 respectivamente. En esta sección se hallará el estimador para σ^2, ya que es indispensable para la construcción de intervalos de confianza y para realizar pruebas de hipótesis necesarias para el modelo de regresión lineal simple, dicho estimador se obtiene a partir de la suma de los cuadrados de los residuales **SCE** y para ello tenemos las siguientes proposiciones.

Proposición 2.6. *La suma de los cuadrados de los residuales* ***SCE*** *puede ser expresada como*

$$\boldsymbol{SCE} = S_{yy} - \widehat{\beta}_1 S_{xy}.$$

Demostración. Recordando que

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

y sustituyendo en la suma de los cuadrados de los residuales **SCE**, tenemos

que

$$\begin{aligned}
\mathbf{SCE} &= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \\
&= \sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 \\
&= \sum_{i=1}^{n}(y_i - \bar{y} + \widehat{\beta}_1\bar{x} - \widehat{\beta}_1 x_i)^2 \\
&= \sum_{i=1}^{n}[y_i - \bar{y} - \widehat{\beta}_1(x_i - \bar{x})]^2 \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 - 2\widehat{\beta}_1(y_i - \bar{y})(x_i - \bar{x}) + \widehat{\beta}_1^2(x_i - \bar{x})^2 \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 - 2\widehat{\beta}_1\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) + \widehat{\beta}_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 - 2\frac{S_{xy}}{S_{xx}}S_{xy} + \frac{S_{xy}^2}{S_{xx}^2}S_{xx} \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 - \frac{S_{xy}^2}{S_{xx}} \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 - \widehat{\beta}_1 S_{xy} \\
&= S_{yy} - \widehat{\beta}_1 S_{xy}.
\end{aligned}$$

∎

Proposición 2.7. *El valor esperado de la suma de los cuadrados de los residuales* ***SCE*** *está dado por*

$$\mathbb{E}(\boldsymbol{SCE}) = (n-2)\sigma^2.$$

Demostración. En efecto, tenemos que

$$\begin{aligned}
\mathbb{E}(\mathbf{SCE}) &= \mathbb{E}\left[\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n}(y_i - \bar{y} + \widehat{\beta}_1\bar{x} - \widehat{\beta}_1 x_i)^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n}\Big((y_i - \bar{y}) - \widehat{\beta}_1(x_i - \bar{x})\Big)^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n}(y_i - \bar{y})^2 + \widehat{\beta}_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2 - 2\widehat{\beta}_1\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right],
\end{aligned}$$

como

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\widehat{\beta}_1\sum_{i=1}^{n}(x_i-\bar{x})^2$$

y

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-n\bar{y}^2,$$

entonces

$$\begin{aligned}\mathbb{E}(\mathbf{SCE})&=\mathbb{E}\left(\sum_{i=1}^{n}y_i^2-n\bar{y}^2-\widehat{\beta}_1^2S_{xx}\right)\\&=\sum_{i=1}^{n}\mathbb{E}(y_i^2)-n\mathbb{E}(\bar{y}^2)-S_{xx}\mathbb{E}(\widehat{\beta}_1^2),\end{aligned}$$

sabemos que si U es una variable aleatoria tenemos que $\mathbb{E}(U^2)=Var(U)+\mathbb{E}^2(U)$, entonces

$$\begin{aligned}\mathbb{E}(\mathbf{SCE})&=\sum_{i=1}^{n}\left\{Var(y_i)+\mathbb{E}^2(y_i)-n\left[Var(\bar{y})+\mathbb{E}^2(\bar{y})\right]-S_{xx}\left[Var(\widehat{\beta}_1)+\mathbb{E}^2(\widehat{\beta}_1)\right]\right\}\\&=n\sigma^2+\sum_{i=1}^{n}(\beta_0+\beta_1x_i)^2-n\left[\frac{\sigma^2}{n}+(\beta_0+\beta_1\bar{x})^2\right]-S_{xx}\left[\frac{\sigma^2}{S_{xx}}+\beta_1^2\right]\\&=n\sigma^2+\sum_{i=1}^{n}(\beta_0+\beta_1x_i)^2-\sigma^2-n(\beta_0+\beta_1\bar{x})^2-\sigma^2-S_{xx}\beta_1^2\\&=(n-2)\sigma^2+\beta_1^2\left(\sum_{i=1}^{n}x_i^2-n\bar{x}^2-S_{xx}\right)\\&=(n-2)\sigma^2+\beta_1^2\left(\sum_{i=1}^{n}(x_i-\bar{x})^2-S_{xx}\right)\\&=(n-2)\sigma^2.\end{aligned}$$

∎

A partir de las proposiciones 2.6 y 2.7 se tiene el siguiente resultado.

Proposición 2.8. *Un estimador insesgado para σ^2 está dado por*

$$\widehat{\sigma}^2=\frac{\boldsymbol{SCE}}{(n-2)}=\frac{\sum_{i=1}^{n}e_i}{(n-2)}=\frac{\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}{(n-2)}.$$

Frecuentemente, al estimador $\widehat{\sigma}^2$ se le denota por las siglas **CME** y suele llamarse *Cuadrado Medio de los Residuales o Errores*. La raíz cuadrada de $\widehat{\sigma}^2$ es a veces llamada *Error Estándar de la Regresión*.

Proposición 2.9. *Bajo el supuestos de que los errores ϵ_i se distribuyen $N(0,\sigma^2)$ en el modelo $y_i=\beta_0+\beta_1x_i+\epsilon_i$, donde $i=1,\ldots,n$, la cantidad*

$$\frac{(n-2)}{\sigma^2}\boldsymbol{CME}$$

sigue una distribución Ji-cuadrada con $n-2$ grados de libertad.

Hasta el momento, el único supuesto que hemos utilizado acerca del los errores aleatorios ϵ_i del modelo de regresión es que se distribuyen normal con media 0 y varianza σ^2 debido a que con frecuencia se presenta esta distribución en la naturaleza. Si se garantiza esta suposición de normalidad, se deduce que las y_i se distribuyen normal con media $\beta_0 + \beta_1 x_i$ y varianza σ^2. Como hemos visto, los estimadores β_0 y β_1 son combinaciones lineales de las observaciones y_i con medias y varianzas que se calcularon anteriormente. Además, se deduce que

$$\frac{(n-2)}{\sigma^2}\mathbf{CME} = \frac{\mathbf{SCE}}{\sigma^2}$$

y en consecuencia

$$\frac{\mathbf{SCE}}{(n-2)} \sim \chi^2_{(n-2)}.$$

Ejemplo 2.2. *A continuación se muestra cómo calcular el estimador $\widehat{\sigma}^2$ utilizando los datos del ejemplo 2.1.*

Solución: A partir de los valores ajustados $\widehat{y}_i$ se calculan los cuadrados de los errores obteniendo

x	y	$\widehat{y}_i$	$(y_i - \widehat{y}_i)^2$
-2	3	2.7	0.09
-1	2	2.1	0.01
0	1	1.5	0.25
1	1	0.9	0.01
2	0.5	0.3	0.04

entonces $\sum_{i=1}^{5}(y_i - \widehat{y}_i)^2 = 0.4$, por lo tanto, el valor del estimador $\widehat{\sigma}^2$ es

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{(n-2)} = \frac{0.4}{5-2} = 1.333.$$

▪

2.5. Predicción de observaciones nuevas

El modelo de regresión ajustado es de gran utilidad para realizar predicciones de observaciones futuras de la variable respuesta. Los riesgos que se tiene al realizar pronósticos con un modelo lineal son debidos a los errores aleatorios del modelo. Si x_0 es el valor observado de un experimento en estudio, el valor

$$\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$$

es el nuevo estimador puntual del nuevo valor de respuesta y_0.

Ejemplo 2.3. *En este ejemplo se muestra cómo puede ser utilizada la recta de regresión ajustada para realizar predicciones dado un valor de la variable regresora x.*

Suponiendo que se obtuvieron los siguientes estadísticos:

$$\bar{x} = 13.9, \quad \bar{y} = 14.6, \quad S_{xx} = 46.8, \quad S_{yy} = 53.3, \quad S_{xy} = 12.2.$$

Calculando

$$\begin{aligned} \widehat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \\ &= \frac{12.2}{46.8} \\ &= 0.2607 \end{aligned}$$

y

$$\begin{aligned} \widehat{\beta}_0 &= \bar{y} - \widehat{\beta}_1 \bar{x} \\ &= 14.6 - (0.2607)13.9 \\ &= 10.9763, \end{aligned}$$

tenemos que la recta de regresión estimada es

$$\widehat{y} = 10.9763 + 0.207x.$$

Si $x_0 = 15.1$ entonces

$$\begin{aligned} \widehat{y} &= 10.9763 + (0.207)(15.1) \\ &= 14.9, \end{aligned}$$

por lo tanto, la predicción de la variable respuesta es 14.9. ▪

2.6. Coeficiente de correlación

Una cuestión importante en los modelos lineales es la variación conjunta de dos o más variables, de las cuales uno se pregunta: ¿Qué tan asociadas de manera lineal se encuentran estas variables?. Los métodos o técnicas que se han desarrollado para medir el grado de asociación entre variables se le conoce como *métodos de correlación*, dando lugar a una *medida de correlación*. Cuando se hace un análisis de correlación entre dos o mas variables, la medida de correlación es llamada comúnmente como *coeficiente de correlación*. Así, el coeficiente de correlación de dos variables aleatorias es un número real que mide el grado de *dependencia lineal* que existe entre dichas variables.

Definición 2.1. *El coeficiente de correlación de dos variables aleatorias X y Y, denotados por $\rho(X,Y)$, se define como el número*

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

En esta definición es necesario suponer que las varianzas de las variables son estrictamente positivas y finitas. Vista como una función de dos variables, el coeficiente de correlación es una función simétrica pero no es lineal pues no separa sumas ni multiplicaciones por escalares o constantes, es decir, en general,

a) $\rho(kX, Y) \neq k\rho(X, Y)$

b) $\rho(X_1 + X_2, Y) \neq \rho(X_1, Y) + \rho(X_2, Y)$

Propiedades del coeficiente de correlación

En general, el coeficiente de correlación satisface las siguientes propiedades.

1. Si las variables X y Y son independientes, entonces $\rho(X, Y) = 0$.

2. $-1 \leq \rho(X, Y) \leq 1$.

3. $|\rho(X, Y)| = 1$ si, y sólo si, existen constantes β_0 y β_1 tales que, con probabilidad uno, $Y = \beta_0 + \beta_1 X$ con $\beta_1 > 0$ si $\rho(X, Y) = 1$, y $\beta_1 < 0$ si $\rho(X, Y) = -1$.

Cuando $\rho(X, Y) = 0$, se dice que las variables X y Y son *no correlacionadas*. Cuando $|\rho(X, Y)| = 1$ se dice que las variables x y y están perfectamente correlacionadas *positivamente* o *negativamente*, de acuerdo al signo de $\rho(X, Y)$. La propiedad 3 nos dice que, mientras el valor de $\rho(X, Y)$ es más cercano a -1 o a 1 la relación es más fuerte, de tal manera que casi se dibuja una línea recta. Si el valor de $\rho(X, Y)$ es cercano a-1 su relación es inversa, es decir, mientras una variable crece, la otra decrece (pendiente negativa). Si el valor de $\rho(X, Y)$ es cercano a 1 su relación es directa, lo cual ambas variables crecen (pendiente positiva).

En el análisis de regresión lineal es de gran interés calcular el *coeficiente de correlación muestral* para medir la dependencia lineal entre variables.
Para la muestra $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, la expresión del coeficiente de correlación muestral, también conocida como *coeficiente de correlación de Pearson* está dada por

$$r_{x,y} = \frac{\sum_{i-1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

el cual satisface que $-1 \leq r_{x,y} \leq 1$. El Cuadro 2.1 describe un resumen más claro del tipo de correlación que puede existir entre la variable x y la variable y.

Correlación positiva		**Correlación negativa**	
$r_{x,y} = 1$	Perfecta	$r_{x,y} = -1$	Perfecta
$0.95 \leq r_{x,y} < 1$	Muy fuerte	$-1 < r_{x,y} \leq -0.95$	Muy fuerte
$0.87 \leq r_{x,y} < 0.95$	Fuerte	$-0.95 < r_{x,y} \leq -0.87$	Fuerte
$0.5 \leq r_{x,y} < 0.87$	Moderada	$-0.87 < r_{x,y} \leq -0.5$	Moderada
$0.1 \leq r_{x,y} < 0.5$	Débil	$-0.5 < r_{x,y} \leq -0.1$	Débil
$0 \leq r_{x,y} < 0.1$	No correlación	$-0.1 < r_{x,y} \leq 0$	No correlación

Cuadro 2.1: Posibles resultados del coeficiente de correlación.

2.7. Coeficiente de determinación R^2

En esta parte se presenta el *coeficiente de determinación* asociado al modelo de regresión lineal. El coeficiente de determinación mide la proporción de la variabilidad de la respuesta y que explica el modelo de regresión.

Definición 2.2. *Se define el* coeficiente de determinación*, denotado por R^2, como la proporción de la variabilidad explicada por el modelo de regresión con respecto a la variabilidad total, es decir*

$$R^2 = \frac{\textit{Variabilidad explicada por el modelo (observaciones que no caen en la recta)}}{\textit{Variabilidad total (todas las observaciones)}}$$

Para dar una expresión numérica del coeficiente de determinación R^2 partimos de la identidad

$$y_i - \bar{y} = (\widehat{y}_i - \bar{y}) + (y_i - \widehat{y}_i).$$

Elevando al cuadrado ambos lados de la ecuación y tomando la suma de todas las observaciones tenemos

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + 2\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})(y_i - \widehat{y}_i).$$

El tercer sumando del lado derecho de esta ecuación se puede escribir de la siguiente manera

$$\begin{aligned}
2\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})(y_i - \widehat{y}_i) &= 2\sum_{i=1}^{n}(\widehat{y}_i y_i - \widehat{y}_i^2 - \bar{y}y_i + \bar{y}\widehat{y}_i) \\
&= 2\sum_{i=1}^{n}[\widehat{y}_i(y_i - \widehat{y}_i) - \bar{y}(y_i - \widehat{y}_i)] \\
&= 2\sum_{i=1}^{n}\widehat{y}_i(y_i - \widehat{y}_i) - 2\bar{y}\sum_{i=1}^{n}(y_i - \widehat{y}_i) \\
&= 2\sum_{i=1}^{n}\widehat{y}_i e_i - 2\bar{y}\sum_{i=1}^{n}e_i \\
&= 0,
\end{aligned}$$

ya que la suma de los residuales siempre es igual a cero (ec. 2.9) y la suma de los residuales ponderados por el valor ajustado $\widehat{y}_i$ correspondiente también es igual a cero (ec. 2.11). En consecuencia, se tiene que

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2. \tag{2.12}$$

El lado izquierdo de la ecuacion (2.12) es la suma corregida de los cuadrados de las observaciones y que mide la variabilidad total de las observaciones. Los dos componentes a la derecha de la igualdad miden, respectivamente, la cantidad de variabilidad en las observaciones de y_i explicada por la recta de regresión, y la variación de los residuales que queda sin explicar por la recta de regresión. Estableciendo la siguiente notación:

SCT $= \sum_{i=1}^{n}(y_i - \bar{y})^2$ la suma de los cuadrados totales o la variabilidad total de la respuesta y.

SCR $= \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2$ la suma explicada por el modelo de regresión o la suma de los cuadrados de la regresión.

SCE $= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ la variabilidad no explicada por el modelo de regresión o la suma de los cuadrados de los residuales u errores la cual ya se ha presentado en apartados anteriores.

Podemos escribir la ecuación (2.12) como

$$\mathbf{SCT} = \mathbf{SCR} + \mathbf{SCE} \tag{2.13}$$

A la ecuación (2.13) se le conoce como *igualdad fundamental del análisis de varianza* para un modelo de regresión el cual se estudiará más adelante.

Por la definición 2.2, el coeficiente de determinación R^2 se puede escribir como

$$R^2 = \frac{\mathbf{SCR}}{\mathbf{SCT}}. \tag{2.14}$$

Despejando **SCR** de la ecuación (2.13) y sustituyendo en (2.14), el coeficiente de determinación se puede ver como

$$R^2 = 1 - \frac{\mathbf{SCE}}{\mathbf{SCT}}.$$

De esta manera, el cociente

$$\frac{\mathbf{SCE}}{\mathbf{SCT}} = \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

es el la proporción de variabilidad de la respuesta y que el modelo no puede explicar.

Propiedades de R^2

El coeficiente de determinación R^2 satisface las siguientes propiedades:

a) $0 \leq R^2 \leq 1$ ya que $0 \leq \mathbf{SCE} \leq \mathbf{SCT}$.

b) Si $R^2 = 1$ entonces tenemos que **SCR**=**SCT** y $\dfrac{\mathbf{SCE}}{\mathbf{SCT}} = 0$.

c) Si $R^2 = 0$ entonces **SCE**=**SCT**.

En particular se buscan valores de R^2 cercanos a 1, ya que esto indica que la mayor parte de la variabilidad de y es determinada o explicada por el modelo de regresión. La magnitud de R^2 también depende del rango de variabilidad de la variable regresora. En general R^2 aumenta a medida que la propagación de los valores de x aumenten, y disminuye a medida que la propagación de los valores de x disminuyan, siempre que el modelo asumido es correcto. El Cuadro 2.2 muestra un resumen del tipo de correlación que puede existir dependiendo del valor de R^2.

Valor de R^2	**Tipo de correlación**
$R^2 = 0$	No correlación
$0 < R^2 < 0.25$	Correlación muy débil
$0.25 \leq R^2 < 0.5$	Correlación débil
$0.5 \leq R^2 < 0.75$	Correlación moderada
$0.75 \leq R^2 < 0.9$	Correlación fuerte
$0.9 \leq R^2 < 1$	Correlación muy fuerte
$R^2 = 1$	Correlación perfecta

Cuadro 2.2: Descripción de los posibles resultados de R^2.

El coeficiente de determinación R^2 también puede verse en términos del coeficiente de correlación muestral $r_{x,y}$ como

$$R^2 = (r_{x,y})^2$$

con

$$r_{x,y} = \frac{\sum_{i-1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Observación 2.1. *Para la comparación de modelos es de gran utilidad usar el coeficiente de determinación R^2, ya que el mejor modelo es aquel que explica la mayoría de la información.*

2.8. Pruebas de hipótesis

Como ya hemos visto, las características generales de la línea recta para el modelo de regresión lineal simple, están determinadas por la intersección con el eje de la variable respuesta β_0 y la pendiente β_1. En esta sección se determinan pruebas de hipótesis sobre los parámetros β_1 y β_0 del modelo.

En particular, para el modelo de regresión lineal simple, bajo el supuesto de que los errores ϵ_i se distribuyen normal con media cero y varianza σ^2 e independientes, se puede hacer uso de la prueba t para realizar pruebas de hipótesis.

Prueba de hipótesis para β_1

Para β_1, supongamos que se desean probar las hipótesis que la pendiente de la recta del modelo de regresión es igual a una costante, por ejemplo, a β^*, las hipótesis correspondientes son

$$H_0 : \beta_1 = \beta_1^* \quad vs \quad H_1 : \beta_1 \neq \beta_1^*,$$

donde dicho contraste corresponde a una prueba de dos colas, es decir, se ha especificado la hipótesis alternativa bilaterial. Como los errores ϵ_i son normales con media 0 y varianza σ^2, las observaciones y_i son normales con media $\beta_0+\beta_1 x_i$ y varianza σ^2. Ahora, $\widehat{\beta}_1$ es una combinación lineal de las observaciones y_i, de tal modo que $\widehat{\beta}_1$ está distribuido normalmente con media β_1 y varianza σ^2/S_{xx}, esto es

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right).$$

Usando esta información, la estadística de prueba está dado por

$$Z = \frac{\widehat{\beta}_1 - \beta_1^*}{\sqrt{\sigma^2/S_{xx}}} \sim N(0,1)$$

bajo la hipótesis nula $H_0 : \beta_1 = \beta_1^*$. Si σ^2 fuera conocida se podría usar Z para realizar el contraste y la región de rechazo para la prueba de dos colas está dada por

$$|Z| \geq z_{\alpha/2}.$$

Como en el caso de la prueba Z, para calcular la estadística de prueba debemos conocer a σ^2 o tener una buena estimación. Cuando esta estimación no existe, que en la mayoría de los casos esto ocurre, puede calcularse una estimación de σ^2 y sustituirse en la estadística Z. De esta forma, como:

a) **CME** es un estimador insesgado para σ^2,

b) $\dfrac{(n-2)}{\sigma^2}\mathbf{CME} \sim \chi^2_{(n-2)}$,

c) **CME** y $\widehat{\beta}_1$ son independientes,

y de acuerdo a la definición de la estadística T tenemos que

$$T_1 = \frac{\widehat{\beta}_1 - \beta_1^*}{\sqrt{\dfrac{\mathbf{CME}}{S_{xx}}}} \sim t_{(n-2)} \tag{2.15}$$

bajo la hipótesis nula $H_0 : \beta_1 = \beta_1^*$. La cantidad de grados de libertad asociada a t_0 es igual a la cantidad de grados de libertad asociados al **CME**. De esta forma la estadística con que se prueba $H_0 : \beta_1 = \beta_1^*$ es t_0. Así, rechazamos H_0 al nivel de significancia α si

$$t_0 > t_{n-2}^{\alpha/2} \quad \text{o} \quad t_0 < -t_{n-2}^{\alpha/2}$$

o bien

$$|t_0| > t_{n-2}^{\alpha/2}.$$

El denominador de la estadística t_0 en la ecuación (2.15) se le conoce con frecuencia como ***error estándar de la pendiente***. Esto es

$$se(\widehat{\beta}_1) = \sqrt{\frac{\mathbf{CME}}{S_{xx}}}.$$

Análogamente, para la ordenada al origen β_0, se puede usar un procedimiento para probar hipótesis acerca de la ordenada al origen β_0. Si se quiere probar

$$H_0 : \beta_0 = \beta_0^* \quad vs \quad H_1 : \beta_0 = \beta_0^*$$

se podría usar la estadística de prueba

$$T_0 = \frac{\widehat{\beta}_0 - \beta_0^*}{\sqrt{\mathbf{CME}\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{(n-2)}$$

en donde el denominador de la estadística se le conoce como ***error estándar de la ordenada al origen***

$$se(\widehat{\beta}_0) = \sqrt{\mathbf{CME}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}.$$

Se rechaza la hipótesis nula $H_0 : \beta_0 = \beta_0^*$ si

$$t_0 > t_{(n-2)}^{\alpha/2} \quad \text{o} \quad t_0 < -t_{(n-2)}^{\alpha/2}$$

o bien

$$|t_0| > t_{(n-2)}^{\alpha/2}.$$

En resumen, tenemos los siguientes resultados.

Prueba de hipótesis para β_i

$H_0 : \beta_i = \beta_i^*$

$$H_1 : \begin{cases} \beta_i > \beta_i^* & \text{region de rechazo cola superior,} \\ \beta_i < \beta_i^* & \text{region de rechazo cola inferior,} \\ \beta_i \neq \beta_i^* & \text{region de rechazo de dos colas.} \end{cases}$$

Estadística de prueba.

$$T_i = \frac{\widehat{\beta}_i - \beta_i^*}{\sqrt{\mathbf{CME} \cdot k_i}}$$

para $i = 0, 1$ y

$$k_0 = \frac{\sum_{i=1}^n x_i^2}{nS_{xx}} \quad \text{y} \quad k_1 = \frac{1}{S_{xx}}.$$

Región de rechazo.

$$C = \begin{cases} \{T_i > t^{\alpha}_{(n-2)}\} & \text{cola superior,} \\ \{T_i < -t^{\alpha}_{(n-2)}\} & \text{cola inferior,} \\ \{|T_i| > t^{\alpha/2}_{(n-2)}\} & \text{dos colas,} \end{cases}$$

donde $t^{\alpha}_{(n-2)}$ es el cuantil de una distribución t con $n-2$ grados de libertad que acumula α de probabilidad.

Ejemplo 2.4. *En este ejemplo, queremos determinar si los datos de un experimento presentan suficiente evidencia para determinar que la pendiente del modelo de regresión que se ajustó es diferente de 0. Suponiendo que los datos mostraron una pendiente $\widehat{\beta}_1 = 0.7$, $S_{xx} = 10$ y que el estimador de σ^2 es $\boldsymbol{CME} = \boldsymbol{SCE}/(n-2) = 0.367$ con $n = 5$.*

Nuestro interés es saber si el parámetro β_1 es distinto de cero, por lo que la prueba de hipótesis de dos colas es

$$H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0.$$

Determinando el valor de la estadística de prueba vemos que

$$T_1 = \frac{\widehat{\beta}_1 - 0}{\sqrt{\dfrac{\mathbf{CME}}{S_{xx}}}} = \frac{0.7 - 0}{\sqrt{0.367/10}} = 3.654.$$

Si consideramos un nivel de significancia de $\alpha = 5\,\%$, el valor del cuantil $t^{0.025}_{(3)}$

```
> qt(p=0.025, df=3, lower.tail=F)
```

```
[1] 3.182
```

y sabemos que la región de rechazo es $|T_1| > t^{0,025}_{(3)}$, es decir, $|3.654| > 3.182446$, entonces, se rechaza la hipótesis nula H_0. Por lo tanto, decimos que los datos muestran que la pendiente del modelo es distinta de cero. ▪

2.9. Prueba de significancia de la regresión

Un tema de suma importancia que involucra un contraste de hipótesis es la determinación de la pendiente del modelo de regresión, es decir, un caso es el contraste

$$H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$$

estas hipótesis se relacionan con la *significancia de la regresión*. Al aceptar la hipótesis nula H_0 implica que no hay relación lineal entre x y y (véase Figura 2.3). Nótese que eso puede implicar que x tiene muy poco valor para explicar la variación de y y que el mejor estimador para cualquier x es $\widehat{y} = \bar{y}$ (Figura 2.3 a), o que la verdadera relación entre x y y no es lineal (Figura 2.3 b).

Observación 2.2. *Por consiguiente, si no se rechaza* $H_0 : \beta_1 = 0$*, equivale a decir que* no hay relación lineal entre y y x.

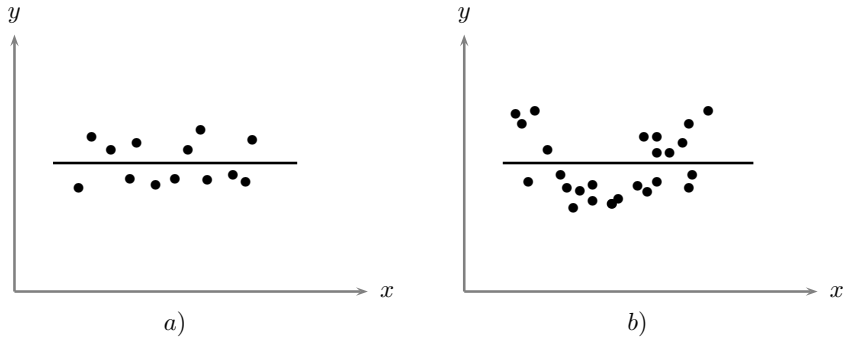


Figura 2.3: Casos en los que no se rechaza $H_0 : \beta_1 = 0$.

Si se rechaza la hipótesis nula H_0, eso implica que x sí tiene valor para explicar la variabilidad de y (véase Figura 2.4), o bien, que la pendiente es distina de cero y si se puede determinar un modelo de regresión. Sin embargo, rechazar $H_0 : \beta_1 = 0$ podría equivaler a que el modelo es adecuado, como se muestra en la Figura 2.4 a) o que aunque hay un efecto lineal de x se podrían obtener mejores resultados agregando términos polinomiales en x (Figura 2.4b).

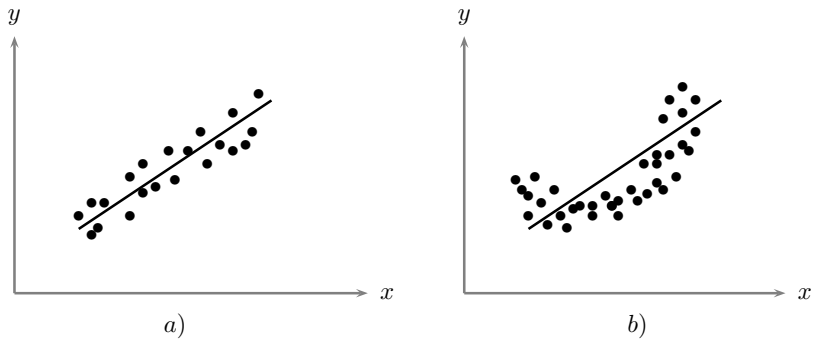


Figura 2.4: Casos en los que sí se rechaza $H_0 : \beta_1 = 0$.

El procedimiento de prueba para $H_0 : \beta_1 = 0$ se puede establecer con dos métodos, en el primero tan sólo se usa la estadística t de la ecuación (2.15),

con $\beta^* = 0$,

$$t_0 = \frac{\widehat{\beta}_1}{\sqrt{\frac{\mathbf{CME}}{S_{xx}}}} \sim t_{(n-2)}$$

y el segundo, se puede usar el método de análisis de varianza que estudiaremos en la siguiente sección.

2.10. Análisis de varianza

También se puede usar el ***método de análisis de varianza*** para estudiar la variación que tiene la variable respuesta y la prueba de significancia de la regresión. Este análisis se basa en una partición de la variabilidad total de la variable respuesta y como se realizó en la sección 2.7. A partir de los siguientes resultados

$$\mathbf{SCE} = S_{yy} - \widehat{\beta}_1 S_{xy}$$

y

$$\mathbf{SCT} = \mathbf{SCR} + \mathbf{SCE}$$

se puede escribir la suma de los cuadrados de la regresión de la siguiente manera

$$\mathbf{SCR} = \widehat{\beta}_1 S_{xy}. \tag{2.16}$$

La cantidad de grados de libertad se determina como sigue: la suma de los cuadrados totales **SCT** tiene $df_T = n - 1$ grados de libertad, por que se pierde un grado de libertad como resultado de la restricción $\sum_{i=1}^{n}(y_i - \bar{y})^2$ para las desviaciones $y_i - \bar{y}$. La suma de cuadrados de la regresión o suma explicada **SCR** tiene $df_R = 1$ grado de libertad, ya que **SCR** queda completamente determinada por un parámetro estimado que es $\widehat{\beta}_1$ (ver ecuación (2.16)). Finalmente, **SCE** tiene $df_E = n - 2$ grados de libertad porque dos restricciones son impuestas sobre las desviaciones de $y_i - \widehat{y}$ como resultado de tomar los estimadores $\widehat{\beta}_0$ y $\widehat{\beta}_1$. Como los grados de libertad cuentan con una propiedad de aditividad tenemos

$$df_T = df_R + df_E.$$

Se puede aplicar la prueba F normal del análisis de varianza para probar la hipótesis $H_0 : \beta_1 = 0$. La estadística de prueba es:

$$F_0 = \frac{\mathbf{SCR}/1}{\mathbf{SCE}/(n-2)} = \frac{\mathbf{CMR}}{\mathbf{CME}}$$

donde $\mathbf{SCR}/1 = \mathbf{CMR}$ se le conoce como *Cuadrado Medio de la Regresión.*

Se rechaza H_0 si:

$$F_0 > F^{\alpha}_{(1,n-2)}$$

Lo anterior puede resumirse en el Cuadro 2.3. La tabla que se presenta es conocida como *tabla ANOVA del análisis de varianza.*

Tabla ANOVA

Fuente de variación	**Suma de Cuadrados**	**Grados de libertad**	**Cuadrados Medios**	F_0
Regresión	$\mathbf{SCR} = \widehat{\beta}_1 \sum_{i-1}^{n} y_i(x_i - \bar{x})$	1	$\mathbf{CMR} = \mathbf{SE}/1$	$\dfrac{\mathbf{CMR}}{\mathbf{CME}}$
Error	$\mathbf{SCE} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \widehat{\beta}_1 \sum_{i=1}^{n} y_i(x_i - \bar{x})$	$n-2$	$\mathbf{CME} = \dfrac{\mathbf{SCE}}{n-2} = \widehat{\sigma}^2$	
Total	$\mathbf{SCT} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$	$n-1$		

Cuadro 2.3: Tabla ANOVA.

2.11. Intervalos de confianza

En esta sección se determinan intervalos de confianza para los parámetros β_0, β_1, σ^2 y adicionalmente se presenta el intervalo de confianza para la respuesta media $E(y)$ para valores dados de x. Para poder determinar intervalos de confianza se requiere mantener el supuesto o hipótesis de que los errores ϵ_i del modelo estén distribuidos normal con media cero y varianza σ^2, e independientes.

Primero se recordará algunos resultados técnicos estudiados en un curso de probabilidad, los cuales facilitarán la construcción de los intervalos de confianza para β_0, β_1 y σ^2.

a) Si $X \sim N(0,1)$, entonces $X^2 \sim \chi^2(1)$.

b) Para $X_1, X_2, \ldots, X_m$ variables aleatorias independientes tales que $X_i \sim \chi^2(n_i)$ para $i = 1, 2, \ldots, m$. Entonces

$$\sum_{i=1}^{m} X_i \sim \chi^2(n_1 + n_2 + \cdots + n_m).$$

c) Para $X_1, X_2, \ldots, X_n$ independientes cada una con distribución $N(\mu, \sigma^2)$. Entonces

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi^2_{(n)}.$$

d) Para X y Y variables aleatorias independientes tales que $X \sim N(0,1)$ y $Y \sim \chi^2(n)$. Entonces

$$\frac{X}{\sqrt{Y/n}} \sim t_{(n)}.$$

Intervalo de confianza para β_0

Para construir el intervalo de confianza para β_0 tenemos que recordar que

a) $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$.

b) $\mathbb{E}(\widehat{\beta}_0) = \beta_0$ (insesgado).

c) $Var(\widehat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$.

Como los errores $\epsilon_i \sim N(0, \sigma^2)$, las observaciones $y_i \sim N(\beta_0 + \beta_1 x, \sigma^2)$ y $\widehat{\beta}_0$ es combinación lineal de las observaciones y_i, la distribución de $\widehat{\beta}_0$ es

$$\widehat{\beta}_0 \sim N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \right)$$

estandarizando

$$\frac{\widehat{\beta}_0 - \beta_0}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}} \sim N(0,1).$$

Si se desconoce el valor de σ^2 se utiliza el estimador $\widehat{\sigma}^2$ y en consecuencia, una cantidad pivotal para β_0 está dada por

$$\frac{\widehat{\beta}_0 - \beta_0}{\sqrt{\mathbf{CME} \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}} \sim t_{(n-2)}$$

así, para el nivel de significancia α

$$\mathbb{P}\left\{ -t^{\alpha/2}_{(n-2)} \leq \frac{\widehat{\beta}_0 - \beta_0}{\sqrt{\mathbf{CME} \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}} \leq t^{\alpha/2}_{(n-2)} \right\} = 1 - \alpha$$

en consecuencia, tenemos que el intervalo de confianza al $100(1-\alpha)\,\%$ para la ordenada al origen del modelo de regresión β_0 es

$$\widehat{\beta}_0 - t^{\alpha/2}_{(n-2)} \sqrt{\mathbf{CME} \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} \leq \beta_0 \leq \widehat{\beta}_0 + t^{\alpha/2}_{(n-2)} \sqrt{\mathbf{CME} \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}.$$

Intervalo de confianza para β_1

Análogamente, para construir el intervalo de confianza para la pendiente del modelo de regresión β_1 utilizamos los siguientes resultados:

a) $\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$.

b) $\mathbb{E}(\widehat{\beta}_1) = \beta_1$ (insesgado).

c) $Var(\widehat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$.

Nuevamente, como los errores $\epsilon_i \sim N(0, \sigma^2)$, las observaciones $y_i \sim N(\beta_0 + \beta_1 x, \sigma^2)$ y $\widehat{\beta}_1$ es combinación lineal de las observaciones y_i, la distribución de $\widehat{\beta}_1$ es

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

estandarizando se tiene que

$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}} \sim N(0,1).$$

Sabemos que $\dfrac{(n-2)}{\sigma^2}\mathbf{CME} \sim \chi^2_{(n-2)}$, de esta forma podemos construir una cantidad pivotal para la pendiente β_1 utilizando la propiedad d), de la siguiente forma,

$$\frac{\dfrac{\widehat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}}}{\sqrt{\dfrac{\dfrac{(n-2)}{\sigma^2}\mathbf{CME}}{(n-2)}}} = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\mathbf{CME}/S_{xx}}} \sim t_{(n-2)}$$

así, para el nivel de significancia α

$$\mathbb{P}\left\{-t^{\alpha/2}_{(n-2)} \leq \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\mathbf{CME}/S_{xx}}} \leq t^{\alpha/2}_{(n-2)}\right\} = 1 - \alpha.$$

En consecuencia, el intervalo de confianza al $100(1-\alpha)\,\%$ para la pendiente β_1 está dado por

$$\widehat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\mathbf{CME}}{S_{xx}}} \leq \beta_1 \leq \widehat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\mathbf{CME}}{S_{xx}}}.$$

Intervalo de confianza para σ^2

Para determinar un intervalo de confianza para σ^2, una cantidad pivotal es

$$\frac{(n-2)}{\sigma^2}\mathbf{CME} \sim \chi^2_{(n-2)},$$

así, para el nivel de significancia α

$$\mathbb{P}\left\{\chi^2_{1-\alpha/2,n-2} \leq \frac{(n-2)}{\sigma^2}\mathbf{CME} \leq \chi^2_{\alpha/2,n-2}\right\} = 1 - \alpha,$$

en consecuancia, el intervalo de confianza al $100(1-\alpha)\,\%$ para σ^2 es

$$\frac{(n-2)\mathbf{CME}}{\chi^2_{\alpha/2,n-2}} \leq \sigma^2 \leq \frac{(n-2)\mathbf{CME}}{\chi^2_{1-\alpha/2,n-2}}.$$

Intervalo de confianza para la respuesta media

El uso de modelos de regresión, por lo general, nos sirven para estimar la media de la variable de respuesta y para un valor x específico. Sea x_0 el valor de la variable regresora, con x_0 en el rango de las x_i utilizadas para el modelo de regresión, por lo que deseamos estimar la respuesta media $\mathbb{E}(y|x_0)$. Un estimador puntual insesgado de $\mathbb{E}(y|x_0)$ se encuentra en el modelo ajustado como

$$\widehat{\mathbb{E}}(y|x_0) \equiv \widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0.$$

Nótese que $\widehat{y}_0$ es una variable aleatoria con distribución normal porque es una combinación lineal de las observaciones y_i, porque se podrá obtener un intervalo de confianza del $100(1-\alpha)\,\%$ para $\mathbb{E}(y|x_0)$. La varianza de $\widehat{y}_0$ es

$$\begin{aligned} Var(\widehat{y}_0) &= Var(\widehat{\beta}_0 + \widehat{\beta}_1 x_0) \\ &= Var[(\bar{y} - \widehat{\beta}_1 \bar{x}) + \widehat{\beta}_1 x_0] \\ &= Var[\bar{y} - \widehat{\beta}_1 (x_0 - \bar{x})] \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right], \end{aligned}$$

entonces la distribución muestral de

$$\frac{\widehat{y}_0 - \mathbb{E}(y|x_0)}{\sqrt{\mathbf{CME}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}}$$

es t con $n-2$ grados de libertad y en consecuencia, el intervalo de confianza del $100(1-\alpha)\,\%$ para la respuesta media en el punto $x = x_0$ es

$$\begin{aligned} \widehat{y}_0 - t_{\alpha/2,n-2} \sqrt{\mathbf{CME}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)} &\leq \mathbb{E}(y|x_0) \\ &\leq \widehat{y}_0 + t_{\alpha/2,n-2} \sqrt{\mathbf{CME}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}. \end{aligned}$$

2.12. Estimación por máxima verosimilitud

El método de mínimos cuadrados puede ser utilizado para estimar los parámetros de un modelo de regresión lineal, independientemente de la distribución de los errores ϵ_i. Este método produce buenos estimadores insesgados para β_0 y β_1. Otros procedimientos estadísticos, tales como la comprobación de hipótesis y la construcción de intervalos de confianza, suponen que los errores tienen distribución normal. Si se conoce cómo se distribuyen los errores, puede usarse un método alternativo para la estimación de los parámetros, como el método de máxima verosimilitud. Considere los datos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Si

suponemos que los errores en el modelo de regresión se distribuyen $N(0,\sigma^2)$ e independientes entre sí, entonces las observaciones y_i son variables aleatorias independientes y con distribución normal con media $\beta_0+\beta_1 x_i$ y varianza σ^2. La función de verosimilitud se encuentra a partir de la distribución conjunta de las observaciones. Teniendo en cuenta la distribución conjunta con las observaciones dadas y los parámetros β_0,β_1 y σ^2 constantes desconocidas tenemos la función de verosimilitud. Para el modelo de regresión lineal simple con errores normales, la función de verosimilitud es

$$\begin{aligned} L(\beta_0,\beta_1,\sigma^2,\underline{y},\underline{x}) &= \prod_{i=1}^{n}(2\pi\sigma^2)^{-1/2}exp\left[-\frac{1}{2\sigma^2}(y_i-\beta_0-\beta_1x_i)^2\right] \\ &= (2\pi\sigma^2)^{-n/2}exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2\right]. \end{aligned}$$

Si se denotan los estimadores de máxima verosimilitud, por $\tilde{\beta}_0,\tilde{\beta}_1$ y $\tilde{\sigma}^2$, que maximizan a L, o equivalentemente, $ln(L)$, tenemos

$$\begin{aligned} lnL(\beta_0,\beta_1,\sigma^2,\underline{y},\underline{x}) &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2-\frac{n}{2}ln(2\pi\sigma^2) \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2-\left(\frac{n}{2}\right)ln2\pi-\left(\frac{n}{2}\right)ln\ \sigma^2 \end{aligned}$$

y los estimadores de máxima verosimilitud $\tilde{\beta}_0,\tilde{\beta}_1$ y $\tilde{\sigma}^2$ deben cumplir

$$\left.\frac{\partial\ lnL}{\partial\beta_0}\right|_{\tilde{\beta}_0,\tilde{\beta}_1,\tilde{\sigma}^2}=\frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}(y_i-\tilde{\beta}_0-\tilde{\beta}_1x_i)=0, \tag{2.17}$$

$$\left.\frac{\partial\ lnL}{\partial\beta_1}\right|_{\tilde{\beta}_0,\tilde{\beta}_1,\tilde{\sigma}^2}=\frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}(y_i-\tilde{\beta}_0-\tilde{\beta}_1x_i)x_i=0, \tag{2.18}$$

y por el principio de invarianza, derivando respecto a σ tenemos

$$\left.\frac{\partial\ lnL}{\partial\sigma}\right|_{\tilde{\beta}_0,\tilde{\beta}_1,\tilde{\sigma}^2}=\frac{1}{\tilde{\sigma}^3}\sum_{i=1}^{n}(y_i+\tilde{\beta}_0+\tilde{\beta}_1x_i)^2-\frac{n}{\tilde{\sigma}}=0. \tag{2.19}$$

De la ecuación (2.17) tenemos que

$$\sum_{i=1}^{n}y_i-n\tilde{\beta}_0-\tilde{\beta}_1\sum_{i=1}^{n}x_i=0,$$

entonces

$$n\tilde{\beta}_0=\sum_{i=1}^{n}y_i-\tilde{\beta}_1\sum_{i=1}^{n}x_i,$$

por lo tanto $\tilde{\beta}_0=\bar{y}-\tilde{\beta}_1\bar{x}$. Sustituyendo esta igualdad en (2.18) tenemos

$$\sum_{i=1}^{n}x_iy_i-(\bar{y}-\tilde{\beta}_1\bar{x})\sum_{i=1}^{n}x_i-\tilde{\beta}_1\sum_{i=1}^{n}x_i^2=0$$

$$\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i + \tilde{\beta}_1 \bar{x}\sum_{i=1}^{n} x_i - \tilde{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0,$$

despejando $\tilde{\beta}_1$ se tiene

$$\begin{aligned}
\tilde{\beta}_1 &= \frac{\bar{y}\displaystyle\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i}{\bar{x}\displaystyle\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i^2} \\
&= \frac{\displaystyle\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i\right)}{\displaystyle\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i} \\
&= \frac{\displaystyle\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \bar{x}}{\displaystyle\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \\
&= \frac{\displaystyle\sum_{i=1}^{n} y_i (x_i - \bar{x})}{\displaystyle\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\
&= \frac{\displaystyle\sum_{i=1}^{n} y_i (x_i - \bar{x})}{\displaystyle\sum_{i=1}^{n} (x_i - \bar{x})^2},
\end{aligned}$$

por lo que $\tilde{\beta}_1 = \dfrac{\displaystyle\sum_{i=1}^{n} y_i (x_i - \bar{x})}{\displaystyle\sum_{i=1}^{n} (x_i - \bar{x})^2} = \dfrac{S_{xy}}{S_{xx}}$.

Por último, de la ecuación (2.19) se tiene

$$\frac{1}{\tilde{\sigma}}\left[\frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2 - n\right] = 0$$

entonces

$$\frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2 = n$$

por lo que

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2.$$

Nótese que los estimadores obtenidos por el método de máxima verosimilitud, $\tilde{\beta}_0$ y $\tilde{\beta}_1$ son idénticos a los estimadores obtenidos por el método de mínimos cuadrados. Además, $\tilde{\sigma}^2$ es un estimador sesgado para $\widehat{\sigma}^2$. La relación que hay entre estos dos estimadores es $\tilde{\sigma}^2 = [(n-2)/n]\widehat{\sigma}^2$. El sesgo es pequeño si n es moderadamente grande. En general, se usa el estimador insesgado $\widehat{\sigma}^2$.
En general, los estimadores de máxima verosimilitud tienen mejores propiedades estadísticas que los estimadores por mínimos cuadrados. Los estimadores de máxima verosimilitud son insesgados (incluyendo $\tilde{\sigma}^2$ que asintóticamente es insesgado, o insesgado cuando n crece) y con varianza mínima en comparación con otros estimadores. Además son estimadores consistentes (la consistencia es una propiedad que indica que los estimadores difieren del verdadero valor del parámetro por una cantidad muy pequeña a medida en que n crece) y son estadísticas suficientes (esto implica que cuentan con toda la información acerca de los parámetros de la muestra original de tamaño n).

2.13. Regresión lineal por el origen

En regresión, algunas situaciones parecen indicar que una línea recta que pasa por el origen se debe ajustar a los datos. Un modelo sin *intersección* es aquel modelo de regresión lineal simple sin el término β_0 y a menudo es apropiado en el análisis de datos en la industria química y otros procesos de fabricación. Por ejemplo, el rendimiento del proceso químico es cero cuando la temperatura de operación del proceso es igual a cero.

Dadas n observaciones $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, el modelo sin intersección es

$$y_i = \beta_1 x_i + \epsilon_i \quad \text{para} \quad i = 1, 2, \ldots, n.$$

Si $\widehat{\beta}_1$ es un estimador del parámetro β_1 entonces la *recta estimada o ajustada* es

$$\widehat{y}_i = \widehat{\beta}_1 x_i \quad \text{para} \quad i = 1, 2, \ldots, n$$

Nuevamente, el objetivo es minimizar las distancias $y_i - \widehat{y}_i$ para $i = 1, 2, \ldots, n$ (véase Figura 2.5) Entonces tenemos que la suma de los cuadrados de los residuales **SCE** está dada por

$$\mathbf{SCE} = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

El principio de mínimos cuadrados elige a $\widehat{\beta}_1$ tal que minimiza a **SCE**, y utilizando nuevamente técnicas de cálculo diferencial. Vemos que

$$\begin{aligned}\mathbf{SCE} &= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \\ &= \sum_{i=1}^{n}(y_i - \widehat{\beta}_1 x_i)^2\end{aligned}$$

Derivando esta ecuación con respecto a $\widehat{\beta}_1$, e igualando a cero

$$\frac{d\ \mathbf{SCE}}{d\widehat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \widehat{\beta}_1 x_i)x_i = 0$$

entonces

$$\sum_{i=1}^{n} y_i x_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \tag{2.20}$$

donde la ecuación (2.20) es la ecuación normal. Despejando $\widehat{\beta}_1$ obtenemos que

$$\widehat{\beta}_1 = \frac{\displaystyle\sum_{i=1}^{n} y_i x_i}{\displaystyle\sum_{i=1}^{n} x_i^2}$$

estimador insesgado para β_1. El estimador para σ^2 es

$$\begin{aligned}\widehat{\sigma}^2 \equiv \mathbf{CME} &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \widehat{\beta}_1 x_i)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i^2 - 2\widehat{\beta}_1 y_i x_i + \widehat{\beta}_i^2 x_i^2) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - 2\widehat{\beta}_1 \sum_{i=1}^{n} y_i x_i + \widehat{\beta}_1^2 \sum_{i=1}^{n} x_i^2\right) \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \widehat{\beta}_1\left(2\sum_{i=1}^{n} y_i x_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2\right)\right] \\ &= \frac{1}{n-1}\left\{\sum_{i=1}^{n} y_i^2 - \widehat{\beta}_1\left[2\sum_{i=1}^{n} y_i x_i - \left(\frac{\displaystyle\sum_{i=1}^{n} y_i x_i}{\displaystyle\sum_{i=1}^{n} x_i^2}\right)\sum_{i=1}^{n} x_i^2\right]\right\} \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - \widehat{\beta}_1 \sum_{i=1}^{n} y_i x_i\right)\end{aligned}$$

por lo tanto $\widehat{\sigma}^2 = \frac{1}{n-1}\left(\sum_{i=1}^n y_i^2 - \widehat{\beta}_1 \sum_{i=1}^n y_i x_i\right)$ con $n-1$ grados de libertad.

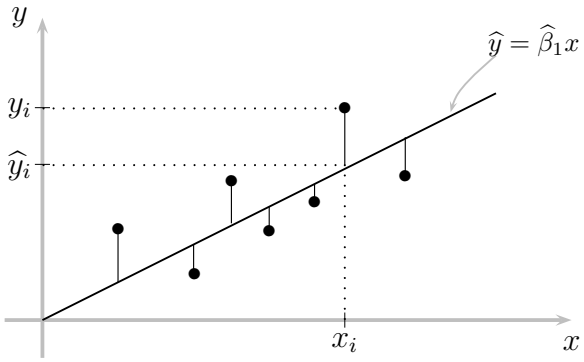


Figura 2.5: Ajuste de una línea recta que pasa por el origen.

Intervalos de confianza

Haciendo el supuesto de normalidad en los errores, podemos probar hipótesis, construir intervalos de confianza y de realizar predicciones para el modelo sin intersección o sin intercepto. El intervalo de confianza al $100(1-\alpha)\,\%$ para β_1 es

$$\widehat{\beta}_1 - t_{\alpha/2,n-1}\sqrt{\frac{\mathbf{CME}}{\sum_{i=1}^n x_i^2}} \leq \beta_1 \leq \widehat{\beta}_1 + t_{\alpha/2,n-1}\sqrt{\frac{\mathbf{CME}}{\sum_{i=1}^n x_i^2}} \tag{2.21}$$

El intervalo de confianza al $100(1-\alpha)\,\%$, para $\mathbb{E}(y|x_0)$, la respuesta media para $x = x_0$, es

$$\widehat{y}_0 - t_{\alpha/2,n-1}\sqrt{\frac{x_0^2\mathbf{CME}}{\sum_{i=1}^n x_i^2}} \leq \mathbb{E}(y|x_0) \leq \widehat{y}_0 + t_{\alpha/2,n-1}\sqrt{\frac{x_0^2\mathbf{CME}}{\sum_{i=1}^n x_i^2}} \tag{2.22}$$

El intervalo de predición a $100(1-\alpha)\,\%$ para una observación futura $x = x_0$, por ejemplo, y_0, es

$$\widehat{y}_0 - t_{\alpha/2,n-1}\sqrt{\mathbf{CME}\left(1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}\right)} \leq y_0 \leq \widehat{y}_0 + t_{\alpha/2,n-1}\sqrt{\mathbf{CME}\left(1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}\right)} \tag{2.23}$$

La confianza del intervalo (2.22) y la confianza del intervalo de predicción de (2.23) aumentan, a medida que x_0 también aumenta. Además, la longitud del intervalo de confianza (2.22) en $x = 0$ es cero, porque el modelo asume que la media de y en $x = 0$. Este comportamiento es muy diferente al modelo cuando

hay intercepto. El intervalo (2.23) tiene una longitud distinta de cero en $x = 0$, ya que el error aleatorio en la observación futura debe tomarse en cuenta.

Es fácil de emplear mal el modelo sin intersección, sobre todo en situaciones donde los datos están en una región de x que inicia en el origen. El diagrama de dispersión a veces sirve de orientación para decidir si es adecuado o no el modelo sin intersección. Alternativamente, podemos usar los dos modelos y elegir entre ellos con base en el mejor ajuste. Si la hipótesis $\beta_0 = 0$ no puede ser rechazada en el modelo con intercepto, esto es una indicación de que el ajuste se puede mejorar utilizando el modelo sin intersección. El cuadrado medio de los residuales es una manera útil para comparar la calidad del ajuste. El modelo que tenga el cuadrado medio de los residuales más pequeño es la mejor opción en el sentido de que minimiza la estimación de la varianza de Y sobre la regresión lineal.

Generalmente R^2 no es una buena comparación estadística de los dos modelos. Para el modelo con intercepto tenemos

$$\begin{aligned} R^2 &= \frac{\mathbf{SCR}}{\mathbf{SCT}} \\ &= \frac{\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \\ &= \frac{\text{variación en } y \text{ explicada por la regresión}}{\text{variación total observada en } y} \end{aligned}$$

Nótese que R^2 indica la proporción de la variabilidad en torno a $\bar{y}$ explicada por la regresión. Para el caso en donde no hay intersección se tiene que la *identidad fundamental* es

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \widehat{y}_i^2 + \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

análogamente, para el modelo sin intersección R^2 es

$$R_0^2 = \frac{\sum_{i=1}^{n} \widehat{y}_i^2}{\sum_{i=1}^{n} y_i^2}.$$

La estadística R_0^2 indica la proporción de la variabilidad en torno al origen (cero) correspondiente a la regresión. De vez en cuando encontramos que R_0^2 es mayor que R^2 a pesar de que el cuadrado medio de los residuales (es una medida razonable de la calidad global del ajuste) para el modelo con intersección es menor que el cuadrado medio de los residuales para el modelo sin intersección. Esto se debe a que R_0^2 se calcula usando las sumas de cuadrados corregidas.

Existen diversas maneras para definir R_0^2 para el modelo sin intersección. Una posibilidad es

$$R_0^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}.$$

Sin embargo, en caso de que $\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ es grande, R_0^2, puede ser negativo. Se recomienda utilizar **CME** como base para la comparación entre los modelos con y sin intersección.

Ejemplo 2.5. *En este ejemplo se ajusta un modelo de regresión que pasa por el origen para el siguiente conjunto de datos.*

y	x
10.15	25
2.96	6
3.00	8
6.88	17
0.28	2
5.06	13
9.14	23
11.86	30
11.69	28
6.04	14
7.57	19
1.74	4
9.38	24
0.16	1
1.84	5

Se realiza la prueba de hipótesis para ver si la pendiente β_1 es distinta de cero para el modelo sin intersección, se ajusta un modelo con intersección y se comparan los errores cuadráticos para ver qué modelo es el más adecuado.

Calculando la pendiente para el modelo sin intersección

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} = \frac{1850.73}{4575} = 0.4026$$

Por lo tanto se tiene la ecuación

$$\widehat{y} = 0.4026x$$

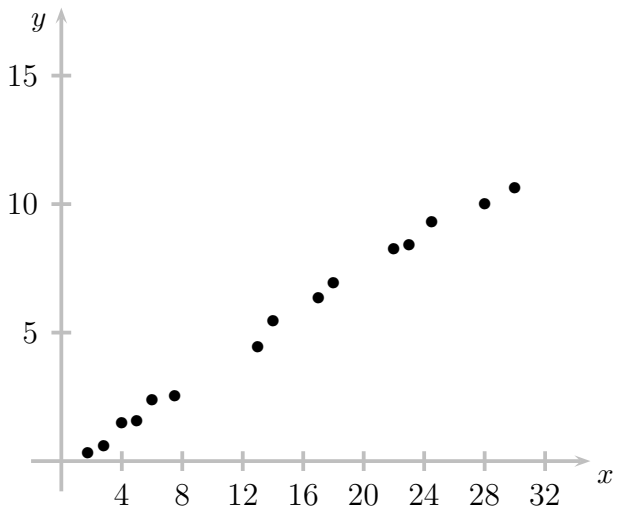


Figura 2.6: Diagrama de dispersión de los datos.

El cuadrado medio de los residuales u errores para este modelo es **CME** = 0.0893 y $R_0^2 = 0.9983$, además la estadística para la prueba $H_0 : \beta_1 = 0$ es

$$t_0 = \frac{\widehat{\beta}_1}{\sqrt{\dfrac{\mathbf{CME}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}}} = 91.133$$

con un nivel de significancia $\alpha = 0.01$. También podemos ajustar a un modelo con intersección a los datos con fines comparativos. Así se tiene

$$\widehat{y} = -0.0938 + 0.4071x$$

La estadística para la prueba $H_0 : \beta_0 = 0$ es $t_0 = -0.65$, que no es significativa, implica que el modelo sin intersección puede proporcionar un mejor ajuste. El cuadrado medio de los residuales para el modelo con intersección es **CME** = 0.0931 y $R^2 = 0.9997$. El **CME** para el modelo sin intersección es menor que el **CME** para el modelo con intersección, así podemos concluir que el modelo sin intersección es mejor. La estadística de R^2 no se pueden comparar directamente como se hace con el cuadrado medio de los residuales.

2.14. Modelo lineal simple mediante matrices

Hasta el momento hemos estudiado el modelo lineal simple usando expresiones algebraicas ordinarias. La manera en que se presentan resultados análogos para modelos de *regresión lineal múltiple* es mediante álgebra de matrices. En esta sección usaremos matrices para representar algunos de los resultados anteriores para el modelo lineal simple y esta notación permitirá continuar con estudios de modelos más generales. Supongamos que tenemos el modelo muestral de regresión lineal simple

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

y sabemos que $\epsilon_i \sim N(0, \sigma^2)$ independientes para $i = 1, 2 \ldots, n$. Ahora escribimos cada uno de los elementos como

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_i + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_i + \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_i + \epsilon_n. \end{aligned} \tag{2.24}$$

entonces tenemos n ecuaciones que determinan a cada observación y_i en función de β_0, β_1 y sus respectivas x_i y ϵ_i. Si definimos las siguientes matrices

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

entonces, las n ecuaciones (2.24) se pueden escribir matricialmente como

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Análogamente, las ecuaciones normales (2.6) y (2.7) de la sección 2.2 se pueden escribir simultáneamente utilizando matrices de la siguiente manera

$$(\boldsymbol{X}'\boldsymbol{X})\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{Y}$$

y las soluciones son

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

Además, se puede demostrar que $\mathbf{SCE} = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ se puede expresar como

$$\mathbf{SCE} = \boldsymbol{Y}'\boldsymbol{Y} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}.$$

Nótese que $\boldsymbol{Y}'\boldsymbol{Y} = \sum_{i=1}^{n} y_i^2$.

2.15. Diagnóstico del modelo

Al seleccionar un modelo, en este caso, el modelo de regresión lineal simple, comúnmente no se cuenta con la certeza de que el modelo sea adeacuado, ya que puede no cumplir con una o más hipótesis, por ejemplo, la linealidad del modelo o la normalidad de los errores. De esta menera, es importante saber si el modelo es adeacuado para los datos.

Se conoce como diagnóstico o comprobación del modelo, llevar a cabo pruebas sobre los supuestos del modelo de regresión lineal los cuales son:

a) La relación entre la variable respuesta y y las variables explicativas es lineal, al menos aproximadamente.

b) El término del error ϵ tiene media cero, $\mathbb{E}(\epsilon) = 0$.

c) El término del error ϵ tiene varianza constante, $Var(\epsilon) = \sigma^2$ (homoscedasticidad).

d) Los errores no están correlacionados, $Cov(\epsilon_i, \epsilon_j) = 0$ para $i \neq j$.

e) Los errores se distribuyen normal, $\epsilon \sim N(0, \sigma^2)$.

Cuando se lleva a cabo un ajuste, hay que considerar siempre la validez de los supuestos del modelo, por lo que se tendría que examinar si el modelo es adecuado. Si no llega a violar dichos supuestos se puede producir un modelo inestable e inadecuado teniendo como resultado un modelo totalmente distinto y tener conclusiones opuestas. Es importante decir que no se pueden detectar incongruencias de los supuestos examinando los estadísticos t, F o el coeficiente de determinación R^2. Así que, ordinariamente, el análisis que se lleva a cabo es sobre los residuales o errores del modelo

$$e_i = y_i - \widehat{y}_i \quad \text{para} \quad i = 1, 2, \ldots, n$$

los cuales se consideran como desviaciones entre los valores observados y el ajuste; son una medida de la variabilidad de la variable respuesta no explicada por el modelo de regresión. También pueden consideranse como los valores observados de los errores del modelo y por esto, el análisis de los residuales es una manera efectiva de descubrir insuficiencias en el modelo.
Considerando lo anterior, podemos afirmar que si nuestro modelo ajustado es correcto, los residuales deberían mostrar un comportamiento que confirme las suposiciones hechas, o al menos, no mostrar una tendencia que invalide alguna de ellas. Así, los residuales después de ser examinados nos deberían permitir concluir:

a) Alguno de los supuestos del modelo parece ser violado.

b) Ninguno de los supuestos del modelo parece ser violado.

Hay que notar que el inciso b) no significa que los supuestos sean correctos, sino que a partir del análisis de los datos, no se cuenta con ninguna razón para decir que algún supuesto no sea correcto.
Los residuales tienen media cero y varianza estimada:

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\mathbf{SCE}}{n-2} = \mathbf{CME}.$$

Algunas veces es útil trabajar con residuales reescalados. Estos residuales son útiles para detectar valores *extremos o aberrantes* (valores raros), mejor conocidos en inglés como *outliers*.

Residuales estandarizados

La varianza aproximada de los residuales se estima utilizando **CME**, así que un cambio de escala de los residuales será el de los residuales estandarizados.

Definición 2.3. *Los residuales estandarizados son definidos como*

$$d_i = \frac{e_i}{\sqrt{\boldsymbol{CME}}} \quad para \quad i = 1, 2, \ldots, n$$

Los residuales estandarizados tienen media cero y varianza muy cercana a uno. Un residuo estandarizado grande ($d_i > 3$) indica potencialmente que el valor es *aberrante o extremo.*

Residuales studentizados

Se puede lograr un mejor cambio de escala de los residuales dividiendo e_i entre la desviación estándar exacta del i-ésimo residuo.

Definición 2.4. *Los residuales studentizados se definen como*

$$r_i = \frac{e_i}{\sqrt{\boldsymbol{CME}\left[1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)\right]}} \quad para \quad i = 1, \ldots, n$$

En muestras pequeñas los residuales studentizados son a menudo más apropiados que los residuales estandarizados, puesto que las diferencias en la varianza de ellos pueden ser mayores, sin embargo, en muestras grandes estos residuales no difieren mucho de los residuales estandarizados. Tienen distribución aproximandamente normal y son más útiles para detectar observaciones aberrantes o influyentes.

A continuación se presentan dos métodos gráficos para examinar los residuales con el propósito de detectar si se cumplen o no los supuestos del modelo de regresión.

Gráficas de residuales

a) **Gráficas de probabilidad normal.** Una gráfica de probabilidad normal, mejor conocida como Q-Q Plot, se utiliza para analizar que los errores sigan una distribución normal, ya que si no, puede generar serios problemas pues las estadísticas t y F, los intervalos de confianza y las predicciones dependen del supuesto de normalidad. Si los errores se distribuyen normal, los puntos mostrarán aproximadamente una línea recta. Si los errores provienen de una distribución cuya cola sea más o menos pesada que la de una normal, el ajuste puede mostrarse en un pequeño grupo de observaciones. En la Figura 2.7 el inciso a) muestra la gráfica de probabilidad ideal, los puntos caen aproximadamente sobre una línea recta. La gráfica del inciso b) muestra puntos donde las colas de la distribución son más ligeras. La del inciso c) muestra el comportamiento de una distribución con colas pesadas y por último, las del inciso d) y e) exiben sesgo positivo y sesgo negativo, respectivamente.

Para analizar el supuesto de normalidad, también se sugiere realizar una análisis con ayuda de histogramas, diagramas de cajas y pruebas no paramétricas para normalidad.

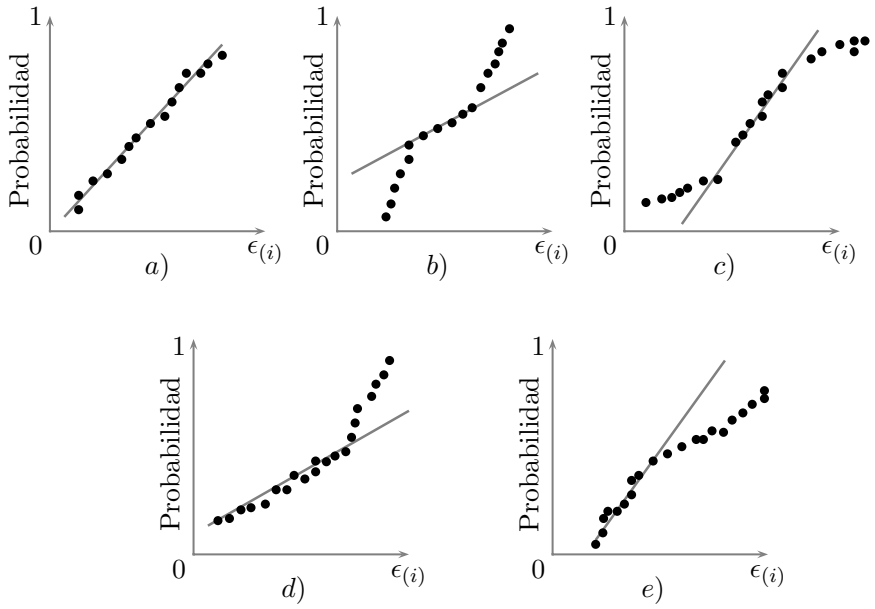


Figura 2.7: Gráficas de probabilidad normal QQ-plots.

b) **Gráficas de residuales contra los valores ajustados.** Este tipo de gráficas son útiles para corroborar el supuesto de varianza constante en los errores ϵ_i del modelo propuesto. Se realizan con un diagrama de dispersión de los residuales e_i o los residuales reescalados d_i y r_i contra los correspondientes valores ajustados $\widehat{y}_i$, los cuales sirven para detectar anomalías o insuficiencias en el modelo.

   Si la gráfica de residuales e_i (o los residuales reescalados d_i y r_i) contra los valores ajustados $\widehat{y}_i$ no tiene ningún patrón, significa que no hay defectos obvios en el modelo y podemos considerar que los errores tienen varianza constante. Si existe algún patrón en los datos esto puede indicar que la varianza no es constante, que no hay linealidad o que tal vez se necesiten otras variables explicativas en el modelo. En la Figura 2.8 se muestran algunos de los patrones que se pueden observar.

2.16. Transformaciones de variables

En la práctica, en muchos de los casos que se presentan no es posible ajustar los datos a una línea recta y satisfacer los supuestos del modelo de regresión, en estos casos es necesario trabajar con transformaciones de las variables para

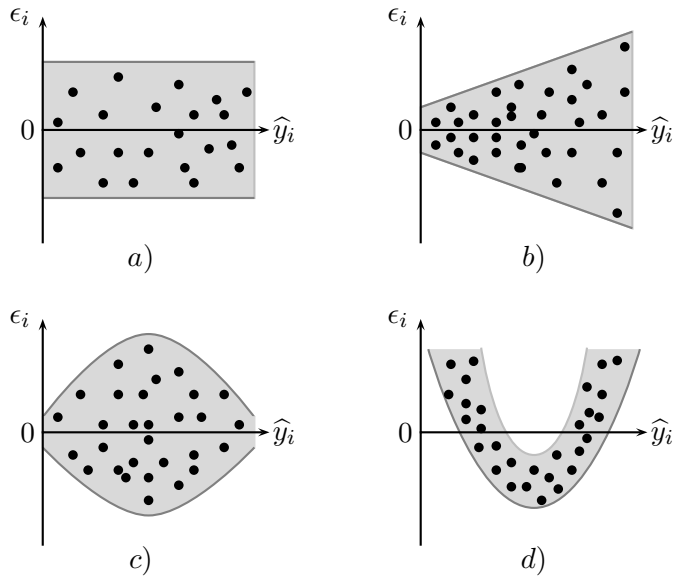


Figura 2.8: Patrones de gráficas de residuales contra valores ajustados. *a*) Satisfactorio; *b*) embudo; *c*) doble arco; *d*) no lineal.

encontrar el mejor ajuste posible. Estas transformaciones se deben a la relación que existe entre las variables, por ejemplo, en el caso de modelos de crecimiento poblacional la relación entre variables no es lineal sino una relación exponencial.

Para estos casos se estudiarán dos tipos de transformaciones, las que permiten linealizar el modelo cuando el supuesto de *linealidad* entre variables no se satisface y las transformaciones que estabilizan la varianza cuando el supuesto de *varianza constante* no se cumple.

Transformaciones para linealizar el modelo

Al realizar el diagnóstico para un modelo de regresión lineal simple, uno de los supuestos básicos que debe cumplir el modelo propuesto que va a describir a los datos es la relación lineal entre y y los valores de la variable regresora x. Cuando este supuesto no se satisface, se puede describir la no linealidad con diagramas de dispersión o con gráficas de residuales. En muchos casos, una función no lineal se puede linealizar haciendo una transformación adecuada. Esos modelos son conocidos como ***modelos no lineales***.

Cuando se observa en el diagrama de dispersión algún tipo de curvatura, se podrá ajustar en el comportamiento de los datos la forma linealizada de la función para representar a estos. Por ejemplo si se tiene la función

$$y = \beta_0 e^{\beta_1 x} \epsilon$$

se puede linealizar o transformar a una ecuación lineal aplicando $y' = ln(y)$ obteniendo

$$ln(y) = ln(\beta_0) + \beta_1 x + ln(\epsilon).$$

En el Cuadro 2.4 se muestra un resumen de las transformaciones más utilizadas en el análisis de regresión lineal simple y en las Figuras 2.9, 2.10, 2.11 y 2.12 se muestra el comportamiento de las gráficas de las cuatro primeras funciones linealizables.

Función linealizable	**Transformación**	**Forma lineal**
$y = \beta_0 x^{\beta_1}$	$y' = log(y),\ x' = log(x)$	$y' = log(\beta_0) + \beta_1 x'$
$y = \beta_0 e^{\beta_1 x}$	$y' = ln(y)$	$y' = ln(\beta_0) + \beta_1 x$
$y = \beta_0 + \beta_1 log(x)$	$x' = log(x)$	$y' = \beta_0 + \beta_1 x'$
$y = \dfrac{x}{\beta_0 x + \beta_1}$	$y' = 1/y,\ x' = 1/x$	$y' = \beta_0 - \beta_1 x'$
$y = \dfrac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$	$y' = ln\left(\dfrac{y}{1-y}\right)$	$y' = \beta_0 + \beta_1 x$

Cuadro 2.4: Funciones linealizables y su transformación.

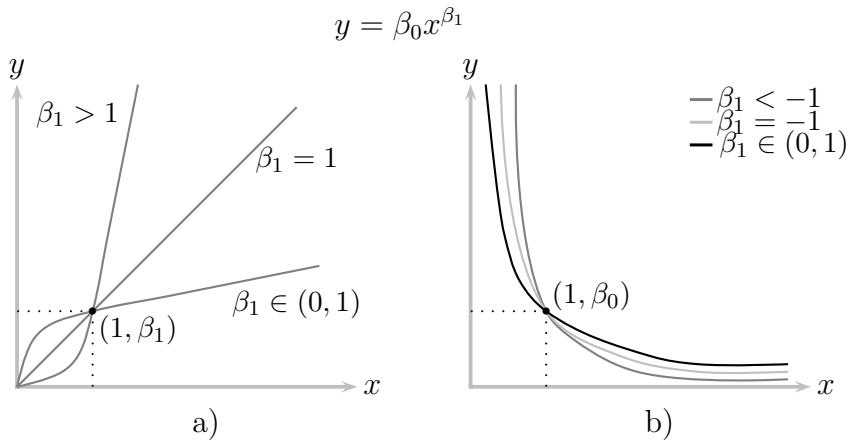


Figura 2.9: Gráficas correspondientes a la función $y = \beta_0 x^{\beta_1}$.

$$y = \beta_0 e^{\beta_1 x}$$

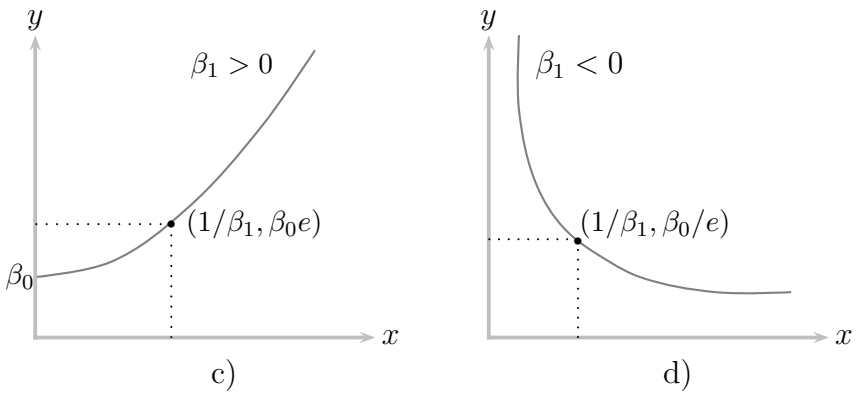


Figura 2.10: Gráficas correspondientes a la función $y = \beta_0 e^{\beta_1 x}$.

$$y = \beta_0 + \beta_1 log x$$

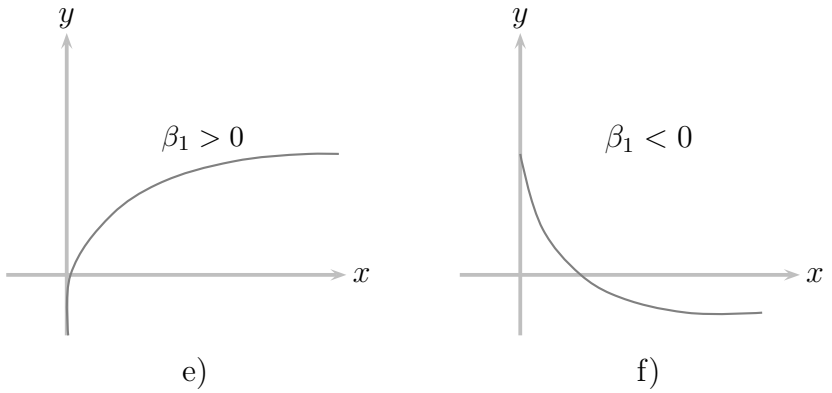


Figura 2.11: Gráficas correspondientes a la función $y = \beta_0 + \beta_1 log x$.

$$y = \frac{x}{\beta_0 x - \beta_1}$$

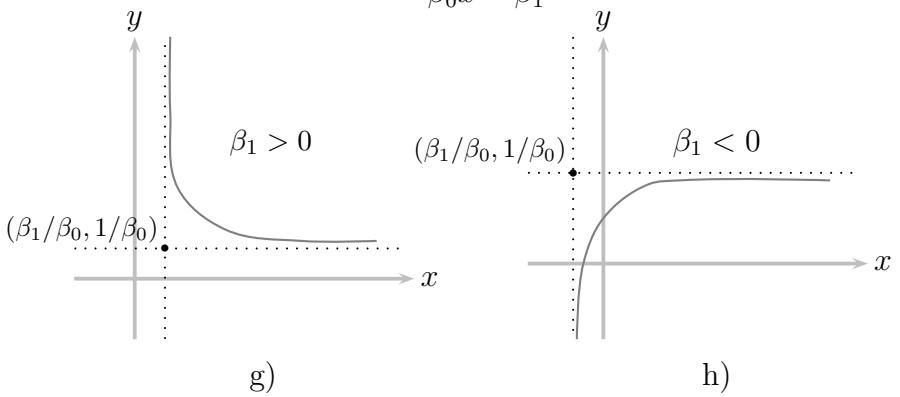


Figura 2.12: Gráficas correspondientes a la función $y = \dfrac{x}{\beta_0 x - \beta_1}$.

Transformaciones para estabilizar la varianza

Una razón frecuente para que un modelo de regresión lineal simple no cumpla con el supuesto de *varianza constante* es cuando la variable dependiente o de

respuesta y puede tener una distribución de probabilidad cuya varianza esté determinada por una transformación de su media, es decir,

$$Var(y) = g(\mathbb{E}(y)).$$

De esta manera la varianza de y cambiará cuando la variable regresora x cambie, o bien, la varianza no es constante. Normalmente la distribución de y no es normal bajo estas condiciones haciendo que sea inválida la prueba de significancia ya que se basa en hipótesis de normalidad. Un ejemplo podría ser cuando y sigue una distribución exponencial, la varianza de y es igual a su media al cuadrado, es decir la relación que existe es cuadrática. Así, si y sigue una distribución exponencial, se podría hacer la transformación $y' = ln(y)$ en función de x.

A las transformaciones que se realizan para mantener la varianza constante en un modelo de regresión lineal se le conocen como *transformaciones estabilizadoras de la varianza*, estas transformaciones son útiles para normalizar las variables y llevar a cabo las pruebas de significancia y realizar un mejor diagnóstico del modelo propuesto. En el Cuadro 2.5 se muestran algunas transformaciones estabilizadoras para la varianza más frecuentes.

Relación entre $Var(y)$ **y** $\mathbb{E}(y)$	**Transformación** (y')
$Var(y) \propto$ constante	$y' = y$
$Var(y) \propto \mathbb{E}(y)$	$y' = \sqrt{y}$
$Var(y) \propto [\mathbb{E}(y)]^2$	$y' = ln(y)$
$Var(y) \propto [\mathbb{E}(y)]^3$	$y' = y^{-1/2}$
$Var(y) \propto [\mathbb{E}(y)]^4$	$y' = y^{-1}$
$Var(y) \propto \mathbb{E}(y)[1 - \mathbb{E}(y)]$	$sen^{-1}(\sqrt{y})$

Cuadro 2.5: Transformaciones usuales estabilizadoras de varianza.

Nótese que en el segundo caso del Cuadro 2.5, cuando la varianza está relacionada con la esperanza de la variable respuesta, se puede aplicar la transformación $y' = \sqrt{y}$ cuando y sigue una distribución Poisson en el modelo de regresión lineal simple. Para el último caso, se puede aplicar $sen^{-1}(\sqrt{y})$ cuando la distribución de y es una proporción $(0 \leq y_i \leq 1)$.

2.17. Ejemplo de un ajuste de regresión utilizando R

En esta sección se mostrará cómo ajustar un modelo de regresión lineal simple utilizando el paquete estadístico R para el siguiente conjunto de datos:

x	y
1	23
2	29
3	49
4	64
4	74
5	87
6	96
6	97
7	109
8	119
9	149
9	145
10	154
10	166

Se realiza la lectura de los datos creando los siguientes vectores:

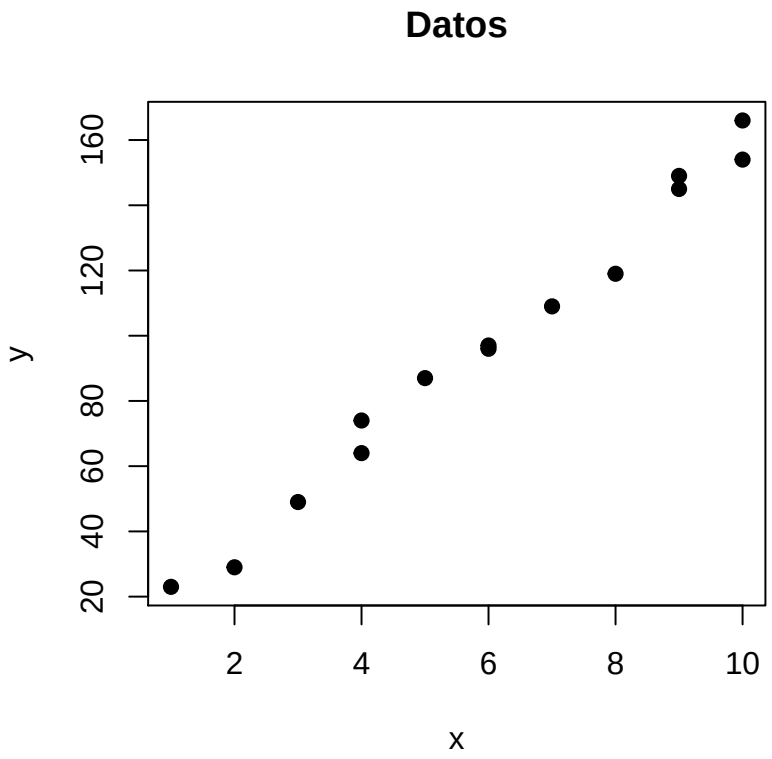
```
# Vectores de datos
> x <- c(1,2,3,4,4,5,6,6,7,8,9,9,10,10)
> y <- c(23,29,49,64,74,87,96,97,109,119,149,145,154,166)
```

Es importante recordar que para comenzar el análisis se recomienda utilizar un diagrama de dispersión de los datos y obtener el coeficiente de correlación para tener un panorama general de la relación lineal entre variables.

Diagrama de dispersión

Utilizando el comando `plot()` podemos ver la dispersión que hay entre variables de la siguiente forma:

```
# Diagrama de dispersión
> plot(x, y, main="Datos", pch=19)
```



En la gráfica se puede observar una relación lineal positiva ya que conforme aumenta la variable x también aumenta la variable y.

Coeficiente de correlación

Ahora utilizaremos la función `cor()` para ver si existe dependencia lineal entre las variables.

```
# Coeficiente de correlación
> cor(x,y)
```

```
[1] 0.9937
```

La correlación es bastante cercana a 1, lo cual indica que la variable x está fuertemente correlacionada positivamente con la variable y.

Recta ajustada

A continuación se ajusta la recta de regresión utilizando la función `lm()` ya definida en R.

```
# Ajuste de regresión
> ajuste <- lm(y ~ x)
```

```
Call:
lm(formula = y ~ x)

Coefficients:
```

```
(Intercept)            x
       4.16        15.51
```

En la salida vemos que el modelo de regresión está dado por

$$\widehat{y} = 4.16 + 15.51x$$

y este modelo se puede interpretar de la siguiente manera:

a) El valor para el estimador de la pendiente $\widehat{\beta}_1$ es de 15.51, el cual indica que en promedio los valores de la variable y aumentan aproximadamente en 15.51 cuando la variable x aumenta en una unidad.

b) El valor del estimador $\widehat{\beta}_0$ es de 4.16, el cual indica el valor de la intersección con el eje ordenado. Nótese que este resultado parece no ser congruente con el diagrama de dispersión por la escala que presenta el diagrama.

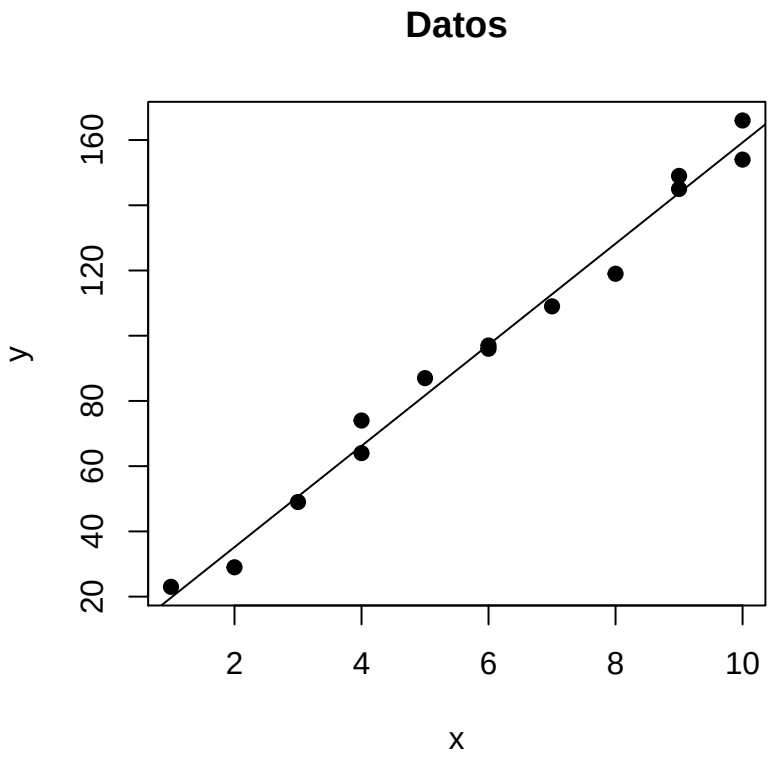
La instrucción `summary()` desgloza un resumen del ajuste mostrando más información sobre éste.

```
# Resumen del ajuste
> summary(ajuste)
```

```
Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
-9.232 -3.341 -0.714  4.777  7.803

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.162      3.355    1.24     0.24
x             15.509      0.505   30.71  8.9e-13 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.39 on 12 degrees of freedom
Multiple R-squared: 0.987,        Adjusted R-squared: 0.986
F-statistic:  943 on 1 and 12 DF,  p-value: 8.92e-13
```

En el resumen se puede observar que el valor del coeficiente de determinación R^2 es de `0.987`. Este resultado indica que el modelo de regresión explica el 98.7 % de la variabilidad de y. Para poder graficar la recta ajustada sobre el conjunto de datos utilizamos la instrucción `abline()` como sigue:

```
# Gráfica de la recta ajustada
> plot(x, y, main="Datos", pch=19)
> abline(ajuste)
```



Intervalos de confianza

Para poder construir intervalos de confianza, en R existe la instrucción `confint()` con la cual podemos determinar el nivel de confianza que deseamos.

```
# Intervalo al 95% de confianza
> confint(ajuste)
```

```
              2.5 % 97.5 %
(Intercept) -3.148  11.47
x           14.409  16.61
```

```
# Intervalo al 90% de confianza
> confint(ajuste, level=0.9)
```

```
                5 %  95 %
(Intercept) -1.818 10.14
x           14.609 16.41
```

Vemos que en ambos intervalos de confianza, para el estimador $\widehat{\beta}_1 = 15.509$ no contienen al cero, además observando en la salida del resumen, tenemos que el valor p para la hipótesis nula $H_0 : \beta_1 = 0$ es `p-value: 8.92e-13` (cero), en consecuencia, para ambos casos, cuando el nivel de significancia α es del 5 % y 10 % rechazamos la hipótesis nula, por lo tanto, decimos que *el valor del parámetro es significativamente distinto a cero.*

Para el estimador $\widehat{\beta}_0 = 4.162$ se tienen intervalos de confianza que sí contienen al cero, además el valor p para la hipótesis nula $H_0 : \beta_0 = 0$ es de 0.239 y para ambos casos, cuando el nivel de significancia α es del 5 % y 10 % no se rechaza ésta hipótesis y se puede decir que el parámetro β_0 es significativamente igual a cero.

Observación 2.3. *Para el parámetro β_0, el diagrama de dispersión, los intervalos de confianza y la prueba de hipótesis indican que un modelo de regresión por el origen deberá ajustar mejor a los datos.*

Para realizar el ajuste de regresión por el origen utilizamos la instrucción `lm(y ~ x -1)` o también se puede usar la instrucción `lm(y ~ x+0)`. Por cuestiones didácticas se terminará el ejemplo con el ajuste inicial y queda para el lector realizar el ajuste del modelo de regresión por el origen.

Análisis de Varianza

En `R` podemos obtener la tabla ANOVA fácilmente con la instrucción `anova.lm()` .

```
# Tabla ANOVA
> anova.lm(ajuste)
```

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
x          1  27420   27420     943 8.9e-13 ***
Residuals 12    349      29
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Los valores de `Sum Sq` representa la suma de los cuadrados y se puede ver que la suma de los cuadrados de la regresión **SCR** es `27420`, miestras que el valor `349` es el valor de la suma de los cuadrados de los residuales **SCE** y la suma de estos dos valores es la suma de los cuadrados totales **SCT**.

Adicionalmente, la tabla indica que los cuadrados medios para la regresión **CMR** es `27420` y para los residuales el cuadrado medio **CME** es `29`.

Los grados de libertad para la prueba F de la suma de los cuadrados de los residuales **SCE** es `12`, además $\widehat{\sigma}^2 = 29$, es decir $\widehat{\sigma} = 5.39$, una desviación estándar pequeña considerando los valores de la variable y. Para la hipótesis nula $H_0 : \beta_1 = 0$ de la prueba F, el valor p es `8.9e-13` (cero) y para un nivel de significancia α =0.05 se rechaza la hipótesis nula, concluyendo que la relación lineal entre las variables x y y es altamente significativa.

Diagnóstico del modelo

En esta parte debemos probar si se cumplen los supuestos establecidos para el

modelo de regresión lineal simple. Hemos dicho que los *residuales* pueden considerarse como la desviación entre los valores observados y el ajuste y son una medida de variabilidad de y no explicada por el modelo de regresión. También pueden considerarse como los valores estimados de los errores ϵ_i del modelo y por esto, el análisis de los residuales es una manera efectiva de describir insuficiencias en el modelo. Para poder calcular los valores ajustados, los residuales, residuales estandarizados y los residuales studentizados utilizando el paquete estadístico `R` se utilizan las siguientes instrucciones:

a) `fitted()` valores ajustados $\widehat{y}_i$.

b) `resid()` residuales e_i.

c) `rstandar()` residuales estandarizados d_i.

d) `rstudent()` residuales studentizados r_i.

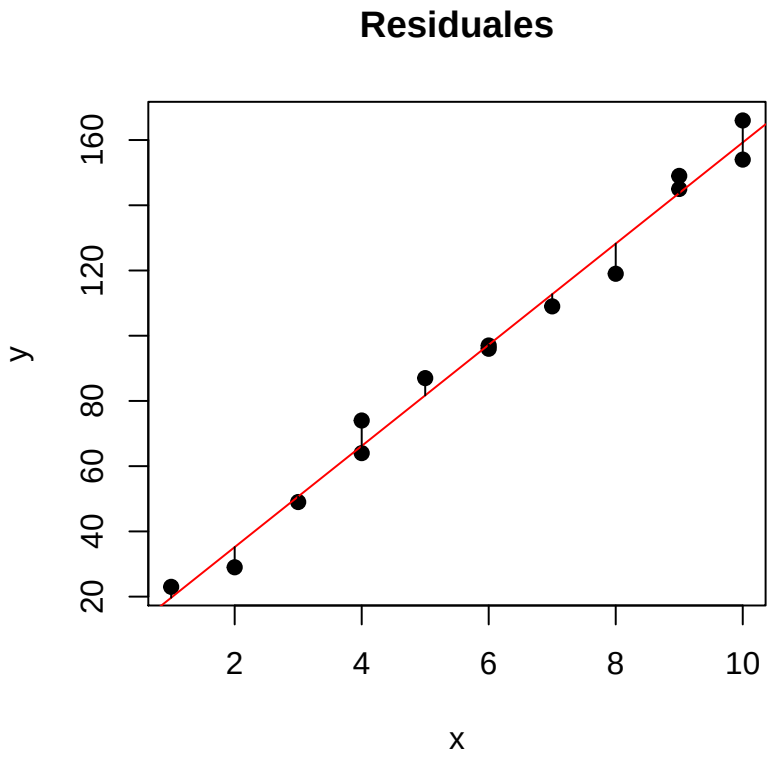
Calculando los valores ajustados $\widehat{y}_i$ obtenemos

```
# Valores ajustados ŷ_i
> yg <- fitted(ajuste)
> yg
```

```
     1      2      3      4      5      6      7      8
 19.67  35.18  50.69  66.20  66.20  81.71  97.21  97.21
     9     10     11     12     13     14
112.72 128.23 143.74 143.74 159.25 159.25
```

Para graficar las distancias entre los valores observados y los valores ajustados se puede realizar con el comando `segments()`.

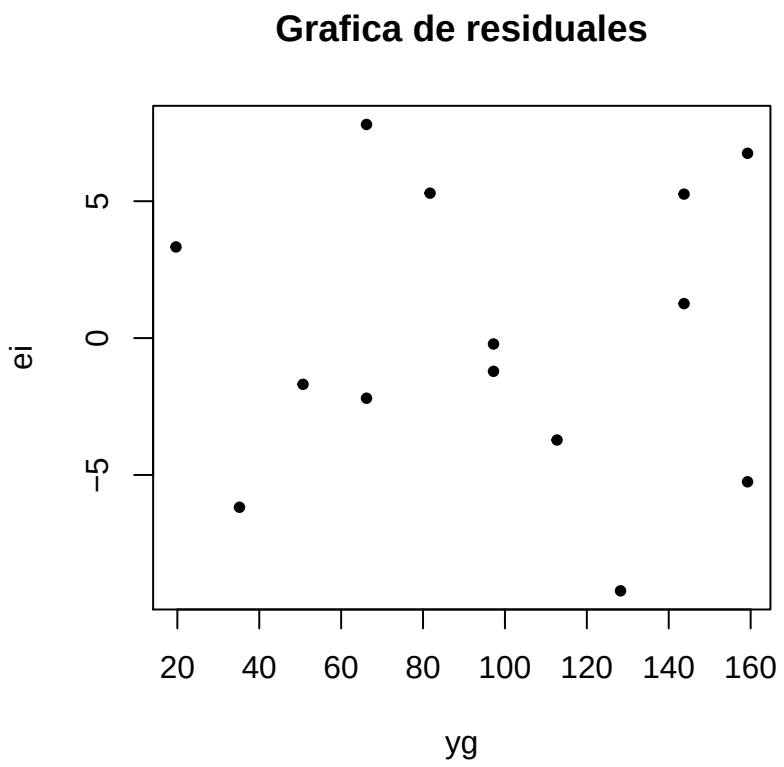
```
> plot(x, y, pch=19, main="Residuales")
> abline(ajuste, col="red")
> segments(x, yg , x , y)
```



Supuesto de varianza constante

Para comprobar el supuesto de varianza constante u homoscedasticidad en el modelo de regresión realizamos la gráfica de los residuales e_i contra los valores ajustados $\widehat{y}_i$.

```
# Residuales e_i
> ei <- resid(ajuste)
# Gáfica de e_i vs ŷ_i
> plot(yg, ei, pch=20, main="Grafica de residuales")
```



En la gráfica vemos que no hay ningún patrón y que el modelo es satisfactorio, por lo que cumple con el supuesto de varianza constante.

Una prueba no paramétrica que permite probar el supuesto de varianza constante es la prueba Breusch-Pagan. En R se puede llevar a cabo la prueba Breusch-Pagan con el comando `bptest()` sobre el objeto `lm()`, la cual requiere de la biblioteca `lmtest`. La hipótesis nula H_0 para esta prueba indica que el ajuste cumple con el supuesto de varianza constante u homoscedasticidad.

```
# Carga la biblioteca lmtest
> library(lmtest)
# Prueba Breusch-Pagan
> bptest(ajuste)

        studentized Breusch-Pagan test

data:  ajuste
BP = 0.5143, df = 1, p-value = 0.4733
```

La salida arroja que la estadística de prueba es `BP=0.5143`, además, como el valor p es mayor que un nivel de significancia del 5 %, podemos decir que nuestro ajuste cumple con el suspuesto de varianza constante u homoscedasticidad.

Supuesto de normalidad

El supuesto de normalidad se puede probar utilizando gráficas cuantitativas como histogramas, QQ-Plots, diagramas de caja y pruebas de bondad de ajuste.

Deseamos comparar ciertas estadísticas descriptivas de los *residuales studentizados* contra las de una distribución normal estándar $N(0,1)$. Para realizar la comparación se pueden utilizar las siguientes medidas:

a) Sesgo igual a cero. El sesgo mide la simetría de la distribución de una variable. Un sesgo mayor a cero indica que hay más residuales positivos que negativos.

b) Kurtosis igual a 3. La kurtosis de una variable indica qué tan picuda es la gráfica de la distribución, reflejando el peso de las colas de la distribución en relación al valor central. Comúnmente, aunque no siempre:

   - Si el valor de la kurtosis es menor a 3 es evidencia de colas más pesadas que la de una normal.
   - En caso contrario, si el valor de la kurtosis es mayor que 3 es evidencia de colas más ligeras que la de una normal.

c) Adicionalmente, se pueden obtener la media, mediana, varianza, desviación estándar, mínimo y máximo.

A continuación se carga la biblioteca `moments` para poder calcular los residuales studentizados del ajuste con la instrucción `rstudent()` y mostrar sus características numéricas de la siguiente manera:

```
# Carga la biblioteca moments
> library(moments)
# Genera los residuales studentiza-
dos
> ri <- rstudent(ajuste)
# Kurtosis
> kurtosis(ri)
```

```
[1] 2.112
```

```
# Asimetría
> skewness(ri)
```

```
[1] -0.1385
```

```
# Media
> mean(ri)
```

```
[1] -0.001444
```

```
# Mediana
> median(ri)
```

```
[1] -0.1319
```

```
# Desviación estándar
> sd(ri)
```

```
[1] 1.095
```

```
# Varianza
> sd(ri)^2
```

```
[1] 1.199
```

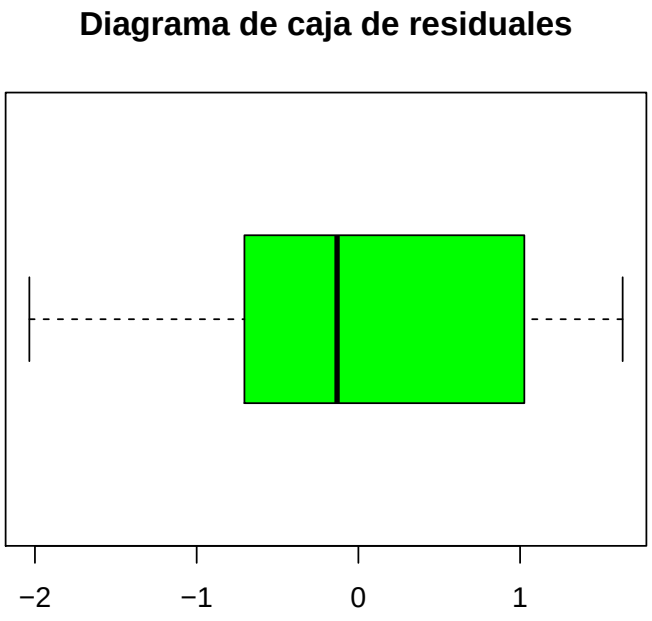
```
# Máximo y mínimo
> max(ri)
> min(ri)
```

```
[1] 1.634

[1] -2.035
```

Diagrama de caja

El diagrama de caja muestra gráficamente la localización de los principales cuantiles de los residuales studentizados. Las siguientes instrucciones muestran cómo realizar dicho gráfico.

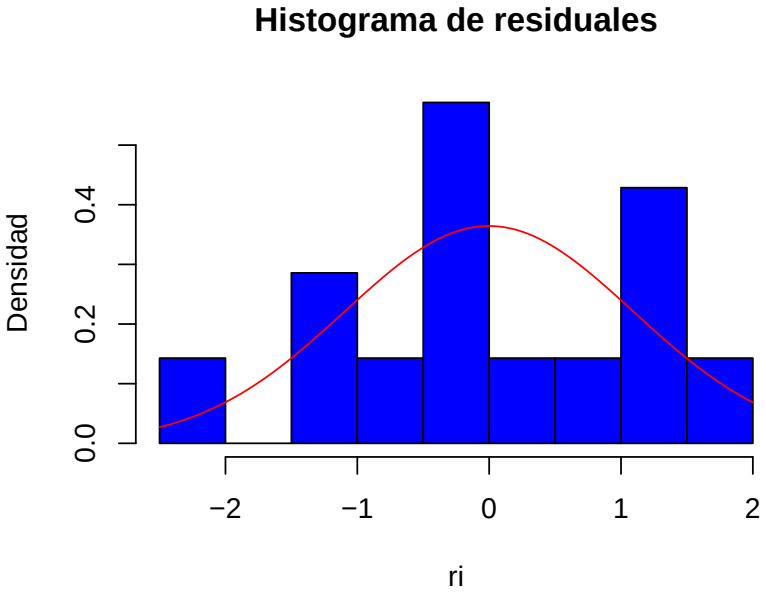
```
# Diagrama de caja de los residuales studentizados
> boxplot(ri,horizontal=TRUE, main="Diagrama de caja de
residuales", col="Green")
```



Histograma

Adicionalmente, se puede graficar un histograma para comprobar el supuesto de normalidad con ayuda del comando `hist()` como se muestra a continuación:

```
# Histogama
> hist(ri, prob=TRUE, breaks=10, col="blue",
main="Histograma de residuales", ylab="Densidad")
# Curva normal
> curve(dnorm(x,mean(ri),sd(ri)), add=TRUE, col="red")
```

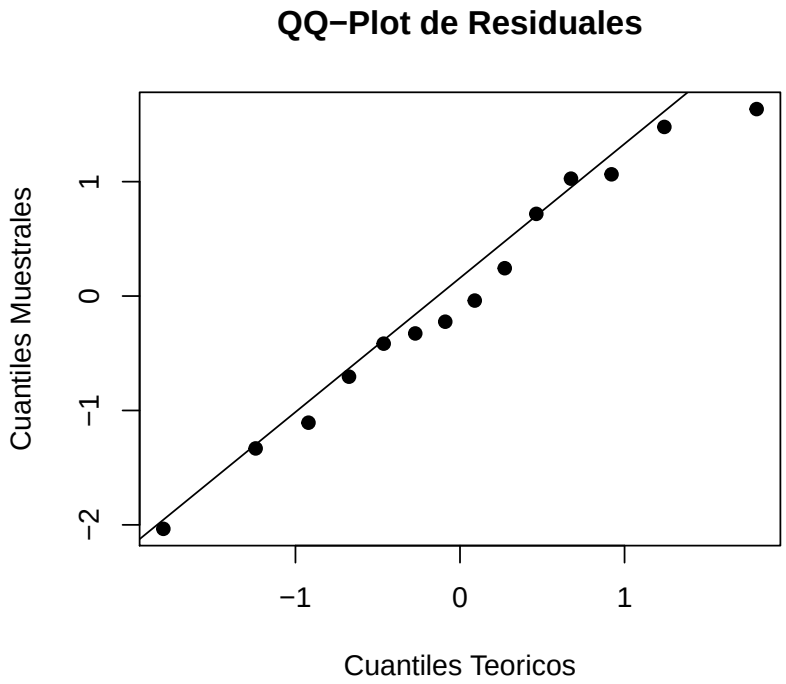


La curva muestra la función de densidad de probabilidad normal con media y desviación estándar de los residuales r_i.

Gráfica de probabilidad normal QQ-Plot

Si las estadísticas anteriores no brindan información suficiente para probar el supuesto de normalidad, podemos apoyarnos en una gráfica de probabilidad normal utilizando el comando `qqnorm()` la cual compara los cuantiles teóricos de una distribución normal contra los cuantiles de la muestra de los residuales.

```
# Gráfica QQ-Plot
> qqnorm(ri,main="QQ-Plot de Residuales", pch=19,
xlab="Cuantiles Teoricos", ylab="Cuantiles Muestrales")
# Recta de cuantiles
> qqline(ri)
```



En la gráfica se aprecia que la mayoría de los puntos forman una línea recta. Esto es un buen indicador de que los residuales sí se distribuyen aproximadamente normal.

Prueba de bondad de ajuste

Una parte importante para probar el supuesto de normalidad es llevar a cabo una prueba de bondad de ajuste. Aquí se utiliza la prueba de Shapiro-Wilk, Lilliefors y la prueba Anderson-Darling para comprobar normalidad sobre los residuales studentizados r_i.

```
# Prueba Shapiro-Wilk
> shapiro.test(ri)
```

```
        Shapiro-Wilk normality test

data:  ri
W = 0.9717, p-value = 0.8982
```

Observando el valor p de la prueba, podemos creer que los datos se distribuyen normal.

```
# Carga la libreria nortest
> library(nortest)
# Prueba Lilliefors
> lillie.test(ri)
```

```
        Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data:  ri
D = 0.1117, p-value = 0.9066
```

```
# Se necesita también la libreria nortest
# Prueba Anderson-Darling
> ad.test(ri)
```

```
        Anderson-Darling normality test

data:  ri
A = 0.1707, p-value = 0.913
```

El resultado que arrojan estas dos últimas pruebas, también permite creer que los residuales r_i siguen una distribución normal, por lo tanto podemos decir que el supuesto de normalidad se cumple.

Supuesto de errores no correlacionados

Para verificar el supuesto de que los residuales o errores ϵ_i no estén correlacionados, es decir, $Cov(\epsilon_i, \epsilon_j) = 0$ para $i \neq j$, es indispensable utilizar la prueba Durbin-Watson. Esta prueba utiliza los residuales de la regresión lineal para minimizar los cuadrados. Las hipótesis a contrastar son:

$$H_0: \text{No existe correlación entre los errores.} \quad vs \quad H_1: \text{Existe correlación entre los errores.}$$

La estadística de prueba, denotada por DW está dada por

$$DW = \frac{\sum_{i>2}(r_i - r_{i-1})^2}{\sum_{i>2} r_i^2}$$

donde r_i es el residuo de la i-ésima observación. La distribución muestral de esta estadística es algo inusual y el rango de la distribución se encuentra entre 0 y 4. Bajo la hipótesis nula, la media de la distribución es cercana a 2. Para valores cercanos a 0 y 4 indica problemas de correlación entre residuales.

En `R`, esta prueba se realiza utilizando el comando `dwtest()` que depende de la biblioteca `lmtest`. La prueba de Durbin-Watson contrasta la hipótesis nula H_0 que especifica que la correlación de los errores es 0 contra la hipótesis alternativa H_1 la cual indica que la correlación es mayor que, menor que, o distinta de 0, y esto puede ser especificado por el argumento `alternative` el cual puede tomar los valores `"greater"`, `"less"` y `"two.sided"` respectivamente. El resultado obtenido es la estadística de prueba DW y el valor p.

```
# Carga la libreria lmtest
> library(lmtest)
# Prueba Durbin-Watson
> dwtest(ajuste, alternative="two.sided")
```

```
        Durbin-Watson test

data:  ajuste
DW = 2.051, p-value = 0.8227
alternative hypothesis: true autocorrelation is not 0
```

El resultado que muestra permite pensar que no hay correlación entre los errores ϵ_i. ▪

2.18. Breve introducción al modelo de regresión lineal múltiple

Un modelo de regresión donde interviene más de una variable regresora se le conoce como *modelo de regresión múltiple*. En esta sección se explicará brevemente cómo se realiza el ajuste y análisis de dicho modelo utilizando el paquete estadístico `R`.
De esta manera, se puede relacionar la respuesta o variable dependiente y con $k-1$ regresores o variables explicativas, de tal manera que el modelo de regresión lineal múltiple está dado por:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \epsilon$$

donde a los parámetros β_j con $j = 0, 1, 2, \ldots, k-1$, se conocen como *coeficientes de regresión*. Nótese que este modelo describe un hiperplano en el espacio de $k-1$ dimensiones de las variables regresoras x_j.

Nuevamente, se supone que el término de error ϵ del modelo tiene los supuestos:

a) $\mathbb{E}(\epsilon) = 0$,

b) $Var(\epsilon) = \sigma^2$,

c) Los errores no están correlacionados.

Cuando se prueban hipótesis o se establecen intervalos de confianza, se debe suponer que la distribución condicional de y dadas las variables $x_1, x_2, \ldots, x_{k-1}$ es normal, con media $\beta_0 + \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1}$ y varianza σ^2.

Si se tienen $n > k-1$ observaciones independientes $y_1, y_2, \ldots, y_n$ de la variable respuesta y. Se puede escribir de la siguiente manera el modelo muestral de regresión para cada y_i como:

$$y_i = \beta_0 + \sum_{j=1}^{k-1} \beta_j x_{ij} + \epsilon_i$$

en donde x_{ij} es el valor de la j-ésima variable independiente, $j = 1, 2, \ldots, k-1$, para la i-ésima observación, $i = 1, 2, \ldots, n$.

Por otro lado, se definen las siguientes matrices como:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk-1} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{k-1} \end{bmatrix} \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

donde

a) $\mathbf{Y}$ es un vector de $n \times 1$ de las observaciones de la variable respuesta,

b) $\mathbf{X}$ es una matriz no singular (tiene inversa) de $n \times k$ de las variables regresoras, y donde la columna de unos corresponde al término constante β_0,

c) $\boldsymbol{\beta}$ es un vector de $k \times 1$ de los coeficientes de regresión,

d) $\boldsymbol{\epsilon}$ es un vector de $n \times 1$ de los errores aleatorios.

El modelo de regresión lineal múltiple en notación matricial se puede expresar como

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Las ecuaciones normales y soluciones por mínimos cuadrados para el modelo de regresión lineal múltiple son:

$$\text{Ecuaciones normales:} \quad (\mathbf{X}'\mathbf{X})\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$$
$$\text{Soluciones:} \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Ejemplo 2.6. *En este ejemplo se presenta como construir el modelo de regresión lineal múltiple utilizando* `R`*. Se desea analizar los gastos (en miles de pesos) de las computadoras personales en un departamento comercial a partir del tiempo de uso (en años) y del número de horas diarias que trabajan (horas/días).*

Se ha tomado una muestra de 5 computadoras personales, de las cuales se han obtenido los siguientes resultados:

Gastos y (miles de pesos)	Antigüedad x_1 (años)	Horas de trabajo x_2 (horas/días)
20.4	1	11
33.0	3	13
36.6	4	15
43.8	6	18
28.6	2	12

Como los datos de este ejemplo son muy pocos, podemos asignar los valores de la variable respuesta y utilizando un vector.

```
# Asigna los valores de la respuesta y
> Y <- c(20.4,33.0,36.6,43.8,28.6)
> Y
```

```
[1] 20.4 33.0 36.6 43.8 28.6
```

Los datos forman la matriz de variables explicativas $\mathbf{X}$ de la siguiente forma:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 11 \\ 1 & 3 & 13 \\ 1 & 4 & 15 \\ 1 & 6 & 18 \\ 1 & 2 & 12 \end{bmatrix}$$

Escribiendo esta matriz en R tenemos el siguiente código:

```
# Matriz X
> X1 <- c(1,1,11)
> X2 <- c(1,3,13)
> X3 <- c(1,4,15)
> X4 <- c(1,6,18)
> X5 <- c(1,2,12)
> X <- rbind(X1,X2,X3,X4,X5)
> X
```

```
   [,1] [,2] [,3]
X1    1    1   11
X2    1    3   13
X3    1    4   15
X4    1    6   18
X5    1    2   12
```

Calculando el producto de matrices para encontrar el vector de los parámetros estimados.

```
# Estimación de β̂
# Multiplicación de X'X
> XtX <- t(X)%*%X
# Inversa (X'X)^-1
> invXtX <- solve(XtX)
# Parámetros estimados β̂
> betag <- invXtX%*%t(X)%*%Y
> betag
```

```
       [,1]
[1,] 57.238
[2,] 10.538
[3,] -4.238
```

Así, podemos escribir el modelo como:

$$y = 57.238 + 10.538x_1 - 4.238x_2.$$

■

Ajuste para un modelo de regresión lineal múltiple en R

Para ajustar un modelo de regresión lineal múltiple utilizando R, se dará un ejemplo con 45 datos simulados para dos variables explicativas x_1, x_2 y una variable respuesta y. Se tiene un archivo con nombre `Data.csv` con el cual se va extraer la información y se mostrará cada una de las instrucciones nencesarias para llevar a cabo el análisis.

Lectura de datos

Se realiza la lectura de los datos de la siguiente manera:

```
> datos <- read.csv('/home/misra/Data.csv',header=T)
```

Se le pide a R que tome en cuenta el encabezado de cada columna ya que en el archivo están distinguidos por `X1`, `X2` y `Y`.

Ajuste de regresión

Usando los comandos directamente de R, se realiza el ajuste utilizando la función `lm()` a través de las siguientes líneas de código:

```
# Ajuste de los datos
> ajuste <- lm(Y ~ X1+X2, data=datos)
> ajuste
```

```
Call:
lm(formula = Y ~ X1 + X2, data = datos)

Coefficients:
(Intercept)           X1           X2
     69.131        0.652       -1.320
```

Pidiendo a R que muestre el resumen de los datos, se obtiene la siguiente salida:

```
# Resumen del Ajuste
> summary(ajuste)
```

```
Call:
lm(formula = Y ~ X1 + X2, data = datos)

Residuals:
   Min     1Q Median     3Q    Max
-21.04  -8.76  -2.15   5.39  40.45
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   69.131     25.055    2.76   0.0085 **
X1             0.652      0.087    7.49  2.5e-09 ***
X2            -1.320      1.235   -1.07   0.2912
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 12.8 on 43 degrees of freedom
Multiple R-squared: 0.655,        Adjusted R-squared: 0.639
F-statistic: 40.8 on 2 and 43 DF,  p-value: 1.14e-10
```

En el resumen que arroja `R` aparece la estadística de la prueba `F-statistic` la cual sirve para ver si las dos variables explicativas son significativas simultáneamente. Como el valor `p-value: 1.14e-10` es menor a 0.05, se rechaza la hipótesis nula de que los dos coeficientes del modelo sean simultáneamente cero.

Por otro lado, de acuerdo a la prueba t, los valores de `Pr(>|t|)` indican qué variables son significativas por separado. En este ejemplo, el término constante `(Intercept)` es significativo a un nivel de significancia de `0.01`, la variable x_1 es significativa a un nivel de `0.001` y la variable x_2 no muestra ser significativa.

El coeficiente de determinación R^2 `Multiple R-square: 0.655` indica que la regresión explica el 65.5 % de la variabilidad de la variable respuesta y a partir de las variables explicativas x_1 y x_2.

Análisis de varianza

Así como en el caso de regresión lineal simple, también se puede calcular la tabla ANOVA para realizar el análisis de varianza para el caso múltiple con la instrucción `anova()`.

```
# Tabla ANOVA
> anova(ajuste)
```

```
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value  Pr(>F)
X1         1  13299   13299   80.56 2.1e-11 ***
X2         1    189     189    1.14    0.29
Residuals 43   7099     165
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

En la salida se muestra que los grados de libertad para la prueba F o el número de variables explicativas sin contar el término constante `Df` es igual `1` para las variables explicativas y para los residuales es `43`.

Los valores de `Sum Sq` representan la suma de los cuadrados y se puede ver que

la suma explicada por el modelo de regresión es `13299` para x_1, para x_2 es `189`, mientras que el valor `7099` es el valor que toma la suma de los cuadrados de los residuales.

Por otro lado, la tabla indica que el cuadrado medio para las variables explicativas x_1 y x_2 es `13299` y `189` respectivamente. El cuadrado medio de los residuales es `165`.

Por último, el valor de la estadística para la prueba F y el valor p para x_1 es `80.56` y `2.1e-11` respectivamente, indicando que es significativa a un nivel de significancia de `0.001`. Para la variable x_2 La prueba F y el valor p son `1.14` y `0.29` respectivamente.

Multicolinealidad

El análisis de multicolinealidad a través de la matriz de correlaciones de las variables explicativas se puede obtener con la instrucción `cor()` y se emplea como se muestra abajo.

```
# Matriz de correlaciones
> cor(datos)
```

```
          Y      X1      X2
Y    1.0000  0.8037 -0.4526
X1   0.8037  1.0000 -0.4572
X2 -0.4526 -0.4572  1.0000
```

La matriz indica que la variable explicativa x_1 está correlacionada con la respuesta y con una correlación positiva de `0.8037`. Para el caso de las variables x_2 y y se tiene una correlación negativa de `-0.4526`.

Diagnóstico del modelo

A continuación se calculan los residuales estandarizados y los valores ajustados para verificar el supuesto de varianza constante.

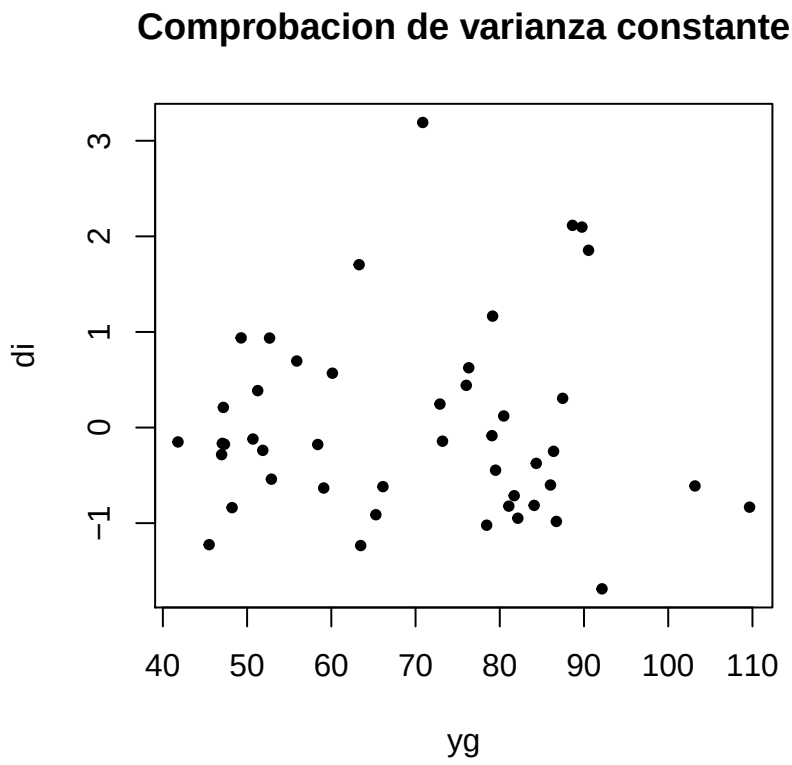
```
# Residuales estandarizados d_i
> di <- rstandard(ajuste)
```

```
# Valores ajustados ŷ_i
> yg <- fitted(ajuste)
```

Supuesto de varianza constante

Para comprobar el supuesto de varianza constante u homoscedasticidad en el modelo de regresión realizamos la gráfica de los residuales d_i contra los valores ajustados $\widehat{y}_i$.

```
# Gráfica de d_i vs ŷ_i
> plot(yg, di, pch=20, main="Comprobacion de varianza
constante")
```



El diagrama de dispersión muestra que probalemente si se cumple el supuesto de varianza constante. Para comprobarlo, se utilizando la prueba Breusch-Pagan.

```
# Carga la biblioteca lmtest
> library(lmtest)
# Prueba Breusch-Pagan
> bptest(ajuste)
```

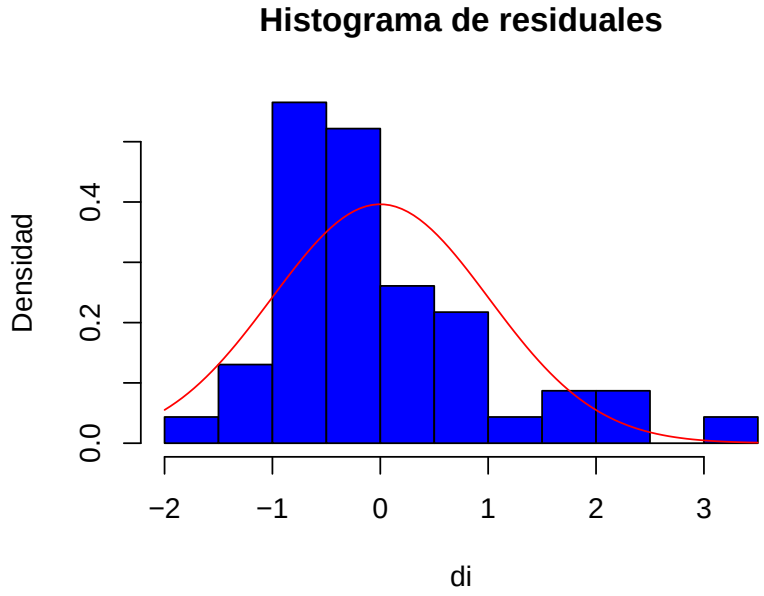
```
        studentized Breusch-Pagan test

data:  ajuste
BP = 2.117, df = 2, p-value = 0.347
```

La salida arroja que la estadística de prueba es `BP=2.117`, además como el valor p `0.347` es mayor que el nivel de significancia del 5 %, no se rechaza la hipótesis H_0 la cual establece homoscedasticidad. Así, decimos que el ajuste cumple con el supuesto de homoscedasticidad o varianza constante.

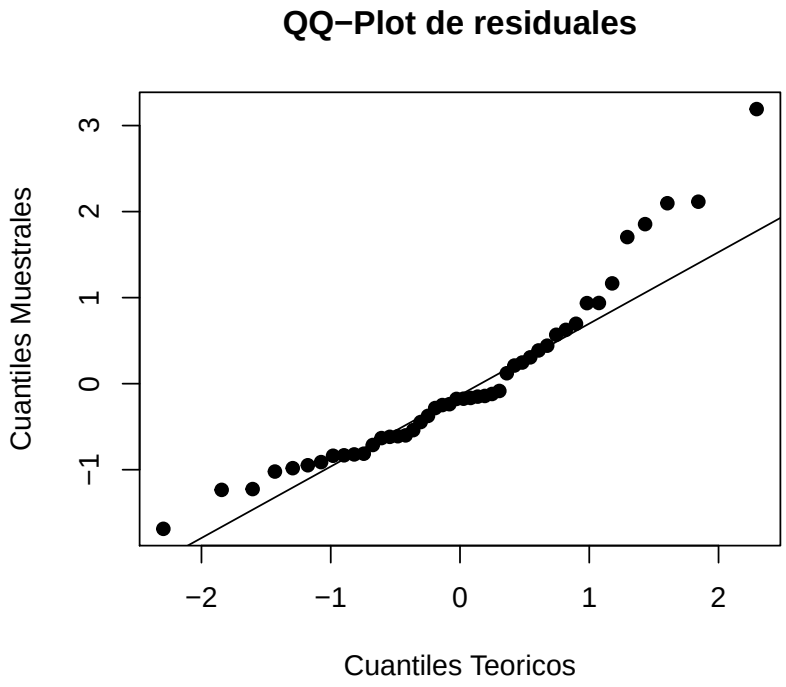
Supuesto de normalidad

Para probar el supuesto de normalidad se utilizarán las instrucciones para mostrar un histograma con su respectiva curva normal y la gráfica QQ-plot.

```
# Histogama
> hist(di, prob=TRUE, breaks=10, col="blue",
main="Histograma de residuales", ylab="Densidad")
# Curva normal
> curve(dnorm(x,mean(di),sd(di)), add=TRUE, col="red")
```




```
# Gráfica QQ-Plot
> qqnorm(di,main="QQ-Plot de residuales", pch=19,
xlab="Cuantiles Teoricos", ylab="Cuantiles Muestrales")
# Recta de cuantiles
> qqline(di)
```



Pruebas de bondad de ajuste

Ahora, tenemos las pruebas para normalidad Shapiro-Wilk, Lilliefors y Anderson-Darling.

```
# Prueba Shapiro-Wilk
> shapiro.test(di)
```

```
        Shapiro-Wilk normality test

data:  di
W = 0.9166, p-value = 0.002867
```

Observando el valor p de la prueba, podemos creer que los datos no se distribuyen normalmente.

```
# Carga la libreria nortest
> library(nortest)
# Prueba Lilliefors
> lillie.test(di)
```

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  di
D = 0.1637, p-value = 0.003402
```

El resultado que arroja esta prueba, también indica que los residuales d_i no siguen una distribución normal. Por último llevamos a cabo la prueba Anderson-Darling para normalidad.

```
# Se necesita también la libreria nortest
# Prueba Anderson-Darling
> ad.test(di)
```

```
        Anderson-Darling normality test

data:  di
A = 1.247, p-value = 0.002681
```

El valor p de esta prueba también indica que los residuales no siguen una distribución normal, por lo tanto podemos decir que el supuesto de normalidad no se cumple.

Supuesto de residuales no correlacionados
Para verificar el supuesto de que los residuales o errores no esten correlacionados se realiza la prueba Durbin-Watson.

```
# Prueba Durbin-Watson
> dwtest(ajuste, alternative="two.sided")
```

```
        Durbin-Watson test

data:  ajuste
DW = 2.041, p-value = 0.9271
alternative hypothesis: true autocorrelation is not 0
```

Por el resultado que muestra el valor p podemos decir que no hay correlación entre los residuales.

▪

Apéndice A

Introducción a R

La estadística es una de las áreas de las matemáticas que se encarga de la organización, estudio y análisis de datos para después tomar decisiones respecto a los resultados obtenidos. En la actualidad para llevar a cabo este procedimiento es indispensable utilizar programas o paquetes estadísticos que faciliten el trabajo del investigador. En este capítulo se presenta una introducción a la estructura y uso del paquete estadístico R ya que es un software estadístico muy poderoso, muy amigable y de distribución libre. Sin embargo, cabe mencionar que R está basado en el lenguaje S y tiene una interfase no muy amigable, por lo que se propone el uso de un editor como R-Studio el cual está disponible para sistemas operativos como Windows y Linux. Aunque existen otros editores como Tinn-R, en el caso de Windows y RKWard para Linux, el uso del editor es totalmente opcional, ya que cada programa está hecho para trabajar por separado. El paquete estadístico R puede obtenerse libremente en el sitio `http://cran.r-project.org`, donde además podrán encontrar manuales e instructivos si se quiere profundizar en el tema. En particular, se recomienda la lectura del texto *An Introduction to R*, o bien, en internet se pueden encontrar gran variedad de notas, tutoriales, etc. bastante útiles para explorar.

A.1. Iniciando R

Existen distintas maneras para iniciar R dependiendo del sistema operativo que se esté utilizando. En Windows basta dar click en el ícono de R que se genera en el escritorio como acceso directo o buscando el programa en el menú de incio. El programa tiene la apariencia como se muestra en la Figura A.1. Para iniciar R en Linux se necesita abrir una terminal o consola y escribir la letra R seguida de teclear ENTER. El programa se presenta en línea de comando listo para utilizarse, como se muestra en la Figura A.2.
El símbolo "`>`" indica que R está listo para recibir un comando para su ejecución. No obstante, se recomienda al usuario descargar de forma gratuita el editor de texto Rstudio desde el sitio `http://www.rstudio.com`. Esta interface o editor de texto permite al usuario un manejo fácil, sencillo y de gran flexibilidad tanto en la instalación de paquetería, presentación de gráficos, entre otras cosas.

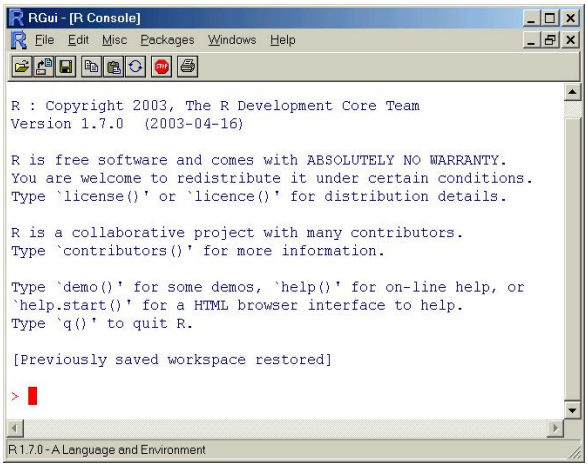


Figura A.1: Presentación de R en Windows.

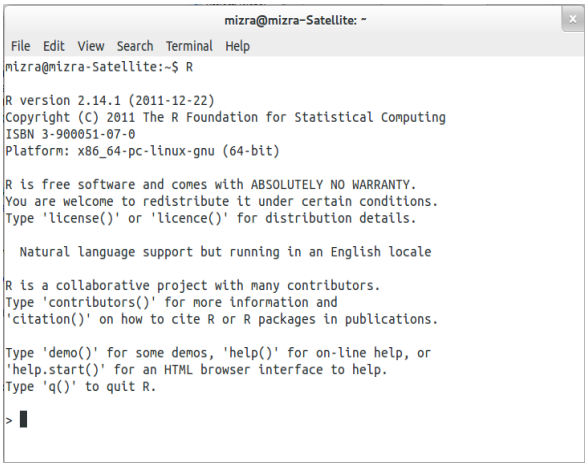


Figura A.2: Presentación de R en Linux.

A.2. Aritmética en R

En R se pueden realizar operaciones aritméticas simples y convencionalmente usuales como en otros programas de cómputo, es decir, se puede sumar, restar, multiplicar, dividir y elevar un número a alguna potencia. Estas operaciones se introducen con los símbolos `+,-,*,/,^` respectivamente. En el siguiente ejemplo se muestran algunos comandos presentando la utilidad de estas operaciones.

Ejemplo A.1. *A continuación se muestra cómo realizar operaciones aritméticas simples en R y la forma en que R responde.*

```
> 5+3*7
```

```
[1] 56
```

```
> (4*4)/16
```

```
[1] 1
```

```
> (2^2)^2
```

```
[1] 16
```

```
> 10-20
```

```
[1] -10
```

Para calcular $(8 \times 3) + 5/3 - 6^2 + \sqrt{4}$ escribimos

```
> (8*3)+(5/3)-(6^2)-sqrt(4)
```

```
[1] -12.33
```

Nótese que cada resultado esta precedido con `[1]`. Por otro lado, en R también podemos encontrar una gama de funciones matemáticas que nos permitirán obtener cálculos de forma sencilla. En el cuadro A.1 se muestran algunas de las operaciones que pueden calcularse. Cada función es escrita con su respectiva descripción seguida por paréntesis.

Descripción	Símbolo
Raíz cuadrada	`sqrt()`
Valor absoluto	`abs()`
Suma	`sum()`
Seno	`sin(radianes)`
Coseno	`cos(radianes)`
Tangente	`tan(radianes)`
Logaritmo	`log()`
Exponencial	`exp()`
Factorial	`factorial()`
Máximo	`max()`
Mínimo	`min()`

Cuadro A.1: Operaciones numéricas básicas en R.

Ejemplo A.2. *En este ejemplo se muestra cómo hacer cálculos con algunas funciones numéricas predeterminadas en R.*

```
> sqrt(36)
```

```
[1] 6
```

```
> abs(-2)
```

```
[1] 2
```

```
# Suma de sqrt(36) y abs(-2)
> sum(sqrt(36), abs(-2))
```

```
[1] 8
```

```
> sin(25)
```

```
[1] -0.1324
```

```
> cos(1)
```

```
[1] 0.5403
```

```
> tan(45)
```

```
[1] 1.62
```

```
> log(2)
```

```
[1] 0.6931
```

```
> exp(4)
```

```
[1] 54.6
```

```
# Factorial de 10
factorial(10)
```

```
[1] 3628800
```

```
# Máximo de sqrt(36) y abs(-3)
max(sqrt(36),abs(-3))
```

```
[1] 6
```

```
# Mínimo de sqrt(36) y abs(-3)
min(sqrt(36),abs(-3))
```

```
[1] 3
```

La función logaritmo `log()` es determinada con base e. Muchas funciones en `R` cuentan con argumentos extras que cambian su comportamiento. Por ejemplo si usted quiere el logaritmo en base 10 tiene que especificar como se muestra a continuación:

```
> log(10, base=10)
```

```
[1] 1
```

```
> log(10,10)
```

```
[1] 1
```

Hay que observar que no es necesario especificar el argumento "`base=`", ya que implícitamente se entiende que el primer argumento es el valor al cual le aplicaremos la función logaritmo y el segundo argumento es la base que se desea.

A.3. Sintaxis

Como en todo lenguaje de programación o en su caso computacional, R maneja una sintaxis y operadores comunes para su uso. Para poder hacer comentarios y especificar cada una de las operaciones que se esta llevando a cabo se utiliza el caracter #.

Ejemplo A.3. *Como en ejemplos anteriores, se muestra el uso del caracter* # *para hacer comentarios en* R.

```
# Este es el valor de Pi
> pi
```

```
[1] 3.142
```

```
# Este es el valor del número e
> exp(1)
```

```
[1] 2.718
```

El valor del número e no está determinado por R ya que depende de la función exp(), en cambio el valor de π =3.141593 sí está determinada como una constante.

Asignación

En R es bastante común trabajar con objetos a los cuales se les asignarán valores, operaciones o funciones. En R existen dos formas de hacer estas asignaciones, la primera es utilizando el símbolo "=" y la segunda es utilizando "<-" ó "->" los cuales se pueden pensar como una flecha en dirección a la derecha y una flecha en dirección a la izquierda respectivamente.

Ejemplo A.4. *En este ejemplo se muestra cómo asignar la operación* $3x + 1$ *a la variable* **y** *y* **z** *cuando* **x** *es igual a 5.*

```
# Se le asigna a 'x' el valor 5 usando =
> x = 5
# Se asigna a 'y' la operación 3x+1 usando <-
> y <- (3*x)+1
# Se asigna a 'z' la operación 3x+1 usando ->
> (3*x)+1 -> z
# Se muestran los valores de 'y'y 'z'
> y
> z
```

```
[1] 16
```

```
[1] 16
```

Cuando se trabaja con decimales, es frecuente reducir el número de los decimales a cantidades pequeñas. Así la función `round(x,2)` redondea a dos decimales el número `x`. Por ejemplo,

```
x <- (8*3)+(5/3)-(7^2)
> round(x)
```

```
[1] -23.33
```

Podemos ver que se está declarando a `x` el valor 5, el cual se utiliza para calcular el valor de `y` y de `z`. A lo largo de este texto las asignaciones que se llevarán a cabo serán utilizando "`<-`".

A.4. Vectores

El paquete estadístico `R` utiliza diferentes *estructuras de datos*. La estructura más simple es el vector, que es una colección ordenada de números. Para crear un vector, por ejemplo `x`, consistente en usar la instrucción `c()` como se muestra en el siguiente ejemplo.

Ejemplo A.5. *Creación del vector* `x` *con entradas* $x_1 = 1, x_2 = 1.5, x_3 = 2, x_4 = 2.5, x_4 = 3$.

```
> x <- c(1, 1.5, 2, 2.5, 3)
> x
```

```
[1] 1.0 1.5 2.0 2.5 3.0
```

Al generar otro vector, por ejemplo,

```
> z <- c(10.4, 5.6, 3.1, 6.4, 21.7)
> z
```

```
[1] 10.4  5.6  3.1  6.4 21.7
```

y si escribimos

```
> 1/z
```

```
[1] 0.09615 0.17857 0.32258 0.15625 0.04608
```

se obtienen los valores inversos de cada entrada del vector `z`. Cabe que mencionar que se puede realizar operaciones utilizando los vectores `x` y `z`, o bien

```
> f <- (2*x)+z
> f
```

```
[1] 12.4  8.6  7.1 11.4 27.7
```

y si se quiere elevar al cuadrado cada entrada del vector `x` simplemente escribimos

```
> x^2
```

```
[1] 1.00 2.25 4.00 6.25 9.00
```

Algunos comandos básicos relacionados con el uso de vectores son los siguientes:

a) `max()` y `min()` arrojan el valor máximo y mínimo de los componentes de un vector.

b) `sort()` ordena de forma ascendente las entradas del vector.

c) `length()` muestra el tamaño o número de elementos del vector.

d) `sum()` muestra la suma de todas las estradas del vector.

e) `prod()` calcula el producto de las entradas del vector.

Ejemplo A.6. *En este ejemplo se calculan todas las características numéricas anteriores para el vector* $y = (2{,}5, 3{,}1, 7{,}8, 3{,}4, 5{,}5, 7{,}9, 6, 3)$.

```
# Crea el vector 'y' usando c()
> y <- c(2.5, 3.1, 7.8, 3.4, 5.5, 7.9, 6,3)
# Calcula el máximo de los componentes de 'y'
> max(y)
```

```
[1] 7.9
```

```
# Calcula el mínimo de los componentes de 'y'
> min(y)
```

```
[1] 2.5
```

```
# Ordena en orden ascendente las componentes de 'y'
> sort(y)
```

```
[1] 2.5 3.0 3.1 3.4 5.5 6.0 7.8 7.9
```

```
# Calcula la longitud de 'y'
> length(y)
```

```
[1] 8
```

```
# Calcula la suma de los componentes de 'y'
> sum(y)
```

```
[1] 39.2
```

```
# Calcula el producto de las componentes de 'y'
> prod(y)
```

```
[1] 160745
```

Del ejemplo anterior nótese que el producto de las entradas del vector `y` es en realidad 160745.013, sin embargo automáticamente `R` está redondeando dicho valor.

Por otro lado, si tenemos dos o más vectores y nuestro interés es organizarlos por columnas podemos utilizar el comando `cbind()`, análogamente si queremos organizarlos por filas utilizamos el comando o instrucción `rbind()`. En el siguiente ejemplo se muestra la utilidad de estas instrucciones.

Ejemplo A.7. *Ordenación de los vectores* $i = (1, 0, 0)$, $j = (0, 1, 0)$ *y* $k = (0, 0, 1)$ *por columnas y por renglones.*

```
> i <- c(1,0,0)
> j <- c(0,1,0)
> k <- c(0,0,1)
# Ordena los vectores i,j y k por columnas
> cbind(i,j,k)
```

```
     i j k
[1,] 1 0 0
[2,] 0 1 0
[3,] 0 0 1
```

```
> rbind(i,j,k)
```

```
  [,1] [,2] [,3]
i    1    0    0
j    0    1    0
k    0    0    1
```

A.5. Sucesiones

Una sucesión numérica en `R` es un vector generado por alguna instrucción ya definida en `R`. Existen varias formas para generar sucesiones numéricas muy útiles como son:

a) El operador `:`

b) El comando `seq()`

c) El comando `rep()`

Si queremos generar una sucesión de números enteros positivos basta utilizar el operador `:` como se muestra en el siguiente ejemplo.

Ejemplo A.8. *Aquí se muestra como utilizar el operador* `:` *para generar sucesiones numéricas de la forma* $i, i + 1, \ldots, i + n$.

```
> s <- 0:10
> s
```

```
 [1]  0  1  2  3  4  5  6  7  8  9 10
```

```
> r <- 15:20
> r
```

```
[1] 15 16 17 18 19 20
```

Si nuestro interés es generar una sucesión numérica con una amplitud o salto distinto a 1 se utiliza el comando `seq()` el cual genera sucesiones más complejas y cuenta con los siguientes argumentos:

```
seq(from = valor inicial, to = valor final, by = amplitud)
```

Ejemplo A.9. *Generar la sucesión de números con valor inicial* 1 *y valor final* 3 *con una amplitud de 0.2.*

```
> a <- seq(1, 3, 0.2)
> a
```

```
 [1] 1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8 3.0
```

Para seleccionar un componente o en su caso un subvector de alguna sucesión numérica se le añade al nombre del vector un par de corchetes `[ ]` indicando entre ellos el lugar o el subvector que se desea extraer. Por ejemplo, el quinto elemento de la sucesión `a` se obtiene de la siguiente manera:

```
> a[5]
```

```
[1] 1.8
```

y para seleccionar los valores que ocupan de la primera hasta la sexta entrada de la sucesión `a` escribimos

```
> a[1:6]
```

```
[1] 1.0 1.2 1.4 1.6 1.8 2.0
```

El último comando que mostraremos para generar una sucesión es utilizando el comando `rep(a,n)` el cual sirve para repetir un número o carácter `a`, `n` veces.

Ejemplo A.10. *Repetir el vector* $i = (1, 0, 0)$ *cuatro veces.*

```
> i <- c(1,0,0)
# Repite el vector i cuatro veces
> rep(i,4)
```

```
 [1] 1 0 0 1 0 0 1 0 0 1 0 0
```

Nótese que en el ejemplo anterior estamos combinando la instrucción `c()` y `rep()`.

Símbolo	Función
<	menor que
>	mayor que
<=	menor o igual
>=	mayor o igual
==	igual que
!=	distinto de
|	uno u otro
&	ambos

Cuadro A.2: Operadores lógicos.

A.6. Operadores lógicos

En R podemos utilizar diversos operadores lógicos para los cuales se usan los símbolos del cuadro A.2.

Ejemplo A.11. *Si se le asigna al vector* `c(1.63, 1.68, 1.75, 1.59, 1.80)` *las alturas de 5 estudiantes y se desea saber cuales de esos 5 estudiantes son menores a 1.70 metros, escribimos*

```
> alturas <- c(1.63, 1.68, 1.75, 1.59, 1.80)
> alturas
```

```
[1] 1.63 1.68 1.75 1.59 1.80
```

```
# Muestra las alturas menores a 1.7
> alturas <1.70
```

```
[1]  TRUE  TRUE FALSE  TRUE FALSE
```

en este caso, podemos ver que las alturas para el primero, segundo y cuarto estudiantes son alumnos que tiene una estatura menor a 1.70 cm. Si nuestro problema es saber cuales de los estudiantes son menores a 1.70 metros o mayores o iguales a 1.80 metros, podemos escribir

```
> (alturas <1.70) | (alturas >= 1.80)
```

```
[1]  TRUE  TRUE FALSE  TRUE  TRUE
```

A.7. Matrices

Además de vectores, un arreglo importante que existe en las matemáticas son las *matrices*. R permite crear y manipular este tipo de arreglos con la instrucción `matrix()` la cual cuenta con los siguientes argumentos:

```
matrix(data, nrow, ncol, byrow)
```

donde `data` son los valores u observaciones que pertenecerán a cada entrada de la matriz, `nrow` es el número de renglones de la matriz, `ncol` el número de columnas de la matriz. El argumento `byrow` es una orden lógica que puede tomar el valor TRUE o FALSE el cual sirve para indicarle a `R` que cada entrada de la matriz será ordenada por renglones o por columnas. En el siguiente ejemplo se muestra cómo se introduce una matriz.

Ejemplo A.12. *Sea la matriz*

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 6 & 7 & 8 & 9 & 10 \end{pmatrix}_{2\times 5}$$

Dicha matriz se introduce en `R` *de la siguiente manera*

```
> A <- matrix(1:10, nrow=2, byrow=T)
```

```
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    2    3    4    5
[2,]    6    7    8    9   10
```

donde "`nrow=2`" especifica sólo dos renglones que se requieren y "`byrow=T`" está pidiendo que cada entrada de la matriz se ordene por renglones. Si escribimos la instrucción con "`byrow=F`" tenemos que cada entrada de la matriz será ordenada por columnas, como se muestra a continuación.

```
> A <- matrix(1:10, nrow=2, byrow=F)
> A
```

```
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    3    5    7    9
[2,]    2    4    6    8   10
```

Nótese que para `R` es igual escribir el argumento "`byrow=F`" o no. Para poder obtener la dimensión de la matriz A utilizamos el comando `dim()`, es decir

```
> dim(A)
```

```
[1] 2 5
```

que en efecto, es una matriz que consta de 2 renglones por 5 columnas.

Ejemplo A.13. *Suponga que tenemos los alumnos A, B y C y sus respectivas calificaciones en cada examen parcial que presentaron.*

Alumno	*Parcial 1*	*Parcial 2*	*parcial 3*
A	*9.0*	*8.6*	*10*
B	*6.7*	*8.2*	*9.1*
C	*7.1*	*9.0*	*9.5*

Generar una matriz con la calificaciones anteriores.

Utilizaremos un vector con todas las calificaciones, y escribimos la matriz adecuadamente de la siguiente manera:

```
> datos <- c(9.0, 8.6, 10, 6.7, 8.2, 9.1, 7.1, 9.0, 9.5)
# Genera la matriz con los datos
> calificaciones <- matrix(datos, nrow=3, byrow=T)
# Muestra la matriz calificaciones
> calificaciones
```

```
     [,1] [,2] [,3]
[1,]  9.0  8.6 10.0
[2,]  6.7  8.2  9.1
[3,]  7.1  9.0  9.5
```

```
> dim(calificaciones)
```

```
[1] 3 3
```

Ahora, si queremos indicar que cada fila se refiere a las calificaciones de los alumnos A, B y C y que cada columna hace referencia a cada parcial creamos los vectores

```
> alumnos <- c(" A ", " B ", " C ")
> parciales <- c("Parcial 1", "Parcial 2", "Parcial 3")
```

Asignamos los nombres de cada alumno a las filas de la matriz

```
> dimnames(calificaciones) <- list(alumnos, NULL)
> calificaciones
```

```
  [,1] [,2] [,3]
A  9.0  8.6 10.0
B  6.7  8.2  9.1
C  7.1  9.0  9.5
```

Asignamos los nombres de cada parcial a cada columna

```
> dimnames(calificaciones) <- list(NULL, parciales)
> calificaciones
```

```
     Parcial 1 Parcial 2 Parcial 3
[1,]       9.0       8.6      10.0
[2,]       6.7       8.2       9.1
[3,]       7.1       9.0       9.5
```

Para asignar simultaneamente los nombres de las filas y de las columnas escribimos

```
> dimnames(calificaciones) <- list(alumnos, parciales)
> calificaciones
```

```
  Parcial 1 Parcial 2 Parcial 3
A       9.0       8.6      10.0
B       6.7       8.2       9.1
C       7.1       9.0       9.5
```

Un comando que es común utilizarlo junto con el comando `matrix()` es `apply()` y tiene los siguientes argumentos:

```
apply(X, MARGIN, FUN)
```

donde el argumento `X` es una matriz, `MARGIN` indica los renglones (`MARGIN=1`) o columnas (`MARGIN=2`) de `X`. Mientras que el argumento `FUN` es una función ya determinada en `R` que se aplicará a los renglones o columnas de la matriz `X`.

En el ejemplo anterior, como se está hablando de alumnos y sus calificaciones, es importante obtener los promedios de cada uno de ellos. Entonces si se quiere calcular las medias de cada renglón utilizamos el comando `apply()` de la siguiente forma

```
> apply(calificaciones, 1, mean)
```

```
    A     B     C
9.200 8.000 8.533
```

donde la función "`mean()`" calcula la media o el promedio de las observaciones. Observe que el segundo argumento, cuando toma el valor 2 obtendrá los promedios de las entradas de la matriz por columna, es decir, los promedios de los parciales.

```
> apply(calificaciones, 2, mean)
```

```
Parcial 1 Parcial 2 Parcial 3
    7.600     8.600     9.533
```

A.8. Operaciones con matrices y vectores

`R` tiene la cualidad de realizar operaciones con matrices y vectores. En el Cuadro A.3 se muestran algunas de las operaciones que se pueden llevar a cabo.

Descripción	Símbolo
Vector de ceros de tamaño `x`	`numeric(x)`
Matriz de ceros de tamanño $n \times m$	`matrix(0,n,m)`
i-ésimo elemento del vector `a`	`a[i]`
j-ésima columna del la matriz A	`A[,j]`
Entrada a_{ij} de la matriz A	`A[i,j]`
Mutiplicación de las matrices A y B	`A%*%B`
Transpuesta de la matriz A	`t(A)`
Matriz inversa de A	`solve(A)`
Solución del sistema de ecuaciones $Ax = b$	`solve(A,b)`
Cálculo de valores y vectores propios de A	`eigen(A)`
Matriz diagonal con elementos del vector `x`	`diag(x)`

Cuadro A.3: Operaciones con matrices y vectores.

Ejemplo A.14. *En este ejemplo se presenta la forma de utilizar algunas operaciones que se muestran en el Cuadro A.3. Dada la matriz*

$$A = \begin{pmatrix} 12 & 23 & 15 \\ 15 & 18 & 31 \\ 17 & 47 & 5 \end{pmatrix}_{3\times 3}$$

En R generamos la matriz A como

```
> d <- c(12,23,15,15,18,31,17,47,5)
> A <- matrix(d, nrow=3, byrow=3)
> A
```

```
     [,1] [,2] [,3]
[1,]   12   23   15
[2,]   15   18   31
[3,]   17   47    5
```

El tercer renglón de la matriz A es

```
> A[3,]
```

```
[1] 17 47  5
```

La entrada $a_{2,2}$ de la matriz A es

```
> A[2,2]
```

```
[1] 18
```

La suma y el producto de los elementos de la segunda columna son

```
> z<-c(sum(A[,2]), prod(A[,2]))
> z
```

```
[1]    88 19458
```

La matriz transpuesta de A, denotada como A', está dada por

```
> t(A)
```

```
     [,1] [,2] [,3]
[1,]   12   15   17
[2,]   23   18   47
[3,]   15   31    5
```

La matriz inversa de A, denotada como A^{-1}, está dada por

```
> solve(A)
```

```
       [,1]    [,2]    [,3]
[1,]  59.43 -25.652 -19.261
[2,] -19.65   8.478   6.391
[3,] -17.35   7.522   5.609
```

Sea el vector $b = (2, 5, 8)$, la solución del sistema de ecuaciones $Ax = b$ viene dada por

```
> b <- c(2,5,8)
> solve(A,b)
```

```
[1] -163.48   54.22   47.78
```

La multiplicación del vector b y la matriz A es

```
> b%*%A
```

```
     [,1] [,2] [,3]
[1,]  235  512  225
```

A.9. Funciones

Como hemos visto, en R la mayor parte de los comandos o instrucciones se realiza a través de funciones con sus propios argumentos entre paréntesis. De esta forma, en R es posible crear distintos tipos de funciones que en algún momento serán útiles para el usuario, ya que esto permite realizar mecanismos para encontrar la solución a distintos problemas que surjan más adelante. La estructura general de una función en R es la siguiente:

```
nombre = function(argumento 1, argumento 2,...)
{cuerpo de la función }
```

Ejemplo A.15. *En este ejemplo se construye una función que calcula la desviación estándar de la sucesión numérica* `seq(0, 5, 0.5)` *con ayuda de las funciones* `sqrt()` *y* `var()` *de la siguiente manera:*

```
> desv = function(x){ sqrt(var(x)) }
```

Ahora generamos la sucesión numérica al cual le aplicamos la función `desv()`

```
> r<-c(seq(0,5, 0.5))
> r
```

```
 [1] 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
```

```
> desv(r)
```

```
[1] 1.658
```

el cual coincide con el mismo valor al utilizar la función `sd()` predeterminada en R

```
> sd(r)
```

```
[1] 1.658
```

Ejemplo A.16. *Se genera la función de densidad de una variable aleatoria X que sigue una distribución binomial con parámetros n y p. La función de densidad de una variable aleatoria que se distribuye $Binom(n, p)$ es la siguiente:*

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \textit{para } x = 0, 1, \ldots, n.$$

En R una manera de calcular $f(x)$ es la siguiente:

```
> Binomial <- function(x,n,p){choose(n,x)*(p^x)*(1-p)^(n-x)}
```

donde la instrucción `choose(n,x)` calcula las combinaciones de n en x. Ahora utilizamos esta función.

```
# ℙ(X = 2) cuando n=3 y p=1/2
> Binomial(2,3,1/2)
```

```
[1] 0.375
```

A.10. Gráficas

R es un paquete muy importante y versátil en cuestiones gráficas. Es posible utilizarlas para mostrar una amplia variedad de gráficos estadísticos o para graficar cualquier función de interés. En esta sección se explica cómo crear un gráfico y cuales son los parámetros más utilizados para su creación. Sin embargo, en R existen dos grupos de órdenes gráficas importantes para la creación de gráficos como son:

a) **Alto nivel.** Son funciones que crean un nuevo gráfico, posiblemente con ejes, etiquetas, títulos, etc.

b) **Bajo nivel.** Son funciones que añaden información a un gráfico existente, tales como puntos adicionales, líneas, curvas y etiquetas.

Existe una serie de argumentos que pueden usarse en los gráficos de alto nivel para poder elegir en cada gráfico la forma, el color, el tamaño, los textos añadidos etc. En el Cuadro A.4 se muestran los argumentos más utilizados para la creación de gráficos.
Una de las principales funciones gráficas de alto nivel es la función `plot()`, es una función genérica, es decir, el tipo de gráfico producido depende de la clase del primer argumento. El siguiente ejemplo ilustra la utilidad de esta función.

Ejemplo A.17. *Considere los siguientes vectores*

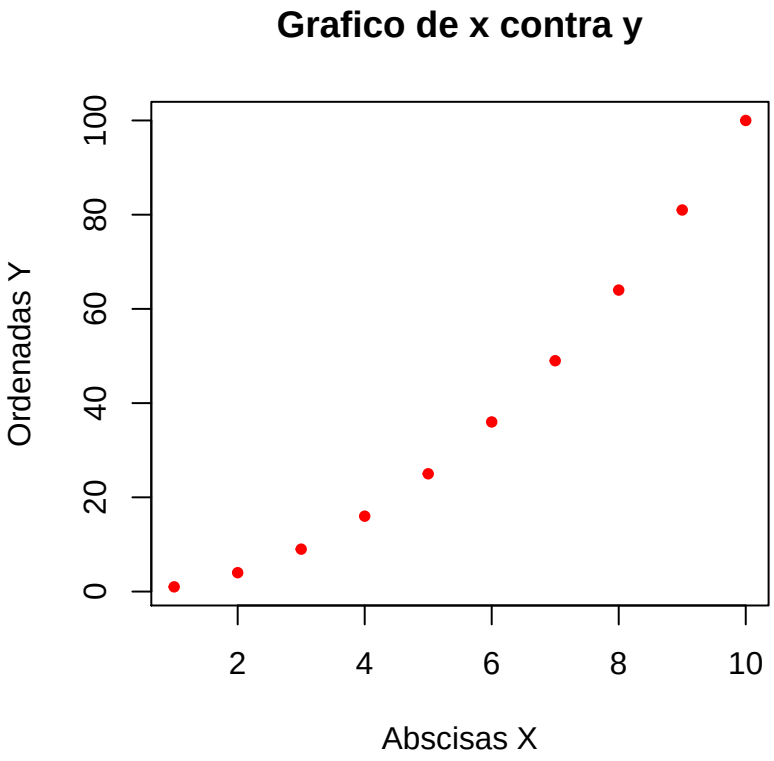
$$a = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$$
$$b = (1, 4, 9, 16, 25, 36, 49, 64, 81, 100)$$

la función `plot()` proporciona un gráfico de dispersión entre los vectores a y b de la siguiente forma:

Función	Descripción
`type="p"`	Representa los datos con puntos (opción por defecto).
`type="l"`	Dibuja una curva conectando los datos.
`type="h"`	Dibuja líneas verticales desde cada punto al eje X.
`type="b"`	Dibuja puntos y líneas uniendo los puntos.
`type="s"`	Dibuja funciones escalonadas donde el punto corresponde al extremo superior de la línea vertical.
`type="S"`	Dibuja funciones escalonadas donde el punto corresponde al extremo inferior de la línea vertical.
`axes=T/F`	Permite generar ejes automaticamente o no.
`main="Título"`	Permite poner un título al gráfico.
`sub="Subtítulo"`	Permite poner un subtítulo al gráfico.
`xlab="Nombre"`	Imprime una etiqueta en el eje de las abscisas.
`ylab="Nombre"`	Imprime una etiqueta en el eje de las ordenadas.
`xlim=c(máx, mín)`	Escala para el eje de las abscisas.
`ylim=c(máx, mín)`	Escla para el eje de las ordenadas.
`lwd=`i	Ancho de la línea ($i = 1, 2, \ldots$)
`lty=1`	Tipo de línea (1=sólida, 2=discontinua, etc.)
`col="Nombre del color"`	Tipo de color (Black, Red, Blue, etc.)

Cuadro A.4: Argumentos para la creación de gráficos.

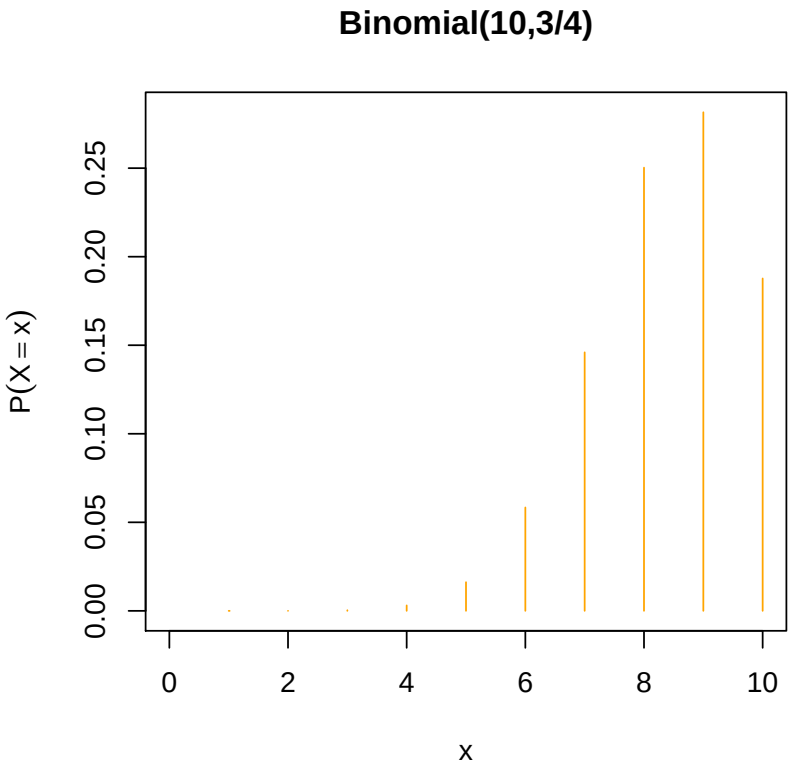
```
# Vectores de datos
> a=c(1,2,3,4,5,6,7,8,9,10)
> b=c(1,4,9,16,25,36,49,64,81,100)
# Gráfico de dispersión
> plot(a,b, main="Grafico de x contra y", xlab="Abscisas X",
+ ylab="Ordenadas Y", type="p", col=red", pch=20)
```



Estamos pidiendo a R que grafique los valores del vectores x contra los valores del vector y, indicando el eje de las abscisas y el eje de las ordenadas, con un título "Gráfico de x contra y", con un puntos de color rojo, estilo 20.

Ejemplo A.18. *Se grafica la función* `binomial()` *que fue creada anteriormente utilizando el comando* `plot()` *con las siguientes líneas de código:*

```
> x <- 0:10
> n <- 10
> p <- 3/4
> plot(Binomial(x,n,p), type="h", col="orange",
+ main="Binomial(10,3/4)", xlim=c(0,10),
+ ylab=expression(P(X==x), cex=0.8),
xlab=expression(x))
```



Una de las instrucciones más importantes en R para representar gráficas mediante curvas fácil y rápidamente es la función `curve()`.

Ejemplo A.19. *A continuación se muestra cómo realizar gráficas en R creando funciones y utilizando la función `curve()`.*

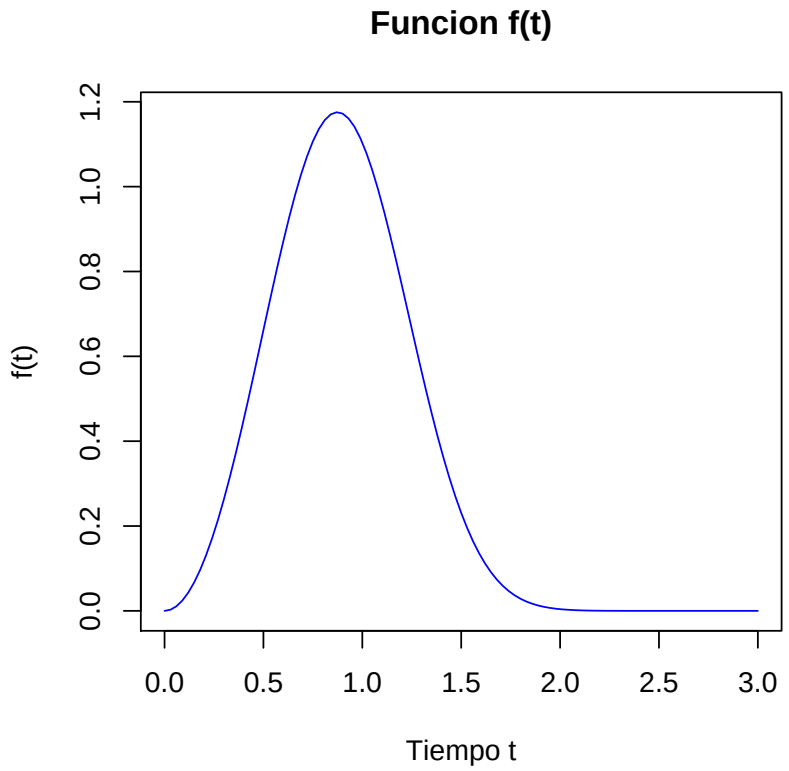
Primero creamos la función que se desea gráficar, en este caso, queremos gráficar la función

$$f(t;\lambda,\alpha) = \lambda\alpha(\lambda t)^{\alpha-1}e^{-(\lambda t)^{\alpha}}.$$

```
# Genera la función f
> f <- function(t){
+ (lam*alpha)*((lam*t)^(alpha-1))*exp(-(lam*t)^alpha)
+ }
```

Asignamos los valores de cada parámetro λ, α y gráficamos la función como sigue:

```
# Valores de los parametros
> lam=1
> alpha=3
# Creacion de la curva
> curve(f(x), col="blue", from=0, to=3,
+ main="Funcion f(t)", xlab="Tiempo t",
+ ylab="f(t)")
```

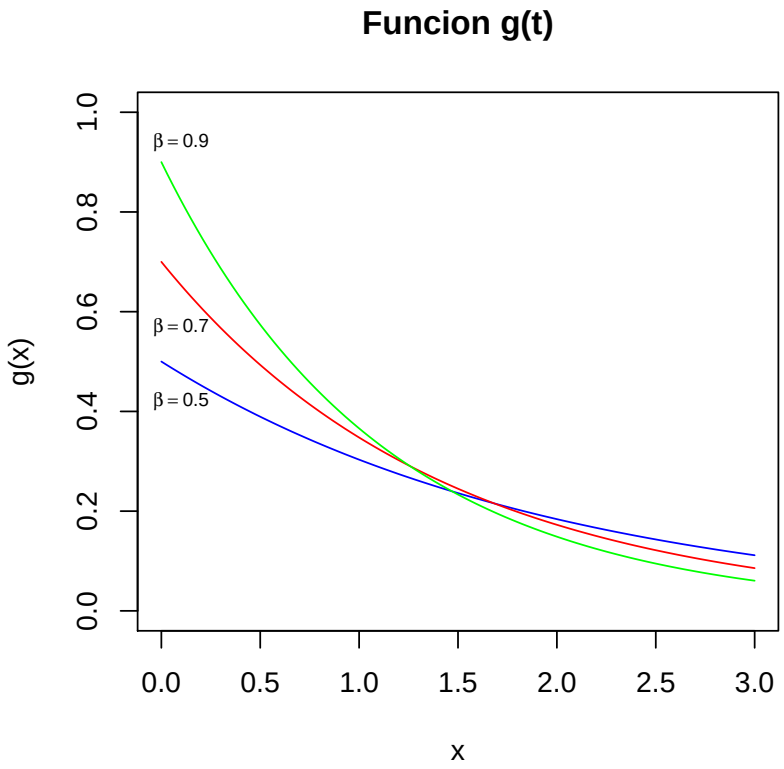


Ejemplo A.20. *Se grafica la función*

$$g(t;\beta) = \beta e^{-\beta t} \quad para\ t > 0$$

cuando β *=0.5, 0.7 y 0.9.*

```
# Genera la funcion g
> g <- function(t){
+ beta*exp(-beta*t)
}
# Curva con β =0.5
> beta=0.5
> curve(g(x), col="blue", main="Funcion g(t)", xlim=c(0,3),
+ ylim=c(0,1))
> text(0.1,0.42, expression(paste(beta==0.5)), cex=0.7)
# Curva con β =0.7
> beta=0.7
> curve(g(x),add=TRUE, col=red")
> text(0.1,0.57, expression(paste(beta==0.7)), cex=0.7)
# Curva con β =0.9
> beta=0.9
> curve(g(x),add=TRUE, col="green")
> text(0.1,0.94, expression(paste(beta==0.9)), cex=0.7)
```



Nótese que se utilizó la instrucción `text()` junto con `expression()` las cuales imprimen en la gráfica los distintos valores de β para las curvas correspondientes.

A.11. Programación en R

Una de las ventajas de R sobre muchos paquetes estadísticos es que su núcleo es un lenguaje de programación con una sintaxis coherente y relativamente moderna. Esto nos permite escribir funciones que simplifican nuestro trabajo y ampliar la funcionalidad de R en nuestros problemas actuales.

Estructuras de programación (loops)

La repetición de algo en términos computacionales se le conoce como *bucle* o en inglés *loop*. R permite crear estructuras repetitivas (loops) y la ejecución condicional de sentencias. En este material se presentan las siguientes estructuras:

a) El bucle *for*.

b) El bucle *while*.

c) La ejecución condicional *if-else*.

Los comandos pueden agruparse entre llaves, utilizando la siguiente sintaxis:

comando1 ; comando2; comando3 ; ....

El bucle for

Para crear un bucle repetitivo (un bucle for), la sintaxis es la siguiente:

```
R for (i in lista-de-valores) {secuencia de comandos}
```

No obstante, los bucles `for` son lentos en `R` (y en Splus), y deben ser evitados en la medida de lo posible.

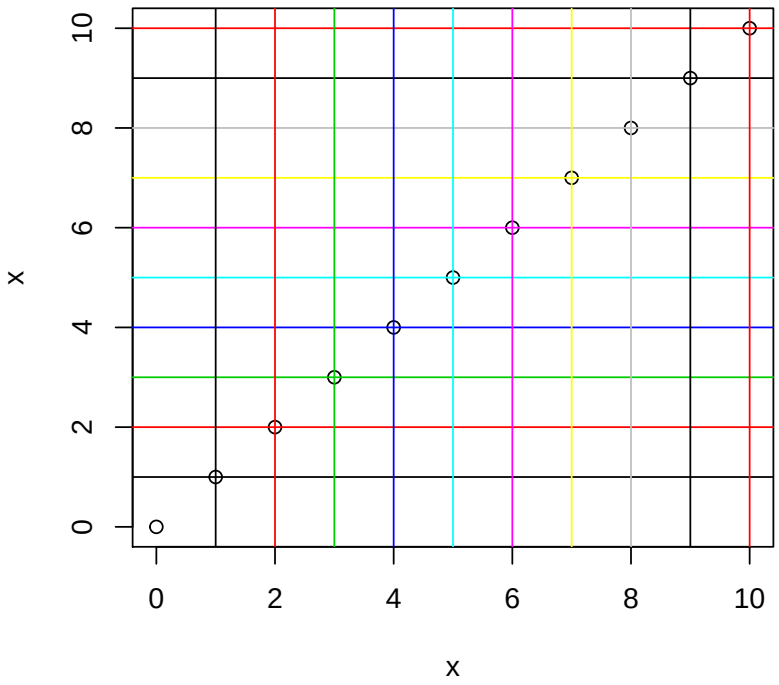
Ejemplo A.21. *En este ejemplo se muestra cómo imprimir los primeros 5 números enteros con ayuda del bucle* ***for()*** *y la instrucción* ***print()****.*

```
> for(i in 1:5)
+ print(i)
```

```
[1] 1
[1] 2
[1] 3
[1] 4
[1] 5
```

Ejemplo A.22. *En este ejemplo se ilustra la creación de un gráfico generando una secuencia de enteros que inicia en el valor -10 y temina en el valor 10 graficando el diagrama de dispersión y pintando distintas lineas horizontales y verticales de distintos colores con la instrucción* ***abline()*** *que se encuentran dentro de dos bucles* ***for()****.*

```
> x = seq(-10,10)
> plot(x, x, xlim=c(0,10), ylim=c(0,10))
> for(i in 1:10)
+ abline(h=i, col=i)
> for(i in 1:10)
+ abline(v=i, col=i)
```



Ejemplo A.23. *Factorial de un número. Las siguientes líneas de código muestran cómo generar un programa para calcular el factorial de un número.*

```
# Factorial de un número
> fact=function(x){
+ ret=1
+ for(i in 1:x){
+ ret=ret*i
+ }
+ return(ret)
+ }
```

Ahora, empleamos la nueva función `fact(x)`.

```
> fact(5)
```

```
[1] 120
```

```
> fact(10)
```

```
[1] 3628800
```

El bucle `while`

La sintaxis para el bucle `while` es la siguiente:

```
while (condición lógica) { expresiones a ejecutar }
```

Ejemplo A.24. *En este ejemplo se le asigna un valor inicial 0 al objeto z y utilizando el bucle* `while()` *se pide que mientras el valor de z sea menor que 5 se vaya sumando 2 unidades y después imprima todos los valores de z que cumplan esa condición.*

```
> z <- 0
> while(z<5){
+ z <- z+2
+ }
> print(z)
```

```
[1] 6
```

Ejemplo A.25. *Este ejemplo muestra el uso del bucle* `while()` *para generar un programa que calcula la suma de los primeros n números naturales.*

```
# Suma de los primeros n naturales
> Otrasuma<-function(x){
+ i=1
+ suma = 0
+ while(i<=x){
+ suma = suma + i
+ i=i+1
+ }
+ print(suma)
+ }
```

Utilizamos la función `Otrasuma(x)`.

```
> Otrasuma(10)
```

```
[1] 55
```

```
> Otrasuma(4)
```

```
[1] 10
```

Condicional `if-else`

La sintaxis para la condicional `if-else` es la siguiente:

```
if(sentencia1, sentencia2,...) else sentencia
```

Ejemplo A.26. *Se generan 50 números aleatorios de una variable aleatoria con distribución exponencial con parámetro* $\lambda = 0.5$ *para el objeto* `s` *y* `r`*. Se crea la función* `Mayor` *la cual compara las medias e imprime el valor de la media de* `s` *si es mayor a la media de* `r`*, de lo contrario imprime la leyenda* `"La media de s no es mayor"`*.*

```
# Genera 50 números aleatorios de una exp(0.5)
> s <- rexp(50, rate=0.5)
> r <- rexp(50, rate=0.5)
# Crea la función Mayor(x) que compara las medias de s y r
> Mayor <- function(x){
+ if(mean(s) >mean(r))
+ print(mean(s))
+ else
+ print("La media de s no es mayor")
+ }
# Imprime el resultado de la función Mayor(x)
> Mayor(x)
```

```
[1] 1.888
```

Ejemplo A.27. *Aquí se muestra la creación de una función que permita calcular la mediana muestral de un conjunto de datos. La mediana muestral se define de la siguiente manera.*

Sea $X_1, \ldots, X_n$ *una muestra aleatoria y considere las estadísticas de orden* $X_{(1)} \leq \ldots \leq X_{(n)}$. *La mediana muestral se define como sigue*

$$Med(X_1, \ldots, X_n) = \begin{cases} X_{(\frac{n+1}{2})} & si\ n\ es\ impar \\ \frac{1}{2}[X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}] & si\ n\ es\ par \end{cases}$$

```
# Función que calcula la mediana muestral
> mediana <- function(x){
# Distingue los valores impar=1 y par=0
+ indicador = length(x)%%2
+ if(indicador==1){
+ mediana = sort(x)[ceiling(length(x)/2)]
+ }
+ else
+ mediana = (sort(x)[length(x)/2]+sort(x)[1+length(x)/2])/2
}
```

La instrucción `ceiling()` tiene un solo argumento numérico y devuelve el valor entero más pequeño no menor que los elementos correspondientes del argumento.

A continuación se compara la nueva función `mediana()` contra la función predeterminada `median()`.

```
# Vector de datos par
> x <- c(3,5,4,8,9,6)
> a <- mediana(x)
> a
```

```
[1] 5.5
```

```
> median(x)
```

```
[1] 5.5
```

```
# Vector de datos impar
> y <-
c(1.2,1.4,1.1,1.8,1.7)
> b <- mediana(y)
> b
```

```
[1] 1.4
```

```
> median(y)
```

```
[1] 1.4
```

Además de la condicional `if-else` existe una instrucción de este tipo sólo que su uso es más simple. Esta instrucción es `ifelse()` y tiene la siguiente estructura:

```
ifelse(test, yes=, no=)
```

donde `test` representa una condición lógica sobre algún objeto. El argumento '`yes`' regresa los valores para los cuales se cumple la condición `test` mientas que '`no`' devuelve los valores para los cuales no se cumple `test`.

Ejemplo A.28. *A continuación se generan 10 números pseudoaleatorios de una distribución $t(3)$ y se pide que si algún número es mayor que cero le asigne 1 y de lo contrario asigne 0.*

```
> X <- rt(n = 10, df = 3)
> ifelse(test = X > 0, yes=1 , no=0)
```

```
 [1] 0 1 0 1 1 0 1 0 0 1
```

A.12. Lectura de datos

En `R` es muy sencillo la lectura de archivos externos. El usuario puede utilizar editores de texto para llevar a cabo este tipo de acciones.

Instrucción `read.table()`. Para poder leer una hoja de datos directamente de un archivo externo, dicho archivo debe reunir las condiciones adecuadas. La forma más sencilla es:

a) La primer línea del archivo debe contener el nombre de cada variable de la hoja de datos.

b) En cada una de las siguientes líneas, el primer elemento es la etiqueta de la fila, y a continuación deben aparecer los valores de cada variable.

Así, el archivo tiene la forma siguiente:

```
   Precio  Superficie  Área
01    52        111      830
02    54        128      710
03    56        147      987
```

Las instrucciones para realizar la lectura de un archivo con extensión `.txt` es la siguiente:

```
# En Windows
> Datos <- read.table("G:/carpeta/archivo.txt")

# En Linux
> Datos <- read.table('/home/carpeta/archivo.txt')
```

En ocasiones no se dispone de etiquetas por renglón. En este caso también es posible la lectura de los datos y automáticamente `R` añade etiquetas predeterminadas. Así, el archivo tendrá la siguiente forma:

```
Precio  Superficie  Área
  52        111      830
  54        128      710
  56        147      987
```

Este archivo podrá leerse utilizando un parámetro adicional, `header=T` el cual indica que la primera línea es de cabecera y que no existen etiquetas por renglones, entonces la lectura de datos se realiza de la siguiente manera:

```
# En Windows
> Datos <- read.table("G:/carpeta/archivo.txt", header=T)

# En Linux
> Datos <- read.table('/home/carpeta/archivo.txt', hea-
der=T)
```

Lectura desde hojas de cálculo

Es de gran interés realizar lecturas de datos desde una hoja de cálculo para exportarla a `R` y llevar a cabo un análisis de dichos datos.

Instrucción `read.csv()`. La instrucción `read.csv()` permite una conección con un hoja de cálculo para la lectura de datos. La hoja de cálculo donde se encuentran capturados los datos de interés debe ser guardada con la extensión `.csv`

Ejemplo A.29. *En este ejemplo se muestra la lectura de datos del archivo mieloma.csv utilizando un sistema operativo Linux.*

```
> data <- read.csv('/home/misra/Datos/mieloma.csv',header=T)
> data
```

```
    t status x1 x2
1  18      1  0  0
2  19      1  0  1
3  28      0  0  0
4  31      1  0  1
5  39      0  0  1
6  19      0  0  1
7  45      0  0  1
8   6      1  0  1
9   8      1  0  1
10 15      1  0  1
11 23      1  0  0
12 28      0  0  0
13  7      1  0  1
14 12      1  1  0
15  9      1  1  0
16  8      1  1  0
17  2      1  1  1
18 26      0  1  0
19 10      1  1  1
20  4      1  1  0
21  3      1  1  0
22  4      1  1  0
23 18      1  1  1
24  8      1  1  1
```

Si se está interesado en sólo utilizar los datos de una sola columna, por ejemplo, los datos de la primer columna, podemos escribir la siguiente instrucción:

```
> data[,1]
```

```
 [1] 18 19 28 31 39 19 45  6  8 15 23 28  7 12  9  8  2 26
[19] 10  4  3  4 18  8
```

Por otro lado, si sólo queremos utilizar algunos datos de nuestra hoja de cálculo, por decir, los primeros 7 datos, una manera sencilla para exportar y utilizar estos datos es *seleccionar los datos que se requieren de nuestra hoja de cálculo y copiarlos*, así como se muestra en la Figura A.3.
Enseguida se escriben las instrucciones

```
# En Windows
> data <- read.table("clipboard", header=T)

# En linux
> data <- read.table(‘clipboard’, header=T)
```

obteniendo

```
> data
```

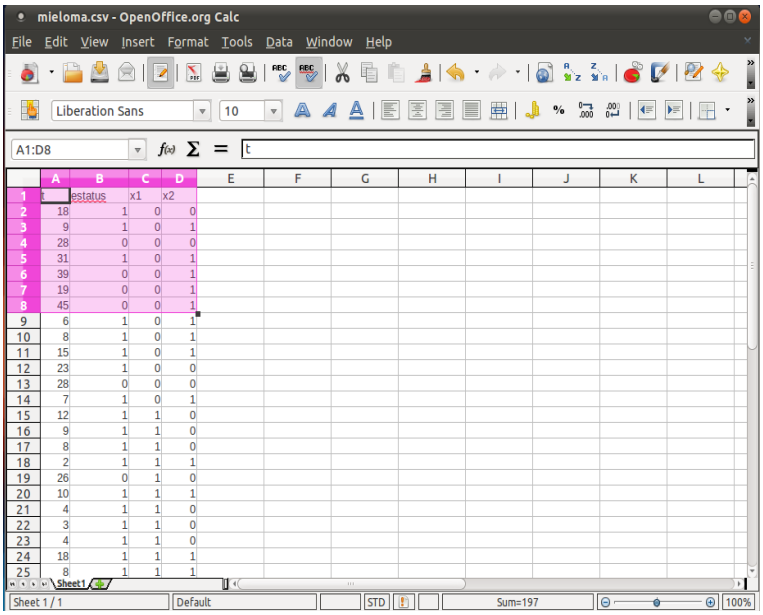


Figura A.3: Exportar datos desde Excel u OpenOffice.

	t	estatus	x1	x2
1	18	1	0	0
2	9	1	0	1
3	28	0	0	0
4	31	1	0	1
5	39	0	0	1
6	19	0	0	1
7	45	0	0	1

A.13. Distribuciones de probabilidad

R soporta algunas distribuciones de probabilidad y existen comandos que permiten calcular la función de *distribución*, $F(x) = P(X \leq x)$, permite calcular cuantiles, la función de *densidad* $f(x) = \mathbb{P}(X = x)$ y generar *números pseudo-aleatorios* de la distribución.

Para calcular probabilidades utilizando una distribución en particular, hay que utilizar el nombre de la distribución precedido de las letras:

a) **'d'** para la función de densidad $f(x)$.

b) **'p'** para la función de distribución $F(x)$.

c) **'q'** para calcular algún cuantil de orden p.

d) **'r'** para generar números pseudoaleatorios.

El Cuadro A.5 resume algunas de las posibles distribuciones que existen en R.

Distribución	Nombre en R	Argumentos (parámetros)
Beta	`beta`	`shape1, shape2`
Binomial	`binom`	`size, prob`
Cauchy	`cauchy`	`location, scale`
Ji-Cuadrada	`chisq`	`df`
Exponencial	`exp`	`rate`
F	`f`	`df1, df1`
Gamma	`gamma`	`shape, scale`
Geométrica	`geom`	`prob`
Hipergeométrica	`hyper`	`m, n, k`
Log-normal	`lnorm`	`meanlog, sdlog`
Logística	`logis`	`location, scale`
Binomial negativa	`nbinom`	`size, prob`
Normal	`norm`	`mean, sd`
Poisson	`pois`	`lambda`
T	`t`	`df`
Uniforme	`unif`	`min, max`
Weibull	`weibull`	`shape, scale`
Wilcoxon	`wilcox`	`m, n`

Cuadro A.5: Distribuciones de probabilidad en R.

Distribución Binomial. La variable aleatoria X sigue una distribución binomial con parámetros $n \in \mathbb{N}$ y $p \in (0,1)$ y se escribe $X \sim binom(n,p)$ si su función de densidad está dada por

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{para } x = 0, 1, \ldots, n.$$

Para poder calcular probabilidades en R utilizando la distribución Binomial podemos hacerlo como se muestra a continuación:

```
# X ~ binom(10, 1/2)
# P(X = 3)
> dbinom(3, size=10, prob=1/2)
```

```
[1] 0.1172
```

```
# P(X ≤ 3)
> pbinom(3, size=10, prob=1/2)
```

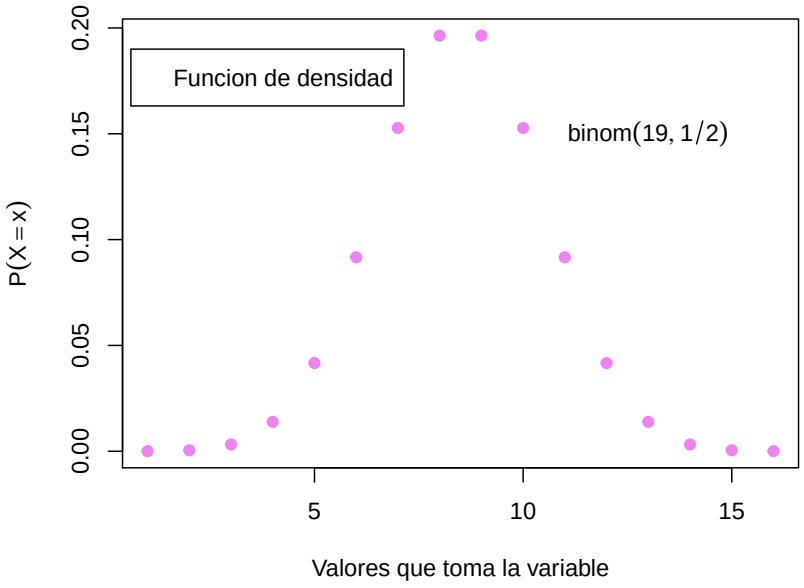
```
[1] 0.1719
```

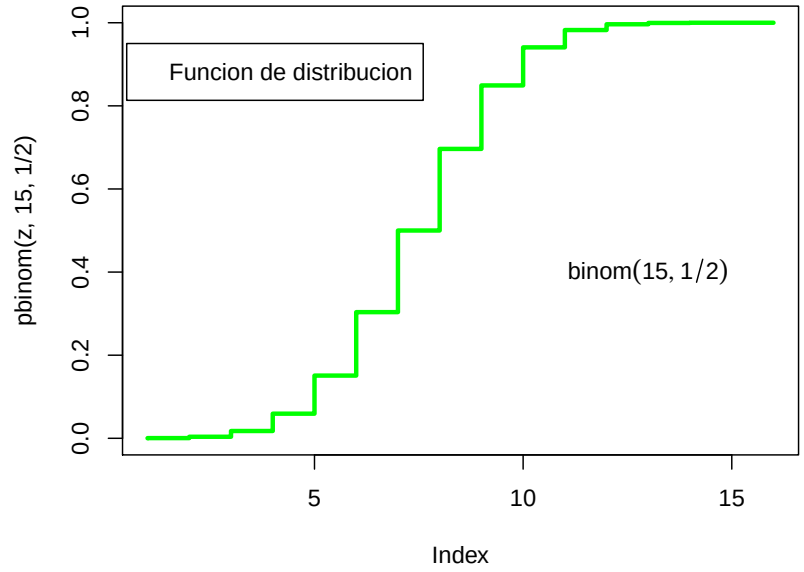
```
# Elimina notación científica y disminuye dígitos
> options(scipen=100, digits=2)
# Sucesión de enteros
> z <- 0:15
# Calcula las probabilidades de z
> dbinom(x=z, size=15, prob=1/2)
```

```
 [1] 0.000031 0.000458 0.003204 0.013885 0.041656 0.091644
 [7] 0.152740 0.196381 0.196381 0.152740 0.091644 0.041656
[13] 0.013885 0.003204 0.000458 0.000031
```

Las siguientes instrucciones generan la gráfica de la función de densidad $f(x)$ y grafica la función de distribución $F(x)$ de la variable aleatoria binomial con $n = 15$ y $p = 1/2$.

```
# Coloca las gráficas en 2 renglones y 1 columna
> par(mfrow=c(2,1))
# Gráfica de la densidad binom(15,1/2)
> plot(dbinom(z,15,1/2), pch=19,
+ col="violet", xlab="Valores que toma la variable",
+ ylab=expression(P(X==x)))
# Leyendas en la gráfica
> text(x=13,y=0.15,expression(binom(19,1/2)),cex=1)
> legend(x=0.6, y=0.19, legend="Funcion de densidad")
# Gráfica de la distribución binom(15,1/2)
> plot(pbinom(z,15,1/2), pch=15, col="green", type="S",
lwd=3)
# Leyendas en la gráfica
> legend(x=0.5, y=0.95, legend="Funcion de distribucion")
> text(13, 0.4, expression(binom(15,1/2)))
```





donde el argumento `expression()` sirve para poner expresiones matemáticas, `lwd=3` dibuja el ancho de la línea y `type="S"` dibuja la gráfica escalonada donde el punto corresponde al extremo inferior de la línea vertical.

Distribución Poisson. La variable aleatoria X sigue una distribución Poisson con parámetro $\lambda > 0$ y se escribe $X \sim Poisson(\lambda)$ si su función de densidad está dada por

$$f(x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!} \quad \text{con } x = 0, 1, \ldots$$

Para calcular probabilidades utilizando la distribución Poisson podemos hacerlo como sigue:

```
# X ~ Poisson(λ = 3/4)
# P(X = 5)
> dpois(x=5, lambda=3/4)
```

```
[1] 0.00093
```

```
# P(X ≤ 5)
> ppois(q=5, lambda=3/4)
```

```
[1] 1
```

```
# P(X > 5) = P(X ≥ 6)
> ppois(q=5, lambda=3/4, lower.tail=F)
```

```
[1] 0.00013
```

El argumento `lower.tail=F` indica que se está tomando la probabilidad de cola superior.

```
# Cuantil q del 25% de probabilidad (cola izquierda)
> qpois(p=0.25, lambda=3/4)
```

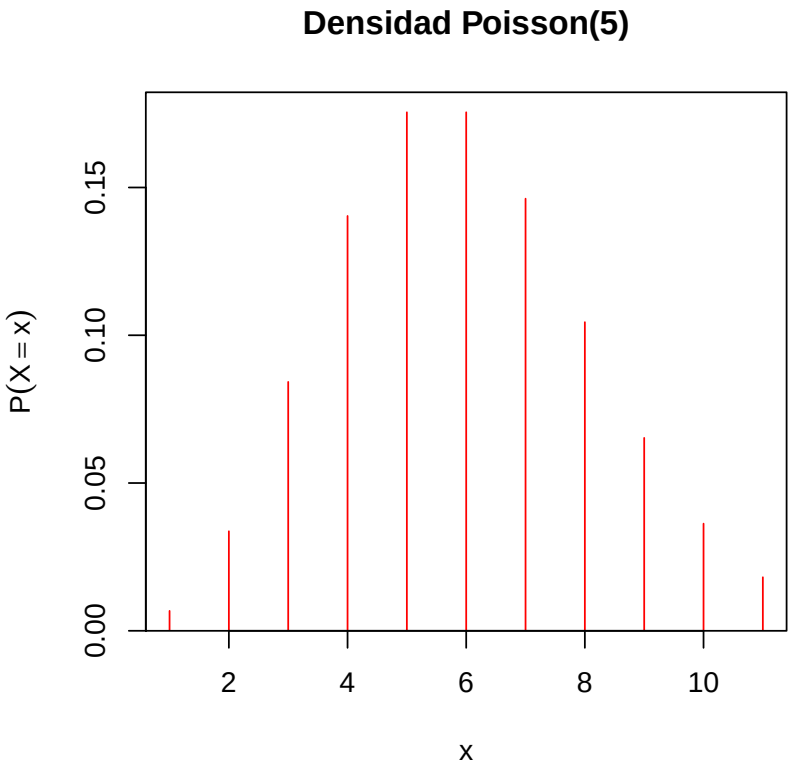
```
[1] 0
```

```
# Cuartiles 25%, 50% y 75% (cola derecha)
> c <- c(0.25, 0.5, 0.75)
> qpois(c, lambda=3/4, lower.tail=F)
```

```
[1] 1 1 0
```

Ahora presentaremos la gráfica de la función de densidad de una variable aleatoria $Poisson(\lambda = 5)$ usando la función `plot()`.

```
> s<-0:10
> plot(dpois(s,lambda=5), type="h", pch=20, col=red",
+ main="Densidad Poisson(5)", xlab="x",
ylab=expression(P(X==x)))
```

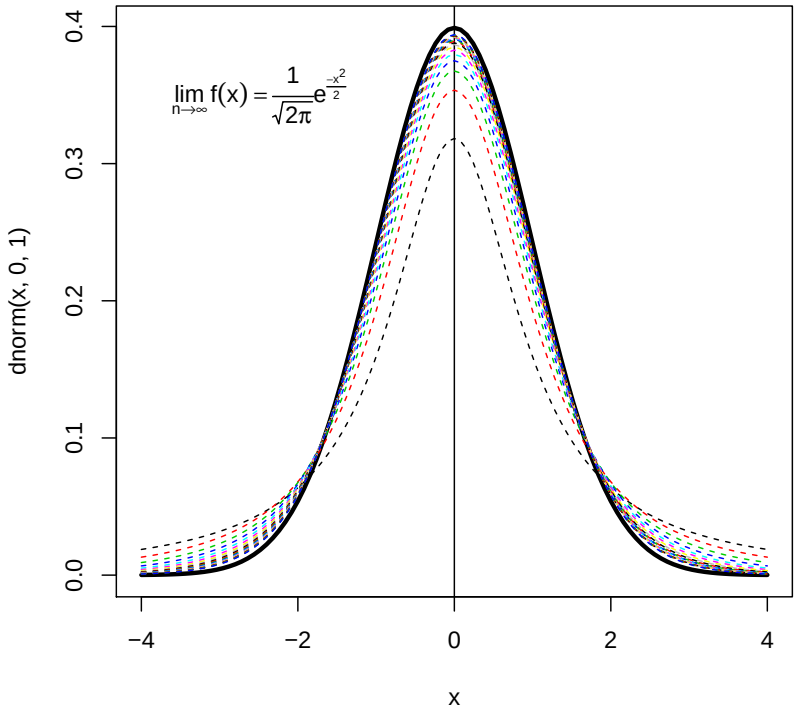


Ejemplo A.30. *En este ejemplo se muestra gráficamente cómo la función de densidad $f(x)$ de una variable aleatoria $t(n)$ tiende a la función de densidad de una distribución normal estándar cuando sus grados de libertad n crecen infinitamente.*

```
> graf<-function(y){
+ curve(dnorm(x,0,1),from=-4,to=4, col="black", lwd=3)
+ text(-2.5,0.35, expression(lim(f(x), n%->% infinity)==
+ frac(1,sqrt(2*pi))*e^frac(-x^2,2)))
+ abline(v=0)
+ for(i in 1:y){
+ curve(dt(x,i), col=i, add=T, lty=2)
+ }
+ }
```

Como resultado tenemos la siguiente gráfica

```
> graf(20)
```



A.14. Estimación por máxima verosimilitud

En esta sección se presenta la función `fitdistr()` la cual proporciona el estimador por máxima verosimilitud del parámetro de una función de probabilidad con base a una muestra de valores. Esta función depende de la biblioteca `MASS`. La función `fitdistr()` tiene la siguiente sintaxis:

```
fitdistr(x, densfun)
```

donde el argumento `x` es un vector numérico y `densfun` especifica la función de densidad asociada al vector `x`. El argumento `desfun` es una cadena de caracteres que puede tomar la siguiente lista de valores: `"beta"`, `"cauchy"`, `"chi-squared"`, `"negative binomial"`, `"exponential"`, `"f"`, `"gamma"`, `"geometric"`, `"lognormal"`, `"logistic"`, `"normal"`, `"Poisson"`, `"t"` y `"weibull"`.

Ejemplo A.31. *A continuación se muestra el uso de la función* ***`fitdistr()`*** *para estimar el parámetro de una muestra que provienen de una distribución Poisson con* $\lambda = 5$.

```
# Carga la biblioteca MASS
> library(MASS)
# Genera 30 números aleatorios de una Poisson(5)
> x <- rpois(n=30, lambda=5)
# Se estima el parámetro
> fitdistr(x, "Poisson")
```

```
  lambda
  5.00
 (0.41)
```

Recordando que un estimador insesgado con varianza mínima uniforme para el parámetro λ de una distribución Poisson es $\bar{X}$, podemos corroborar el valor estimado calculando la media.

```
> mean(x)
```

```
[1] 5
```

A.15. Ayuda en R

Una de las características que tiene R es que existen distintas formas para acceder a la ayuda que éste nos proporciona. Los siguientes comandos son las diferentes maneras para pedirle ayuda a R

a) Documentación en línea de la función ayuda.

```
> help()
```

b) Documentación en línea para algún tema o función en específico.

```
> help(tema)
```

c) Igual que la forma anterior.

```
> help.search("tema")
```

d) Igual que la forma anterior.

```
> ?tema
```

e) Accede a un índice de búsqueda relacionada con el tema solicitado.

```
> ??tema
```

Apéndice B

Pruebas de hipótesis

B.1. Introducción

En este apéndice se da una breve introducción del concepto de pruebas de hipótesis para el caso paramétrico y los elementos con los que cuenta este concepto. Posteriormente, se definen los tipos de errores que se pueden tener al realizar una prueba de hipótesis y la probabilidad de cometer estos errores. Por último, se definen la función potencia, el tamaño de una prueba y el valor p. Para un estudio mas detallado se recomienda consultar [1] o [16].

Supongamos que se afirma que el parámetro de una distribución tiene un cierto valor. ¿Como decidimos que efectivamente el valor dado es el valor real del parámetro? Por ejemplo, supongamos que nos dicen que la media de una distribución normal es μ =3.4. ¿En qué forma podemos probar la "hipótesis" de que μ =3.4? Al tomar una muestra de la población en estudio, se encuentra que la media de la muestra es $\bar{X}$ =2.9, entonces debemos decidir entre aceptar o rechazar que $\bar{X}$ =2.9 coincide con la hipótesis μ =3.4 dentro de cierto "nivel de confianza". Para tomar una decisión como ésta se realiza un *contraste* o *prueba de hipótesis*, la cual es una regla que nos permite optar por un valor u otro del parámetro desconocido θ que se define dentro de un *espacio paramétrico* denotado como Θ.

Antes de realizar un contraste de hipótesis se define una *hipótesis nula*, denotada como H_0, que se considera a priori cierta y ésta se contrasta frente a otra conocida como *hipótesis alternativa*, denotada como H_1 o H_a. Ambas hipótesis originan una *regla de decisión* que define dos regiones complementarias conocidas como *región de rechazo* o *región crítica* y la otra como *región de aceptación*. El problema es determinar una regla para decidir cuándo rechazar la hipótesis nula H_0 favoreciendo la hipótesis alternativa H_1, de esta manera, esta regla será determinada por una *estadística de prueba* que bajo el supuesto de que la hipótesis nula es cierta, representará el resultado con base en los datos de la muestra (evidencia muestral). Si se toma una muestra aleatoria $X_1, X_2, \ldots, X_n$ seleccionada de una distribución $f(x;\theta)$ con θ desconocido, el espacio mues-

tral del vector aleatorio $(X_1, X_2, \ldots, X_n)$, es el conjunto de todos los posibles resultados de $(X_1, X_2, \ldots, X_n)$. Así, la estadística de prueba especifica un procedimiento de contraste donde un subconjunto contiene los valores del vector $(X_1, X_2, \ldots, X_n)$ para los cuales se aceptará H_0 y el otro conjunto contendrá los valores de $(X_1, X_2, \ldots, X_n)$ para los cuales se rechazará H_0.

Definición B.1. *Una hipótesis estadística es una afirmación o conjetura acerca de la distribución de una o más variables aleatorias. Una hipótesis es* simple *si se especifica por completo la distribución de probabilidad en cuestión, en caso contrario, la hipótesis se llama* compuesta.

En la Figura B.1 se ilustra de manera gráfica los tipos de hipótesis y la forma en que cada una especifica o no a la distribución.

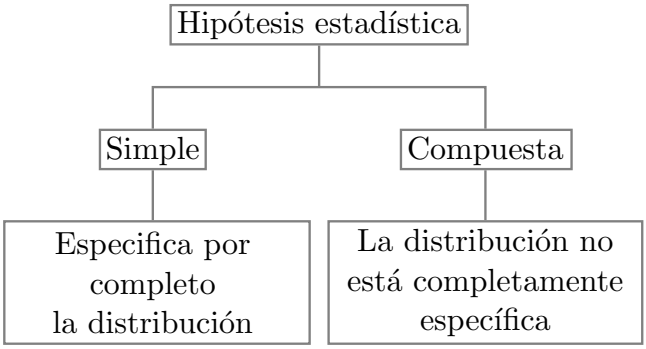


Figura B.1: Tipos de hipótesis estadísticas.

Ejemplo B.1.

a) Si $X \sim \chi^2(k)$, entonces la afirmación "$k \neq k_0$" con k_0 conocida es una hipótesis compuesta.

b) Si $X \sim N(\mu, 4)$, entonces la afirmación "$\mu = 1$" es una hipótesis simple .

c) Si $X \sim Poisson(\lambda)$, entonces la afirmación "$\lambda = 10$" es una hipótesis simple.

d) Si $X \sim Exp(\lambda)$, entonces la afirmación "$\lambda > 5$" es una hipótesis compuesta.

Una prueba de hipótesis consta de los siguientes elementos:

a) Una *hipótesis nula* H_0 que expresa una conjetura sobre el parámetro de la distribución.

b) Una *hipótesis alternativa* H_1.

c) Una *estadística de prueba*, la cual permite decidir si se rechaza la hipótesis nula H_0 a favor de la hipótesis alternativa H_1.

d) Una *región de rechazo* o *región crítica* denotada como $\mathscr{RR}$ o $\mathscr{C}$, que indica para qué valores de la estadística de prueba se debe rechazar la hipótesis H_0.

e) Una *región de aceptación* denotada como $\mathscr{A}$, que indica para qué valores de la estadística de prueba se acepta la hipótesis H_0.

Definición B.2. *Una prueba de hipótesis es una regla para decidir si se acepta la hipótesis nula H_0 o se rechaza en favor de la hipótesis alternativa H_1.*

Como ya hemos dicho, la región crítica $\mathscr{C}$ especifica los valores de la estadística de prueba para la cual la hipótesis nula H_0 se rechaza a favor de la hipótesis alternativa H_1. Análogamente, si la estadística de prueba no cae en dicha región, entonces no se rechaza la hipótesis nula H_0. Esto último indica que no hay evidencia suficiente que sustente una hipótesis diferente a H_0. En resumen tenemos que:

a) Tanto la hipótesis nula H_0 como la hipótesis alternativa H_1 pueden ser simples o compuestas.

b) Si el valor de la estadística de prueba cae en la región de rechazo, se rechaza H_0 (en términos prácticos se acepta H_1).

c) Si el valor la estadística de prueba no cae en la región de rechazo, no se rechaza H_0 (no hay evidencia suficiente para rechazarla y en términos prácticos se acepta).

De este modo tenemos los contrastes:

a) H_0 : Simple *vs* H_1 : Simple.

b) H_0 : Simple *vs* H_1 : Compuesta.

c) H_0 : Compuesta *vs* H_1 : Compuesta.

Ejemplo B.2. *Un veterinario A afirma que la media del peso de cerdos de cierta edad después de aplicar una nueva dieta por tres semanas debe de ser de 80 kilos o menos. Sin embargo, un veterinario B piensa que de acuerdo a su experiencia la media del peso de los cerdos debe haber aumentado aún más. El veterinario B usará una muestra de cerdos para verificar la afirmación del veterinario A.*

a) ¿Cuál de las siguientes pruebas de hipótesis debe usarse para probar la afirmación del veterinario A? Justifique su respuesta.

- $H_0 : \mu \geq 80$ vs $H_1 : \mu < 80$.
- $H_0 : \mu \leq 80$ vs $H_1 : \mu > 80$.
- $H_0 : \mu = 80$ vs $H_1 : \mu \neq 80$.

b) ¿Qué conclusión puede darse cuando no se puede rechazar H_0?

c) ¿Qué concluye el veterinario B cuando si se puede rechazar H_0?

Solución:

a) Con la afirmación del veterinario A se propone la hipótesis nula como H_0 : "La media de los pesos de los cerdos debe ser de 80 kilos o menos". Como consecuencia, la hipótesis alternativa puede verse como H_1 : "La media de los pesos de los cerdos es mayor a los 80 kilos". Entonces, la hipótesis adecuada que debe usarse es la segunda.

b) Si no se puede rechazar H_0, entonces se dice que no existe evidencia suficiente para contradecir al veterinario A, es decir, que la media de los pesos de los cerdos después de aplicar una nueva dieta probablemente es de 80 kilos o menos.

c) Si se rechaza H_0, entonces la información que obtuvo el veterinario B a partir de la muestra, permite decir que la afirmación del veterinario A no es la más adecuada. En consecuencia, al parecer esta nueva dieta es más eficiente para engordar a los cerdos.

▪

B.2. Tipos de errores

Los *errores* que se pueden cometer al tomar una decisión de acuerdo a una prueba de hipótesis se clasifican en dos tipos.

Definición B.3. *Se definen los errores tipo I y tipo II de la siguiente manera:*

Error tipo I. *Consiste en rechazar H_0 cuando es verdadera.*
Error tipo II. *Consiste en aceptar H_0 cuando es falsa.*

En la Figura B.2 se muestra los posibles resultados de una prueba de hipótesis.

	H_0 verdadera	H_0 falsa
Aceptar H_0	Decisión correcta	Error tipo II
Rechazar H_0	Error tipo I	Decisión correcta

Figura B.2: Tipos de errores de una prueba de hipótesis.

A la probabilidad de cometer el error tipo I se denota como α y se escribe:

$$\alpha = \mathbb{P}[\text{"Error tipo I"}],$$

mientras que a la probabilidad de cometer el error tipo II se denota como β y se escribe:

$$\beta = \mathbb{P}[\text{“Error tipo II”}].$$

Nótese que la suma de los errores $\alpha + \beta$ no es 1, ya que cometer el error tipo I y cometer el error tipo II no son eventos complementarios, sin embargo, α y β se desea que sean lo más pequeñas posibles.

Definición B.4. *Suponiendo una prueba de hipótesis:*

a) A la probabilidad de cometer el error tipo I (α) se llama tamaño de la región crítica ó nivel de significancia.

b) El complemento de α, es decir, aceptar H_0 siendo que H_0 es verdadera (decisión correcta) tiene probabilidad $1 - \alpha$ y se conoce como nivel de confianza de la prueba.

c) A la probabilidad β del error tipo II se conoce como característica de operación *de la prueba.*

d) Al complemento de la probabilidad de cometer el error tipo II, es decir, $1 - \beta$ (decisión también correcta) se le llama potencia de la prueba. El número $1 - \beta$ es la “potencia” de rechazar una hipótesis falsa.

En la Figura B.3 se muestra un resumen de las definiciones anteriores.

	H_0 verdadera	H_0 falsa
Aceptar H_0	Decisión correcta $1 - \alpha$ nivel de confianza	Error tipo II β Característica de operación
Rechazar H_0	Error tipo I α nivel de significancia	Decisión correcta $1 - \beta$ potencia de la prueba

Figura B.3: Complementos de los errores.

Ejemplo B.3. *Sea $X_1, X_2, \ldots, X_5$ una muestra aleatoria de una distribución Bernoulli(p). La variable aleatoria $Y = \sum_{i=1}^{5} X_i$ se distribuye binom$(5, p)$. Dada la hipótesis*

$$H_0 : p = 1/2 \quad \text{vs} \quad H_1 : p > 1/2$$

a) Hallar α para $\mathscr{C} = \{Y \geq 4\}$ y $\mathscr{C} = \{Y = 5\}$.

b) Calcular $\beta(0.8)$ y $\beta(0.9)$ para $\mathscr{C} = \{Y \geq 4\}$ y $\mathscr{C} = \{Y = 5\}$.

Solución:

a) Para la región crítica $\mathscr{C} = \{Y \geq 4\}$ tenemos que

$$\begin{aligned}
\alpha &= \mathbb{P}[\text{error tipo I}] \\
&= \mathbb{P}[\text{rechazar } H_0 \text{ cuando } H_0 \text{ es verdadera}] \\
&= \mathbb{P}[Y \geq 4 \text{ cuando } p = 1/2] \\
&= \binom{5}{4}(1/2)^4(1/2)^1 + \binom{5}{5}(1/2)^5(1/2)^0 \\
&= 5(1/32) + 1/32 \\
&= 3/16 \\
&= 0.1875.
\end{aligned}$$

Para la región crítica $\mathscr{C} = \{Y = 5\}$

$$\begin{aligned}
\alpha &= \mathbb{P}[\text{error tipo I}] \\
&= \mathbb{P}[\text{rechazar } H_0 \text{ cuando } H_0 \text{ es verdadera}] \\
&= \mathbb{P}[Y = 5 \text{ cuando } p = 1/2] \\
&= \binom{5}{5}(1/2)^5(1/2)^0 \\
&= 1/32 \\
&= 0.03125.
\end{aligned}$$

b) Para calcular β, nótese que depende del parámetro p, ya que la hipótesis H_1 es compuesta.

Para la región crítica $\mathscr{C} = \{Y \geq 4\}$

$$\begin{aligned}
\beta(p) &= \mathbb{P}[\text{error tipo II}] \\
&= \mathbb{P}[\text{aceptar } H_0 \text{ cuando } H_0 \text{ es falsa}] \\
&= \mathbb{P}[Y \leq 3 \text{ cuando } p > 1/2]
\end{aligned}$$

entonces

$$\begin{aligned}
\beta(0.8) &= \mathbb{P}[Y \leq 3 \text{ cuando } p = 0.8] \\
&= 1 - \mathbb{P}[Y > 3 \text{ cuando } p = 0.8] \\
&= 1 - \left\{ \binom{5}{4}(0.8)^4(0.2)^1 + \binom{5}{5}(0.8)^5(0.2)^0 \right\} \\
&= 1 - \{5(0.08192) + 0.32768\} \\
&= 1 - (0.4096 + 0.32768) \\
&= 0.26272
\end{aligned}$$

y

$$\begin{aligned}\beta(0.9) &= \mathbb{P}[Y \leq 3 \text{ cuando } p = 0.9] \\ &= 1 - \mathbb{P}[Y > 3 \text{ cuando } p = 0.9] \\ &= 1 - \left\{ \binom{5}{4}(0.9)^4(0.1)^1 + \binom{5}{5}(0.9)^5(0.1)^0 \right\} \\ &= 1 - \{5(0.06561) + 0{,}59049\} \\ &= 1 - (0.91854) \\ &= 0.08145.\end{aligned}$$

Para la región crítica $\mathscr{C} = \{Y = 5\}$

$$\begin{aligned}\beta(p) &= \mathbb{P}[\text{error tipo II}] \\ &= \mathbb{P}[\text{aceptar } H_0 \text{ cuando } H_0 \text{ es falsa}] \\ &= \mathbb{P}[Y \leq 4 \text{ cuando } p > 1/2]\end{aligned}$$

entonces

$$\begin{aligned}\beta(0.8) &= \mathbb{P}[Y \leq 4 \text{ cuando } p = 0.8] \\ &= 1 - \mathbb{P}[Y = 5 \text{ cuando } p = 0.8] \\ &= 1 - \binom{5}{5}(0.8)^5(0.2)^0 \\ &= 1 - 0.32758 \\ &= 0.67232\end{aligned}$$

y

$$\begin{aligned}\beta(0.9) &= \mathbb{P}[Y \leq 4 \text{ cuando } p = 0.9] \\ &= 1 - \mathbb{P}[Y = 5 \text{ cuando } p = 0.9] \\ &= 1 - \binom{5}{5}(0.9)^5(0.1)^0 \\ &= 1 - 0.59049 \\ &= 0.40951.\end{aligned}$$

▪

B.3. Función potencia

En esta sección introducimos el concepto de función potencia, la cual nos permitirá generalizar algunos conceptos, por ejemplo el nivel de significancia.

Definición B.5. *Suponiendo la prueba de hipótesis*

$$H_0 : \theta \in \Theta_0 \quad vs \quad H_1 : \theta \in \Theta_1$$

donde Θ_0 y Θ_1 son subconjuntos del espacio paramétrico Θ. La función potencia de una prueba se define como:

$$\Pi(\theta) = \mathbb{P}[Rechazar\ H_0 | \theta].$$

En otras palabras, la función potencia es la probabilidad de que una muestra $X_1, ..., X_n$ se encuentre en la región de rechazo de la prueba, $\mathscr{C}$, esto es

$$\Pi(\theta) = P[(X_1, ..., X_n) \in \mathscr{C}|\theta].$$

Puesto que $\Pi(\theta)$ depende de cada posible valor del parámetro θ, la función potencia ideal sería:

$$\Pi(\theta) = \begin{cases} 0 & \text{para } \theta \in \Theta_0 \text{ (hipótesis nula)}, \\ 1 & \text{para } \theta \in \Theta_1 \text{ (hipótesis alternativa)}, \end{cases}$$

es decir,

$$\mathbb{P}[\text{Rechazar } H_0|H_0] = 0 \quad \text{cuando } \theta \in \Theta_0$$

y

$$\mathbb{P}[\text{Rechazar } H_0|H_1] = 1 \quad \text{cuando } \theta \in \Theta_1.$$

Lo anterior es indicativo de que no se rechaza H_0 cuando es cierta, y se rechaza cuando es falsa, lo cual significa que se está tomando la decisión correcta. Esto es porque la probabilidad de rechazar H_0 cuando es verdadera ($\theta \in \Theta_0$) es nula y la probabilidad de rechazar H_0 cuando es falsa ($\theta \in \Theta_1$) es uno.

B.4. Tamaño de la prueba

En esta sección se define el llamado tamaño de una prueba y corresponde a una definición más general del nivel de significancia α que puede usarse sin importar el tipo de hipótesis nula o alternativa usadas.

Definición B.6. *Considere una prueba de hipótesis*

$$H_0 : \theta \in \Theta_0 \quad vs \quad H_1 : \theta \in \Theta_1$$

con función potencia $\Pi(\theta)$*. El tamaño de la prueba se define como*

$$\sup_{\theta \in \Theta_0} \Pi(\theta)$$

y corresponde a la máxima probabilidad de rechazar H_0 *cuando* H_0 *es verdadera.*

Verificamos a continuación que el tamaño de la prueba y el nivel de significancia coinciden. Por definición,

$$\sup_{\theta \in \Theta_0} \Pi(\theta) = \sup \mathbb{P}(\text{“Rechazar} H_0''|\text{“}H_0 \text{ es cierta”}) = \sup \mathbb{P}(\text{“Error tipo I”}) = \alpha.$$

En otras palabras, el nivel de significancia corresponde a la probabilidad obtenida cuando se maximiza el error tipo I sobre todas los posibles valores del parámetro según el rango que éste puede tomar de acuerdo al conjunto Θ_0. En el siguiente ejemplo se muestra una prueba de hipótesis en la cual H_0 y H_1 son compuestas y se obtiene el nivel de significancia o tamaño de la prueba a partir de la función potencia.

Ejemplo B.4. *Sea $X_1, X_2, \ldots, X_n$ muestra aleatoria de tamaño $n = 16$ de una población $N(\theta, 16)$, donde el parámetro θ es desconocido. Considere la prueba de hipótesis*

$$H_0 : \theta \leq 15 \quad vs \quad H_1 : \theta > 15$$

donde se rechaza H_0 si $\bar{X} > 15 + 4/\sqrt{n}$. Encontrar el tamaño de la prueba.

Solución: La región crítica $\mathscr{C} = \{(x_1, \ldots, x_{16}) : \bar{x} > 15 + 4/\sqrt{n}\}$ y por el contraste de hipótesis tenemos que $\Theta_0 = \{\theta : \theta \leq 15\}$ y $\Theta_1 = \{\theta : \theta > 15\}$. A partir de la definición de la función potencia tenemos

$$\begin{aligned}
\Pi(\theta) &= \mathbb{P}[\text{Rechazar } H_0|\theta] \\
&= \mathbb{P}\left[\bar{X} > 15 + 4/\sqrt{n} \ |\theta\right] \\
&= \mathbb{P}\left[\frac{\bar{X} - \theta}{4/\sqrt{n}} > \frac{15 + \frac{4}{\sqrt{n}} - \theta}{4/\sqrt{n}}\right] \\
&\approx \mathbb{P}[Z > 16 - \theta],
\end{aligned}$$

en donde $Z \sim N(0, 1)$. Por lo tanto la función potencia está dada por

$$\Pi(\theta) = 1 - \mathbb{P}(Z \leq 16 - \theta).$$

Podemos evaluar la función potencia para distintos valores del parámetro θ dentro de los espacios Θ_0 y Θ_1 obteniendo

$\theta \in \Theta_0$	$\Pi(\theta)$	$\theta \in \Theta_1$	$\Pi(\theta)$
13	0.0013	15.5	0.3085
13.5	0.0062	16	0.5000
14	0.0228	16.5	0.6915
14.5	0.0067	17	0.8413
15	**0.1587**	17.5	0.9332

La función potencia para este ejemplo se muestra en la Figura B.4.

Denotando a la probabilidad acumulada $\mathbb{P}(Z \leq z)$ como $\Phi(z)$, observando la forma creciente de la función potencia y que $\theta \in \Theta_0$ a lo más puede ser igual a 15, entonces tenemos que el supremo de la función potencia en Θ_0 se alcanza en $\theta = 15$, así que

$$\begin{aligned}
\sup_{\theta \leq 15}[1 - \Phi(16 - \theta)] &= 1 - \phi(1) \\
&= 1 - 0.8413 \\
&= 0.1587.
\end{aligned}$$

Por lo que el tamaño de la prueba es $\alpha = 0.1587$. En palabras, esto quiere decir que la probabilidad de cometer el error tipo I es de 0.1587, así que en 15.87 % de las veces la decisión es incorrecta (rechazar H_0 cuando es cierta) y en el restante 84.13 % de las veces la decisión es la correcta (no rechazar H_0 cuando es cierta) en el sentido del error tipo I.

▪

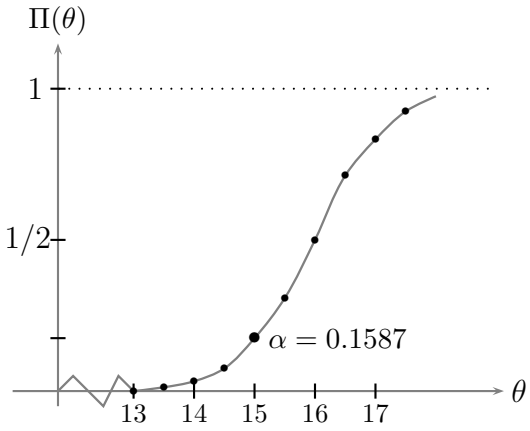


Figura B.4: Gráfica de la función potencia asociada al Ejemplo B.4.

B.5. Valor p

Como se ha visto, la probabilidad de cometer el error tipo I es conocido como nivel de significancia, o simplemente el nivel de la prueba. Aunque se recomiendan valores pequeños de α, el tamaño real de α que se utiliza en el análisis es seleccionado arbitrariamente. Por ejemplo, si un investigador elige llevar a cabo una prueba de hipótesis usando $\alpha = 5\,\%$, mientras que otro investigador podría elegir $\alpha = 1\,\%$, es posible que para estos dos investigadores que se encuentran analizando los mismos datos puedan llegar a conclusiones opuestas, es decir, mientras uno concluye que la hipótesis nula H_0 debe ser rechazada con un nivel de significancia $\alpha = 5\,\%$, el otro investigador decide que H_0 debe ser aceptada con $\alpha = 1\,\%$.

Una vez que se decide realizar una prueba de hipótesis es posible encontrar el valor p asociado a dicha prueba, donde esta cantidad representa el valor más pequeño de α para el cual la hipótesis nula H_0 puede ser rechazada.

Definición B.7. *El valor p o nivel de significancia observado, denotado por v_p, de una prueba estadística es el valor más pequeño de α para el cual la hipótesis nula H_0 puede ser rechazada.*

El valor p es la probabilidad de cometer el error tipo I y se calcula en base al valor observado de la estadística de prueba. El valor p mide la fuerza de la evidencia en contra de H_0. Por lo tanto, si el valor propuesto de α es mayor o igual que el valor p, la hipótesis nula H_0 se rechaza. En la Figura B.5 se muestra parte de la gráfica de una función de densidad $f(x)$ y se muestra cuándo rechazar la hipótesis nula H_0 utilizando el valor p.
De esta manera el valor p es la probabilidad de obtener una discrepancia mayor o igual que la observada cuando H_0 es cierta. La estadística de prueba, denotada como T, es una medida con base a la muestra observada que indica cuándo es razonable rechazar H_0 cuando H_0 es cierta. El valor observado de la estadística T, y se escribe t_{obs}, se utiliza para definir el valor p en cada situación.

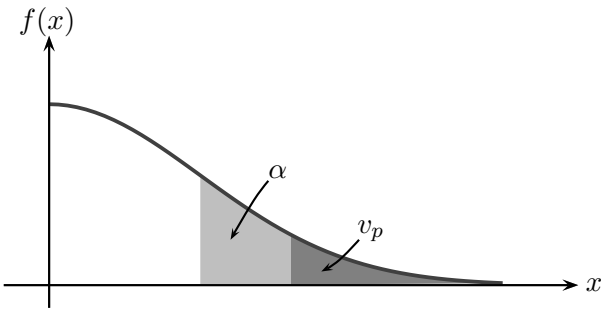


Figura B.5: Se rechaza H_0 cuando el valor $p \leq \alpha$.

Una prueba de cola superior es aquella en la que se especifican valores del parámetro θ desconocido mayores al valor dado θ_0 en la hipótesis alternativa H_1, es decir $H_1 : \theta > \theta_0$. Análogamente, una prueba de cola inferior es en la que se especifican valores para el parámetro θ desconocido menores al valor dado θ_0 en la hipótesis alternativa H_1, es decir $H_1 : \theta < \theta_0$. A dichas pruebas, también se les conoce como *pruebas unilaterales*. De manera similar, una prueba de dos colas es aquella en la que el valor θ es igual al valor θ_0, o bien $H_1 : \theta = \theta_0$.

En una prueba de cola superior el valor p se define como

$$\mathbb{P}(T \geq t_{obs}|H_0).$$

En una prueba de cola inferior el valor p viene dado por

$$\mathbb{P}(T \leq t_{obs}|H_0).$$

Para una prueba de dos colas distinguiremos si la distribución es simétrica o no.

a) Si la *distribución es simétrica*, se define el valor p como dos veces la probabilidad de la cola en la que se encuentre el valor observado.

   Si $t_{obs} > 0$ en un contraste de dos colas el valor p está dado por

   $$v_p = 2\mathbb{P}(T \geq t_{obs}|H_0)$$

   Si $t_{obs} < 0$ entonces

   $$v_p = 2\mathbb{P}(T \leq t_{obs}|H_0)$$

b) Si la *distribución no es simétrica*, el valor p está dado por

   $$v_p = 2\,\text{mín}\{\mathbb{P}(T \leq t_{obs}|H_0),\ \mathbb{P}(T \geq t_{obs}|H_0)\}$$

Observación B.1. *El valor p no se elige como el nivel de significancia α, sino que se determina a partir de la muestra observada.*

Ejemplo B.5. *En un cierto hospital se registraron 200 partos en el mes de enero, de los cuales 115 fueron niñas. Una pregunta que surge es: ¿La proporción de niños que nacen es la misma que la proporción de niñas?*

Solución: En el hospital se obtuvo $X_1, X_2, \ldots, X_{200}$ una muestra aleatoria de tamaño $n = 200$ y cada nacimiento se pueden modelar con una variable aleatoria con distribución $Bernoulli(p)$ donde

$$X_i = \begin{cases} 1 & \text{si es niña} \\ & \\ 0 & \text{si es niño} \end{cases} \qquad \text{para toda} \quad i = 1, \ldots, 200.$$

y p desconocido. De esta forma $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ sigue una distribución $binom(n, p)$ con media p y varianza $p(1-p)/n$. Por lo tanto,

$$\frac{\widehat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \sim N(0, 1).$$

Dos posibles contrastes de hipótesis de interés son:

a) $H_0 : p = 1/2$ *vs* $H_1 : p > 1/2$ (la proporción de nacimientos de niñas y niños es la misma contra la proporción de niñas que nacen es mayor).

b) $H_0 : p = 1/2$ *vs* $H_1 : p \neq 1/2$ (la proporción de nacimientos de niñas y niños es la misma contra la proporción de nacimientos de niñas y niños es distinta).

El problema es encontrar una regla para decidir cuando rechazar H_0 en favor de H_1 con base en los datos de la muestra aleatoria. Cuando H_0 es cierta, es decir, cuando $p = 1/2$, tenemos que

$$Z = \frac{\widehat{p} - 1/2}{\sqrt{\dfrac{1/2(1-1/2)}{n}}} \sim N(0, 1).$$

La estadística Z es una medida de la distancia entre $\widehat{p}$ y su valor esperado $1/2$ cuando H_0 es cierta. Entonces es razonable rechazar H_0 cuando la estadística Z sea grande. Se tiene que $\widehat{p} = 115/200$ =0.575 entonces el valor observado de la estadística Z es $z_{obs} = 2.12132$.
La prueba de hipótesis del inciso a) corresponde a una prueba de cola superior y el valor p es $\mathbb{P}(Z \geq 2.12132)$, entonces

```
# ℙ(Z ≥ 2.12132)
> pnorm(q=2.12132 ,mean=0, sd=1, lower.tail=F)
```

```
[1] 0.016947
```

Si el nivel de significancia $\alpha = 5\,\%$ se rechaza la hipótesis nula H_0 ya que el valor p es menor que α. En consecuencia estamos aceptando que nacen más niñas que niños y esta conclusión va acorde con los resultados de la muestra. Para poder aceptar la hipótesis nula $H : p = 1/2$ se tendría que tener un nivel de significancia aproximadamente menor o igual a $1.6\,\%$, en otras palabras, la probabilidad de rechazar la hipótesis nula cuando es cierta tiene que ser menor o igual a 0.016, o bien, que la probabilidad de cometer el error tipo I tiene que ser muy pequeña.

Para la prueba del inciso b) corresponde a una prueba de dos colas y el valor p es

$$2\,\text{mín}\{\mathbb{P}(Z \leq 2.12132), \mathbb{P}(Z \geq 2.12132)\}$$

entonces

```
# ℙ(Z ≤ 2.12132)
> i <- pnorm(q=2.12132, mean=0, sd=1)
# ℙ(Z ≥ 2.12132)
> s <- pnorm(q=2.12132, mean=0, sd=1,lower.tail=F)
> 2*min(i,s)
```

```
[1] 0.033895
```

Si $\alpha = 5\,\%$, entonces se rechaza la hipótesis H_0 ya que el valor p es menor que el nivel de significancia α. Por lo tanto, se considera que las proporciones de los nacimientos de niñas y niños son distintos. ∎

Apéndice C

Tablas estadísticas

C.1. Probabilidades de la distribución binomial

		p									
n	x	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
2	0	0.903	0.810	0.723	0.640	0.563	0.490	0.423	0.360	0.303	0.250
	1	0.095	0.180	0.255	0.320	0.375	0.420	0.455	0.480	0.495	0.500
	2	0.003	0.010	0.023	0.040	0.063	0.090	0.123	0.160	0.203	0.250
3	0	0.857	0.729	0.614	0.512	0.422	0.343	0.275	0.216	0.166	0.125
	1	0.135	0.243	0.325	0.384	0.422	0.441	0.444	0.432	0.408	0.375
	2	0.007	0.027	0.057	0.096	0.141	0.189	0.239	0.288	0.334	0.375
	3	0.000	0.001	0.003	0.008	0.016	0.027	0.043	0.064	0.091	0.125
4	0	0.815	0.656	0.522	0.410	0.316	0.240	0.179	0.130	0.092	0.063
	1	0.171	0.292	0.368	0.410	0.422	0.412	0.384	0.346	0.299	0.250
	2	0.014	0.049	0.098	0.154	0.211	0.265	0.311	0.346	0.368	0.375
	3	0.000	0.004	0.011	0.026	0.047	0.076	0.111	0.154	0.200	0.250
	4	0.000	0.000	0.001	0.002	0.004	0.008	0.015	0.026	0.041	0.063
5	0	0.774	0.590	0.444	0.328	0.237	0.168	0.116	0.078	0.050	0.031
	1	0.204	0.328	0.392	0.410	0.396	0.360	0.312	0.259	0.206	0.156
	2	0.021	0.073	0.138	0.205	0.264	0.309	0.336	0.346	0.337	0.313
	3	0.001	0.008	0.024	0.051	0.088	0.132	0.181	0.230	0.276	0.313
	4	0.000	0.000	0.002	0.006	0.015	0.028	0.049	0.077	0.113	0.156
	5	0.000	0.000	0.000	0.000	0.001	0.002	0.005	0.010	0.018	0.031
6	0	0.735	0.531	0.377	0.262	0.178	0.118	0.075	0.047	0.028	0.016
	1	0.232	0.354	0.399	0.393	0.356	0.303	0.244	0.187	0.136	0.094
	2	0.031	0.098	0.176	0.246	0.297	0.324	0.328	0.311	0.278	0.234
	3	0.002	0.015	0.041	0.082	0.132	0.185	0.235	0.276	0.303	0.313
	4	0.000	0.001	0.005	0.015	0.033	0.060	0.095	0.138	0.186	0.234
	5	0.000	0.000	0.000	0.002	0.004	0.010	0.020	0.037	0.061	0.094
	6	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.004	0.008	0.016
7	0	0.698	0.478	0.321	0.210	0.133	0.082	0.049	0.028	0.015	0.008
	1	0.257	0.372	0.396	0.367	0.311	0.247	0.185	0.131	0.087	0.055
	2	0.041	0.124	0.210	0.275	0.311	0.318	0.298	0.261	0.214	0.164
	3	0.004	0.023	0.062	0.115	0.173	0.227	0.268	0.290	0.292	0.273
	4	0.000	0.003	0.011	0.029	0.058	0.097	0.144	0.194	0.239	0.273
	5	0.000	0.000	0.001	0.004	0.012	0.025	0.047	0.077	0.117	0.164
	6	0.000	0.000	0.000	0.000	0.001	0.004	0.008	0.017	0.032	0.055
	7	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.004	0.008
8	0	0.663	0.430	0.272	0.168	0.100	0.058	0.032	0.017	0.008	0.004
	1	0.279	0.383	0.385	0.336	0.267	0.198	0.137	0.090	0.055	0.031
	2	0.051	0.149	0.238	0.294	0.311	0.296	0.259	0.209	0.157	0.109
	3	0.005	0.033	0.084	0.147	0.208	0.254	0.279	0.279	0.257	0.219
	4	0.000	0.005	0.018	0.046	0.087	0.136	0.188	0.232	0.263	0.273
	5	0.000	0.000	0.003	0.009	0.023	0.047	0.081	0.124	0.172	0.219
	6	0.000	0.000	0.000	0.001	0.004	0.010	0.022	0.041	0.070	0.109
	7	0.000	0.000	0.000	0.000	0.000	0.001	0.003	0.008	0.016	0.031
	8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.004

Tabla C.1 (continuación)

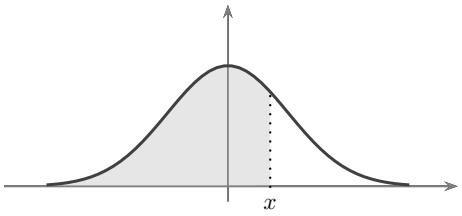
		p									
n	x	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
9	0	0.630	0.387	0.232	0.134	0.075	0.040	0.021	0.010	0.005	0.002
	1	0.299	0.387	0.368	0.302	0.225	0.156	0.100	0.060	0.034	0.018
	2	0.063	0.172	0.260	0.302	0.300	0.267	0.216	0.161	0.111	0.070
	3	0.008	0.045	0.107	0.176	0.234	0.267	0.272	0.251	0.212	0.164
	4	0.001	0.007	0.028	0.066	0.117	0.172	0.219	0.251	0.260	0.246
	5	0.000	0.001	0.005	0.017	0.039	0.074	0.118	0.167	0.213	0.246
	6	0.000	0.000	0.001	0.003	0.009	0.021	0.042	0.074	0.116	0.164
	7	0.000	0.000	0.000	0.000	0.001	0.004	0.010	0.021	0.041	0.070
	8	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.004	0.008	0.018
	9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.002
10	0	0.599	0.349	0.197	0.107	0.056	0.028	0.013	0.006	0.003	0.001
	1	0.315	0.387	0.347	0.268	0.188	0.121	0.072	0.040	0.021	0.010
	2	0.075	0.194	0.276	0.302	0.282	0.233	0.176	0.121	0.076	0.044
	3	0.010	0.057	0.130	0.201	0.250	0.267	0.252	0.215	0.166	0.117
	4	0.001	0.011	0.040	0.088	0.146	0.200	0.238	0.251	0.238	0.205
	5	0.000	0.001	0.008	0.026	0.058	0.103	0.154	0.201	0.234	0.246
	6	0.000	0.000	0.001	0.006	0.016	0.037	0.069	0.111	0.160	0.205
	7	0.000	0.000	0.000	0.001	0.003	0.009	0.021	0.042	0.075	0.117
	8	0.000	0.000	0.000	0.000	0.000	0.001	0.004	0.011	0.023	0.044
	9	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.004	0.010
	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001
11	0	0.569	0.314	0.167	0.086	0.042	0.020	0.009	0.004	0.001	0.000
	1	0.329	0.384	0.325	0.236	0.155	0.093	0.052	0.027	0.013	0.005
	2	0.087	0.213	0.287	0.295	0.258	0.200	0.140	0.089	0.051	0.027
	3	0.014	0.071	0.152	0.221	0.258	0.257	0.225	0.177	0.126	0.081
	4	0.001	0.016	0.054	0.111	0.172	0.220	0.243	0.236	0.206	0.161
	5	0.000	0.002	0.013	0.039	0.080	0.132	0.183	0.221	0.236	0.226
	6	0.000	0.000	0.002	0.010	0.027	0.057	0.099	0.147	0.193	0.226
	7	0.000	0.000	0.000	0.002	0.006	0.017	0.038	0.070	0.113	0.161
	8	0.000	0.000	0.000	0.000	0.001	0.004	0.010	0.023	0.046	0.081
	9	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.005	0.013	0.027
	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.005
	11	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
12	0	0.540	0.282	0.142	0.069	0.032	0.014	0.006	0.002	0.001	0.000
	1	0.341	0.377	0.301	0.206	0.127	0.071	0.037	0.017	0.008	0.003
	2	0.099	0.230	0.292	0.283	0.232	0.168	0.109	0.064	0.034	0.016
	3	0.017	0.085	0.172	0.236	0.258	0.240	0.195	0.142	0.092	0.054
	4	0.002	0.021	0.068	0.133	0.194	0.231	0.237	0.213	0.170	0.121
	5	0.000	0.004	0.019	0.053	0.103	0.158	0.204	0.227	0.222	0.193
	6	0.000	0.000	0.004	0.016	0.040	0.079	0.128	0.177	0.212	0.226
	7	0.000	0.000	0.001	0.003	0.011	0.029	0.059	0.101	0.149	0.193
	8	0.000	0.000	0.000	0.001	0.002	0.008	0.020	0.042	0.076	0.121
	9	0.000	0.000	0.000	0.000	0.000	0.001	0.005	0.012	0.028	0.054
	10	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.007	0.016
	11	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.003
	12	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Tabla C.1 (continuación)

		p									
n	x	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
13	0	0.513	0.254	0.121	0.055	0.024	0.010	0.004	0.001	0.000	0.000
	1	0.351	0.367	0.277	0.179	0.103	0.054	0.026	0.011	0.004	0.002
	2	0.111	0.245	0.294	0.268	0.206	0.139	0.084	0.045	0.022	0.010
	3	0.021	0.100	0.190	0.246	0.252	0.218	0.165	0.111	0.066	0.035
	4	0.003	0.028	0.084	0.154	0.210	0.234	0.222	0.184	0.135	0.087
	5	0.000	0.006	0.027	0.069	0.126	0.180	0.215	0.221	0.199	0.157
	6	0.000	0.001	0.006	0.023	0.056	0.103	0.155	0.197	0.217	0.209
	7	0.000	0.000	0.001	0.006	0.019	0.044	0.083	0.131	0.177	0.209
	8	0.000	0.000	0.000	0.001	0.005	0.014	0.034	0.066	0.109	0.157
	9	0.000	0.000	0.000	0.000	0.001	0.003	0.010	0.024	0.050	0.087
	10	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.006	0.016	0.035
	11	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.004	0.010
	12	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002
	13	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
14	0	0.488	0.229	0.103	0.044	0.018	0.007	0.002	0.001	0.000	0.000
	1	0.359	0.356	0.254	0.154	0.083	0.041	0.018	0.007	0.003	0.001
	2	0.123	0.257	0.291	0.250	0.180	0.113	0.063	0.032	0.014	0.006
	3	0.026	0.114	0.206	0.250	0.240	0.194	0.137	0.085	0.046	0.022
	4	0.004	0.035	0.100	0.172	0.220	0.229	0.202	0.155	0.104	0.061
	5	0.000	0.008	0.035	0.086	0.147	0.196	0.218	0.207	0.170	0.122
	6	0.000	0.001	0.009	0.032	0.073	0.126	0.176	0.207	0.209	0.183
	7	0.000	0.000	0.002	0.009	0.028	0.062	0.108	0.157	0.195	0.209
	8	0.000	0.000	0.000	0.002	0.008	0.023	0.051	0.092	0.140	0.183
	9	0.000	0.000	0.000	0.000	0.002	0.007	0.018	0.041	0.076	0.122
	10	0.000	0.000	0.000	0.000	0.000	0.001	0.005	0.014	0.031	0.061
	11	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.003	0.009	0.022
	12	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.006
	13	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001
	14	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
15	0	0.463	0.206	0.087	0.035	0.013	0.005	0.002	0.000	0.000	0.000
	1	0.366	0.343	0.231	0.132	0.067	0.031	0.013	0.005	0.002	0.000
	2	0.135	0.267	0.286	0.231	0.156	0.092	0.048	0.022	0.009	0.003
	3	0.031	0.129	0.218	0.250	0.225	0.170	0.111	0.063	0.032	0.014
	4	0.005	0.043	0.116	0.188	0.225	0.219	0.179	0.127	0.078	0.042
	5	0.001	0.010	0.045	0.103	0.165	0.206	0.212	0.186	0.140	0.092
	6	0.000	0.002	0.013	0.043	0.092	0.147	0.191	0.207	0.191	0.153
	7	0.000	0.000	0.003	0.014	0.039	0.081	0.132	0.177	0.201	0.196
	8	0.000	0.000	0.001	0.003	0.013	0.035	0.071	0.118	0.165	0.196
	9	0.000	0.000	0.000	0.001	0.003	0.012	0.030	0.061	0.105	0.153
	10	0.000	0.000	0.000	0.000	0.001	0.003	0.010	0.024	0.051	0.092
	11	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.007	0.019	0.042
	12	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.005	0.014
	13	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.003
	14	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	15	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Se tabulan los valores de la función de densidad $f(x) = \mathbb{P}(X = x)$ para $x = 0, 1, \ldots, n$ para una variable aleatoria $X \sim \mathit{binom}(n, p)$. Para calcular la función de distribución $F(x) = \mathbb{P}(X \leq x)$ se toma $\sum_{i=0}^{x} \mathbb{P}(X = i)$.

C.2. Probabilidades de la distribución normal estándar

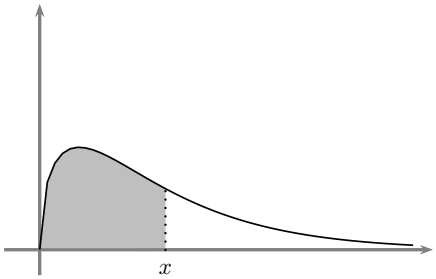


$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Tabla elaborada utilizando `R`.

C.3. Cuantiles de la distribución Ji-cuadrada

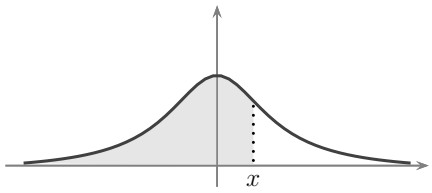


$$F(x) = \int_0^x \frac{1}{\Gamma(n/2)} \left(\frac{1}{2}\right)^{n/2} t^{n/2-1} e^{-t/2} dt$$

	$F(x) = 0.50$	0.75	0.90	0.95	0.975	0.99	0.995	0.999
$n = 1$	0.455	1.323	2.706	3.841	5.024	6.635	7.879	10.828
2	1.386	2.773	4.605	5.991	7.378	9.210	10.597	13.816
3	2.366	4.108	6.251	7.815	9.348	11.345	12.838	16.266
4	3.357	5.385	7.779	9.488	11.143	13.277	14.860	18.467
5	4.351	6.626	9.236	11.070	12.833	15.086	16.750	20.515
6	5.348	7.841	10.645	12.592	14.449	16.812	18.548	22.458
7	6.346	9.037	12.017	14.067	16.013	18.475	20.278	24.322
8	7.344	10.219	13.362	15.507	17.535	20.090	21.955	26.124
9	8.343	11.389	14.684	16.919	19.023	21.666	23.589	27.877
10	9.342	12.549	15.987	18.307	20.483	23.209	25.188	29.588
11	10.341	13.701	17.275	19.675	21.920	24.725	26.757	31.264
12	11.340	14.845	18.549	21.026	23.337	26.217	28.300	32.909
13	12.340	15.984	19.812	22.362	24.736	27.688	29.819	34.528
14	13.339	17.117	21.064	23.685	26.119	29.141	31.319	36.123
15	14.339	18.245	22.307	24.996	27.488	30.578	32.801	37.697
16	15.338	19.369	23.542	26.296	28.845	32.000	34.267	39.252
17	16.338	20.489	24.769	27.587	30.191	33.409	35.718	40.790
18	17.338	21.605	25.989	28.869	31.526	34.805	37.156	42.312
19	18.338	22.718	27.204	30.144	32.852	36.191	38.582	43.820
20	19.337	23.828	28.412	31.410	34.170	37.566	39.997	45.315
21	20.337	24.935	29.615	32.671	35.479	38.932	41.401	46.797
22	21.337	26.039	30.813	33.924	36.781	40.289	42.796	48.268
23	22.337	27.141	32.007	35.172	38.076	41.638	44.181	49.728
24	23.337	28.241	33.196	36.415	39.364	42.980	45.559	51.179
25	24.337	29.339	34.382	37.652	40.646	44.314	46.928	52.620
26	25.336	30.435	35.563	38.885	41.923	45.642	48.290	54.052
27	26.336	31.528	36.741	40.113	43.195	46.963	49.645	55.476
28	27.336	32.620	37.916	41.337	44.461	48.278	50.993	56.892
29	28.336	33.711	39.087	42.557	45.722	49.588	52.336	58.301
30	29.336	34.800	40.256	43.773	46.979	50.892	53.672	59.703
40	39.335	45.616	51.805	55.758	59.342	63.691	66.766	73.402
50	49.335	56.334	63.167	67.505	71.420	76.154	79.490	86.661
60	59.335	66.981	74.397	79.082	83.298	88.379	91.952	99.607
70	69.334	77.577	85.527	90.531	95.023	100.425	104.215	112.317
80	79.334	88.130	96.578	101.879	106.629	112.329	116.321	124.839
90	89.334	98.650	107.565	113.145	118.136	124.116	128.299	137.208
100	99.334	109.141	118.498	124.342	129.561	135.807	140.169	149.449
z_p	0.000	0.675	1.282	1.645	1.960	2.326	2.576	3.090

Para $n > 100$ se utiliza la aproximación $x_p = \frac{1}{2}(z_p + \sqrt{2n-1})^2$, o más exacto $x_p = n\left(1 - \frac{2}{9n} + z_p\sqrt{\frac{2}{9n}}\right)^3$, donde z_p es el cuantil de orden p de la distribución normal estándar. Las entradas de esta tabla corresponden a los cuantiles de orden p de una variable aleatoria ji-cuadrada con n grados de libertad, seleccionando $\mathbb{P}(X \leq x_p) = p$ y $\mathbb{P}(X > x_p) = 1 - p$.

C.4. Cuantiles de la distribución t

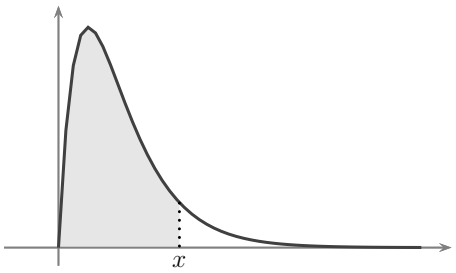


$$F(x) = \int_{-\infty}^{x} \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\,\Gamma(n/2)}(1+t^2/n)^{-(n+1)/2}dt$$

	$F(x) = 0.6$	0.75	0.9	0.95	0.975	0.99	0.995	0.9975
$n = 1$	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.321
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971
50	0.255	0.679	1.299	1.676	2.009	2.403	2.678	2.937
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	2.915
70	0.254	0.678	1.294	1.667	1.994	2.381	2.648	2.899
80	0.254	0.678	1.292	1.664	1.990	2.374	2.639	2.887
90	0.254	0.677	1.291	1.662	1.987	2.368	2.632	2.878
100	0.254	0.677	1.290	1.660	1.984	2.364	2.626	2.871
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807

Tabla elaborada utilizando `R`. Las entradas de esta tabla corresponden a los cuantiles x_p de la distribución t para distintos grados de libertad n. Para encontrar cuantiles $x_p < 0.5$ pueden ser calculados como $x_p = -x_{1-p}$.

C.5. Cuantiles de la distribución F



$$F(x) = \int_0^{\infty} \frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)} \left(\frac{m}{n}\right)^{m/2} x^{m/2-1} \left(1+\frac{m}{n}x\right)^{-(m+n)/2}$$

$F(x) = 0.75$

$m \backslash n$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	5.83	7.50	8.20	8.58	8.82	8.98	9.10	9.19	9.26	9.32	9.37	9.41	9.44	9.47	9.49
2	2.57	3.00	3.15	3.23	3.28	3.31	3.34	3.35	3.37	3.38	3.39	3.39	3.40	3.41	3.41
3	2.02	2.28	2.36	2.39	2.41	2.42	2.43	2.44	2.44	2.44	2.45	2.45	2.45	2.45	2.46
4	1.81	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08
5	1.69	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89
6	1.62	1.76	1.78	1.79	1.79	1.78	1.78	1.78	1.77	1.77	1.77	1.77	1.77	1.76	1.76
7	1.57	1.70	1.72	1.72	1.71	1.71	1.70	1.70	1.69	1.69	1.69	1.68	1.68	1.68	1.68
8	1.54	1.66	1.67	1.66	1.66	1.65	1.64	1.64	1.63	1.63	1.63	1.62	1.62	1.62	1.62
9	1.51	1.62	1.63	1.63	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.58	1.58	1.57	1.57
10	1.49	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55	1.55	1.54	1.54	1.54	1.53
11	1.47	1.58	1.58	1.57	1.56	1.55	1.54	1.53	1.53	1.52	1.52	1.51	1.51	1.51	1.50
12	1.46	1.56	1.56	1.55	1.54	1.53	1.52	1.51	1.51	1.50	1.49	1.49	1.49	1.48	1.48
13	1.45	1.55	1.55	1.53	1.52	1.51	1.50	1.49	1.49	1.48	1.47	1.47	1.47	1.46	1.46
14	1.44	1.53	1.53	1.52	1.51	1.50	1.49	1.48	1.47	1.46	1.46	1.45	1.45	1.44	1.44
15	1.43	1.52	1.52	1.51	1.49	1.48	1.47	1.46	1.46	1.45	1.44	1.44	1.43	1.43	1.43
16	1.42	1.51	1.51	1.50	1.48	1.47	1.46	1.45	1.44	1.44	1.43	1.43	1.42	1.42	1.41
17	1.42	1.51	1.50	1.49	1.47	1.46	1.45	1.44	1.43	1.43	1.42	1.41	1.41	1.41	1.40
18	1.41	1.50	1.49	1.48	1.46	1.45	1.44	1.43	1.42	1.42	1.41	1.40	1.40	1.40	1.39
19	1.41	1.49	1.49	1.47	1.46	1.44	1.43	1.42	1.41	1.41	1.40	1.40	1.39	1.39	1.38
20	1.40	1.49	1.48	1.47	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.39	1.38	1.38	1.37

Tabla C.5 (continuación)

$F(x) = 0.80$

$m \diagdown n$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	9.47	12.00	13.06	13.64	14.01	14.26	14.44	14.58	14.68	14.77	14.84	14.90	14.95	15.00	15.04
2	3.56	4.00	4.16	4.24	4.28	4.32	4.34	4.36	4.37	4.38	4.39	4.40	4.40	4.41	4.42
3	2.68	2.89	2.94	2.96	2.97	2.97	2.97	2.98	2.98	2.98	2.98	2.98	2.98	2.98	2.98
4	2.35	2.47	2.48	2.48	2.48	2.47	2.47	2.47	2.46	2.46	2.46	2.46	2.45	2.45	2.45
5	2.18	2.26	2.25	2.24	2.23	2.22	2.21	2.20	2.20	2.19	2.19	2.18	2.18	2.18	2.18
6	2.07	2.13	2.11	2.09	2.08	2.06	2.05	2.04	2.03	2.03	2.02	2.02	2.01	2.01	2.01
7	2.00	2.04	2.02	1.99	1.97	1.96	1.94	1.93	1.93	1.92	1.91	1.91	1.90	1.90	1.89
8	1.95	1.98	1.95	1.92	1.90	1.88	1.87	1.86	1.85	1.84	1.83	1.83	1.82	1.82	1.81
9	1.91	1.93	1.90	1.87	1.85	1.83	1.81	1.80	1.79	1.78	1.77	1.76	1.76	1.75	1.75
10	1.88	1.90	1.86	1.83	1.80	1.78	1.77	1.75	1.74	1.73	1.72	1.72	1.71	1.70	1.70
11	1.86	1.87	1.83	1.80	1.77	1.75	1.73	1.72	1.70	1.69	1.69	1.68	1.67	1.67	1.66
12	1.84	1.85	1.80	1.77	1.74	1.72	1.70	1.69	1.67	1.66	1.65	1.65	1.64	1.63	1.63
13	1.82	1.83	1.78	1.75	1.72	1.69	1.68	1.66	1.65	1.64	1.63	1.62	1.61	1.61	1.60
14	1.81	1.81	1.76	1.73	1.70	1.67	1.65	1.64	1.63	1.62	1.61	1.60	1.59	1.58	1.58
15	1.80	1.80	1.75	1.71	1.68	1.66	1.64	1.62	1.61	1.60	1.59	1.58	1.57	1.56	1.56
16	1.79	1.78	1.74	1.70	1.67	1.64	1.62	1.61	1.59	1.58	1.57	1.56	1.55	1.55	1.54
17	1.78	1.77	1.72	1.68	1.65	1.63	1.61	1.59	1.58	1.57	1.56	1.55	1.54	1.53	1.53
18	1.77	1.76	1.71	1.67	1.64	1.62	1.60	1.58	1.56	1.55	1.54	1.53	1.53	1.52	1.51
19	1.76	1.75	1.70	1.66	1.63	1.61	1.58	1.57	1.55	1.54	1.53	1.52	1.51	1.51	1.50
20	1.76	1.75	1.70	1.65	1.62	1.60	1.58	1.56	1.54	1.53	1.52	1.51	1.50	1.50	1.49

Tabla C.5 (continuación)

$F(x) = 0.90$

$m \diagdown n$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.47	60.71	60.90	61.07	61.22
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.40	9.41	9.41	9.42	9.42
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.22	5.21	5.20	5.20
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.91	3.90	3.89	3.88	3.87
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.28	3.27	3.26	3.25	3.24
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.92	2.90	2.89	2.88	2.87
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.68	2.67	2.65	2.64	2.63
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.52	2.50	2.49	2.48	2.46
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.40	2.38	2.36	2.35	2.34
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.30	2.28	2.27	2.26	2.24
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.23	2.21	2.19	2.18	2.17
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.17	2.15	2.13	2.12	2.10
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.12	2.10	2.08	2.07	2.05
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.07	2.05	2.04	2.02	2.01
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.04	2.02	2.00	1.99	1.97
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	2.01	1.99	1.97	1.95	1.94
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.98	1.96	1.94	1.93	1.91
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.95	1.93	1.92	1.90	1.89
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.93	1.91	1.89	1.88	1.86
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.91	1.89	1.87	1.86	1.84

Tabla C.5 (continuación)

$F(x) = 0.95$

$m \diagdown n$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	242.98	243.91	244.69	245.36	245.95
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.40	19.41	19.42	19.42	19.43
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71	8.70
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87	5.86
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64	4.62
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.94
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.51
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.22
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	3.01
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86	2.85
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.62
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.53
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.46
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42	2.40
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.40	2.37	2.35
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.35	2.33	2.31
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.31	2.29	2.27
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.28	2.26	2.23
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.25	2.22	2.20

Tabla C.5 (continuación)

$F(x) = 0.975$

$m \diagdown n$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28	968.63	973.03	976.71	979.84	982.53	984.87
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.41	39.42	39.43	39.43
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.37	14.34	14.30	14.28	14.25
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.79	8.75	8.71	8.68	8.66
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.57	6.52	6.49	6.46	6.43
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.41	5.37	5.33	5.30	5.27
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.71	4.67	4.63	4.60	4.57
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.24	4.20	4.16	4.13	4.10
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.91	3.87	3.83	3.80	3.77
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.66	3.62	3.58	3.55	3.52
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.47	3.43	3.39	3.36	3.33
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.32	3.28	3.24	3.21	3.18
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.20	3.15	3.12	3.08	3.05
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.09	3.05	3.01	2.98	2.95
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	3.01	2.96	2.92	2.89	2.86
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.93	2.89	2.85	2.82	2.79
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.87	2.82	2.79	2.75	2.72
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.81	2.77	2.73	2.70	2.67
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.76	2.72	2.68	2.65	2.62
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.72	2.68	2.64	2.60	2.57

Tabla C.5 (continuación)

$F(x) = 0.99$

$m \backslash n$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6083.32	6106.32	6125.86	6142.67	6157.28
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.41	99.42	99.42	99.43	99.43
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.13	27.05	26.98	26.92	26.87
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.45	14.37	14.31	14.25	14.20
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.96	9.89	9.82	9.77	9.72
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.66	7.60	7.56
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47	6.41	6.36	6.31
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67	5.61	5.56	5.52
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11	5.05	5.01	4.96
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71	4.65	4.60	4.56
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40	4.34	4.29	4.25
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16	4.10	4.05	4.01
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.91	3.86	3.82
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.75	3.70	3.66
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.61	3.56	3.52
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.62	3.55	3.50	3.45	3.41
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.46	3.40	3.35	3.31
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43	3.37	3.32	3.27	3.23
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30	3.24	3.19	3.15
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23	3.18	3.13	3.09

Tabla elaborada utilizando `R`. Se tabulan los cuantiles $f^{1-\alpha}_{(n,m)}$ tal que $\mathbb{P}\left(X \leq f^{1-\alpha}_{(n,m)}\right) = 1-\alpha$, donde la variable X sigue una distribución F con n y m grados de libertad. La distribución F tiene la siguiente propiedad:

$$f^{1-\alpha}_{(n,m)} = \frac{1}{f^{\alpha}_{(m,n)}}$$

C.6. Probabilidades de la distribución del número total de corridas R

En cada entrada de la etiqueta p se tabula la probabilidad acumulada de cola izquierda del valor de R, el número total de corridas en una sucesión de $n = n_1 + n_2$ símbolos de dos tipos para $n_1 \leq n_2$.

Cola izquierda

n_1	n_2	r	p	n_1	n_2	r	p	n_1	n_2	r	p	n_1	n_2	r	p
2	2	2	0.333	2	18	2	0.011	3	14	2	0.003	4	10	2	0.002
2	3	2	0.200			3	0.105			3	0.025			3	0.014
		3	0.500			4	0.284			4	0.111			4	0.068
2	4	2	0.133	3	3	2	0.100			5	0.350			5	0.203
		3	0.400			3	0.300	3	15	2	0.002			6	0.419
2	5	2	0.095	3	4	2	0.057			3	0.022	4	11	2	0.001
		3	0.333			3	0.200			4	0.091			3	0.011
2	6	2	0.071	3	5	2	0.036			5	0.331			4	0.055
		3	0.286			3	0.143	3	16	2	0.002			5	0.176
2	7	2	0.056			4	0.429			3	0.020			6	0.374
		3	0.250	3	6	2	0.024			4	0.082	4	12	2	0.001
2	8	2	0.044			3	0.107			5	0.314			3	0.009
		3	0.222			4	0.345	3	17	2	0.002			4	0.045
2	9	2	0.036	3	7	2	0.017			3	0.018			5	0.154
		3	0.200			3	0.083			4	0.074			6	0.335
		4	0.491			4	0.283			5	0.298	4	13	2	0.001
2	10	2	0.030	3	8	2	0.012	4	4	2	0.029			3	0.007
		3	0.182			3	0.067			3	0.114			4	0.037
		4	0.455			4	0.236			4	0.371			5	0.136
2	11	2	0.026	3	9	2	0.009	4	5	2	0.016			6	0.302
		3	0.167			3	0.055			3	0.071	4	14	2	0.001
		4	0.423			4	0.200			4	0.262			3	0.006
2	12	2	0.022			5	0.491			5	0.500			4	0.031
		3	0.154	3	10	2	0.007	4	6	2	0.010			5	0.121
		4	0.396			3	0.045			3	0.048			6	0.274
2	13	2	0.019			4	0.171			4	0.190	4	15	2	0.001
		3	0.143			5	0.455			5	0.405			3	0.005
		4	0.371	3	11	2	0.005	4	7	2	0.006			4	0.027
2	14	2	0.017			3	0.038			3	0.033			5	0.108
		3	0.133			4	0.148			4	0.142			6	0.249
		4	0.350			5	0.423			5	0.333	4	16	2	0.000
2	15	2	0.015	3	12	2	0.004	4	8	2	0.004			3	0.004
		3	0.125			3	0.033			3	0.024			4	0.023
		4	0.331			4	0.130			4	0.109			5	0.097
2	16	2	0.013			5	0.396			5	0.279			6	0.227
		3	0.118	3	13	2	0.004	4	9	2	0.003	5	5	2	0.008
		4	0.314			3	0.029			3	0.018			3	0.040
2	17	2	0.012			4	0.114			4	0.085			4	0.167
		3	0.111			5	0.371			5	0.236			5	0.357
		4	0.298							6	0.471				

Tabla C.6 (continuación)

Cola izquierda

n_1	n_2	r	p	n_1	n_2	r	p	n_1	n_2	r	p	n_1	n_2	r	p
5	6	2	0.004	5	14	2	0.000	6	11	2	0.000	7	9	2	0.000
		3	0.024			3	0.002			3	0.001			3	0.001
		4	0.110			4	0.011			4	0.009			4	0.010
		5	0.262			5	0.044			5	0.036			5	0.035
5	7	2	0.003			6	0.125			6	0.108			6	0.108
		3	0.015			7	0.299			7	0.242			7	0.231
		4	0.076			8	0.496			8	0.436			8	0.427
		5	0.197	5	15	2	0.000	6	12	2	0.000	7	10	2	0.000
		6	0.424			3	0.001			3	0.001			3	0.001
5	8	2	0.002			4	0.009			4	0.007			4	0.006
		3	0.010			5	0.037			5	0.028			5	0.024
		4	0.054			6	0.108			6	0.087			6	0.080
		5	0.152			7	0.272			7	0.205			7	0.182
		6	0.347			8	0.460			8	0.383			8	0.355
5	9	2	0.001	6	6	2	0.002	6	13	2	0.000	7	11	2	0.000
		3	0.007			3	0.013			3	0.001			3	0.001
		4	0.039			4	0.067			4	0.005			4	0.004
		5	0.119			5	0.175			5	0.022			5	0.018
		6	0.287			6	0.392			6	0.070			6	0.060
5	10	2	0.001	6	7	2	0.001			7	0.176			7	0.145
		3	0.005			3	0.008			8	0.338			8	0.296
		4	0.029			4	0.043	6	14	2	0.000			9	0.484
		5	0.095			5	0.121			3	0.001	7	12	2	0.000
		6	0.239			6	0.296			4	0.004			3	0.000
		7	0.455			7	0.500			5	0.017			4	0.003
5	11	2	0.000	6	8	2	0.001			6	0.058			5	0.013
		3	0.004			3	0.005			7	0.151			6	0.046
		4	0.022			4	0.028			8	0.299			7	0.117
		5	0.077			5	0.086	7	7	2	0.001			8	0.247
		6	0.201			6	0.226			3	0.004			9	0.428
		7	0.407			7	0.413			4	0.025	7	13	2	0.000
5	12	2	0.000	6	9	2	0.000			5	0.078			3	0.000
		3	0.003			3	0.003			6	0.209			4	0.002
		4	0.017			4	0.019			7	0.383			5	0.010
		5	0.063			5	0.063	7	8	2	0.000			6	0.035
		6	0.170			6	0.175			3	0.002			7	0.095
		7	0.365			7	0.343			4	0.015			8	0.208
5	13	2	0.000	6	10	2	0.000			5	0.051			9	0.378
		3	0.002			3	0.002			6	0.149	8	8	2	0.000
		4	0.013			4	0.013			7	0.296			3	0.001
		5	0.053			5	0.047							4	0.009
		6	0.145			6	0.137							5	0.032
		7	0.330			7	0.287							6	0.100
						8	0.497							7	0.214
														8	0.405

Tabla C.6 (continuación)

Cola izquierda

n_1	n_2	r	p	n_1	n_2	r	p	n_1	n_2	r	p	n_1	n_2	r	p
8	9	2	0.000	9	9	2	0.000	10	10	2	0.000	11	11	2	0.000
		3	0.001			3	0.000			3	0.000			3	0.000
		4	0.005			4	0.003			4	0.001			4	0.000
		5	0.020			5	0.012			5	0.004			5	0.002
		6	0.069			6	0.044			6	0.019			6	0.007
		7	0.157			7	0.109			7	0.051			7	0.023
		8	0.319			8	0.238			8	0.128			8	0.063
		9	0.500			9	0.399			9	0.242			9	0.135
8	10	2	0.000	9	10	2	0.000			10	0.414			10	0.260
		3	0.000			3	0.000	10	11	2	0.000			11	0.410
		4	0.003			4	0.002			3	0.000	11	12	2	0.000
		5	0.013			5	0.008			4	0.001			3	0.000
		6	0.048			6	0.029			5	0.003			4	0.000
		7	0.117			7	0.077			6	0.012			5	0.001
		8	0.251			8	0.179			7	0.035			6	0.005
		9	0.419			9	0.319			8	0.092			7	0.015
8	11	2	0.000	9	11	2	0.000			9	0.185			8	0.044
		3	0.000			3	0.000			10	0.335			9	0.099
		4	0.002			4	0.001			11	0.500			10	0.202
		5	0.009			5	0.005	10	12	2	0.000			11	0.335
		6	0.034			6	0.020			3	0.000	12	12	2	0.000
		7	0.088			7	0.055			4	0.000			3	0.000
		8	0.199			8	0.135			5	0.002			4	0.000
		9	0.352			9	0.255			6	0.008			5	0.001
8	12	2	0.000			10	0.430			7	0.024			6	0.003
		3	0.000	9	12	2	0.000			8	0.067			7	0.009
		4	0.001			3	0.000			9	0.142			8	0.030
		5	0.006			4	0.001			10	0.271			9	0.070
		6	0.025			5	0.003			11	0.425			10	0.150
		7	0.067			6	0.014							11	0.263
		8	0.159			7	0.040							12	0.421
		9	0.297			8	0.103								
		10	0.480			9	0.205								
						10	0.362								

Tabla C.6 (continuación)

En cada entrada de la etiqueta p se tabula la probabilidad acumulada de cola derecha del valor de R, el número total de corridas en una sucesión de $n = n_1 + n_2$ símbolos de dos tipos para $n_1 \leq n_2$.

Cola derecha

n_1	n_2	r	p	n_1	n_2	r	p	n_1	n_2	r	p	n_1	n_2	r	p
2	2	4	0.333	4	8	9	0.071	5	11	11	0.058	6	12	12	0.075
2	3	5	0.100			8	0.212			10	0.154			11	0.217
		4	0.500			7	0.467			9	0.374			10	0.395
2	4	5	0.200	4	9	9	0.098	5	12	11	0.075	6	13	13	0.034
2	5	5	0.286			8	0.255			10	0.181			12	0.092
2	6	5	0.357	4	10	9	0.126	5	12	9	0.421			11	0.257
2	7	5	0.417			8	0.294	5	13	11	0.092			10	0.439
2	8	5	0.467	4	11	9	0.154			10	0.208	6	14	13	0.044
3	3	6	0.100			8	0.330			9	0.465			12	0.111
		5	0.300	4	12	9	0.181	5	14	11	0.111			11	0.295
3	4	7	0.029			8	0.363			10	0.234			10	0.480
		6	0.200	4	13	9	0.208	5	15	11	0.129	7	7	14	0.001
		5	0.457			8	0.393			10	0.258			13	0.004
3	5	7	0.071	4	14	9	0.234	6	6	12	0.002			12	0.025
		6	0.286			8	0.421			11	0.013			11	0.078
3	6	7	0.119	4	15	9	0.258			10	0.067			10	0.209
		6	0.357			8	0.446			9	0.175			9	0.383
3	7	7	0.167	4	16	9	0.282			8	0.392	7	8	15	0.000
		6	0.417			8	0.470	6	7	13	0.001			14	0.002
3	8	7	0.212	5	5	10	0.008			12	0.008			13	0.012
		6	0.467			9	0.040			11	0.034			12	0.051
3	9	7	0.255			8	0.167			10	0.121			11	0.133
3	10	7	0.294			7	0.357			9	0.267			10	0.296
3	11	7	0.330	5	6	11	0.002			8	0.500			9	0.486
3	12	7	0.363			10	0.024	6	8	13	0.002	7	9	15	0.001
3	13	7	0.393			9	0.089			12	0.016			14	0.006
3	14	7	0.421			8	0.262			11	0.063			13	0.025
3	15	7	0.446			7	0.478			10	0.179			12	0.084
3	16	7	0.470	5	7	11	0.008			9	0.354			11	0.194
3	17	7	0.491			10	0.045	6	9	13	0.006			10	0.378
4	4	8	0.029			9	0.146			12	0.028	7	10	15	0.002
		7	0.114			8	0.348			11	0.098			14	0.010
		6	0.371	5	8	11	0.016			10	0.238			13	0.043
4	5	9	0.008			10	0.071			9	0.434			12	0.121
		8	0.071			9	0.207	6	10	13	0.010			11	0.257
		7	0.214			8	0.424			12	0.042			10	0.451
		6	0.500	5	9	11	0.028			11	0.136	7	11	15	0.004
4	6	9	0.024			10	0.098			10	0.294			14	0.017
		8	0.119			9	0.266	6	11	13	0.017			13	0.064
		7	0.310			8	0.490			12	0.058			12	0.160
4	7	9	0.045	5	10	11				11	0.176			11	0.318
		8	0.167			10	0.126			10	0.346	7	12	15	0.007
		7	0.394			9	0.322	6	12	13	0.025			14	0.025

Tabla C.6 (continuación)

Cola derecha

n_1	n_2	r	p	n_1	n_2	r	p	n_1	n_2	r	p
7	12	13	0.089	9	9	18	0.000	10	13	13	0.320
		12	0.199			17	0.000			12	0.500
		11	0.376			16	0.003	10	12	21	0.000
7	13	15	0.010			15	0.012			20	0.000
		14	0.034			14	0.044			19	0.001
		13	0.116			13	0.109			18	0.006
		12	0.238			12	0.238			17	0.020
		11	0.430			11	0.399			16	0.056
8	8	16	0.000	9	10	19	0.000			15	0.125
		15	0.001			18	0.000			14	0.245
		14	0.009			17	0.001			13	0.395
		13	0.032			16	0.008	11	11	22	0.000
		12	0.100			15	0.026			21	0.000
		11	0.214			14	0.077			20	0.000
		10	0.405			13	0.166			19	0.002
8	9	17	0.000			12	0.319			18	0.007
		16	0.001			11	0.490			17	0.023
		15	0.004	9	11	19	0.000			16	0.063
		14	0.020			18	0.001			15	0.135
		13	0.061			17	0.003			14	0.260
		12	0.157			16	0.015			13	0.410
		11	0.298			15	0.045	11	12	23	0.000
		10	0.500			14	0.115			22	0.000
8	10	17	0.000			13	0.227			21	0.000
		16	0.002			12	0.395			20	0.001
		15	0.010	10	10	20	0.000			19	0.004
		14	0.036			19	0.000			18	0.015
		13	0.097			18	0.000			17	0.041
		12	0.218			17	0.001			16	0.099
		11	0.379			16	0.004			15	0.191
8	11	17	0.001			15	0.019			14	0.335
		16	0.004			14	0.051			13	0.493
		15	0.018			13	0.128	12	12	24	0.000
		14	0.057			12	0.242			23	0.000
		13	0.138			11	0.414			22	0.000
		12	0.278	10	11	21	0.000			21	0.001
		11	0.453			20	0.000			20	0.003
8	12	17	0.001			19	0.000			19	0.009
		16	0.007			18	0.003			18	0.030
		15	0.029			17	0.010			17	0.070
		14	0.080			16	0.035			16	0.150
		13	0.183			15	0.085			15	0.263
		12	0.337			14	0.185			14	0.421

Fuente: Gibbons J.D and Chakraborti S. (2003). *Nonparametric Statistical Inference*. Fourth Edition. Marcel Dekker, Inc. New York.

C.7. Cuantiles de la prueba estadística Kruskal-Wallis para muestras de tamaño pequeño

Tamaño de la muestra	$w_{0.90}$	$w_{0.95}$	$w_{0.99}$
2,2,2	3.7143	4.5714	4.5714
3,2,1	3.8571	4.2857	4.2857
3,2,2	4.4643	4.5000	5.3571
3,3,1	4.0000	4.5714	5.1429
3,3,2	4.2500	5.1389	6.2500
3,3,3	4.6000	5.0667	6.4889
4,2,1	4.0179	4.8214	4.8214
4,2,2	4.1667	5.1250	6.0000
4,3,1	3.8889	5.0000	5.8333
4,3,2	4.4444	5.4000	6.3000
4,3,3	4.7000	5.7273	6.7091
4,4,1	4.0667	4.8667	6.1667
4,4,2	4.4455	5.2364	6.8727
4,4,3	4.7730	5.5758	7.1364
4,4,4	4.5000	5.6538	7.5385
5,2,1	4.0500	4.4500	5.2500
5,2,2	4.2933	5.0400	6.1333
5,3,1	3.8400	4.8711	6.4000
5,3,2	4.4946	5.1055	6.8218
5,3,3	4.4121	5.5152	6.9818
5,4,1	3.9600	4.8600	6.8400
5,4,2	4.5182	5.2682	7.1182
5,4,3	4.5231	5.6308	7.3949
5,4,4	4.6187	5.6176	7.7440
5,5,1	4.0364	4.9091	6.8364
5,5,2	4.5077	5.2462	7.2692
5,5,3	4.5363	5.6264	7.5429
5,5,4	4.5200	5.6429	7.7914
5,5,5	4.5000	5.6600	7.9800

Se rechaza la hipótesis nula al nivel de significancia α si la estadística de Kruskal-Wallis dada en la ecuación (1.16), excede el cuantil $1 - \alpha$ dado en la tabla.

Fuente: Conover W.J. *Practical Nonparametric Statistics*, Third edition. USA. John Wiley & Sons, Inc. 1999.

C.8. Cuantiles de la estadística de prueba de Kolmogorov

Prueba de una cola											
	$p = 0.90$	0.95	0.975	0.99	0.995		$p = 0.90$	0.95	0.975	0.99	0.995
Prueba de dos colas											
	$p = 0.80$	0.90	0.95	0.98	0.99		$p = 0.80$	0.90	0.95	0.98	0.99
$n = 1$	0.900	0.950	0.975	0.990	0.995	$n = 21$	0.226	0.259	0.287	0.321	0.344
2	0.684	0.776	0.842	0.900	0.929	22	0.221	0.253	0.281	0.314	0.337
3	0.565	0.636	0.708	0.785	0.829	23	0.216	0.247	0.275	0.307	0.330
4	0.493	0.565	0.624	0.689	0.734	24	0.212	0.242	0.269	0.301	0.323
5	0.447	0.509	0.563	0.627	0.669	25	0.208	0.238	0.264	0.295	0.317
6	0.410	0.468	0.519	0.577	0.617	26	0.204	0.233	0.259	0.290	0.311
7	0.381	0.436	0.483	0.538	0.576	27	0.200	0.229	0.254	0.284	0.305
8	0.358	0.410	0.454	0.507	0.542	28	0.197	0.225	0.250	0.279	0.300
9	0.339	0.387	0.430	0.480	0.513	29	0.193	0.221	0.246	0.275	0.295
10	0.323	0.369	0.409	0.457	0.489	30	0.190	0.218	0.242	0.270	0.290
11	0.308	0.352	0.391	0.437	0.468	31	0.187	0.214	0.238	0.266	0.285
12	0.296	0.338	0.375	0.419	0.449	32	0.184	0.211	0.234	0.262	0.281
13	0.285	0.325	0.361	0.404	0.432	33	0.182	0.208	0.231	0.258	0.277
14	0.275	0.314	0.349	0.390	0.418	34	0.179	0.205	0.227	0.254	0.273
15	0.266	0.304	0.338	0.377	0.404	35	0.177	0.202	0.224	0.251	0.269
16	0.258	0.295	0.327	0.366	0.392	36	0.174	0.199	0.221	0.247	0.265
17	0.250	0.286	0.318	0.355	0.381	37	0.172	0.196	0.218	0.244	0.262
18	0.244	0.279	0.309	0.346	0.371	38	0.17	0.194	0.215	0.241	0.258
19	0.237	0.271	0.301	0.337	0.361	39	0.168	0.191	0.213	0.238	0.255
20	0.232	0.265	0.294	0.329	0.352	40	0.165	0.189	0.210	0.235	0.252
				Aproximación para $n > 40$			$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Esta tabla proporciona los cualtiles $w_n^{1-\alpha}$ de la estadística de prueba de Kolmogorov D, D^+ y D^-. Estos cuantiles son exactos para $n \leq 40$ en la prueba de dos colas. Se obtiene una mejor aproximación para $n > 40$ si el denominador $\sqrt{n}$ se sustituye por $(n + \sqrt{n/10})^{1/2}$.
Fuente: Conover W.J. *Practical Nonparametric Statistics*, Third edition. USA. John Wiley & Sons, Inc. 1999.

C.9. Cuantiles de la estadística de prueba de Lilliefors para normalidad

	$p = 0.80$	0.85	0.90	0.95	0.99
Tamaño de la muestra $n = 4$	0.303	0.320	0.344	0.374	0.414
5	0.290	0.302	0.319	0.344	0.398
6	0.268	0.280	0.295	0.321	0.371
7	0.252	0.264	0.280	0.304	0.353
8	0.239	0.251	0.266	0.290	0.333
9	0.227	0.239	0.253	0.275	0.319
10	0.217	0.228	0.241	0.262	0.303
11	0.209	0.219	0.232	0.252	0.291
12	0.201	0.210	0.223	0.243	0.281
13	0.193	0.203	0.215	0.233	0.270
14	0.187	0.196	0.209	0.227	0.264
15	0.181	0.190	0.202	0.219	0.256
16	0.176	0.184	0.195	0.212	0.248
17	0.170	0.179	0.190	0.207	0.241
18	0.166	0.174	0.185	0.201	0.234
19	0.162	0.171	0.181	0.197	0.230
20	0.159	0.167	0.177	0.192	0.223
21	0.155	0.163	0.173	0.188	0.219
22	0.152	0.160	0.170	0.185	0.214
23	0.149	0.156	0.165	0.181	0.210
24	0.145	0.153	0.162	0.177	0.205
25	0.144	0.151	0.159	0.173	0.202
26	0.141	0.147	0.156	0.170	0.198
27	0.138	0.145	0.153	0.166	0.193
28	0.136	0.142	0.151	0.165	0.191
29	0.134	0.140	0.149	0.162	0.188
30	0.132	0.138	0.146	0.159	0.183
≥ 31	$\frac{0.741}{d_n}$	$\frac{0.775}{d_n}$	$\frac{0.819}{d_n}$	$\frac{0.895}{d_n}$	$\frac{0.1035}{d_n}$
$d_n = \sqrt{n} - 0.01 + 0.83\sqrt{n}$					

Las entradas de esta tabla proporcionan el cualtil w_p de la estadística de prueba de Lilliefors D_1. Se rechaza H_0 al nivel de significancia α si $D_1 > w_{1-\alpha}$ para una muestra en particular de tamaño n.

Fuente: Conover W.J. *Practical Nonparametric Statistics*, Third edition. USA. John Wiley & Sons, Inc. 1999.

C.10. Cuantiles de la estadística de prueba de Lilliefors para la distribución exponencial

$p = 0.05$	0.10	0.20	0.30	0.50	0.70	0.80	0.90	0.95	0.99	0.999
$n = 2$	0.3127	0.3200	0.3337	0.3617	0.4337	0.5034	0.5507	0.5934	0.6133	0.6284	0.6317
3	0.2299	0.2544	0.2899	0.3166	0.3645	0.4122	0.4508	0.5111	0.5508	0.6003	0.6296
4	0.2072	0.2281	0.2545	0.2766	0.3163	0.3685	0.4007	0.4442	0.4844	0.5574	0.6215
5	0.1884	0.2052	0.2290	0.2483	0.2877	0.3317	0.3603	0.4045	0.4420	0.5127	0.5814
6	0.1726	0.1882	0.2102	0.2290	0.2645	0.3045	0.3320	0.3732	0.4085	0.4748	0.5497
7	0.1604	0.1750	0.1961	0.2136	0.2458	0.2838	0.3098	0.3481	0.3811	0.4459	0.5181
8	0.1506	0.1646	0.1845	0.2006	0.2309	0.2671	0.2914	0.3274	0.3590	0.4208	0.4913
9	0.1426	0.1561	0.1746	0.1897	0.2186	0.2529	0.2758	0.3101	0.3404	0.3995	0.4679
10	0.1359	0.1486	0.1661	0.1805	0.2082	0.2407	0.2626	0.2955	0.3244	0.3813	0.4473
12	0.1249	0.1364	0.1524	0.1657	0.1912	0.2209	0.2411	0.2714	0.2981	0.3511	0.4132
14	0.1162	0.1268	0.1418	0.1542	0.1778	0.2054	0.2242	0.2525	0.2774	0.3272	0.3858
16	0.1091	0.1191	0.1332	0.1448	0.1669	0.1929	0.2105	0.2371	0.2606	0.3076	0.3632
18	0.1032	0.1127	0.1260	0.1369	0.1578	0.1824	0.1990	0.2242	0.2465	0.2911	0.3441
20	0.0982	0.1073	0.1199	0.1303	0.1501	0.1735	0.1893	0.2132	0.2345	0.2771	0.3277
22	0.0939	0.1025	0.1146	0.1245	0.1434	0.1657	0.1809	0.2038	0.2241	0.2649	0.3135
24	0.0901	0.0984	0.1099	0.1195	0.1376	0.1590	0.1735	0.1954	0.2150	0.2542	0.3010
26	0.0868	0.0947	0.1058	0.1150	0.1324	0.1530	0.1670	0.1881	0.2069	0.2447	0.2899
28	0.0838	0.0914	0.1021	0.1110	0.1278	0.1477	0.1611	0.1815	0.1997	0.2362	0.2799
30	0.0811	0.0885	0.0988	0.1074	0.1236	0.1428	0.1559	0.1756	0.1932	0.2286	0.2709
35	0.0754	0.0822	0.0918	0.0997	0.1148	0.1326	0.1447	0.1630	0.1793	0.2123	0.2517
40	0.0707	0.0771	0.0861	0.0935	0.1077	0.1243	0.1356	0.1528	0.1681	0.1990	0.2361
45	0.0668	0.0729	0.0814	0.0884	0.1017	0.1174	0.1281	0.1443	0.1588	0.1880	0.2231
50	0.0636	0.0693	0.0774	0.0840	0.0966	0.1116	0.1217	0.1371	0.1509	0.1787	0.2121
60	0.0582	0.0635	0.0708	0.0769	0.0885	0.1021	0.1114	0.1255	0.1381	0.1635	0.1943
70	0.0541	0.0589	0.0658	0.0714	0.0821	0.0946	0.1033	0.1164	0.1281	0.1517	-
80	0.0507	0.0553	0.0616	0.0669	0.0769	0.0887	0.0968	0.1090	0.1200	0.1421	-
90	0.0479	0.0522	0.0582	0.0632	0.0726	0.0838	0.0914	0.1029	0.1132	0.1341	-
$n = 100$	0.0455	0.0496	0.0553	0.0600	0.0690	0.0796	0.0868	0.0977	0.1075	0.1274	-
Aproximación para $n > 100$	$\frac{0.4550}{\sqrt{n}}$	$\frac{0.4959}{\sqrt{n}}$	$\frac{0.5530}{\sqrt{n}}$	$\frac{0.6000}{\sqrt{n}}$	$\frac{0.6898}{\sqrt{n}}$	$\frac{0.7957}{\sqrt{n}}$	$\frac{0.8678}{\sqrt{n}}$	$\frac{0.9773}{\sqrt{n}}$	$\frac{1.0753}{\sqrt{n}}$	$\frac{1.2743}{\sqrt{n}}$	-

C.11. Distribución de la estadística de Shapiro-Wilk para normalidad

En cada entrada se tabulan los cuantiles w_α tales que $\mathbb{P}(w > w_\alpha) = \alpha$.

	α								
n	0.01	0.02	0.05	0.1	0.5	0.9	0.95	0.98	0.99
3	0.753	0.756	0.767	0.789	0.959	0.998	0.999	1.000	1.000
4	0.687	0.707	0.748	0.792	0.935	0.987	0.992	0.996	0.997
5	0.686	0.715	0.762	0.806	0.927	0.979	0.986	0.991	0.993
6	0.713	0.743	0.788	0.826	0.927	0.974	0.981	0.986	0.989
7	0.730	0.760	0.803	0.838	0.928	0.972	0.979	0.985	0.988
8	0.749	0.778	0.818	0.851	0.932	0.972	0.978	0.984	0.987
9	0.764	0.791	0.829	0.859	0.935	0.972	0.978	0.984	0.986
10	0.781	0.806	0.842	0.869	0.938	0.972	0.978	0.983	0.986
11	0.792	0.817	0.850	0.876	0.940	0.973	0.979	0.984	0.986
12	0.805	0.828	0.859	0.883	0.943	0.973	0.979	0.984	0.986
13	0.814	0.837	0.866	0.889	0.945	0.974	0.979	0.984	0.986
14	0.825	0.846	0.874	0.895	0.947	0.975	0.980	0.984	0.986
15	0.835	0.855	0.881	0.901	0.950	0.975	0.980	0.984	0.987
16	0.844	0.863	0.887	0.906	0.952	0.976	0.981	0.985	0.987
17	0.851	0.869	0.892	0.910	0.954	0.977	0.981	0.985	0.987
18	0.858	0.874	0.897	0.914	0.956	0.978	0.982	0.986	0.988
19	0.863	0.879	0.901	0.917	0.957	0.978	0.982	0.986	0.988
20	0.868	0.884	0.905	0.920	0.959	0.979	0.983	0.986	0.988
21	0.873	0.888	0.908	0.923	0.960	0.980	0.983	0.987	0.989
22	0.878	0.892	0.911	0.926	0.961	0.980	0.984	0.987	0.989
23	0.881	0.895	0.914	0.928	0.962	0.981	0.984	0.987	0.989
24	0.884	0.898	0.916	0.930	0.963	0.981	0.984	0.987	0.989
25	0.888	0.901	0.918	0.931	0.964	0.981	0.985	0.988	0.989
26	0.891	0.904	0.920	0.933	0.965	0.982	0.985	0.988	0.989
27	0.894	0.906	0.923	0.935	0.965	0.982	0.985	0.988	0.990
28	0.896	0.908	0.924	0.936	0.966	0.982	0.985	0.988	0.990
29	0.898	0.910	0.926	0.937	0.966	0.982	0.985	0.988	0.990
30	0.900	0.912	0.927	0.939	0.967	0.983	0.985	0.988	0.990
31	0.902	0.914	0.929	0.940	0.967	0.983	0.986	0.988	0.990
32	0.904	0.915	0.930	0.941	0.968	0.983	0.986	0.988	0.990
33	0.906	0.917	0.931	0.942	0.968	0.983	0.986	0.989	0.990
34	0.908	0.919	0.933	0.943	0.969	0.983	0.986	0.989	0.990
35	0.910	0.920	0.934	0.944	0.969	0.984	0.986	0.989	0.990
36	0.912	0.922	0.935	0.945	0.970	0.984	0.986	0.989	0.990
37	0.914	0.924	0.936	0.946	0.970	0.984	0.987	0.989	0.990
38	0.916	0.925	0.938	0.947	0.971	0.984	0.987	0.989	0.990
39	0.917	0.927	0.939	0.948	0.971	0.984	0.987	0.989	0.991
40	0.919	0.928	0.940	0.949	0.972	0.985	0.987	0.989	0.991
41	0.920	0.929	0.941	0.950	0.972	0.985	0.987	0.989	0.991
42	0.922	0.930	0.942	0.951	0.972	0.985	0.987	0.989	0.991
43	0.923	0.932	0.943	0.951	0.973	0.985	0.987	0.990	0.991
44	0.924	0.933	0.944	0.952	0.973	0.985	0.987	0.990	0.991
45	0.926	0.934	0.945	0.953	0.973	0.985	0.988	0.990	0.991
46	0.927	0.935	0.945	0.953	0.974	0.985	0.988	0.990	0.991
47	0.928	0.936	0.946	0.954	0.974	0.985	0.988	0.990	0.991
48	0.929	0.937	0.941	0.954	0.974	0.985	0.988	0.990	0.991
49	0.929	0.937	0.947	0.955	0.974	0.985	0.988	0.990	0.991
50	0.930	0.938	0.947	0.955	0.974	0.985	0.988	0.990	0.991

Fuente: Conover W.J. *Practical Nonparametric Statistics*, Third edition. USA. John Wiley & Sons, Inc. 1999.

C.12. Coeficientes de la estadística de Shapiro-Wilk

Se tabulan los valores de las constantes $a_{i,n}$ donde $j = 1,2,\ldots,[n/2]$ y $n = 2,3,\ldots$

	n									
j	1	2	3	4	5	6	7	8	9	10
1	-	0.7071	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739
2	-	-	0.0000	0.1677	0.2413	0.2806	0.3031	0.3164	0.3244	0.3291
3	-	-	-	-	0.0000	0.0875	0.1401	0.1743	0.1976	0.2141
4	-	-	-	-	-	-	0.0000	0.0561	0.0947	0.1224
5	-	-	-	-	-	-	-	-	0.0000	0.0399

	n									
j	11	12	13	14	15	16	17	18	19	20
1	0.5601	0.5475	0.5359	0.5251	0.5150	0.5056	0.4968	0.4886	0.4808	0.4732
2	0.3315	0.3325	0.3325	0.3318	0.3306	0.3090	0.3273	0.3253	0.3232	0.3211
3	0.2260	0.2347	0.2412	0.2495	0.2495	0.2521	0.2540	0.2553	0.2561	0.2565
4	0.1429	0.1586	0.1707	0.1802	0.1878	0.1988	0.1988	0.2027	0.2059	0.2085
5	0.0695	0.0922	0.1099	0.1240	0.1353	0.1447	0.1524	0.1587	0.1641	0.1686
6	0.0000	0.0303	0.0539	0.0727	0.0880	0.1005	0.1109	0.1197	0.1271	0.1334
7	-	-	0.0000	0.0240	0.0433	0.0593	0.0725	0.0837	0.0932	0.1013
8	-	-	-	-	0.0000	0.0196	0.0359	0.0496	0.0612	0.0711
9	-	-	-	-	-	-	0.0000	0.0163	0.0303	0.0422
10	-	-	-	-	-	-	-	-	0.0000	0.0140

	n									
j	21	22	23	24	25	26	27	28	29	30
1	0.4643	0.4590	0.4542	0.4493	0.4450	0.4407	0.4366	0.4328	0.4291	0.4254
2	0.3185	0.3156	0.3126	0.3098	0.3069	0.3043	0.3018	0.2992	0.2968	0.2944
3	0.2578	0.2571	0.2563	0.2554	0.2543	0.2533	0.2522	0.251	0.2499	0.2487
4	0.2119	0.2131	0.2139	0.2145	0.2148	0.2151	0.2152	0.2151	0.2150	0.2148
5	0.1736	0.1764	0.1787	0.1807	0.1822	0.1836	0.1848	0.1857	0.1864	0.187
6	0.1399	0.1443	0.1480	0.1512	0.1539	0.1563	0.1584	0.1601	0.1616	0.163
7	0.1092	0.1150	0.1201	0.1245	0.1283	0.1316	0.1346	0.1372	0.1395	0.1415
8	0.0804	0.0878	0.0941	0.0997	0.1046	0.1089	0.1128	0.1162	0.1192	0.1219
9	0.0530	0.0618	0.0696	0.0764	0.0823	0.0876	0.0923	0.0965	0.1002	0.1036
10	0.0263	0.0368	0.0459	0.0539	0.0610	0.0672	0.0728	0.0778	0.0822	0.0862
11	0.0000	0.0122	0.0228	0.0321	0.0403	0.0476	0.0540	0.0598	0.0650	0.0697
12	-	-	0.0000	0.0107	0.0200	0.0284	0.0358	0.0424	0.0483	0.0537
13	-	-	-	-	0.0000	0.0094	0.0178	0.0253	0.0320	0.0381
14	-	-	-	-	-	-	0.0000	0.0084	0.0159	0.0227
15	-	-	-	-	-	-	-	-	0.0000	0.0076

Tabla C.12 (continuación)

Se tabulan los valores de las constantes $a_{i,n}$ donde $j = 1,2,\ldots,[n/2]$ y $n = 2,3,\ldots$

	n									
j	31	32	33	34	35	36	37	38	39	40
1	0.4220	0.4188	0.4156	0.4127	0.4096	0.4068	0.4040	0.4015	0.3989	0.3964
2	0.2921	0.2898	0.2876	0.2854	0.2834	0.2813	0.2794	0.2774	0.2755	0.2737
3	0.2475	0.2463	0.2451	0.2439	0.2427	0.2415	0.2403	0.2391	0.2380	0.2368
4	0.2145	0.2141	0.2173	0.2132	0.2127	0.2121	0.2116	0.2110	0.2104	0.2098
5	0.1874	0.1878	0.1880	0.1882	0.1883	0.1883	0.1883	0.1881	0.1880	0.1878
6	0.1641	0.1651	0.1660	0.1667	0.1673	0.1678	0.1683	0.1686	0.1689	0.1691
7	0.1433	0.1449	0.1463	0.1475	0.1487	0.1496	0.1505	0.1513	0.1524	0.1526
8	0.1243	0.1265	0.1284	0.1301	0.1317	0.1331	0.1344	0.1356	0.1366	0.1376
9	0.1066	0.1093	0.1118	0.114	0.1160	0.1179	0.1196	0.1211	0.1225	0.1237
10	0.0899	0.0931	0.0961	0.0988	0.1013	0.1036	0.1056	0.1075	0.1092	0.1108
11	0.0739	0.0777	0.0812	0.0844	0.0873	0.0900	0.0924	0.0947	0.0967	0.0986
12	0.0585	0.0629	0.0669	0.0706	0.0739	0.0770	0.0798	0.0824	0.0848	0.0870
13	0.0435	0.0485	0.0530	0.0572	0.0610	0.0645	0.0677	0.0706	0.0733	0.0759
14	0.0289	0.0344	0.0395	0.0441	0.0484	0.0523	0.0559	0.0592	0.0622	0.0651
15	0.0144	0.0206	0.0262	0.0314	0.0361	0.0404	0.0444	0.0481	0.0515	0.0546
16	0.0000	0.0068	0.0187	0.0187	0.0239	0.0287	0.0331	0.0372	0.0409	0.0444
17	-	-	0.0000	0.0062	0.0119	0.0172	0.0220	0.0264	0.0305	0.0343
18	-	-	-	-	0.0000	0.0057	0.0110	0.0158	0.0203	0.0244
19	-	-	-	-	-	-	0.0000	0.0053	0.0101	0.0146
20	-	-	-	-	-	-	-	-	0.0000	0.0049

	n									
j	41	42	43	44	45	46	47	48	49	50
1	0.3940	0.3917	0.3894	0.3872	0.3850	0.3830	0.3808	0.3789	0.3770	0.3751
2	0.2719	0.2701	0.2684	0.2667	0.2651	0.2635	0.262	0.2604	0.2589	0.2574
3	0.2357	0.2345	0.2334	0.2323	0.2313	0.2302	0.2291	0.2281	0.2271	0.226
4	0.2091	0.2085	0.2078	0.2072	0.2065	0.2058	0.2052	0.2045	0.2038	0.2032
5	0.1876	0.1874	0.1871	0.1868	0.1865	0.1862	0.1859	0.1855	0.1851	0.1847
6	0.1693	0.1694	0.1695	0.1695	0.1695	0.1695	0.1695	0.1693	0.1692	0.1691
7	0.1531	0.1535	0.1539	0.1542	0.1545	0.1548	0.155	0.1551	0.1553	0.1554
8	0.1384	0.1392	0.1398	0.1405	0.1410	0.1415	0.142	0.1423	0.1427	0.143
9	0.1249	0.1259	0.1269	0.1278	0.1286	0.1293	0.13	0.1306	0.1312	0.1317
10	0.1123	0.1136	0.1149	0.1160	0.1170	0.118	0.1189	0.1197	0.1205	0.1212
11	0.1004	0.1020	0.1035	0.1049	0.1062	0.1073	0.1085	0.1095	0.1105	0.1113
12	0.0891	0.0909	0.0927	0.0943	0.0959	0.0972	0.0986	0.0998	0.1010	0.102
13	0.0782	0.0804	0.0824	0.0842	0.0860	0.0876	0.0892	0.0906	0.0919	0.0932
14	0.0677	0.0701	0.0724	0.0745	0.0765	0.0783	0.0801	0.0817	0.0832	0.0846
15	0.0575	0.0602	0.0628	0.0651	0.0673	0.0694	0.0713	0.0731	0.0748	0.0764
16	0.0476	0.0506	0.0534	0.0560	0.0584	0.0607	0.0628	0.0648	0.0667	0.0685
17	0.0379	0.0411	0.0442	0.0471	0.0497	0.0522	0.0546	0.0568	0.0588	0.0608
18	0.0283	0.0318	0.0352	0.0383	0.0412	0.0439	0.0465	0.0489	0.0511	0.0532
19	0.0188	0.0227	0.0263	0.0296	0.0328	0.0357	0.0385	0.0411	0.0436	0.0459
20	0.0094	0.0136	0.0175	0.0211	0.0245	0.0277	0.0307	0.0335	0.0361	0.0386
21	0.0000	0.0045	0.0087	0.0126	0.0163	0.0197	0.0229	0.0259	0.0288	0.0314
22	-	-	0.0000	0.0042	0.0081	0.0118	0.0153	0.0185	0.0215	0.0244
23	-	-	-	-	0.0000	0.0039	0.0076	0.0111	0.0143	0.0174
24	-	-	-	-	-	-	0.0000	0.0037	0.0071	0.0104
25	-	-	-	-	-	-	-	-	0.0000	0.0035

Fuente: Conover W.J. *Practical Nonparametric Statistics*, Third edition. USA. John Wiley & Sons, Inc. 1999.

Referencias

[1] Casella G. and Roger L. Berger. (2002), *Statistical Inference*. Second Edition. Duxbury, Thomson Learning.

[2] Chatterjee S. and Price, B. (1991), *Regression Analysis by Example*. Second Edition. Wiley, New York.

[3] Conover W. J. (1999), *Practical Nonparametric Statistics*. Third Edition. USA. John Wiley & Sons, Inc.

[4] Crawshaw J. and Chambers J. (2002), *A Concise Course in Advanced Level Statistics.* Fourth Edition. United Kingdom. Nelson Thornes.

[5] Daniel W. (1990), *Applied Nonparametric Statistics*. Second Edition. USA. PWS Kent.

[6] Dalgaar P. (2008), *Introductory Statistics with R*. Second Edition. New York, USA. Springer.

[7] Draper N. and Smith, H. (1981), *Applied Regression Analysis*. Second Edition, Wiley, New York.

[8] Everitt B.S. (1977), *The Analysis of Contingency Tables*. Chapman and Hall, London.

[9] Faraway J.J. (2009), *Linear Models with R*. Taylor and Francis Group, 2009.

[10] Friedman, M. (1937), The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32, 675-701.

[11] Gibbons J.D. and Chakraborti, S. (2003), *Nonparametric Statistical Inference*. Fourth Edition. Marcel Dekker, Inc. New York.

[12] Iman, R.L. and Devenport, J.M. (1976), *New approximations to the exact distribution of the Kruskal-Wallis test statistic*. Communications in Statistics - Theory and Methods, A5, 1335-1348.

[13] Kolmogorov, A.N. (1933), *Sulla determinazione empirica di una legge di distribuzione*. Giornale dell' Istituto Italiano degli Attuari, 4, 83-91.

[14] Kruskal, W.H. (1952), *A nonparametric test for the several sample problem*. The Annals of Mathematical Statistics, Vol. 23, pp. 525-540.

[15] Kruskal, W.H., and Wallis, W.A. (1952), *Use of ranks on one-criterion variance analysis*. Journal of the American Statistical Association, 47,583-621.

[16] Mood A. M, Graybill F. A and Boes D. C. (1974), *Introduction to the Theory of Statistics*. Third Edition. McGraw Hill.

[17] Montgomery D.C, Peck, E.A. and Vining G.G. (2001), *Introduction to Linear Regression Analysis*. Third Edition. John Wiley and Sons, Inc. New York.

[18] Neave H.R. and Worthington. (1988), *Distibution-Free Tests*. Unwin Hyman, London.

[19] Rawlings J.O. (1988), *Applied Regression Analysis, a Research Tool*, Wadswortg & Brooks, USA.

[20] Rincón L. (2007), *Curso Intermedio de Probabilidad*. Primera edición. Las Prensas de Ciencias, México.

[21] Shapiro, S.S, and Wilk, M.B. (1965). *An analysis of variance test for normaly (complete samples)*. Biometrika, 52, 591-611 (6.2).

[22] Shapiro, S.S, and Wilk, M.B. (1968). *Approximations for the null distribution of the W statistic*. Technometrics, 10, 861-866 (6.2).

[23] Smirnov, N.V. (1939), *Estimate of deviation between empirical distribution functions in two independent sample*. (Russian) Bulletin Moscow University, 2 (2), 3-16.

[24] Smirnov, N.V. (1948), *Table for estimating goodness of fit of empirical distributions*. The Annals of Mathematical Statistics, 19, 279-281.

[25] Verzani, John. (2005), *Using R for Introductory Statistics*. Chapman & Hall/CRC Press.

Índice alfabético