Comprendiendo la Estadística Inferencial

Giovanni Sanabria Brenes

Contenidos

Presentación

Dado un conjunto de datos, la Estadística Descriptiva pretende por un lado obtener representaciones descriptivas de los datos como valores numéricos que indica su distribución, cuadros y gráficos. Por otro lado, la Estadística Inferencial busca resultados sobre un conjunto de datos mayor a partir de los datos dados, como por ejemplo determinar el porcentaje de todos los ciudadanos que apoyan un candidato a partir de los resultados de una encuesta.

El presente material brinda un estudio general de la Estadística Inferencial y es fruto de la experiencia adquirida durante varios semestres impartiendo cursos y del desarrollo de varias propuestas sobre la enseñanza del tema. Así, se brindan ejemplos que facilitan la comprensión de ciertos conceptos, por ejemplo la confianza de un intervalo, el nivel de significancia de una prueba y la ecuación de una regresión lineal.

El libro "Comprendiendo la Estadística Inferencial" realiza una combinación adecuada entre la rigurosidad matemática y la intuición, y está dirigido a estudiantes de las carreras de Ingeniería en Computación y de Enseñanza de la Matemática. Sin embargo, ignorando algunos desarrollos teóricos, ciertas partes del texto pueden ser utilizadas por estudiantes de otras carreras universitarias.

Este libro cuenta con una serie de ejercicios en su mayoría con una breve respuesta que el lector debe justificar, además cada apartado cuento con uno o varios ejercicios de repaso.

Para el estudio de los tópicos presente en el libro se requiere que el lector tenga conocimientos básicos en probabilidad, para el repaso de estos conocimientos previos se recomienda el libro "Comprendiendo las Probabilidades" el cuál es compatible en su notación con el presente.

Capítulo I

Elementos de Estadística Inferencial

Se abordan los conceptos básicos de Estadística Inferencial de manera formal e intutiva.Se estudia la aplicación del teorema del Límite Central en esta rama de la estadística. Finalmente se definen las distribuciones muestrales necesarias para los siguientes capítulos.

1 Introducción

Dado un conjunto de datos, la estadística descriptiva pretende por un lado obtener representaciones descriptivas de los datos como valores numéricos que indica su distribución, cuadros y gráficos. Por otro lado, la estadística inferencial busca resultados sobre un conjunto de datos mayor a partir de los datos dados, como por ejemplo determinar el porcentaje de todos los ciudadanos que apoyan un candidato a partir de los resultados de una encuesta.

En el presente documento se centra en el estudio de la estadística inferencial que se desarrollará a continuación.

2 Conceptos básicos

La **estadística inferencial** tiene como objetivo generalizar los resultados de un subconjunto de datos a todo el conjunto. Seguidamente se define los conceptos básicos de esta rama de la matemática:

Definición 1 ***Población****: conjunto de datos que se desea estudiar. Estos datos deben verse como valores de una misma variable, la cual se utiliza para designar la población.*

Definición 2 ***Muestra****: subconjunto de datos que se seleccionan de la población.*

Así, la estadística inferencial busca generalizar los resultados obtenidos en una muestra a toda la población. Si la muestra es igual a la población, la generalización o estudio se le llama **censo** y es exacta. En caso contrario, se debe buscar que la muestra sea lo más representativa de la población, para que la generalización sea confiable y se disminuya el sesgo.

Ejemplo 1 *Se quiere realizar un estudio para determinar el porcentaje de ciudadanos de un país C que están a favor de un Tratado de Libre Comercio (TLC) con un país E. La población sería el conjunto de opiniones (a favor o en contra) de todos los ciudanos del país C sobre el TLC. Si se define X como la opinión de un ciudadano sobre el TLC, se dice que la población esta dada por X. Como muestra se tomo el conjunto de ciudadanos empresarios del país C que exportan al país E. ¿Es la muestra representativa de la población?.*

Para que la muestra sea representativa de la población se introduce el concepto de muestra aleatoria.

Definición 3 ***(Muestra Aleatoria)*** *Considere la población dada por la variable aleatoria X con función de distribución f_X. Una muestra aleatoria de tamaño n esta formada por n de estas variables $(X_1, X_2, X_3, ..., X_n)$. Estas variables cumplen:*

1. *Todas sigue la misma distribución.*
2. *Son mutuamente independientes.*

Una vez que se tiene certeza de que la muestra es aleatoria, el paso siguiente es definir la característica de la población que se desea estudiar (parámetro) y la herramienta a aplicar a la muestra (estadístico) para realizar el estudio.

Definición 4 ***Parámetro****: valor numérico que se le asigna a la población.*

El parámetro es valor numérico que se desconoce debido a varios factores, entre ellos: el tamaño de la población y la accesibilidad a los datos que componen la población. La estadística inferencial busca acotar este parámetro por medio de un análisis de la muestra. Seguidamente se define los parámetros más comunes.

Definición 5 *Considere la población $\{x_1, x_2, x_3, ..., x_N\}$ dada por la variable aleatoria X. Se define los siguientes parámetros:*

1. *media poblacional:*
$$\mu = \frac{\sum X}{N} = \frac{\sum_{i=1}^{N} x_i}{N}.$$
2. *Varianza poblacional:*
$$\sigma^2 = \frac{\sum (X-\mu)^2}{N} = \frac{\sum_{i=1}^{N} (x_i-\mu)^2}{N}.$$
3. *Desviación estándar poblacional:*
$$\sigma = \sqrt{\sigma^2}.$$
4. *proporción poblacional o porcentaje de éxitos:*
$$P = \frac{b}{N}$$
donde b es el número de datos éxitosos de la población (datos que cumplen una determinada condición).

Ejemplo 2 *Se debe estudiar el gasto mensual en consumo de energía eléctrica de las familias un pequeño residencial de un cantón cercano a San José en el mes de Octubre. El residencial esta forma por 40 familias y seguidamente se presentan los gatos en miles de colones de cada familia en este mes, redondeadas al décimo inferior*

10,5	11,0	10,1	8,9	7,3
8,5	10,6	10,0	8,7	7,0
6,2	10,5	10,0	18,5	10,2
7,5	19,2	9,8	8,0	6,3
14,2	10,5	9,5	10,3	15,8
5,2	9,4	11,1	7,8	6,0
12,2	6,9	7,9	15,5	5,4
9,5	10,2	9,4	7,4	12,7

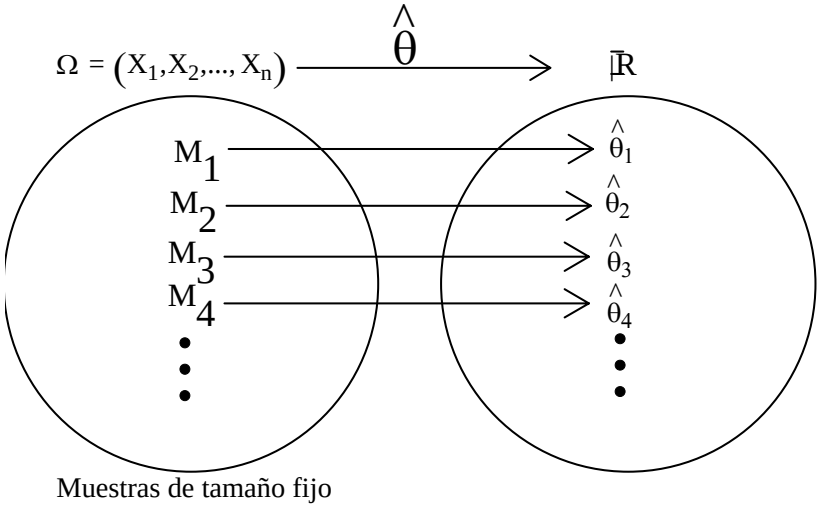
Aqui la población sería el consumo de energía eléctrica de cada familia del residencial en el mes de Octubre. Note que la población se conoce totalmente.

1. *Determine la media poblacional.* $R/ \quad \mu = 9,8925$

2. *Halle la Desviación Estándar y la Varianza poblacional.* $R/\quad \sigma^2 = 10,01219375, \quad \sigma = 3,164205074$

La **Estadística Descriptiva** se dedica a calcular e interpretar los parámetros sobre una población de datos, estos junto con tablas, gráficos y diagramas le permiten dar una descripción de la población de datos. Cuando no es factible la determinación de estos parámetros, la Estadística Inferencial permite hallar aproximaciones a los parámetros utilizando muestras. ¿Qué herramienta utiliza para lograr la aproximación?

Definición 6 ***Estadístico o estimador****: variable aleatoria que asigna un valor (llamado estimación) a cada muestra de tamaño fijo. A cada parámetro* θ *se le asigna un estadístico* $\widehat{\theta}$ *:*



La distribución de probabilidad de un estadístico es llamada ***distribución muestral.***

Seguidamente se presentan los estadísticos asociados a los parámetros más comunes.

Definición 7 *Considere la población dada por la variable aleatoria* X *y una muestra aleatoria de esta población* $M = (X_1, X_2, X_3, ..., X_n)$ *. Se define los siguientes estadísticos:*

1. *El estadístico asociado a la media poblacional* μ *se denomina media muestral* $\overline{X}$*:*

$$\overline{X}(M) = \frac{\sum_{i=1}^{n} X_i}{n}.$$

2. *El estadístico asociado a la varianza poblacional* σ^2 *se denomina varianza muestral* S^2*:*

$$S^2(M) = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}.$$

3. *El estadístico asociado a la desviación estándar poblacional σs e denomina desviación estándar muestral S:*

$$S(M) = \sqrt{S^2(M)}.$$

4. *El estadístico asociado a la proporción poblacional P se denomina proporción muestral $\widehat{P}$:*

$$\widehat{P}(M) = \frac{B(M)}{n}$$

donde $B(M)$ es el número de datos de la muestra M que cumplen una determinada condición.

Ejemplo 3 *Se desea determine la edad promedio de los estudiantes de la Universidad Bienestar Seguro en el II semestre del* 2007. *En este caso la población es el conjunto formado por todas las edades de los estudiantes de la Universidad Bienestar Seguro en el II semestre del* 2007, *el parámetro a estudiar es*

$$\mu: \text{ edad promedio de los estudiantes ...}$$

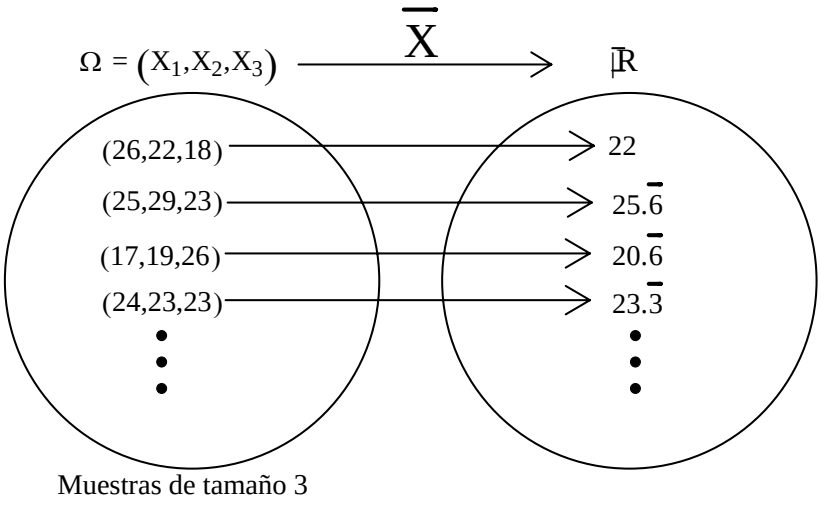
Como herramienta para aproximarse a μ se utilizará el estadístico

$$\overline{X}: \text{ edad promedio en muestras de tamaño } 3.$$

Si Jorge, Karla y Anthony son tres estudiantes de la universidad en el II semestre, elegidos al azar, y tiene edades de $26, 22$ *y* 18 *años respectivamente entonces*

$$\overline{X}(\{26, 22, 18\}) = \frac{26 + 22 + 18}{3} = 22$$

Si se continua tomando muestras, se obtiene que:



Muestras de tamaño 3

Si se sigue tomando muestras de tamaño 3, unas 300, se puede realizar una distribución de frecuencias de los datos obtenidos y determinar empíricamente la distribución de probabilidad de $\overline{X}$. Sin embargo, ¿Es útil el $\overline{X}$ escogido para inferir el valor de μ?

Más adelante, se estudiará que el tamaño de las muestras influye sobre la distribución muestra de $\overline{X}$.

¿Cómo se utiliza el estadístico? Por medio del estadístico $\widehat{\theta}$ se puede hacer inferencias sobre el parámetro θ de la población, las inferencias puede ser:

1. **Estimación.** Consiste en un utilizar $\widehat{\theta}$ para hallar una aproximación a θ (estimación puntual) o un intervalo que contenga a θ con cierto grado de certeza (intervalos de confianza).

2. **Pruebas de hipótesis**. Se utilizar $\widehat{\theta}$ para determinar si una afirmación sobre el valor de θ es muy probable.

¿Cuando un estadístico es un buen estimador? Para ello, se necesita que el estadístico tenga como valor esperado el parámetro respectivo, este concepto se define a continuación.

Definición 8 *El estadístico $\widehat{\theta}$ asociado al parámetro θ es insesgado si y solo si:*

$$E\left(\widehat{\theta}\right) = \theta$$

Los teoremas siguientes indican que los estadísticos estudiados son insesgados.

Teorema 1 *Considere la población dada por la variable aleatoria X con media poblacional μ y varianza poblacional σ^2. Dada una muestra aleatoria de esta población $M = (X_1, X_2, X_3, ..., X_n)$, se tiene que*

1. *Se cumple que*
$$E(X_i) = \mu, \quad Var(X_i) = \sigma^2 \quad y \quad E\left(X_i^2\right) = \sigma^2 + \mu^2.$$

2. *La media muestral es insesgada:*
$$E\left(\overline{X}\right) = \mu.$$

3. $Var\left(\overline{X}\right) = \dfrac{\sigma^2}{n}$.

4. $E\left(\overline{X}^2\right) = \dfrac{\sigma^2}{n} + \mu^2$

5. $S^2 = \dfrac{1}{n-1}\left(\sum\limits_{i=1}^{n} X_i^2 - n\overline{X}^2\right)$

6. *La varianza muestral es insesgada:*
$$E\left(S^2\right) = \sigma^2$$

Prueba. *Se probará la 1, 5 y* 6, *las demás quedan de ejercicio.*

1. *Por la definición de muestra aleatoria se tiene que cada X_i es una copia de la variable X, entonces*

$$E(X_i) = \mu \quad y \quad Var(X_i) = \sigma^2.$$

Además, por teorema $Var(X_i) = E\left(X_i^2\right) - [E(X_i)]^2$, por lo tanto

$$E\left(X_i^2\right) = Var(X_i) + [E(X_i)]^2 = \sigma^2 + \mu^2$$

5. *Note que*

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i^2 - 2X_i\overline{X} + \overline{X}^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2 - 2\overline{X}\sum_{i=1}^{n}X_i + \sum_{i=1}^{n}\overline{X}^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2 - 2\overline{X}n\overline{X} + n\overline{X}^2\right) \qquad \textit{por definición: } \sum_{i=1}^{n}X_i = n\overline{X} \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2 - n\overline{X}^2\right) \end{aligned}$$

6. *Se tiene que*

$$\begin{aligned} E\left(S^2\right) &= \frac{1}{n-1}E\left(\sum_{i=1}^{n}X_i^2 - n\overline{X}^2\right) && \textit{por 5. y propiedad de la esperanza} \\ &= \frac{1}{n-1}\left[E\left(\sum_{i=1}^{n}X_i^2\right) - nE\left(\overline{X}^2\right)\right] && \textit{por propiedad de la esperanza} \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n}E\left(X_i^2\right) - nE\left(\overline{X}^2\right)\right] && \textit{por propiedad de la esperanza} \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(\sigma^2+\mu^2\right) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right] && \textit{por 1. y 4.} \\ &= \frac{1}{n-1}\left[n\left(\sigma^2+\mu^2\right) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right] \\ &= \sigma^2 \end{aligned}$$

Teorema 2 *El estadístico $\widehat{P}$ es insesgado.*

Prueba. *Recuerde que $\widehat{P} = \dfrac{B}{n}$ es un estimador para el porcentaje p de éxitos en una población, donde B es el número de éxitos en muestras aleatorias de tamaño n, por lo tanto*

$$B \sim B\left(n, p\right)$$

y entonces

$$E\left(\widehat{P}\right) = E\left(\frac{B}{n}\right) = \frac{1}{n}E\left(B\right) = \frac{1}{n}\cdot np = p \qquad \blacksquare$$

Por otro lado, entre dos estadísticos insesgados asociados al mismo parámetro, es mejor estimador aquel que tenga una menor varianza. Esto pues entre menor sea la varianza, hay más probabilidad de que el estadístico se concentre en la media.

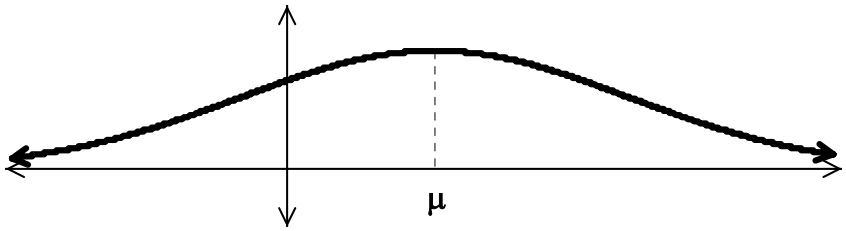
Seguidamente se estudiará la distribución muestral de $\overline{X}$.

3 Distribución muestral de $\overline{X}$ y el Teorema del Límite Central

3.1 Distribución normal

Recuerde que si $X \sim N\left(\mu, \sigma^2\right)$ entonces $Z = \dfrac{X-\mu}{\sigma} \sim N\left(0,1\right)$, es decir Z sigue una distribución normal estándar. La función acumulada de la distribución normal estándar se denota por

$$F_Z\left(k\right) = P\left(Z \leq k\right) = \phi\left(k\right)$$



Para determinar los valores de $\phi\left(k\right)$ se utilizan una tabla:

k	centésimas de k
	$\vdots$
Unidades y décimas de k	$\cdots \quad \cdots \quad \phi\left(k\right)$

Dado que la distribución normal estándar es simétrica respecto a cero, se contará únicamente con los valores Z negativos.

Ejemplo 4 *Sea* $Z \sim N\left(0,1\right)$ *Utilizando la tabla*

$$P\left(Z < -1.52\right) = 0.0643$$

y por simetría se tiene que

$$P\left(Z > 1.52\right) = P\left(Z < -1.52\right) = 0.0643$$

Además

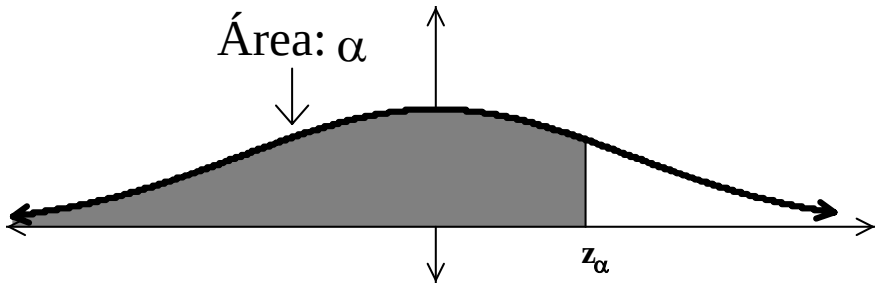
$$P\left(Z > -1.52\right) = 1 - P\left(Z < -1.52\right) = 1 - 0.0643 = 0.9357$$

y

$$P\left(Z < 1.52\right) = 1 - P\left(Z > 1.52\right) = 1 - 0.0643 = 0.9357$$

Definición 9 *Si* $Z \sim N\left(0,1\right)$ *entonces se define* ***el valor z:***

$$z_\alpha : \textit{valor de } Z \textit{ que acumula una área a la izquierda de } \alpha\%.$$



Es decir $P(Z \leq z_\alpha) = \alpha$.

Note que si α es menor a 0.5 entonces z_α es negativo y viceversa. De igual forma si α es mayor a 0.5 entonces z_α es positivo y viceversa.

Teorema 3 *Se tiene que* $z_\alpha = -z_{1-\alpha}$

Ejemplo 5 *Utilizando directamente la tabla se tiene que*

$$z_{0.1} \approx -1.28, \quad z_{0.05} \approx -1.645, \quad z_{0.025} \approx -1.96$$

y por simétria se tiene que

$$z_{0.9} = -z_{0.1} \approx 1.28, \quad z_{0.95} = -z_{0.05} \approx 1.645, \quad z_{0.975} = -z_{0.025} \approx 1.96$$

3.2 Suma y promedio de normales

Dado que por definición una muestra aleatoria $(X_1, X_2, ..., X_n)$ esta formada por variables aleatorias que sigue una misma distribución y son mutuamente independientes, son aplicables los teoremas que nos hablan sobre la distribución de $\overline{X}$ y S_n.

Teorema 4 ***(La suma y promedio de normales es normal para muestras aleatorias)*** *Considere la población dada por la variable aleatoria* X *que sigue una distribución normal con media poblacional* μ *y varianza poblacional* σ^2. *Dada una muestra aleatoria de esta población* $(X_1, X_2, X_3, ..., X_n)$ *se tiene:*

1. *La media muestral* $\overline{X}$ *sigue una distribución normal:*

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

2. *La suma muestral* $S_n = X_1 + X_2 + X_3 + ... + X_n$ *sigue una distribución normal:*

$$S_n \sim N\left(n\mu, n\sigma^2\right)$$

Ejemplo 6 *Un técnico de la empresa ECI señala que las llamadas telefónicas en cierta localidad tienen una distribución normal con media de 180 segundos y una desviación estándar de* 300 *segundos.*

1. *En diez llamadas, ¿Cuál es la probabilidad de que su duración promedio sea menor a* 160 *segundos?*

   Considere la variable

   $$X: \text{ duración de una llamada de la localidad}$$

   Se tiene que $X \sim N\left(180, 300^2\right)$. *Por el Teorema (Promedio de normales es normal) se tiene que* $\overline{X}$, *la duración promedio de diez llamadas, cumple que*

   $$\overline{X} \sim N\left(180, \frac{300^2}{10}\right)$$

   Por lo tanto, la probabilidad de que la duración promedio de las diez llamadas sea menor a 160 *segundos es*

   $$P\left(\overline{X} < 160\right) = P\left(Z < \frac{160-180}{300/\sqrt{10}}\right) = \phi\left(\frac{160-180}{300/\sqrt{10}}\right) = \phi\left(-0.210\,819\right) = 0.4168$$

2. *En diez llamadas, ¿Cuál es la probabilidad de que su duración total sea mayor a* 31 *minutos?*

   Como $X \sim N\left(180, 300^2\right)$, *por el Teorema (Suma de normales es normal) se tiene que* S_{10}, *la duración total de diez llamadas, cumple que*

   $$S_{10} \sim N\left(1800, 300^2 \cdot 10\right)$$

   Por lo tanto, la probabilidad de que la duración total de las diez llamadas sea mayor a 31 *minutos* (1860 *segundos*) *ea*

   $$\begin{aligned} P\left(S_{10} > 1860\right) &= 1 - \phi\left(\frac{1860-1800}{300\sqrt{10}}\right) \\ &= 1 - \phi\left(0.06324\,56\right) = 1 - 0.5239 = 0.4761 \end{aligned}$$

Ejemplo 7 *El tiempo que tarda una persona resolviendo un examen típico de admisión a la Universidad Bienestar Seguro tiene una distribución Normal con media 3 horas. Se sabe que una muestra de 20 estudiantes, la probabilidad de que tarden en promedio más* 3 *horas* 15 *minutos en resolver el examen es del* 5%. *Determine la desviación estándar de la duración del examen típico.*

Considere la variable

$$X: \text{duración del examen típico en horas.}$$

Note que $X \sim N\left(3, \sigma^2\right)$. *Se toma una muestra aleatoria de duraciones de* 20 *estudiantes en el examen:* $(X_1, X_2, ..., X_{20})$. *Por el Teorema Promedio de Normales es Normal, se tiene que*

$$\overline{X} \underset{aprox}{\sim} N\left(3, \frac{\sigma^2}{20}\right)$$

Así, la probabilidad de que los estudiantes muestreados tarden en promedio más 3.25 *horas en resolver el examen es*

$$0.05 = P\left(\overline{X} > 3.25\right) = 1 - \phi\left(\frac{3.25-3}{\sigma/\sqrt{20}}\right) \Longrightarrow \phi\left(\frac{3.25-3}{\sigma/\sqrt{20}}\right) = 0.95$$

Utilizando la tabla de la normal estandar, se debe cumplir que

$$\frac{3.25-3}{\sigma/\sqrt{20}} = 1.645 \quad \Longrightarrow \quad \sigma = 0.679\,656.$$

Por lo tanto, la desviación estándar de la duración del examen típico es de aproximadamente 0.68 *horas.*

Ejercicio 1 *El tiempo que tarda una persona en llenar un solicitud de empleo en una pagina web sigue una distribución normal con media* 15 *minutos y desviación estándar de 3 minutos. Si 10 personas llenan la solicitud, ¿Cuál es la probabilidad de que en promedio duren menos de* 17 *minutos?*
R/ 0.9826

3.3 El teorema del límite central

¿Qué pasa cuando la población no sigue una distribución Normal? Veamos el siguiente ejemplo.

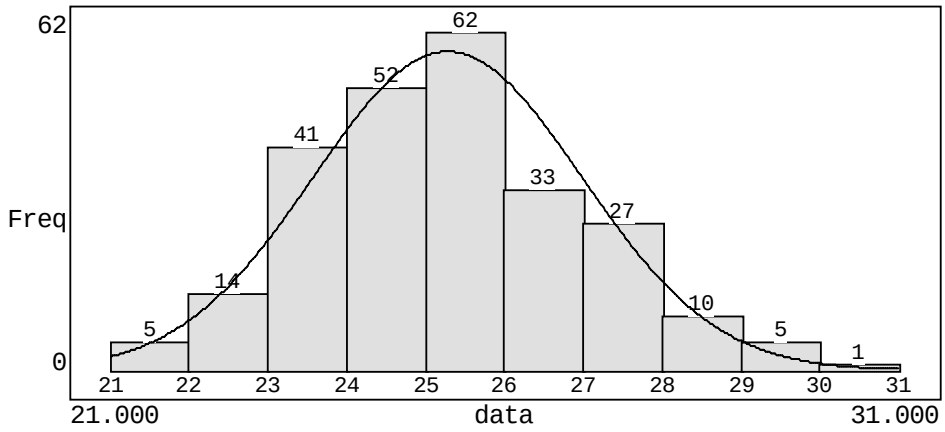
Ejemplo 8 *En el 2004 se realizó una encuesta de opinión para evaluar el servicio que le ofrece la soda comedor del ITCR. Se entrevistaron 147 personas y entre las variables del estudio, se encontraba la edad del cliente de la soda. Para esta variable se contaron en 133 valores (algunos de los entrevistados no dieron su edad) Se quiere estudiar el parámetro*

μ : *edad promedio de las personas que utilizan el servicio comedor del ITCR en el 2004.*

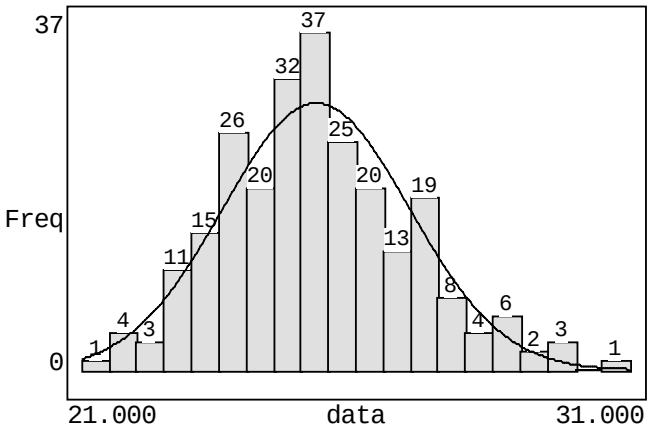
Para ello considere el estadísco

$\overline{X}$: *promedio muestral sobre muestras de tamaño 30.*

A partir de los datos recogidos, se tomaron 250 muestras aleatorias de tamaño 30 (¿qué limitaciones existen?) y se determino el valor de $\overline{X}$ para cada una de las muestras. A partir de los datos obtenidos de $\overline{X}$ se realizó una distribución frecuencia con 10 clases y se obtuvo el siguiente histograma:



Observe, por ejemplo, que de los 250 valores de $\overline{X}$ hay 52 entre 24 y 25. Si se realiza una distribución de frecuencias con 20 clases, el histograma será:



Note que la distribución de $\overline{X}$ se aproxima a una distribución normal, y tomando más muestras de tamaño 30 y realizan el histograma se puede apreciar mejor esta distribución.

Teorema 5 ***(Teorema del Límite Central para muestras aleatorias)*** *Considere la población dada por la variable aleatoria X que con media poblacional μ y varianza poblacional σ^2. Dada una muestra aleatoria de esta población $(X_1, X_2, X_3, ..., X_n)$ Para n suficientemente grande ("$n \to \infty$") se tiene que*

1. *La media muestral $\overline{X}$ sigue una distribución normal:*

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

2. *La suma muestral $S_n = X_1 + X_2 + X_3 + ... + X_n$ sigue una distribución normal:*

$$S_n \sim N\left(n\mu, n\sigma^2\right)$$

Similarmente, en la práctica, se considere que para muestras de tamaño mayor igual a 30, la distribución normal es una buen aproximación a la distribución de $\overline{X}$ y S_n :

$$\overline{X} \underset{aprox}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right) \quad y \quad S_n \underset{aprox}{\sim} N\left(n\mu, n\sigma^2\right).$$

Ejemplo 9 *El doctor Adrían tiene un consultorio privado y ha estimado que en promedio cobra 15000 colones por consulta con una desviación estándar de 7000 colones. Hace poco un cliente se quejo de que el cobro de 18000 colones por su consulta fue excesivo. Después de mucho discutir, el doctor establece que si en una muestra de 50 facturas por consulta se revela un ingreso promedio de consulta*

inferior al monto cobrado al cliente, le devolverá todo el dinero. ¿Considera que el arreglo del doctor, lo favorece?

Considere la variable

$$X: \text{ ingreso en colones por consulta}$$

Se toma una muestra del ingreso de 40 *consultas:* $(X_1, X_2, ..., X_{40})$ *Como el tamaño de la muestra es de* $n = 40$ *por el Teorema del Límite Central se tiene que* $\overline{X}$*, ingreso promedio muestral, cumple que*

$$\overline{X} \underset{aprox}{\sim} N\left(15000, 7000^2/50\right)$$

Por lo tanto, la probabilidad de que le devuelva al cliente el dinero es

$$P\left(\overline{X} < 18000\right) = \phi\left(\frac{18000 - 15000}{7000/\sqrt{50}}\right) = \phi\left(3.030\,46\right) = 0.9988$$

Este arreglo perjudica al doctor.

Ejercicio 2 *Una clase de melocotones de cierto huerto tienen un diámetro medio de*10 *cm con una desviación estándar igual a 0.7 cm. En una muestra de* 40 *melocotones, determine la probabilidad de que el diámetro promedio de los cuarenta melocotones sea superior a* $9,8$ *cm.* *R/* 0.9648

3.4 Generación de valores de una variable con distribución normal

Debido a la importancia que tiene la distribución normal en la Estadística, en veces es utíl generar algunos valores de una determinada variable normal para, por ejemplo, ejemplificar un resultado. Existe varias maneras de simular los valores de una variables, las cuales se basan en obtener valores de una uniforme continua entre 0 y 1, pues son equivalentes a los datos por la función random de una calculadora o computadora. Seguidamente se explicará el método de simulación basado en el Teorema del Límite Central (TLC).

Sean $X_1, X_2, X_3, ..., X_n$ variables aleatorias mutuamente independientes que siguen una distribución uniforme en $[0, 1]$. Entonces

$$E\left(X_i\right) = \frac{1}{2} \quad y \quad Var\left(X_i\right) = \frac{1}{12}, \quad \text{para } i = 1, 2, 3, ..., n$$

Si $n \geq 30$ por el TLC se tiene que

$$S_n = X_1 + X_2 + X_3 + ... + X_n \underset{aprox}{\sim} N\left(\frac{n}{2}, \frac{n}{12}\right)$$

Por lo tanto, para obtener valores aproximados de la variable $Z \sim N\left(0, 1\right)$ basta pasar a normal estándar la distribución aproximada de S_n. Así se obtiene que

$$Z = \frac{S_n - \frac{n}{2}}{\sqrt{\frac{n}{12}}} \sim N\left(0, 1\right)$$

$\sqrt{\frac{40}{12}} = \frac{1}{3}\sqrt{3}\sqrt{10}: \frac{10}{3}$

Ejemplo 10 *Generemos 5 valores de la variable normal estándar Z. Utilizando* $n = 40$ *se tiene que*

$$Z = \frac{S_{40} - 20}{\sqrt{\frac{10}{3}}}$$

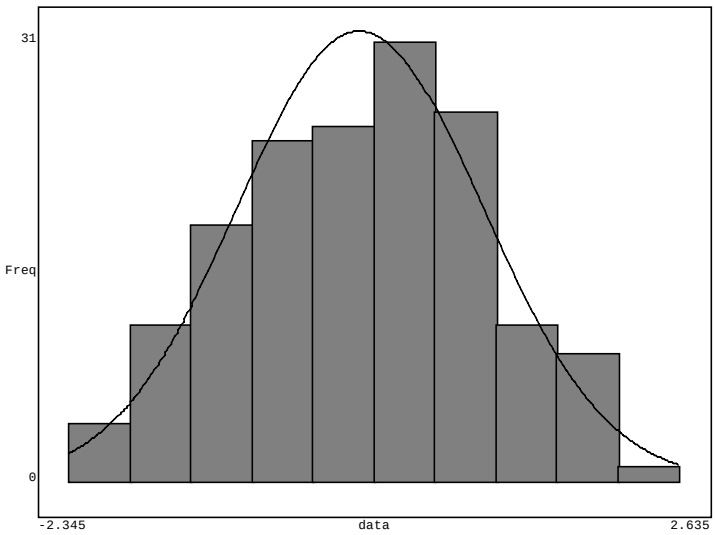
Así, con ayuda una calculadora o computadora se determine los valores de Z. Note que para cada valor de Z se calcula la suma de 40 valores de la función random, a este resultado se le resta 20 y se divide por $\sqrt{\frac{10}{3}}$. *Utilizando Excel, la función random es ALEATORIO(), así para obtener un valor de Z se debe escribir en un celda*

$$= (ALEATORIO() + AEATORIO() + ... + ALEATORIO() - 20)/RAIZ(10/3)$$

Una vez obtenido este valor, recuerde que esta fórmula se puede copiar a otras celdas, utilizando el Mouse, para obtener otros valores de Z. En nuestro caso, se obtienen los valores:

$$0,543551593 \qquad 0,260990164 \qquad -0,514336612 \qquad -2,200284197 \qquad -1,733486719$$

Por otro lado, obteniendo 160 valores y realizando el histograma se obtiene



¿Cómo generar valores de una variable normal no estándar? Suponga que se quienen generar valores de la variable

$$X \sim N\left(\mu, \sigma^2\right)$$

Dado que $Z = \dfrac{X-\mu}{\sigma} \sim N(0,1)$ entonces $X = \sigma Z + \mu$ y como $Z = \dfrac{S_n - \dfrac{n}{2}}{\sqrt{\dfrac{n}{12}}}$ se obtiene que

$$X = \sigma \left(\frac{S_n - \dfrac{n}{2}}{\sqrt{\dfrac{n}{12}}} \right) + \mu$$

Esta ecuación permite generar valores aproximados de X, para un n fijo mayor a 30.

4 Algunas distribuciones importantes

4.1 Distribución Gamma

Definición 10 *Se define la función Gamma por*

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$

Teorema 6 *Se tiene que*

1. $\Gamma(1) = 1$
2. $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$
3. *Para todo* $n \in \mathbb{N} : \Gamma(n) = (n-1)!$

Definición 11 *Sea X una v.a.c. Se dice que X sigue una distribución gamma si y solo si su función de probabilidad esta dada por*

$$f_X(x) = \begin{cases} \dfrac{x^{\alpha-1} e^{\frac{-x}{\beta}}}{\beta^\alpha \Gamma(\alpha)} & si \quad x \geq 0 \\ 0 & en\ otro\ caso \end{cases}$$

donde α y β son constantes. Se escribe

$$X \sim Gamma(\alpha, \beta)$$

Teorema 7 *Sea X una v.a.c. tal que $X \sim Gamma(\alpha, \beta)$, entonces:*

1. $E(X) = \alpha\beta$
2. $Var(X) = \alpha\beta^2$

Si $X \sim Gamma(\alpha, \beta)$ para realizar los calculos F_X se utiliza la tabla de la Gamma Imcompleta:

$$F_X(k) = F\left(\frac{k}{\beta}, \alpha\right)$$

donde

$y \diagdown \alpha$			α
			$\vdots$
y	$\ldots$	$\ldots$	$F(y, \alpha)$

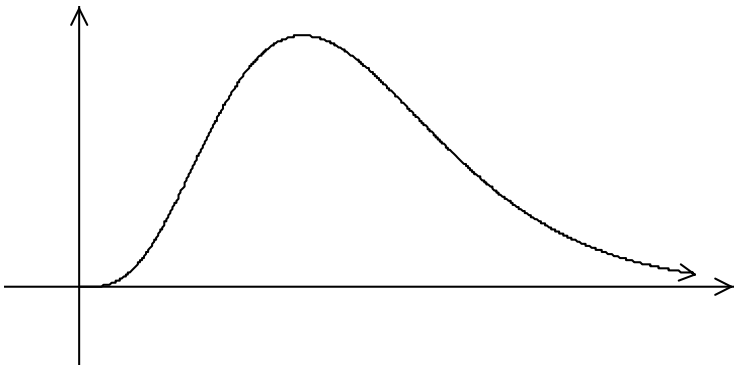
En nuestro caso, el interés en esta distribución se centra en su utilidad para definir la distribución Chi Cuadrado.

4.2 Distribución Chi Cuadrado

Definición 12 *Sea X una v.a.c. Se dice que X sigue una distribución chi cuadrada con v grados de libertad si y solo si*

$$X \sim Gamma(v/2, 2)$$

Se denota $X \sim \chi^2(v)$



Los grados de libertad indican las piezas de información independientes en el cálculo de X.

Teorema 8 *Sea X una v.a.c. tal que $X \sim \chi^2(v)$, entonces:*

1. $E(X) = v$
2. $Var(X) = 2v$

Prueba. *Ejercicio.* ■

Definición 13 *Si χ^2 tiene una distribución chi cuadrado con v grados de libertad entonces se define* ***el valor chi***

$$\chi^2_{\alpha,v} : \textit{valor de } \chi^2 \textit{ que acumula una área a la izquierda de } \alpha\%.$$

Es decir $P\left(\chi^2 \leq \chi^2_{\alpha,v}\right) = \alpha$.

Para determinar el valor de $\chi^2_{\alpha,v}$ se suelen utilizar tablas.

Ejemplo 11 *Calcular el valor de* $\chi^2_{0.2,14}$. *Buscando en la tabla se tiene que*

$v \diagdown \alpha$			0.2
			$\vdots$
14	...	...	14.6853

Por lo tanto $\chi^2_{0.2,14} = 9.46733$.

Ejemplo 12 *Si* $\chi^2 \sim \chi^2(10)$, *determine aproximadamente la* $P(X > 4)$.

Debido a la tabla, se necesita expresar la probabilidad solicitada en terminos de la acumulada de χ^2 :

$$P\left(\chi^2 > 4\right) = 1 - P\left(\chi^2 \leq 4\right)$$

Luego en la fila de $v = 10$ *se busca el valor más cercano a* 4, *este es* 3.94030, *por lo tanto*

$$P\left(\chi^2 > 4\right) = 1 - P\left(\chi^2 \leq 4\right) \approx 1 - 0.05 = 0.95$$

También se puede acotar está probabilidad. Note que en la fila de $v = 10$, 4 *se encuentra entre* 3.94030 *y* 4.86518, *así* $P\left(\chi^2 \leq 4\right) \in]0.05, 0.1[$, *por lo tanto*

$$P\left(\chi^2 > 4\right) \in]0.9, 0.95[.$$

Ejercicio 3 *Suponga que* $\chi^2 \sim \chi^2(24)$. *Acote el valor de* $P\left(\chi^2 > 17\right)$. $R/$ $]0.8, 0.9[$

El siguiente teorema indica la utilidad de esta distribución.

Teorema 9 *Considere la población dada por la variable aleatoria* X *que sigue una distribución normal con varianza poblacional* σ^2. *Dada una muestra aleatoria de esta población* $(X_1, X_2, X_3, ..., X_n)$. *Entonces la variable*

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

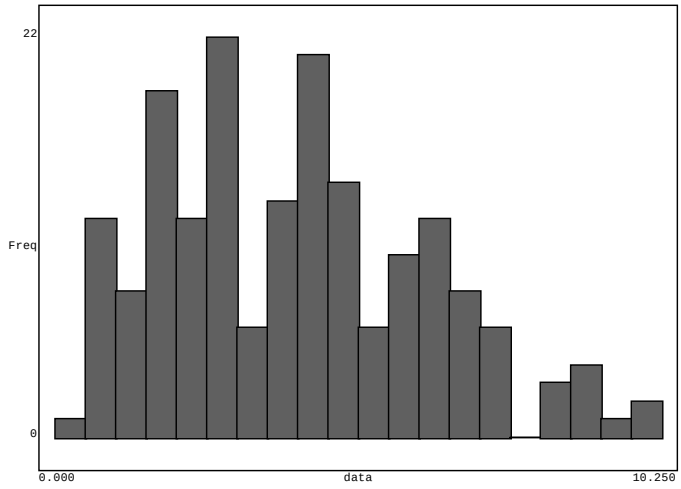
sigue una distribución χ^2 *con* $v = n - 1$ *grados de libertad.*

Ejemplo 13 *Los siguientes datos son valores de una variable* Z *que sigue una distribución normal estándar:*

$0,543551593$	$0,260990164$	$-0,514336612$	$-2,200284197$	$-1,733486719$
$0,853214429$	$-0,559109129$	$-0,993302277$	$-1,477762554$	$0,836494738$
$-0,682542739$	$-0,24714637$	$0,693164286$	$0,556075608$	$-0,105383202$
$-0,174740261$	$0,883016138$	$-0,053566134$	$1,07158394$	$0,411039337$

A partir de estos valores se tomaron 180 muestras aleatorias de tamaño 6 y se registro el valor $\chi^2 = \frac{(n-1)S^2}{\sigma^2} = 5S^2$ *para cada muestra. Realizando un histograma con los valores registrados se*

obtiene:



Note que la distribución Chi-Cuadrado se ajusta bien a este histograma.

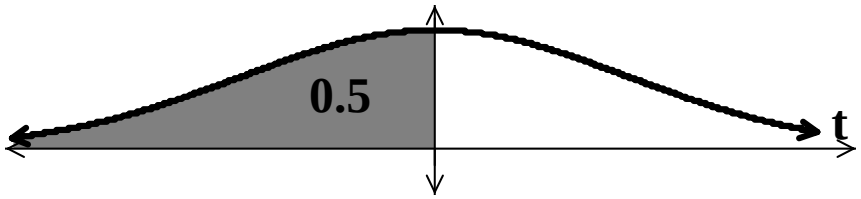
4.3 Distribución t

Definición 14 *Sea T una v.a.c. Se dice que T sigue una distribución t con v grados de libertad si y solo si su función de distribución esta dada por*

$$f_T(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\sqrt{\pi v}}\left(1+\frac{t^2}{v}\right)^{-(v+1)/2}$$

Se denota $T \sim t(v)$.

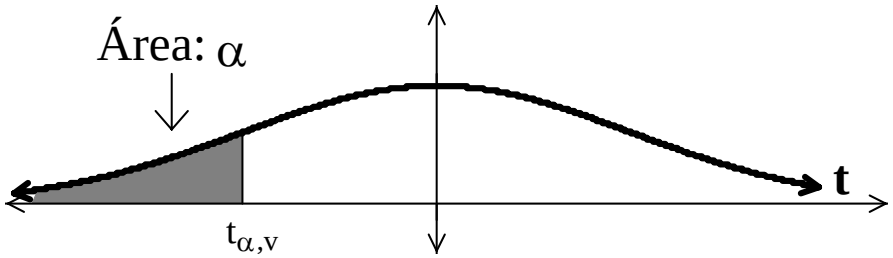
La distribución t es simétrica con respecto a su media que es 0.



Definición 15 *Si T tiene una distribución t con v grados de libertad entonces se define* ***el valor t****:*

$t_{\alpha,v}$ *: valor de T que acumula una área a la izquierda de $\alpha\%$.*

Es decir $P(T \leq t_{\alpha,v}) = \alpha$



y por la simetría de t : $P(T > -t_{\alpha,v}) = \alpha$.

Para determinar el valor de $t_{\alpha,v}$ .se suelen utilizar tablas.

Teorema 10 *Dado que* $t_{\alpha,v} = -t_{1-\alpha,v}$ *entonces* $|t_{\alpha,v}| = |t_{1-\alpha,v}|$.

Así, se puede utilizar una de las siguientes tablas para hallar $|t_{\alpha,v}|$:

$v \diagdown 1-\alpha$			$1-\alpha$
			$\vdots$
v	$\ldots$	$\ldots$	$\lvert t_{1-\alpha,v}\rvert$

$v \diagdown \alpha$			α
			$\vdots$
v	$\ldots$	$\ldots$	$\lvert t_{\alpha,v}\rvert$

Generalmente se dispone solo de una de las dos tablas. La tabla a utilizar, en nuestro caso, tienen los valores de $|t_{\alpha,v}|$ para $\alpha < 0.5$. Dado que la distribución t está centrada en cero entonces $t_{\alpha,v}$ en la tabla a utilizar es negativo, por lo tanto si x es positivo entonces

$$P(T > x) = P(T < -x)$$

Así, si $T \sim t(v)$ y x es positivo, para determinar aproximadamente $P(T > x)$ se busca en la fila del v el valor más cercano a x, y el α asociado a ese valor es la probabilidad solicitada.

Ejemplo 14 *Calcular el valor de* $t_{0.975,14}$. *Buscando en la tabla se tiene que*

$v \diagdown 1-\alpha$			0.025
			$\vdots$
14	$\ldots$	$\ldots$	2.14479

Por lo tanto $|t_{0.975,14}| = 2.14479$, *como este valor acumula una área a la izquierda del* 97.5% *entonces es positivo:* $t_{0.975,14} = 2.14479$.

Ejercicio 4 *Si* $t \sim t(48)$, *determine el valor de a que cumple que* $0.04 = P(t > a)$ $R/$ 1.78854.

Ejercicio 5 *Si* $t \sim t(12)$, *determine aproximadamente*

1. $P(t > 2.2)$ $R/$ 0.025
2. $P(t < -1.4)$ $R/$ 0.1
3. $P(t < 1.8)$ $R/$ 0.95

El siguiente relaciona las distribuciones estudiadas.

Teorema 11 *Sean Z y V variables aleatorias continuas independientes tales que*

$$Z \sim N(0,1) \quad y \quad V \sim \chi^2(v)$$

Considere la v.a.c. dada por

$$T = \frac{Z}{\sqrt{V/v}}$$

Se tiene que T sigue una distribución t con v grados de libertad.

El resultado siguiente presenta la utilidad de la distribución t.

Teorema 12 *Considere la población dada por la variable aleatoria X que sigue una distribución normal con media poblacional μ y varianza poblacional σ^2. Dada una muestra aleatoria de esta población $(X_1, X_2, X_3, ..., X_n)$. Entonces la variable*

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

sigue una distribución t con $v = n - 1$ grados de libertad.

Esquema de prueba *Siga los siguentes pasos para obtener la demostración del teorema :*

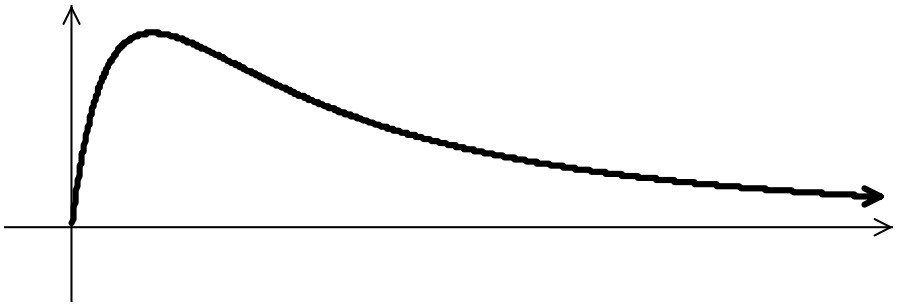
1. *Considere las variables*
$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \quad y \quad V = \frac{(n-1)S^2}{\sigma^2}$$
2. *Determine la distribución de las variables Z y V*
3. *Aplique el teorema anterior*

4.4 Distribución F

Definición 16 *Sea F una v.a.c. Se dice que F sigue una distribución f con v_1 y v_2 grados de libertad si y solo si su función de distribución esta dada por*

$$f_F(x) = \begin{cases} \dfrac{\Gamma\left(\dfrac{v_1 + v_{21}}{2}\right)\left(\dfrac{v_1}{v_2}\right)^{v_1/2}}{\Gamma\left(\dfrac{v_1}{2}\right)\Gamma\left(\dfrac{v_2}{2}\right)} \dfrac{x^{v_1/2-1}}{\left(1 + \dfrac{v_1 x}{v_2}\right)^{(v_1+v_2)/2}} & si \quad x > 0 \\ 0 & en \quad otro\ caso \end{cases}$$

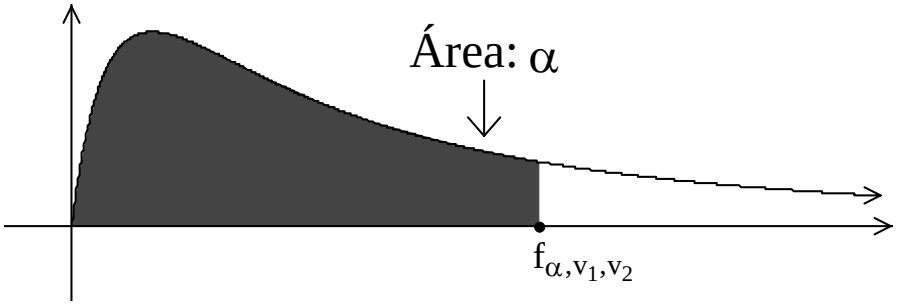
La distribución F es asimétrica:



Definición 17 *Si F tiene una distribución F con v_1 y v_2 grados de libertad entonces se define* ***el valor f****:*

$$f_{\alpha,v_1,v_2} : \textit{valor de } F \textit{ que acumula una área a la izquierda de } \alpha\%.$$

Es decir $P\left(F \leq f_{\alpha,v_1,v_2}\right) = \alpha$



Teorema 13 *Se tiene que*

$$f_{1-\alpha,v_1,v_2} = \frac{1}{f_{\alpha,v_2,v_1}}$$

Ejemplo 15 *Determine el valor de $f_{0.95,6,10}$ y de $f_{0.05,10,6}$. Utilizando la tabla*

$\alpha = 0.95$

$v_1 \diagdown v_2$			10
			$\vdots$
6	...	...	3.217

Por lo tanto $f_{0.95,6,10} = 3.217$ *y* $f_{0.05,10,6} = \dfrac{1}{3.217} = 0.310\,849$

Note que, en las tablas suministradas de la distribución F, el valor de α es mayor a 0.9 y los valores de f son mayores a 1. Entonces, por el teorema anterior, para $\alpha \leq 0.1$ se tiene que $f_{\alpha,v_1,v_2} < 1$. Esto es útil para determinar probabilidades (valores de α) de la distribución F.

Ejemplo 16 *Si* $F \sim f(14, 20)$ *Acote el valor de* $P(F < 2.5)$.

Se busca el valor de α *que cumple que* $f_{\alpha,14,20} = 2.5$, *dado que este valor* f *es mayor que 1, los valores disponibles son*

$$f_{0.9,14,20} = 1.859, \quad f_{0.95,14,20} = 2.225, \quad f_{0.975,14,20} = 2.603, \quad f_{0.99,14,20} = 3.130.$$

Note que 2.5 *está entre* $f_{0.95,14,20} = 2.225$ *y* $f_{0.975,14,20} = 2.603$, *por lo tanto*

$$P(F < 2.5) \in]0.95, 0.975[$$

y entonces

$$P(F > 2.5) = 1 - P(F < 2.5) \in]0.025, 0.05[$$

Por lo general, como se evidencia más adelante, bastará con acotar el valor de las probabilidades para F en lugar determinar su valor para tomar alguna decisión.

Ejemplo 17 *Si* $F \sim f(14, 20)$ *Acote el valor de* $P(F > 1.289)$.

Para calcular $P(F > 1.289)$ *se necesita expresar la probabilidad en términos de la acumulada* :

$$P(F > 1.289) = 1 - P(F < 1.289)$$

Note que se busca el valor de α *que cumple que* $f_{\alpha,14,20} = 1.289$, *dado que este valor* f *es mayor que 1, los valores disponibles son*

$$f_{0.9,14,20} = 1.859, \quad f_{0.95,14,20} = 2.225, \quad f_{0.975,14,20} = 2.603, \quad f_{0.99,14,20} = 3.130.$$

Note que 1.289 *está entre* $f_{0.9,14,20} = 1.859$ *y* $f_{0.1,14,20} < 1$ *(pues* $\alpha \leq 0.1$*), por lo tanto*

$$P(F < 1.289) \in]0.1, 0.9[$$

y entonces

$$P(F > 1.289) = 1 - P(F < 1.289) \in]0.1, 0.9[.$$

Ejemplo 18 *Si* $F \sim f(10, 16)$ *Calcule* $P(F < 0.3)$.

Esta probabilidad ya está expresada en términos de la acumulada. Se busca el valor de α *que cumple que* $f_{\alpha,10,16} = 0.3$, *dado que este valor* f *es menor que 1, los valores disponibles son*

$$\begin{aligned}
f_{0.1,10,16} &= \frac{1}{f_{0.9,16,10}} = \frac{1}{2.094} = 0.477\,555, \\
f_{0.05,10,16} &= \frac{1}{f_{0.95,16,10}} = \frac{1}{2.828} = 0.353\,607, \\
f_{0.025,10,16} &= \frac{1}{f_{0.975,16,10}} = \frac{1}{3.496} = 0.286\,041, \\
f_{0.01,10,16} &= \frac{1}{f_{0.99,16,10}} = \frac{1}{4.520} = 0.221\,239
\end{aligned}$$

Dado que 0.3 *está entre* $f_{0.025,10,16}$ *y* $f_{0.05,10,16}$ *entonces*

$$P(F < 0.3) \in]0.025, 0.5[.$$

Ejercicio 6 *Suponga que* $F \sim f(8,20)$. *Acote lo mejor posible, utilizando las tablas, el valor de* $P(F > 0.2)$. $R/\quad]0.99, 0.975[$.

Teorema 14 *Considere las poblaciónes dadas por la variables aleatorias* X, Y *que siguen una distribución normal con varianzas poblacionales de* σ_1^2 *y* σ_2^2 *repectivamente. Sean* S_1^2 *y* S_2^2 *varianzas muestras de muestas aleatorias independientes de tamaño* n_1 *y* n_2*, tomadas de cada población respectivamente. Se tiene que*

$$F = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

sigue una distribución F *con* $v_1 = n_1 - 1$ *y* $v_2 = n_2 - 1$ *grados de libertad.*

5 Ejercicios

1. Cierto tipo de llantas tienen una duración media de 30000 kilómetros con una desviación de 2500 km. Don Juan tiene una flota de diez taxis, que utilizan ese tipo de llantas. Don Juan decide cambiarle las llantas a toda su flota.

   (a) ¿cuál es la probabilidad de que la de la duración promedio de las llantas adquiridas por don Juan sea inferior a 29500 kilómetros? $R/\quad 0.1029$

   (b) Se sabe que la probabilidad de que la duración promedio de las llantas adquiridas por don Juan sea mayor a c es de por lo menos 10%. Determine el valor máximo de c.$R/\quad 30505.96$

2. Una bolsa de zanahoria tiene un peso promedio de 2 kg y una desviación de 100 gramos. Se empacan cajas que contienen 70 bolsas. Determine la probabilidad de que en una caja el peso promedio por bolsa sea inferior a 1990 gramos. $R/\quad 0.2005$

3. El tiempo que un cajero del supermercado MENOS MENOS tarda con cada cliente tiene una distribución normal con un promedio de 7 minutos y una desviación estándar de 2 minutos. Si un cajero atendió a 8 clientes ¿Cuál es la probabilidad de que el tiempo que tardó en atenderlos sea inferior a 72 minutos? $R/\quad 0.9977$

4. Un paquete de azúcar DULCE pesa en promedio 2 kilos con una varianza de 0.01 kg^2. Estos paquetes son empacados en sacos de 30 paquetes cada uno.¿Cuál es la probabilidad de que el peso de un saco de azúcar sea inferior a 59 kilos? $R/\quad 0.0336$

5. Determine

   (a) $\chi^2_{0.025,24}$ $R/\quad 12.4012$

   (b) $t_{0.2,44}$ $R/\quad -0.849867$

   (c) $t_{0.98,6}$ $R/\quad 2.61224$

   (d) $t_{0.96,39}$ $R/\quad 1.797\,55$

   (e) $f_{0.99,11,17}$ $R/\quad : 4.180\,6$

   (f) $f_{0.975,4,6}$ $R/\quad 6.227$

   (g) $f_{0.025,4,6}$ $R/\quad 0.16$

(h) $z_{0.94}$ $R/\quad 1.55$

(i) $\chi^2_{0.2,22}$ $R/\quad 16.314$

(j) $t_{0.9,40}$ $R/\quad 1.30308$

(k) $f_{0.1,10,16}$ $R/\quad 0.4478$

(l) $z_{0.15}$ $R/\quad -1.04$

(m) $\chi^2_{0.8,20}$ $R/\quad 25.0375$

(n) $t_{0.1,36}$ $R/\quad -1.30551$

(o) $f_{0.9,10,12}$ $R/\quad 2.188$

(p) $z_{0.86}$ $R/\quad 1.08$

(q) $t_{0.1,30}$ $R/\quad -1.31042$

(r) $f_{0.05,12,22}$ $R/\quad 0.396\,354$

6. Para cada una de las variables indicadas, determine la probabilidad solicitada.

(a) $t \sim t(12), \quad P(t > 2)$ $R/\quad 0.04$

(b) $t \sim t(14), \quad P(t < 1.7)$ $R/\quad 0.95$

(c) $t \sim t(16), \quad P(t > -1.3)$ $R/\quad 0.9$

(d) $t \sim t(18), \quad P(t < -1.9)$ $R/\quad 0.04$

(e) $\chi^2 \sim \chi^2(20), \quad P\left(\chi^2 > 15\right)$ $R/\quad 0.8$

(f) $\chi^2 \sim \chi^2(22), \quad P\left(\chi^2 < 25\right)$ $R/\quad 0.8$

(g) $\chi^2 \sim \chi^2(24), \quad P\left(\chi^2 > 40\right)$ $R/\quad 0.025$

7. Para cada una de las variables indicadas, acote la probabilidad solicitada.

(a) $F \sim f(48, 20), \quad P(F > 2)$ $R/\quad]0.025, 0.05[$

(b) $F \sim f(2, 10), \quad P(F > 0.03)$ $R/\quad]0.95, 0.975[$

(c) $t \sim t(12), \quad P(t > 1.8)$ $R/\quad]0.04, 0.05[$

(d) $\chi^2 \sim \chi^2(20), \quad P\left(\chi^2 < 19\right)$ $R/\quad]0.2, 0.8[$

(e) $F \sim f(14, 20), \quad P(F > 0.57)$ $R/\quad]0.1, 0.9[$

(f) $t \sim t(10), \quad P(t < -2)$ $R/\quad]0.025, 0.04[$

(g) $\chi^2 \sim \chi^2(20), \quad P\left(\chi^2 > 10\right)$ $R/\quad]0.95, 0.975[$

(h) $F \sim f(12, 18), P(F > 0.7)$ $R/\quad]0.1, 0.9[$

(i) $\chi^2 \sim \chi^2(26), \quad P\left(\chi^2 > 18\right)$ $R/\quad]0.8, 0.9[$

(j) $F \sim f(6, 17), \quad P(F > 0.21).$ $R/\quad]0.95, 0.975[$

8. **Proyecto**: Generando Valores de la normal y Chi Cuadrado utilizando Excel u otro programa.

(a) Genere 500 valores de la variable $Z \sim N(0, 1)$.

(b) Realice un histograma con los valores generados de Z

(c) A partir del valores generados de Z genere 200 valores para la distribución $\chi^2(15)$ .Sugerencia: utilice el teorema 9, note que debe tomar 200 muestras aleatorias de tamaño 16..

(d) Realice un histograma con los valores χ^2 generados

(e) Genere 300 valores de la variable $X \sim N(5, 4)$.

(f) Realice un histograma con los valores X generados

Capítulo II

Estimación de parámetros

Se inicia con un estudo de los conceptos generales de estimación: estimación puntual, estimación por intervalo, precisión y confiabilidad. Luego se aborda la estimación de máxima verosimilitud y finalmente se estudian las deferentes técnicas para realizar estimaciones por intervalo para una y dos poblaciones.

1 Introducción

Anteriormente se estudiaron los conceptos generales de estadística diferencial, como la constante parámetro, la variable estadístico y su distribución de probabilidad llamada distribución muestral. El presente documento desarrolla una de los principales objetivos de la estadística inferencial, la estimación del valor del parámetro a partir de una muestra, el estadístico y una distribución muestral.

El concepto de estimación se desarrolla en los primeros años de vida a través de los sentidos, por ejemplo, se utiliza la vista para estimar la altura de una puerta y el tacto para estimar el peso de una bolsa de papas. Continuando con el primer ejemplo, la estimación de la altura de la puerta puede ser puntual indicado que mide aproximadamente 2.9 metros, o por intervalo indicado que su valor probablemente está entre 2.8 m y 3 m. Dado que los sentidos nos puede engañar se suele utilizar instrumentos de medición para hacer estimaciones más precisas.

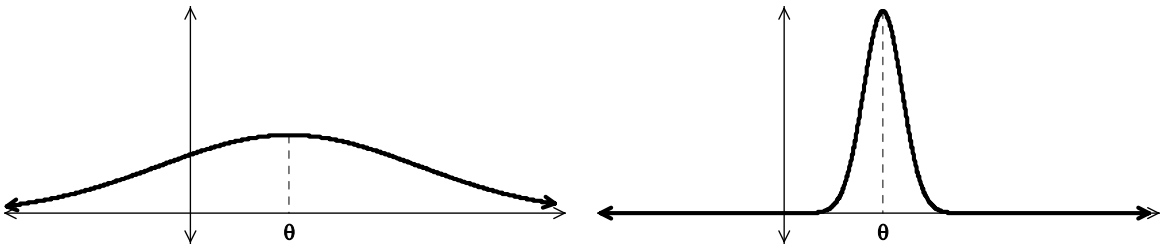
La estadística inferencia, para el caso de estimación por intervalo, propone reglas más objetivas para determinar la estimación del parámetro, donde se puede medir la confiabilidad del intervalo y controlar su tamaño manipulando el tamaño de muestra.

Seguidamente se abordarán los conceptos involucrados en los tipos de estimación.

2 Tipos de Estimación

La estimación estadística busca aproximar un parámetro θ a partir de un estadístico $\widehat{\Theta}$. Para que está estimación sea buena su valor esperado debe ser el parámetro buscado y además los valores aleatorios que genere debe ser cercanos al parámetro. Es decir, el estadístico $\widehat{\Theta}$ es un buen estimador si es insesgado $\left(E\left(\widehat{\Theta}\right)=\theta\right)$ y tener varianza pequeña, esto hace que los valores más probables de $\widehat{\Theta}$ se concentre en θ.

Ejemplo 19 *Considere los siguientes estadísticos insesgados que siguen una distribución normal:*



Distribución de $\widehat{\Theta_1}$ *Distribución de* $\widehat{\Theta_2}$

Si bien ambos estadísticos son insesgados, note que $\widehat{\Theta_2}$ *es el mejor, ya que tiene menor varianza y entonces sus valores se concentrar alrededor de* θ.

Ejercicio 7 *Sea* X *y* Y *dos variables aleatorias independientes tales que*

$$E(X)=5\theta, Var(X)=2, E(Y)=3\theta, Var(Y)=1.$$

Considere los siguientes estimadores de un parámetro θ de una población:

$$\widehat{\theta}_1 = \frac{X+Y}{8}, \qquad \widehat{\theta}_2 = \frac{X-Y}{2}.$$

1. *Determine si cada uno de estos estimadores es insesgado o no.*
2. *¿Cuál de estos es el mejor estimador del parámetro θ? Justifique su respuesta.*

¿Como utilizar un buen estimador $\widehat{\Theta}$ para estimar θ? Hay dos maneras de utilizar el estadístico para estimar el parámetro:

1. **Estimación puntual**: se determina un valor único $\widehat{\theta}$ de $\widehat{\Theta}$ a partir de una muestra de tamaño adecuado. Se dice que $\widehat{\theta}$ es una estimación de θ.
2. **Estimación por intervalo**: se determina un intervalo probable de valores de $\theta : I =]\theta_i, \theta_s[$. El **nivel de confianza** del intervalo es la probabilidad de que θ esté en $I : P(\theta \in I)$. La **precisión del intervalo** esta relacionado inversamente con su ancho: $\theta_s - \theta_i$; es decir si el intervalo es muy preciso entonces tiene menos ancho. Una relación entre estos dos últimos conceptos son:

   Mayor precisión $\Leftrightarrow$ menor confianza

   Por ejemplo, ¿Cuál es peso promedio de un elefante africano de sabana? Un posible respuesta es que este valor promedio u esta entre 1 y 20 toneladas. Esta estimación por intervalo es muy confiable (hay casi un 100% de probabilidad de que se encuentre en ese intervalo, pues su peso no pasa las 10 tonelas) pero poco precisa (¡el intervalo es enorme!).

3 Estimación puntual

3.1 Estimación de los principales parámetros

Si el estadístico $\widehat{\Theta}$ esta asociado al parámetro θ se puede tomar el valor $\widehat{\theta}$ de $\widehat{\Theta}$, en una muestra aleatoria de tamaño n, como estimación puntual de θ. El problema en este tipo de estimaciones es que no se puede medir el error de estimación, sin embargo entre mayor sea el tamaño la muestra se espera que mejor sea la estimación.

Ejemplo 20 *Las duraciones de ocho baterías cargadas de computadora de marca DUTEC son 151, 153, 175, 134, 170, 172, 156 y 114 minutos. Entonces una estimación puntual de la duración promedio de una batería DUTEC es*

$$\overline{x} = \frac{151+153+175+134+170+172+156+114}{8} = 153.125$$

Suponga que la empresa DUTEC asegura que la duración promedio de sus baterías es de 200. De acuerdo a la estimación obtenida se puede sospechar que la afirmación de DUTEC es poco probable.

Sin embargo, esta sospecha debe ser aceptada o rechazada con métodos estadísticos que se estudian más adelante. Repasando los conceptos estudiados, note que esta situación se tiene que

$$\begin{array}{cc} \textit{Población } X: & \textit{duración de una batería DUTEC cargada} \\ \textit{Parámetro } \mu: & \textit{duración promedio de una batería DUTEC cargada} \\ \textit{Estimador } \overline{X}: & \textit{duración promedio de ocho baterías DUTEC cargada} \\ & \textit{Tamaño de muestra: } n = 8 \\ & \textit{Estimación: } \overline{x} = 153.125 \end{array}$$

De acuerdo a lo estudiado en el primer capítulo, dada una muestra observada $(x_1, x_2, ..., x_n)$de una muestra aleatoria M se tienen las siguientes estimaciones puntuales de los respectivos parámetros:

Parámetro	Estimador	Estimación puntual
μ	$\overline{X}$	$\overline{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$
σ^2	S^2	$s^2 = \dfrac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}$
σ	S	$s = \sqrt{s^2}$
p	$\widehat{P}$	$\widehat{p} = \dfrac{b}{n}$, b es el número de éxitos en la muestra

Suponga que se tiene una población X_1 con media poblacional μ_1 y una población X_2 con media poblacional μ_2. Si de cada población se toman muestras de tamaño n_1 y n_2 respectivamente, entonces un estimador insesgado de la diferencia de promedios $(\mu_1 - \mu_2)$ es: $\overline{X}_1 - \overline{X}_2$. Así, si se toman valores de las muestras aletorias: $(x_1, x_2, ..., x_{n_1})$ y $(y_1, y_2, ..., y_{n_2})$, entonces una estimación puntual de la diferencia de promedios $(\mu_1 - \mu_2)$ es:

$$\overline{x}_1 - \overline{x}_2 = \frac{\sum_{i=1}^{n_1} x_i}{n_1} - \frac{\sum_{i=1}^{n_2} y_i}{n_2}.$$

Similarmente, considerando la misma notación, un estimador insesgado para la diferencia de dos proporciones $(p_1 - p_2)$ es $\widehat{P}_1 - \widehat{P}_2$ y una estimación puntual es

$$\widehat{p}_1 - \widehat{p}_2 = \frac{b_1}{n_1} - \frac{b_2}{n_2}$$

Ejercicio 8 *Una bebida afirma en su publicidad por televisión que su empleo diario, durante un mes, produce una pérdida promedio de al menos cinco libras de peso. Para analizar esta afirmación, se toma un grupo control de ocho personas y se les suministro el producto diariamente por un mes obteniendo los siguientes datos:*

Peso Inicial (lb)	*165*	*195*	*188*	*170*	*185*	*163*	*155*	*177*
Peso Final (lb)	*164*	*190*	*187*	*163*	*185*	*159*	*148*	*174*

1. *Determine a partir de la muestra una estimación puntual de la pérdida de peso promedio.R/* 3.5
2. *Determine a partir de la muestra una estimación puntual de la diferencia entre el peso promedio inicial y el peso promedio final.* *R/* 3.5
3. *¿Considera correcta la información hecha en publicidad?*

Este tipo de estimación aunque es muy intuitiva tiene el problema de que no se puede determinar su confianza.

3.2 Estimación de máxima verosimitud

Este tipo de estimación puntual tiene como requisito que se conozca la distribución de la población en términos del parámetro desconocido.

Definición 18 *La **función de verosimilitud** es la probabilidad de que una muestra aleatoria observada ocurra en función del parámetro desconocido.*

Debido a la definición de muestra aleatoria tamaño n: n variables $(X_1, X_2, X_3, ..., X_n)$ mutuamente independientes se obtiene el siguiente teorema.

Teorema 15 *Dada una muestra observada $m = (x_1, x_2, ..., x_n)$ de una población con parámetro θ y función de distribución f que depende de θ, se tiene que la función de verosimiltud es*

$$g(\theta) = f(x_1) f(x_2) \cdots f(x_n)$$

Justificación. *Se tiene una muestra aleatoria de tamaño $n : (X_1, X_2, ..., X_n)$ y una muestra observada $(x_1, x_2, ..., x_n)$. De acuerdo a la definición de función de verosimitud, está es*

$$\begin{aligned} g(\theta) &= P(M = m) \\ &= P(X_1 = x_1 \cap X_2 = x_2 \cap ... \cap X_n = x_n) \\ &= P(X_1 = x_1) P(X_2 = x_2) \cdots P(X_n = x_n) \quad \textit{(Pues } X_1, ..., X_n \textit{ son independientes)} \\ &= f(x_1) f(x_2) \cdots f(x_n) \end{aligned}$$

Note que está justificación es una prueba válida cuando la población es discreta, y nos permite tener una idea intuitiva de la función de verosimilitud. ■

La estimación de máxima verosimitud parte de la premisa de que la muestra observada es la muestra que tiene mayor probabilidad de ser observada, por lo tanto una estimación del parámetro es aquella que maximiza la función de verosimilitud g. Así, la estimación de θ consiste en maximar dicha función.

Recuerde que para maximar una función en una variable se debe derivar dicha función para hallar los posibles extremos de la función. Dado que la función g es una multiplicación de varios términos, se suele utilizar el método de derivación logarítmica para obtener su derivada.

Ejemplo 21 *La variable aleatoria X tiene densidad de probabilidad dada por*

$$f(x) = \frac{x}{a^2 e^{x/a}}, \text{ para } x \geq 0$$

Tres observaciones de X son $x_1 = 12; x_2 = 16$ y $x_3 = 15$. Encuentre la estimación de máxima verosimilitud del parámetro a.

1. ***Determinación de la función de verosimilitud*** *g. En este caso la muestra observada es $(12, 16, 15)$, por lo tanto la función de verosimitud es*

$$g(a) = f(12) f(16) f(15) = \frac{12}{a^2 e^{12/a}} \frac{16}{a^2 e^{16/a}} \frac{15}{a^2 e^{15/a}} = \frac{2880}{a^6 e^{43/a}}$$

2. ***Simplificación de*** $\ln g(a)$. *Se busca el valor de a que maximice g. Para ello lo mejor es primeramente aplicar logaritmo a g :*

$$\ln g(a) = \ln 2880 - 6 \ln a - \frac{43}{a}$$

3. ***Calcular*** $g'(a)$. *Derivando la expresión anterior se tiene que*

$$\frac{g'(a)}{g(a)} = -\frac{6}{a} + \frac{43}{a^2} \quad \Longrightarrow \quad g'(a) = \left(-\frac{6}{a} + \frac{43}{a^2}\right) g(a)$$

4. ***Estimación de*** *a. Para que $g'(a) = 0$ basta que $-\frac{6}{a} + \frac{43}{a^2} = 0$ o que $g(a) = 0$, dado que se busca el valor máximo de $g(a)$ y que $g(a) \geq 0$, pues es una multiplicación de probabilidades, se descarta la ecuación $g(a) = 0$, ya que está busca el valor de a que hace mínimo a g. Así,*

$$-\frac{6}{a} + \frac{43}{a^2} = 0 \Longrightarrow a = \frac{43}{6}$$

Por lo tanto, la estimación de máxima verosimilitud del parámetro a es $\frac{43}{6}$. El lector puede verificar que en este valor la función g alcanza sui máximo.

Ejercicio 9 *Una variable aleatoria X tiene una distribución con densidad $f(x) = \alpha^3 x^{\alpha^3 - 1}$ para $0 \leq x \leq 1$, con α constante. Dadas las observaciones $x_1 = 0.5$, $x_2 = 0.4$ y $x_3 = 0.25$, encuentre la estimación de máxima verosimilitud de α.* $R/ \quad \alpha = \sqrt[3]{\frac{-3}{\ln 0.05}}$

Cuando son varios los parámetros desconocidos, la maximización de la función de verosimitud se logra cuando las derivadas parciales son nulas, por lo tanto, las estimaciones se obtienen al resolver el sistema de las derivadas parciales iguales a cero.

Por ejemplo, si son dos parametros desconocidos α y β, la función de distribución será $g(\alpha, \beta)$. Y la estimación de máxima verosimitud de los parámetros es, en condiciones ideales, solución del sistema

$$\begin{cases} \dfrac{\partial g}{\partial \alpha} = 0 \\ \dfrac{\partial g}{\partial \beta} = 0 \end{cases}.$$

Ejercicio 10 *Una variable aleatoria X tiene una distribución con densidad*

$$f(x) = \frac{1}{\sqrt{2\pi v}} e^{-(x-u)^2/(2v)}$$

Se observan los valores de X : $152, 93, -14$ y 65.

1. *Pruebe que la función de verosimilitud es*

$$g(u, v) = \frac{1}{4\pi^2 v^2} e^{-v^{-1}\left(2u^2 - 296u + 18\,087\right)}$$

2. *Determine las estimaciones de máxima verosimilitud de u y v.* $R/ \quad u = 74, \quad v = 3567.5$

3.3 Ejercicios

1. Sea X y Y dos variables aleatorias independientes tales que

$$X \sim N(5\theta + 5, 9), \quad Y \sim B\left(20, \frac{1}{4}\right)$$

Considere los siguientes estimadores de un parámetro θ de una determina población:

$$\widehat{\theta}_1 = \frac{X - Y}{5}, \qquad \widehat{\theta}_2 = \frac{X + Y}{5}$$

¿Cuál de estos es el mejor estimador del parámetro θ? Justifique su respuesta.$R/$ Es mejor $\widehat{\theta}_1$.

2. Sea X y Y dos variables aleatorias independientes tales que[1]

$$X \sim B\left(10\theta, \frac{1}{2}\right), \quad Y \sim B\left(20\theta, \frac{1}{5}\right)$$

Considere los siguientes estimadores de un parámetro θ de una determina población:

$$\widehat{\theta}_1 = X - Y, \qquad \widehat{\theta}_2 = \frac{X + Y}{9}$$

¿Cuál de estos es el mejor estimador del parámetro θ? Justifique su respuesta. $R/ \quad \widehat{\theta}_2$

3. Una variable aleatoria X tiene una distribución con densidad $f(x) = \lambda e^{-\lambda x}$ para$x \geq 0$, con λ una constante positiva. Dadas las observaciones $x_1 = 0.3$, $x_2 = 0.2$ y x_3, se obtuvo que la estimación de máxima verosimilitud de λ es 4. Determine el valor de x_3.. $R/ \quad x_3 = 0.25$

4. Sea X y Y dos variables aleatorias independientes tales que

$$X \sim N(5\theta, 9), \quad Y \sim N(6\theta, 9.2)$$

Considere los siguientes estimadores de un parámetro θ de una determina población:

$$\widehat{\theta}_1 = \frac{X}{5}, \qquad \widehat{\theta}_2 = Y - X, \qquad \widehat{\theta}_3 = \frac{X + Y}{10}$$

[1] Recuerde que si $W \sim B(n, p)$ significa que W sigue una Distribución Binomial donde n es el número de intentos y p la probabilidad de obtener un éxito.

(a) Determine si cada uno de estos estimadores es insesgado o no. $R/$ $\widehat{\theta}_1$ y $\widehat{\theta}_2$ son insesgados

(b) ¿Cuál de estos es el mejor estimador del parámetro θ? Justifique su respuesta. $R/$ Es mejor $\widehat{\theta}_1$

5. Una variable aleatoria X tiene una distribución con densidad $f(x) = \frac{k}{3}\left(\frac{x}{3}\right)^{k-1}$ para $0 \leq x \leq 3$, con k constante. Dadas las observaciones $x_1 = 0.3$, $x_2 = 0.1$ y $x_3 = 0.9$, encuentre la estimación de máxima verosimilitud de k. $R/$ $k \approx 0.434\,29$

6. Una variable aleatoria X tiene una distribución con densidad $f(x) = (\alpha+1)x^{\alpha}$ para $0 \leq x \leq 1$, con α constante. Dadas las observaciones $x_1 = 0.2$, $x_2 = 0.1$ y $x_3 = 0.5$, encuentre la estimación de máxima verosimilitud de α. $R/$ $a \approx -0.348\,558$

7. Una variable aleatoria X tiene una distribución con densidad $f(x) = \frac{k}{5}\left(\frac{x}{5}\right)^{k-1}$ para $0 \leq x \leq 5$, con k constante. Dadas las observaciones $x_1 = 2.5$, $x_2 = 0.1$ y $x_3 = 5$, encuentre la estimación de máxima verosimilitud de k. $R/$ $k \approx 0.651\,442$

8. Una variable aleatoria X tiene una distribución con densidad $f(x) = \frac{k^2}{3}\left(\frac{x}{3}\right)^{k^2-1}$ para $0 \leq x \leq 3$, con k constante. Dadas las observaciones $x_1 = 0.3$, $x_2 = 0.1$ y $x_3 = 0.9$, encuentre las estimaciones de máxima verosimilitud de k. $R/$ $k \approx \pm 0.659\,01$

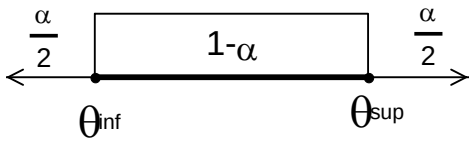
9. Una variable aleatoria X tiene una distribución con densidad $f(x) = f\,(x) = \lambda^2 e^{-\lambda^2(x-5)}$ para $x \geq 5$, con λ constante. Dadas las observaciones $x_1 = 6$, $x_2 = 6.5$ y $x_3 = 7.5$, encuentre las estimaciones de máxima verosimilitud de λ. $R/$ $k \approx \pm 0.774\,597$

4 Estimación por intervalo: Intervalo de confianza (IC)

4.1 Introducción

Definición 19 *Dado el estadístico $\widehat{\Theta}$ insesgado asociado al parámetro θ de una población, tal que existe un intervalo $I\left(\widehat{\Theta}\right) = \left]\widehat{\Theta}_{\text{inf}}, \widehat{\Theta}_{\text{sup}}\right[$ que depende de $\widehat{\Theta}$ y cumple que*

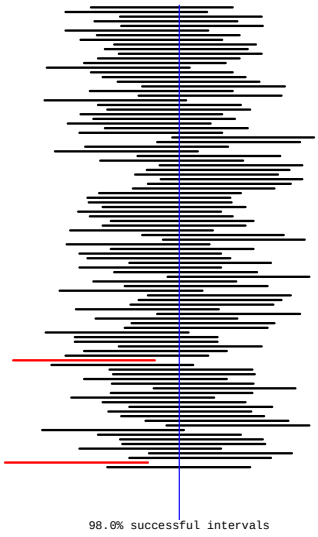
$$P\left(\theta \in I\left(\widehat{\Theta}\right)\right) = 1-\alpha. \quad y$$
$$P\left(\theta \leq \widehat{\Theta}_{\text{inf}}\right) = \frac{\alpha}{2}.$$



*Un **intervalo de confianza (IC)** del $(1-\alpha)\,100\%$ para θ es el intervalo I evaluado en una buena estimación puntual $\widehat{\theta}$ de $\widehat{\Theta} : I\left(\widehat{\theta}\right)$.*

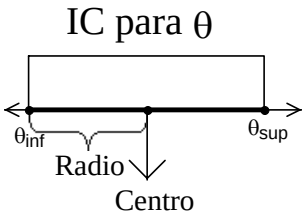
La definición anterior, indica que un intervalo de confianza depende del valor de una estimación $\widehat{\theta}$ de θ. ¿Qué significado tiene la confianza de $(1-\alpha)\,100\%$? Note que cada estimación de θ genera un IC, así si tomamos 100 estimaciones de $\widehat{\Theta}$, se obtienen 100 intervalos de confianza, de los cuales se esperaría que alrededor del $(1-\alpha)\,100\%$ contengan el parámetro.

Ejemplo 22 *Por medio de Winstats se puede simular* 100 *intervalos de confianza con una precisión de 20 unidades y un nivel de confianza de 95%, obteniedo*



La línea azul indica el valor del parámetro. Se puede observar que hay 98 intervalos que contienen el parámetro.

Definición 20 *Dada un intervalo de confianza* $I =]\theta_{\inf}, \theta_{\sup}[$ *se define el centro del IC por* $\dfrac{\theta_{\inf} + \theta_{\sup}}{2}$ *y el radio o error de estimación por* $\dfrac{\theta_{\sup} - \theta_{\inf}}{2}$.



Ejemplo 23 *Suponga que el estadístico* $\widehat{\Theta}$ *se distribuye normalmente*

$$\widehat{\Theta} \sim N(\theta, 4)$$

para muestras de un tamaño dado. Entonces $Z = \dfrac{\widehat{\Theta} - \theta}{2} \sim N(0,1)$. *Utilizando la tabla normal se tiene que*

$$P(z_{0.05} < Z < -z_{0.05}) = 0.9$$

Sustituyendo $Z = \dfrac{\widehat{\Theta} - \theta}{2}$:

$$\begin{aligned} & P\left(z_{0.05} < \frac{\widehat{\Theta} - \theta}{2} < -z_{0.05}\right) = 0.9 \\ \Longrightarrow \quad & P\left(2z_{0.05} < \widehat{\Theta} - \theta < -2z_{0.05}\right) = 0.9 \\ \Longrightarrow \quad & P\left(2z_{0.05} - \widehat{\Theta} < -\theta < -2z_{0.05} - \widehat{\Theta}\right) = 0.9 \\ \Longrightarrow \quad & P\left(\widehat{\Theta} + 2z_{0.05} < \theta < \widehat{\Theta} - 2z_{0.05}\right) = 0.9 \end{aligned}$$

Así, se tiene que

$$P\left(\theta \in I\left(\widehat{\Theta}\right)\right) = 90\%, \text{ con } I\left(\widehat{\Theta}\right) = \left]\widehat{\Theta} + 2z_{0.05}, \widehat{\Theta} - 2z_{0.05}\right[.$$

Por lo tanto, tomando una estimación $\widehat{\theta}$*, un IC del* 90% *tiene extremos*

$$\widehat{\theta} + 2z_{0.05}.$$

Lo realizado en el ejemplo anterior, se generaliza en el siguiente teorema.

Teorema 16 *Considere la población dada por la variable aleatoria* X *y el estadístico* $\widehat{\Theta}$ *asociado al parámetro de población* θ *para muestras de tamaño* n *con*

$$Var\left(\widehat{\Theta}\right) = \overline{\sigma}^2$$

Si $\widehat{\Theta}$ *es insesgado* $\left(E\left(\widehat{\Theta}\right) = \theta\right)$ *y su función de distribución es simétrica con respecto a* θ *entonces un intervalo de confianza de* $100\,(1-\alpha)\,\%$ *para* θ *tiene extremos*

$$\widehat{\theta} \pm A_{\alpha/2} \cdot \overline{\sigma}$$

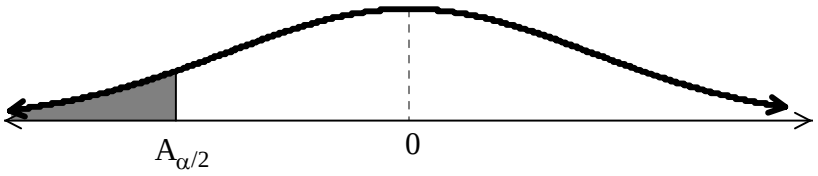
donde: $\widehat{\theta}$ *es el valor de* $\widehat{\Theta}$ *para una muestra de tamaño* n *y* $A_{\alpha/2}$ *cumple que*

$$P\left(\frac{\widehat{\Theta} - \theta}{\overline{\sigma}} < A_{\alpha/2}\right) = \frac{\alpha}{2}$$

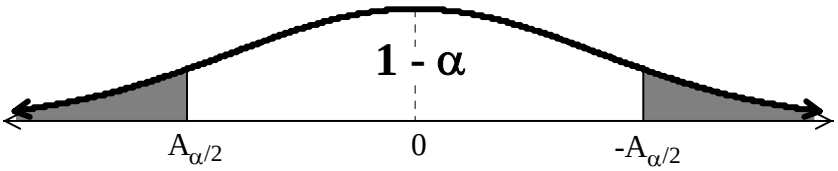
En este caso, $\widehat{\theta}$ *es el centro del intervalo y* $\left|A_{\alpha/2} \cdot \overline{\sigma}\right|$ *es su radio.*

Prueba. *Considere la variable aleatoria* $A = \dfrac{\widehat{\Theta} - \theta}{\overline{\sigma}}$*, note que* $E(A) = 0$ *y como la función de distribución de* $\widehat{\Theta}$ *es simétrica con respecto a* θ *entonces la función de distribución de* A *es*

simétrica respecto a 0 :



Como $P\left(A < A_{\alpha/2}\right) = \dfrac{\alpha}{2}$ *entonces* $P\left(A > -A_{\alpha/2}\right) = \dfrac{\alpha}{2}$:



Por lo tanto se tiene que

$$P\left(A_{\alpha/2} < A < -A_{\alpha/2}\right) = 1 - \left(\frac{\alpha}{2} + \frac{\alpha}{2}\right) = 1 - \alpha,$$

es decir

$$\begin{aligned} & P\left(A_{\alpha/2} < \frac{\widehat{\Theta} - \theta}{\overline{\sigma}} < -A_{\alpha/2}\right) = 1 - \alpha \\ \Leftrightarrow \quad & P\left(A_{\alpha/2} \cdot \overline{\sigma} < \widehat{\Theta} - \theta < -A_{\alpha/2} \cdot \overline{\sigma}\right) = 1 - \alpha \\ \Leftrightarrow \quad & P\left(A_{\alpha/2} \cdot \overline{\sigma} - \widehat{\Theta} < -\theta < -A_{\alpha/2} \cdot \overline{\sigma} - \widehat{\Theta}\right) = 1 - \alpha \\ \Leftrightarrow \quad & P\left(\widehat{\Theta} + A_{\alpha/2} \cdot \overline{\sigma} < \theta < \widehat{\Theta} - A_{\alpha/2} \cdot \overline{\sigma}\right) = 1 - \alpha \end{aligned}$$

Por lo tanto para un valor $\widehat{\theta}$ *de* $\widehat{\Theta}$ *encontramos que hay una probabilidad de* $1 - \alpha$ *de que*

$$\theta \in \left]\widehat{\theta} + A_{\alpha/2} \cdot \overline{\sigma}, \widehat{\theta} - A_{\alpha/2} \cdot \overline{\sigma}\right[. \qquad \blacksquare$$

4.2 Estimación con una población

4.2.1 Intervalo de confianza para un promedio

Teorema 17 ***(IC para un promedio si el promedio muestral es normal y se conoce la varianza poblacional)*** *Considere la población dada por la variable aleatoria X que con media poblacional μ y varianza poblacional σ^2. Si $\overline{X}$ sigue una distribución normal para muestras de tamaño n y se conoce σ entonces*

1. *Un intervalo de confianza del $100\,(1-\alpha)\,\%$ para μ tiene extremos*

$$\overline{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

donde $\overline{x}$ es el valor de $\overline{X}$ para una muestra de tamaño n y $z_{\alpha/2}$ cumple que

$$\phi\left(-z_{\alpha/2}\right) = \frac{\alpha}{2}$$

2. *Para encontrar un intervalo de confianza del $100\,(1-\alpha)\,\%$ para μ con un radio menor o igual r, el tamaño de la muestra debe ser*

$$n \geq \left(\frac{z_{\alpha/2} \cdot \sigma}{r}\right)^2$$

Prueba. ***1.*** *Se tiene que*

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

Por el teorema 16 se tiene que un intervalo de confianza del $100\,(1-\alpha)\,\%$ para μ tiene extremos

$$\overline{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

2. *Para que el intervalo de confianza del $100\,(1-\alpha)\,\%$ para μ tenga un radio menor o igual r, se debe cumplir que*

$$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq r \quad \Longrightarrow \quad \sqrt{n} \geq \frac{z_{\alpha/2} \cdot \sigma}{r} \quad \Longrightarrow \quad n \geq \left(\frac{z_{\alpha/2} \cdot \sigma}{r}\right)^2. \quad \blacksquare$$

Observación. Para que $\overline{X}$ siga una distribución normal se debe cumplir una de las siguientes opciones:

1. La población dada por la variable aleatoria X sigua una distribución normal (Teorema: Promedio de normales es normal).

2. El tamaño de las muestras sea mayor a 30 (Resultado empírico por el Teorema del Límite central). En este caso, si no se conoce σ, se utiliza s que es una buena estimación de σ y se aplica el teorema anterior.

Ejemplo 24 *El peso de una bolsa de frijoles marca SABORES sigue una distribución normal de media desconocida y varianza* $0.04\ kg^2$. *Un inspector de la Oficina del Consumidor tomó una muestra de 20 bolsas y obtuvo un peso promedio de* $1.9\ kg$.

1. *Determine un IC del* 90% *para el peso promedio de la bolsa de frijoles SABORES.*

   Sea X *el peso de una bolsa de frijoles SABORES y considere el parámetro* μ *(peso promedio de una bolsa de frijoles SABORES). Se tiene que* $\sigma^2 = 0.04$, $n = 20$ *y la media muestral observada es* $\overline{x} = 1.9$. *Note que*

   $$X \sim N(\mu, 0.04) \Longrightarrow \overline{X} \sim N\left(\mu, \frac{0.04}{20}\right)$$

   donde μ *se desconoce. Así, un IC del* 90% *para* μ *tiene extremos*

   $$\overline{x} \pm z_{0.05} \cdot \frac{\sigma}{\sqrt{n}} = 1.9 \pm 1.645 \cdot \frac{0.02}{\sqrt{20}} \approx 1.9 \pm 7.356\,66 \times 10^{-3}$$

   Por lo tanto, el IC para μ *es*

   $$I =]1.892\,64, 1.907\,3[$$

2. *El empaque del producto asegura que el peso promedio de una bolsa de frijoles SABORES es de* $2kg$. *¿El inspector considera aceptable esta información?*

   Note que $2 \notin I$, *por lo tanto es poco probable (hay solo un* 10% *de probabilidad) que el peso promedio sea de* $2kg$. *Por lo tanto, el inspector debe considerar inapropiada la información dada a los consumidores.*

Ejemplo 25 *Se tiene interés en estimar la vida útil de un producto nuevo. ¿Qué tamaño de muestra mímimo debe tomarse para estimar la media con un error de estimación máximo igual a* $\frac{1}{10}$ *de desviación estándar con una confiabilidad de* 90%*?*

De acuerdo a la fórmula del tamaño de muestra, se tiene que esté debe ser:

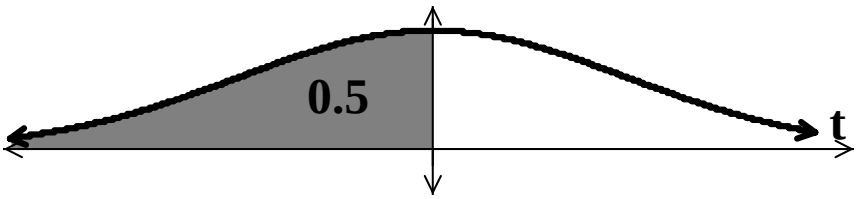
$$n \geq \left(\frac{z_{\alpha/2} \cdot \sigma}{r}\right)^2 = \left(\frac{z_{0.05} \cdot \sigma}{\frac{\sigma}{10}}\right)^2 = \left(10 \cdot z_{\alpha/2}\right)^2$$
$$= (10 \cdot -1.645)^2 = 270.603$$

El tamaño de muestra debe ser como mínimo 271.

Teorema 18 *Considere la población dada por la variable aleatoria* X *que sigue una distribución normal con media poblacional* μ. *Se tiene que*

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

tiene una distribución t con $v = n - 1$ grados de libertad, la cual es simétrica con respecto a su media que es 0.



Recuerde que $t_{\alpha,v}$ es el valor de T que acumula una área a la derecha de $\alpha\%$, es decir $P(T > t_{\alpha,v}) = \alpha$ y por la simetría de t : $P(T < -t_{\alpha,v}) = \alpha$.

Teorema 19 ***(IC para un promedio si la población es normal y se desconocen la varianza poblacional)*** *Considere la población dada por la variable aleatoria X que sigue una distribución normal con media poblacional μ. Si $\overline{X}$ sigue una distribución normal para muestras de tamaño n y se desconoce σ entonces un intervalo de confianza del $100(1-\alpha)\%$ para μ tiene extremos*

$$\overline{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$

donde $\overline{x}$ es el valor de $\overline{X}$ para una muestra de tamaño n y $t_{\alpha/2,n-1}$ es el valor de la distribución t de Student con $n-1$ grados de libertad.

Prueba. *Ejercicio. Sugerencia: utilice el teorema 16* ■

Ejemplo 26 *Una bebida afirma en su publicidad por televisión que su empleo diario, durante un mes, produce una pérdida promedio de cinco libras de peso. Para analizar esta afirmación, se toma un grupo control de ocho personas y se les suministro el producto diariamente por un mes obteniendo los siguientes datos:*

Peso Inicial (lb)	*165*	*195*	*188*	*170*	*185*	*163*	*155*	*177*
Peso Final (lb)	*164*	*190*	*187*	*163*	*185*	*159*	*148*	*174*

1. *Suponiendo que la pérdida de peso sigue una distribución normal, encuentre un intervalo de confianza de 95% para la pérdida de peso promedio.*

   Sea X : la pérdida de peso durante un mes en una persona que utiliza el producto en libras. La muestra observada de X es

Pérdida de peso (X):	*1*	*5*	*1*	*7*	*0*	*4*	*7*	*3*

   Por lo tanto,

$$\overline{x} = \frac{28}{8} = 3.5, \quad s \approx 2.7255, \quad t_{\alpha/2,n} = t_{0.025,7} = 2.36462$$

   Dado que $X \sim N$ entonces el IC buscado tiene extremos:

$$\overline{x} \pm t_{\alpha/2,7} \cdot \frac{s}{\sqrt{n}} = 3.5 \pm 2.36462 \cdot \frac{2.7255}{\sqrt{8}}$$

   Así el IC del 95% para la pérdida de peso promedio es:

$$]1.\,221\,43, 5.\,778\,57[\,.$$

2. *¿Los datos se aponen a la información hecha en publicidad?*

No, pues 5 *pertenece al IC.*

Ejercicio 11 *Las duraciones de ocho baterías cargadas de computadora de marca DUTEC son 151, 153, 175, 134, 170, 172, 156 y 114 minutos.*

1. *Suponiendo que las duraciones se distribuyen normalmente. Encuentre un intervalo de confianza de 90% para la duración promedio de las baterías.* *R/* $]139.19, 167.06[$
2. *Suponiendo que la desviación estándar de las duraciones es 20 min, ¿de qué tamaño debió ser una muestra para que el intervalo de confianza de 90% tuviera radio menor que 7.5 min?R/* 20

Ejercicio 12 *Una muestra aleatoria de diez estudiantes dio las siguientes cifras en horas para el tiempo que pasan estudiando durante la semana previa a los exámenes finales.*

$$28; 57; 42; 35; 61; 39; 55; 46; 49; 38.$$

Suponga que el tiempo de estudio durante la semana previa a los exámenes finales se distribuye normalmente.Calcule un intervalo de confianza para el tiempo medio con un nivel de confianza del 95%. *R/* $]37.459\,5, 52.540\,5[$

Ejercicio 13 *Recientemente el Ministerio de Seguridad a ampliado la cantidad de efectivos dedicados al combate de las drogas. Desde su ampliación han sido capturados 800 traficantes de droga de la ciudad. El valor promedio de las drogas decomisadas a estos 800 narcotraficantes es de 500000 dólares con una desviación estándar de 45000 dólares. Calcule un intervalo de confianza del* 90% *para el valor medio de los estupefacientes que están en manos de los narcotraficantes de la ciudad.* *R/* $]497297, 502703[$

4.2.2 Intervalo de confianza para una proporción

Teorema 20 *Sea p la proporción de determinados éxitos en una población. Dada la proporción muestral* $\widehat{P} = \dfrac{B}{n}$ *para muestras de tamaño n (estadístico asociado a la proporción poblacional p) donde B es la variable que indica el número de éxitos en la muestra, se tiene que*

1. *B sigue una distribución binomial:*
$$B \sim B(n, p)$$
2. $E\left(\widehat{P}\right) = p$ *es decir* $\widehat{P}$ *es insesgado.*
3. $Var\left(\widehat{P}\right) = \dfrac{pq}{n}$, *donde q es* $1 - p$.

Prueba. *Recuerde que* $\widehat{P} = \dfrac{B}{n}$ *es un estimador para el porcentaje p de éxitos en una población, donde B es el número de éxitos en muestras aleatorias de tamaño n, por lo tanto*

$$B \sim B(n, p)$$

y entonces

$$\begin{aligned} E\left(\widehat{P}\right) &= E\left(\frac{B}{n}\right) = \frac{1}{n}E\left(B\right) = \frac{1}{n}\cdot np = p. \\ Var\left(\widehat{P}\right) &= Var\left(\frac{B}{n}\right) = \frac{1}{n^2}Var\left(B\right) = \frac{1}{n^2}npq = \frac{pq}{n}. \end{aligned}$$ ■

Teorema 21 *Para n suficientemente grande se tiene que $\widehat{P}$ sigue una distribución Normal con media p y varianza $\frac{pq}{n}$:*

$$\widehat{P} \sim N\left(p, \frac{pq}{n}\right)$$

Prueba *Sea B la variable aleatoria tal que $\widehat{P} = \frac{B}{n}$. Por el teorema aproximación de la Binomial con la Normal se tiene que n suficientemente grande*

$$B \sim N\left(np, npq\right)$$

por lo tanto $\widehat{P} \sim N\left(p, \frac{pq}{n}\right)$. ■

Teorema 22 *Considere una población dada con una proporción poblacional p. Si $\widehat{P}$ sigue una distribución normal para muestras de tamaño n entonces*

$$P\left(\widehat{P} - z_{\alpha/2}\cdot\sqrt{\frac{pq}{n}} < p < \widehat{P} + z_{\alpha/2}\cdot\sqrt{\frac{pq}{n}}\right) = 1-\alpha$$

Es decir para un valor $\widehat{p}$ de $\widehat{P}$ para una muestra de tamaño n encontramos que hay una probabilidad de $1-\alpha$ de que

$$p \in \left]\widehat{p} - z_{\alpha/2}\cdot\sqrt{\frac{pq}{n}}, \widehat{p} + z_{\alpha/2}\cdot\sqrt{\frac{pq}{n}}\right[.$$

Prueba. *Por demostración del teorema 16.* ■

El teorema anterior, tiene el problema que para hallar un IC para p, se necesita conocer p, lo que es absurdo. Las siguientes observaciones permiten solucionar este problema.

Observación. Note que si n es suficientemente grande, entonces:

1. $\widehat{P}$ sigue una distribución normal, de acuerdo al teorema 21.

2. Una buena estimación de p y q es respectivamente $\widehat{p}$ y $\widehat{q}$.

¿Cuándo n es suficientemente grande? Para esto, se acepta empíricamente que n es grande si

$$n\widehat{p} \geq 5 \quad \wedge \quad n\widehat{q} \geq 5.$$

Combinando la observación y el teorema anterior se obtiene el siguiente resultado.

(Intervalo de confianza para una proporción con muestras grandes) Considere una población dada con una proporción poblacional p y el estadístico $\widehat{P}$ para muestras de tamaño n. Sea $\widehat{p}$ el valor de $\widehat{P}$ para una muestra de tamaño n Si $n\widehat{p} \geq 5$ y $n\widehat{q} \geq 5$ entonces un intervalo de confianza del $100\,(1-\alpha)\,\%$ para p tiene extremos

$$\widehat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\widehat{p}\widehat{q}}{n}}.$$

Ejemplo 27 *En una ciudad de 50 amas de casa, 18 no utilizan el detergente X. Determine un IC del 95% para la verdadera proporción de amas de casa que no utilizan el detergente X.*

$$\begin{gathered}
\textit{Sea } p \textit{ la proporción de amas de casa que no usan el detergente } X \\
\textit{Se tiene que } n = 50, \widehat{p} = \frac{18}{50}, \widehat{q} = \frac{32}{50}, \alpha = 0.05 \\
\textit{Como } n\widehat{p} = 18 \geq 5 \textit{ y } n\widehat{q} = 32 \geq 5 \textit{ el IC tiene extremos:} \\
\widehat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\widehat{p}\widehat{q}}{n}} = \frac{18}{50} \pm 1.96 \cdot \sqrt{\frac{18 \cdot 32}{50^3}} \\
\textit{Así el IC del } 95\% \textit{ para } p \textit{ es: }]0.226\,95, 0.493\,05[
\end{gathered}$$

Note que este IC es grande, pues la proporción de amas de casa que no utilizan el detergente podría ser pequena (cercana al 25%) ó muy alta (cercana al 50%). Para la compañia que elabora el detergente puede ser negativo difundir estos resultados. Se requiere un IC más preciso.

Teorema 23 *Considere una población dada con una proporción poblacional p. Si $\widehat{P}$ sigue una distribución normal para muestras de tamaño n, entonces para encontrar un intervalo de confianza del $100\,(1-\alpha)\,\%$ para p con un radio menor o igual r, el tamaño de la muestra debe ser*

$$n \geq \left(\frac{z_{\alpha/2} \cdot \sqrt{pq}}{r} \right)^2.$$

Prueba. *De acuerdo al teorema anterior, un IC del $100\,(1-\alpha)\,\%$ para p tiene radio $z_{\alpha/2} \cdot \sqrt{\frac{pq}{n}}$.*

Asi se quiere que

$$z_{\alpha/2} \cdot \sqrt{\frac{pq}{n}} \leq r \Longrightarrow \frac{z_{\alpha/2} \cdot \sqrt{pq}}{\sqrt{n}} \leq r \Longrightarrow \sqrt{n} \geq \frac{z_{\alpha/2} \cdot \sqrt{pq}}{r}.$$

Por lo tanto, $n \geq \left(\frac{z_{\alpha/2} \cdot \sqrt{pq}}{r} \right)^2$. ■

Observación. Para el tamaño de muestra, dado que p y q son desconocidos se puede:

1. Utilizar buenas estimaciones $\widehat{p}$ y $\widehat{q}$ de p y q obtenidas previamente para un valor previo de n (es decir, se debe cumplir que $n\widehat{p} \geq 5$ y $n\widehat{q} \geq 5$). Así, se utilizan estas estimaciones en el calculo del tamaño de muestra

$$n \geq \left(\frac{z_{\alpha/2} \cdot \sqrt{\widehat{p}\widehat{q}}}{r} \right)^2$$

Luego, para garantizar el uso de la normal estandar en el cálculo anterior, el mínimo valor de n, denotado por m, debe cumplir

$$m\widehat{p} \geq 5 \quad y \quad m\widehat{q} \geq 5.$$

Si m no cumple estas condiciones, se toma el tamaño de muestra n como aquel que satisface $n\widehat{p} \geq 5$ y $n\widehat{q} \geq 5$.

2. Dado que $pq = p - p^2$ es una parábola cóncava hacia abajo con vértice $(0.5, 0.25)$, se puede usar el hecho que $pq \leq 0.25$. Esto es muy útil si no se conoce una estimación p y q, ya que se toma como tamaño de muestra el n que cumple que

$$n \geq \left(\frac{z_{\alpha/2} \cdot \sqrt{0.25}}{r} \right)^2$$

El lector puede notar que esta condición es más exigente que la dada en el teorema, es decir

$$n \geq \left(\frac{z_{\alpha/2} \cdot \sqrt{0.25}}{r} \right)^2 \Longrightarrow n \geq \left(\frac{z_{\alpha/2} \cdot \sqrt{pq}}{r} \right)^2.$$

Ejemplo 28 *Considere el ejemplo anterior. Dado que el IC determinado es muy grande, ¿De qué tamaño debe ser la muestra si se desea tener una confianza de al menos el 95% de que el error estimado al estimar la proporción sea menor que 0.02, sin importar el verdadero valor de p?*

Dado que el valor de p no importa, se elige la segunda opción de la observación anterior para determinar el tamaño de muestra:

$$n \geq \left(\frac{z_{\alpha/2} \cdot \sqrt{0.25}}{r} \right)^2 = \left(\frac{1.96 \cdot \sqrt{0.25}}{0.02} \right)^2 = 2401$$

Note que para la compañia puede ser poco factible obtener el IC con lo solicitado, por la cantidad de amas de casa que se debe encuestar. Quizás una mejor solución sería aumentar poco a poco el tamaño máximo del radio hasta hallar un tamaño de muestra con el cuál la compañía vea factible trabajar.

Ejemplo 29 *Seguidamente se presenta una muestra de notas obtenidas en el examen de admisión 2010 de la Universidad Bienestar Seguro:*

$$72, 87, 28, 55, 92, 75, 83, 70, 30, 60, 53, 91, 90, 70, 70, 70, 55, 85$$

Con base en estos datos, el Rector de la universidad determinó de manera correcta un IC para el porcentaje de estudiantes que aprobaron el examen (Se aprueba con 70), obteniendo

$$IC :]0.448\,889\,, 0.884\,444\,[\,.$$

1. *Determine aproximadamente el nivel de confianza del IC que halló el Rector.*

   Sea p la proporción de estudiantes que aprobaron el examen. Note que de las 18 *notas observadas, solo* 12 *aprobaron el examen por lo tanto*

$$n = 18, \quad \widehat{p} = \frac{12}{18}, \quad \widehat{q} = \frac{6}{18}$$

Como $n\widehat{p} \geq 5$ *y* $n\widehat{q} \geq 5$ *el IC tiene extremos:*

$$\widehat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\widehat{p}\widehat{q}}{n}} = \frac{12}{18} \pm z_{\alpha/2} \cdot \sqrt{\frac{12 \cdot 6}{18^3}}$$

Dado que $z_{\alpha/2}$ *es negativo entonces el extremo derecho del IC es*

$$\frac{12}{18} - z_{\alpha/2} \cdot \sqrt{\frac{12 \cdot 6}{18^3}} = 0.884\,444$$

Despejando $z_{\alpha/2}$ *se obtiene que*

$$z_{\alpha/2} = -1.96 \Longrightarrow \alpha/2 = 0.025 \Longrightarrow \alpha = 0.05.$$

Por lo tanto, el IC es del 95%.

2. *El IC que halló el rector es muy grande. ¿De qué tamaño debe ser la muestra si se desea tener una confianza del 90% para estimar el porcentaje de estudiantes que aprobaron el examen con un error de estimación menor que 0.05?*

   Opción #1*. Dado que* $\widehat{p} = \dfrac{12}{18}$ *y* $\widehat{q} = \dfrac{6}{18}$ *son buenas estimaciones de* p *y* q *respectivamente, entonces el tamaño de muestra debe cumplir*

$$n \geq \left(\frac{z_{\alpha/2} \cdot \sqrt{\widehat{p}\widehat{q}}}{r} \right)^2 = \left(\frac{1.645 \cdot \sqrt{\dfrac{12 \cdot 6}{18^2}}}{0.05} \right)^2 = 240.536$$

   Así, se toma $n = 241$*. Este valor de* n *cumple que* $n\widehat{p} \approx 160.667 \geq 5$ *y* $n\widehat{q} \approx 80.333 \geq 5$.

   Opción #2*. Si importar el valor de* p *y* q*, el tamaño de muestra debe cumplir*

$$n \geq \left(\frac{z_{\alpha/2} \cdot \sqrt{0.25}}{r} \right)^2 = \left(\frac{1.645 \cdot \sqrt{0.25}}{0.05} \right)^2 = 270.603$$

   Así, se toma $n = 271$*. Note que este valor es mucho mayor al obtenido en la opción* 1*, esto se debe a que la opción* 2 *pone una mayor condición al* n *de la requerida, pero evita utilizar valores aproximados de* p *y* q*.*

Ejercicio 14 *Una muestra aleatoria de 50 estudiantes del Tec da los siguientes resultados:*

	Mujeres	*Hombres*
De San José:	*12*	*9*
De otras provincias:	*10*	*19*

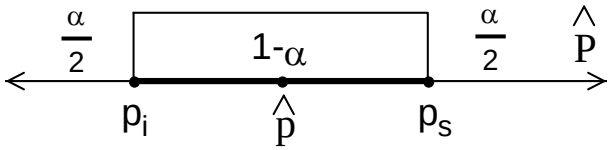
Encuentre un intervalo de confianza de 95% para la proporción de estudiantes de otras provincias.
R/ $]0.443\,19, 0.716\,81[$

¿Qué sucede si las muestras son pequeñas? En este caso se debe recordar que

a) $B = n\widehat{P} \sim B(n,p)$ y por lo tanto

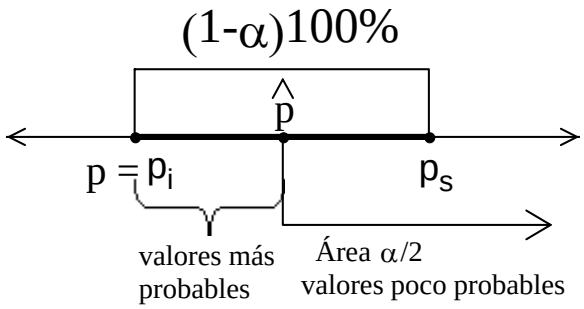
$$P(B=k) = C(n,k)\, p^k (1-p)^{n-k}.$$

b) Si el IC de $(1-\alpha)\,100\%$ para p se obtiene de un intervalo de la forma $]p_i, p_s[$ que depende de $\widehat{P}$ y cumple

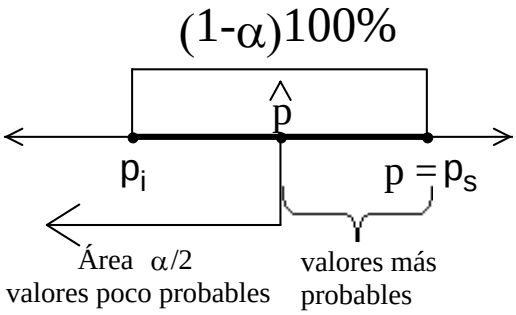


Sin embargo, para determinar el IC se requiere una nueva interpretación de él: los valores más probables de $\widehat{P}$ deben estar lo más cerca posible a p en el caso en que p tome como valor los extremos del IC. Es decir, si p toma valores extremos, la probabilidad de $(1-\alpha)\,100\%$ de que $p \in]p_i, p_s[$ debe concentrarse hacia ese extremo. De acuerdo a esto, se tiene dos casos:

1. Si p es el valor límite $p = p_i$ entonces la probabilidad de que el estadístico $\widehat{P}$ tome un valor superior a una estimación $\widehat{p}$ es de $\alpha/2$.



2. Si p es el valor límite $p = p_s$ entonces la probabilidad de que el estadístico $\widehat{P}$ tome un valor inferior a una estimación $\widehat{p}$ es de $\alpha/2$.



Por lo tanto, para hallar el IC se hallan empíricamente los valores p_i y p_s que satisfacen que

$$P\left(\widehat{P} > \widehat{p} \mid p = p_i\right) = \frac{\alpha}{2} \quad y \quad P\left(\widehat{P} < \widehat{p} \mid p = p_s\right) = \frac{\alpha}{2}$$

Lo cual es equivalente a

$$P\left(B \leq n\widehat{p} \mid p = p_i\right) = 1 - \frac{\alpha}{2} \quad y \quad P\left(B < n\widehat{p} \mid p = p_s\right) = \frac{\alpha}{2}$$

Estas probabilidades se pueden expresar en términos de p_i y p_s respectivamente, utilizando la distribución binomial. Se recomienda utilizar un software para darle valores a p_i y p_s hasta obtener aproximadamente las igualdades.

Ejemplo 30 *En una muestra de 50 personas de un barrio, 4 padecen de asma. Determine un IC del 90% para la proporción de personas del barrio que padecen de asma.*

En este caso se tiene que $n = 50$, $\widehat{p} = \dfrac{4}{50}$ *y* $\alpha = 0.1$. *Como* $n\widehat{p} = 4 < 5$ *entonces la muestra nos grande. Por lo tanto, se busca los valores de* p_i *y* p_s *que cumplen que*

$$P\left(B \leq 4 \mid p = p_i\right) = 0.95 \quad y \quad P\left(B < 4 \mid p = p_s\right) = 0.05$$

donde $B \sim B\left(50, p\right)$. *Estas ecuaciones son equivalentes a*

$$\sum_{k=0}^{4} C\left(50, k\right) p_i^k \left(1 - p_i\right)^{50-k} = 0.95 \quad y \quad \sum_{k=0}^{3} C\left(50, k\right) p_s^k \left(1 - p_s\right)^{50-k} = 0.05$$

Así, se pueden dar valores a p_i *y* p_s *con una calculadora hasta obtener aproximadamente las igualdades. Sin embargo, este proceso puede ser lento. Una mejor opción es utilizar Excel. Sea A igual a la suma de la primera ecuación, entonces digitando lo siguiente en Excel*

	A	B
1		
2	pi	A
3	0,2	=DISTR.BINOM(4;50;A3;VERDADERO)
4		

basta alterar el valor de la celda A3 hasta que el valor de A sea de aproximadamente 0.95, *realizando esto se obtiene que* $p_i \approx 0.0402$. *Similarmente se obtiene que el valor de* p_s *es* $p_s \approx 0.1475$. *Por lo tanto, el IC buscado es*

$$]0.0402, 0.1475[\,.$$

Ejercicio 15 *En una muestra de 40 niños de 5 años de una localidad se encontró que 3 saben nadar.*

1. *Plantee las ecuaciones que permiten determinar un IC del 80% para la proporción de niños nadadores de 5 anos de la localidad.*

2. *Utilice Excel para determinar un IC del 80% para la proporción de niños nadadores.*

   R/ $]0.0495, 0.1535[$

4.2.3 Intervalo de confianza para una varianza

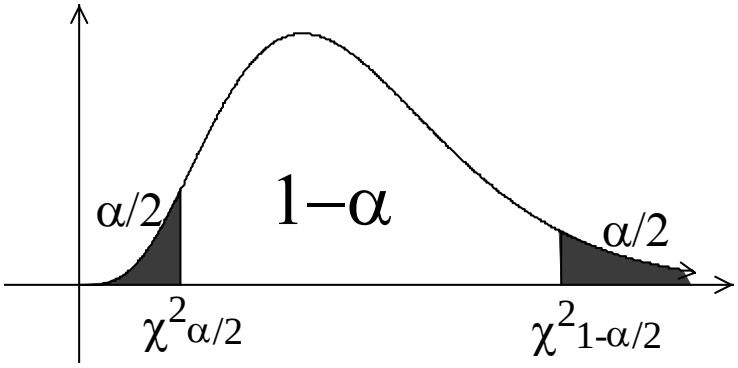
Recuerde el siguiente teorema.

Teorema 24 *Considere la población dada por la variable aleatoria X que sigue una distribución normal con varianza poblacional σ^2. Dada una muestra aleatoria de esta población $(X_1, X_2, X_3, ..., X_n)$. Entonces la variable*

$$\chi^2 = \frac{(n-1)\,S^2}{\sigma^2}$$

sigue una distribución χ^2 con $v = n-1$ grados de libertad.

Recuerde que $\chi^2_{\alpha/2,n-1}$ deja una área de $\frac{\alpha}{2}$ a la izquierda y $\chi^2_{1-\alpha/2,n-1}$ deja una área de $\frac{\alpha}{2}$ a la derecha:



entonces

$$P\left(\chi^2_{\alpha/2,n-1} < \chi^2 = \frac{(n-1)\,S^2}{\sigma^2} < \chi^2_{1-\alpha/2,n-1}\right) = 1-\alpha$$

Por lo tanto,

$$P\left(\frac{(n-1)\,S^2}{\chi^2_{1-\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)\,S^2}{\chi^2_{\alpha/2,n-1}}\right) = 1-\alpha$$

Así, de acuerdo a la definición de IC, dado u n valor s de S entonces un IC del $100\,(1-\alpha)\,\%$ para σ^2 es $\left]\dfrac{(n-1)\,s^2}{\chi^2_{1-\alpha/2,n-1}} \quad , \quad \dfrac{(n-1)\,s^2}{\chi^2_{\alpha/2,n-1}}\right[$.

Teorema 25 *Considere la población dada por la variable aleatoria X que sigue una distribución normal con varianza poblacional σ^2. Dada una muestra aleatoria de esta población $(X_1, X_2, X_3, ..., X_n)$ con variancia s^2. Un intervalo de confianza del $100\,(1-\alpha)\,\%$ para σ^2 es*

$$\left]\frac{(n-1)\,s^2}{\chi^2_{1-\alpha/2,n-1}} \quad , \quad \frac{(n-1)\,s^2}{\chi^2_{\alpha/2,n-1}}\right[$$

Ejemplo 31 *Un médico afirma que los pesos de niñas de 3 años de cierta provincia son muy similares. Para analizar esta afirmación se toma el peso en kg de 10 niñas:*

$$14.5, 11.6, 12.8, 15.1, 14.2, 13.7, 12.9, 13.8, 14.1, 11.9.$$

Suponga que el peso de niñas de 3 años sigue una distribución normal. Determine un IC del 95% para la desviación estandar de los pesos de niñas de 3 años.

Sea σ la desviación estandar de los pesos de niñas de 3 años. Se tiene que

$$n = 10, \overline{x} = 13.46, s^2 = 1.282666667, \alpha = 0.05$$

Además,

$$\chi^2_{1-\alpha/2,n-1} = \chi^2_{0.975,9} = 19.0228 \quad y \quad \chi^2_{\alpha/2,n-1} = \chi^2_{0.025,9} = 2.70039$$

Con estos datos, un IC del 95% para σ^2 es

$$\left]\frac{9 \cdot 1.282666667}{19.0228}, \frac{9 \cdot 1.282666667}{2.70039}\right[=]0.606\,851\,, 4.\,274\,94[$$

Por lo tanto, sacando raiz a cada extremo del intervalo, un IC del 95% para σ es

$$= \left]\sqrt{0.606\,851}, \sqrt{4.274\,94}\right[=]0.779\,006\,, 2.\,067\,59[$$

En este caso, se puede observar que hay una alta probabilidad de que la desviación estandar sea menor a 2.07kg, la cual se puede considerar baja en comparación a otros grupos de edad.

Ejercicio 16 *Las duraciones de ocho baterías cargadas de computadora de marca DUTEC son 151, 153, 175, 134, 170, 172, 156 y 114 minutos. Suponiendo que las duraciones se distribuyen normalmente, encuentre un intervalo de confianza de 90% para la desviación estándar poblacional*

R/ $]14.6727, 37.38[$

4.2.4 Ejercicios

1. Seguidamente se presenta una muestra de notas obtenidas por estudiantes de Ingeniería en Computación del ITCR, en el curso Estadística II semestre del 2009:

75	75	75	85	70	70	80	85	75	60	70	60
80	60	80	75	75	70	70	50	80	75	70	

   Suponga que la nota de Estadística sigue una Distribución Normal.

   (a) Encuentre un IC del 95% para la nota promedio del curso de estadística en el 2009. *R/* $]68.\,647\,2, 76.\,135\,4[$

   (b) Encuentre un IC del 90% para el porcentaje de estudiantes del curso de estadística en el 2009.que aprobaron el curso con una nota mayor igual a 80. *R/* $]0.110\,252\,, 0.411\,487[$

   (c) Un profesor afirma que menos de la mitad de los estudiantes que llevaron estadística en el 2009 tuvieron una nota inferior a 80. ¿Apoya esta afirmación? Justifique. *R/* Si

2. Un IC del 90% para la estatura promedio de los niños de 5 años de cierto distrito es $]90, 125[$ en centímetros. Responda a las siguientes afirmaciones con VERDADERO o FALSO. Justifique su respuesta.. (6 puntos)

   (a) 90% de las estaturas de los niños de 5 años de la ciudad están entre 90 cm y 125 cm. *R/ Falso*

   (b) La probabilidad de que el intervalo incluya la estatura promedio muestral es del 90%. *R/ Falso*

   (c) Hay un 90% de probabilidad de que el intervalo incluya la estatura promedio de los niños de 5 años de la ciudad. *R/ Verdadero*

   (d) Si con los datos muestrales utilizados en este IC se calcula un IC con una confianza mayor, este será de menor tamaño. *R/ Falso*

3. Las duraciones de ocho baterías de computadora son 151, 153, 175, 134, 170, 172, 156 y 114 minutos. Suponga que las duraciones se distribuyen normalmente. Miguel, un estudiante de estadística determinó correctamente un IC para la desviación estándar poblacional, obteniendo

   $$IC :]14.6727, 37.38[$$

   Determine el nivel de confianza del IC determinado por Miguel. *R/* $\alpha = 0.1$, el nivel de confianza es del 90%

4. Seguidamente se presenta una muestra de notas obtenidas por de Ingeniería en Computación del ITCR, carné 2004, en el curso Matemática Discreta:

75	50	35	45	85	85	85	70	30
60	85	85	90	70	70	70	55	85

   Suponga que la nota de Matemática Discreta para estudiantes carné 2004 sigue una Distribución Normal. Encuentre un IC del 95% para la nota promedio de Matemática Discreta para estudiantes carné 2004. *R/* $]59.0691, 77.5976[$

5. Sea E el margen de error en un intervalo de confianza (también llamado radio del intervalo) para la media de una variable aleatoria X con desviación estándar σ, basado en una muestra de tamaño n≥30. Si se triplica el tamaño de la muestra manteniendo el nivel de confianza, ¿cuál es la razón del nuevo margen de error con respecto a E? ¿Reduce o aumenta el nuevo error? *R/* $\frac{E_{nuevo}}{E_{ant}} = \frac{1}{\sqrt{3}}$, se reduce el error.

6. En una ciudad de 125 apartamentos observados, 102 no permiten que los inquilinos tengan mascotas. Determine un IC del 90% para la verdadera proporción de apartamentos de la ciudad que prohiben mascotas. *R/* $]0.758\,988, 0.873\,012[$

7. Al estimar una proporción p se toma una muestra de tamaño 50, donde $n\widehat{p} \geq 5$ y $n\widehat{q} \geq 5$, y se obtiene un IC del 95% con un radio de 0.133. Determine los posibles valores de $\widehat{p}$. *R/* 0.640 607 o 0.359 393

8. En una ciudad, se entrevistaron a 598 personas de 18 años sin hijos obteniendo que 366 de estas personas afirman que desean tener hijos. Determine un IC del 90% para la verdadera proporción de personas de 18 anos sin hijos, de la ciudad, que quieren tener hijos.$R/$ $]0.579\,261, 0.644\,819\,[$

9. Considere una población X tal que
$$X \sim N\left(\mu, \sigma^2\right)$$
Suponga que se desconoce la varianza poblacional σ^2. Un IC para μ de 95% basado en muestras de tamaño 16 es $]42.7, 49.3\,[$. Determine la varianza muestral. $R/$ $s^2 = 38.\,352\,9$

10. En el restaurante Buen Gusto, un cliente se quejó del excesivo tiempo que se tarda en servir los alimentos una vez que se ordena. El gerente ha asegurado que la duración promedio de servir los alimentos después de ordenar es inferior a 7 minutos. Se han registrado diez duraciones en minutos: $2, 3, 6, 4, 7, 3, 8, 9, 5, 4$.

    (a) Suponiendo que el tiempo que se tarda en servir los alimentos una vez que se ordena se distribuyen normalmente, encuentre un intervalo de confianza de 90% para la duración promedio de servir los alimentos después de ordenar. $R/$ $]3.\,748\,88, 6.\,451\,12[$

    (b) ¿Considera aceptable la afirmación del gerente? Justifique $R/$ Si

11. Una persona denuncia ante la Oficina del Consumidor que el peso de una bolsa de arroz Blanco, con peso marcado de 2 kg, suele ser muy variable. Ante este un Inspector de la Oficina del Consumidor desea determinar si la desviación estándar del peso de estas bolsas es superior a $50g$, si esto sucede sancionará a la empresa Blanco S.A. Así, se toma una muestra al azar de 10 bolsas de arroz Blanco, obteniendo los siguientes peso en kg:
$$1.98, 2.15, 1.95, 1.90, 1.82, 1.95, 2.11, 1.88, 1.92, 2.06$$
Suponga que el peso de una bolsa de arroz Blanco sigue una distribución normal.

    (a) Determine un IC del 95% para la desviación estándar de los pesos de una bolsa de arroz blanco. $R/$ $]0.072\,126\,1\,, 0.191\,433[$

    (b) ¿Aceptaría que la empresa Blanco reciba una sanción de parte de la Oficina del Consumidor? Justifique su respuesta. $R/$ Si

12. Si $]40.7, 51.3[$ es el intervalo de confianza del 95 % para la media de una variable aleatoria normalmente distribuida con variancia desconocida, basado en una muestra de tamaño 16, halle el valor de la variancia muestral. $R/$ $81.\,143\,2$

13. Una fábrica de pantalones investiga la preferencia de su marca en dos ciudades. En la ciudad A se encuentra que 60 de 300 prefieren esa marca de pantalones, y en la ciudad B son 38 de 250 los que la prefieren.

    (a) Encuentre un IC del 95% para la diferencia de proporciones de preferencia entre las dos ciudades. $R/$ $]-0.015\,478\,5, 0.111\,479[$

    (b) ¿Se puede suponer que la proporción es mayor en la primer ciudad? $R/$ No

(c) ¿De qué tamaño deben ser las muestras para que el IC de 95% tenga un radio menor a 0.05? $R/$ 444

14. Se toma una muestra aleatoria de 50 cascos utilizados por los corredores de motocicleta y se someten a una prueba de impacto. Se observó que 18 de los casos no resisten la prueba pues presentan cierto daño.

    (a) Encuentre un IC del 95% para la verdadera proporción de cascos que no resisten la prueba de impacto. $R/$ $]0.226\,95, 0.493\,05[$

    (b) ¿Considera que es positivo para la compañia que se difunda este IC? $R/$ No

    (c) ¿Será necesario tomar una muestra más grande? $R/$ Si

    (d) ¿De qué tamaño debe ser la muestra si se desea tener una confianza de al menos el 95% de que el error estimado al estimar la proporción sea menor que 0.02, sin importar el verdadero valor de p? $R/$ 2401

    (e) ¿Es factible obtener el IC con el tamaño encontrado? $R/$ No

15. En el 2010, un grupo de estudiantes del ITCR se propuso analizar la cantidad de horas que duerme un estudiante de Ingeniería en Computación durante en el tiempo lectivo. Se tomó una muestra de 71 estudiantes de Ingeniería en Computación. El promedio observado de horas que duerme cada uno en promedio es de 5.17 horas como una desviación de 1.88 horas. Estos datos son basados en información brindada por los mismos estudiantes, no se realizó una observación directa.

    (a) Determine un IC del 90% para promedio de horas que duermen en promedio los estudiantes de Ingeniería en Computación en tiempo lectivo. $R/$ $]4.803144, 5.536856[$

    (b) Se dice una persona duerme una cantidad saludable de horas, si duerme en promedio de 6 a 8 horas por la noche. ¿Considera que los estudiantes de Ingeniería en Computación duerme una cantidad saludable de horas, durante en el tiempo lectivo? Justifique.$R/$ No

16. Sea E el radio o margen de error en un intervalo de confianza para la media μ de una variable aleatoria X con desviación estándar conocida σ, basado en una muestra de tamaño $2n \geq 30$. Si se quiere disminuir el error en dos tercios utilizando el mismo nivel de confianza, ¿Cuántas veces se debe aumentar el tamaño de la muestra? $R/$ 9 veces

17. Al estimar una proporción p se toma una muestra de tamaño n y se obtiene una proporción observada $\widehat{p}$. Si $n\widehat{p} \geq 5$, $n\left(1-\widehat{p}\right) \geq 5$, y se halla un IC del 90% para p, igual a

$$]0.143784, 0.30692[$$

    (a) Halle aproximadamente el valor de $\widehat{p}$. $R/$ 0.225 352

    (b) Determine aproximadamente el tamaño de la muestra utilizada para hallar el IC.$R/$ $n \approx 71$

18. Considere una población X tal que
$$X \sim N\left(\mu, \sigma^2\right)$$
Suponga que se desconoce la varianza poblacional σ^2. Un IC para μ basado en muestras de tamaño 20 es $]26.514\,1\,, 33.48594[$ en unidades u, donde la varianza muestral observada en la muestra es de 50 u^2. Determine aproximadamente el nivel de confianza del IC. $R/$ Confianza: 96%

4.3 Estimación con dos poblaciones

4.3.1 Intervalo de confianza para una diferencia entre dos promedios

Teorema 26 ***(IC para la diferencia de promedios si los promedios son normales y se conocen las varianzas poblacionales)*** *Considere la población X_1 con media poblacional μ_1 y varianza poblacional σ_1^2. Considere la población X_2 con media poblacional μ_2 y varianza poblacional σ_2^2. Si $\overline{X}_1$ y $\overline{X}_2$ siguen una distribución normal para muestras de tamaño n_1 y n_2 respectivamente, y se conocen σ_1 y σ_2 entonces*

1. *Un intervalo de confianza del $100\,(1-\alpha)\,\%$ para $\mu_1 - \mu_2$ tiene extremos*
$$(\overline{x}_1 - \overline{x}_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$
donde $\overline{x}_1$ es el valor de $\overline{X}_1$ para una muestra de tamaño $n_1, \overline{x}_2$ es el valor de $\overline{X}_2$ para una muestra de tamaño n_2 y $z_{\alpha/2}$ cumple que
$$\phi\left(-z_{\alpha/2}\right) = \frac{\alpha}{2}$$

2. *Para encontrar un intervalo de confianza del $100\,(1-\alpha)\,\%$ para $\mu_1 - \mu_2$ con un radio menor o igual r, el tamaño de las muestras de cada población debe ser*
$$n \geq \left(\frac{z_{\alpha/2} \cdot \sqrt{\sigma_1^2 + \sigma_2^2}}{r}\right)^2$$

Prueba. ***1.*** *Se tiene que*
$$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$
Sea $Z = \dfrac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$, *entonces por el teorema 16 se tiene que un intervalo de confianza del $100\,(1-\alpha)\,\%$ para μ tiene extremos*
$$(\overline{x}_1 - \overline{x}_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

2. *Sea n el tamaño de ambas muestras, para que el intervalo de confianza del $100\,(1-\alpha)\,\%$ para $\mu_1-\mu_2$ tenga un radio menor o igual r, se debe cumplir que*

$$z_{\alpha/2}\cdot\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{n}}\leq r \implies \sqrt{n}\geq\frac{z_{\alpha/2}\cdot\sqrt{\sigma_1^2+\sigma_2^2}}{r}$$

$$\implies n\geq\left(\frac{z_{\alpha/2}\cdot\sqrt{\sigma_1^2+\sigma_2^2}}{r}\right)^2 \quad \blacksquare$$

Observación. Recuerde que para que $\overline{X_1},\overline{X}_2$ se distribuyen normalmente, se debe cumplir una de las siguientes opciones para cada una de estos estadísticos:

1. Su poblaciones siguen una distribución normal (Teorema: Promedio de normales es normal)

2. El tamaño de la muestra sea mayor a 30 (Resultado empírico por el Teorema del Límite central). En este caso se puede utilizar la desviación estandar muestral como una buena aproximación de la desviación estandar poblacional.

Por ejemplo, si la población $X_1\sim N$ y $n_2\geq 30$ entonces $\overline{X_1},\overline{X}_2$ se distribuyen normalmente.

Ejemplo 32 *Una universidad analizó las capacidades de rendimiento de 2 tipos de variedades de cierto producto agrícola, así obtuvo los siguientes datos utilizando parcelas de igual tamaño:*

Variedad	*Tamaño de muestra*	*Rendimiento promedio (Kg por parcela)*	*Desviación muestral*
A	32	43	15
B	30	47	22

1. *Determine un IC del 90% para la diferencia entre el rendimiento promedio de las variedades A y B.*

   Note que

   $$n_1=32, n_2=30, \overline{x}_1=43, \overline{x}_2=47, s_1=15, s_2=22, \alpha=0.1.$$

   Como n_1 y n_2 son mayores iguales a 30 entonces, $\overline{X_1},\overline{X}_2$ se distribuyen normalmente y el IC tiene extremos

   $$(43-47)\pm 1.645\cdot\sqrt{\frac{15^2}{32}+\frac{22^2}{30}}=-4\pm 7.917\,32.$$

   Por lo tanto, el IC solicitado es

   $$I=\,]-11.917\,3, 3.917\,32[\,.$$

2. *El ministro de agricultura afirma que la variedad 2 tiene mejor rendimiento. ¿Aceptaría esta afirmación?*

   Note que I tiene valores positivos y negativos, pues $0\in I$. Por lo tanto, no se considera aceptable la afirmación del ministro.

Ejemplo 33 *Un IC de* 90% *para la diferencia de promedios* $(\mu_1 - \mu_2)$ *es* $]165.5, 192.9[$. *Si las muestras utilizadas en el cálculo del IC son ambas d tamaño* 50, *determine el valor de* $\overline{x}_1 - \overline{x}_2$ *y de* $\sigma_1^2 + \sigma_2^2$.

Dado que $n_1 = n_2 = 50$ *entonces el IC tiene extremos*

$$(\overline{x}_1 - \overline{x}_2) \pm z_{0.05} \cdot \sqrt{\frac{\sigma_1^2}{50} + \frac{\sigma_2^2}{50}}.$$

Por lo tanto, el centro del IC es

$$\overline{x}_1 - \overline{x}_2 = \frac{165.5 + 192.9}{2} = 179.2$$

y el radio es

$$\left| z_{0.05} \cdot \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{50}} \right| = \frac{192.9 - 165.5}{2} = 13.7$$

de donde

$$1.645\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{50}} = 13.7 \Longrightarrow \sigma_1^2 + \sigma_2^2 \approx 3468.0019$$

Por lo tanto $\overline{x}_1 - \overline{x}_2 = 179.2$ *y* $\sigma_1^2 + \sigma_2^2 \approx 3468.0019$.

Otros resultados son:

Teorema 27 ***(IC para la diferencia de promedios si las poblaciones son normales, se desconocen las varianzas poblacionales y se suponen iguales)****Considere la población* 1 *dada por la variable aleatoria* X_1 *con media poblacional* μ_1 *y la población* 2 *dada por la variable aleatoria* X_2 *con media poblacional* μ_2. *Si las poblaciones* X_1 *y* X_2 *siguen una distribución normal y se desconocen las desviaciones poblacionales* σ_1 *y* σ_2 *pero se suponen iguales, entonces un intervalo de confianza del* $100\,(1-\alpha)\,\%$ *para* $\mu_1 - \mu_2$ *tiene extremos*

$$(\overline{x}_1 - \overline{x}_2) \pm t_{\alpha/2,v} \cdot \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

donde $\overline{x}_1, s_1$ *son valores de* $\overline{X}_1$ *y* S_1 *para una muestra de tamaño* n_1
$\overline{x}_2, s_2$ *son valores de* $\overline{X}_2$ *y* S_2 *para una muestra de tamaño* n_2
$s_p^2 = \dfrac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$
$v = n_1 + n_2 - 2$
$t_{\alpha/2,v}$ *es el valor de la distribución* t *de Student con* v *grados de libertad*

Teorema 28 ***(IC para la diferencia de promedios si las poblaciones son normales, se desconocen las varianzas poblacionales y no se suponen iguales)****Considere la población* 1 *dada por la variable aleatoria* X_1 *con media poblacional* μ_1 *y la población* 2 *dada por la variable aleatoria* X_2 *con media poblacional* μ_2. *Si las poblaciones* X_1 *y* X_2 *siguen una distribución normal y*

se desconocen las desviaciones poblacionales σ_1 y σ_2 pero no se suponen iguales, entonces un intervalo de confianza del $100\,(1-\alpha)\,\%$ para $\mu_1 - \mu_2$ tiene extremos

$$(\overline{x}_1 - \overline{x}_2) \pm t_{\alpha/2,v} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

donde $\overline{x}_1, s_1$ *son valores de* $\overline{X}_1$ *y* S_1 *para una muestra de tamaño* n_1
$\overline{x}_2, s_2$ *son valores de* $\overline{X}_2$ *y* S_2 *para una muestra de tamaño* n_2

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}$$

$t_{\alpha/2,v}$ *es el valor de la distribución* t *de Student con* v *grados de libertad*

Ejemplo 34 *Una ferretería tiene dos marcas de pintura color verde para portones: A y B. El vendedor asegura que la pintura B es más cara por que seca más rápidamente que la pintura A. Se obtienen mediciones de ambos tipos de pintura. Los tiempos de secado(en minutos) son los siguientes:*

Pintura A:	*120*	*132*	*123*	*122*	*140*	*110*	*120*	*107*		
Pintura B:	*126*	*124*	*116*	*125*	*109*	*130*	*125*	*117*	*129*	*120*

1. *Encuentre un intervalo de confianza del 95% para la diferencia entre los tiempos de secado promedio, suponiendo que las desviaciones estándar de éstos son iguales y que el tiempo de secado de ambas pinturas está distribuido de manera normal.*

   Sean

   X_1 : *tiempo de secado de la pintura A*
   X_2 : *tiempo de secado de la pintura B*

   Note que

   $$n_1 = 8, n_2 = 10, \overline{x}_1 = 121.75, s_1 = 10.7004, \quad \overline{x}_2 = 122.1, s_1 = 6.53962$$

   Note que si bien $\overline{X}_1, \overline{X}_2 \sim N$ *pues las poblaciones son normales, no se concen los valores de* σ_1 *y* σ_2*, además estos no se pueden aproximar por* s_1 *y* s_2 *pues las muestras son pequeñas. Sin embargo, como* σ_1 *y* σ_2 *se suponen iguales, se procede de acuerdo al teorema 27. Note que*

   $$s_p^2 = \frac{7s_1^2 + 9s_2^2}{16} = 74.\,149\,3$$

   y el IC tiene extremos

   $$(\overline{x}_1 - \overline{x}_2) \pm t_{\alpha/2,v} \cdot \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = (121.75 - 122.1) \pm 2.11991 \cdot \sqrt{\frac{(74.\,149\,3)}{8} + \frac{(74.\,149\,3)}{10}}$$

   Así el IC del 95% *es*

   $$]-9.\,008\,89, 8.\,308\,89[\,.$$

2. *¿Existe alguna evidencia que respalde la afirmación del vendedor?*

   $$No,\ el\ IC\ tiene\ valores\ positivos\ y\ negativos.$$

   Note que hay un 95% de probabilidad de que $\mu_1 - \mu_2 \in\ ]-9.\,008\,89, 8.\,308\,89[$. *Si el IC fuera positivo entonces es muy probable que* $\mu_1 - \mu_2 > 0$ *y entonces* $\mu_1 > \mu_2$, *así se tendría que la pintura A secaría más rápido que la B, en promedio. Similarmente si el IC fuera negativo ,la pintura B secaría más rápido que la A, en promedio, lo cual respaldaría la afirmación del vendedor. Sin embargo, el IC determinado contiene al cero, por lo que no se puede inferir ningún resultado.*

Por otro lado, es importante diferenciar entre un IC para el promedio de una diferencia y un IC para la diferencia de promedios. El primer caso se refiere a un IC para un promedio visto en la sección (4.2.1) , en este caso las variables involucradas se funden en una sola variable, la variable diferencia, de la cuál interesa estimar su promedio, y para la muestra se debe realizar observaciones pareadas (es decir para cada individuo elegido en la muestra se determina el valor de las variables involucradas con el fin de obtener un valor de la variable diferencia).

El segundo caso, se trata del IC estudia en este apartado, en el cuál las poblaciones son independientes por lo que no es necesario que las observaciones sean pareadas.

Ejercicio 17 *Las notas obtenidas por 10 estudiantes de Computación del Instituto Tecnológico de Costa Rica en el año 2003-2004 en los cursos de Probabilidad y Estadística son:*

Probabilidad:	70	30	30	73	60	75	65	60	40	55
Estadística:	100	85	90	95	90	95	85	85	80	90

Suponiendo que la distribución de notas es normal:

1. *Encuentre un IC del 95% para el promedio de diferencias entre la nota de probabilidad y de Estadística.* *R/ Sea* $d = Nota_{est} - Nota_{prob}$, *el IC para d es* $]23.577, 43.822[$.
2. *Encuentre un IC del 95% para la diferencia entre la nota media en probabilidad y la nota media en Estadística.* *R/ El IC para* $\mu_{est} - \mu_{prob}$: $]21.228, 46.172[$
3. *¿Aceptaría el hecho de que el rendimiento promedio en probabilidad es menor que el rendimiento promedio en estadística?* *R/* Si

4.3.2 Intervalo de confianza para una diferencia entre dos proporciones

Teorema 29 *Dadas dos poblaciones tales que*

Población	*Proporción poblacional*	*Tamaño de muestra*	*Proporción muestral*
X_1	p_1	n_1	$\widehat{P}_1 = \dfrac{B_1}{n_1}$
X_2	p_2	n_2	$\widehat{P}_2 = \dfrac{B_2}{n_2}$

donde B_i *es el número de exitos en muestras de tamaño* n_i. *Si* B_1 *y* B_2 *son independientes, se tiene que*

1. $\widehat{P}_1 - \widehat{P}_2$ *es un estimador insesgado de* $p_1 - p_2$*, es decir*

$$E\left(\widehat{P}_1 - \widehat{P}_2\right) = p_1 - p_2$$

2. $Var\left(\widehat{P}_1 - \widehat{P}_2\right) = \dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}$

3. *Si* $\widehat{P}_1 \sim N$ *y* $\widehat{P}_2 \sim N$ *entonces Un "intervalo de confianza" del* $100\,(1-\alpha)\,\%$ *para* $p_1 - p_2$ *tiene extremos*

$$\widehat{p}_1 - \widehat{p}_2 \pm z_{\alpha/2} \cdot \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

4. *Para encontrar un intervalo de confianza del* $100\,(1-\alpha)\,\%$ *para* $p_1 - p_2$ *con un radio menor o igual* r*, el tamaño de las muestras de cada población debe ser*

$$n \geq \left(\frac{z_{\alpha/2} \cdot \sqrt{p_1 q_1 + p_2 q_2}}{r}\right)^2$$

Prueba. *Seguidamente se demuestran cada uno de los resultados.*

1. *Dado que* $\widehat{P}_1$ *y* $\widehat{P}_2$ *son estimadores insesgados de* p_1 *y* p_2 *respectivamente entonces*

$$E\left(\widehat{P}_1 - \widehat{P}_2\right) = E\left(\widehat{P}_1\right) - E\left(\widehat{P}_2\right) = p_1 - p_2.$$

2. *Se tiene que*

$$\begin{aligned}
Var\left(\widehat{P}_1 - \widehat{P}_2\right) &= Var\left(\frac{B_1}{n_1} - \frac{B_2}{n_2}\right) \\
&= Var\left(\frac{B_1}{n_1}\right) + Var\left(-\frac{B_2}{n_2}\right) && (B_1 \text{ y } B_2 \text{ son independientes}) \\
&= \left(\frac{1}{n_1}\right)^2 Var\left(B_1\right) + \left(\frac{-1}{n_2}\right)^2 Var\left(B_2\right) && (\text{propiedades de la varianza}) \\
&= \left(\frac{1}{n_1}\right)^2 n_1 p_1 q_1 + \left(\frac{-1}{n_2}\right)^2 n_2 p_2 q_2 && (B_1 \sim B\left(n_1, p_1\right) \text{ y } B_2 \sim B\left(n_2, p_2\right)) \\
&= \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}
\end{aligned}$$

3. *Si* $\widehat{P}_1 \sim N$ *y* $\widehat{P}_2 \sim N$ *entonces*

$$\widehat{P}_1 - \widehat{P}_2 \sim N\left(p_1 - p_2, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right)$$

y además

$$P\left(\widehat{P}_1 - \widehat{P}_2 - z_{\alpha/2} \cdot \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} < p_1 - p_2 < \widehat{P}_1 - \widehat{P}_2 + z_{\alpha/2} \cdot \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right) = 1 - \alpha$$

Es decir para un valor $\widehat{p}_i$ de $\widehat{P}_i$ para una muestra de tamaño n_i encontramos que un "intervalo de confianza" del $100\,(1-\alpha)\,\%$ para $p_1 - p_2$ es tiene extremos

$$\left]\widehat{p}_1 - \widehat{p}_2 - z_{\alpha/2}\cdot\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}, \widehat{p}_1 - \widehat{p}_2 + z_{\alpha/2}\cdot\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}\right[.$$

4. *Tomando muestras de igual tamaño, de acuerdo al resultado anterior, el radio del IC es*

$$\left|z_{\alpha/2}\cdot\sqrt{\frac{p_1q_1}{n}+\frac{p_2q_2}{n}}\right|,$$

se quiere que este radio sea menor a r :

$$\left|z_{\alpha/2}\cdot\sqrt{\frac{p_1q_1}{n}+\frac{p_2q_2}{n}}\right| \leq r \quad\Longrightarrow\quad \frac{z_{\alpha/2}\cdot\sqrt{p_1q_1+p_2q_2}}{r} \leq \sqrt{n}$$

$$\Longrightarrow\quad n \geq \left(\frac{z_{\alpha/2}\cdot\sqrt{p_1q_1+p_2q_2}}{r}\right)^2 \quad \blacksquare$$

Observaciones. Considere los dos últimos resultados del teorema anterior.

1. (**IC de $p_1 - p_2$ para muestras grandes**) El resultado (3) no brinda un *IC* práctico. Sin embargo, recuerde que si
$$n_i\widehat{p}_i \geq 5 \quad\wedge\quad n_i\widehat{q}_i \geq 5$$
entonces:
   (a) Aceptamos empíricamente que cada $\widehat{P}_i$ sigue una distribución normal, de acuerdo al teorema 21, pues se considera que el tamaño de las muestras n_i es grande.
   (b) Dado que n es grande entonces una buena estimación de p_i y q_i es respectivamente $\widehat{p}_i$ y $\widehat{q}_i$.

   Así, si las muestras son independientes, $n_i\widehat{p}_i \geq 5$ y $n_i\widehat{q}_i \geq 5$ entonces un intervalo de confianza del $100\,(1-\alpha)\,\%$ para $p_1 - p_2$ tiene extremos
$$\widehat{p}_1 - \widehat{p}_2 \pm z_{\alpha/2}\cdot\sqrt{\frac{\widehat{p}_1\widehat{q}_1}{n_1}+\frac{\widehat{p}_2\widehat{q}_2}{n_2}}$$

2. (**Tamaño de muestras**) En el resultado (4) sobre el tamaño de muestras, dado que p_i y q_i son desconocidos se puede proceder de una de las siguientes maneras:
   (a) Si ya se tienen buenas estimaciones $\widehat{p}_i$ y $\widehat{q}_i$ de p_i y q_i respectivamente (es decir $n_i\widehat{p}_i \geq 5$ y $n_i\widehat{q}_i \geq 5$ para $i = 1, 2$), entonces se utilizan estas estimaciones en el calculo del tamaño de muestra
$$n \geq \left(\frac{z_{\alpha/2}\cdot\sqrt{\widehat{p}_1\widehat{q}_1+\widehat{p}_2\widehat{q}_2}}{r}\right)^2$$

Luego, para garantizar el uso de la normal estandar en el cálculo anterior, el mínimo valor de n, denotado por m, debe cumplir

$$m\widehat{p}_i \geq 5 \quad y \quad m\widehat{q}_i \geq 5, \text{ para } i = 1, 2.$$

Si m no cumple estas condiciones, se toma el tamaño de muestra n como aquel que satisface $n\widehat{p}_i \geq 5$ y $n\widehat{q}_i \geq 5$, para $i = 1, 2$.

(b) Usar el hecho de que $p_i q_i \leq 0.25$,por lo tanto, se busca el n que cumpla

$$n \geq \left(\frac{z_{\alpha/2} \cdot \sqrt{0.25 + 0.25}}{r} \right)^2.$$

Estos valores de n es un hecho que satisfacen la condición del resultado 4, pues la condición propuesta es una condición más exigente que la da por este resultado. En esta opción puede ocurrir que el valor de n encontrado sea mucho más grande que el n requerido.

Ejemplo 35 *Un producto electrónico es fabricado por dos empresas A y B. Se compararon las proporciones de productos con defectos de fabricación para ambas empresas, reuniéndose los datos siguientes*

Empresa A:	*7 productos defectuosos de 100*
Empresa B:	*6 productos defectuosos de 80*

En la tienda COMPUBUENA vende ambos productos y uno de los vendedores asegura que el producto de A tiene menos defectos de fabricación que el producto de B. Sin embargo, los compradores suelen dudar sobre la información dada por el vendedor, ya que el producto de A tiene un precio mucho mayor con respecto al precio del otro producto.

1. *Encuentre un intervalo de confianza del 95% para la diferencia en las dos proporciones.*

   Sean

   p_1 : *la proporción de artículos defectuosos de la empresa A.*
   p_2 : *la proporción de artículos defectuosos de la empresa B.*

   Se tienen los siguientes datos:

   $$n_1 = 100, n_2 = 80, \widehat{p}_1 = \frac{7}{100}, \widehat{q}_1 = \frac{93}{100}, \widehat{p}_2 = \frac{6}{80}, \widehat{q}_2 = \frac{74}{80}.$$

   Como $n_i\widehat{p}_i \geq 5$ *y* $n_i\widehat{q}_i \geq 5$ *el IC tiene extremos:*

   $$\widehat{p}_1 - \widehat{p}_2 \pm z_{\alpha/2} \cdot \sqrt{\frac{\widehat{p}_1\widehat{q}_1}{n_1} + \frac{\widehat{p}_2\widehat{q}_2}{n_2}} = \frac{7}{100} - \frac{6}{80} \pm 1.96\sqrt{\frac{7 \cdot 93}{100^3} + \frac{6 \cdot 74}{80^3}}$$

   Así el IC del 95% *es*

   $$I = \,]-0.08136\,93, 0.07136\,93\,[$$

2. *¿Los datos apoyan la afirmación que realiza el vendedor?*

   $$NO,\ \text{pues } I \text{ tiene valores positivos y negativos.}$$

   Esto no significa que el vendedor este mintiendo, ni que su afirmación es falsa ni que se rechaza. Esto quiere decir que la afirmación del vendedor no es posible aceptarla ni rechazarla con los datos obtenidos. Se requiere un IC más pequeño, pues este tiene un radio aproximadamente de

   $$\frac{0.07136\,93 + 0.08136\,93}{2} \approx 0.07636\,93$$

3. *¿De qué tamaño deben ser las muestras si se desea tener una confianza de al menos el 95% de que el error estimado al estimar la diferencia en las dos proporciones sea menor que 0.02?*

   Dado que $\widehat{p}_1 = \frac{7}{100}, \widehat{q}_1 = \frac{93}{100}, \widehat{p}_2 = \frac{6}{80}$ *y* $\widehat{q}_2 = \frac{74}{80}$ *son buenas estimaciones de* p_1, q_1, p_2 *y* q_2, *respectivamente, entonces se busca un* n *que cumpla que*

   $$n \geq \left(\frac{z_{0.025} \cdot \sqrt{\widehat{p}_1\widehat{q}_1 + \widehat{p}_2\widehat{q}_2}}{r} \right)^2 = \left(\frac{1.96 \cdot \sqrt{\frac{7 \cdot 93}{100^2} + \frac{6 \cdot 74}{80^2}}}{0.02} \right)^2 \approx 1291.\,50$$

   Así se toma $n = 1292$. *Note que para este valor de* n : $n\widehat{p}_1 = 90.\,44$, $n\widehat{q}_1 = 1201.\,56$, $n\widehat{p}_2 = 96.\,9$, $n\widehat{q}_2 = 1195.\,1$, *todos mayores a 5, esto justifica el uso de la normal estandar en el cálculo de n. Si bien este n es grande, en este caso la muestra no destruye el objeto muestreado y con muestras de tamaño* 1292 *se halla un IC que pueda ayudar a tomar una mejor decisión con respecto a la afirmación del vendedor.*

Ejercicio 18 *Al estudiar 300 personas en la ciudad A, 128 prefieren el jabón X sobre las demás marcas de jabones desodorantes. En la ciudad B, 49 de 400 personas prefieren dicho jabón*

1. *Obtenga el intervalo de confianza de 95% para estimar la diferencia en las dos proporciones.* *R/* $]0.2396, 0.3687[$
2. *Interprete el IC e indique las medidas que puede tomar la empresa que fabrica el jabón X.*

Ejercicio 19 *Se compararon las proporciones de piezas defectuosas producidas por dos rnáquinas, reuniéndose los datos siguientes*

Maquina 1:	*10 piezas defectuosas de 150*
Maquina 2:	*14 piezas defectuosas de 150*

Determine un intervalo de confianza de 90% para estimar la diferencia en las dos proporciones. *R/* $]-0.0778, 0.0248[$

4.3.3 Intervalo de confianza para la razón entre dos variancias

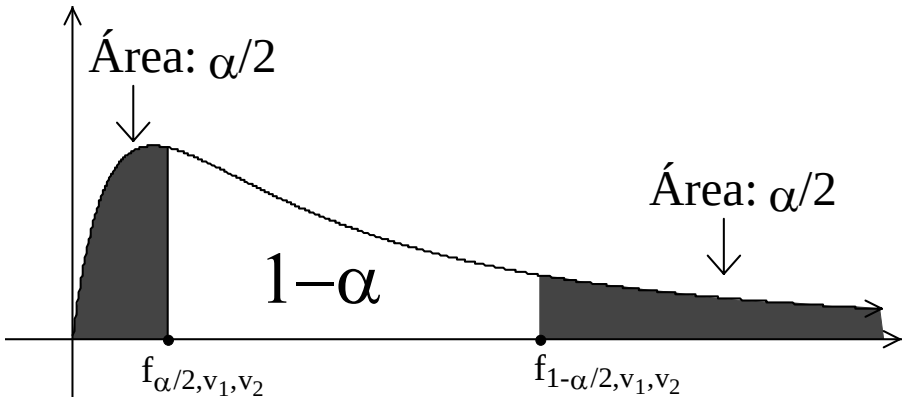
Recuerde el siguiente teorema.

Teorema 30 *Considere las poblaciónes dadas por la variables aleatorias X, Y que siguen una distribución normal con varianzas poblacionales de σ_1^2 y σ_2^2 repectivamente. Sean S_1^2 y S_2^2 varianzas muestras de muestras aleatorias independientes de tamaño n_1 y n_2, tomadas de cada población respectivamente. Se tiene que*

$$F = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

sigue una distribución F con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad.

Recuerde que $f_{\alpha/2,n_1-1,n_2-1}$ deja una área de $\frac{\alpha}{2}$ a la izquierda y $f_{1-\alpha/2,n_1-1,n_2-1}$ deja una área de $\frac{\alpha}{2}$ a la derecha:



entonces

$$P\left(f_{\alpha/2,n_1-1,n_2-1} < F = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} < f_{1-\alpha/2,n_1-1,n_2-1}\right) = 1 - \alpha$$

Por lo tanto,

$$P\left(\frac{S_2^2 f_{\alpha/2,n_1-1,n_2-1}}{S_1^2} < \frac{\sigma_2^2}{\sigma_1^2} < \frac{S_2^2 f_{1-\alpha/2,n_1-1,n_2-1}}{S_1^2}\right) = 1 - \alpha$$

Dado que $f_{\alpha/2,n_1-1,n_2-1} = \dfrac{1}{f_{1-\alpha/2,n_2-1,n_1-1}}$ se obtiene el siguiente resultado

Teorema 31 *Considere las poblaciónes 1 y 2 que siguen una distribución normal con varianzas poblacionales de σ_1^2 y σ_2^2 repectivamente. Sean S_1^2 y S_2^2 las varianzas muestras de muestas aleatorias independientes de tamaño n_1 y n_2, tomadas de cada población respectivamente.Un intervalo de confianza del $100\,(1-\alpha)\,\%$ para $\dfrac{\sigma_2^2}{\sigma_1^2}$ es*

$$\left]\frac{s_2^2}{s_1^2\ f_{1-\alpha/2,n_2-1,n_1-1}}, \frac{s_2^2 f_{1-\alpha/2,n_1-1,n_2-1}}{s_1^2}\right[$$

Ejemplo 36 *Un curso es impartido tradicionalmente por 2 profesores A y B. Se tiene que 15 de los estudiantes del profesor A tienen una nota final promedio de 67.1 con una desviación estándar de 15, y 21 estudiantes del profesor B tienen un promedio de 62.8 con una desviación estándar de 18. Suponga que ambas notas siguen una distribución normal.*

1. *Encuentre un intervalo de confianza de 90% para el cociente de las desviaciones estándar de las notas obtenidas por ambos profesores.*

   Sean

   $$\begin{array}{rcl} \sigma_1 & : & \textit{desviación estándar de las notas del profesor } A. \\ \sigma_2 & : & \textit{desviación estándar de las notas del profesor } B. \end{array}$$

   Se tiene que

   $$\alpha = 0.1, n_1 = 15, s_1 = 15, n_2 = 21, s_2 = 18$$

   Además,

   $$\begin{aligned} f_{1-\alpha/2,n_1-1,n_2-1} &= f_{0.95,14,20} = 2.225 \\ f_{1-\alpha/2,n_2-1,n_1-1} &= f_{0.95,20,14} = 2.388 \\ \frac{1}{f_{0.95,20,14}} &= \frac{1}{} = 0.418\,76 \\ f_{0.05,20,14} &= \frac{1}{f_{0.95,14,20}} = \frac{1}{2.225} = 0.449\,438 \end{aligned}$$

   Por lo tanto, un intervalo de confianza del 90% *para* $\dfrac{\sigma_2^2}{\sigma_1^2}$ *es*

   $$\begin{aligned} \left]\frac{s_2^2}{s_1^2\ f_{1-\alpha/2,n_2-1,n_1-1}}, \frac{s_2^2 f_{1-\alpha/2,n_1-1,n_2-1}}{s_1^2}\right[&= \left]\frac{(18)^2}{(15)^2 \cdot 2.388}, \frac{(18)^2 \cdot 2.225}{(15)^2}\right[\\ &=]0.603\,014\,, 3.\,204[\end{aligned}$$

   Aplicando raiz cuadrada a ambos extremos del intervalo, se tiene que un intervalo de confianza del 90% *para* $\dfrac{\sigma_2}{\sigma_1}$ *es*

   $$]0.776\,540\,, 1.\,789\,97[\,.$$

2. *¿Puede suponerse que las variancias son iguales? Explique.*

   $$\textit{Si, pues } 1 \ \in]0.603\,014\,, 3.\,204[$$

   Note que si las varianzas son iguales entonces $\dfrac{\sigma_2}{\sigma_1} = 1$.

3. *Encuentre un intervalo de confianza de 95% para la diferencia entre los promedios de notas obtenidos por los dos profesores.*

Sean:

$$\begin{aligned} X_1: &\quad \textit{la nota obtenida con el profesor } A \\ X_2: &\quad \textit{la nota obtenida con el profesor } B \end{aligned}$$

Se tiene que

$$\overline{x}_1 = 67.1, s_1 = 15, \quad \overline{x}_2 = 62.8, s_1 = 18$$

Dado que las muestras tienen tamaño menor a 30 *y se desconocen las variancias pero se suponen iguales (por* 2*), el IC tiene extremos:*

$$(\overline{x}_1 - \overline{x}_2) \pm t_{\alpha/2,v} \cdot \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}.$$

donde

$$\begin{aligned} s_p^2 &= \frac{14s_1^2 + 20s_1^2}{34} = \frac{14\,(15)^2 + 20\,(18)^2}{34} \approx 283.\,235 \\ t_{\alpha/2,v} &= t_{0.025,34} = 2.03224. \end{aligned}$$

Así, si tiene que el IC tiene extremos

$$(67.1 - 62.8) \pm 2.03224 \cdot \sqrt{\frac{283.\,235}{15} + \frac{283.\,235}{21}} \approx 4.3 \pm 11.\,562\,3.$$

Así un IC del 95% *para* $\mu_1 - \mu_2$ *es*$]-7.\,262\,31, 15.\,862\,3[$.

4. *Considera que hay un profesor que tiene mejor rendimiento promedio que el otro. Justifique.*

No, pues el IC anterior tiene valores positivos y negativos

Ejercicio 20 *El Instituto del Consumidor desea comparar la variabilidad en la cantidad de cierto componente químico de un medicamento elaborado por las compañías A y B. Ambos medicamentos se distribuyen en forma de tabletas y contiene en promedio 250 mg del componente químico. En una muestra de 26 tabletas de cada compañía, se encontraron las desviaciones estándares de 1.25 y 1.18 en la cantidad del componente químico para cada compañia respectivamente. (Suponga que cada población se distribuye normalmente)*

1. *Calcule un intervalo de confianza de 95% para* $\frac{\sigma_A^2}{\sigma_B^2}$. *R/* $]0.503, 2.5025[$
2. *¿Está en contra de que* $\sigma_A^2 > \sigma_B^2$*?* *R/* *No*
3. *¿Aceptaría que* $\sigma_A^2 > \sigma_B^2$*?* *R/* *No*

4.3.4 Ejercicios

1. La oficina del Consumidor desea determinar si hay diferencia entre el peso promedio de dos tipos de bolsas de arroz A y B , que se venden con peso nominal de dos kilos. Para ello, realizó una inspección, obteniendo los siguientes resultados:

Tipo	Tamaño de la muestra	Media muestral	Desviación muestral
A	50	$2.05kg$	$0.2kg$
B	45	$1.97kg$	$0.12kg$

   (a) Encuentre un IC del 95% para la diferencia entre el peso promedio de la bolsa del arroz A y el peso promedio de la bolsa del arroz B. $R/$ $\left]1.\,440\,59 \times 10^{-2} \quad , \quad 0.145\,594\right[$

   (b) Considera que el peso promedio de la bolsa del arroz A es diferente al peso promedio de la bolsa del arroz B. Justifique. $R/$ Si, se puede considerar que $u_1 > u_2$

2. Detergentes BIEN LIMPIO investiga la preferencia de su marca frente a otras. Se encuesta a 200 personas en el San José y se encuentra que 50 de ellas están a favor de su marca, y en Cartago 25 personas de un total de 150 la prefieren.

   (a) Encuentre un intervalo de confianza de 95% para la diferencia de proporciones entre las preferencias en San José y en Cartago $R/$ $]-0.001608, 0.167\,608[$

   (b) Determine de qué tamaños deben ser las muestras para encontrar un intervalo de confianza de 96% con un radio menor que 0:04 si por cada encuestado de Cartago se encuestarán dos en San José $R/$ $n > 611.04$

   (c) ¿Puede afirmarse que los capitalinos prefieren más la marca que los cartaguineses? Justifique $R/$ No

3. En una comunidad existen dos colegios, el alcalde de la comunidad afirma que los estudiantes del primer colegio de sétimo año tiene estaturas más similares que los estudiantes del otro colegio. Ante esta afirmación un estudiante del ITCR decide hallar un intervalo de confianza de 90% para el cociente de las varianzas $\left(\frac{\sigma_2^2}{\sigma_1^2}\right)$, tomando una muestra de 21 estudiantes en el primer colegio y de 19 estudiantes en el segundo colegio, obteniendo

$$IC :]0.1212, 0.5859[$$

   (a) Determine aproximadamente el conciente de las varianzas muestrales para las muestras escogidas. $R/$ $0.265\,557$

   (b) ¿Los datos apoyan la afirmación que realiza el alcalde? $R/$ No

   (c) Suponga que en la comunidad un colegio es diurno y otro nocturno. ¿Cuál considere que es el colegio nocturno? Justifique su respuesta. $R/$ Colegio 1

4. Las estaturas de 6 niños de 1° y 3° grado son:

Grado	Estaturas
1°	$121, 115, 118, 122, 119, 118$
3°	$135, 132, 130, 136, 135, 133$

Suponga que ambas estaturas se distribuyen normalmente.Determine aproximadamente un IC de 90% para el cociente de las varianzas de las estaturas de ambas poblaciones.
$R/$ $]0.155\,412, 4.\,401\,03[$.

5. Un IC de 90% para la diferencia de promedios $(\mu_1 - \mu_2)$ es $]166.5, 191.9[$. Determine el valor de $\overline{x}_1 - \overline{x}_2$ utilizado en el cálculo del IC $R/$ 179. 2.

6. En el 2009 se realizó un pequeño estudio para analizar qué porcentaje de estudiantes de la carrera de Ingeniería en Computación de la UNAS realizan actividad física para cuidar su salud. En una encuesta a 50 estudiantes (20 hombres - 30 mujeres), obteniendo que de los hombres solo 7 realizan actividad física. De acuerdo con los datos, un intervalo de confianza del 95% para la diferencia entre la proporción de hombres que realizan ejercicios y la proporción de hombres que realizan ejercicio es :
$$IC \text{ para } p_h - p_m :]-0.624\,12, -0.07588[$$

   (a) Determine aproximadamente el número de mujeres encuestadas que realizan ejercicio.$R/$ 21

   (b) Considera aceptable indicar que el porcentaje de mujeres estudiantes de UNAS que realizan ejercicio, es mayor que el porcentaje de estudiantes varones. $R/$ Si

7. Las carreras de Electrónica y Computación tiene gran demanda laboral, esto hace que muchos estudiantes ingresen al mundo laboral antes de graduarse. Sin embargo, el profesor afirma que, para estudiantes de último de carrera, es mayor la proporción de estudiantes de Computación que laboran que la de estudiantes de Electrónica. En una pequeña encuesta se obtuvo la siguiente información

Estudiantes de último año	# de encuestados	Laboran
Computación	20	12
Electrónica	30	14

   Suponga que se quiere determinar un intervalo de confianza para la diferencia de proporciones entre los estudiantes de último año de Computación y de Electrónica que laboran. ¿De qué tamaño debe ser las muestras si se desea un IC con una confianza de al menos el 95% de que el error estimado al estimar la diferencia en las dos proporciones sea menor que 0.1? $R/$ 188

8. Para determinar un intervalo para la diferencia de medias $(\mu_1 - \mu_2)$ se tomaron muestras de igual tamaño $n_1 = n_2 = 50$. Con los datos obtenidos, Silvia determinó un intervalo de confianza I_1 del 80% para $\mu_1 - \mu_2$. Utilizando los mismos datos, Luis determinó un intervalo de confianza I_2 del 90% para $\mu_1 - \mu_2$. ¿Cuál de las siguientes afirmaciones es verdadera? Justifique su respuesta explícitamente. $R/$ la a.

   (a) Se cumple que $I_1 \subseteq I_2$

   (b) Se cumple que $I_2 \subseteq I_1$

   (c) No se cumple ninguna de las anteriores.

   (d) 90% para μ tiene extremos

9. La universidad Bienestar Seguro tiene 2 fórmulas para examen de admisión que pretende utilizar durante los próximos 3 años. Sin embargo, un profesor de estadística de la universidad afirma que la fórmula 1 va a tener un mejor rendimiento promedio que la fórmula 2. Ante esto, la universidad aplicó las fórmulas a un grupo de estudiantes, obteniendo los siguientes resultados:

Fórmula	Tamaño de la muestra	Media muestral	Desviación muestral
1	21	65	24
2	17	63	15

   (a) Encuentre un intervalo de confianza de 90% para el cociente de las varianzas de las notas obtenidas en las fórmulas de examen. *R/* $]0.178\,857, 0.889\,063[$

   (b) ¿Puede suponerse que las variancias son iguales? Explique. *R/* No

   (c) Encuentre un intervalo de confianza de 95% para la diferencia entre los promedios de las notas obtenidas en las fórmulas de examen. *R/* $]-10.\,959\,2, 14.\,959\,2[$

   (d) ¿Es aceptable la afirmación del profesor? Justifique *R/* No

10. Sea desea investigar la duración de dos tipos A y B de baterías AA no recargables, las cuales tiene precios similares. Un estudiante, que utiliza mucho baterías AA, señala que la duración promedio de las baterías A supera en más de 10 minutos a la duración promedio de las baterías B. Se tomaron muestras de duraciones de ambos tipos de baterías, la información se resume en la siguiente tabla

Batería AA	tamaño de muestra	$\overline{x}$	s
tipo A	21	5.3 horas	0.8 horas
tipo B	19	5.1 horas	0.6 horas

   (a) Encuentre un intervalo de confianza de 90% para el cociente de las desviaciones estándar poblaciones de las duraciones de las baterías. *R/* $]0.511\,377\,, 1.\,110\,15[$

   (b) ¿Puede suponerse que las variancias son iguales? Explique. *R/* Si

   (c) Determine un IC del 90% para la diferencia de duraciones promedio de baterías *R/* $]-0.180\,233,\ 0.580\,233\,[$ para $\mu_A - \mu_B$

   (d) ¿es aceptable la afirmación del estudiante? *R/* *No*

11. El pueblo C tiene problemas con el fumado en adolescentes. En una muestra de 60 hombres adolescentes se observó que 40 fuman, y en una muestra de n_2 mujeres adolescentes se observó una proporción $\widehat{p}_2$ de fumadoras, con $n_2\widehat{p}_2 \geq 5$ y $n_2\widehat{q}_2 \geq 5$. Con estos datos se tiene que un IC del 95% para la diferencia de la proporciones, proporción de hombres adolescentes que fuman menos la proporción de mujeres adolescentes, es

$$I = \,]\,-0.355\,219, -0.044\,780\,8\,[$$

   (a) Determine el centro del IC y el valor aproximado de $\widehat{p}_2$ *R/* $\widehat{p}_2 = 0.866\,667$

   (b) Halle el valor aproximado de n_2. *R/* $n_2 \approx 45$

   (c) El sacerdote del pueblo afirma que en esto tiempos son más las mujeres adolescentes que fuman que hombres adolescentes. Utilizando el IC, ¿Los datos apoyan la afirmación del sacerdote? *R/* Si

Capítulo III

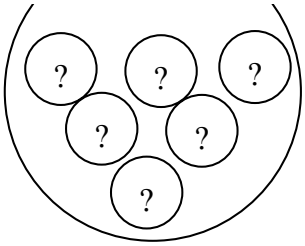
Pruebas de hipótesis

Se inicia con un estudo de los conceptos generales de pruebas hipótesis: hipótesis nula, hipótesis alternativa, tipos de errores, nivel de significancia, regiones de aceptación y rechazo, y pontencia de una prueba. Luego se abordan los dos métodos para contrastar hipótesis: determinación de regiones y el valor P, y se establece la relación entre ambos. Finalmente, se estudian las principales pruebas de hipótesis para una y dos poblaciones, y las pruebas de independencia, bondad de ajuste y análisis de variancia.

1 ¿Qué es una prueba de hipótesis?

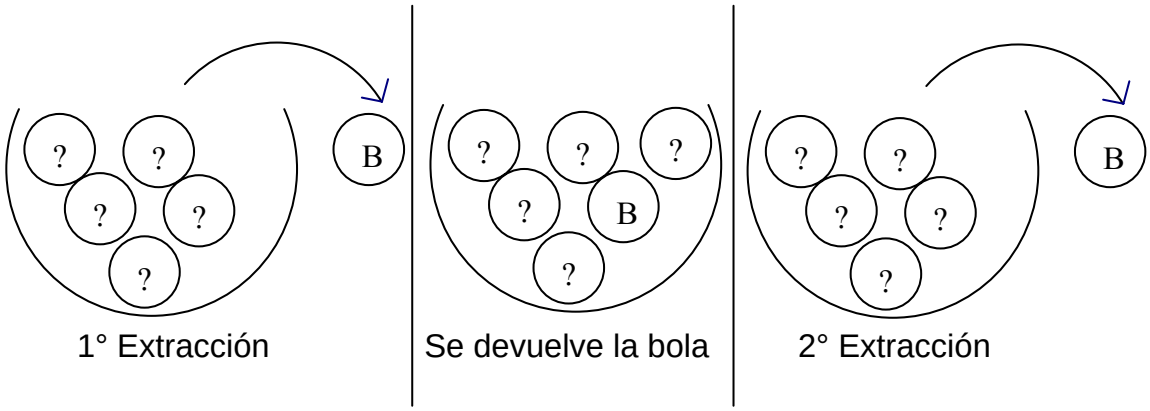
1.1 Introducción

Suponga que se tiene una bolsa con 6 bolas de las cuales no se conoce su color:



Una persona realiza la siguiente afirmación: "**en la bolsa hay 5 bolas verdes y una blanca**" y le solicitan a usted que rechace o acepte la afirmación, para ello le dan la posibilidad de realizar una muestra aleatoria de tamaño 2. Es decir le permiten elegir 2 bolas al azar con reposición de la bolsa.

Suponga que al realizar las dos extracciones con reposición obtiene dos bolas blancas:



Con base en la información obtenida en la muestra, responda las siguientes preguntas:

1. ¿Acepta ó rechaza la afirmación de la persona?
2. ¿Hay alguna posibilidad de que se cumpla la afirmación?
3. ¿Considera más probable que la afirmación se correcta o incorrecta?
4. Discuta la diferencia, con respecto a la afirmación, entre la toma de una decisión (aceptar o rechazar la afirmación) y la asignación de un valor de verdad a la afirmación (afirmación correcta o incorrecta).

Suponga que la bolsa fue desechada y nunca se determino la veracidad de la afirmación, entonces ¿Puede asegurar que su decisión fue la correcta o la más probable?.

Cuando se toman decisiones con respecto a la información brindada por una muestra se realiza una prueba de hipótesis. Discuta la diferencia entre una prueba de hipótesis y una prueba matemática.

1.2 Tipos de hipótesis

La investigación define las hipótesis como las afirmaciones que son objeto de estudio y que se desea determinar su veracidad. La prueba de hipótesis es un método de investigación que a partir de una afirmación, sobre el valor de un parámetro, define dos hipótesis opuestas (una de las cuales es la afirmación) y busca evidencias para aceptarlas o rechazarlas. Seguidamente se definen los dos tipos de hipótesis.

Definición 21 ***Hipótesis Nula*** *("Acusado es inocente"): es una aseveración en el sentido de que un parámetro θ tenga un valor específico θ_0 y se denota H_0:*

$$H_0 : \theta = \theta_0$$

donde θ_0 es un valor del estadístico $\widehat{\Theta}$ llamado valor nulo. El procedimiento consiste en suponer que H_0 es verdadera y buscar evidencias en contra del supuesto. Para que H_0 se acepte no debe haber una fuerte evidencia en su contra.

Definición 22 ***Hipótesis Alternativa*** *("Acusado es culpable"): es una aseveración que se acepta si se rechaza H_0, se denota por H_1 y tiene la forma:*

$$H_1 : \theta < \theta_0, \quad H_1 : \theta > \theta_0 \quad o \quad H_1 : \theta \neq \theta_0$$

Observaciones: Se tiene que

1. De acuerdo a la ley de tricotomía:

   $$\theta = \theta_0, \quad \theta < \theta_0 \quad o \quad \theta > \theta_0$$

   H_0 y H_1 deben abordar estas afirmaciones y ser opuestas.

2. Aunque H_0 se redacta en forma de igualdad, en realidad suele ser su caso extremo para que se cumpla. Por ejemplo, si se desea estudiar la afirmación $\theta \leq \theta_0$, entonces, dado que la afirmación debe estar sola en una de las hipótesis para poder analizarla, se tiene que

   $$H_0 : \theta = \theta_0 \quad (\leq) \qquad H_1 : \theta > \theta_0$$

   Observe que la afirmación, dado que contiene la igualdad ($\theta \leq \theta_0$), se le asigna a H_0. Así, H_0 se acepta si $\theta \leq \theta_0$ y su caso extremo para que se acepte es que $\theta = \theta_0$. Por otro lado, en H_1 se le asigna lo contrario a la afirmación.

3. Recuerde que *"el acusado es inocente hasta que se demuestre lo contrario"*. Por lo tanto, hasta el momento el procedimiento para realizar una prueba de hipótesis consiste en:

   (a) Definir H_0 y H_1 a partir de la afirmación a analizar

(b) Asumir H_0 es verdadero, es decir se asume que *"el acusado es inocente".*

(c) Tomar una muestra para hallar un valor $\widehat{\theta}$ del estadístico $\widehat{\Theta}$

(d) Utilizar la estimación $\widehat{\theta}$ y la distribución de $\widehat{\Theta}$ para determinar evidencias en contra de H_0, es decir *"se buscan evidencias que se demuestre lo contrario".*

(e) Decisión. Si se determinar evidencias significativas en contra de H_0, entonces se rechaza H_0. De lo contrario se acepta H_0.

(f) Conclusión. Si se acepta H_0, no significa que se acepta que *"el acusado es inocente"* sino, al igual que en un juicio, se concluye que no se hallaron evidencias significativas en contra de H_0. Por otro lado si se rechaza H_0, se concluye que existe evidencia en contra de H_0.

Ejemplo 37 *Para cada una de las siguientes situaciones identifique el parámetro, indique la afirmación a analizar y, con base en ésta, identifique la hipótesis nula y alternativa.*

1. *Se quiere demostrar que el nivel promedio de carbono del centro de San José es mayor que* 4.9 *partes por millón.*

   Sea μ *el nivel promedio de carbono del centro de San José.*

   $$Afirmación : \mu > 4.9$$
   $$H_0 : \mu = 4.9 \quad (\leq) \qquad H_1 : \mu > 4.9$$

2. *Se desea determinar si el porcentaje de costarricenses que duermen con el televisor encendido es menor a* 35%.

   Sea p *el porcentaje de costarricenses que duermen con el televisor encendido.*

   $$Afirmación : p < 0.35$$
   $$H_0 : p = 0.35 \quad (\geq) \qquad H_1 : p < 0.35$$

3. *Una máquina dispensadora de refrescos, cuando está bien ajustada, llena los vasos con una desviación estándar no mayor que 6 ml. Se desea determinar si existe evidencia para sospechar que la máquina del Super La Minita está desajustada.*

   Sea σ *la desviación estándar actual de la máquina dispensadora de refrescos del Super La Minita*

   $$Afirmación : \sigma \leq 6$$
   $$H_0 : \sigma = 6 \quad (\leq) \qquad H_1 : \sigma > 6$$

4. *Se desea concluir que el promedio de estaturas de los estudiantes de la escuela SONRISITAS supera en al menos 3cm el promedio de las estudiantes.*

   Sean

   μ_h : *el promedio de estaturas de los estudiantes de la escuela SONRISITAS en cm*
   μ_m : *el promedio de estaturas de las estudiantes de la escuela SONRISITAS en cm*

   Entonces

   $$Afirmación : \mu_h - \mu_m \geq 3$$
   $$H_0 : \mu_h - \mu_m = 3 \quad (\geq) \qquad H_1 : \mu_h - \mu_m < 3$$

Definición 23 *Una **prueba de una cola** es aquella en la cual H_1 tiene la forma:*

$$H_1 : \theta < \theta_0 \quad o \quad H_1 : \theta > \theta_0$$

Definición 24 *Una **prueba de dos colas** es aquella en la cual H_1 tiene la forma:*

$$H_1 : \theta \neq \theta_0$$

Ejercicio 21 *Para cada una de las siguientes afirmaciones redacte las hipótesis nula y alternativa*

1. *Se afirma que el promedio de edad de los vendedores del Mercado Central de Cartago es de por lo menos 30 años.* $R/ \quad H_1 : \mu < 30$
2. *Se desea determinar si la proporción de apartamentos de San José que prohíben mascotas no es superior al 30%.* $R/ \quad H_1 : p > 0.3$
3. *Se quiere concluir que la desviación estándar del peso de las bolsas de arroz Blanco con peso nominal de 2 kg es de 40 gramos.* $R/ \quad H_1 : \sigma \neq 40$

1.3 Regiones de aceptación y rechazo

La aceptación de H_0 se da si al tomar una muestra no se encuentra evidencia significativa en su contra. Es decir la aceptación de H_0 no está directamente ligada a la posibilidad de que se cumpla H_0 (aunque si indirectamente), más bien su aceptación se da cuando no existan datos en su contra. Por lo tanto el concepto de aceptación en pruebas de hipótesis es distinto al de la vida cotidiana.

Ejemplo 38 *Juan afirma que Ana tiene puesta una camisa de color rosada. Se tiene que*

$$\begin{aligned} H_0 &: \quad \textit{la camisa que tiene puesta Ana es rosada} \\ H_1 &: \quad \textit{la camisa que tiene puesta Ana no es rosada} \end{aligned}$$

Se entrevistan a cuatro personas que vieron Ana: Luis indica que la camisa de Ana no es de color verde, Rebeca dice que no es roja, Alberto que no recuerda el color y Sara indica que no es amarrilla. Por lo tanto, dado que la información obtenida no brinda evidencia en contra de H_0, se acepta que la camisa es rosada, esto pese a que directamente no se obtuvo información de que fuera de dicho color. Así,

$$\textit{Se acepta } H_0$$

y lo correcto es concluir que no se hallan evidencias en contra de que sea rosada

Ejemplo 39 *Considere el ejemplo anterior. Suponga que Anthony afirma que Ana tiene puesta una camisa de color negra. Así,*

$$\begin{aligned} H_0 &: \quad \textit{la camisa que tiene puesta Ana es negra} \\ H_1 &: \quad \textit{la camisa que tiene puesta Ana no es negra} \end{aligned}$$

De acuerdo a la información indicada en el ejemplo anterior, no hay evidencia significativa para rechazar H_0. Por lo tanto,

$$\textit{Se acepta que la camisa es negra} \qquad (1)$$

Sin embargo, en el ejemplo anterior, se obtuvo que

$$\textit{Se acepta que la camisa es rosada} \qquad (2)$$

¿Se contradicen (1) *y* (2)*? No, en realidad se acepta que la camisa es negra y que es rosada, al no existir evidencia en contra de ambas afirmaciones. Es decir, la aceptación se da porque no se halló un argumento sólido para negarla.*

Los ejemplos anteriores modelan el concepto de aceptación en una prueba de hipótesis e indicar que es posible aceptar dos afirmaciones contradictorias si no evidencia significativa para rechazarlas.

¿Cómo aceptar o rechazar H_0? La hipótesis nula realiza una afirmación sobre el valor de un parámetro θ. Recuerde que $\widehat{\Theta}$, un buen estimador o estadístico asociado a θ, brinda estimaciones sobre el valor θ a partir de muestras de tamaño fijo. Así, la idea es utilizar una estimación $\widehat{\theta}$ de θ y la distribución muestral de $\widehat{\Theta}$ para determinar evidencia en contra de H_0. Por lo tanto, el valor de la estimación de θ es trascendental para aceptar o rechazar H_0.

Ejemplo 40 *Se afirma que el promedio de edad de los vendedores del Mercado Central de Cartago es de por lo menos* 30 *años.*

Para analizar esta afirmación, sea μ la edad promedio de los vendedores del Mercado Central de Cartago. Así,

$$Afirmación : \mu \geq 30$$
$$H_0 : \mu = 30 \quad (\geq) \qquad H_1 : \mu < 30$$

El estadístico utilizado es $\overline{X}$. Recuerde que este estadístico brinda estimaciones de μ para muestras de tamaño 40, por ejemplo.

1. *Suponga que en una muestra de tamaño 40 se halló el valor $\overline{x} = 32$. Como $\overline{x}$ es una estimación de μ, esto indica que μ es cercano a* 32*, y recuerde que H_0 se cumple si $\mu \geq 30$. Así, el valor $\overline{x} = 32$ no representa una evidencia en contra de H_0, por lo tanto*

   $$\textit{Se acepta } H_0.$$

2. *Suponga que en una muestra de tamaño 40 se halló el valor $\overline{x} = 29$.Esto indica que μ es cercano a* 29*. Así, se ha determinado evidencia en contra de H_0. Sin embargo, ¿esta evidencia es significativa?*

   (a) *Recuerde que $\overline{x} = 29$ es solo una estimación de μ. Se sabe que μ es cercano a 29, entonces μ puede ser* 30 *el cual se supone cercano a* 29*. Por lo tanto, se puede considerar que la evidencia no es significativa y*

   $$\textit{Se acepta } H_0$$

   (b) *Por otro lado, se puede considerar que* 30 *está alejando de* 29*. Así, la evidencia es significativa y*

   $$\textit{Se rechaza } H_0$$

(c) *¿Es 29 cercano o no a* 30? *Esto dependerá de la distribución del estadístico* $\overline{X}$ *y del riesgo de equivocarse que se está dispuesto a asumir. Como se verá más adelante, si se considera que como* $\overline{x} = 29.9$ *entonces* $\mu < 30$ *(29.9 está alejado de* 30*) entonces se está aumentando el riesgo de cometer un error. En general, si estimaciones muy cercanas a* 30 *pero menores se consideran significativas para rechazar* H_0*, se aumenta el riesgo de equivocarse en la decisión tomada.*

Definición 25 *Dado que*

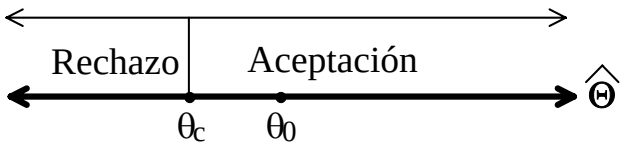
$$H_0 : \theta = \theta_0$$

se define

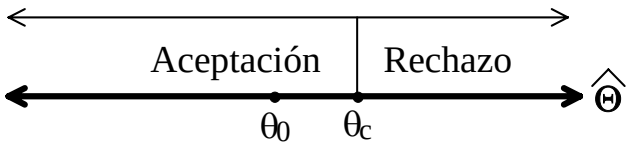
1. ***Región de aceptación****: Valores de* $\widehat{\Theta}$*, estimaciones de* θ*, que aceptan* H_0
2. ***Región de rechazo****: Valores de* $\widehat{\Theta}$*, estimaciones de* θ*, que rechazan* H_0
3. ***Valor crítico****: separa una región de rechazo y una de aceptación. Se denota* θ_c*.*

Seguidamente se presentan las regiones de aceptación y rechazo para cada prueba de hipótesis:

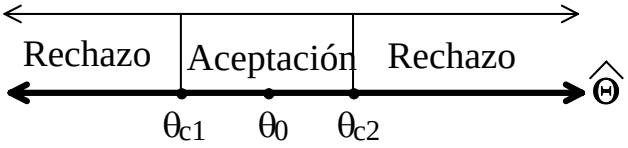
$H_0 : \theta = \theta_0 \quad (\geq)$
$H_1 : \theta < \theta_0$



$H_0 : \theta = \theta_0 \quad (\leq)$
$H_1 : \theta > \theta_0$



$H_0 : \theta = \theta_0$
$H_1 : \theta \neq \theta_0$



El lector puede notar que hay una tolerancia en cada una de las pruebas para considerar que $\theta = \theta_0$, tolerancia ejemplificada en el ejemplo anterior. De manera general, considere la primera prueba donde H_0 es

$$H_0 : \theta = \theta_0 \quad (\geq)$$

Entonces, para valores del estadístico (estimador de θ) muy cercanos a θ_0, aunque menores a este valor nulo, se considera que $\theta = \theta_0$ y por lo tanto debe formar parte de la región de aceptación.

Esta tolerancia, que diferencia el valor crítico θ_c y el valor nulo θ_0, se justifica debido a la utilización del estadístico $\widehat{\Theta}$ cuyos valores son estimaciones de θ tomadas de una muestra. Así, si las estimaciones de θ son muy cercanas a θ_0 se puede aceptar que $\theta = \theta_0$.

La distancia entre el valor nulo y el valor crítico dependerá del riesgo que se este dispuesto a tomar al realizar la prueba de hipótesis. Entre más lejanos están estos valores, se considera que estimaciones cercanas a θ_0 aceptan H_0 y hay menos posibilidad de cometer el error de rechazar H_0 siendo verdadera. Esto se estudiará en el siguiente apartado.

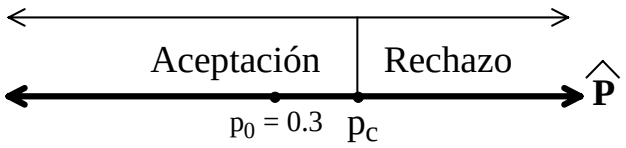
Así, la obtención del valor crítico dependerá de la distribución del estadístico y del riesgo que se está dispuesto a asumir al rechazar H_0 siendo verdadera. Esto se verá en el apartado (1.5) .

Ejemplo 41 *Se desea determinar si la proporción p de apartamentos de San José que prohíben mascotas no es superior al* 30%.

Se tiene que

$$Afirmación : p \leq 0.3$$
$$H_0 : p = 0.3 \quad (\leq) \qquad H_1 : p > 0.3$$

El estadístico a utilizar es $\widehat{P}$. *Las regiones de aceptación y rechazo son*



donde p_c *es el valor crítico.*

Ejercicio 22 *Para cada una de las siguientes situaciones, indique el parámetro a utilizar y la afirmación, redacte las hipótesis nula y alternativa, además identifique el estadístico a utilizar y las regiones de aceptación y rechazo.*

1. *Se afirma que la edad promedio de los niños de la guardería Osito Feliz es mayor a* 5.$R/ \quad H_1 : \mu > 5$, *valor crítico* μ_c
2. *Se desea determinar si el porcentaje de escolares que utilizan zapatos DURAN es de a lo sumo el* 40%. $R/ \quad H_1 : p > 0.4$, *valor crítico* p_c
3. *Se quiere concluir que la varianza en la duración de las llantas FIRE es igual a* 1000 km^2 $R/ \quad H_1 : \sigma^2 \neq 1000$, *valores críticos* σ^2_{c1} *y* σ^2_{c2}

1.4 Errores en una prueba de hipótesis

Al dictar sentencia en un juicio se pueden cometer dos tipos de errores: declarar culpable al acusado y que éste sea inocente, declarar inocente al acusado que éste sea culpable Estos dos errores son un modelo para explicar los error que se pueden cometer cuando se ha realizado una prueba de hipótesis.

Definición 26 *(Errores de una prueba de hipótesis) Se pueden presentar los siguientes errores:*

1. ***Error tipo I ("Condenar a un inocente")*** *la hipótesis nula ("el acusado es inocente") se rechaza siendo verdadera. La probabilidad del error tipo I se denota por α y es la probabilidad de rechazar H_0 ("se declara culpable al acusado") sabiendo que H_0 es verdadera ("sabiendo el acusado es inocente"):*
$$\alpha = P\left(H_1|H_0\right)$$
2. ***Error tipo II ("Dejar libre al culpable")*** *la hipótesis nula se acepta ("Acusado es inocente") siendo falsa. la probabilidad del error tipo II se denota por β y es la probabilidad de aceptar H_0 ("se declara inocente al acusado") sabiendo que H_1 es verdadera ("sabiendo el acusado es culpable"):*
$$\beta = P\left(H_0|H_1\right)$$

Analicemos la probabilidad del error tipo I **("Condenar a un inocente")**:

$$\alpha = P\left(H_1|H_0\right)$$

Note que H_1 se acepta (que es equivalente a rechazar H_0) para estimaciones de θ dadas por el estadístico $\widehat{\Theta}$ que se encuentran en la región de rechazo. Por otra lado dado que H_0 es verdadera, el caso extremo para que H_0 sea verdadero es que $\theta = \theta_0$. Así

$$\alpha = P\left(H_1|H_0\right) = P\left(\widehat{\Theta} \text{ este en la región de rechazo } \middle| \, \theta = \theta_0\right)$$

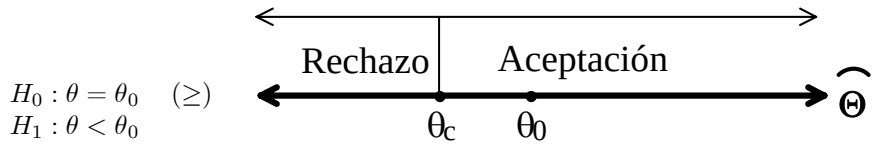
En cuando a la probabilidad del error tipo II **("Dejar libre al culpable")**:

$$\beta = P\left(H_0|H_1\right)$$

Se tiene que H_0 se acepta para estimaciones de θ dadas por el estadístico $\widehat{\Theta}$ que se encuentran en la región de aceptación. Por otro lado, la probabilidad de este tipo de error solo podrá ser analizada para hipótesis alternativas específicas: $H_1' : \theta = \theta_1$, donde θ_1 es un valor específico de θ para el cual H_1 se cumple. Así la probabilidad del error tipo II para la hipótesis alternativa específica: $H_1' : \theta = \theta_1$ es

$$\beta = P\left(H_0|H_1'\right) = P\left(\widehat{\Theta} \text{ este en la región de aceptación } \middle| \, \theta = \theta_1\right)$$

Ejemplo 42 *Considere la siguiente prueba de hipótesis*



En este caso, la probabilidad del error tipo I se redacta

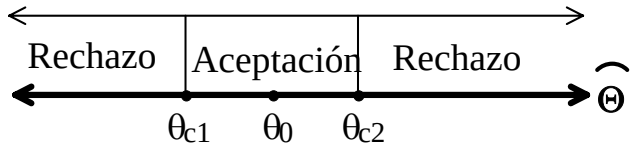
$$\alpha = P\left(H_1|H_0\right) = P\left(\widehat{\Theta} < \theta_c | \theta = \theta_0\right)$$

y la probabilidad del error tipo II para la hipótesis alternativa específica $H_1' : \theta = \theta_1$ se redacta

$$\beta = P\left(H_0|H_1'\right) = P\left(\widehat{\Theta} > \theta_c | \theta = \theta_1\right)$$

Ejemplo 43 *Considere la siguiente prueba de hipótesis*

$H_0 : \theta = \theta_0$
$H_1 : \theta \neq \theta_0$



En este caso, la probabilidad del error tipo I se redacta

$$\alpha = P\left(H_1|H_0\right) = P\left(\widehat{\Theta} < \theta_{c1} \quad \vee \quad \widehat{\Theta} > \theta_{c2}|\theta = \theta_0\right)$$

y además

$$\frac{\alpha}{2} = P\left(\widehat{\Theta} < \theta_{c1}|\theta = \theta_0\right)$$

Por otro lado, la probabilidad del error tipo II para la hipótesis alternativa específica $H_1' : \theta = \theta_1$ *se redacta*

$$\beta = P\left(H_0|H_1'\right) = P\left(\theta_{c1} < \widehat{\Theta} < \theta_{c2}|\theta = \theta_1\right)$$

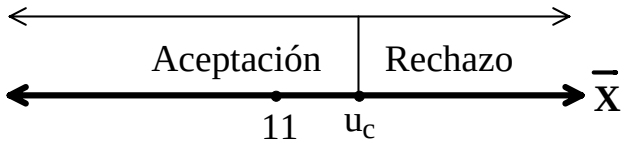
Ejemplo 44 *Suponga que la edad promedio* μ *de los niños de una cierta comunidad es mayor a* 11, *y suponga que la hipótesis alternativa específica es* $H_1' : u = 12$. *Redacte las hipótesis nula y alternativa, identifique el estadístico a utilizar y las regiones de aceptación y rechazo. Además redacte las probabilidades del error tipo I y del error tipo II para la hipótesis alternativa específica dada.*

La afirmación es

$$afirmación : \mu > 11$$

Recuerde que el estadístico asociado a μ *es* $\overline{X}$. *Así se tiene que*

$H_0 : \mu = 11 \quad (\leq)$
$H_1 : \mu > 11$



donde las probabilidades de los errores son

$$\begin{aligned} \alpha &= P\left(H_1|H_0\right) = P\left(\overline{X} > \mu_c|\mu = 11\right) \\ \beta &= P\left(H_0|H_1'\right) = P\left(\overline{X} < \mu_c|\mu = 12\right) \end{aligned}$$

Ejercicio 23 *Para cada una de las siguientes afirmaciones redacte las hipótesis nula y alternativa, identifique el estadístico a utilizar y las regiones de aceptación y rechazo. Además redacte las probabilidades del error tipo I y del error tipo II para la hipótesis alternativa específica dada.*

1. *Se afirma que* $\mu < 5$, $H_1' : u = 4.5$.
2. *Se desea determinar si* $p \leq 0.4$, $H_1' : p = 0.37$.

3. *Se quiere concluir que* $\sigma^2 \neq 9$, $H_1' : \sigma^2 = 10$.

4. *Se afirma que* $\mu_1 - u_2 > 5$, $H_1' : \mu_1 - u_2 = 6$.

El ejemplo siguiente da evidencia de que α es la medida de la región de rechazo cuando se asume que H_0 es verdadera.

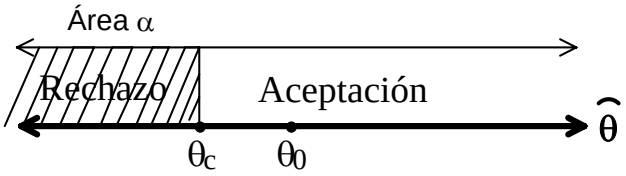
Ejemplo 45 *Considere la prueba*

$$H_0 : \theta = \theta_0 \quad (\geq) \qquad H_1 : \theta < \theta_0$$

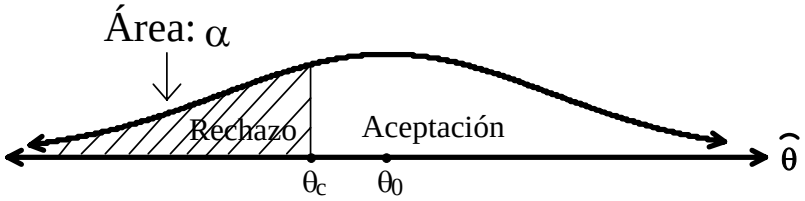
En este caso, la probabilidad del error tipo I es

$$\alpha = P\left(H_1 | H_0\right) = P\left(\widehat{\Theta} < \theta_c | \theta = \theta_0\right)$$

Así, α *es el área acumulada bajo la distribución de* $\widehat{\Theta}$ *hasta el valor crítico si se supone que* $\theta = \theta_0$*:*



Por simplicidad se bosqueja la distribución de $\widehat{\Theta}$ *como una línea horizontal. Sin embargo, el lector debe ser consciente de este hecho, por ejemplo si* $\widehat{\Theta}$ *es normal entonces realmente se tiene que*



Teorema 32 *La probabilidad del error Tipo I es la medida de la región o regiones de rechazo, asumiendo la hipótesis nula.*

El error tipo I es mayor cuando el valor crítico θ_c es cercano al valor nulo θ_0, pues aumenta la posibilidad de que estimaciones de θ cercanas a θ_0 sean consideradas significativas para rechazar H_0, corriendo el riesgo de que H_0 se verdadera. Véase esto en el siguiente ejemplo.

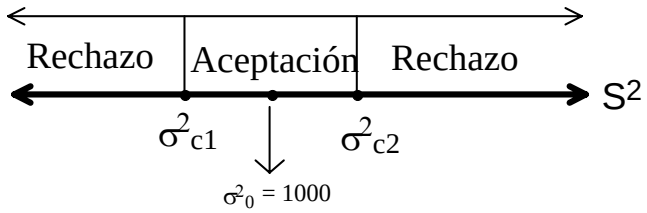
Ejemplo 46 *Se quiere concluir que la varianza* σ^2 *en la duración de las llantas FIRE es igual a* 1000 km^2.

En este caso se tiene que la afirmación es

$$afirmación : \sigma^2 = 1000$$

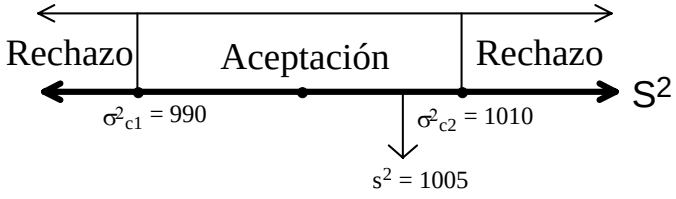
Recuerde que el estadístico asociado a σ^2 *es* S^2. *Así se tiene que*

$H_0 : \sigma^2 = 1000 \quad (\leq)$
$H_1 : \sigma^2 \neq 1000$



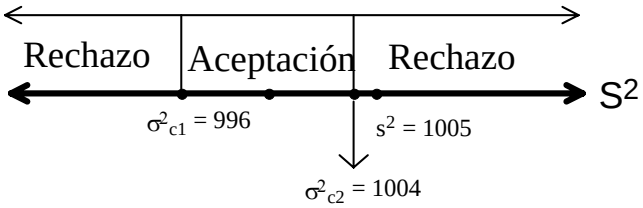
Suponga que en una muestra se determinó que una estimación de $\sigma^2 : s^2 = 1005$. *¿Es esta estimación significativa para rechazar* H_0?

1. *Si los valores críticos están alejados del valor nulo* σ_0^2. *Suponga que por algún método, que se verá más adelante, se determinan los valores críticos* $\sigma_{c1}^2 = 990$ *y* $\sigma_{c2}^2 = 1010$:



Dado que s^2 *está en la región de aceptación, entonces no se considera significativo para rechazar* H_0. *Por lo tanto se acepta* H_0.

2. *Si los valores críticos están cercanos del valor nulo* σ_0^2. *Suponga que los valores críticos son* $\sigma_{c1}^2 = 996$ *y* $\sigma_{c2}^2 = 1004$:



Dado que s^2 *está en la región de rechazo, entonces se considera evidencia significativa en contra de* H_0. *Por lo tanto se rechaza* H_0.

Comparando (1) *y* (2), *se puede notar que el error tipo I es mayor en* (2) *esto pues:*

(a) *En* (2) *aumenta la probabilidad de que estimaciones como* $1005, 1004.1$ *sean consideradas significativas para rechazar* H_0, *esto no ocurre en* (1). *Así, en* (2) *hay más posibilidad de rechazar* H_0 *y por lo tanto de cometer el error tipo I.*

(b) *Utilizando el teorema anterior, las regiones de rechazo son mayor en* (2) *que en* (1) , *por lo tanto el error tipo I es mayor en* (2) .

En cierta medida los errores son opuestos, si se tratada de evitar el error tipo I (**"***Condenar a un inocente***"**) se puede caer en el error tipo II (**"***Dejar libre al culpable***"**). Esto hace que, solo se pueda controlar uno de los errores, el más grave.

¿Cuál de los errores es más grave? En un juicio se prefiere dejar libre a un culpable que condenar un inocente, es decir se considera más grave condenar a un inocente. Recuerde que *"el acusado es inocente hasta que se demuestre lo contrario",* entonces se controla el error tipo I, se acepta inicialmente H_0 (**"***el acusada es inocente***"**) y si no hay realmente evidencia significativa para rechazar H_0, se prefiere **"***dejar libre al acusado***"** y evitar el error tipo I.

Así, en pruebas de hipótesis se controla el error tipo I por medio del nivel de significancia, que se define seguidamente.

Definición 27 ***Nivel de significancia****: es el valor máximo aceptable para* α, *la probabilidad de que ocurra el error tipo I.*

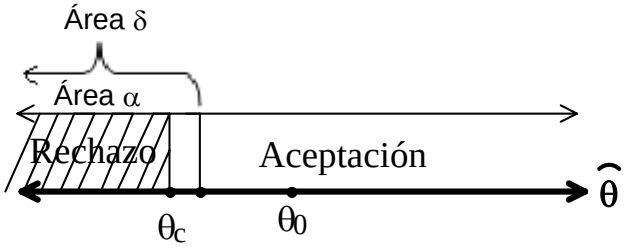
Por lo general, el nivel de significancia de una prueba es dado. Note que formalmente este valor no es la probabilidad α del error tipo I, sino una cota superior para α. Sin embargo, si se considera α igual al nivel de significancia, entonces θ_c será el valor crítico permitido que genere el máximo error tipo I aceptable y por lo tanto el más cercano a θ_0. Además, bajo esta consideración, el nivel de significancia es la medida de la región de rechazo.

Ejemplo 47 *Considere la prueba*

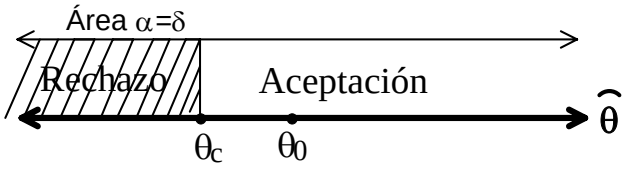
$$H_0 : \theta = \theta_0 \quad (\geq) \qquad H_1 : \theta < \theta_0$$

Denotemos con δ *el nivel de significancia y con* α *la probabilidad del error tipo I. Se tiene que*

Si $\alpha < \delta$:



Si $\alpha = \delta$:



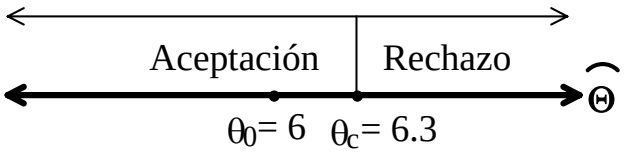
Note que entre más cerca este θ_c de θ_0, como se comento anteriormente, aumenta la probabilidad de cometer el error tipo I. Esto se debe a que se reduce la tolerancia para aceptar que $\theta = \theta_0$ a partir de estimaciones de θ. Por otro lado, lo más cerca que puede estar θ_c de θ_0 es cuanto el error tipo I ha llegado a su máximo permitido, esto sucede si $\alpha = \delta$.

Definición 28 ***Potencia de una prueba*** *para una hipótesis alternativa específica H_1' es la probabilidad de que no ocurra el error tipo II.*

El nivel de significancia es escogido dependiendo de las exigencias con respecto al error tipo I. Como se ha mencionado, las pruebas de hipótesis que se estudiarán asumen un nivel de significancia y por lo tanto controlan el error tipo I. Por el contrario el error tipo II no puede controlarse en general y solo puede calcularse para valores específicos de H_1.

Ejemplo 48 *Considere la siguiente prueba de hipótesis*

$H_0 : \theta = 6 \quad (\leq)$
$H_1 : \theta > 6$



donde $\theta_c = 6.3$ y además $\widehat{\Theta} \sim N(\theta, 0.7)$. En este caso, la probabilidad del error tipo I es

$$\begin{aligned} \alpha &= P(H_1|H_0) = P\left(\widehat{\Theta} > 6.3|\theta = 6\right) = P\left(Z > \frac{6.3-6}{\sqrt{0.7}}\right) \\ &= P(Z > 0.358\,569) \approx 0.359424 = 35.9424\%. \end{aligned}$$

Por otro lado, la probabilidad del error tipo II para la hipótesis alternativa específica $H_1' : \theta = 6.5$ es

$$\begin{aligned} \beta &= P(H_0|H_1') = P\left(\widehat{\Theta} < 6.3|\theta = 6.5\right) = P\left(Z < \frac{6.3-6.5}{\sqrt{0.7}}\right) \\ &= P(Z < -0.239\,046) \approx 0.405165 = 40.5165\% \end{aligned}$$

Ejercicio 24 *Se afirma que $\theta > 4$ y suponga que el estadístico $\widehat{\Theta}$ se distribuye normalmente*

$$\widehat{\Theta} \sim N(\theta, 0.5)$$

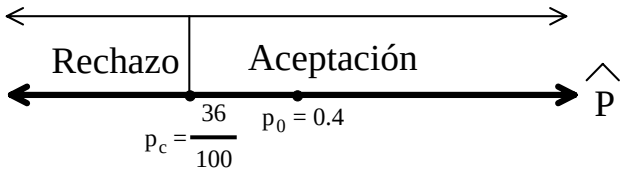
para muestras de un tamaño dado y que el valor crítico es $\theta_c = 4.2$. Plantee las hipótesis, identifique las regiones, y determine las probabilidades del error tipo I y del error tipo II para la hipótesis alternativa específica $H_1' : \theta = 4.5$. $R/ \quad \alpha = 0.3897, \beta = 0.3372$

Ejemplo 49 *Se cree que al menos el 40% de las amas de casa de una ciudad utilizan el detergente X. Se toma una muestra aleatoria de 100 amas de casa y si más de 36 utilizan el detergente, se aceptará que el porcentaje poblacional es de al menos el 40%. Recuerde que $\widehat{P}$ es insesgado, y suponga que $\widehat{P} \sim N\left(p, \frac{pq}{100}\right)$.*

1. *Determine las regiones de aceptación y rechazo*

$$\begin{array}{c} Afirmación : p \geq 0.4 \\ H_0 : p = 0.4 \quad (\geq) \qquad H_1 : p < 0.4 \end{array}$$

El tamaño de muestra es $n = 100$ *y el estadístico es* $\widehat{P}$. *Note que se dice que si la estimación dada por* $\widehat{P}$ *es mayor a* 36 *de* 100, *entonces se acepta* H_0. *Por lo tanto el valor crítico es* $p_c = \dfrac{36}{100}$:



2. *¿cuál es la probabilidad del error tipo I de la prueba?*

$$\alpha = P\left(H_1|H_0\right) = P\left(\widehat{P} < \frac{36}{100}|p = 0.4\right)$$

Dado que $\widehat{P} \sim N\left(p, \dfrac{pq}{100}\right)$ *y que en esta probabilidad se asume que* $p = 0.4$ *(y que* $q = 1-p = 0.6$*) entonces*

$$\alpha = P\left(Z < \frac{0.36 - 0.4}{\sqrt{\dfrac{0.4 \cdot 0.6}{100}}}\right) = P\left(Z < -0.816\,497\right) \approx 0.206108 = 20.6108\%.$$

3. *Si el porcentaje de amas de casa que utilizan el detergente* X *es en realidad 35%, ¿cuál es la potencia de la prueba?*

En este caso la hipótesis alternativa específica es

$$H_1' : p = 0.35.$$

La probabilidad del error tipo II asociado a está hipótesis es

$$\beta = P\left(H_0|H_1'\right) = P\left(\widehat{P} > 0.36 \middle| p = 0.35\right)$$

Dado que $\widehat{P} \sim N\left(p, \dfrac{pq}{100}\right)$ *y que en esta probabilidad se asume que* $p = 0.35$ *(y que* $q = 1 - p = 0.65$*) entonces*

$$\beta = P\left(Z > \frac{0.36 - 0.35}{\sqrt{\dfrac{0.35 \cdot 0.65}{100}}}\right) = P\left(Z > 0.209\,657\right) \approx 0.416834 = 41.6834\%$$

Por lo tanto la potencia de la prueba es

$$1 - \beta \approx 0.583\,166 = 58.3\,166\%$$

Ejercicio 25 *Se cree que el 70% de los habitantes de una ciudad están a favor de un proyecto de ley. Se toma una muestra aleatoria de 50 habitantes, y si más de 30 pero menos de 40 de ellos están a favor, se aceptará que el porcentaje a favor es de 70%. Recuerde que $\widehat{P}$ es insesgado, y suponga que $\widehat{P} \sim N\left(p, \frac{pq}{50}\right)$.*

1. *¿cuál es la probabilidad del error tipo I de la prueba?* *R/* 0.1236
2. *Si el porcentaje a favor es en realidad 55%, ¿cuál es la potencia de la prueba?* *R/* 0.7614

Como se mencionó los errores son en cierta medida opuestos. Dada que el valor contralado es α, entonces sucede uno de los siguientes efectos: si se disminuye α entonces aumenta β, si se aumenta α entonces disminuye β. Sin embargo, se pueden disminuir ambos si se aumenta el tamaño de la muestra, de la cuál se obtiene la estimación de θ. Esto se verá más adelante para ciertas pruebas de hipótesis.

1.5 Contraste de hipótesis

¿Cómo realizar una prueba de hipótesis? Para realizar el contraste de hipótesis existen dos enfoques que se explican seguidamente.

1.5.1 Enfoque clásico o determinación de regiones

Este enfoque consiste en hallar el o los valores críticos a partir del nivel de significancia dado. El enfoque sigue los siguientes pasos

Paso 1. Redactar H_0 y H_1

Paso 2. Especificar el criterio de prueba o contraste:

a) Nivel de significancia α
b) Distribución muestral de $\widehat{\Theta}$
c) Identificar las regiones de aceptación y rechazo tomando α como la probabilidad del error tipo I

Como se discutió antes, al tomar el nivel de significancia como la probabilidad el error tipo I, se está analizan el caso límite en que es más probable cometer el error de rechazar $H_0 : \theta = \theta_0$ para estimaciones de θ cercanas a θ_0.

Paso 3. Determinar el o los valores críticos, de acuerdo a el criterio de contraste (paso 2) . Para ello, se redacta la probabilidad del error tipo I.

Paso 4. Datos muestrales: hallar un valor $\widehat{\theta}$ de $\widehat{\Theta}$

Paso 5. Decisión. Sea R la región de rechazo

$$\text{Si } \widehat{\theta} \in R \text{ entonces se rechaza } H_0$$
$$\text{Si } \widehat{\theta} \notin R \text{ entonces se acepta } H_0$$

Paso 6. Conclusión

Si se rechaza H_0, se dice que si hay evidencia para rechazarla
Si se acepta H_0, se dice que no hay evidencia significativa en contra de H_0

Ejemplo 50 *Dada una población se afirma que uno de sus parámetros θ no es menor a 4. Suponga que el estadístico $\widehat{\Theta}$ asociado a θ se distribuye normalmente*

$$\widehat{\Theta} \sim N(\theta, 0.25)$$

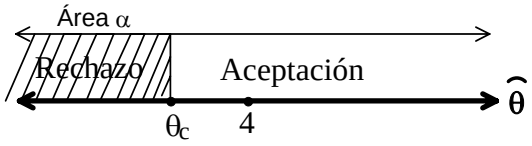
para muestras de un tamaño n. Además para una muestra de ese tamaño se observo un valor del estadístico de $\widehat{\theta} = 3$. Realice el contraste de hipótesis utilizando el enfoque clásico con un nivel de significancia de 0.1.

1. ***Redacción*** *de H_0 y H_1*

$$\textit{Afirmación: } \theta \geq 4$$
$$H_0 : \theta = 4\,(\geq), \quad H_1 : \theta < 4$$

2. ***Criterio de contraste:***

   a) *Nivel de significancia $\alpha = 0.1$*
   b) *Distribución muestral de $\widehat{\Theta} \sim N$*
   c) *Regiones:*



3. ***Determinación del valor crítico***

$$0.1 = \alpha = P(H_1|H_0) = P\left(\widehat{\Theta} < \theta_c|\theta = 4\right) = P\left(Z < \frac{\theta_c - 4}{\sqrt{0.25}}\right)$$

Así

$$P\left(Z < \frac{\theta_c - 4}{\sqrt{0.25}}\right) = 0.1 \Longrightarrow \frac{\theta_c - 4}{\sqrt{0.25}} = -1.28 \Longrightarrow \theta_c \approx 3.36$$

Por lo tanto las regiones son

$$\textit{Región de Aceptación: } A =]3.36, +\infty[, \quad \textit{Región de Rechazo: } R =]-\infty, 3.36[$$

4. ***Datos muestrales:*** $\widehat{\theta} = 3$.

5. ***Decisión***. *Como* $\widehat{\theta} = 3 \in R$ *entonces se rechaza* H_0.

6. ***Conclusión***. *Hay evidencia significativa para rechazar* H_0. *Dado que la afirmación se encuentra en* H_0, *entonces hay evidencia significativa para rechazar la afirmación.*

Ejercicio 26 *Se afirma que* $\theta > 4$ *y suponga que el estadístico* $\widehat{\Theta}$ *se distribuye normalmente*

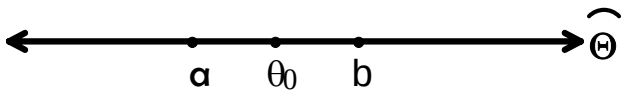
$$\widehat{\Theta} \sim N(\theta, 0.5)$$

para muestras de un tamaño dado. En una muestra de ese tamaño se observo un valor del estadístico de $\widehat{\theta} = 5$. *Realice el contraste de hipótesis con un nivel de signicancia de* 0.05.

1.5.2 Valor P de la prueba

Definición 29 ***El valor* P** *de una prueba de hipótesis es la probabilidad de que el estadístico* $\widehat{\Theta}$ *tenga el valor observado en la muestra o un valor extremo (en dirección de* H_1*) cuando* H_0 *es verdadera. Es decir es el nivel de significancia más bajo en el que el valor observado es significativo para rechazar* H_0.

En otras palabras, el valor P, indica el valor mínimo de la posibilidad de que ocurra el error tipo I cuando el valor observado se considera parte de la región de rechazo. Note que el valor P está relacionado con el nivel de significancia. La redacción del valor P depende del tipo de prueba:

Hipótesis	**Valor P**
$H_0 : \theta = \theta_0 \quad (\geq)$ $H_1 : \theta < \theta_0$	$P\left(\widehat{\Theta} \leq \widehat{\theta} \mid H_0\right)$
$H_0 : \theta = \theta_0 \quad (\leq)$ $H_1 : \theta > \theta_0$	$P\left(\widehat{\Theta} \geq \widehat{\theta} \mid H_0\right)$
$H_0 : \theta = \theta_0$ $H_1 : \theta \neq \theta_0$	Sean a y b los valores dador por el valor observado $\widehat{\theta}$ y su simétrico respecto a θ_0 :  Valor $P = P\left(\widehat{\Theta} < a \mid H_0\right) + P\left(\widehat{\Theta} > b \mid H_0\right)$ Así, si $\widehat{\Theta}$ tiene una distribución simétrica respecto a θ_0 entonces: Valor $P = \begin{cases} 2P\left(\widehat{\Theta} < \widehat{\theta} \mid H_0\right) & si \quad \widehat{\theta} < \theta_0 \\ 2P\left(\widehat{\Theta} > \widehat{\theta} \mid H_0\right) & si \quad \widehat{\theta} > \theta_0 \end{cases}$

Ejemplo 51 *Se afirma que* $\mu > 5$. *y se observa una muestra un valor de* $\overline{x} = 4.5$. *Entonces se tiene que*

$$\begin{gathered} \textit{Afirmación}: \mu > 5 \\ H_0 : \mu = 5\,(\leq), \quad H_1 : \mu > 5 \\ \textit{Valor } P = P\left(\widehat{\Theta} \geq \widehat{\theta} \mid H_0\right) = P\left(\overline{X} \geq 4.5 | \mu = 5\right) \end{gathered}$$

Ejemplo 52 *Se afirma que* $\sigma^2 = 9$. *y se observa una muestra un valor de* $s = 2.8$. *Entonces se tiene que*

$$\begin{gathered} \textit{Afirmación: } \sigma^2 = 9 \\ H_0 : \sigma^2 = 9, \quad H_1 : \sigma^2 \neq 9 \\ \textit{El valor observado es } s^2 = 7.84 \textit{ y valor simétrico a este} \\ \textit{respecto a } 9 \textit{ es } 10.16 \textit{ entonces} \\ \textit{Valor } P = P\left(S^2 < 7.84 \mid \sigma^2 = 9\right) + P\left(S^2 > 10.16 \mid \sigma^2 = 9\right) \end{gathered}$$

Ejercicio 27 *Para las siguientes afirmaciones redacte las hipótesis y el valor P.*

1. *Se afirma que* $\mu < 15$. *Valor observado* $\overline{x} = 14.1$
2. *Se desea determinar si* $p \geq 0.4$. *Valor observado* $\widehat{p} = 0.5$
3. *Se afirma que* $\mu_1 - u_2 > 5$. *Valor observado* $\overline{x}_1 - \overline{x}_2 = 4.9$

Para este enfoque se siguen los siguientes pasos

Paso 1. Redactar H_0 y H_1

Paso 2. Valor observado (datos muestrales): hallar un valor $\widehat{\theta}$ de $\widehat{\Theta}$

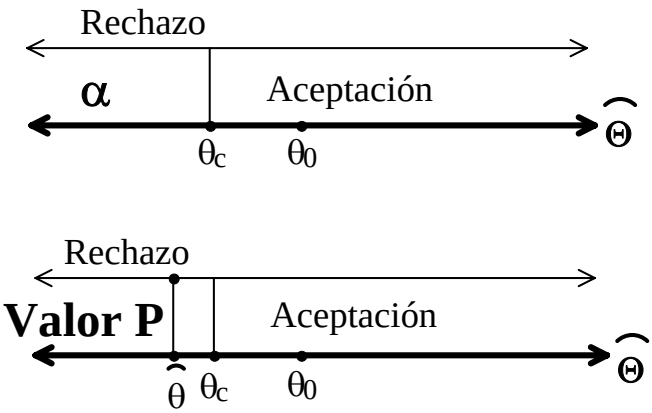
Paso 3. Redacción y determinación del valor P.

Paso 4. Decisión.

a) **Si se conoce el nivel de significancia** α.

$a.1$) **Si el valor** P **es menor o igual a** α. Aceptando H_0, se tiene que la probabilidad de que $\widehat{\Theta}$ tenga el valor observado en dirección a H_1 es menor que α, el área de la región de rechazo, por ejemplo

$$H_0 : \theta = \theta_0 \quad (\geq), \qquad H_1 : \theta < \theta_0$$



Por lo tanto el valor observado está en la región de rechazo y se rechaza H_0.

$a.2$) **Si el valor P es mayor a α**. En este caso, por justificaciones similares, se acepta H_0.

En resumen:

Si el Valor $P \leq \alpha$ entonces se rechaza H_0
Si el Valor $P > \alpha$ entonces se acepta H_0

b) **Si no se conoce el nivel de significancia α**. Se utiliza $\alpha = 0.05$.

Paso 5. Conclusión

Si se rechaza H_0, se dice que si hay evidencia para rechazarla
Si se acepta H_0, se dice que no hay evidencia significativa a favor de H_1

Ejemplo 53 *Dada una población se afirma que uno de sus parámetros θ no es menor a 4. Suponga que el estadístico $\widehat{\Theta}$ asociado a θ se distribuye normalmente*

$$\widehat{\Theta} \sim N(\theta, 0.25)$$

para muestras de un tamaño n. Además para una muestra de ese tamaño se observo un valor del estadístico de $\widehat{\theta} = 3$. Realice el contraste de hipótesis utilizando Valor P con un nivel de significancia de 0.1.

1. ***Redacción** de H_0 y H_1*

$$\text{Afirmación: } \theta \geq 4$$
$$H_0 : \theta = 4\,(\geq), \quad H_1 : \theta < 4$$

2. ***Valor observado**: $\widehat{\theta} = 3$.*

3. ***Valor** P :*

$$\text{Valor } P = P\left(\widehat{\Theta} \leq 3 | \theta = 4\right) = P\left(Z \leq \frac{3-4}{\sqrt{0.25}}\right) = P\left(Z \leq -2\right) = 0.022750$$

4. ***Decisión**. Como $Valor\ P < 0.1 = \alpha$ entonces se rechaza H_0.*

5. ***Conclusión**. Hay evidencia para rechazar H_0, y como la afirmación se encuentra en H_0, se rechaza la afirmación.*

Ejercicio 28 *Se afirma que $\theta > 4$ y suponga que el estadístico $\widehat{\Theta}$ se distribuye normalmente*

$$\widehat{\Theta} \sim N(\theta, 0.5)$$

para muestras de un tamaño dado. En una muestra de ese tamaño se observo un valor del estadístico de $\widehat{\theta} = 5$. Realice el contraste de hipótesis utilizando el valor P.

1.5.3 ¿Cuál de los dos enfoques utilizar para contrastar la hipótesis?

Ambos enfoques son equivalentes, sin embargo al compararlos se tiene que

Enfoque	**Utilización de valor muestral $\widehat{\theta}$**	**Utilización del nivel de significancia α**
Determinación de regiones	Al final, para ver en qué región se encuentra $\widehat{\theta}$	Al inicio, para calcular las regiones
Valor P	Al inicio, para calcular el Valor P	Al final, al comparar α y el Valor P

Así se tiene que

1. Se utiliza el valor P cuando se conoce el valor muestral y no se tiene certeza sobre que nivel de significancia utilizar. Por ejemplo, si el valor P es 0.007, esto me indica que puedo utilizar un nivel de significancia bastante pequeño para rechazar la hipótesis como 0.01, en lugar de utilizar un nivel de significancia a priori del 10% por ejemplo. Note que al reportar un valor pequeño de α en una prueba que se rechaza, la conclusión tiene una mejor aceptación. Particularmente el valor P es útil cuando se tiene un valor muestral difícil de volver a conseguir, ya sea por razones económicas o de tiempo por ejemplo, y se quiere manipular el nivel de significancia.

2. Se utiliza la determinación de regiones cuando aún no se tiene el valor muestral y se tiene certeza del nivel de significancia a utilizar. Así, con el nivel de confianza se pueden determinar las regiones de prueba y luego tomar la muestra para obtener un valor muestral y ubicarlo en alguna de las regiones. Se puede dar un efecto negativo y no ético de la estadística, que es la manipulación del valor muestral para ajustarlo a conveniencia.

Para fines prácticos de este documento, se utiliza el valor P cuando el ejercicio no indique el nivel de significancia a utilizar. En caso contrario, el lector puede escoger el enfoque a utilizar.

1.6 Cambio de estadístico

Definición 30 *Sea $\widehat{\Theta}$ un estadístico asociado a θ para muestras de tamaño n. Considere el estadistico $\overline{\Theta} = f\left(\widehat{\Theta}\right)$ asociado a $f(\theta)$ para muestras de tamaño n. Se dice que $\widehat{\Theta}$ y $\overline{\Theta}$ tienen el mismo comportamiento si y solo si f es una función estrictamente creciente*

Teorema 33 *Sean $\widehat{\Theta}$ un estadístico asociado a θ para muestras de tamaño n, y $\overline{\Theta} = m\widehat{\Theta} + b$ con m constante positiva y b constante. Entonces $\widehat{\Theta}$ y $\overline{\Theta}$ tienen el mismo comportamiento.*

Prueba. *Note que $\overline{\Theta} = f\left(\widehat{\Theta}\right)$ con $f(x) = mx + b$, como $m > 0$ entonces f es estrictamente creciente. Por lo tanto, $\widehat{\Theta}$ y $\overline{\Theta}$ tienen el mismo comportamiento.* ■

Ejemplo 54 *Recuerde que $\overline{X}$ es un estadístico asociado a μ. Note que $\overline{X}$ y $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ tiene el mismo comportamiento.*

En efecto, note que

$$Z = \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sigma} = \frac{\sqrt{n}}{\sigma}\overline{X} + \frac{-\mu\sqrt{n}}{\sigma}$$

Dado que $\frac{\sqrt{n}}{\sigma} > 0$ *pues* $\sqrt{n} > 0$ *y* $\sigma > 0$ *entonces, por el teorema anterior,* $\overline{X}$ *y* Z *tienen el mismo comportamiento.*

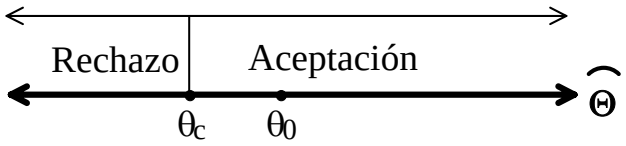
Si $\widehat{\Theta}$ y $\overline{\Theta} = f\left(\widehat{\Theta}\right)$ tienen el mismo comportamiento, esto puede ser utíl para cambiar de estadístico en ambos enfoques, si se conoce la distribución del nuevo estadístico $\overline{\Theta}$.

Ejemplo 55 *Considere la siguiente prueba de hipótesis*

$$H_0 : \theta = \theta_0 \quad (\geq) \qquad H_1 : \theta < \theta_0$$

Sea $\widehat{\Theta}$ *un estadístico asociado a* θ *para muestras de tamaño* n*, y suponga que* $\widehat{\Theta}$ *y* $\overline{\Theta} = f\left(\widehat{\Theta}\right)$ *tienen el mismo comportamiento. Entonces* f *es estrictamente creciente.*

1. *Utilizando regiones.*



Note que las regiones de aceptación y rechazo utilizando el estadístico $\widehat{\Theta}$ *son respectivamente:*

$$A_1 =]-\infty, \theta_c[\,, \quad R_1 =]\theta_c, +\infty[$$

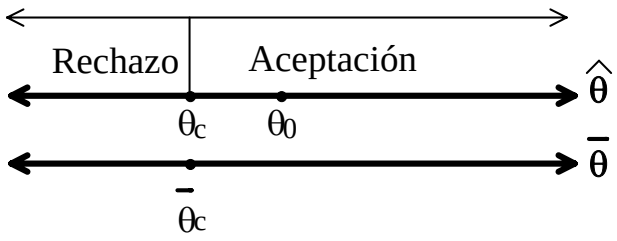
Sea $\overline{\theta}_c = f\left(\theta_c\right)$*, se define las regiones de aceptación y rechazo utilizando el estadístico* $\overline{\Theta}$ *por:*

$$A_2 = \left]-\infty, \overline{\theta}_c\right[, \quad R_2 = \left]\overline{\theta}_c, +\infty\right[$$

Dado que f *es estrictamente creciente note que*

$$\begin{aligned} \widehat{\Theta} \quad &\in \quad A_1 \Leftrightarrow \widehat{\Theta} < \theta_c \Leftrightarrow f\left(\widehat{\Theta}\right) < f\left(\theta_c\right) \Leftrightarrow \overline{\Theta} < \overline{\theta}_c \\ &\Leftrightarrow \quad \overline{\Theta} = f\left(\widehat{\Theta}\right) \in A_2 \qquad m \in (1) \end{aligned}$$

Similarmente, $\widehat{\Theta} \in R_1 \Leftrightarrow \overline{\Theta} \in R_2$*. Esto nos permite cambiar de estadístico:*



Así, se puede utilizar el estadístico $\overline{\Theta}$, *y hallar sus regiones* A_2 *y* R_2. *Luego, dado el valor muestral* $\widehat{\theta}$ *se calcula* $f\left(\widehat{\theta}\right)$ *que es el valor observado para* $\overline{\Theta}$, *si* $f\left(\widehat{\theta}\right) \in A_2$, *por* (1) *se tiene que* $\widehat{\theta} \in A_1$ *y se acepta* H_0. *Así,*

$$\begin{array}{c} \text{si } f\left(\widehat{\theta}\right) \in A_2, \text{ entonces se acepta } H_0 \\ \text{y si } f\left(\widehat{\theta}\right) \in R_2, \text{ entonces se rechaza } H_0 \end{array}$$

2. *Utilizando valor P. Sea* $\widehat{\theta}$ *el valor observado en una muestra de tamaño n. Como*

$$H_0 : \theta = \theta_0 \quad (\geq) \qquad H_1 : \theta < \theta_0$$

entonces el valor P es

$$P\left(\widehat{\Theta} \leq \widehat{\theta} | \theta = \theta_0\right)$$

Note que dado que f es estrictamente creciente:

$$\widehat{\Theta} \leq \widehat{\theta} \quad \Leftrightarrow \quad f\left(\widehat{\Theta}\right) \leq f\left(\widehat{\theta}\right) \quad \Leftrightarrow \quad \overline{\Theta} \leq f\left(\widehat{\theta}\right)$$

Así, el valor P se puede redefinir utilizando el nuevo estadístico:

$$\text{Valor } P = P\left(\overline{\Theta} \leq \widehat{\theta} | \theta = \theta_0\right)$$

Ejemplo 56 *Dada una población se afirma que uno de sus parámetros* θ *no es menor a* 4. *Suponga que el estadístico* $\widehat{\Theta}$ *asociado a* θ *se distribuye normalmente*

$$\widehat{\Theta} \sim N\left(\theta, 0.25\right)$$

para muestras de un tamaño n. Además para una muestra de ese tamaño se observo un valor del estadístico de $\widehat{\theta} = 3$. *Realice el contraste de hipótesis utilizando el enfoque clásico con un nivel de significancia de* 0.1.

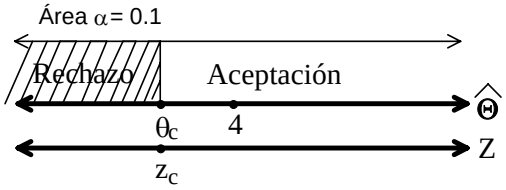
1. ***Redacción*** *de* H_0 *y* H_1

$$\begin{array}{c} \textit{Afirmación: } \theta \geq 4 \\ H_0 : \theta = 4\,(\geq), \quad H_1 : \theta < 4 \end{array}$$

2. ***Criterio de contraste****: El nivel de significancia* $\alpha = 0.1$. *Dado que* $\widehat{\Theta} \sim N\left(\theta, 0.25\right)$ *entonces*

$$Z = \frac{\widehat{\Theta} - \theta}{\sqrt{0.25}} \sim N\left(0, 1\right)$$

Como $\widehat{\Theta}$ *y Z tiene el mismo comportamiento, entonces se puede utilizar Z como estadístico de la prueba. Así las regiones son*



donde $z_c = \dfrac{\theta_c - \theta}{\sqrt{0.25}}$

3. ***Determinación del valor crítico.*** *Note que* z_c *acumula una área de* 0.1*, por lo tanto*

$$z_c = z_{0.1} = -1.28$$

Así las regiones de aceptación y rechazo utilizando Z *son*

$$\textit{Región de Aceptación: } A =]-1.28, +\infty[\,, \quad \textit{Región de Rechazo: } R =]-\infty, -1.28[$$

4. ***Datos muestrales****: El valor observado es* $\widehat{\theta} = 3$ *para* $\widehat{\Theta}$*. Por lo tanto el valor observado para* Z*, asumiendo* $H_0 : \theta = 4$*, es*

$$z_{obs} = \frac{\widehat{\theta} - \theta}{\sqrt{0.25}} = \frac{3-4}{\sqrt{0.25}} = -2$$

5. ***Decisión****. Como* $z_{obs} = -2 \in R$ *entonces se rechaza* H_0*.*

6. ***Conclusión****. Hay evidencia significativa para rechazar* H_0*. Dado que la afirmación se encuentra en* H_0*, entonces hay evidencia significativa para rechazar la afirmación.*

Ejemplo 57 *Dada una población se afirma que uno de sus parámetros* θ *no es menor a* 4*. Suponga que el estadístico* $\widehat{\Theta}$ *asociado a* θ *se distribuye normalmente*

$$\widehat{\Theta} \sim N(\theta, 0.25)$$

para muestras de un tamaño n*. Además para una muestra de ese tamaño se observo un valor del estadístico de* $\widehat{\theta} = 3$*. Realice el contraste de hipótesis utilizando Valor P con un nivel de significancia de* 0.1*.*

1. ***Redacción*** *de* H_0 *y* H_1

$$\textit{Afirmación: } \theta \geq 4$$
$$H_0 : \theta = 4\,(\geq)\,, \quad H_1 : \theta < 4$$

Dado que $\widehat{\Theta} \sim N(\theta, 0.25)$ *entonces*

$$Z = \frac{\widehat{\Theta} - \theta}{\sqrt{0.25}} \sim N(0, 1)$$

Como $\widehat{\Theta}$ *y* Z *tiene el mismo comportamiento, entonces se puede utilizar* Z *como estadístico de la prueba.*

2. ***Valor observado****: El valor observado es* $\widehat{\theta} = 3$ *para* $\widehat{\Theta}$*. Por lo tanto el valor observado para* Z*, asumiendo* $H_0 : \theta = 4$*, es*

$$z_{obs} = \frac{\widehat{\theta} - \theta}{\sqrt{0.25}} = \frac{3-4}{\sqrt{0.25}} = -2$$

3. ***Valor** P. Utilizando el estadístico Z se tiene que*

$$Valor\ P = P\left(Z \leq z_{obs}\right) = P\left(Z \leq -2\right) = 0.022750$$

4. ***Decisión**. Como $Valor\ P < 0.1 = \alpha$ entonces se rechaza H_0.*

5. ***Conclusión**. Hay evidencia para rechazar H_0, y como la afirmación se encuentra en H_0, se rechaza la afirmación.*

1.7 Ejercicios

1. Para cada una de las siguientes afirmaciones redacte las hipótesis nula y alternativa, identifique el estadístico a utilizar y las regiones de aceptación y rechazo. Además redacte las probabilidades del error tipo I y del error tipo II para la hipótesis alternativa específica dada.

   (a) Se afirma que $\mu \leq 6$, $H_1' : u = 7.5$.

   (b) Se desea determinar si $p > 0.7$, $H_1' : p = 0.8$.

   (c) Se quiere concluir que $\sigma^2 \neq 3$, $H_1' : \sigma^2 = 4$.

   (d) Se afirma que $p_1 - p_2 < 0$, $H_1' : p_1 - p_2 = 0.3$.

2. Considere la prueba de hipótesis

$$H_0 : \theta = \theta_0 \quad (\leq) \qquad H_1 : \theta > \theta_0$$

   Suponga que Juan realiza el contraste de hipótesis con un nivel de significancia del 0.1, obteniendo el valor crítico θ_{c1}. Por otro lado, Ana realiza el contraste de hipótesis con un nivel de significancia del 0.05 utilizando los mismos datos que Juan, obteniendo el valor crítico θ_{c2}. ¿Qué valor es mayor entre θ_{c1} y θ_{c2}?

3. Considere la prueba de hipótesis

$$H_0 : \theta = \theta_0 \qquad H_1 : \theta \neq \theta_0$$

   Suponga que Juan halla la región de aceptación A_1 utilizando un nivel de significancia del 0.1. Por otro lado, Ana halla la región de aceptación A_2 con un nivel de significancia del 0.05 utilizando los mismos datos que Juan. ¿Cuál de las siguientes afirmaciones es correcta:

$$a) A_1 \subseteq A_2 \quad b) A_2 \subseteq A_1 \quad c) \text{ Ninguna de las anteriores.}$$

4. Se afirma que $\theta \leq 7$ y suponga que el estadístico $\widehat{\Theta}$ se distribuye normalmente

$$\widehat{\Theta} \sim N\left(\theta, 0.5\right)$$

   para muestras de un tamaño dado y que el valor crítico es $\theta_c = 7.2$. Plantee las hipótesis, identifique las regiones, y determine las probabilidades del error tipo I y del error tipo II para la hipótesis alternativa específica $H_1' : \theta = 9$.

5. Se afirma que $\theta < 5$ y suponga que el estadístico $\widehat{\Theta}$ se distribuye normalmente

$$\widehat{\Theta} \sim N(\theta, 0.5)$$

para muestras de un tamaño dado. En una muestra de ese tamaño se observo un valor del estadístico de $\widehat{\theta} = 4.9$.

   (a) Realice el contraste de hipótesis utilizando la determinación de regiones sin utilizar cambio de estadístico

   (b) Realice el contraste de hipótesis utilizando el valor P sin utilizar cambio de estadístico.

   (c) Realice el contraste de hipótesis utilizando la determinación de regiones utilizando Z como estadístico.

   (d) Realice el contraste de hipótesis utilizando el valor P utilizando Z como estadístico.

6. Se afirma que $\theta = 6$ y suponga que el estadístico $\widehat{\Theta}$ se distribuye normalmente

$$\widehat{\Theta} \sim N(\theta, 0.5)$$

para muestras de un tamaño dado. En una muestra de ese tamaño se observo un valor del estadístico de $\widehat{\theta} = 6.1$.

   (a) Realice el contraste de hipótesis utilizando la determinación de regiones sin utilizar cambio de estadístico

   (b) Realice el contraste de hipótesis utilizando el valor P sin utilizar cambio de estadístico.

   (c) Realice el contraste de hipótesis utilizando la determinación de regiones utilizando Z como estadístico.

   (d) Realice el contraste de hipótesis utilizando el valor P utilizando Z como estadístico.

7. Justifique por qué los siguientes pares de estadísticos tiene el mismo comportamiento, donde es el tamaño de la muestra

   (a) $\overline{X}$ y $T = \dfrac{\overline{X} - \mu_0}{s/\sqrt{n}}$.

   (b) $\widehat{P}$ y $Z = \dfrac{\widehat{P} - p_0}{\sqrt{p_0 q_{0/n}}}$

   (c) S y S^2, para ello considere la función $f(x) = x^2$.

   (d) S^2 y $\chi^2 = \dfrac{(n-1)S^2}{\sigma_0^2}$

2 Pruebas de hipótesis con un parámetro

2.1 Pruebas con un promedio

La hipótesis nula, para este caso, es de la forma

$$H_0 : \mu = \mu_0$$

Considere la población dada por la variable aleatoria X que con media poblacional μ y varianza poblacional σ^2. por resultados anteriores recuerde que

1. Si $\overline{X}$ sigue una distribución normal para muestras de tamaño n y se conoce σ entonces

$$\overline{X} \sim N\left(\mu, \sigma^2\right) \quad \Longrightarrow \quad Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N\left(0, 1\right)$$

2. Si la población X sigue una distribución normal con media poblacional μ y se desconoce σ entonces

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t\left(n-1\right)$$

Por lo tanto, bajo la hipótesis $H_0 : \mu = \mu_0$, se tiene que:

1. Si $\overline{X}$ sigue una distribución normal para muestras de tamaño n (la población X es normal o $n \geq 30$) y se conoce σ entonces

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

sigue una distribución normal estándar.[2]

2. Si la población X sigue una distribución normal con media poblacional μ y se desconoce σ entonces

$$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

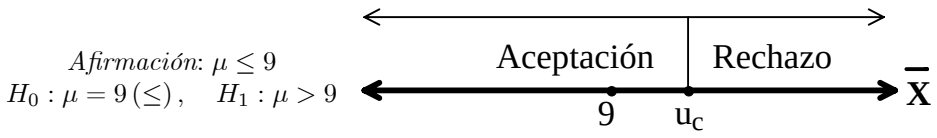
tiene una distribución t con $v = n - 1$ grados de libertad.

Ejemplo 58 ***(¿La directiva tiene razón?)*** *El sindicato de empleados de una empresa ubicada en una zona rural ha solicitado que la hora de entrada sea a las 7:15am y no a las 7am, esto debido a que la mayoría de los empleados no viven cerca de la empresa. La directiva de la empresa ha denegado la solicitud pues considera que no más de 9 kilómetros recorre diariamente el empleado promedio para llegar a la empresa. Se desea contrastar la afirmación emitida por dicha directiva con un nivel de significancia de 0.05. Para ello, en una muestra de 50 empleados, se observó un recorrido promedio de* 9.1 *km con una desviación estándar de* 5 *km.*

[2] En realidad, si $X \sim N$ entonces $\overline{X}$ sigue una distribución normal. Pero si $n \geq 30$ entonces $\overline{X}$ sigue una distribución que se aproxima a la normal, esto se denota $\overline{X} \underset{aprox}{\sim} N$.

1. *Plantee la hipótesis nula, la hipótesis alternativa, y las regiones de aceptación y rechazo.*

   Sea μ la distancia promedio, en kilómetros, que recorre diariamente el empleado para llegar a la empresa.



2. *Determine las regiones de aceptación y rechazo.*

   Dado que el tamaño de muestra es $n = 50$ entonces

$$\overline{X} \underset{aprox}{\sim} N \quad y \quad \sigma \approx s = 5$$

   Además $\alpha = 0.05$, por lo tanto

$$\begin{aligned} & \alpha = P\left(H_1 | H_0\right) \\ \Longrightarrow \qquad & 0.05 = P\left(\overline{X} \geq \mu_c | \mu = 9\right) = P\left(Z > \frac{\mu_c - 9}{5/\sqrt{50}}\right) \\ \Longrightarrow \qquad & \frac{\mu_c - 9}{5/\sqrt{50}} = z_{0.95} = 1.645 \\ \Longrightarrow \qquad & \mu_c = 10.167 \end{aligned}$$

   Región de rechazo: $R =]10.\,163\,2, +\infty[$ y Región de aceptación: $A =]-\infty, 10.\,163\,2[$.

3. *¿Hay evidencia en contra de la afirmación de la directiva?*

   El valor observado es $\overline{x} = 9.1$, como $\overline{x} \in A$ entonces se acepta H_0. Dado que la afirmación de la directiva fue ubicada en H_0, entonces se acepta la afirmación, por lo tanto no se encontró evidencia significativa en contra de lo expresado por la directiva.

Ejemplo 59 ***(¿El sindicato tiene razón?)*** *Considere el ejemplo anterior. Ante lo expresado por la directiva, el sidicato señala que el empleado promedio recorre al menos 9.2 km para llegar a la empresa. Utilizando los mismos datos del ejemplo anterior, ¿aceptaría la afirmación del sindicato?*

Note que la afirmación de la directiva ($\mu \leq 9$) es opuesta a la afirmación del sidicato ($\mu \geq 9.2$). Como la afirmación de la directiva se considera aceptable (ejemplo anterior), se puede cometer el error de rechazar la afirmación del sindicato. En pruebas de hipótesis, cada afirmación debe ser analizada por separado. Analicemos la afirmación del sindicato

$$\begin{gathered} \text{Afirmación: } \mu \geq 9.2 \\ H_0 : \mu = 9.2\ (\geq)\,, \quad H_1 : \mu < 9.2 \end{gathered}$$

Como $n = 50$ entonces $\overline{X} \underset{aprox}{\sim} N$ y $\sigma \approx s = 5$. Entonces el valor P de la prueba es

$$\begin{aligned} \text{Valor } P \quad & = P\left(\overline{X} \leq \overline{x} | H_0\right) = P\left(\overline{X} \leq 9.1 | \mu = 9.2\right) \\ & = P\left(Z \leq \frac{9.1 - 9.2}{5/\sqrt{50}}\right) = P\left(Z \leq -0.141\,421\right) \\ & = 0.444330 \end{aligned}$$

Dado que Valor $P > 0.05$, se acepta H_0 y entonces se acepta la afirmación. Por lo tanto, no hay evidencia significativa para rechazar la afirmación del sindicato

Ejemplo 60 ***(¿La directiva o el sindicato?)*** *En los ejemplos anteriores, se aceptan las afirmaciones de la directiva de la empresa y del sindicato, pero estas afirmaciones son contradictorias, ¿Qué está sucediendo?*

Lo que sucede es que la información muestral es insificiente para rechazar alguna de las afirmaciones. Además se debe recordar que el concepto de aceptación en pruebas de hipótesis es distinto al usado comúnmente, ambas afirmaciones se aceptan por que no hay evidencias cignificativas en su contra. Aquí se recomienda hacer el estudio con una nueva muestra.

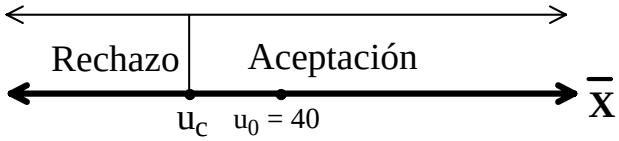
Ejemplo 61 *Una muestra aleatoria de diez estudiantes dio las siguientes cifras en horas para el tiempo que pasan estudiando durante la semana previa a los exámenes finales.*

$$28; 57; 42; 35; 61; 39; 55; 46; 49; 38.$$

Suponga que el tiempo de estudio durante la semana previa a los exámenes finales se distribuye normalmente. Un grupo de profesores considera que el tiempo medio debería ser como mínimo de 40 horas.

1. *Plantee la hipótesis nula, la hipótesis alternativa para la afirmación de los profesores, y las regiones de aceptación y rechazo.*

Afirmación: $u \geq 40$
$H_0 : u = 40\,(\geq), \quad H_1 : u < 40$



Note que, bajo H_0, $T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$ sigue una distribución t con $v = 9$ grados de libertad, pues la población es normal y $n = 10 < 30$.

2. *Pruebe si los profesores están en lo cierto con un nivel de significancia de 0.05, determinando las regiones de aceptación y rechazo.*

$$\alpha = P\left(H_1 | H_0\right) \quad \Longrightarrow \quad 0.05 = P\left(\overline{X} \leq u_c | u = 40\right)$$

Al asumir H_0, como $T = \dfrac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t\,(9)$, $n = 10$ y realizando el cálculo de la desviación muestral $s \approx 10.540\,9$, se tiene que

$$0.05 = P\left(T \leq \frac{u_c - 40}{10.540\,9/\sqrt{10}}\right) \Longrightarrow \frac{u_c - 40}{10.540\,9/\sqrt{10}} = t_{0.05,9}$$

Dado que $t_{0.05,9} = -1.83311$ *entonces*

$$\frac{u_c - 40}{10.540\,9/\sqrt{10}} = -1.83311 \Longrightarrow u_c = 33.889\,6$$

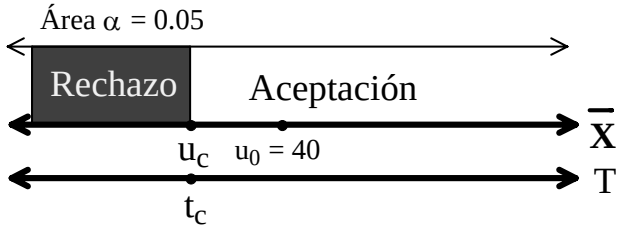
Región de rechazo: $R =]-\infty, 33.889\,6[$

Región de aceptación: $A =]33.889\,6, +\infty[$

Como $\overline{x} = 45 \in A$, *entonces se acepta* H_0 *y se acepta la afirmación. No hay evidencia significativa en contra de la afirmación.*

3. *Pruebe si los profesores están en lo cierto con un nivel de significancia de 0.05, determinando las regiones de aceptación y rechazo para el estadístico* T.

Dado que $T = \dfrac{\overline{X} - \mu_0}{s/\sqrt{n}}$, *entonces* T *y* $\overline{X}$ *tiene el mismo comportamiento:*



Como $T \sim t\,(9)$ *y el* t_c *acumula una área del* 5% *entonces* $t_c = t_{0.05,9} = -1.83311$, *por lo tanto las regiones para* T *son*

Región de rechazo: $R_t =]-\infty, -1.83311[$

Región de aceptación: $A_t =]-1.83311, +\infty[$

Dado que el valor de $\overline{X}$ *observado es* $\overline{x} = 45$ *entonces el valor* t *observado es*

$$t_{obs} = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{45 - 40}{10.540\,9/\sqrt{10}} \approx 1.5$$

Como $1.5 \in A_t$, *entonces se acepta* H_0 *y se acepta la afirmación.*

4. *Pruebe si los profesores están en lo cierto utilizando el valor P.*

Dado que $H_1 : u < 40$ *y el valor observado es* $\overline{x} = 45$ *entonces*

$$\textit{Valor } P = P\left(\overline{X} \leq \overline{x} | u = 40\right) = P\left(T \leq \frac{45 - 40}{10.540\,9/\sqrt{10}}\right) \approx P\,(T \leq 1.5)$$

Recuerde que $T \sim t\,(9)$, *entonces con ayuda de la tabla se puede acotar está probabilidad:*

$$\textit{Valor } P \approx 1 - \underbrace{P\,(T \geq 1.5)}_{\in]0.05, 0.1[} \in]0.9, 0.95[$$

Como el Valor $P > 0.9 > 0.05$ *entonces se acepta* H_0 *y se acepta la afirmación.*

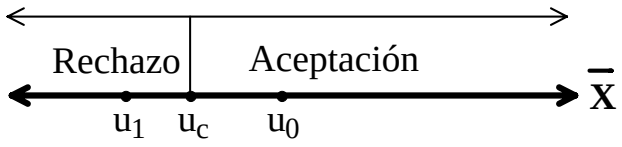
Por otro lado, suponga que se quiere hallar el tamaño de muestra n para probar

$$H_0 : \mu = \mu_0, \quad H_1 : \mu < \mu_0$$

con un nivel de significancia de α y una potencia mínima de $1-\beta$ para la hipótesis alternativa específica

$$H_1' : \mu = \mu_1$$

donde α y β son menores al 50%. Entonces las regiones de aceptación y rechazo son



y las probabilidades de que sucedan los errores tipo I y II son[3]:

$$\begin{aligned}
& \alpha \geq P\left(\overline{X} < \mu_c | \mu = \mu_0\right) \quad y \quad \beta \geq P\left(\overline{X} > \mu_c | \mu = \mu_1\right) \\
\Longrightarrow \quad & \alpha \geq P\left(Z < \frac{\mu_c - \mu_0}{\sigma/\sqrt{n}}\right) \quad y \quad \beta \geq P\left(Z > \frac{\mu_c - \mu_1}{\sigma/\sqrt{n}}\right) \\
\Longrightarrow^{4} \quad & \phi\left(z_\alpha\right) \geq \phi\left(\frac{\mu_c - \mu_0}{\sigma/\sqrt{n}}\right) \quad y \quad 1 - \beta \leq P\left(Z \leq \frac{\mu_c - \mu_1}{\sigma/\sqrt{n}}\right) \\
\Longrightarrow \quad & \phi\left(z_\alpha\right) \geq \phi\left(\frac{\mu_c - \mu_0}{\sigma/\sqrt{n}}\right) \quad y \quad \phi\left(z_{1-\beta}\right) \leq \phi\left(\frac{\mu_c - \mu_1}{\sigma/\sqrt{n}}\right) \\
\Longrightarrow \quad & z_\alpha \geq \frac{\mu_c - \mu_0}{\sigma/\sqrt{n}} \quad y \quad z_{1-\beta} \leq \frac{\mu_c - \mu_1}{\sigma/\sqrt{n}}
\end{aligned}$$

Se considere que μ_1 está en la región de rechazo por ser parte de la hipótesis alternativa específica, entonces $\mu_1 < \mu_c < \mu_0$. Así ambos extremos de la primer desigualdad son negativos, pues $\alpha < 0.5$, entonces

$$|z_\alpha| \leq \left|\frac{\mu_c - \mu_0}{\sigma/\sqrt{n}}\right| = \frac{\mu_0 - \mu_c}{\sigma/\sqrt{n}} \qquad (1)$$

En cuanto a la segunda desigualdad, como $z_{1-\beta} = -z_\beta$ entonces $z_\beta \geq -\left(\frac{\mu_c - \mu_1}{\sigma/\sqrt{n}}\right) = \frac{\mu_1 - \mu_c}{\sigma/\sqrt{n}}$ y como ambos extremos som negativos

$$|z_B| \leq \left|\frac{\mu_1 - \mu_c}{\sigma/\sqrt{n}}\right| = \frac{\mu_c - \mu_1}{\sigma/\sqrt{n}} \qquad (2)$$

[3] Note que formalmente α es la máxima probabilidad del error tipo 1. Está probabilidad es $P\left(\overline{X} \leq \mu_c | \mu = \mu_0\right)$, la cual debe ser menor igual que α. Además la potencia es mínimo $1-\beta$, entonces β es la máxima probabilidad del error tipo 2, es decir β es mayor igual a $P\left(\overline{X} \geq \mu_c | \mu = \mu_1\right)$.

[4] Recuerde que ϕ denota la distribución acumulada de la normal estándar

Sumando (1) y (2) se tiene que

$$|z_\alpha| + |z_B| \leq \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} \Longrightarrow |z_\alpha| + |z_B| \leq \frac{\sqrt{n}\,(\mu_0 - \mu_1)}{\sigma}$$

Por lo tanto,

$$n \geq \frac{(|z_\alpha| + |z_\beta|)^2\,\sigma^2}{(\mu_1 - \mu_0)^2}$$

Así, se realiza la demostración de el teorema siguiente para uno de los tres casos de pruebas de hipótesis. Los otros casos se demuestran de manera similar y se dejan como ejercicio al lector.

Teorema 34 *Si la población se distribuye aproximadamente normal con varianza poblacional σ^2, para probar*

$$H_0 : \mu = \mu_0$$

con un nivel de significancia de α y una potencia mínima de $1-\beta$ para la hipótesis alternativa específica

$$H_1' : \mu = \mu_1,$$

se debe tomar una muestra de tamaño n que cumpla

1. $n \geq \dfrac{(|z_\alpha| + |z_\beta|)^2\,\sigma^2}{(\mu_1 - \mu_0)^2}$ *si la prueba es de una cola.*
2. $n \geq \dfrac{\left(\left|z_{\alpha/2}\right| + |z_\beta|\right)^2\,\sigma^2}{(\mu_1 - \mu_0)^2}$ *si la prueba es de dos colas.*

Ejemplo 62 *Una empresa empaca cierto producto agrícola en sacos. Para la distribución del producto se requiere que el peso promedio de cada saco sea de al menos 100 libras, esto para disminuir la utilización de sacos y facilitar su transporte. La experiencia ha indicado que la desviación estándar del peso de un saco es de 2 libras. En una muestra aleatoria de nueve sacos, el peso promedio observado es de 98 libras. Suponga que el peso de un saco se distribuye normalmente.*

1. *Plantee la hipótesis nula, la hipótesis alternativa ¿Cuál es valor P de la prueba?*

   Sea μ el peso promedio de un saco, se tiene que

   $$\begin{gathered} \textit{Afirmación}\text{: } \mu \geq 100 \\ H_0 : \mu = 100\ (\geq), \quad H_1 : \mu < 100 \end{gathered}$$

   Además se tienen los siguientes datos

   $$\overline{x} = 98, \quad n = 9, \quad \sigma = 2$$

   Como la población es normal y se conoce σ entonces, bajo H_0, $\overline{X} \sim N\left(\mu_0, \sigma^2\right)$ por lo tanto el valor P es

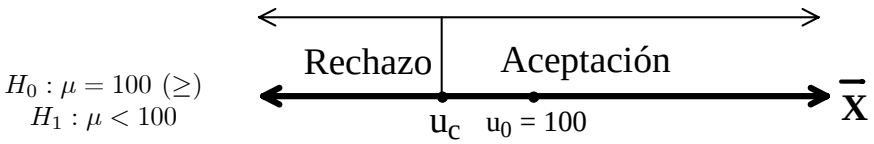
   $$\textit{Valor } P = P\left(\overline{X} \leq 98 | u = 100\right) = P\left(Z \leq -3\right) \approx 0.001349967$$

2. *¿Debe considerarse el peso promedio de un saco aceptable para su distribución con* $\alpha = 0.05$*?*

   Como el Valor $P < \alpha$ *se rechaza* H_0 *y se rechaza la afirmación, no se considera el peso promedio de cada saco como aceptable.*

3. *¿Cuál es la probabilidad de aceptar la hipótesis nula con* $\alpha = 0.05$ *si el peso promedio verdadero es de* 104 *libras ?*

   Note que se quiere averiguar $P(H_0|\mu = 104)$*, esto concuerda con* β*, la probabilidad del error tipo II, tomando* $H_1' : \mu = 104$*. Dado que*



   entonces

$$\beta = P(H_0|H_1') = P\left(\overline{X} \geq \mu_c|\mu = 104\right)$$

   Para averiguar β *se requiere conocer el valor crítico* μ_c*. Como* $\alpha = 0.05$ *entonces*

$$\alpha = P(H_1|H_0) \Longrightarrow 0.05 = P\left(\overline{X} \leq \mu_c|\mu = 100\right)$$

$$\text{Como } \overline{X} \sim N : 0.05 = P\left(Z \leq \frac{\mu_c - 100}{2/\sqrt{9}}\right), \text{ pues } n = 9 \text{ y } \sigma = 2$$

$$\Longrightarrow \frac{u_c - 100}{2/3} = -1.645 \Longrightarrow \quad \boxed{u_c = 98.9033}.$$

   Así,

$$\beta = P\left(\overline{X} \geq 98.9033|\mu = 104\right) = P\left(Z \geq \frac{98.9033 - 104}{2/\sqrt{9}}\right) = P(Z \geq -7.645\,05) \approx 1$$

   Note que el error tipo II es muy grande, esto se debe a que la hipótesis alternativa específica $H_1' : u = 104$ *realmente no favore a* $H_1 : u < 100$*.*

4. *¿Que tamaño debe tomarse para realizar la prueba con un nivel de significancia de* 5% *y una potencia de* 85% *si el verdadero promedio es* 99 *libras?*

   Se tiene que

$$H_0 : \mu = 100, H_1' : \mu = 99, \alpha = 0.05, \beta = 0.15, \sigma = 2$$

   entonces $\mu_0 = 100$*,* $\mu_1 = 99$*,* $z_\alpha = -1.645$*,* $z_\beta = -1.04$ *y el tamaño de la muestra debe ser*

$$n \geq \frac{(|z_\alpha| + |z_\beta|)^2 \sigma^2}{(\mu_1 - \mu_0)^2} = \frac{(1.645 + 1.04)^2 2^2}{(99 - 100)^2} \approx 28.836\,9$$

   Basta tomar un tamaño de muestra de 29*. Note que como la población es normal y se conoce* σ *no se requiere que* $n \geq 30$ *para justificar el uso de la normal estándar.*

Ejemplo 63 *Cosidere el ejemplo* (58) *. En este ejemplo la directiva de una empresa considera que no más de 9 kilómetros recorre diariamente el empleado promedio para llegar a la empresa. Así*

$$\begin{gathered} \textit{Afirmación: } \mu \leq 9 \\ H_0 : \mu = 9 \, (\leq), \quad H_1 : \mu > 9 \end{gathered}$$

Además, en una muestra de 50 empleados, se observó un recorrido promedio de 9.1 *km con una desviación estándar de* 5 *km. ¿Que tamaño debe tomarse para realizar la prueba con un nivel de significancia de* 5% *y una potencia de* 90% *si el verdadero promedio es* 13 *km?*

Se tiene que

$$H_0 : \mu = 9, H_1' : \mu = 13, \alpha = 0.05, \beta = 0.1.$$

entonces $\mu_0 = 9$, $\mu_1 = 13$, $z_\alpha = -1.645$, $z_\beta = -1.28$. *Además como* $s = 5$ *para una muestra de tamaño* 50 *entonces* $\sigma \approx s = 5$. *Así, el tamaño de la muestra debe ser*

$$n \geq \frac{\left(|z_\alpha| + |z_\beta|\right)^2 \sigma^2}{\left(\mu_1 - \mu_0\right)^2} = \frac{\left(1.645 + 1.28\right)^2 5^2}{\left(13 - 9\right)^2} \approx 13.3636$$

Como se debe garantizar el uso de la normal estándar y no se sabe si la población es normal entonces se debe tomar un tamaño de muestra de 30.

Ejercicio 29 *Un cliente se queja de que las baterías de las computadoras portátiles CDF tiene una duracción promedio inferior a la indicada, en el manual, de* 3 *horas. Las duraciones de ocho baterías de computadora portatil CDF, cargadas completamente, son* 181, 173, 185, 164, 170, 177, 186 *y* 178 *minutos.*

1. *Plantee la hipótesis nula, la hipótesis alternativa ¿Cuál es valor P de la prueba?*
2. *Determine las regiones de aceptación y rechazo, utilizando un nivel de significancia del* 5%.
3. *A un nivel de significancia del* 5%, *¿contradicen los datos recolectados la afirmación del cliente? R/*

Ejercicio 30 *Los sistemas de emergencia para aviones son impulsados por un combustible sólido. Una de las características importantes de este producto es la rapidez de combustión. El sistema de emergencia se considera efectivo si la rapidez promedio de combustión es de alrededor de* $50cm/s$. *Suponga que la desviación estándar de esta rapidez es de* $3cm/s$.

1. *Plantee la hipótesis nula, la hipótesis alternativa para lo requerido por el sistema.*
2. *Recientemente, el sistema de emergencia de un avión no funcionó correctamente, esto hace suponer que la rapídez promedio de combustión no está cumpliendo lo requerido. Ante esto, se realizan* 25 *mediciones aleatorias y se obtiene una rapidez promedio de combustión de* $51.5cm/s$. *Suponiendo que la rapidez de combustión sigue una distribución normal ¿Contradicen estos datos la suposición de que la rapídez promedio de combustión no está cumpliendo lo requerido?*
3. *Realice nuevamente la prueba, determinando las regiones de aceptación y rechazo, con un nivel de significancia del* 5%.
4. *¿Que tamaño debe tomarse para realizar la prueba con un nivel de significancia de* 5% *y una potencia de* 90% *si la verdadera rapidez promedio de combustión es de* 51 cm/s*?*

2.2 Pruebas de hipótesis con una proporción

2.2.1 Muestras grandes

Considere una población dada con una proporción poblacional p de cierta carácteristica y el estadístico $\widehat{P}$ asociado a p para muestras de tamaño n. Recuerde que si $np \geq 5$ y $nq \geq 5$ entonces

$$\widehat{P} \underset{aprox}{\sim} N\left(p, \frac{pq}{n}\right).$$

En una prueba de hipótesis con una proporción la hipótesis nula es de la forma

$$H_0 : p = p_0$$

Así, bajo la hipótesis nula $H_0 : p = p_0$, si $np_0 \geq 5$ y $nq_0 \geq 5$ entonces

$$\widehat{P} \underset{aprox}{\sim} N\left(p_0, \frac{p_0 q_0}{n}\right) \Longrightarrow Z = \frac{\widehat{P} - p_0}{\sqrt{p_0 q_{0/n}}} \underset{aprox}{\sim} N(0,1).$$

Ejemplo 64 *Un investigador afirma que al menos el 10% de los cascos para motocicleta marca FAST tienen defectos de fabricación que pueden provocar daños a quien lo usa. Una muestra aleatoria de 200 cascos revela que 16 de ellos contienen tales defectos.*

1. *¿Cuál es valor P de la prueba?*

   Sea p el porcentaje de cascos FAST con defectos.

   $$\begin{array}{c} \textit{Afirmación}: p \geq 0.1 \\ H_0 : p = 0.1\ (\geq), \quad H_1 : p < 0.1 \end{array}$$

   El valor observado es $\widehat{p} = \dfrac{16}{200} = 0.08$. *El valor P es*

   $$\textit{Valor } P = P\left(\widehat{P} \leq \widehat{p} | H_0\right) = P\left(\widehat{P} \leq 0.08 | p = 0.1\right)$$

   Note que $n = 200$ *y* $p_0 = 0.1$, *como* $np_0 = 20 \geq 5$ *y* $nq_0 = 180 \geq 5$ *entonces* $\widehat{P} \underset{aprox}{\sim} N\left(p_0, \dfrac{p_0 q_0}{n}\right)$ *y*

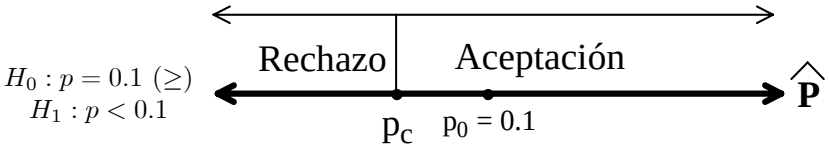
   $$\textit{Valor } P = P\left(Z \leq \frac{0.08 - 0.1}{\sqrt{\dfrac{0.1 \cdot 0.9}{200}}}\right) = P\left(Z \leq -0.9428\right) = 0.1736$$

2. *¿Hay evidencia que respalde la afirmación del investigador con* $\alpha = 0.05$*?*

   $$\textit{Como el Valor } P > \alpha \textit{ entonces se acepta } H_0$$

   y por lo tanto se acepta la afirmación. Se concluye que no hay evidencia significativa en contra de la afirmación

3. *Determine las regiones de aceptación y rechazo con* $\alpha = 0.05$.



Como $\alpha = 0.05$ *entonces*

$$\alpha = P\left(H_1|H_0\right) \Longrightarrow 0.05 = P\left(\widehat{P} \leq p_c|p = 0.1\right)$$

Además $n = 200$. *Al asumir que* $p = 0.1$, *se tiene que* $np = 20 \geq 5$ *y* $nq = 180 \geq 5$ *entonces* $\widehat{P} \underset{aprox}{\sim} N$ *y*

$$0.05 = P\left(\widehat{P} \leq p_c|p = 0.1\right) = P\left(Z \leq \frac{p_c - 0.1}{\sqrt{0.1 \cdot 0.9/200}}\right)$$
$$\Longrightarrow \frac{p_c - 0.1}{\sqrt{0.1 \cdot 0.9/200}} = z_{0.05} = -1.645 \Longrightarrow p_c = 0.06510\,43$$

Así se tiene que:
La región de rechazo es $R =]-\infty, 0.06510\,43[$
La región de aceptación es $A =]0.06510\,43, +\infty[$

Note que utilizando las regiones, como el valor observado $\widehat{p} = 0.08$ *está en la región de aceptación, entonces se acepta* H_0.

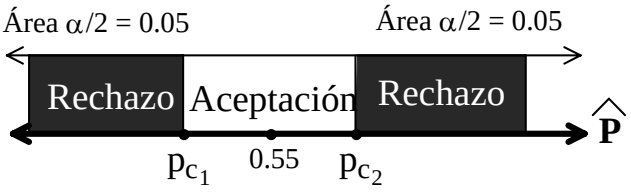
Ejemplo 65 *El candidato del partido PLAN asegura que el* 55% *de los ciudadanos votará por él en las próximas elecciones. La encuestadora UNI obtuvo que de* 1000 *habitantes,* 520 *votarán por dicho candidato. Utilizando un nivel de significancia del* 10%, *¿existe evidencia significativa en contra de la afirmación del candidato?*

Sea p *el porcentaje de ciudadanos que apoyan al candidato. Note que*

$$Afirmación : p = 0.55$$
$$H_0 : p = 0.55 \qquad H_1 : p \neq 0.55$$

Datos: $n = 1000, \widehat{p} = 0.520$, $p_0 = 0.55$, $\alpha = 0.1$.

1. *Utilizando regiones para* $\widehat{P}$.



Se tiene que

$$\frac{\alpha}{2} = P\left(\widehat{P} < p_{c_1} | H_0\right) \Longrightarrow 0.05 = P\left(\widehat{P} < p_{c_1} | p = 0.55\right)$$

Suponiendo que $p = 0.55$, *se tiene que* $np = 550 \geq 5$ *y* $nq = 450 \geq 5$, *por lo tanto*

$$\begin{aligned} 0.05 \quad &= \quad P\left(\widehat{P} < p_{c_1} | p = 0.55\right) = P\left(Z < \frac{p_{c_1} - 0.55}{\sqrt{\frac{0.55 \cdot 0.45}{1000}}}\right) \\ &\Longrightarrow \quad \frac{p_{c_1} - 0.55}{\sqrt{\frac{0.55 \cdot 0.45}{1000}}} = z_{0.05} = -1.645 \Longrightarrow \boxed{p_{c_1} = 0.524\,121} \end{aligned}$$

Dado que p_{c_1} *y* p_{c_2} *son simétricos respecto*[5] *a* p_0, *entonces*

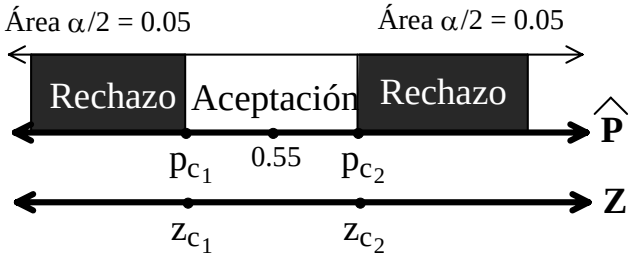
$$p_{c_2} = p_0 + (p_0 - p_{c_1}) = 0.55 + 0.024121 = 0.574\,121\,.$$

Por lo tanto, las regiones para $\widehat{P}$ *son*

$$\textit{Región Aceptación } A =]0.524\,121, 0.574\,121[\textit{ Región Rechazo } R = \mathbb{R} - A.$$

Como $\widehat{p} = 0.520 \in R$, *entonces se rechaza* H_0. *Hay evidencia significativa en contra de la afirmación del candidato. Note que* p_{c_2} *no era necesario averiguarlo, pues* $\widehat{p} < p_{c_1} \Longrightarrow \widehat{p} \in R$.

2. *Utilizando regiones para Z. Dado que bajo* H_0, $np = 550 \geq 5$ *y* $nq = 450 \geq 5$, *se tiene que* $\widehat{P} \underset{aprox}{\sim} N\left(p_0, \frac{p_0 q_0}{n}\right)$, *entonces* $Z = \dfrac{\widehat{P} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \underset{aprox}{\sim} N(0,1)$. *Como* $\widehat{P}$ *y* Z *tienen el mismo comportamiento entonces*



donde $z_{c_1} = z_{0.05} = -1.645$ *y por simetría se tiene que* $z_{c_2} = 1.645$. *Así las regiones para* Z *son*

$$\textit{Región Aceptación } A_2 =]-1.645, 1.645[\textit{ Región Rechazo } R_2 = \mathbb{R} - A.$$

[5] Esto pues, bajo H_0, $\widehat{P} \underset{aprox}{\sim} N\left(p_0, \frac{p_0 q_0}{n}\right)$, entonces la distribución de $\widehat{P}$ es simétrica respecto a su media p_0.

Como el valor onservado para $\widehat{P}$ *es* $\widehat{p} = 0.520$, *entonces el valor observado de* Z *es*

$$z_{obs} = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{0.520 - 0.55}{\sqrt{\frac{0.55 \cdot 0.45}{1000}}} = -1.\,906\,93$$

Como $z_{obs} \in R_2$ *entonces se rechaza* H_0 *y se obtiene la misma conclusión.*

3. *Utilizando el valor P. Como* $H_1 : p \neq 0.55$ *y* $\widehat{p} = 0.520$ *entonces dado que el valor P va en dirección a* H_1 *entonces* $\widehat{P} \leq 0.520$. *Así,*

$$\begin{aligned} \textit{Valor } P &= 2 \cdot P\left(\widehat{P} \leq 0.520 | p = 0.55\right) \\ &= 2 \cdot P\left(Z \leq \frac{0.520 - 0.55}{\sqrt{\frac{0.55 \cdot 0.45}{1000}}}\right), \quad \textit{pues } np = 550 \textit{ y } nq = 450. \\ &= 2 \cdot P\left(Z \leq -1.\,906\,93\right) \\ &\approx 2 \cdot 0.028067 = 0.056\,134 \end{aligned}$$

Como el Valor $P < \alpha = 0.1$, *se rechaza* H_0 *y se obtiene la misma conclusión.*

Teorema 35 *Tomando una muestra suficientemente grande para que* $\widehat{P}$ *se distribuya aproximadamente normal, para probar*

$$H_0 : p = p_0$$

con un nivel de significancia de α *y una potencia mínima de* $1-\beta$ *para la hipótesis alternativa específica*

$$H_1' : p = p_1$$

puede tomarse una muestra de tamaño

1. $n \geq \dfrac{\left(|z_\alpha| \sqrt{p_0 q_0} + |z_\beta| \sqrt{p_1 q_1}\right)^2}{\left(p_1 - p_0\right)^2}$ *si la prueba es de una cola.*

2. $n \geq \dfrac{\left(\left|z_{\alpha/2}\right| \sqrt{p_0 q_0} + |z_\beta| \sqrt{p_1 q_1}\right)^2}{\left(p_1 - p_0\right)^2}$ *si la prueba es de dos colas.*

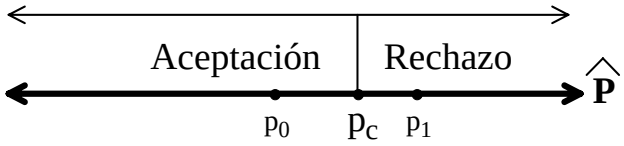
Prueba. *Se probará uno de los dos casos de* (1), *los otros se dejan de ejercicio. Suponga que se quiere probar*

$$H_0 : p = p_0, \quad H_1 : p > p_0$$

con un nivel de significancia de α *y una potencia mínima de* $1 - \beta$ *para la hipótesis alternativa específica*

$$H_1' : p = p_1$$

donde α y β son menores al 50%. Entonces las regiones de aceptación y rechazo son



y las probabilidades de que sucedan los errores tipo I y II están acotadas por:

$$\alpha \geq P\left(H_1|H_0\right) \quad y \quad \beta \geq P\left(H_0|H_1'\right)$$
$$\Longrightarrow \quad \alpha \geq P\left(\widehat{P} > p_c|p = p_0\right) \quad y \quad \beta \geq P\left(\widehat{P} < p_c|p = p_1\right)$$

Asumiendo que la muestra es suficientemente grande de modo que $np_0 \geq 5$, $nq_0 \geq 5$, $np_1 \geq 5$ y $nq_1 \geq 5$ se tiene que

$$\alpha \geq P\left(Z > \frac{p_c - p_0}{\sqrt{p_0q_0/n}}\right) \quad y \quad \beta \geq P\left(Z < \frac{p_c - p_1}{\sqrt{p_1q_1/n}}\right)$$
$$\Longrightarrow \quad 1 - \alpha \leq P\left(Z < \frac{p_c - p_0}{\sqrt{p_0q_0/n}}\right) \quad y \quad \phi\left(z_\beta\right) \geq \phi\left(\frac{p_c - p_1}{\sqrt{p_1q_1/n}}\right)$$
$$\Longrightarrow \quad \phi\left(z_{1-\alpha}\right) \leq \phi\left(\frac{p_c - p_0}{\sqrt{p_0q_0/n}}\right) \quad y \quad z_\beta \geq \frac{p_c - p_1}{\sqrt{p_1q_1/n}}$$
$$\Longrightarrow \quad z_{1-\alpha} \leq \frac{p_c - p_0}{\sqrt{p_0q_0/n}} \quad y \quad z_\beta \geq \frac{p_c - p_1}{\sqrt{p_1q_1/n}}$$

Note que como $z_{1-\alpha} = -z_\alpha$ entonces de la primera desigualdad se tiene que

$$-z_\alpha \leq \frac{p_c - p_0}{\sqrt{p_0q_0/n}} \Longrightarrow z_\alpha \geq \frac{p_0 - p_c}{\sqrt{p_0q_0/n}}$$

Como z_α y $p_0 - p_c$ son negativos, pues $\alpha < 0.5$ y $p_0 < p_c < p_1$, entonces al aplicar el valor absoluto

$$|z_\alpha| \leq \left|\frac{p_0 - p_c}{\sqrt{p_0q_0/n}}\right| = \frac{p_c - p_0}{\sqrt{p_0q_0/n}} \Longrightarrow \boxed{\frac{|z_\alpha|\sqrt{p_0q_0}}{\sqrt{n}} \leq p_c - p_0} \qquad (1)$$

Similarmente, como $z_\beta \geq \dfrac{p_c - p_1}{\sqrt{p_1q_1/n}}$ se tiene que

$$|z_\beta| \leq \left|\frac{p_c - p_1}{\sqrt{p_1q_1/n}}\right| = \frac{p_1 - p_c}{\sqrt{p_1q_1/n}} \Longrightarrow \boxed{\frac{|z_\beta|\sqrt{p_1q_1}}{\sqrt{n}} \leq p_1 - p_c} \qquad (2)$$

Sumando (1) *y* (2) *se tiene que*

$$\frac{|z_\alpha|\sqrt{p_0q_0}}{\sqrt{n}} + \frac{|z_\beta|\sqrt{p_1q_1}}{\sqrt{n}} \leq p_1 - p_0$$

Por lo tanto,

$$\sqrt{n} \geq \frac{|z_\alpha|\sqrt{p_0q_0} + |z_\beta|\sqrt{p_1q_1}}{p_1 - p_0} \Longrightarrow n \geq \frac{\left(|z_\alpha|\sqrt{p_0q_0} + |z_\beta|\sqrt{p_1q_1}\right)^2}{(p_1 - p_0)^2} \quad \blacksquare$$

De acuerdo a la prueba del teorema anterior, el tamaño de muestra n a determinar debe cumplir las siguientes condiciones

$$np_0 \geq 5, nq_0 \geq 5, np_1 \geq 5 y nq_1 \geq 5$$

Esto garantiza que $\widehat{P} \underset{aprox}{\sim} N$ bajo H_0 y bajo H_1'. Asi el tamaño de muestra de cumplir estas condiciones y la dada por el teorema.

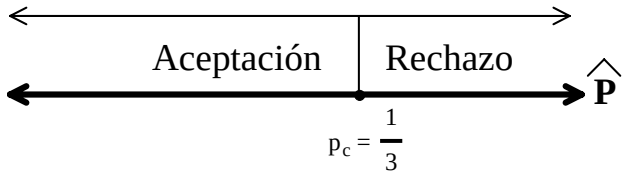
Ejemplo 66 *Se toma una muestra aleatoria de 300 habitantes de cierta ciudad, y se les pregunta si están a favor de que el país C realice un tratado de libre comercio con China. Realizando la prueba de hipótesis se ha determinado que si menos de 100 personas responde de manera afirmativa, entonces se concluye que a lo sumo el 30% de los habitantes está a favor del tratado.*

1. *Determine las regiones de aceptación y rechazo.*

   Sea p el porcentaje de habitantes a favor del tratado.

   $$Afirmación : p \leq 0.3$$
   $$H_0 : p = 0.3 \quad (\leq) \qquad H_1 : p > 0.3$$

   El tamaño de muestra es $n = 300$ *y el estadístico es* $\widehat{P}$. *Note que se dice que si la estimación dada por* $\widehat{P}$ *es menor a* $\frac{100}{300} = \frac{1}{3}$, *entonces se acepta* H_0. *Por lo tanto el valor crítico es* $p_c = \frac{1}{3}$:



   Así, se tiene que

   $$\textit{Región Aceptación } A = \left[0, \frac{100}{300}\right[\textit{, Región Rechazo } R = \left]\frac{100}{300}, 1\right].$$

2. *Encuentre la probabilidad del error tipo I*

$$\alpha = P\left(H_1|H_0\right) = P\left(\widehat{P} > \frac{1}{3}|p = 0.3\right)$$

Note que $n = 300$. *Bajo* $H_0 : p = 0.3$, *como* $np = 90$ *y* $nq = 210$ *son mayores a* 5, *entonces*

$$\alpha = P\left(Z > \frac{1/3 - 0.3}{\sqrt{\frac{0.3 \cdot 0.7}{300}}}\right) = P\left(Z > 1.\,259\,88\right) = 0.103835.$$

3. *¿Cuál es la probabilidad del error tipo II si sólo el 40% de los habitantes está a favor del tratado?*

Se tiene que $H_1' : p = 0.4$, *entonces*

$$\beta = P\left(H_0|H_1'\right) = P\left(\widehat{P} < \frac{1}{3}|p = 0.4\right)$$

Bajo $H_1' : p = 0.4$, *como* $np = 120$ *y* $nq = 180$ *son mayores a* 5, *entonces*

$$\beta = P\left(Z \leq \frac{1/3 - 0.4}{\sqrt{\frac{0.4 \cdot 0.6}{300}}}\right) = P\left(Z \leq -2.\,357\,02\right) = 0.009137.$$

4. *Una casa encuestadora, desea realizar la prueba con un nivel de significancia de* 5% *y una potencia de del* 2% *si la verdadera proporción es del* 40%. *¿Qué tamaño de muestra debe tomar?*

Se tiene que

$$H_0 : p = 0.3, H_1' : p = 0.4, \alpha = 0.05, \beta = 0.02.$$

entonces $p_0 = 0.2$, $p_1 = 0.4$, $z_\alpha = -1.645$, $z_\beta = -2.05$. *Así, el tamaño de la muestra debe ser*

$$n \geq \frac{\left(|z_\alpha|\sqrt{p_0 q_0} + |z_\beta|\sqrt{p_1 q_1}\right)^2}{\left(p_1 - p_0\right)^2} = \frac{\left(1.645\sqrt{0.3 \cdot 0.7} + 2.05\sqrt{0.4 \cdot 0.6}\right)^2}{\left(0.4 - 0.3\right)^2} \approx 309.1$$

Se debe tomar una muestra de 310 *habitantes. Note que si* $n = 310$ *entonces* $np_0 = 93$, $nq_0 = 217$, $np_1 = 124$ *y* $nq_1 = 186$, *los cuales son mayores a* 5.

Ejercicio 31 *Se quiere determinar si al menos el 85% de los ciudadanos que transitan frecuentemente por San José en vehículo está a favor de utilizar más el transporte público para reducir las presas. Se toma una muestra aleatoria de 500 ciudadanos con dichas características y se les pregunta si están a favor de utilizar más el transporte público para reducir las presas. En el estudio se determinó que el valor crítico de la prueba es* $p_c = 0.8$.

1. *De las 500 personas encuestadas, ¿cuántas deben responder de manera afirmativa para concluir que al menos el 85% de los habitantes está a favor de utilizar más el transporte público?R/ Más de 400*

2. *Encuentre la probabilidad del error tipo I .* $R/$ 0.0008741

3. *¿Cuál es la probabilidad β del error tipo II si sólo el 80% de dicha población esta dispuesta a utilizar más el transporte público?* $R/$

Ejercicio 32 *Un fabricante de lentes evalúa una nueva máquina pulidora. El fabricante aprobará la máquina si el porcentaje de lentes pulidos que contienen defectos en la superficie no es mayor del 2%. Se toma una muestra aleatoria de 250 lentes y se encuentra que seis de ellos tiene defectos.*

1. *Proponga y plantee la prueba para determinar si la máquina será aceptada.* $R/$ $H_1 : p > 0.02$

2. *Realice la prueba e indique si el fabricante aprobará o no la nueva máquina.* $R/$ *Valor* $P = 0.326355$

2.2.2 Muestras pequeñas

Si la muestra es pequeña (bajo H_0 se tiene que $np_0 < 5$ o $nq_0 < 5$)se debe recordar que la variable $B = n\widehat{P}$ es binomial. Por lo tanto, bajo la hipótesis $H_0 : p = p_0$, si $np_0 < 5$ o $nq_0 < 5$ se tiene que

$$B \sim B(n, p_0)$$

Ejemplo 67 *En una muestra aleatoria de 15 estudiantes del TEC, solo uno de ellos apoyan al Club Sport Cartaginés.*

1. *¿Es esto evidencia fuerte de que menos de una quinta parte de los estudiantes del TEC apoyan al Cartaguinés?*

   Sea p el porcentaje de estudiantes del TEC que apoyan al Club Sport Cartaginés.

   $$\begin{array}{c} \textit{Se afirrma que } p < 0.2 \\ H_0 : p = 0.2\,(\geq) \quad H_1 : p < 0.2 \end{array}$$

   Se tiene los datos

   $$n = 15, p_0 = 0.2, \widehat{p} = \frac{1}{15}$$

   Como $np_0 = 3 < 5$ entonces se utiliza la binomial en la prueba. Así. bajo H_0, se tiene que

   $$\begin{array}{rcl} B & = & n\widehat{P} = 15\widehat{P} \sim B(n, p_0) = B(15, 0.2) \\ & \Longrightarrow & P(B = k|H_0) = C(15, k)\left(\frac{1}{5}\right)^k \left(\frac{4}{5}\right)^{15-k} \end{array}$$

   y el valor P es

   $$\textit{Valor } P = P\left(\widehat{P} \leq \widehat{p}|H_0\right) = P\left(\widehat{P} \leq \frac{1}{15}|p = \frac{1}{5}\right)$$

   Note que si $\widehat{P} \leq \frac{1}{15}$, multiplicando por $n = 15$ a ambos lados se tiene que $15\widehat{P} \leq 1$ es decir $B \leq 1$. Así,

   $$\begin{array}{c} \textit{Valor } P = P\left(B \leq 1|p = \frac{1}{5}\right) = P\left(B = 0|p = \frac{1}{5}\right) + P\left(B = 1|p = \frac{1}{5}\right) \\ = C(15, 0)\left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^{15} + C(15, 1)\left(\frac{1}{5}\right)^1 \left(\frac{4}{5}\right)^{14} = 0.167\,126 \end{array}$$

Como Valor $P > 0.05$, no hay evidencia fuerte de que menos de la quinta parte de los estudiantes del TEC apoyan al Cartaguinés.

2. *¿De qué tamaño debería ser una muestra para probar la hipótesis de que menos de una quinta parte de los estudiantes del TEC apoyan al Cartaguinés con un nivel de significancia de 5% y un potencia de 85% cuando una sexta parte de los estudiantes apoyan al Cataginés?*

   Se tiene que

   $$H_0 : p = \frac{1}{5}, H_1' : p = \frac{1}{6}, \alpha = 0.05, \beta = 0.15.$$

   entonces $p_0 = \frac{1}{5}$, $p_1 = \frac{1}{6}$, $z_\alpha = -1.645$, $z_\beta = -1.04$. *Así, el tamaño de la muestra debe ser*

   $$n \geq \frac{\left(|z_\alpha| \sqrt{p_0 q_0} + |z_\beta| \sqrt{p_1 q_1}\right)^2}{(p_1 - p_0)^2} = \frac{\left(1.645\sqrt{\frac{1}{5}\frac{4}{5}} + 1.04\sqrt{\frac{1}{6}\frac{5}{6}}\right)^2}{\left(\frac{1}{6} - \frac{1}{5}\right)^2} = 983.923$$

   Si $n = 984$ *entonces* $np_0 = 196.8$, $nq_0 = 787.2$, $np_1 = 164$ *y* $nq_1 = 820$, *los cuales son mayores a* 5. *Por lo tanto se puede tomar una muestra de 984 estudiantes.*

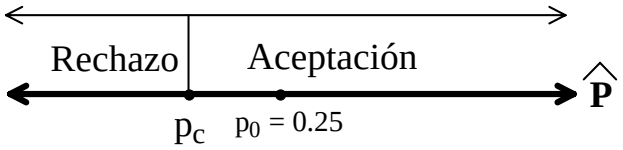
Ejemplo 68 *En una muestra aleatoria de 24 alumnos de computación del ITCR, 5 fueron al Estadio Ricardo Saprissa el mes pasado. Se cree que menos de la cuarta parte de los estudiantes de computación asistieron el mes pasado a este estadio.*

1. *Pruebe la afirmación anterior con un nivel de significancia de 5% determinando las regiones de aceptación y rechazo.*

   Sea p el porcentaje de estudiantes de computación que asistieron el mes pasado a este estadio.

   Se afirrma que $p < 0.25$

   $Ho : p = 0.25\ (\geq) \quad H_1 : p < 0.25$



   Datos:

   $$n = 24, p_0 = 0.25, \widehat{p} = \frac{5}{24}, \alpha = 0.05$$

Como $np_0 = 6$ *y* $nq_0 = 18$ *son ambos mayores que 5:*

$$\begin{aligned}
& \alpha = P\left(H_1|H_0\right) \\
\Longrightarrow \quad & 0.05 = P\left(\widehat{P} < p_c|p = 0.25\right) = P\left(Z < \frac{p_c - 0.25}{\sqrt{\frac{0.25 \cdot 0.75}{24}}}\right) \\
\Longrightarrow \quad & \frac{p_c - 0.25}{\sqrt{\frac{0.25 \cdot 0.75}{24}}} = -1.645 \\
\Longrightarrow \quad & p_c = 0.104\,601
\end{aligned}$$

Así, se tiene que

$$\textit{Región Aceptación } A =]0.104601, 1[\,, \textit{ Región Rechazo } R = [0, 0.104601[\,.$$

Como $\widehat{p} \in A$ *se acepta* H_0 *y se rechaza la afirmación. Existe evidencia significativa en contra de que menos de la cuarta parte de los estudiantes de computación asistieron el mes pasado a este estadio.*

2. *¿Cuál es la probabilidad* β *del error tipo II de la prueba si sólo la quinta parte de los estudiantes de computación asistieron el mes pasado a este estadio?*

$$\begin{gathered}
\textit{Se tiene que } H_1' : p = 0.2, \textit{ entonces} \\
\beta = P\left(H_0|H_1'\right) = P\left(\widehat{P} > 0.104601|p = 0.2\right)
\end{gathered}$$

Dado que $n = 24$, *entonces bajo* $H_1' : p = 0.2$ *se tiene que* $np = 4.8 < 5$, *por lo tanto se utiliza la binomial*

$$B = n\widehat{P} = 24\widehat{P} \sim B\left(24, 0.2\right) \Longrightarrow P\left(B = k|H_1'\right) = C\left(24, k\right) 0.2^k 0.8^{15-k}$$

Así,

$$\begin{aligned}
\beta \quad & = P\left(\widehat{P} > 0.104601|p = 0.2\right) \\
& = P\left(B > 2.510\,42|p = 0.2\right), \textit{ pues } 0.104601 \cdot 24 = 2.510\,42 \\
& = 1 - P\left(B = 0|p = 0.2\right) - P\left(B = 1|p = 0.2\right) - P\left(B = 2|p = 0.2\right) \\
& = 1 - C\left(24, 0\right) 0.2^0 0.8^{24} - C\left(24, 1\right) 0.2^1 0.8^{23} - C\left(24, 2\right) 0.2^2 0.8^{22} \\
& = 0.885\,483
\end{aligned}$$

3. *¿De qué tamaño debería ser una muestra para probar la afirmación con un nivel de significancia de 5% y un potencia de 85% cuando solo el 10% de los estudiantes de computación asistieron el mes pasado a este estadio?*

Se tiene que

$$H_0 : p = 0.25, H_1' : p = 0.1, \alpha = 0.05, \beta = 0.15.$$

entonces $p_0 = 0.25$, $p_1 = 0.1$, $z_\alpha = -1.645$, $z_\beta = -1.04$. *Así, el tamaño de la muestra debe ser*

$$n \geq \frac{\left(|z_\alpha| \sqrt{p_0 q_0} + |z_\beta| \sqrt{p_1 q_1}\right)^2}{(p_1 - p_0)^2} = \frac{\left(1.645\sqrt{0.25 \cdot 0.75} + 1.04\sqrt{0.1 \cdot 0.9}\right)^2}{(0.25 - 0.1)^2} = 46.631\,2$$

Sin embargo, si se toma $n = 47$ *entonces* $np_1 = 4.7 < 5$, *y esto impide el uso del teorema sobre el cálculo del tamaño de muestra. Por lo tanto para que*

$$np_1 \geq 5 \Longrightarrow n \geq \frac{5}{p_1} = \frac{5}{0.1} = 50.$$

Así, tomando $n = 50$, *se cumple que* $np_0 = 12.5, nq_0 = 37.5, np_1 = 5$ *y* $nq_1 = 45$.*Se debe tomar una muestra de al menos 50 estudiantes.*

Ejercicio 33 *En el colegio Bienestar Seguro se han detectado varios estudiantes con problemas de sobrepeso en el presente año. El director ha afirmado que más del* 70% *de los estudiantes tiene problemas de sobrepeso. Se quiere contrastar la hipótesis con un nivel de significancia del* 10% *y una muestra de 12 estudiantes.*

1. *Determine aproximadamente el valor crítico* p_c *que separa las regiones de aceptación y rechazo.* $R/ \quad \frac{10}{12} < p_c < \frac{11}{12}$.
2. *Suponga que en una muesta de* 12 *estudiantes,* 8 *tiene sobrepeso. ¿Contradice estos datos la afirmación del director?* $R/$ Si

2.3 Pruebas de hipótesis con una variancia

Considere la población dada por la variable aleatoria X que sigue una distribución normal con desviación estándar σ, suponga que se toma una muestra aleatoria de tamaño n. La hipótesis nula, para este caso, es de la forma

$$H_0 : \sigma = \sigma_0 \text{ que es equivalente a } H_0 : \sigma^2 = \sigma_0^2$$

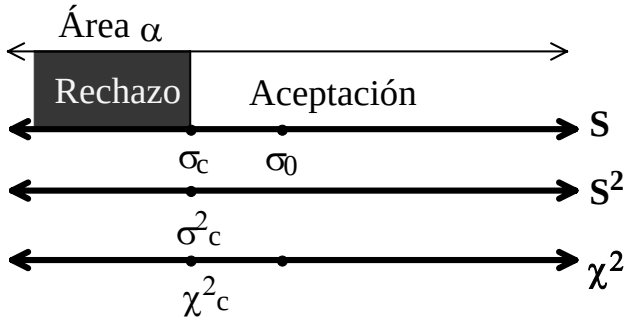
Por resultados anteriores se tiene que $\chi^2 = \dfrac{(n-1)\,S^2}{\sigma^2} \sim \chi^2\,(n-1)$. En particular, bajo la hipótesis $H_0 : \sigma = \sigma_0$, se tiene que

$$\chi^2 = \frac{(n-1)\,S^2}{\sigma_0^2}$$

sigue una distribución χ^2 con $v = n - 1$ grados de libertad.

Dado que $S > 0$ entonces S^2 tiene el mismo comportamiento que S. Además, como $\dfrac{(n-1)}{\sigma_0^2} > 0$, el estadístico χ^2 tiene el mismo comportamiento que el estimador S^2. Por ejemplo, para $H_1 : \sigma < \sigma_0$ se

tienen las siguientes regiones para los distintos estadísticos



donde $\sigma_c^2 = (\sigma_c)^2$ y $\chi_c^2 = \dfrac{(n-1)\,\sigma_c^2}{\sigma_0^2}$. Además, recuerde que $\alpha = P(H_1|H_0)$ es la medida de la región de rechazo asumiendo H_0, por lo tanto χ_c^2 es el valor que acumula una probabilidad de α, es decir

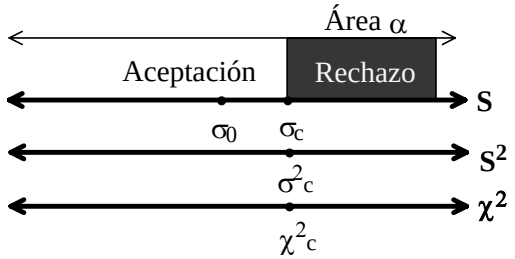
$$\chi_c^2 = \chi_{\alpha,n-1}^2,$$

y el valor P es

$$P\left(\chi^2 < \chi_{obs}^2\right), \text{ donde } \chi_{obs}^2 = \frac{(n-1)\,s^2}{\sigma_0^2}.$$

Similarmente se tienen los otros casos

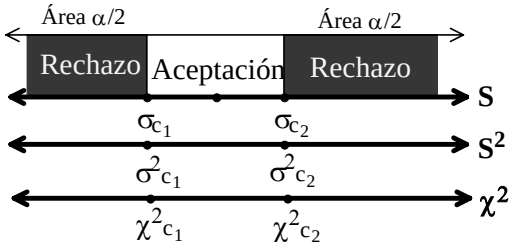
$H_1 : \sigma > \sigma_0$



$$\chi_c^2 = \chi_{1-\alpha,n-1}^2$$

$$\text{Valor } P = P\left(\chi^2 > \chi_{obs}^2\right)$$

$H_1 : \sigma \neq \sigma_0$



$$\chi_{c_1}^2 = \chi_{\alpha/2,n-1}^2 \quad y \quad \chi_{c_2}^2 = \chi_{1-\alpha/2,n-1}^2$$

$$\text{Valor}^6\, P = \begin{cases} 2P\left(\chi^2 > \chi_{obs}^2\right) & si \quad \chi_{obs}^2 > n-1 \\ 2P\left(\chi^2 < \chi_{obs}^2\right) & si \quad \chi_{obs}^2 < n-1 \end{cases}$$

Ejemplo 69 *Se estima que el tiempo que requieren los estudiantes de estadística para resolver el primer examen parcial se distribuye normalmente. Si en este año una muestra de 20 estudiantes dio*

[6] Primero note el valor χ^2 cuando $S = \sigma_0$ es $\chi_0^2 = \dfrac{(n-1)\,\sigma_0^2}{\sigma_0^2} = n-1$. Además, dado que χ^2 no es simétrica respecto a χ_0^2 entonces el valor P indicado no es exacto es una aproximación. Por lo tanto, se prefiere utilizar regiones para este tipo de pruebas.

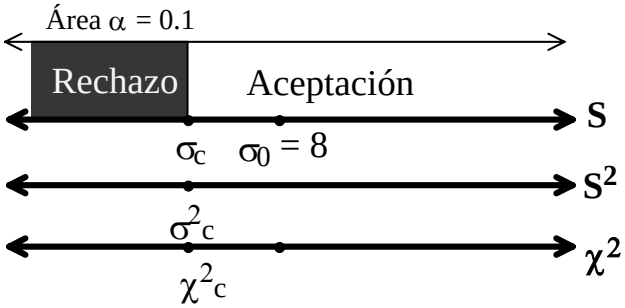
una desviación estándar de 5 minutos, ¿hay evidencia de que en realidad la desviación estándar de las notas es menor que 8 minutos? Utilice $\alpha = 10\%$.

Sea σ *la desviación estándar de las notas.*

$$Afirmación : \sigma < 8$$
$$H_0 : \sigma = 8 \ (\geq) \quad H_1 : \sigma < 8$$

Datos: $n = 20$, $s = 5$, $\alpha = 0.1$.

1. *Utilizando regiones*



donde $\chi^2_c = \chi^2_{\alpha,n-1} = \chi^2_{0.1,19} = 11.6509$. *Se puede realizar la prueba de varias formas según el estadístico a utilizar:*

Estadístico	***Regiones***	***Valor observado***	***Contraste***
χ^2	$A_1 =]11.6509, +\infty[$ $R_1 =]0, 11.6509[$	$\chi^2_{obs} = \dfrac{(n-1)\,s^2}{\sigma_0^2} = 7.4219$	$\chi^2_{obs} \in R_1$ *Se rechaza* H_0
S^2	$\chi^2_c = \dfrac{(n-1)\,\sigma_c^2}{\sigma_0^2}$ $\Longrightarrow 11.6509 = \dfrac{19\sigma_c^2}{8^2}$ $\Longrightarrow \sigma_c^2 = 39.2421$ $A_2 =]39.2421, +\infty[$ $R_2 =]0, 39.2421[$	$s^2 = 25$	$s^2 \in R_2$ *Se rechaza* H_0
S	$\sigma_c^2 = 39.2421 \Longrightarrow \sigma_c = 6.264$ $A_3 =]6.264, +\infty[$ $R_3 =]0, 6.264[$	$s = 5$	$s \in R$ *Se rechaza* H_0

2. *Utilizando el valor P. Dado que* $H_1 : \sigma < 8$ *entonces el Valor P, según el estadístico a utilizar, es*

Estadístico	***Valor observado***	***Valor*** P
S	$s = 5$	$P(S \leq 5)$
S^2	$s^2 = 25$	$P\left(S^2 \leq 25\right)$
χ^2	$\chi^2_{obs} = 7.4219$	$P\left(\chi^2 \leq 7.4219\right)$

Note que el primer y segundo caso se reduce al tercer. Así, independiente del estadístico a utlizar, se tiene que $Valor\ P = P\left(\chi^2 \leq 7.4219\right)$. *Utilizando la tabla con* $v = 19$ *se obtiene que*

$$Valor\ P = P\left(\chi^2 \leq 7.4219\right) \in]0, 0.01[$$

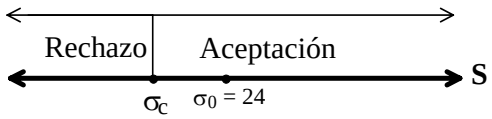
Como el $Valor\ P < \alpha$ *entonces se rechaza* H_0.

Así, en cualquiere de las posibles opciones para realizar la prueba se tiene que hay evidencia significativa de que la desviación estándar del tiempo es menor de 8 minutos.

Ejemplo 70 *Para una determinada población se afirma que* $\sigma^2 \geq 24$. *Un estudiante de Computación realizó correctamente el contraste de hipótesis tomando una muestra de tamaño* 40 *al nivel de significancia de* 0.05, *obteniendo que no existe evidencia en contra de la afirmación ¿Que tan pequeña puede ser la desviación estándar de los datos de la muestra?*

Sea σ *la desviación estándar de la población. Dado que se quiere el valor mínimo de* s, *se utilizará el estadístico* S. *Así, se tiene que*

Afirmación: $\sigma^2 \geq 24$
$H_0 : \sigma^2 = 24\,(\geq), \quad H_1 : \sigma^2 < 24$



Se tiene que $n = 40$ *y* $\alpha = 0.05$, *entonces*

$$\begin{gathered}
\alpha = P\left(H_1|H_0\right) = P\left(S^2 < \sigma_c^2|\sigma^2 = 24\right) \\
\Longrightarrow 0.05 = P\left(\chi^2 < \frac{39\sigma_c^2}{24}\right) \ con\ v = 39 \\
\Longrightarrow \frac{39\sigma_c^2}{24} = 25.6954 \Longrightarrow \sigma_c = 3.976\,5 \\
Región\ de\ rechazo\ para\ S\text{:}\ R = [0 \quad , \quad 3.976\,5[\\
Región\ de\ aceptación\ para\ S\text{:}\ A =]3.976\,5 \quad , \quad +\infty[
\end{gathered}$$

Como se acepta H_0 *entonces* $s > 3.9765$.

Ejercicio 34 *Una máquina dispensadora de refrescos, cuando está bien ajustada, llena los vasos con una desviación estándar no mayor que 6 ml. Si se observan diez vasos con contenidos 261, 273, 265, 268, 263, 266, 258, 251, 261 y 271 (todos en ml), ¿hay evidencia para sospechar que la máquina está desajustada?* $R/ \quad Valor\ P \in]0.2, 0.8[$.

Ejercicio 35 *Suponga que se ha utilizado una muestra de tamaño 30 para contrastar:* $H_0 : \sigma^2 = 17$ *versus* $H_1 : \sigma^2 > 17$, *al nivel de significancia de* 0.05. *¿Que tan grande debe ser el valor de* S^2 *en una muestra antes de que sea rechazada la hipotenusa nula?* $R/ \quad s^2 < 24.9472$

2.4 Ejercicios

1. Cuando el año pasado (2008) se anuncio la recesión de la economía de Estados Unidos, poco después se hizo evidente una crisis dado que muchas empresas de los Estados Unidos estaban

en dificultades financieras. Cierto economista estimó que las empresas endeudadas que buscan liberarse de las presiones de los acreedores debían, en promedio, más de 2250 millones de dólares. En un estudio de 20 empresas endeudadas se reveló que debían en promedio 2517 millones de dólares, con una desviación de 900 millones de dólares.

(a) ¿Cuál es el valor P aproximado con base en la evidencia? $R/$ 0.10

(b) ¿Existe evidencia para respaldar la afirmación del economista al nivel del 5%?.
$R/$ Se acepta H_0, Hay una leve evidencia en contra de la afirmación.

2. La Universidad Bienestar Seguro afirma que más del 65% de los titulados universitarios dejan su primer trabajo antes de 2 años de laborar. Un estudio realizado por la carrera de Informática de la universidad encontró que 350 de 500 recién graduados que fueron entrevistados se mantuvieron en su primer empleo menos de 2 años. Utilice los datos para probar si la universidad está en lo cierto con un nivel de significancia de 0.03, determinando las regiones de aceptación y rechazo. $R/$ $\widehat{p} \in R =]0.690\,592 \quad , \quad 1[$. Se rechaza H_0 y se acepta la afirmación

3. Se cree que 80% de los habitantes de una ciudad están a favor de un proyecto de ley. Se toma una muestra aleatoria de 70 habitantes, y si más de 49 pero menos de 63 de ellos están a favor, se aceptará que el porcentaje a favor es el 80%.

(a) Si el porcentaje a favor es realmente 80%, ¿cuál es la significancia de la prueba? $R/$ $\alpha = 0.0366178$

(b) Si el porcentaje a favor es en realidad 70%, ¿cuál es la potencia de la prueba?$R/$ 0.500 131

4. El peso de una bolsa de frijoles marca SABORES sigue una distribución normal de media desconocida y varianza 0.04 kg^2. Un inspector de la Oficina del Consumidor tomó una muestra de 20 bolsas y obtuvo un peso promedio de 1.9 kg. El dueño de la empresa SABORES asegura que el peso promedio de una bolsa de frijoles SABORES es superiora 2 kg. ¿Contradicen los datos la afirmación del dueño de la empresa? $R/$ Si

5. Juan considere que la proporción de mujeres del TEC es menor al 40%. Ana considera que el porcentaje de estudiantes del TEC que son de Cartago es mayor al 45%. Además Luis señala que en el TEC hay igual cantidad de mujeres que de Hombres. En una muestra aleatoria de 50 estudiantes del TEC da los siguientes resultados:

	Mujeres	Hombres
De Cartago:	11	10
De otras provincias:	10	19

(a) Determine si existe evidencia significativa para cada una de las afirmaciones, utilizando el valor P.

(b) Determine si existe evidencia significativa para cada una de las afirmaciones, utilizando regiones con un nivel de significancia del 10%.

6. En una muestra de 40 niños de 7 años de una localidad se encontró que 3 saben nadar. El profesor de Educación Física de la escuela de la localidad afirma que menos del 15% de los niños de 7 años saben nadar. ¿Contradicen los datos la afirmación del profesor? $R/$ No

7. El Ministerio de Salud considera que menos del 10% de las personas de cierta ciudad padecen asma. En una muestra de 50 personas de un barrio, 4 padecen de asma. ¿Hay evidencia en contra de la afirmación indicada por el Ministerio de Salud? Utilice un nivel de significancia del 10% y determine aproximadamente las regiones. $R/$ Si hay evidencia en contra.

8. Un médico afirma que los pesos de niñas de 3 años de cierta provincia son muy similares, específicamente señala que la desviación estándar de los pesos no llega a ser de $3kg$. Para analizar esta afirmación se toma el peso en kg de 10 niñas:

   $$14.5, 11.6, 12.8, 15.1, 14.2, 13.7, 12.9, 13.8, 14.1, 11.9.$$

   Suponga que el peso de niñas de 3 años sigue una distribución normal. Utilice $\alpha = 0.1$

   (a) Determine las regiones de aceptación y rechazo para χ^2.

   (b) Determine las regiones de aceptación y rechazo para S^2.

   (c) Determine el valor P de la prueba. $R/$ No

   (d) ¿Hay evidencia en contra de la afirmación indicada por el doctor?

   (e) Determine la probabilidad de aceptar H_0 sabiendo que la desviación estándar de los pesos de lo niños es en realidad $\sigma = 2.95$.

9. Un cliente se queja de que la duracción de la batería de su computadora, cargada completamente, varía demasiado. La empresa que fabrica las baterías indica que las duraciones de la batéria que elaboran tiene una desviación estándar inferior a 35 minutos. Las duraciones de ocho baterías de computadora son 151, 153, 175, 134, 170, 172, 156 y 114 minutos. Suponga que las duraciones se distribuyen normalmente Con un nivel de significancia del 10%, ¿considera que el cliente puede demandar a la empresa? $R/$ Si

10. Seguidamente se presenta una muestra de notas obtenidas por estudiantes de Ingeniería en Computación del ITCR, en el curso Estadística, el II semestre del 2009:

75	75	75	85	70	70	80	85	75	60	70	60
80	60	80	75	75	70	90	50	80	75	70	

    Suponga que la nota de Estadística sigue una Distribución Normal. Un profesor afirma que menos de la mitad de los estudiantes que llevaron estadística en el 2009 tuvieron una nota inferior a 80. Una profesora, al ver las notas de la muestra, indica que la nota promedio de estádistica en el II semestre del 2009 es de 80.

    (a) Determine el valor P para la afirmación de la profesora.

    (b) ¿Apoya la afirmación de la profesora? Justifique. $R/$

    (c) Determine las regiones de aceptación y rechazo para la afirmación del profesor, utilizando $\alpha = 0.1$

    (d) ¿Apoya la afirmación del profesor? Justifique. $R/$ Si

11. Seguidamente se presenta una muestra de notas obtenidas por de Ingeniería en Computación del ITCR, carné 2004, en el curso Matemática Discreta:

75	50	35	45	85	85	85	70	30
60	85	85	90	70	70	70	55	85

Suponga que la nota de Matemática Discreta para estudiantes carné 2004 sigue una Distribución Normal. ¿Hay evidencia en contra de que la nota promedio de este curso, para estudiantes carné 2004, sea igual a 75? $R/$ No

12. En una ciudad, se entrevistaron a 598 personas de 18 años sin hijos obteniendo que 306 de estas personas afirman que desean tener hijos. ¿Considera que la mitad de los adultos de 18 años sin hijos, desean tener hijos? Justifique. $R/$

13. En el restaurante Buen Gusto, un cliente se quejó del excesivo tiempo que se tarda en servir los alimentos una vez que se ordena. El gerente ha asegurado que la duración promedio de servir los alimentos después de ordenar es inferior a 7 minutos. Se han registrado diez duraciones en minutos: $2, 3, 6, 4, 7, 3, 8, 9, 5, 4$. Suponga que el tiempo que se tarda en servir los alimentos una vez que se ordena se distribuyen normalmente.

    (a) Determine las regiones de aceptación y rechazo para la afirmación del gerente, utilizando un nivel de significancia del 5%.

    (b) ¿Considera aceptable la afirmación del gerente? Justifique $R/$ Si

14. Sean $H_0 : \mu = 120$ y $H_1 : \mu < 120$. Para contrastar dicha hipótesis se toma una muestra aleatoria de 40 observaciones con una desviación estándar de 10.3.

    (a) ¿Cuál es el nivel de significancia si la región de rechazo es $R_1 =]-\infty, 117.2[$? $R/$ $\alpha = 0.0427$.

    (b) ¿Cuál es el nivel de significancia si la región de rechazo es $R_2 =]-\infty, 115.87[$? $R/$ $\alpha = 0.005543$.

15. Un profesor del ITCR afirma que menos del 25% de los estudiantes fuman. Ante este problema, Juan tomó una muestra aleatoria de 24 estudiantes y realizó el contraste de hipótesis con un nivel de significancia del 5% obteniendo correctamente que el valor P es aproximadamente de 0.31918.

    (a) Determine aproximadamente el número de estudiantes fumadores que Juan observó en la muestra. $R/$ 5

    (b) ¿Considere aceptable la afirmación del profesor? $R/$ $Valor\ P > 0.05$, hay evidencia en contra de la afirmación del profesor.

16. Un convenio entre la asociación de trabajadores de una empresa y el gerente exige una producción media diaria de al menos 50 unidades. En 150 días de estudio se observó una producción media de 47.3 unidades, con una desviación de 5.7 unidades.

(a) Plantee las hipótesis nula y alternativa para esta prueba y halle las regiones de aceptación y rechazo con un nivel de significancia de 0.05. $R/ \quad u_c = 49.2344$

(b) Determine si se cumple la exigencia del convenio con un nivel de significancia de $0.05. R/$ $\overline{x} \in R$, hay evidencia de que no se cumple con la clausura.

(c) Si la desviación estándar de la producción diaria es de 6 unidades. ¿De qué tamaño debe ser una muestra para que la prueba tenga una significancia de 2% y una potencia de 95% cuando el verdadero promedio es de 48 unidades ? $R/ \quad 123$

17. La empresa de llantas Mundiales ha sacado un nuevo modelo de llantas para automóviles al mercado asegurando que su duración tiene una desviación estándar menor a 5000 km. Suponga que la duración de las llantas se distribuye normalmente. Se quiere contrastar la afirmación a un nivel de significancia del 10% con una muestra aleatoria de 25 llantas del nuevo modelo.

    (a) Determine las regiones de aceptación y rechazo de la prueba para el estadístico $S. R/ \quad A =]4.038\,71, +\infty[\,, R = [0, 4.038\,71[$

    (b) ¿Acote la probabilidad del error tipo II de la prueba si duración de las llantas del nuevo modelo tiene una desviación estándar de 4700 km? $R/ \quad \beta \in]0.8, 0.9[\,.$

18. Un médico afirma que la edad promedio de personas con cierta enfermedad en San José es de 32 años. Para investigar dicha afirmación se tomó una muestra de 25 personas de la capital que poseen la enfermedad y se obtuvo una edad promedio de 33.1 años con una desviación de 3.8 años.

    (a) Utilizando un nivel de significancia del 10%, determine las regiones de aceptación y rechazo de $\overline{X}$ para contrastar la afirmación del médico. $R/ \quad A =]30.699\,7, 33.300\,3[$

    (b) ¿Existe evidencia en contra de la afirmación del médico? $R/$ Se acepta la afirmación del médico

19. Un médico afirma que la edad promedio de personas con sobrepeso en cierta ciudad es de 35 años. Para investigar dicha afirmación se tomó una muestra de 20 personas con sobrepeso de dicha ciudad y se observó un edad promedio de 36.7 años con una desviación de 4.5 años. Al realizar contraste de hipótesis se obtuvo que uno de los promedios críticos es $u_c = 33.1389$.

    (a) Determine aproximadamente el nivel de significancia utilizado en el contraste de hipótesis realizado. $R/ \quad \alpha \approx 0.08$

    (b) ¿Existe evidencia en contra de la afirmación del médico? $R/ \quad u_{c_2} = 36.8611$. Se acepta la afirmación del médico

20. Un profesor de la Universidad Futuro Garantizado cree que más de la mitad de los estudiantes matriculados se retiran de al menos una materia. En una muestra de 80 estudiantes, se observó que 45 se habían retirado de al menos una materia.

    (a) Determine el valor P de la prueba de hipótesis para contrastar la creencia del profesor. $R/$ Valor $P \approx 0.131357$

(b) ¿Considere aceptable la creencia del profesor? $R/$ Hay evidencia en contra de la afirmación del profesor.

(c) ¿De qué tamaño debe ser una muestra para que la prueba tenga una significancia de 2% y una potencia de 90% cuando el verdadero porcentaje de estudiantes que se retiran de al menos una materia es del 60% ? $R/$ $n = 273$, note que np_0, np_1, nq_0, nq_1 son mayores a 5

21. Una empresa dedicada a la fabricación de pantalones para mujeres desea introducir su producto en el mercado nacional. En estudios previos se ha determinado que las mujeres entre 22 y 40 años son las que tiene mayor poder adquisitivo. La empresa desea determinar si la medida de la cintura de este grupo de mujeres no varía mucho, pues esto va a influir en la cantidad y variedad de tallas ofrecidas al mercado. Específicamente, la empresa desea determinar si la medida de la cintura de estas de mujeres tiene una desviación estándar σ no mayor a 6 cm. Suponga que la medida de la cintura de las mujeres, entre 22 años y 40 años, se distribuye normalmente. En una muestra de 30 mujeres entre 22 y 40 años se observó que la medida de sus cinturas tiene una desviación estándar de 7.3 centímetros. Se quiere contrastar si σ no es mayor a 6 cm a un nivel de significancia del 10%.

(a) Determine las regiones de aceptación y rechazo de la prueba para el estadístico S^2.$R/$ $A =]0, 48.\,522\,4[$, $R =]48.\,522\,4, +\infty[$

(b) ¿Es razonable suponer que σ no es mayor a 6 cm? $R/$ Existe evidencia en contra de que $\sigma \leq 6$

(c) ¿Acote la probabilidad del error tipo II de la prueba si $\sigma = 8.1$ cm? $R/$ $\beta \in]0.1, 0.2[$

22. Una pequeña tienda para su funcionamiento requiere de una venta promedio diaria de por lo menos 50 mil colones para cubrir sus gastos. En 90 días de estudio se obtuvo una venta promedio de 49 500 colones, con una desviación de 5700 colones.

(a) Plantee las hipótesis nula y alternativa para esta prueba y halle las regiones de aceptación y rechazo con un nivel de significancia de 0.05. $R/$ Región de rechazo: $]-\infty, 49.\,234\,4[$, Región de aceptación: $]49.\,234\,4, +\infty[$

(b) Determine si la tienda debería dejar de funcionar con un nivel de significancia de 0.05. $R/$ hay evidencia de que no se cumple con la clausura.

23. Una muestra aleatoria de doce estudiantes de sétimo añodel colegio San Beto tiene las siguientes estaturas.en centímetros

$$142, 158, 135, 129, 159, 148, 153, 139, 133, 151, 131, 140.$$

Suponga que la estatura de los estudiantes de sétimo se distribuye normalmente. Un grupo de doctores considera que la estatura promedio de los estudiantes de sétimo del colegio es de a lo sumo 140 cm. Determine si los datos aportan evidencia en contra de lo indicado por los doctores con un nivel de significancia de 0.05, determinando las regiones de aceptación y rechazo. $R/$ $u_c = 145.\,414$. No hay evidencia en contra de la afirmación

24. Un psicólogo asegura que el 35% de la población estudiantil padece de depresión. En una muestra de 150 estudiantes se encontró que 60 padecen depresión.

(a) Determine el valor P de la prueba que permite contrastar la afirmación del psicólogo.$R/$ $Valor\ P \approx 0.200\,546$

(b) Utilizando un nivel de significancia del 5%, si el porcentaje de la población estudiantil que padece depresión es en realidad 40%, ¿cuál es la potencia de la prueba? $R/$ $p_{c_1} = 0.273\,669, p_{c2} = 0.426\,331, 1 - \beta \approx 0.255\,416$

3 Pruebas de hipótesis con dos parámetros

3.1 Pruebas con dos promedios

Considere la siguiente notación

Población	Párametros		Tamaño de muestra	Estadísticos	
	Media Poblacional	Variancia Poblacional		Media Poblacional	Variancia muestral
X_1	μ_1	σ_1^2	n_1	$\overline{X}_1$	S_1^2
X_2	μ_2	σ_2^2	n_2	$\overline{X}_2$	S_2^2

Recuerde que $\overline{X}_1 - \overline{X}_2$ es un estimador insesgado de $\mu_1 - \mu_2$, además $Var\left(\overline{X}_1 - \overline{X}_2\right) = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$. La hipótesis nula, para este tipo de pruebas, es de la forma

$$H_0 : \mu_1 - \mu_2 = d_0$$

donde d_0 es llamada la diferencia nula. Por resultados anteriores y bajo la hipótesis nula $H_0 : \mu_1 - \mu_2 = d_0$, se sabe que:

1. Si $\overline{X}_1$ y $\overline{X}_2$ siguen una distribución normal para muestras de tamaño n_1 y n_2 respectivamente, y se conocen σ_1 y σ_2 entonces

$$\overline{X}_1 - \overline{X}_2 \sim N\left(d_0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \Longrightarrow Z = \frac{\overline{X}_1 - \overline{X}_2 - d_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

   Recuerde que está opción se utiliza en alguno de los siguentes casos:

   (a) n_1 y n_2 son mayores iguales a 30. No es necesario conocer σ_1 y σ_2, pues $\sigma_1 \approx s_1$ y $\sigma_2 \approx s_2$.

   (b) Las poblaciones X_1 y X_2 siguen una distribución normal y se conoce σ_1 y σ_2.

   (c) La población X_1 sigue una distribución normal, $n_2 \geq 30$, y se conoce σ_1.No es necesario conocer σ_2, pues $\sigma_2 \approx s_2$.

   (d) La población X_2 sigue una distribución normal, $n_1 \geq 30$, y se conoce σ_2.No es necesario conocer σ_1, pues $\sigma_1 \approx s_1$.

2. Si las poblaciones X_1 y X_2 siguen una distribución normal y se desconocen las desviaciones poblacionales σ_1 y σ_2 pero se suponen iguales, entonces

$$T = \frac{\overline{X}_1 - \overline{X}_2 - d_0}{\sqrt{\dfrac{s_p^2}{n_1} + \dfrac{s_p^2}{n_2}}}$$

sigue una distribución t de Student con $v = n_1 + n_2 - 2$ grados de libertad, donde s_p^2 es una estimación dada por el estadístico

$$S_p^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}, \quad v = n_1 + n_2 - 2$$

Note que, al suponer que $\sigma_1 = \sigma_2$, S_p^2 es un estimador insesgado de $\sigma_1^2 = \sigma_2^2$:

$$\begin{aligned} E\left(S_p^2\right) &= E\left(\frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}\right) \\ &= \frac{(n_1 - 1) E\left(S_1^2\right) + (n_2 - 1) E\left(S_2^2\right)}{n_1 + n_2 - 2} \end{aligned}$$

Dado que S_1^2 y S_2^2 son estimadores insesgados de σ_1^2 y σ_2^2 respectivamente, entonces

$$\begin{aligned} E\left(S_p^2\right) &= \frac{(n_1 - 1) \sigma_1^2 + (n_2 - 1) \sigma_2^2}{n_1 + n_2 - 2} \\ &= \frac{(n_1 - 1) \sigma_1^2 + (n_2 - 1) \sigma_1^2}{n_1 + n_2 - 2}, \text{ pues se supone que } \sigma_1^2 = \sigma_2^2 \\ &= \sigma_1^2 = \sigma_2^2 \end{aligned}$$

3. Si las poblaciones X_1 y X_2 siguen una distribución normal y se desconocen las desviaciones poblacionales σ_1 y σ_2 pero no se suponen iguales, entonces

$$T = \frac{\overline{X}_1 - \overline{X}_2 - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

sigue una distribución t de Student con v grados de libertad, donde

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}$$

Ejemplo 71 *Se obtuvieron las estaturas de 20 mujeres y 30 hombres seleccionados aleatoriamente de la población de alumnos de cierta escuela. En la siguiente tabla se resumen los datos encontrados*

	Número	*Media*	*Desviación*
Hombres (m):	*30*	*69.8*	*1.92*
Mujeres (f):	*20*	*63.8*	*2.18*

Suponga que la estatura de las mujeres y la estatura de los hombres se distribuyen normalmente. ¿Puede concluirse, con un nivel de significancia de 4%*, que el promedio de estaturas de los hombres*

de la escuela supera en más de 3cm el promedio de las mujeres? Suponga que $\sigma_1 = \sigma_2$.

En este caso la afirmación es $u_m - u_f > 3$ *y las hipótesis son*

$$H_0 : u_m - u_f = 3 \ (\leq), \quad H_1 : u_m - u_f > 3$$

Datos:

$$n_1 = 30, \quad n_2 = 20, \quad \overline{x}_1 = 69.8, \quad \overline{x}_2 = 63.8, \quad s_1^2 = 1.92, \quad s_2^2 = 2.18.$$

Como las poblaciones son normales, se desconocen las varianzas poblacionales, $n_2 < 30$ *y se supone que* $\sigma_1 = \sigma_2$ *entonces*

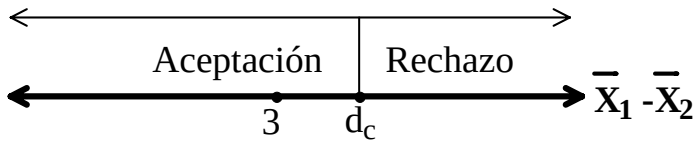
$$T = \frac{\overline{X}_1 - \overline{X}_2 - d_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \sim t\,(v)$$

donde el valor de S_p^2 *observado es*

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2} = \frac{19 \cdot 2.18^2 + 29 \cdot 1.92^2}{48} = 4.108\,4$$

y $v = n_1 + n_2 - 2 = 48$. *Seguidamente se presentan varias formas de resolver el problema.*

1. *Utilizando regiones con estadístico* $\overline{X}_1 - \overline{X}_2$. *Las regiones para* $\overline{X}_1 - \overline{X}_2$ *se definen a continuación*



Recuerde que para hallar el valor crítico se hace uso del nivel de significancia $\alpha = 0.04$*:*

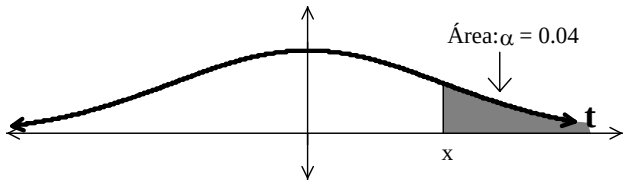
$$\begin{aligned} 0.04 &= P\,(H_1|H_0) = P\left(\overline{X}_1 - \overline{X}_2 > d_c | u_m - u_f = 3\right) \\ \Longrightarrow \quad 0.04 &= P\left(t > \frac{d_c - 3}{\sqrt{s_p^2/n_1 + s_p^2/n_2}}\right) = P\left(t > \frac{d_c - 3}{0.585\,121}\right) \end{aligned}$$

Note que $x = \dfrac{d_c - 3}{0.585\,121}$ *es positivo pues* $d_c > 3$:

$$\Longrightarrow \frac{d_c - 3}{0.585\,121} = t_{0.96,48}$$

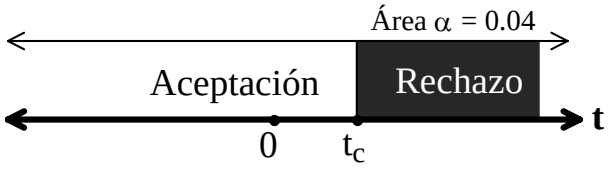
Como $t_{0.96,48} = -t_{0.04,48} = 1.78854$

Entonces $d_c = 4.\,04648$



Como el valor observado $\overline{x}_1 - \overline{x}_2 = 69.8 - 63.8 = 6$ *está en la región de rechazo se rechaza* H_0.

2. *Utilizando el enfoque clásico con estadístico T. Las regiones para t se definen a continuación*



El valor t crítico es

$$t_c = t_{0.96,48} = -t_{0.04,48} = 1.78854$$

por lo tanto la región de aceptación para t es $]-\infty, 1.78854[$ *y de rechazo* $]1.78854, +\infty[$. *Por otro lado el valor t observado es*

$$t_{obs} = \frac{69.8 - 63.8 - 3}{\sqrt{\frac{4.108\,4}{20} + \frac{4.108\,4}{30}}} = 5.1271$$

Como $t_{obs} \in]1.78854, +\infty[$ *entonces se rechaza* H_0.

3. *Utilizando el valor P y estadístico* $\overline{X}_1 - \overline{X}_2$. *Dado que* $H_1 : u_m - u_f > 3$ *y el valor observado del estadístico* $\overline{X}_1 - \overline{X}_2$ *es*

$$d_{obs} = \overline{x}_1 - \overline{x}_2 = 6$$

entonces se tiene que

$$\begin{aligned} \textit{Valor } P &= P\left(\overline{X}_1 - \overline{X}_2 \geq 6 | u_m - u_f = 3\right) \\ &= P\left(t \geq \frac{6-3}{0.5851}\right) = P\left(t \geq 5.\,127\,33\right) 4 \end{aligned}$$

Utilizando la tabla de t con $v = 48$ *se tiene que*

$$\textit{Valor } P = P\left(t \geq 5.\,127\,33\right) < 0.005 < \alpha = 0.04.$$

Por lo tanto se rechaza H_0.

4. *Utilizando el valor P y estadístico T. Bajo la hipótesis nula* $H_0 : u_m - u_f = 3$, *se tiene que el valor observado de T es*

$$t_{obs} = \frac{\overline{x}_1 - \overline{x}_2 - d_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{69.8 - 63.8 - 3}{\sqrt{\frac{4.108\,4}{20} + \frac{4.108\,4}{30}}} = 5.1271$$

y entonces

$$\textit{Valor } P = P\left(T \geq t_{obs}\right) = P\left(t \geq 5.\,127\,33\right) < 0.005 < \alpha = 0.04$$

por lo tanto se rechaza H_0.

Así, en los 4 casos se rechaza H_0 *y se acepta la afirmación. Puede concluirse que no hay evidencia en contra de que el promedio de estaturas de los hombres de la escuela supera en más de 3cm el promedio de las mujeres*

Por otro lado, recuerde que si quiere disminuir las probabilidades de ocurrencia de los errores, se debe elegir un tamaño de muestra mayor. Este tamaño esta dado por el siguiente teorema.

Teorema 36 *Tomando una muestras de tamaños apropiados $n_1 = n_2 = n$ de modo que $\overline{X}_1 - \overline{X}_2$ se distribuya aproximadamente normal, para probar*

$$H_0 : \mu_1 - \mu_2 = d_0$$

con un nivel de significancia de α y una potencia mínima de $1-\beta$ para la hipótesis alternativa específica

$$H_1' : \mu_1 - \mu_2 = d_1$$

puede tomarse una muestra de tamaño

1. $n \geq \dfrac{(|z_\alpha| + |z_\beta|)^2 \left(\sigma_1^2 + \sigma_2^2\right)}{(d_1 - d_0)^2}$ *si la prueba es de una cola.*

2. $n \geq \dfrac{\left(\left|z_{\alpha/2}\right| + |z_\beta|\right)^2 \left(\sigma_1^2 + \sigma_2^2\right)}{(d_1 - d_0)^2}$ *si la prueba es de dos colas.*

Prueba. *Ejercicio.* ■

Ejercicio 36 *Una muestra de países latinoamericanos y europeos revelo los siguientes resultados, donde n es el número de países, $\overline{x}$ el promedio de las esperanzas de vida en años y s la desviación estándar de las esperanzas de vida en años:*

Países	n	$\overline{x}$	s
Europeos	35	76.16	1.85
Latinoamericanos	31	60.7	1.08

Pruebe la hipótesis de que la esperanza de vida en los países Europa es superior a la de los países Latinoamericanos en por lo menos 16 años. *R/ Valor $P \approx 0.070781$, se acepta la afirmación*

3.2 Pruebas con dos proporciones

Para $i = 1, 2$, considere la población i con proporción poblacional p_i de cierta característica y estadístico $\widehat{P}_i = \dfrac{B_i}{n}$ para muestras aleatorias de tamaño n_i, donde B_i es el número de sujetos de la muestra que poseen la característica. Asuma que $\widehat{P}_1$ y $\widehat{P}_2$ son independientes. La hipótesis nula, para este caso, es de la forma

$$H_0 : p_1 - p_2 = d_0$$

donde d_0 es llamada la diferencia nula. Se sabe que si las muestras son grandes entonces

$$\widehat{P}_1 - \widehat{P}_2 \sim N\left(d_0, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right)$$

Sin embargo, se presenta el siguiente PROBLEMA: bajo la hipótesis nula $H_0 : p_1 - p_2 = d_0$, no se sabe el valor de p_1 ni de p_2. Se puede hallar una solución satisfactoria para el caso en que $d_0 = 0$ y otra solución un poco imperfecta para los otros casos.

3.2.1 Prueba de que dos proporciones son iguales

Para el caso en que $d_0 = 0$, la hipótesis nula es de la forma

$$H_0 : p_1 = p_2$$

En este caso se define la proporción muestral común por

$$\widehat{P} = \frac{n_1\widehat{P}_1 + n_2\widehat{P}_2}{n_1 + n_2} = \frac{B_1 + B_2}{n_1 + n_2}$$

que es la proporción total de éxitos en las dos muestras. Recuerde que si $n_i p_i \geq 5$ y $n_i q_i \geq 5$ para $i = 1, 2$, entonces

$$\widehat{P}_1 - \widehat{P}_2 \sim N\left(p_1 - p_2, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right)$$

Note que bajo la hipótesis nula $H_0 : p_1 = p_2$, el estadístico $\widehat{P}$ es un estimador insesgado tantode p_1 como de p_2, pues

$$E\left(\widehat{P}\right) = E\left(\frac{n_1\widehat{P}_1 + n_2\widehat{P}_2}{n_1 + n_2}\right) = \frac{n_1 E\left(\widehat{P}_1\right) + n_2 E\left(\widehat{P}_2\right)}{n_1 + n_2} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

entonces bajo $H_0 : p_1 = p_2$:

$$E\left(\widehat{P}\right) = \frac{n_1 p_1 + n_2 p_1}{n_1 + n_2} = p_1 = p_2.$$

Así, para muestras de tamaño n_1 y n_2 el valor observado $\widehat{p}$ de $\widehat{P}$ es una estimación de p_1 y p_2 siempre bajo H_0. Por lo tanto, bajo $H_0 : p_1 = p_2$, si las distribuciones muestrales $\widehat{P}_1$ y $\widehat{P}_2$ son independientes y si las muestras son grandes[7] ($n_i\widehat{p} \geq 5$ y $n_i\widehat{q} \geq 5$) entonces,

$$\widehat{P}_1 - \widehat{P}_2 \sim N\left(0, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right) \underset{aprox}{\sim} N\left(0, \frac{\widehat{p}\widehat{q}}{n_1} + \frac{\widehat{p}\widehat{q}}{n_2}\right) \tag{1}$$

donde $\widehat{p} = \dfrac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1 + n_2}$ y $\widehat{q} = 1 - \widehat{q}$. Es decir

$$Z = \frac{\widehat{P}_1 - \widehat{P}_2}{\sqrt{\dfrac{\widehat{p}\widehat{q}}{n_1} + \dfrac{\widehat{p}\widehat{q}}{n_2}}} \underset{aprox}{\sim} N(0, 1)$$

Además, para el tamaño de la muestra se tiene el siguiente teorema.

[7] Recuerde que las muestras son grandes si $n_i p_i \geq 5$ y $n_i q_i \geq 5$, para $i = 1, 2$. Bajo H_0, como $\widehat{p}$ es una estimación de p_1 y p_2, entonces lasmuestras son grandes si $n_i\widehat{p} \geq 5$ y $n_i\widehat{q} \geq 5$, para $i = 1, 2$.

Teorema 37 *Tomando una muestras de tamaños apropiados $n_1 = n_2 = n$ de modo que $\widehat{P}_1 - \widehat{P}_2$ se distribuya aproximadamente normal, para probar*

$$H_0 : p_1 = p_2$$

con un nivel de significancia de α y una potencia mínima de $1-\beta$ cuando las proporciones verdaderas son $p_1 = p_1'$ y $p_2 = p_2'$, puede tomarse una muestra de tamaño

1. $n \geq \dfrac{\left(|z_\alpha| \sqrt{\dfrac{(p_1' + p_2')(q_1' + q_2')}{2}} + |z_\beta| \sqrt{p_1'q_1' + p_2'q_2'} \right)^2}{(p_1' - p_2')^2}$ *para muestras de una cola.*

2. $n \geq \dfrac{\left(|z_{\alpha/2}| \sqrt{\dfrac{(p_1' + p_2')(q_1' + q_2')}{2}} + |z_\beta| \sqrt{p_1'q_1' + p_2'q_2'} \right)^2}{(p_1' - p_2')^2}$ *para muestras de dos colas.*

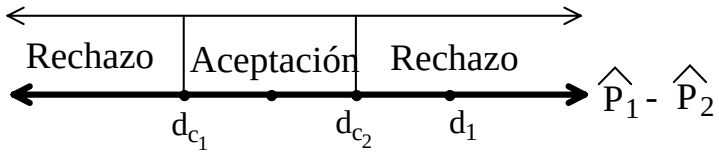
Prueba. *Se probará el caso 2, el otro caso se deja de ejerccio. Suponga que se quiere probar*

$$H_0 : p_1 = p_2, \quad H_1 : p_1 \neq p_2$$

con un nivel de significancia de α y una potencia mínima de $1-\beta$ para la hipótesis alternativa específica

$$H_1' : p_1 = p_1', p_2 = p_2'$$

donde α y β son menores al 50% y $p_1' - p_2' = d_1 > d_{c2}$. Entonces las regiones de aceptación y rechazo son



y las probabilidades de que sucedan los errores tipo I y II están acotadas por:

$$\alpha \geq P(H_1|H_0) \quad y \quad \beta \geq P(H_0|H_1')$$

Asumiendo que la muestra es suficientemente grande de modo que

$$\widehat{P}_1 - \widehat{P}_2 \sim N\left(p_1 - p_2, \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2} \right)$$

Se tiene que

$$\frac{\alpha}{2} \geq P\left(\widehat{P}_1 - \widehat{P}_2 > d_{c2} | p_1 = p_2\right) = P\left(Z > \frac{d_{c2}}{\sqrt{2\frac{\widehat{p}\widehat{q}}{n}}} \right)$$

$$\Longrightarrow \quad \phi\left(z_{1-\alpha/2}\right) \leq \phi\left(\frac{d_{c2}}{\sqrt{2\widehat{p}\widehat{q}/n}} \right) \Longrightarrow z_{1-\alpha/2} \leq \frac{d_{c2}}{\sqrt{2\widehat{p}\widehat{q}/n}}$$

Como $z_{1-\alpha/2} = -z_{\alpha/2}$ *entonces*

$$z_{\alpha/2} \geq \frac{-d_{c2}}{\sqrt{2\widehat{p}\widehat{q}/n}} \Longrightarrow |z_{\alpha/2}| \sqrt{\frac{2\widehat{p}\widehat{q}}{n}} \leq d_{c2} \qquad (1)$$

Por otro lado, note que

$$P\left(\widehat{P}_1 - \widehat{P}_2 < d_{c2} | H_1'\right) \approx 0,$$

pues la hipótesis espécifica beneficia a la región de rechazo de la derecha $(d_1 > d_{c2})$*. Si está hipótesis se asume, es poco probable que* $\widehat{P}_1 - \widehat{P}_2$ *este en la región de la izquierda. Así*

$$\begin{aligned}
\beta &\geq P\left(H_0 | H_1'\right) = P\left(d_{c1} < \widehat{P}_1 - \widehat{P}_2 < d_{c2} | H_1'\right) \\
&\approx P\left(d_{c1} < \widehat{P}_1 - \widehat{P}_2 < d_{c2} | H_1'\right) + \underbrace{P\left(\widehat{P}_1 - \widehat{P}_2 < d_{c2} | H_1'\right)}_{\approx 0} \\
&\approx P\left(\widehat{P}_1 - \widehat{P}_2 < d_{c2} | p_1 = p_1' \wedge p_2 = p_2'\right) = P\left(Z < \frac{d_{c2} - (p_1' - p_2')}{\sqrt{\frac{p_1'q_1'}{n} + \frac{p_2'q_2'}{n}}}\right)
\end{aligned}$$

Por lo tanto

$$z_\beta \geq \frac{d_{c2} - (p_1' - p_2')}{\sqrt{\frac{p_1'q_1'}{n} + \frac{p_2'q_2'}{n}}} \Longrightarrow |z_\beta| \sqrt{\frac{p_1'q_1'}{n} + \frac{p_2'q_2'}{n}} \leq (p_1' - p_2') - d_{c2} \qquad (2)$$

De 1 *y* 2 *se obtiene que*

$$|z_{\alpha/2}| \sqrt{\frac{2\widehat{p}\widehat{q}}{n}} + |z_\beta| \sqrt{\frac{p_1'q_1'}{n} + \frac{p_2'q_2'}{n}} \leq p_1' - p_2'$$

Además $\widehat{p} = \dfrac{n\widehat{p}_1 + n\widehat{p}_2}{n+n} = \dfrac{\widehat{p}_1 + \widehat{p}_2}{2}$ *y* $\widehat{q} = 1 - \widehat{p} = \dfrac{\widehat{q}_1 + \widehat{q}_2}{2}$*, así*

$$|z_{\alpha/2}| \sqrt{\frac{(p_1' + p_2')(q_1' + q_2')}{2n}} + |z_\beta| \sqrt{\frac{p_1'q_1' + p_2'q_2'}{n}} \leq p_1' - p_2'$$

y despejando n *se obtiene el resultado.* ■

Ejemplo 72 *De una muestra de 100 personas de la ciudad A, 34 tienen sobrepeso. Una muestra de 200 personas de la ciudad B indica que 50 tienen sobrepeso. Ante estos resultados, una persona afirma que hay una mayor proporción de personas con sobrepeso en la ciudad A, con respecto a la otra ciudad.*

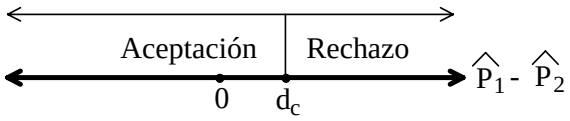
1. *Con un nivel de significancia de 0.08, ¿hay evidencia que respalde la afirmación realizada por la persona?*

Sean

p_1 : *la proporción de personas con sobrepeso en la ciudad A*
p_2 : *la proporción de personas con sobrepeso en la ciudad B.*

Se tiene que

Afirmación: $p_1 - p_2 > 0$,
$H_0 : p_1 - p_2 = 0\,(\leq)$
$H_1 : p_1 - p_2 > 0$



Note que

$$d_{obs} = \widehat{p}_1 - \widehat{p}_2 = \frac{34}{100} - \frac{50}{200} = 0.09, \quad \widehat{p} = \frac{84}{300} = 0.28 \Longrightarrow \widehat{q} = 0.72$$

Además para determinar d_c *se tiene que*

$$0.08 = \alpha = P\left(H_1|H_0\right) = P\left(\widehat{P}_1 - \widehat{P}_2 > d_c|p_1 = p_2\right)$$

Como $n_1\widehat{p} = 28, n_1\widehat{q} = 72, n_2\widehat{p} = 56$ *y* $n_2\widehat{q} = 144$ *son mayores a* 5, *entonces*

$$0.08 = P\left(Z > \frac{d_c}{\sqrt{\widehat{p}\widehat{q}\left(\frac{1}{100} + \frac{1}{200}\right)}}\right) = P\left(Z > \frac{d_c}{0.05499\,09}\right)$$
$$\Longrightarrow \frac{d_c}{0.05499\,09} = z_{0.92} = -z_{0.08} = 1.41 \Longrightarrow d_c = 0.07753\,72$$

Las regiones para $\widehat{P}_1 - \widehat{P}_2$ *son*

Región de aceptación: $A =]-\infty, 0.07753\,72[$, *Región de rechazo:* $R =]-\infty, 0.07753\,72[$.

Como $d_{obs} \in R$ *entonces se rechaza* H_0,*si se respalda la afirmación de la persona.*

2. *¿Cuál es el Valor P de la Prueba?*

Como $n_1\widehat{p} = 28, n_1\widehat{q} = 72, n_2\widehat{p} = 56$ *y* $n_2\widehat{q} = 144$ *son mayores a* 5 *entonces, bajo* H_0 *se tiene que*

$$Z = \frac{\widehat{P}_1 - \widehat{P}_2}{\sqrt{\frac{\widehat{p}\widehat{q}}{n_1} + \frac{\widehat{p}\widehat{q}}{n_2}}} \underset{aprox}{\sim} N\left(0, 1\right)$$

Como el valor observado de $\widehat{P}_1 - \widehat{P}_2$ *es* $d_{obs} = \widehat{p}_1 - \widehat{p}_2 = 0.09$, *entonces el valor observado de* Z *es*

$$z_{obs} = \frac{d_{obs}}{\sqrt{\widehat{p}\widehat{q}\left(\frac{1}{100} + \frac{1}{200}\right)}} = \frac{0.09}{0.05499\,09} = 1.\,636\,63$$

Como $\widehat{P}_1 - \widehat{P}_2$ *y* Z *tienen el mismo comportamiento, y* $H_1 : p_1 - p_2 > 0$ *entonces*

$$Valor\ P = P\left(Z \geq z_{obs}\right) \approx 1 - 0.9484 = 0.051\,6$$
Note que utilizando el Valor P también se rechaza H_0

Ejemplo 73 *Un profesor del ITCR afirma que la promoción en el curso de Cálculo para Computación es similar a la promoción en el curso de álgebra para Computación. En una muestra de 120 estudiantes carné 2004-2006 que cursaron Cálculo para Computación, 88 lo aprobaron. Por otro lado, en una muestra de 120 estudiantes carné 2004-2006 que cursaron Álgebra para Computación, 99 lo aprobaron. ¿Hay evidencia que respalde la afirmación realizada por el profesor ? Utilice* $\alpha = 0.05$

Sean

p_1 : *la promoción en Cálculo para Computación*
p_2 : *la promoción en Álgebra para Computación.*

El profesor afirma que $p_1 = p_2$ *entonces*

$$H_0 : p_1 = p_2, \quad H_1 : p_1 \neq p_2$$

Además

$$n_1 = n_2 = 120, \widehat{p}_1 = \frac{88}{120}, \widehat{p}_2 = \frac{99}{120}, \quad \widehat{p} = \frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1 + n_2} = \frac{187}{240} \text{ y } \widehat{q} = \frac{53}{240}$$

Como $n_1\widehat{p} = n_2\widehat{p} = 93.5$ *y* $n_1\widehat{q} = n_2\widehat{q} = 26.5$ *son mayores a* 5 *entonces*

$$Z = \frac{\widehat{P}_1 - \widehat{P}_2}{\sqrt{\frac{\widehat{p}\widehat{q}}{n_1} + \frac{\widehat{p}\widehat{q}}{n_2}}} \underset{aprox}{\sim} N(0,1)$$

El valor observado de $\widehat{P}_1 - \widehat{P}_2$ *es*

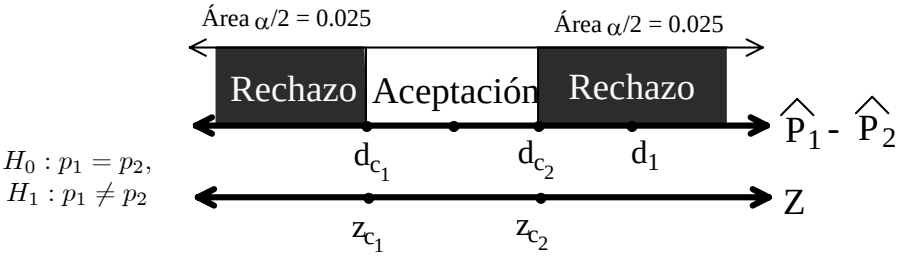
$$d_{obs} = \widehat{p}_1 - \widehat{p}_2 = \frac{88 - 99}{120} = \frac{-11}{120}$$

y el valor observado Z *es*

$$z_{obs} = \frac{\frac{-11}{120}}{\sqrt{\frac{2 \cdot \frac{187}{240}\frac{53}{240}}{120}}} = -1.711\,75$$

Se realizará la prueba de dos formas.

1. *Regiones utilizando* Z. *Recuerde que* $\alpha = 0.05$.



Note $z_{c_1} = z_{0.025} = -1.96$ *y por simétria* $z_{c_2} = 1.96$. *Así las regiones para* Z *son*

$$\textit{Región Aceptación } A =]-1.96, 1.96[\textit{ Región Rechazo } R = \mathbb{R} - A.$$

Como $z_{obs} = -1.711\,75 \in A$, *entonces se acepta* H_0. *No hay evidencia significativa de que* $p_1 = p_2$.

2. *Valor P utilizando Z. Como* $H_1 : p_1 \neq p_2$ *y* $z_{obs} = -1.711\,75$ *entonces note que*

$$\begin{aligned} \textit{Valor } P &= 2 \cdot P\left(Z < -1.71175\right) = 2 \cdot 0.043633 \\ &= 0.087\,266 > 0.05 = \alpha \end{aligned}$$

Por lo tanto se acepta H_0. *y hay una leve evidencia de que* $p_1 = p_2$ *Note que si* α *fuera* 0.1 *entonce se rechazaría* H_0.

Ejercicio 37 *De una muestra de 100 acciones de la Bolsa de Valores A, 32 tuvieron ganancia el lunes pasado. Un muestra de 100 acciones de la B indica que 27 obtuvieron ganancia ese día. Ante estos resultados, una persona afirma que hay una mayor proporción de acciones que obtuvieron ganancia en la Bolsa A, con respecto a la otra bolsa, el lunes pasado. Utilice* $\alpha = 0.04$

1. *¿Hay evidencia que respalde la afirmación realizada por la persona ?* *R/* $z_c = 1.75$, $z_{obs} = 0.7753$, *Valor* $P = 0.21917$, *no hay evidencia*
2. *¿De qué tamaño debería haberse tomado las muestras si se desea una potencia de* 90% *cuando el* 30% *de las personas obtuvo ganancia en la Bolsa A y el* 25% *en B?* *R/* *Se deben tomar muestras mayores a* 1463.

3.2.2 Prueba de que dos proporciones se diferencian en $d_0 \neq 0$

Para este caso, no hay una regla definida. Una opción, poco justificada, es utilizar las estimaciones dadas por $\widehat{p}_i$ y $\widehat{q}_i$ para p_i y q_i respectivamente, para $i = 1, 2$. Así, si las muestras son independientes y grandes ($n_i\widehat{p}_i \geq 5$ y $n_i\widehat{q}_i \geq 5$) entonces

$$\widehat{P}_1 - \widehat{P}_2 \underset{aprox}{\sim} N\left(d_0, \frac{\widehat{p}_1\widehat{q}_1}{n_1} + \frac{\widehat{p}_2\widehat{q}_2}{n_2}\right).$$

3.3 Pruebas con dos variancias

Considere la siguiente notación

Población	Variancia Poblacional	Tamaño de muestra	Variancia muestral
X_1	σ_1^2	n_1	S_1^2
X_2	σ_2^2	n_2	S_2^2

Suponga que las poblaciones X_1 y X_2 siguen una distribución normal, y los estadísticos S_1^2 y S_2^2 son independientes. Recuerde que

$$F = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

sigue una distribución f con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad. En las pruebas de dos variancias , la hipótesis nula es de la forma

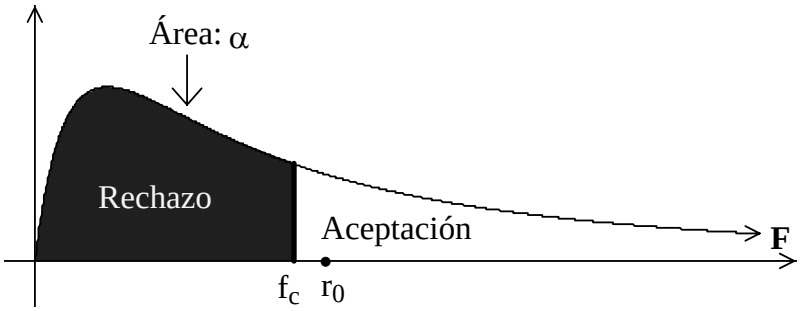
$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = r_0$$

donde r_0 es llamada la razón nula. Entonces, bajo H_0, se tiene que

$$F = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} = \frac{S_1^2}{r_0 S_2^2} \sim f\left(n_1 - 1, n_2 - 1\right)$$

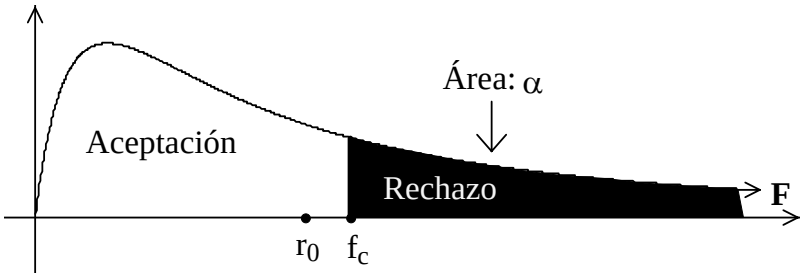
sigue una distribución F con $v_1 =$ y $v_2 =$ grados de libertad. A continuación se analizan los diferentes casos de H_1 :

1. Si $H_1 : \frac{\sigma_1^2}{\sigma_2^2} < r_0$ entonces



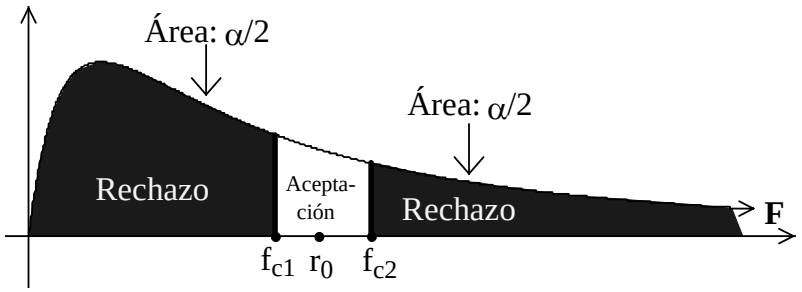
donde $f_c = f_{\alpha, n_1 - 1, n_2 - 1}$ y el Valor P es $P\left(F < f_{obs}\right)$ con $f_{obs} = \frac{s_1^2}{r_0 s_2^2}$.

2. Si $H_1 : \frac{\sigma_1^2}{\sigma_2^2} > r_0$ entonces



donde $f_c = f_{1-\alpha, n_1 - 1, n_2 - 1}$ y el Valor P es $P\left(F > f_{obs}\right)$.

3. Si $H_1 : \dfrac{\sigma_1^2}{\sigma_2^2} \neq r_0$ entonces



donde $f_{c1} = f_{\alpha/2,n_1-1,n_2-1}$, $f_{c2} = f_{1-\alpha/2,n_1-1,n_2-1}$. Dado que el Valor P sería la probabilidad de que F tome el valor f_{obs} o un valor extremo de forma que $F \neq r_0$, así si

(a) $f_{obs} > r_0$ entonces el valor P es aproximadamente[8] $2P\left(F > f_{obs}\right)$.

(b) $f_{obs} < r_0$ entonces el valor P es aproximadamente $2P\left(F < f_{obs}\right)$.

Ejemplo 74 *Se obtuvieron las estaturas de 21 mujeres y 31 hombres seleccionados aleatoriamente de la población de alumnos de cierta escuela. En la siguiente tabla se resumen los datos encontrados*

	Número	*Media*	*Desviación*
Mujeres (f)*:*	*21*	*63.8*	*2.18*
Hombres (m)*:*	*31*	*69.8*	*1.92*

Sean: σ_f^2 *la varianza de las estaturas de las mujeres y* σ_m^2 *la varianza de las estaturas de los hombres. Determine, mediante una prueba de hipótesis, si se debe suponer que* $\sigma_f^2 = \sigma_m^2$ *o no.*

$$H_0 : \sigma_f^2 = \sigma_m^2, \quad H_1 : \sigma_f^2 \neq \sigma_m^2$$

Note que $\sigma_f^2 = \sigma_m^2$ *es equivalente a* $\dfrac{\sigma_f^2}{\sigma_m^2} = 1$*, por lo tanto* $r_0 = 1$*. Así, el valor* f *observado es*

$$f_{obs} = \frac{s_1^2}{r_0 s_2^2} = \frac{2.18^2}{1.92^2} = 1.289\,17, \text{ además } (v_1, v_2) = (20, 30)$$

Como $f_{obs} > r_0 = 1$ *y el valor* P *es una probabilidad en dirección a* H_1 *entonces*

$$\text{Valor } P = 2 \cdot P\left(F > f_{obs}\right) = 2 \cdot \left(1 - P\left(F < f_{obs}\right)\right) = 2 \cdot \left(1 - P\left(F < 1.289\right)\right)$$

Utilizando la tabla, como $f_{0.9,20,30} = 1.667$ *entonces* $P\left(F < 1.289\right) \in \left]0.1, 0.9\right[$*, entonces*

$$\text{Valor } P = 2 \cdot \left(1 - P\left(F < 1.289\right)\right) > 2 \cdot (1 - 0.1) = 0.2.$$

Así, el $Valor\ P > 0.2 > 0.005$*, por lo tanto se acepta* H_0*. Se comcluye que no hay evidencia en contra de que* $\sigma_f^2 = \sigma_m^2$*. Por lo tanto se puede suponer que* $\sigma_f^2 = \sigma_m^2$*.*

[8] Note que F no es una distribución simétrica respecto a r_0, por lo tanto lo que se indica es el valor P si la distribución fue simétrica. Se recomienda que si este valor P es muy cercano al nivel de significancia, utilizar mejor el enfoque de regiones para constrastar la hipótesis.

Recuerde que algunas pruebas de hipótesis e intervalos de confianza sobre diferencia de promedios requieren apriori suponer si las variancias poblacionales son iguales o no. El ejemplo anterior, brinda un camino para determinar la suposición a tomar. Así, se recomienda en este tipo de investigaciones hacer un prueba de hipótesis para determinar si las variancias se suponen iguales o distintas.

Ejemplo 75 *Un vendedor de accesorios de computadoras afirma que los discos duro modelo A tienen una vida útil promedio que supera a la vida útil promedio de los modelo B en más de 1 año. Para investigar se obtuvo la siguiente información donde $\overline{x}$ es la duración promedio muestral (en años) y s la desviación estándar muestral (en años):*

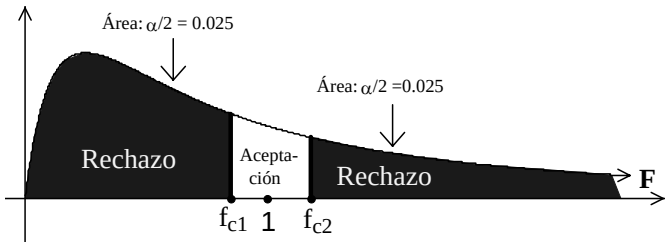
Disco Duro	*tamaño de muestra*	$\overline{x}$	s
Modelo A	15	6.5	2.3
Modelo B	17	5.4	1.6

1. *Pruebe la hipótesis de que las variancias de las duraciones de ambos modelos son iguales a un nivel de significancia del* 5%.

   Note que $\alpha = 0.05$ *y* $r_0 = 1$.

   $H_0 : \sigma_A^2 = \sigma_B^2$

   $H_1 : \sigma_A^2 \neq \sigma_B^2$

   El valor f observado es

   $$f_{obs} = \frac{2.3^2}{1.6^2} = 2.066\,41, \quad y \quad (v_1, v_2) = (14, 16)$$

   Los valores críticos son

   $$\begin{aligned} f_{c1} &= f_{0.025,14,16} = \frac{1}{2.923} = 0.342\,114 \\ f_{c2} &= f_{0.975,14,16} = 2.817 \end{aligned}$$

   Por lo tanto las regiones para F son

   $$\textit{Región Aceptación } A =]0.342\,114, 2.817[\,, \textit{ Región Rechazo } R = \mathbb{R} - A.$$

   Como $f_{obs} = 2.066\,41 \in A$, *entonces se acepta* H_0 *y se puede suponer que* $\sigma_A^2 = \sigma_B^2$.

2. *¿Cuál es el valor P en la prueba anterior?*

   Como $f_{obs} > 1$ *y el valor P es una probabilidad en dirección a* H_1 *entonces*

   $$\begin{aligned} P &= 2 \cdot P\left(F > f_{obs}\right) = 2 \cdot \left(1 - P\left(F \leq f_{obs}\right)\right) \\ &= 2 \cdot \left(1 - \underbrace{P\left(F \leq 2.066\,41\right)}_{\in]0.9, 0.95[}\right) > 2 \cdot (1 - 0.9) = 0.2 \end{aligned}$$

Note que el valor $P > \alpha = 0.05$, *lo cual era de esperar, pues se aceptó* H_0.

3. *De acuerdo con el punto (1.), con un nivel de significancia de 5% ¿es aceptable la afirmación del vendedor?*

Sean

μ_A : *vida útil promedio de los discos duro modelo A*
μ_B : *vida útil promedio de los discos duro modelo B*

El vendedor afirma que $\mu_A - \mu_B > 1$, *entonces*

$$H_0 : \mu_A - \mu_B = 1\,(\leq) \quad H_1 : \mu_A - \mu_B > 1$$

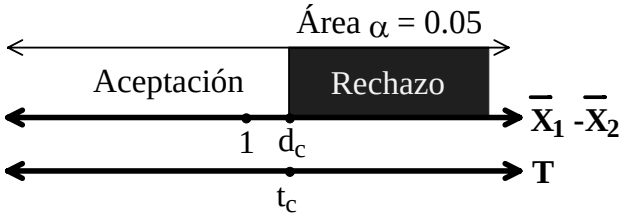
Note que $n_1 = 15$, $n_2 = 17$, $d_{obs} = \overline{x}_A - \overline{x}_B = 6.5 - 5.4 = 1.1$. *Como se supone que* $\sigma_A^2 = \sigma_B^2$, *por la prueba anterior, entonces*

$$T = \frac{\overline{X}_1 - \overline{X}_2 - d_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \sim t\,(v)$$

donde $v = n_1 + n_2 - 2 = 30$, *además*

$$s_p^2 = \frac{(n_1 - 1)\,s_1^2 + (n_2 - 1)\,s_2^2}{n_1 + n_2 - 2} = \frac{14\,(2.3)^2 + 16\,(1.6)^2}{15 + 17 - 2} = 3.834.$$

Las regiones de la prueba para T *tienen la forma*



donde el valor crítico es

$$t_c = t_{0.95,30} = 1.69726$$

Por lo tanto, las regiones para T *son*

$$\textit{Región Aceptación } A =]-\infty, 1.69726[\,, \textit{ Región Rechazo } R =]1.69726, +\infty[\,.$$

El valor de T *observado es*

$$t_{obs} = \frac{\overline{x}_D - \overline{x}_S - d_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{1.1 - 1}{\sqrt{\frac{3.834}{15} + \frac{3.834}{17}}}$$

Como $t_{obs} \in A$ *se acepta* H_0, *hay evidencia significativa en contra de la afirmación del vendedor.*

Ejercicio 38 *Se diseñó un estudio para comparar la influencia que tiene el uso de la computadora en un curso de estadística elemental para estudiantes de colegio, para ello se tomaron dos grupos. El Grupo 1 tomó el curso asistido por computadora a través de la utilización de paquetes estadísticos, mientras que el Grupo 2 recibió el curso sin utilización de la computación. Al final del curso se aplicó un test a ambos grupos, los resultados fueron:*

	n	$\overline{x}$	s
Grupo 1 (con computación)	*11*	*60.3*	*3.8*
Grupo 2 (sin computación):	*15*	*67.2*	*2.1*

1. *Pruebe la hipótesis de que las variancias son iguales al nivel de significancia de* 0.05. *R/ No se puede suponer que son iguales, región de aceptación* $A =]0.281, 3.147[$.

2. *De acuerdo con el punto* (1.)*: ¿Indican los datos que el resultado promedio de quienes tuvieron experiencia computacional fue significativamente menor que el de aquellos sin tal experienciaR/ Valor* $P \in]0.005, 0.01[$*, se rechaza* H_0.

3.4 Ejercicios

1. Una muestra de países desarrollados y subdesarrollados del mundo dio los siguientes resultados, donde n es el número de países, $\overline{x}$ el promedio de las esperanzas de vida y s la desviación estándar de las esperanzas de vida (en años):

Países	n	$\overline{x}$	s
Desarrollados	25	78.16	1.85
Subdesarrollados	21	58.7	1.08

   Suponga que las esperanzas de vida se distribuyen normalmente y que las varianzas de estás son diferentes. Con un nivel de significancia de 5% , pruebe la hipótesis de que la esperanza de vida en los países desarrollados es superior a la de los países subdesarrollados en a lo sumo 18 años. $R/$ Valor $P < 0.005$, no hay evidencia de que la esperanza de vida es superior en a lo sumo 18 años.

2. La oficina del Consumidor desea determinar si hay diferencia entre el peso promedio de dos tipos de bolsas de arroz A y B, que se venden con peso nominal de dos kilos. Para ello, realizó una inspección, obteniendo los siguientes resultados:

Tipo	Tamaño de la muestra	Media muestral	Desviación muestral
A	49	$2.05kg$	$0.2kg$
B	21	$1.97kg$	$0.12kg$

   (a) Pruebe la hipótesis de que las variancias son iguales al nivel de significancia de $0.05.R/$ Se rechaza H_0, se supone que $\sigma_A^2 \neq \sigma_B^2$.

   (b) De acuerdo con el punto a.) ¿Indican los datos que el peso promedio de la bolsa de arroz A es mayor al peso promedio de la bolsa de arroz $B?R/$ Valor $P < 0.025$, Se rechaza H_0, los datos indican que $\mu_A > \mu_B$

(c) Suponiendo que las bolsas de marca A y B tienen el mismo valor y la misma calidad, ¿Qué marca de arroz, A ó B, debe comprar los consumidores? $R/$ La marca A

3. Una persona afirma que es mayor la probabilidad de que un estudiante se duerma en una clase que recibe a las $7am$ que en una clase que recibe por la tarde a las $1pm$. La Federación de Estudiantes decide contrastar la aseveración para lo cuál reúne dos grandes muestras.

Hora del problema	# muestrado	# que se durmieron
$7am$	500	30
$1pm$	600	28

Realícese el contraste de hipótesis al nivel de significación 0.02. .
$.R/$ $d_c = 0.02779\,32$. No hay eviedencia que indique que hay una mayor probababilidad de que un estudiante se duerma en la mañana.

4. Una muestra de países desarrollados y subdesarrollados del mundo dio los siguientes resultados, donde n es el número de países, $\overline{x}$ el promedio de las esperanzas de vida y s la desviación estándar de las esperanzas de vida (en años):

Países	n	$\overline{x}$	s
Desarrollados	25	78.16	1.85
Subdesarrollados	21	58.7	1.08

Suponga que las esperanzas de vida se distribuyen normalmente. Pruebe la hipótesis de que las variancias son iguales al nivel de significancia de 0.05. $R/$ Se rechaza H_0, se supone que $\sigma_A^2 \neq \sigma_B^2$.

5. Un profesor de la Universidad Bienestar Seguro afirma que, en los últimos 5 años, un estudiante que repite el curso de probabilidad tiene menor posibilidad de aprobarlo que alguien que lo lleva por primera vez. Revisando algunos expedientes para estos años se halló la siguiente información

Estudiantes	Aprobaron	Reprobaron
Repitentes	15	60
1° vez en el curso	30	70

¿Hay evidencia que respalde la afirmación realizada por el profesor ?
$R/$ Valor $P = 0.066807$. Hay una leve evidencia en contra de la afirmación del profesor.

6. Una universidad analizó las capacidades de rendimiento de 2 tipos de variedades de cierto producto agrícola, así obtuvo los siguientes datos utilizando parcelas de igual tamaño:

Variedad	Tamaño de muestra	Rendimiento promedio (Kg por parcela)	Desviación muestral
A	32	43	15
B	30	47	22

El ministro de agricultura afirma que la variedad 2 tiene mejor rendimiento. A un nivel de significancia del 5%, ¿Considera aceptable la afirmación del ministro? $R/$ No

7. Una ferretería tiene dos marcas de pintura color verde para portones: A y B. El vendedor asegura que la pintura B es más cara por que seca más rápidamente que la pintura A. Se obtienen mediciones de ambos tipos de pintura. Los tiempos de secado(en minutos) son los siguientes:

Pintura A:	120	132	123	122	140	110	120	107		
Pintura B:	126	124	116	125	109	130	125	117	129	120

Suponga que el tiempo de secado de ambas pinturas está distribuido de manera normal

(a) Pruebe la hipótesis de que las variancias son iguales. $R/$

(b) ¿Existe evidencia en contra de la afirmación del vendedor? $R/$

8. Un producto electrónico es fabricado por dos empresas A y B. Se compararon las proporciones de productos con defectos de fabricación para ambas empresas, reuniéndose los datos siguientes

Empresa A:	7 productos defectuosos de 100
Empresa B:	6 productos defectuosos de 80

En la tienda COMPUBUENA vende ambos productos y uno de los vendedores asegura que el producto de A tiene menos defectos de fabricación que el producto de B. Sin embargo, los compradores suelen dudar sobre la información dada por el vendedor, ya que el producto de A tiene un precio mucho mayor con respecto al precio del otro producto.

(a) ¿Los datos contradice la afirmación que realiza el vendedor?

(b) ¿De qué tamaño debe ser la muestra si se desea realizar la prueba sobre la afirmación del vendedor con un nivel de significancia del 5% y una potencia del 90% cuando en realidad el porcentaje de productos defectuosos es del 10% para ambas empresas.

9. Un curso es impartido tradicionalmente por 2 profesores A y B. Se tiene que 15 de los estudiantes del profesor A tienen una nota final promedio de 67.1 con una desviación estándar de 15, y 21 estudiantes del profesor B tienen un promedio de 62.8 con una desviación estándar de 18. Suponga que ambas notas siguen una distribución normal.

(a) Pruebe la hipótesis de que las variancias de las notas obtenidas por ambos profesores son iguales, determinando las regiones de aceptación y rechazo con $\alpha = 0.05$.

(b) ¿Los datos contradice el hecho de que el curso de ambos profesores tiene rendimiento promedio muy similar? $R/\ No$

10. Detergentes BIEN LIMPIO investiga la preferencia de su marca frente a otras. Se encuesta a 200 personas en el San José y se encuentra que 50 de ellas están a favor de su marca, y en Cartago 25 personas de un total de 150 la prefieren. El gerente de BIEN LIMPIO desea verificar si los capitalinos prefieren más la marca que los cartaguineses.

(a) A un nivel de significancia del 4%, determine las regiones de aceptación y rechazo para la afirmación del gerente.

(b) Determine el valor P de la prueba.

(c) ¿los datos apoyan lo expresado por el gerente?

11. En una comunidad existen dos colegios, el alcalde de la comunidad afirma que los estudiantes del primer colegio de sétimo año tiene estaturas más similares que los estudiantes del otro colegio. Las estaturas de 21 estudiantes del primer colegio muestran una desviación estándar de $8cm$. Además en 19 estaturas de estudiantes del segundo colegio, se observa una desviación estándar de $5cm$. ¿Los datos apoyan la afirmación que realiza el alcalde?

12. Las estaturas de 6 niños de 1° y 3° grado son:

Grado	Estaturas
1°	$121, 115, 118, 122, 119, 118$
3°	$135, 132, 130, 136, 135, 133$

Suponga que ambas estaturas se distribuyen normalmente.¿Hay evidencia de que los estudiantes de 3° grado tienen estaturas más similares que los estudiantes de primer grado?

13. Sea desea investigar la duración de dos tipos A y B de baterías AA no recargables, las cuales tiene precios similares. Un estudiante, que utiliza mucho baterías AA, señala que la duración promedio de las baterías A supera en más de 10 minutos a la duración promedio de las baterías B. Se tomaron muestras de duraciones de ambos tipos de baterías, la información se resume en la siguiente tabla

Batería AA	tamaño de muestra	$\overline{x}$	s
tipo A	21	5.3 horas	0.8 horas
tipo B	19	5.1 horas	0.6 horas

Suponga que las variancias son iguales. Con un nivel de significancia de 10% ¿es aceptable la afirmación del estudiante? $R/$ $t_{obs} = 0.147\,800$, $t_c = 1.30423$, valor $P > 0.4$, hay evidencia en contra de la afirmación.

14. Las carreras de Electrónica y Computación tiene gran demanda laboral, esto hace que muchos estudiantes ingresen al mundo laboral antes de graduarse. Sin embargo, el profesor afirma que, para estudiantes de último de carrera, es mayor la proporción de estudiantes de Computación que laboran que la de estudiantes de Electrónica. En una pequeña encuesta se obtuvo la siguiente información

Estudiantes de último año	# de encuestados	Laboran
Computación	20	12
Electrónica	30	14

(a) Determine el Valor P para contrastar la afirmación del profesor. $R/$ Valor $P \approx 0.178786$

(b) ¿Hay evidencia que respalde la afirmación realizada por el profesor ? $R/$ Hay evidencia en contra de la afirmación

15. La universidad Bienestar Seguro tiene 2 fórmulas para examen de admisión que pretende utilizar durante los próximos 3 años. Sin embargo, un profesor de estadística de la universidad afirma

que la fórmula 1 va a tener un mejor rendimiento promedio que la fórmula 2. Ante esto, la universidad aplicó las fórmulas a un grupo de estudiantes, obteniendo los siguientes resultados:

Fórmula	Tamaño de la muestra	Media muestral	Desviación muestral
1	21	65	24
2	17	63	15

(a) ¿Puede suponerse que las variancias son iguales? Explique. $R/$ No

(b) A un nivel de significancia del 6%, ¿Existe evidencia a favor de la afirmación del profesor? (Determine las regiones). $R/$ No

16. Los siguientes datos representan los tiempos de duración de las películas, en minutos, que producen dos compañías cinematográficas

Compañía	Tiempo (*minutos*)
A	$103, 94, 110, 87, 98$
B	$97, 82, 123, 92, 88, 118, 175$

Suponga que las duraciones se distribuyen normalmente.

(a) Se desea saber si las variancias se consideran distintas. Pruebe esta hipótesis con un nivel de significancia del 5%. $R/$ $f_{c1} = 0.16$, $f_{c2} = 6.227$, $f_{obs} = 0.00736$, se consideran las variancias distintas.

(b) De acuerdo con el punto a.) ¿Indican los datos que la duración promedio de las películas de la compañía A es inferior en más de 10 minutos a la duración promedio de las películas de B ? $R/$ $t_{obs} = -0.181\,160\,, v \approx 7$,Valor $P \in]0.4, 1[$. Si hay evidencia

4 Otras pruebas de hipótesis

4.1 Bondad de ajuste

La Bondad de Ajuste es una prueba de hipótesis que determina si existe evidencia significativa en contra de que una población se distribuye de cierto modo, utilizando la información dada por una muestra.

Ejemplo 76 *Se lanza una moneda hasta obtener 2 escudos, sea X el número de lanzamientos realizados. Al repetir el experimento 128 veces se obtienen los siguientes resultados:*

X	2	3	4	5	6	7	8	9	10
Frecuencia Observada	30	34	25	17	12	5	2	2	1

¿Se puede suponer la distribución de X es $f_x(k) = \dfrac{k-1}{2^k}$?
Note que la probabilidad de que el número de lanzamientos realizados sea 2, según f_x, es

$$P(X=2) = f_x(2) = \frac{1}{4}.$$

Por lo tanto, si $X \sim f_x$ se esperaría que de 128 valores aleatorios de X, en un cuarto de ellos se obtenga un 2. Así, la frecuencia esperada de obtener 2 es

$$e_1 = 128 \cdot \frac{1}{4} = 32,$$

la cuál se debe comparar con la frecuencia observada en la muestra que es $o_1 = 30$. Similarmente se obtienen las demás frecuenciá esperadas según f_x :

i	1	2	3	4	5	6	7	8	9
x_i	2	3	4	5	6	7	8	9	10
o_i	30	34	25	17	12	5	2	2	1
e_i	32	32	24	16	10	6	3.5	2	1.125

Parece lógico que si las frecuencias esperadas son muy similares a las observadas, entonces se considera aceptable que $X \sim f_x$. Pero, ¿cómo medir dicha similitud?

Considere la población dada por la variable X, y los parámetros desconocidos $p_1, p_2, ..., p_k$ que indican las probabilidades de que X tome k valores (si X es una *v.a.d.*) o de que X pertenezca a k clases de valores de X (si X es *v.a.c.*). Es decir, p_i es la probabilidad de que X tome el valor *iésimo* (si X es discreto) ó que X se encuentre en la clase *iésima* (si X es continua). Suponga que para tales valores o clases se conocen las frecuencias esperadas $e_1, e_2, ..., e_k$ según la función de distribución de probabilidad f_x, para una muestra de tamaño n. Una prueba de bondad de ajuste tiene como hipótesis nula:

$$H_0: \ X \text{ sigue la distribución } f_x : p_1 = \frac{e_1}{n}, p_2 = \frac{e_2}{n}, ..., p_k = \frac{e_k}{n}.$$

Bajo la hipótesis nula se tiene que $e_i = np_i$ y si $e_i \geq 5$ para $i = 1, 2, 3, ..., k$, el estadístico

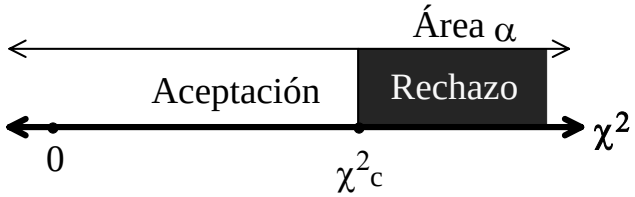
$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i}$$

tiene una distribución χ^2 con $v = k - 1$ grados de libertad, donde $O_1, O_2, ..., O_k$ son estimadores de $np_1, ..., np_k$. Estos estimadores indican las frecuencias observadas en cada muestra de tamaño n.

Así, al tomar una muestra de tamaño n se obtienen las frecuencias observadas: $o_1, o_2, ..., o_k$, estimaciones de np_i dada por el estimador O_i, para $i = 1, 2, ..., k$. Para esta muestra si el valor

$$\chi^2_{obs} = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

es pequeño, entonces o_i es similar a e_i para cada $i = 1, 2, ..., k$ y por lo tanto habría un buen ajuste, es decir se aceptaría H_0. Así, nuestras regiones de aceptación y rechazo para una prueba de Bondad de Ajuste son:



Donde el valor crítico $\chi^2_c = \chi^2_{1-\alpha,k-1}$ Además el valor P es $P\left(\chi^2 > \chi^2_{obs}\right)$

Ejemplo 77 *Una encuesta realizada a 150 familias de una zona rural revélo los siguientes datos sobre el número de televisores por familia*

# de televisores	*0*	*1*	*2*	*3*	*4*
# de familias	*23*	*35*	*32*	*27*	*33*

Pruebe la hipótesis de que la distribución del número de televisores por familia, de la zona rural, es uniforme de 0 *a* 4.

Sea X : *el número de televisores que posee una familia de la zona rural considerada. Se tiene que*

$$\begin{aligned} H_0 &: X \text{ sigue la distribución uniforme de } 0 \text{ a } 4 \\ H_1 &: X \text{ no sigue la distribución uniforme de } 0 \text{ a } 4 \end{aligned}$$

El tamaño de muestra es $n = 150$ *y* $k = 5$. *Como la variable* X *es discreta entonces la función de distribución de* X *es*

$$f_X(x) = \frac{1}{5}, \text{ con } x = 0, 1, 2, 3, 4.$$

Por lo tanto la frecuencia esperada es invariante según el número de televisores pues $e_i = np_i = 150 \cdot \frac{1}{5} = 30$. *Así,*

X :	*0*	*1*	*2*	*3*	*4*
frec. observada (o_i) :	*23*	*35*	*32*	*27*	*33*
frec. esperada (e_i) :	*30*	*30*	*30*	*30*	*30*

Note que la suma de los e_i es igual a $n = 150$. Como se cumple que $e_i = 30 \geq 5$ para $i = 1, 2, .., 5$ entonces, bajo H_0, se cumple que

$$\chi^2 = \sum_{i=1}^{5} \frac{(O_i - e_i)^2}{e_i}$$

tiene una distribución χ^2 con $n - 1 = 4$ grados de libertad. El valor χ^2 observado es

$$\chi^2_{obs} = \sum_{i=1}^{5} \frac{(o_i - e_i)^2}{e_i} = \frac{(23-30)^2 + (35-30)^2 + (32-30)^2 + (27-30)^2 + (33-30)^2}{30} = 3.2$$

El Valor P es

$$\text{Valor } P = P\left(\chi^2 > \chi^2_{obs}\right) = 1 - \underbrace{P\left(\chi^2 \leq 3.2\right)}_{\in]0.2,0.8[} \in]0.2, 0.8[$$

Note que $P\left(\chi^2 \leq 3.2\right) \in]0.2, 0.8[$ pues $\chi^2_{0.2,4} = 1.64878$ y $\chi^2_{0.8,4} = 5.98862$. Como el Valor $P > 0.05$ entonces se acepta H_0. Por lo tanto, no hay evidencia en contra de que X siga una distribución uniforme.

Ejemplo 78 *Se lanza una moneda hasta obtener 2 escudos, sea X el número de lanzamientos realizados. Al repetir el experimento 128 veces se obtienen los siguientes resultados:*

X	2	3	4	5	6	7	8	9	10
Frecuencia Observada	30	34	25	17	12	5	2	2	1

Pruebe la hipótesis, al nivel de significancia del 5%, de que la distribución de X es $P(X = k) = \dfrac{k-1}{2^k}$.

Se tiene que

$$H_0 \quad : \quad X \text{ sigue la distribución } P(X = k) = \frac{k-1}{2^k}$$
$$H_1 \quad : \quad X \text{ no sigue la distribución } P(X = k) = \frac{k-1}{2^k}$$

Note que $n = 128$, $e_1 = np_1 = n \cdot P(X = 2) = 128 \cdot \dfrac{1}{2^2} = 32$, similarmente se obtiene las otras frecuencias esperadas. Así

i	1	2	3	4	5	6	7	8	9
x_i	2	3	4	5	6	7	8	9	10
o_i	30	34	25	17	12	5	2	2	1
e_i	32	32	24	16	10	6	3.5	2	1.125

Sin embargo los e_i no suman 128, esto se debe a que existe una frecuencia esperada de que el número de lanzamientos sea mayor a 10 :

i	1	2	3	4	5	6	7	8	9	10
x_i	2	3	4	5	6	7	8	9	10	11 ó más
o_i	30	34	25	17	12	5	2	2	1	0
e_i	32	32	24	16	10	6	3.5	2	1.125	1.375

Como los e_i debe ser mayores iguales a 5, se fusionan las clases del 8 en adelante:

i	1	2	3	4	5	6	7
x_i	2	3	4	5	6	7	8 *o más*
o_i	30	34	25	17	12	5	5
e_i	32	32	24	16	10	6	8
$\frac{(e_i - o_i)^2}{e_i}$	0.125	0.125	0.04167	0.0625	0.4	0.1667	1. 125

Por lo tanto

$$\chi^2_{obs} = 2.\,045\,87, \quad \chi^2_c = \chi^2_{1-\alpha,k-1} = \chi^2_{0.95,6} = 12.5916$$

Como $\chi^2_{obs} < \chi^2_c$ se acepta que la distribución de X es la dada al no existir evidencia significativa en su contra.

Ejercicio 39 *Varios clientes se han quejado de que el CASINO DINERO FACIL usa dados cargados. Para investigar la acusación se lanza un par de sus dados 45 veces, obteniendo los siguientes datos sobre la suma de los puntos en cada tirada:*

Suma:	2	3	4	5	6	7	8	9	10	11	12
Frecuencia	0	1	3	7	10	14	2	2	3	2	1

Pruebe la hipótesis de que los dados están cargados. *R/* $\chi^2_{obs} = 15.308$, *Valor* $P < 0.025$. *Hay evidencia de que los dados están cargados.*

Para el caso en que la variable X es continua, se recomienda hacer una prueba visual antes de una prueba formal, ya que la primera podría descartar a priori la segunda prueba.

1. Prueba visual. Se puede realizar un histograma de frecuencia relativas para los datos muestrales para intuir la forma de la distribución. Además, con ayuda de software, se puede superponer la distribución intuida sobre el histograma para realizar un análisis visual más minucioso. Esta pruebe nos permite descartar el ajuste o someterlo a una prueba formal.

2. Prueba formal. Para realizar esta prueba se deben agrupar los datos si no lo están.

   (a) Datos individuales. Si los datos no están agrupados se recomienda realizar de 5 a 10 clases con frecuencia esperadas $e_i \geq 5$ y tratar de que estas frecuencias sean lo más uniformes posible.

   (b) Datos agrupados. Si los datos ya están agrupados en clase y solo se conoce la frecuencia observada en cada clase, se deben determinar las frecuencias esperadas de cada clase: e_i =(área de clase)·(tamaño de la muestra) $= p_i \cdot n$. En caso de que algún e_i sea menor a 5, se deben fusionar clases por proximidad hasta garantizar que los $e_i \geq 5$.

Ejemplo 79 *El curso Estruturas de Datos II es impartido en el segundo año de la carrera de computación de la Universidad Futuro Seguro. Un profesor considera que la edad de los estudiantes que*

cursan el curso sigue una distribución normal. En una muestra de 200 estudiantes se obtuvieron las siguientes frecuencias

Edad (X)	*Frecuencia observada*
$x \leq 18$	30
$18 \leq x < 19$	100
$19 \leq x < 20$	50
$20 \leq x < 21$	15
$x \geq 21$	5

Los datos observados tienen $\overline{x} = 18.825$ *y desviación estándar* $s = 0.907464855$ *¿Considera correcta la información del profesor?*

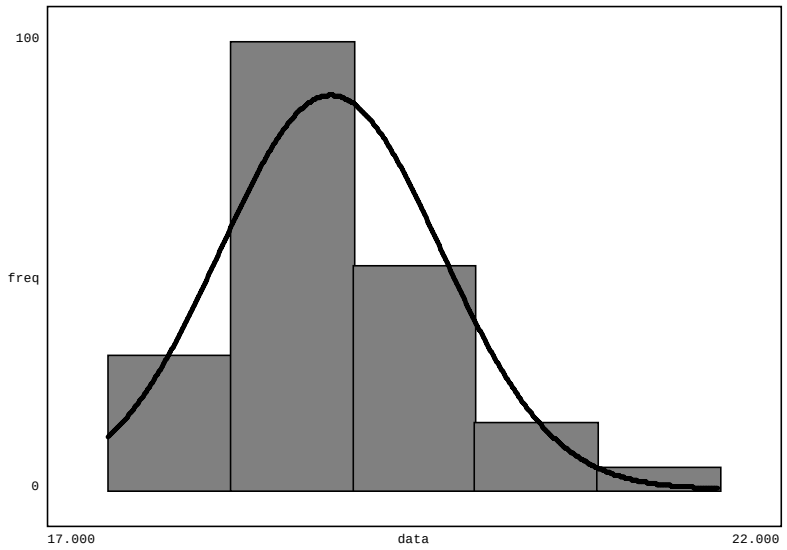
Prueba Visual. *Se utilizará el programa Winstats* [9]*. Suponda que se cuenta con una tabla en Microsoft Excel que tiene las 200 edades*[10]*. Guarde este archivo como Texto (MS-DOS). Abra el programa Winstats y en la pestaña* ***Window*** *eliga la opción* ***1-Var data****. En la nueva ventana eliga la opción* ***Text input*** *de la pestaña* ***File*** *y cargue el archivo de texto con los datos. Luego, en la pestaña* ***Stats*** *seleccione el menú* ***Histogram****, en este menú:*

1. *Eliga la opción* ***Interval****, para cambiar el intervalo del histigrama. Debe eligir un intervalo de manera que a la hora de dividirlo en 5 clases, éstas coincidan con las dadas en el ejercicio. En este caso se elige el intervalo* $[17, 22]$ *asumiendo que todos los 200 datos se encuentran en ese intervalo. En práctica Winstats define las clases a partir de los datos mínimo y máximo, y de la cantidad de clases.*
2. *Eliga la opción* ***Number of groups****, para indicar el número de clases. Debe utilizar el deslizador para cambiar el número de clases a 5 y no digitarlo en el espacio indicado.*
3. *Finalmente, eliga la opción Histogram para ver el histograma en una nueva ventana. En esta ventana, eliga la opción* ***Normal overlay*** *de la pestaña* ***Edit*** *para sobreponer la "curva*

[9] Este software es gratuito, no requiere instalación y lo puede conseguir en la dirrección http://math.exeter.edu /rparris /winstats.html

[10] El lector puede simular estas edades y realizar su tabla: escribe 20 números menores a 18, luego 80 números menores a 19 y mayores o iguales a 18, etc.

Normal"[11] *al histograma:*



La normal que sobrepone Winstats es la que mejor abarca el área del histograma, esta Normal tiene media $\overline{x}$ *y varianza* s^2*. Otros pogramas como Fathom permite ajustar manualmente la normal. Al observar el gráfico, se nota que la normal se ajusta levemente al gráfico, por lo que se requiere hacer una prueba formal.*

Prueba Formal. *Sea* X *la edad de los estudiantes que cursan Estructuras de Datos II. Recuerde que* $\overline{x} = 18.825$ *y* $s = 0.907464855$*, entonces* $N\left(\overline{x}, s^2\right) = N\left(18.825, 0.823\,492\right)$*. Se tiene que*

$$\begin{aligned} H_0 &: X \sim N\left(18.825, 0.823\,492\right) \\ H_1 &: X \nsim N\left(18.825, 0.823\,492\right) \end{aligned}$$

El tamaño de muestra es $n = 200$ *y* $k = 5$*. Para determinar las frecuencias esperadas, bajo* H_0*, note que*

$$\begin{aligned} P\left(X \leq 18\right) &= P\left(Z \leq \frac{18 - 18.82}{0.907464855}\right) = P\left(Z \leq -0.903\,616\right) \approx 0.114060 \\ P\left(x < 19\right) &= P\left(Z < \frac{19 - 18.82}{0.907464855}\right) = P\left(Z < 0.198\,355\right) \approx 1 - 0.420740 = 0.579\,26 \\ P\left(x < 20\right) &= P\left(Z < \frac{20 - 18.82}{0.907464855}\right) = P\left(Z < 1.\,300\,33\right) \approx 1 - 0.096801 = 0.903\,199 \\ P\left(x < 21\right) &= P\left(Z < \frac{21 - 18.82}{0.907464855}\right) = P\left(Z < 2.\,402\,30\right) \approx 1 - 0.008198 = 0.991\,802 \\ P\left(x \geq 21\right) &\approx 1 - 0.991\,802 = 0.008198 \end{aligned}$$

[11] Si el histograma es de valores absolutos, lo que se sobrepone al histograma es la curva amplificada.

Por lo tanto

$$\begin{aligned} e_1 &= n \cdot p_1 = 200 \cdot P(X \leq 18) \approx 200 \cdot 0.114060 = 22.812 \\ e_2 &= n \cdot p_2 = 200 \cdot P(18 \leq x < 19) \approx 200\,(0.57926 - 0.114060) = 93.04 \\ e_3 &= n \cdot p_3 = 200 \cdot P(19 \leq x < 20) \approx 200\,(0.903\,199 - 0.579\,26) = 64.787\,8 \\ e_4 &= n \cdot p_4 = 200 \cdot P(20 \leq x < 21) \approx 200\,(0.991\,802 - 0.903\,199) = 17.720\,6 \\ e_5 &= n \cdot p_5 = 200 \cdot P(x \geq 21) \approx 200 \cdot 0.008198 = 1.639\,6 \end{aligned}$$

Como los e_i debe ser mayores iguales a 5, se fusionan la clase 5 se fusiona a la clase 4 :

i	1	2	3	4
Clase	$x \leq 18$	$18 \leq x < 19$	$19 \leq x < 20$	$x \geq 20$
o_i	30	100	50	20
e_i	22.812	93.04	64.787 8	19.360 2
$\frac{(e_i - o_i)^2}{e_i}$	2.264 92	0.520 653	3.375 31	0.021 143 6

Por lo tanto $\chi^2_{obs} = 6.182\,03$, $v = k - 1 = 3$ y el Valor P es

$$\text{Valor } P = P\left(\chi^2 > \chi^2_{obs}\right) = 1 - \underbrace{P\left(\chi^2 \leq 6.182\,03\right)}_{\in]0.8,0.9[} \in]0.1, 0.2[$$

Note que $P\left(\chi^2 \leq 6.182\,03\right) \in]0.8, 0.9[$ pues $\chi^2_{0.8,3} = 4.64163$ y $\chi^2_{0.9,4} = 6.25139$. Como el Valor $P > 0.05$ entonces se acepta H_0. Por lo tanto, no hay evidencia en contra de que X siga una distribución normal.

Ejercicio 40 *Un centro de Salud considera que el peso de la población adulta que padecen cierta enfermedad sigue una distribución normal con media y desviación estándar de 140 y 10 libras, respectivamente. Se revisan los registros y se obtiene los datos de 300 adultos que padecen la enfermedad. Los datos se agruparon en las clases siguiente:*

Pesos (x)	*Frecuencia observada*
$x \leq 120$	5
$120 \leq x < 130$	20
$130 \leq x < 140$	120
$140 \leq x < 150$	105
$150 \leq x < 160$	35
160 *y mayor*	15

¿Muestran los datos observados evidencia significativa para rechazar la hipótesis de que los pesos están distribuidos normalmente?

4.2 Independencia

Suponga que se tienen dos variables cualitativas X, Y con atributos $x_1, x_2, ..., x_m$ y $y_1, y_2, ..., y_p$. Se toma una muestra de tamaño n y se realiza la **tabla de contingencia**, la cual indica la frecuencia observada o_{ij} de los individuos de la muestra que tienen los atributos x_i y y_j :

	y_1	y_2	$\cdots$	yp	Total
x_1	o_{11}	o_{12}	$\cdots$	o_{1p}	TX_1
x_2	o_{21}	o_{22}	$\cdots$	o_{2p}	TX_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_m	o_{11}	o_{11}	$\cdots$	o_{11}	TX_m
Total	TY_1	TY_2	$\cdots$	TY_p	n

donde TX_i es el total de individuos observados que poseen el atributo x_i, similarmente TY_i es el total de individuos observados que poseen el atributo y_i.

Ejemplo 80 *Los empleados de la empresa SSS trabajan en un horario de 8am a 12md y de 1pm a 5pm, su deseempeño en la mañana y en la tarde ha sido valorado por un inspector como Bueno, Regular o Malo. Así, se tienen dos variables cualitativas*

$$\begin{aligned} X: &\quad \textit{dessempeño en la mañana} \\ Y: &\quad \textit{dessempeño en la tarde} \end{aligned}$$

La variable X tiene atributos: bueno (x_1), regular (x_2) y malo (x_3); similarmente la variable Y. Suponga que se averigua el desempeño de 528 empleados y se resumen los datos en una tabla de contingencia:

		Tarde			
		Bueno	*Regular*	*Malo*	*Total*
Mañana	*Bueno*	56	71	12	139
	Regular	47	163	38	248
	Malo	14	42	85	141
Total		117	276	135	528

Esta tabla indica que de los 528 empleados, hay 56 que tuvieron un buen desempeño en ambos intervalos de trabajo, hay 38 que tuvuero un desempeño regular en la mañana y malo en la tarde, por ejemplo. También se observar que 276 empleados tuvieron un desempeño regular en la tarde.

Continuando con la teoría, note que TX_i es la frecuencia con que fue observada el atributo x_i en la muestra, por lo tanto la probabilidad de que X tome el valor de x_i se puede aproximar con la proporción observada

$$P(X = x_i) \approx \frac{TX_i}{n}, \text{ similarmente } P(Y = y_j) \approx \frac{TY_j}{n} \text{ y } P(X = x_i \text{ y } Y = y_j) \approx \frac{o_{ij}}{n}$$

Se quiere contrastar la prueba de hipótesis

$$\begin{aligned} H_0 &: X, Y \text{ son independientes.} \\ H_1 &: X, Y \text{ no son independientes.} \end{aligned}$$

Suponga que la probabilidad real de observar los atributos x_i y y_i en un grupo de n individuos es

$$P\left(X = x_i \text{ y } Y = y_j\right) = \frac{b_{ij}}{n}$$

donde b_{ij} es la frecuencia real de individuos que poseen los atributos x_i y y_j. Note que o_{ij} es una estimación de b_{ij} dada por el estadistico que se denotará O_{ij} que brinda la frecuencia observada de individuos que tienen los atributos x_i y y_j en muestras de tamaño n. Por otro lado, bajo H_0, se tiene que

$$P\left(X = x_i \text{ y } Y = y_j\right) = \frac{e_{ij}}{n}$$

donde e_{ij} es la frecuencia esperada de individuos que poseen los atributos x_i y y_j en un grupo de n individuos, asumiendo H_0. Por lo tanto, al asumir que X, Y son independiente (bajo H_0) se tiene que

$$\frac{e_{ij}}{n} = P\left(X = x_i \text{ y } Y = y_j\right) = P\left(X = x_i\right) P\left(Y = y_j\right) \approx \frac{TX_i}{n}\frac{TY_j}{n}$$

de donde se tiene que

$$e_{ij} = nP\left(X = x_i \text{ y } Y = y_j\right) \approx \frac{TX_i \cdot TY_j}{n}$$

Así, una prueba de independencia de las variables X, Y es una prueba de hipótesis donde la hipótesis nula es

$$H_0\left(X, Y \text{ son independientes}\right) : b_{ij} = e_{ij} \text{ que es equivalente a que } b_{ij} \approx \frac{TX_i \cdot TY_j}{n}$$

para $i = 1, 2, 3, ..., m$ y $j = 1, 2, ..., p$. Para realizar estas pruebas se usa el siguiente resultado: Bajo la hipótesis nula y con $e_{ij} \geq 5$, el estadístico

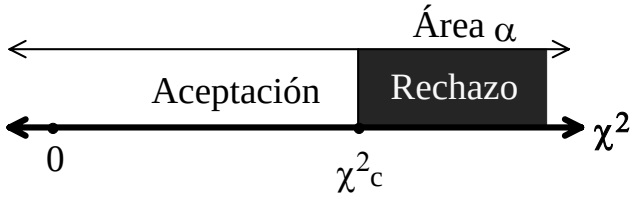
$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{p}\frac{\left(O_{ij} - e_{ij}\right)^2}{e_{ij}}$$

tiene una distribución χ^2 con $v = (m-1)(p-1)$ grados de libertad, donde, como se ha indicado, $O_{11}, ...; O_{mp}$ son estimadores de $b_{11}, ..., b_{mp}$.

Si en una muestra se obtienen las frecuencias observadas o_{ij} y el valor

$$\chi^2_{obs} = \sum_{i=1}^{m}\sum_{j=1}^{p}\frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}}$$

es muy pequeño, entonces o_{ij} es muy similar a e_{ij} y por lo tanto habría independencia. Así, nuestras regiones de aceptación y rechazo para una prueba de bonda de ajuste son:



Donde el valor crítico $\chi^2_c = \chi^2_{1-\alpha,(m-1)(p-1)}$ Además el valor P es $P\left(\chi^2 > \chi^2_{obs}\right)$.

Ejemplo 81 *Los empleados de la empresa SSS trabajan en un horario de 8am a 12md y de 1pm a 5pm, su deseempeño en la mañana y en la tarde ha sido valorado por un inspector como Bueno, Regular o Malo. Seguidamente se presenta la relación entre el desempeño de un grupo de trabajadores según el intervalo de trabajo:*

		Tarde			
		Bueno	*Regular*	*Malo*	*Total*
Mañana	*Bueno*	56	71	12	139
	Regular	47	163	38	248
	Malo	14	42	85	141
Total		117	276	135	528

Pruebe la hipótesis de que el desempeño en la tarde es independiente del desempeño en la mañana.

Considere las variables cualitativas

$$\begin{aligned} X: &\quad \textit{dessempeño en la mañaña} \\ Y: &\quad \textit{dessempeño en la tarde} \end{aligned}$$

Se tiene que

$$\begin{aligned} H_0 &: X, Y \textit{ son independientes.} \\ H_1 &: X, Y \textit{ no son independientes.} \end{aligned}$$

Para determinar las frecuencias esperadas note que

$$\begin{aligned} e_{11} &\approx \frac{TX_1 \cdot TY_2}{n} = \frac{139 \cdot 117}{528} \approx 30.80 \\ e_{12} &\approx \frac{TX_1 \cdot TY_2}{n} = \frac{139 \cdot 276}{528} \approx 72.66 \end{aligned}$$

similarmente se obtiene las otras frecuencias esperadas, obtiniendo la tabla de frecuencias esperadas:

		Tarde		
		Bueno	*Regular*	*Malo*
	Bueno	30.8	72.66	35.54
Mañana	*Regular*	54.955	129.64	63.41
	Malo	31.244	73.705	36.05

El lector puede verificar que la suma de estas frecuencias esperadas es muy cercana a $n = 528$. A partir de la tabla de frecuencias es peradas y la tabla de frecuencias observadas (tabla dada), se obtiene el valor observada de la distribución χ^2 :

$$\chi^2_{obs} = \frac{(56-30.8)^2}{30.8} + \frac{(47-54.955)^2}{54.955} + \frac{(14-31.244)^2}{31.244} + \frac{(71-72.66)^2}{72.66} + \frac{(163-129.64)^2}{129.64} + \frac{(42-73.705)^2}{73.705} + \frac{(12-35.54)^2}{35.54} + \frac{(38-63.41)^2}{63.41} + \frac{(85-36.05)^2}{36.05} = 145.79$$

La distribución χ^2 *sigue una distribución* $\chi^2(v)$ *donde* $v = (3-1)(3-1) = 4$ *entonces*

$$\text{Valor } P = P\left(\chi^2 > \chi^2_{obs}\right) = P\left(\chi^2 > 145.79\right) = 1 - \underbrace{P\left(\chi^2 < 145.79\right)}_{\in]0.99,1[} < 0.01$$

Como el valor $P < 0.05$ *entonces se rechaza* H_0*. Se concluye que existe evidencia de que el desempeño de la tarde dependa del desempeño en la mañana.*

Ejemplo 82 *Se desea estudiar la dependencia entre un determinado cáncer y los hábitos de fumado, para ello se tomaron los datos de 360 individuos:*

fumador	*No fumador*	*Fumador leve*	*muy fumador*
Con cáncer	*58*	*67*	*45*
Sin cáncer	*74*	*60*	*56*

A un nivel de significancia del 10%*, ¿existe evidencia de qué la presencia o ausencia del cáncer es dependiente de los hábitos de fumar?*

Sean

H_0 : *la presencia o ausencia del cáncer es independiente de los hábitos de fumar*

H_1 : *la presencia o ausencia del cáncer depende de los hábitos de fumar*

Se tiene que

Las frecuencias esperadas

	No	*moderado*	*Fumador*
Con cáncer	62.3333	59.9722	47.6944
Sin cáncer	69.6667	67.0278	53.3056

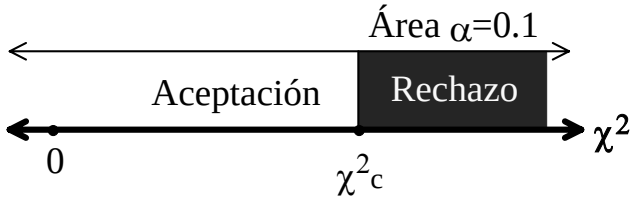
los valores $\dfrac{(O_{ij} - e_{ij})^2}{e_{ij}}$ *para chi*

	No	*moderado*	*Fumador*
Con cáncer	0.3012	0.8235	0.1522
Sin cáncer	0.2695	0.7369	0.1362

Note que todas las frecuencias esperadas son mayores o iguales a 5 *y* $v = 1 \cdot 2 = 2$*. Por lo tanto*

$$\chi^2_{obs} = 2.419597295 \quad y$$

Las regiones de aceptación y rechazo son de la forma



donde $\chi^2_c = \chi^2_{0.9,2} = 4.60517$*. Así las regiones son*

$$A = [0, 4.60517[\,, \quad R =]4.60517, +\infty[$$

pues la distribución χ^2 *no es negativa. Como* $\chi^2_{obs} \in A$ *entonces se acepta* H_0*. Hay evidencia de independencia.*

Ejercicio 41 *Un supervisor desea determinar si el número de artículos fabricados con defectos depende del día de la semana en que son producidos. Reunió la información siguiente.*

Día de la semana	*Lunes*	*Martes*	*Miércoles*	*Jueves*	*Viernes*
Sin defectos	*85*	*90*	*95*	*95*	*90*
Defectuosos	*15*	*10*	*5*	*5*	*10*

¿Existe evidencia suficiente de que el número de artículos defectuosos es independiente del día de la semana en que se fabrican con $\alpha = 0.05$*? R/* Si, $\chi^2_c = 9.4877$, $\chi^2_{obs} = 8.547$ *y Valor* $P \in]0.05, 0.1[$.

4.3 Análisis de variancia (ANOVA)

El ANOVA tiene como objetivo analizar la relación entre una variable cuantitativa X y una variable cualitativa Y de k atributos. Cada atributo i define una una población dada por la variable cuantitativa X_i : variable X restringuida al atributo i. Así, se tienens k poblaciones $X_1, X_2, ..., X_k$ (llamadas tratamientos) que se suponen normales, independientes, con variancias similares y con medias poblacionales $\mu_1, \mu_2, ..., \mu_n$. Se desea determinar si X no varía según el atributo de Y, es decir si las poblaciones son equivalentes y entonces los tratamientos son igualmente efectivos. Para ello, se plantea las hipótesis

H_0 (X no varía según el atributo de Y, poblaciones equivalentes) : $\mu_1 = \mu_2 = ... = \mu_k$
H_1 (poblaciones no equivalentes) : al menos dos de las medias no son iguales

Ejemplo 83 *Se quiere determinar si la dosis de un determinado medicamento (baja, media y alta) influye en el tiempo de sueño de los clientes que lo consumen. Se plantea el problema de determinar si los tiempos de sueño varían o no según la dosis del medicamento.*

Aquí la variable cuantitativa es

$$X : \textit{tiempo de sueño de los clientes que consumen el medicamento,}$$

y la variable cualitativa es

$$Y : \textit{ dosis del medicamento,}$$

la cual tiene atributos: dosis baja (y_1), *dosis media* (y_2) *y dosis alta* (y_3) *. Así, se tiene tres poblaciones o tratamientos:*

$$\begin{aligned} X_1 &: \textit{tiempo de sueño de los clientes que consumen la dosis baja} \\ X_2 &: \textit{tiempo de sueño de los clientes que consumen la dosis media} \\ X_3 &: \textit{tiempo de sueño de los clientes que consumen la dosis alta} \end{aligned}$$

y los valores μ_1, μ_2 *y* μ_3 *denotan los tiempos promedios de sueño para cada dosis respectivamente. Entonces el problema* "los tiempos de sueño varían o no según la dosis del medicamento" *se puede modelar por medio del contraste de hipótesis:*

$$\begin{aligned} H_0 \textit{ (tiempos no varían según la dosis)} &: \mu_1 = \mu_2 = \mu_3 \\ H_1 \textit{ (tiempos varían según la dosis)} &: \textit{ al menos dos de las medias no son iguales} \end{aligned}$$

Para contrastar la hipótesis se toma una muestra de tamaño $N = \sum_{i=1}^{k} n_i$ donde

1. n_i : es el número de observaciones en el tratamiento $i - ésimo$.
2. y_{ij} :es la $j - ésima$ observación del tratamiento $i - ésimo$.
3. $T_i = \sum_{j=1}^{n_i} y_{ij}$ es la suma de las observaciones en el tratamiento $i - ésimo$.
4. $\overline{y_i} = \dfrac{T_i}{n_i}$ es el promedio de las observaciones en el tratamiento $i - ésimo$.

5. $T = \sum_{i=1}^{k} T_i$ es la suma total de observaciones.

6. $\overline{y} = \dfrac{T}{N}$ es el promedio total observado.

Ejemplo 84 *Suponga que en el ejemplo anterior, varios voluntarios son expuestos a la aplicación de tres dosis (baja, media y alta) del medicamento, y se observa el número de minutos que duerme, los datos son los siguientes:*

Dosis		
Baja	*Media*	*Alta*
67	*96*	*74*
69	*98*	*24*
72	*130*	*15*
79	*65*	*33*
		17

Entonces

	Baja	*Media*	*Alta*
n_i	4	4	5
T_i	287	389	163
$\overline{y}_i$	$\dfrac{287}{4}$	$\dfrac{389}{4}$	$\dfrac{163}{5}$

La suma total de datos es $T = \sum_{i=1}^{3} T_i = 839$, *el numero de datos es* $N = 13$ *y el promedio de los datos es* $\overline{y} = \dfrac{839}{13}$.

La prueba de hipótesis ANOVA utiliza las siguientes medidas de dispersión.

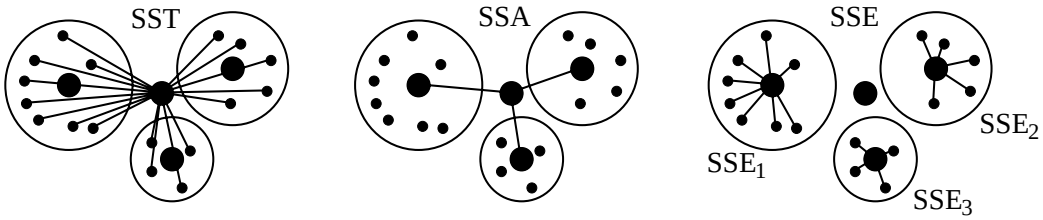
1. $SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \overline{y}\right)^2$ es la suma del cuadrado de las variaciones de las observaciones con respecto al promedio total observado. Esta medida indica que tan dispersos están los datos observados, es llamada también la variación total.

2. $SSA = \sum_{i=1}^{k} n_i \left(\overline{y}_i - \overline{y}\right)^2$ es la suma de la multiplicación de la cantidad de datos por el cuadrado de las variaciones de los promedios de cada tratamiento con respecto al promedio total observado. Esta medida indica que tan dispersos están los promedios observados de cada tratamiento con respecto al promedio total. Es decir mide la variación inter-tratamientos.

3. $SSE_i = \sum_{j=1}^{n_i} \left(y_{ij} - \overline{y_i}\right)^2$ es la suma del cuadrado de las variaciones de las observaciones del tratamiento i con respecto a su promedio. Esta medida indica que tan dispersos están los datos observados dentro del tratamiento i.

4. $SSE = \sum_{i=1}^{k} SSE_i$es la suma de los SSE_i. Note que se SSE es pequeño entonces cada SSE_i es pequeño y entonces las variaciones de los datos entre cada tratamiento son pequeñas. Así, SSE mide la variación intra-tratamientos.

Al representar cada población o tratamiento con un círculo cuyo centro sea el promedio muestral y las observaciones con puntos, se tienen las siguientes representaciones de las medidas de variabilidad expuestas:



Note que

1. SST nos brinda la variabilidad total de las observaciones independiente de la población a la que pertenezcan.

2. SSA brinda la variabilidad entre las observaciones por población. Si SSA es grande, significa que observaciones de poblaciones distintas son muy diferentes y los círculos se alejan unos de otros. Por el contrario, si SSA es pequeño los círculo se traslapan y observaciones de un mismo tratamiento son similares.

3. El diagrama SSE permite apreciar la variabilidad de las observaciones dentro de la población a la que pertenecen. Si SSE es muy pequeño, cada SSE_i se rá pequeño, y los círculos se comprimen, por lo tanto las observaciones de una misma población son muy similares y el promedio muestral sería un buen representante de las observaciones que pertenecen a la misma población. Si SSE es grande, los círculos crecen y se pueden traslapar, obteniendo poblaciones similares.

Así, para que las poblaciones sean similares según los datos muestrales, basta que SSA sea pequeño y SSE grande, esto se logra si $\dfrac{SSA}{SSE}$ es pequeño. El teorema siguiente establece una relación entre estas medidas de variabilidad.

Teorema 38 *Bajo la notación anterior se tiene que*

1. $SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 - \dfrac{T^2}{N}$

2. $SSA = \sum_{i=1}^{k} \dfrac{T_i^2}{n_i} - \dfrac{T^2}{N}$.

3. $SST = SSA + SSE$, *es decir*

> *"La variación inter-tratamientos más la variación intra-tratamientos es igual a la variación total"*

Prueba. *Se probará* (1) *y* (3) *. La demostración del resultado* 3 *se deja de ejercicio.*

1. *Se tiene que*

$$\begin{aligned}
SST &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}^2 - 2y_{ij}\overline{y} + \overline{y}^2\right) \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - 2\overline{y}\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}}_{T:\ suma\ de\ los\ datos} + \overline{y}^2\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i} 1}_{N:\ \#\ de\ datos} \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - 2\overline{y}\cdot T + \overline{y}^2 N \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - 2\left(\frac{T}{N}\right)T + \left(\frac{T}{N}\right)^2 N \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{T^2}{N}
\end{aligned}$$

3. *Se tiene que*

$$\begin{aligned}
SSE &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_i})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}^2 - 2y_{ij}\overline{y_i} + \overline{y_i}^2\right) \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - 2\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}\overline{y_i} + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\overline{y_i}^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - 2\sum_{i=1}^{k}\overline{y_i}\left(\sum_{j=1}^{n_i} y_{ij}\right) + \sum_{i=1}^{k}\overline{y_i}^2\left(\sum_{j=1}^{n_i} 1\right) \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - 2\sum_{i=1}^{k}\overline{y_i}T_i + \sum_{i=1}^{k}\overline{y_i}^2 n_i \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - 2\sum_{i=1}^{k}\left(\frac{T_i}{n_i}\right)T_i + \sum_{i=1}^{k}\left(\frac{T_i}{n_i}\right)^2 n_i \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{k}\frac{T_i^2}{n_i}
\end{aligned}$$

Por (1) *y* (2) *se tiene que*

$$SST - SSA = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{k}\frac{T_i^2}{n_i},$$

por lo tanto $SST - SSA = SSE$. ■

Como se mencionó anteriormente, para aceptar la hipótesis nula

$$H_0 \text{ (poblaciones equivalentes)} : \mu_1 = \mu_2 = ... = \mu_k,$$

se debe tener que SSA sea pequeño y SSE grande. Esto sucede se tendría, por el punto 3 del teorema anterior, que la variación total es explicada en su mayoría por SSE, es decir por la variación intra-tratamientos y no por la variación inter-tratamientos, la cuál sería pequeña.

Así, para aceptar H_0 basta que $\dfrac{SSA}{SSE}$ sea pequeño, pero ¿qué tan pequeño? Los conceptos y el resultado siguientes, nos permiten medir el tamaño de $\dfrac{SSA}{SSE}$ probabilísticamente.

Definición 31 *Bajo la notación anterior se definen los estadísticos*[12]

1. $S_1^2 = \dfrac{SSA}{k-1}$ *es la variancia inter-tratamientos muestral.*
2. $S^2 = \dfrac{SSE}{N-k}$ *es la variancia intra-tratamientos muestral.*

Teorema 39 *Bajo* H_0 *y la notación anterior, se tiene que el estadístico*

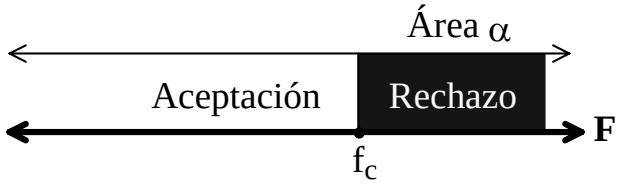
$$F = \frac{S_1^2}{S^2} = \underbrace{\frac{N-k}{k-1}}_{Constante}\frac{SSA}{SSE}$$

sigue una distribución F *con* $v_1 = k-1$ *y* $v_2 = N-k$ *grados de libertad.*

De acuerdo a lo anterior si en una muestra, el valor F observado

$$f_{obs} = \frac{s_1^2}{s^2}$$

es pequeño (SSA pequeño o SSE es grande), entonces se tendría que las poblaciones son equivalentes. Así, nuestras regiones de aceptación y rechazo para una prueba ANOVA son:



Donde el valor crítico $f_c = f_{1-\alpha,k-1,N-k}$. Además el valor P es $P\left(F > f_{obs}\right)$.

[12] Por simplicidad se denotan estos estádisticos igual a su valor observado en una muestra. De lo contrario, se debería diferenciar el estadístico SSA de el valor SSA en una muestra.

Ejemplo 85 *Se quiere determinar si la dosis de un determinado medicamento influye en el tiempo de sueño de los clientes que lo consumen. Varios voluntarios son expuestos a la aplicación de tres dosis (baja, media y alta) del medicamento, y se observa el número de minutos que duerme. Los datos son los siguientes:*

Dosis		
Baja	*Media*	*Alta*
67	*96*	*74*
69	*98*	*24*
72	*130*	*15*
79	*65*	*33*
		17

1. *¿Puede afirmarse que los tiempos de sueño no varían según la dosis del medicamento a un nivel de significancia del 5%?*

   Sean μ_1, μ_2 y μ_3 los tiempos promedios de sueño para cada dosis respectivamente. Considere las hipótesis

   $$H_0 \text{ (tiempos no varían según la dosis)} \ : \ \mu_1 = \mu_2 = \mu_3$$
   $$H_1 \text{ (tiempos varían según la dosis)} \ : \ \text{al menos dos de las medias no son iguales}$$

   Note que

	Baja	*Media*	*Alta*
n_i	4	4	5
T_i	287	389	163

   La suma total de datos es $T = \sum_{i=1}^{3} T_i = 839$ y $N = 13$. La suma de los datos al cuadrado es

   $$\sum_{i=1}^{3}\sum_{j=1}^{n_i} y_{ij}^2 = 68275$$

   Además

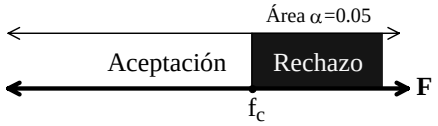
   $$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{T^2}{N} = 68275 - \frac{839^2}{13} = \frac{183\,654}{13} \approx: 14127.2$$
   $$SSA = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - \frac{T^2}{N} = \frac{287^2}{4} + \frac{389^2}{4} + \frac{163^2}{5} - \frac{839^2}{13} = \frac{1246\,509}{130} \approx 9588.53$$
   $$SSE = SST - SSA = \frac{183\,654}{13} - \frac{1246\,509}{130} = \frac{45\,387}{10} \approx 4538.7$$

   Por lo tanto el valor observado del estadísco de prueba es

   $$f_{obs} = \frac{S_1^2}{S^2} = \frac{N-k}{k-1}\frac{SSA}{SSE} = \frac{13-3}{3-1}\frac{\frac{1246\,509}{130}}{\frac{45\,387}{10}} = \frac{230\,835}{21\,853} \approx 10.563\,1$$

El valor crítico de la prueba es

$$f_c = f_{1-\alpha,k-1,N-k} = f_{0.95,2,10} = 4.103$$



Como $f_{obs} > f_c$ entonces f_{obs} se encuentra en el área de rechazo, por lo tanto se rechaza H_0, es decir se concluye que no existe evidencia sufiente de que los tiempos de sueño no varían según la dosis.

2. *Acote el Valor P de la prueba*

   Este valor es

$$Valor\ P = P\left(F > f_{obs}\right) = P\left(F > 10.563\,1\right) = 1 - P\left(F \leq 10.563\,1\right)$$

   donde $F \sim F\left(2, 10\right)$. Note que $10.5631 > 1$ entonces utilizando la tabla, como $f_{0.99,2,10} = 7.559 < 10.563\,1$ entonces $P\left(F \leq 10.563\,1\right) \in \left]0.99, 1\right[$, entonces

$$Valor\ P = 1 - P\left(F \leq 10.563\,1\right) \in \left]0, 0.1\right[$$

   Note que como $Valor\ P < \alpha = 0.05$, entonces se rechaza H_0 y se obtiene la misma conclusión dada en el punto anterior.

Ejercicio 42 *Todos los domingos durante 2 horas, don Juan sale a pescar y utiliza una de sus tres cañas de pescar. Don Juan desea determinar si el número de peces que pesca por domingo es independiente de la caña que utilice. Para ello, registró durante 15 domingos el número de peces obtenidos y la caña utilizada, los datos son los siguientes:*

Caña 1	*Caña 2*	*Caña 3*
12	*10*	*16*
10	*17*	*14*
18	*16*	*16*
12	*13*	*11*
14		*20*
		21

1. *Plantee la hipótesis nula y alternativa. ¿Cuál es el valor P de la prueba?* $R/$ $f_{obs} = 1.287$, $Valor\ P \in \left]0.1, 0.9\right[$

2. *Determine las regiones de aceptación y rechazo, utilizando un nivel de significancia de* 0.05? $R/$ $f_c = 3.885$

3. *Al nivel de significancia de* 0.05*, ¿Existe la evidencia suficiente de que el número de peces obtenidos cada domingo es independiente de la caña utilizada?* $R/$ *Si hay evidencia.*

4.4 Ejercicios

1. Las calificaciones de un curso de estadística para un semestre particular fueron las siguientes

Calificación	100	80	60	40	20
Frecuencia Observada	14	18	32	20	16

   Pruebe la hipótesis al nivel de significancia del 5% de que la distribución de las calificaciones es uniforme. *R/*

2. Una muestra de 1000 votantes se clasifica de acuerdo cons sus ingresos; como bajo, promedio y alto y si están a favor o encontra de un nueva reforma de ley. Las frecuencias observadas se dan por la siguiente tabla de contingencia:

Reforma de ley	Bajo	Promedio	Alto
A favor	182	213	203
En contra	154	138	110

   Puede concluirse con estos datos que la opinión de un votante referente a la reforma de ley es independiente de su nivel de ingreso. *R/*

3. Durante un semestre un estudiante recibió calificaciones en diversas materias como se muestra en la siguiente tabla:

Matemáticas	72	80	83	75	
Ciencias	81	74	77		
Inglés	88	82	90	87	80
Economía	74	71	77	70	

   Determine si existe diferencia significativa entre el rendimiento promedio de las materias que cursó este semestre el estudiante. *R/* Valor $P = 0.0046$, los datos apoyan que si existe diferencia.

4. Se compararon estudiantes avanzados en la carrera de computación de tres universidades distintas con respecto a sus conocimientos en sistemas operativos. A cada estudiante se le aplicó una prueba y se registraron sus notas, los datos son los siguientes:

Universidad		
U1	U2	U3
67	96	74
90	80	94
72	71	65
79	86	53
		87

   ¿Puede afirmarse que no hay diferencia entre los conocimientos sobre sistemas operativos en las tres universidades, a un nivel de significancia del 5%? *R/* $f_{obs} \approx 0.500\,742$, $f_c = 4.103$, se acepta H_0

5. Una encuesta a 320 familias con 5 hijos o hijas, reveló la siguiente información:

Número de hijas	0	1	2	3	4	5
Número de familias	18	56	110	88	40	8

Pruebe la hipótesis, al nivel de significancia del 5%, de que la distribución del número de hijas en familias con 5 hijos es $P(X=k)=\dfrac{C(5,k)}{2^k}$.$R/$ $\chi^2_{obs}=11.96, \chi^2_c=11.0705$. Hay evidencia en contra de que la distribución de X es la dada.

6. El semestre anterior se impartieron varios grupos del curso de Estadística en la Universidad Bienestar Seguro, distribuidos en las tres modalidades de asistencia: virtual, semi-virtual y presencial. El primer examen fue colegido (los grupos realizaron el mismo examen) y los resultados de algunos estudiantes fueron:

Presencial	Semi-virtual	Virtual
67	96	74
90	80	94
72	71	65
79	86	

¿Puede afirmarse que las notas promedio poblacionales del primer examen no varía segun la modalidad de asistencia? $R/$ El Valor $P>0.1$. Hay evidencias a favor de que los promedios no varían según la modalidad.

7. El Viceministro de Comercio no está seguro si deben estar tantas personas contratadas para atender las quejas presentadas a la Oficina del Consumidor durante todo el año, pues considera que en ciertos cuatrimestres del año no se deberían presentar tan quejas. Ante esto, se obtuvo del registro de la oficina el número de quejas recibidas durante 15 distintos cuatrimestres elegidos al azar. Los resultados fueron los siguientes

I cuatrimestre ($enero-abril$)	II cuatrimestre ($mayo-agosto$)	III cuatrimestre ($set.-dic.$)
120	100	150
107	155	144
128	160	106
135	130	111
	104	150
		80

¿Existe la evidencia suficiente de que el número quejas por cuatrimestre es independiente del número de cuatrimestre del año?

8. Se lanza un dado 150 veces y se registra los resultados obtenidos. Estos se resumen en la siguiente tabla

Resultado:	1	2	3	4	5	6
frecuencia:	18	22	28	26	21	35

¿Hay evidencia para pensar que el dado es ilegal?

9. Considere el experimente de lanzar un dado varias veces hasta obtener un número par y se anota el número de lanzamientos realizados. Suponga que se realiza este experimento 200 veces, obteniendo los siguientes resultados

# de lanzamientos realizados (X):	1	2	3	4	5	6
frecuencia:	95	56	30	13	4	2

Realice la prueba de hipótesis de que la distribución de X es $f_X(k) = \dfrac{1}{2^k}$.

10. Seguidamente se presentan una serie de experimentos. Simule estos experimentos 200 veces utilizando Excel u otro software, luego registre las frecuencias obtenidas para la variable indicada. Finalmente con los datos simulados pruebe la hipótesis de que X sigue la distribución f_X indicada.

    (a) Se lanza un dado hasta obtener un resultado mayor a 4. X : # de lanzamientos realizados. $f_X(k) = \left(\dfrac{2}{3}\right)^{k-1} \dfrac{1}{3}$.

    (b) Se lanza un dado hasta obtener dos resultados mayores a 4. X : # de lanzamientos realizados. $f_X(k) = \left(\dfrac{2}{3}\right)^{k-1} \dfrac{1}{9}$.

    (c) Se lanza un dado 20 veces. X : # de lanzamientos realizados en los cuales se obtuvo un resultado mayor a 4. $f_X(k) = C(20,k) \left(\dfrac{1}{3}\right)^{k} \left(\dfrac{2}{3}\right)^{20-k}$.

    (d) Se lanza un dado tres veces y cada vez que se lanza, si el resultado es par, se borra esa cara y se pone un número impar. X : # de lanzamientos en los cuales se obtuvo un resultado par.

$$f_X(k) = \begin{cases} \dfrac{1}{8} & si \quad k = 0 \\ \dfrac{37}{72} & si \quad k = 1 \\ \dfrac{1}{3} & si \quad k = 2 \\ \dfrac{1}{36} & si \quad k = 3 \end{cases}$$

11. En un estudio, se registro la edad de 250 personas contratadas este año para laborar en empresas agrícolas. Los datos se agruparon en las clases siguientes:

Edad (x)	Frecuencia observada
$x \leq 19$	10
$19 \leq x < 22$	35
$22 \leq x < 25$	90
$25 \leq x < 28$	80
$28 \leq x < 31$	25
31 y mayor	10

¿Muestran los datos observados evidencia significativa para rechazar la hipótesis de que la edad sigue una distribución Normal con media de 25 años y desviación estándar de dos años?

12. En una encuesta sobre una soda comedor de la Universidad Futuro Garantizado, se les preguntó a 250 clientes su opinión sobre el higiene de las instalaciones e higiene en la presentación de los productos. Los resultados se resumen en la siguiente tabla de contingencia.

		Higiene productos		
		Bueno	Regular	Malo
Higiene de las instalaciones	*Bueno*	10	30	10
	Regular	15	50	40
	Malo	15	20	70

¿Existe evidencia en contra de que la opinión que tiene un cliente sobre la higiene de las instalaciones es independiente de su opinión sobre la higiene en la presentación de los productos? $R/ \quad \chi^2_{obs} = 37.927$, Valor $P < 0.01$. Existe evidencia en contra de la independencia de variables.

13. En una encuesta sobre una soda comedor de la Universidad Futuro Garantizado, se les preguntó a 250 clientes su opinión sobre el tamaño de las porciones de comida dada y la atención brindada. Los resultados se resumen en la siguiente tabla de contingencia.

		Atención			
		Excelente	Buena	Regular	Mala
Tamaño de las porciones	Muy pequeñas	0	2	3	10
	pequeñas	5	4	9	15
	adecuadas	15	18	15	10
	grandes	25	15	8	5
	muy grandes	45	26	17	3

A un nivel de significancia del 10%, ¿Existe evidencia en contra de que la opinión que tiene un cliente sobre la atención brindada es independiente de su opinión sobre el tamaño de las porciones?

14. Seguidamente se presenta la edad, en años cumplidos, de 15 estudiantes recién graduados seleccionados aleatoriamente de tres carreras de la Universidad Futuro Garantizado (las cuales tiene la misma duración):

Carrera	Edades
Computación	$30, 27, 28, 32, 25, 27$
Administración	$24, 25, 24, 27, 25$
Forestal	$23, 24, 25, 28$

A un nivel de significancia del 10%, ¿Hay evidencia de que la edad promedio de graduación sean distintas en las tres carreras?

15. En la siguiente tabla se muestra la relación entre el rendimiento académico de estudiantes en

matemáticas y ciencias:

		Calificaciones en matemática		
		Altas	*medias*	*bajas*
Calificaciones en **Ciencias**	*altas*	30	71	20
	medias	50	144	20
	bajas	35	90	40

Pruebe la hipótesis de que el desempeño en física es independiente del desempeño en matemáticas. $R/ \quad \chi^2_{obs} = 15.982, v = 4, Valor\ P \approx 0.$hay dependencia.

16. En una encuesta sobre la soda comedor EL COMELON, se les preguntó a 200 clientes su opinión sobre la variedad de los alimentos y su nivel de ingreso. Los resultados se resumen en la siguiente tabla de contingencia.

		Nivel de ingreso		
		Bajo	Medio	Alto
Variedad	Poco	3	10	27
	Regular	15	20	50
	Mucha	21	40	14

¿Existe evidencia de que la opinión que tiene un cliente sobre la variedad de los alimentos es independiente de su nivel de ingreso? $R/ \quad \chi^2_{obs} = 36,8627451$,Valor $P < 0.01$, existe evidencia en contra de que las variables sean independientes.

17. Un curso es impartido normalmente por tres profesores. Un estudiante considera que la nota en el curso depende del profesor que lo imparta. Ante esto la Oficina de Registro seleccionó al azar el promedio de 11 estudiantes que aprobaron el curso, los cuales se muestran a continuación

Profesor A	Profesor B	Profesor C
77	96	74
88	84	84
81	75	75
82	80	
78		

Con base a estos datos, ¿Puede afirmarse que la nota promedio en el curso no varía según el profesor que lo imparte?$R/ \quad f_{obs} \approx 0.758\,938$, Valor $P > 0.1$. Hay evidencias a favor de que los promedios no varían según el profesor

Capítulo IV

Análisis de Regresión

Se estudia los principales anális de regresión: lineal simple, no lineal simple, múltiple lineal y múltiple no lineal.

1 Introducción

PROBLEMA: Sean X, Y dos variables cuantitativas continuas, se quiere determinar cómo varía Y en función de X. Suponga que se toma una muestra de valores de estas variables: n puntos (x_i, y_i) con $i = 1, 2, 3, ..., n$. ¿Cómo tener una idea sobre la relación entre X y Y a partir de los valores muestrales?

En general, la Estadística Inferencial establece diversos métodos para estudiar la relación entre dos variables dependiendo de su naturaleza:

Relación entre	**Método**
Dos variables cualitativas	Pruebas de Independencia
Una cualitativa y otra cuantitativa	Análisis de varianzas (ANOVA)
Dos variables cuantitativas	Métodos de Regresión

Los métodos de regresión pretende establecer una ecuación que evidencie la relación entre cada valor de una variable X y el valor esperado de una variable Y para ese valor de X. Incluso se pueden establecer métodos de regresión entre más de dos variables (Regresión Múltiple).

Definición 32 *Dadas dos variables cuantitativas continuas Y, X se define Y_x como la variable aleatoria Y que corresponde a un valor fijo x de X. Es decir Y_x es la variable Y condicianada al valor x de X :*

$$Y_x = Y|\,(X = x)\,.$$

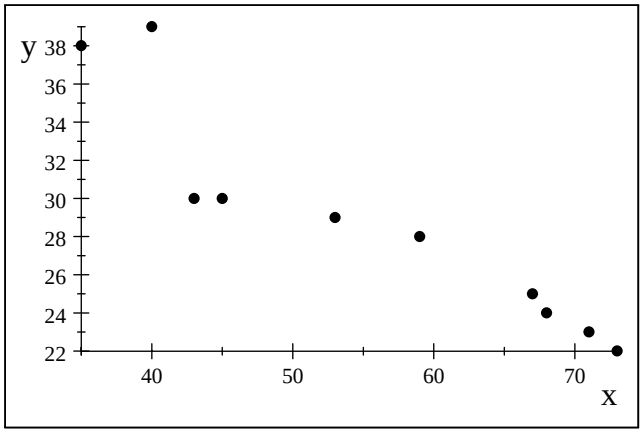
Dado que las variables son cuantitativas, la gráfica de los puntos muestales nos permiten tener una idea de la forma de la relación que existe entre ellas.

Definición 33 *Un **Diagrama de Dispersión** es la gráfica formada por los puntos de la muestra. Este nos da la idea de la relación existen entre los valores de la muestra*

Ejemplo 86 *En un curso de maestría de cierta universidad, se obtuvo para una muestra de 10 estudiantes, su edad Y, y el promedio X obtenido en el curso:*

X :	68	73	35	40	59	53	67	45	71	43
Y :	24	22	38	39	28	29	25	30	23	30

Realice el diagrama de dispersión para estas variables.



La gráfica eviendencia una relación inversa entre X y Y. pues a mayor edad menor nota.

2 Regresión lineal simple

2.1 Definición

Definición 34 *Se dice que existe una* ***regresión lineal simple*** *entre dos variables cuantitativas continuas Y, X, si existen constantes α β tales que*

$$E(Y_x) = \mu_{Yx} = \alpha + \beta x$$

donde:

1. *α es la intersección con el eje de las ordenadas, y tiene la misma unidad de medida que Y. Note que $E(Y_{x=0}) = \alpha$, es decir α el valor medio de Y cuando $X = 0$.*
2. *β pendiente o aumento promedio en Y debido a cada unidad en X y tiene la misma unidad de medida que $\dfrac{Y}{X}$.*

Por simplicidad se denota $E(Y_x)$ por y, así la ecuación anterior

$$y = \alpha + \beta x$$

es llamada la **Ecuación de Regresión Lineal** o ecuación del modelo de regresión lineal de Y con función de X.

Ejemplo 87 *Dado un grupo de niños, se quiere analizar su edad (X) en años y su estatura (Y) en cm. Entonces $Y_{x=5}$ es la variable que indica la estatura de niños de 5 años, y $E(Y_{x=5})$ es la constante que indica la estatura esperada de los niños de 5 años. Suponga que existe un regresión lineal simple entre X y Y :*

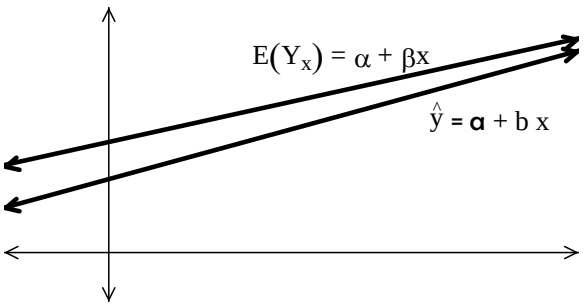
$$\mu_{Yx} = 20 + 10x$$

Así $20cm$ sería la estatura promedio de los niños recién nacidos y la estatura aumenta en promedio $10\dfrac{cm}{año}$.

2.2 Estimación puntual de los coeficientes de regresión

Si existe una regresión lineal entre dos variables cuantitativas continuas Y, X se recurre a una muestra para hacer una estimación de los parámetros:

Parámetro	Estimación
α	a
β	b
$\mu_{Yx} = \alpha + \beta x$	$\widehat{y}(x) = a + bx$

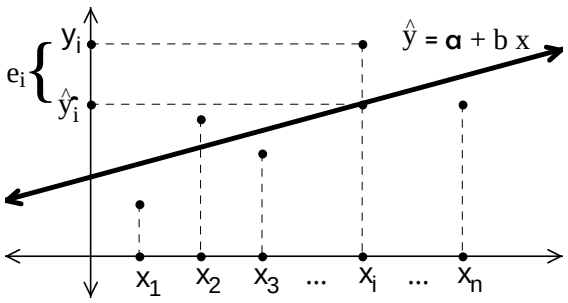


La recta $\widehat{y}(x) = a + bx$ es llamada la **línea de mejor ajuste**. Note que la recta $\widehat{y}(x) = a + bx$ es una estimación de la recta teórica $\mu_{Yx} = \alpha + \beta x$.

¿Cómo hallar las estimaciones a **y** b? Note que dado el punto observado (x_i, y_i), se tiene que y_i es el valor de Y observado para $X = x_i$, y $\widehat{y}(x_i)$ es la estimación del valor esperado de Y cuando $X = x_i$. Así, una forma de determinar la recta de mejor ajuste es suponiendo que lo observado (y_i) debe estar lo más cercano posible a lo esperado $(\widehat{y}(x_i))$. Por lo tanto, se busca la recta $\widehat{y}(x) = a + bx$ que mejor se ajuste a los valores de la muestra: (x_i, y_i) con $i = 1, 2, 3, ..., n$; para ello se debe cumplir que los errores

$$|e_i| = |y_i - \widehat{y}(x_i)| = |y_i - a - bx_i| \text{ para } i = 1, 2, 3, ..., n$$

sean mínimos:



Un método para minimizar estos valores es logrando que **la suma del cuadrado de los errores**

$$SCE = f(a, b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

sea mínima. La función SCE alcanza su mínimo en su único punto crítico que se halla igualando a cero las derivadas parciales

$$\begin{aligned} \frac{\partial(SCE)}{\partial a} &= -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \\ \frac{\partial(SCE)}{\partial b} &= -2\sum_{i=1}^{n} x_i(y_i - a - bx_i) = 0 \end{aligned}$$

Estas ecuaciones son equivalentes a

$$\begin{cases} \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \end{cases}$$

Resolviendo este sistema se obtiene que

$$\begin{aligned} b &= \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\cdot\overline{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x})^2} \\ a &= \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n} = \overline{y} - b\overline{x} \end{aligned}$$

El método utilizado para hallar la estimación a y b es llamado **el método de mínimos cuadrados**.

Teorema 40 *Sean X, Y dos variables cuantitativas continuas. Dada una muestra de n valores de estas variables: (x_i, y_i) con $i = 1, 2, 3, ..., n$, las estimaciones puntuales, por el método de mínimos cuadrados, de los coeficientes de regresión lineal α y β son respectivamente*

$$b = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\cdot\overline{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x})^2} \qquad y \qquad a = \overline{y} - b\overline{x}$$

Ejemplo 88 *Un estudiante de computación, que tiene una pequeña empresa de elaboración de software, llevó a cabo un estudio para determinar la relación entre el tiempo que tarda elaborando un software en días (X) y el ingreso obtenido por su venta en dólares (Y). Se recolectaron los valores de estas variables para 14 software, obteniendo los siguientes datos:*

X :	2	3	5	9	10	16	19	20	24	27	32	41	55	60
Y :	100	200	250	400	500	850	930	900	1300	1360	1500	2050	2800	2900

1. *Encuentre la ecuación de regresión lineal para el ingreso como función del tiempo de elaboración de un software*

   De acuerdo a los datos:

x	y	x^2	y^2	xy
2	100	4	10000	200
3	200	9	40000	600
5	250	25	62500	1250
9	400	81	160000	3600
10	500	100	250000	5000
16	850	256	722500	13600
19	930	361	864900	17670
20	900	400	810000	18000

	24	1300	576	1690000	31200
	27	1360	729	1849600	36720
	32	1500	1024	2250000	48000
	41	2050	1681	4202500	84050
	55	2800	3025	7840000	154000
	60	2900	3600	8410000	174000
Suma :	323	16040	11871	29162000	587890

Así, se tiene que

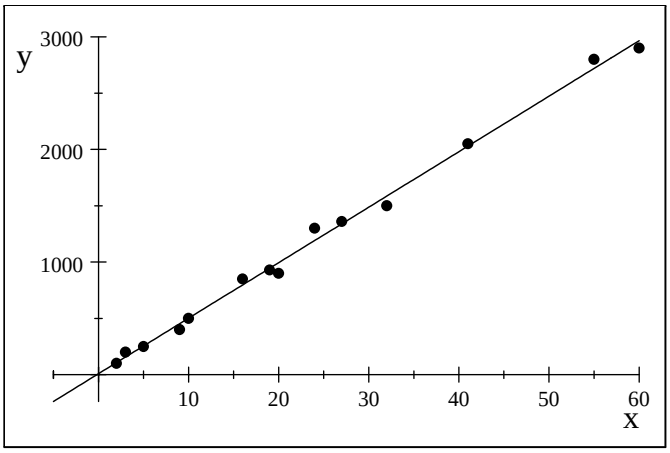
$$b = \frac{\frac{1}{n}\sum xy - \overline{x}\cdot\overline{y}}{\frac{1}{n}\sum x^2 - (\overline{x})^2} = \frac{\frac{1}{14}(587890) - \frac{323}{14}\frac{16040}{14}}{\frac{1}{14}(11871) - \left(\frac{323}{14}\right)^2} \approx 49.2935.$$

$$a = \overline{y} - b\overline{x} \approx \frac{16040}{14} - (49.2935)\frac{323}{14} \approx 8.44282$$

Por lo tanto la ecuación de mínimos cuadrados es

$$\widehat{y} = 49.2935x + 8.44282$$

Esta recta brinda aproximadamente el valor esperado del software cuando se tardó x días en elaborarlo. Al gráficar esta recta en el Diagrama de Dispersión se puede observar lo bien que se ajusta a los puntos muestrales:



2. *¿Aproximadamente, al elaborar un software, cuánto es el valor promedio del ingreso fijo (ingreso que no depende del tiempo de elaboración)?*

   Recuerde que el valor esperado del ingreso por elaborar un software está dado por

$$\mu_{Y\,x} = \alpha + \beta x$$

donde α es el ingreso promedio fijo y βx es el ingreso promedio variable. Así, el valor promedio del ingreso fijo es α cuya estimación es

$$a = 8.\,442\,82 \text{ dólares.}$$

3. *¿Aproximadamente, al elaborar un software, cuánto es el aumento promedio del ingreso por día de elabolación?*

   Note que β es el aumento promedio del ingreso por día de elabolación. Su aproximación es

$$b = 49.\,293\,5 \text{ dolares/día}$$

Ejemplo 89 *Considere los datos de la siguiente tabla:*

$X:$	2	3	5	9	13
$Y:$	8	10	18	30	40

A partir de estos datos, estime el coeficiente β de la ecuación de regresión $\mu_{Y\,x} = 2 + \beta x$ utilizando el método de mínimos cuadrados.

En este caso, sería un error utilizar el teorema anterior para determinar la estimación b de β. Esto por que este teorema realiza una estimación conjunta de α y β, es decir las estimaciones que brinda no son independientes. Así, se busca el b tal que $\widehat{y}(x) = a + bx$ y la suma del cuadro de los errores

$$SCE = \sum_{i=1}^{5} e_i^2 = \sum_{i=1}^{5} \left(y_i - \widehat{y}(x_i)\right)^2 = \sum_{i=1}^{5} \left(y_i - 2 - bx_i\right)^2$$

sea mínima. Note que SCE es una función que depende de b, es decir

$$f(b) = SCE = \sum_{i=1}^{5} \left(y_i - 2 - bx_i\right)^2$$

Si se desarrolla la suma y se sustituyen los datos de la tabla, se obtiene que f es una función cuadrática cóncava hacia arriba, por lo tanto su mínimo se encuentra el vértice. Dado que es muy largo desarrollar f para calcular la coordena X del vértice de la parábola que representa, haciendo uso del cálculo diferencial, se sabe que la función cuadrática f alcanza mínimo cuando $f'(b) = 0$, donde

$$f'(b) = \sum_{i=1}^{5} \left[2\left(y_i - 2 - bx_i\right) \cdot -x_i\right] = -2\left(\sum_{i=1}^{5} x_i y_i - 2\sum_{i=1}^{5} x_i - b\sum_{i=1}^{5} x_i^2\right).$$

A partir de los datos: $\sum_{i=1}^{5} x_i = 32$, $\sum_{i=1}^{5} x_i^2 = 288$ y $\sum_{i=1}^{5} x_i y_i = 926$, entonces

$$f'(b) = -2\left(926 - 2 \cdot 32 - 288b\right) = 576b - 1724.$$

Así, $f'(b) = 0$ si $b = \dfrac{1724}{576} = \dfrac{431}{144}$. Por lo tanto la estimación de β es

$$b = \frac{431}{144}.$$

Ejercicio 43 *Se espera que, por lo general, el número de horas de estudio* $(\mathbf{X})$ *en la preparación para hacer un examen tenga una correlación directa con la calificación obtenida* $(\mathbf{Y})$ *alcanzada en tal examen. Se obtuvieron las horas de estudio así como las calificaciones (en escala de 0 a 100) obtenidas por diez estiudiantes seleccionados al azar de un grupo, los datos están resumidos en la siguiente tabla:*

$\sum x$	$\sum y$	$\sum y^2$	$\sum x^2$	$\sum xy$
118	591	39013	1648	7956

1. *Determine la ecuación de mínimos cuadrados para la calificación como función de las horas de estudio.* $R/ \quad b = 3.842723005, a = 13.75586854$
2. *¿Aproximadamente, cuál es la calificación promedio de los estudiantes que no estudiaron para el examen?* $R/ \quad a = 13.75586854$
3. *¿Cuánto es el aumento promedio aproximado en la calificación por hora de estudio?* $R/ \quad b = 3.842723005$

Ejercicio 44 *Considere los datos de la siguiente tabla:*

$x:$	1	5	10	20
$y:$	4	12	21	43

A partir de estos datos, estime el coeficiente β *de la ecuación de regresión* $\mu_{Y\,x} = \beta + \beta x$ *utilizando el método de mínimos cuadrados.* $R/ \quad b = \frac{567}{281}$

Otras estimaciones puntuales a considerar son:

1. Estimación de la variancia poblacional σ_x^2 de X :

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

2. Estimación de la variancia poblacional σ_y^2 de Y :

$$s_y^2 = \frac{\sum\limits_{i=1}^{n} (y_i - \overline{y})^2}{n-1}$$

Notación 1 *En adelante se utilizará la siguiente notación:*

1. $S_{xx} = (n-1)\, s_x^2$.
2. $S_{yy} = (n-1)\, s_y^2$.
3. $S_{xy} = \sum\limits_{i=1}^{n} \left[(x_i - \overline{x})\, (y_i - \overline{y}) \right]$.

Teorema 41 *Se cumple que*

1. $S_{xx} = \sum x^2 - \dfrac{\left(\sum x\right)^2}{n}$.
2. $S_{yy} = \sum y^2 - \dfrac{\left(\sum y\right)^2}{n}$.
3. $S_{xy} = \sum xy - \dfrac{\left(\sum x\right)\left(\sum y\right)}{n}$
4. $SCE = S_{yy} - bS_{xy}$.

Prueba. *Se probará la* (1) *y la* (4), *las demás quedan de ejercicio.*

1. *Note que*

$$\begin{aligned} S_{xx} &= (n-1)\, s_x^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} \left(x_i^2 - 2x_i\overline{x} + \overline{x}^2\right) \\ &= \sum_{i=1}^{n} x_i^2 - 2\overline{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \overline{x}^2 = \sum_{i=1}^{n} x_i^2 - 2\overline{x}\,(n\overline{x}) + n\overline{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - -n\overline{x}^2 = \sum x^2 - \frac{\left(\sum x\right)^2}{n}. \end{aligned}$$

2. *Se tiene que*

$$\begin{aligned} SCE &= \sum_{i=1}^{n} (y_i - a - bx_i)^2 = \sum_{i=1}^{n} (y_i - (\overline{y} - b\overline{x}) - bx_i)^2 \\ &= \sum_{i=1}^{n} ((y_i - \overline{y}) + b\,(\overline{x} - x_i))^2 \\ &= \sum_{i=1}^{n} \left((y_i - \overline{y})^2 + 2b\,(y_i - \overline{y})\,(\overline{x} - x_i) + b^2\,(\overline{x} - x_i)^2\right) \\ &= \sum_{i=1}^{n} (y_i - \overline{y})^2 + 2b\sum_{i=1}^{n} (y_i - \overline{y})\,(\overline{x} - x_i) + b^2 \sum_{i=1}^{n} (\overline{x} - x_i)^2 \\ &= S_{yy} - 2bS_{xy} + b^2 S_{xx} \end{aligned}$$

Dado que $S_{xx} = \sum_{i=1}^{n} x_i^2 - n\,(\overline{x})^2$ *y* $S_{xy} = \sum_{i=1}^{n} x_i y_i - n\,(\overline{x} \cdot \overline{y})$ *se tiene que*

$$bS_{xx} = \frac{\dfrac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x} \cdot \overline{y}}{\dfrac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x})^2} \cdot S_{xx} = \frac{\dfrac{1}{n} S_{xy}}{\dfrac{1}{n} S_{xx}} \cdot S_{xx} = S_{xy}$$

Por lo tanto

$$SCE = S_{yy} - 2bS_{xy} + b^2 S_{xx} = S_{yy} - 2bS_{xy} + bS_{xy} = S_{yy} - bS_{xy} \quad \blacksquare$$

2.3 Estimación por intervalo de los coeficientes de regresión

Recuerde que α y β son parámetros a estimar que cumplen que $E(Y_x) = \alpha + \beta x$. El siguiente teorema nos brinda los estimadores de estos parámetros.

Teorema 42 *Sean X, Y dos variables cuantitativas continuas. Suponga que existen α y β tales que $E(Y_x) = \alpha + \beta x$. Dada una muestra de n valores de estas variables: (x_i, y_i) con $i = 1, 2, 3, ..., n$, se tiene que:*

1. *Sea $\overline{Y} = \dfrac{\sum_{i=1}^{n} Y_{x_i}}{n}$, se tiene que $E\left(\overline{Y}\right) = \alpha + \beta\overline{x}$*
2. *Un estimador insesgado de β es*

$$B = \frac{\sum_{i=1}^{n} \left[(x_i - \overline{x})\left(Y_{x_i} - \overline{Y}\right)\right]}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

3. *Un estimador insesgado de α es*

$$A = \overline{Y} - B\overline{x}$$

Prueba. *Se prueba la (2), las otras quedan de ejercicio.*

$$\begin{aligned} E(B) &= \frac{\sum_{i=1}^{n} \left[(x_i - \overline{x})\left(E(Y_{x_i}) - E\left(\overline{Y}\right)\right)\right]}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} \left[(x_i - \overline{x}) E(Y_{x_i})\right] - E\left(\overline{Y}\right) \sum_{i=1}^{n} (x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \\ &= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(\alpha + \beta x_i) - (\alpha + \beta\overline{x}) \left(\sum_{i=1}^{n} (x_i - \overline{x})\right)}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} (x_i - \overline{x}) \beta x_i - \beta\overline{x} \sum_{i=1}^{n} (x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \\ &= \frac{\beta \sum_{i=1}^{n} \left((x_i - \overline{x}) x_i - \overline{x}(x_i - \overline{x})\right)}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{\beta \sum_{i=1}^{n} (x_i - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \beta \qquad \blacksquare \end{aligned}$$

Seguidamente se define los errores a considerar.

Definición 35 *Se define para cada valor x_i :*

1. ***(Estimador del Error)****La variable aleatoria residuo o error para x_i*

$$E_i = Y_{x_i} - A - Bx_i$$

2. *La variable aleatoria error o perturbación del modelo por*

$$\varepsilon_i = Y_{x_i} - \alpha - \beta x_i$$

Note que $e_i = y_i - \widehat{y}(x_i) = y_i - a - bx_i$ es una estimación del error x_i dada por el estimador E_i. Además ε_i es el eror que se da entre la variable $Y|X = x_i$ y su esperanza. Note que $Y_{x_i} = \alpha + \beta x_i + \varepsilon_i$.

Teorema 43 *Se tiene que:*

1. $E(\varepsilon_i) = 0$
2. $Var(\varepsilon_i) = Var(Y_{x_i})$.
3. $E(E_i) = 0$.
4. $B - \beta = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})\varepsilon_i}{S_{xx}}$

Prueba. *Se prueba la* (3) *y la* (4), *las demás se dejan de ejercicio.*

3. *Note que*

$$E_i = Y_{x_i} - A - Bx_i = \alpha + \beta x_i + \varepsilon_i - A - Bx = (\alpha - A) + (\beta - B)x_i + \varepsilon_i,$$

por lo tanto

$$E(E_i) = E(\alpha - A) + x_i E(\beta - B) + E(\varepsilon_i) = E(\varepsilon_i) = 0.$$

4. *Note que*

$$\begin{aligned}
\sum_{i=1}^{n}\left[(x_i - \overline{x})\left(Y_{x_i} - \overline{Y}\right)\right] &= \sum_{i=1}^{n}\left[\left(x_i Y_{x_i} - \overline{x}Y_{x_i} - x_i\overline{Y} + \overline{x}\overline{Y}\right)\right] = \sum_{i=1}^{n} x_i Y_{x_i} - \overline{x}n\overline{Y} - n\overline{x}\overline{Y} + n\overline{x}\overline{Y} \\
&= \sum_{i=1}^{n} x_i Y_{x_i} - \overline{x}n\overline{Y} = \sum_{i=1}^{n} x_i Y_{x_i} - \overline{x}\sum_{i=1}^{n} Y_{x_i} = \sum_{i=1}^{n}\left[(x_i - \overline{x})Y_{x_i}\right] \\
&= \sum_{i=1}^{n}\left[(x_i - \overline{x})(\varepsilon_i + \alpha + \beta x_i)\right] = \sum_{i=1}^{n}(x_i - \overline{x})\varepsilon_i + \sum_{i=1}^{n}(x_i - \overline{x})(\alpha + \beta x_i) \\
&= \sum_{i=1}^{n}(x_i - \overline{x})\varepsilon_i + \alpha\sum_{i=1}^{n}(x_i - \overline{x}) + \beta\sum_{i=1}^{n}(x_i - \overline{x})x_i \\
&= \sum_{i=1}^{n}(x_i - \overline{x})\varepsilon_i + \beta\sum_{i=1}^{n}(x_i - \overline{x})x_i \\
&= \sum_{i=1}^{n}(x_i - \overline{x})\varepsilon_i + \beta\left(\sum_{i=1}^{n}(x_i - \overline{x})x_i - \overline{x}\underbrace{\sum_{i=1}^{n}(x_i - \overline{x})}_{0}\right) \\
&= \sum_{i=1}^{n}(x_i - \overline{x})\varepsilon_i + \beta\sum_{i=1}^{n}(x_i - \overline{x})^2
\end{aligned}$$

por lo tanto

$$B = \frac{\sum_{i=1}^{n}\left[(x_i-\overline{x})\left(Y_{x_i}-\overline{Y}\right)\right]}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})\,\varepsilon_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2} + \beta = \frac{\sum_{i=1}^{n}(x_i-\overline{x})\,\varepsilon_i}{S_{xx}} + \beta$$

de donde se obtiene el resultado. ■

Hasta el momento se cuenta con A y B, estimadores insesgados de α y β respectivamente. Dado que se quiere estimar por intervalo estos parámetros, se requiere la distribución de A y B.

Teorema 44 *Si para cada valor x_i de X, se tiene que la variable Y_{x_i} cumple que (**hipótesis de regresión**):*

a) Y_{x_i} *sigue una distribución normal*

b) $E(Y_{x_i}) = \alpha + \beta x_i$

c) $Var(Y_{x_i}) = \sigma^2$ *(note que no depende del valor de x_i)*

d) Las variables Y_{x_i} son independientes

Es decir, si se cumple que para cada x_i :

$$Y_{x_i} \sim N\left(\alpha + \beta x_i, \sigma^2\right)$$

y las variables Y_{x_i} son independientes, entonces

1. *Se tiene que $\varepsilon_i \sim N\left(0, \sigma^2\right)$ y los ε_i son independientes.*
2. $Var(B) = \dfrac{\sigma^2}{S_{xx}}$
3. $Var(A) = \sigma^2\left(\dfrac{1}{n} + \dfrac{\overline{x}^2}{S_{xx}}\right) = \dfrac{\sigma^2\sum_{i=1}^{n}x_i^2}{nS_{xx}}$.
4. $B \sim N\left(\beta, \dfrac{\sigma^2}{S_{xx}}\right)$.
5. $A \sim N\left(\alpha, \sigma^2\left(\dfrac{1}{n} + \dfrac{\overline{x}^2}{S_{xx}}\right)\right)$

Prueba. *A continuación se probarán los resultados del* (2) *al* (4) *. El resultado* (1) *y* (5) *. se deja de ejercicio.*

2. *Recuerde que* $Var(B) = E\left((B-\beta)^2\right)$, *por el teorema anterior se tiene que*

$$\begin{aligned} Var(B) &= E\left((B-\beta)^2\right) = E\left(\left(\frac{\sum_{i=1}^{n}(x_i-\overline{x})\varepsilon_i}{S_{xx}}\right)^2\right) \\ &= \frac{1}{(S_{xx})^2}E\left(\sum_{i=1}^{n}(x_i-\overline{x})^2\varepsilon_i^2 + 2\sum_{i\neq j}(x_i-\overline{x})\varepsilon_i(x_j-\overline{x})\varepsilon_j\right) \\ &= \frac{1}{(S_{xx})^2}\left(\sum_{i=1}^{n}(x_i-\overline{x})^2 E\left(\varepsilon_i^2\right) + 2\sum_{i\neq j}(x_i-\overline{x})(x_j-\overline{x})E(\varepsilon_i\varepsilon_j)\right) \end{aligned}$$

Como los e_i *son independientes tiene media cero entonces* $E(\varepsilon_i\varepsilon_j) = E(\varepsilon_i)E(\varepsilon_j) = 0$ *y* $E\left(\varepsilon_i^2\right) = Var(\varepsilon_i) = \sigma^2$. *Por lo tanto*

$$Var(B) = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2\sigma^2}{(S_{xx})^2} = \frac{\sigma^2}{S_{xx}}$$

3. *Se tiene que*

$$\begin{aligned} Var(A) &= E\left((A-\alpha)^2\right) = E\left(\left(\overline{Y}-B\overline{x}-\alpha\right)^2\right) = E\left(\left(\frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i+\alpha+\beta x_i) - B\overline{x} - \alpha\right)^2\right) \\ &= E\left(\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i + \alpha + \beta\overline{x} - B\overline{x} - \alpha\right)^2\right) = E\left(\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i + \overline{x}(\beta - B)\right)^2\right) \\ &= E\left(\frac{1}{n^2}\left(\sum_{i=1}^{n}\varepsilon_i\right)^2 + 2\overline{x}(\beta-B)\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i + \overline{x}^2(\beta-B)^2\right) \\ &= E\left(\frac{1}{n^2}\left(\sum_{i=1}^{n}\varepsilon_i^2 + 2\sum_{i\neq j}\varepsilon_i\varepsilon_j\right) + 2\overline{x}(\beta-B)\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i + \overline{x}^2(\beta-B)^2\right) \\ &= \frac{\sigma^2}{n} + \frac{2}{n}\overline{x}E\left((\beta-B)\sum_{i=1}^{n}\varepsilon_i\right) + \overline{x}^2E\left((\beta-B)^2\right) \end{aligned}$$

Dado que $B-\beta=\dfrac{\sum_{i=1}^{n}(x_i-\overline{x})\,\varepsilon_i}{S_{xx}}$ *entonces*

$$\begin{aligned} E\left((\beta-B)\sum_{i=1}^{n}\varepsilon_i\right) &= \frac{1}{S_{xx}}E\Big(\underbrace{\left(\sum_{i=1}^{n}(x_i-\overline{x})\,\varepsilon_i\right)\sum_{i=1}^{n}\varepsilon_i}_{\substack{\text{Sumados con factores } \varepsilon_i\varepsilon_j \\ \text{tienen esperanza } 0,\ \text{excepto si } i=j}}\Big) \\ &= \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})\,E\left(\varepsilon_i^2\right)=\frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})\,\sigma^2 \\ &= \frac{\sigma^2}{S_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})=0 \end{aligned}$$

por lo tanto

$$\begin{aligned} Var(A) &= \frac{\sigma^2}{n}+\overline{x}^2E\left((\beta-B)^2\right)=\frac{\sigma^2}{n}+\overline{x}^2Var(B) \\ &= \frac{\sigma^2}{n}+\overline{x}^2\frac{\sigma^2}{S_{xx}}=\sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right) \end{aligned}$$

Lo cual, realizan la suma, es igual a

$$Var(A)=\sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)=\frac{\sigma^2\sum_{i=1}^{n}x_i^2}{nS_{xx}}.$$

4. *Note que B es combinación lineal de las variables $Y_{x_1}, Y_{x_2}, ..., Y_{x_n}$ las cuales son normales e independientes por lo tanto B sigue una distribución normal.* ■

Las hipótesis del teorema anterior se llamarán en adelante **hipótesis de regresión** y los resultados siguientes dependerán de estás hipótesis. Si bien el teorema anterior establece la normalidad de los estimadores A y B, el problema es que esta normalidad depende de σ el cual se desconoce pero se puede establecer un estimador para este parámetro.

Teorema 45 *Bajo las hipótesis de regresión se tiene que un estimador insesgado de $\sigma^2=Var(Y_{x_i})$ es*

$$S^2=\frac{\sum_{i=1}^{n}(E_i)^2}{n-2}$$

Por lo que una estimación de σ^2 es

$$s^2=\frac{\sum_{i=1}^{n}(e_i)^2}{n-2}=\frac{SCE}{n-2}=\frac{S_{yy}-bS_{xy}}{n-2}$$ [13]

[13] Note que al minimizar SCE se minimiza el estimador de la varianza de Y_x, lo cual logra Y_{x_i} sea un buen estimador de su media $\alpha+\beta x_i$

Prueba. *Note que* $E\left(\sum_{i=1}^{n}(E_i)^2\right)$ *es igual a*

$$\begin{aligned}
& E\left(\sum_{i=1}^{n}\left(Y_{x_i}-A-Bx_i\right)^2\right)\\
= & \sum_{i=1}^{n}E\left(\left((\alpha-A)+(\beta-B)\,x_i+\varepsilon_i\right)^2\right)\\
= & \sum_{i=1}^{n}E\left((\alpha-A)^2+(\beta-B)^2\,x_i^2+\varepsilon_i^2+2\,(\alpha-A)\,(\beta-B)\,x_i+2\,(\beta-B)\,x_i\varepsilon_i+2\,(\alpha-A)\,\varepsilon_i\right)\\
= & \sum_{i=1}^{n}Var\,(A)+x_i^2Var\,(B)+E\left(\varepsilon_i^2\right)+2x_iE\left((\alpha-A)\,(\beta-B)\right)+2x_iE\left((\beta-B)\,\varepsilon_i\right)\\
& +2E\left((\alpha-A)\,\varepsilon_i\right))
\end{aligned}$$

La esperanza $E\left((\alpha-A)\,(\beta-B)\right)$ *es llamada la covarianza de* A *y* B*, la cual se denota* $cov\,(A,B)$*. Como ejercicio demuestre que*

$$cov\,(A,B)=-\overline{x}Var\,(B)$$

Además, note que

$$E\left((\beta-B)\,\varepsilon_i\right)=E\left(-\frac{\sum_{j=1}^{n}(x_j-\overline{x})\,\varepsilon_j}{S_{xx}}\varepsilon_i\right)=-\frac{(x_i-\overline{x})\,E\left(\varepsilon_i^2\right)}{S_{xx}}=-\frac{(x_i-\overline{x})\,\sigma^2}{S_{xx}}$$

y también

$$\begin{aligned}
E\left((\alpha-A)\,\varepsilon_i\right) &= E\left(\left(\alpha-\overline{Y}-B\overline{x}\right)\varepsilon_i\right)=E\left(\left(\alpha-\frac{1}{n}\sum_{i=1}^{n}\left(\varepsilon_i+\alpha+\beta x_i\right)-B\overline{x}\right)\varepsilon_i\right)\\
&= E\left(\left(\alpha-\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i-\alpha-\beta\overline{x}-B\overline{x}\right)\varepsilon_i\right)\\
&= E\left(\left(-\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i-\beta\overline{x}-B\overline{x}\right)\varepsilon_i\right)=-\frac{1}{n}E\left(\varepsilon_i^2\right)-\overline{x}E\left((\beta+B)\,\varepsilon_i\right)\\
&= -\frac{1}{n}\sigma^2-\overline{x}E\left(B\varepsilon_i\right)=-\frac{1}{n}\sigma^2-\overline{x}E\left((B-\beta)\,\varepsilon_i+\beta\varepsilon_i\right)\\
&= -\frac{1}{n}\sigma^2-\overline{x}\frac{(x_i-\overline{x})\,\sigma^2}{S_{xx}}.
\end{aligned}$$

Por lo tanto $E\left(\sum_{i=1}^{n}(E_i)^2\right)$ *es igual a*

$$\begin{aligned}
&\sum_{i=1}^{n}\left(Var(A)+x_i^2Var(B)+\sigma^2+-2x_i\overline{x}Var(B)-2x_i\frac{(x_i-\overline{x})\sigma^2}{S_{xx}}+-\frac{2}{n}\sigma^2\right.\\
&\left.-2\overline{x}\frac{(x_i-\overline{x})\sigma^2}{S_{xx}}\right)\\
=\;& Var(A)n+Var(B)\sum_{i=1}^{n}x_i^2+n\sigma^2+-2n\overline{x}^2Var(B)-2\frac{\sigma^2}{S_{xx}}\sum_{i=1}^{n}x_i(x_i-\overline{x})\\
&-2\sigma^2-2\overline{x}\frac{\sigma^2}{S_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})\\
=\;& Var(A)n+Var(B)\sum_{i=1}^{n}x_i^2+n\sigma^2+-2n\overline{x}^2Var(B)-2\frac{\sigma^2}{S_{xx}}\left(\sum_{i=1}^{n}x_i^2-n\overline{x}^2\right)+-2\sigma^2
\end{aligned}$$

Dado que $Var(B)=\frac{\sigma^2}{S_{xx}}$ *y* $Var(A)=\frac{\sigma^2\sum_{i=1}^{n}x_i^2}{nS_{xx}}=\frac{Var(B)}{n}\sum_{i=1}^{n}x_i^2$, *entonces* $E\left(\sum_{i=1}^{n}(E_i)^2\right)$ *es igual a*

$$\begin{aligned}
&Var(A)n+Var(B)\sum_{i=1}^{n}x_i^2+n\sigma^2+-2n\overline{x}^2Var(B)-2\frac{\sigma^2}{S_{xx}}\left(\sum_{i=1}^{n}x_i^2-n\overline{x}^2\right)+-2\sigma^2\\
=\;& 2Var(B)\sum_{i=1}^{n}x_i^2+n\sigma^2+-2n\overline{x}^2Var(B)-2Var(B)\left(\sum_{i=1}^{n}x_i^2-n\overline{x}^2\right)+-2\sigma^2\\
=\;& (n-2)\sigma^2
\end{aligned}$$

Por lo tanto

$$E\left(S^2\right)=\frac{1}{n-2}\cdot E\left(\sum_{i=1}^{n}(E_i)^2\right)=\sigma^2. \qquad \blacksquare$$

El siguiente teorema establece la distribución a utilizar en la estimación por intervalo de α y β. El resultado se obtiene de los dos teoremas anteriores y de resultados ya vistos [14].

[14]

Teorema 46 *Considere la población dada por la variable aleatoria X que sigue una distribución normal con varianza poblacional* σ^2. *Dada una muestra aleatoria de esta población* $(X_1, X_2, X_3, ..., X_n)$, *la variable*

$$\chi^2=\frac{(n-1)S^2}{\sigma^2}$$

sigue una distribución χ^2 *con* $v=n-1$ *grados de libertad. en sentido si existe una relación lineal*

Teorema 47 *Si* $Z\sim N(0,1)$ *y* $V\sim\chi^2(v)$ *son dos variables independientes entonces*

$$T=\frac{Z}{\sqrt{V/v}}$$

Teorema 48 *Bajo las hipótesis de regresión se tiene que*

$$T = \frac{(B-\beta)\sqrt{S_{xx}}}{S}$$

sigue una distribución t con $n-2$ grados de libertad. Y también

$$T = \frac{A-\alpha}{S\sqrt{\frac{\sum x^2}{nS_{xx}}}}$$

sigue una distribución t con $n-2$ grados de libertad.

Prueba. *Dado que $Y_{x_i} \sim N\left(\alpha+\beta x_i, \sigma^2\right)$ y un estimador de σ^2 es S^2 entonces por el teorema (46)*

$$\chi^2 = \frac{(n-2)\,S^2}{\sigma^2} \quad (*)$$

sigue una distribución χ^2 con $v = n-2$ grados de libertad. Además como $B \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$ entonces

$$\frac{(B-\beta)}{\frac{\sigma^2}{S_{xx}}} = \frac{(B-\beta)\,S_{xx}}{\sigma^2} \sim N(0,1) \qquad (**)$$

De $(*)$ y $(**)$ por el teorema (47)[:]

$$T = \frac{\frac{(B-\beta)\,S_{xx}}{\sigma^2}}{\sqrt{\frac{(n-2)\,S^2}{\sigma^2\,(n-2)}}} = \frac{(B-\beta)\sqrt{S_{xx}}}{S}$$

sigue una distribución t con $n-2$ grados de libertad. La segunda parte se deja de ejercicio. ■

Los resultados anteriores junto con las hipótesis de regresión, nos permiten establecer intervalos de confianza para los parámetros α yβ.

Teorema 49 *Bajo las hipótesis de regresión, se tiene que IC de $(1-\delta)\,100\%$ para*

1. *α tiene extremos $a \pm t_{\delta/2,n-2}\left(s\sqrt{\frac{\sum x^2}{nS_{xx}}}\right)$.*

2. *β tiene extremos $b \pm t_{\delta/2,n-2}\left(s\sqrt{\frac{1}{S_{xx}}}\right)$.*

sigue una distribución t con v grados de libertad

Prueba. *Ejercicio.* ■

Ejemplo 90 *Un estudiante de computación, que tiene una pequeña empresa de elaboración de software, llevó a cabo un estudio para determinar la relación entre el tiempo que tarda elaborando un software en días (X) y el ingreso obtenido por su venta en dólares (Y). Se recolectaron los valores de estas variables para 14 software, obteniendo los siguientes datos:*

$$\sum x = 323 \qquad \sum y = 16040 \qquad \sum x^2 = 11871$$
$$\sum y^2 = 29162000 \qquad \sum xy = 587890$$

Asuma las hipótesis de regresión. En un ejemplo anterior se determinó que la ecuación de regresión lineal para el ingreso como función del tiempo de elaboración de un software es

$$\widehat{y} = 49.293\,5x + 8.442\,82$$

1. *Determine un intervalo de confianza del 90% para el valor promedio del ingreso fijo (ingreso que no depende del tiempo de elaboración)*

   De acuerdo a los datos, se tiene que

   $$\begin{array}{cc} \delta = 0.1 & S_{xy} = 217824.2857 \\ n = 14 & SCE = 47429.80037 \\ S_{xx} = 4418.928571 & s^2 = 3952.483364 \\ S_{yy} = 10784742.86 & s = 62.86877893 \end{array}$$
   $$t_{\delta/2,n-2} = t_{0.05,12} = -1.78229$$

   Note que $a = 8.442\,82$, entonces un IC para α tiene extremos

   $$a + t_{\delta/2,n-2}\left(s\sqrt{\frac{\sum x^2}{nS_{xx}}}\right) \approx 0.346\,416.$$
   $$a - t_{\delta/2,n-2}\left(s\sqrt{\frac{\sum x^2}{nS_{xx}}}\right) \approx 16.539\,2.$$

   Por lo tanto un IC del 90% para el valor promedio del ingreso fijo es

   $$]0.346\,416, 16.539\,2[$$

2. *Determine un intervalo de confianza del 90% para el aumento promedio del ingreso por día de elabolación*

   Note que $b = 49.293\,5$. Un IC para β tiene extremos

   $$b + t_{\delta/2,n-2}\left(s\sqrt{\frac{1}{S_{xx}}}\right) \approx 47.607\,9.$$
   $$b - t_{\delta/2,n-2}\left(s\sqrt{\frac{1}{S_{xx}}}\right) \approx 50.979\,1.$$

Por lo tanto un IC del 90% *para el aumento promedio del ingreso en dólares por día de elabolación es*

$$]47.\,607\,9, 50.\,979\,1[$$

Ejercicio 45 *Se solicitó a una muestra de diez estudiantes que viene a la Universidad en bicicleta, que registraran la distancia recorrida y el tiempo que utilizarán para llegar a la universidad el próximo lunes. Los datos reunidos se muestran en la siguiente tabla:*

Recorrido en km (x)	*Tiempo en minutos* (y)
1	5
3	10
5	15
5	20
7	15
7	25
8	20
10	25
10	35
12	35

1. *Determine la ecuación de mínimos cuadrados para el tiempo como función del recorrido.R/* $\widehat{y} = 2.6641x + 2.3842$
2. *Suponga que se cumplen las hipótesis de regresión. Encuentre un intervalo de confianza de 98% para el aumento del tiempo esperado por kilométro recorrido. R/* $]1.485590925, 3.842609075[$

2.4 Pruebas de hipótesis de la relación lineal entre el predictor y el resultado

Muchas veces es importante determinar si la regresión lineal es significativa, es decir si en realidad hay una relación lineal entre la variable predictor (X) y la variable resultado (μ_Y), para ello considere el parámetro β. Recuerde que la ecuación de regresión es

$$\mu_{Y\,x} = \alpha + \beta x$$

Note que si β es 0 entonces no existe relación lineal entre X y Y.

Así, una manera de determinar si la regresión lineal es significativa es por medio de la prueba de hipótesis:

$$\begin{aligned} H_0 \text{ (no existe relación lineal entre } X \text{ y } Y) &: \beta = 0 \\ H_1 \text{ (si existe relación lineal entre } X \text{ y } Y) &: \beta \neq 0 \end{aligned}$$

Esta prueba es un caso particular de la prueba

$$H_0 : \beta = \beta_0$$

la cual para ser contrastada, se utiliza el hecho que, bajo H_0, por el teorema (48) se tiene que

$$T = \frac{(B - \beta_0)\sqrt{S_{xx}}}{S}$$

sigue una distribución t con $n-2$ grados de libertad. Así, a un nivel de significancia δ, se tiene que el valor observado de T es

$$t_{obs} = \frac{(b - \beta_0)\sqrt{S_{xx}}}{s}$$

y los posibles casos de la prueba son :

Hipótesis alternativa	Valores críticos	Valor P
$H_1 : \beta < \beta_0$	$t_c = t_{\delta/2,n-2}$	$P(T < t_{obs})$
$H_1 : \beta > \beta_0$	$t_c = t_{1-\delta/2,n-2}$	$P(T > t_{obs})$
$H_1 : \beta \neq \beta_0$	$t_{c1} = t_{\delta/2,n-2}$ $t_{c2} = t_{1-\delta/2,n-2}$	$2 \cdot P(T > \lvert t_{obs} \rvert)$

Ejemplo 91 *Un estudiante de computación, que tiene una pequeña empresa de elaboración de software, llevó a cabo un estudio para determinar la relación entre el tiempo que tarda elaborando un software en días (x) y el ingreso obtenido por su venta en dólares (y). Se recolectaron los valores de estas variables para 14 software, obteniendo los siguientes datos:*

$$\sum x = 323 \qquad \sum y = 16040 \qquad \sum x^2 = 11871$$
$$\sum y^2 = 29162000 \qquad \sum xy = 587890$$

Asuma las hipótesis de regresión. En ejemplos anteriores se determinó que la ecuación de regresión lineal para el ingreso como función del tiempo de elaboración de un software es

$$\widehat{y} = 49.293\,5x + 8.442\,82$$

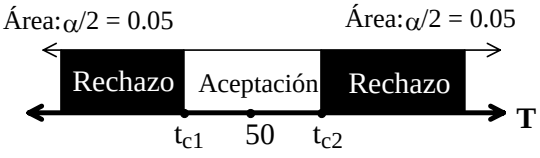
y además

$$S_{xx} = 4418.928571 \quad y \quad s = 62.86877893$$

Con un nivel de signicancia del 10%*, ¿Hay evidencia de que el aumento promedio del ingreso por día de elabolación sea de* 50 *dólares por día?*

Se tiene que

$H_0 : \beta = 50$
$H_1 : \beta \neq 50$



Los valores críticos son

$$t_{c1} = t_{0.05,12} = -1.78229 \quad y \quad t_{c2} = t_{0.95,12} = 1.78229$$

Así, la región de aceptación para el estadístico t es $A =]-1.78229, 1.78229[$. El valor t observado es

$$t_{obs} = \frac{(b-\beta_0)\sqrt{S_{xx}}}{s} = \frac{(49.2935-50)\sqrt{4418.928571}}{62.86877893} \approx -0.747026$$

Como $t_{obs} \in A$, se acepta H_0. No hay evidencia en contra de que $\beta = 50$.

Ejercicio 46 *Considere el ejercicio (45), donde se pregunto a una muestra de diez estudiantes, que viajan en bicicleta, la distancia recorrida (x) y el tiempo que utilizaron para llegar a la universidad (y). Según los datos, ¿Hay evidencia de que exista una relación lineal entre las variables x, y ? R/ $t_{obs} = 6.541293, v = 8$, Valor $P < 0.01$. No hay evidencia en contra de la relación lineal.*

2.5 Intervalos de predicción

Note $\widehat{y} = a + bx_i$ es una estimación del parámetro $E(Y_{x_i})$, donde x_i no necesariamente debe estar en los datos muestrales. En otros palabras, dada la muestra de puntos $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, se tiene un valor de $X : x_0$ que no se halla en la muestra, para el cual se quiere predecir el valor esperado de Y_{x_0}, entonces una **estimación puntual de** $E(Y_{x_0})$ es

$$\widehat{y} = a + bx_0$$

¿Y cómo logramos una estimación por intervalo de $E(Y_{x_0})$? El siguiente teorema nos brinda los ingredientes para responder esta pregunta.

Teorema 50 *Suponga que se cumplen las hipótesis de regresión. Considere la variable*

$$\widehat{Y}_0 = E(Y_{x_0}) = A + Bx_0$$

Se tiene que

1. *$\widehat{Y}_0$ es un estimador insesgado de $\widehat{y} = a + bx_0$*
2. *La varianza de $\widehat{Y}_0$ es*
$$\sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right)$$
3. *$\widehat{Y}_0$ sigue una distribución normal.*
4. *La variable*
$$T = \frac{\widehat{Y} - (a + bx_0)}{S\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}}$$
sigue una distribución t con $n-2$ grados de libertad.

Prueba. *Se probará la #2, las demás quedan de ejercicio. Note que*

$$\begin{aligned} Var\left(\widehat{Y}\right) &= Var\left(A+Bx_0\right) = Var\left(\overline{Y}-B\overline{x}+Bx_0\right) = Var\left(\overline{Y}-B\left(\overline{x}-x_0\right)\right) \\ &= Var\left(\overline{Y}\right)+\left(\overline{x}-x_0\right)^2 Var\left(B\right) == \frac{\sigma^2}{n}+\left(\overline{x}-x_0\right)^2\frac{\sigma^2}{S_{xx}} \qquad \blacksquare \end{aligned}$$

El teorema anterior permite establecer una estimación por intervalo de $E\left(Y_{x_0}\right)$, la cual se enuncia en el siguiente teorema.

Teorema 51 *Bajo las hipótesis de regresión, se tiene que IC de* $\left(1-\delta\right)100\%$ *para* $E\left(Y_{x_0}\right)$ *tiene extremos*

$$\underbrace{\left(a+bx_0\right)}_{\widehat{y}_0} \pm t_{\delta/2,n-2}\left(s\sqrt{\frac{1}{n}+\frac{\left(x_0-\overline{x}\right)^2}{S_{xx}}}\right).$$

Prueba. *Sea* $s\sqrt{\frac{1}{n}+\frac{\left(x_0-\overline{x}\right)^2}{S_{xx}}} = \overline{s}$ *y* $T = \frac{\widehat{Y}-\left(a+bx_0\right)}{\overline{s}}$. *Por el teorema anterior*

$$\begin{aligned} & P\left(\left(a+bx_0\right)+t_{\delta/2,n-2}\left(\overline{s}\right) < \widehat{Y}_0 < \left(a+bx_0\right)-t_{\delta/2,n-2}\left(\overline{s}\right)\right) \\ = \; & P\left(t_{\delta/2,n-2} < T < -t_{\delta/2,n-2}\right) \\ = \; & 1-\left(\frac{\delta}{2}+\frac{\delta}{2}\right) = 1-\delta \qquad \blacksquare \end{aligned}$$

Por otro lado, dado un valor x_0 que no se halla en la muestra tomada, suele ser más importante estimar el valor desconocido y_0. El problema es que y_0 no es un parámetro de la distribución, por lo tanto en lugar de obtener un IC para y_0 se obtiene un **intervalo de predicción** de y_0. Para determinar este intervalo se utilizarán los resultados del siguiente teorema.

Teorema 52 *Considere para el valor* x_0 *la variable aleatoria residuo o error*

$$Y_{x_0}-\widehat{Y}_0$$

Bajo las hipótesis de regresión se tiene que

1. $Y_{x_0}-\widehat{Y}_0 = \alpha+\beta x_0+\varepsilon_0-\left(A+Bx_0\right).$
2. $E\left(Y_{x_0}-\widehat{Y}_0\right)=0$
3. $Var\left(Y_{x_0}-\widehat{Y}_0\right) = \sigma^2\left(1+\frac{1}{n}+\frac{\left(x_0-\overline{x}\right)^2}{S_{xx}}\right)$ *dado que* $\widehat{Y}_0, Y_{x_0}$ *son independientes*
4. $Y_{x_0}-\widehat{Y}_0$ *sigue una distribución normal:*

$$Y_{x_0}-\widehat{Y}_0 \sim N\left(0,\sigma^2\left(1+\frac{1}{n}+\frac{\left(x_0-\overline{x}\right)^2}{S_{xx}}\right)\right)$$

5. *La variable*

$$T = \frac{Y_{x_0} - \widehat{Y}_0}{S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}}$$

sigue una distribución t con $n-2$ grados de libertad.

6. *Se tiene que*

$$P\left(\widehat{Y}_0 - t_{\alpha/2,n-2}\ S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}} < Y_{x_0} < \widehat{Y}_0 + t_{\alpha/2,n-2}\ S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}\right) = 1-\delta$$

Prueba. *Ejercicio.* ■

El último resultado nos permite definir un intervalo de predicción para y_0.

Definición 36 *Bajo las hipótesis de regresión, se tiene que un intervalo de predicción (IP) de $(1-\delta)\,100\%$ para el valor desconocido y_0 asociado a x_0 tiene extremos*

$$\widehat{y}_0 \pm t_{\delta/2,n-2}\ s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}.$$

Note que el IP para y_0 tiene mayor ancho que el IC de $E(Y_{x_0})$. Esto se debe a que IP depende del error del modelo y del error asociado a la estimación del valor desconocido y_0.

Ejemplo 92 *Un estudiante de computación, que tiene una pequeña empresa de elaboración de software, llevó a cabo un estudio para determinar la relación entre el tiempo que tarda elaborando un software en días (X) y el ingreso obtenido por su venta en dólares (Y). Se recolectaron los valores de estas variables para 14 software, obteniendo los siguientes datos:*

$$\sum x = 323 \qquad \sum y = 16040 \qquad \sum x^2 = 11871$$
$$\sum y^2 = 29162000 \qquad \sum xy = 587890$$

Asuma las hipótesis de regresión. En un ejemplo anterior se determinó que la ecuación de regresión lineal para el ingreso como función del tiempo de elaboración de un software es

$$y = 49.\,293\,5x + 8.\,442\,82$$

1. *Determine un intervalo de confianza del 90% para el ingreso esperado por venta de software con un tiempo de elaboración de 30 días.*

   De acuerdo a los datos, se tiene que

$$\begin{array}{cc} \delta = 0.1 & S_{xy} = 217824.2857 \\ n = 14 & SCE = 47429.80037 \\ \sum x = 323 & s^2 = 3952.483364 \\ S_{xx} = 4418.928571 & s = 62.86877893 \\ S_{yy} = 10784742.86 & t_{\delta/2,n-2} = -1.78229 \end{array}$$

Note que $x_0 = 30 \Longrightarrow \widehat{y}_0 = ax_0 + b = 1487.25$. *Por lo tanto, un IC para* $E(Y_{x_0})$ *tiene extremos*

$$\widehat{y}_0 \pm t_{\delta/2,n-2}\left(s\sqrt{\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}}\right) \approx 1455.11$$

$$\widehat{y}_0 \pm t_{\delta/2,n-2}\left(s\sqrt{\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}}\right) \approx 1519.39$$

Por lo tanto un IC del 90% *para el ingreso esperado por venta de software con un tiempo de elaboración de 30 días es*

$$IC \text{ para } E(y_0) :]1455.11, 1519.39[\,.$$

2. *Determine un intervalo de predicción del* 90% *para el ingreso por venta de software con un tiempo de elaboración de 30 días.*

   Se tiene que $x_0 = 30$ *y* $\widehat{y}_0 = 1487.25$, *por lo tanto el IP tiene extremos*

$$\widehat{y}_0 + t_{\delta/2,n-2}\ s\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}} \approx 1370.68$$

$$\widehat{y}_0 - t_{\delta/2,n-2}\ s\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}} \approx 1603.82$$

   Así, un intervalo de predicción del 90% *para el ingreso por venta de software con un tiempo de elaboración de 30 días es*

$$IC \text{ para } y_0 :]1370.68,\quad 1603.82[\,.$$

Ejercicio 47 *Un comerciante al menudeo llevó a cabo un estudio para determinar la relación entre los gastos de publicidad semanal* (X)*y las ventas* (Y)*. Ambas. Se recolectaron los valores de estas variables en dólares durante doce semanas, obteniendo los siguientes datos:*

$$\sum x = 430 \qquad \sum y = 5181 \qquad \sum x^2 = 16902$$
$$\sum y^2 = 2276541 \qquad \sum xy = 193230$$

Asuma las hipótesis de regresión.

1. *Encuentre la ecuación de regresión lineal para las ventas semanales como función de los gastos de publicidad* $R/\quad b = 5.073086365,\quad a = 249.9644053$

2. *¿Aproximadamente, cuánto es el valor promedio de las ventas semanales si no se realizan gastos en publicidad?* $R/\quad 249.9644053$

3. *Determine un intervalo del* 95% *para la venta semanal esperada con gastos de* \$35 *en publicidad.* $R/\quad]420,4481621,\quad 434.596694[$

4. *Determine un intervalo del* 95% *para las ventas semanales con gastos de* \$35 *en publicidad.* $R/\quad]402.0811977,\quad 452.9636583[$

5. *Pruebe con un nivel de significancia de 5%, si las ventas semanales se incrementa en promedio* $\$4,5$ *por cada dólar gastado en publicidad.* $R/$ *Valor* $P = 0,071049808$, *si hay evidencia de que el incremento es de* \$4.5

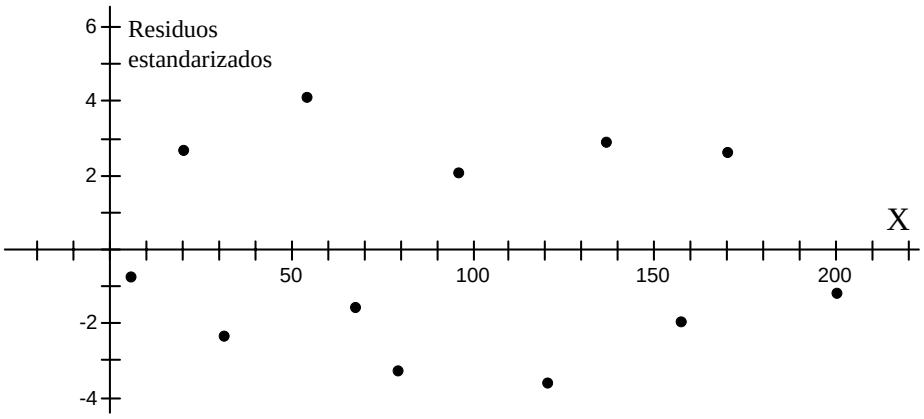
2.6 Coeficiente de correlación

Empíricamente, una condición visual necesaria para que exista una buena regresión lineal se obtiene por medio del **Diagrama de Residuos.** Este gráfico consiste en un diagrama de dispersión formado por los puntos $\left(x_i, \frac{e_i}{\sigma_e}\right)$ para $i = 1, 2, ..., n$, donde $\frac{e_i}{\sigma_e}$ son los residuos estandarizados, es decir los residuos divididos por su desviación estándar

$$\sigma_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}}$$

Esta fórmula de σ_e se debe a que la media de los residuos es cero. Dado que los residuos estandarizados tienen media cero y varianza 1 entonces los puntos del Diagrama de Residuos se encuentran alrededor del eje X. Si la regresión lineal es buena, los residuos estandarizados deben distribuirse uniformemente concentrándose alrededor del eje X, empíricamente se dice que debe estar entre las rectas $y = 3, y = -3$.

Ejemplo 93 *Dadas dos variables X y Y, suponga que se realiza los cálculos para estimar los coeficientes de regresión lineal de Y como función de X y se obtiene el siguiente Diagrama de Residuos:*



De acuerdo a este diagrama, ¿Considera que existe una buena regresión lineal entre las variables?

No, dado que algunos puntos están distanciados del eje X, pues se sale de la banda $[-3, 3]$. Por lo tanto no se considera que exista una buena regresión de las variables.

Sin embargo, una manera más objetiva de medir la calidad de la regresión lineal es por medio del coeficiente de correlación.

Definición 37 *Sean X, Y dos variables cuantitativas continuas. Se define* ***el coeficiente de correlación poblaciónal*** *al como el parámetro*

$$\rho = \beta \frac{\sigma_x}{\sigma_y}$$

Note que

1. Si $\rho = 0$ entonces $\beta = 0$, asumiendo que la variable X no es constante. Así si $\rho \approx 0$ la recta $y = \alpha + \beta x$ es casi horizontal por lo que si X varía, la recta no explica la variación de Y. Así, no hay una buena regresión lineal.

2. Por otro lado note que dado que

$$\rho^2 = \beta^2 \frac{\sigma_x^2}{\sigma_y^2} = 1 - \frac{\sigma_y^2 - \beta^2 \sigma_x^2}{\sigma_y^2}. \qquad (*)$$

   Si $\rho = \pm 1$, se obtiene que $\rho^2 = 1$ y sustituyen en $(*)$ se sigue que

$$\sigma_y^2 = \beta^2 \sigma_x^2 \Longrightarrow Var(Y) = Var(\beta X) = Var(\alpha + \beta X),$$

   esto indica que la variación de Y es dada por la variación de la recta, y entonces hay una fuerte relación lineal entre X y Y. El siguiente teorema brinda una estimación puntual de ρ.

Definición 38 *Sean X, Y dos variables cuantitativas continuas. Se define* ***el coeficiente de correlación muestral*** *a la variable que asocia a cada muestra de n valores de estas variables: (x_i, y_i) con $i = 1, 2, 3, ..., n$, el valor:*

$$r = b \frac{s_x}{s_y}$$

donde $s_x^2 = \dfrac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$ *y* $s_y^2 = \dfrac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1}$.

Teorema 53 *Sean X, Y dos variables cuantitativas continuas. El coeficiente de correlación muestral para cada muestra de n valores de estas variables: (x_i, y_i) con $i = 1, 2, 3, ..., n$, es:*

$$r = b\sqrt{\frac{S_{xx}}{S_{yy}}} = b \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n(\overline{x})^2}{\sum_{i=1}^{n} y_i^2 - n(\overline{y})^2}} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n(\overline{x})^2\right)\left(\sum_{i=1}^{n} y_i^2 - n(\overline{y})^2\right)}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Interpretación del coeficiente de correlación muestral. Sean X, Y dos variables cuantitativas continuas. Sea r el coeficiente de correlación muestral, se tiene que:

1. **Si $r \approx 0$ no hay correlación lineal**. Dada que r es una estimación de ρ, se tiene que $\rho \approx 0$ y entonces, como se mencionó anteriormente, no hay una buena regresión lineal.. Además, como $r \approx 0$ es por qué $b \approx 0$ o $s_x \approx 0$. Si $b \approx 0$, la recta de mejor ajuste es casi constante y la recta no puede explicar las variaciones en Y debido a los valores de X. Si $s_x \approx 0$ los valores de X son muy similares por lo que no pueden explicar la variación de los valores de Y. Si $-0.2 \leq r \leq 0.2$ se considera que no existe relación lineal entre las variables.

2. **Si $r \approx \pm 1$ hay una buena correlación lineal**. En este caso $\rho \approx r \approx \pm 1$ y entonces hay una fuerte relación lineal entre X y Y.

(a) **Si $r \approx 1$ hay correlación lineal positiva**. Note que si r es cercano a 1 entonces b es positivo por lo tanto la recta de mejor ajuste es creciente. Así a mayores valores de X corresponden mayores valores de Y, y a menores valores de X corresponden menores valores Y. Si $r \geq 0.8$ se considera que existe una correlación positiva aceptable.

(b) **Si $r \approx -1$ hay correlación lineal negativa.** Note que si r es cercano a -1 entonces b es negativo por lo tanto la recta de mejor ajuste es decreciente. Así a mayores valores de X corresponden menores valores de Y, y viceversa. Si $r \leq -0.8$ se considera que existe una correlación negativa aceptable.

Definición 39 *Sean X, Y dos variables cuantitativas continuas. Sea r el coeficiente de correlación muestral, se define* ***el coeficiente de determinación muestral*** *como r^2, el cual es una estimación a la proporción de la variación de Y que puede explicarse por su relación lineal con X.*

Ejemplo 94 *Un estudiante de computación, que tiene una pequeña empresa de elaboración de software, llevó a cabo un estudio para determinar la relación entre el tiempo que tarda elaborando un software en días (X) y el ingreso obtenido por su venta en dólares (Y). Se recolectaron los valores de estas variables para 14 software, obteniendo los siguientes datos:*

$$\sum x = 323 \qquad \sum y = 16040 \qquad \sum x^2 = 11871$$
$$\sum y^2 = 29162000 \qquad \sum xy = 587890$$

Asuma las hipótesis de regresión. En un ejemplo anterior se determinó que la ecuación de regresión lineal para el ingreso como función del tiempo de elaboración de un software es

$$y = 49.\,293\,5x + 8.\,442\,82.$$

1. *Utilice el coeficiente de correlación para valorar la calidad de la regresión.*

   Dado que

$$\begin{aligned} b &= 49.\,293\,5 \\ S_{xx} &= \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 11871 - \frac{(323)^2}{14} \approx 4418.928571. \\ S_{yy} &= \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 29162000 - \frac{(16040)^2}{14} \approx 10784742.86. \end{aligned}$$

   Entonces

$$r = b\sqrt{\frac{S_{xx}}{S_{yy}}} \approx 49.\,293\,5\sqrt{\frac{4418.928571}{10784742.86}} \approx 0.997\,799$$

   Como $r \approx 1$, entonces existe una fierte correlación lineal positiva entre X y Y.

2. *¿Aproximadamente, que porcentaje de variación en los ingresos se debe a otros factores aparte del tiempo de elaboración?*

   Note que

$$1 - r^2 = 1 - (0.997\,799)^2 = 0.00439716.$$

Por lo tanto, el porcentaje de variación en el ingreso que se debe a otros factores aparte del tiempo de elaboración es cercano al 0.439716%. *Esto confirma que existe una fuerte correlación de las variables* X, Y.

Ejemplo 95 *Se espera que, por lo general, el número de horas de estudio* x *en la preparación de un examen tenga una correlación directa(positiva) con la calificación* y *alcanzada en tal examen. Se obtuvieron las horas de estudio así como las calificaciones obtenidas por diez estiudiantes seleccionados al azar de un grupo, los datos están resumidos en la siguiente tabla:*

$$\sum x = 118 \qquad \sum y = 591 \qquad \sum x^2 = 1648$$
$$\sum y^2 = 39013 \qquad \sum xy = 7956$$

La ecuación de mínimos cuadrados es $y = a + bx$ *con* $b = 3.842723005, \quad a = 13.75586854$. *¿Aproximadamente qué porcentaje de la variación en la calificación se debe a otros factores aparte del tiempo de estudio?*

El coeficiente de correlación muestral es (ejercicio):

$$r = 0.961233$$

por lo tanto $r^2 = 0.923969384 \Longrightarrow 1 - r^2 \approx 0.076\,030\,6$. *Aproximadamente el* 7.64% *de la variación en la calificación se debe a otros factores aparte del tiempo de estudio.*

Ejercicio 48 *Considere nuevamente el ejercicio: Un comerciante al menudeo llevó a cabo un estudio para determinar la relación entre los gastos de publicidad semanal* $(v)y$ *las ventas* (w). *Ambas. Se recolectaron los valores de estas variables en dólares durante doce semanas, obteniendo los siguientes datos:*

$$\sum v = 430 \qquad \sum w = 5181 \qquad \sum v^2 = 16902$$
$$\sum w^2 = 2276541 \qquad \sum vw = 193230$$

1. *¿Aproximadamente, que porcentaje de variación en las ventas semanales se debe a otros factores aparte de los gastos de publicidad?* $R/$ 0.030343318
2. *¿Considera importante invertir en publicidad? Explique.* $R/$ Si

2.7 Ejercicios

1. Considere los datos de la siguiente tabla:

$X:$	2	6	8	10	12	14	16
$Y:$	1	5	8	11	15	20	25

   (a) Encuentre la ecuación de regresión lineal para Y como función de X. $R/ \quad \widehat{y} = 1.7234x - 4.5984$

   (b) ¿Cuál es el valor aproximado esperado de Y cuando $X = 20$? $R/$ 29.86885246

   (c) Determine un IC del 90% para α $R/ \quad]-7.732636431, -1.464084881[$

   (d) Determine un IC del 95% para β $R/ \quad]1.349347231, 2.09737408[$

(e) ¿Considera que la recta $\mu_{Yx} = \alpha + \beta x$ es creciente e interseca al eje Y negativo?
$R/$ Si, de acuerdo a los IC anteriores.

(f) Halle un intervalo de confianza del 90% para el valor esperado de Y cuando $X = 15$.
$R/$ $]19.22383126, 23.2802671[$

(g) Encuentre un intervalo de predicción de 95% para Y cuando $X = 7$.
$R/$ $]2.63598165, 12.29434622[$

(h) Determine el coeficiente de correlación muestral y valore la regresión lineal realizada.
$R/$ $r \approx 0.982642941$, hay una fuerte correlación positiva

(i) Aproximadamente, ¿qué porcentaje de variación de Y se debe a otros factores a parte de la variación de X? $R/$ 3.4413%

(j) Pruebe si existe una relación lineal significativa entre las variables. $R/$ $t_{obs} = 11.8446005$, Valor $P < 0.01$. No hay evidencia en contra de que exista dicha relación.

(k) A un nivel de significancia del 10%, ¿existe evidencia en contra de que $\beta > 1.5$? $R/$ $t_c = 1.475884037, t_{obs} = 1.535150362$, no hay evidencia en contra de que $\beta > 1.5$

2. Considere los datos de la siguiente tabla:

$X:$	-2	0	5	9	15	20
$Y:$	40	35	30	25	20	15

(a) Encuentre la ecuación de regresión lineal para Y como función de X.
$R/$ $\widehat{y} = -1.0836x + 35.988$

(b) ¿Cuál es el valor aproximado esperado de Y cuando $X = 10$? $R/$ 25.15220354

(c) Determine un IC del 95% para α $R/$ $]33.90093052, 38.07544385[$

(d) Determine un IC del 90% para β $R/$ $]-1.228400414, -0.938796315[$

(e) ¿Considera que la recta $\mu_{Yx} = \alpha + \beta x$ es decreciente e interseca al eje Y positivo?
$R/$ Si,, por los IC anteriores.

(f) Halle un intervalo de confianza del 96% para el valor esperado de Y cuando $X = 10$.
$R/$ $]23.49967209, 26.804735[$

(g) Encuentre un intervalo de predicción de 90% para Y cuando $X = 1$.
$R/$ $]-4.660623038, 57.60340233[$

(h) Determine el coeficiente de correlación muestral y valore la regresión lineal realizada.
$R/$ $r \approx -0.992233095$, hay una fuerte correlación negativa

(i) Aproximadamente, ¿qué porcentaje de variación de Y se debe a otros factores a parte de la variación de X? $R/$ 1.5473486%

(j) Pruebe, a un nivel de significancia del 5%, si existe una relación lineal significativa entre las variables. $R/$ $t_{c1} = -2,776445105, t_{obs} = -15,95326648$. Hay evidencia de que si existe dicha relación.

(k) ¿existe evidencia en contra de que $\beta \leq -1$? $R/$ $t_{obs} = -1,230776114$, Valor $P \in]0.8, 0.9[$. Se puede considerar que $\beta \leq -1$

3. La tabla siguiente presenta las notas de aprobación en matemática y química de 10 estudiantes elegidos al azar entre un grupo muy numeroso de un colegio determinado , el objetivo es determinar si las notas de matemática dependen de los resultados obtenidos en química.

Matemática:	75	80	93	65	87	71	98	68	84	77
Química:	80	87	95	73	91	80	95	75	90	81

   (a) Encuentre la ecuación de regresión lineal para la nota de Química como función de los nota de Matemática. $R/ \quad \widehat{y} = 0.723x + 26.98$

   (b) ¿Aproximadamente, que porcentaje de variación en las notas de Química se debe a otros factores aparte de las notas de matemáticas? $R/ \quad 5.1\%$

4. Considere los datos de la siguiente tabla:

$X:$	-10	-7	-5	-3	-1
$Y:$	-20	-23	-21	-27	-26

   (a) Encuentre la ecuación de regresión lineal para Y como función de X. $R/ \quad \widehat{y} = -0.7254x - 27.172$

   (b) ¿Cuál es el valor aproximado esperado de Y cuando $X = 0$? $R/ \quad -27.17213115$

   (c) Determine un IC del 96% para α $R/ \quad]-33.09737553, -21.24688676[$

   (d) Determine un IC del 80% para β $R/ \quad]-1.184831089, -0.265988583[$

   (e) Halle un intervalo de confianza del 90% para el valor esperado de Y cuando $X = -6$. $R/ \quad]-24.94864753, -20.69069674[$

   (f) Encuentre un intervalo de predicción de 95% para Y cuando $X = -12$. $R/ \quad]-27.60638189, -9.328044343[$

   (g) Determine el coeficiente de correlación muestral y valore la regresión lineal realizada. $R/ \quad r \approx -0.830848578$, hay una leve correlación negativa

   (h) Aproximadamente, ¿qué porcentaje de variación de Y se debe a otros factores a parte de la variación de X? $R/ \quad 30.969064\%$

   (i) Pruebe, a un nivel de significancia del 10%, si existe una relación lineal significativa entre las variables. $R/ \quad t_{c1} = -2,353363435, t_{obs} = -2,585940147$. No hay evidencia en contra de que existe dicha relación.

   (j) A un nivel de significancia del 10%, ¿existe evidencia en contra de que $\beta = -1$? $R/ \quad t_{c1} = -2.353363435, t_{obs} = 0.9788587$. Se acepta H_0.

5. Considere los datos de la siguiente tabla:

$X:$	-10	-5	0	5	10	15	20	25
$Y:$	-10	-6	-9	-1	-2	1	15	2

   (a) Encuentre la ecuación de regresión lineal para Y como función de X. $R/ \quad \widehat{y} = 0.519x - 5.1429$

(b) ¿Cuál es el valor aproximado esperado de Y cuando $X = -3$? $R/ \quad -6.7$

(c) Determine un IC del 80% para α $R/ \quad]-8.251289861, -2.034424424[$

(d) Determine un IC del 90% para β $R/ \quad]0.212664306, 0.825430932[$

(e) Halle un intervalo de confianza del 99% para el valor esperado de Y cuando $X = 6$.
$R/ \quad]-8.782638599, 4.725495742[$

(f) Encuentre un intervalo de predicción de 98% para Y cuando $X = -6$.
$R/ \quad]-26.55406942, 10.0397837[$

(g) Determine el coeficiente de correlación muestral y valore la regresión lineal realizada.
$R/ \quad r \approx 0.802273437$, hay una leve correlación positiva

(h) Aproximadamente, ¿qué porcentaje de variación de Y se debe a otros factores a parte de la variación de X? $R/ \quad 35.6357332\%$

(i) Pruebe si existe una relación lineal significativa entre las variables. $R/ \quad t_{obs} = 3,291964844$, Valor $P < 0.02$. No hay evidencia en contra de que exista dicha relación.

(j) ¿existe evidencia en contra de que $\beta \geq 0.4$? $R/ \quad t_{obs} = 0.755037808$, Valor $P \in]0.6, 0.8[$. No hay evidencia en contra de que $\beta \geq 0.4$

6. Considere los datos de la siguiente tabla:

$X:$	3	6	9	12
$y:$	11	18	22	35

A partir de estos datos, estime el coeficiente β de la ecuación de regresión $\mu_{Yx} = 5+\beta x$ utilizando el método de mínimos cuadrados. $R/ \quad 2.255555556$

7. Considere los datos de la siguiente tabla:

$X:$	-8	-6	-4	-2	2
$Y:$	33	22	14	2	-5

A partir de estos datos, estime el coeficiente β de la ecuación de regresión $\mu_{Yx} = \beta x-7$ utilizando el método de mínimos cuadrados. $R/ \quad -4.774193548$

8. Considere los datos de la siguiente tabla:

$X:$	-11	-7	-2	8	20	30
$Y:$	-70	-50	-20	40	110	175

A partir de estos datos, estime el coeficiente β de la ecuación de regresión $\mu_{Yx} = 10 + 2\beta x$ utilizando el método de mínimos cuadrados. $R/ \quad 2.779583875$

9. Sean X, Y dos variables cuantitativas continuas. Dada una muestra de n valores de estas variables: (x_i, y_i) con $i = 1, 2, 3, ..., n$. Pruebe que la estimación puntual del coeficiente β de la ecuación de regresión $\mu_{Yx} = c + \beta x$, donde c es una constante conocida, por el método de mínimos cuadrados es

$$b = \frac{\sum_{i=1}^{n} x_i y_i - c \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}$$

10. Considere los datos de la siguiente tabla:

$X:$	2	5	7	11	15
$Y:$	-30	-75	-100	-170	-230

A partir de estos datos, estime el coeficiente β de la ecuación de regresión $\mu_{Yx} = \beta x$ utilizando el método de mínimos cuadrados. $R/ \quad -15.2240566$

11. Considere los datos de la siguiente tabla:

$X:$	-4	-3	-2	-1	0
$Y:$	2	0	-3	-4	-5

A partir de estos datos, estime el coeficiente β de la ecuación de regresión $\mu_{Yx} = 3\beta + \beta x$ utilizando el método de mínimos cuadrados. $R/ \quad -1.866666667$

12. Considere los datos de la siguiente tabla:

$X:$	2	8	20	35
$Y:$	15	55	140	245

A partir de estos datos, estime el coeficiente β de la ecuación de regresión $\mu_{Yx} = 3x + \beta x$ utilizando el método de mínimos cuadrados. $R/ \quad 3.996455995$

13. En la siguiente tabla se muestra las edades y la presión sanguínea de 9 personas:

Edad	56	42	72	36	63	47	49	68	60
Presión sanguínea	147	125	160	118	149	128	145	152	155

Suponiendo que se cumplen las hipótesis de regresión.

(a) Encuentre la ecuación de regresión lineal para la presión sanguínea como función de la edad.

$R/ \quad \widehat{y} = 80.092 + 1.13219x$

(b) ¿Aproximadamente, cuál es la presión sanguínea promedio de recién nacidos? $R/ \quad 80.092$

(c) ¿Aproximadamente, cuánto es el aumento promedio de la presión sanguínea por año de edad? $R/ \quad 1.13219$

(d) Encuentre un intervalo de predicción de 95% para la presión sanguínea cuando se tiene 45 años. $R/ \quad]116.324, 145.758[$

(e) Aproximadamente, ¿qué porcentaje de variación de la presión sanguínea se debe a otros factores a parte de la edad de las mujeres? $R/ \quad 13.09\%$

14. En la siguiente tabla se muestra la calificación obtenida (Y) por 10 personas en una prueba de admisión a una determinada universidad, el año pasado, y el número de horas (X) que durmió la noche previa al examen.

# de horas que durmió	2	3	4.5	5	5	5.5	6	6	7	8
nota obtenida	50	55	65	60	67	70	75	69	76	85

Suponiendo que se cumplen las hipótesis de regresión.

(a) Encuentre la ecuación de regresión lineal para la calificación de un estudiante como función del número de horas que durmió la noche previa al examen.. $R/ \quad \hat{y} = 37.7580 + 5.66192x$

(b) ¿Aproximadamente, cuál es la nota promedio de un estudiante que no durmió en la noche previa al examen.? $R/ \quad 37.7580$

(c) ¿Aproximadamente, cuánto es el aumento promedio de la calificación de un estudiante por hora que duerme en la noche previa al examen? $R/ \quad 5.66192$

(d) Encuentre un intervalo de predicción de 95% para la nota de los estudiantes que durmieron 4 horas en la noche previa al examen. ¿Se puede suponer que estos estudiantes reprobaron el examen (se aprueba con 70)? $R/ \quad 53,25663097 < \hat{y}_{!x=45} < 67,55475693$

(e) Aproximadamente, ¿qué porcentaje de variación de la calificación se debe a otros factores distintos al número de horas que duerme el estudiante? (2 puntos) $R/ \quad 6.9\%$.

15. Una fabrica recolectó la siguiente información de 8 de sus trabajadores sobre el número de minutos que llegaron tarde al trabajo (X) y el número de piezas defectuosas que fabricaron ese día (Y):

X	2	5	10	15	20	25	30	40
Y	2	4	9	12	15	20	24	30

Suponiendo que se cumplen las hipótesis de regresión.

(a) Encuentre la ecuación de regresión lineal para el número de piezas defectuosas que fabrica un empleado por día como función del número de minutos que llega tarde al trabajo. $R/ \quad \hat{y} = 0.732887615 + 0.749230606x$

(b) ¿Aproximadamente, ¿cuál es el número promedio de piezas defectuosas que realiza un empleado que llega puntual a su trabajo? $R/ \quad 0.732887615$ piezas en promedio

(c) ¿Aproximadamente, ¿cuánto es el aumento promedio en el número de piezas defectuosas que realiza un empleado por cada minuto que llega tarde al trabajo? $R/ \quad 0.749230606$ piezas por minuto

(d) Encuentre un intervalo de predicción de 95% para el número de piezas defectuosas que realiza un empleado en un día que llega 35 minutos tarde al trabajo. $R/ \quad 25.00546394 < \hat{y}_{x=35} < 28.90645371$

(e) Si un trabajo realiza más de 20 piezas defectuosas en un día será sancionado. ¿Considera probable que si un trabajador llega 35 minutos tarde al trabajo será sancionado? Justifique su respuesta. $R/ \quad$ Si

(f) Aproximadamente, ¿qué porcentaje de variación de en el número de piezas defectuosas que elabora un empleado en un día se debe a otros factores distintos al número de minutos que llega tarde a trabajar? $R/ \quad$ Aproximadamente el 0,42%.

(g) Un trabajador tiene un salario base diario de 20 mil colones del cuál se le descuenta 200 colones por cada pieza defectuosa que realiza. Es decir, el salario diario de un trabajador es $S = 20000 - 200Y$. Determine un intervalo de confianza del 90% para el salario diario promedio de un trabajador que llega 13 minutos tarde a su trabajo. $R/ \quad$ El IC para $E(S)$ es $]18008.15747 \quad , \quad 17802.68833[$

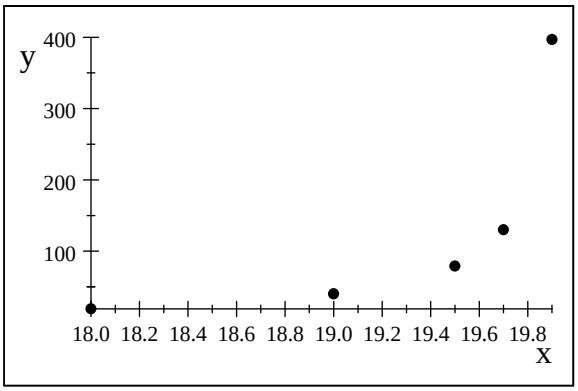
3 Regresión no lineal simple

¿Qué sucede si existe una fuerte relación entre dos variables cuantitativas y esta no es lineal? Veamos el siguiente ejemplo

Ejemplo 96 *Considere los datos muestrales para 2 variables x, y dados en la siguiente tabla:*

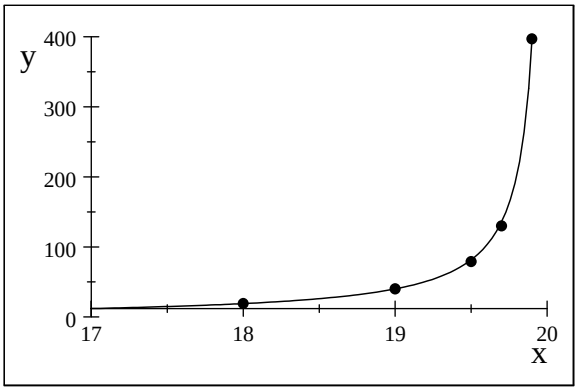
$x:$	18	19	19.5	19.7	19.9
$Y:$	19	40	79	130	397

Realizando el diagrama de dispersión se obtiene



Por el diagrama se descarta que exista una relación lineal entre las variables. Sin embargo se puede suponer que existe una relación hiperbólica (luego se justificará):

$$E\left(Y_x\right) = y = \frac{x}{\alpha x + \beta}$$



¿Como estimar los nuevos parámetros? Una forma es transformar el problema de regresión en lineal, para ello note que

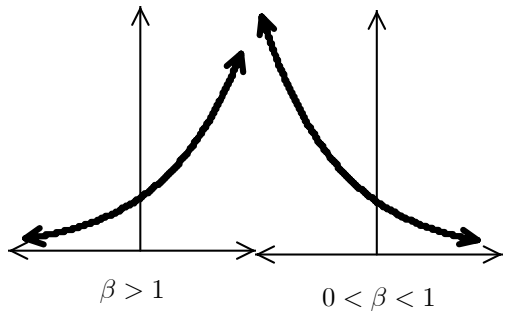
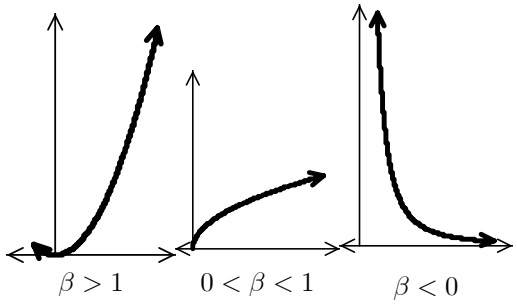
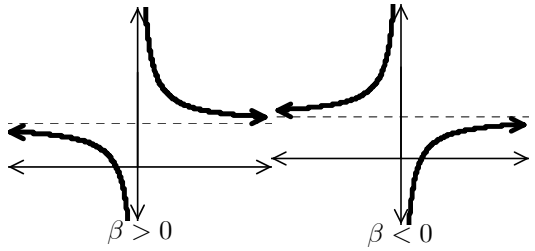
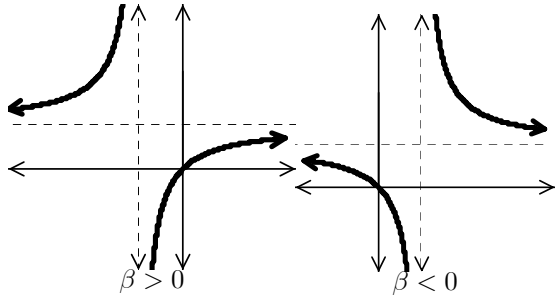
$$y = \frac{x}{\alpha x + \beta} \Longrightarrow \frac{1}{y} = \frac{\alpha x + \beta}{x} = \alpha + \beta\left(\frac{1}{x}\right)$$

Así si se define $y_1 = \dfrac{1}{y}$ *y* $x_1 = \dfrac{1}{x}$*, nuestro modelo de regresión se convierte en un modelo lineal*

$$y_1 = \alpha + \beta x_1$$

Donde a partir de los datos observados (x_i, y_i) *con* $i = 1, 2, 3, ..., n$, *se obtiene los datos muestrales para las variables* (x_1, y_1) : $\left(\frac{1}{x_i}, \frac{1}{y_i}\right)$ *con* $i = 1, 2, 3, ..., n$. *Estos nuevos datos permiten estimar los parámetros* α *y* β *con los métodos estudiados.*

Seguidamente se presenta los modelos de regresión más importante que se van a considerar, por simplicidad en la notación se denota $E(Y_x)$ por $y_.$. Se deja como ejercicio justificar la transformación de estos modelos al modelo lineal.

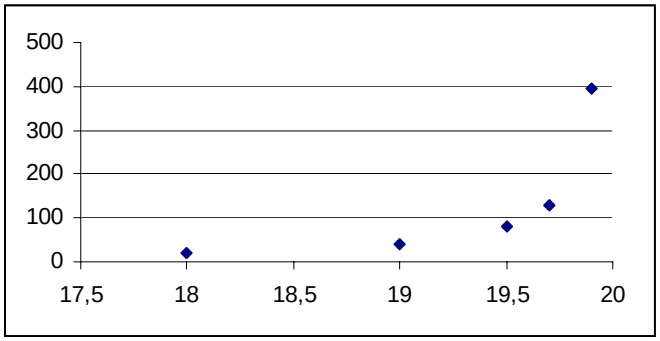
Regresión	Gráfica	Transformación
Exponencial $y = \alpha\beta^x$ El valor de $y \neq 0$ Asíntota horizontal eje X		$\ln y = \ln \alpha + x \ln \beta$ $y_1 = \ln y, x_1 = x$ $\alpha_1 = \ln \alpha, \beta_1 = \ln \beta$ Se tiene el modelo: $y_1 = \alpha_1 + \beta_1 x_1$
Potencial $y = \alpha x^\beta$ Pasa por $(0, 0)$		$\ln y = \ln \alpha + \beta \ln x$ $y_1 = \ln y, x_1 = \ln x$ $\alpha_1 = \ln \alpha, \beta_1 = \beta$ Se tiene el modelo: $y_1 = \alpha_1 + \beta_1 x_1$
Recíproca $y = \alpha + \dfrac{\beta}{x}$ Asíntota vertical eje Y Asíntota horizontal $\neq$ eje X		$y = \alpha + \beta\left(\dfrac{1}{x}\right)$ $y_1 = y, x_1 = \dfrac{1}{x}$ $\alpha_1 = \alpha, \beta_1 = \beta$ Se tiene el modelo $y_1 = \alpha_1 + \beta_1 x_1$
Hiperbólica $y = \dfrac{x}{\alpha x + \beta}$ Tiene Asíntiotas Vertical y Horizontales que no son los ejes.		$\dfrac{1}{y} = \alpha + \beta\left(\dfrac{1}{x}\right)$ $y_1 = \dfrac{1}{y}, x_1 = \dfrac{1}{x}$ $\alpha_1 = \alpha, \beta_1 = \beta$ Se tiene el modelo $y_1 = \alpha_1 + \beta_1 x_1$

Ejemplo 97 *Considere los datos en la siguiente tabla:*

x	18	19	19.5	19.7	19.9
Y	19	40	79	130	397

1. *Escoja y justifique un modelo de regresión para Y como función de x, sabiendo que la variable x por su naturaleza toma valores menores que 20.*

   Realicemos el diagrama de dispersión



   El modelo que más se adapta es el hiperbólico, pues se puede suponer que conforme $x \longrightarrow 20^-$ los valores de y crecen, es decir se espera que exista una asíntota vertical $x = 20$. Se descarta el modelo exponencial por que no tiene asíntota vertical y los otros modelos (Potencial y Recíproca) solo tiene asíntota vertical $x = 0$.

2. *Encuentre los coeficientes de la ecuación de regresión para el modelo escogido.*

   La ecuación de regresión es

$$y = E\left(Y_x\right) = \frac{x}{\alpha x + \beta}$$

   y la trasformación utilizada es $\frac{1}{y} = \alpha + \beta\left(\frac{1}{x}\right)$. *Por lo tanto si se define* $x_1 = \frac{1}{x}$ *y* $y_1 = \frac{1}{y}$, $\alpha_1 = \alpha$ *y* $\beta_1 = \beta$. *Así, la ecuación se transforma en un modelo lineal*

$$y_1 = \alpha_1 + \beta_1 x_1$$

   Seguidamente se encuentran las estimaciones de $\alpha_1 = \alpha$ *y* $\beta_1 = \beta$ *(* $a_1 = a$ *y* $b_1 = b$ *respectivamente)*

x	y	$x_1 = \frac{1}{x}$	$y_1 = \frac{1}{y}$	x_1^2	y_1^2	$x_1 y_1$
18	19	0.055555556	0.052631579	0.00309	0.00277	0.00292
19	40	0.052631579	0.025	0.00277	0.00063	0.00132
19.5	79	0.051282051	0.012658228	0.00263	0.00016	0.00065
19.7	130	0.050761421	0.007692308	0.00258	0.00006	0.00039
19.9	397	0.050251256	0.002518892	0.00253	0.00001	0.00013
Suma :		0.260481863	0.100501006	0.013588262	0.00362083	0.005405956

De lo anterior se obtiene que

$$b_1 = \frac{\frac{1}{n}\sum x_1 y_1 - \overline{x}_1 \cdot \overline{y}_1}{\frac{1}{n}\sum x_1^2 - (\overline{x}_1)^2} \approx \frac{\frac{1}{5}(0.005406) - \frac{0.260482}{5}\frac{0.100501}{5}}{\frac{1}{5}(0.013588) - \left(\frac{0.260482}{5}\right)^2} \approx 9.403241079$$

$$a_1 = \overline{y}_1 - b\overline{x}_1 = \frac{0.100501}{5} - (9.403241079)\frac{0.260482}{5} \approx -0.46977455$$

Por lo tanto la ecuación de regresión para el modelo escogido es

$$y = \frac{x}{-0.46977455x + 9.403241079}$$

3. *Determine el valor esperado aproximado de Y cuando $x = 17$.*

   El valor esperado de Y es aproximadamente

$$y = \frac{17}{-0.46977455 \cdot 17 + 9.403241079} \approx 11.\,996\,6.$$

4. *Encuentre un Intervalo de confianza del 95% para α y β (los parámetros del modelo escogido en a).*

   Note que α y β son los parámetros de la regresión lineal:

$$y_1 = \alpha_1 + \beta_1 x_1$$

   Por lo tanto, por lo visto en la sección (2.3) se tiene que

$$\begin{array}{cc} \delta = 0.05 & S_{x_1y_1} = 0.000170218 \\ n = 5 & SCE = 0.0000001385068 \\ b = b_1 = 9.403241079 & s^2 = 0.000000046169 \\ a = a_1 = -0.46977455 & s = 0.00021487 \\ S_{x_1x_1} = 0.0000181 & t_{\delta/2,n-2} = 3.182446305 \\ S_{y_1y_1} = 0.00160074 & \end{array}$$

   Por lo tanto los extremos del IC para $\alpha_1 = \alpha$ son

$$a_1 + t_{\delta/2,n-2}\left(s\sqrt{\frac{\sum x_1^2}{nS_{x_1x_1}}}\right) \approx -0.461396.$$

$$a_1 - t_{\delta/2,n-2}\left(s\sqrt{\frac{\sum x_1^2}{nS_{x_1x_1}}}\right) \approx -0.478153101.$$

   y los extremos del IC para $\beta_1 = \beta$ son

$$b_1 + t_{\delta/2,n-2}\left(s\sqrt{\frac{1}{S_{x_1x_1}}}\right) \approx 9.24252033.$$

$$b_1 - t_{\delta/2,n-2}\left(s\sqrt{\frac{1}{S_{x_1x_1}}}\right) \approx 9.563961829.$$

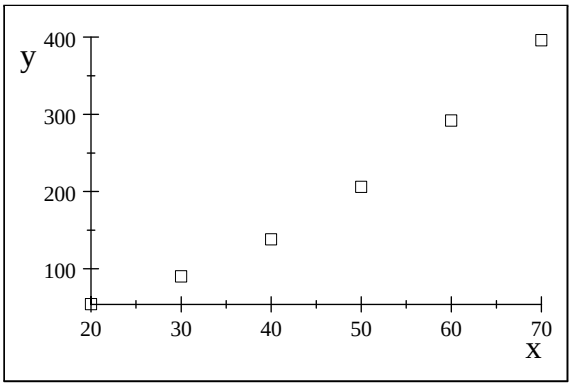
Por lo tanto, los IC del 95% *para* α *y para* β *son respectivamente*

$$]-0.461396, \quad -0.478153101[\qquad]9.24252033, \quad 9.563961829[$$

Ejemplo 98 *Seguidamente se muestra la distancia de frenado D (en pies) de un automóvil que viaja a la velocidad V (en millas por hora) a partir del instante en que se percibe el peligro.*

Velocidad V (mi / h)	20	30	40	50	60	70
Distancia de frenado D (pies)	54	90	138	206	292	396

1. *Seleccione y justique un modelo de regresión de la distancia de frenado dependiendo de la velocidad que lleva el automóvil.*



Note que la gráfica debe pasar por $(0,0)$, *pues si la velocidad es* 0, *el auto está detenido y la distancia de frenado es* 0. *Así, según la gráfica y que pasa por* $(0,0)$, *el modelo seleccionado es el*
potencial $y = \alpha x^{\beta}$

2. *Encuentre la ecuación de regresión del modelo seleccionado en la parte* (1).

Como $y = \alpha x^{\beta}$, *aplicando logaritmo se tiene que*

$$\ln y = \ln\left(\alpha x^{\beta}\right) = \ln\alpha + \beta \ln x$$

Tomando $y_1 = \ln y$, $x_1 = \ln x$, $\alpha_1 = \ln\alpha$ *y* $\beta_1 = \beta$, *la ecuación anterior se puede expresar por*

$$y_1 = \alpha_1 + \beta_1 x_1$$

Está última ecuación es lineal con variables x_1, y_1. *Los valores muestrales de estás nuevas variables son:*

x	y	$x_1 = \ln x$	$y_1 = \ln y$	$(x_1)^2$	$(y_1)^2$	$x_1 y_1$
20	54	2.995732274	3.988984047	8.974411855	15.91199372	11.94992825
30	90	3.401197382	4.49980967	11.56814363	20.24828707	15.30474087
40	138	3.688879454	4.927253685	13.60783163	24.27782888	18.17604488
50	206	3.912023005	5.327876169	15.30392399	28.38626447	20.84277414
60	292	4.094344562	5.676753802	16.76365739	32.22553373	23.24258606
70	396	4.248495242	5.981414211	18.04971182	35.77731597	25.41200982
Suma :		22.34067192	30.40209158	84.26768032	156.8272238	114.928084

Entonces por el método de mínimos cuadrados se tiene que las estimaciones de α_1 y β_1 son respectivamente

$$\begin{aligned} b_1 &= \frac{\frac{1}{n}\sum x_1 y_1 - \overline{x}_1 \cdot \overline{y}_1}{\frac{1}{n}\sum x_1^2 - (\overline{x}_1)^2} \approx \frac{\frac{1}{6}(114.928) - \frac{22.34067}{6}\frac{30.40209}{6}}{\frac{1}{6}(84.26768) - \left(\frac{22.34067}{6}\right)^2} \approx 1.59456 \\ a_1 &= \overline{y}_1 - b\overline{x}_1 \approx \frac{30.40209}{6} - (1.59456)\frac{22.34067}{6} \approx -0.87022. \end{aligned}$$

Como $a_1 = \ln a$ y $b_1 = b$ entonces las estimaciones de α y β son respectivamente

$$a = e^{-0.87022} = 0.418\,86 \text{ y } b = 1.59456.$$

Así, la ecuación de regresión es

$$y = 0.418\,86x^{1.59456}$$

3. *Encuentre un intervalo de confianza de 90% para α y β (los parámetros del modelo escogido en 1)*

$\delta = 0.1$	$S_{x_1x_1} = 1.083410023$
$n = 6$	$S_{y_1y_1} = 2.779361722$
$b_1 = 1.594556411$	$S_{x_1y_1} = 1.727558398$
$a_1 = -0.870228343$	$SCE = 0.024672402$
$b = b_1 = 1.594556411$	$s^2 = 0.006168101$
$a = e^{a_1} = 0.418855895$	$s = 0.078537256$
$t_{\delta/2,n-2} = 2.131846782$	

Recuerde que los extremos de los IC para α_1 y para β_1 son respectivamente de la forma

$$a_1 \pm t_{\delta/2,n-2}\left(s\sqrt{\frac{\sum x_1^2}{nS_{x_1x_1}}}\right), \quad b_1 \pm t_{\delta/2,n-2}\left(s\sqrt{\frac{1}{S_{x_1x_1}}}\right)$$

Por lo tanto

Extremos del IC para $\alpha_1 = \ln\alpha$		*Extremos del IC para α*
izquierdo: -1.47305185	$\Longrightarrow$	*izquierdo:* $e^{-1.47305185} = 0.229224857$
derecho: -0.267404836		*derecho:* $e^{-0.267404836} = 0.765363162$

Extremos del IC para $\beta_1 = \beta$		*Extremos del IC para β*
izquierdo: 1.433701145	$\Longrightarrow$	*izquierdo:* 1.433701145
derecho: 1.755411678		*derecho:* 1.755411678

Algunas observaciones:

1. En el modelo exponencial: $y = \alpha\beta^x$ se tiene que
   (a) α es el valor experado de Y cuando $X = 0$.

(b) β es el factor de aumento promedio de Y debido a cada unidad de aumento de X. Por ejemplo, si el modelo es $y = 5 \cdot 1.28^x$, entonces 1.28 es el factor de aumento promedio de Y debido a cada unidad de aumento de X, es decir por cada unidad de X, la variable Y tiene un aumento del 28%. En general, $(\beta - 1)\,100\%$ es el porcentaje de aumento (si $\beta - 1 > 0$) o de dismunición (si $\beta - 1 < 0$) en Y por cada unidad de aumento en X.

2. En los modelos de regresión no lineal existen dos caminos para estimar los coeficientes: el método de mínimos cuadrados (visto en la sección anterior) y el método de transformación al modelo de regresión lineal. Estos métodos generan estimaciones distintas. En este documento se selecionó el segundo camino pues brinda una presentación teórica más simple.

3. Hay problemas con el IC de μ_{Yx_0} en ciertos modelos no lineales. Por ejemplo para el modelo potencial $y = \alpha x^\beta$ la ecuación de trasformación es

$$y_1 = \alpha_1 + \beta_1 x_1$$

donde $y_1 = \ln y$ y $x_1 = \ln x$. Note que si $x = x_0$ entonces $x_1 = \ln x_0$, así un IC para $E(y_1)$ ciuando $x_1 = \ln x_0$ tiene extremos

$$(a_1 + b_1 \ln x_0) \pm t_{\delta/2,n-2} \left(s \sqrt{\frac{1}{n} + \frac{(\ln x_0 - \overline{x_1})^2}{S_{x_1 x_1}}} \right).$$

Sin embargo, aunque $y = e^{y_1}$, no se cumple que $E(y) = e^{E(y_1)}$. Por lo tanto no se puede hallar un IC para $E(y)$ cuando $x = x_0$, a partir del IC para $E(y_1)$.

Ejemplo 99 *Considere los datos de la siguiente tabla:*

$x:$	1	2	4	6
$Y:$	5	7	20	40

A partir de estos datos, estime el coeficiente β de la ecuación de regresión $y = 3 + x^\beta$ trasformando el modelo a un modelo lineal y utilizando el método de mínimos cuadrados.

Sea $y = 3 + x^\beta$, entonces $y - 3 = x^\beta$, así

$$\ln(y - 3) = \beta \ln x$$

Por lo tanto, sea $y_1 = \ln(y - 3)$ y $x_1 = \ln x$. Así, la ecuación de regresión para el modelo lineal es

$$y_1 = \beta x_1$$

Para estimar β se utiliza el método de mínimos cuadrados, así se busca el b tal que $\widehat{y}_1 = b x_1$ y la suma del cuadro de los errores

$$SCE = \sum_{i=1}^{4} e_i^2 = \sum_{i=1}^{4} \left((y_1)_i - \widehat{y}\left((x_1)_i\right) \right)^2 = \sum_{i=1}^{4} \left((y_1)_i - b\,(x_1)_i \right)^2$$

sea mínima, donde los valores muestrales de las variables x_1, y_1 *son*

$$\begin{array}{rccccc} x_1: & 0 & 0.693\,147 & 1.386\,29 & 1.791\,76 \\ y_1: & 0 & 1.386\,29 & 2.833\,21 & 3.610\,92 \end{array}$$

Note que SCE *es una función que depende de* b *:*

$$f(b) = SCE = \sum_{i=1}^{4} \left((y_1)_i - b\,(x_1)_i\right)^2,$$

de hecho f *es una función cuadrática cóncava hacia arriba. Por lo tanto su mínimo se alcanza cuando* $f'(b) = 0$*, note que*

$$\begin{aligned} f'(b) &= \sum_{i=1}^{4} \left[2\left((y_1)_i - b\,(x_1)_i\right) \cdot -(x_1)_i\right] \\ &= -2\left(\sum_{i=1}^{4}\left[(y_1)_i\,(x_1)_i\right] - b\sum_{i=1}^{4}\left[(x_1)_i\right]^2\right) \\ &\approx -2\,(11.358455 - 5.612657b) \end{aligned}$$

Así $f'(b) = 0$ *si* $b \approx 2.\,023\,72$*. Por lo tanto, una estimación de* β *es*

$$b = 2.\,023\,72.$$

Ejercicio 49 *Considere los datos en la siguiente tabla:*

$$\begin{array}{rcccccc} x: & 0 & 4 & 5 & 6 & 7 & 8 \\ Y: & 2 & 37 & 80 & 172 & 360 & 756 \end{array}$$

1. *Escoja y justifique un modelo de regresión para y como función de x.* *R/ Exponencial*
2. *Encuentre una ecuación de regresión para el modelo escogido.* *R/* $y = 1.9648\,(2.\,102\,65)^x$

Ejercicio 50 *Los siguientes datos, que representan las velocidades promedio para las diferentes distancias de cinco corredores, todos con más de 70 años de edad:*

$$\begin{array}{rccccc} \textit{Distancia } (d) & 1.6 & 3 & 5 & 10 & 42 \\ \textit{Vel. prom. } (V) & 4.7 & 4.2 & 4.1 & 3.9 & 3.8 \end{array}$$

1. *Determine la ecuación de regresión no lineal simple (potencial) de v en función de d.* *R/* $v = 4.604d^{-0.05}$
2. *Determine un IC del* 95% *para* α *(* $v = \alpha d^{\beta}$*, asuma las hipótesis de regresión)*
   R/ $]4.10688482,\quad 5.16208568[$

3.1 Ejercicios

1. Considere los datos de la siguiente tabla:

$$\begin{array}{lccccc} X: & -4 & -2 & 3 & 5 & 7 \\ Y: & 50 & 30 & 10 & 6 & 2 \end{array}$$

   (a) Determine una ecuación para el modelo de regresión exponencial $y = \alpha\beta^x$.
   $R/ \quad \widehat{y} = 18.3154\,(0.7627)^x$

   (b) ¿Cuál es el valor aproximado esperado de Y cuando $X = 2$? $R/ \quad 10.6543$

   (c) Determine un IC del 80% para α $R/ \quad]14.83071265, 22.61890121[$

   (d) Determine un IC del 90% para β $R/ \quad]0.713405605, 0.815402581[$

   (e) ¿Considera que el valor esperado de Y disminuye por cada unidad de X?
   $R/$ Si, por IC anterior

   (f) Encuentre un intervalo de predicción de 95% para Y cuando $X = -3$.
   $R/ \quad]14.89999143, 114.3725251[$

   (g) Pruebe si el porcentaje de disminución de Y, por cada unidad de X, es del 30%.
   $R/ \quad t_{obs} = 3.021509968$, Valor $P \in]0.05, 0.1[$. Se acepta que $\beta = 0.7$ levemente.

2. Considere los datos de la siguiente tabla:

$$\begin{array}{lcccccc} X: & 2 & 4 & 6 & 8 & 10 & 12 \\ Y: & 2 & 6 & 8 & 10 & 11 & 11 \end{array}$$

   (a) Determine una ecuación para el modelo de regresión exponencial $y = \alpha\beta^x$.
   $R/ \quad \widehat{y} = 2.4285\,(1.16292)^x$

   (b) ¿Cuál es el valor aproximado esperado de Y cuando $X = 11$? $R/ \quad 12.7754$

   (c) Determine un IC del 90% para α $R/ \quad]1.161728788, 5.076424094[$

   (d) Determine un IC del 98% para β $R/ \quad]0.984664462, 1.373441746[$

   (e) ¿Considera que el valor esperado de Y cuando $X = 0$ es menor a 6 unidades? $R/$ Si,

   (f) ¿Considera que el valor esperado de Y aumenta por cada unidad de X? $R/$ No

   (g) Encuentre un intervalo de predicción de 90% para Y cuando $X = 9$.
   $R/ \quad]3.933192406, 22.68840805[$

   (h) Pruebe, a un nivel de significacia del 10%, si el porcentaje de aumento de Y por cada unidad de X es menor al 30%. $R/ \quad t_{obs} = -2.509378034$, $t_c = -1.533206273$, no hay evidencia significativa en contra de que $\beta < 1.3$.

3. Considere los datos de la siguiente tabla:

$$\begin{array}{lccccc} X: & 2 & 3 & 7 & 10 & 17 \\ Y: & 100 & 90 & 40 & 10 & 5 \end{array}$$

   (a) Determine una ecuación para el modelo de regresión potencial $y = \alpha x^\beta$.
   $R/ \quad \widehat{y} = 379.5215x^{-1.4638}$

(b) ¿Cuál es el valor aproximado esperado de Y cuando $X = 15$? $R/$ 7.2063

(c) Determine un IC del 90% para α $R/$ $]118.9566454, 1210.832691[$

(d) Determine un IC del 96% para β $R/$ $]-2.34891188, -0.578618665[$

(e) ¿Considera que el valor esperado de Y cuando X es menor a 1300 unidades? $R/$ Si

(f) Encuentre un intervalo de predicción de 95% para Y cuando $X = 5$.
$R/$ $]7.600620173, 170.3600581[$

(g) Pruebe si $\beta < -0.5$. $R/$ $t_{obs} = -3,791171671$, Valor $P \in]0.01, 0.02[$. Se acepta que $\beta < -0.5$

4. Considere los datos de la siguiente tabla:

$X:$	1	5	10	15	20	25
$Y:$	1	10	19	22	25	26

(a) Determine una ecuación para el modelo de regresión potencial $y = \alpha x^{\beta}$.
$R/$ $\widehat{y} = 1.3141x^{1.0266}$

(b) ¿Cuál es el valor aproximado esperado de Y cuando $X = 17$? $R/$ 24.0889

(c) Determine un IC del 96% para α $R/$ $]0.54127242, 3.190133922[$

(d) Determine un IC del 92% para β $R/$ $]0.739104881, 1.314136273[$

(e) Encuentre un intervalo de predicción de 90% para Y cuando $X = 6$.
$R/$ $]3.8595739, 17.71811942[$

(f) Pruebe, a un nivel de significancia del 5%, si $\beta = 1.5$. $R/$ $t_{obs} = -3,840951577$, $t_{c1} = -2,776445105$,hay evidencia significativa en contra de que $\beta = 1.5$

5. Considere los datos de la siguiente tabla:

$X:$	0.5	1	5	9	14
$Y:$	200	100	40	20	1

(a) Seleccione y justique un modelo de regresión de Y con función de X. $R/$ Potencial

(b) Encuentre la ecuación de regresión del modelo seleccionado. $R/$ $\widehat{y} = 115.1562x^{-1.2418}$

(c) ¿Cuál es el valor aproximado esperado de Y cuando $X = 10$? $R/$ 6.5997

(d) Determine un IC del 92% para α $R/$ $]18.42803518, 719.6078194[$

(e) Determine un IC del 95% para β $R/$ $]-2.539948745, 0.0564166[$

(f) Encuentre un intervalo de predicción de 95% para Y cuando $X = 10$.
$R/$ $]0.085298721, 510.6228498[$

(g) Pruebe, a un nivel de significancia del 10%, si β es positivo. $R/$ $t_{obs} = -3,044143118$, $t_{c1} = 1,637744352$,hay evidencia significativa en contra de que β sea positivo.

6. Considere los datos de la siguiente tabla:

$X:$	0.01	1	2	10	40
$Y:$	2000	70	60	50	51

(a) Seleccione y justique un modelo de regresión de Y con función de X. $R/$ Recíproco

(b) Encuentre la ecuación de regresión del modelo seleccionado. $R/$ $\widehat{y} = 49.8256 + \dfrac{19.5018}{x}$

(c) ¿Cuál es el valor aproximado esperado de Y cuando $X = 30$? $R/$ 50.4756

(d) Determine un IC del 90% para α $R/$ $]48.41815035, 51.23300384[$

(e) Determine un IC del 95% para β $R/$ $]19.45926114, 19.54437211[$

(f) Halle un intervalo de confianza del 90% para el valor esperado de Y cuando $X = 5$. $R/$ $]50.23986138, 57.21201946[$

(g) Encuentre un intervalo de predicción de 95% para Y cuando $X = 0.2$. $R/$ $]143.1309767, 151.5383438[$

7. Considere los datos de la siguiente tabla:

$X:$	-200	-50	-25	-10	-1	-0.3
$Y:$	-30	-31	-35	-40	-145	-415

(a) Seleccione y justique un modelo de regresión de Y con función de X. $R/$ Recíproco

(b) Encuentre la ecuación de regresión del modelo seleccionado.$R/$ $\widehat{y} = -29.2359 + \dfrac{115.7284}{x}$

(c) ¿Cuál es el valor aproximado esperado de Y cuando $X = -20$? $R/$ -35.0223

(d) Determine un IC del 95% para α $R/$ $]-30.24333151, -28.22842988[$

(e) Determine un IC del 90% para β $R/$ $]115.1841543, 116.2725552[$

(f) Halle un intervalo de confianza del 95% para el valor esperado de Y cuando $X = -30$. $R/$ $]-34.08868515, -32.09829989[$

(g) Encuentre un intervalo de predicción de 90% para Y cuando $X = -0.8$. $R/$ $]-175.6563186, -172.1363297[$

8. Considere los datos de la siguiente tabla:

$X:$	-7	-5	-1	3	6	9	15
$Y:$	80	75	60	50	20	10	8

(a) Seleccione y justique un modelo de regresión de Y con función de X. $R/$ Exponencial

(b) Encuentre la ecuación de regresión del modelo seleccionado. $R/$ $\widehat{y} = 43.7191\,(0,888657792)^x$

(c) ¿Cuál es el valor aproximado esperado de Y cuando $X = 12$? $R/$ 10.6044

(d) Determine un IC del 96% para α $R/$ $]30.56616985, 62.53198884[$

(e) Determine un IC del 95% para β $R/$ $]0.851440388, 0.92750201[$

(f) ¿Considera que el valor esperado de Y cuando $X = 0$ es menor a 60 unidades? $R/$ No

(g) ¿Considera que el valor esperado de Y disminuye por cada unidad de X? $R/$ Si

(h) Encuentre un intervalo de predicción de 92% para Y cuando $X = 10$.
$R/$ $]6.077920609, 29.66738759[$

(i) Pruebe, si el porcentaje de disminución de Y por cada unidad de X es mayor o igual al 15%. $R/$ $t_{obs} = 2.672313403$, Valor $P \in]0.02, 0.025[$. Existe evidencia en contra de que $\beta \leq 0.85$.

9. Considere los datos de la siguiente tabla:

$X:$	3.2	3.4	4	10	100	500
$Y:$	-1600	-850	-400	-145	-103	-100.5

(a) Seleccione y justique un modelo de regresión de Y con función de X. $R/$ Hiperbólico

(b) Encuentre la ecuación de regresión del modelo seleccionado. $R/$ $\widehat{y} = \dfrac{x}{-0.01x + 0.02997}$

(c) ¿Cuál es el valor aproximado esperado de Y cuando $X = 50$? $R/$ -106.5969

(d) Determine un IC del 96% para α $R/$ $]-0.01007626, -0.009884657[$

(e) Determine un IC del 95% para β $R/$ $]0.029537345, 0.030394958[$

(f) Encuentre un intervalo de predicción de 90% para Y cuando $X = 40$.
$R/$ $]-106.7913887, -109.907515[$

10. Considere los datos de la siguiente tabla:

$x:$	1	2	3	4
$Y:$	7	11	20	36

A partir de estos datos, estime el coeficiente β de la ecuación de regresión $y = 3 + 2\beta^x$ trasformando el modelo a un modelo lineal y utilizando el método de mínimos cuadrados.
$R/$ 2.020434385

11. Considere los datos de la siguiente tabla:

$x:$	1	3	10	100
$Y:$	2	5	15	40

A partir de estos datos, estime el coeficiente β de la ecuación de regresión $y = \dfrac{x}{0.02x + \beta}$ trasformando el modelo a un modelo lineal y utilizando el método de mínimos cuadrados.
$R/$ 0.485828816

12. El gerente de una empresa toma una muestra de 8 trabajadores para determinar si hay relación entre el número de años cumplidos que tiene un empleado de laborar en la empresa (X) y los días que falta al trabajo en el año (Y), obteniendo los siguientes resultados:

Años laborados (X)	0	5	10	20	25
Días que falta (Y)	5	8	14	25	32

(a) Seleccione y justique un modelo de regresión del número de días que falta al trabajo en el año con función de los años laborados.

$R/$ Exponencial $y = \alpha\beta^x$.

(b) Encuentre la ecuación de regresión del modelo seleccionado en la parte (**a**) .

$R/$ $y = 5,579100043\,(1,076401484)^x$

(c) Encuentre un intervalo de confianza de 90% para α y β (los parámetros del modelo escogido en **a**) $R/$ IC para $\alpha :]4.4155, 7.0494[$, IC para $\beta :]1.0599, 1.0931[$.

13. Los siguientes datos indican el valor (Y), en miles de colones, de cierto marca de computadoras portátiles con características determinadas, después de X años de uso:

Años de uso (X)	0.5	1	2	3	4
valor (Y)	850	600	400	290	150

(a) Seleccione y justifique un modelo de regresión del valor de este tipo de computadoras en función de los años de uso. $R/$ exponencial

(b) Encuentre la ecuación de regresión del modelo seleccionado en la parte (**a**) .

$R/$ $y = 1029.977976\,(0.628744589)^x$

(c) ¿Aproximadamente, cuál es el valor promedio de este tipo de portátiles nuevas?$R/$ 1029977.976 colones

(d) Encuentre un intervalo de confianza de 90% para α (uno de los parámetros del modelo escogido en **a,** según la notación vista) $R/$ $]846.4924561, 1253.235777[$

(e) Encuentre un intervalo de confianza de 90% para β $R/$ $]0.580541892, 0.68094958[$

14. Se invierten 200 000 colones a una tasa de interés poco variable. Varios meses después, el valor de la inversión, en miles de colones, es:

meses (x)	0	6	12	24
valor (y)	200	270	365	670

(a) Escoja y justifique un modelo de regresión no lineal para el valor de la inversión en función del número de meses. Justifique su respuesta. $R/$ Exponencial

(b) Encuentre una ecuación de regresión para el modelo escogido. $R/$ $y = 199.7\,(1.\,051\,27)^x$

(c) Encuentre un Intervalo de confianza del 95% para α (uno de los parámetros del modelo escogido en a) $R/$ $]198.5130, 200.921[$

15. Considere las siguientes observaciones:

x	1	3	5	6	7
y	21	120	980	2930	8755

(a) Determine la ecuación de regresión no lineal simple $(y = 10 + \alpha\beta^x)$ de y en función de x.

$R/$ $y = 10 + 3.747570459\,(3.035145944)^x$

(b) ¿Cuál es el valor esperado de y cuando $x = 4$? $R/$ 328.0300243

(c) Encuentre un intervalo de confianza de 90% para α $R/$ $]3.441725607, 4.080593851[$

(d) Encuentre un intervalo de confianza de 90% para β $R/$ $]2.982856645, 3.088351873[$

(e) Encuentre un intervalo de predicción de 95% para y cuando $x = 4$.
$R/$ $]290.8424619, 370.1417523[$

16. Los siguientes datos representan la velocidad para un corredor en diferentes pendientes, para una distancia de 300m:

Pendiente (%)	-2	-1.5	0.2	1.2	1.9	2.5
Velocidad promedio (m/s)	17	14	8	5	3	2

Nota: La pendiente se define como la diferencia de altura, del punto inicial al final, dividida por la distancia recorrida.

(a) Seleccione y justifique un modelo de regresión no lineal de la velocidad en función de la pendiente. $R/$ exponencial

(b) Encuentre la ecuación de regresión del modelo seleccionado en la parte (**a**).
$R/$ $y = 7.397469686\,(0.632404909)^x$

(c) ¿Cuál es la velocidad promedio aproximada del corredor en caminos horizontales de 300m?
$R/$ 7.397469686 m/s

(d) Si el corredor tardó un minuto 10 segundos en recorrer 300m, ¿cuánta altura se estima que se elevó? Considera la estimación dada como buena. $R/$ 357.3682912 m, No

(e) Encuentre un intervalo de confianza de 90% para α (uno de los parámetros del modelo escogido en **a**) $R/$ $]6.46594771, 8.463192128[$

(f) Encuentre un intervalo de confianza de 90% para β (uno de los parámetros del modelo escogido en **a**) $R/$ $]0.584598232, 0.68412107[$

(g) Determine un intervalo de predicción de 95% para la velocidad en caminos horizontales de 300m. $R/$ $]4.699486313, 11.64437007[$

(h) Determine un intervalo de predicción de 95% para el tiempo que tarda recorrido 300m con una pendiente de -1%. $R/$ $]15.97090364, 41.18481278[$

17. Demuestre que $\sum_{i=1}^{n} e_i = 0$.

18. Una investigación indica que cierta enfermedad tuvo su aparición en los primeros días del año 1998. Seguidamente se presenta algunos datos sobre el porcentaje de personas infectadas de la enfermedad en un país C, X años después del 1° de enero del 1998:

Años (X)	0.2	2	5	10	12
Porcentaje (Y)	3%	12.7%	13.4%	13.9%	14%

(a) Justifique por qué el modelo Hiperbólico es un buen modelo de regresión para explicar la relación entre las variables X,Y. $R/$ Hiperbólica con $\beta > 0$

(b) Encuentre la ecuación de regresión del modelo Hiperbólico para Y como función de X.
$R/\quad y = \dfrac{x}{0.062475762x + 0.053998736}$

(c) Encuentre un intervalo de confianza de 90% para α $R/\quad]0.053623507,\quad 0.071328017[$

(d) Encuentre un intervalo de predicción del 95% para el porcentaje de personas del país C que tendrán la enfermedad el 1° de enero del 2011 $R/\quad]10.8345, 24.4132[$

4 Regresión múltiple

Un modelo de regresión multiple se da cuando el valor esperado de la variable cuantitativa Y está en función de más de una variable. Es decir existen k variables $x_1, x_2, ..., x_k$ y una función f tal que

$$E\left(Y_{x_1,x_2,...,x_k}\right) = f\left(x_1, x_2, ..., x_k\right)$$

En los siguientes apartados se estudiará la regrresión múltiple cuando f es una función lineal y ciertos casos en las cuales f no es lineal.

4.1 Regresión lineal múltiple

Definición 40 *Dadas las variables $Y, x_1, x_2, ..., x_k$ si existen constantes $\beta_0, \beta_1, ..., \beta_k$ tales que*

$$E\left(Y_{x_1,x_2,...,x_k}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

se dice que hay una regresión lineal múltiple.

Para estimar puntualmente estas constantes (b_i es una estimación de β_i) se parte del valor de las variables para n individuos:

		Individuos			
	1	2	3	...	n
x_1	x_{11}	x_{12}	x_{13}	...	x_{1n}
x_2	x_{21}	x_{22}	x_{23}	...	x_{2n}
Variables x_3	x_{31}	x_{32}	x_{33}	...	x_{3n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_k	x_{k1}	x_{k2}	x_{k3}	...	x_{kn}
Y	y_1	y_2	y_3	...	y_n

Así se busca $B = (b_1, b_2, ..., b_k)$ de manera que los datos obtenidos para cada individuo satisfagan la ecuación

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k.$$

Como son n individuos entonces se buscan algún valor de $b_1, b_2, ..., b_k$ que sea solución del sistema de n ecuaciones

$$\begin{cases} y_1 = b_0 + b_1 x_{11} + b_2 x_{21} + ... + b_k x_{k1} \\ y_2 = b_0 + b_1 x_{12} + b_2 x_{22} + ... + b_k x_{k2} \\ \vdots \\ y_n = b_0 + b_1 x_{1n} + b_2 x_{2n} + ... + b_k x_{kn} \end{cases}$$

Este sistema se puede denotar matricialmente por

$$X^t B^t = Y^t \qquad (1)$$

donde

$$X = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ x_{31} & x_{32} & x_{33} & \dots & x_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & x_{k3} & \dots & x_{kn} \end{pmatrix} \quad Y = \begin{pmatrix} y_1 & y_2 & y_3 & \dots & y_n \end{pmatrix}$$

Multiplicando a ambos lados por X se obtiene que

$$\left(XX^t\right) B^t = XY^t \qquad (2)$$

Dado que el tamaño de X es $(k+1) \times n$ entonces (XX^t) tiene tamaño $(k+1) \times (k+1)$, además B^t y XY^t tiene tamaño $(k+1) \times 1$. Así, el sistema (2), a diferencia del sistema (1), es un sistema cuadrado (hay $k+1$ incógnitas y $k+1$ ecuaciones) que bajo las condiciones usuales determina B de manera única.

Teorema 54 *Dada la regresión lineal múltiple*

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

Para hallar los estimadores b_i de los β_i se resuelve el sistema

$$\left(XX^t\right) B^t = XY^t$$

Ejemplo 100 *Se tienen las siguientes observaciones*

x_1	0	1	1	0	-1
x_2	1	0	-1	0	1
Y	1	4	6	3	0

1. *Estime la ecuación de regresión de Y como función lineal de x_1,x_2.*

   En este caso, se tiene que

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & -1 \\ 1 & 0 & -1 & 0 & 1 \end{pmatrix} \quad y \quad Y = (1, 4, 6, 3, 0)$$

   Por lo tanto

$$XX^t = \begin{pmatrix} 5 & 1 & 1 \\ 1 & 3 & -2 \\ 1 & -2 & 3 \end{pmatrix}, \qquad XY^t = \begin{pmatrix} 14 \\ 10 \\ -5 \end{pmatrix}$$

El sistema a resolver es

$$\begin{pmatrix} 5 & 1 & 1 \\ 1 & 3 & -2 \\ 1 & -2 & 3 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 14 \\ 10 \\ -5 \end{pmatrix}$$

y su solución es

$$\begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \\ -2 \end{pmatrix}$$

Por lo tanto, la estimación de la ecuación de regresión que expresa a y como función lineal de x_1,x_2.es

$$y = 3 + x_1 - 2x_2$$

2. *Determine el valor esperado aproximado de Y cuando $x_1 = 2$ y $x_2 = 3$.*

   El valor esperado de Y es aproximadamente

$$y = 3 + x_1 - 2x_2 = 3 + 2 - 2 \cdot 3 = -1.$$

Ejercicio 51 *Se tienen las siguientes observaciones*

x_1	1	4	2	3	-2
x_2	2	1	-1	3	2
y	4	15	12	8	-6

Encuentre una estimación a la ecuación que expresa y como función lineal de x_1,x_2. *R/* $y = \frac{3791}{951} + \frac{3020}{951}x_1 - \frac{1676}{951}x_2$

4.2 Regresión no lineal múltiple

Algunos modelos de regresión no lineal múltiple pueden ser trasformados a regresión lineal múltiple.

Ejemplo 101 *Dadas las siguientes observaciones*

x	3	8	13	16	20
y	5	9	4	6	2
Z	-46	-110	10	-12	64

1. *Estime los coeficientes en la ecuación $z = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 y^2$.*

   Sea $x_1 = x, x_2 = y, x_3 = y^2$. Se quiere estimar los β_i que cumplen que

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

   Dado que este modelo es de regresión lineal múltiple, se procede de acuerdo al teorema:

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 3 & 8 & 13 & 16 & 20 \\ 5 & 9 & 4 & 6 & 2 \\ 25 & 81 & 16 & 36 & 4 \end{pmatrix}, Y = \begin{pmatrix} -46 & -110 & 10 & -12 & 64 \end{pmatrix}$$

Entonces

$$XX^t = \begin{pmatrix} 5 & 60 & 26 & 162 \\ 60 & 898 & 275 & 1587 \\ 26 & 275 & 162 & 1142 \\ 162 & 1587 & 1142 & 8754 \end{pmatrix} \quad y \quad XY^t = \begin{pmatrix} -94 \\ 200 \\ -1124 \\ -10\,076 \end{pmatrix}$$

El sistema a resolver es

$$\begin{pmatrix} 5 & 60 & 26 & 162 \\ 60 & 898 & 275 & 1587 \\ 26 & 275 & 162 & 1142 \\ 162 & 1587 & 1142 & 8754 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} -94 \\ 200 \\ -1124 \\ -10\,076 \end{pmatrix}$$

La solución es

$$b_0 = 2, \quad b_1 = 4, \quad b_2 = -7, \quad b_3 = -1$$

Por lo tanto,la estimación de la ecuación de regresión es

$$z = 2 + 4x - 7y - y^2$$

2. *Determine el valor esperado aproximado de Z cuando $x = 2, y = 7$.*

   El valor esperado de Z es aproximadamente

$$z = 2 + 4x - 7y - y^2 = 2 + 8 - 49 - 49 = -88.$$

Ejemplo 102 *Dadas las siguientes observaciones, estime los coeficientes en la ecuación $z = \beta_0 x^{\beta_1} y^{\beta_2}$:*

x	1	1.5	2	1
y	0.5	1	2	2
Z	0.06	6.7	512	64

Note que

$$\ln z = \ln \beta_0 + \beta_1 \ln x + \beta_2 \ln y$$

Así, mediante el cambio de variable $x_1 = \ln x, x_2 = \ln y, z_1 = \ln z$, nuestro problema de regresión se transforma en estimar los $\overline{B}_i$ tales que :

$$z_1 = \overline{\beta}_0 + \overline{\beta}_1 x_1 + \overline{\beta}_2 x_2$$

donde $\overline{\beta}_0 = \ln \beta_0, \overline{\beta}_1 = \beta_1, \overline{\beta}_2 = \beta_2$ y se tienen las observaciones

x_1	0	0.405465108	0.693147181	0
x_2	−0.693147181	0	0.693147181	0.693147181
z_1	−2.813410717	1.902107526	6.238324625	4.158883083

Dado que este modelo es de regresión lineal múltiple, se procede de acuerdo al teorema:

$$\begin{aligned} X &= \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 0.405465108 & 0.693147181 & 0 \\ -0.693147181 & 0 & 0.693147181 & 0.693147181 \end{pmatrix}, \\ Y &= (-2.813410717, 1.902107526, 6.238324625, 4.158883083) \end{aligned}$$

Entonces

$$XX^t = \begin{pmatrix} 4 & 1.098\,61 & 0.693\,147 \\ 1.098\,61 & 0.644\,855 & 0.480\,453 \\ 0.693\,147 & 0.480\,453 & 1.441\,36 \end{pmatrix} \quad y \quad XY^t = \begin{pmatrix} 9.485\,9 \\ 5.095\,31 \\ 9.156\,90 \end{pmatrix}$$

El sistema a resolver es

$$\begin{pmatrix} 4 & 1.098\,61 & 0.693\,147 \\ 1.098\,61 & 0.644\,855 & 0.480\,453 \\ 0.693\,147 & 0.480\,453 & 1.441\,36 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 9.485\,9 \\ 5.095\,31 \\ 9.156\,90 \end{pmatrix}$$

$$\overline{b}_0 = 0.674\,424, \quad \overline{b}_1 = 3.007\,83, \quad \overline{b}_2 = 5.026\,02$$

Por lo tanto la estimación de los β_i :

$$b_0 = e^{0.674\,424} = 1.962\,9, \quad b_1 = 3.007\,83, \quad b_2 = 5.026\,02$$

y la estimación de la ecuación de regresión es

$$z = 1.962\,9x^{3.007\,83}y^{5.026\,02}$$

Ejercicio 52 *En un estudio, se determinó el valor de las variables:* y, x_1, x_2 *para* n *individuos. Se desea estimar los coeficientes de regresión en el modelo:*

$$u_{Y \mid x_1,x_2} = \beta_0 - \beta_1 x_1 - 2\beta_1 x_2 + 3\beta_2 x_1 x_2$$

Explique de qué manera se puede realizar la estimación de los coeficientes transformando el modelo a uno de regresión lineal múltiple.

4.3 Ejercicios

1. Se llevó a cabo un experimento para determinar la distancia de frenado a diferentes velocidades de un modelo nuevo de automóvil. Se registraron los siguientes datos

$Velocodad: v\ (km/h)$	35	50	65	80	95	110
Distancia de frenado: $d\ (m)$	16	26	41	62	88	119

   (a) Ajuste una curva de regresión lineal múltiple de la forma: $u_{D \mid v} = \beta_0 + \beta_1 v + \beta_2 v^2$.
   $R/ \quad d = 0.011\,825v^2 + -0.33944v + 13.359$

   (b) Estime la distancia de frenado cuando el vehículo viaja a una velocidad de 70 km/h.
   $R/ \quad d = 47.541$

2. En un estudio, se determinó el valor de las variables: y, x_1, x_2 para n individuos. Se desea estimar los coeficientes de regresión en el modelo:

$$u_{Y \mid x_1,x_2} = \beta_0 + \beta_1 x_1 - 2\beta_2 x_2 + \beta_3 x_1 x_2 - \beta_4 x_1^5$$

   Explique de qué manera se puede realizar la estimación de los coeficientes transformando el modelo a uno de regresión múltiple.

3. En un estudio, se determinó el valor de las variables: y, x_1, x_2 para n individuos. Se desea estimar los coeficientes de regresión en el modelo:

$$u_{Y \mid x_1, x_2} = \beta_0 x_1^{\beta_1} \beta_2^{x_2}$$

Explique de que manera se puede realizar la estimación de los coeficientes transformando el modelo a uno de regresión lineal múltiple.

4. Considere las siguientes observaciones:

x	1	2	3	4
y	2	1	-1	-2
Z	19	95	54	57

Estime los coeficientes β_0, β_1 y β_2 en la ecuación $z = \beta_0 x^{\beta_1} \beta_2^y$.

$R/ \quad z = 2.275\,9 x^{3.851\,69} 2.888\,83^y$

5. Considere las siguientes observaciones:

x	1	1	1.2	2
y	1	0	0	1
Z	14	11	11	31

Estime los coeficientes β_0, β_1 y β_2 en la ecuación $z = 10 + \beta_0 x^{\beta_1} \beta_2^y$.

$R/ \quad z = 10 + 0.815\,484 x^{2.237\,51} 5.175\,42^y$

6. Un profesor de Estudios Sociales considera que la nota obtenida en su examen final (Y) depende linealmente del número de horas que durme un alumno en la noche antes del examen (x_1) y del número de horas de estudio en la semana previa al examen (x_2). Seguidamente se presentan los datos obtenidos de 4 estudiantes

x_1	4	1	6	8
x_2	4	8	12	15
Y	45	60	80	96

(a) Estime los coeficientes β_0, β_1 y β_2 en la ecuación de regresión lineal múltiple $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. $R/ \quad y = 24.533 + 0.770\,285\, x_1 + 4.313\,65 x_2$

(b) De acuerdo a la ecuación estimada, ¿Qué tiene más peso en la nota: las horas de estudio ó las horas que duerme la noche antes del examen? $R/$ Las horas de estudio.

7. Considere las siguientes observaciones:

x	0	1	2	3
y	5	3	2	1

Estime los coeficientes β_0, β_1 y β_2 en la ecuación $y = \dfrac{1}{\beta_0 + \beta_1 x + \beta_2 x^2}$.

$R/ \quad b_0 = 0.214\,995\,, b_1 = -0.018\,355\,, b_2 = 0.091\,675$

Tablas de distribuciones

Seguidamente se presenta las tablas para las siguientes distribuciones:

1. Distribución Binomial
2. Distribución Normal
3. Distribución t
4. Distribución χ^2
5. Distribución F

Distribución Acumulada de la Binomial: B(k;n,p)

		p										
n	**k**	**0,05**	**0,1**	**0,2**	**0,25**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**
1	**0**	0,9500	0,9000	0,8000	0,7500	0,7000	0,6000	0,5000	0,4000	0,3000	0,2000	0,1000
	1	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
2	**0**	0,9025	0,8100	0,6400	0,5625	0,4900	0,3600	0,2500	0,1600	0,0900	0,0400	0,0100
	1	0,9975	0,9900	0,9600	0,9375	0,9100	0,8400	0,7500	0,6400	0,5100	0,3600	0,1900
	2	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
3	**0**	0,8574	0,7290	0,5120	0,4219	0,3430	0,2160	0,1250	0,0640	0,0270	0,0080	0,0010
	1	0,9928	0,9720	0,8960	0,8438	0,7840	0,6480	0,5000	0,3520	0,2160	0,1040	0,0280
	2	0,9999	0,9990	0,9920	0,9844	0,9730	0,9360	0,8750	0,7840	0,6570	0,4880	0,2710
	3	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
4	**0**	0,8145	0,6561	0,4096	0,3164	0,2401	0,1296	0,0625	0,0256	0,0081	0,0016	0,0001
	1	0,9860	0,9477	0,8192	0,7383	0,6517	0,4752	0,3125	0,1792	0,0837	0,0272	0,0037
	2	0,9995	0,9963	0,9728	0,9492	0,9163	0,8208	0,6875	0,5248	0,3483	0,1808	0,0523
	3	1,0000	0,9999	0,9984	0,9961	0,9919	0,9744	0,9375	0,8704	0,7599	0,5904	0,3439
	4	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
5	**0**	0,7738	0,5905	0,3277	0,2373	0,1681	0,0778	0,0313	0,0102	0,0024	0,0003	0,0000
	1	0,9774	0,9185	0,7373	0,6328	0,5282	0,3370	0,1875	0,0870	0,0308	0,0067	0,0005
	2	0,9988	0,9914	0,9421	0,8965	0,8369	0,6826	0,5000	0,3174	0,1631	0,0579	0,0086
	3	1,0000	0,9995	0,9933	0,9844	0,9692	0,9130	0,8125	0,6630	0,4718	0,2627	0,0815
	4	1,0000	1,0000	0,9997	0,9990	0,9976	0,9898	0,9688	0,9222	0,8319	0,6723	0,4095
	5	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
6	**0**	0,7351	0,5314	0,2621	0,1780	0,1176	0,0467	0,0156	0,0041	0,0007	0,0001	0,0000
	1	0,9672	0,8857	0,6554	0,5339	0,4202	0,2333	0,1094	0,0410	0,0109	0,0016	0,0001
	2	0,9978	0,9842	0,9011	0,8306	0,7443	0,5443	0,3438	0,1792	0,0705	0,0170	0,0013
	3	0,9999	0,9987	0,9830	0,9624	0,9295	0,8208	0,6563	0,4557	0,2557	0,0989	0,0159
	4	1,0000	0,9999	0,9984	0,9954	0,9891	0,9590	0,8906	0,7667	0,5798	0,3446	0,1143
	5	1,0000	1,0000	0,9999	0,9998	0,9993	0,9959	0,9844	0,9533	0,8824	0,7379	0,4686
	6	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
7	**0**	0,6983	0,4783	0,2097	0,1335	0,0824	0,0280	0,0078	0,0016	0,0002	0,0000	0,0000
	1	0,9556	0,8503	0,5767	0,4449	0,3294	0,1586	0,0625	0,0188	0,0038	0,0004	0,0000
	2	0,9962	0,9743	0,8520	0,7564	0,6471	0,4199	0,2266	0,0963	0,0288	0,0047	0,0002
	3	0,9998	0,9973	0,9667	0,9294	0,8740	0,7102	0,5000	0,2898	0,1260	0,0333	0,0027
	4	1,0000	0,9998	0,9953	0,9871	0,9712	0,9037	0,7734	0,5801	0,3529	0,1480	0,0257
	5	1,0000	1,0000	0,9996	0,9987	0,9962	0,9812	0,9375	0,8414	0,6706	0,4233	0,1497
	6	1,0000	1,0000	1,0000	0,9999	0,9998	0,9984	0,9922	0,9720	0,9176	0,7903	0,5217
	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
8	**0**	0,6634	0,4305	0,1678	0,1001	0,0576	0,0168	0,0039	0,0007	0,0001	0,0000	0,0000
	1	0,9428	0,8131	0,5033	0,3671	0,2553	0,1064	0,0352	0,0085	0,0013	0,0001	0,0000
	2	0,9942	0,9619	0,7969	0,6785	0,5518	0,3154	0,1445	0,0498	0,0113	0,0012	0,0000
	3	0,9996	0,9950	0,9437	0,8862	0,8059	0,5941	0,3633	0,1737	0,0580	0,0104	0,0004
	4	1,0000	0,9996	0,9896	0,9727	0,9420	0,8263	0,6367	0,4059	0,1941	0,0563	0,0050
	5	1,0000	1,0000	0,9988	0,9958	0,9887	0,9502	0,8555	0,6846	0,4482	0,2031	0,0381
	6	1,0000	1,0000	0,9999	0,9996	0,9987	0,9915	0,9648	0,8936	0,7447	0,4967	0,1869
	7	1,0000	1,0000	1,0000	1,0000	0,9999	0,9993	0,9961	0,9832	0,9424	0,8322	0,5695
	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Distribución Acumulada de la Binomial: B(k;n,p)

		p										
n	**k**	**0,05**	**0,1**	**0,2**	**0,25**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**
9	**0**	0,6302	0,3874	0,1342	0,0751	0,0404	0,0101	0,0020	0,0003	0,0000	0,0000	0,0000
	1	0,9288	0,7748	0,4362	0,3003	0,1960	0,0705	0,0195	0,0038	0,0004	0,0000	0,0000
	2	0,9916	0,9470	0,7382	0,6007	0,4628	0,2318	0,0898	0,0250	0,0043	0,0003	0,0000
	3	0,9994	0,9917	0,9144	0,8343	0,7297	0,4826	0,2539	0,0994	0,0253	0,0031	0,0001
	4	1,0000	0,9991	0,9804	0,9511	0,9012	0,7334	0,5000	0,2666	0,0988	0,0196	0,0009
	5	1,0000	0,9999	0,9969	0,9900	0,9747	0,9006	0,7461	0,5174	0,2703	0,0856	0,0083
	6	1,0000	1,0000	0,9997	0,9987	0,9957	0,9750	0,9102	0,7682	0,5372	0,2618	0,0530
	7	1,0000	1,0000	1,0000	0,9999	0,9996	0,9962	0,9805	0,9295	0,8040	0,5638	0,2252
	8	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9980	0,9899	0,9596	0,8658	0,6126
	9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
10	**0**	0,5987	0,3487	0,1074	0,0563	0,0282	0,0060	0,0010	0,0001	0,0000	0,0000	0,0000
	1	0,9139	0,7361	0,3758	0,2440	0,1493	0,0464	0,0107	0,0017	0,0001	0,0000	0,0000
	2	0,9885	0,9298	0,6778	0,5256	0,3828	0,1673	0,0547	0,0123	0,0016	0,0001	0,0000
	3	0,9990	0,9872	0,8791	0,7759	0,6496	0,3823	0,1719	0,0548	0,0106	0,0009	0,0000
	4	0,9999	0,9984	0,9672	0,9219	0,8497	0,6331	0,3770	0,1662	0,0473	0,0064	0,0001
	5	1,0000	0,9999	0,9936	0,9803	0,9527	0,8338	0,6230	0,3669	0,1503	0,0328	0,0016
	6	1,0000	1,0000	0,9991	0,9965	0,9894	0,9452	0,8281	0,6177	0,3504	0,1209	0,0128
	7	1,0000	1,0000	0,9999	0,9996	0,9984	0,9877	0,9453	0,8327	0,6172	0,3222	0,0702
	8	1,0000	1,0000	1,0000	1,0000	0,9999	0,9983	0,9893	0,9536	0,8507	0,6242	0,2639
	9	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9990	0,9940	0,9718	0,8926	0,6513
	10	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
11	**0**	0,5688	0,3138	0,0859	0,0422	0,0198	0,0036	0,0005	0,0000	0,0000	0,0000	0,0000
	1	0,8981	0,6974	0,3221	0,1971	0,1130	0,0302	0,0059	0,0007	0,0000	0,0000	0,0000
	2	0,9848	0,9104	0,6174	0,4552	0,3127	0,1189	0,0327	0,0059	0,0006	0,0000	0,0000
	3	0,9984	0,9815	0,8389	0,7133	0,5696	0,2963	0,1133	0,0293	0,0043	0,0002	0,0000
	4	0,9999	0,9972	0,9496	0,8854	0,7897	0,5328	0,2744	0,0994	0,0216	0,0020	0,0000
	5	1,0000	0,9997	0,9883	0,9657	0,9218	0,7535	0,5000	0,2465	0,0782	0,0117	0,0003
	6	1,0000	1,0000	0,9980	0,9924	0,9784	0,9006	0,7256	0,4672	0,2103	0,0504	0,0028
	7	1,0000	1,0000	0,9998	0,9988	0,9957	0,9707	0,8867	0,7037	0,4304	0,1611	0,0185
	8	1,0000	1,0000	1,0000	0,9999	0,9994	0,9941	0,9673	0,8811	0,6873	0,3826	0,0896
	9	1,0000	1,0000	1,0000	1,0000	1,0000	0,9993	0,9941	0,9698	0,8870	0,6779	0,3026
	10	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9995	0,9964	0,9802	0,9141	0,6862
	11	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
12	**0**	0,5404	0,2824	0,0687	0,0317	0,0138	0,0022	0,0002	0,0000	0,0000	0,0000	0,0000
	1	0,8816	0,6590	0,2749	0,1584	0,0850	0,0196	0,0032	0,0003	0,0000	0,0000	0,0000
	2	0,9804	0,8891	0,5583	0,3907	0,2528	0,0834	0,0193	0,0028	0,0002	0,0000	0,0000
	3	0,9978	0,9744	0,7946	0,6488	0,4925	0,2253	0,0730	0,0153	0,0017	0,0001	0,0000
	4	0,9998	0,9957	0,9274	0,8424	0,7237	0,4382	0,1938	0,0573	0,0095	0,0006	0,0000
	5	1,0000	0,9995	0,9806	0,9456	0,8822	0,6652	0,3872	0,1582	0,0386	0,0039	0,0001
	6	1,0000	0,9999	0,9961	0,9857	0,9614	0,8418	0,6128	0,3348	0,1178	0,0194	0,0005
	7	1,0000	1,0000	0,9994	0,9972	0,9905	0,9427	0,8062	0,5618	0,2763	0,0726	0,0043
	8	1,0000	1,0000	0,9999	0,9996	0,9983	0,9847	0,9270	0,7747	0,5075	0,2054	0,0256
	9	1,0000	1,0000	1,0000	1,0000	0,9998	0,9972	0,9807	0,9166	0,7472	0,4417	0,1109
	10	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9968	0,9804	0,9150	0,7251	0,3410
	11	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9978	0,9862	0,9313	0,7176
	12	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Distribución Acumulada de la Binomial: B(k;n,p)

		p										
n	**k**	**0,05**	**0,1**	**0,2**	**0,25**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**
13	**0**	0,5133	0,2542	0,0550	0,0238	0,0097	0,0013	0,0001	0,0000	0,0000	0,0000	0,0000
	1	0,8646	0,6213	0,2336	0,1267	0,0637	0,0126	0,0017	0,0001	0,0000	0,0000	0,0000
	2	0,9755	0,8661	0,5017	0,3326	0,2025	0,0579	0,0112	0,0013	0,0001	0,0000	0,0000
	3	0,9969	0,9658	0,7473	0,5843	0,4206	0,1686	0,0461	0,0078	0,0007	0,0000	0,0000
	4	0,9997	0,9935	0,9009	0,7940	0,6543	0,3530	0,1334	0,0321	0,0040	0,0002	0,0000
	5	1,0000	0,9991	0,9700	0,9198	0,8346	0,5744	0,2905	0,0977	0,0182	0,0012	0,0000
	6	1,0000	0,9999	0,9930	0,9757	0,9376	0,7712	0,5000	0,2288	0,0624	0,0070	0,0001
	7	1,0000	1,0000	0,9988	0,9944	0,9818	0,9023	0,7095	0,4256	0,1654	0,0300	0,0009
	8	1,0000	1,0000	0,9998	0,9990	0,9960	0,9679	0,8666	0,6470	0,3457	0,0991	0,0065
	9	1,0000	1,0000	1,0000	0,9999	0,9993	0,9922	0,9539	0,8314	0,5794	0,2527	0,0342
	10	1,0000	1,0000	1,0000	1,0000	0,9999	0,9987	0,9888	0,9421	0,7975	0,4983	0,1339
	11	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9983	0,9874	0,9363	0,7664	0,3787
	12	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9987	0,9903	0,9450	0,7458
	13	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
14	**0**	0,4877	0,2288	0,0440	0,0178	0,0068	0,0008	0,0001	0,0000	0,0000	0,0000	0,0000
	1	0,8470	0,5846	0,1979	0,1010	0,0475	0,0081	0,0009	0,0001	0,0000	0,0000	0,0000
	2	0,9699	0,8416	0,4481	0,2811	0,1608	0,0398	0,0065	0,0006	0,0000	0,0000	0,0000
	3	0,9958	0,9559	0,6982	0,5213	0,3552	0,1243	0,0287	0,0039	0,0002	0,0000	0,0000
	4	0,9996	0,9908	0,8702	0,7415	0,5842	0,2793	0,0898	0,0175	0,0017	0,0000	0,0000
	5	1,0000	0,9985	0,9561	0,8883	0,7805	0,4859	0,2120	0,0583	0,0083	0,0004	0,0000
	6	1,0000	0,9998	0,9884	0,9617	0,9067	0,6925	0,3953	0,1501	0,0315	0,0024	0,0000
	7	1,0000	1,0000	0,9976	0,9897	0,9685	0,8499	0,6047	0,3075	0,0933	0,0116	0,0002
	8	1,0000	1,0000	0,9996	0,9978	0,9917	0,9417	0,7880	0,5141	0,2195	0,0439	0,0015
	9	1,0000	1,0000	1,0000	0,9997	0,9983	0,9825	0,9102	0,7207	0,4158	0,1298	0,0092
	10	1,0000	1,0000	1,0000	1,0000	0,9998	0,9961	0,9713	0,8757	0,6448	0,3018	0,0441
	11	1,0000	1,0000	1,0000	1,0000	1,0000	0,9994	0,9935	0,9602	0,8392	0,5519	0,1584
	12	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9991	0,9919	0,9525	0,8021	0,4154
	13	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9992	0,9932	0,9560	0,7712
	14	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
15	**0**	0,4633	0,2059	0,0352	0,0134	0,0047	0,0005	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,8290	0,5490	0,1671	0,0802	0,0353	0,0052	0,0005	0,0000	0,0000	0,0000	0,0000
	2	0,9638	0,8159	0,3980	0,2361	0,1268	0,0271	0,0037	0,0003	0,0000	0,0000	0,0000
	3	0,9945	0,9444	0,6482	0,4613	0,2969	0,0905	0,0176	0,0019	0,0001	0,0000	0,0000
	4	0,9994	0,9873	0,8358	0,6865	0,5155	0,2173	0,0592	0,0093	0,0007	0,0000	0,0000
	5	0,9999	0,9978	0,9389	0,8516	0,7216	0,4032	0,1509	0,0338	0,0037	0,0001	0,0000
	6	1,0000	0,9997	0,9819	0,9434	0,8689	0,6098	0,3036	0,0950	0,0152	0,0008	0,0000
	7	1,0000	1,0000	0,9958	0,9827	0,9500	0,7869	0,5000	0,2131	0,0500	0,0042	0,0000
	8	1,0000	1,0000	0,9992	0,9958	0,9848	0,9050	0,6964	0,3902	0,1311	0,0181	0,0003
	9	1,0000	1,0000	0,9999	0,9992	0,9963	0,9662	0,8491	0,5968	0,2784	0,0611	0,0022
	10	1,0000	1,0000	1,0000	0,9999	0,9993	0,9907	0,9408	0,7827	0,4845	0,1642	0,0127
	11	1,0000	1,0000	1,0000	1,0000	0,9999	0,9981	0,9824	0,9095	0,7031	0,3518	0,0556
	12	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9963	0,9729	0,8732	0,6020	0,1841
	13	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9995	0,9948	0,9647	0,8329	0,4510
	14	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9995	0,9953	0,9648	0,7941
	15	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Distribución Acumulada de la Binomial: B(k;n,p)

		p										
n	**k**	**0,05**	**0,1**	**0,2**	**0,25**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**
16	**0**	0,4401	0,1853	0,0281	0,0100	0,0033	0,0003	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,8108	0,5147	0,1407	0,0635	0,0261	0,0033	0,0003	0,0000	0,0000	0,0000	0,0000
	2	0,9571	0,7892	0,3518	0,1971	0,0994	0,0183	0,0021	0,0001	0,0000	0,0000	0,0000
	3	0,9930	0,9316	0,5981	0,4050	0,2459	0,0651	0,0106	0,0009	0,0000	0,0000	0,0000
	4	0,9991	0,9830	0,7982	0,6302	0,4499	0,1666	0,0384	0,0049	0,0003	0,0000	0,0000
	5	0,9999	0,9967	0,9183	0,8103	0,6598	0,3288	0,1051	0,0191	0,0016	0,0000	0,0000
	6	1,0000	0,9995	0,9733	0,9204	0,8247	0,5272	0,2272	0,0583	0,0071	0,0002	0,0000
	7	1,0000	0,9999	0,9930	0,9729	0,9256	0,7161	0,4018	0,1423	0,0257	0,0015	0,0000
	8	1,0000	1,0000	0,9985	0,9925	0,9743	0,8577	0,5982	0,2839	0,0744	0,0070	0,0001
	9	1,0000	1,0000	0,9998	0,9984	0,9929	0,9417	0,7728	0,4728	0,1753	0,0267	0,0005
	10	1,0000	1,0000	1,0000	0,9997	0,9984	0,9809	0,8949	0,6712	0,3402	0,0817	0,0033
	11	1,0000	1,0000	1,0000	1,0000	0,9997	0,9951	0,9616	0,8334	0,5501	0,2018	0,0170
	12	1,0000	1,0000	1,0000	1,0000	1,0000	0,9991	0,9894	0,9349	0,7541	0,4019	0,0684
	13	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9979	0,9817	0,9006	0,6482	0,2108
	14	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9967	0,9739	0,8593	0,4853
	15	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9967	0,9719	0,8147
	16	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
17	**0**	0,4181	0,1668	0,0225	0,0075	0,0023	0,0002	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,7922	0,4818	0,1182	0,0501	0,0193	0,0021	0,0001	0,0000	0,0000	0,0000	0,0000
	2	0,9497	0,7618	0,3096	0,1637	0,0774	0,0123	0,0012	0,0001	0,0000	0,0000	0,0000
	3	0,9912	0,9174	0,5489	0,3530	0,2019	0,0464	0,0064	0,0005	0,0000	0,0000	0,0000
	4	0,9988	0,9779	0,7582	0,5739	0,3887	0,1260	0,0245	0,0025	0,0001	0,0000	0,0000
	5	0,9999	0,9953	0,8943	0,7653	0,5968	0,2639	0,0717	0,0106	0,0007	0,0000	0,0000
	6	1,0000	0,9992	0,9623	0,8929	0,7752	0,4478	0,1662	0,0348	0,0032	0,0001	0,0000
	7	1,0000	0,9999	0,9891	0,9598	0,8954	0,6405	0,3145	0,0919	0,0127	0,0005	0,0000
	8	1,0000	1,0000	0,9974	0,9876	0,9597	0,8011	0,5000	0,1989	0,0403	0,0026	0,0000
	9	1,0000	1,0000	0,9995	0,9969	0,9873	0,9081	0,6855	0,3595	0,1046	0,0109	0,0001
	10	1,0000	1,0000	0,9999	0,9994	0,9968	0,9652	0,8338	0,5522	0,2248	0,0377	0,0008
	11	1,0000	1,0000	1,0000	0,9999	0,9993	0,9894	0,9283	0,7361	0,4032	0,1057	0,0047
	12	1,0000	1,0000	1,0000	1,0000	0,9999	0,9975	0,9755	0,8740	0,6113	0,2418	0,0221
	13	1,0000	1,0000	1,0000	1,0000	1,0000	0,9995	0,9936	0,9536	0,7981	0,4511	0,0826
	14	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9988	0,9877	0,9226	0,6904	0,2382
	15	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9979	0,9807	0,8818	0,5182
	16	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9977	0,9775	0,8332
	17	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
18	**0**	0,3972	0,1501	0,0180	0,0056	0,0016	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,7735	0,4503	0,0991	0,0395	0,0142	0,0013	0,0001	0,0000	0,0000	0,0000	0,0000
	2	0,9419	0,7338	0,2713	0,1353	0,0600	0,0082	0,0007	0,0000	0,0000	0,0000	0,0000
	3	0,9891	0,9018	0,5010	0,3057	0,1646	0,0328	0,0038	0,0002	0,0000	0,0000	0,0000
	4	0,9985	0,9718	0,7164	0,5187	0,3327	0,0942	0,0154	0,0013	0,0000	0,0000	0,0000
	5	0,9998	0,9936	0,8671	0,7175	0,5344	0,2088	0,0481	0,0058	0,0003	0,0000	0,0000
	6	1,0000	0,9988	0,9487	0,8610	0,7217	0,3743	0,1189	0,0203	0,0014	0,0000	0,0000
	7	1,0000	0,9998	0,9837	0,9431	0,8593	0,5634	0,2403	0,0576	0,0061	0,0002	0,0000
	8	1,0000	1,0000	0,9957	0,9807	0,9404	0,7368	0,4073	0,1347	0,0210	0,0009	0,0000
	9	1,0000	1,0000	0,9991	0,9946	0,9790	0,8653	0,5927	0,2632	0,0596	0,0043	0,0000

Distribución Acumulada de la Binomial: B(k;n,p)

		p										
n	**k**	**0,05**	**0,1**	**0,2**	**0,25**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**
18	**10**	1,0000	1,0000	0,9998	0,9988	0,9939	0,9424	0,7597	0,4366	0,1407	0,0163	0,0002
	11	1,0000	1,0000	1,0000	0,9998	0,9986	0,9797	0,8811	0,6257	0,2783	0,0513	0,0012
	12	1,0000	1,0000	1,0000	1,0000	0,9997	0,9942	0,9519	0,7912	0,4656	0,1329	0,0064
	13	1,0000	1,0000	1,0000	1,0000	1,0000	0,9987	0,9846	0,9058	0,6673	0,2836	0,0282
	14	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9962	0,9672	0,8354	0,4990	0,0982
	15	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9993	0,9918	0,9400	0,7287	0,2662
	16	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9987	0,9858	0,9009	0,5497
	17	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9984	0,9820	0,8499
	18	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
19	**0**	0,3774	0,1351	0,0144	0,0042	0,0011	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,7547	0,4203	0,0829	0,0310	0,0104	0,0008	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,9335	0,7054	0,2369	0,1113	0,0462	0,0055	0,0004	0,0000	0,0000	0,0000	0,0000
	3	0,9868	0,8850	0,4551	0,2631	0,1332	0,0230	0,0022	0,0001	0,0000	0,0000	0,0000
	4	0,9980	0,9648	0,6733	0,4654	0,2822	0,0696	0,0096	0,0006	0,0000	0,0000	0,0000
	5	0,9998	0,9914	0,8369	0,6678	0,4739	0,1629	0,0318	0,0031	0,0001	0,0000	0,0000
	6	1,0000	0,9983	0,9324	0,8251	0,6655	0,3081	0,0835	0,0116	0,0006	0,0000	0,0000
	7	1,0000	0,9997	0,9767	0,9225	0,8180	0,4878	0,1796	0,0352	0,0028	0,0000	0,0000
	8	1,0000	1,0000	0,9933	0,9713	0,9161	0,6675	0,3238	0,0885	0,0105	0,0003	0,0000
	9	1,0000	1,0000	0,9984	0,9911	0,9674	0,8139	0,5000	0,1861	0,0326	0,0016	0,0000
	10	1,0000	1,0000	0,9997	0,9977	0,9895	0,9115	0,6762	0,3325	0,0839	0,0067	0,0000
	11	1,0000	1,0000	1,0000	0,9995	0,9972	0,9648	0,8204	0,5122	0,1820	0,0233	0,0003
	12	1,0000	1,0000	1,0000	0,9999	0,9994	0,9884	0,9165	0,6919	0,3345	0,0676	0,0017
	13	1,0000	1,0000	1,0000	1,0000	0,9999	0,9969	0,9682	0,8371	0,5261	0,1631	0,0086
	14	1,0000	1,0000	1,0000	1,0000	1,0000	0,9994	0,9904	0,9304	0,7178	0,3267	0,0352
	15	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9978	0,9770	0,8668	0,5449	0,1150
	16	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9996	0,9945	0,9538	0,7631	0,2946
	17	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9992	0,9896	0,9171	0,5797
	18	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9989	0,9856	0,8649
	19	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
20	**0**	0,3585	0,1216	0,0115	0,0032	0,0008	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,7358	0,3917	0,0692	0,0243	0,0076	0,0005	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,9245	0,6769	0,2061	0,0913	0,0355	0,0036	0,0002	0,0000	0,0000	0,0000	0,0000
	3	0,9841	0,8670	0,4114	0,2252	0,1071	0,0160	0,0013	0,0000	0,0000	0,0000	0,0000
	4	0,9974	0,9568	0,6296	0,4148	0,2375	0,0510	0,0059	0,0003	0,0000	0,0000	0,0000
	5	0,9997	0,9887	0,8042	0,6172	0,4164	0,1256	0,0207	0,0016	0,0000	0,0000	0,0000
	6	1,0000	0,9976	0,9133	0,7858	0,6080	0,2500	0,0577	0,0065	0,0003	0,0000	0,0000
	7	1,0000	0,9996	0,9679	0,8982	0,7723	0,4159	0,1316	0,0210	0,0013	0,0000	0,0000
	8	1,0000	0,9999	0,9900	0,9591	0,8867	0,5956	0,2517	0,0565	0,0051	0,0001	0,0000
	9	1,0000	1,0000	0,9974	0,9861	0,9520	0,7553	0,4119	0,1275	0,0171	0,0006	0,0000
	10	1,0000	1,0000	0,9994	0,9961	0,9829	0,8725	0,5881	0,2447	0,0480	0,0026	0,0000
	11	1,0000	1,0000	0,9999	0,9991	0,9949	0,9435	0,7483	0,4044	0,1133	0,0100	0,0001
	12	1,0000	1,0000	1,0000	0,9998	0,9987	0,9790	0,8684	0,5841	0,2277	0,0321	0,0004
	13	1,0000	1,0000	1,0000	1,0000	0,9997	0,9935	0,9423	0,7500	0,3920	0,0867	0,0024
	14	1,0000	1,0000	1,0000	1,0000	1,0000	0,9984	0,9793	0,8744	0,5836	0,1958	0,0113
	15	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9941	0,9490	0,7625	0,3704	0,0432

Distribución Acumulada de la Binomial: B(k;n,p)

		p										
n	**k**	**0,05**	**0,1**	**0,2**	**0,25**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**
20	**16**	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9987	0,9840	0,8929	0,5886	0,1330
	17	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9964	0,9645	0,7939	0,3231
	18	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9995	0,9924	0,9308	0,6083
	19	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9992	0,9885	0,8784
	20	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
21	**0**	0,3406	0,1094	0,0092	0,0024	0,0006	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,7170	0,3647	0,0576	0,0190	0,0056	0,0003	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,9151	0,6484	0,1787	0,0745	0,0271	0,0024	0,0001	0,0000	0,0000	0,0000	0,0000
	3	0,9811	0,8480	0,3704	0,1917	0,0856	0,0110	0,0007	0,0000	0,0000	0,0000	0,0000
	4	0,9968	0,9478	0,5860	0,3674	0,1984	0,0370	0,0036	0,0002	0,0000	0,0000	0,0000
	5	0,9996	0,9856	0,7693	0,5666	0,3627	0,0957	0,0133	0,0008	0,0000	0,0000	0,0000
	6	1,0000	0,9967	0,8915	0,7436	0,5505	0,2002	0,0392	0,0036	0,0001	0,0000	0,0000
	7	1,0000	0,9994	0,9569	0,8701	0,7230	0,3495	0,0946	0,0123	0,0006	0,0000	0,0000
	8	1,0000	0,9999	0,9856	0,9439	0,8523	0,5237	0,1917	0,0352	0,0024	0,0000	0,0000
	9	1,0000	1,0000	0,9959	0,9794	0,9324	0,6914	0,3318	0,0849	0,0087	0,0002	0,0000
	10	1,0000	1,0000	0,9990	0,9936	0,9736	0,8256	0,5000	0,1744	0,0264	0,0010	0,0000
	11	1,0000	1,0000	0,9998	0,9983	0,9913	0,9151	0,6682	0,3086	0,0676	0,0041	0,0000
	12	1,0000	1,0000	1,0000	0,9996	0,9976	0,9648	0,8083	0,4763	0,1477	0,0144	0,0001
	13	1,0000	1,0000	1,0000	0,9999	0,9994	0,9877	0,9054	0,6505	0,2770	0,0431	0,0006
	14	1,0000	1,0000	1,0000	1,0000	0,9999	0,9964	0,9608	0,7998	0,4495	0,1085	0,0033
	15	1,0000	1,0000	1,0000	1,0000	1,0000	0,9992	0,9867	0,9043	0,6373	0,2307	0,0144
	16	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9964	0,9630	0,8016	0,4140	0,0522
	17	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9993	0,9890	0,9144	0,6296	0,1520
	18	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9976	0,9729	0,8213	0,3516
	19	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9944	0,9424	0,6353
	20	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9994	0,9908	0,8906
	21	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
22	**0**	0,3235	0,0985	0,0074	0,0018	0,0004	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,6982	0,3392	0,0480	0,0149	0,0041	0,0002	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,9052	0,6200	0,1545	0,0606	0,0207	0,0016	0,0001	0,0000	0,0000	0,0000	0,0000
	3	0,9778	0,8281	0,3320	0,1624	0,0681	0,0076	0,0004	0,0000	0,0000	0,0000	0,0000
	4	0,9960	0,9379	0,5429	0,3235	0,1645	0,0266	0,0022	0,0001	0,0000	0,0000	0,0000
	5	0,9994	0,9818	0,7326	0,5168	0,3134	0,0722	0,0085	0,0004	0,0000	0,0000	0,0000
	6	0,9999	0,9956	0,8670	0,6994	0,4942	0,1584	0,0262	0,0019	0,0000	0,0000	0,0000
	7	1,0000	0,9991	0,9439	0,8385	0,6713	0,2898	0,0669	0,0070	0,0002	0,0000	0,0000
	8	1,0000	0,9999	0,9799	0,9254	0,8135	0,4540	0,1431	0,0215	0,0011	0,0000	0,0000
	9	1,0000	1,0000	0,9939	0,9705	0,9084	0,6244	0,2617	0,0551	0,0043	0,0001	0,0000
	10	1,0000	1,0000	0,9984	0,9900	0,9613	0,7720	0,4159	0,1207	0,0140	0,0003	0,0000
	11	1,0000	1,0000	0,9997	0,9971	0,9860	0,8793	0,5841	0,2280	0,0387	0,0016	0,0000
	12	1,0000	1,0000	0,9999	0,9993	0,9957	0,9449	0,7383	0,3756	0,0916	0,0061	0,0000
	13	1,0000	1,0000	1,0000	0,9999	0,9989	0,9785	0,8569	0,5460	0,1865	0,0201	0,0001
	14	1,0000	1,0000	1,0000	1,0000	0,9998	0,9930	0,9331	0,7102	0,3287	0,0561	0,0009
	15	1,0000	1,0000	1,0000	1,0000	1,0000	0,9981	0,9738	0,8416	0,5058	0,1330	0,0044
	16	1,0000	1,0000	1,0000	1,0000	1,0000	0,9996	0,9915	0,9278	0,6866	0,2674	0,0182
	17	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9978	0,9734	0,8355	0,4571	0,0621

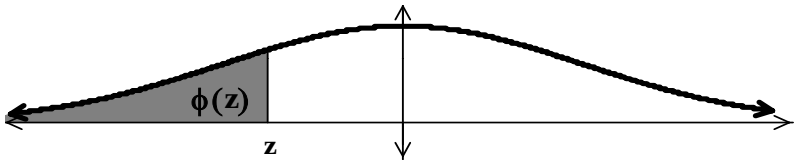
Distribución Acumulada de la Binomial: B(k;n,p)

		p										
n	**k**	**0,05**	**0,1**	**0,2**	**0,25**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**
22	**18**	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9996	0,9924	0,9319	0,6680	0,1719
	19	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9984	0,9793	0,8455	0,3800
	20	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9959	0,9520	0,6608
	21	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9996	0,9926	0,9015
	22	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
23	**0**	0,3074	0,0886	0,0059	0,0013	0,0003	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,6794	0,3151	0,0398	0,0116	0,0030	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,8948	0,5920	0,1332	0,0492	0,0157	0,0010	0,0000	0,0000	0,0000	0,0000	0,0000
	3	0,9742	0,8073	0,2965	0,1370	0,0538	0,0052	0,0002	0,0000	0,0000	0,0000	0,0000
	4	0,9951	0,9269	0,5007	0,2832	0,1356	0,0190	0,0013	0,0000	0,0000	0,0000	0,0000
	5	0,9992	0,9774	0,6947	0,4685	0,2688	0,0540	0,0053	0,0002	0,0000	0,0000	0,0000
	6	0,9999	0,9942	0,8402	0,6537	0,4399	0,1240	0,0173	0,0010	0,0000	0,0000	0,0000
	7	1,0000	0,9988	0,9285	0,8037	0,6181	0,2373	0,0466	0,0040	0,0001	0,0000	0,0000
	8	1,0000	0,9998	0,9727	0,9037	0,7709	0,3884	0,1050	0,0128	0,0005	0,0000	0,0000
	9	1,0000	1,0000	0,9911	0,9592	0,8799	0,5562	0,2024	0,0349	0,0021	0,0000	0,0000
	10	1,0000	1,0000	0,9975	0,9851	0,9454	0,7129	0,3388	0,0813	0,0072	0,0001	0,0000
	11	1,0000	1,0000	0,9994	0,9954	0,9786	0,8364	0,5000	0,1636	0,0214	0,0006	0,0000
	12	1,0000	1,0000	0,9999	0,9988	0,9928	0,9187	0,6612	0,2871	0,0546	0,0025	0,0000
	13	1,0000	1,0000	1,0000	0,9997	0,9979	0,9651	0,7976	0,4438	0,1201	0,0089	0,0000
	14	1,0000	1,0000	1,0000	0,9999	0,9995	0,9872	0,8950	0,6116	0,2291	0,0273	0,0002
	15	1,0000	1,0000	1,0000	1,0000	0,9999	0,9960	0,9534	0,7627	0,3819	0,0715	0,0012
	16	1,0000	1,0000	1,0000	1,0000	1,0000	0,9990	0,9827	0,8760	0,5601	0,1598	0,0058
	17	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9947	0,9460	0,7312	0,3053	0,0226
	18	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9987	0,9810	0,8644	0,4993	0,0731
	19	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9948	0,9462	0,7035	0,1927
	20	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9990	0,9843	0,8668	0,4080
	21	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9970	0,9602	0,6849
	22	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9941	0,9114
	23	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
24	**0**	0,2920	0,0798	0,0047	0,0010	0,0002	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,6608	0,2925	0,0331	0,0090	0,0022	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,8841	0,5643	0,1145	0,0398	0,0119	0,0007	0,0000	0,0000	0,0000	0,0000	0,0000
	3	0,9702	0,7857	0,2639	0,1150	0,0424	0,0035	0,0001	0,0000	0,0000	0,0000	0,0000
	4	0,9940	0,9149	0,4599	0,2466	0,1111	0,0134	0,0008	0,0000	0,0000	0,0000	0,0000
	5	0,9990	0,9723	0,6559	0,4222	0,2288	0,0400	0,0033	0,0001	0,0000	0,0000	0,0000
	6	0,9999	0,9925	0,8111	0,6074	0,3886	0,0960	0,0113	0,0005	0,0000	0,0000	0,0000
	7	1,0000	0,9983	0,9108	0,7662	0,5647	0,1919	0,0320	0,0022	0,0000	0,0000	0,0000
	8	1,0000	0,9997	0,9638	0,8787	0,7250	0,3279	0,0758	0,0075	0,0002	0,0000	0,0000
	9	1,0000	0,9999	0,9874	0,9453	0,8472	0,4891	0,1537	0,0217	0,0010	0,0000	0,0000
	10	1,0000	1,0000	0,9962	0,9787	0,9258	0,6502	0,2706	0,0535	0,0036	0,0000	0,0000
	11	1,0000	1,0000	0,9990	0,9928	0,9686	0,7870	0,4194	0,1143	0,0115	0,0002	0,0000
	12	1,0000	1,0000	0,9998	0,9979	0,9885	0,8857	0,5806	0,2130	0,0314	0,0010	0,0000
	13	1,0000	1,0000	1,0000	0,9995	0,9964	0,9465	0,7294	0,3498	0,0742	0,0038	0,0000
	14	1,0000	1,0000	1,0000	0,9999	0,9990	0,9783	0,8463	0,5109	0,1528	0,0126	0,0001
	15	1,0000	1,0000	1,0000	1,0000	0,9998	0,9925	0,9242	0,6721	0,2750	0,0362	0,0003

Distribución Acumulada de la Binomial: B(k;n,p)

		p										
n	**k**	**0,05**	**0,1**	**0,2**	**0,25**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**
24	**16**	1,0000	1,0000	1,0000	1,0000	1,0000	0,9978	0,9680	0,8081	0,4353	0,0892	0,0017
	17	1,0000	1,0000	1,0000	1,0000	1,0000	0,9995	0,9887	0,9040	0,6114	0,1889	0,0075
	18	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9967	0,9600	0,7712	0,3441	0,0277
	19	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9992	0,9866	0,8889	0,5401	0,0851
	20	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9965	0,9576	0,7361	0,2143
	21	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9993	0,9881	0,8855	0,4357
	22	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9978	0,9669	0,7075
	23	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9953	0,9202
	24	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
25	**0**	0,2774	0,0718	0,0038	0,0008	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,6424	0,2712	0,0274	0,0070	0,0016	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,8729	0,5371	0,0982	0,0321	0,0090	0,0004	0,0000	0,0000	0,0000	0,0000	0,0000
	3	0,9659	0,7636	0,2340	0,0962	0,0332	0,0024	0,0001	0,0000	0,0000	0,0000	0,0000
	4	0,9928	0,9020	0,4207	0,2137	0,0905	0,0095	0,0005	0,0000	0,0000	0,0000	0,0000
	5	0,9988	0,9666	0,6167	0,3783	0,1935	0,0294	0,0020	0,0001	0,0000	0,0000	0,0000
	6	0,9998	0,9905	0,7800	0,5611	0,3407	0,0736	0,0073	0,0003	0,0000	0,0000	0,0000
	7	1,0000	0,9977	0,8909	0,7265	0,5118	0,1536	0,0216	0,0012	0,0000	0,0000	0,0000
	8	1,0000	0,9995	0,9532	0,8506	0,6769	0,2735	0,0539	0,0043	0,0001	0,0000	0,0000
	9	1,0000	0,9999	0,9827	0,9287	0,8106	0,4246	0,1148	0,0132	0,0005	0,0000	0,0000
	10	1,0000	1,0000	0,9944	0,9703	0,9022	0,5858	0,2122	0,0344	0,0018	0,0000	0,0000
	11	1,0000	1,0000	0,9985	0,9893	0,9558	0,7323	0,3450	0,0778	0,0060	0,0001	0,0000
	12	1,0000	1,0000	0,9996	0,9966	0,9825	0,8462	0,5000	0,1538	0,0175	0,0004	0,0000
	13	1,0000	1,0000	0,9999	0,9991	0,9940	0,9222	0,6550	0,2677	0,0442	0,0015	0,0000
	14	1,0000	1,0000	1,0000	0,9998	0,9982	0,9656	0,7878	0,4142	0,0978	0,0056	0,0000
	15	1,0000	1,0000	1,0000	1,0000	0,9995	0,9868	0,8852	0,5754	0,1894	0,0173	0,0001
	16	1,0000	1,0000	1,0000	1,0000	0,9999	0,9957	0,9461	0,7265	0,3231	0,0468	0,0005
	17	1,0000	1,0000	1,0000	1,0000	1,0000	0,9988	0,9784	0,8464	0,4882	0,1091	0,0023
	18	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9927	0,9264	0,6593	0,2200	0,0095
	19	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9980	0,9706	0,8065	0,3833	0,0334
	20	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9995	0,9905	0,9095	0,5793	0,0980
	21	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9976	0,9668	0,7660	0,2364
	22	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9996	0,9910	0,9018	0,4629
	23	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9984	0,9726	0,7288
	24	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9962	0,9282
	25	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Distribución Acumulada de la Normal Estandar: $\phi(z)$



z	0,00	0,005	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
-3,9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
-3,8	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
-3,7	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
-3,6	0,0002	0,0002	0,0002	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
-3,5	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002
-3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002
-3,3	0,0005	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
-3,2	0,0007	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
-3,1	0,0010	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
-3,0	0,0013	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
-2,9	0,0019	0,0018	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
-2,8	0,0026	0,0025	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
-2,7	0,0035	0,0034	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
-2,6	0,0047	0,0046	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
-2,5	0,0062	0,0061	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
-2,4	0,0082	0,0081	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
-2,3	0,0107	0,0106	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
-2,2	0,0139	0,0137	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
-2,1	0,0179	0,0176	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
-2,0	0,0228	0,0225	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
-1,9	0,0287	0,0284	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
-1,8	0,0359	0,0355	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
-1,7	0,0446	0,0441	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
-1,6	0,0548	0,0542	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
-1,5	0,0668	0,0662	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
-1,4	0,0808	0,0800	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
-1,3	0,0968	0,0959	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
-1,2	0,1151	0,1141	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
-1,1	0,1357	0,1346	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
-1,0	0,1587	0,1574	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
-0,9	0,1841	0,1827	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
-0,8	0,2119	0,2104	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
-0,7	0,2420	0,2404	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
-0,6	0,2743	0,2726	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
-0,5	0,3085	0,3068	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
-0,4	0,3446	0,3427	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
-0,3	0,3821	0,3802	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
-0,2	0,4207	0,4188	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
-0,1	0,4602	0,4582	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
-0,0	0,5000	0,4980	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641

Distribución Acumulada t : $P(t \leq t_{\alpha,v}) = \alpha$ (Los valores de la tabla son en valor absoluto)

	α										
v	**0,005**	**0,010**	**0,020**	**0,025**	**0,030**	**0,040**	**0,050**	**0,100**	**0,200**	**0,300**	**0,400**
1	63,6567	31,8205	15,8945	12,7062	10,5789	7,91582	6,31375	3,07768	1,37638	0,72654	0,32492
2	9,92484	6,96456	4,84873	4,30265	3,89643	3,31976	2,91999	1,88562	1,06066	0,61721	0,28868
3	5,84091	4,5407	3,48191	3,18245	2,95051	2,60543	2,35336	1,63774	0,97847	0,58439	0,27667
4	4,60409	3,74695	2,99853	2,77645	2,60076	2,33287	2,13185	1,53321	0,94096	0,56865	0,27072
5	4,03214	3,36493	2,75651	2,57058	2,42158	2,19096	2,01505	1,47588	0,91954	0,55943	0,26718
6	3,70743	3,14267	2,61224	2,44691	2,31326	2,10431	1,94318	1,43976	0,9057	0,55338	0,26483
7	3,49948	2,99795	2,51675	2,36462	2,24088	2,04601	1,89458	1,41492	0,89603	0,54911	0,26317
8	3,35539	2,89646	2,44898	2,306	2,18915	2,00415	1,85955	1,39682	0,88889	0,54593	0,26192
9	3,24984	2,82144	2,39844	2,26216	2,15038	1,97265	1,83311	1,38303	0,8834	0,54348	0,26096
10	3,16927	2,76377	2,35931	2,22814	2,12023	1,9481	1,81246	1,37218	0,87906	0,54153	0,26018
11	3,10581	2,71808	2,32814	2,20099	2,09614	1,92843	1,79588	1,36343	0,87553	0,53994	0,25956
12	3,05454	2,681	2,30272	2,17881	2,07644	1,91231	1,78229	1,35622	0,87261	0,53862	0,25903
13	3,01228	2,65031	2,2816	2,16037	2,06004	1,89887	1,77093	1,35017	0,87015	0,5375	0,25859
14	2,97684	2,62449	2,26378	2,14479	2,04617	1,8875	1,76131	1,34503	0,86805	0,53655	0,25821
15	2,94671	2,60248	2,24854	2,13145	2,03429	1,87774	1,75305	1,34061	0,86624	0,53573	0,25789
16	2,92078	2,58349	2,23536	2,11991	2,024	1,86928	1,74588	1,33676	0,86467	0,53501	0,2576
17	2,89823	2,56693	2,22385	2,10982	2,015	1,86187	1,73961	1,33338	0,86328	0,53438	0,25735
18	2,87844	2,55238	2,2137	2,10092	2,00707	1,85534	1,73406	1,33039	0,86205	0,53382	0,25712
19	2,86093	2,53948	2,2047	2,09302	2,00002	1,84953	1,72913	1,32773	0,86095	0,53331	0,25692
20	2,84534	2,52798	2,19666	2,08596	1,99371	1,84433	1,72472	1,32534	0,85996	0,53286	0,25674
21	2,83136	2,51765	2,18943	2,07961	1,98804	1,83965	1,72074	1,32319	0,85907	0,53246	0,25658
22	2,81876	2,50832	2,18289	2,07387	1,98291	1,83542	1,71714	1,32124	0,85827	0,53208	0,25643
23	2,80734	2,49987	2,17696	2,06866	1,97825	1,83157	1,71387	1,31946	0,85753	0,53175	0,2563
24	2,79694	2,49216	2,17154	2,0639	1,97399	1,82805	1,71088	1,31784	0,85686	0,53144	0,25617
25	2,78744	2,48511	2,16659	2,05954	1,9701	1,82483	1,70814	1,31635	0,85624	0,53115	0,25606
26	2,77871	2,47863	2,16203	2,05553	1,96651	1,82186	1,70562	1,31497	0,85567	0,53089	0,25595
27	2,77068	2,47266	2,15782	2,05183	1,9632	1,81913	1,70329	1,3137	0,85514	0,53065	0,25586
28	2,76326	2,46714	2,15393	2,04841	1,96014	1,81659	1,70113	1,31253	0,85465	0,53042	0,25577
29	2,75639	2,46202	2,15033	2,04523	1,95729	1,81424	1,69913	1,31143	0,85419	0,53021	0,25568
30	2,75	2,45726	2,14697	2,04227	1,95465	1,81205	1,69726	1,31042	0,85377	0,53002	0,25561
31	2,74404	2,45282	2,14383	2,03951	1,95218	1,81	1,69552	1,30946	0,85337	0,52984	0,25553
32	2,73848	2,44868	2,1409	2,03693	1,94987	1,80809	1,69389	1,30857	0,853	0,52967	0,25546
33	2,73328	2,44479	2,13816	2,03452	1,9477	1,80629	1,69236	1,30774	0,85265	0,5295	0,2554
34	2,72839	2,44115	2,13558	2,03224	1,94567	1,80461	1,69092	1,30695	0,85232	0,52935	0,25534
35	2,72381	2,43772	2,13316	2,03011	1,94375	1,80302	1,68957	1,30621	0,85201	0,52921	0,25528
36	2,71948	2,43449	2,13087	2,02809	1,94195	1,80153	1,6883	1,30551	0,85172	0,52908	0,25523
37	2,71541	2,43145	2,12871	2,02619	1,94024	1,80012	1,68709	1,30485	0,85144	0,52895	0,25518
38	2,71156	2,42857	2,12667	2,02439	1,93863	1,79878	1,68595	1,30423	0,85118	0,52883	0,25513
39	2,70791	2,42584	2,12474	2,02269	1,93711	1,79751	1,68488	1,30364	0,85094	0,52871	0,25508
40	2,70446	2,42326	2,12291	2,02108	1,93566	1,79631	1,68385	1,30308	0,8507	0,52861	0,25504
50	2,67779	2,40327	2,10872	2,00856	1,92444	1,787	1,67591	1,29871	0,84887	0,52776	0,2547
60	2,66028	2,39012	2,09936	2,0003	1,91703	1,78085	1,67065	1,29582	0,84765	0,5272	0,25447
70	2,6479	2,38081	2,09273	1,99444	1,91177	1,77647	1,66691	1,29376	0,84679	0,5268	0,25431
80	2,63869	2,37387	2,08778	1,99006	1,90784	1,77321	1,66412	1,29222	0,84614	0,5265	0,25419
90	2,63157	2,3685	2,08394	1,98667	1,9048	1,77068	1,66196	1,29103	0,84563	0,52626	0,2541
100	2,62589	2,36422	2,08088	1,98397	1,90237	1,76866	1,66023	1,29007	0,84523	0,52608	0,25402
∞	2,57583	2,32635	2,05375	1,95996	1,88079	1,75069	1,64487	1,28156	0,84162	0,5244	0,25335

Valores de la Distribución χ^2 : $\chi^2_{\alpha,v}$, donde $P(\chi^2 \leq \chi^2_{\alpha,v}) = \alpha$

	α										
v	**0,005**	**0,010**	**0,020**	**0,025**	**0,030**	**0,040**	**0,050**	**0,100**	**0,200**	**0,300**	**0,400**
1	0,00004	0,00016	0,00063	0,00098	0,00141	0,00252	0,00393	0,01579	0,06418	0,14847	0,275
2	0,01003	0,0201	0,04041	0,05064	0,06092	0,08164	0,10259	0,21072	0,44629	0,71335	1,02165
3	0,07172	0,11483	0,18483	0,2158	0,2451	0,30015	0,35185	0,58437	1,00517	1,42365	1,86917
4	0,20699	0,29711	0,4294	0,48442	0,53505	0,62715	0,71072	1,06362	1,64878	2,1947	2,75284
5	0,41174	0,5543	0,75189	0,83121	0,90306	1,03132	1,14548	1,61031	2,34253	2,99991	3,6555
6	0,67573	0,87209	1,13442	1,23734	1,32961	1,49237	1,63538	2,20413	3,07009	3,82755	4,57015
7	0,98926	1,23904	1,56429	1,68987	1,80163	1,99711	2,16735	2,83311	3,82232	4,67133	5,49323
8	1,34441	1,6465	2,03248	2,17973	2,31007	2,53665	2,73264	3,48954	4,59357	5,52742	6,42265
9	1,73493	2,0879	2,53238	2,70039	2,84849	3,10467	3,32511	4,16816	5,38005	6,39331	7,35703
10	2,15586	2,55821	3,05905	3,24697	3,41207	3,69654	3,9403	4,86518	6,17908	7,26722	8,29547
11	2,60322	3,05348	3,60869	3,81575	3,99716	4,30875	4,57481	5,57778	6,98867	8,14787	9,23729
12	3,07382	3,57057	4,17829	4,40379	4,6009	4,93855	5,22603	6,3038	7,80733	9,03428	10,182
13	3,56503	4,10692	4,76545	5,00875	5,22101	5,58377	5,89186	7,0415	8,63386	9,92568	11,1291
14	4,07467	4,66043	5,3682	5,62873	5,85563	6,24264	6,57063	7,78953	9,46733	10,8215	12,0785
15	4,60092	5,22935	5,98492	6,26214	6,50322	6,91371	7,26094	8,54676	10,307	11,7212	13,0297
16	5,14221	5,81221	6,61424	6,90766	7,16251	7,59577	7,96165	9,31224	11,1521	12,6243	13,9827
17	5,69722	6,40776	7,255	7,56419	7,83241	8,28778	8,67176	10,0852	12,0023	13,5307	14,9373
18	6,2648	7,01491	7,90622	8,23075	8,51199	8,98887	9,39046	10,8649	12,857	14,4399	15,8932
19	6,84397	7,63273	8,56704	8,90652	9,20044	9,69828	10,117	11,6509	13,7158	15,3517	16,8504
20	7,43384	8,2604	9,2367	9,59078	9,89708	10,4154	10,8508	12,4426	14,5784	16,2659	17,8088
21	8,03365	8,8972	9,91456	10,2829	10,6013	11,1395	11,5913	13,2396	15,4446	17,1823	18,7683
22	8,64272	9,54249	10,6	10,9823	11,3125	11,8703	12,338	14,0415	16,314	18,1007	19,7288
23	9,26042	10,1957	11,2926	11,6886	12,0303	12,6072	13,0905	14,848	17,1865	19,0211	20,6902
24	9,88623	10,8564	11,9918	12,4012	12,7543	13,3498	13,8484	15,6587	18,0618	19,9432	21,6525
25	10,5197	11,524	12,6973	13,1197	13,484	14,0978	14,6114	16,4734	18,9398	20,867	22,6156
26	11,1602	12,1981	13,4086	13,8439	14,219	14,8509	15,3792	17,2919	19,8202	21,7924	23,5794
27	11,8076	12,8785	14,1254	14,5734	14,9592	15,6087	16,1514	18,1139	20,703	22,7192	24,544
28	12,4613	13,5647	14,8475	15,3079	15,7042	16,3711	16,9279	18,9392	21,588	23,6475	25,5093
29	13,1211	14,2565	15,5745	16,0471	16,4538	17,1377	17,7084	19,7677	22,4751	24,577	26,4751
30	13,7867	14,9535	16,3062	16,7908	17,2076	17,9083	18,4927	20,5992	23,3641	25,5078	27,4416
31	14,4578	15,6555	17,0423	17,5387	17,9656	18,6827	19,2806	21,4336	24,2551	26,4397	28,4087
32	15,134	16,3622	17,7827	18,2908	18,7275	19,4608	20,0719	22,2706	25,1478	27,3728	29,3763
33	15,8153	17,0735	18,5271	19,0467	19,4931	20,2424	20,8665	23,1102	26,0422	28,3069	30,3444
34	16,5013	17,7891	19,2754	19,8063	20,2622	21,0273	21,6643	23,9523	26,9383	29,2421	31,313
35	17,1918	18,5089	20,0274	20,5694	21,0348	21,8154	22,465	24,7967	27,8359	30,1782	32,2821
36	17,8867	19,2327	20,7829	21,3359	21,8106	22,6065	23,2686	25,6433	28,735	31,1152	33,2517
37	18,5858	19,9602	21,5419	22,1056	22,5895	23,4005	24,0749	26,4921	29,6355	32,0532	34,2216
38	19,2889	20,6914	22,304	22,8785	23,3714	24,1973	24,8839	27,343	30,5373	32,9919	35,192
39	19,9959	21,4262	23,0693	23,6543	24,1562	24,9968	25,6954	28,1958	31,4405	33,9315	36,1628
40	20,7065	22,1643	23,8376	24,433	24,9437	25,7989	26,5093	29,0505	32,345	34,8719	37,134
41	21,4208	22,9056	24,6087	25,2145	25,7339	26,6034	27,3256	29,9071	33,2506	35,8131	38,1055
42	22,1385	23,6501	25,3827	25,9987	26,5266	27,4104	28,144	30,7654	34,1574	36,755	39,0774
43	22,8595	24,3976	26,1594	26,7854	27,3219	28,2196	28,9647	31,6255	35,0653	37,6975	40,0496
44	23,5837	25,148	26,9386	27,5746	28,1195	29,0311	29,7875	32,4871	35,9743	38,6408	41,0222
45	24,311	25,9013	27,7203	28,3662	28,9194	29,8447	30,6123	33,3504	36,8844	39,5847	41,995
46	25,0413	26,6572	28,5045	29,1601	29,7215	30,6604	31,439	34,2152	37,7955	40,5292	42,9682
47	25,7746	27,4158	29,291	29,9562	30,5258	31,4781	32,2676	35,0814	38,7075	41,4744	43,9417
48	26,5106	28,177	30,0798	30,7545	31,3322	32,2977	33,0981	35,9491	39,6205	42,4201	44,9154
49	27,2493	28,9406	30,8708	31,5549	32,1406	33,1192	33,9303	36,8182	40,5344	43,3664	45,8895
50	27,9907	29,7067	31,6639	32,3574	32,9509	33,9426	34,7643	37,6886	41,4492	44,3133	46,8638

Valores de la distribución F: $F_{0.9,v1,v2}$, donde $P(F \leq F_{0.9,v1,v2}) = 0.9$

	v_2																
v_1	1	2	4	6	8	10	12	14	16	18	20	25	30	60	80	100	∞
1	39,86	8,526	4,545	3,776	3,458	3,285	3,177	3,102	3,048	3,007	2,975	2,918	2,881	2,791	2,769	2,756	2,706
2	49,50	9,000	4,325	3,463	3,113	2,924	2,807	2,726	2,668	2,624	2,589	2,528	2,489	2,393	2,370	2,356	2,303
3	53,59	9,162	4,191	3,289	2,924	2,728	2,606	2,522	2,462	2,416	2,380	2,317	2,276	2,177	2,154	2,139	2,084
4	55,83	9,243	4,107	3,181	2,806	2,605	2,480	2,395	2,333	2,286	2,249	2,184	2,142	2,041	2,016	2,002	1,945
5	57,24	9,293	4,051	3,108	2,726	2,522	2,394	2,307	2,244	2,196	2,158	2,092	2,049	1,946	1,921	1,906	1,847
6	58,20	9,326	4,010	3,055	2,668	2,461	2,331	2,243	2,178	2,130	2,091	2,024	1,980	1,875	1,849	1,834	1,774
7	58,91	9,349	3,979	3,014	2,624	2,414	2,283	2,193	2,128	2,079	2,040	1,971	1,927	1,819	1,793	1,778	1,717
8	59,44	9,367	3,955	2,983	2,589	2,377	2,245	2,154	2,088	2,038	1,999	1,929	1,884	1,775	1,748	1,732	1,670
9	59,86	9,381	3,936	2,958	2,561	2,347	2,214	2,122	2,055	2,005	1,965	1,895	1,849	1,738	1,711	1,695	1,632
10	60,19	9,392	3,920	2,937	2,538	2,323	2,188	2,095	2,028	1,977	1,937	1,866	1,819	1,707	1,680	1,663	1,599
11	60,47	9,401	3,907	2,920	2,519	2,302	2,166	2,073	2,005	1,954	1,913	1,841	1,794	1,680	1,653	1,636	1,570
12	60,71	9,408	3,896	2,905	2,502	2,284	2,147	2,054	1,985	1,933	1,892	1,820	1,773	1,657	1,629	1,612	1,546
13	60,90	9,415	3,886	2,892	2,488	2,269	2,131	2,037	1,968	1,916	1,875	1,802	1,754	1,637	1,609	1,592	1,524
14	61,07	9,420	3,878	2,881	2,475	2,255	2,117	2,022	1,953	1,900	1,859	1,785	1,737	1,619	1,590	1,573	1,505
15	61,22	9,425	3,870	2,871	2,464	2,244	2,105	2,010	1,940	1,887	1,845	1,771	1,722	1,603	1,574	1,557	1,487
16	61,35	9,429	3,864	2,863	2,455	2,233	2,094	1,998	1,928	1,875	1,833	1,758	1,709	1,589	1,559	1,542	1,471
17	61,46	9,433	3,858	2,855	2,446	2,224	2,084	1,988	1,917	1,864	1,821	1,746	1,697	1,576	1,546	1,528	1,457
18	61,57	9,436	3,853	2,848	2,438	2,215	2,075	1,978	1,908	1,854	1,811	1,736	1,686	1,564	1,534	1,516	1,444
19	61,66	9,439	3,849	2,842	2,431	2,208	2,067	1,970	1,899	1,845	1,802	1,726	1,676	1,553	1,523	1,505	1,432
20	61,74	9,441	3,844	2,836	2,425	2,201	2,060	1,962	1,891	1,837	1,794	1,718	1,667	1,543	1,513	1,494	1,421
21	61,81	9,444	3,841	2,831	2,419	2,194	2,053	1,955	1,884	1,829	1,786	1,710	1,659	1,534	1,503	1,485	1,410
22	61,88	9,446	3,837	2,827	2,413	2,189	2,047	1,949	1,877	1,823	1,779	1,702	1,651	1,526	1,495	1,476	1,401
23	61,95	9,448	3,834	2,822	2,409	2,183	2,041	1,943	1,871	1,816	1,773	1,695	1,644	1,518	1,487	1,468	1,392
24	62,00	9,450	3,831	2,818	2,404	2,178	2,036	1,938	1,866	1,810	1,767	1,689	1,638	1,511	1,479	1,460	1,383
25	62,05	9,451	3,828	2,815	2,400	2,174	2,031	1,933	1,860	1,805	1,761	1,683	1,632	1,504	1,472	1,453	1,375
26	62,10	9,453	3,826	2,811	2,396	2,170	2,027	1,928	1,855	1,800	1,756	1,678	1,626	1,498	1,465	1,446	1,368
27	62,15	9,454	3,823	2,808	2,392	2,166	2,022	1,923	1,851	1,795	1,751	1,672	1,621	1,492	1,459	1,440	1,361
28	62,19	9,456	3,821	2,805	2,389	2,162	2,019	1,919	1,847	1,791	1,746	1,668	1,616	1,486	1,453	1,434	1,354
29	62,23	9,457	3,819	2,803	2,386	2,159	2,015	1,916	1,843	1,787	1,742	1,663	1,611	1,481	1,448	1,428	1,348
30	62,26	9,458	3,817	2,800	2,383	2,155	2,011	1,912	1,839	1,783	1,738	1,659	1,606	1,476	1,443	1,423	1,342
31	62,30	9,459	3,816	2,798	2,380	2,152	2,008	1,909	1,835	1,779	1,734	1,655	1,602	1,471	1,438	1,418	1,336
32	62,33	9,460	3,814	2,795	2,378	2,150	2,005	1,905	1,832	1,776	1,731	1,651	1,598	1,466	1,433	1,413	1,331
33	62,36	9,461	3,812	2,793	2,375	2,147	2,002	1,902	1,829	1,772	1,728	1,648	1,595	1,462	1,429	1,408	1,326
34	62,39	9,462	3,811	2,791	2,373	2,144	2,000	1,899	1,826	1,769	1,724	1,644	1,591	1,458	1,424	1,404	1,321
35	62,42	9,463	3,810	2,789	2,371	2,142	1,997	1,897	1,823	1,766	1,721	1,641	1,588	1,454	1,420	1,400	1,316
36	62,44	9,463	3,808	2,787	2,369	2,140	1,995	1,894	1,820	1,764	1,718	1,638	1,585	1,450	1,416	1,396	1,311
37	62,46	9,464	3,807	2,786	2,367	2,138	1,992	1,892	1,818	1,761	1,716	1,635	1,582	1,447	1,413	1,392	1,307
38	62,49	9,465	3,806	2,784	2,365	2,135	1,990	1,889	1,815	1,758	1,713	1,632	1,579	1,444	1,409	1,388	1,303
39	62,51	9,466	3,805	2,783	2,363	2,134	1,988	1,887	1,813	1,756	1,711	1,630	1,576	1,440	1,406	1,385	1,299
40	62,53	9,466	3,804	2,781	2,361	2,132	1,986	1,885	1,811	1,754	1,708	1,627	1,573	1,437	1,403	1,382	1,295
42	62,57	9,467	3,802	2,778	2,358	2,128	1,982	1,881	1,807	1,749	1,704	1,622	1,568	1,432	1,397	1,375	1,288
44	62,60	9,469	3,800	2,776	2,355	2,125	1,979	1,878	1,803	1,746	1,700	1,618	1,564	1,426	1,391	1,370	1,281
46	62,63	9,469	3,798	2,774	2,353	2,122	1,976	1,874	1,800	1,742	1,696	1,614	1,560	1,421	1,386	1,364	1,275
48	62,66	9,470	3,797	2,772	2,350	2,120	1,973	1,871	1,796	1,739	1,693	1,611	1,556	1,417	1,381	1,359	1,269
50	62,69	9,471	3,795	2,770	2,348	2,117	1,970	1,869	1,793	1,736	1,690	1,607	1,552	1,413	1,377	1,355	1,263
60	62,79	9,475	3,790	2,762	2,339	2,107	1,960	1,857	1,782	1,723	1,677	1,593	1,538	1,395	1,358	1,336	1,240
70	62,87	9,477	3,786	2,756	2,333	2,100	1,952	1,849	1,773	1,714	1,667	1,583	1,527	1,382	1,344	1,321	1,222
80	62,93	9,479	3,782	2,752	2,328	2,095	1,946	1,843	1,766	1,707	1,660	1,576	1,519	1,372	1,334	1,310	1,207
100	63,01	9,481	3,778	2,746	2,321	2,087	1,938	1,834	1,757	1,698	1,650	1,565	1,507	1,358	1,318	1,293	1,185
∞	63,33	9,491	3,761	2,722	2,293	2,055	1,904	1,797	1,718	1,657	1,607	1,518	1,456	1,291	1,245	1,214	1,000

Valores de la distribución F: $F_{0.95,v1,v2}$, donde $P(F \leq F_{0.95,v1,v2}) = 0.95$

	v_2																
v_1	1	2	4	6	8	10	12	14	16	18	20	25	30	60	80	100	∞
1	161,4	18,51	7,709	5,987	5,318	4,965	4,747	4,600	4,494	4,414	4,351	4,242	4,171	4,001	3,960	3,936	3,841
2	199,5	19,00	6,944	5,143	4,459	4,103	3,885	3,739	3,634	3,555	3,493	3,385	3,316	3,150	3,111	3,087	2,996
3	215,7	19,16	6,591	4,757	4,066	3,708	3,490	3,344	3,239	3,160	3,098	2,991	2,922	2,758	2,719	2,696	2,605
4	224,6	19,25	6,388	4,534	3,838	3,478	3,259	3,112	3,007	2,928	2,866	2,759	2,690	2,525	2,486	2,463	2,372
5	230,2	19,30	6,256	4,387	3,687	3,326	3,106	2,958	2,852	2,773	2,711	2,603	2,534	2,368	2,329	2,305	2,214
6	234,0	19,33	6,163	4,284	3,581	3,217	2,996	2,848	2,741	2,661	2,599	2,490	2,421	2,254	2,214	2,191	2,099
7	236,8	19,35	6,094	4,207	3,500	3,135	2,913	2,764	2,657	2,577	2,514	2,405	2,334	2,167	2,126	2,103	2,010
8	238,9	19,37	6,041	4,147	3,438	3,072	2,849	2,699	2,591	2,510	2,447	2,337	2,266	2,097	2,056	2,032	1,938
9	240,5	19,38	5,999	4,099	3,388	3,020	2,796	2,646	2,538	2,456	2,393	2,282	2,211	2,040	1,999	1,975	1,880
10	241,9	19,40	5,964	4,060	3,347	2,978	2,753	2,602	2,494	2,412	2,348	2,236	2,165	1,993	1,951	1,927	1,831
11	243,0	19,40	5,936	4,027	3,313	2,943	2,717	2,565	2,456	2,374	2,310	2,198	2,126	1,952	1,910	1,886	1,789
12	243,9	19,41	5,912	4,000	3,284	2,913	2,687	2,534	2,425	2,342	2,278	2,165	2,092	1,917	1,875	1,850	1,752
13	244,7	19,42	5,891	3,976	3,259	2,887	2,660	2,507	2,397	2,314	2,250	2,136	2,063	1,887	1,845	1,819	1,720
14	245,4	19,42	5,873	3,956	3,237	2,865	2,637	2,484	2,373	2,290	2,225	2,111	2,037	1,860	1,817	1,792	1,692
15	245,9	19,43	5,858	3,938	3,218	2,845	2,617	2,463	2,352	2,269	2,203	2,089	2,015	1,836	1,793	1,768	1,666
16	246,5	19,43	5,844	3,922	3,202	2,828	2,599	2,445	2,333	2,250	2,184	2,069	1,995	1,815	1,772	1,746	1,644
17	246,9	19,44	5,832	3,908	3,187	2,812	2,583	2,428	2,317	2,233	2,167	2,051	1,976	1,796	1,752	1,726	1,623
18	247,3	19,44	5,821	3,896	3,173	2,798	2,568	2,413	2,302	2,217	2,151	2,035	1,960	1,778	1,734	1,708	1,604
19	247,7	19,44	5,811	3,884	3,161	2,785	2,555	2,400	2,288	2,203	2,137	2,021	1,945	1,763	1,718	1,691	1,587
20	248,0	19,45	5,803	3,874	3,150	2,774	2,544	2,388	2,276	2,191	2,124	2,007	1,932	1,748	1,703	1,676	1,571
21	248,3	19,45	5,795	3,865	3,140	2,764	2,533	2,377	2,264	2,179	2,112	1,995	1,919	1,735	1,689	1,663	1,556
22	248,6	19,45	5,787	3,856	3,131	2,754	2,523	2,367	2,254	2,168	2,102	1,984	1,908	1,722	1,677	1,650	1,542
23	248,8	19,45	5,781	3,849	3,123	2,745	2,514	2,357	2,244	2,159	2,092	1,974	1,897	1,711	1,665	1,638	1,529
24	249,1	19,45	5,774	3,841	3,115	2,737	2,505	2,349	2,235	2,150	2,082	1,964	1,887	1,700	1,654	1,627	1,517
25	249,3	19,46	5,769	3,835	3,108	2,730	2,498	2,341	2,227	2,141	2,074	1,955	1,878	1,690	1,644	1,616	1,506
26	249,5	19,46	5,763	3,829	3,102	2,723	2,491	2,333	2,220	2,134	2,066	1,947	1,870	1,681	1,634	1,607	1,496
27	249,6	19,46	5,759	3,823	3,095	2,716	2,484	2,326	2,212	2,126	2,059	1,939	1,862	1,672	1,626	1,598	1,486
28	249,8	19,46	5,754	3,818	3,090	2,710	2,478	2,320	2,206	2,119	2,052	1,932	1,854	1,664	1,617	1,589	1,476
29	250,0	19,46	5,750	3,813	3,084	2,705	2,472	2,314	2,200	2,113	2,045	1,926	1,847	1,656	1,609	1,581	1,467
30	250,1	19,46	5,746	3,808	3,079	2,700	2,466	2,308	2,194	2,107	2,039	1,919	1,841	1,649	1,602	1,573	1,459
31	250,2	19,46	5,742	3,804	3,075	2,695	2,461	2,303	2,188	2,102	2,033	1,913	1,835	1,642	1,595	1,566	1,451
32	250,4	19,46	5,739	3,800	3,070	2,690	2,456	2,298	2,183	2,096	2,028	1,908	1,829	1,636	1,588	1,559	1,444
33	250,5	19,47	5,735	3,796	3,066	2,686	2,452	2,293	2,178	2,091	2,023	1,902	1,823	1,630	1,582	1,553	1,436
34	250,6	19,47	5,732	3,792	3,062	2,681	2,447	2,289	2,174	2,087	2,018	1,897	1,818	1,624	1,576	1,547	1,429
35	250,7	19,47	5,729	3,789	3,059	2,678	2,443	2,284	2,169	2,082	2,013	1,892	1,813	1,618	1,570	1,541	1,423
36	250,8	19,47	5,727	3,786	3,055	2,674	2,439	2,280	2,165	2,078	2,009	1,888	1,808	1,613	1,564	1,535	1,417
37	250,9	19,47	5,724	3,783	3,052	2,670	2,436	2,277	2,161	2,074	2,005	1,884	1,804	1,608	1,559	1,530	1,411
38	251,0	19,47	5,722	3,780	3,049	2,667	2,432	2,273	2,158	2,070	2,001	1,879	1,800	1,603	1,554	1,525	1,405
39	251,1	19,47	5,719	3,777	3,046	2,664	2,429	2,270	2,154	2,066	1,997	1,876	1,796	1,599	1,549	1,520	1,399
40	251,1	19,47	5,717	3,774	3,043	2,661	2,426	2,266	2,151	2,063	1,994	1,872	1,792	1,594	1,545	1,515	1,394
42	251,3	19,47	5,713	3,769	3,037	2,655	2,420	2,260	2,144	2,056	1,987	1,865	1,785	1,586	1,536	1,506	1,384
44	251,4	19,47	5,709	3,765	3,033	2,650	2,415	2,255	2,139	2,050	1,981	1,858	1,778	1,578	1,528	1,498	1,375
46	251,6	19,47	5,706	3,761	3,028	2,645	2,410	2,250	2,133	2,045	1,976	1,853	1,772	1,572	1,521	1,491	1,366
48	251,7	19,47	5,702	3,757	3,024	2,641	2,405	2,245	2,128	2,040	1,970	1,847	1,766	1,565	1,514	1,484	1,358
50	251,8	19,48	5,699	3,754	3,020	2,637	2,401	2,241	2,124	2,035	1,966	1,842	1,761	1,559	1,508	1,477	1,350
60	252,2	19,48	5,688	3,740	3,005	2,621	2,384	2,223	2,106	2,017	1,946	1,822	1,740	1,534	1,482	1,450	1,318
70	252,5	19,48	5,679	3,730	2,994	2,610	2,372	2,210	2,093	2,003	1,932	1,807	1,724	1,516	1,463	1,430	1,293
80	252,7	19,48	5,673	3,722	2,986	2,601	2,363	2,201	2,083	1,993	1,922	1,796	1,712	1,502	1,448	1,415	1,273
100	253,0	19,49	5,664	3,712	2,975	2,588	2,350	2,187	2,068	1,978	1,907	1,779	1,695	1,481	1,426	1,392	1,243
∞	254,3	19,50	5,628	3,669	2,928	2,538	2,296	2,131	2,010	1,917	1,843	1,711	1,622	1,389	1,325	1,283	1,000

Valores de la distribución F: $F_{0.975,v1,v2}$, donde $P(F \leq F_{0.975,v1,v2}) = 0.975$

	v_2																
v_1	1	2	4	6	8	10	12	14	16	18	20	25	30	60	80	100	∞
1	647,8	38,51	12,22	8,813	7,571	6,937	6,554	6,298	6,115	5,978	5,871	5,686	5,568	5,286	5,218	5,179	5,024
2	799,5	39,00	10,65	7,260	6,059	5,456	5,096	4,857	4,687	4,560	4,461	4,291	4,182	3,925	3,864	3,828	3,689
3	864,2	39,17	9,979	6,599	5,416	4,826	4,474	4,242	4,077	3,954	3,859	3,694	3,589	3,343	3,284	3,250	3,116
4	899,6	39,25	9,605	6,227	5,053	4,468	4,121	3,892	3,729	3,608	3,515	3,353	3,250	3,008	2,950	2,917	2,786
5	921,8	39,30	9,364	5,988	4,817	4,236	3,891	3,663	3,502	3,382	3,289	3,129	3,026	2,786	2,730	2,696	2,567
6	937,1	39,33	9,197	5,820	4,652	4,072	3,728	3,501	3,341	3,221	3,128	2,969	2,867	2,627	2,571	2,537	2,408
7	948,2	39,36	9,074	5,695	4,529	3,950	3,607	3,380	3,219	3,100	3,007	2,848	2,746	2,507	2,450	2,417	2,288
8	956,7	39,37	8,980	5,600	4,433	3,855	3,512	3,285	3,125	3,005	2,913	2,753	2,651	2,412	2,355	2,321	2,192
9	963,3	39,39	8,905	5,523	4,357	3,779	3,436	3,209	3,049	2,929	2,837	2,677	2,575	2,334	2,277	2,244	2,114
10	968,6	39,40	8,844	5,461	4,295	3,717	3,374	3,147	2,986	2,866	2,774	2,613	2,511	2,270	2,213	2,179	2,048
11	973,0	39,41	8,794	5,410	4,243	3,665	3,321	3,095	2,934	2,814	2,721	2,560	2,458	2,216	2,158	2,124	1,993
12	976,7	39,41	8,751	5,366	4,200	3,621	3,277	3,050	2,889	2,769	2,676	2,515	2,412	2,169	2,111	2,077	1,945
13	979,8	39,42	8,715	5,329	4,162	3,583	3,239	3,012	2,851	2,730	2,637	2,476	2,372	2,129	2,071	2,036	1,903
14	982,5	39,43	8,684	5,297	4,130	3,550	3,206	2,979	2,817	2,696	2,603	2,441	2,338	2,093	2,035	2,000	1,866
15	984,9	39,43	8,657	5,269	4,101	3,522	3,177	2,949	2,788	2,667	2,573	2,411	2,307	2,061	2,003	1,968	1,833
16	986,9	39,44	8,633	5,244	4,076	3,496	3,152	2,923	2,761	2,640	2,547	2,384	2,280	2,033	1,974	1,939	1,803
17	988,7	39,44	8,611	5,222	4,054	3,474	3,129	2,900	2,738	2,617	2,523	2,360	2,255	2,008	1,948	1,913	1,776
18	990,3	39,44	8,592	5,202	4,034	3,453	3,108	2,879	2,717	2,596	2,501	2,338	2,233	1,985	1,925	1,890	1,751
19	991,8	39,45	8,575	5,184	4,016	3,435	3,090	2,861	2,698	2,576	2,482	2,318	2,213	1,964	1,904	1,868	1,729
20	993,1	39,45	8,560	5,168	3,999	3,419	3,073	2,844	2,681	2,559	2,464	2,300	2,195	1,944	1,884	1,849	1,708
21	994,3	39,45	8,546	5,154	3,985	3,403	3,057	2,828	2,665	2,543	2,448	2,284	2,178	1,927	1,866	1,830	1,689
22	995,4	39,45	8,533	5,141	3,971	3,390	3,043	2,814	2,651	2,529	2,434	2,269	2,163	1,911	1,850	1,814	1,672
23	996,3	39,45	8,522	5,128	3,959	3,377	3,031	2,801	2,637	2,515	2,420	2,255	2,149	1,896	1,835	1,798	1,655
24	997,2	39,46	8,511	5,117	3,947	3,365	3,019	2,789	2,625	2,503	2,408	2,242	2,136	1,882	1,820	1,784	1,640
25	998,1	39,46	8,501	5,107	3,937	3,355	3,008	2,778	2,614	2,491	2,396	2,230	2,124	1,869	1,807	1,770	1,626
26	998,8	39,46	8,492	5,097	3,927	3,345	2,998	2,767	2,603	2,481	2,385	2,219	2,112	1,857	1,795	1,758	1,612
27	999,6	39,46	8,483	5,088	3,918	3,335	2,988	2,758	2,594	2,471	2,375	2,209	2,102	1,845	1,783	1,746	1,600
28	1000	39,46	8,476	5,080	3,909	3,327	2,979	2,749	2,584	2,461	2,366	2,199	2,092	1,835	1,772	1,735	1,588
29	1001	39,46	8,468	5,072	3,901	3,319	2,971	2,740	2,576	2,453	2,357	2,190	2,083	1,825	1,762	1,725	1,577
30	1001	39,46	8,461	5,065	3,894	3,311	2,963	2,732	2,568	2,445	2,349	2,182	2,074	1,815	1,752	1,715	1,566
31	1002	39,47	8,455	5,058	3,887	3,304	2,956	2,725	2,560	2,437	2,341	2,174	2,066	1,806	1,743	1,706	1,556
32	1002	39,47	8,449	5,052	3,881	3,297	2,949	2,718	2,553	2,430	2,334	2,166	2,058	1,798	1,735	1,697	1,546
33	1003	39,47	8,443	5,046	3,874	3,291	2,943	2,711	2,546	2,423	2,327	2,159	2,051	1,790	1,726	1,688	1,537
34	1003	39,47	8,438	5,041	3,869	3,285	2,937	2,705	2,540	2,416	2,320	2,152	2,044	1,782	1,719	1,680	1,528
35	1004	39,47	8,433	5,035	3,863	3,279	2,931	2,699	2,534	2,410	2,314	2,146	2,037	1,775	1,711	1,673	1,520
36	1004	39,47	8,428	5,030	3,858	3,274	2,925	2,694	2,529	2,405	2,308	2,140	2,031	1,768	1,704	1,666	1,512
37	1005	39,47	8,423	5,025	3,853	3,269	2,920	2,689	2,523	2,399	2,302	2,134	2,025	1,762	1,697	1,659	1,505
38	1005	39,47	8,419	5,021	3,848	3,264	2,915	2,684	2,518	2,394	2,297	2,128	2,019	1,756	1,691	1,652	1,497
39	1005	39,47	8,415	5,017	3,844	3,260	2,911	2,679	2,513	2,389	2,292	2,123	2,014	1,750	1,685	1,646	1,490
40	1006	39,47	8,411	5,012	3,840	3,255	2,906	2,674	2,509	2,384	2,287	2,118	2,009	1,744	1,679	1,640	1,484
42	1006	39,47	8,404	5,005	3,832	3,247	2,898	2,666	2,500	2,375	2,278	2,109	1,999	1,733	1,668	1,629	1,471
44	1007	39,48	8,397	4,998	3,825	3,240	2,891	2,658	2,492	2,367	2,270	2,100	1,991	1,724	1,658	1,618	1,459
46	1007	39,48	8,391	4,992	3,818	3,233	2,884	2,651	2,485	2,360	2,263	2,093	1,982	1,715	1,649	1,609	1,448
48	1008	39,48	8,386	4,986	3,812	3,227	2,877	2,644	2,478	2,353	2,256	2,085	1,975	1,706	1,640	1,600	1,438
50	1008	39,48	8,381	4,980	3,807	3,221	2,871	2,638	2,472	2,347	2,249	2,079	1,968	1,699	1,632	1,592	1,428
60	1010	39,48	8,360	4,959	3,784	3,198	2,848	2,614	2,447	2,321	2,223	2,052	1,940	1,667	1,599	1,558	1,388
70	1011	39,48	8,346	4,943	3,768	3,182	2,831	2,597	2,429	2,303	2,205	2,032	1,920	1,643	1,574	1,532	1,357
80	1012	39,49	8,335	4,932	3,756	3,169	2,818	2,583	2,415	2,289	2,190	2,017	1,904	1,625	1,555	1,512	1,333
100	1013	39,49	8,319	4,915	3,739	3,152	2,800	2,565	2,396	2,269	2,170	1,996	1,882	1,599	1,527	1,483	1,296
∞	1018	39,50	8,257	4,849	3,670	3,080	2,725	2,487	2,316	2,187	2,085	1,906	1,787	1,482	1,400	1,347	1,000

Valores de la distribución F: $F_{0.99,v1,v2}$, donde $P(F \leq F_{0.99,v1,v2}) = 0.99$

	v_2																
v_1	1	2	4	6	8	10	12	14	16	18	20	25	30	60	80	100	∞
1	4052	98,50	21,20	13,75	11,26	10,04	9,330	8,862	8,531	8,285	8,096	7,770	7,562	7,077	6,963	6,895	6,635
2	4999	99,00	18,00	10,92	8,649	7,559	6,927	6,515	6,226	6,013	5,849	5,568	5,390	4,977	4,881	4,824	4,605
3	5403	99,17	16,69	9,780	7,591	6,552	5,953	5,564	5,292	5,092	4,938	4,675	4,510	4,126	4,036	3,984	3,782
4	5625	99,25	15,98	9,148	7,006	5,994	5,412	5,035	4,773	4,579	4,431	4,177	4,018	3,649	3,563	3,513	3,319
5	5764	99,30	15,52	8,746	6,632	5,636	5,064	4,695	4,437	4,248	4,103	3,855	3,699	3,339	3,255	3,206	3,017
6	5859	99,33	15,21	8,466	6,371	5,386	4,821	4,456	4,202	4,015	3,871	3,627	3,473	3,119	3,036	2,988	2,802
7	5928	99,36	14,98	8,260	6,178	5,200	4,640	4,278	4,026	3,841	3,699	3,457	3,304	2,953	2,871	2,823	2,639
8	5981	99,37	14,80	8,102	6,029	5,057	4,499	4,140	3,890	3,705	3,564	3,324	3,173	2,823	2,742	2,694	2,511
9	6022	99,39	14,66	7,976	5,911	4,942	4,388	4,030	3,780	3,597	3,457	3,217	3,067	2,718	2,637	2,590	2,407
10	6056	99,40	14,55	7,874	5,814	4,849	4,296	3,939	3,691	3,508	3,368	3,129	2,979	2,632	2,551	2,503	2,321
11	6083	99,41	14,45	7,790	5,734	4,772	4,220	3,864	3,616	3,434	3,294	3,056	2,906	2,559	2,478	2,430	2,248
12	6106	99,42	14,37	7,718	5,667	4,706	4,155	3,800	3,553	3,371	3,231	2,993	2,843	2,496	2,415	2,368	2,185
13	6126	99,42	14,31	7,657	5,609	4,650	4,100	3,745	3,498	3,316	3,177	2,939	2,789	2,442	2,361	2,313	2,130
14	6143	99,43	14,25	7,605	5,559	4,601	4,052	3,698	3,451	3,269	3,130	2,892	2,742	2,394	2,313	2,265	2,082
15	6157	99,43	14,20	7,559	5,515	4,558	4,010	3,656	3,409	3,227	3,088	2,850	2,700	2,352	2,271	2,223	2,039
16	6170	99,44	14,15	7,519	5,477	4,520	3,972	3,619	3,372	3,190	3,051	2,813	2,663	2,315	2,233	2,185	2,000
17	6181	99,44	14,11	7,483	5,442	4,487	3,939	3,586	3,339	3,158	3,018	2,780	2,630	2,281	2,199	2,151	1,965
18	6192	99,44	14,08	7,451	5,412	4,457	3,909	3,556	3,310	3,128	2,989	2,751	2,600	2,251	2,169	2,120	1,934
19	6201	99,45	14,05	7,422	5,384	4,430	3,883	3,529	3,283	3,101	2,962	2,724	2,573	2,223	2,141	2,092	1,905
20	6209	99,45	14,02	7,396	5,359	4,405	3,858	3,505	3,259	3,077	2,938	2,699	2,549	2,198	2,115	2,067	1,878
21	6216	99,45	13,99	7,372	5,336	4,383	3,836	3,483	3,237	3,055	2,916	2,677	2,526	2,175	2,092	2,043	1,854
22	6223	99,45	13,97	7,351	5,316	4,363	3,816	3,463	3,216	3,035	2,895	2,657	2,506	2,153	2,070	2,021	1,831
23	6229	99,46	13,95	7,331	5,297	4,344	3,798	3,444	3,198	3,016	2,877	2,638	2,487	2,134	2,050	2,001	1,810
24	6235	99,46	13,93	7,313	5,279	4,327	3,780	3,427	3,181	2,999	2,859	2,620	2,469	2,115	2,032	1,983	1,791
25	6240	99,46	13,91	7,296	5,263	4,311	3,765	3,412	3,165	2,983	2,843	2,604	2,453	2,098	2,015	1,965	1,773
26	6245	99,46	13,89	7,280	5,248	4,296	3,750	3,397	3,150	2,968	2,829	2,589	2,437	2,083	1,999	1,949	1,755
27	6249	99,46	13,88	7,266	5,234	4,283	3,736	3,383	3,137	2,955	2,815	2,575	2,423	2,068	1,983	1,934	1,739
28	6253	99,46	13,86	7,253	5,221	4,270	3,724	3,371	3,124	2,942	2,802	2,562	2,410	2,054	1,969	1,919	1,724
29	6257	99,46	13,85	7,240	5,209	4,258	3,712	3,359	3,112	2,930	2,790	2,550	2,398	2,041	1,956	1,906	1,710
30	6261	99,47	13,84	7,229	5,198	4,247	3,701	3,348	3,101	2,919	2,778	2,538	2,386	2,028	1,944	1,893	1,696
31	6264	99,47	13,83	7,218	5,188	4,236	3,690	3,337	3,090	2,908	2,768	2,527	2,375	2,017	1,932	1,881	1,684
32	6267	99,47	13,81	7,207	5,178	4,227	3,681	3,327	3,080	2,898	2,758	2,517	2,365	2,006	1,921	1,870	1,671
33	6270	99,47	13,80	7,198	5,168	4,217	3,671	3,318	3,071	2,889	2,748	2,508	2,355	1,996	1,910	1,859	1,660
34	6273	99,47	13,79	7,189	5,159	4,209	3,663	3,309	3,062	2,880	2,739	2,499	2,346	1,986	1,900	1,849	1,649
35	6276	99,47	13,79	7,180	5,151	4,200	3,654	3,301	3,054	2,871	2,731	2,490	2,337	1,976	1,890	1,839	1,638
36	6278	99,47	13,78	7,172	5,143	4,193	3,647	3,293	3,046	2,863	2,723	2,482	2,329	1,968	1,881	1,830	1,628
37	6280	99,47	13,77	7,164	5,136	4,185	3,639	3,286	3,039	2,856	2,715	2,474	2,321	1,959	1,873	1,821	1,619
38	6283	99,47	13,76	7,157	5,129	4,178	3,632	3,279	3,031	2,849	2,708	2,467	2,313	1,951	1,864	1,813	1,610
39	6285	99,47	13,75	7,150	5,122	4,172	3,626	3,272	3,025	2,842	2,701	2,460	2,306	1,943	1,856	1,805	1,601
40	6287	99,47	13,75	7,143	5,116	4,165	3,619	3,266	3,018	2,835	2,695	2,453	2,299	1,936	1,849	1,797	1,592
42	6291	99,48	13,73	7,131	5,104	4,153	3,607	3,254	3,006	2,823	2,683	2,440	2,286	1,922	1,835	1,783	1,576
44	6294	99,48	13,72	7,120	5,093	4,143	3,597	3,243	2,995	2,812	2,671	2,429	2,275	1,910	1,822	1,770	1,562
46	6297	99,48	13,71	7,110	5,083	4,133	3,587	3,233	2,985	2,802	2,661	2,419	2,264	1,898	1,810	1,757	1,548
48	6300	99,48	13,70	7,100	5,074	4,124	3,578	3,224	2,976	2,793	2,652	2,409	2,254	1,887	1,799	1,746	1,535
50	6303	99,48	13,69	7,091	5,065	4,115	3,569	3,215	2,967	2,784	2,643	2,400	2,245	1,877	1,788	1,735	1,523
60	6313	99,48	13,65	7,057	5,032	4,082	3,535	3,181	2,933	2,749	2,608	2,364	2,208	1,836	1,746	1,692	1,473
70	6321	99,48	13,63	7,032	5,007	4,058	3,511	3,157	2,908	2,724	2,582	2,337	2,181	1,806	1,714	1,659	1,435
80	6326	99,49	13,61	7,013	4,989	4,039	3,493	3,138	2,889	2,705	2,563	2,317	2,160	1,783	1,690	1,634	1,404
100	6334	99,49	13,58	6,987	4,963	4,014	3,467	3,112	2,863	2,678	2,535	2,289	2,131	1,749	1,655	1,598	1,358
∞	6366	99,50	13,46	6,880	4,859	3,909	3,361	3,004	2,753	2,566	2,421	2,169	2,006	1,601	1,494	1,427	1,000

Bibliografía

1. Acuña, Luis. Estadística Aplicada con Fathom. Editorial Tecnológica de Costa Rica, 2004
2. Antibí, André. Didáctica de las Matemáticas: Métodos de Resolución de problemas. Serie Cabecar, Costa Rica 2000.
3. Calderon, Silvia; Morales, Mario. "Análisis Combinatorio". Editorial Tecnológica de Costa Rica, 1992.
4. Devore, J.,"Probabilidad y Estádistica para Ingeniería y Ciencias" 4a ed. International Thomson Editores, México, 1998.
5. Gómez, Miguel. 2000. Elementos de estadísitca descriptiva. EUNED, San José, Costa Rica.
6. Mora, E. "Curso Intermedio de Probabilidades". Universidad de Costa Rica, 2001.
7. Murrillo, M. "Introducción a la matemática Discreta". Editorial Tecnológica de Costa Rica, 2004.
8. Sanabria G. "Tópicos precedentes al estudio de la Teoría de Probabilidades, II semestre del 2007", Publicaciones ITCR.
9. Sanabria, Giovanni. 2010. "Una propuesta para la enseñanza de los Elementos de Análisis Combinatorio". Memorias del Primer Encuentro Nacional en la Enseñanza de la Probabilidad y la Estadística (1° ENEPE), Puebla – México del 16 al 18 de junio del 2010.
10. Walpole, R; Myers, R; Myers, S. "Probabilidad y estadistica para ingenieros", 6a ed, Prentice-Hall Hispanoamericana. S.A., 1999.