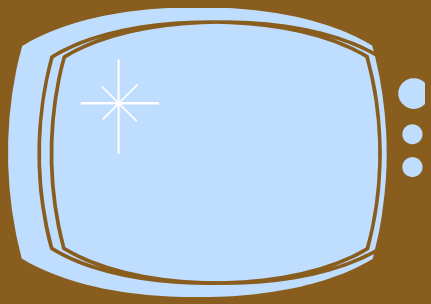


Diseño de muestreo

PIE 3

ALFREDO MARIO BARONIO
ANA MARIA VIANCO

Cuadernos de Econometría

Diseño de muestreo - PIE 3. Cuadernos de Econometría. Alfredo Mario Baronio y Ana María Vianco. 1ª edición. 2015

Xxx p; xxx cm

ISBN

1. *Estadístico, 2 Estimación 3 Contraste. 4 Probabilístico*

Fecha de catalogación: xxx 2015

Diseño de muestreo - PIE 3. Cuadernos de Econometría.
Alfredo Mario Baronio y Ana María Vianco.

Primera edición xxx 2015

ISBN XXXXXXX

Tirada xxx ejemplares

Esta edición es financiada con subsidios otorgados al proyecto Producción de Datos y Econometría Aplicada por la Secretaría de Ciencia y Técnica de la UNRC y el Instituto de Investigaciones de la UNVM.

Contenido

En los cuadernos anteriores, se establece que el principal "material" del PIE -proceso de investigación econométrica- son los datos y que, para obtenerlos, el investigador debe plantear una tabla de datos de unidades de observación por variables. Se introduce el concepto de que estas variables tienen características poblacionales que las distinguen a unas de otra y poseen distribuciones de probabilidad. En este cuaderno se estudia cómo a las características poblacionales de una variable, medidas de posición y dispersión –denominadas parámetros-, son estimados a partir de características– obtenidos en la muestra aleatoria de las unidades de observación medidas de posición y dispersión muestral –denominadas estadísticos. También, las condiciones estadísticas que deben cumplir los estimadores, la forma de utilizarlos para realizar estimaciones de los parámetros -puntuales o por intervalo- y cómo contrastar si los valores estimados se corresponden con los "naturales" de aquellos o lo supuestos por el investigador. Finalmente, el diseño de las fuentes de información primarias y los tipos de muestreos probabilísticos que permiten obtener unidades de observación válidas para realizar la inferencia estadística de los parámetros poblacionales.

1. DISTRIBUCIONES EN EL MUESTREO

Se retoma el concepto de parámetros y estadísticos para estudiar su distribución en el muestreo; para ello se trabaja con la media aritmética, estableciendo cuáles son sus propiedades y sus formas de estimación, teniendo en cuenta la dstribución de probabilidad asociada a la variable de interés y el tamaño de la muestra considerada.

1.1. Estadísticos y parámetros

Hasta aquí se ha desarrollado el concepto de probabilidad y sus distribuciones asociadas, ahora se estudia la forma en que los *estadísticos* de la muestra se pueden utilizar para hacer inferencias sobre los *parámetros* de la población.

El término *observación* se usará para representar cualquier clase de recolección numérica de información. Los datos, de los cuales se obtiene información por medio de la observación, pueden estar contenidos en cualquier *población* o *muestra* extraída de la misma. Estos son dos conceptos fundamentales para aplicar el PIE y realizar un análisis cuantitativo de los datos recogidos fueron definidos en el cuaderno anterior. Para obtener los datos habrá que diseñar las

fuentes de información –primaria o secundaria– que los contienen -tiempo o individuos-.

Como ya se mencionó, uno de los principales recursos de la investigación econométrica es la *inferencia estadística*. Es decir, utilizar *estadísticos* obtenidos sobre las variables con los datos de la muestra, para estimar el valor de los parámetros en la población.

En la práctica, se selecciona aleatoriamente una muestra de unidades de observación de tamaño determinado entre los N elementos de la población. Supóngase que se propone seleccionar n unidades y que se observarán las características X, Y y Z. La tabla de datos se puede representar según la Figura 1.1, donde:

- Los valores x_i, y_i, z_i, para todo $i = 1, \dots, n$ representan *los datos* que se van a obtener cuando se lleve a cabo la investigación;
- Las medidas de posición y dispersión que se obtendrán, últimas dos filas de la tabla, son estadísticos, por ejemplo el estadístico:

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

  correspondiente a la *media aritmética* de la variable X en la muestra. Este estadístico se utilizará para estimar al parámetro

  $$\mu_X = \frac{\sum_{i=1}^{n} x_i}{N}$$

  correspondiente a la media de la variable X en la población;
- N representa el tamaño de la población -son los elementos de la población- y n el de la muestra -son las unidades de observación seleccionadas en la población-.

Las n unidades de observación se van a obtener aplicando un diseño de muestreo. A fin de poder utilizar la media de la muestra para estimar la media de la población, se deberían examinar todas las muestras posibles de la población y calcular una media para cada muestra. Esto es, se debería repetir el experimento anterior tantas veces como muestras con reposición se podrían obtener de los N

elementos de la población. La distribución de estos resultados se denomina *distribución en el muestreo*. En la práctica con la selección de una *sola* muestra y utilizando la teoría de la probabilidad, se pueden hacer inferencias a la población.

	X	Y	Z
1	x_1	y_1	z_1
2	x_2	y_2	z_2
3	x_3	y_3	z_3
⋮	⋮	⋮	⋮
i	x_i	y_i	z_i
⋮	⋮	⋮	⋮
n	x_n	y_n	z_n
Medida de posición	$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$	$\bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$	$\bar{Z} = \frac{\sum_{i=1}^{n} z_i}{n}$
Medida de dispersión	$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$	$S^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{n-1}$	$S^2 = \frac{\sum_{i=1}^{n}(z_i - \bar{Z})^2}{n-1}$

Figura 1.1 Tabla de datos experimental

Esta parte del proceso se denomina *inferencia estadística* y es uno de los objetivos de la investigación econométrica. La inferencia se hace utilizando estadísticos de la muestra; como se sabe, los estadísticos son "mediciones" que se realizan en la muestra para estimar las mismas "mediciones" en la población que se denominan parámetros.

El investigador, a través de la tabla de datos planteada en la Figura 1.1, podrá obtener las medidas de posición y dispersión de la muestra (estadísticos) y buscará estimar valores probables de las medidas de posición o dispersión poblacionales (parámetros). Las conclusiones a las que arribará serán respecto a la población y no respecto de la muestra.

Ejemplo 1.1. Un encuestador político se interesa en los resultados de la muestra sólo como medio para estimar la *proporción* de votos que recibirá cada candidato entre la población de votantes.

En la Figura 1.2 se observa la relación entre estadísticos y parámetros

Nombre	Parámetros	Estimador
Media	μ_X	$\bar{X}$
Diferencia entre las medias de dos poblaciones	$\mu_{X_1} - \mu_{X_2}$	$\bar{X}_1 - \bar{X}_2$
Proporción	π	P
Varianza	σ_X^2	S_X^2
Desviación estándar	σ_X	S_X

Por ejemplo, el estadístico $\bar{X}$, que surge de una muestra, es estimador del parámetro μ_X en la población. El estadístico $\bar{X}_1 - \bar{X}_2$, diferencia de las medias provenientes de dos muestras, en distintas poblaciones, es estimador del parámetro $\mu_{X_1} - \mu_{X_2}$ que indica la diferencia en las medias proveniente de dos poblaciones.

Figura 1.2: Parámetros poblacionales y estadísticos muestrales

En la práctica se selecciona solo una muestra; por lo tanto, se debe examinar el concepto de distribución en el muestreo, con el propósito de poder utilizar la teoría de la probabilidad para hacer inferencias en cuanto a los parámetros de la población.

Se estudiarán las distribuciones en el muestreo sin abordar específicamente el problema de cómo se obtiene una muestra aleatoria, tema que será tratado más adelante. Pero, además, para hacer inferencia desde los estadísticos a los parámetros poblacionales, habrá que estudiar la teoría de la estimación estadística y contrastes de hipótesis. Para ello, se requiere previamente conocer las distribuciones en el muestreo de la media y de las distribuciones de probabilidad exactas: Chi-cuadrado, F de Snedecor y t de Student.

1.2. Distribución en el muestreo de la media

Existen diversas medidas con que se pueden caracterizar los datos de la muestra. Dentro de las medidas de posición, tanto el *modo* como la *mediana* y la *media* calculadas desde los datos muestrales de una variable, pueden ser considerados estadísticos útiles para estimar la media de esa variable en la población.

No obstante, el mejor estimador para *inferir* la media poblacional es la *media aritmética* obtenida sobre una determinada variable a partir de los datos correspondientes a las unidades de observación de la muestra. Esto es así pues, este estadístico tiene tres propiedades deseables y muy importantes en todo proceso de estimación: es *insesgado*, *eficiente* y *consistente*.

La media aritmética, obtenida de los datos de una muestra aleatoria, es una *variable aleatoria* ya que se ha calculado a partir de un proceso de estimación aleatorio. Por lo tanto, al ser una variable aleatoria tiene una distribución de probabilidad, una esperanza matemática y una varianza.

Esto es así ya que *si se repite* sucesivamente un experimento muestral, en un espacio y tiempo determinado, se obtendrán sucesivos valores diferentes para las variables consideradas, lo que producirá diferentes estadísticos para estimar los parámetros. Es decir, se tendrán tantas medias aritméticas como veces se repita el experimento, por lo que la media es una variable y como proviene de un proceso de selección aleatorio, será una variable aleatoria. Por lo tanto, tendrá media, varianza y distribución de probabilidad asociada.

Para el caso particular de la media aritmética de los datos de una muestra aleatoria, se demostrará empíricamente que:

$$\mathrm{E}(\overline{\mathrm{X}})=\mu_{\mathrm{X}} \qquad (1.1)$$

Nótese que se está diciendo que el promedio de las medias aritméticas, obtenidas de las sucesivas muestras, es igual al parámetro poblacional que estima y no que la media aritmética de la muestra es igual al parámetro que estima.

Esto es, la esperanza matemática de la media aritmética de los datos obtenidos a partir de una muestra aleatoria, es igual al parámetro poblacional que se está estimando (la media de la población). Al cumplir con esta propiedad, se dice que la media aritmética es un *estimador insesgado* de la media poblacional.

Por lo tanto, si se toma una cantidad determinada de muestras aleatorias de una población, n muestras, y de cada una se calcula, a partir de los datos observados de una variable, una media aritmética, se obtendrán n medias aritméticas muestrales. La esperanza matemática de todas estas medias aritméticas muestrales es la media de la población,

$$\sum_{i=1}^{n} \frac{\overline{X}_i}{n} = \mu_X \qquad (1.2)$$

Ejemplo 1.2. Experimento (extraído de Kmenta, J. pp 27 y ss.) Bajo el supuesto de que la variable X representa *el número de visitas al dentista en un año determinado por individuos de una cierta población*, y que es una variable aleatoria discreta con distribución uniforme. Esto es, X puede asumir solo valores enteros y cada uno de estos valores se observa con la misma probabilidad. La figura 1.3 muestra la tabla de datos con la frecuencia relativa en la población y el gráfico que la ilustra.

Identificando con $[a]$ el valor mínimo y con $[b]$ el valor máximo de X, la media de una distribución uniforme discreta es

$$E(X) = \mu_X = \frac{a+b}{2} = \frac{0+9}{2} = 4.5$$

y la desviación estándar:

$$\sigma_X = \sqrt{\frac{[(b-a)+1]^2 - 1}{12}} = 2.87$$

De modo que la media y la desviación estándar de la población bajo estudio asumen los valores 4.5 y 2.87.

Valores de X	Frecuencia relativa en la población
0	0.1
1	0.1
2	0.1
3	0.1
4	0.1
5	0.1
6	0.1
7	0.1
8	0.1
9	0.1
	1.0

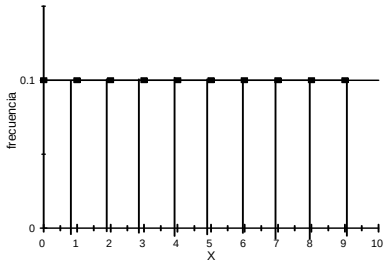


Figura 1.3. Tabla de datos del experimento

De esta población se extraen 100 muestras posibles de tamaño 5. Es decir, se seleccionan 5 personas al azar de la población y se le pregunta cuántas veces visitó al dentista en el año. En una de las muestras las respuestas de las personas fueron:

PERSONA	NUMERO DE VISITAS
Primera	2
Segunda	0
Tercera	4
Cuarta	1
Quinta	1

Sobre esta muestra se calcula la media aritmética:

$$\frac{2+0+4+1+1}{5} = 1.6$$

Este experimento se repitió 100 veces y se calcularon 100 medias aritméticas. Los resultados del experimento se muestran en la Figura 1.4

Aunque la variable (y, por lo tanto, los datos individuales) solo pueden tomar valores enteros entre 0 y 9, las medias de las muestras no

serán, generalmente, enteros y la distribución muestral será una distribución de frecuencias cuyas clases estarán definidas por intervalos y no por puntos. Se sabe que para representar cada clase se puede elegir un valor, como, por ejemplo, el centro de cada intervalo.

Valores de la Media de la muestra X		Frecuencia	
Intervalo	Punto Medio	Absoluta	Relativa
0.5 a 1.499	1	1	0.01
1.5 a 2.499	2	5	0.05
2.5 a 3.499	3	12	0.12
3.5 a 4.499	4	31	0.31
4.5 a 5.499	5	28	0.28
5.5 a 6.499	6	15	0.15
6.5 a 7.499	7	5	0.05
7.5 a 8.499	8	3	0.03
8.5 a 9.499	9	0	0.00
		100	1.00

Figura 1.4

Las principales características de esta distribución son

Media $= \sum_{i=1}^{9} \bar{X}_i f_i = 1\,(0{,}01) + 2\,(0{,}05) + 3\,(0{,}12) + 4\,(0{,}31) + 5\,(0{,}28) +$

$$+ 6\,(0{,}15) + 7\,(0{,}05) + 8\,(0{,}03) + 9\,(0{,}00) = 4{,}60$$

Desviación estándar $= \sqrt{\sum_{i=1}^{9} (\bar{X}_i - 4{,}60)^2 f_i} = 1{,}3638$

Los resultados indican que un 59% de los valores estimados $(0{,}31 + 0{,}28)$ están comprendidos en el intervalo ± 1 alrededor del valor real $4{,}5$ y que el 86% de las estimaciones están comprendidas en el intervalo ± 2 alrededor de dicho valor.

Realizado el mismo experimento con la *variante* de tomar 100 muestras de tamaño 10 se obtuvieron las siguientes características

Media $\sum_{i=1}^{9} \bar{X}_i f_i = 4{,}57$

Desviación estándar= $\sqrt{\sum_{i=1}^{9} (\bar{X}_i - 4{,}57)^2 f_i} = 1{,}0416$

Los resultados del experimento permiten realizar las siguientes generalizaciones:

- la media de la muestra, $\bar{X}$, es un estimador *insesgado* de la media de la población, μ_X
- a medida que aumenta el tamaño de la muestra, la distribución muestral de $\bar{X}$ se va concentrando cada vez más en torno a la media de la población y, por tanto, $\bar{X}$ es un estimador *consistente* de μ_X
- el estimador $\bar{X}$ es más estable de una muestra a otra, que la mediana o el modo, esto lo convierte en un estimador *eficiente*.

Se ha trabajado con distribuciones muestrales experimentales. Si se lo hubiese hecho con las distribuciones teóricas se hubieran obtenido idénticos resultados pero con mucho menos trabajo ya que sólo se tomaría una muestra.

Función generatriz de momentos de la media muestral

La media aritmética de la muestra es un estimador insesgado de la media de la población; como tal, es un estadístico y se vio que los mismos son variables aleatorias. Por lo tanto, como cualquier variable tienen ciertas características que la distingue de otras variables, como son: media, varianza y distribución de probabilidad. En esta sección se estudia la forma de derivar esas características.

La media de una variable aleatoria se conoce como el momento natural de orden 1. Partiendo de la fórmula de cálculo de la media aritmética, se puede derivar ese momento usando la función generatriz de momentos.

Se sabe que:

$$\overline{X}=\sum_{i=1}^{n}\frac{x_i}{n} \qquad (1.3)$$

Entonces si se toma el numerador de esta fórmula, cabe preguntarse:

¿A que es igual la función generatriz de momentos de la suma de variables aleatorias *independientes* (es decir, la probabilidad de ocurrencia de una variable no depende de la probabilidad de ocurrencia de la otra) $x_1, x_2, \ldots . x_n$?

OBSERVACIÓN: $x_1, x_2, \ldots . x_n$ son los datos para las n unidades de observación correspondientes a la variable aleatoria X. Estos datos provienen de una muestra aleatoria de unidades de observación y, por lo tanto, son también variables aleatorias, ya que si se toma otra muestra, las unidades de observación serán otras y los datos de la variable X ($x_1, x_2, \ldots . x_n$) serán otros. Como se dijo anteriormente, en la práctica se toma una sola muestra aleatoria; es decir, estos datos se corresponden con una sola muestra, representando características *no aleatorias* para las unidades de observación seleccionadas. Entonces, como las características de la población se obtienen a partir de las características de la muestra, cada función generatriz de momentos de los datos coincide con la de la variable observada.

Como ya se estudió anteriormente, la función generatriz de momentos de una *variable aleatoria continua* es:

$$M_X(\theta) = E\left(e^{\theta X}\right) = \int_{-\infty}^{\infty} e^{\theta X} f(X) dX \qquad (1.4)$$

Una de las propiedades es que la función generatriz de momentos de la sumatoria de variables aleatorias independientes en el parámetro θ es:

$$M_{x_1+x_2+\cdots+x_n}(\theta) = M_{x_1}(\theta) * M_{x_2}(\theta) * \cdots * M_{x_n}(\theta)$$

Pero, en muestreo aleatorio, las características de la población se obtienen a partir de las características de la muestra, entonces:

$$M_{x_i}(\theta) = M_X^n(\theta)$$

Por ende $M_X^n(\theta)$ es la función generatriz de momento de la variable poblacional X elevado a la potencia n.

En síntesis,

$$M_{x_1+x_2+\cdots+x_n}(\theta) = M_{x_1}(\theta) * M_{x_2}(\theta) * \cdots * M_{x_n}(\theta) = M_X^n(\theta) \qquad (1.5)$$

Por otra parte, se sabe también que la *función generatriz de momentos del producto de una constante por una variable* para el parámetro θ, $M_{CX}(\theta)$, es la función generatriz de momentos del producto del parámetro (θ) por la constante (C) para la variable X

$$M_{CX}(\theta) = M_X(C\theta) \qquad (1.6)$$

Con las propiedades anteriores se puede hallar la función generatriz de momentos de la media aritmética de la muestra en el parámetro θ $(M_{\bar{X}}(\theta))$

$$M_{\bar{X}}(\theta) = M_{\frac{1}{n}(x_1+x_2+\cdots+x_n)}(\theta) = M_{x_1+x_2+\cdots+x_n}\left(\frac{\theta}{n}\right)$$

Con este último resultado y teniendo en cuenta las propiedades

$$M_{\bar{X}}(\theta) = M_X^n\left(\frac{\theta}{n}\right) \qquad (1.7)$$

La función generatriz de momentos de la media $\bar{X}$ en el parámetro θ, es la función generatriz de momentos de la variable X elevado a la n en el parámetro $\frac{\theta}{n}$.

La expresión (1.7) relaciona la función generatriz de momentos de la media muestral con la variable observada en la población, a partir de los datos correspondientes a las unidades de observación seleccionadas por medio de una muestra, en esa población.

1.3. Teorema central del límite

Con la distribución de probabilidad de X (variable aleatoria poblacional) y muestras aleatorias de tamaño n, de donde se obtiene $\bar{X}$ (variable aleatoria muestral), se analizará cómo se distribuye $\bar{X}$. La respuesta no es única, pues depende de cuál sea la distribución de probabilidad de la variable X. Atendiendo a esto, se tendrán dos alternativas, según se esté en presencia de:

a) La variable X con distribución de probabilidad normal
b) La variable X con cualquier tipo de distribución de probabilidad y con tamaño de muestra lo suficientemente grande $(n > 30)$

La primera pate del *teorema central del límite* dice que: "Si la variable aleatoria X asociada a una población tiene distribución normal de media μ_X y desviación estándar σ_X, entonces la variable aleatoria $\bar{X}$ asociada a la distribución de medias de las muestras también es normal, de media $E(\bar{X}) = \mu_X$ y desvío $\sigma_{\bar{X}} = \sigma_X/\sqrt{n}$ ".

Para demostrar el teorema se calcula el momento natural de orden 1 y el momento natural de orden 2. Para esto se deriva la función generatriz de momentos respecto de θ, y se evalúa θ en 0. Entonces,

$$m_1' = \frac{\partial M_{\bar{X}}(\theta)}{\partial \theta} = \frac{\partial M_X^n\left(\frac{\theta}{n}\right)}{\partial \theta}$$

$$m_2' = \frac{\partial^2 M_{\bar{X}}(\theta)}{\partial \theta^2} = \frac{\partial^2 M_X^n\left(\frac{\theta}{n}\right)}{\partial \theta^2}$$

En el cuaderno anterior se demostró que la función generatriz de momentos de la variable $X \sim N(\mu_X, \sigma_X)$ es:

$$M_X(\theta) = e^{\theta\mu_X + \frac{1}{2}\theta^2\sigma_X^2}$$

Por lo tanto, $M_X^n\left(\frac{\theta}{n}\right)$ es

$$M_X^n\left(\frac{\theta}{n}\right) = \left[e^{\frac{\theta}{n}\mu_X + \frac{1}{2}\sigma_X^2\frac{\theta^2}{n^2}}\right]^n$$

De esta forma, la función generatriz de momentos de la media muestral $\bar{X}$ cuando $X \sim N(\mu_X, \sigma_X)$ es

$$M_{\bar{X}}(\theta) = M_X^n\left(\frac{\theta}{n}\right) = \left[e^{\frac{\theta}{n}\mu_X + \frac{1}{2}\sigma_X^2\frac{\theta^2}{n^2}}\right]^n = \left[e^{\frac{\theta}{n}\mu_X} e^{\frac{1}{2}\sigma_X^2\frac{\theta^2}{n^2}}\right]^n = \left(e^{\frac{\theta}{n}\mu_X}\right)^n \left(e^{\frac{1}{2}\sigma_X^2\frac{\theta^2}{n^2}}\right)^n$$

Entonces,

$$M_{\bar{X}}(\theta) = e^{\theta\mu_X + \frac{1}{2}\sigma_X^2\frac{\theta^2}{n}} \qquad (1.8)$$

Ahora se puede calcular el momento natural de orden 1 y obtener la esperanza matemática de $\bar{X}$

$$m_1' = \frac{\partial M_{\bar{X}}(\theta)}{\partial\theta} = \frac{\partial M_X^n\left(\frac{\theta}{n}\right)}{\partial\theta} = e^{\theta\mu_X + \frac{1}{2}\sigma_X^2\frac{\theta^2}{n}}\left(\mu_X + \sigma_X^2\frac{\theta}{n}\right)$$

Evaluando en $\theta = 0$

$$m_1' = e^{0\mu_X + \frac{1}{2}\sigma_X^2\frac{0^2}{n}}\left(\mu_X + \sigma_X^2\frac{0}{n}\right) = \mu_X$$

De modo que el momento de orden 1 es

$$m_1' = E(\bar{X}) = \mu_X \qquad (1.9)$$

Con lo que se demuestra la propiedad enunciada más arriba, esto es que la media aritmética de la muestra es un estimador insesgado de la media poblacional.

Para hallar el momento natural de orden 2, se deriva la derivada que dio origen al momento de orden 1:

$$m_2' = \frac{\partial^2 M_{\bar{X}}(\theta)}{\partial \theta^2} = e^{\theta \mu_X + \frac{1}{2}\sigma_X^2 \frac{\theta^2}{n}} \left(\mu_X + \sigma_X^2 \frac{\theta}{n}\right)^2 + e^{\theta \mu_X + \frac{1}{2}\sigma_X^2 \frac{\theta^2}{n}} \left(\frac{\sigma_X^2}{n}\right)$$

Evaluando en $\theta = 0$

$$m_2' = \frac{\partial^2 M_{\bar{X}}(\theta)}{\partial \theta^2} = e^0 \left(\mu_X + \sigma_X^2 \frac{0}{n}\right)^2 + e^0 \left(\frac{\sigma_X^2}{n}\right)$$

Entonces el momento natural de orden 2 es

$$m_2' = \frac{\partial^2 M_{\bar{X}}(\theta)}{\partial \theta^2} = (\mu_X)^2 + \left(\frac{\sigma_X^2}{n}\right)$$

La varianza de $\bar{X}$ es,

$$V(\bar{X}) = m_2' - (m_1')^2$$

$$V(\bar{X}) = \sigma_{\bar{X}}^2 = (\mu_X)^2 + \left(\frac{\sigma_X^2}{n}\right) - (\mu_X)^2 = \frac{\sigma_X^2}{n} \qquad (1.10)$$

y el desvío de $\bar{X}$ será

$$\sigma_{\bar{X}} = \sqrt{\frac{\sigma_X^2}{n}} = \frac{\sigma_X}{\sqrt{n}} \qquad (1.11)$$

Los resultados alcanzados en (1.9) y (1.11) confirman la aseveración del teorema.

En síntesis, si $X \sim N(\mu_X, \sigma_X)$ la función generatriz de momentos de la variable aleatoria X en el parámetro θ es $M_X(\theta) = e^{\theta \mu_X + \frac{1}{2}\theta^2 \sigma_X^2}$ y la función generatriz de momentos de la variable aleatoria $\bar{X}$ es $M_{\bar{X}}(\theta) = e^{\theta \mu_X + \frac{1}{2}\sigma_X^2 \frac{\theta^2}{n}}$ siendo $\bar{X} \sim N(\mu_X, \frac{\sigma_X}{\sqrt{n}})$.

La segunda parte del teorema central del límite dice que: "Si X es una variable aleatoria con cualquier distribución de probabilidad de media μ_X y desviación estándar σ_X, para la cual existe su función generatriz de momentos, entonces la variable aleatoria $\bar{X}$ tiende asintóticamente a la distribución normal de media μ_X y desviación estándar $\frac{\sigma_X}{\sqrt{n}}$ ".

Para demostrar, qué sucede si X tiene alguna distribución distinta a la normal pero tiene una media μ_X y un desvío σ_X y la muestra es lo suficientemente grande $(n > 30)$, se define una variable aleatoria estandarizada correspondiente a la distribución de medias de las muestras:

$$\mathrm{t} = \frac{\bar{\mathrm{x}} - \mu_{\mathrm{x}}}{\sigma_{\bar{\mathrm{x}}}} = \frac{\bar{\mathrm{x}} - \mu_{\mathrm{x}}}{\sigma_{\mathrm{x}}/\sqrt{\mathrm{n}}} = Z_{\bar{x}}$$

La función generatriz de momentos de t en el parámetro θ

$M_t(\theta) = M_{\frac{\bar{\mathrm{x}} - \mu_{\mathrm{x}}}{\sigma_{\mathrm{x}}/\sqrt{\mathrm{n}}}}(\theta)$ donde $\sigma_{\mathrm{x}}/\sqrt{\mathrm{n}}$ es constante

Aplicando la propiedad $M_{CX}(\theta) = M_X(C\theta)$

$M_t(\theta) = M_{\bar{X} - \mu_X}\left(\frac{\theta}{\sigma_{\mathrm{x}}/\sqrt{\mathrm{n}}}\right)$ donde μ_X es constante

Además, se vio que $M_{C+X}(\theta) = \mathrm{e}^{C\theta} M_X(\theta)$, con lo cual

$$M_t(\theta) = e^{\frac{-\mu_X \theta}{\sigma_X/\sqrt{n}}} \; M_X\left(\frac{\theta}{\sigma_X/\sqrt{n}}\right)$$

Teniendo en cuenta además que $M_{\bar{x}}(\theta) = M_x^n\left(\frac{\theta}{n}\right)$

Entonces,

$$M_t(\theta) = \mathrm{e}^{\frac{-\mu_x\theta\sqrt{n}}{\sigma_x}} M_x^n\left(\frac{\theta/n}{\sigma_x/\sqrt{n}}\right) = \mathrm{e}^{\frac{-\mu_x\theta\sqrt{n}}{\sigma_x}} M_x^n\left(\frac{\theta}{\sigma_x\sqrt{n}}\right) \qquad (1.12)$$

Esto es así porque

$$\frac{\theta/n}{\sigma_x/\sqrt{n}} = \frac{\theta\sqrt{n}}{\sigma_x n} = \frac{\theta n^{1/2}}{\sigma_x n} = \frac{\theta n^{1/2-1}}{\sigma_x} = \frac{\theta n^{-1/2}}{\sigma_x} = \frac{\theta}{\sigma_x\sqrt{n}}$$

Al tomar logaritmo natural en (1.12)

$$\ln[M_t(\theta)] = \frac{-\mu_x\theta\sqrt{n}}{\sigma_x} + n.\ln\left[M_x\left(\frac{\theta}{\sigma_x.\sqrt{n}}\right)\right] \qquad (1.13)$$

OBSERVACIÓN: Hay que tener en cuenta que

$M_x(\theta) = \mathrm{E}\{\mathrm{e}^{\theta x}\}$

y que $\mathrm{e}^{\theta x}$ puede desarrollarse en series de potencias, esto significa que el número e elevado a cualquier potencia se puede expresar de la siguiente manera:

$$\mathrm{e}^{w} = 1 + w + \frac{w^2}{2!} + \frac{w^3}{3!} + \cdots$$

Entonces, haciendo $w = \theta X$

$$M_x(\theta) = E\left\{1 + \theta X + \frac{(\theta X)^2}{2!} + \frac{(\theta X)^3}{3!} + \cdots\right\}$$

La esperanza es distributiva respecto de la suma

$$M_x(\theta) = E\{1\} + E\{\theta X\} + E\left\{\frac{(\theta X)^2}{2!}\right\} + E\left\{\frac{(\theta X)^3}{3!}\right\} + \cdots$$

Pero luego $E\{CX\} = C.E\{X\}$

$$M_x(\theta) = E\{1\} + \theta E\{X\} + \frac{\theta^2}{2!}E\{X^2\} + \frac{\theta^3}{3!}E\{X^3\} + \cdots$$

De modo que la función generatriz de la variable X es

$$M_x(\theta) = 1 + \theta\mu_1' + \frac{\theta^2}{2!}\mu_2' + \frac{\theta^3}{3!}\mu_3' + \cdots \qquad (1.14)$$

Donde se ha tenido en cuenta que $E\{1\} = E\{C\} = C = 1$, y que, $E\{X\} = \mu_1'$ es el momento con respecto al origen de la variable aleatoria X de órdenes crecientes.

Derivando sucesivamente la expresión (1.14)

$$\frac{d}{d\theta}M_X(\theta) = \frac{d}{d\theta}\left[1 + \theta\mu_1' + \frac{\theta^2}{2!}\mu_2' + \frac{\theta^3}{3!}\mu_3' + \frac{\theta^4}{4!}\mu_4' + \cdots\right]$$

$$= 0 + \mu_1' + \frac{2}{2}\theta\mu_2' + \frac{3}{3.2}\theta^2\mu_3' + \frac{4}{4.3.2}\theta^3\mu_4' + \cdots$$

Si $\theta = 0$ entonces $\frac{d}{d\theta}M_X(\theta) = \mu_1'$

A continuación se busca la segunda derivada

$$\frac{d^2}{d\theta^2}M_X(\theta) = \frac{d}{d\theta}\left[\mu_1' + \frac{2}{2}\theta\mu_2' + \frac{3}{3.2}\theta^2\mu_3' + \frac{4}{4.3.2}\theta^3\mu_4' + \cdots\right]$$

$$= 0 + \mu_2' + \frac{2}{2}\theta\mu_3' + \frac{3}{3.2}\theta^2\mu_4' + \cdots$$

Si $\theta = 0$ entonces $\frac{d^2}{d\theta^2} M_X(\theta) = \mu_2'$

Luego la tercer derivada

$$\frac{d^3}{d\theta^3} M_X(\theta) = \frac{d}{d\theta}\left[\mu_2' + \frac{2}{2}\theta\mu_3' + \frac{3}{3.2}\theta^2\mu_4' + \cdots\right]$$

$$= 0 + \mu_3' + \theta\mu_4' + \cdots$$

Si $\theta = 0$ entonces $\frac{d^3}{d\theta^3} M_X(\theta) = \mu_3'$

Las derivaciones anteriores permiten generalizar el resultado en la siguiente expresión

$$\frac{d^k}{d\theta^k} M_X(\theta)|_{\theta=0} = \mu_k'$$

La derivada k-ésima respecto al parámetro θ, k veces, evaluada en $\theta = 0$, es el k-ésimo momento de la variable aleatoria con respecto al origen.

En (1.13) se tiene que

$$\ln[M_t(\theta)] = \frac{-\mu_x\theta\sqrt{n}}{\sigma_x} + n.\ln\left[M_x\left(\frac{\theta}{\sigma_x.\sqrt{n}}\right)\right]$$

Remplazando aquí la expresión encontrada en (1.13) para $M_x\left(\frac{\theta}{\sigma_x.\sqrt{n}}\right)$, se tiene

$$\ln[M_t(\theta)] = \frac{-\mu_x\theta\sqrt{n}}{\sigma_x} + n.\ln\left[1 + \frac{\theta}{\sigma_x.\sqrt{n}}\mu_1' + \frac{1}{2!}\frac{\theta^2}{\sigma_x^2 n}\mu_2' + \frac{1}{3!}\frac{\theta^3}{\sigma_x^3 n^{3/2}}\mu_3' + \frac{1}{4!}\frac{\theta^4}{\sigma_x^4 n^{4/2}}\mu_4' + \cdots\right] \quad (1.14)$$

OBSERVACIÓN: Si $n > 30$ se lo considera lo suficientemente grande como para que el contenido del corchete sea una suma de términos que tiende a $1 +$ algo cercano a 0. Esto se puede desarrollar usando una expansión en serie de potencia donde

$$\ln(1+Z) = Z - \frac{Z^2}{2} + \frac{Z^3}{3} - \frac{Z^4}{4} + \cdots \qquad |Z| < 1$$

Z será igual a

$$Z = \frac{\theta}{\sqrt{n}\,\sigma_x}\mu_1' + \frac{1}{2!}\frac{\theta^2}{n\,\sigma_x^2}\mu_2' + \frac{1}{3!}\frac{\theta^3}{n^{3/2}\,\sigma_x^3}\mu_3' + \frac{1}{4!}\frac{\theta^4}{n^{4/2}\,\sigma_x^4}\mu_4' + \cdots$$

que es parte de la expresión que se encuentra entre corchetes en (1.14).

Si se desarrolla la serie en potencia

$$\begin{aligned}\ln(1+Z) = &\left[\frac{\theta}{\sqrt{n}\sigma_x}\mu_1' + \frac{1}{2!}\frac{\theta^2}{n\sigma_x^2}\mu_2' + \frac{1}{3!}\frac{\theta^3}{n^{3/2}\,\sigma_x^3}\mu_3' + \frac{1}{4!}\frac{\theta^4}{n^{4/2}\,\sigma_x^4}\mu_4' + \cdots\right] \\ &- \frac{1}{2}\left[\frac{\theta}{\sqrt{n}\sigma_x}\mu_1' + \frac{1}{2!}\frac{\theta^2}{n\sigma_x^2}\mu_2' + \frac{1}{3!}\frac{\theta^3}{n^{3/2}\,\sigma_x^3}\mu_3' + \frac{1}{4!}\frac{\theta^4}{n^{4/2}\,\sigma_x^4}\mu_4' + \cdots\right]^2 \\ &+ \frac{1}{3}\left[\frac{\theta}{\sqrt{n}\sigma_x}\mu_1' + \frac{1}{2!}\frac{\theta^2}{n\sigma_x^2}\mu_2' + \frac{1}{3!}\frac{\theta^3}{n^{3/2}\,\sigma_x^3}\mu_3' + \frac{1}{4!}\frac{\theta^4}{n^{4/2}\,\sigma_x^4}\mu_4' + \cdots\right]^3 \\ &- \frac{1}{4}\left[\frac{\theta}{\sqrt{n}\sigma_x}\mu_1' + \frac{1}{2!}\frac{\theta^2}{n\sigma_x^2}\mu_2' + \frac{1}{3!}\frac{\theta^3}{n^{3/2}\,\sigma_x^3}\mu_3' + \frac{1}{4!}\frac{\theta^4}{n^{4/2}\,\sigma_x^4}\mu_4' + \cdots\right]^4 \\ &+ \cdots\end{aligned}$$

Luego se tienen que desarrollar los cuadrados, los cubos, las cuartas potencias… Para resolver el cuadrado puede considerarse que se está en presencia de un binomio $(A + B)$

$$\Big[\underbrace{\frac{\theta}{\sqrt{n}\sigma_x}\mu_1'}_{A} + \underbrace{\frac{1}{2!}\frac{\theta^2}{n\sigma_x^2}\mu_2' + \frac{1}{3!}\frac{\theta^3}{n^{3/2}\,\sigma_x^3}\mu_3' + \frac{1}{4!}\frac{\theta^4}{n^{4/2}\,\sigma_x^4}\mu_4' + \cdots}_{B}\Big]^2$$

El desarrollo del cuadrado da lugar a la expresión

$$\frac{\theta^2}{n\sigma_x^2}(\mu_1')^2 + 2\frac{\theta}{\sqrt{n}\,\sigma_x}\mu'_1\left[\frac{1}{2!}\frac{\theta^2}{n\sigma_x^2}\mu_2' + \frac{1}{3!}\frac{\theta^3}{n^{3/2}\,\sigma_x^3}\mu_3' + \frac{1}{4!}\frac{\theta^4}{n^{4/2}\,\sigma_x^4}\mu_4' + \cdots\right] + \left[\frac{1}{2!}\frac{\theta^2}{n\sigma_x^2}\mu_2' + \frac{1}{3!}\frac{\theta^3}{n^{3/2}\,\sigma_x^3}\mu_3' + \frac{1}{4!}\frac{\theta^4}{n^{4/2}\,\sigma_x^4}\mu_4' + \cdots\right]^2$$

De igual manera hay que trabajar para resolver las potencias restantes. Entonces, se tiene que

$$\begin{aligned}\ln(1+Z) = {} & \frac{\theta}{\sqrt{n}\,\sigma_X}\mu_1' + \frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\mu_2' + \frac{1}{3!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_3' + \frac{1}{4!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_4' \\ & + \frac{1}{5!}\frac{\theta^5}{n^{5/2}\sigma_X^5}\mu_5' + \cdots \\ & - \frac{1}{2}\left\{\frac{\theta^2}{n\sigma_X^2}(\mu_1')^2 + \frac{2}{2!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_1'\mu_2' + \frac{2}{3!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_1'\mu_3'\right. \\ & + \frac{2}{4!}\frac{\theta^5}{n^{5/2}\sigma_X^5}\mu_1'\mu_4' + \frac{1}{6!}\frac{\theta^6}{n^{6/2}\sigma_X^6}\mu_1'\mu_5' \\ & \left. + \cdots\left[\frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\mu_2' + \frac{1}{3!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_3' + \frac{1}{4!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_4' + \cdots\right]^2\right\} \\ & + \frac{1}{3}\left[\frac{\theta}{\sqrt{n}\,\sigma_X}\mu_1' + \frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\mu_2' + \frac{1}{3!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_3' + \frac{1}{4!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_4' + \cdots\right]^3 \\ & + \cdots\end{aligned}$$

Agrupando los términos por factor común

$$\ln(1+Z) = \frac{\theta}{\sqrt{n}\,\sigma_X}\mu_1' + \frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}[\mu_2' - (\mu_1')^2] + \frac{\theta^3}{n^{3/2}\sigma_X^3}\left[\frac{\mu_3'}{3!} - \frac{\mu_1'\mu_2'}{2!}\right]$$
$$+ \frac{\theta^4}{n^{4/2}\sigma_X^4}\left[\frac{\mu_4'}{4!} - \frac{\mu_1'\mu_3'}{3!}\right] + \frac{\theta^5}{n^{5/2}\sigma_X^5}\left[\frac{\mu_5'}{5!} - \frac{\mu_1'\mu_4'}{4!}\right]$$
$$+ \frac{\theta^6}{n^{6/2}\sigma_X^6}\left[\frac{\mu_6'}{6!} - \frac{\mu_1'\mu_5'}{5!}\right]$$
$$+ \cdots\left[\frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\mu_2' + \frac{1}{3!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_3' + \frac{1}{4!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_4' + \cdots\right]^2$$
$$+ \frac{1}{3}\left[\frac{\theta}{\sqrt{n}\,\sigma_X}\mu_1' + \frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\mu_2' + \frac{1}{3!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_3' + \frac{1}{4!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_4' + \cdots\right]^3$$
$$+ \cdots$$

Reemplazando el resultado del $\ln(1+Z)$ en (1.14)

$$\ln[M_t(\theta)] = \frac{-\mu_X\theta\sqrt{n}}{\sigma_X}$$
$$+ n\left\{\frac{\theta}{\sqrt{n}\,\sigma_X}\mu_1' + \frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}[\mu_2' - (\mu_1')^2] + \frac{\theta^3}{n^{3/2}\sigma_X^3}\left[\frac{\mu_3'}{3!} - \frac{\mu_1'\mu_2'}{2!}\right]\right.$$
$$+ \frac{\theta^4}{n^{4/2}\sigma_X^4}\left[\frac{\mu_4'}{4!} - \frac{\mu_1'\mu_3'}{3!}\right] + \frac{\theta^5}{n^{5/2}\sigma_X^5}\left[\frac{\mu_5'}{5!} - \frac{\mu_1'\mu_4'}{4!}\right] + \frac{\theta^6}{n^{6/2}\sigma_X^6}\left[\frac{\mu_6'}{6!} - \frac{\mu_1'\mu_5'}{5!}\right]$$
$$+ \cdots\left[\frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\mu_2' + \frac{1}{3!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_3' + \frac{1}{4!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_4' + \cdots\right]^2$$
$$\left.+ \frac{1}{3}\left[\frac{\theta}{\sqrt{n}\,\sigma_X}\mu_1' + \frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\mu_2' + \frac{1}{3!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_3' + \frac{1}{4!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_4' + \cdots\right]^3 + \cdots\right\}$$

Al reordenar y simplificar la expresión:

$$\ln[M_t(\theta)] = \frac{-\mu_X\sqrt{n}}{\sigma_X}\theta + \frac{\sqrt{n}}{\sigma_X}\theta\mu_1' + \frac{\mu_2' - (\mu_1')^2}{2!\,\sigma_X^2}\theta^2 + \frac{1}{n^{1/2}\sigma_X^3}\left[\frac{\mu_3'}{3!} - \frac{\mu_1'\mu_2'}{2!}\right]\theta^3 + \frac{1}{n^{2/2}\sigma_X^4}\left[\frac{\mu_4'}{4!} - \frac{\mu_1'\mu_3'}{3!}\right]\theta^4 + \frac{1}{n^{3/2}\sigma_X^5}\left[\frac{\mu_5'}{5!} - \frac{\mu_1'\mu_4'}{4!}\right]\theta^5 + \cdots términos\ en\ \theta^k\ para\ k > 5$$

Si se tiene en cuenta que $\mu_X = \mu_1'$ y que $\sigma_X^2 = \mu_2' - (\mu_1')^2$, entonces

$$\ln[M_t(\theta)] = \left[-\frac{\mu_X\sqrt{n}}{\sigma_X} + \frac{\mu_X\sqrt{n}}{\sigma_X}\right]\theta + \frac{\sigma_X^2}{2!\,\sigma_X^2}\theta^2 + \frac{1}{n^{1/2}\sigma_X^3}\left[\frac{\mu_3'}{3!} - \frac{\mu_1'\mu_2'}{2!}\right]\theta^3 + \frac{1}{n^{2/2}\sigma_X^4}\left[\frac{\mu_4'}{4!} - \frac{\mu_1'\mu_3'}{3!}\right]\theta^4 + \frac{1}{n^{3/2}\sigma_X^5}\left[\frac{\mu_5'}{5!} - \frac{\mu_1'\mu_4'}{4!}\right]\theta^5 + \cdots términos\ en\ \theta^k\ para\ k > 5$$

Simplificando la expresión

$\ln[M_t(\theta)] = \frac{\theta^2}{2}$ + términos en θ^k con $k > 2$

Pero los términos en θ^k cuando $k > 2$ tienden a 0 cuando $n \to \infty$, por lo que se puede afirmar que

$$\lim_{n\to\infty} \ln[M_t(\theta)] = \frac{\theta^2}{2}$$

Es decir que $\lim_{n\to\infty} M_t(\theta) = e^{\frac{\theta^2}{2}}$ porque $\ln e^{\frac{\theta^2}{2}} = \frac{\theta^2}{2}$ (1.15)

Si se compara (1.15) con la función generatriz de momentos de una variable aleatoria normal estandarizada, se observa que la función generatriz de momentos de la variable aleatoria estándar t tiende a la función generatriz de momentos de la distribución normal estándar a medida que el tamaño de la muestra aumenta.

En el Teorema Central del Límite radica la importancia de la distribución normal, ya que basta con que el tamaño de la muestra sea suficientemente grande, mayor a 30, para poder asegurar la buena aproximación de la distribución de medias de la muestra a la distribución normal, independientemente de la propia distribución que tenga la población de donde se extraigan las muestras.

Con este teorema queda resuelto el problema de la distribución de probabilidad de las medias de las muestras, cuando estas provienen de una población normal o de una población con cualquier distribución de probabilidad y el tamaño de las muestras es grande. En ambos casos se recurre a la distribución normal.

Para Gomez Villegas (2005), la aplicación del teorema central del límite como justificación de que los errores de medida se distribuyen aproximadamente según una normal es considerada como una de las más importantes aportaciones científicas. En realidad, en los siglos XVII y XVIII, el teorema central del límite se conoció como la ley de frecuencias de los errores. Muchos científicos pensaron que la ley de frecuencias de los errores suponía un gran avance. Al respecto, el autor cita el comentario de Francis Galton (*La herencia natural* publicado en 1889) *"difícilmente conozco algo que alimente tanto mi imaginación como el maravilloso orden cósmico que se deriva de la "Ley de frecuencias de los errores". Si los griegos hubieran conocido esta ley, seguro que la habrían endiosado. Reina con serenidad y en completa auto modestia entre la confusión más salvaje. Cuanto más vigentes están la ley de la calle y la aparente anarquía, más perfecto es su balanceo. Es la ley suprema de la sinrazón"*.

Factor de corrección para poblaciones finitas

Las poblaciones pueden ser finitas o infinitas. Si de una población finita se obtienen muestras con reemplazo, puede considerarse teóricamente que la población es infinita ya que cualquier número de muestras pueden obtenerse de la población sin agotarla. Además, para casi cualquier propósito práctico, el muestreo de poblaciones

finitas de tamaño muy grande puede considerase como muestreo de poblaciones infinitas.

De acuerdo al concepto de muestreo aleatorio, el muestreo sin reemplazo de poblaciones finitas no conduce a muestras aleatorias, debido a que al dejar permanentemente fuera el elemento ya extraído, influye en la probabilidad de extracción del siguiente; o sea, que las pruebas sucesivas para obtener una muestra no son independientes. Esto no ocurre si la población es infinita.

Resumiendo lo anterior puede decirse que, para obtener muestras aleatorias, el muestreo debe ser de cualquier tipo en poblaciones infinitas y sólo con reemplazo en poblaciones finitas, a excepción del caso de poder considerar a la población infinita por ser de tamaño muy grande.

Con respecto al teorema, cabe indicar que cuando la población es finita y el muestreo se realiza sin reemplazo, hay que incluir -en la determinación de σ_X- el valor del tamaño de la población, es decir,

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} . \sqrt{\frac{N-n}{N-1}}$$

donde $\frac{N-n}{N-1}$ recibe el nombre de factor de corrección para poblaciones finitas en muestreo sin reemplazo. Obsérvese que este factor de corrección tiende a cero cuando $n \to N$.

Ejemplo 1.3. Olivera Zalazar y Zúñiga Barrera (1987) plantean que la vida útil de los focos fabricados por una empresa tiene distribución normal de media 200 horas y desviación estándar 25 horas. En este contexto, piden calcular la probabilidad de que la duración media de 25 focos escogidos al azar sea superior a 208 horas.

De acuerdo al enunciado del problema, la población de las duraciones de los focos de la empresa tiene distribución normal de parámetros.

$\mu_x = 200$ horas y $\sigma_X = 25$ horas

y puede considerarse que es de tamaño infinito. Entonces, la distribución de las medias de las muestras también tendrá distribución normal de parámetros:

$n = 25$ focos

$\mu_{\bar{X}} = \mu_x = 200$ horas

$\sigma_{\bar{X}} = \frac{\sigma_x}{\sqrt{n}} = \frac{25}{\sqrt{25}} = 5$ horas

Estandarizando la variable aleatoria $\bar{X}$, se tiene

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - 200}{5}$$

Si $\bar{X} = 208$, $z = \frac{208-200}{5} = 1{,}6$

Por lo tanto, la probabilidad de que la media de los 25 focos de la muestra tenga un valor mayor a 208 horas valdrá:

$$P(\bar{X} > 208) = P(z > 1{,}6)$$
$$= 1 - P(z \leq 1{,}6) = 1 - 0{,}9452 = 0{,}0548$$

Sólo el $5{,}48\%$ de los focos de la muestra tendrán una vida superior a las 208 horas, es decir, un solo foco.

Ejemplo 1.4. Olivera Zalazar y Zúñiga Barrera (1987) citan el caso de los pesos de 600 ejes producidos en un torno, que tienen distribución normal de media 53 kg y desviación estándar de 2,5 kg. Los ejes se empacan en cajas de 10 unidades que soportan, cada una, hasta 540 kg de peso. Si se envían 35 de estas cajas, calcular cuántas cajas cabe esperar se rompan por exceso de peso.

La población de pesos de los ejes tiene distribución normal de acuerdo al enunciado explícito del problema. Sus parámetros son:

$N = 600$, $\mu_x = 53Kg$, $\sigma_x = 2{,}5Kg$

Como la población es normal, la distribución de medias de las muestras también es normal y sus parámetros valen;

$n = 10$, $\mu_{\bar{x}} = \mu_x = 53Kg$

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{2{,}5}{\sqrt{10}}\sqrt{\frac{600-10}{600-1}} = 0{,}785Kg$$

Una caja de ejes se romperá por exceso de peso, si el peso conjunto de los 10 ejes es mayor de 540 kg; y el peso conjunto de los 10 ejes es mayor de 540 kg si el peso medio de los 10 ejes es superior a 54kg (540/10).

Entonces, se trata de calcular cuántas de las 35 muestras tienen media mayor de 54 kg; esto se hará por medio del cálculo de la probabilidad $P(\bar{X} > 54)$, en donde $\bar{X}$ tiene distribución normal de media 53 y desviación estándar 0,785.

Estandarizando $\bar{X} = 54$, se tiene

$$z = \frac{\bar{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{54-53}{0{,}785} = 1{,}274$$

Por lo tanto,

$$P(\bar{X} > 54) = P(z > 1{,}274)$$
$$= 1 - P(z \leq 1{,}274) = 1 - 0{,}8987 = 0{,}1013$$

El $10{,}13\%$ de las 35 cajas se romperán, o sea, de 3 a 4 cajas.

Ejemplo 1.5. Olivera Zalazar y Zúñiga Barrera (1987) trabajan con los tabiques comprimidos que se usan en una construcción que tienen un peso medio de 5,50 kg y una desviación estándar de 0,85 kg. Estos se elevan en lotes, al lugar en donde se emplean, por medio de una grúa cuyo límite de seguridad es de 200 kg. Con esta información se necesita calcular el tamaño máximo de los lotes de manera de que la

probabilidad de exceder el límite de seguridad de la grúa sea menor de 5%.

Como en el ejemplo anterior, se vuelve a tener una población de pesos cuyos parámetros son

$\mu_x = 5{,}50Kg$ y $\sigma_x = 0{,}85Kg$

con distribución de probabilidad desconocida.

La simple comparación entre el peso de un tabique y el límite de seguridad de la grúa, hace que los lotes de tabiques que se cargan den muestras de tamaño grande. Esto permite asegurar que la distribución de medias de las muestras de pesos de los tabiques es aproximadamente normal de parámetros

$\mu_{\bar{x}} = \mu_x = 5{,}50Kg$ y $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{0{,}85}{\sqrt{n}}$

en donde sólo se sabe que n es grande (mayor de 30).

El peso de la muestra de n tabiques excede el límite de seguridad de la grúa de 200kg, si la media de la muestra es superior a $200/n$. Se busca saber cuál es el tamaño de la muestra del lote (cantidad de tabiques) que hace que el límite de seguridad sea superado. El problema se plantea como:

$$P\left(\bar{X} > \frac{200}{n}\right) < 0{,}05$$

Estandarizando x resulta $z = \frac{\bar{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{X} - 5{,}50}{0{,}85/\sqrt{n}}$

Si $\bar{X} = \frac{200}{n}$

Entonces $Z = \frac{200/n - 5{,}50}{0{,}85/\sqrt{n}} = \frac{200 - 5{,}50n}{0{,}85\sqrt{n}}$

De modo que

$$P\left(\bar{X} > \frac{200}{n}\right) = P\left(z > \frac{200 - 5{,}50n}{0{,}85\sqrt{n}}\right) < 0{,}05$$

lo que es lo mismo $P\left(z \leq \frac{200 - 5{,}50n}{0{,}85\sqrt{n}}\right) \geq 0{,}95$

De la tabla de probabilidades de la normal se obtiene que ésta es efectivamente mayor de $0{,}95$ si el valor crítico de z es mayor o igual a $1{,}645$.

Luego se obtiene

$$\frac{200 - 5{,}50n}{0{,}85\sqrt{n}} \geq 1{,}645$$

$$200 - 5{,}50n \geq 1{,}398\sqrt{n}$$

Elevando ambos miembros al cuadrado

$$(200 - 5{,}50n)^2 \geq (1{,}398\sqrt{n})^2$$

$$40000 - 2200n + 30{,}25n^2 \geq 1{,}955n$$

$$30{,}25n^2 - 2201{,}955n + 40000 \geq 0$$

Resolviendo la ecuación se obtienen las raíces $34{,}86$ y $37{,}93$ Cada una de estas raíces representa dos tamaños de muestras diferentes.

Si $Z \leq \frac{200-5{,}50n}{0{,}85\sqrt{n}}$, cuando $n = 34{,}86 \implies z \leq -1{,}645$

$$n = 37{,}93 \Longrightarrow z \leq 1{,}645$$

Luego

$P(z < -1{,}645) = 0{,}05$

$P(z < 1{,}645) = 0{,}95$

Se busca que el tamaño del lote no supere el límite de seguridad de la grúa fijado en el 5% Por esta razón, el tamaño de muestra requerido es

$\mathrm{n} \leq 34{,}86$

2. DISTRIBUCIÓN GAMMA

Se incluye una introducción a la Teoría de las Pequeñas Muestras, haciendo énfasis en distribuciones especiales que el investigador necesita para contrastar sus hipótesis de investigación. Estas distribuciones están basadas en la función Gamma del análisis matemático.

2.1 Características y momentos

La distribución Gamma servirá de punto de partida para obtener las funciones de densidad de las distribuciones llamadas exactas, que comprende a:

Distribución chi cuadrado de Pearson, χ^2

Distribución t de Student, t

Distribución F de Snedecor o Fischer, F

A continuación se utilizará genéricamente la variable X en representación de cualquier variable de la tabla de datos; de esta manera, X representará a cualquier $X_j, \forall j = 1, \ldots, k$.

La función gamma del análisis matemático es:

$$\Gamma(\lambda) = \int_0^{\infty} e^{-X} X^{\lambda-1} dX; \quad \lambda > 0 \qquad (2.40)$$

En la Figura 2.1 se observa gráficamente el comportamiento de la función Gamma

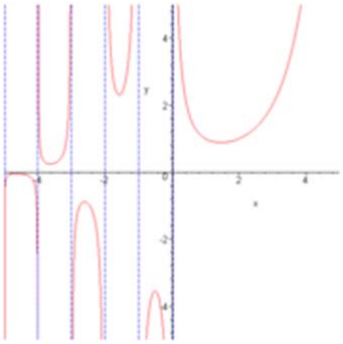


Figura 2.1: Función Gamma

Resolviendo esta integral se obtiene

$$\Gamma(\lambda) = (\lambda - 1)! \qquad (2.41)$$

Se debe tener presente, en los desarrollos posteriores, el valor notable:

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi} \qquad (2.42)$$

OBSERVACIÓN: Para resolver la integral (2.40) se integra por partes haciendo:

$$e^{-X} dX = dv \quad y \quad X^{\lambda-1} = u$$

Entonces se aplica la fórmula de integración por partes y se tiene

$$\Gamma(\lambda) = \int_0^{-\infty} u\, dv = uv \int_0^{-\infty} v\, du$$

$$= -e^{-X} X^{\lambda-1}|_0^{\infty} - \int_0^{\infty} -e^{-X}(\lambda-1)X^{\lambda-2} dX$$

$$= 0 + (\lambda-1) \underbrace{\int_0^{\infty} e^{-X} X^{\lambda-2} dX}_{\Gamma(\lambda-1)}$$

$$= (\lambda-1)\, \Gamma(\lambda-1)$$

Por lo tanto,

$$\Gamma(\lambda) = (\lambda-1)\Gamma(\lambda-1) = (\lambda-1)\,(\lambda-2)\Gamma(\lambda-2)$$

$$= (\lambda-1)\,(\lambda-2)\cdots\Gamma(1)$$

Pero

$$\Gamma(1) = \int_0^{\infty} e^{-X} X^{1-1} dX = \int_0^{\infty} e^{-X} dX = -e^{-X}|_0^{\infty} = -\frac{1}{e^{-X}}|_0^{\infty}$$
$$= \underbrace{\left(-\frac{1}{\infty}\right)}_{0} - \underbrace{\left(-\frac{1}{e^0}\right)}_{1} = 1$$

Entonces queda demostrado (2.41), ya que

$$\Gamma(\lambda) = (\lambda-1)(\lambda-2)\cdots = (\lambda-1)!$$

Por otra parte

$$\Gamma\left(\frac{1}{2}\right) = \int_0^{\infty} e^{-X} X^{\frac{1}{2}-1}\, dX = \int_0^{\infty} e^{-X} X^{-\frac{1}{2}} dX$$

Haciendo la sustitución,

$$X = z_1^2 \;\Rightarrow\; dX = 2z_1 dz_1 \;\; y \;\; z_1 = \sqrt{X}$$

Se tiene

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty e^{-X} X^{-\frac{1}{2}} dX = \int_0^\infty (z_1^2)^{-\frac{1}{2}} e^{-z_1^2} 2z_1 dz_1 =$$

$$= 2\int_0^\infty (z_1)^{-1} e^{-z_1^2} z_1 dz_1 = 2\int_0^\infty e^{-z_1^2} dz_1$$

De donde,

$$\Gamma\left(\frac{1}{2}\right) = 2\int_0^\infty e^{-z_1^2} dz_1$$

Elevando al cuadrado ambos miembros queda,

$$\Gamma\left(\frac{1}{2}\right)^2 = \left(2\int_0^\infty e^{-z_1^2} dz_1\right)\left(2\int_0^\infty e^{-z_1^2} dz_1\right)$$

$$\Gamma\left(\frac{1}{2}\right)^2 = 4\int_0^\infty \int_0^\infty e^{-z_1^2} e^{-z_2^2} dz_1 dz_2 =$$

$$= 4\int_0^\infty \int_0^\infty e^{-(z_1^2+z_2^2)} dz_1 dz_2$$

Pasando a coordenadas polares y haciendo la sustitución correspondiente de acuerdo con

$$r^2 = z_1^2 + z_2^2$$

Se sabe de (2.24) que la función de transformación es $z_1 = r\,cos\,w,\ z_2 = r\,sen\,w$

$$\Gamma\left(\frac{1}{2}\right)^2 = 4\int_0^{\frac{\pi}{2}} \int_0^\infty e^{-r^2} r dr\, dw$$

Donde $rdr\,dw$ es el resultado del jacobiano de la transformación.

Pero como,

$\int_0^\infty e^{f(X)} f'(X) dX = e^{f(X)} + C$, siendo C una constante de integración

Se tiene que

$$\int_0^\infty e^{-r^2} r dr = -\frac{1}{2}\int_0^\infty e^{-r^2}(2r) dr = -\frac{1}{2} e^{-r^2}|_0^\infty$$

O sea que en el límite cuando B tiende a infinito, esto es

$$\lim_{B\to\infty} -\frac{1}{2} e^{-r^2}|_0^B = \lim_{B\to\infty}\left[-\frac{1}{2e^{B^2}} + \frac{1}{2e^0}\right] \Rightarrow \Gamma\left(\frac{1}{2}\right)^2 = 4\int_0^{\frac{\pi}{2}} \frac{1}{2} dw$$

Finalmente,

$$\Gamma\left(\frac{1}{2}\right)^2 = 2\int_0^{\frac{\pi}{2}} dw = 2w|_0^{\frac{\pi}{2}} = 2\left(\frac{\pi}{2} - 0\right) = \pi$$

Con lo que queda demostrado (2.42) ya que

$$\Gamma\left(\frac{1}{2}\right)^2 = \pi \Rightarrow \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

Hay dos integrales de las que resulta útil conocer su resolución por la aplicación que de ellas se hace más adelante:

1) En la integral:

$\int_0^\infty e^{-\alpha X} X^{\lambda-1} dX;\ \alpha > 0,\ \lambda > 0$ (2.43)

se realiza la siguiente sustitución: $X\alpha = u$

de donde se deduce: $X = \frac{u}{\alpha};\quad dX = \frac{du}{\alpha}$

sustituyendo en (2.43)

$$\int_0^\infty e^{-u} \frac{u^{\lambda-1}}{\alpha^{\lambda-1}} \frac{du}{\alpha} = \int_0^\infty e^{-u} \frac{u^{\lambda-1}}{\alpha^\lambda} du$$

si se saca $\frac{1}{\alpha^{\lambda}}$ fuera de la integral y se compara con (2.40) queda:

$$\frac{1}{\alpha^{\lambda}}\int_0^{\infty} e^{-u}u^{\lambda-1}du = \frac{1}{\alpha^{\lambda}}\,\Gamma(\lambda)$$

o sea que:

$$\int_0^{\infty} e^{-\alpha X}X^{\lambda-1}dX = \frac{\Gamma(\lambda)}{\alpha^{\lambda}} \qquad (2.44)$$

2) En la integral

$$\int_0^{\infty} e^{-\alpha X^2}dX;\ \alpha > 0$$

Sustituyendo $X^2 = u$, se tiene

$$X = u^{1/2};\quad dX = \frac{1}{2}u^{-\frac{1}{2}}du$$

al sustituir se obtiene:

$$\frac{1}{2}\int_0^{\infty} e^{-\alpha u}u^{-\frac{1}{2}}du = \frac{\Gamma(\frac{1}{2})}{2\alpha^{\frac{1}{2}}} = \frac{\sqrt{\pi}}{2\alpha^{\frac{1}{2}}} \qquad (2.45)$$

La distribución gamma de la variable aleatoria X, con parámetros α y λ, se define mediante la siguiente función de densidad:

$$f(X) = \begin{cases} \frac{\alpha^{\lambda}}{\Gamma(\lambda)}e^{-\alpha X}X^{\lambda-1}; & \alpha > 0, \lambda > 0, X > 0 \\ 0; & X \leq 0 \end{cases} \qquad (2.46)$$

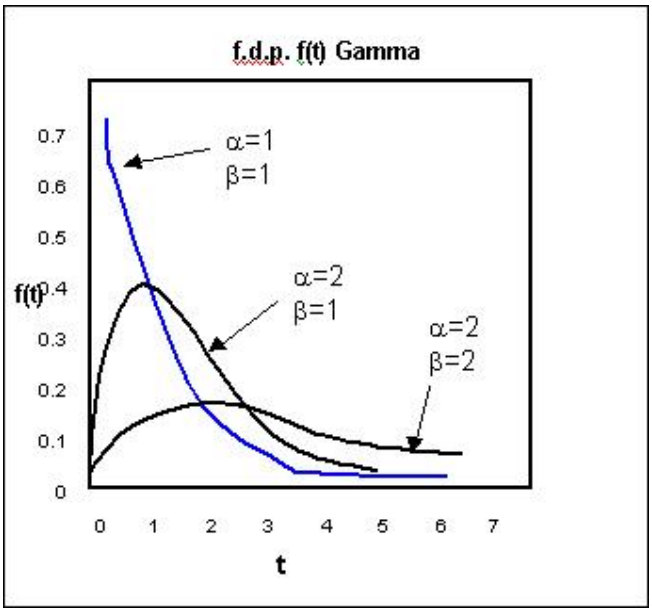


Figura 2.2: Distribución Gamma

Si se integra la función de densidad entre los límites de su recorrido, el resultado debe ser 1. El recorrido de esta función es de 0 a ∞, si se realiza la integral

$$\int_0^\infty f(X)dX = \int_0^\infty \frac{\alpha^\lambda}{\Gamma(\lambda)} e^{-\alpha X} X^{\lambda-1} dX = \frac{\alpha^\lambda}{\Gamma(\lambda)} \int_0^\infty e^{-\alpha X} X^{\lambda-1} dX$$

Teniendo en cuenta el resultado de (2.44)

$$\int_0^\infty f(X)dX = \frac{\alpha^\lambda}{\Gamma(\lambda)} \frac{\Gamma(\lambda)}{\alpha^\lambda} = 1$$

Con este resultado se confirma que la expresión hallada en (2.46) es la función de densidad de la distribución Gamma.

La función generatriz de momentos de esta distribución con función de densidad $f(X)$, es

$$M_X(\theta) = E\left(e^{\theta X}\right) = \int_0^\infty e^{\theta X} f(X) dX$$

que, haciendo uso de (2.46), queda:

$$\frac{\alpha^\lambda}{\Gamma(\lambda)} \int_0^\infty e^{-\alpha X} e^{\theta X} X^{\lambda-1} dX = \frac{\alpha^\lambda}{\Gamma(\lambda)} \int_0^\infty e^{-X(\alpha-\theta)} X^{\lambda-1} dX$$

La integral es análoga a (2.44) pero con $\alpha - \theta$ en lugar de α, luego,

$$M_X(\theta) = \frac{\alpha^\lambda \Gamma(\lambda)}{\Gamma(\lambda)(\alpha-\theta)^\lambda} = \frac{\alpha^\lambda}{(\alpha-\theta)^\lambda} = \frac{(\alpha-\theta)^{-\lambda}}{\alpha^{-\lambda}} = \left(\frac{\alpha-\theta}{\alpha}\right)^{-\lambda} = \left(\frac{\alpha}{\alpha} - \frac{\theta}{\alpha}\right)^{-\lambda} =$$
$$= \left(1 - \frac{\theta}{\alpha}\right)^{-\lambda}$$

La función generatriz de momentos de una variable X con distribución gamma, en el parámetro θ, es

$$M_X(\theta) = \left(1 - \frac{\theta}{\alpha}\right)^{-\lambda} \qquad (2.47)$$

Derivando sucesivamente la función generatriz se obtienen los momentos naturales de orden 1 y orden 2 (m_1'y m_2') que posibilitan el cálculo de la media y la varianza de la distribución gamma.

$$m_1' = \frac{\partial M_X(\theta)}{\partial \theta} = \frac{\partial \left(\frac{\alpha-\theta}{\alpha}\right)^{-\lambda}}{\partial \theta} = -\lambda \left(\frac{\alpha-\theta}{\alpha}\right)^{-\lambda-1} \frac{1}{\alpha}(-1)$$

Haciendo $\theta = 0$

$$m_1' = -\lambda \left(\frac{\alpha-0}{\alpha}\right)^{-\lambda-1} \left(-\frac{1}{\alpha}\right) = -\lambda \left(\frac{\alpha}{\alpha}\right)^{-\lambda-1} \left(-\frac{1}{\alpha}\right) =$$

$$= -\lambda(1)^{-\lambda-1} \left(-\frac{1}{\alpha}\right) = -\lambda(1) \left(-\frac{1}{\alpha}\right)$$

De donde, $m_1' = \frac{\lambda}{\alpha} = E(X)$

$$m_2' = \frac{\partial^2 M_X(\theta)}{\partial \theta^2} = \frac{\partial \left[-\lambda \left(\frac{\alpha-\theta}{\alpha} \right)^{-\lambda-1} \left(-\frac{1}{\alpha} \right) \right]}{\partial \theta} =$$

$$= \frac{\partial \left[\left(\frac{\alpha-\theta}{\alpha} \right)^{-\lambda-1} \left(\frac{\lambda}{\alpha} \right) \right]}{\partial \theta} =$$

$$= (-\lambda - 1) \left(\frac{\alpha-\theta}{\alpha} \right)^{-\lambda-2} \left(\frac{\lambda}{\alpha} \right) \left(-\frac{1}{\alpha} \right)$$

Haciendo $\theta = 0$

$$m_2' = (-\lambda - 1) \left(\frac{\alpha-0}{\alpha} \right)^{-\lambda-2} \left(\frac{\lambda}{\alpha} \right) \left(-\frac{1}{\alpha} \right) =$$

$$= (-\lambda - 1) \left(\frac{\alpha}{\alpha} \right)^{-\lambda-2} \left(\frac{\lambda}{\alpha} \right) \left(-\frac{1}{\alpha} \right) =$$

$$= (-\lambda - 1)(1)^{-\lambda-2} \left(\frac{\lambda}{\alpha} \right) \left(-\frac{1}{\alpha} \right) =$$

$$= (-\lambda - 1) \left(-\frac{\lambda}{\alpha^2} \right)$$

De donde, $m_2' = \frac{\lambda^2+\lambda}{\alpha^2}$

Por lo tanto,

$$V(X) = m_2' - (m_1')^2 = \frac{\lambda^2 + \lambda}{\alpha^2} - \frac{\lambda^2}{\alpha^2} = \frac{\lambda}{\alpha^2}$$

2.2. Distribuciones exactas

Las distribuciones Chi – Cuadrado de Pearson, F de Senedecor y t de Student, llamadas distribuciones exactas, son desarrolladas a continuación sobre la base de la distribución Gamma.

Gomez Villegas (2005) realiza una Aproximación histórica sobre estas distribuciones y dice que "*la distribución χ^2 fue introducida por Karl Pearson alrededor de 1900, al estudiar el problema de ajuste entre una distribución poblacional y unos datos muestrales. En 1904, la generalización de este problema al caso en que existan parámetros desconocidos en la distribución, dio lugar a que William Sealy Gosset (1876-1937) introdujera la distribución de Student, así llamada por ser Student el pseudónimo utilizado por Gosset para comunicar los resultados de sus investigaciones con muestra pequeñas, pues la empresa de elaboración de cerveza en la que trabajaba impedía la divulgación de los resultados obtenidos por sus empleados. Snedecor (1881-1974), estadístico estadounidense, introdujo la distribución que lleva su nombre alrededor de 1907. Snedecor trabajó en el laboratorio de estadística del University College fundado por Karl Pearson, junto a Student y Fischer. En agradecimiento a la importante aportación de Fischer a la Inferencia Estadística, Snedecor le dedicó la distribución del cociente de distribuciones χ^2 y la denotó mediante la letra F.*" (p. 43)

Distribución Chi-cuadrado (χ^2) de Pearson

Dada una variable poblacional $X \sim N(\mu, \sigma)$ cuyo recorrido, representado por los datos $x_1, x_2, \ldots, x_n$, se obtuvo de las n unidades de observación de una muestra aleatoria simple; el estadístico (variable aleatoria):

$$Y = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2} \qquad (2.48)$$

Se distribuye como una Chi-Cuadrado con n grados de libertad, y se representa

$$Y \sim \chi^2_n$$

Al desarrollar el sumatorio en (2.48), se tiene que

$$Y = \frac{(x_1 - \mu)^2}{\sigma^2} + \frac{(x_2 - \mu)^2}{\sigma^2} + \cdots + \frac{(x_n - \mu)^2}{\sigma^2} = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

que conduce a la forma alternativa de escribir (2.48):

$$Y = \sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$$

Donde, $Z_i \sim N(0,1)$

La función generatriz de momentos de la variable $Y \sim \chi_n^2$ es:

$$M_Y(\theta) = M_{\sum Z_i^2}(\theta) = M_{Z_1^2 + Z_2^2 + \cdots + Z_n^2}(\theta)$$

Aplicando la propiedad de función generatriz de momentos para la suma de variables aleatorias

$$M_Y(\theta) = M_{Z_1^2}(\theta)\, M_{Z_2^2}(\theta) \cdots M_{Z_n^2}(\theta) = [M_{Z^2}(\theta)]^n \qquad (2.49)$$

Entonces como $Z_i \sim N(0,1)$, se tiene

$$M_{Z^2}(\theta) = E\left(e^{\theta Z^2}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\theta Z^2} e^{-\frac{Z^2}{2}} dZ$$

Como se trata de una función par, integramos entre 0 e ∞ y multiplicamos por 2

$$\frac{2}{\sqrt{2\pi}} \int_0^{\infty} e^{-\left(\frac{1}{2} - \theta\right) Z^2} dZ \qquad (2.50)$$

Esta integral es análoga a la que se obtuvo en (2.45), para $\alpha = \frac{1}{2} - \theta$, por lo tanto

$$M_{Z^2}(\theta) = \frac{2}{\sqrt{2\pi}} \quad \frac{\sqrt{\pi}}{2\sqrt{\frac{1}{2}-\theta}} \quad = \frac{1}{\sqrt{2}\sqrt{\frac{1}{2}-\theta}} = \quad \frac{1}{\sqrt{1-2\theta}}$$

Entonces, la función generatriz de momentos de Z^2 en el parámetro θ es

$$M_{Z^2}(\theta) = \frac{1}{\sqrt{1-2\theta}} \qquad (2.51)$$

Por lo tanto, la función generatriz de momentos de la variable Y, dada en (2.49), es

$$M_Y(\theta) = [M_{Z^2}(\theta)]^n = \sqrt{\left(\frac{1}{1-2\theta}\right)^n} \qquad (2.52)$$

(2.52) es la función generatriz de momentos de una variable aleatoria Y con distribución chi cuadrado de n grados de libertad.

La media y la desviación típica de la distribución χ^2 se obtienen derivando la función generatriz de momentos, hallada en (2.52):

$$M_Y(\theta) = \left(\frac{1}{1-2\theta}\right)^{\frac{n}{2}}$$

$$m_1' = \frac{\partial M_Y(\theta)}{\partial \theta} = \frac{n}{2}\left(\frac{1}{1-2\theta}\right)^{\frac{n}{2}-1}\left(-\frac{-2}{(1-2\theta)^2}\right) =$$

$$= n\left(\frac{1}{1-2\theta}\right)^{\frac{n-2}{2}}\left(\frac{1}{(1-2\theta)^2}\right) =$$

$$= \left(\frac{n}{(1-2\theta)^2}\right)\left(\frac{1}{1-2\theta}\right)^{\frac{n-2}{2}} =$$

Haciendo $\theta = 0$, se obtiene el momento natural de orden 1: $m_1' = n$

Para obtener el momento natural de orden 2,

$$m_2' = \frac{\partial^2 M_Y(\theta)}{\partial \theta^2} = \frac{\partial^2}{\partial \theta^2}\left[\left(\frac{n}{(1-2\theta)^2}\right)\left(\frac{1}{1-2\theta}\right)^{\frac{n-2}{2}}\right]$$

$$m_2' = \left[-\frac{n\,2\,(1-2\theta)(-2)}{(1-2\theta)^4}\right]\left(\frac{1}{1-2\theta}\right)^{\frac{n-2}{2}} + \frac{\left[\frac{(n-2)}{2}\left(\frac{1}{1-2\theta}\right)^{\frac{n-2}{2}-1}(-(-2))\right]}{(1-2\theta)^2}$$

Haciendo $\theta = 0$, se obtiene el momento natural de orden 2:

$$m_2' = n^2 + 2n$$

Es decir, la esperanza matemática de una variable aleatoria con distribución Chi cuadrado es igual a la cantidad de unidades de observación que dieron origen al recorrido de la variable poblacional;

$$E(Y) = m_1' = n$$

Mientras que la varianza se obtiene de hacer:

$$V(Y) = m_2' - (m_1')^2 = n^2 + 2n - (n)^2 = 2n$$

que es igual a dos veces la esperanza matemática.

Por otra parte, comparando (2.52) con (2.47)

$$M_Y(\theta) = \left(\frac{1}{1-2\theta}\right)^{\frac{n}{2}}; \qquad Y \sim \chi_n^2$$

$$M_X(\theta) = \left(1 - \frac{\theta}{\alpha}\right)^{-\lambda}; \qquad X \sim \Gamma(\alpha, \lambda)$$

Se observa que, la función generatriz de momentos de esta distribución tiene una estructura similar a una distribución Gamma

con parámetros $\alpha = \frac{1}{2}$ y $\lambda = \frac{n}{2}$; esto es, haciendo el reemplazo de esos valores en (2.47), se obtiene:

$$M_X(\theta) = \left(1 - \frac{\theta}{1/2}\right)^{-\frac{n}{2}} = (1-2\theta)^{-\frac{n}{2}} = \left(\frac{1}{1-2\theta}\right)^{\frac{n}{2}} = M_Y(\theta) \qquad (2.53)$$

En consecuencia, teniendo en cuenta (2.46), la función de densidad de una variable aleatoria X con distribución Gamma era

$$f(X) = \frac{\alpha^{\lambda}}{\Gamma(\lambda)} e^{-\alpha X} X^{\lambda-1}$$

Ahora si los parámetros asumen los valores $\alpha = \frac{1}{2}$ y $\lambda = \frac{n}{2}$, se obtiene

$$f(X) = \frac{\left(\frac{1}{2}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} e^{-\frac{1}{2}X} X^{\frac{n}{2}-1} = \frac{e^{-\frac{X}{2}} X^{\frac{n}{2}-1}}{\sqrt{2^n}\,\Gamma\left(\frac{n}{2}\right)}$$

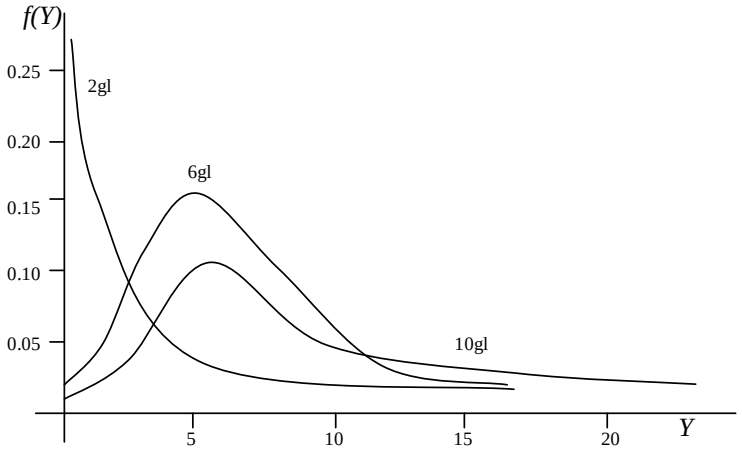


Figura 2.3: Distribución Chi Cuadrado

En vista de lo anterior, la función de densidad de la variable aleatoria Y con distribución χ^2_n es:

$$f(Y) = \frac{e^{-\left(\frac{Y}{2}\right)}{}_Y\left(\frac{n}{2}-1\right)}{\sqrt{2^n}\Gamma\left(\frac{n}{2}\right)} \qquad (2.54)$$

La Figura 2.3 presenta funciones de distribución de Chi cuadrado con diferentes grados de libertad. Cuando los grados de libertad tienden a ∞, $\chi^2_{n\to\infty} \to Z \sim N(0,1)$

Propiedad reproductiva

Dadas dos variables independientes, Y_1 e Y_2, donde ambas se distribuyen χ^2 con n_1 y n_2 grados de libertad, respectivamente; la suma de ellas es otra variable χ^2 con $n_1 + n_2$ grados de libertad.

En efecto,

$$M_{Y_1+Y_2}(\theta) = M_{Y_1}(\theta)M_{Y_2}(\theta) = \sqrt{\left(\frac{1}{1-2\theta}\right)^{n_1}}\sqrt{\left(\frac{1}{1-2\theta}\right)^{n_2}}$$
$$= \sqrt{\left(\frac{1}{1-2\theta}\right)^{n_1}\left(\frac{1}{1-2\theta}\right)^{n_2}}$$

que según (2.52) es la función generatriz de momentos de $\chi^2_{n_1+n_2}$

Ejemplo 2.1. La variable X tiene distribución Chi-cuadrado con 20 grados de libertad:

$$X \sim \chi^2_{20}$$

¿Cuál es la probabilidad de que la variable X asuma valores inferiores o iguales a 34.17?

Para hallar esta probabilidad es necesario utilizar una tabla para valores críticos de la Chi cuadrado. La tabla tiene los grados de libertad a la izquierda y la probabilidad acumulada desde el valor crítico a infinito en el extremo superior.

La variable tiene 20 grados de libertad, por ende es necesario ubicar la fila correspondiente a 20 grados. Luego debe localizarse sobre la fila el

punto 34.17. Una vez hallado, se debe leer arriba (en áreas de extremo superior) la probabilidad, en este caso 0.025.

$$P(X \geq 34.17) = 0.025$$

Otra forma de notación es $P(\chi^2_{20} \geq 34.17) = 0.025$

La probabilidad de los menores a 34.17, que es lo que se busca, se halla de la siguiente manera:

$$P(X \geq 34.17) = 0.025 = 1 - P(X \geq 34.17) = 1 - 0.025 = 0.975$$

En la actualidad, es posible conocer la probabilidad con software que incluyan paquetes estadísticos, tal el caso de Microsoft Excel 2010. En el menú principal de Excel se encuentra la opción *Fórmulas*, *Insertar función*; en la categoría *Estadísticas* se encuentran la totalidad de funciones disponibles .

La distribución chi cuadrado cuenta con cuatro funciones en el entorno de Excel 2010, cada una de ellas brinda información sobre la probabilidad o el valor crítico asociado para una distribución de dos colas o de cola derecha.

Para conocer la probabilidad correspondiente a un valor, se utiliza la función =DISTRI.CHICUADRADO(x;grados_de_libertad;acumulado; donde x es el valor para el que se quiere encontrar la probabilidad ($x = 34{,}17$), los grados de libertad 20, acumulado debe indicarse con *Verdadero* o *Falso*. Cuando acumulado es verdadero informa la probabilidad de la distribución acumulada hasta el punto $x = 34{,}17$; cuando es falso proporciona el valor de la función de densidad.

Con la función =DISTRI.CHICUADRAD.CD(x;grados de probabilidad), se conoce la probabilidad existente en la cola derecha de la distribución; debe interpretarse la consigna grados de probabilidad como grados de libertad.

Cuando la incógnita sea el valor que asegura un determinado nivel de probabilidad, debe trabajarse con las funciones =INV.CHICUAD(probabilidad;grados de libertad) o =INV.CHICUAD.CD(probabilidad;grados de libertad).

Distribución F de Snedecor o Fischer

Sea X una variable chi-cuadrado con m grados de libertad e Y otra variable chi-cuadrado con n grados de libertad, suponiendo que X e Y son independientes, entonces la variable aleatoria

$$F = \frac{X/m}{Y/n} = \frac{\chi_m^2}{\chi_n^2} \sim F_{m,n} \qquad (2.55)$$

se distribuye F de Snedecor con m y n grados de libertad en el numerador y denominador respectivamente.

De acuerdo al resultado alcanzado en (2.54), la función de densidad de X es

$$g(X) = \frac{e^{-\left(\frac{X}{2}\right)} X^{\left(\frac{m}{2}-1\right)}}{\sqrt{2^m}\,\Gamma\left(\frac{m}{2}\right)}$$

y la función de densidad de Y es

$$g(Y) = \frac{e^{-\left(\frac{Y}{2}\right)} Y^{\left(\frac{n}{2}-1\right)}}{\sqrt{2^n}\,\Gamma\left(\frac{n}{2}\right)}$$

Por ende, la función de densidad conjunta de las variables X e Y, $g(X,Y)$, es:

$$g(X,Y) = \frac{e^{-\left(\frac{X}{2}\right)} X^{\left(\frac{m}{2}-1\right)}}{\sqrt{2^m}\,\Gamma\left(\frac{m}{2}\right)} \frac{e^{-\left(\frac{Y}{2}\right)} Y^{\left(\frac{n}{2}-1\right)}}{\sqrt{2^n}\,\Gamma\left(\frac{n}{2}\right)}$$

Donde $g(X) \sim \chi_m^2$ y $g(Y) \sim \chi_n^2$

Agrupando términos con igual base se obtiene que

$$g(X,Y)=\frac{X^{\left(\frac{m}{2}-1\right)}Y^{\left(\frac{n}{2}-1\right)}e^{-\left(\frac{X+Y}{2}\right)}}{2^{\frac{m+n}{2}}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}$$

de donde la función de densidad conjunta es

$$\iint g(X,Y)dXdY=\iint\frac{X^{\left(\frac{m}{2}-1\right)}Y^{\left(\frac{n}{2}-1\right)}e^{-\left(\frac{X+Y}{2}\right)}}{2^{\frac{m+n}{2}}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}dXdY \qquad (2.56)$$

A partir de (2.55), se reexpresa X en función de F e Y

$$X=\frac{mYF}{n};dX=\frac{mY}{n}dF\Rightarrow\frac{dX}{dF}=\frac{mY}{n}$$

reemplazando en (2.56), los valores de X por sus iguales

$$\iint g(F,Y)dFdY=\iint\frac{m^{\left(\frac{m}{2}-1\right)}Y^{\left(\frac{m}{2}-1\right)}F^{\left(\frac{m}{2}-1\right)}Y^{\left(\frac{n}{2}-1\right)}e^{-\left(\frac{mYF}{2n}\right)-\left(\frac{Y}{2}\right)mY}}{2^{\frac{m+n}{2}}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)n^{\left(\frac{m}{2}-1\right)}n}dFdY$$

$$=\iint\frac{m^{\left(\frac{m}{2}-1\right)}Y^{\left(\frac{m}{2}-1\right)}F^{\left(\frac{m}{2}-1\right)}Y^{\left(\frac{n}{2}-1\right)}e^{-\left(Y\frac{n+mF}{2n}\right)mY}}{2^{\frac{m+n}{2}}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)n^{\left(\frac{m}{2}-1\right)}n}dFdY \qquad (2.57)$$

al integrar $g(F,Y)dFdY$ respecto de Y se obtiene la función de densidad de probabilidad marginal de F, esto es

$$\int_0^\infty g(F,Y)dY=h(F)dF$$

Para resolver la integral, se deben juntar potencias de igual base en el numerador y el denominador de (2.57).

1) $m^{\left(\frac{m}{2}-1\right)}m=m^{\left(\frac{m}{2}-1+1\right)}=m^{\left(\frac{m}{2}\right)}$
2) $n^{\left(\frac{m}{2}-1\right)}n=n^{\left(\frac{m}{2}\right)}$
3) $Y^{\left(\frac{m}{2}-1\right)}Y^{\left(\frac{n}{2}-1\right)}Y=Y^{\left(\frac{m+n}{2}-1\right)}$

De esta forma,

$$h(F)dF = \int_0^\infty g(F,Y)dY = \int_0^\infty \frac{m^{\left(\frac{m}{2}\right)}F^{\frac{m}{2}-1}}{n^{\left(\frac{m}{2}\right)}2^{\frac{m+n}{2}}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} e^{-\left(Y\frac{n+mF}{2n}\right)}Y^{\left(\frac{m+n}{2}-1\right)}dY \quad (2.58)$$

De donde:

$$h(F)dF = \frac{m^{\left(\frac{m}{2}\right)}F^{\frac{m}{2}-1}}{n^{\left(\frac{m}{2}\right)}2^{\frac{m+n}{2}}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \int_0^\infty e^{-\left(Y\frac{n+mF}{2n}\right)}Y^{\left(\frac{m+n}{2}-1\right)}\,dY$$

que es análoga a la integral de la Distribución Gamma hallada en (2.44)

$$\int_0^\infty e^{-\alpha X}X^{\lambda-1}dX = \frac{\Gamma(\lambda)}{\alpha^\lambda}$$

donde: $\alpha = \frac{n+mF}{2n}$ y $\lambda = \frac{m+n}{2}$

Entonces:

$$h(F) = \frac{m^{\left(\frac{m}{2}\right)}F^{\frac{m}{2}-1}}{n^{\left(\frac{m}{2}\right)}2^{\frac{m+n}{2}}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\left(\frac{n+mF}{2n}\right)^{\frac{m+n}{2}}}$$

$$= \frac{m^{\left(\frac{m}{2}\right)}F^{\frac{m}{2}-1}}{n^{\left(\frac{m}{2}\right)}2^{\frac{m+n}{2}}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \frac{\Gamma\left(\frac{m+n}{2}\right)2^{\frac{m+n}{2}}n^{\frac{m+n}{2}}}{(n+mF)^{\frac{m+n}{2}}} =$$

$$= \frac{\Gamma\left(\frac{m+n}{2}\right)m^{\left(\frac{m}{2}\right)}F^{\left(\frac{m-2}{2}\right)}n^{\left(\frac{m+n}{2}-\frac{m}{2}\right)}(n+mF)^{-\left(\frac{m+n}{2}\right)}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}$$

Por lo tanto,

$$h(F) = \frac{\Gamma\left(\frac{m+n}{2}\right) m^{\left(\frac{m}{2}\right)}{}_F{}^{\left(\frac{m-2}{2}\right)} n^{\left(\frac{n}{2}\right)}{}_{(n+mF)}{}^{-\left(\frac{m+n}{2}\right)}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \qquad (2.59)$$

es la función de densidad de la distribución F de Snedecor

Las características de esta distribución son

$$E(F) = \frac{n}{n-2}; \quad n > 2$$

$$V(F) = \frac{n^2(2n+2m-4)}{m(n-2)^2(n-4)}; \quad n > 4$$

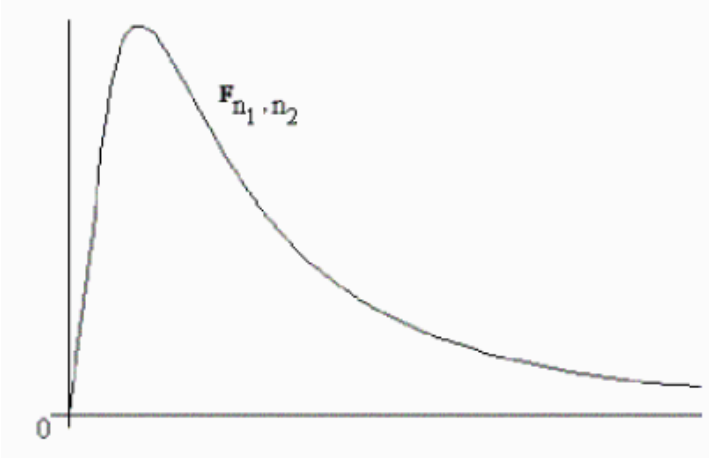


Figura 2.4 Distribución $\boldsymbol{F}$

Ejemplo 2.2. ¿Cuál es la probabilidad de una variable F, que se distribuye F de Snedecor con 4 grados de libertad en el numerador y 6 grados de libertad en el denominador, para valores igual o mayores a 9,15?

Para hallar esta probabilidad es necesario utilizar la tabla para valores críticos de la distribución F. La tabla tiene los grados de libertad del denominador a la izquierda y los grados de libertad en el numerador arriba. La tabla en el extremo superior muestra la probabilidad acumulada desde el valor crítico a infinito.

La variable tiene 4 grados de libertad en el numerador y 6 grados de libertad en el denominador, por ende es necesario ubicar la celda en la cual, la columna y la fila correspondientes a estos grados de libertad, se crucen. En la Tabla donde se encuentre el punto crítico, allí se hallará la probabilidad.

La probabilidad asociada a los puntos mayores al valor crítico 9.15 es 0.01. Su notación es:

$$P(F_{4,6} > 9.15) = 0{,}01$$

Las funciones a utilizar en Excel para encontrar la probabilidad son =DISTR.F.CD(x,grados de libertad1;grados de libertad2) y DISTR.F.N(x,grados de libertad1;grados de libertad2;acumulado). Tal como en el caso de chi-cuadrado, acumulado puede ser *Verdadero* o *Falso*; lo primero hace referencia a la distribución de probabilidad acumulativa y lo segundo a la función de densidad.

Para encontrar el valor crítico a partir de un nivel de probabilidad conocido, se trabaja con las funciones =INV.F(probabilidad;grados de libertad1;grados de libertad2) y =INV.F.CD(probabilidad;grados de libertad1;grados de libertad2).

Relaciones entre las distribuciones Normal y χ^2

Teorema: Se tiene $x_1, x_2, \dots, x_n$ datos provenientes de una muestra aleatoria de una variable X con esperanza μ y varianza σ^2. Se define un estimador de la varianza

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1} \qquad (2.60)$$

en donde $\bar{X}$ es el promedio muestral, se tiene que:

a) $E(S^2) = \sigma^2$

b) Si X está distribuida normalmente, $\frac{n-1}{\sigma^2}S^2$ tiene una distribución χ^2 con $(n-1)$ grados de libertad

Para demostrar la parte a) del teorema, a la suma de cuadrados de los desvíos respecto de la media, se le resta y suma μ

$$\sum_{i=1}^{n}(x_i - \bar{X})^2 = \sum_{i=1}^{n}(x_i - \mu + \mu - \bar{X})^2$$

Agrupando convenientemente y desarrollando el cuadrado

$$\sum_{i=1}^{n}[(x_i - \mu)^2 + 2(x_i - \mu)(\mu - \bar{X}) + (\mu - \bar{X})^2]$$

Distribuyendo el sumatorio

$$\sum_{i=1}^{n}(x_i - \mu)^2 + 2(\mu - \bar{X})\sum_{i=1}^{n}(x_i - \mu) + n(\mu - \bar{X})^2 =$$

$$= \sum_{i=1}^{n}(x_i - \mu)^2 + 2(\mu - \bar{X})\left(\sum_{i=1}^{n}x_i - n\mu\right) + n(\mu - \bar{X})^2$$

Multiplicando y dividiendo el segundo término por (-n)

$$= \sum_{i=1}^{n}(x_i - \mu)^2 + 2(\mu - \bar{X})\left(\sum_{i=1}^{n}x_i - n\mu\right)\frac{-n}{-n} + n(\mu - \bar{X})^2$$

$$= \sum_{i=1}^{n}(x_i - \mu)^2 - 2n(\mu - \bar{X})(\mu - \bar{X}) + n(\mu - \bar{X})^2$$

$$= \sum_{i=1}^{n} (x_i - \mu)^2 - 2n(\mu - \bar{X})^2 + n(\mu - \bar{X})^2$$

De modo que,

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

Luego, se toma esperanza

$$E(S^2) = E\left[\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}\right] = E\left[\frac{\sum_{i=1}^{n}(x_i - \mu)^2 - n(\bar{X} - \mu)^2}{n-1}\right]$$

OBSERVACIÓN: la diferencia de $\bar{X} - \mu$ al encontrarse al cuadrado es siempre positiva, por ello y por conveniencia se invierte el orden de los minuendos

Aplicando las propiedades de esperanza matemática

$$= \frac{1}{n-1} E\left[\sum_{i=1}^{n} (x_i - \mu)^2 - n(\bar{X} - \mu)^2\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} E(x_i - \mu)^2 - nE(\bar{X} - \mu)^2\right]$$

De modo que

$$E(S^2) = \frac{n(\sigma^2 - \sigma_{\bar{X}}^2)}{n-1} \qquad (2.61)$$

Entonces,

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

En síntesis

$$E(S^2) = \sigma^2 \tag{2.62}$$

con lo que se demuestra la primera parte del Teorema.

El estimador

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n} \tag{2.63}$$

es un *estimador sesgado* de la varianza poblacional; siendo el sesgo igual a $\frac{n-1}{n}$. Para demostrarlo se parte consignando

$$E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$$

de donde

$$\frac{n}{n-1}E(\hat{\sigma}^2) = \sigma^2$$

Esto puede escribirse reemplazando el estimador por su igual

$$\frac{n}{n-1}E\left(\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}\right) = \sigma^2$$

esto es,

$$E\left(\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}\right) = \sigma^2 \tag{2.64}$$

Pero $\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1} = S^2$; es decir,

$$E(S^2) = \sigma^2 \tag{2.65}$$

con lo que queda demostrado que el sesgo es $\frac{n-1}{n}$. De este modo S^2 es un *estimador insesgado* de la varianza poblacional y el *factor de corrección* que permite evitar el sesgo es $\frac{n-1}{n}$.

Se demostrará ahora la parte b) del teorema. El enunciado dice que si

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

Entonces

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \qquad (2.66)$$

Sea $\bar{X}$ un estimador insesgado de μ, se define

$$Z = \frac{\bar{X} - \mu}{\sigma} \sim N(0,1)$$

Y,

$$\sum_{i=1}^{n} Z_i^2 \sim \chi^2_{n-1}$$

Por lo tanto,

$$\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}$$

Si se divide numerador y denominador por $n-1$

$$\frac{(n-1)\frac{\sum_{i=1}^{n}(x_i-\bar{X})^2}{n-1}}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \qquad (2.67)$$

En general, una suma de variables aleatorias independientes con distribución normal, elevadas al cuadrado, se distribuye χ^2 con grados de libertad determinados por la cantidad de observaciones menos la cantidad de estimadores de parámetros, en este caso 1 $(\bar{X})$.

Ejemplo 2.3. Considerando el caso especial de $n = 2$

$$\sum_{i=1}^{2}(x_i - \bar{X})^2 = (x_1 - \bar{X})^2 + (x_2 - \bar{X})^2$$

$$= \left(x_1 - \frac{x_1 + x_2}{2}\right)^2 + \left(x_2 - \frac{x_1 + x_2}{2}\right)^2$$

$$= \frac{(2x_1 - x_1 - x_2)^2}{4} + \frac{(2x_2 - x_1 - x_2)^2}{4}$$

$$= \frac{(x_1 - x_2)^2 + (x_1 - x_2)^2}{4} = \frac{(x_1 - x_2)^2}{2}$$

Puesto que x_1 y x_2 están independientemente distribuidos $N(\mu, \sigma^2)$ se encuentra que $(x_1 - x_2)$ tiene una distribución $N(0, 2\sigma^2)$, luego

$$\frac{\sum_{i=1}^{2}(x_i - \bar{X})^2}{\sigma^2} = \left[\frac{(x_1 - x_2) - 0}{\sqrt{2}\sigma}\right]^2 = \frac{1}{\sigma^2}S^2$$

tiene la distribución χ^2 con 1 grado de libertad.

Distribución *t* de Student

Sea Z una variable aleatoria normal estándar

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

Y χ^2 una variable aleatoria chi cuadrado con $n - 1$ grados de libertad

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

Se define

$$t = \frac{Z}{\sqrt{\chi^2_{n-1}/_{n-1}}} \qquad (2.68)$$

(2.68) sigue una distribución t de Student con $n-1$ grados de libertad.

Reemplazando en (2.68), las variables Z y χ^2

$$t = \frac{\frac{X-\mu}{\sigma}}{\sqrt{\frac{(n-1)}{\sigma^2}S^2 \Big/ n-1}}$$

Se dice que el estadístico t se distribuye como una t de Student con $n-1$ grados de libertad: $t \sim t_{n-1}$.

La distribución t se caracteriza por

$$E(t) = 0;\ \ n > 0$$

$$V(t) = \frac{n}{n-2};\ \ n > 2$$

La distribución t es simétrica con respecto a 0 y asintóticamente tiende a una distribución normal tipificada.

Figura 2.5 Distribución $\boldsymbol{t}$

Relación entre las distribuciones t y F

Si se eleva al cuadrado la expresión de t, el resultado se puede escribir como

$$t^2 = \frac{Z^2/1}{\chi^2_{n-1}/n-1}$$

en donde Z^2, al ser el cuadrado de una variable normal tipificada, tiene una distribución χ^2_1.

Así pues, $t^2 = F(1, n-1)$; es decir, el cuadrado de una variable t con $n-1$ grados de libertad es una F con $(1, n-1)$ grados de libertad.

Ejemplo 2.4. Dada una variable X, que se distribuye t de Student con 10 grados de libertad, ¿cuál es la probabilidad de asumir valores iguales o superiores a 1.3722?

Para hallar esta probabilidad es necesario utilizar una tabla para valores críticos de la distribución t. La tabla tiene los grados de libertad a la izquierda y la probabilidad acumulada desde el punto crítico a infinito arriba en el extremo superior.

La variable tiene 10 grados de libertad, debe ubicarse la fila correspondiente a estos grados de libertad y en ellos ubicar el punto crítico 1.3722. Luego leer en la fila correspondiente a Áreas de extremo superior la probabilidad desde 1.3722 hasta infinito.

La probabilidad de que la variable asuma valores inferiores a 1.3722 es:

$$P(t_{10} \leq 1.3722) = 1 - P(t_{10} \geq 1.3722) = 1 - 0.10 = 0.90$$

En Excel se dispone de las funciones =DIST.T.2C(x;grados de libertad), =DISTR.T.CD(x;grados de libertad), =DISTR.T.N(x;grados de libertad;acumulado), =INV.T(probabilidad;grados de libertad), =INV.T.2C(probabilidad;grados de libertad)

---

3. ESTIMACIÓN Y CONTRASTE DE PARÁMETROS

Siguiendo con el proceso de investigación econométrica, se estudiará la forma en que los parámetros de una variable en la población se estiman a partir de los estadísticos de dicha variable en la muestra. Se verá que la estimación puede ser puntual o por intervalo, pero para ello se necesita establecer el tamaño de la muestra, el número de unidades de observación a ser obtenida del total de elementos que conforman la población. En la primera parte de este cuaderno se estudió las características de la media muestral en tanto estadístico útil para estimar la media de la población. En ésta se estudian las propiedades que deben cumplir los estadísticos para ser considerados "buenos" estimadores de los parámetros de una variable en la población. Por último, se verá cómo evaluar, a través de contrastes de hipótesis, los valores estimados de los parámetros respecto de sus valores "naturales" o establecidos por el investigador de acuerdo a las hipótesis iniciales de su problema a investigar.

3.1. Tamaño de la muestra

Una pregunta práctica en gran parte de la investigación econométrica tiene que ver con la determinación del tamaño de la muestra. La decisión del *tamaño de la muestra* está directamente relacionada con el costo de la investigación y por tanto debe ser justificada.

Las características poblacionales de interés son los parámetros de la población, es decir la *media*, la *varianza* y la *desviación estándar* de las variables a estudiar que conforman las características a analizar sobre las unidades de observación en la tabla de datos.

Normalmente, estas medidas son desconocidas y se deberá estimar su valor, lo más aproximadamente posible, tomando una muestra de unidades de observación a partir de los elementos de la población.

Del mismo modo que la variable en la *población* tiene un conjunto de características, la variable en cada *muestra* también tiene un conjunto de características. Una de ellas es el promedio o media muestral $\bar{X}$ que, por ser una característica muestral, cambiará si se obtuviera una nueva muestra y se utiliza para estimar la media de la variable en la población, μ.

Otra característica es la varianza de la variable en la muestra, S^2. Esta será pequeña si las respuestas de la muestra (datos) son similares y grande si se encuentran dispersas. S^2 es la varianza de la variable en la muestra que se usa para estimar la varianza de la variable en la población σ^2.

Al comienzo de este cuaderno, se demostró que la media de la variable en la muestra, $\bar{X}$, tiene una distribución de probabilidad normal (con media igual a μ y varianza igual a $\sigma_{\bar{x}}^2$), cualquiera sea la distribución de probabilidad de la variable en la población; aunque si esta no es normal, se requiere que la muestra sea lo suficientemente grande de acuerdo al Teorema del Límite Central.

Por consiguiente, $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

La varianza en $\bar{X}$ será más grande a medida que la varianza de la variable en la población, σ^2, sea más grande. A medida que aumente el tamaño de la muestra, la variación en $\bar{X}$ disminuirá.

De acuerdo a esto, existe un nivel de confianza de 0,95 (95%) de que $\bar{X}$ caiga dentro de $\pm$1,96 desvíos estándar de la media de la variable en la población μ. La cantidad 1,96 es el valor de la distribución normal estándar para un 95% de probabilidad.

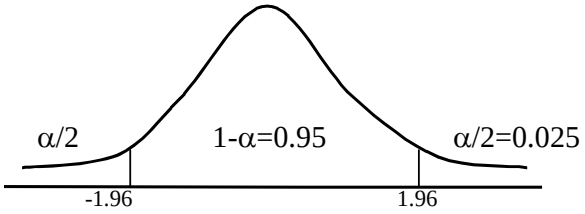


Figura 3.1

Ejemplo 3.1. En el estudio de la variable número de visitas al dentista en un año determinado por individuos de una población (Ejemplo 1.2.), si se han obtenido 100 muestras diferentes de 10 individuos, aproximadamente el 95% de las medias muestrales resultantes $(\bar{X})$ estarán dentro de $\pm 1{,}96$ desvíos estándar $(\sigma_{\bar{X}} = 1{,}0416)$ de la media de la población $(\mu = 4{,}5)$. Es decir, se cometerá un error de estimación o de muestreo igual a: $\bar{X} \pm 1{,}96\sigma_{\bar{X}}$ error que se hará cada vez más pequeño al tomar mayor cantidad de muestras de la población.

Ahora bien, como se ha expresado anteriormente, el investigador tomará solo una muestra del total de muestras disponibles en la población. Al realizar esta acción estará cometiendo un error de estimación, denominado *error de muestreo* (e). Es decir, la media aritmética $\bar{X}$ obtenida en la muestra diferirá de la verdadera media en la población μ por un error,

$$\bar{X} - \mu = e$$

Entonces bajo el concepto del Teorema del Límite Central, el error de muestreo se define como el producto del desvío de $\bar{X}$ por el valor Z resultante del nivel de confianza con que el investigador desea obtener la estimación, esto es, si $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$, entonces

$$\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{e}{\sigma_{\bar{X}}} = Z \sim N(0,1)$$

Por lo tanto,

$$e = Z\sigma_{\bar{X}} = Z\frac{\sigma}{\sqrt{n}} \tag{3.1}$$

De donde se deduce que la fórmula para calcular el *tamaño de la muestra* es

$$n = \frac{Z^2\sigma^2}{e^2} \tag{3.2}$$

Ejemplo 3.2. Si existe la necesidad de estar un 95% seguro de que al estimar la media de la variable número de visitas al dentista en un año determinado por individuos de una población, el *error de muestreo* no exceda de 0,5 personas conociendo que la desviación estándar de la variable en la población es 2,87; el error de muestreo (e) es igual a 0,5 y puesto que el nivel de confianza es del 95%, Z asume el valor 1,96. Por tanto, el tamaño de la muestra debe ser:

$$n = \frac{(1{,}96)^2\,(2{,}87)^2}{(0{,}5)^2} \cong 127$$

Por consiguiente, el tamaño de la muestra, en *poblaciones infinitas* –poblaciones que no se puedan contar-, es independiente del tamaño de la población y dependiente de:

Nivel de Confianza	$> Z \Rightarrow > n$
Desviación estándar de la población	$> \sigma \Rightarrow > n$
Error de muestreo	$> e \Rightarrow < n$

Error, riesgo y tamaño de la muestra

Si $\hat{\theta}$ es un estimador insesgasdo del parámetro θ, $E(\hat{\theta}) = \theta$, y $\hat{\theta} \sim N(\theta, \sigma_{\hat{\theta}})$, entonces

El error de estimación es la diferencia entre el valor obtenido en la estimación puntual $(\hat{\theta}_0)$, y el valor del parámetro, θ

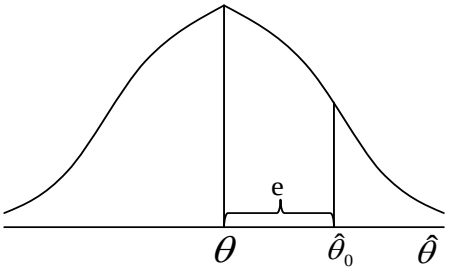


Figura 3.2

$$e = \hat{\theta}_0 - \theta$$

Si el error máximo aceptable por el investigador es e, habrá un riesgo de cometer un error superior a él. Es decir, un riesgo de que $|\hat{\theta}_0 - \theta| > e$. El riesgo es la probabilidad de que la diferencia entre el parámetro y su estimador supere la cuantía de e.

$$P(|\hat{\theta}_0 - \theta| > e) = \textit{ riesgo de cometer un error superior a } e \qquad (3.3)$$

Ejemplo 3.3. Suponga que $X \sim N(500, 30)$, $n = 100$, $|e| = 6$. Se quiere saber cuál es el riesgo de cometer un error en la estimación de μ.

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

por lo tanto

$$Z = \frac{6\sqrt{100}}{30} = \pm 2$$

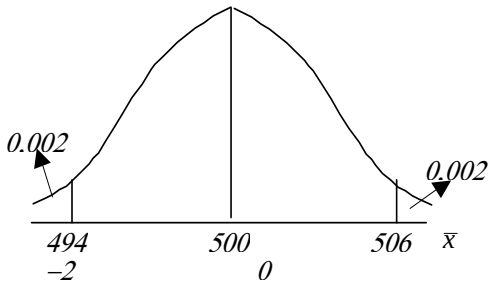


Figura 3.3

La probabilidad (o sea, el riesgo) de que al efectuar la estimación puntual se cometa un error en valor absoluto mayor a 6 es

$$P(|Z| > 2) = 1 - P(-2 < Z < 2) = 1 - (0{,}9772 - 0{,}0228) = 0{,}0456$$
$$= riesgo$$

Se pueden establecer las siguientes relaciones entre riesgo, error y tamaño de la muestra:

a) Riesgo

$Z = \frac{e\sqrt{n}}{\sigma}$

- el riesgo es función inversa de Z. Cuando Z crece el riesgo disminuye.
- Z depende del error, de n y de la varianza poblacional, por lo tanto el riesgo, que es función de Z, también depende de ellos.
- si aumenta el error e, aumenta Z, disminuye el riesgo
- si aumenta n, aumenta Z, disminuye el riesgo
- si aumenta σ^2, disminuye Z, aumenta el riesgo

b) Error

$e = Z\frac{\sigma}{\sqrt{n}}$

- si aumenta n, disminuye el error
- si aumenta σ^2, aumenta el error

c) Tamaño muestral

$n = \frac{Z^2\sigma^2}{e^2}$

- manteniendo el mismo error y el mismo riesgo, si aumenta σ^2 aumenta n. Se puede calcular n para estimar μ si se conocen error, riesgo y σ^2 poblacional.

Desviación estándar desconocida

El procedimiento que se acaba de mostrar *supone* que la desviación estándar de la población es conocida. Pero, en la mayor parte de las situaciones prácticas no es conocida y debe ser estimada a partir de

- Usar una desviación estándar de la muestra obtenida de una encuesta anterior comparable o de una encuesta piloto
- Estimar σ subjetivamente
- Usar proporciones

Los dos primeros casos se resuelven de la manera desarrollada, reemplazando el valor de σ en las fórmulas anteriores. Para el tercer caso, el procedimiento es usar una proporción de la muestra (p) para *estimar* una proporción desconocida de la población, (π). Recuerde que p es un estimador insesgado de π como lo es $\bar{X}$ de μ, por lo tanto se puede trabajar con el error de muestreo usando proporciones en lugar del error de muestreo definido en (3.1).

El *error de muestreo usando proporciones* es,

$$e = Z\sigma_p \qquad (3.4)$$

donde σ_p es el desvío de la proporción en la muestra.

OBSERVACIÓN: En una tabla de datos una variable cualitativa expresa si un elemento de la población o una unidad de la muestra poseen o no un determinado atributo. Para poder trabajar se codifica, asignándole el valor 1 o 0 a cada unidad de observación, según posea o no el atributo.

Sea n el número de unidades de observación extraídas de una población de tamaño N, en la que se observa la variable $X_j =$ sexo. Esta es una variable dicotómica que asume el valor 1 si el individuo observado es masculino y el valor 0 si es femenino, siendo masculino y femenino los atributos. El recorrido de la variable X será: $R_X = \{x_i / x_i = 0 \wedge x_i = 1\}$.

Se supone que la tabla de datos para esta situación es la ilustrada en la Figura 3.4.

	$X =$ sexo	
	M	F
1	1	0
2	1	0
3	0	1
4	0	1
⋮		
i	0	1
⋮		
N	1	0

Figura 3.4. Tabla lógica para la variable sexo

Según la información de la tabla planteada se puede tomar los siguientes valores para la variable:

$x_i = 1$ si el i-ésimo elemento posee el atributo "femenino"

$x_i = 0$ si el i-ésimo elemento posee el atributo "masculino"

En consecuencia, el valor de la variable se convierte en un conjunto de ceros y unos. Se define el siguiente total:

$$f = \sum_{i=1}^{n} x_i = 0 + 0 + 1 + 1 + \cdots + 0$$

De esta manera f es el total de individuos de sexo femenino en la muestra. La proporción de mujeres en la muestra es entonces:

$$p = \frac{f}{n}$$

Donde p es la proporción de unidades de observación que poseen un atributo (en este caso ser de sexo femenino) en la

muestra. Por lo tanto, se trata de un estadístico que tiene la función de *estimar* la verdadera proporción de elementos que posee un determinado atributo (en este caso ser de sexo femenino) en la población o sea un parámetro poblacional, definido como:

$$\pi = \frac{F}{N}$$

Donde F es el total de individuos de sexo femenino en la población.

Las proporciones π y p no son más que *promedios* de una variable dicotómica. A veces se multiplican por 100, convirtiéndose en porcentajes.

La varianza del estimador p se puede obtener a partir de la fórmula de la varianza poblacional de la variable dicotómica X, de la siguiente manera:

$$\sigma_X^2 = \left[\frac{\sum_{i=1}^{N}(x_i - \mu_X)^2}{N-1}\right] = \left[\frac{\sum_{i=1}^{N} x_i^2 - \mathrm{N}\mu_X^2}{N-1}\right]$$

Por motivos de notación, los resultados se presentan en términos de una expresión ligeramente diferente, en donde el divisor $(N-1)$ se usa en lugar de N. Esta convención, de acuerdo a lo expresado por W. Cochran (1998) *"la usan quienes enfocan la teoría del muestreo por medio del análisis de la varianza. Su ventaja es que la mayoría de los resultados toman una forma ligeramente más simple. Siempre y cuando se use consistentemente la misma notación, todos los resultados son equivalentes en ambas notaciones"* (p.47).

Pero, en este caso

$$\left[\frac{\sum_{i=1}^{N} x_i^2}{N-1}\right] = \frac{0^2 + 0^2 + 1^2 + 1^2 + \cdots + 0^2}{N-1} = \frac{F}{N-1}$$

Y también,

$$\mu_X = \pi$$

Remplazando,

$$\sigma_X^2 = \left[\frac{\sum_{i=1}^{N} x_i^2 - N\mu_X^2}{N-1}\right] = \left[\frac{F}{N-1} - \frac{N\pi^2}{N-1}\right] =$$

Pero como $F = N\pi$, entonces $\sigma_X^2 = \frac{1}{N-1}[N\pi - N\pi^2]$

De donde,

$$\sigma_X^2 = \frac{N}{N-1}[\pi(1-\pi)]$$

En forma similar, un estimador de la varianza de X en la muestra será:

$$S_X^2 = \frac{n}{n-1}[p(1-p)]$$

De la misma forma que la $V(\bar{X}) = \sigma_{\bar{X}}^2$ era, por el teorema central del límite, igual a la $V(X)/n$; la varianza de la proporción muestral será:

$$V(p) = \sigma_p^2 = \frac{\sigma_X^2\left(\frac{N-n}{N}\right)}{n} = \frac{\frac{N}{N-1}[\pi(1-\pi)]\left(\frac{N-n}{N}\right)}{n} = \frac{[\pi(1-\pi)](N-n)}{n(N-1)}$$
$$= \frac{[\pi(1-\pi)]}{n}\frac{N-n}{N-1}$$

Donde se ha tenido en cuenta el factor de corrección para poblaciones finitas $\left(\frac{N-n}{N}\right)$.

Una estimación insesgada de la $V(p)$, derivada de la muestra, es

$$S_p^2 = \frac{S_X^2\left(\frac{N-n}{N}\right)}{n} = \frac{\frac{n}{n-1}[p(1-p)]\left(\frac{N-n}{N}\right)}{n} = \frac{[p(1-p)]}{n-1}\frac{N-n}{N}$$

De lo anterior se tiene que si N es muy grande en relación a n, de tal manera que el factor de corrección para poblaciones finitas es muy pequeño, una estimación insesgada de la varianza de p es:

$$S_p^2 = \frac{p(1-p)}{n-1}$$

En algunos trabajos se usa la fórmula anterior dividiendo por n en lugar de $(n-1)$; pero esto es un error, ya que, aún en poblaciones infinitas, resulta una estimación sesgada de la verdadera varianza de p.

En (3.4) se estableció que:

$$e = Z\sigma_p$$

Por lo tanto, en vista de la observación anterior

$$e = Z\sqrt{\frac{\pi(1-\pi)}{n}}\sqrt{\frac{N-n}{N-1}}$$

Al resolver para n se obtiene,

$$n = \frac{\frac{z^2\pi(1-\pi)}{e^2}}{1+\frac{1}{N}\left[\frac{z^2\pi(1-\pi)}{e^2}-1\right]} \qquad (3.5)$$

Para fines prácticos, si la corrección para poblaciones finitas es despreciable, se utiliza:

$$n = \frac{Z^2\pi(1-\pi)}{e^2} \tag{3.6}$$

El "peor de los casos", en donde la varianza de la población es máxima, ocurre cuando la proporción poblacional es igual a $0{,}50$ porque, con $\pi = 0{,}50$, la varianza es $\pi(1-\pi) = 0{,}25$ y este es el máximo valor de varianza que se puede tener en una variable dicotómica.

Debido a que la proporción poblacional es desconocida, un procedimiento común consiste en suponer el "peor de los casos". La fórmula para el tamaño de la muestra se simplifica entonces a:

$$n = \frac{Z^2 0{,}25}{e^2} \tag{3.7}$$

Ejemplo 3.4. Si la proporción de la población debe ser estimada dentro de un error, de 0,05 (ó 5 %) a un nivel de confianza de 0,95, el tamaño necesario de la muestra es:

$$n = \frac{(1{,}96)^2 0{,}25}{(0{,}05)^2} \cong 384$$

Suponiendo que la población, de donde se desea extraer la muestra, cuenta con 3000 elementos, será necesario tener en cuenta la corrección para poblaciones finitas (cpf). En este caso, el valor de $n = 384$ es considerado como una primera aproximación y, al aplicar la fórmula general, el tamaño muestral se reduce.

$$n = \frac{\frac{z^2\pi(1-\pi)}{e^2}}{1 + \left\{\frac{1}{N}\left[\frac{z^2\pi(1-\pi)}{e^2} - 1\right]\right\}} = \frac{384}{1 + \{(384-1)/3000\}} \cong 341$$

Un instrumento de encuesta o un experimento no se basa sólo en una pregunta. Algunas veces, pueden estar involucradas cientos de ellas, pero no será necesario pasar por este proceso en todas las preguntas. Generalmente se toman los interrogantes más representativos y se determina el tamaño de la muestra a partir de ellas. En el análisis deben incluirse los cuestionamientos más importantes para el estudio y que tengan la varianza esperada más alta.

p	*Porcentajes*											$(cpf)con\ N = 400$
	5	**10**	**15**	**20**	**25**	**30**	**35**	**40**	**45**	**50**	$\boldsymbol{n}$	n
	4,27	5,88	7,00	7,84	8,49	8,98	9,35	9,60	9,75	9,80	**100**	*80*
	3,02	4,16	4,95	5,54	6,00	6,35	6,61	6,79	6,89	6,93	**200**	*134*
	2,47	3,39	4,04	4,53	4,90	5,19	5,40	5,54	5,63	5,66	**300**	*172*
	2,14	2,94	3,50	3,92	4,24	4,49	4,67	4,80	4,88	4,90	**400**	*200*
	1,91	2,63	3,13	3,51	3,80	4,02	4,18	4,29	4,36	4,38	**500**	*222*
	1,74	2,40	2,86	3,20	3,46	3,67	3,82	3,92	3,98	4,00	**600**	*240*
	1,61	2,22	2,65	2,96	3,21	3,39	3,53	3,63	3,69	3,70	**700**	*255*
	1,51	2,08	2,47	2,77	3,00	3,18	3,31	3,39	3,45	3,46	**800**	*267*
	1,42	1,96	2,33	2,61	2,83	2,99	3,12	3,20	3,25	3,27	**900**	*277*
	1,35	1,86	2,21	2,48	2,68	2,84	2,96	3,04	3,08	3,10	**1000**	*286*
	1,10	1,52	1,81	2,02	2,19	2,32	2,41	2,48	2,52	2,53	**1500**	*316*
	0,96	1,31	1,56	1,75	1,90	2,01	2,09	2,15	2,18	2,19	**2000**	*333*
	0,85	1,18	1,40	1,57	1,70	1,80	1,87	1,92	1,95	1,96	**2500**	*345*
	0,78	1,07	1,28	1,43	1,55	1,64	1,71	1,75	1,78	1,79	**3000**	*353*
	0,72	0,99	1,18	1,33	1,43	1,52	1,58	1,62	1,65	1,66	**3500**	*359*
	0,68	0,93	1,11	1,24	1,34	1,42	1,48	1,52	1,54	1,55	**4000**	*364*
	0,60	0,83	0,99	1,11	1,20	1,27	1,32	1,36	1,38	1,39	**5000**	*370*
	0,43	0,59	0,70	0,78	0,85	0,90	0,93	0,96	0,98	0,98	**10000**	*385*
$(1-p)$	**95**	**90**	**85**	**80**	**75**	**70**	**65**	**60**	**55**	**50**		

Porcentajes

Figura 3.5. Cálculo de la precisión absoluta

Se puede calcular la *precisión absoluta* al 95% sobre la estimación de π en función de la proporción observada p y el tamaño de la muestra n, como se muestra en la Figura 3.5

$$Precisión\ absoluta\ al\ 95\% = \ 1{,}96 \cdot \left[\sqrt{\frac{p(1-p)}{n}}\right]$$

Se observa que la precisión es tanto mejor cuando el tamaño de la muestra crece, pero también, cuando la proporción observada se aleja del 50% (el "peor de los casos").

En efecto, si se representan las variaciones de $p(1-p)$ en función de $(0 \leq p \leq 1)$, se obtiene el gráfico que se muestra en la Figura 3.6

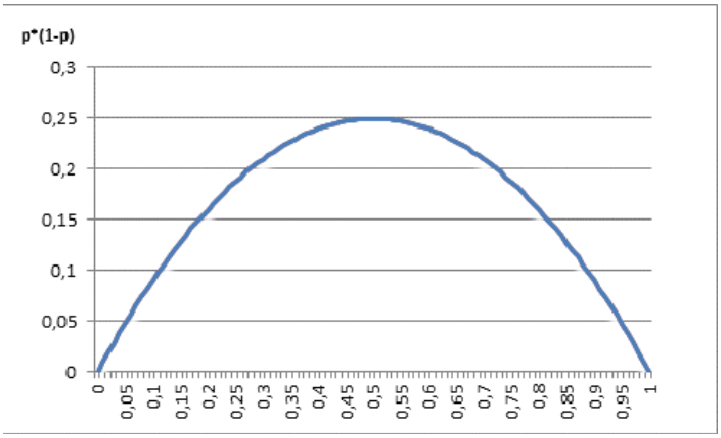


Figura 3.6. Variaciones $p(1-p)$

Entonces para cada p, $p(1-p) \leq \frac{1}{4}$

Por lo tanto, el intervalo de confianza al 95% más grande que se puede obtener para la estimación de una proporción corresponde a la columna del 50% de la Figura 3.7 y es igual a $p \pm \frac{1}{\sqrt{n}}$.

Ejemplo 3.5. Para una muestra de tamaño 400 se tiene: $0{,}50 \pm \frac{1}{\sqrt{400}} = 0{,}50 \pm 0{,}05$, esto es un intervalo de confianza para p igual a $[0{,}45; 0{,}55]$ lo que implica una precisión de aproximadamente 4,90%.

3.2 Estimación puntual

En la discusión que sigue, el tema de interés es la estimación de parámetros (generalmente un solo parámetro) de distribuciones de probabilidad de forma conocida, por lo que la forma de la función de densidad $f(X, \theta)$ también es conocida.

Todas las estimaciones se hacen con base en la información contenida en la muestra. Esta muestra tendrá siempre un tamaño fijo; es decir, no se tiene en cuenta, en el análisis de los métodos de estimación, que el tamaño de la muestra es también una variable aleatoria que depende de los mismos datos de los parámetros a medida que van apareciendo. Esto se puede interpretar como que, a medida que se obtienen las observaciones, el tamaño de la muestra puede ir ajustándose ante esas "apariciones" aleatorias que alteran el valor de las proporciones. Pero esto se abandona ya que el tamaño de la muestra se fija apriorísticamente de acuerdo al procedimiento analizado en el punto anterior; de tal forma que las unidades de observación a seleccionar quedan fijas y la tabla de datos tiene una estructura conocida en cuanto a la cantidad de muestras que serán necesarias para llevar adelante el proceso de investigación.

En la estimación puntual se construye una sola función de los datos de la muestra para estimar el valor de un parámetro poblacional. Dada $f(\bar{X}, \theta)$ conocida, existe una muestra n de tamaño fijo proveniente de una población con distribución conocida.

Los estimadores puntuales tienen las propiedades de: insesgamiento, invarianza, consistencia, eficiencia y suficiencia y son requeridos cuando la misma estimación debe insertarse como un número en una

estructura matemática más compleja, como es común en la economía moderna.

Propiedades de los estimadores

Las propiedades para la media aritmética introducidas al comienzo del cuaderno se pueden generalizar para cualquier estimador; se puede decir que existe un número infinito de estimadores del parámetro desconocido θ, pero solo algunos serán aceptables, por lo que se necesita contar con criterios que permitan elegir entre ellos.

En general, serán de interés aquellos estimadores con una alta probabilidad de proporcionar estimaciones que estén cerca del verdadero valor del parámetro. Como los estimadores son variables aleatorias, los criterios de elección se van a basar en las propiedades de su función de distribución.

En primer lugar, se estudiarán las *propiedades en muestras finitas* que se cumplen para cualquier tamaño de la muestra, incluso para las pequeñas; posteriormente, las *propiedades asintóticas*, que se satisfacen cuando el tamaño muestral tiende a infinito.

Propiedades en muestras finitas

a) Insesgamiento

Se dice que un estimador $\hat{\theta}$ es insesgado, sí $E(\hat{\theta}) = \theta$. Cuando $E(\hat{\theta}) \neq \theta$, el estimador es sesgado y se define el *sesgo* como:

$$sesgo\ (\hat{\theta}) = E(\hat{\theta}) - \theta$$

Esta primera propiedad fue analizada cuando se desarrolló la distribución muestral de la media aritmética de la muestra, donde se demostró empíricamente que

$$E(\bar{X}) \cong \mu$$

siendo μ (parámetro poblacional) la media de la población.

Generalizando, si existe una población descrita por la función de densidad $f(\hat{\theta}, \theta)$, donde f es conocida y el valor del parámetro θ es desconocido. Sea $(\theta_1, \theta_2, \ldots, \theta_n)$ una muestra aleatoria obtenida de esta población, entonces el estadístico $\hat{\theta} = f(\theta_1, \theta_2, \ldots, \theta_n)$ es un *estimador insesgado* del parámetro θ si $E(\hat{\theta}) = \theta$ para todo n y para cualquier valor de θ. En cualquier otro caso $\hat{\theta}$ es sesgado.

Ejemplo 3.6. Estimadores insesgados

$$E(x_1) = \mu; \; E(\bar{X}) = \mu$$

Si μ existe, tanto una sola observación de la variable aleatoria X, x_1, como su media, $\bar{X}$, son estimadores insesgados de μ. Sus varianzas respectivas son σ^2 y σ^2/n. Por lo tanto es preferible $\bar{X}$ a x_1, siempre y cuando $n > 1$, ya que $\sigma^2/n < \sigma^2$.

x_1 es una observación muestral que puede asumir cualquiera de los N valores distintos de la población. De donde

$$E(x_1) = x_1 \underbrace{P(x_1 = x_1)}_{1/N} + x_2 \underbrace{P(x_1 = x_2)}_{1/N} + \cdots + x_N \underbrace{P(x_1 = x_n)}_{1/N} = \frac{1}{N}\sum_{i=1}^{N} x_i = \mu$$

b) $E(\bar{X} - \bar{Y}) = E(X) - E(Y)$

Esto es, la diferencia de las medias de dos muestras aleatorias independientes es un estimador insesgado de la diferencia de las medias en la población.

c) $E(S^2) = \sigma^2$

Dado

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}$$

$$E(\hat{\sigma}^2) = \frac{1}{n} E\left\{\sum_{i=1}^{n}[x_i - E(X) + E(X) - \bar{X}]^2\right\} = \frac{1}{n}\sum_{i=1}^{n}\{E[x_i - E(X)]^2 - E[\bar{X} - E(X)]^2\}$$

---

donde, $E[x_i - E(X)]^2 = \sigma^2 \quad y \; E[\bar{X} - E(X)]^2 = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$

Reemplazando

$$E(\hat{\sigma}^2) = \frac{1}{n}\left\{n\sigma^2 - n\frac{\sigma^2}{n}\right\} = \sigma^2 - \frac{\sigma^2}{n} = \frac{n\sigma^2 - \sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

Por lo tanto, $\hat{\sigma}^2$ es un estimador sesgado de σ^2.

Pero S^2 es insesgado, ya que multiplicando y dividiendo por n

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}\frac{n}{n} = \frac{n}{n-1}\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n} = \frac{n}{n-1}\hat{\sigma}^2$$

Entonces

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

Por lo tanto, $E(S^2) = \sigma^2$

b) Eficiencia

Otro elemento que hay que tener en cuenta es la varianza del estimador $V(\hat{\theta})$; siendo de interés los estimadores que tengan la menor varianza.

Pero seguir solamente este criterio puede llevar a elecciones absurdas.

Ejemplo 3.7. Si θ es un escalar, el estimador $\hat{\theta} = 50$, tiene varianza cero y, sin embargo, no existe ninguna garantía de que esté próximo al verdadero valor del parámetro. Por ello, en muchas ocasiones, la elección entre estimadores se restringe a la clase de estimadores insesgados.

Un estimador $\hat{\theta}$ es *eficiente* dentro de las clases de los estimadores insesgados del parámetro θ si su varianza es la menor entre todas las varianzas de los estimadores insesgados.

Error cuadrático medio

El criterio de eficiencia restringe la elección de un estimador a la clase de estimadores insesgados. Puede ocurrir que de esta forma se ignoren estimadores que son sesgados, pero que tienen una varianza menor que otros estimadores insesgados.

Si se debe elegir entre un estimador $\tilde{\theta}$ insesgado de varianza grande y otro estimador θ^* sesgado pero con varianza pequeña, es posible que θ^* proporcione estimaciones más cercanas al verdadero valor θ que $\tilde{\theta}$.

Para comparar estos estimadores, es preciso utilizar el criterio del *Error Cuadrático Medio* que, en el caso particular de que θ sea un escalar, se define como:

$$ECM(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = E\left[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta\right]^2 = E\left\{\left[\hat{\theta} - E(\hat{\theta})\right] + \left[E(\hat{\theta}) - \theta\right]\right\}^2$$

$$= E\left\{\left[\hat{\theta} - E(\hat{\theta})\right]^2 + 2\left[\hat{\theta} - E(\hat{\theta})\right]\left[E(\hat{\theta}) - \theta\right] + \left[E(\hat{\theta}) - \theta\right]^2\right\} =$$

$$= E\left[\hat{\theta} - E(\hat{\theta})\right]^2 + 2\left[E(\hat{\theta}) - \theta\right]\underbrace{E\left[\hat{\theta} - E(\hat{\theta})\right]}_{0} + \left[E(\hat{\theta}) - \theta\right]^2 =$$

$$E\left[\hat{\theta} - E(\hat{\theta})\right]^2 + \left[E(\hat{\theta}) - \theta\right]^2 = varianza\ \hat{\theta} + sesgo^2$$

El criterio de elegir aquel estimador de θ que minimice el Error Cuadrático Medio, tiene en cuenta tanto el sesgo como la varianza.

Si se considera solo la clase de estimadores insesgados, entonces este criterio es equivalente al criterio de eficiencia.

Ejemplo 3.8. El Error Cuadrático Medio del estadístico $\bar{X}$ es,

$$ECN(\bar{X}) = \underbrace{E[\bar{X} - E(\bar{X})]}_{varianza\ de\ \bar{X}} + \underbrace{[E(\bar{X}) - \mu]^2}_{sesgo\ de\ \bar{X}}$$

c) Invarianza

Si $\hat{\theta}$ es un estimador de θ, se puede pensar que $\hat{\theta}^2$ es un estimador de θ^2. Por lo tanto un estimador es invariante cuando $g(\hat{\theta})$ es un estimador de $g(\theta)$.

Ejemplo 3.9. $\bar{X}$ no es un estimador invariante de μ ya que

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

Se sabe que

$V(X) = m'_2 - m'^2_1 = E(X^2) - [E(X)]^2$

De la misma manera, si en lugar de la variable X se tiene la variable $\bar{X}$, entonces

$V(\bar{X}) = E(\bar{X}^2) - [E(\bar{X})]^2$

Ahora bien, según el teorema del límite central, la media y la varianza de $\bar{X}$, son μ y $\frac{\sigma^2}{n}$, respectivamente, entonces

$\frac{\sigma^2}{n} = E(\bar{X}^2) - \mu^2$

De este modo

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

Por ende, $\bar{X}$ no es un estimador invariante.

d) Suficiencia

Es un estimador que proporciona toda la información contenida en la muestra, respecto del parámetro a estimar.

Si se considera el parámetro θ y dos estimadores θ_1 y θ_2; la función de densidad conjunta de θ_1 y θ_2 es $f(\theta_1, \theta_2; \theta)$. Si dado θ_1 la función de densidad condicional $g(\theta_2|\theta_1; \theta)$ de θ_2 para cualquier estimador θ_2 no depende de θ, el estimador θ_1 se llamará suficiente para θ.

Ejemplo 3.10. Si la variable $X \sim N(\mu, \sigma^2)$, se dice que $\bar{X}$ es un estimador suficiente de μ ya que la función de densidad de la muestra puede factorizarse como sigue:

$$f(x_1, x_2, \dots, x_n) = g(\bar{X}; \mu)\, h(x_1, x_2, \dots, x_n)$$

Donde g depende del estadístico $\bar{X}$ y del parámetro μ, y h es independiente de μ. Es decir, esl estadístico $\bar{X}$ proporciona toda la información de la muestra sobre μ.

Propiedades asintóticas

Las propiedades estudiadas hasta ahora están relacionadas con la distribución de un estimador independientemente del tamaño de muestra utilizado. Es decir, cuando un estimador es insesgado, lo es tanto para muestras pequeñas como grandes.

En ocasiones se observa que la distribución de un estimador cambia para distintos tamaños muestrales y, en otros casos, se plantean problemas econométricos en los que es muy difícil, o imposible, determinar las propiedades de los estimadores en muestras pequeñas.

En estas situaciones, la elección de un estimador se basa en sus *propiedades asintóticas*; es decir, las propiedades que satisface el estimador conforme el tamaño muestral T crece y tiende a infinito.

a) Consistencia

La propiedad de consistencia va a garantizar que el estimador genere, con una probabilidad alta, estimaciones próximas al verdadero valor del parámetro si la muestra es lo suficientemente grande.

Sea $\hat{\theta}_t$, el estimador de θ basado en una muestra de tamaño T, es consistente sí y solo sí

$$\forall \varepsilon > 0 \qquad \lim_{t\to\infty} P\left[\left|\hat{\theta}_t - \theta\right| \leq \varepsilon\right] = 1$$

Se denota como $p\lim_{t\to\infty} \hat{\theta}_t = \theta$ o $plim\, \hat{\theta} = \theta$.

Intuitivamente, si un estimador es consistente, a medida que el tamaño de muestra crece, es más probable que las estimaciones estén cerca del verdadero valor del parámetro.

Ejemplo 3.11. $\bar{X}$ es un estimador consistente de μ ya que dado $E(\bar{X}) = \mu$ y $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$

Si $n \to \infty$ entonces $\frac{\sigma^2}{n} \to 0$, por lo tanto $\sigma_{\bar{X}}^2 \to 0$ y $\bar{X} \to \mu$.

---

b) Insesgadez asintótica

El estimador $\hat{\theta}_t$ de θ, se dice que es asintóticamente insesgado sí y solo sí:

$$\lim_{t\to\infty} E(\hat{\theta}_t) = \theta$$

Las propiedades de consistencia e insesgadez asintótica no implican la una a la otra; un estimador consistente no tiene por qué ser asintóticamente insesgado, y viceversa.

a) Si el estimador $\hat{\theta}_t$ es una variable aleatoria que no posee primeros momentos, puede que no sea asintóticamente insesgado, aunque sea consistente.

Ejemplo 3.12. Un estimador $\hat{\theta}_t$ que solo toma dos valores (θ y t^k) con las siguientes probabilidades:

$P(\hat{\theta}_t = 0) = 1 - t^{-1}$ y $P(\hat{\theta}_t = t^k) = t^{-1}$

De forma que $\forall \varepsilon > 0$:

$$P\left[\left|\hat{\theta}_t - \theta\right| > \varepsilon\right] = P(\hat{\theta}_t = t^k) = t^{-1}$$

Este estimador es consistente porque:

$$\lim_{t\to\infty} P\left[\left|\hat{\theta}_t - \theta\right| \leq \varepsilon\right] = 1 - \lim_{t\to\infty} P\left[\left|\hat{\theta}_t - \theta\right| > \varepsilon\right] = 1 - \lim_{t\to\infty} t^{-1} = 1 - 0 = 1$$

Y, sin embargo, es fácil comprobar que es sesgado en muestras finitas:

$E(\hat{\theta}_t) = \theta(1 - t^{-1}) + t^k t^{-1} = \theta - \theta t^{-1} + t^{k-1}$

Y no es asintóticamente insesgado para valores de $k \geq 1$:

$$\lim_{t\to\infty} E(\hat{\theta}_t) = \theta + \lim_{t\to\infty} t^{k-1} \neq 0$$

---

b) El caso contrario, un estimador asintóticamente insesgado que no es consistente.

Ejemplo 3.13. El estimador

$$\hat{\theta}_t \sim N(\theta, 1)\ \forall t$$

Es insesgado para todo tamaño muestral y, por lo tanto, es asintóticamente insesgado. Sin embargo, no es consistente, ya que la probabilidad

$$P\left[\left|\hat{\theta}_t - \theta\right| \leq \varepsilon\right]$$

No decrece cuando t aumenta. El problema es que al aumentar el número de observaciones no disminuye la varianza del estimador, es decir, no proporciona más información sobre θ.

Estos ejemplos, extraídos de Fernández Sainz et. al. (1995) dan una idea de dos *condiciones suficientes* para que un estimador sea consistente:

$$1.\ \lim_{t\to\infty} E\left(\hat{\theta}_t\right) = \theta$$

$$2.\ \lim_{t\to\infty} V\left(\hat{\theta}_t\right) = 0$$

Es decir, si un estimador $\hat{\theta}_t$ es asintóticamente insesgado y su varianza tiende a cero conforme el tamaño muestral tiende a infinito, entonces es consistente.

Estas condiciones son *suficientes*, pero no *necesarias*. Es posible que un estimador sea consistente sin que se cumplan estas condiciones como en el ejemplo anterior, donde el estimador no era asintóticamente insesgado porque su valor medio tendía a infinito al aumentar el tamaño muestral.

c) Distribución asintótica

Cuando no se puede derivar analíticamente la distribución exacta de un estimador o estadístico, una solución es estudiar su distribución de probabilidad para la muestra que tiende a infinito. Si, a medida que el tamaño de muestra va aumentando, la distribución del estimador se aproxima cada vez más a una distribución específica conocida; entonces, para tamaños de muestra grande, se puede usar esta distribución del estimador.

Una sucesión de variables aleatorias $\hat{\theta}_t$ se dice que *converge en distribución* a $\hat{\theta}$, si la función de distribución F_t de $\hat{\theta}_t$ converge a la función de distribución F de $\hat{\theta}$ en todos los puntos de continuidad de F.

Se denota $\hat{\theta}_t \xrightarrow{d} \hat{\theta}$ y F se denomina distribución asintótica de $\hat{\theta}_t$. A veces también se denota $\hat{\theta}_t \xrightarrow{d} F$, o también $\hat{\theta}_t \overset{a}{\sim} F$.

3.3. Estimación por Intervalo

En la estimación por intervalo, generalmente, se construyen dos funciones basadas en los mismos datos muestrales, dentro de los cuales se encontrará, en un sentido probabilístico, el valor desconocido del parámetro. Esto consiste en obtener un cierto intervalo aleatorio que contiene el parámetro a estimar. Este intervalo aleatorio se denomina intervalo de confianza y su magnitud se establece de la siguiente manera

$$P\{L_i \leq \theta \leq L_s\} = 1-\alpha$$

Donde:

θ es el parámetro a estimar

L_i límite inferior del intervalo

L_s límite superior del intervalo

$1-\alpha$ nivel de confianza, esto es la probabilidad que de cada 100 veces que se realice la estimación, $(1-\alpha)\cdot 100$ veces el parámetro a estimar va a caer dentro de este intervalo. Es una probabilidad conocida.

Los límites de confianza L_i y L_s son variables aleatorias que dependen de las observaciones muestrales y del nivel de confianza $1-\alpha$.

En la estimación puntual se estima θ, puntualmente, a través de $\hat{\theta}$ con un riesgo α de cometer un error de estimación.

En la estimación por intervalo se estima θ a través de un intervalo aleatorio que tiene una probabilidad conocida de contener el parámetro.

Para entender aún más el concepto, se introduce el ejemplo de la Figura 3.8. Se supone que el estimador $\hat{\theta} \sim N$, o asintóticamente normal, y que la amplitud del intervalo es $\pm 2\sigma_{\hat{\theta}}$; además, que se hacen varias estimaciones puntuales $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n$. La probabilidad de que este intervalo contenga al parámetro θ se obtiene de la tabla normal y es 0,9544. Se observa que los intervalos de confianza construidos a partir de $\hat{\theta}_2$ y $\hat{\theta}_3$ contienen al parámetro, no así el construido a partir de $\hat{\theta}_4$.

Por lo tanto, el 95,44% de las veces, el intervalo de amplitud $\pm 2\sigma_{\hat{\theta}}$ contendrá al parámetro y en el 4,56% no lo contendrá. Si la amplitud del intervalo fuera $\pm 2{,}4\sigma_{\hat{\theta}}$ la probabilidad sería 0,9836.

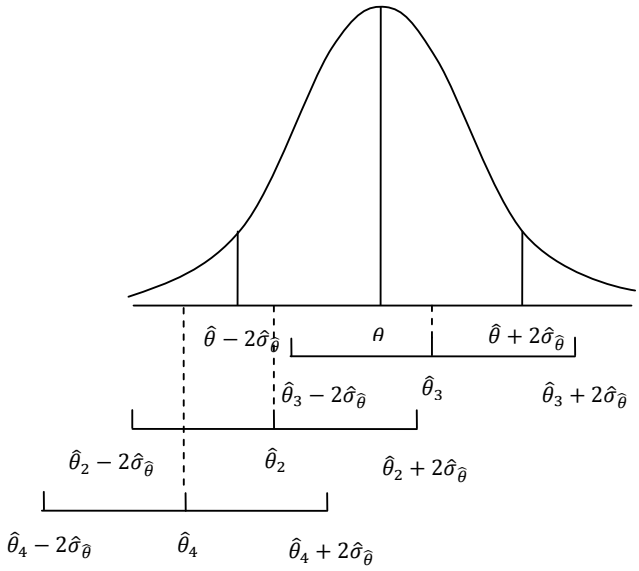


Figura 3.8

Elementos de la estimación por intervalos

1) θ; parámetro a estimar
2) $x_1, x_2, \dots, x_n$; n observaciones muestrales de una población cuya función de densidad es g(X/θ).
3) $\hat{\theta}$; un estimador de θ y como tal variable aleatoria $\hat{\theta} = \hat{\theta}\,(x_1, x_2, \dots, x_n)$.
4) $K = K(\hat{\theta}, \theta)$; es un estadístico que por ser función de $\hat{\theta}$ es una variables aleatoria. $K(\hat{\theta}, \theta)$ se elige de manera que tenga una función de probabilidad conocida y tabulada.
5) g(K); es la función de densidad de la variable $K(\hat{\theta}, \theta)$.
6) $1 - \alpha = P\{K_1 \leq K(\hat{\theta}, \theta) \leq K_2\}$; es el nivel de confianza.
7) K_1, K_2; son los coeficientes de confianza y se obtienen de la tabla de la distribución g(K). Su valor depende del nivel de confianza $(1 - \alpha)$. Los coeficientes de confianza son funciones de $(1 - \alpha)$, esto es, al fijar $(1 - \alpha)$ quedan determinados K_1 y K_2
8) L_i, L_s; son límites del intervalo de confianza para estimar θ. Se obtienen al despejar θ en la desigualdad $K_1 \leq K(\hat{\theta}, \theta) \leq K_2$, de tal manera que $L_i \leq \theta \leq L_s$. L_i, L_s son variables aleatorias que dependen de las observaciones muestrales y de los coeficientes de confianza y son función del nivel de confianza y del tamaño de la muestra.

Intervalo de confianza cuando $\hat{\theta}$ se distribuye normal o aproximadamente normal

Se supone $\hat{\theta} \sim N(\theta, \sigma_{\hat{\theta}})$. El estadístico $K(\hat{\theta}, \theta)$ que se utilizará es $Z \sim N(0, 1)$

$$K\left(\hat{\theta}, \theta\right) = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} = Z \sim N(0, 1)$$

La probabilidad de que el estadístico se encuentre entre dos valores conocidos, k_1 y k_2, es de $1-\alpha$.

$$P\left\{K_1 \leq \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}} \leq K_2\right\} = 1-\alpha$$

En la desigualdad del primer miembro, multiplicando por $(-\sigma_{\hat{\theta}})$:

$$P\left(-K_1\sigma_{\hat{\theta}} \geq \theta - \hat{\theta} \geq -K_2\sigma_{\hat{\theta}}\right) = 1-\alpha$$

Luego sumando $\hat{\theta}$ y reordenando, queda

$$P\left(\hat{\theta} - K_2\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} - K_1\sigma_{\hat{\theta}}\right) = 1-\alpha$$

los límites de confianza son

$$L_i = \hat{\theta} - K_2\sigma_{\hat{\theta}}; \quad K_2 = Z_{1-\alpha/2}$$

$$L_s = \hat{\theta} - K_1\sigma_{\hat{\theta}}; \quad K_1 = Z_{\alpha/2}$$

Dada la simetría de la curva normal, en todos los casos se verifica que $Z_{\alpha/2} = -Z_{1-\alpha/2}$

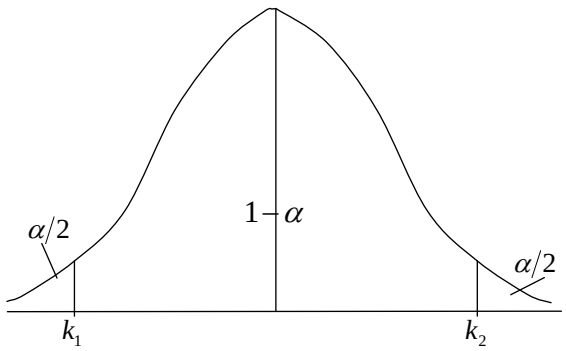


Figura 3.9

o sea que: $L_i = \hat{\theta} - Z_{1-\alpha/2}\sigma_{\hat{\theta}}$; $\quad L_s = \hat{\theta} + Z_{1-\alpha/2}\sigma_{\hat{\theta}}$

Intervalo de confianza para estimar μ

Para estimar el intervalo de confianza de la media en la población, hay que tener en cuenta la varianza de la población bajo estudio y el tamaño de la muestra.

Si la población es normal de parámetro (μ, σ), $\bar{X}$ se distribuye normal con parámetro $(\mu, \sigma/\sqrt{n})$; es decir,

Si $X \sim N(\mu, \sigma) \quad \rightarrow \quad \bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

El estadístico es

$$K(\bar{X}/\mu) = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

El cual se asocia a una distribución de probabilidad, de acuerdo al conocimiento que se tenga de la varianza en la población y del tamaño de la muestra utilizado.

Por esto es que se tienen 3 casos: desvío conocido, desvío desconocido pero muestra grande y desvío desconocido con muestra pequeña. A continuación se detallan los tres casos.

Caso 1. σ es conocida,

El estadístico $K(\bar{X}/\mu) \sim N(0,1)$

resolviendo, como se indicó anteriormente, los límites del intervalo se definen como

$$L_i = \bar{X} - Z_{1-\alpha/2}\,\sigma/\sqrt{n}$$

$$L_s = \bar{X} + Z_{1-\alpha/2}\,\sigma/\sqrt{n}$$

Caso 2. σ es desconocida y la muestra es grande

El estadístico se distribuye como una normal

$$K(\bar{X}/\mu) \rightarrow Z(0,1)$$

el desvío desconocido debe estimarse por

$$s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$$

Cuando la muestra es grande $S \rightarrow \sigma$. Al resolver, los límites del intervalo de confianza se definen como

$$L_i = \bar{X} - Z_{1-\alpha/2}\,s/\sqrt{n}$$

$$L_s = \bar{X} + Z_{1-\alpha/2}\,s/\sqrt{n}$$

Caso 3. σ es desconocida y la muestra es pequeña

El estadístico K se distribuye como una t de Student

$$K(\bar{X},\mu) \sim t$$

Trabajando desde las observaciones muestrales, se las estandariza, eleva al cuadrado y suma para definir una variable Y

$$Y = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{\sigma^2}$$

Donde $Y \sim \boldsymbol{\chi}^2_{n-1}$

Multiplicando y dividiendo por $n-1$:

$$Y = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{\sigma^2}\frac{n-1}{n-1} = (n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{n-1}$$

En consecuencia, dada una muestra de tamaño n y siendo

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{(\bar{X} - \mu)\sqrt{n}}{\sigma} \sim N(0,1)$$

se obtiene

$$t_{n-1} = \frac{Z}{\sqrt{\frac{Y}{n-1}}} = \frac{\frac{(\bar{X}-\mu)\sqrt{n}}{\sigma}}{\sqrt{\frac{(n-1)\frac{s^2}{\sigma^2}}{(n-1)}}} = \frac{\frac{(\bar{X}-\mu)\sqrt{n}}{\sigma}}{\frac{s}{\sigma}}$$

Esto significa que

$$t_{n-1} = \frac{(\bar{X} - \mu)\sqrt{n}}{s}$$

El nivel de confianza es: $\mathrm{P}\left\{K_1 \leq \frac{(\bar{X}-\mu)\sqrt{n}}{s} \leq K_2\right\} = 1 - \alpha$

Donde los valores críticos k_1 y k_2 vienen definidos por la distribución t

$K_1 = t_{n-1;\, \alpha/2}$

$K_2 = t_{n-1;\, 1-\alpha/2}$

por lo tanto los límites L_1 y L_2 del intervalo de confianza serán

$L_i = \bar{X} - t_{n-1;\, 1-\alpha/2}\, S/\sqrt{n}$

$$L_s = \bar{X} - t_{n-1;\,\alpha/2}\, S/\sqrt{n}$$

En resumen, la media de la muestra $(\overline{\mathrm{X}})$ es usada para estimar la media desconocida de la población (μ). Debido a que $\overline{\mathrm{X}}$ varía de muestra en muestra, no es igual a la media de la población (μ); es decir, existe un error de la muestra. Para reflejar el alcance de este error muestral es útil una estimación de intervalo en torno de $\overline{\mathrm{X}}$.

Tamaño del intervalo

El tamaño del intervalo dependerá de la confianza que se desee, de que contenga a la media de la población verdadera y desconocida.

Para tener una confianza de 0,95 de que la estimación del intervalo contenga a la media de la población verdadera, la estimación de intervalo es:

$\bar{X} \pm$ error de la muestra $(e) = \bar{X} \pm Z\frac{\sigma}{\sqrt{n}} =$ estimación de intervalo de μ

Ejemplo 3.14. En el estudio del ingreso promedio de los clientes de un banco que poseen la tarjeta AA, se obtuvo

$\bar{X} = 793{,}00$

$\sigma = 531{,}20$

El intervalo es

$$\bar{X} \pm 1{,}96\frac{\sigma}{\sqrt{n}} = 793{,}00 \pm 1{,}96\frac{531{,}20}{\sqrt{10}} = 793{,}00 \pm 329{,}24$$

e incluye la media poblacional verdadera $\mu = 831{,}58$.

Aproximadamente el 95% de las muestras generarán una estimación de intervalo que incluirá a la verdadera media de la población.

El tamaño de la estimación del intervalo depende de cuatro factores.

- nivel de confianza.
- desviación estándar de la población.
- tamaño de la muestra.
- conocimiento de la desviación estándar de la población $\sigma_X = \sigma$.

Si la desviación estándar de la población no es conocida, es necesario estimarla con la desviación estándar de la muestra, S. Si, además, el tamaño de la muestra es pequeño, la estimación del intervalo de 0,95, es:

$$\bar{X} \pm t\frac{S}{\sqrt{n}}$$

Donde, se reemplaza la distribución normal con la distribución t. Si el tamaño de muestra es grande, aun cuando no se conozca el desvío, se estima el intervalo con la distribución normal

---

Ejemplo 3.15. Siguiendo con el caso del Ejemplo 3.14, ahora se tiene:

$\mathrm{S} = 453{,}56$

$\mathrm{n} = 10$

Entonces el intervalo es:

$$\bar{\mathrm{X}} \pm t\frac{s}{\sqrt{n}} = 793{,}00 \pm 2{,}26\frac{453{,}56}{\sqrt{10}} = 793{,}00 \pm 324{,}15$$

A medida que n se vuelva más grande, la distribución de t se aproxima a la distribución normal. Por ejemplo, si $n = 20 \Rightarrow t = 2{,}09$; $n = 60 \Rightarrow t = 2{,}00$

Intervalo de confianza para la varianza de una variable con distribución poblacional normal

El parámetro a estimar es ahora σ^2 y su estimador es

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$$

El estadístico que se utiliza es

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

El nivel de confianza $1-\alpha$ determina los valores críticos k_1 y k_2 entre los que se encuentra el estadístico K

$$P\left\{K_1 \leq (n-1)\frac{S^2}{\sigma^2} \leq K_2\right\} = 1 - \alpha$$

para despejar σ^2 se deben seguir los siguientes pasos:

1. se toman recíprocas y se cambia el sentido de la desigualdad entre llaves, esto es:

$$\frac{1}{K_1} \geq \frac{\sigma^2}{(n-1)S^2} \geq \frac{1}{K_2}$$

2. se multiplican los miembros de la desigualdad por $(n-1)S^2$

$$\frac{(n-1)S^2}{K_2} \leq \sigma^2 \leq \frac{(n-1)S^2}{K_1}$$

El intervalo de confianza para estimar σ^2 con nivel de confianza $1-\alpha$ será

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

K_1 y K_2 son valores particulares de χ^2 que vienen determinados por el nivel de confianza $1-\alpha$ deseado y se encuentran en la tabla

$$\mathrm{K}_1 = \chi^2_{\mathrm{n}-1;\,\alpha/2}$$

$$\mathrm{K}_2 = \chi^2_{\mathrm{n}-1;\,1-\alpha/2}$$

Los límites del intervalo se definen

$$\mathrm{L}_\mathrm{i} = \frac{(n-1)\mathrm{S}^2}{\chi^2_{n-1;\,1-\alpha/2}}$$

$$\mathrm{L}_\mathrm{s} = \frac{(n-1)\mathrm{S}^2}{\chi^2_{n-1;\,\alpha/2}}$$

Ejemplo 3.16. Por requerimiento del Departamento Técnico de la empresa INTI SA se desea estimar un intervalo de confianza para la varianza del llenado de las botellas de una bebida gaseosa a un nivel de confianza de 0,95. Se conoce que el llenado de las botellas se distribuye aproximadamente normal. Se toma una muestra de 19 botellas al azar con los siguientes resultados.

$$\sum \mathrm{X}_\mathrm{i} = 185{,}2\mathrm{dc}^3 \qquad \sum \mathrm{X}_\mathrm{i}^2 = 3545{,}8\mathrm{dc}^3$$

El intervalo de confianza para la varianza, σ^2, es:

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

El enunciado del problema no indica el valor de S^2, pero ofrece algunos resultados para la variable bajo estudio que permiten su cálculo.

Si $S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{X}^2}{n-1}$

Reemplazando el resultado de la suma y de la suma de los elementos cuadráticos

$$S^2 = \frac{3545{,}8 - 19\left(\frac{185{,}2}{19}\right)^2}{18} = 96{,}70$$

La distribución chi cuadrado con 18 grados de libertad, para un nivel de confianza de 0,95, arroja los siguientes estadísticos:

$\chi^2_{n-1;\,\alpha/2} = \chi^2_{18;\,0{,}025} = 8{,}23$

$\chi^2_{n-1;\,1-\alpha/2} = \chi^2_{18;\,0{,}975} = 31{,}53$

Reemplazando el valor de S^2 y los valores críticos de la distribución χ^2 en el intervalo de confianza

$$P\left(\frac{18 * 96{,}70}{31{,}53} \leq \sigma^2 \leq \frac{18 * 96{,}70}{8{,}23}\right) = 0{,}95$$

$$P(55{,}26 \leq \sigma^2 \leq 211{,}49) = 0{,}95$$

El resultado indica que la variabilidad en el llenado de las botellas de una bebida gaseosa se encuentra entre $55{,}26\ \mathrm{dc}^3$ y $211{,}49\ \mathrm{dc}^3$

Intervalo de confianza para la diferencia de medias

El parámetro a estimar es la diferencia en las medias de dos poblaciones, $(\mu_1 - \mu_2)$. Para realizar la estimación se cuenta con muestras provenientes de ambas poblaciones de las cuales se obtienen las respectivas medias, $\overline{\mathrm{X}}_1$ y $\overline{\mathrm{X}}_2$.

Se sabe que si las muestras son independientes, $\overline{\mathrm{X}}_1$ y $\overline{\mathrm{X}}_2$ son independientes, entonces

1) $\mathrm{E}(\overline{\mathrm{X}}_1 - \overline{\mathrm{X}}_2) = \mathrm{E}(\overline{\mathrm{X}}_1) - \mathrm{E}(\overline{\mathrm{X}}_2) = \mu_1 - \mu_2$

2) $\mathrm{V}(\overline{\mathrm{X}}_1 - \overline{\mathrm{X}}_2) = \mathrm{V}(\overline{\mathrm{X}}_1) + \mathrm{V}(\overline{\mathrm{X}}_2) = \frac{\sigma_1^2}{\mathrm{n}_1} + \frac{\sigma_2^2}{\mathrm{n}_2}$

3) Si $\overline{\mathrm{X}}_1 \sim \mathrm{N}\left(\mu_1, \sqrt{\frac{\sigma_1^2}{\mathrm{n}_1}}\right)$ y $\overline{\mathrm{X}}_2 \sim \mathrm{N}\left(\mu_2, \sqrt{\frac{\sigma_2^2}{\mathrm{n}_2}}\right)$

entonces

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2;\ \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Al igual que en el caso de intervalos para la media, μ, se deben considerar alternativas de acuerdo al conocimiento que se tenga de las varianzas en cada población y al tamaño de muestra utilizado. A continuación se detalla cada uno de los casos.

Caso 1. σ_1^2 y σ_2^2 conocidas.

Cuando las varianzas son conocidas en ambas poblaciones, se utiliza un estadístico $\mathrm{Z} \sim \mathrm{N}(0,1)$

$$\mathrm{Z} = \frac{\overline{\mathrm{X}}_1 - \overline{\mathrm{X}}_2 - (\mu_1 - \mu_2)}{\sigma_{\overline{\mathrm{X}}_1 - \overline{\mathrm{X}}_2}} = \frac{\overline{\mathrm{X}}_1 - \overline{\mathrm{X}}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{\mathrm{n}_1} + \frac{\sigma_2^2}{\mathrm{n}_2}}}$$

Operando de la misma manera que en los casos anteriores, los límites para estimadores que se distribuyen normal para la diferencia de dos medias de poblaciones con varianzas conocidas se definen como

$$\mathrm{P}(\mathrm{Z}_1 \leq \mathrm{Z} \leq \mathrm{Z}_2) = 1 - \alpha$$

Remplazando Z por su igual

$$\mathrm{P}\left(\mathrm{Z}_1 \leq \frac{\overline{\mathrm{X}}_1 - \overline{\mathrm{X}}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq \mathrm{Z}_2\right) = 1 - \alpha$$

Multiplicando por $-\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$ y sumando $(\bar{X}_1-\bar{X}_2)$, se tiene el intervalo de confianza $1-\alpha$ para la diferencia de las medias de dos poblaciones

$$\mathrm{P}\left((\bar{X}_1-\bar{X}_2)-\mathrm{Z}_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\leq\mu_1-\mu_2\leq(\bar{X}_1-\bar{X}_2)+\mathrm{Z}_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)=1-\alpha$$

De modo que los límites serán:

$$\mathrm{L_i}=\bar{\mathrm{X}}_1-\bar{\mathrm{X}}_2-\mathrm{Z}_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

$$\mathrm{L_s}=\bar{\mathrm{X}}_1-\bar{\mathrm{X}}_2+\mathrm{Z}_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

Caso 2. σ_1^2 y σ_2^2 desconocidas y muestras grandes.

Los estimadores de las varianzas poblacionales, S_1^2 y S_2^2, tienden a σ_1^2 y σ_2^2, respectivamente, por lo que el estadístico K se distribuye como una normal tipificada. Entonces:

$$\mathrm{Z}=\frac{\bar{\mathrm{X}}_1-\bar{\mathrm{X}}_2-(\mu_1-\mu_2)}{\sqrt{\frac{\mathrm{S}_1^2}{n_1}+\frac{\mathrm{S}_2^2}{n_2}}}$$

Teniendo en cuenta lo desarrollado anteriormente, el intervalo de confianza se plantea

$$\mathrm{P}\left((\bar{X}_1-\bar{X}_2)-\mathrm{Z}_{1-\frac{\alpha}{2}}\sqrt{\frac{\mathrm{S}_1^2}{n_1}+\frac{\mathrm{S}_2^2}{n_2}}\leq\mu_1-\mu_2\leq(\bar{X}_1-\bar{X}_2)+\mathrm{Z}_{1-\frac{\alpha}{2}}\sqrt{\frac{\mathrm{S}_1^2}{n_1}+\frac{\mathrm{S}_2^2}{n_2}}\right)=1-\alpha$$

De modo que los límites vienen determinados por

$$L_i = \overline{X}_1 - \overline{X}_2 - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$$L_s = \overline{X}_1 - \overline{X}_2 + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Caso 3. σ_1^2 y σ_2^2 desconocidas pero iguales y muestras pequeñas.

En este caso se debe utilizar la distribución de probabilidad t, que es el cociente entre una normal y la raíz cuadrada de una chi cuadrada dividida por sus grados de libertad. Al tener dos muestras provenientes de dos poblaciones, el estadístico Z se define como

$$Z(0,1) = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}$$

La χ^2 relaciona las varianzas de la muestra con las de la población teniendo en cuenta que la suma de variables chi cuadrado dan por resultado otra chi cuadrado.

$$\chi^2_{gl} = \chi^2_{n_1-1} + \chi^2_{n_2-1}$$

Utilizando el estadístico definido en el cálculo del intervalo de confianza para la varianza, se tiene que la χ^2 que reúne las dos muestras provenientes de las dos poblaciones se definen como:

$$\chi^2_{gl} = (n_1 - 1)\frac{S_1^2}{\sigma^2} + (n_2 - 1)\frac{S_2^2}{\sigma^2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$$

Esto significa que $t_{n_1+n_2-2} =$

$$t_{n_1+n_2-2} = \frac{Z(0,1)}{\sqrt{\frac{\chi^2_{n_1-1+n_2-1}}{n_1+n_2-2}}} = \frac{\frac{\overline{X}_1-\overline{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma^2}{n_1}+\frac{\sigma^2}{n_2}}}}{\sqrt{\frac{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{\sigma^2}}{n_1+n_2-2}}}$$

Simplificando en la expresión anterior $\sqrt{\sigma^2}$ queda

$$t_{n_1+n_2-2} = \frac{\frac{\overline{X}_1-\overline{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}}$$

Luego, multiplicando extremos por medios

$$t_{n_1+n_2-2} = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{n_1+n_2}{n_1 n_2}}\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}}$$

Se obtiene el estadístico que permite hallar los límites

$$L_i = \overline{X}_1 - \overline{X}_2 - t_{n_1+n_2-2;1-\frac{\alpha}{2}}\sqrt{\frac{n_1+n_2}{n_1 n_2}\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$$

$$L_s = \overline{X}_1 - \overline{X}_2 + t_{n_1+n_2-2;1-\frac{\alpha}{2}}\sqrt{\frac{n_1+n_2}{n_1 n_2}\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$$

De modo que, la probabilidad será

$$P\left((\bar{X}_1 - \bar{X}_2) - t_{n_1+n_2-2;1-\frac{\alpha}{2}} \sqrt{\frac{n_1+n_2}{n_1 n_2} \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} \leq \mu_1 - \mu_2 \right.$$
$$\left. \leq (\bar{X}_1 - \bar{X}_2) + t_{n_1+n_2-2;1-\frac{\alpha}{2}} \sqrt{\frac{n_1+n_2}{n_1 n_2} \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} \right)$$
$$= 1 - \alpha$$

Caso 4. σ_1^2 y σ_2^2 desconocidas y distintas en muestras pequeñas.

Si σ_1^2 y σ_2^2 son desconocidas y distintas, la distribución de

$$K = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Depende de σ_1^2/σ_2^2 y, en principio, K no puede utilizarse para el cálculo del intervalo; en su lugar existen diferentes soluciones al problema. Gomez Villegas (2005) utiliza la propuesta por Hsu que consiste en aproximar la distribución por una t, los grados de libertad corresponden al mínimo tamaño de muestra (n_1 o n_2) menos 1 ($(n_1, n_2) - 1$). El intervalo es

$$P\left(\bar{X}_1 - \bar{X}_2 - K_{2,(n_1,n_2)-1} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + K_{2,(n_1,n_2)-1} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right)$$
$$= 1 - \alpha$$

Para cualquiera de los casos, si en su recorrido el intervalo de confianza contiene el cero, las poblaciones bajo estudio presentan valores iguales para las medias de la variable en consideración; y, cuando no lo contiene, son distintas.

Ejemplo 3.17. Para un determinado producto de consumo popular, el promedio de ventas en una muestra de 10 comercios fue de $342500

con un desvío estimado de $20000. Para un segundo producto, el promedio de ventas en una muestra de 12 comercios fue de $325000 con un desvío estimado de $17500.

Se supone que los montos de las ventas por local tienen una distribución normal para ambos productos. La Cámara Empresaria quiere saber, a un nivel de confianza de 0,95, si hay diferencias significativas en el nivel promedio de ventas por comercio.

La información brindada en el enunciado puede ordenarse de la siguiente manera

		Producto 1	Producto 2
Ventas promedio	$\bar{X}$	342500	325000
Variabilidad	S	20000	17500
Tamaño de la muestra	n	10	12

Dado que no se conoce el desvío en la población y la muestra es pequeña, debe aplicarse el tercer caso:

$$P\left((\overline{X}_1 - \overline{X}_2) - t_{n_1+n_2-2;1-\frac{\alpha}{2}} \sqrt{\frac{n_1+n_2}{n_1 n_2} \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} \leq \mu_1 - \mu_2 \right.$$

$$\left. \leq (\overline{X}_1 - \overline{X}_2) + t_{n_1+n_2-2;1-\frac{\alpha}{2}} \sqrt{\frac{n_1+n_2}{n_1 n_2} \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} \right)$$

$$= 1 - \alpha$$

La distribución t, al nivel de confianza de 0,95 y con 20 grados de libertad, tiene los siguientes estadísticos:

$$t_{n_1+n_2-2;1-\alpha/2} = t_{10+12-2;0,975} = t_{20;0,975} = \pm 2{,}086$$

El error es igual a:

$$e = t_{n_1+n_2-2;1-\frac{\alpha}{2}}\sqrt{\frac{n_1+n_2}{n_1 n_2}\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$$

$$= \pm 2{,}086\sqrt{\frac{10+12}{10*12}\frac{9*20000^2+11*17500^2}{10+12-2}}$$

$$= \pm 2{,}0860\sqrt{63880208{,}3} = \pm 16672{,}38$$

Reemplazando en el intervalo de confianza

$$P((342500-325000)-16672{,}38 \leq \mu_1-\mu_2 \leq (342500-325000)+16672{,}38) = 1-\alpha$$

$$P(17500-16672{,}38 \leq \mu_1-\mu_2 \leq 17500+16672{,}38) = 0{,}95$$

$$P(827{,}63 \leq \mu_1-\mu_2 \leq 34172{,}38) = 0{,}95$$

El intervalo de confianza alcanzado indica que el producto 1 tiene un promedio de ventas más elevado que el producto 2.

Intervalo de confianza para el cociente de varianzas poblacionales

Al igual que con la diferencia de medias, el cociente de varianzas se utiliza para comparar dos poblaciones. Si el intervalo de confianza contiene el 1 se considera que las dos poblaciones son iguales y si no lo contiene, son distintas.

El parámetro a estimar es σ_2^2/σ_1^2 y las poblaciones bajo estudio deben ser normales.

En cada una de las poblaciones se toman muestras, de tamaño n_1 y n_2, con las cuales se estimará el cociente de varianzas (σ_2^2/σ_1^2) a partir del cálculo de las varianzas de cada muestra $(S_1^2\ y\ S_2^2)$.

En el intervalo de confianza de la varianza el estadístico a utilizar se distribuye χ^2 con $n-1$ grados de libertad

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

En el cociente de varianzas, se tiene un cociente de χ^2 que da por resultado el estadístico F

$$\mathrm{F}_{\mathrm{n}_1-1;\,\mathrm{n}_2-1} = \frac{\frac{(\mathrm{n}_1-1)\frac{\mathrm{S}_1^2}{\sigma_1^2}}{(\mathrm{n}_1-1)}}{\frac{(\mathrm{n}_2-1)\frac{\mathrm{S}_2^2}{\sigma_2^2}}{(\mathrm{n}_2-1)}} = \frac{\mathrm{S}_1^2\,\sigma_2^2}{\mathrm{S}_2^2\,\sigma_1^2}$$

El intervalo de confianza entre dos valores críticos que garantizan la probabilidad $1-\alpha$ se define como

$$P\left(K_1 \leq \frac{\mathrm{S}_1^2\,\sigma_2^2}{\mathrm{S}_2^2\,\sigma_1^2} \leq \mathrm{K}_2\right) = 1-\alpha$$

de donde resulta

$$P\left(K_1\frac{S_2^2}{S_1^2} \leq \frac{\sigma_2^2}{\sigma_1^2} \leq K_2\frac{S_2^2}{S_1^2}\right) = 1-\alpha$$

$K_1\, y\, K_2$ se encuentran en la tabla F y están asociados a la confianza $1-\alpha$

$K_1 = F_{n_1-1;\,n_2-1;\,\alpha/2}$

$K_2 = F_{n_1-1;\,n_2-1;\,1-\alpha/2}$

los límites del intervalo son

$L_i = F_{n_1-1;\,n_2-1;\,\alpha/2}\dfrac{S_2^2}{S_1^2}$

$L_s = F_{n_1-1;\,n_2-1;\,1-\alpha/2}\dfrac{S_2^2}{S_1^2}$

De modo que la probabilidad

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

Ejemplo 3.18. Se utilizarán los datos del ejemplo de intervalo de confianza para las diferencias de medias.

		Producto 1	Producto 2
Ventas promedio	$\bar{X}$	342500	325000
Variabilidad	S	20000	17500
Tamaño de la muestra	n	10	12

Para calcular el intervalo de confianza de la diferencia de varianzas, en las ventas promedios de dos productos, se utiliza el intervalo

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

Los estadísticos de la distribución F para 9 grados de libertad en el numerador y 11 grados de libertad en el denominador, para un nivel de confianza de 0,95, son:

$$F_{n_1-1;\,n_2-1;\,\alpha/2}=F_{9,11;\,0,025}=\frac{1}{F_{11,9;\,0,975}}=\frac{1}{3,91}$$

$$F_{n_1-1;\,n_2-1;\,1-\alpha/2}=F_{9,11;\,0,975}=3,59$$

Reemplazando en el intervalo de confianza

$$\mathrm{P}\left(\frac{1}{3,91}\,\frac{17500^2}{20000^2}\leq\frac{\sigma_2^2}{\sigma_1^2}\leq 3,59\,\frac{17500^2}{20000^2}\right)=0,95$$

$$\mathrm{P}\left(0,1958\leq\frac{\sigma_2^2}{\sigma_1^2}\leq 2,7486\right)=0,95$$

Esto significa que las varianzas son iguales

Intervalo de confianza para estimar la proporción de ocurrencia de un evento en la población

Se utiliza para variables cualitativas y permite conocer la frecuencia con la que se presenta un evento en la población. El parámetro $\theta = \pi$ y el estimador $\hat{\theta} = \mathrm{p}$; es decir, la frecuencia con la que se presenta en la población (π) y con la que se presenta en la muestra (p).

El estadístico $k = k(p/\pi)$ tiene asociado una función de densidad $g(k)$ conocida de igual manera que con la estimación de la media poblacional

$$k = \frac{p - \pi}{\sqrt{\frac{p(1-p)}{n}}}$$

Generalmente, no se conoce el desvío en la población por lo que se estima a partir del resultado de la muestra $\sqrt{p * (1-p)/n}$

El intervalo de confianza $(1-\alpha)$ para el estadístico K se define asociando los valores críticos k_1 y k_2 oportunos

$$P\left(K_1 \leq \frac{p - \pi}{\sqrt{\frac{p(1-p)}{n}}} \leq K_2\right) = 1 - \alpha$$

Resolviendo como se realizó anteriormente

$$P\left(K_1\sqrt{\frac{p(1-p)}{n}} \leq p - \pi \leq K_2\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$$

El intervalo de confianza para estimar la ocurrencia del evento en la población con una probabilidad $1-\alpha$ se define así

$$P\left(p - K_2\sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + K_1\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$$

Si n es grande $(n > 30)$ $k \sim \mathrm{N}(0,1)$

Si n es pequeña $(n \leq 30)$ $k \sim \mathrm{t}_{n-1}$

Intervalo de confianza para la igualdad de proporciones poblacionales

El parámetro $\theta = \pi_1 = \pi_2 \rightarrow \pi_1 - \pi_2 = 0$

Estimador $\hat{\theta} = \mathrm{p}_1 = \mathrm{p}_2 \rightarrow p_1 - p_2 = 0$

Para cada una de las muestras se tiene

$$\mathrm{E}(\mathrm{p}_1) = \pi_1 \qquad \mathrm{V}(\mathrm{p}_1) = \frac{\mathrm{p}_1(1-\mathrm{p}_1)}{\mathrm{n}_1}$$

$$\mathrm{E}(\mathrm{p}_2) = \pi_2 \qquad \mathrm{V}(\mathrm{p}_2) = \frac{\mathrm{p}_2(1-\mathrm{p}_2)}{\mathrm{n}_2}$$

Esto significa que

$$\mathrm{E}(\mathrm{p}_1 - \mathrm{p}_2) = \mathrm{E}(\mathrm{p}_1) - \mathrm{E}(\mathrm{p}_2) = \pi_1 - \pi_2$$

$$\mathrm{V}(\mathrm{p}_1 - \mathrm{p}_2) = \mathrm{V}(\mathrm{p}_1) + \mathrm{V}(\mathrm{p}_2) = \frac{\mathrm{p}_1(1-\mathrm{p}_1)}{\mathrm{n}_1} + \frac{\mathrm{p}_2(1-\mathrm{p}_2)}{\mathrm{n}_2}$$

Si n_1 y n_2 son mayores que 30, $(p_1 - p_2) \sim N(\pi_1 - \pi_2;\ \sigma_{\pi_1 - \pi_2})$

La variable estandarizada será

$$K = \frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

$$P\left((p_1 - p_2) - K_2 \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \leq \pi_1 - \pi_2 \right.$$

$$\left. \leq (p_1 - p_2) - K_1 \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \right) = 1 - \alpha$$

Berenson y Levine (1993) definen la diferencia entre dos modalidades de la misma variable. Para esto genera una estimación total de la proporción de la población combinando las proporciones muestrales; el valor

$$\bar{p} = \frac{p_1 + p_2}{n_1 + n_2}$$

la varianza de este valor $\bar{p}$ es

$$\bar{p}(1 - \bar{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

Y el intervalo de confianza

$$P\left(\bar{p} \pm Z \sqrt{\bar{p}(1 - \bar{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right) = 1 - \alpha$$

3.4. Pruebas de hipótesis

La prueba de hipótesis, también conocida como dócima o contraste de hipótesis, proporciona una oportunidad para analizar la información. Cuando un hallazgo empírico emerge del análisis de información basado en una muestra, una hipótesis sencilla debe ocurrírsele a todo investigador:

¿Representa el hallazgo empírico solo un accidente de muestreo?

Una hipótesis estadística es una conjetura con respecto a uno o más parámetros poblacionales y el método de contraste es el que permite tomar decisiones en base a datos muestrales y riesgos conocidos.

La prueba de hipótesis, o contraste de hipótesis, consiste en:

1. Formular hipótesis nula, H_0
2. Formular hipótesis alternativa, H_1
3. Establecer nivel de confianza para H_0
4. Definir estadístico de contraste
5. Aplicar regla de decisión

La hipótesis nula indica que en la población no se han producido cambios y la hipótesis alternativa es una afirmación que enfrenta lo establecido por la hipótesis nula. Se fija un nivel de confianza $1-\alpha$ de que la hipótesis nula sea cierta. Este nivel de confianza fija los puntos críticos dentro de los cuales se acepta o se rechaza la hipótesis nula.

Tipos de dócima

Pueden plantearse tres situaciones:

Dócima bilateral

Dócima lateral izquierda

Dócima lateral derecha

Dócima bilateral

En la dócima bilateral, las hipótesis nulas se plantean con igualdades entre el parámetro a docimar y un valor particular que puede alcanzar este parámetro. La hipótesis alternativa contradice a la nula y siempre se plantea como desigualdad.

$H_0: \theta = \theta_0$

$H_1: \theta \neq \theta_0$

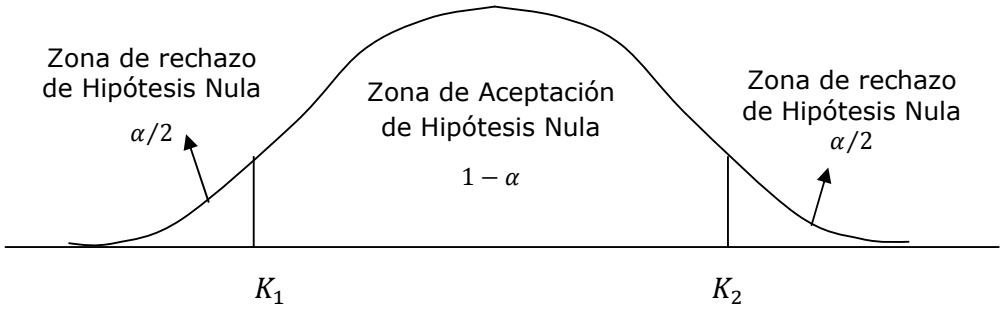


Figura 3.10

Los estadísticos de contraste, K_1 y K_2, son los puntos críticos que definen la región de aceptación y de rechazo de la H_0 y se obtienen de la distribución de probabilidad asociada a la prueba al nivel de confianza $1-\alpha$ fijado. α es la probabilidad de rechazar la H_0 siendo cierta. Esto se lo conoce como Error tipo I. Si se acepta la H_0 siendo falsa se comete el Error tipo II y se lo conoce como β, esto es la Potencia de la prueba.

De acuerdo al valor del estadístico empírico y a los valores que delimitan la zona de aceptación y la de rechazo se concluye en aceptar la H_0 o en rechazarla. La característica de la dócima bilateral es la de centrar la probabilidad de la zona de aceptación, de modo que la zona de rechazo quede dividida en dos partes conteniendo los valores extremos.

Dócima lateral

Las dócimas laterales plantean la hipótesis nula como desigualdad y la alternativa es otra desigualdad que complementa y opone a la hipótesis nula. La zona de aceptación no está centrada y la zona de rechazo se reúne en una sola cola.

Cuando es lateral izquierda:

$H_0: \theta \geq \theta_0$

$H_1: \theta < \theta_0$

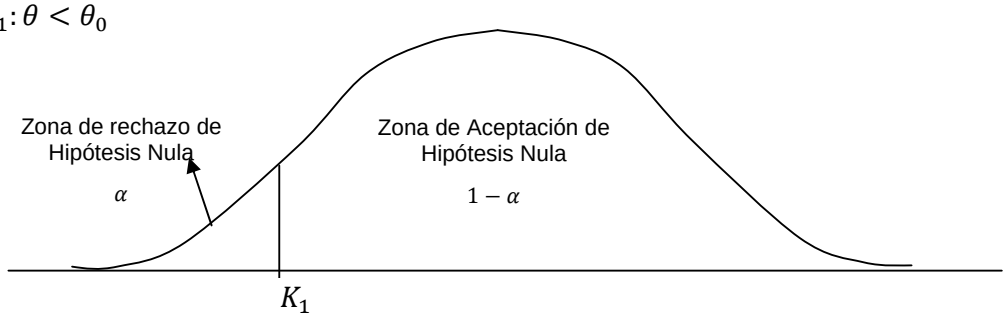


Figura 3.11

Cuando es lateral derecha:

$H_0: \theta \leq \theta_0$

$H_1: \theta > \theta_0$

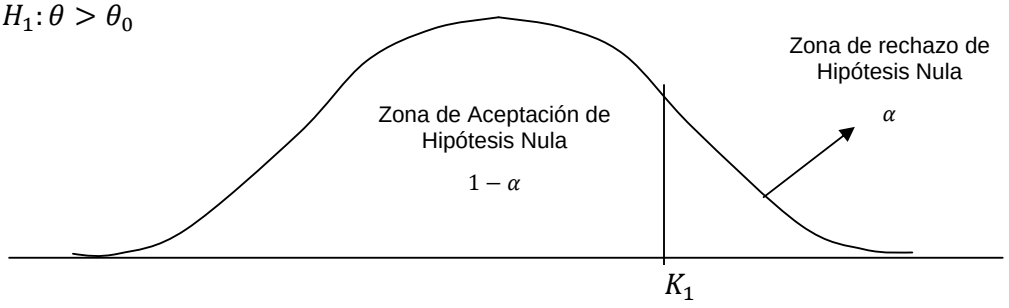


Figura 3.12

En la Figura 3.13 se reúnen las diferentes situaciones de pruebas de hipótesis, para variables cuantitativas, con el estadístico a utilizar.

Ejemplo 3.19. Un grupo de investigadores está interesado en conocer la concentración de una enzima en cierta población de seres humanos. Toman una muestra de 20 individuos de la población bajo estudio y calculan que la media muestral es de 22 y la varianza muestral $S^2 = 45$. Se supone que la muestra proviene de una población que sigue una distribución normal con varianza desconocida. Los investigadores saben que, habitualmente, los niveles de concentración de esta enzima en las poblaciones es de 25. Una hipótesis a ser investigada es que los niveles de concentración de la enzima en esta población es distinto de 25.

1. La **Hipótesis Nula** indica que la concentración de enzima es igual a 25 y que si todas las personas de la población fueran contactadas, se encontraría que el nivel de enzima es 25. $H_0: \mu = 25$

2. Otro argumento, que será denominado **Hipótesis Alternativa,** es que la concentración media de la enzima de la población es diferente de 25. Es importante destacar que se está identificando la hipótesis alternativa con la conclusión a la que se quiere llegar; de manera que, si los datos permiten rechazar la hipótesis nula, la conclusión de los

investigadores tendrá mayor peso, dado que la probabilidad complementaria de rechazar una hipótesis nula verdadera es pequeña. $H_1: \mu \neq 25$

Dócima	Estadístico
Media con σ^2 poblacional conocida	$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$
Media con σ^2 desconocida, n grande	$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim N(0,1)$
Media con σ^2 desconocida, n pequeña, probabilidad Normal	$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$
Proporción	$Z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$
Diferencias de medias con σ^2 conocida y n de cualquier tamaño	$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$
Diferencia de medias con σ^2 desconocida muestras grandes	$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0,1)$
Diferencia de medias, σ^2 desconocida pero iguales y n pequeño	$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{n_1 + n_2}{n_1 n_2} \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}} \sim t_{n_{1+n_2-2}}$
Diferencia de medias, σ^2 desconocida y distintas, n pequeño	$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_{mínimo\ (n_1, n_2)-1}$
Igualdad de Proporciones	$K = \frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0,1)$
σ^2 de población normal	$W = (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$
Igualdad de varianza	$\frac{S_1^2\, \sigma_2^2}{S_2^2\, \sigma_1^2} \sim F_{n_1-1;\, n_2-1}$

Figura 3.13. Estadísticos para la dócima

Pero, ¿qué tan convincente es la evidencia? Después de todo, solo se conocen los resultados de una muestra de 20 personas. La diferencia

entre 25 (el nivel de concentración de enzima conocido), y 22 (el promedio de la muestra), podría ser más bien un caso de error de muestreo que diferencias reales en la existencia de enzimas en la población.

3. Para evaluar la evidencia estadísticamente, se establece una probabilidad denominada **nivel de confianza** $(1-\alpha)$ de 0,95 (ó 95%) de que la hipótesis nula es verdadera; es decir que, el resultado es sólo una opinión parcial o sea un accidente de muestreo.

4. A continuación, se debe definir un estadístico que permita contrastar la hipótesis nula a un nivel de confianza de 0,95. Dado que no se conoce el desvío de la variable en la población y que la única información con que se cuenta es la muestral y el tamaño de muestra es menor a 30, el estadístico a utilizar es:

$$\mathrm{t} = \frac{\overline{\mathrm{X}} - \mu}{\mathrm{S}_{\overline{\mathrm{X}}}} = \frac{22 - 25}{\sqrt{45}/\sqrt{20}} = -2{,}00$$

donde, t es un estadístico con distribución t de Student y refleja, en unidades tipificadas, en cuánto difiere la media de la muestra $\bar{X}$ de la media de la población μ**.**

El valor teórico obtenido de una tabla de distribución t de Student para 19 $(n-1)$ grados de libertad y para un valor α de 0,05, es $\mathrm{t} = 2{,}093$

Luego, se compara el valor empírico de este estadístico con el valor teórico

5. La regla de decisión es, acepte la hipótesis nula, al nivel de confianza establecido, si el valor empírico de t es, en valor absoluto, menor al valor teórico. Caso contrario se debe rechazar.

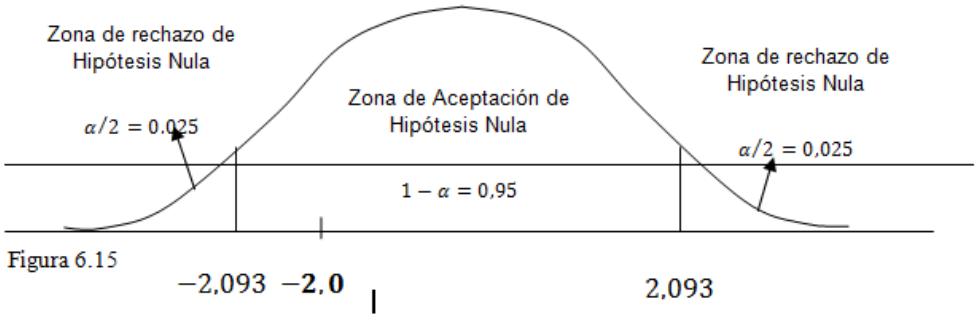


Figura 6.15

En este caso, se debe aceptar la hipótesis nula, es decir que no basta la evidencia de la muestra para decir que la población tiene una concentración de enzima diferente a 25.

Ejemplo 3.20. El fabricante de un medicamento afirma que el mismo es efectivo en el 90% de los casos o más en que se aplicó para aliviar una alergia. En una muestra de 200 pacientes se encontró que produjo alivio en 160 casos. ¿Se justifica la afirmación del fabricante a un nivel del 1%?

Este caso se corresponde con una dócima lateral derecha donde:

$H_0: \pi \leq 90$

$H_1: \pi > 90$

En la muestra el estimador es:

$$p = \frac{160}{200} = 0,80$$

Dado que la muestra es grande, $n = 200$, el estadístico a utilizar es una $Z \sim N(0,1)$. Al nivel de confianza de 0,99, el valor crítico de la distribución normal es 2,33. Este divide la zona de aceptación de la zona de rechazo de la hipótesis nula.

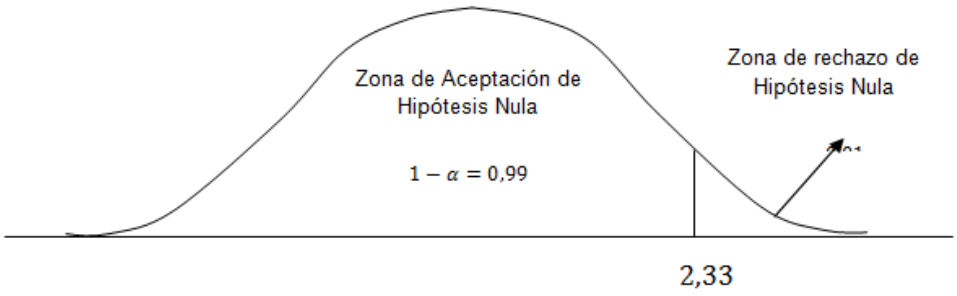


Figura 3.14

Para realizar el contraste se debe calcular el valor empírico del estadístico Z.

$$Z^* = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0{,}80-0{,}90}{\sqrt{\frac{0{,}90*0{,}90}{200}}} = \frac{-0{,}1}{0{,}02121} = -4{,}7148$$

El valor empírico cae en la zona de aceptación de la hipótesis nula. Esto significa que la efectividad es menor a 0,90, por lo tanto el fabricante no tiene razón.

También es válido calcular el intervalo de confianza para la hipótesis nula y ver a qué región pertenece el valor muestral. El conjunto de hipótesis se mantiene:

$H_0: \pi \leq 90$

$H_1: \pi > 90$

Debe calcularse ahora el límite superior del intervalo

$$L_s = \pi + Z_1\sqrt{\frac{\pi(1-\pi)}{n}} = 0{,}90 + 2{,}33 * 0{,}0212 = 0{,}949$$

El valor de 0,949, para el límite superior del intervalo, separa la zona de aceptación de la zona de rechazo de la hipótesis nula.

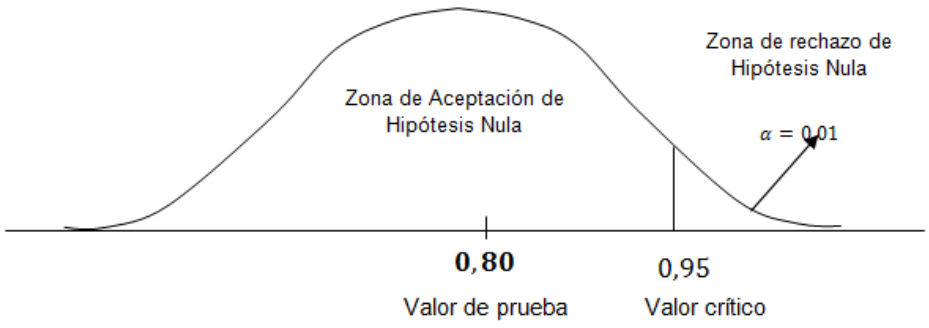


Figura 3.15

El valor empírico, o valor de prueba:

$p = \frac{160}{200} = 0{,}80$

cae en la zona de aceptación de la hipótesis nula.

Ejemplo 3.21. Al asignar subsidios federales entre distintas localidades, una determinada ciudad se clasificó como área de renta elevada. Las autoridades locales se mostraron disconformes y presentaron pruebas, basada en una muestra aleatoria de 25 familias, de que la renta familiar media de la ciudad era de $7145 y se encontraba en niveles similares a la media nacional. Un estudio a escala nacional indicó que la renta familiar media era de $6500 y la desviación estándar $920.

¿Son consistentes los resultados muestrales presentados por las autoridades locales con su afirmación de que la ciudad bajo análisis no es más próspera que las demás ciudades similares del país?.

1. el parámetro a docimar es μ, representa la renta familiar promedio

2. la población es normal $\mu = 6500$

$\sigma = 920$

3. $\mathrm{H}_0: \mu = 6500$

$\mathrm{H}_1: \mu \neq 6500$

4. Se conoce el desvío de la variable en la población, aun cuando la muestra es pequeña, el estadístico se distribuye normal

$$\mathrm{K}(\overline{\mathrm{X}}, \mu) = \mathrm{Z} = \frac{\overline{\mathrm{X}} - \mu}{\sigma_{\overline{\mathrm{X}}}}$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{920}{\sqrt{25}} = 184$$

5. El nivel de confianza $1 - \alpha = 0{,}95$. Con este nivel de confianza y la distribución normal, la zona de aceptación de la hipótesis nula se encuentra entre los valores críticos –1,96 y 1,96

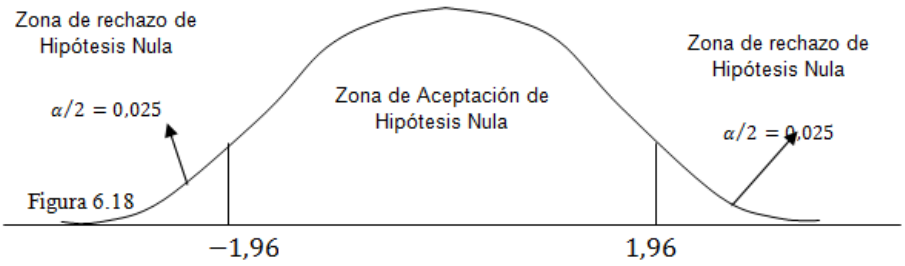


Figura 3.16

6. El valor empírico del estadístico Z es

$$Z^* = \frac{\overline{\mathrm{X}}^* - 6500}{184} = \frac{7145 - 6500}{184} = 3{,}51$$

7. Z^* es mayor a 1,96, se rechaza la H_0 y se concluye que los ingresos medios de esta ciudad son diferentes al de otras ciudades similares.

También es válido en el punto 6 calcular los límites del intervalo de confianza

$$\overline{X}^{*} = \pm Z\frac{\sigma}{\sqrt{n}} + 6500 = \pm 1{,}96 * 184 + 6500$$

$$\overline{X}_1{}^{*} = 6139{,}36 \qquad \overline{X}_2{}^{*} = 6860{,}64$$

El valor muestral $\overline{X} = 7145$ cae en la zona de rechazo de la hipótesis nula por lo que la afirmación de las autoridades locales no es cierta.

OBSERVACIÓN. Gomez Villegas (2005) comenta que Fisher y el binomio Neyman y Egon Pearson fueron los que contribuyeron a la creación y desarrollo de la estimación mediante regiones de confianza.

Ronald Fisher nació el 17 de febrero de 1890 en East Finchley (Londres), en 1909 obtuvo una beca para el Caius College de Cambridge donde se diplomó en 1913. Allí estudió matemática, mecánica, biometría y genética. En 1912 publicó el primero de sus 395 artículos en el que introduce un criterio para ajustar curvas de frecuencia, siendo el germen del método de máxima verosimilitud; apoyado en la función de verosimilitud, obtiene la distribución fiducial que constituye su aporte a la determinación de las regiones de confianza. Un trabajo publicado en 1915 sobre la distribución del coeficiente de correlación le dio fama en la comunidad científica y lo puso en contacto con Student. A la jubilación de Karl Pearson en 1933, se ocupa del Departamento de Eugenesia en el University College. Fue catedrático de genética en Cambridge, entre 1943 y 1957, año en que se traslada a Adelaida (Australia), donde muere en 1962. En los años 20 desarrolla el diseño de experimentos y la teoría de la estimación estadística. Estableció la distinción entre muestra y población e introdujo los conceptos de suficiencia y anciliaridad, definió la cantidad de información que lleva su nombre, desarrolló el método de estimación de la máxima verosimilitud y obtuvo las propiedades asintóticas del mismo.

La abundante producción científica no le impidió participar en varias instituciones estadísticas; fue presidente de la Royal

Statistical Society, de la Sociedad de Biometría y presidente del Instituto Internacional de Estadística. Además de esto, tuvo tiempo para polemizar con algunos colegas.

Jerzy Neyman nació el 16 de abril de 1894 en Bendery, Rusia, en el seno de una familia de origen polaco. Durante la revolución rusa de 1917 a 1921 su vida corre peligro por lo que se traslada a Polonia donde obtiene un doctorado en 1923. En 1924 obtiene una beca para trabajar en el University College de Londres bajo la dirección de Karl Pearson, allí coincide con Fisher y Egon Pearson, hijo de Karl. En colaboración con Egon Pearson y Fisher desarrolla la teoría de contrastes de hipótesis y de la estimación por regiones de confianza hasta que choca con este último. En 1934, Egon Pearson le propone un puesto de profesor en el University College que ocupa hasta 1938 cuando se traslada a la Universidad de California en Berkeley, para crear el Departamento de Estadística. Muere en Berkeley el 5 de agosto de 1981. Se interesó en teoría de la probabilidad, física estadística y procesos estocásticos; quizás su contribución estadística más importante fue la creación de los tests estadísticos -en colaboración con Egon Pearson- y la elaboración de la moderna teoría de muestras.

La ruptura entre Fisher y Neyman se produce por las diferencias entre la aproximación fiducial y la de regiones de confianza. Neyman nunca respondió a las críticas, salvo en 1961 cuando celebró las bodas de plata de su disidencia con Fisher.

Ross (2007) considera que el concepto de nivel de significación se debió originariamente al estadístico inglés Ronald Fisher, quien formuló el concepto de hipótesis nula como aquella que uno intenta desacreditar. En palabras de Fisher *"Puede decirse que todos los experimentos se diseñan para poder asignar una probabilidad al hecho de que los resultados se opongan a la hipótesis nula"*. La idea de la hipótesis alternativa se debe a los trabajos conjuntos del estadístico de origen polaco Jerzy Neyman y de su habitual colaborador Egon Pearson (Hijo de Karl). Fisher no aceptó la idea de especificar una hipótesis alternativa, considerando que en la mayoría de las aplicaciones

científicas no era posible especificar las alternativas; esto dio origen a una gran disputa entre Fisher, por un lado, y Neyman-Pearson, por el otro. Debido tanto al temperamento de Fisher, al que no le gustaba demasiado explicar las cosas, como al hecho de que éste ya mantenía una discusión previa con Neyman sobre los beneficios relativos de los estimadores por intervalos de confianza, propuestos por Neyman, frente a los estimadores de confianza fiduciarios (hoy en desuso) propuestos por Fisher, la disputa se convirtió en extremadamente personal y mordaz. En una ocasión, Fisher calificó la posición de Neyman como "terrorífica para la liberta intelectual en el mundo occidental".

Fisher es famoso por sus disputas científicas, también mantuvo un acalorado debate con Karl Pearson acerca de los méritos de dos procedimientos diferentes para obtener estimadores puntuales, conocidos como el método de los momentos y el método de la máximaverosimilitud, respectivamente. Fisher, que fue el fundador del área de la genética de poblaciones, también discutió con Sewell Wright, otro influyente genetista de poblaciones, acerca del papel desempeñado por el azar en la determinación de las frecuencias de genes futuras. Curiosamente, Wrigth mantuvo que la causalidad era el factor clave en los proceso de evolución a largo plazo.

Error tipo 2 y potencia de la prueba

El contraste de hipótesis finaliza en la aceptación o rechazo de la hipótesis nula, donde se observan cuatro escenarios posibles. Si el resultado del contraste lleva a aceptar la hipótesis nula siendo cierta, no hay error; como tampoco lo hay si se rechaza la hipótesis nula siendo la alternativa la cierta. Pero, cuando se rechaza la hipótesis nula siendo cierta o se acepta la nula siendo falsa, se está en presencia de los errores tipo 1 y tipo 2. La Figura 3.17 ilustra los cuatro escenarios.

Elección	H_0 cierta	H_1 cierta
H_0	No hay error (verdadero positivo)	Error de tipo II (β o falso negativo)
H_1	Error de tipo I (α o falso positivo)	No hay error (verdadero negativo)

Figura 3.17. Escenarios posibles en el contraste de hipótesis

El error de tipo I está determinada por el valor de significación α asociado a la prueba y el error tipo II se denota por β; cada uno mide:

El error tipo I la $P\left(Rechazar\ H_0 /_{H_0\ es\ cierta}\right) = \alpha$

El error tipo II la $P\left(Aceptar\ H_0 /_{H_0\ es\ falsa}\right) = \beta$

Se comete un error Tipo II al no rechazar una hipótesis nula falsa. Con el fin de calcular la probabilidad de que esto ocurra es útil considerar la probabilidad como el área bajo la distribución muestral, con base en la media verdadera de la población superpuesta a la distribución muestral supuesta en la hipótesis nula, y en la región de no rechazo. Esto se ilustra en la Figura 3.18.

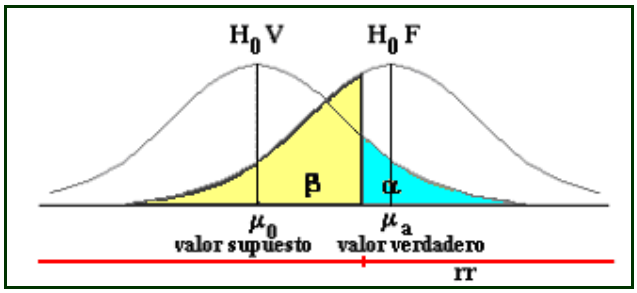


Figura 3.18 Error Tipo I y error Tipo II. Fuente: Marques Dos Santos, UNAM 2005

La probabilidad del error Tipo II se calcula asumiendo que la hipótesis nula es falsa, ya que ésta se define como *la probabilidad de no rechazar una hipótesis nula falsa*. El procedimiento para calcular el error Tipo II, para un valor específico de μ supuesto en H_0 es el siguiente:

1) Establecer la región de no rechazo para H_0, utilizando la media planteada en la H_0 y los datos del problema, determinando los valores de los límites entre los que se encuentra el valor de la H_0 (o el valor del límite por encima o por debajo del cual la H_0 es cierta).

2) Identificar los puntos críticos correspondientes a α en términos de la distribución de probabilidad utilizada.

3) Establecer la región de aceptación para la media alternativa (correspondiente a H_1 verdadera o H_0 falsa) como se ilustra en la Figura 3.19, determinando los valores críticos.

4) Calcular la probabilidad β.

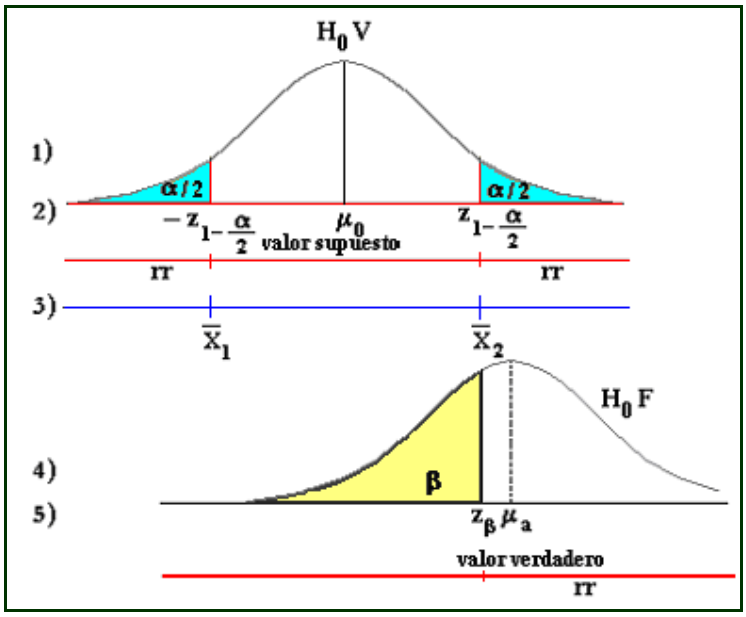


Figura 3.19 Pasos para calcular la probabilidad de Error Tipo II. Fuente: Marques Dos Santos, UNAM 2005

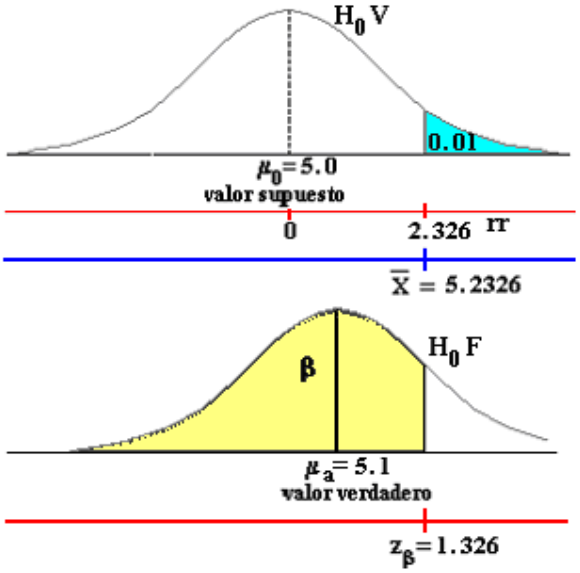
Ejemplo 3.22: En cierto estudio de $n = 36$ casos se encuentra que $\bar{X} = 5{,}22$. Se sabe que $\sigma = 0{,}6$, el nivel de significación se fijó en $\alpha = 0{,}01$ y la hipótesis nula es $H_0: \mu \leq 5{,}0$. ¿Cuál es la probabilidad del error Tipo II si el límite máximo verdadero es de 5.1?

Las hipótesis planteadas son:

$$H_0: \mu \leq 5{,}0$$

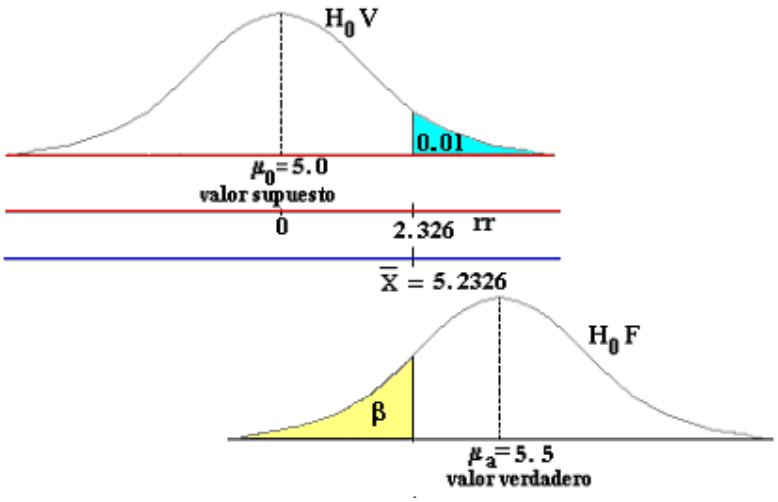
$$H_1: \mu \leq 5{,}0$$

El cálculo del error tipo II se realiza haciendo:



1. Intervalo de confianza para la $\boldsymbol{H_0}$ verdadera $(\boldsymbol{H_0V})$
2. valor crítico de la distribución de probabilidad asociada a $\alpha = 0{,}01$ $(z = 2{,}326)$
3. Intervalo de confianza para la media muestral, $\bar{X} = 5{,}1$ $(H_0\boldsymbol{F})$. Límite superior $\bar{X} = 5{,}2326$.
4. valor crítico de la distribución de probabilidad asociada a la hipótesis nula falsa $(H_0\boldsymbol{F})$ $(z = 1{,}326)$
6. Hallar el valor de β: $P(z < 1{,}326) = 0{,}9082$.

¿Cuál es la probabilidad del error tipo II si el límite máximo verdadero es de 5,5?



1. Intervalo de confianza para la $\boldsymbol{H_0V}$. Límite superior $\bar{X} = \mathbf{5{,}23}$.
2. valor crítico de la distribución de probabilidad en $\alpha = 0{,}01$ $(z = 2{,}326)$
3. valor crítico de la distribución de probabilidad para $H_0\boldsymbol{F}$ $(z = -2{,}674)$

4. Valor de β: $P(z < -2{,}674) = 0{,}0035$.

La potencia de la prueba es la probabilidad de rechazar la H_0 cuando es falsa:

$$P\left(Rechaza\ H_0 /_{H_0\ es\ falsa}\right) = 1 - \beta$$

Cuando es necesario diseñar un contraste de hipótesis, es deseable hacerlo de tal manera que las probabilidades de ambos tipos de error fueran tan pequeñas como sea posible. Sin embargo, con una muestra determinada, disminuir la probabilidad del error de tipo I conduce a incrementar la probabilidad del error de tipo II.

Usualmente, se diseñan los contrastes de tal manera que el nivel de significación no supere el 0,10. El recurso para aumentar la potencia del contraste, esto es, disminuir β la probabilidad de error de tipo II, es aumentar el tamaño de la muestra, lo que en la práctica signifca un incremento de los costos del estudio que se quiere realizar.

4. DISEÑO DE FUENTES DE INFORMACIÓN

El investigador debe distinguir dos tipos de fuentes de información que puede aplicar a su trabajo econométrico; las fuentes de información primaria y las fuentes de información secundaria. Las primeras justifican el estudio de las técnicas de muestreo. Hasta aquí el investigador conoce qué problema debe investigar y cómo plantear su tabla de datos. Generalmente, el investigador posee datos de fuentes de información secundaria, que debe recolectar, procesar y organizar como se verá en el próximo cuaderno. Si el investigador no cuenta con esta clase de información deberá realizar un relevamiento aplicando los métodos y tipos de muestreo que se desarrollan en este punto.

4.1. Elementos de muestreo

El diseño de muestreo es una de las formas que el investigador posee para realizar la investigación econométrica por medio de fuentes de información primaria. Otras formas de llevar a cabo el diseño de estas fuentes son:

- la observación directa, examinar una situación sin modificarla;
- la experimentación, construir una situación controlada por el investigador;
- el estudio de huellas o rastros, observación diferida de determinadas consecuencias del fenómeno analizado.

Estas formas de diseño no son taxativas, existen otras como las historias de vida, el análisis de contenido, medida de actitudes, evaluación de programas y simulación por computadora.

Muestreo aparece en el tercer paso de una investigación estadística y es aquella teoría que establece los procedimientos que permiten generalizar sobre la población a través del estudio de una parte de la misma; es decir, a través del estudio de la muestra.

Cada observación, o elemento tomado de la población, contiene cierta cantidad de información acerca del parámetro o parámetros de interés. Ya que la información cuesta dinero, el investigador debe determinar cuánta información debe comprar: poca información impedirá realizar buenas estimaciones; mientras que, mucha información ocasiona un despilfarro de dinero. La cantidad de información depende del número de elementos muestreados y de la cantidad de variación de los datos, ambos fenómenos pueden ser controlados a través del diseño de la encuesta y el tamaño de la muestra.

Esta introducción a los métodos estadísticos para la selección de muestras aleatorias en marcos poblacionales determinados, no se dedica al estudio de métodos específicos de muestreo; sino que trata de elaborar cuáles serán los objetivos de la aplicación de metodologías estadísticas a problemas de investigación. Constituyen las respuestas al porqué, cómo y dónde, algunas razones y motivaciones para usar buenos métodos de muestreo, una visión panorámica de los problemas básicos y de los métodos para resolverlos, y una indicación de la manera que el concepto de muestreo de poblaciones cabe dentro de los métodos de muestra y dentro de la búsqueda general de conocimiento científico.

El diseño de muestra tiene dos aspectos: un proceso de selección, que consiste en la regla y operaciones mediante las cuales se incluyen en la muestra algunos miembros de la población, y un proceso de estimación para calcular los estadísticos de la muestra, que son estimaciones muestrales de los parámetros (valores) de la población.

En la teoría del muestreo son importantes las siguientes definiciones:

Elemento: objeto sobre el cual se toman las mediciones.

Población: conjunto de elementos acerca de los cuales se desea hacer alguna inferencia, es el universo de referencia.

Unidades de muestreo: conjunto no superpuesto de elementos de la población que cubren la población completa. Si bien en el estudio se necesita encuestar individuos, es cierto que un hogar significa un conjunto de individuos (es decir un conjunto de elementos) y el proceso de seleccionar hogares, y dentro de los hogares seleccionar el elemento, puede resultar más eficiente siempre y cuando la persona no sea encuestada dos veces. Es posible que el número de elementos y el número de unidades de muestreo coincidan, esto es así cuando se muestrean individuos en lugar de hogares.

Marco: es una lista de unidades de muestreo. La selección del elemento (individuo) puede hacerse directamente del marco, esto es si se poseen listas de individuos. También pueden darse marcos múltiples: primero seleccionando viviendas y dentro de las viviendas individuos.

Muestra: conjunto de unidades seleccionadas de un marco o de varios marcos. De la muestra se obtendrán los datos objeto de la investigación que se utilizará para describir la población y realizar estimaciones sobre ella.

Ejemplo 4.1. Una importante firma industrial alimenticia quiere conocer la proporción de habitantes de cierta población que han consumido una nueva línea de polvos para helados de reciente aparición en el mercado.

Elemento será la persona que habita en esa población y que fue seleccionada por algún método de muestreo para que respondiera acerca de su conocimiento o no del producto.

Población son todos los habitantes de la población que superen un cierto límite de edad a definir por el investigador.

Perez Lopez (2005) manifiesta que *"Al hablar de métodos de muestreo nos referimos al conjunto de técnicas estadísticas que estudian la forma de seleccionar una muestra lo suficientemente representativa de una población cuya información permita inferir las propiedades o características de toda la población cometiendo un error medible y acotable. A partir de la muestra, seleccionada mediante un determinado método de muestreo, se estiman las características poblacionales (media, total, proporción, etc) con un error cuantificable y controlable. Las estimaciones se realizan a través de funciones matemáticas de la muestra, denominadas estimadores, que se convierten en variables aleatorias al considerar la variabilidad de las muestras. Los errores se cuantifican mediante varianzas, desviaciones típicas o errores cuadráticos medios de los estimadores, que miden la precisión de estos. La metodología que permite inferir resultados, predicciones y generalizaciones sobre la población estadística, basándose en la información contenida en las muestras representativas previamente elegidas por métodos de muestreo formales, se denomina inferencia estadística"*.

Selección de Muestras

El muestreo estudia los métodos para seleccionar y observar una parte de la población con el fin de hacer inferencias acerca de toda la población. Una muestra puede tener varias ventajas sobre un censo completo:

a) economía;
b) rapidez y oportunidad;
c) posibilidad de hacerse (si la observación es destructiva, el empleo de un censo no es práctico);
d) calidad y precisión (en algunas situaciones, no hay dinero suficiente para pagar el personal adiestrado y los supervisores necesarios para realizar un buen censo, o aun para obtener una muestra grande).

Los censos completos poseen ventajas especiales en algunas situaciones:

a) se pueden obtener datos para unidades pequeñas;
b) la aceptación pública es más fácil de alcanzar para datos completos;
c) la colaboración y la respuesta del público se pueden obtener más fácilmente.

Criterios del diseño de la muestra

1) Orientación hacia la meta. El diseño completo, tanto en la selección como en la estimación, debe orientarse a los objetivos de la investigación, hechos a la medida del diseño de la encuesta y ajustados a las condiciones de la encuesta.

2) La medibilidad es una característica de los diseños que permite calcular, a partir de la propia muestra, estimaciones válidas o aproximaciones de su variabilidad de muestreo. Esto se suele expresar en las encuestas con los errores estándares, pero a veces, pueden utilizarse otras expresiones de la función de verosimilitud o de la distribución de muestreo. Esta es la base necesaria para la inferencia estadística que sirve como puente, científico y objetivo, entre el resultado de la muestra y el valor desconocido de la población.

3) La practicidad se refiere a los problemas que deben resolverse para llevar a cabo el diseño esencialmente como se propuso. Una muestra probabilística no puede crearse por suposición, ni estará dada, como sucede en los problemas teóricos. El método de los muestreadores de cuota a sus entrevistadores: "vayan y obtengan una muestra aleatoria", es sumamente impráctico; ni el entrevistador ni el que lo envía pueden hacerlo. Se requiere de cuidado para traducir el modelo de selección teórico a un conjunto de instrucciones de oficina y campo. Estas instrucciones deben ser simples, claras, prácticas y completas. Por ejemplo, para identificar un segmento de muestra, no se le debe pedir al entrevistador que localice una línea marcada arbitrariamente en un mapa; sus deberes deben confinarse a localizar calles y direcciones.

4) La economía se refiere a cumplir los objetivos de la encuesta con un costo mínimo y al grado en que se alcanza este objetivo.

En líneas generales, hay dos formas de obtener una muestra: de manera informal y casual, o bien, de manera probabilística.

4.2. Muestras informales y casuales

Las muestras no probabilísticas constituyen un problema en la inferencia pues no hay manera de estimar qué tan representativas son esas muestras seleccionadas. Los procedimientos usados en este tipo de muestreo son:

- muestras de juicio: el entrevistador selecciona a cualquier sujeto que desee.
- muestras de cuotas: son muestras de juicios pero con previa asignación de cuotas por sexo, edad, clase social, raza, entre otros, que tratan de simular características conocidas de la población.
- trozo de pastel: en este caso el entrevistador no interviene en el proceso de selección pues consiste en una autoselección, personas que responden a un cupón, que concurren a un centro de exhibición, el público de un teatro en particular, por ejemplo.

Los peatones pueden ser interrogados en cuanto a sus opiniones de un nuevo producto. Si la respuesta de todos en la población es uniforme, todos ellos lo odian o lo aman, tal enfoque puede ser satisfactorio.

4.3. Muestras Probabilística

Todos los miembros de la población tienen una probabilidad conocida de estar en la muestra. Una muestra probabilística tiene las ventajas de:

- permitir al investigador demostrar la representatividad de la muestra.
- permitir un planteamiento explícito en cuanto a la cantidad de variación que será introducida, porque se usa una muestra en lugar de un censo de la población.
- hacer posible la identificación más explícita de las probables desviaciones.

El objetivo de un muestreo probabilístico es hacer una inferencia acerca de la población con base en la información contenida en la muestra. Existen dos factores que pueden afectar la información contenida en la muestra*:*

- El tamaño de la muestra
- La cantidad de variación en los datos (que puede ser controlada por el *método* de selección de una muestra)

Las unidades de muestreo contienen los elementos y se usan para seleccionarlos en la muestra. En el muestreo de elementos, cada unidad de muestreo contiene solamente un elemento; pero en el muestreo de conglomerados cualquier unidad de muestreo, llamada conglomerado, puede contener varios elementos.

Ejemplo 4.2. Una muestra de estudiantes se puede obtener de una muestra de aulas; o una muestra de viviendas de una muestra de manzanas de la ciudad.

Una misma encuesta puede usar diferentes clases de unidades de muestreo; en muestreo polietápico se usa una jerarquía de unidades de muestreo o conglomerados, de manera que el elemento pertenezca únicamente a una unidad de muestreo en cada etapa.

Ejemplo 4.3. Puede tomarse una muestra de los habitantes de una Región al seleccionar sucesivamente los municipios, las manzanas, las viviendas y, finalmente, las personas.

Las unidades de listado se usan para identificar y seleccionar unidades de muestreo a partir de listas. A veces se necesitan procedimientos detallados para convertir listados en unidades de muestreo, como por ejemplo, para convertir un listado de direcciones en viviendas y hogares. Los problemas pueden ser serios si los elementos no se identifican unívocamente con los listados. Por ejemplo, una muestra de familias tomada de los listados de teléfonos puede involucrar serias dificultades.

Entre los métodos de selección de una muestra probabilística se encuentran:

1) Muestreo aleatorio simple
2) Muestreo sistemático
3) Muestreo estratificado
4) Muestreo por conglomerados
5) Muestreo por etapas múltiples

Muestreo Aleatorio Simple

El muestreo aleatorio simple es un enfoque en el cual cada miembro de la población, y por tanto cada muestra posible, tiene una probabilidad igual de ser seleccionado.

Las muestras aleatorias simples pueden ser seleccionadas mediante el uso de una tabla de números aleatorios. La manera de hacer inferencias es estimar ciertos parámetros de la población utilizando la información de la muestra.

Frecuentemente, el objetivo es estimar una media poblacional (parámetro μ) o un total poblacional (parámetro τ); otras veces se requiere estimar la proporción poblacional (parámetro π).

Para lograr la estimación se usa el promedio muestral (estadístico $\hat{\mu}$), el estimador del total poblacional (estadístico $\hat{\tau}$) y la proporción muestral (estadístico $\hat{\pi}$ o p), donde:

$$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}; \qquad \hat{\tau} = N\hat{\mu};$$

siendo,

N, tamaño de la población

n, tamaño de la muestra

x_i, datos muestrales, i = 1, ... , n

En el muestreo aleatorio simple, el tamaño de la muestra -es decir, la cantidad de observaciones necesarias para estimar un parámetro poblacional con un límite para el error de estimación (asumir un riesgo determinado de cometer dicho error)- viene dado por

$$n = \frac{z^2\sigma^2}{e^2}$$

Muestreo Sistemático

Consiste en esparcir sistemáticamente la muestra a lo largo de la lista de miembros de la población. Si la población tiene 10.000 individuos y se desea un tamaño de muestra de 1.000, cada décima persona es seleccionada para la muestra.

Aunque en casi todos los ejemplos prácticos tal procedimiento generaría una muestra equivalente a una muestra aleatoria simple, el

investigador debe estar consciente de las regularidades dentro de la lista.

Se debe hallar primero la frecuencia de extracción de elementos, es decir, cada cuántos elementos se extrae uno. Esto se logra haciendo

$$\frac{N}{n} = K$$

Para determinar cuál es el primer elemento de la muestra, se selecciona de una tabla de números aleatorios un valor inferior a K. A este elemento se lo denomina arranque aleatorio (a). El segundo individuo será el a+K, el tercero a+K+K y así sucesivamente.

La ventaja de este método de selección es su practicidad; la desventaja surge a partir de la determinación de los elementos *a* y K, donde algunas unidades de observación pasan a tener probabilidad cierta de ser seleccionadas y otras probabilidad nula.

Muestreo Estratificado.

En el muestreo aleatorio simple, una muestra aleatoria se selecciona de una lista, o de un marco muestral, que representa a la población.

Al desarrollar un plan de muestreo, es aconsejable buscar subgrupos naturales que sean más homogéneos que la población total. Tales subgrupos se denominan "estratos".

Este tipo de muestreo es conveniente cuando la población o universo puede ser dividido en categorías estratos o grupos que reúnen cierto interés analítico y que por razones teóricas y empíricas presentan diferencias entre ellos.

La estratificación puede producir un límite más pequeño para el error de estimación que el que se generaría por una muestra aleatoria del mismo tamaño. Este resultado es particularmente cierto si las mediciones dentro de los estratos son homogéneas.

El costo por observación en la encuesta puede ser reducido mediante la estratificación de los elementos de la población en grupos convenientes.

Se pueden obtener estimaciones de los parámetros poblacionales para subgrupos de la población. Los subgrupos deben ser entonces estratos identificables.

Ejemplo 4.4. Se necesita información sobre las actitudes de los estudiantes hacia una nueva instalación atlética dentro de la Universidad. Se conoce que existen tres grupos de estudiantes con características diferenciadas:

A) los que viven en residencias estudiantiles tienen actitudes muy homogéneas hacia la instalación propuesta, la variación o la varianza en sus actitudes es muy pequeña.

B) residentes en la ciudad, son menos homogéneos.

C) residentes fuera de la ciudad, varían ampliamente en sus opiniones.

En tal situación, en lugar de permitir que la muestra provenga de la totalidad de los tres grupos aleatoriamente, será más prudente tomar un menor número de miembros del grupo de residentes y extraer más del grupo ajeno al campo. Se particiona la lista de los estudiantes en los tres grupos y se extrae una muestra aleatoria simple de cada uno de los tres grupos.

Una muestra estratificada puede ser

Proporcional, donde la fracción de muestreo es igual en cada estrato de la muestra que la existente en la población

No proporcional

La estimación de la media de la población, en el muestreo estratificado, es un promedio ponderado de las medias de las muestras encontradas en cada estrato:

$$\mu_X = \sum_i \pi_i \, \bar{x}_i$$

donde, $\bar{x}_i$ = la media de la muestra para el estrato i

π_i = la proporción de la población en el estrato i

Ejemplo 4.5. El propósito de la investigación es estudiar el rendimiento escolar según sea su extracción de clase social. Para ello, el investigador se sitúa en una escuela a la cual concurren 500 alumnos y de los cuales le informan la composición de clase social en la Figura 4.1.

Población		Muestra	
Clases sociales	Elementos	Proporcional	No proporcional
Alta	50	5	25
Media	300	30	25
Baja	150	15	25
Total	500	50	75

Figura 4.1. Composición de clase social

Si se quiere que el tamaño de la muestra sea del 10% de la población se aplica esa fracción de muestreo a cada estrato, dando lugar a un muestreo proporcional. Ahora bien en el estrato Clase Alta solo se tienen 5 casos y, puede ocurrir, que no alcancen para realizar ciertas estimaciones. Si se establece que se necesitan 25 casos en cada clase para poder realizar el estudio se está en presencia del muestreo no proporcional, donde el tamaño muestral es de 75 y los resultados a obtener dentro de cada estrato se deben ponderar por el peso del estrato en la población. En el caso del muestreo proporcional esto no ocurre porque es autoponderado.

¿Cómo se determina la mejor aplicación del presupuesto de muestreo a diversos estratos?

Este problema clásico de muestreo fue solucionado en 1935 por Jerzy Neyman, a partir de la siguiente expresión:

$$n_i = \frac{\pi_i \sigma_i / \sqrt{c_i}}{\sum_i \pi_i \sigma_i / \sqrt{c_i}} n$$

donde,

n_i = el tamaño de la muestra para el estrato i
π_i = la proporción de la población en el estrato i
σ_i = la desviación estándar de la población en el estrato i
c_i = el costo de una entrevista en el estrato i
Σ_i = la suma a lo largo de todos los estratos
n = tamaño de la muestra

Ejemplo 4.6. La Figura 4.2 presenta información de la encuesta sobre el uso mensual de las cajeros automáticos. La población se encuentra estratificada por ingreso. El segmento de ingresos *altos* tiene la variación más alta y el costo de entrevista más alto. Los estratos de ingresos medios y bajos tienen el mismo costo de entrevista, pero difieren con respecto a la desviación estándar del uso de los procesadores bancarios.

Estrato (i)	Proporción (π_i)	Desviación estándar (σ_i)	Costo (c_i) entrevista	$\pi_i\sigma_i/\sqrt{c_i}$	n_i
Bajo	0.3	1	25	0.06	177
Medio	0.5	2	25	0.20	588
Alto	0.2	4	100	0.08	235

Figura 4.2. Uso de cajeros automáticos

Para asignar 1000 casos entre los diferentes estratos, se tiene en cuenta la $\sum(\pi_i\sigma_i/\sqrt{c_i}) = 0.34$ y se calcula:

$$n_{bajo} = \frac{\pi_i\sigma_i/\sqrt{c_i}}{\sum(\pi_i\sigma_i/\sqrt{c_i})} * 1000 = \frac{0.06}{0.34} * 1000 = 177$$

$$n_{medio} = \frac{\pi_i\sigma_i/\sqrt{c_i}}{\sum(\pi_i\sigma_i/\sqrt{c_i})} * 1000 = \frac{0.20}{0.34} * 100 = 588$$

$$n_{alto} = \frac{\pi_i\sigma_i/\sqrt{c_i}}{\sum(\pi_i\sigma_i/\sqrt{c_i})} * 1000 = \frac{0.08}{0.34} * 1000 = 235$$

Las cantidades muestrales de la última columna, presentan la división de la muestra de 1.000 personas en los tres estratos. Al estrato de

ingresos altos le corresponden 235 personas que representan el 23.5% de la muestra; sin embargo, la proporción de personas de altos ingresos en la población es de 20.0%. Si se hubiera seleccionado una muestra aleatoria simple de tamaño 1000, 200 personas pertenecerían al estrato de altos ingresos y no sería un número suficiente para estimar los parámetros en la población.

La disponibilidad presupuestaria determina, en alguna medida, el tamaño de la muestra. Este es ajustado hacia arriba hasta que alcanza el límite presupuestal, de modo que el presupuesto debe ser calculado como:

Presupuesto = $\Sigma_i c_i n_i$

Conocido el tamaño de la muestra, se determina el error muestral y se decide si es excesivo o no. La fórmula del error muestral es:

Error muestral = z $\sigma_{\bar{x}}$

donde,

$$\sigma_{\bar{x}} = \sqrt{\sum_i \frac{\pi_i^2 \sigma_i^2}{n_i}}$$

Si el error muestral es excesivo se debe ampliar el presupuesto, lo que permite tomar un tamaño de muestra mayor; de no ser posible el proyecto debe desecharse.

Muestreo De Conglomerados

En el muestreo de conglomerados, la población se divide nuevamente en subgrupos. En esta técnica se selecciona una muestra aleatoria de subgrupos y todos los miembros de los subgrupos forman parte de la muestra. Este método es útil cuando se pueden identificar aquellos subgrupos que sean representativos de la totalidad de la población.

Ejemplo 4.7. Se tomó una muestra de estudiantes universitarios de segundo año, que cursaban estadística en todas las facultades de Argentina, para estudiar la prevención del SIDA. Se contaba en ese momento con 200 comisiones de estadística, cada una de las cuales tenía, en promedio, 30 estudiantes.

El muestreo utilizado fue el de conglomerados: se seleccionaron 15 comisiones y, dentro en ellas, la totalidad de los alumnos. De modo que el tamaño de muestra fue de 450 alumnos.

Si se hubiera decidido tomar una muestra aleatoria simple reuniendo 450 alumnos en el total de 200 cursos de estadística, el costo sería significativamente mayor.

La gran pregunta, desde luego, es si los cursos son representativos de la población y la respuesta es "no necesariamente". Si los cursos de las áreas de ingresos superiores tienen diferentes opiniones acerca de la prevención del SIDA, que los cursos con estudiantes de ingresos más bajos, el supuesto que fundamenta al enfoque, no se mantendría.

La gran ventaja del muestreo de conglomerados es que su costo es más bajo. Los subgrupos o conglomerados son seleccionados de modo que, el costo para obtener la información deseada dentro del conglomerado, sea mucho más pequeño que si se obtuviera una muestra aleatoria simple.

Diseños De Etapas Múltiples

Si se usan otros diseños muestrales, la lógica para generar el tamaño óptimo de la muestra aún se mantendrá; sin embargo, la fórmula puede complicarse.

Ejemplo 4.8. En un diseño por áreas, el primer paso puede ser el de seleccionar comunidades al azar. De este modo, el procedimiento puede ser escoger porciones de radios censales, luego manzanas y finalmente familias.

En tal diseño, la expresión para determinar el error estándar de $\bar{X}$ se vuelve sumamente compleja. La situación consiste en repetir la totalidad del plan de muestreo y obtener dos, tres o cuatro estimaciones independientes de $\bar{X}$. Estas diferentes estimaciones pueden ser usadas para estimar el error estándar de $\bar{X}$.

Bajo el supuesto de que se conocen los tamaños de las poblaciones, formalmente, el diseño en etapas múltiples parte de considerar una población de unidades secundarias de tamaño N repartida en M unidades primarias. Sea k la unidad de observación perteneciente a la unidad primaria sobre la que se desea estudiar alguna característica en particular. Para $k = 1, \dots M$, N_k es el tamaño de la unidad primaria k que se supone conocido; entonces

$$\sum_{k=1}^{M} N_k = N$$

Sea Y una variable real de interés. Se denota

$\tau, \mu\ y\ \sigma^2$ el total, la media y la varianza de Y sobre la población;

$\tau_k, \mu_k\ y\ \sigma_k^2$ el total, la media y la varianza de Y sobre la unidad primaria k $(k = 1, \dots, M)$.

Se plantea la media y la varianza del total para cada unidad primaria de la siguiente manera:

$$\mu_\tau = \frac{1}{M}\sum_{k=1}^{M} \tau_k$$

$$\sigma_\tau^2 = \frac{1}{M}\sum_{k=1}^{M} (\tau_k - \mu_\tau)^2$$

Se tiene entonces que

$$\mu = \frac{M}{N}\mu_\tau$$

Donde, $\tau_k = N_k\mu_k$. Por lo tanto, estimar μ se reduce a estimar μ_τ.

Para ello se considera una muestra aleatoria simple de M de probabilidades iguales sin reemplazo de m unidades primarias. Luego, en cada unidad primaria de la muestra, $k \in M$, se toma una muestra aleatoria simple de probabilidades iguales sin reemplazo de n_k individuos.

Se denotan $\bar{Y}_k$ y S_k^2 la media y la varianza corregida de Y sobre esta muestra. Se estima entonces μ_τ mediante

$$\hat{\mu}_\tau = \frac{1}{m}\sum_{k\epsilon M}\hat{\tau}_k$$

Donde $\hat{\tau}_k = N_k\bar{y}_k$

Y se deduce el estimador de μ:

$$\hat{\mu} = \frac{M}{N}\frac{1}{m}\sum_{k\epsilon M}\hat{\tau}_k$$

Es un estimador insesgado de μ cuya varianza se estima sin sesgo mediante:

$$\hat{V}(\hat{\mu}) = \frac{M^2}{N^2}\left[S_\tau^2\left(1-\frac{m}{M}\right) + \frac{1}{Mm}\sum_{k\epsilon M}N_k^2\frac{S_k^2}{n_K}\left(1-\frac{n_k}{N_k}\right)\right]$$

Donde: $S_\tau^2 = \frac{1}{m-1}\sum_{k\epsilon M}(\hat{\tau}_k - \hat{\mu}_\tau)^2$

Ejemplo 4.9. Supóngase que se considera una población de 3.000.000 habitantes (N) repartida en 3.000 municipios (M). Sea Y el consumo de un cierto producto en diciembre de 1995. Se extrae una muestra aleatoria simple con probabilidades iguales sin remplazo de 30

municipios; luego, para la extracción de las unidades secundarias se implementan dos métodos.

Primer método: En cada uno de los municipios de la muestra se extrae una muestra aleatoria simple, de habitantes con probabilidades iguales sin remplazo, con una tasa de muestreo constante igual a 1/100. El número de unidades secundarias encuestadas, en cada municipio, es entonces aleatorio y su esperanza es igual a 10.

Segundo método. En cada uno de los municipios de la muestra se extrae una muestra aleatoria simple con probabilidades iguales sin remplazo de 10 habitantes. (Aunque es útil a los fines prácticos, este procedimiento no es el mejor, ya que no se mantiene constante la probabilidad de un municipio a otro y la probabilidad de ser elegido para un individuo es diferente).

Las columnas segunda a cuarta de la Figura 4.3, registran los resultados sobre las 30 muestras. N_k es la población total en cada municipio integrante de la muestra

$$m = 30 \quad k = 1,2,\dots,m = 1,2,\dots,30$$

$$\sum_{k=1}^{m} N_k = 400 + 400 + \cdots + 1.900 = 30.100$$

En la séptima columna, n_k representa la cantidad de elementos muestreados en cada unidad primaria (municipio), siendo la tasa de muestreo constante e igual a 1 de cada 100 unidades secundarias (personas dentro de cada unidad primaria). Esta columna es insumo para el primer método.

Las demás columnas se utilizan para el cálculo de $\hat{\mu}$ y $V(\hat{\mu})$.

U.O.	N_k	$\bar{y}_k$	s_k	$\frac{N_k \bar{y}_k}{10^3}$	$\left(\frac{N_k \bar{y}_k}{10^3}\right)^2$	n_k	1° METODO $\frac{(N_k s_k)^2}{n_k} \times \frac{1}{10^6}$	2° METODO $\frac{(N_k s_k)^2}{10} \times \frac{1}{10^6}$
1	400	750	200	300	90000	4	1600.0	640.0
2	400	800	180	320	102400	4	1296.0	518.4
3	400	750	180	280	78400	4	1296.0	518.4
4	500	850	220	425	180625	5	2420.0	1210.0
5	500	750	180	375	140625	5	1620.0	810.0
6	600	800	200	480	230400	6	2400.0	1440.0
7	600	750	180	450	202500	6	1944.0	1166.4
8	600	700	150	420	176400	6	1350.0	810.0
9	600	750	120	450	202500	6	864.0	518.4
10	700	800	200	560	313600	7	2800.0	1960.0
11	700	700	130	490	240100	7	1183.0	828.1
12	800	650	180	520	270400	8	2592.0	2073.6
13	800	650	220	520	270400	8	3872.0	3097.6
14	900	500	140	450	202500	9	1764.0	1587.6
15	900	700	170	630	396900	9	2601.0	2340.9
16	1000	650	190	650	422500	10	3610.0	3610.0
17	1000	550	140	550	302500	10	1960.0	1960.0
18	1000	650	180	650	422500	10	3240.0	3240.0
19	1100	600	190	660	435600	11	3971.0	4368.1
20	1100	650	150	715	511225	11	2475.0	2722.5
21	1200	700	210	840	705600	12	5292.0	6350.4
22	1200	700	220	840	705600	12	5808.0	6969.6
23	1300	550	160	715	511225	13	3328.0	4326.4
24	1400	600	150	840	705600	14	3150.0	4410.0
25	1500	450	140	675	455625	15	2940.0	4410.0
26	1600	550	180	880	774400	16	5184.0	8294.4
27	1700	500	160	850	722500	17	4352.0	7398.4
28	1800	550	170	990	980100	18	5202.0	9363.6
29	1900	550	180	1045	1092025	19	6156.0	11696.4
30	1900	600	200	1140	1299600	19	7600.0	14440.0
$\sum$	30100			18710	13144350	301	93870.0	113079.2
Σ/30	1003,33			623.67	438145		3129.0	3769.3

Figura 4.3. Selección de municipios

Se tiene

$$\hat{\mu} = \frac{M}{N}\frac{1}{m}\sum_{k \in M} N_k \bar{y}_k$$

A efectos de ilustrar los cálculos en la tabla, se considera $\frac{M}{N} = \frac{1}{10^3}$.

En este caso para el cálculo de $V(\hat{\mu})$, las tasas de muestreo del primer nivel $\left(1-\frac{m}{M}\right)$ y del segundo nivel $\left(1-\frac{n_k}{N_k}\right)$ son despreciables, pues se obtiene 0,99 en ambos niveles.

$$\hat{V}(\hat{\mu}) = \frac{M^2}{N^2}\left[\frac{s_\tau^2}{m} + \frac{1}{Mm}\sum\left(N_k\right)^2 \frac{s_k^2}{n_k}\right]$$

donde

$$\frac{s_\tau^2}{m} = \frac{1}{m-1}\left[\frac{1}{m}\sum_{k\in M}\left(N_k\bar{y}_k\right)^2 - \left(\frac{1}{m}\sum_{k\in M} N_k\bar{y}_k\right)^2\right]$$

OBSERVACIÓN. Se sabe que $S_\tau^2 = \frac{1}{m-1}\sum_{k\in M}(\hat{\tau}_k - \hat{\mu}_\tau)^2$. Entonces

$$S_\tau^2 = \frac{1}{m-1}\sum_{k=1}^{m}(\hat{\tau}_k^2 - 2\hat{\tau}_k\hat{\mu}_\tau + \hat{\mu}_\tau^2)$$

$$S_\tau^2 = \frac{1}{m-1}\sum_{k=1}^{m}\hat{\tau}_k^2 - m\hat{\mu}_k^2$$

Pero $\hat{\tau}_k = N_k\bar{y}_k$ y $\hat{\mu}_\tau = \frac{1}{m}\sum_{k=1}^{m}\hat{\tau}_k$.

Por lo tanto

$$S_\tau^2 = \frac{1}{m-1}\sum_{k=1}^{m}(N_k\bar{y}_k)^2 - m\left(\frac{1}{m}\sum_{k=1}^{m}\hat{\tau}_k\right)^2$$

$$S_\tau^2 = \frac{1}{m-1}\sum_{k=1}^{m}(N_k\bar{y}_k)^2 - m\left(\frac{1}{m}\sum_{k=1}^{m}N_k\bar{y}_k\right)^2$$

Finalmente, al dividir ambos miembros de la igualdad por m se obtiene

$$\frac{s_\tau^2}{m} = \frac{1}{m-1}\left[\frac{1}{m}\sum_{k\in M}(N_k\bar{y}_k)^2 - \left(\frac{1}{m}\sum_{k\in M}N_k\bar{y}_k\right)^2\right]$$

En este ejemplo, con $n_k = N_k/100$ para el primer método y $n_k = 10$ para el segundo, se obtienen los siguientes resultados:

Primer método	Segundo método
$\hat{\mu} = 623{,}67$	$\hat{\mu} = 623{,}67$
$\hat{V}(\hat{\mu}) = 1697{,}07$	$\hat{V}(\hat{\mu}) = 1697{,}29$

En este ejemplo se obtienen resultados similares con los dos métodos porque el segundo término en el cálculo de la varianza es despreciable. Entonces el intervalo de confianza para $\hat{\mu}$ es similar para cada uno de los métodos.

$\hat{\mu} \pm 1{,}96\sqrt{\hat{V}(\hat{\mu})}$	$623{,}67 \pm 81$	$[543; 705]$

Ejemplo 4.9. Se trata de un país de 22.200.000 habitantes distribuidos en 4000 municipios de acuerdo a la Figura 4.4. Se decide constituir una muestra de 2.220 habitantes; es decir, una tasa de muestreo de 1/10.000, utilizando, en cada categoría de municipios, un muestreo bietápico. Este consiste en una muestra de municipios, luego muestra de habitantes en los municipios extraídos, de tal manera que cada habitante tenga la misma probabilidad igual a 1/10.000 de pertenecer a la muestra.

Se puede procesar según dos métodos.

Primer Método

Se realiza un muestreo aleatorio estratificado proporcional de municipios con una tasa uniforme de 1/100, luego, en los municipios

extraídos, un muestreo aleatorio simple de un habitante para cada 100 habitantes.

Se obtiene entonces para cada muestra de municipios:

$m_1 = 30$ $m_2 = 7$ $m_3 = 3$

y, en promedio, para cada municipio de la muestra de cada categoría:

$n_1 = 10$ $n_2 = 60$ $n_3 = 500$

lo que da, como previsto, para cada categoría, un número medio de habitantes en la muestra de:

300, 420 y 1500 respectivamente

Con este método, los tamaños de las muestras de habitantes son aleatorios y el manejo del trabajo de los encuestadores se vuelve difícil. Además el número de "puntos de encuesta", que corresponden aquí al número de municipios de la muestra, es débil: solamente 40.

MUNICIPIOS DE MENOS DE 2000 HABITANTES	MUNICIPIOS DE 2000 A 10000 HABITANES	MUNICIPIOS DE MAS DE 10000 HABITANTES
3000 MUNICIPIOS 3.000.000 HABITANTES (MEDIA: 1.000 HAB.)	700 MUNICIPIOS 4.200.000 HABITANTES (MEDIA: 6.000 HAB.)	300 MUNICIPIOS 15.000.000 HABITANTES (MEDIA: 50.000 HAB.)

Figura 4.4. Distribución de municipios en la población

Segundo Método

Se decide encuestar a 10 habitantes por municipio de la muestra, lo que, cuando se toma en cuenta la duración del cuestionario, corresponde a un día de trabajo de un encuestador.

Hay pues que constituir una muestra de 222 municipios. Se constituye esta muestra de municipios con probabilidades desiguales, proporcionales al tamaño del municipio.

Se obtiene entonces, para la muestra de municipios:

$m_1 = 30$ $m_2 = 42$ $m_3 = 150$

y, como tenemos que encuestar 10 habitantes por municipio de la muestra, al final tenemos, tal como previsto:

300, 420 y 1500 habitantes encuestados en cada categoría

Con este método, se facilita la gestión del trabajo de los encuestadores y el número de puntos de encuestas crece (222).

Para un municipio, su probabilidad de pertenecer a la muestra de municipios es la misma dentro de una categoría pero distinta de una categoría a otra.

Tenemos $\pi_i = 1/100$ para la primera categoría, $\pi_i = 6/100$ para la segunda y $\pi_i = 50/100$ para la tercera.

De hecho se trató, en el ejemplo anterior, de estimar la media de Υ sobre el primer estrato; se obtuvo, con cualquiera de los dos métodos:

$\hat{\mu} = 624 \quad \hat{V}(\hat{\mu}) = 1697$

Se trata ahora de estimar μ utilizando el estimador de estratificación.

Se supone que los resultados sobre los tres estratos son los de la Figura 4.5. Se obtiene pues, para el conjunto (última columna):

$$\hat{\mu} = \sum \frac{N_k}{N} \hat{\mu}_k \qquad \hat{V}(\hat{\mu}) = \sum \left(\frac{N_k}{N} \right)^2 \hat{V}(\hat{\mu}_k)$$

	Primer estrato	Segundo estrato	Tercer estrato	Conjunto
N_k	3.000.000	4.200.000	15.000.000	22.200.000
N_k / N	30/222	42/222	150/222	1
$\hat{\mu}_k$	624	700	750	744
$\hat{V}(\hat{\mu}_k)$	1700	1600	500	317

Figura 4.5. Muestreo de municipios

Pues la estimación de μ por intervalo de confianza al 95%:

$\hat{\mu} \pm 1.96\sqrt{\hat{V}(\hat{\mu})}$ $\quad 744 \pm 35 \quad$ $[709; 779]$

Se realizó aquí un muestreo de 3 etapas. Como se trata de una estratificación a la primera etapa, al final se tiene, en cada estrato, un muestreo bietápico: una muestra de municipios, luego, en cada municipio de la muestra, una muestra de habitantes.

CASOS DE ESTUDIO, PREGUNTAS Y PROBLEMAS

Caso 1: Edades y estaturas del curso Inferencia Estadística.

El tema de estudio es la distribución de las edades y las estaturas en el curso de Inferencia Estadística del presente ciclo lectivo.

El objetivo general es conocer las medidas antropométricas de la población; específicamente se busca hallar la media de estatura y la media de edad, adicionalmente utilizar la teoría de distribuciones en el muestreo para inferir datos a la población.

Se parte de las siguientes premisas:

- la existencia de datos extremos en la muestra aumenta la dispersión de la media
- a medida que aumenta el número de muestras, la media de las muestras tiende a la media de la población

Para realizar este estudio se debe:

a) Generar una tabla de datos de n observaciones por 2 variables cuantitativas; donde. la unidad de observación sean los alumnos de Inferencia Estadística, y las variables sus respectivas edad y estatura.
b) Cada alumno debe
   - 1.seleccionar cinco muestras de tamaño cinco.
   - 2.calcular la media de cada muestra
   - 3.calcular la media de las cinco muestras
c) Generar una tabla con las medias de cada una de las cinco muestras generadas por cada alumno (el número de

observaciones de esta tabla será nx5) y calcular la media de las medias.

d) Generar una tabla con la media de cada alumno (el número de observaciones de esta tabla será n) y calcular la media de las medias.
e) En la tabla generada en a), calcular la media de la población.
f) Comparar las medias obtenidas en c), d) y e).

Caso 2: Edades y Estaturas

Mediciones realizadas en cursos de Inferencia Estadística informaron las edades y estaturas de los integrantes. La tabla reúne la información del tamaño de muestra y las medidas promedio

Inferencia Estadística	Curso año 1998 n=12		Curso año 2007 n=12	
	Edad	Estatura	Edad	Estatura
Media	20,3750	1,73125	21,8330	1,70500
Desvío	2,2500	0,09258	4,8586	0,09719

Obtiene para cada curso, con un nivel de significación del 0,05,

- el intervalo de confianza para la media
- el intervalo de confianza para el desvío

¿Hay similitudes entre las dos poblaciones? ¿Por qué? Indique si es necesario obtener información adicional

Con un nivel de significación de 0,05, pruebe la hipótesis de que la estatura promedio de los cursantes 2007 de Inferencia Estadística supera 1,71cm.

Caso 3: Actividad Económica en Río Cuarto

La actividad del comercio minorista, excluido los sectores de alimentos y bebidas y textil, comprende alrededor del 18% de las empresas de la ciudad. Una encuesta realizada por muestreo estratificado, entre 1998 y 2004, permitió monitorear la evolución de las ventas del sector; esta información se reúne en la tabla.

Los estratos se definieron en función de la venta mensual que realizara la empresa

- Hasta $100000
- Más de $100000

Evolución de ventas del sector comercio minorista								
	Empresas con ventas hasta $100.000				Empresas con ventas mayores a $100.000			
Fecha	Media	Desvío	n	N	Media	Desvío	n	N
Julio 98	10290,33	7316,15	11	1014	459803,80	428747,81	20	40
Julio 99	9235,69	7752,72	11	1023	467755,06	400134,52	20	40
Julio 00	10470,89	7211,74	11	1059	437535,36	395982,72	20	40
Julio 01	6833,18	5696,52	11	1078	244791,27	230112,77	20	40
Julio 02	16613,58	14523,07	12	1054	251042,89	234032,71	20	40
Julio 03	18865,16	18848,22	13	1128	354017,00	309265,69	22	40
Julio 04	32928,64	30721,44	13	1230	526558,14	426254,49	22	40

(*) Excluye los sectores alimentos, bebidas y textiles.

FUENTE: Indice de Evolución Económica. Programa Institucional de Investigación y Extensión. Facultad de Ciencias Económicas. Universidad Nacional de Río Cuarto.

Con un nivel de confianza de 0,90, calcula para el mes de julio de 2004

- El intervalo de confianza para la media poblacional
- El intervalo de confianza para el desvío poblacional

Estima las ventas totales del sector comercio minorista para el mes de julio de 2004.

¿Hay diferencias significativas en los niveles y la variabilidad de las ventas de julio de 2004 respecto de las registradas en julio de 1998?

Caso 4: Demanda de diario regional

Una editorial del interior del país realiza un estudio con el objetivo de estimar la demanda semanal de un nuevo diario regional. Los resultados que arroja el estudio, sobre una muestra de 400 casos, se resumen en las tablas que se adjuntan.

Región de cobertura del nuevo diario

Total de habitantes	101289
Total de hogares	27980
FUENTE: INDEC.	

Le gustaría un diario regional?

SI	87,67
NO	5,02
Ns/Nc	7,31
Total de hogares que leen	100,00
FUENTE: Encuesta de opinión. Estudios & Mercados Consultora. Diciembre 2003.	

Hábitos de los hogares hacia la lectura de diarios

Cuándo lee el diario? (en % sobre el total de hogares)	
Todos los días (5 a 7)	8,5
Algunos días (2 a 4)	21,0
Fin de semana	7,5
Algunas veces	17,25
De otra forma	0,25
Ns/Nc	1,75
No lee diarios	43,75
Total de hogares	100,00

FUENTE: Encuesta de opinión. Estudios & Mercados Consultora. Diciembre 2003.

Para leer el diario… (en % sobre el total de hogares que leen)	
Lo compra	55,25
Se lo prestan	25,11
Lo lee en su trabajo	10,96
De otra forma	8,22
Ns/Nc	0,46
Total de hogares que leen	100,00

FUENTE: Encuesta de opinión. Estudios & Mercados Consultora. Diciembre 2003.

Los directivos de la editorial necesitan que se realice una estimación por intervalo de confianza, con un nivel de significación del 0,05, del número promedio de diarios a vender por semana.

El responsable del diario afirma que editará un nuevo diario regional si tiene evidencias empíricas suficientes que le garanticen, con un nivel de significación del 0,10, que el 75% de los hogares compran habitualmente algún diario.

Caso 5: Cooperativa de Alimentos

Kinnear, T. y Taylor, J. (1993) relatan la historia de J.L, Gerente General de una Cooperativa de Alimentos. En determinado momento, el gerente se dio cuenta que había perdido contacto con los patrones de compra de los miembros de la cooperativa; simplemente, la cooperativa parecía mucho más grande ahora en comparación con los primeros años. El Gerente se preguntaba si podría hacer uso de algún tipo de dato que estuviese a su alcance con el propósito de ampliar su comprensión de los hábitos de compra de los miembros. Esperaba poder utilizar este conocimiento para planificar mejor la mezcla y el volumen de artículos que la cooperativa ofrecía.

J.L. y un pequeño grupo de voluntarios fundaron la cooperativa en 1974. Esta cooperativa había aumentado de diez miembros iniciales en enero de 1974 a 500 miembros en setiembre de 1990. La empresa estaba localizada en una vieja bodega del noroeste de Milán. Milán era una comunidad de 7500 personas localizada en el sureste de Michigan, aproximadamente a 40 millas al suroeste de Detroit. La cooperativa obtuvo sus miembros de una gran cantidad de comunidades alrededor del Milán, incluyendo Ann Arbor y Monroe.

Antecedentes. El objetivo de la cooperativa era proveer productos alimenticios de alta calidad a un precio por debajo de los que se ofrecían en los supermercados locales. Para este objetivo, la cooperativa utilizaba las cajas de empaque como estanterías, solicitaba la colaboración de los mismos compradores para que marcaran sus precios en los artículos, vendía sólo las mejores marcas del mercado y, por lo general, no ofrecía los "lujos" relacionados con los supermercados tradicionales. Para poder comprar en la cooperativa, las personas tenían que ser socios. La cuota de afiliación era de $25 anuales. Cualquier ganancia obtenida por la cooperativa durante un año,

se devolvía a los socios en forma de crédito para sus compras. J.l. estaba convencida de que los socios compraban la mayor parte de sus alimentos en la cooperativa.

Preocupaciones de J.L. durante los primeros años de la cooperativa. J.L. se sentía orgullosa de conocer a todos sus miembros. Había gastado una cantidad considerable de tiempo en la tienda y sentía que conocía lo que las personas compraban y cuánto dinero estaban gastando. A medida que creció el número de socios, sus deberes administrativos la mantenían más tiempo en la oficina. Por tanto, ya no podía decir que conocía a todos los socios, ni que tenía idea de sus patrones de gastos. Deseaba conocer mejor estos aspectos de su negocio y pensó que tal vez parte de los datos que se habían recolectado con anterioridad sobre los miembros podrían proporcionarle respuestas.

Datos disponibles

En junio de 1990 se utilizó un cuestionario para recolectar datos sobre los socios. Durante ese mes todos los miembros vinieron a la cooperativa por lo menos una vez. Por tanto existían datos sobre cada uno de los socios. Los datos consistían en las características demográficas de los miembros y en sus gastos semanales en alimentos.

Los datos se encontraban en las tarjetas que los socios habían llenado en el momento de la entrevista. J.L. tenía estas tarjetas en un archivador en su oficina. A continuación se encuentra una descripción del contenido de las tarjetas. Los valores reales de las tarjetas se hallan tabulados en las páginas siguientes.

Con el propósito de poder conocer mejor a los socios, J.L. quiere saber inicialmente el promedio semanal de gastos en alimentos, puesto que no dispone de demasiado tiempo, quiere hacer esto sin tener que mirar las quinientas tarjetas. Sin embargo, también desea asegurarse de que el promedio que calcule sea

exacto. Para esto contrata a un consultor y este le recomienda realizar una muestra probabilística sobre las 500 tarjetas que se encuentran tabuladas en la tabla. Las tarjetas tienen la siguiente información:

Descripción de las variables

A = número de identificación de la unidad familiar; 1 - 500

B = gastos reales semanales en alimentos

C = número de personas en la unidad familiar; 1 - 9

1 = una persona
2 = dos personas
3 = tres personas
4 = cuatro personas
5 = cinco personas
6 = seis personas
7 = siete personas
8 = ocho personas
9 = nueve personas

D = ingreso familiar anual real

E = educación del jefe del hogar; 1 - 5

1 = menos del 8 grado
2 = entre los grados 9 y 11
3 = con título en secundaria
4 = algunos años de educación secundaria
5 = con título universitario

F = edad actual del jefe del hogar

G = gasto semanal en alimentos, codificados en 7 categorías; 1 - 7

1 = menos de $ 15

2 = $ 15 a $ 29.99
3 = $ 30 a $ 44.99
4 = $ 45 a $ 59.99
5 = $ 60 a $ 74.99
6 = $ 75 a $ 89.99
7 = $90 ó más

A	B	C	D	E	F	G	H	I	J	K
1	12	1	2500	1	56	1	1	1	1	5
2	16.5	1	2800	1	70	2	1	1	1	6
3	18	1	2000	1	20	2	1	1	1	1
4	17	1	4500	1	60	2	1	1	2	5
5	46.5	1	8000	1	40	4	1	1	3	3
6	45	1	7000	1	51	4	1	1	3	4
7	15	1	3500	1	76	2	1	1	2	7
8	60	2	2800	1	20	5	1	1	1	1
9	15	2	2500	1	51	2	1	1	1	4
10	18	2	4000	1	32	2	1	1	2	2
11	22.5	2	5000	1	47	2	1	1	2	4
12	20	2	8000	1	35	2	1	1	3	3
13	97	2	5500	1	58	7	1	1	2	5
14	57	2	6000	1	27	4	1	1	3	2
15	39	2	3000	1	38	3	1	1	2	3
16	30	2	4000	1	40	3	1	1	2	3
17	42	2	3000	1	19	3	1	1	2	1
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
497	115	6	24000	5	36	7	2	2	5	3
498	75	7	28000	5	37	6	2	2	6	3
499	105	7	20000	5	39	7	2	2	5	3
500	75	8	33000	5	42	6	2	2	6	3

Tabla extraída de Kinnear, T. Taylor, J. Investigación de Mercados. Un enfoque aplicado. Mc.Graw Hill. 1993. Copia disponible en www.econometricos.com.ar .

H = algún niño menor de 6 años en el hogar; 1 - 2

1 = no
2 = sí

I = algún hijo entre 6 y 18 años en el hogar; 1 - 2

1 = no
2 = sí

J = ingreso anual familiar, codificado en seis categorías; 1 - 6

1 = menos de $ 3000
2 = $ 3000 - $ 5999.99
3 = $ 6000 - $ 9999.99
4 = $ 10000 - $ 14999.99
5 = $ 15000 - $ 24999.99
6 = $ 25000 ó más

K = edad del jefe del hogar, codificado en siete categorías; 1 - 7

1 = menos de 25
2 = 25 - 34
3 = 35 - 44
4 = 45 - 54
5 = 55 - 64
6 = 65 - 74
7 = 75 ó más

Se pide:

a) Seleccionar una *muestra aleatoria simple* de tamaño 50. Construir la tabla con las observaciones seleccionadas. Calcular los estadísticos que permitan describir la información.

b) Seleccionar una *muestra sistemática* de tamaño 50. Construir la tabla con las observaciones seleccionadas. Calcular los estadísticos que permitan describir la información.

c) Seleccionar una *muestra estratificada proporcional* según la presencia de niños menores de 6 años en el hogar de tamaño 25. Construir la tabla con las observaciones

seleccionadas. Calcular los estadísticos que permitan describir la información.

Seleccionar una *muestra estratificada no proporcional* según la presencia de niños menores de 6 años en el hogar donde cada estrato sea de tamaño 25. Construir la tabla con las observaciones seleccionadas. Calcular los estadísticos que permiten describir la información.

Preguntas

1. De los siguientes conceptos, ¿cuáles indican muestras? y ¿cuáles población?

a) grupos de medidas llamados *parámetros*
b) uso de *inferencia estadística*
c) hacer un *censo*
d) juzgar la calidad de un embarque de fruta inspeccionando varias de las cajas
e) universo
f) grupo de medidas denominadas *estadísticas*
g) aplicación de conceptos de probabilidad
h) inspección de todos los artículos que se fabrican

2. ¿Porqué es válida la siguiente igualdad?

$$\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

siendo: $\sum_{i=1}^{n} x_i^2$ la suma del cuadrado de cada observación y $(\sum_{i=1}^{n} x_i)^2$ el cuadrado de la suma total

Demostrar

3. ¿Porqué la proporción de 0,5 en una variable dicotómica se corresponde con una varianza máxima? Demuestre.

Problemas

1. Calcular las siguientes probabilidades

$P(\chi^2_{12} \leq 26.2)$ $P(\chi^2_{12} > 6.3)$

$P(5.23 \leq \chi^2_{12} \leq 14.8)$ $P(\chi^2_{12} < x_0) = 0.90$

2. Dada una variable G que se distribuye como una t de Student con 15 grados de libertad, calcular

$P(G < 1.341)$ $P(-2.131 < G < 2.131)$

$P(G > -1.753)$ $P(G > 0.691)$

3. Dada una variable aleatoria F se desea saber

$P(\frac{1}{3.85} < F_{10,8} < 4.30)$ $P(F_{10,8} > x_0) = 0.95$

$P(3.35 < F_{10,8} < 4.30)$ $P(F_{10,8} > 4.30)$

4. Las calificaciones de examen final, en un curso de introducción a la estadística, tiene distribución normal con una media de 73 y una desviación estándar de 8.

1) ¿cuál es la probabilidad de obtener, como máximo, una calificación de 91 en este examen?
2) ¿qué porcentaje de estudiantes calificó entre 65 y 89?
3) ¿qué porcentaje de estudiantes calificó entre 81 y 89?
4) ¿cuál debe ser la calificación final del examen, si solo el 5% de los estudiantes examinados tuvieron la calificación más alta?

El profesor decide aprobar al 10% de los alumnos que tengan la mayor nota, sin que importe su calificación.

5) El alumno que haya obtenido una calificación de 81 en este examen, ¿aprueba?
6) Con una calificación de 68 en un examen en que la media es 62 y la desviación es 3,el alumno ¿aprueba?.
7) ¿Cuál de las dos situaciones es más conveniente para el alumno? Mostrar y explicar estadísticamente.

5. El diámetro de las pelotas de ping-pong fabricada por una industria ubicada en la Provincia de Córdoba, tienen una distribución aproximadamente normal con una media de 1,30 pulgadas y una desviación estándar de 0,04 pulgadas.

Cuál es la probabilidad de que una pelota de ping pong seleccionada aleatoriamente tenga un diámetro de

1) entre 1,28 y 1,30 pulgadas?
2) entre qué dos valores simétricamente distribuidos alrededor de la media caerá el 60% de las pelotas de ping-pong (en términos del diámetro)?

Si se seleccionaron muchas muestras de tamaño 16

3) cuáles se esperaría que fueran la media y el desvío estándar de la media?
4) qué distribución seguirían las medias de la muestra?
5) qué proporción de las medias de muestras estarían entre 1,28 y 1,30 pulgadas?
6) entre qué valores se encontraría el 60% de las medias de las muestras?

Qué es más probable que ocurra

7) una pelota individual mayor a 1,34 pulgadas?
8) una media muestral por arriba de 1,32 pulgada en una muestra de tamaño 4?
9) una media muestral por arriba de 1,31 pulgadas en una muestra de tamaño 16?

6. El gerente de control de calidad de una fábrica de lámparas eléctricas desea estimar la duración promedio de un embarque de lámparas (focos). Se selecciona una muestra aleatoria de 64 focos. Los resultados indican una duración promedio de la muestra de 540 horas con una desviación estándar de 120 horas.

Realiza una estimación de la duración promedio real de los focos de este embarque:

a) con intervalo de confianza de 90%,
b) con intervalo de confianza de 95%,
c) con intervalo de confianza de 99%.

Observar cómo cambia el tamaño del intervalo.

7. El administrador de la sucursal de un banco de ahorro local desea estimar la cantidad promedio que se tiene en las cuentas

de ahorro de los clientes del banco. Seleccionó una muestra aleatoria de 30 depositantes y los resultados indicaron un promedio de muestra de $4750,00 y un desvío estándar de $1200,00

a) Establecer una estimación por intervalo de confianza del 95%, de la cantidad promedio que se tiene en todas las cuentas de ahorro
b) Si un cliente tiene $4000,00 ¿puede considerárselo fuera de lo normal? ¿Porqué?.

8. Una línea de colectivos piensa establecer una ruta desde un barrio hasta el centro de la ciudad. Se selecciona una muestra aleatoria de 50 posibles usuarios y 18 indicaron que utilizarían esta ruta de colectivos. Establecer una estimación por intervalo con 95% de confianza de la proporción real de usuarios para esta nueva ruta de autobuses.

9. De dos grupos de pacientes A y B, compuestos de 50 y 100 individuos respectivamente, al primero le fue dado un nuevo tipo de píldoras para dormir y al segundo le fue dado el tipo convencional.

Los pacientes del grupo A durmieron 8,32 horas en promedio, con una desviación típica estimada de 0,24 horas.

Los pacientes del grupo B durmieron 6,75 horas en promedio, con una desviación típica estimada de 0,30 horas.

Encontrar los límites de confianza para la diferencia en las varianzas del número medio de horas de sueño inducidas por los dos tipos de píldoras, para un nivel de significación del 0,10 y el 0,02.

10. El saldo que arroja la cuenta deudores varios en el balance general de una empresa es de $120.000. Se conoce además que el total de deudores es de 500. Tomada una muestra al azar de 50 de esos deudores se encontró que debían en total $11.900. Es confiable a un nivel del 1% la cifra que reflejan los registros contables si se supone población aproximadamente normal y desvío poblacional de 100?

11. En una fábrica de alambres se sabe, por registros estadísticos anteriores, que la resistencia media de un tipo de alambre era de 12,46kg con una desviación típica de 1,8 kg. Se toma una muestra de 25 pedazos de alambre y se encuentra que la media era de 31,01 y la varianza estimada de 4,5. Como se planea efectuar una docima de la media, necesitan saber si la varianza no se ha alterado. Docimar a un nivel de significación del 5% si tal hecho ocurre. Suponer que la población es normal.

12. En una muestra de 400 lámparas producidas por una empresa dedicada al ramo se encontró que la vida media de cada una es de 1585 horas. Se conoce además que la desviación típica de la población es de 120 horas. El Departamento de Ingeniería asegura que la vida media de las lámparas producidas debe ser de 1600 horas. A un nivel de significación del 1%, ¿corresponde aceptar la aseveración del Departamento de Ingeniería a la luz de los resultados muestrales?

13. Un constructor está considerando dos lugares alternativos para un centro comercial regional. Como los ingresos de los hogares de la comunidad son una consideración importante en esa selección, desea probar la hipótesis nula de que no existe diferencia entre el ingreso promedio por hogar en las dos comunidades. Consistente con esta hipótesis supone que la desviación estándar del ingreso por hogar es también igual en las dos comunidades. Para una muestra de 30 hogares de la

primera comunidad encuentra que el ingreso promedio es de $35500 con desviación estándar muestral de $1800. Para una muestra de 40 hogares de la segunda comunidad encuentra que el ingreso promedio es de $34600 con desviación estándar muestral de $2400. Probar la hipótesis nula para el nivel de significación de 0,05.

14. Aceros S.A. fabrica barras de acero. El proceso de producción hace barras con una longitud promedio de, cuanto menos, 2,8 pies cuando el proceso funciona correctamente. Se selecciona una muestra de 25 barras en la línea de producción. La muestra indica una longitud promedio de 2,43 pies y una desviación estándar de 0,20 pies. La compañía desea determinar si la máquina necesita algún ajuste. A un nivel de significación del 5% ¿que decisión se debe tomar?

15. De 100 graduados en Contador Público, una muestra aleatoria de 12 estudiantes indica una calificación promedio de 6,75 con una desviación estándar muestral de 1,00. Para 50 egresados de Licenciatura en Administración de Empresas, una muestra aleatoria de 10 estudiantes tiene un promedio de calificación de 7,25 con una desviación estándar muestral de 0,75. Se supone que las calificaciones tienen distribución normal. Probar la hipótesis de que la calificación promedio para las dos categorías de estudiantes es distinta utilizando el nivel de significación del 5%.

16. Se plantea la hipótesis de que la desviación estándar del salario por hora de los trabajadores a destajo, en una determinada industria, cuanto menos es de $3000. Para una muestra de 15 trabajadores elegidos al azar, se encuentra que la desviación estándar es de 2000. Se supone que las cifras de ingresos de los trabajadores de la población tienen una distribución normal. Teniendo en cuenta este resultado muestral

¿puede rechazarse la hipótesis nula utilizando el nivel de significación del 5%?.

17. Una fábrica de galletitas incorporó un nuevo proceso de elaboración de bizcochos dulces que trae aparejado ciertas ventajas. Pero un detalle muy importante es el peso de las galletas. El gerente de fabricación desea saber el grado de variación del nuevo proceso respecto del anterior en cuanto a peso de cada galleta. Tomada una muestra de cada proceso se obtuvo

$n = 16$ $\qquad$ $n = 13$

$\sum_{i=1}^{16} \mathrm{X}_i = 319{,}20$ $\qquad$ $\sum_{i=1}^{13} \mathrm{Y}_i = 261{,}30$

$\sum_{i=1}^{16} \mathrm{X}_i^2 = 6373{,}11$ $\qquad$ $\sum_{i=1}^{13} \mathrm{Y}_i^2 = 5255{,}54$

Sabiendo que ambas poblaciones son aproximadamente normales. A un nivel de confianza del 99%, ¿hay diferencias significativas en cuanto a las varianzas de los pesos de los bizcochos entre ambos procesos?

18. El gerente de marketing de una cadena de tiendas de flores y plantas cree que, durante el mes de mayo, habrá una gran demanda de bulbos de gladiolos en las tiendas de la periferia de la ciudad.

Si las ventas de bulbos exceden de $100 por semana se dispondrán para la venta en todas las tiendas suburbanas de la cadena.

Se selecciona una muestra aleatoria de 16 tiendas y los resultados de la prueba en las tiendas indicaron ventas promedio de $120 con una desviación estándar de la muestra de $25

Con un nivel de significación del 1% ¿se deben vender los bulbos de gladiolos en todas las tiendas suburbanas de la cadena?

19. En una compañía se quiere determinar los gastos médicos familiares anuales promedio de los empleados. La gerencia de la compañía quiere tener 95% de confianza de que el promedio de la muestra está correcto con aproximación ±$50 de los gastos familiares reales. Un estudio piloto indica que la desviación estándar se puede estimar en $1400. ¿Qué tan grande es el tamaño de la muestra necesario?

20. Se desea estimar la suma promedio de ventas con aproximación de ±$100 con 99% de confianza y se supone que la desviación estándar es de $200, ¿qué tamaño de muestra se necesita?

21. Un grupo de estudio quiere estimar la facturación mensual promedio por luz eléctrica en el mes de julio en una ciudad grande. En base a estudios efectuados en otras ciudades se supone que la desviación estándar es de $20. El grupo querría estimar la facturación promedio de julio con aproximación ±$5 del promedio real con 99% de confianza ¿Qué tamaño de muestra se necesita?

22. Una empresa quiere realizar una encuesta sobre el conjunto de su personal, compuesto de 10 000 personas. Estudios preliminares han demostrado:

- que las variables que se intenta analizar en la encuesta

son contrastadas según las categorías del personal, por lo que se tiene interés en estratificar en función de esas categorías. Por motivos de simplificación, considera que hay tres grandes categorías que formarán los estratos;

- que las variables están fuertemente relacionadas con la edad de los individuos.

Se propone un plan de muestreo como si se quisiera estudiar la edad de los individuos: si una estrategia es mejor que otras para estimar la media de la edad, se puede pensar que también será la mejor para estimar los verdaderos parámetros de interés. Como se conoce la edad de los miembros del personal, se puede razonar haciendo comparaciones exactas.

Se dispone de la siguiente información

Categorías	Peso en el conjunto de personal	Desviación estándar de las edades
1	20%	18.0
2	30%	12.0
3	50%	3.6
Total	100%	16.0

a) Sea μ la edad media y $\hat{\mu}$ el estimador procedente de una muestra aleatoria simple, sin reemplazo con probabilidades iguales, de $n = 100$ individuos.

   ¿Cuál es la desviación estándar de $\hat{\mu}$?

b) Se decide que la muestra de 100 individuos tiene que ser estratificada en función de las tres categorías del personal.

   ¿Cuál es la repartición proporcional?

   ¿Cuál es la desviación estándar de $\hat{\mu}$?

Comparar los resultados con los de a).

c) ¿Cuál sería la repartición óptima de la muestra?

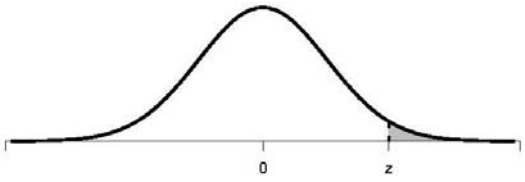
¿Cuál es la desviación estándar de $\hat{\mu}$ que resulta ahora?

Comparar sus resultados con el resultado de b).

Tablas estadísticas

Normal estándar acumulada

Densidad: $\phi(x) = \frac{e^{\frac{-x^2}{2}}}{\sqrt{2\pi}}$ Acumulada en la cola derecha: $\alpha = 1 - \Phi(z) = \int_z^{\infty} \phi(x)dx$



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
3.7	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
3.8	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
3.9	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

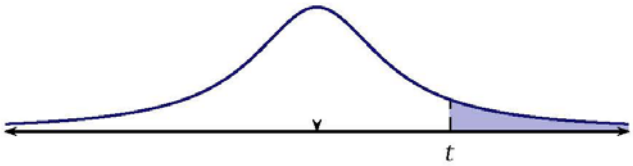
Tablas Estadísticas. Versión: Marzo, 2014. Producción: Datos obtenidos a partir del entorno de programación R. Configuración tipográfica mediante LaTeX realizada por el autor.

www.faviodercole.com.ar
contacto@faviodercole.com.ar

Cuantiles de distribución acumulada *t* de Student

$$f(x|\nu) = \frac{1}{\sqrt{\nu\pi}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad \alpha = 1 - F(x|\nu) = \int_t^\infty f(x|\nu)dx$$

donde ν son grados de libertad.

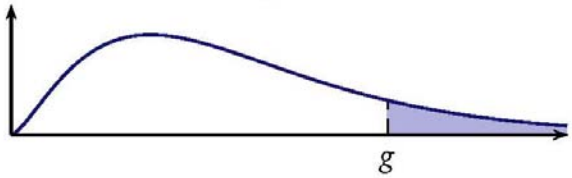


$\nu\backslash\alpha$	.005	.010	.020	.025	.05	.10	.15	.20	.25	.30	.35	.40	.45
1	63.656	31.820	15.894	12.706	6.3138	3.0777	1.9626	1.3764	1.0000	0.7265	0.5095	0.3249	0.1584
2	9.9248	6.9646	4.8487	4.3027	2.9200	1.8856	1.3862	1.0607	0.8165	0.6172	0.4447	0.2887	0.1421
3	5.8409	4.5407	3.4819	3.1824	2.3534	1.6377	1.2498	0.9785	0.7649	0.5844	0.4242	0.2767	0.1366
4	4.6041	3.7469	2.9985	2.7764	2.1318	1.5332	1.1896	0.9410	0.7407	0.5686	0.4142	0.2707	0.1338
5	4.0321	3.3649	2.7565	2.5706	2.0150	1.4759	1.1558	0.9195	0.7267	0.5594	0.4082	0.2672	0.1322
6	3.7074	3.1427	2.6122	2.4469	1.9432	1.4398	1.1342	0.9057	0.7176	0.5534	0.4043	0.2648	0.1311
7	3.4995	2.9980	2.5168	2.3646	1.8946	1.4149	1.1192	0.8960	0.7111	0.5491	0.4015	0.2632	0.1303
8	3.3554	2.8965	2.4490	2.3060	1.8595	1.3968	1.1081	0.8889	0.7064	0.5459	0.3995	0.2619	0.1297
9	3.2498	2.8214	2.3984	2.2622	1.8331	1.3830	1.0997	0.8834	0.7027	0.5435	0.3979	0.2610	0.1293
10	3.1693	2.7638	2.3593	2.2281	1.8125	1.3722	1.0931	0.8791	0.6998	0.5415	0.3966	0.2602	0.1289
11	3.1058	2.7181	2.3281	2.2010	1.7959	1.3634	1.0877	0.8755	0.6974	0.5399	0.3956	0.2596	0.1286
12	3.0545	2.6810	2.3027	2.1788	1.7823	1.3562	1.0832	0.8726	0.6955	0.5386	0.3947	0.2590	0.1283
13	3.0123	2.6503	2.2816	2.1604	1.7709	1.3502	1.0795	0.8702	0.6938	0.5375	0.3940	0.2586	0.1281
14	2.9768	2.6245	2.2638	2.1448	1.7613	1.3450	1.0763	0.8681	0.6924	0.5366	0.3933	0.2582	0.1280
15	2.9467	2.6025	2.2485	2.1314	1.7531	1.3406	1.0735	0.8662	0.6912	0.5357	0.3928	0.2579	0.1278
16	2.9208	2.5835	2.2354	2.1199	1.7459	1.3368	1.0711	0.8647	0.6901	0.5350	0.3923	0.2576	0.1277
17	2.8982	2.5669	2.2238	2.1098	1.7396	1.3334	1.0690	0.8633	0.6892	0.5344	0.3919	0.2573	0.1276
18	2.8784	2.5524	2.2137	2.1009	1.7341	1.3304	1.0672	0.8620	0.6884	0.5338	0.3915	0.2571	0.1274
19	2.8609	2.5395	2.2047	2.0930	1.7291	1.3277	1.0655	0.8610	0.6876	0.5333	0.3912	0.2569	0.1274
20	2.8453	2.5280	2.1967	2.0860	1.7247	1.3253	1.0640	0.8600	0.6870	0.5329	0.3909	0.2567	0.1273
21	2.8314	2.5176	2.1894	2.0796	1.7207	1.3232	1.0627	0.8591	0.6864	0.5325	0.3906	0.2566	0.1272
22	2.8188	2.5083	2.1829	2.0739	1.7171	1.3212	1.0614	0.8583	0.6858	0.5321	0.3904	0.2564	0.1271
23	2.8073	2.4999	2.1770	2.0687	1.7139	1.3195	1.0603	0.8575	0.6853	0.5317	0.3902	0.2563	0.1271
24	2.7969	2.4922	2.1715	2.0639	1.7109	1.3178	1.0593	0.8569	0.6848	0.5314	0.3900	0.2562	0.1270
25	2.7874	2.4851	2.1666	2.0595	1.7081	1.3163	1.0584	0.8562	0.6844	0.5312	0.3898	0.2561	0.1269
26	2.7787	2.4786	2.1620	2.0555	1.7056	1.3150	1.0575	0.8557	0.6840	0.5309	0.3896	0.2560	0.1269
27	2.7707	2.4727	2.1578	2.0518	1.7033	1.3137	1.0567	0.8551	0.6837	0.5306	0.3894	0.2559	0.1268
28	2.7633	2.4671	2.1539	2.0484	1.7011	1.3125	1.0560	0.8546	0.6834	0.5304	0.3893	0.2558	0.1268
29	2.7564	2.4620	2.1503	2.0452	1.6991	1.3114	1.0553	0.8542	0.6830	0.5302	0.3892	0.2557	0.1268
30	2.7500	2.4573	2.1470	2.0423	1.6973	1.3104	1.0547	0.8538	0.6828	0.5300	0.3890	0.2556	0.1267
35	2.7238	2.4377	2.1332	2.0301	1.6896	1.3062	1.0520	0.8520	0.6816	0.5292	0.3885	0.2553	0.1266
40	2.7045	2.4233	2.1229	2.0211	1.6839	1.3031	1.0500	0.8507	0.6807	0.5286	0.3881	0.2550	0.1265
45	2.6896	2.4121	2.1150	2.0141	1.6794	1.3006	1.0485	0.8497	0.6800	0.5281	0.3878	0.2549	0.1264
50	2.6778	2.4033	2.1087	2.0086	1.6759	1.2987	1.0473	0.8489	0.6794	0.5278	0.3875	0.2547	0.1263
55	2.6682	2.3961	2.1036	2.0040	1.6730	1.2971	1.0463	0.8482	0.6790	0.5275	0.3873	0.2546	0.1262
60	2.6603	2.3901	2.0994	2.0003	1.6706	1.2958	1.0455	0.8477	0.6786	0.5272	0.3872	0.2545	0.1262
65	2.6536	2.3851	2.0958	1.9971	1.6686	1.2947	1.0448	0.8472	0.6783	0.5270	0.3870	0.2544	0.1262
70	2.6479	2.3808	2.0927	1.9944	1.6669	1.2938	1.0442	0.8468	0.6780	0.5268	0.3869	0.2543	0.1261
75	2.6430	2.3771	2.0901	1.9921	1.6654	1.2929	1.0436	0.8464	0.6778	0.5266	0.3868	0.2542	0.1261
80	2.6387	2.3739	2.0878	1.9901	1.6641	1.2922	1.0432	0.8461	0.6776	0.5265	0.3867	0.2542	0.1261
85	2.6349	2.3710	2.0857	1.9883	1.6630	1.2916	1.0428	0.8459	0.6774	0.5264	0.3866	0.2541	0.1260
90	2.6316	2.3685	2.0839	1.9867	1.6620	1.2910	1.0424	0.8456	0.6772	0.5263	0.3866	0.2541	0.1260
95	2.6286	2.3662	2.0823	1.9853	1.6611	1.2905	1.0421	0.8454	0.6771	0.5262	0.3865	0.2541	0.1260
100	2.6259	2.3642	2.0809	1.9840	1.6602	1.2901	1.0418	0.8452	0.6770	0.5261	0.3864	0.2540	0.1260
500	2.5857	2.3338	2.0591	1.9647	1.6479	1.2832	1.0375	0.8423	0.6750	0.5247	0.3855	0.2535	0.1257
1000	2.5808	2.3301	2.0564	1.9623	1.6464	1.2824	1.0370	0.8420	0.6747	0.5246	0.3854	0.2534	0.1257
5000	2.5768	2.3271	2.0543	1.9604	1.6452	1.2817	1.0365	0.8417	0.6745	0.5244	0.3853	0.2534	0.1257
∞	2.5758	2.3263	2.0537	1.9600	1.6449	1.2816	1.0364	0.8416	0.6745	0.5244	0.3853	0.2533	0.1257

Cuantiles de distribución acumulada χ^2

$$f(x|\nu) = \begin{cases} 2^{-\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)^{-1} x^{\frac{\nu}{2}-1}e^{-\frac{x}{2}} & \text{para } x \geq 0 \\ 0 & \text{para } x < 0 \end{cases} \qquad \alpha = 1 - F(x|\nu) = \int_g^{\infty} f(x|\nu)dx$$

donde ν es grado de libertad.

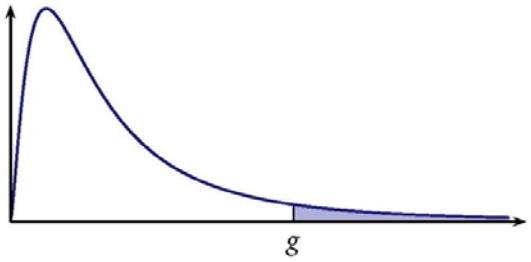


$g\backslash\alpha$	0.005	0.010	0.025	0.050	0.100	0.150	0.850	0.900	0.950	0.975	0.990	0.995
1	7.8794	6.6349	5.0239	3.8415	2.7055	2.0723	0.0358	0.0158	0.0039	0.0010	0.0002	0.0000
2	10.597	9.2103	7.3778	5.9915	4.6052	3.7942	0.3250	0.2107	0.1026	0.0506	0.0201	0.0100
3	12.838	11.345	9.3484	7.8147	6.2514	5.3170	0.7978	0.5844	0.3518	0.2158	0.1148	0.0717
4	14.860	13.277	11.143	9.4877	7.7794	6.7449	1.3665	1.0636	0.7107	0.4844	0.2971	0.2070
5	16.750	15.086	12.833	11.070	9.2364	8.1152	1.9938	1.6103	1.1455	0.8312	0.5543	0.4117
6	18.548	16.812	14.449	12.592	10.645	9.4461	2.6613	2.2041	1.6354	1.2373	0.8721	0.6757
7	20.278	18.475	16.013	14.067	12.017	10.748	3.3583	2.8331	2.1673	1.6899	1.2390	0.9893
8	21.955	20.090	17.535	15.507	13.362	12.027	4.0782	3.4895	2.7326	2.1797	1.6465	1.3444
9	23.589	21.666	19.023	16.919	14.684	13.288	4.8165	4.1682	3.3251	2.7004	2.0879	1.7349
10	25.188	23.209	20.483	18.307	15.987	14.534	5.5701	4.8652	3.9403	3.2470	2.5582	2.1559
11	26.757	24.725	21.920	19.675	17.275	15.767	6.3364	5.5778	4.5748	3.8157	3.0535	2.6032
12	28.300	26.217	23.337	21.026	18.549	16.989	7.1138	6.3038	5.2260	4.4038	3.5706	3.0738
13	29.819	27.688	24.736	22.362	19.812	18.202	7.9008	7.0415	5.8919	5.0088	4.1069	3.5650
14	31.319	29.141	26.119	23.685	21.064	19.406	8.6963	7.7895	6.5706	5.6287	4.6604	4.0747
15	32.801	30.578	27.488	24.996	22.307	20.603	9.4993	8.5468	7.2609	6.2621	5.2293	4.6009
16	34.267	32.000	28.845	26.296	23.542	21.793	10.309	9.3122	7.9616	6.9077	5.8122	5.1422
17	35.718	33.409	30.191	27.587	24.769	22.977	11.125	10.085	8.6718	7.5642	6.4078	5.6972
18	37.156	34.805	31.526	28.869	25.989	24.155	11.946	10.865	9.3905	8.2307	7.0149	6.2648
19	38.582	36.191	32.852	30.144	27.204	25.329	12.773	11.651	10.117	8.9065	7.6327	6.8440
20	39.997	37.566	34.170	31.410	28.412	26.498	13.604	12.443	10.851	9.5908	8.2604	7.4338
21	41.401	38.932	35.479	32.671	29.615	27.662	14.439	13.240	11.591	10.283	8.8972	8.0337
22	42.796	40.289	36.781	33.924	30.813	28.822	15.279	14.041	12.338	10.982	9.5425	8.6427
23	44.181	41.638	38.076	35.172	32.007	29.979	16.122	14.848	13.091	11.689	10.196	9.2604
24	45.559	42.980	39.364	36.415	33.196	31.132	16.969	15.659	13.848	12.401	10.856	9.8862
25	46.928	44.314	40.646	37.652	34.382	32.282	17.818	16.473	14.611	13.120	11.524	10.520
26	48.290	45.642	41.923	38.885	35.563	33.429	18.671	17.292	15.379	13.844	12.198	11.160
27	49.645	46.963	43.195	40.113	36.741	34.574	19.527	18.114	16.151	14.573	12.879	11.808
28	50.993	48.278	44.461	41.337	37.916	35.715	20.386	18.939	16.928	15.308	13.565	12.461
29	52.336	49.588	45.722	42.557	39.087	36.854	21.247	19.768	17.708	16.047	14.256	13.121
30	53.672	50.892	46.979	43.773	40.256	37.990	22.110	20.599	18.493	16.791	14.953	13.787
35	60.275	57.342	53.203	49.802	46.059	43.640	26.460	24.797	22.465	20.569	18.509	17.192
40	66.766	63.691	59.342	55.758	51.805	49.244	30.856	29.051	26.509	24.433	22.164	20.707
45	73.166	69.957	65.410	61.656	57.505	54.810	35.290	33.350	30.612	28.366	25.901	24.311
50	79.490	76.154	71.420	67.505	63.167	60.346	39.754	37.689	34.764	32.357	29.707	27.991
55	85.749	82.292	77.380	73.311	68.796	65.855	44.245	42.060	38.958	36.398	33.570	31.735
60	91.952	88.379	83.298	79.082	74.397	71.341	48.759	46.459	43.188	40.482	37.485	35.534
65	98.105	94.422	89.177	84.821	79.973	76.807	53.293	50.883	47.450	44.603	41.444	39.383
70	104.21	100.43	95.023	90.531	85.527	82.255	57.844	55.329	51.739	48.758	45.442	43.275
75	110.29	106.39	100.84	96.217	91.061	87.688	62.412	59.795	56.054	52.942	49.475	47.206
80	116.32	112.33	106.63	101.88	96.578	93.106	66.994	64.278	60.391	57.153	53.540	51.172
85	122.32	118.24	112.39	107.52	102.08	98.511	71.589	68.777	64.749	61.389	57.634	55.170
90	128.30	124.12	118.14	113.15	107.57	103.90	76.195	73.291	69.126	65.647	61.754	59.196
95	134.25	129.97	123.86	118.75	113.04	109.29	80.813	77.818	73.520	69.925	65.898	63.250
100	140.17	135.81	129.56	124.34	118.50	114.66	85.441	82.358	77.929	74.222	70.065	67.328
500	585.21	576.49	563.85	553.13	540.93	532.80	467.30	459.93	449.15	439.94	429.39	422.30
1000	1118.9	1107.0	1089.5	1074.7	1057.7	1046.4	953.71	943.13	927.59	914.26	898.91	888.56
5000	5261.3	5235.6	5197.9	5165.6	5128.6	5103.7	4896.4	4872.3	4836.7	4805.9	4770.3	4746.2

Cuantiles de distribución F de Snedecor

$$f(x|\nu_1,\nu_2) = \mathrm{B}\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)^{-1}\left(\frac{\nu_1 x}{\nu_1 x+\nu_2}\right)^{\frac{\nu_1}{2}}\left(1-\frac{\nu_1 x}{\nu_1 x+\nu_2}\right)^{\frac{\nu_2}{2}}x^{-1} \qquad \alpha = 1-F(x|\nu_1,\nu_2) = \int_g^{\infty} f(x|\nu_1,\nu_2)dx$$

donde ν_1 son los grados de libertad del numerador (filas en la tabla)
y ν_2 son los grados de libertad del denomindor (columnas en la tabla).



$\alpha = ,01$	1	2	3	4	5	6	7	8	9	10	11	12	14	16	18	20	30	40	50	100	250	500
1	4052	98.50	34.12	21.20	16.26	13.75	12.25	11.26	10.56	10.04	9.646	9.330	8.862	8.531	8.285	8.096	7.562	7.314	7.171	6.895	6.737	6.686
2	4999	99.00	30.82	18.00	13.27	10.92	9.547	8.649	8.022	7.559	7.206	6.927	6.515	6.226	6.013	5.849	5.390	5.179	5.057	4.824	4.691	4.648
3	5403	99.17	29.46	16.69	12.06	9.780	8.451	7.591	6.992	6.552	6.217	5.953	5.564	5.292	5.092	4.938	4.510	4.313	4.199	3.984	3.861	3.821
4	5625	99.25	28.71	15.98	11.39	9.148	7.847	7.006	6.422	5.994	5.668	5.412	5.035	4.773	4.579	4.431	4.018	3.828	3.720	3.513	3.395	3.357
5	5764	99.30	28.24	15.52	10.97	8.746	7.460	6.632	6.057	5.636	5.316	5.064	4.695	4.437	4.248	4.103	3.699	3.514	3.408	3.206	3.091	3.054
6	5859	99.33	27.91	15.21	10.67	8.466	7.191	6.371	5.802	5.386	5.069	4.821	4.456	4.202	4.015	3.871	3.473	3.291	3.186	2.988	2.875	2.838
7	5928	99.36	27.67	14.98	10.46	8.260	6.993	6.178	5.613	5.200	4.886	4.640	4.278	4.026	3.841	3.699	3.304	3.124	3.020	2.823	2.711	2.675
8	5981	99.37	27.49	14.80	10.29	8.102	6.840	6.029	5.467	5.057	4.744	4.499	4.140	3.890	3.705	3.564	3.173	2.993	2.890	2.694	2.583	2.547
9	6022	99.39	27.35	14.66	10.16	7.976	6.719	5.911	5.351	4.942	4.632	4.388	4.030	3.780	3.597	3.457	3.067	2.888	2.785	2.590	2.479	2.443
10	6056	99.40	27.23	14.55	10.05	7.874	6.620	5.814	5.257	4.849	4.539	4.296	3.939	3.691	3.508	3.368	2.979	2.801	2.698	2.503	2.392	2.356
11	6083	99.41	27.13	14.45	9.963	7.790	6.538	5.734	5.178	4.772	4.462	4.220	3.864	3.616	3.434	3.294	2.906	2.727	2.625	2.430	2.319	2.283
12	6106	99.42	27.05	14.37	9.888	7.718	6.469	5.667	5.111	4.706	4.397	4.155	3.800	3.553	3.371	3.231	2.843	2.665	2.562	2.368	2.257	2.220
13	6126	99.42	26.98	14.31	9.825	7.657	6.410	5.609	5.055	4.650	4.342	4.100	3.745	3.498	3.316	3.177	2.789	2.611	2.508	2.313	2.202	2.166
14	6143	99.43	26.92	14.25	9.770	7.605	6.359	5.559	5.005	4.601	4.293	4.052	3.698	3.451	3.269	3.130	2.742	2.563	2.461	2.265	2.154	2.117
15	6157	99.43	26.87	14.20	9.722	7.559	6.314	5.515	4.962	4.558	4.251	4.010	3.656	3.409	3.227	3.088	2.700	2.522	2.419	2.223	2.111	2.075
16	6170	99.44	26.83	14.15	9.680	7.519	6.275	5.477	4.924	4.520	4.213	3.972	3.619	3.372	3.190	3.051	2.663	2.484	2.382	2.185	2.073	2.036
17	6181	99.44	26.79	14.11	9.643	7.483	6.240	5.442	4.890	4.487	4.180	3.939	3.586	3.339	3.158	3.018	2.630	2.451	2.348	2.151	2.038	2.002
18	6192	99.44	26.75	14.08	9.610	7.451	6.209	5.412	4.860	4.457	4.150	3.909	3.556	3.310	3.128	2.989	2.600	2.421	2.318	2.120	2.007	1.970
19	6201	99.45	26.72	14.05	9.580	7.422	6.181	5.384	4.833	4.430	4.123	3.883	3.529	3.283	3.101	2.962	2.573	2.394	2.290	2.092	1.979	1.942
20	6209	99.45	26.69	14.02	9.553	7.396	6.155	5.359	4.808	4.405	4.099	3.858	3.505	3.259	3.077	2.938	2.549	2.369	2.265	2.067	1.953	1.915
21	6216	99.45	26.66	13.99	9.528	7.372	6.132	5.336	4.786	4.383	4.077	3.836	3.483	3.237	3.055	2.916	2.526	2.346	2.242	2.043	1.928	1.891
22	6223	99.45	26.64	13.97	9.506	7.351	6.111	5.316	4.765	4.363	4.057	3.816	3.463	3.216	3.035	2.895	2.506	2.325	2.221	2.021	1.906	1.869
23	6229	99.46	26.62	13.95	9.485	7.331	6.092	5.297	4.746	4.344	4.038	3.798	3.444	3.198	3.016	2.877	2.487	2.306	2.202	2.001	1.886	1.848
24	6235	99.46	26.60	13.93	9.466	7.313	6.074	5.279	4.729	4.327	4.021	3.780	3.427	3.181	2.999	2.859	2.469	2.288	2.183	1.983	1.867	1.829
25	6240	99.46	26.58	13.91	9.449	7.296	6.058	5.263	4.713	4.311	4.005	3.765	3.412	3.165	2.983	2.843	2.453	2.271	2.167	1.965	1.849	1.810
26	6245	99.46	26.56	13.89	9.433	7.280	6.043	5.248	4.698	4.296	3.990	3.750	3.397	3.150	2.968	2.829	2.437	2.256	2.151	1.949	1.832	1.794
27	6249	99.46	26.55	13.88	9.418	7.266	6.029	5.234	4.685	4.283	3.977	3.736	3.383	3.137	2.955	2.815	2.423	2.241	2.136	1.934	1.816	1.778
28	6253	99.46	26.53	13.86	9.404	7.253	6.016	5.221	4.672	4.270	3.964	3.724	3.371	3.124	2.942	2.802	2.410	2.228	2.123	1.919	1.801	1.763
29	6257	99.46	26.52	13.85	9.391	7.240	6.003	5.209	4.660	4.258	3.952	3.712	3.359	3.112	2.930	2.790	2.398	2.215	2.110	1.906	1.788	1.749
30	6261	99.47	26.50	13.84	9.379	7.229	5.992	5.198	4.649	4.247	3.941	3.701	3.348	3.101	2.919	2.778	2.386	2.203	2.098	1.893	1.774	1.735
35	6276	99.47	26.45	13.79	9.329	7.180	5.944	5.151	4.602	4.200	3.895	3.654	3.301	3.054	2.871	2.731	2.337	2.153	2.046	1.839	1.718	1.678
40	6287	99.47	26.41	13.75	9.291	7.143	5.908	5.116	4.567	4.165	3.860	3.619	3.266	3.018	2.835	2.695	2.299	2.114	2.007	1.797	1.674	1.633
45	6296	99.48	26.38	13.71	9.262	7.115	5.880	5.088	4.539	4.138	3.832	3.592	3.238	2.990	2.807	2.666	2.269	2.083	1.975	1.763	1.638	1.596
50	6303	99.48	26.35	13.69	9.238	7.091	5.858	5.065	4.517	4.115	3.810	3.569	3.215	2.967	2.784	2.643	2.245	2.058	1.949	1.735	1.608	1.566
55	6308	99.48	26.33	13.67	9.218	7.073	5.839	5.047	4.498	4.097	3.791	3.551	3.197	2.949	2.765	2.624	2.225	2.037	1.927	1.712	1.583	1.540
60	6313	99.48	26.32	13.65	9.202	7.057	5.824	5.032	4.483	4.082	3.776	3.535	3.181	2.933	2.749	2.608	2.208	2.019	1.909	1.692	1.561	1.517
65	6317	99.48	26.30	13.64	9.188	7.043	5.810	5.019	4.470	4.069	3.763	3.522	3.168	2.920	2.736	2.594	2.193	2.004	1.893	1.674	1.543	1.498
70	6321	99.48	26.29	13.63	9.176	7.032	5.799	5.007	4.459	4.058	3.752	3.511	3.157	2.908	2.724	2.582	2.181	1.991	1.880	1.659	1.526	1.481
75	6324	99.49	26.28	13.61	9.166	7.022	5.789	4.998	4.449	4.048	3.742	3.501	3.147	2.898	2.714	2.572	2.170	1.980	1.868	1.646	1.511	1.465
80	6326	99.49	26.27	13.61	9.157	7.013	5.781	4.989	4.441	4.039	3.734	3.493	3.138	2.889	2.705	2.563	2.160	1.969	1.857	1.634	1.498	1.452
85	6329	99.49	26.26	13.60	9.149	7.005	5.773	4.981	4.433	4.032	3.726	3.485	3.130	2.882	2.697	2.555	2.152	1.960	1.848	1.624	1.486	1.439
90	6331	99.49	26.25	13.59	9.142	6.998	5.766	4.975	4.426	4.025	3.719	3.478	3.124	2.875	2.690	2.548	2.144	1.952	1.839	1.614	1.476	1.428
95	6332	99.49	26.25	13.58	9.136	6.992	5.760	4.969	4.420	4.019	3.713	3.472	3.117	2.868	2.684	2.541	2.137	1.945	1.832	1.606	1.466	1.418
100	6334	99.49	26.24	13.58	9.130	6.987	5.755	4.963	4.415	4.014	3.708	3.467	3.112	2.863	2.678	2.535	2.131	1.938	1.825	1.598	1.457	1.408
250	6353	99.50	26.17	13.51	9.064	6.923	5.692	4.901	4.353	3.951	3.645	3.404	3.048	2.797	2.612	2.468	2.057	1.860	1.742	1.501	1.343	1.285
500	6360	99.50	26.15	13.49	9.042	6.902	5.671	4.880	4.332	3.930	3.624	3.382	3.026	2.775	2.589	2.445	2.032	1.833	1.713	1.466	1.297	1.232
1000	6363	99.50	26.14	13.47	9.031	6.891	5.660	4.869	4.321	3.920	3.613	3.372	3.015	2.764	2.577	2.433	2.019	1.819	1.698	1.447	1.272	1.201
10000	6366	99.50	26.13	13.46	9.022	6.881	5.651	4.860	4.312	3.910	3.604	3.362	3.005	2.754	2.567	2.422	2.008	1.806	1.685	1.429	1.247	1.168

$\alpha = ,025$	1	2	3	4	5	6	7	8	9	10	11	12	14	16	18	20	30	40	50	100	250	500
1	647.8	38.51	17.44	12.22	10.01	8.813	8.073	7.571	7.209	6.937	6.724	6.554	6.298	6.115	5.978	5.871	5.568	5.424	5.340	5.179	5.085	5.054
2	799.5	39.00	16.04	10.65	8.434	7.260	6.542	6.059	5.715	5.456	5.256	5.096	4.857	4.687	4.560	4.461	4.182	4.051	3.975	3.828	3.744	3.716
3	864.2	39.17	15.44	9.979	7.764	6.599	5.890	5.416	5.078	4.826	4.630	4.474	4.242	4.077	3.954	3.859	3.589	3.463	3.390	3.250	3.169	3.142
4	899.6	39.25	15.10	9.605	7.388	6.227	5.523	5.053	4.718	4.468	4.275	4.121	3.892	3.729	3.608	3.515	3.250	3.126	3.054	2.917	2.837	2.811
5	921.8	39.30	14.88	9.364	7.146	5.988	5.285	4.817	4.484	4.236	4.044	3.891	3.663	3.502	3.382	3.289	3.026	2.904	2.833	2.696	2.618	2.592
6	937.1	39.33	14.73	9.197	6.978	5.820	5.119	4.652	4.320	4.072	3.881	3.728	3.501	3.341	3.221	3.128	2.867	2.744	2.674	2.537	2.459	2.434
7	948.2	39.36	14.62	9.074	6.853	5.695	4.995	4.529	4.197	3.950	3.759	3.607	3.380	3.219	3.100	3.007	2.746	2.624	2.553	2.417	2.338	2.313
8	956.7	39.37	14.54	8.980	6.757	5.600	4.899	4.433	4.102	3.855	3.664	3.512	3.285	3.125	3.005	2.913	2.651	2.529	2.458	2.321	2.243	2.217
9	963.3	39.39	14.47	8.905	6.681	5.523	4.823	4.357	4.026	3.779	3.588	3.436	3.209	3.049	2.929	2.837	2.575	2.452	2.381	2.244	2.165	2.139
10	968.6	39.40	14.42	8.844	6.619	5.461	4.761	4.295	3.964	3.717	3.526	3.374	3.147	2.986	2.866	2.774	2.511	2.388	2.317	2.179	2.100	2.074
11	973.0	39.41	14.37	8.794	6.568	5.410	4.709	4.243	3.912	3.665	3.474	3.321	3.095	2.934	2.814	2.721	2.458	2.334	2.263	2.124	2.045	2.019
12	976.7	39.41	14.34	8.751	6.525	5.366	4.666	4.200	3.868	3.621	3.430	3.277	3.050	2.889	2.769	2.676	2.412	2.288	2.216	2.077	1.997	1.971
13	979.8	39.42	14.30	8.715	6.488	5.329	4.628	4.162	3.831	3.583	3.392	3.239	3.012	2.851	2.730	2.637	2.372	2.248	2.176	2.036	1.955	1.929
14	982.5	39.43	14.28	8.684	6.456	5.297	4.596	4.130	3.798	3.550	3.359	3.206	2.979	2.817	2.696	2.603	2.338	2.213	2.140	2.000	1.919	1.892
15	984.9	39.43	14.25	8.657	6.428	5.269	4.568	4.101	3.769	3.522	3.330	3.177	2.949	2.788	2.667	2.573	2.307	2.182	2.109	1.968	1.886	1.859
16	986.9	39.44	14.23	8.633	6.403	5.244	4.543	4.076	3.744	3.496	3.304	3.152	2.923	2.761	2.640	2.547	2.280	2.154	2.081	1.939	1.857	1.830
17	988.7	39.44	14.21	8.611	6.381	5.222	4.521	4.054	3.722	3.474	3.282	3.129	2.900	2.738	2.617	2.523	2.255	2.129	2.056	1.913	1.830	1.803
18	990.3	39.44	14.20	8.592	6.362	5.202	4.501	4.034	3.701	3.453	3.261	3.108	2.879	2.717	2.596	2.501	2.233	2.107	2.033	1.890	1.806	1.779
19	991.8	39.45	14.18	8.575	6.344	5.184	4.483	4.016	3.683	3.435	3.243	3.090	2.861	2.698	2.576	2.482	2.213	2.086	2.012	1.868	1.784	1.757
20	993.1	39.45	14.17	8.560	6.329	5.168	4.467	3.999	3.667	3.419	3.226	3.073	2.844	2.681	2.559	2.464	2.195	2.068	1.993	1.849	1.764	1.736
21	994.3	39.45	14.16	8.546	6.314	5.154	4.452	3.985	3.652	3.403	3.211	3.057	2.828	2.665	2.543	2.448	2.178	2.051	1.976	1.830	1.745	1.717
22	995.4	39.45	14.14	8.533	6.301	5.141	4.439	3.971	3.638	3.390	3.197	3.043	2.814	2.651	2.529	2.434	2.163	2.035	1.960	1.814	1.728	1.700
23	996.3	39.45	14.13	8.522	6.289	5.128	4.426	3.959	3.626	3.377	3.184	3.031	2.801	2.637	2.515	2.420	2.149	2.020	1.945	1.798	1.712	1.684
24	997.2	39.46	14.12	8.511	6.278	5.117	4.415	3.947	3.614	3.365	3.173	3.019	2.789	2.625	2.503	2.408	2.136	2.007	1.931	1.784	1.697	1.669
25	998.1	39.46	14.12	8.501	6.268	5.107	4.405	3.937	3.604	3.355	3.162	3.008	2.778	2.614	2.491	2.396	2.124	1.994	1.919	1.770	1.683	1.655
26	998.8	39.46	14.11	8.492	6.258	5.097	4.395	3.927	3.594	3.345	3.152	2.998	2.767	2.603	2.481	2.385	2.112	1.983	1.907	1.758	1.670	1.641
27	999.6	39.46	14.10	8.483	6.250	5.088	4.386	3.918	3.584	3.335	3.142	2.988	2.758	2.594	2.471	2.375	2.102	1.972	1.895	1.746	1.658	1.629
28	1000	39.46	14.09	8.476	6.242	5.080	4.378	3.909	3.576	3.327	3.133	2.979	2.749	2.584	2.461	2.366	2.092	1.962	1.885	1.735	1.647	1.617
29	1001	39.46	14.09	8.468	6.234	5.072	4.370	3.901	3.568	3.319	3.125	2.971	2.740	2.576	2.453	2.357	2.083	1.952	1.875	1.725	1.636	1.606
30	1001	39.46	14.08	8.461	6.227	5.065	4.362	3.894	3.560	3.311	3.118	2.963	2.732	2.568	2.445	2.349	2.074	1.943	1.866	1.715	1.625	1.596
35	1004	39.47	14.06	8.433	6.197	5.035	4.332	3.863	3.529	3.279	3.086	2.931	2.699	2.534	2.410	2.314	2.037	1.905	1.827	1.673	1.581	1.551
40	1006	39.47	14.04	8.411	6.175	5.012	4.309	3.840	3.505	3.255	3.061	2.906	2.674	2.509	2.384	2.287	2.009	1.875	1.796	1.640	1.546	1.515
45	1007	39.48	14.02	8.394	6.158	4.995	4.291	3.821	3.487	3.237	3.042	2.887	2.654	2.488	2.364	2.266	1.986	1.852	1.772	1.614	1.518	1.486
50	1008	39.48	14.01	8.381	6.144	4.980	4.276	3.807	3.472	3.221	3.027	2.871	2.638	2.472	2.347	2.249	1.968	1.832	1.752	1.592	1.495	1.462
55	1009	39.48	14.00	8.370	6.132	4.969	4.264	3.795	3.460	3.209	3.014	2.859	2.625	2.458	2.333	2.235	1.953	1.816	1.735	1.573	1.474	1.441
60	1010	39.48	13.99	8.360	6.123	4.959	4.254	3.784	3.449	3.198	3.004	2.848	2.614	2.447	2.321	2.223	1.940	1.803	1.721	1.558	1.457	1.423
65	1010	39.48	13.99	8.353	6.114	4.951	4.246	3.776	3.441	3.190	2.994	2.839	2.605	2.437	2.312	2.213	1.929	1.791	1.709	1.544	1.442	1.407
70	1011	39.48	13.98	8.346	6.107	4.943	4.239	3.768	3.433	3.182	2.987	2.831	2.597	2.429	2.303	2.205	1.920	1.781	1.698	1.532	1.429	1.394
75	1011	39.48	13.97	8.340	6.101	4.937	4.232	3.762	3.426	3.175	2.980	2.824	2.590	2.422	2.296	2.197	1.911	1.772	1.689	1.522	1.417	1.381
80	1012	39.49	13.97	8.335	6.096	4.932	4.227	3.756	3.421	3.169	2.974	2.818	2.583	2.415	2.289	2.190	1.904	1.764	1.681	1.512	1.407	1.370
85	1012	39.49	13.97	8.330	6.091	4.927	4.222	3.751	3.416	3.164	2.969	2.812	2.578	2.410	2.283	2.184	1.897	1.757	1.674	1.504	1.397	1.360
90	1013	39.49	13.96	8.326	6.087	4.923	4.218	3.747	3.411	3.160	2.964	2.808	2.573	2.405	2.278	2.179	1.892	1.751	1.667	1.496	1.389	1.351
95	1013	39.49	13.96	8.323	6.083	4.919	4.214	3.743	3.407	3.155	2.960	2.803	2.569	2.400	2.273	2.174	1.886	1.746	1.661	1.489	1.381	1.343
100	1013	39.49	13.96	8.319	6.080	4.915	4.210	3.739	3.403	3.152	2.956	2.800	2.565	2.396	2.269	2.170	1.882	1.741	1.656	1.483	1.374	1.336
250	1016	39.49	13.92	8.282	6.041	4.876	4.170	3.698	3.361	3.109	2.912	2.755	2.519	2.349	2.220	2.120	1.826	1.680	1.592	1.407	1.282	1.235
500	1017	39.50	13.91	8.270	6.028	4.862	4.156	3.684	3.347	3.094	2.898	2.740	2.503	2.333	2.204	2.103	1.806	1.659	1.569	1.378	1.245	1.192
1000	1018	39.50	13.91	8.264	6.022	4.856	4.149	3.677	3.340	3.087	2.890	2.733	2.495	2.324	2.195	2.094	1.797	1.648	1.557	1.363	1.224	1.166
10000	1018	39.50	13.90	8.258	6.016	4.850	4.143	3.671	3.334	3.081	2.884	2.726	2.488	2.317	2.188	2.086	1.788	1.638	1.546	1.349	1.204	1.140

$\alpha = ,05$	1	2	3	4	5	6	7	8	9	10	11	12	14	16	18	20	30	40	50	100	250	500
1	161.4	18.51	10.13	7.709	6.608	5.987	5.591	5.318	5.117	4.965	4.844	4.747	4.600	4.494	4.414	4.351	4.171	4.085	4.034	3.936	3.879	3.860
2	199.5	19.00	9.552	6.944	5.786	5.143	4.737	4.459	4.256	4.103	3.982	3.885	3.739	3.634	3.555	3.493	3.316	3.232	3.183	3.087	3.032	3.014
3	215.7	19.16	9.277	6.591	5.409	4.757	4.347	4.066	3.863	3.708	3.587	3.490	3.344	3.239	3.160	3.098	2.922	2.839	2.790	2.696	2.641	2.623
4	224.6	19.25	9.117	6.388	5.192	4.534	4.120	3.838	3.633	3.478	3.357	3.259	3.112	3.007	2.928	2.866	2.690	2.606	2.557	2.463	2.408	2.390
5	230.2	19.30	9.013	6.256	5.050	4.387	3.972	3.687	3.482	3.326	3.204	3.106	2.958	2.852	2.773	2.711	2.534	2.449	2.400	2.305	2.250	2.232
6	234.0	19.33	8.941	6.163	4.950	4.284	3.866	3.581	3.374	3.217	3.095	2.996	2.848	2.741	2.661	2.599	2.421	2.336	2.286	2.191	2.135	2.117
7	236.8	19.35	8.887	6.094	4.876	4.207	3.787	3.500	3.293	3.135	3.012	2.913	2.764	2.657	2.577	2.514	2.334	2.249	2.199	2.103	2.046	2.028
8	238.9	19.37	8.845	6.041	4.818	4.147	3.726	3.438	3.230	3.072	2.948	2.849	2.699	2.591	2.510	2.447	2.266	2.180	2.130	2.032	1.976	1.957
9	240.5	19.38	8.812	5.999	4.772	4.099	3.677	3.388	3.179	3.020	2.896	2.796	2.646	2.538	2.456	2.393	2.211	2.124	2.073	1.975	1.917	1.899
10	241.9	19.40	8.786	5.964	4.735	4.060	3.637	3.347	3.137	2.978	2.854	2.753	2.602	2.494	2.412	2.348	2.165	2.077	2.026	1.927	1.869	1.850
11	243.0	19.40	8.763	5.936	4.704	4.027	3.603	3.313	3.102	2.943	2.818	2.717	2.565	2.456	2.374	2.310	2.126	2.038	1.986	1.886	1.827	1.808
12	243.9	19.41	8.745	5.912	4.678	4.000	3.575	3.284	3.073	2.913	2.788	2.687	2.534	2.425	2.342	2.278	2.092	2.003	1.952	1.850	1.791	1.772
13	244.7	19.42	8.729	5.891	4.655	3.976	3.550	3.259	3.048	2.887	2.761	2.660	2.507	2.397	2.314	2.250	2.063	1.974	1.921	1.819	1.759	1.740
14	245.4	19.42	8.715	5.873	4.636	3.956	3.529	3.237	3.025	2.865	2.739	2.637	2.484	2.373	2.290	2.225	2.037	1.948	1.895	1.792	1.732	1.712
15	245.9	19.43	8.703	5.858	4.619	3.938	3.511	3.218	3.006	2.845	2.719	2.617	2.463	2.352	2.269	2.203	2.015	1.924	1.871	1.768	1.707	1.686
16	246.5	19.43	8.692	5.844	4.604	3.922	3.494	3.202	2.989	2.828	2.701	2.599	2.445	2.333	2.250	2.184	1.995	1.904	1.850	1.746	1.684	1.664
17	246.9	19.44	8.683	5.832	4.590	3.908	3.480	3.187	2.974	2.812	2.685	2.583	2.428	2.317	2.233	2.167	1.976	1.885	1.831	1.726	1.664	1.643
18	247.3	19.44	8.675	5.821	4.579	3.896	3.467	3.173	2.960	2.798	2.671	2.568	2.413	2.302	2.217	2.151	1.960	1.868	1.814	1.708	1.645	1.625
19	247.7	19.44	8.667	5.811	4.568	3.884	3.455	3.161	2.948	2.785	2.658	2.555	2.400	2.288	2.203	2.137	1.945	1.853	1.798	1.691	1.628	1.607
20	248.0	19.45	8.660	5.803	4.558	3.874	3.445	3.150	2.936	2.774	2.646	2.544	2.388	2.276	2.191	2.124	1.932	1.839	1.784	1.676	1.613	1.592
21	248.3	19.45	8.654	5.795	4.549	3.865	3.435	3.140	2.926	2.764	2.636	2.533	2.377	2.264	2.179	2.112	1.919	1.826	1.771	1.663	1.598	1.577
22	248.6	19.45	8.648	5.787	4.541	3.856	3.426	3.131	2.917	2.754	2.626	2.523	2.367	2.254	2.168	2.102	1.908	1.814	1.759	1.650	1.585	1.563
23	248.8	19.45	8.643	5.781	4.534	3.849	3.418	3.123	2.908	2.745	2.617	2.514	2.357	2.244	2.159	2.092	1.897	1.803	1.748	1.638	1.573	1.551
24	249.1	19.45	8.639	5.774	4.527	3.841	3.410	3.115	2.900	2.737	2.609	2.505	2.349	2.235	2.150	2.082	1.887	1.793	1.737	1.627	1.561	1.539
25	249.3	19.46	8.634	5.769	4.521	3.835	3.404	3.108	2.893	2.730	2.601	2.498	2.341	2.227	2.141	2.074	1.878	1.783	1.727	1.616	1.550	1.528
26	249.5	19.46	8.630	5.763	4.515	3.829	3.397	3.102	2.886	2.723	2.594	2.491	2.333	2.220	2.134	2.066	1.870	1.775	1.718	1.607	1.540	1.518
27	249.6	19.46	8.626	5.759	4.510	3.823	3.391	3.095	2.880	2.716	2.588	2.484	2.326	2.212	2.126	2.059	1.862	1.766	1.710	1.598	1.530	1.508
28	249.8	19.46	8.623	5.754	4.505	3.818	3.386	3.090	2.874	2.710	2.582	2.478	2.320	2.206	2.119	2.052	1.854	1.759	1.702	1.589	1.521	1.499
29	250.0	19.46	8.620	5.750	4.500	3.813	3.381	3.084	2.869	2.705	2.576	2.472	2.314	2.200	2.113	2.045	1.847	1.751	1.694	1.581	1.513	1.490
30	250.1	19.46	8.617	5.746	4.496	3.808	3.376	3.079	2.864	2.700	2.570	2.466	2.308	2.194	2.107	2.039	1.841	1.744	1.687	1.573	1.505	1.482
35	250.7	19.47	8.604	5.729	4.478	3.789	3.356	3.059	2.842	2.678	2.548	2.443	2.284	2.169	2.082	2.013	1.813	1.715	1.657	1.541	1.470	1.447
40	251.1	19.47	8.594	5.717	4.464	3.774	3.340	3.043	2.826	2.661	2.531	2.426	2.266	2.151	2.063	1.994	1.792	1.693	1.634	1.515	1.443	1.419
45	251.5	19.47	8.587	5.707	4.453	3.763	3.328	3.030	2.813	2.648	2.517	2.412	2.252	2.136	2.048	1.978	1.775	1.675	1.615	1.494	1.421	1.396
50	251.8	19.48	8.581	5.699	4.444	3.754	3.319	3.020	2.803	2.637	2.507	2.401	2.241	2.124	2.035	1.966	1.761	1.660	1.599	1.477	1.402	1.376
55	252.0	19.48	8.576	5.693	4.437	3.746	3.311	3.012	2.794	2.628	2.498	2.392	2.231	2.114	2.025	1.955	1.749	1.648	1.587	1.463	1.386	1.360
60	252.2	19.48	8.572	5.688	4.431	3.740	3.304	3.005	2.787	2.621	2.490	2.384	2.223	2.106	2.017	1.946	1.740	1.637	1.576	1.450	1.372	1.345
65	252.4	19.48	8.569	5.683	4.426	3.734	3.299	2.999	2.781	2.615	2.484	2.378	2.216	2.099	2.009	1.939	1.731	1.628	1.566	1.440	1.360	1.333
70	252.5	19.48	8.566	5.679	4.422	3.730	3.294	2.994	2.776	2.610	2.478	2.372	2.210	2.093	2.003	1.932	1.724	1.621	1.558	1.430	1.350	1.322
75	252.6	19.48	8.563	5.676	4.418	3.726	3.290	2.990	2.771	2.605	2.473	2.367	2.205	2.087	1.998	1.927	1.718	1.614	1.551	1.422	1.341	1.312
80	252.7	19.48	8.561	5.673	4.415	3.722	3.286	2.986	2.768	2.601	2.469	2.363	2.201	2.083	1.993	1.922	1.712	1.608	1.544	1.415	1.332	1.303
85	252.8	19.48	8.559	5.670	4.412	3.719	3.283	2.983	2.764	2.597	2.466	2.359	2.197	2.078	1.988	1.917	1.707	1.602	1.539	1.408	1.325	1.295
90	252.9	19.48	8.557	5.668	4.409	3.716	3.280	2.980	2.761	2.594	2.462	2.356	2.193	2.075	1.985	1.913	1.703	1.597	1.534	1.402	1.318	1.288
95	253.0	19.49	8.555	5.666	4.407	3.714	3.277	2.977	2.758	2.591	2.459	2.353	2.190	2.071	1.981	1.910	1.699	1.593	1.529	1.397	1.312	1.281
100	253.0	19.49	8.554	5.664	4.405	3.712	3.275	2.975	2.756	2.588	2.457	2.350	2.187	2.068	1.978	1.907	1.695	1.589	1.525	1.392	1.306	1.275
250	253.8	19.49	8.537	5.643	4.381	3.686	3.248	2.947	2.726	2.558	2.426	2.318	2.154	2.034	1.942	1.869	1.652	1.542	1.475	1.331	1.232	1.194
500	254.1	19.49	8.532	5.635	4.373	3.678	3.239	2.937	2.717	2.548	2.415	2.307	2.142	2.022	1.929	1.856	1.637	1.526	1.457	1.308	1.202	1.159
1000	254.2	19.49	8.529	5.632	4.369	3.673	3.234	2.932	2.712	2.543	2.410	2.302	2.136	2.016	1.923	1.850	1.630	1.517	1.448	1.296	1.185	1.138
10000	254.3	19.50	8.527	5.628	4.365	3.669	3.230	2.928	2.707	2.538	2.405	2.297	2.131	2.010	1.917	1.844	1.623	1.510	1.439	1.284	1.168	1.116

$\alpha = ,1$	1	2	3	4	5	6	7	8	9	10	11	12	14	16	18	20	30	40	50	100	250	500
1	39.86	8.526	5.538	4.545	4.060	3.776	3.589	3.458	3.360	3.285	3.225	3.177	3.102	3.048	3.007	2.975	2.881	2.835	2.809	2.756	2.726	2.716
2	49.50	9.000	5.462	4.325	3.780	3.463	3.257	3.113	3.006	2.924	2.860	2.807	2.726	2.668	2.624	2.589	2.489	2.440	2.412	2.356	2.324	2.313
3	53.59	9.162	5.391	4.191	3.619	3.289	3.074	2.924	2.813	2.728	2.660	2.606	2.522	2.462	2.416	2.380	2.276	2.226	2.197	2.139	2.106	2.095
4	55.83	9.243	5.343	4.107	3.520	3.181	2.961	2.806	2.693	2.605	2.536	2.480	2.395	2.333	2.286	2.249	2.142	2.091	2.061	2.002	1.967	1.956
5	57.24	9.293	5.309	4.051	3.453	3.108	2.883	2.726	2.611	2.522	2.451	2.394	2.307	2.244	2.196	2.158	2.049	1.997	1.966	1.906	1.870	1.859
6	58.20	9.326	5.285	4.010	3.405	3.055	2.827	2.668	2.551	2.461	2.389	2.331	2.243	2.178	2.130	2.091	1.980	1.927	1.895	1.834	1.798	1.786
7	58.91	9.349	5.266	3.979	3.368	3.014	2.785	2.624	2.505	2.414	2.342	2.283	2.193	2.128	2.079	2.040	1.927	1.873	1.840	1.778	1.741	1.729
8	59.44	9.367	5.252	3.955	3.339	2.983	2.752	2.589	2.469	2.377	2.304	2.245	2.154	2.088	2.038	1.999	1.884	1.829	1.796	1.732	1.695	1.683
9	59.86	9.381	5.240	3.936	3.316	2.958	2.725	2.561	2.440	2.347	2.274	2.214	2.122	2.055	2.005	1.965	1.849	1.793	1.760	1.695	1.657	1.644
10	60.19	9.392	5.230	3.920	3.297	2.937	2.703	2.538	2.416	2.323	2.248	2.188	2.095	2.028	1.977	1.937	1.819	1.763	1.729	1.663	1.624	1.612
11	60.47	9.401	5.222	3.907	3.282	2.920	2.684	2.519	2.396	2.302	2.227	2.166	2.073	2.005	1.954	1.913	1.794	1.737	1.703	1.636	1.597	1.583
12	60.71	9.408	5.216	3.896	3.268	2.905	2.668	2.502	2.379	2.284	2.209	2.147	2.054	1.985	1.933	1.892	1.773	1.715	1.680	1.612	1.572	1.559
13	60.90	9.415	5.210	3.886	3.257	2.892	2.654	2.488	2.364	2.269	2.193	2.131	2.037	1.968	1.916	1.875	1.754	1.695	1.660	1.592	1.551	1.537
14	61.07	9.420	5.205	3.878	3.247	2.881	2.643	2.475	2.351	2.255	2.179	2.117	2.022	1.953	1.900	1.859	1.737	1.678	1.643	1.573	1.532	1.518
15	61.22	9.425	5.200	3.870	3.238	2.871	2.632	2.464	2.340	2.244	2.167	2.105	2.010	1.940	1.887	1.845	1.722	1.662	1.627	1.557	1.515	1.501
16	61.35	9.429	5.196	3.864	3.230	2.863	2.623	2.455	2.329	2.233	2.156	2.094	1.998	1.928	1.875	1.833	1.709	1.649	1.613	1.542	1.499	1.485
17	61.46	9.433	5.193	3.858	3.223	2.855	2.615	2.446	2.320	2.224	2.147	2.084	1.988	1.917	1.864	1.821	1.697	1.636	1.600	1.528	1.485	1.471
18	61.57	9.436	5.190	3.853	3.217	2.848	2.607	2.438	2.312	2.215	2.138	2.075	1.978	1.908	1.854	1.811	1.686	1.625	1.588	1.516	1.473	1.458
19	61.66	9.439	5.187	3.849	3.212	2.842	2.601	2.431	2.305	2.208	2.130	2.067	1.970	1.899	1.845	1.802	1.676	1.615	1.578	1.505	1.461	1.446
20	61.74	9.441	5.184	3.844	3.207	2.836	2.595	2.425	2.298	2.201	2.123	2.060	1.962	1.891	1.837	1.794	1.667	1.605	1.568	1.494	1.450	1.435
21	61.81	9.444	5.182	3.841	3.202	2.831	2.589	2.419	2.292	2.194	2.117	2.053	1.955	1.884	1.829	1.786	1.659	1.596	1.559	1.485	1.440	1.425
22	61.88	9.446	5.180	3.837	3.198	2.827	2.584	2.413	2.287	2.189	2.111	2.047	1.949	1.877	1.823	1.779	1.651	1.588	1.551	1.476	1.431	1.416
23	61.95	9.448	5.178	3.834	3.194	2.822	2.580	2.409	2.282	2.183	2.105	2.041	1.943	1.871	1.816	1.773	1.644	1.581	1.543	1.468	1.422	1.407
24	62.00	9.450	5.176	3.831	3.191	2.818	2.575	2.404	2.277	2.178	2.100	2.036	1.938	1.866	1.810	1.767	1.638	1.574	1.536	1.460	1.414	1.399
25	62.05	9.451	5.175	3.828	3.187	2.815	2.571	2.400	2.272	2.174	2.095	2.031	1.933	1.860	1.805	1.761	1.632	1.568	1.529	1.453	1.406	1.391
26	62.10	9.453	5.173	3.826	3.184	2.811	2.568	2.396	2.268	2.170	2.091	2.027	1.928	1.855	1.800	1.756	1.626	1.562	1.523	1.446	1.399	1.384
27	62.15	9.454	5.172	3.823	3.181	2.808	2.564	2.392	2.265	2.166	2.087	2.022	1.923	1.851	1.795	1.751	1.621	1.556	1.517	1.440	1.393	1.377
28	62.19	9.456	5.170	3.821	3.179	2.805	2.561	2.389	2.261	2.162	2.083	2.019	1.919	1.847	1.791	1.746	1.616	1.551	1.512	1.434	1.386	1.370
29	62.23	9.457	5.169	3.819	3.176	2.803	2.558	2.386	2.258	2.159	2.080	2.015	1.916	1.843	1.787	1.742	1.611	1.546	1.507	1.428	1.380	1.364
30	62.26	9.458	5.168	3.817	3.174	2.800	2.555	2.383	2.255	2.155	2.076	2.011	1.912	1.839	1.783	1.738	1.606	1.541	1.502	1.423	1.375	1.358
35	62.42	9.463	5.163	3.810	3.165	2.789	2.544	2.371	2.242	2.142	2.062	1.997	1.897	1.823	1.766	1.721	1.588	1.521	1.481	1.400	1.350	1.333
40	62.53	9.466	5.160	3.804	3.157	2.781	2.535	2.361	2.232	2.132	2.052	1.986	1.885	1.811	1.754	1.708	1.573	1.506	1.465	1.382	1.330	1.313
45	62.62	9.469	5.157	3.799	3.152	2.775	2.528	2.354	2.224	2.124	2.043	1.977	1.876	1.801	1.744	1.698	1.562	1.493	1.452	1.367	1.314	1.296
50	62.69	9.471	5.155	3.795	3.147	2.770	2.523	2.348	2.218	2.117	2.036	1.970	1.869	1.793	1.736	1.690	1.552	1.483	1.441	1.355	1.301	1.282
55	62.75	9.473	5.153	3.792	3.143	2.765	2.518	2.343	2.213	2.112	2.031	1.965	1.862	1.787	1.729	1.683	1.544	1.474	1.432	1.344	1.289	1.270
60	62.79	9.475	5.151	3.790	3.140	2.762	2.514	2.339	2.208	2.107	2.026	1.960	1.857	1.782	1.723	1.677	1.538	1.467	1.424	1.336	1.279	1.260
65	62.84	9.476	5.150	3.787	3.138	2.759	2.511	2.336	2.205	2.103	2.022	1.956	1.853	1.777	1.718	1.672	1.532	1.461	1.418	1.328	1.271	1.251
70	62.87	9.477	5.149	3.786	3.135	2.756	2.508	2.333	2.202	2.100	2.019	1.952	1.849	1.773	1.714	1.667	1.527	1.455	1.412	1.321	1.263	1.243
75	62.90	9.478	5.148	3.784	3.133	2.754	2.506	2.330	2.199	2.097	2.016	1.949	1.846	1.769	1.711	1.664	1.523	1.451	1.407	1.315	1.256	1.236
80	62.93	9.479	5.147	3.782	3.132	2.752	2.504	2.328	2.196	2.095	2.013	1.946	1.843	1.766	1.707	1.660	1.519	1.447	1.402	1.310	1.250	1.229
85	62.95	9.479	5.146	3.781	3.130	2.750	2.502	2.326	2.194	2.092	2.011	1.944	1.840	1.764	1.704	1.657	1.515	1.443	1.398	1.305	1.245	1.223
90	62.97	9.480	5.145	3.780	3.129	2.749	2.500	2.324	2.192	2.090	2.009	1.942	1.838	1.761	1.702	1.655	1.512	1.439	1.395	1.301	1.240	1.218
95	62.99	9.481	5.145	3.779	3.127	2.748	2.498	2.322	2.191	2.089	2.007	1.940	1.836	1.759	1.700	1.652	1.509	1.436	1.391	1.297	1.235	1.213
100	63.01	9.481	5.144	3.778	3.126	2.746	2.497	2.321	2.189	2.087	2.005	1.938	1.834	1.757	1.698	1.650	1.507	1.434	1.388	1.293	1.231	1.209
250	63.20	9.487	5.138	3.768	3.114	2.732	2.481	2.304	2.171	2.068	1.985	1.918	1.812	1.734	1.673	1.625	1.477	1.401	1.353	1.249	1.176	1.148
500	63.26	9.489	5.136	3.764	3.109	2.727	2.476	2.298	2.165	2.062	1.979	1.911	1.805	1.726	1.665	1.616	1.467	1.389	1.340	1.232	1.154	1.122
1000	63.30	9.490	5.135	3.762	3.107	2.725	2.473	2.295	2.162	2.059	1.975	1.907	1.801	1.722	1.661	1.612	1.462	1.383	1.333	1.223	1.141	1.106
10000	63.32	9.491	5.134	3.761	3.105	2.722	2.471	2.293	2.160	2.056	1.972	1.904	1.798	1.719	1.657	1.608	1.457	1.378	1.327	1.215	1.129	1.089

Tabla de Contenido

Referencias

- Berenson, M., & Levine, D. M. (1993). *Estadística para Administración y Economía. Conceptos y Aplicaciones.* Mexico: McGraw-Hill/ Intermericano de México S.A.
- Berenson, M., & Levine, D. M. (1996). *Estadística Básica en Administración.* México: Prentice Hall.
- Carrizo, J. F., & Carrizo, J. F. (1977). *Nociones de Inferencia Estadística.* Córdoba.
- Chou, Y.-L. (1977). *Análisis Estadístico (Segunda ed.).* México: Nueva Editorial Interamericana.
- Cochran, W. G. (1987). *Técnicas de Muestreo.* Editorial CECSA.
- Dagum, C., & Bee de Dagum, E. M. (1971). *Introducción a la econometría (10 ed.).* México: Editorial Siglo XXI.
- Daniel, W. W. (1999). *Bioestadística, base para el análisis de las ciencias de la salud. (Tercera ed.).* México: Editorial Limussa.
- Dixon, W. J., & Massey, F. J. (1957). *Introduction to Statistical Analysis.* Nueva York: McGraw-Hill.
- Espasa, A., & Cancelo, J. R. (1993). *Métodos cuantitativos para el análisis de la coyuntura.* Madrid: Alianza Editorial SA.
- Fine, D. (1997). *Metodología de la Encuesta.* Chile: Programa Presta – ULB – Belgica – Universidad de Concepción.
- Freeman, H. (1963). *Introducción a la Inferencia Estadística.* México: Trillas.
- Gomez Villegas, M. A. (2005). *Inferencia Estadística.* España: Ediciones Díaz de Santo.
- Gujarati, D. N. (2004). *Econometría (cuarta ed.).* México: McGraw-Gill Interamericana editores, S.A. de C.V.
- Gujarati, D., & Porter, D. (2010). *Econometría (Quinta Edición).* México: McGraw Hill.
- Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, M. d. (2010). *Metodología de la Investigación (Quinta ed.).* México: McGraw-Hill/ Interamericana editores, S.A. de C.V.

- Hildebrand, D. K., & Ott, L. (1997). *Estadística Aplicada a la Administración y a la Economía.* Wilmington Delaware, E.U.A.: Addison-Wesley Iberoamericana, S.A.
- Johnston, J., & Dinardo, J. (2001). *Métodos De Econometría.* Barcelona: Editorial Vicens Vives.
- Kazmier, L., & Díaz Mata, A. (1993). *Estaística aplicada a la Administración y a la Economía.* México: Mc. Graw Hill.
- Kinnear, T., & Taylor, J. (1993). *Investigación de Mercado.* Mc. Graw Hill.
- Kmenta, J. (s.f.). *Elementos de Econometría.* Vicens Universidad.
- Leontief, W. (1971). *Theoretical Assumptions and Non-Observed Facts.*
- Loria, E. (2007). *Econometría con aplicaciones.* México: Pearson Prentice Hall.
- Lucas, R. E. (1976). *Econometric Policy Evaluation: A Critique.*
- Mao, J. C. (1980). *Análisis Financiero (Cuarta ed.).* Buenos Aires: El Ateneo.
- Marques Dos Santos, María José (2005) Cálculo de la probabilidad del error tipo I y tipo II y de la potencia de la prueba. En Home Page José Luis García Cué. Probabilidad y Estadística. Colegio de Postgraduados, FES Zaragoza UNAM.
- Mendenhall, W., Wackerly, D., & Scheaffer, R. (1990). *Estadística Matemática con Aplicaciones.* México: Grupo Editorial Iberoamérica.
- Meyer, P. L. (1973). *Probabilidad y aplicaciones estadísticas.* México: Fondo Educativo Interamericano SA.
- Mood, A. M., & Graybill, F. A. (1972). *Introducción a la Estadística.* Editorial Aguilar.
- Olivera Zalazar, A., & Zúñiga Barrera, S. (1987). *Muestreo. Serie de Probabilidad y Estadística 5.* México: Limusa.
- Padua, J. (1996). *Técnicas de Investigación Aplicadas a las Ciencias Sociales (Septima reinpresión).* México: Fondo de Cultura Económica.
- Pérez López, C. (2005). *Muestreo estadístico. Conceptos y Problemas resueltos.* España: Editorial Pearson Prentice Hall.
- Pugachev, V. S. (1973). *Introducción a la Teoría de las Probabilidades.* Moscú: Editorial Mir.
- Ross, S. M. (2007). *Introducción a la Estadística.* Barcelona: Editorial Reverté.
- Spiegel, M. R. (1976). *Teoría y problemas de probabilidad y estadística.* México: Mc. Graw Hill.

- Tramutola, C. D. (1971). *Modelos probabilísticos y decisiones financieras.* Capital Federal: E.C.Moderan.
- Wackerly, D., Mendenhall, W., & Scheaffer, R. (2008). *Estadística Matemática con Aplicaciones (Séptima edición).* México: Cencage Learning.