Capítulo 6 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 el capítulo anterior se estudiaron las características de la media muestral en tanto estadístico útil para estimar la media de la población. En éste se estudiarán 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 se pueden 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.

6.1 Teoría estadística

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 la o 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 $\overline{X}$ que, por ser una característica muestral, cambiará si se obtuviera una nueva muestra.

Otra característica es la varianza de la variable en la muestra, S^2. Esta será pequeña si las respuestas de la muestra 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.

Ejemplo. El interés está centrado en conocer la media (o promedio) de la variable ingreso de los 57 clientes de un banco que poseen la tarjeta de crédito AA. La población son los 57 clientes y los datos se ordenan en 8 intervalos de clase de acuerdo al ingreso que posea cada cliente.

Intervalo de clase	Punto medio de la clase m_i	Frecuencia de clase Absoluta f_i	$m_i f_i$	$(m_i - \mu)$	$(m_i - \mu)f$	$(m_i - \mu)^2 f_i$
De \$150.00 a \$300.99	225.00	6	1350.00	-606.58	-3639.48	2207635.78
De \$301.00 a \$500.99	400.00	13	5200.00	-431.58	-5610.54	2421396.85
De \$501.00 a \$700.99	600.00	13	7800.00	-231.58	-3010.54	697180.85
De \$701.00 a \$1000.99	850.00	8	6800.00	18.42	147.36	2714.37
De \$1001.00 a \$1250.99	1125.00	5	5625.00	293.42	1467.10	430476.48
De \$1251.00 a \$1500.99	1375.00	3	4125.00	543.42	1630.26	885915.89
De \$1501.00 a \$2000.99	1750.00	6	10500.00	918.42	5510.52	5060971.78
Más de \$2000.99	2000.00	3	6000.00	1168.42	3505.26	4095615.89
Sumas		57	47400		0.00	15801907.89

Figura 6.1. Cálculo de la media desde el intervalo de clase

La medida promedio son los ingresos de los 57 clientes del banco que poseen la tarjeta AA, lo cual se denomina media de la variable ingreso de la población.

$$\mu = \frac{\sum_{i=1}^{N} m_i f_i}{\sum_{i=1}^{N} f_i} = 831.58$$

Calculada la media de la variable ingreso en la población, se pude calcular la varianza de la misma. Esta es la suma de las distancias cuadráticas promedio que separa el ingreso observado para cada cliente respecto del ingreso promedio de la población, μ; esto es, de la media de la variable ingreso.

$$\sigma^2 = \frac{\sum_{i=1}^{N} (m_i - \mu)^2 f_i}{\sum_{i=1}^{N} f_i} = 277226{,}45$$

El desvío estándar de la variable ingreso en la población es la raíz cuadrada de la varianza

$$\sigma = \sqrt{\sigma^2} = \sqrt{277226{,}45} \cong 526.52$$

Ejemplo. Se toma una muestra aleatoria, de tamaño 10, de los clientes del banco que poseen la tarjeta de crédito AA. Las diez personas seleccionadas y el ingreso respectivo se muestran en la Figura 6.2.

La media de la variable en la muestra es

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{7930.00}{10} = 793.00$$

La varianza de la variable en la muestra

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1} = \frac{1851410.00}{9} = 205712.22$$

El desvío de la variable en la muestra

$$s = \sqrt{s^2} \cong 453.56$$

	x_i	$x_i - \mu$	$(x - \mu)^2$
1	1700.00	907.00	822649.00
2	900.00	107.00	11449.00
3	550.00	-243.00	59049.00
4	1300.00	507.00	257049.00
5	300.00	-493.00	243049.00
6	600.00	-193.00	37249.00
7	350.00	-443.00	196249.00
8	1100.00	307.00	94249.00
9	450.00	-343.00	117649.00
10	680.00	-113.00	12769.00
$\sum$	7930.00	0.00	1851410.00

Figura 6.2. Cálculo de la media desde las observaciones muestrales

S^2 es una estimación sin sesgo de la varianza de la variable en la población porque incluye el término n-1.

Si el tamaño de la población, denominado N, es pequeño con relación al tamaño de la muestra n, debe ser añadido un factor de corrección de población finita

$$\frac{N-n}{N} = 1 - f$$

Donde $f = \frac{n}{N}$ se lo denomina fracción de muestreo.

De modo que la varianza de la variable en la muestra es

$$S^2 = \frac{N-n}{N}\left[\frac{\sum(X_i - \bar{X})^2}{n-1}\right]$$

Si N es más de 10 veces el tamaño de la muestra, el factor de corrección tiene rara vez significado.

Desde luego, todas las muestras no generarán el mismo valor de $\bar{X}$ (ó s). El punto de importancia es que $\bar{X}$ variará de muestra a muestra.

La varianza de $\overline{X}$ es medida por su error estándar:

$\sigma_{\overline{X}}$ = error estándar de $\overline{x}$ = $\sigma / \sqrt{n}$

La varianza en $\overline{X}$ 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 $\overline{X}$ disminuirá.

En el capítulo anterior se demostró que la media de la variable en la muestra, $\overline{X}$, tiene una distribución de probabilidad normal (con media igual a μ y varianza igual a $\sigma^2_{\overline{X}}$), cualquiera sea la distribución de probabilidad de la variable en la población; aunque si esta no es normal se requerirá que la muestra sea lo suficientemente grande de acuerdo al Teorema del Límite Central.

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

De acuerdo a esto, existe un nivel de confianza de 0.95 (95%) de que $\overline{x}$ caiga dentro de $\pm$1.96 errores 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.

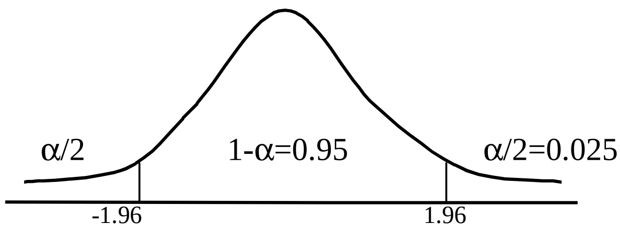


Figura 6.3

Ejemplo. En el estudio de la variable ingresos de los clientes del banco que poseen la tarjeta de crédito AA, si se han obtenido 100 muestras diferentes de 10 individuos, aproximadamente el 95% de las medias muestrales resultantes $(\bar{X})$ estarán dentro de ± 1.96 errores estándar $(\sigma_{\bar{X}} = 166.50)$ de la media de la población $(\mu = 831.58)$

6.2 Tamaño de la muestra

Para determinar el tamaño de la muestra debe especificarse

1) El tamaño del error de la muestra que se desea.

2) El nivel de confianza, por ejemplo, del 95%.

El error de la muestra es

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

de donde se deduce el tamaño de la muestra

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

Ejemplo. Si existe la necesidad de estar un 95% seguro de que, al estimar la media de una variable en la población, el *error muestral* no exceda de 100 pesos conociendo que la desviación estándar de la variable en la población es 480.50; el error muestral (e) es igual a 100 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:

$$n = \frac{(1.96)^2(480.50)^2}{100^2} \cong 89$$

El tamaño de la muestra es independiente del tamaño de la población y dependiente de:

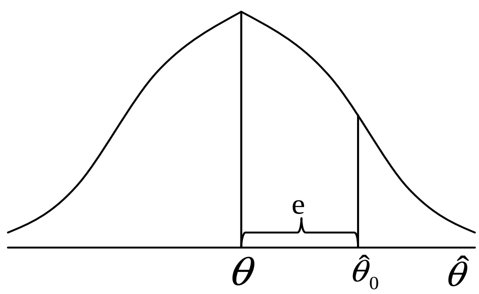
Nivel de Confianza	$> z \Rightarrow > n$
Desviación estándar de la población	$> \sigma \Rightarrow > n$
Error muestral	$> e \Rightarrow < n$

Error, riesgo y tamaño muestral

Dado el estimador $\hat{\theta}$ que se distribuye normal con parámetros $(\theta, \sigma_{\hat{\theta}})$ y es insesgado del parámetro θ

$$\hat{\theta} \sim N(\theta, \sigma_{\hat{\theta}}) \qquad E(\hat{\theta}) = \theta\,.$$

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



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

Figura 6.4

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 $|\hat{\theta} - \theta| > e$. El riesgo es la probabilidad de que la diferencia entre el parámetro y su estimador supere la cuantía de e.

$P(|\hat{\theta} - \theta| > e) =$ riesgo de cometer un error superior a *e*.

Ejemplo. Suponga que $X \sim N(500{,}30)$, $n = 100$, $|e| > 6$ y $\hat{\mu} = \overline{X}$. Se quiere saber cuál es el riesgo de cometer un error de estimación de μ superior a 6.

$$z = \frac{\overline{x} - \mu}{\sigma_{\overline{X}}} = \frac{e}{\sigma / \sqrt{n}} = \frac{e\sqrt{n}}{\sigma}$$

por lo tanto

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

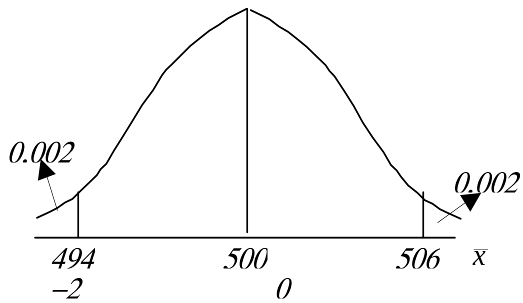


Figura 6.5

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

$P(|z| > 2) = 1 - P(-2 \leq z \leq 2) = 1 - (0.9772 - 0.0228) = 0.0456 = riesgo$

Relaciones entre error, riesgo y n

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

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

- 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. Para el tercer caso, el procedimiento es usar una proporción de la muestra *p* para *estimar* una proporción desconocida de la población, π.

El error muestral usando proporciones es de

$$e = z\sigma_p$$

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 $X_i =$ sexo. Esta es una variable dicotómica: masculino y femenino son los atributos. Supóngase que la tabla de datos para esta situación sea la de la figura 6.6.

	$X_i =$ sexo	
	M	F

1	1	0
2	1	0
3	0	1
4	0	1
$\vdots$		
i	0	1
$\vdots$		
n	1	0

Figura 6.6. Tabla lógica para la variable Sexo

Según la información de la tabla planteada,
$X_i = 1$ 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 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 "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,

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

De donde,

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

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

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

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:

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

Donde se ha tenido en cuenta el factor de corrección para poblaciones finitas.
Una estimación insesgada de la $V(p)$, derivada de la muestra, es

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

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 por $(n-1)$; pero esto es un error, ya que, aún en poblaciones infinitas, resulta una estimación sesgada de la verdadera varianza de p.

Se había establecido 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,

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

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

$$n = \frac{z^2\ \pi(1-\pi)}{e^2}$$

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 π = 0.50, la varianza es π(1 - π) = 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:

$$n = \frac{z^2\ (0.25)}{e^2}$$

De este modo, 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.

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

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.

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 6.7

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

p *porcentajes* (cpf)*con*

	5	10	15	20	25	30	35	40	45	50	n	$N = 400$ 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 6.7. Cálculo de la precisión absoluta

Se observa que la precisión es tanto mejor cuanto 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 6.8

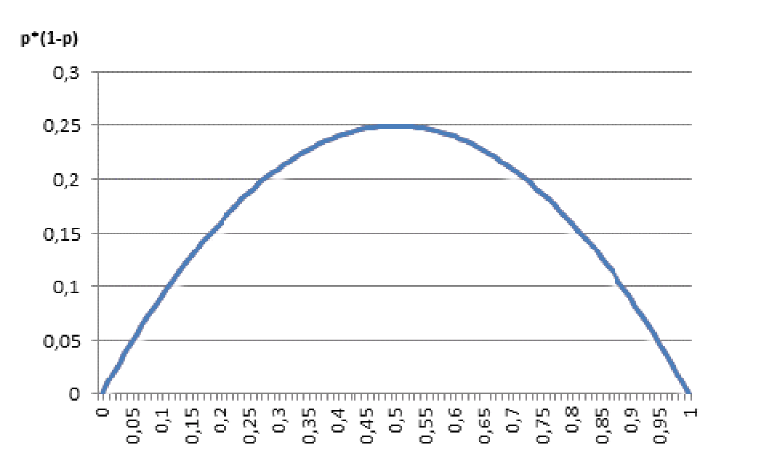


Figura 6.8. Variaciones p(1-p)

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 6.7 y es igual a $p \pm \frac{1}{\sqrt{n}}$

6.3 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 $f(\overline{x},\theta)$ 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 en el capítulo anterior 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 (**$\hat{\theta}$**) = θ.** Cuando **E (**$\hat{\theta}$**) ≠ θ**, el estimador es sesgado y se define el sesgo como:

$$Sesgo\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta$$

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

$$E\left(\overline{X}\right) \cong \mu$$

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

Generalizando, hay una población descrita por la función de densidad $f\left(\overline{x}, \theta\right)$, donde *f* es conocida y el valor del parámetro θ es desconocido.

Sea $x_1, x_2, \ldots, x_n$ una muestra aleatoria obtenida de esta población, el estadístico $t(x_1, x_2, \ldots, x_n)$ es un *estimador insesgado* del parámetro θ si $E(t) = \theta$ para todo *n* y para cualquier valor de θ. En cualquier otro caso θ es sesgado.

Ejemplo. Estimadores insesgados

a) $\mathrm{E}(\mathrm{X}_1) = \mu$, $\mathrm{E}(\overline{\mathrm{X}}) = \mu$

Si μ existe, tanto una sola observación de la variable aleatoria X, X_1, como su media, $\overline{X}$, son estimadores insesgados de μ. Sus varianzas respectivas son σ^2 y σ^2/n. Por lo tanto es preferible $\overline{X}$ 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

$$E(X_1) = X_1 P(x_1 = X_1) + X_2 P(x_1 = X_2) + \ldots + X_N P(x_1 = X_N) = \frac{1}{N}\sum_{i=1}^{N} X_i = \mu$$

b) $\mathrm{E}(\overline{\mathrm{x}} - \overline{\mathrm{y}}) = \mathrm{E}(\mathrm{x}) - \mathrm{E}(\mathrm{y})$

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

Dado $\hat{\sigma}^2 = \dfrac{\sum(\mathrm{x}_i - \overline{\mathrm{x}})^2}{\mathrm{n}}$

$$E(\hat{\sigma}^2) = \frac{1}{n} E\left\{\sum\left[x_i - E(x) + E(x) - \overline{x}\right]^2\right\}$$

$$= \frac{1}{n}\sum\left\{E[x_i - E(x)]^2 - E[\overline{x} - E(x)]^2\right\}$$

donde, $\mathrm{E}[\mathrm{x}_i - \mathrm{E}(\mathrm{x})]^2 = \sigma^2$ y $\mathrm{E}[\overline{\mathrm{x}} - \mathrm{E}(\mathrm{x})]^2 = \sigma^2_{\overline{\mathrm{x}}} = \dfrac{\sigma^2}{\mathrm{n}}$

Reemplazando

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

Por lo tanto, $\hat{\sigma}^2$ es un estimador sesgado de σ^2. Pero S^2 es <u>insesgado</u> ya que

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

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

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, 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:

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

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.

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. $\overline{X}$ no es un estimador invariante de μ ya que

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

Observación: Se sabe que

$$\begin{aligned}V(X)&=m_2-m_1^2\\&=E\left(X^2\right)-\left[E(X)\right]^2\end{aligned}$$

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

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

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

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

De este modo

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

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

d) Suficiencia

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

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

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

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

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. $\overline{X}$ es un estimador consistente de μ ya que dado $E(\overline{X}) = \mu$, $\sigma^2_{\overline{X}} = \dfrac{\sigma^2}{n}$.

Si $n \to \infty$, $\dfrac{\sigma^2}{n} \to 0$ por lo tanto $\sigma^2_{\overline{X}} \to 0$ y $\overline{X} \to \mu$.

---

b) Insesgadez asintótica

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

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

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

$$P\left(\hat{\theta}_t = \theta\right) = 1 - T^{-1} \qquad \text{y} \qquad P\left(\hat{\theta}_t = T^k\right) = T^{-1}$$

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

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

Este estimador es consistente porque:

$$\lim_{t\to\infty} P\left\{\left|\hat{\theta}_t - \theta\right| \le \varepsilon\right\} = 1 - \lim_{T\to\infty} P\left\{\left|\hat{\theta}_t - \theta\right| > \varepsilon\right\} = 1 - \lim_{T\to\infty}\frac{1}{T} = 1 - 0 = 1$$

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

$$E\left(\hat{\theta}\right) = \theta\left(1 - T^{-1}\right) + T^k \ T^{-1} = \theta - \theta \ T^{-1} + T^{k-1}$$

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

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

---

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

Ejemplo. El estimador

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

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 $\boldsymbol{\theta}$.

Estos ejemplos, 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 $\hat{\theta}$ es asintóticamente insesgado y su varianza tiende a cero conforme el tamaño muestral tiende a finito, 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.

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.

Convergencia en Distribución. Una sucesión de variables aleatorias $\hat{\theta}_t$ se dice que *converge en distribución* a $\hat{\theta}$, sí 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 $\hat{\theta}_t \overset{a}{\sim} F$.

a) Teorema de Mann-Wald

Sea **X** una matriz de dimensión (**T x K**) y **u** un vector de dimensión **T** de variables aleatorias idéntica e independientemente distribuidas, tales que:

I) $E(u) = 0_t, \; E(uu') = \sigma^2 I_t$

II) $E(X_i' u) = 0, \forall i = 1, \cdots, k,$ donde X_i es la i – ésima columna de X

III) $p\lim \frac{X' X}{T} = Q_{xx}$ finita, simétrica y definida positiva

Entonces se tiene:

$$a) \;\; p\lim \frac{X' u}{T} = 0_k$$

$$b) \;\; \frac{X' u}{T^{1/2}} \xrightarrow{d} N(0_k, \; \sigma^2 Q_{xx})$$

b) Teorema de Slutzky

Si $p\lim \hat{\theta}_t = \theta$ y g ($\hat{\theta}_t$) es una función continua de $\hat{\theta}_t$, entonces:

$$p\lim g(\hat{\theta}_t) = g(p\lim \hat{\theta}_t) = g(\theta)$$

Es decir, el límite en probabilidad de una función continua viene dado por la función evaluada en el límite de probabilidad.

c) Teorema de Cramer

Sea $\mathbf{Y_t} = \mathbf{Z_t} \mathbf{q_t}$, donde $\mathbf{Z_t}$ es una función continua de matrices aleatorias y $\mathbf{q_t}$ es un vector de variables aleatorias. Si se satisface que:

$$I) \;\; p\lim Z_t = Z \qquad \textit{finita}$$

$$II) \;\; p\lim q_t = q \qquad \textit{finito}$$

$$III) \;\; q \overset{a}{\sim} N(\mu, \; M)$$

entonces:

$$I)\quad p\lim\ Z_t\ q_t\ =\ Zq$$

$$II)\quad Z_T\ q_T\ \xrightarrow{d}\ Zq$$

$$III)\quad Zq\ \overset{a}{\sim}\ N\left(Z\mu,\ ZMZ'\right)$$

6.4. Criterios de Estimación

Dada una secuencia de vectores de variables aleatorias

$\mathbf{Z_T = (Y_{t.} X_t);\ t = 1, \ldots, T.}$

Si la función de densidad de estas variables viene dada por

$\mathbf{f(Y_{1,} \ldots, Y_{T,}\ X_{1,} \ldots, X_{T;}\theta)}$

Donde $\boldsymbol{\theta}$ es un vector de parámetros desconocidos que caracteriza esta función de densidad.

Supongáse que el objetivo es realizar inferencias de $\boldsymbol{\theta}$ sobre la base de los valores observados de $\mathbf{(Y_{1,} \ldots, Y_{T,}\ X_{1,} \ldots, X_T)}$.

Ahora bien, en Econometría interesa la distribución condicionada de un conjunto de variables aleatorias dado otro conjunto de variables aleatorias.

Por ejemplo, analizar el consumo de la población para una renta dada. Si la propensión marginal a consumir es el parámetro de interés y se determina independientemente del proceso que ha generado la renta, se puede plantear la distribución del consumo condicionado a la renta.

La función de densidad conjunta de esta muestra se puede escribir como:

$$\mathrm{f}\left(\mathrm{Y}_1, \cdots, \mathrm{Y}_\mathrm{T}, \mathrm{X}_1, \cdots, \mathrm{X}_\mathrm{T}; \theta\right) = \mathrm{f}\left(\mathrm{Y}_1, \cdots, \mathrm{Y}_\mathrm{T} \mid \mathrm{X}_1, \cdots, \mathrm{X}_\mathrm{T}; \beta\right) \mathrm{f}\left(\mathrm{X}_1, \cdots, \mathrm{X}_\mathrm{T}; \gamma\right)$$

Si el vector $\boldsymbol{\beta}$ recoge los parámetros de interés y no hay relación entre $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ se puede ignorar $f\left(X_1, \cdots, X_t; \gamma\right)$ a la hora de hacer inferencia sobre $\boldsymbol{\beta}$.

El vector $Y = \left(Y_1,\ \cdots,\ Y_T\right)'$, será el vector de variables endógenas o dependientes.

El vector $X = \left(X_1,\ \cdots,\ X_T\right)'$, el vector de variables exógenas o independientes.

La función de distribución de interés vendrá dada por $\mathbf{f(Y | X, \beta)}$ y considerando las variables independientes de **X** como fijas (no estocásticas) en muestras repetidas, se reducirá a $\mathbf{f(Y, \beta)}$.

Criterio de máxima verosimilitud

El principio de estimación de máxima verosimilitud se basa en la idea de que es más probable que la muestra observada haya sido generada por un modelo caracterizado por valores determinados de los parámetros y no por otros alternativos.

La estimación por máxima verosimilitud de un vector de parámetros **β** consiste en elegir aquellos valores particulares $\hat{\beta}_{MV}$ que con mayor probabilidad caracterizarían el proceso que ha generado la muestra disponible.

La función de verosimilitud, tiene la misma forma que la función de densidad conjunta de la muestra, pero, a diferencia de ésta que es una función de las variables aleatorias **(Y_T, X_T)**, la función de verosimilitud, es una función de los valores que puedan tomar los parámetros**, β**, dada una muestra concreta.

Para especificar, por lo tanto, la función de verosimilitud es necesario suponer una forma explícita para la función de densidad conjunta de la muestra. Sea

$$f\left(Y_1, \cdots, Y_t \mid \ X_1, \cdots, X_t; \beta\right)$$

la función de densidad que genera la muestra dado un vector **β**. Análogamente, tomando como dada la muestra disponible, la función de verosimilitud, es una función de **β**:

$$f\left(\beta \mid Y_1, \ \cdots, \ Y_t, \qquad X_1, \ \cdots, \ X_T\right) = f\left(\beta \mid Y, X\right) \ = \ L(\beta)$$

En el caso de que $\left(Y_1, \ \cdots, \ Y_T\right)$, sean observaciones independientes, se puede escribir la función de verosimilitud como:

$$L(\beta) = \prod_{T=1}^{T} \ f\left(Y_T \mid X_T; \ \beta\right)$$

El estimador máximo-verosímil $\hat{\beta}_{MV}$, es aquel valor de los parámetros que, dada la muestra, maximiza la función **L(β)**. En la práctica, en lugar de considerar L(β), se considera el logaritmo natural de L(β), In L(β), dado que el valor de β que maximiza L(β), es el mismo que maximiza Ln (β) y es, analíticamente, más sencillo trabajar en logaritmos.

El estimador máximo verosímil es un estimador eficiente. Para demostrarlo se define la *matriz de información* que es la matriz de la esperanza matemática de las segundas derivadas de la función de verosimilitud, con signo negativo:

$$I(\beta) \ = \ -E\left[\frac{\partial^2 \ InL}{\partial\beta \ \partial\beta^{'}}\right]$$

y la inversa de la matriz de información $[I(\beta)]^{-1}$, se denomina cota de Cramer-Rao.

Si β^* es cualquier estimador insesgado de **β**, se puede demostrar que $V(\beta^*) - [I(\beta^*)]^{-1}$, es una matriz semidefinida positiva.

Este resultado implica que, bajo ciertas condiciones de regularidad, si el estimador de interés, $\hat{\beta}$, es un estimador insesgado y su varianza, $V(\hat{\beta})$, alcanza la cota de Cramer-Rao, entonces $\hat{\beta}$, es un estimador eficiente, ya que no es posible que otro estimador insesgado tenga una varianza menor.

En este mismo contexto se define la *eficiencia asintótica*. Si un estimador $\hat{\beta}$ es consistente y la matriz de covarianzas de su distribución asintótica, **IA(β),** es tal que:

$$IA(\beta) = \lim_{T\to\infty} \left[\frac{1}{T}\ I(\beta)\right]^{-1}$$

Entonces $\hat{\beta}$ es un estimador asintóticamente eficiente.

Bajo ciertas condiciones de regularidad, el estimador máximo-verosímil tiene propiedades asintóticas deseables: es asintóticamente insesgado, consistente, asintóticamente eficiente y su distribución asintótica es una distribución normal.

Su único inconveniente es que es preciso suponer una distribución específica

$$f\left(Y_1, \cdots, Y_t \mid X_1, \cdots, X_t; \beta\right) .$$

Una mala especificación de esta función de distribución implica la pérdida de estas buenas propiedades asintóticas.

En general, las propiedades asintóticas del estimador máximo-verosímil, son muy atractivas en casos en los que es imposible encontrar estimadores con buenas propiedades para muestras finitas, situación esta que se produce frecuentemente en la práctica.

El procedimiento para obtener los estimadores maximoverosímiles consiste en:

1°. Definir la función de verosimilitud $L(\theta_1, \theta_2, \ldots, \theta_k)$

2°. Tomar *ln* de *L:* $\ln[L(\theta_1, \theta_2, \ldots, \theta_k)]$

3°. Los estimadores de $\theta_1, \theta_2, \ldots, \theta_k$ serán aquellas expresiones para las cuales la función *L* sea máxima. Para esto es necesario

a) derivar *ln(L)* respecto a cada θ_i
b) igualar a cero

c) reemplazar los θ_i por los estimadores $\hat{\theta}_i$
d) resolver el sistema de ecuaciones

Ejemplo. Dada una población normal de parámetro (μ,σ), la función de verosimilitud para obtener buenos estimadores para la media μ y para la varianza σ^2, será

$$f(\mu/\sigma^2)=\prod_{i=1}^{n} f(x_i)=\prod_{i=1}^{n}\frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

1° Se define la función de verosimilitud

$$L(\mu,\sigma^2)=\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}\prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{\sum(x_i-\mu)^2}{2\sigma^2}}$$

$$=\left(\frac{1}{\sigma^2 2\pi}\right)^{-\left(\frac{n}{2}\right)} e^{-\frac{\sum(x_i-\mu)^2}{2\sigma^2}}$$

2° Se toma *ln*

$$\ln L(\mu,\sigma^2)=-\frac{n}{2}\ln(\sigma^2 2\pi)-\frac{\sum(x_i-\mu)^2}{2\sigma^2}=-\frac{n}{2}\ln(\sigma^2)-\frac{n}{2}\ln(2\pi)-\frac{\sum(x_i-\mu)^2}{2\sigma^2}$$

3° Se deriva el $\ln L(\mu,\sigma^2)$ con respecto a μ y σ^2, y se iguala a cero

$$\begin{cases}\dfrac{\partial\ln(\mathrm{L})}{\partial\mu}=-2\dfrac{\sum(\mathrm{x_i}-\mu)}{2\sigma^2}(-1)=\dfrac{\sum(\mathrm{x_i}-\mu)}{\sigma^2}=0\\[2ex]\dfrac{\partial\ln(\mathrm{L})}{\partial\sigma^2}=-\dfrac{\mathrm{n}}{2}\dfrac{1}{\sigma^2}+\dfrac{\sum(\mathrm{x_i}-\mu)^2}{2(\sigma^2)^2}=0\end{cases}$$

Estas son las derivadas de primer orden, a partir de las que se obtienen los puntos extremos en la optimización. Para esto, cada derivada se iguala a 0 y se reemplaza μ por $\bar{x}$ y σ^2 por $\hat{\sigma}^2$, resultando un sistema de dos ecuaciones con dos incógnitas:

$$\begin{cases}\dfrac{\sum(\mathrm{x_i}-\bar{\mathrm{x}})}{\hat{\sigma}^2}=0\\[2ex]-\dfrac{\mathrm{n}}{2}\dfrac{1}{\hat{\sigma}^2}+\dfrac{\sum(\mathrm{x_i}-\bar{\mathrm{x}})^2}{2(\hat{\sigma}^2)^2}=0\end{cases}$$

En (1) se multiplica ambos miembros por $\hat{\sigma}^2$ y se resuelve el sumatorio, de modo que

$$\sum x_i-n\bar{x}=0\Rightarrow\bar{x}=\frac{\sum x_i}{n}$$

En (2) se despeja $\hat{\sigma}^2$ y se obtiene:

$$\frac{-n\hat{\sigma}^2+\sum(x_i-\bar{x})^2}{2\left(\hat{\sigma}^2\right)^2}=0\Rightarrow\sum(x_i-\bar{x})^2=n\hat{\sigma}^2\therefore\hat{\sigma}^2=\frac{\sum(x_i-\bar{x})^2}{n}$$

La media muestral es un Estimador Máximo Verosímil (EMV) de la media poblacional. El EMV de σ^2 es $\hat{\sigma}^2$ que, por lo demostrado anteriormente, es un estimador sesgado.

Criterio mínimo-cuadrático

Si se especifica una forma funcional determinada para el valor medio de la distribución de interés **f(Y|X, β)**

E(Y|X, β) = h(X; β)

El criterio mínimo-cuadrático para la obtención de un estimador de **β** se basa en elegir aquel valor de **β** que, dada la muestra disponible: $\left(Y_1,\cdots,Y_t \mid X_1,\cdots,X_t;\beta\right)$ minimice la siguiente función de distancia entre **Y** y **h(X; β)**

$$S(\beta) = (Y-h(X;\ \beta))'\ (Y\ -\ h(X;\ \beta))$$

Si **Y** es un vector (**Tx1**) y **X** es una matriz (**T x K**), este criterio se puede escribir como la suma de cuadrados de las desviaciones entre **Y** y su valor medio **h**:

$$S(\beta) = \sum_{t=1}^{T}\ (Y_t-h(X_t;\ \beta))^2$$

Dependiendo de la función de distancia elegida, se determinará un criterio mínimo-cuadrático diferente. Si la función es lineal, **E(Y_T|X_T, β) = X_T' β,** la función objetivo es:

$$S(\beta) = \sum_{t=1}^{T}\ \left(Y_t-\ X_T'\ \beta\right)^2$$

y el criterio mínimo-cuadrático general recibe el nombre específico de *mínimos cuadrados ordinarios (MCO)*.

Si por el contrario, se desea dar más peso a las observaciones Y_T menos alejadas del valor medio **E(Y_T|X_T, β),** se puede especificar una matriz de pesos **W** (**T x T**), que recoja distintas ponderaciones para distintas desviaciones, siendo por lo tanto la función criterio de la forma:

$$S(\beta) = (Y-h(X;\ \beta))'\ \ W\ (Y\ -\ h(X;\ \beta))$$

El método de estimación que se basa en elegir los valores de **β** que minimizan la función criterio anterior, se denomina *mínimos cuadrados generalizados (MCG).*

6.5 Estimación por Intervalo

En la estimación por intervalo, generalmente, se construyen dos funciones basadas en los mismos datos muestrales, dentro de los cuales, en un sentido probabilístico, el valor desconocido del parámetro ha de encontrarse. Esto consiste en obtener un cierto intervalo aleatorio que contiene el parámetro a estimar. Este intervalo aleatorio se denomina intervalo de confianza y se establece su magnitud 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 $(1-\alpha)100$ 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, suponga que el estimador $\hat{\theta}$ ~ *N*, o asintóticamente normal, y que la amplitud del intervalo sea $\pm 2\sigma_{\hat{\theta}}$. Suponga, 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 contenidos a partir de $\hat{\theta}_2$ y $\hat{\theta}_3$ contiene 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.

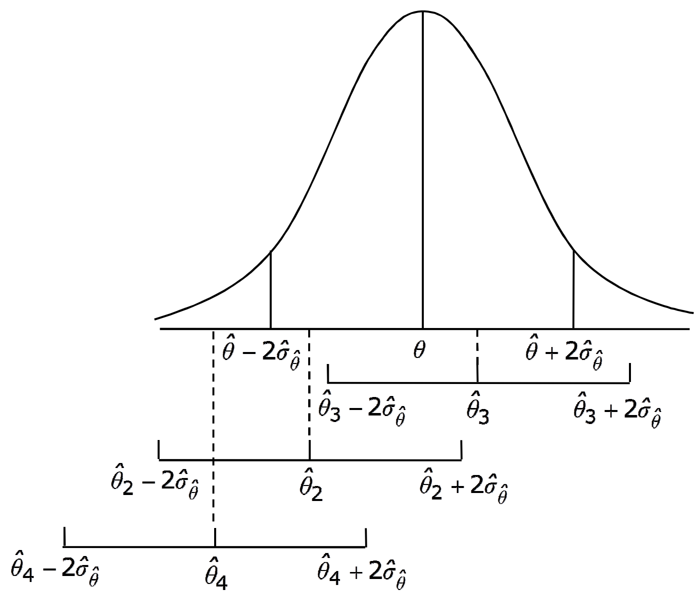


Figura 6.9

Elementos de la estimación por intervalos

1) θ; parámetro a estimar
2) $x_1, x_2, \ldots, x_n$; n observaciones muestrales de una población cuya función de densidad es $g(x/\theta)$.
3) $\hat{\theta}$; un estimador de θ y como tal variable aleatoria, $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$.
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

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

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

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 $(-\sigma_{\hat{\theta}})$:

$$\mathrm{P}\left(-\mathrm{K}_1 \sigma_{\hat{\theta}} \geq \theta - \hat{\theta} \geq -\mathrm{K}_2 \sigma_{\hat{\theta}}\right) = 1 - \alpha$$

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

$$\mathrm{P}\left(\hat{\theta} - \mathrm{K}_2 \sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} - \mathrm{K}_1 \sigma_{\hat{\theta}}\right) = 1 - \alpha$$

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 $z_{\alpha/2} = -z_{1-\alpha/2}$

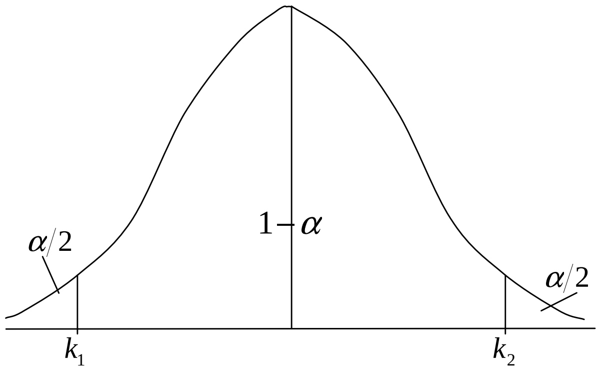


Figura 6.10

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 para la media de 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), $\overline{\mathrm{X}}$ se distribuye normal con parámetro $(\mu, \sigma/\sqrt{\mathrm{n}})$; 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

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

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.

si σ es conocida,

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

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

$$L_i = \overline{X} - z_{1-\alpha/2}\,\sigma/\sqrt{n}\,;$$
$$L_S = \overline{X} + z_{1-\alpha/2}\,\sigma/\sqrt{n}$$

Si σ es desconocida y la muestra es grande

El estadístico se distribuye como una normal

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

el desvío desconocido debe estimarse por

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

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

$$L_i = \overline{X} - z_{1-\alpha/2}\,s/\sqrt{n}\,;$$
$$L_S = \overline{X} + z_{1-\alpha/2}\,s/\sqrt{n}$$

Si σ es desconocida y la muestra es pequeña

El estadístico K se distribuye como una t

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

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

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

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

Multiplicando y dividiendo por *n-1*:

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

En consecuencia, dada una muestra de tamaño *n* y siendo

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

se obtiene

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

Esto significa que

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

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

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

$$L_i = \overline{X} - t_{n-1;1-\alpha/2}\, s/\sqrt{n}$$
$$L_s = \overline{X} - t_{n-1;\alpha/2}\, s/\sqrt{n}$$

En resumen, la media de la muestra ($\overline{X}$) es usada para estimar la media desconocida de la población (μ). Debido a que $\overline{X}$ 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 $\overline{x}$.

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:

$$\overline{x} \pm \text{error de la muestra (e)} = \overline{x} \pm z\frac{\sigma}{\sqrt{n}} = \text{estimación de intervalo de } \mu$$

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

$\overline{X} = 793.00$

$\sigma = 531.20$

El intervalo es

$$\overline{x} \pm 1.96\frac{\sigma}{\sqrt{n}} = 793.00 \pm 1.96\frac{531.20}{\sqrt{10}} = 793.00 \pm 329.24$$

e incluye la media poblacional verdadera (μ = 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 dependerá 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, sería:

$$\overline{x} \pm t\frac{s}{\sqrt{n}}$$

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

$$\overline{X} = 793.00$$

$$s = 453.56$$

$$n = 10$$

Entonces el intervalo es:

$$\overline{x} \pm t\frac{s}{\sqrt{n}} = 793.00 \pm 2.26\frac{453.56}{\sqrt{10}} = 793.00 \pm 324.15$$

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

$$S^2 = \frac{\sum_{i=1}^{n}(X_1 - \overline{X})^2}{n-1}$$

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

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

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

$$K_1 = \chi^2_{n-1;\alpha/2}$$

$$K_2 = \chi^2_{n-1;1-\alpha/2}$$

Los límites del intervalo se definen

$$L_i = \frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}}$$

$$L_s = \frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}$$

Ejemplo. 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.

$\Sigma X_i = 185.2 dc^3 \qquad \Sigma X_i^2 = 3545.8 dc^3$

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

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

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

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^2en el intervalo de confianza

$$P\left(\frac{18*96.70}{31.53} \le \sigma^2 \le \frac{18*96.70}{8.23}\right) = 0.95$$

$$P\left(55.26 \le \sigma^2 \le 211.49\right) = 0.95$$

El resultado indica que la variabilidad en el llenado de las botellas de una bebida gaseosa se encuentra entre $55.26 dc^3$ y $211.49 dc^3$

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, $\overline{X}_1$ y $\overline{X}_2$.

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

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

2) $V(\overline{X}_1 - \overline{X}_2) = V(\overline{X}_1) + V(\overline{X}_2) = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$

3) Si $\overline{X}_1 \sim N\left(\mu_1, \sqrt{\dfrac{\sigma_1^2}{n_1}}\right)$ y $\overline{X}_2 \sim N\left(\mu_2, \sqrt{\dfrac{\sigma_2^2}{n_2}}\right)$, entonces

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

Al igual que en el caso de intervalos para la media, μ, se deben considerar intervalos alternativos 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 $Z \sim N(0,1)$

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sigma_{\overline{X}_1 - \overline{X}_2}} = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

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

$$P(Z_1 \le Z \le Z_2) = 1 - \alpha$$

Remplazando Z por su igual

$$P\left(Z_1 \le \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \le Z_2 \right) = 1 - \alpha$$

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

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

De modo que los límites serán:

$$L_i = \overline{X}_1 - \overline{X}_2 - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$L_s = \overline{X}_1 - \overline{X}_2 + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

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:

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

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

$$P\left((\overline{X}_1 - \overline{X}_2) - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \le \mu_1 - \mu_2 \le (\overline{X}_1 - \overline{X}_2) + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right) = 1 - \alpha$$

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

$$t_{n_1+n_2-2} = \frac{\dfrac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}}{\sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}}$$

Luego, multiplicando extremos por medios

$$t_{n_1+n_2-2} = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{n_1+n_2}{n_1 n_2}}\sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}}$$

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

$$P\left(\left(\overline{X}_1 - \overline{X}_2\right) - t_{n_1+n_2-2;1-\frac{\alpha}{2}} \sqrt{\frac{n_1+n_2}{n_1 n_2}\,\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \le \mu_1 - \mu_2 \le \right.$$

$$\left. \left(\overline{X}_1 - \overline{X}_2\right) + t_{n_1+n_2-2;1-\frac{\alpha}{2}} \sqrt{\frac{n_1+n_2}{n_1 n_2}\,\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \right) = 1 - \alpha$$

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

$$k=\frac{\overline{X}_1-\overline{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

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. Una de ellas se debe a Hsu y consiste en aproximar la distribución por una t con grados de libertad fijadas en el mínimo $(n_1,n_2)-1$ obteniéndose el intervalo

$$P\left(\overline{X}_1-\overline{X}_2-k_{2,(n_1,n_2)-1}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\le\mu_1-\mu_2\le\overline{X}_1-\overline{X}_2+k_{2,(n_1,n_2)-1}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right)=1-\alpha$$

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. 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 tienda 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	$\overline{X}$	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:

$$P\left(\left(\overline{X}_1-\overline{X}_2\right)-t_{n_1+n_2-2;1-\frac{\alpha}{2}}\sqrt{\frac{n_1+n_2}{n_1n_2}\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\le\mu_1-\mu_2\le\right.$$
$$\left.\left(\overline{X}_1-\overline{X}_2\right)+t_{n_1+n_2-2;1-\frac{\alpha}{2}}\sqrt{\frac{n_1+n_2}{n_1n_2}\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\right)=1-\alpha$$

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

$$P\big((342500-325000)-16672.38 \leq \mu_1-\mu_2 \leq (342500-325000)+16672.38\big)=1-\alpha$$
$$P(17500-16672.38 \leq \mu_1-\mu_2 \leq 17500+16672.38)=0.95$$
$$P(827.63 \leq \mu_1-\mu_2 \leq 34172.38)=0.95$$

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 $\left(\sigma_2^2/\sigma_1^2\right)$ a partir del cálculo de las varianzas de cada muestra $\left(s_1^2 \text{ y } s_2^2\right)$.

En el intervalo de confianza de la varianza el estadístico a utilizar se distribuye χ^2 con n-1 grados de libertad

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

En el cociente de varianzas, se tiene un cociente de χ^2 que da por resultado el estadístico F

$$F_{n_1-1;n_2-1} = \frac{\dfrac{(n_1-1)\dfrac{S_1^2}{\sigma_1^2}}{(n_1-1)}}{\dfrac{(n_2-1)\dfrac{S_2^2}{\sigma_2^2}}{(n_2-1)}} = \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2}$$

El intervalo de confianza entre dos valores críticos que garantizan la probabilidad $1-\alpha$ se define como

$$P\left(K_1 \leq \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} \leq K_2 \right) = 1-\alpha$$

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. Se utilizarán los datos del ejemplo de intervalo de confianza para las diferencias de medias.

		Producto 1	Producto 2
Ventas promedio	$\overline{X}$	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

$$\mathrm{P}\left(\frac{1}{3.91}\frac{17500^2}{20000^2} \leq \frac{\sigma_2^2}{\sigma_1^2} \leq 3.59\frac{17500^2}{20000^2}\right) = 0.95$$

$$P\left(0.1958 \leq \frac{\sigma_2^2}{\sigma_1^2} \leq 2.7486\right) = 0.95$$

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 dell 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 $(n>30)$ $k \sim N(0,1)$

Si n es pequeña $(n \leq 30)$ $k \sim t_{n-1}$

Intervalo de confianza para la igualdad de proporciones poblacionales

El parámetro $\theta = \pi_1 = \pi_2 \rightarrow \pi_1 - \pi_2 = 0$

Estimador $\hat{\theta} = p_1 = p_2$

Para cada una de las muestras tendremos

$E(p_1) = \pi_1$ $\qquad V(p_1) = \frac{p_1(1-p_1)}{n_1}$

$$E(p_2) = \pi_2 \qquad V(p_2) = \frac{p_2(1-p_2)}{n_2}$$

Esto significa que

$$E(p_1 - p_2) = E(p_1) - E(p_2) = \pi_1 - \pi_2$$

$$V(p_1 - p_2) = V(p_1) + V(p_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

Si n_1 y n_1 son mayores que 30 $(p_1 - p_2) \sim N(\pi_1 - \pi_2; \sigma_{\Pi 1-\Pi 2})$

La variable estandarizada será

$$k = \frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

$$P\left((p_1 - p_2) - k_2\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \le \pi_1 - \pi_2 \le (p_1 - p_2) - k_1\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right) = 1 - \alpha$$

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; define 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

$$P\left(\bar{p} \pm Z\sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right) = 1 - \alpha$$

6.6 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.

$H_0 : \theta = \theta_0$
$H_1 : \theta \neq \theta_0$

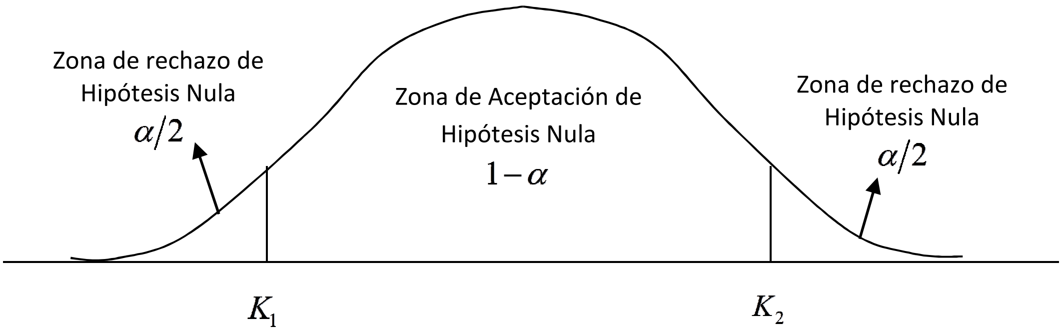


Figura 6.11

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 palntean 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:

$H_0 : \theta \geq \theta_0$
$H_1 : \theta < \theta_0$

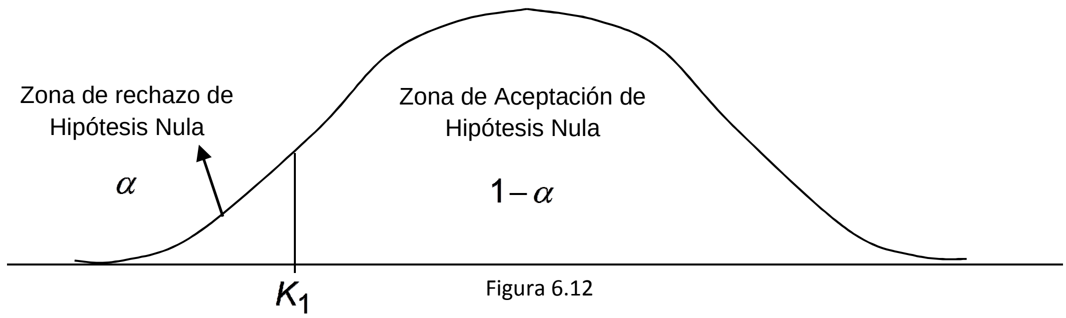


Figura 6.12

Cuando es lateral derecha:

$$H_0 : \theta \leq \theta_0$$
$$H_1 : \theta > \theta_0$$

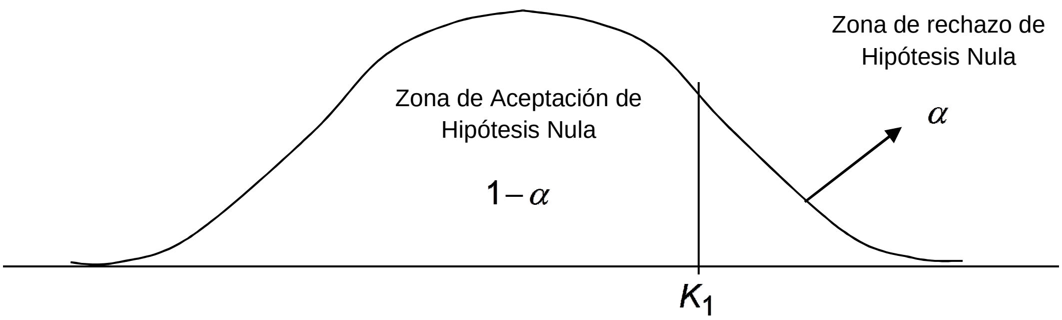


Figura 6.13

En la Figura 6.14 se reúnen las diferentes situaciones de pruebas de hipótesis, para variables cuantitativas, con el estadístico a utilizar.

Ejemplo. 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. Suponga, la **Hipótesis Nula** de 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.

$$H_0 : \mu = 25$$

Dócima	Estadístico
Media con σ^2 poblacional conocida	$Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$
Media con σ^2 desconocida, n grande	$Z = \dfrac{\overline{X} - \mu}{s/\sqrt{n}} \sim N(0,1)$
Media con σ^2 desconocida, n pequeña, probabilidad Normal	$t = \dfrac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$
Proporción	$Z = \dfrac{p - \pi}{\sqrt{\dfrac{\pi(1-\pi)}{n}}}$
Diferencias de medias con σ^2 conocida y n de cualquier tamaño	$Z = \dfrac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$
Diferencia de medias con σ^2 desconocida muestras grandes	$Z = \dfrac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \sim N(0,1)$
Diferencia de medias, σ^2 desconocida pero iguales y n pequeño	$t = \dfrac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{n_1 + n_2}{n_1 n_2} \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}} \sim t_{n_1 + n_2 - 2}$
Diferencia de medias, σ^2 desconocida y distintas, n pequeño	$t = \dfrac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \sim t_{\text{mínimo}(n_1, n_2) - 1}$
Igualdad de Proporciones	$k = \dfrac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} \sim N(0,1)$
σ^2 de población normal	$W = (n-1)\dfrac{s^2}{\sigma^2} \sim \chi^2_{n-1}$
Igualdad de varianza	$\dfrac{s_1^2}{s_2^2}\dfrac{\sigma_2^2}{\sigma_1^2} \sim F_{n_1 - 1; n_2 - 1}$

Figura 6.14. Estadísticos para la dócima

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.

$$H_1 : \mu \neq 25$$

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 mas 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:

$$t = \frac{\overline{x} - \mu}{s_{\overline{x}}} = \frac{22 - 25}{\sqrt{45} / \sqrt{20}} = -2.00$$

donde, t es un estadístico con distribución t de Student y refleja, en unidades tipificadas, en cuánto difiere la media de la muestra $\overline{x}$ 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 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.

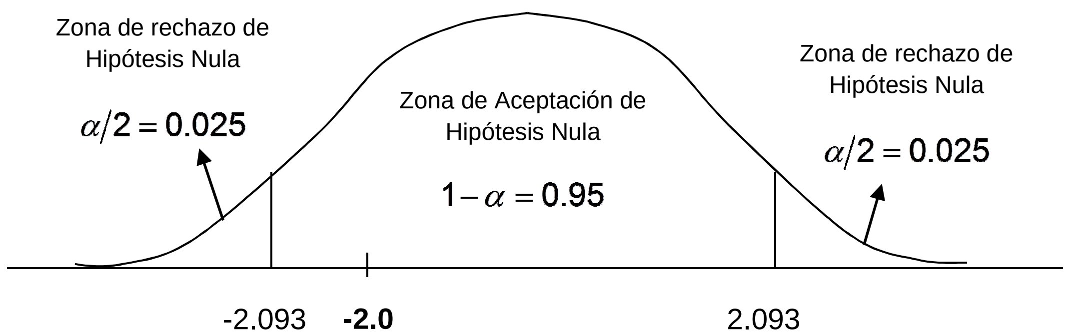


Figura 6.15

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. 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.

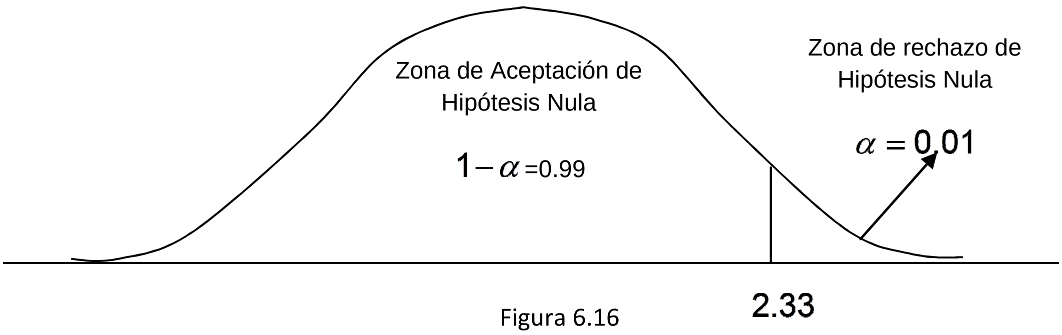


Figura 6.16

Para realizar el contraste se debe calcular el valor empírico del estadístico Z.

$$Z^* = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0{,}80-0{,}90}{\sqrt{\frac{0{,}90*0{,}1}{200}}} = \frac{-0{,}1}{0{,}02121} = -4.7148$$

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.

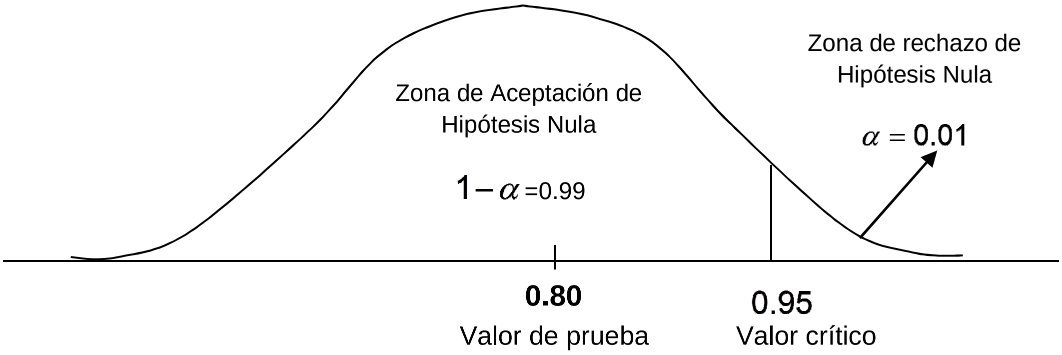


Figura 6.17

El valor empírico, o valor de prueba:

$$p = \frac{160}{200} = 0{,}80$$

cae en la zona de aceptación de la hipótesis nula.

Ejemplo. 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. $H_0 : \mu = 6500$

$H_1 : \mu \neq 6500$

4. Se conoce el desvío de la variable en la población, aun cuando la muestra es pequeña, el estadístico se distribuye normal

$$K(\overline{X}, \mu) = Z = \frac{\overline{X} - \mu}{\sigma_{\bar{x}}}$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{920}{\sqrt{25}} = 184$$

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

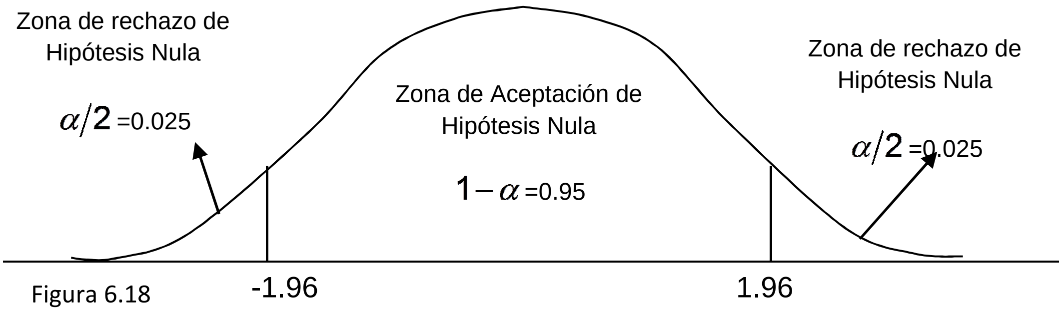


Figura 6.18

6. El valor empírico del estadístico Z es

$$Z^* = \frac{\bar{X}^* - 6500}{184} = \frac{7145 - 6500}{184} = 3.51$$

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

$$\bar{X}^* = \pm Z\frac{\sigma}{\sqrt{n}} + 6500 = \pm 1.96 * 184 + 6500$$

$$\bar{X}_1^* = 6139.36, \qquad \bar{X}_2^* = 6860.64$$

El valor muestral $\bar{X} = 7145$ cae en la zona de rechazo de la hipótesis nula por lo que la afirmación de las autoridades locales no es cierta.

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 temepramento 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.

CASOS DE ESTUDIO, PREGUNTAS Y PROBLEMAS

Caso 6.1: 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? ¿Porqué? 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 6.2: 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 6.3: 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 realices 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.

Preguntas

¿Porqué la proporción de 0.5 en una variable dicotómica se corresponde con una varianza máxima? Demuestre.

Problemas

6.1. 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%.

Observa cómo cambia el tamaño del intervalo.

6.2. El administrador de una 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) Establezca 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?.

6.3. 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. Establece una estimación por intervalo con 95% de confianza de la proporción real de usuarios para esta nueva ruta de autobuses.

6.4. 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 un tipo convencional.

Para los pacientes del primer grupo el número medio de horas de dormir fue de 8.32 con una desviación típica estimada de 0.24 horas

Para los pacientes del grupo B el número medio de horas de dormir fue de 6.75 con una desviación típica estimada de 0.30 horas

Encuentra 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.

6.5. El saldo que arroja la cuenta deudores varios en el balance general de una empresa es de $120000. 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 $11900. Es confiable a un nivel del 1% la cifra que reflejan los registros contables si suponemos población aproximadamente normal y desvío poblacional de 100?

6.6. 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 encontró 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. Docime a un nivel de significación del 5% si tal hecho ocurre. Suponga que la población es normal.

6.7. 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. ¿Aceptaría Ud. esta aseveración a la luz de los resultados muestrales a un nivel de significación del 1%?

6.8. 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. Prueba la hipótesis nula para el nivel de significación de 0.05.

6.9. 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 tomaría?

6.10. De 100 graduados en contaduría, una muestra aleatoria de 12 estudiantes tiene un promedio de calificación de 2.70 (el máximo es 4) con una desviación estándar muestral de 0.40. Para 50 egresados de sistemas de información una muestra aleatoria de 10 estudiantes tiene un promedio de calificación de 2.90 con una desviación estándar muestral de 0.30. Se supone que las calificaciones tienen distribución normal. Pruebe la hipótesis de que la calificación promedio para las dos categorías de estudiantes es distinta utilizando el nivel de significación del 5%.

6.11. Se plantea la hipótesis de que la desviación estándar del salario por hora de los trabajadores a destajo en una determinada industria es de, cuanto menos, $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%?.

6.12. Una fábrica de galletitas incorporó un nuevo proceso de elaboración de bizcochos dulces que les 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

$$n = 16 \qquad\qquad n = 13$$

$$\sum_{i=1}^{16} X_i = 319.20 \qquad\qquad \sum_{i=1}^{13} Y_i = 261.30$$

$$\sum_{i=1}^{16} X_i^2 = 6373.11 \qquad\qquad \sum_{i=1}^{13} Y_i^2 = 5255.54$$

Sabiendo que ambas poblaciones son aproximadamente normales. ¿Hay diferencias significativas en cuanto a las varianzas de los pesos de los bizcochos entre ambos procesos a un nivel de confianza del 99%?

6.13. El gerente de marketing de una cadena de tiendas de flores y plantas cree que habrá una gran demanda de bulbos de gladiolos en las tiendas de la periferia de la ciudad durante el mes de mayo.

Si las ventas de bulbos exceden de $100 por semana se venderán 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 deberían vender los bulbos de gladiolos en todas las tiendas suburbanas de la cadena?

Bibliografía

- **Berenson, Mark y Levine, Daniel.** *Estadística Básica En Administración*. México: Prentice Hall, 1996.
- **____.** *Estadística Para Administración Y Economía. Conceptos Y Aplicaciones.* México: Mc. Graw Hill, 1993.
- **Carrizo, José Fernando y Carrizo, José Fernando.** "Nociones De Inferencia Estadística," Córdoba: Instituto de Econometría y Estadística. Facultad de Ciencias Económicas. Universidad Nacional de Córdoba, 1977.
- **Cochran, W. G. .** *Técnicas De Muestreo*. Editorial CECSA, 1987.
- **Chou, Ya-Lun.** *Análisis Estadístico*. México: Nueva Editorial Interamericana, 1977.
- **Dagum, C y Bee de Dagum E.** *Introducción a La Econometría*. México: Editorial Siglo XXI, 1971.
- **Daniel, W.** *Bioestadística, Base Para El Análisis De Las Ciencias De La Salud.* México: Editorial Limussa, 1999.
- **Dixon, W.J. y Massey, F.J.** *Introduction to Statistical Analysis*. Nueva York: Mc Graw Hill, 1957.
- **Gomez Villegas, Miguel Angel.** *Inferencia Estadística*. España: Ediciones Díaz de Santo, 2005.
- **Gujarati, Damodar.** *Econometría*. México: Mc.Graw Hill, 2004.
- **Hildebrand, D. y Lyman Ott R.** *Estadística Aplicada a La Administración Y a La Economía*. Wilmington, USA: Addison Wesley Iberoamericana, 1997.
- **Johnston, J. y Dinardo, J. .** *Métodos De Econometría.* Barcelona: Editorial Vicens Vives, 2001.
- **Kazmier, L y Diaz Mata, A.** *Estaística Aplicada a La Administración Y a La Economía*. México: McGraw Hill, 1993.
- **Meyer, P. L.** *Probabilidad Y Aplicaciones Estadísticas*. México: Fondo Educativo Interamericano SA, 1973.
- **Mood, A. y Graybill, F.** *Introducción a La Estadística*. Editorial Aguilar, 1972.
- **Padua, J.** *Técnicas De Investigación Aplicadas a Las Ciencias Sociales*. México: Fondo de Cultura Económica, 1996.
- **Ross, Sheldon M.** *Introducción a La Estadística*. Barcelona: Editorial Reverté, 2007.
- **Tramutola, C.D.** *Modelos Probabilísticos Y Decisiones Financieras*. Capital Federal: E.C.Moderan, 1971.