DIRECTORIO DEL CURSO FUNDAMENTOS DE LAS TECNICAS DE MUESTRO ESTADISTICO 1980

1. M. EN C. ADELA ABAD DE SERVIN
   Profesora e Investigadora
   Escuela Nacional de Estudios Profesionales
   Acatlán Asesor de la Jefatura de Planeación
   UNAM
   Tel. 373.23.18

2. M. EN C. EDMUNDO BERUMEN TORRES
   Director General de Bioestadística
   Secretaría de Salubridad y Asistencia
   Reforma No. 503
   México, D.F.
   Tel. 286.33.96

3. DR. OCTAVIO A. RASCON CHAVEZ
   Profesor
   División de Estudios de Posgrado
   Facultad de Ingeniería
   UNAM
   México 20, D.F.
   Tel. 548.09.50

4. M. EN C. LUIS ALEJANDRO SERVIN ANDRADE
   Asesor de la Subjefatura de Evaluación
   Departamento de Matemáticas Aplicadas
   IMSS
   Toledo No. 21-7° Piso
   México, D.F.
   Tel. 511.30.27

5. M. EN I. AUGUSTO VILLARREAL ARANDA (COORDINADOR)
   Gerente de Producción
   EIVY'S PERFUMERIA DE MEXICO, S.A.
   Asia No. 31
   México 21, D.F.
   Tel. 554.45.31

'edcs.

FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO 1980

Fecha	Tema	Hora	Profesor
Junio 9	INTRODUCCION	18 a 21 h	M.en C. Edmundo Berumen Torres.
Junio 11 " 13 y " 16	MUESTREO ALEATORIO SIMPLE	18 a 21 h c/día.	Dr. Octavio A. Rascón Chávez
Junio 18	MUESTREO ALEATORIO SIMPLE	18 a 19:30 h	" " "
	MUESTREO ALEATORIO SIMPLE	19:30 a 21 h	M. en I. Augusto Villarreal Aranda.
Junio 20 y " 23	TAMAÑO DE LA MUESTRA	18 a 21 h c/día.	" " "
Junio 25 y " 27	MUESTREO ALEATORIO SIMPLE PARA RAZONES	18 a 21 h c/día.	M. en C. Adela Abad de Servín.
Junio 30 Julio 2 y Julio 4	MUESTREO ESTRATIFICADO	18 a 21 h c/día.	" " " " "
Julio 7 y " 9	MUESTREO POR CONGLOMERADOS	18 a 21 h c/día.	M. en C. Luis Alejandro Servín A.
Julio 11 y " 14	MUESTREO SISTEMATICO	18 a 21 h c/día.	M. en C. Edmundo Berumen Torres.

'edcs.

centro de educación continua
división de estudios de posgrado
facultad de ingeniería unam

FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO

E J E M P L O S

DR. OCTAVIO A. RASCON CHAVEZ

JUNIO, 1980

EJEMPLO

En un estudio sobre los nombres de las personas, se tomó una muestra aleatoria de 10 nombres de una colección de 100. Además del sexo, se consideraron las variables aleatorias Número de letras = Y, Número de vocales = Z, Número de consonantes = W. Los datos obtenidos son los siguientes:

Persona	Nombre	Nº de letras, Y	Nº de vocales, Z	Nº de consonantes, W	Sexo
1	Mario	5	3	2	M=0
2	Juan	4	2	2	M=0
3	Raúl	4	2	2	M=0
4	Gloria	6	3	3	F=1
5	Cosme	5	2	3	M=0
6	Carlos	6	2	4	M=0
7	Luz	3	1	2	F=1
8	Ernesto	7	3	4	M=0
9	Rosa	4	2	2	F=1
10	Norma	5	2	3	F=1

2. Estimar el número medio de letras por nombre, calcular el error estándar y el intervalo de confianza del 95%.

$$\hat{Y} = \bar{y} = \sum_{i=1}^{10} y_i / n = (5+4+4+6+5+6+3+7+4+5)/10 = 4.9$$

$$\hat{V}(\bar{y}) = (1-f)\hat{S}^2/n = \left(1-\frac{10}{100}\right)\frac{\sum_{i=1}^{10}(y_i-\bar{y})^2}{10-1}/10 = 0.129 \left(\frac{\text{letras}}{\text{nombre}}\right)^2$$

$$\text{Error estándar} = \sqrt{\hat{V}(\bar{y})} = \sqrt{0.129} = 0.359 \ \frac{\text{letras}}{\text{nombre}}$$

$$\text{Intervalo de confianza} = \bar{y} \mp t_c \sqrt{\hat{V}(\bar{y})} = 4.9 \mp 2.26 \times 0.359$$

(t_c se obtiene para $10-1=9$ grados de libertad)

$$\text{I. de C.} = \{4.09, 5.71\}$$

6. Estimar el número total de letras en los 100 nombres, U, tanto en forma puntual como con un intervalo de 90%.

$$\hat{U} = N\bar{y} = 100 \times 4.9 = 490 \text{ letras}$$

$$\text{I.C.} = 490 \mp t_c \underbrace{\sqrt{\hat{V}(N\bar{y})}}_{\hat{\sigma}(N\bar{y})} = 490 \mp 1.83 \times \underbrace{\frac{100 \times 0.359}{t_c}}_{\hat{\sigma}(N\bar{y})} = \{424.3, 555.7\}$$

c. Estimar el porcentaje de personas con sexo femenino y el error estándar de la estimación.

$$\hat{p} = p = 4/10 = 0.4 = 40\%$$

$$\hat{\sigma}(\hat{p}) = \sqrt{\frac{N-n}{(n-1)N}\hat{p}\hat{q}} = \sqrt{\frac{100-10}{(10-1)100}0.4\times0.6} = 0.155$$

El intervalo de confianza del 95% es

$$I.C. = 0.155 \pm 2.26\times0.155 = \{1.91, 2.61\}$$

d. Estimar el número de vocales por consonante.

$$\hat{R} = \frac{\bar{z}}{\bar{w}} = \frac{22/10}{27/10} = 0.81 \text{ vocales/consonante}$$

$$\hat{\sigma}^2(\hat{R}) = \frac{1-f}{n\bar{w}^2}\cdot\frac{\sum_{i=1}^{10}(z_i - \hat{R}w_i)^2}{n-1} = \frac{1-\frac{10}{100}}{10(2.7)^2}\cdot\frac{\sum_{i=1}^{10}(z_i - 0.81w_i)^2}{10-1}$$

$$= \frac{0.09}{(2.7)^2}\cdot\frac{\sum_{i=1}^{10}z_i^2 - 2\times0.81\sum z_i w_i + (0.81)^2\sum_{i=1}^{10}w_i^2}{9}$$

$$= \frac{0.09}{(2.7)^2}\cdot\frac{52 - 1.62(61) + (0.81)^2(79)}{9} = 0.006875$$

$$\hat{\sigma}(\hat{R}) = \sqrt{0.006875} = 0.0829$$

El intervalo de confianza del 95% es

$$I.C. = 0.81 \pm 1.83\times0.0829 = \{0.66, 0.96\}$$

EJEMPLO

En un estudio de mercadotecnia se tomó una muestra de 20 precios de carne en 20 tiendas distintas para estimar su variabilidad. Los datos arrojaron un promedio $\bar{x} = \$92.00$ y una desviación estándar $s = \$8.00$. Encontrar el 95% de nivel de confianza del 95% de la varianza.

$$I.C. = \left\{ \frac{20(8)^2}{32.9}, \frac{20(8)^2}{8.91} \right\} = \{38.91, 143.66\}\ \$^2$$

El intervalo de confianza para la desviación estandar es

$$\{\sqrt{38.91}, \sqrt{143.66}\} = \{6.24, 11.99\}\ \$$

centro de educación continua
división de estudios de posgrado
facultad de ingeniería unam.

FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO

ESTIMACION DE TOTALES, PROPORCIONES Y RAZONES

DR. OCTAVIO A. RASCON CHAVEZ

JUNIO, 1980

ESTIMACION DE TOTALES

Puesto que $\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{y}{n} = \hat{\mu}$

es un estimador de la media, μ, de una población, entonces el total de las observaciones en la muestra es

$$y = x_1 + x_2 + \cdots + x_n = n\bar{x}$$

Por lo tanto, un estimador del total, Y, de la población

$$Y = X_1 + X_2 + \cdots + X_N$$

es $$\hat{Y} = N\bar{x}$$

en donde N = tamaño de la población.

Este es un <u>estimador insesgado</u> ya que

$$E[\hat{Y}] = E[N\bar{x}] = N\,E[\bar{x}] = N\mu = Y$$

La <u>variancia</u> de $\hat{Y}$ es

$$\sigma^2(\hat{Y}) = \sigma^2(N\bar{x}) = N^2\sigma^2(\bar{x}) = N^2\frac{\sigma^2(x)}{n}(1-f)$$

en donde $f = n/N$ = proporción de muestreo.

El <u>intervalo de confianza</u> es

$$\left\{N\bar{x} - z_c\sigma(\hat{Y}),\ N\bar{x} + z_c\sigma(\hat{Y})\right\} \quad ---(1)$$

en donde z_c es el valor crítico correspondiente al nivel de significancia α, el cual se obtiene de la <u>distribución normal</u>.

Si la muestra es pequeña, en vez de usar la distribución normal se usa la t de Student y, en tal caso, el intervalo de confianza es

$$\left\{ N\bar{x} - t_c \sqrt{N^2(1-f)} \frac{S(x)}{\sqrt{n-1}},\ N\bar{x} + t_c N\sqrt{1-f}\,\frac{S(x)}{\sqrt{n-1}} \right\}$$

en donde $S(x)$ es la desviación estándar de la muestra

$$S(x) = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

y t_c es el valor crítico correspondiente a un nivel de significancia α, que se obtiene de la distribución t de Student.

En la ecuación (1), $\sigma(\hat{Y})$ se estima usando $\sigma(\hat{Y}) = S(\hat{Y}) = NS(x)\sqrt{\frac{1-f}{n}}$.

O sea, usando a $S(x)$ como estimador de $\sigma(x)$.

ESTIMACION DE PROPORCIONES

Las proporciones se estiman mediante las frecuencias relativas correspondientes con que aparece la observación de interés en la muestra. Así, si p es la proporción que es preciso estimar, su estimador, $\hat{p}$, será

$$\hat{p} = p = \frac{n(a)}{n} \times 100 \text{ , en porcentaje}$$

La varianza de p es

$$\sigma^2(p) = \frac{NPQ}{N-1}\frac{1-f}{N} \text{ , donde } Q = 1-P$$

Se puede demostrar que el estimador de $\sigma^2(p)$ es

$$\sigma^2(p) = \hat{V}(p) = \frac{N-n}{(n-1)N} pq \text{ , donde } q = 1-p$$

Asimismo, el estimador del <u>total</u> de elementos que tiene la población <u>con la característica</u> en cuestión es

$$\hat{A} = N\hat{p}$$

La varianza de este estimador es

$$\sigma^2(\hat{A}) = V(\hat{A}) = \frac{N^2PQ}{N-1}\frac{1-f}{n}$$

la cual se estima con $\hat{V}(\hat{A}) = \frac{N(N-n)}{n-1} pq$.

Para muestras grandes el intervalo de confianza es:

$$\left\{\hat{A} - z_c\sqrt{\hat{V}(\hat{A})}, \hat{A} + z_c\sqrt{\hat{V}(\hat{A})}\right\}$$

ESTIMACION DE RAZONES O COCIENTES

Sean W y X dos variables aleatorias. La razón de la primera respecto a la segunda se estima como:

$$\hat{R} = \frac{\bar{w}}{\bar{x}}$$

Se puede demostrar que la varianza de este estimador es

$$V(\hat{R}) = \sigma^2(\hat{R}) \doteq \frac{1-f}{n\bar{X}^2} \frac{\sum_{i=1}^{n} (w_i - R y_i)^2}{N-1}$$

el cual se estima con

$$\hat{V}(\hat{R}) = \frac{1-f}{n\bar{x}^2} \frac{\sum_{i=1}^{n} (w_i - \hat{R} y_i)^2}{n-1}$$

Para muestras grandes, el intervalo de confianza es

$$\left\{ \hat{R} - z_c \hat{V}(\hat{R}),\ \hat{R} + z_c \hat{V}(\hat{R}) \right\}$$

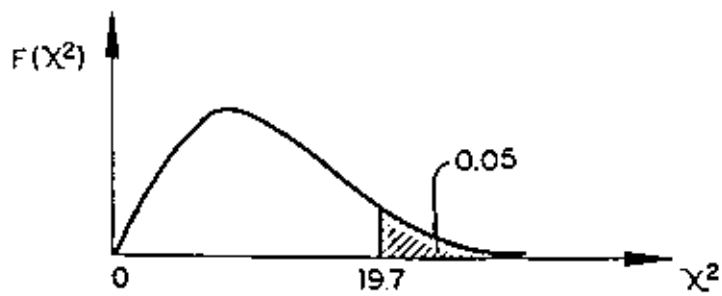
PROBLEMA 17

EN UN ESTUDIO CON FINES ANTROPOLOGICOS SE OBTUVO UNA MUESTRA ALEATORIA DEL TAMAÑO DE LA CABEZA DE LOS INDIGENAS ORIGINARIOS DE CIERTA REGION TROPICAL. LOS DATOS AGRUPADOS SE PRESENTAN EN LA SIGUIENTE TABLA. PROBAR LA HIPOTESIS DE QUE ESTOS DATOS CORRESPONDEN A UNA VARIABLE CON DISTRIBUCION NORMAL.

INTERVALO DE VALORES, mm	FRECUENCIA OBSERVADA, f_i	FRECUENCIA ESPERADA, e_i	$f_i - e_i$	$(f_i - e_i)^2$	$\frac{(f_i - e_i)^2}{e_i}$
<171.5	0	0.4	0.4	0.16	0.40
171.5-175.5	3	2.4	0.6	0.36	0.15
175.5-179.5	9	10.5	-1.5	1.25	0.12
179.5-183.5	29	33.1	-4.1	16.81	0.51
183.5-187.5	76	71.3	4.7	22.09	0.31
187.5-191.5	104	104.2	-0.2	0.04	0.00
191.5-195.5	110	108.8	1.8	3.24	0.03
195.5-199.5	88	77.3	10.7	114.49	1.48
199.5-203.5	30	37.5	-7.5	56.25	1.50
203.5-207.5	6	13.0	-7.0	49.00	3.77
207.5-211.5	4	3.0	1.0	1.00	0.33
211.5-215.5	2	0.5082	1.4918	2.23	4.58
215.5-219.5	1	0.0462	0.9538	0.910	19.69
> 219.5	0	≈ 0	TOTAL:		32.67

$$\chi^2 = \sum_{i=1}^{N} \frac{(f_i - e_i)^2}{e_i}, \qquad \chi^2 = 32.67 > 19.7 = \chi^2_{0.95,11} = \chi_c$$

POR LO QUE LA HIPOTESIS NULA NO PUEDE RECHAZARSE CON UN 95% DE NIVEL DE CONFIANZA.



EJEMPLO 19

Sacar una muestra de 50 números de una tabla aleatoria, entre - 0 y 1 y probar la hipótesis de que corresponde a una distribución uniforme. Usar $\alpha = 0.05$

Solución.

Utilizando los renglones 1, 3, 5, 7, 9 de la tabla de números - aleatorios presentada en Vol. 1 de Estadística Descriptiva, multiplicando x 10^{-5} dichos números y eliminando los 3 últimos números se obtiene la siguiente muestra:

0.16 - 0.81 - 0.04 - 0.53 - 0.79 - 0.21 - 0.83 - 0.92 - 0.36 - 0.31
0.59 - 0.73 - 0.47 - 0.47 - 0.87 - 0.99 - 0.00 - 0.88 - 0.71 - 0.18
0.20 - 0.23 - 0.30 - 0.03 - 0.23 - 0.14 - 0.15 - 0.45 - 0.22 - 0.19
0.09 - 0.74 - 0.68 - 0.96 - 0.20 - 0.42 - 0.78 - 0.05 - 0.22 - 0.24
0.54 - 0.35 - 0.19 - 0.11 - 0.31 - 0.76 - 0.17 - 0.03 - 0.44 - 0.64

Agrupando datos en 10 intervalos tenemos:

Intervalo	f_i	e_i	f_i-e_i	$(f_i-e_i)^2$	$(f_i-e_i)^2/e_i$
0.000-0.105	6	5	1	1	0.20
0.105-0.205	10	5	5	25	5
0.205-0.305	7	5	2	4	0.80
0.305-0.405	4	5	-1	1	0.20
0.405-0.505	5	5	0	0	0
0.505-0.605	3	5	-2	4	0.80
0.605-0.705	2	5	-3	9	1.80
0.705-0.805	6	5	1	1	0.20
0.805-0.905	4	5	-1	1	0.20
0.905-1.005	3	5	2	4	0.80

$\Sigma = 10.0$

$$\chi^2 = \sum_{i=1}^{n} \frac{(f_i - e_i)^2}{e_i} = 10.0$$

$$\chi^2_{0.95,9} = 16.9 > 10$$

Por lo tanto se acepta la hipótesis que los números corresponden a una distribución uniforme, con un nivel de significancia de 0.05.

ESTIMACION DE UNA PROPORCION POR INTERVALOS

De la probabilidad $P(\theta - z_c\sigma_S \leq S \leq \theta + z_c\sigma_S) = 1-\alpha$ (1)

en donde α = nivel de significancia

$1-\alpha$ = nivel de confianza

z_c = valor crítico que delimita la zona de probabilidad $1-\alpha$

se obtiene la estimación de θ mediante un intervalo de confianza.

Considerando a p como la proporción de "éxitos" se tiene $\mu_x=p$, $\sigma^2(x) = p(1-p)$, $\bar{x}$ = número éxitos/n, $\mu_{\bar{x}} = p$ y $\sigma^2(\bar{x}) = p(1-p)/n$.

La ecuación 1 se puede escribir en la forma

$P(|S-\theta| \leq z_c\sigma_S) = 1-\alpha$, por lo que si $S = \bar{x}$:

$P(|\bar{x}-p| \leq z_c\sqrt{p(1-p)/n}\ ) = 1-\alpha \Leftrightarrow P((\bar{x}-p)^2 \leq z_c^2\, p(1-p)/n)$

Pero $(\bar{x}-p)^2 = z_c^2\, p(1-p)/n$ se puede escribir como

$(1 + z_c^2/n)p^2 - 2p(\bar{x} + z_c^2/(2n)) + \bar{x}^2 = 0$

Resolviendo esta ecuación de segundo grado en p se obtiene que el intervalo de confianza es

$$\Gamma\left(\bar{x} + \frac{z_c^2}{2n} - z_c\sqrt{\frac{\bar{x}(1-\bar{x})}{n} + \frac{z_c^2}{4n^2}}\ \right) \leq p \leq \Gamma\left(\bar{x} + \frac{z_c^2}{2n} + z_c\sqrt{\frac{\bar{x}(1-\bar{x})}{n} + \frac{z_c^2}{4n^2}}\ \right)$$

donde $\Gamma = \dfrac{1}{1+z_c^2/n}$.

Si n es grande y $\dfrac{z_c^2}{2n} << \bar{x}$, los límites de confianza son aproximadamente $\bar{x} \pm z_c\sqrt{\bar{x}(1-\bar{x})/n}$

Si la población es finita y el muestreo es sin reemplazo, se usa

$\sigma_S^2 \doteq \bar{x}(1-\bar{x})\dfrac{1-f}{n-1}$.

centro de educación continua
división de estudios de posgrado
facultad de ingeniería unam

FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO

INFERENCIA ESTADISTICA

M. EN I. AUGUSTO VILLARREAL ARANDA

JUNIO, 1980

INFERENCIA ESTADISTICA

Por: M en I Augusto Villarreal Aranda*

1. Introducción

La parte de la estadística que proporciona las reglas para inferir ciertas características de una población a partir de muestras extraídas de ella, junto con indicaciones probabilísticas de la veracidad de tales inferencias, se llama *inferencia estadística*.

En la inferencia estadística se estudian las relaciones existentes entre una población, las muestras obtenidas de ella, y las técnicas para estimar parámetros, tales como la media y la variancia, o bien para determinar si las diferencias entre dos muestras son debidas al azar, etc.

2. Distribuciones muestrales

Si se consideran todas las muestras posibles de tamaño

* *Secretario Académico*, División de Estudios Superiores, Facultad de Ingeniería, UNAM y *Profesor investigador*, Instituto de Ingeniería, UNAM

n que pueden extraerse de una población, y para cada una se calcula el valor del promedio aritmético, este seguramente variará de una muestra a otra, ya que depende de los valores de los datos que se hayan obtenido en cada muestra. Por lo tanto, el promedio aritmético es en sí una variable aleatoria, como también lo son, por la misma razón, el rango y la variancia de la muestra.

A todo elemento que es función de los valores de los datos que se tienen en una muestra se le denomina *estadística;* toda estadística es, entonces, una variable aleatoria cuya distribución de probabilidades se conoce como *distribución muestral.* Si, por ejemplo, la estadística considerada es la variancia de la muestra, su densidad de probabilidades se llama *distribución muestral de la variancia.*

En forma similar se pueden obtener las distribuciones muestrales de la desviación estándar, del rango, etc., cada una de las cuales tendrá sus propios parámetros, lo que permite hablar de la media y la desviación estándar de la variancia, etc.

3. Muestreo con y sin remplazo

Cuando se efectúa un muestreo en una población de tal manera que cada elemento de la misma se pueda escoger más de una vez, se dice que el muestreo es *con remplazo;* en caso contrario, el muestreo es *sin remplazo.* Si de una urna se quiere extraer una muestra de bolas de colores, se puede proceder de dos maneras: se saca al azar una bola, se anota su color y se regresa a la urna antes de obtener otra, y así sucesivamente; en este caso el muestreo es *con remplazo.* La segunda forma consiste en extraer

al azar todas las bolas que constituyen la muestra sin regresarlas a la urna, siendo entonces un muestreo *sin remplazo*.

4. Distribucion muestral del promedio aritmético

Supóngase que se extraen sin remplazo todas las muestras posibles de tamaño n de una población finita de tamaño $N_p > n$. Si la media y la desviación estándar de la distribución muestral del promedio aritmético se denotan con $\mu_{\bar{X}}$ y $\sigma_{\bar{X}}$, y la media y la desviación estándar de la población con μ y σ, respectivamente, entonces es posible demostrar que se cumplen las siguientes ecuaciones

$$\mu_{\bar{X}} = \mu$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N_p - n}{N_p - 1}}$$

Además, si la población es infinita (o el muestreo es con remplazo), los resultados anteriores se reducen a

$$\mu_{\bar{X}} = \mu$$

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

puesto que

$$\lim_{N_p \to \infty} \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N_p - n}{N_p - 1}} = \frac{\sigma}{\sqrt{n}}$$

Para valores grandes de n ($n \geq 30$) se demuestra, empleando el teorema del límite central, que la distribución muestral del promedio aritmético es aproximadamente una distribución normal con media $\mu_{\bar{X}}$ y desviación estándar $\sigma_{\bar{X}}$, independientemente de cuál sea la densidad de probabilidades de X, la variable aleatoria asociada a la población. Si esta variable tiene distribución normal, la distribución muestral del promedio aritmético también es normal, aun para valores pequeños de n ($n < 30$).

Ejemplo 4.1

Supóngase que se tiene una población finita formada por los datos 1,2,3,4,5. Se desea conocer la media y la desviación estándar de la distribución muestral del promedio aritmético, considerando las muestras de tamaño 3 obtenidas sin remplazo.

Primer procedimiento.

Siendo la población finita y el muestreo sin remplazo, es posible obtener la distribución muestral correspondiente para calcular después sus parámetros, considerando que el número total de muestras distintas de tamaño 3 que pueden obtenerse a partir de una población de 5 elementos es

$$\frac{5!}{3!(5-3)!} = 10$$

Dichas muestras son las siguientes, junto con sus promedios aritméticos correspondientes:

	$\bar{X}_i$		$\bar{X}_i$
1, 2, 3	6/3	3, 4, 5	12/3
1, 2, 4	7/3	3, 4, 1	8/3
1, 2, 5	8/3	4, 5, 1	10/3
2, 3, 4	9/3	4, 5, 2	11/3
2, 3, 5	10/3	5, 1, 3	9/3

Para calcular la media y la desviación estándar, se emplea la siguiente tabla

$\bar{X}_i$	6/3	7/3	8/3	8/3	9/3	9/3	10/3	10/3	11/3	12/3
$\bar{X}_i^2$	36/9	49/9	64/9	64/9	81/9	81/9	100/9	100/9	121/9	144/9

$$\sum_{i=1}^{10} \bar{X}_i = 90/3 \qquad \sum_{i=1}^{10} \bar{X}_i^2 = 840/9$$

$$\mu_{\bar{X}} = \bar{\bar{X}} = \frac{1}{10} \sum_{i=1}^{10} \bar{X}_i = \frac{1}{10} \cdot \frac{90}{3} = 3$$

$$\sigma_{\bar{X}}^2 = \frac{1}{10} \sum_{i=1}^{10} \bar{X}_i^2 - \bar{\bar{X}}^2 = \frac{1}{10} \cdot \frac{840}{9} - (3)^2 =$$

$$= 9.333 - 9.000 = 0.333 \Rightarrow \sigma_{\bar{X}} = \sqrt{0.333} = 0.577$$

Es decir, $\mu_{\bar{X}} = 3$ y $\sigma_{\bar{X}} = 0.577$

Segundo procedimiento.

Por tratarse de una población finita, se verifica que

$$\mu_{\bar{X}} = \mu \qquad \text{y} \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N_p - n}{N_p - 1}}$$

en donde $N_p = 5$, $n = 3$ y $\mu = 3$.

El valor de σ^2 de la población es

$$\sigma^2 = \frac{1+4+9+16+25}{5} - (3)^2 = \frac{55}{5} - 9 = 11-9 = 2$$

Por lo tanto, $\sigma = \sqrt{2} = 1.4145$ y

$$\sigma_{\bar{X}} = \frac{1.4145}{\sqrt{3}} \sqrt{\frac{5 - 3}{5 - 1}} = (0.8164)(0.7071) = 0.577$$

Es decir, $\mu_{\bar{X}} = 3$ y $\sigma_{\bar{X}} = 0.577$

Comparando los resultados, se puede observar que ambos procedimientos conducen a la obtención de los mismos valores de $\mu_{\bar{X}}$ y $\sigma_{\bar{X}}$ para la distribución muestral del promedio aritmético.

Ejemplo 4.2

En una bodega se tienen cinco mil varillas de acero; el valor medio del peso, X, de cada varilla es de 5.02 kg, y la desviación estándar 0.3 kg. Hallar la probabilidad de que una muestra de cien varillas, escogida al azar, tenga un peso total

a. entre 496 y 500 kg

b. de más de 510 kg.

Para la distribución muestral del promedio, se tiene que $\mu_{\bar{X}} = \mu = 5.02$ kg y, por tratarse de una población finita,

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N_p - n}{N_p - 1}} = \frac{0.30}{\sqrt{100}} \sqrt{\frac{5000 - 100}{5000 - 1}} = 0.027$$

a. El peso total de la muestra estará entre 496 y 500 kg si el peso promedio de las cien varillas se encuentra entre 4.96 y 5.00 kg. Puesto que la muestra es mayor de 30 elementos se puede considerar como aproximadamente normal a la distribución muestral, y los valores estándar correspondientes a $\bar{X} = 4.96$ y a $\bar{X} = 5.00$ se obtienen mediante la transformación

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

es decir,

$$Z_1 = \frac{4.96 - 5.02}{0.027} = -2.22$$

$$Z_2 = \frac{5.00 - 5.02}{0.027} = -0.74$$

En la fig 4.1 se puede apreciar que

$$P[496 \leq X \leq 500] = P[-2.22 \leq Z \leq -0.74] =$$

$$= P[-2.22 \leq Z \leq 0] - P[-0.74 \leq Z \leq 0]$$

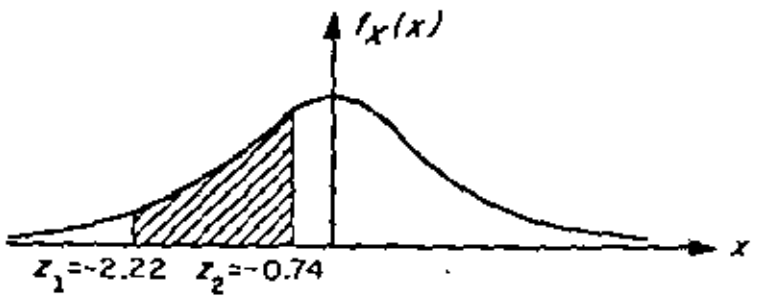


Fig 4.1 Distribución normal correspondiente al ejemplo

Recurriendo a la tabla de áreas bajo la curva normal estándar entre 0 y Z queda finalmente

$$P[496 \leq X \leq 500] = 0.4868 - 0.2704 = 0.2164$$

b. El peso total de la muestra excederá de 510 kg si el peso promedio de las cien varillas pasa de 5.10 kg.

Estandarizando dicho valor, queda

$$Z_3 = \frac{5.10 - 5.02}{0.027} = 2.96$$

Calculando el área bajo la curva normal a la derecha de este valor (fig 4.2), se tiene que

$$P[X \geq 510] = P[Z \geq 2.96] = P[Z > 0] - P[0 \leq Z \leq 2.96] =$$

$$= 0.5 - 0.4985 = 0.0015$$

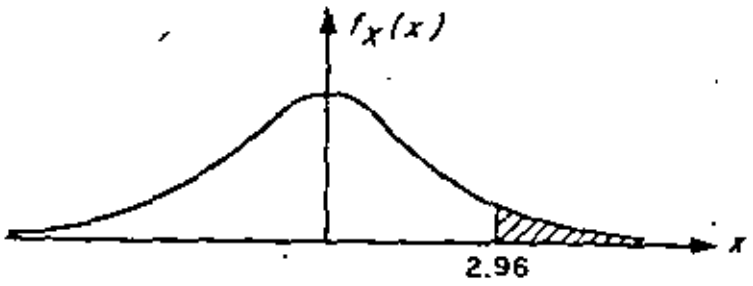


Fig 4.2 Distribución normal correspondiente al ejemplo

5. Distribución muestral de diferencias de promedios aritméticos

Con frecuencia se presenta el caso en el que se tienen datos de dos poblaciones con variables aleatorias asociadas X y Y, respectivamente, surgiendo la duda de si estas se pueden considerar como una sola, es decir, si X = Y. Para probar estadísticamente esta hipótesis (como se verá más adelante), es necesario obtener las distribuciones muestrales de la diferencia de los promedios y de las variancias de las muestras de ambas variables.

Sean $\bar{X}$ y $\bar{Y}$ los promedios aritméticos obtenidos de muestras aleatorias de tamaño n_X y n_Y de dos poblaciones con características X y Y, respectivamente. Se puede demostrar que la distribución muestral de la diferencia de los promedios correspondientes a poblaciones infinitas con medias μ_X y μ_Y y desviaciones estándar σ_X y σ_Y, tiene los siguientes parámetros:

$$\mu_{\bar{X} - \bar{Y}} = \mu_{\bar{X}} - \mu_{\bar{Y}} = \mu_X - \mu_Y$$

$$\sigma_{\bar{X} - \bar{Y}} = \sqrt{\sigma^2_{\bar{X}} + \sigma^2_{\bar{Y}}} = \sqrt{\frac{\sigma^2_X}{n_X} + \frac{\sigma^2_Y}{\ell}}$$

si las muestras son independientes.

Esta distribución también es aplicable a poblaciones finitas si el muestreo es con remplazo. Para el caso de poblaciones finitas en las cuales el muestreo se hace sin remplazo, los parámetros de la distribución muestral de la diferencia de los promedios aritméticos son

$$\mu_{\bar{X}-\bar{Y}} = \mu_{\bar{X}} - \mu_{\bar{Y}} = \mu_X - \mu_Y$$

$$\sigma_{\bar{X}-\bar{Y}} = \sqrt{\sigma^2_{\bar{X}} + \sigma^2_{\bar{Y}}} = \sqrt{\frac{\sigma^2_X}{n_X}\,\frac{N_X - n_X}{N_X - 1} + \frac{\sigma^2_X}{n_Y}\cdot\frac{N_Y - n_Y}{N_Y - 1}}$$

suponiendo que las muestras sean independientes.

Ejemplo 5.1

Considérese que de una población X se obtienen tres muestras posibles, cuyos correspondientes promedios aritméticos son 3, 7 y 8. De otra población Y se extraen dos muestras posibles, con promedios 2 y 4, respectivamente. Se deben obtener los parámetros de la distribución muestral de las diferencias de los promedios aritméticos.

Primer procedimiento

Todas las posibles diferencias de promedios aritméticos de X con los de Y serían

$$\begin{array}{lll} 3-2 & 7-2 & 8-2 \\ 3-4 & 7-4 & 8-4 \end{array} \Longrightarrow \begin{array}{lll} 1 & 5 & 6 \\ -1 & 3 & 4 \end{array}$$

Es decir,

$$\mu_{\bar{X}-\bar{Y}} = \frac{-1+1+3+4+5+6}{6} = \frac{18}{6} = 3$$

$$\sigma^2_{\bar{X}-\bar{Y}} = \frac{(-1-3)^2 + (1-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2 + (6-3)^2}{6} =$$

$$= \frac{34}{6} = \frac{17}{3}$$

Segundo procedimiento

Se sabe que

$$\mu_{\bar{X}-\bar{Y}} = \mu_{\bar{X}} - \mu_{\bar{Y}} \; ; \; \sigma^2_{\bar{X}-\bar{Y}} = \sigma^2_{\bar{X}} + \sigma^2_{\bar{Y}}$$

Por ello,

$$\mu_{\bar{X}} = \frac{3+7+8}{3} = \frac{18}{3} = 6$$

$$\mu_{\bar{Y}} = \frac{2+4}{2} = \frac{6}{2} = 3$$

$$\sigma^2_{\bar{X}} = \frac{(3-6)^2 + (7-6)^2 + (8-6)^2}{3} = \frac{14}{3}$$

$$\sigma^2_{\bar{Y}} = \frac{(2-3)^2 + (4-3)^2}{2} = \frac{2}{2} = 1$$

$$\mu_{\bar{X}-\bar{Y}} = 6 - 3 = 3$$

$$\sigma^2_{\bar{X}-\bar{Y}} = \frac{14}{3} + 1 = \frac{17}{3}$$

Se observa que ambos procedimientos conducen a los mismos resultados.

Ejemplo 5.2

Las varillas de acero que fabrica una compañía A tienen un peso medio de 6.5 kg y una desviación estándar de 0.4, en tanto que las producidas por una empresa B tienen un peso medio de 6.3 kg y una desviación estándar de 0.3 kg. Si se toman muestras aleatorias de 100 varillas de cada fábrica, ¿cuál es la probabilidad de que las de la compañía A tengan un peso promedio de por lo menos

a. 0.35 kg

b. 0.10 kg

mayor que el de la compañía B?

Se puede suponer en este caso que las distribuciones muestrales involucradas son normales, en virtud de que el tamaño de ambas muestras es mayor de 30 elementos. También se puede suponer que ambas poblaciones son infinitas, y siendo $\bar{X}_A$ y $\bar{X}_B$ los pesos promedios de las muestras de las fábricas A y B, respectivamente, entonces

$$\mu_{\bar{X}_A - \bar{X}_B} = \mu_{\bar{X}_A} - \mu_{\bar{X}_B} = 6.5 - 6.3 = 0.20 \text{ kg}$$

$$\sigma_{\bar{X}_A - \bar{X}_B} = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} = \sqrt{\frac{(0.4)^2}{100} + \frac{(0.3)^2}{100}} = 0.05 \text{ kg}$$

La variable estandarizada de la diferencia de los promedios es

$$Z = \frac{(\bar{X}_A - \bar{X}_B) - \mu_{\bar{X}_A - \bar{X}_B}}{\sigma_{\bar{X}_A - \bar{X}_B}} = \frac{(\bar{X}_A - \bar{X}_B) - 0.20}{0.05}$$

a. Estandarizando la diferencia de 0.35 kg se llega a

$$Z_1 = \frac{0.35 - 0.20}{0.05} = \frac{0.15}{0.05} = 3$$

La probabilidad deseada es el área bajo la curva normal a la derecha de $Z = 3$, es decir

$$P\left[\bar{X}_A > \bar{X}_B + 0.35\right] = P\left[Z > 3\right] = 0.500 - 0.4987 = 0.0013$$

b. Al estandarizar la diferencia de 0.10 kg, la variable Z resulta

$$Z_2 = \frac{0.10 - 0.20}{0.05} = \frac{-0.1}{0.05} = -2$$

La probabilidad requerida es el área bajo la curva normal a la derecha de $Z = -2$, es decir

$$P\left[\bar{X}_A > \bar{X}_B + 0.10\right] = P\left[Z > -2\right] = 0.5 + 0.4772 = 0.9772$$

6. Teoría estadística de la estimación

En la práctica profesional a menudo resulta necesario inferir información acerca de una población mediante el uso de muestras extraídas de ella; una parte básica de dicha inferencia consiste en *estimar* los valores de los parámetros de la población (media, variancia, etc.) a partir de las estadísticas correspondientes de la muestra, como se explica a continuación.

7. Estimadores puntuales. Clasificación

Si un estimador de un parámetro de la población consiste en un solo valor de una estadística, se le conoce como *estimador puntual* del parámetro.

Cuando la media de la distribución muestral de una estadística es igual al parámetro que se está estimando de la población, entonces la estadística se conoce como *estimador insesgado* del parámetro; si no sucede así, entonces se denomina *estimador sesgado*. Ambos estimadores son puntuales, y sus valores correspondientes se llaman estimaciones insesgadas o sesgadas, respectivamente. Dicho de otra manera, si S es una estadística cuya distribución muestral tiene media μ_S, y el parámetro correspondiente de la población es θ, se dice que S es un estimador insesgado de θ si

$$\mu_S = \theta$$

Por otra parte, si la estadística S_n de la muestra tiende a ser igual al parámetro θ de la población a medida que se

hace más grande el tamaño de la muestra, entonces la estadística recibe el nombre de *estimador consistente* del parámetro.

Empleando símbolos, si

$$\lim_{n\to\infty} S_n = \theta$$

resulta que la estadística S_n es un estimador consistente. Por ejemplo, el promedio aritmético es un estimador insesgado y consistente de la média, y la variancia de la muestra es un estimador sesgado y consistente de la variancia de la población.

Si las distribuciones muestrales de varias estadísticas tienen el mismo valor de la media, se dice que la estadística que cuenta con la menor variancia es un *estimador eficiente* de dicha media, en tanto que las estadísticas restantes se conocen como *estimadores ineficientes* del parámetro.

Por ejemplo, las distribuciones muestrales del promedio aritmético y de la mediana cuentan con medias que son, en ambos casos, iguales a la media de la población. Sin embargo, la variancia de la distribución muestral del promedio aritmético es menor que la de la distribución de la mediana, por lo que el promedio aritmético obtenido de una muestra aleatoria proporciona un estimador eficiente de la media de la población, en tanto que la mediana obtenida de la muestra proporciona un estimador ineficiente de dicho parámetro.

8. Estimación de intervalos de confianza para los parámetros de una población

La estimación de un parámetro de una población mediante un par de números entre los cuales se encuentra, con cierta probabilidad, el valor de dicho parámetro, se llama estimación del intervalo del mismo.

Sea S una estadística obtenida de una muestra de tamaño n para estimar el valor del parámetro θ, y sea σ_S la desviación estándar (conocida o estimada) de su distribución muestral. La probabilidad, $1 - \alpha$, de que el valor de θ se localice en el intervalo de $S - z_c \sigma_S$ a $S + z_c \sigma_S$, donde z_c es una constante, se escribe en la forma

$$P[S - z_c \sigma_S \leq \theta \leq S + z_c \sigma_S] = 1 - \alpha$$

Si se fija el valor de $1 - \alpha$, se puede obtener el valor de z_c necesario para que se satisfaga la ecuación anterior, con lo cual queda definido el *intervalo de confianza* del parámetro θ, $(S - z_c \sigma_S, S + z_c \sigma_S)$, correspondiente al nivel de *confianza* $1 - \alpha$.

La constante z_c que fija el intervalo de confianza se conoce como *valor crítico*. Si la distribución de S es normal, el valor de z_c correspondiente a uno de α se obtiene de la tabla de áreas bajo la curva normal o de la tabla 8.1 siguiente.

TABLA 8.1 VALORES DE z_c PARA DISTINTOS NIVELES DE CONFIANZA

Nivel de confianza, en porcentaje	z_c
99.73	3.00
99.00	2.58
98.00	2.33
96.00	2.05
95.45	2.00
95.00	1.96
90.00	1.64
80.00	1.28
68.27	1.00
50.00	0.674

Ejemplo 8.1

Sea el promedio aritmético $\bar{X}$ una estadística con distribución normal. Las probabilidades o niveles de confianza de que $\mu_{\bar{X}}$ (o μ de la población) se encuentre localizada entre los límites $\bar{X} \pm \sigma_{\bar{X}}$, $\bar{X} \pm 2\,\sigma_{\bar{X}}$ y $\bar{X} \pm 3\,\sigma_{\bar{X}}$ son 68.26, 95.44 y 99.73%, respectivamente, obteniéndose dichos valores de la tabla de áreas bajo la curva normal. Lo anterior significa que el intervalo $\bar{X} \pm 3\,\sigma_{\bar{X}}$ contendrá a $\mu_{\bar{X}}$ en el 99.73 por ciento de las muestras de tamaño n, por lo que los intervalos de confianza de 68.26, 95.44 y 99.73 por ciento para estimar a μ son $(\bar{X} - \sigma_{\bar{X}}, \bar{X} + \sigma_{\bar{X}})$ $(\bar{X} - 2\,\sigma_{\bar{X}}, \bar{X} + 2\,\sigma_{\bar{X}})$ y $(\bar{X} - 3\,\sigma_{\bar{X}}, \bar{X} + 3\,\sigma_{\bar{X}})$, lo cual se aprecia en la *fig* 8.1 siguiente.

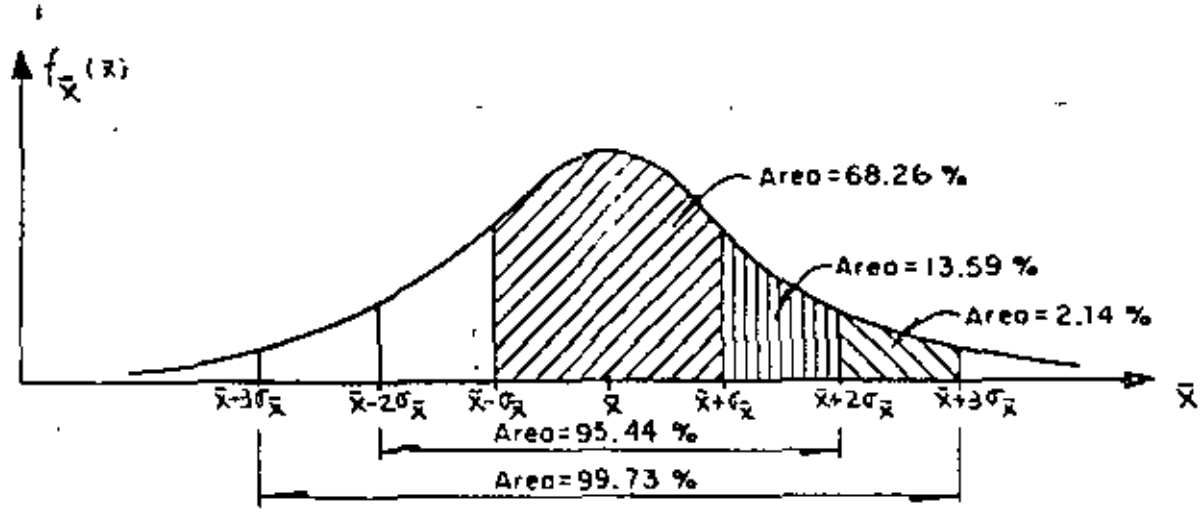


Fig 8.1

9. Estimación de intervalos de confianza para la media

Los límites de confianza para la media de una población con variable aleatoria X asociada están dados por

$$\bar{X} \pm z_c \, \sigma_{\bar{X}}$$

en donde z_c depende del nivel de confianza deseado. Si $\bar{X}$ tiene distribución normal, z_c puede obtenerse en forma directa de la tabla 8.1. Por ejemplo, los límites de confianza de 95 y 99 por ciento para estimar la media, μ, de la población son $\bar{X} \pm 1.96\sigma_{\bar{X}}$ y $\bar{X} \pm 2.58\ \sigma_{\bar{X}}$, respectivamente. Al obtener estos límites hay que usar el valor calculado de $\bar{X}$ para la muestra correspondiente.

Entonces, los límites de confianza para la media de la población quedan dados por

$$\bar{X} \pm z_c \frac{\sigma}{\sqrt{n}}$$

en caso de que el muestreo se haga a partir de una población infinita o de que se efectúe con remplazo a partir de una población finita, o por

$$\bar{X} \pm z_c \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N_p - n}{N_p - 1}}$$

si el muestreo es sin remplazo a partir de una población finita de tamaño N_p.

Ejemplo 9.1

Las mediciones de los diámetros de una muestra aleatoria de 100 tubos de albañal mostraron una media de 32 cm y una desviación estándar de 2 cm. Obténganse los límites de confianza de

a. 95 por ciento

b. 97 por ciento

para el diámetro medio de todos los tubos.

a. De la tabla 8.1, los límites de confianza del 95 por ciento son

$$\bar{X} \pm 1.96\sigma/\sqrt{n} = 32 \pm 1.96(2/\sqrt{100}) = 32 \pm 0.392 \text{ cm}$$

o sea 31.608 y 32.392, en donde se ha empleado el valor de S_X para estimar el de σ de la población, puesto que la muestra es suficientemente grande (mayor de 30 elementos). Esto significa

que con una probabilidad de 95 por ciento, el valor de μ_X se encuentra entre 31.608 y 32.392 cm.

b. Si $Z = z_c$ es tal que el área bajo la curva normal a la derecha de z_c es el 1.5 por ciento del área total, entonces el área entre 0 y z_c es 0.5 - 0.015 = 0.485, por lo que de la tabla de áreas bajo la curva normal se obtiene $z_c = 2.17$. Por lo tanto, los límites de confianza del 97 por ciento son:

$$\bar{X} \pm 2.17\sigma/\sqrt{n} = 32 \pm 2.17(2/\sqrt{100}) = 32 \pm 0.434 \text{ cm}$$

y el intervalo de confianza respectivo es (31.566 cm, 32.434 cm).

Ejemplo 9.2

Una muestra aleatoria de 50 calificaciones de cierto examen de admisión tiene un promedio aritmético de 72 puntos, con desviación estándar igual a 10. Si el examen se aplicó a 1018 personas, obtener

a. El intervalo de confianza del 95% para la media del total de calificaciones.

b. El tamaño de muestra necesario para que el error en la estimación de la media no exceda de 2 puntos, considerando el mismo nivel de confianza.

c. El nivel de confianza para el cual la media de la población sea 72 ± 1 puntos.

a. Si se estima a σ de la población con S_X de la muestra y se considera que la población es finita, los límites de confianza son, puesto que $\bar{X} = 72$, $Z_c = 1.96$, $S_X = 10$, $N_p = 1018$ y $n = 50$,

$$72 \pm 1.96 \frac{10}{\sqrt{50}} \sqrt{\frac{1018 - 50}{1018 - 1}}$$

$$72 \pm 1.96 \ (1.4142) \ (0.9755)$$

$$72 \pm 2.704$$

y el intervalo de confianza respectivo es

$$(69.296, \ 74.704)$$

b. Puesto que el error en la estimación de la media es, para población finita,

$$\text{Error en la estimación} = Z_c \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N_p - n}{N_p - 1}}$$

en este caso se tendría

$$Z_c \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N_p - n}{N_p - 1}} < 2$$

o sea, para un nivel de confianza de 95%,

$$1.96 \frac{10}{\sqrt{n}} \sqrt{\frac{1018 - n}{1018 - 1}} < 2$$

$$\frac{19.6}{\sqrt{n}} \sqrt{\frac{1018 - n}{1018 - }} < 2$$

Elevando al cuadrado la desigualdad, queda

$$\frac{384.16}{n} \quad \frac{1018 - n}{1017} < 4$$

o sea

$$87.85 < n$$

Por lo cual, se requieren al menos 88 elementos en la muestra para que el error en la estimación no exceda de 2 puntos, para $1 - \alpha = 0.95$.

c. Los límites de confianza son, en este caso

$$72 \pm Z_c \frac{10}{\sqrt{50}} \sqrt{\frac{1018 - 50}{1018 - 1}}$$

$$72 \pm Z_c \ (1.4142)(0.9755)$$

o sea

$$72 \pm 1.3795 \ Z_c$$

Puesto que se desea que el valor de la media sea 72 ± 1 puntos, se verifica que

$$1 = 1.3795 \ Z_c$$

Es decir

$$Z_c = \frac{1}{1.3795} = 0.725$$

El área bajo la curva normal estándar entre 0 y $z_c = 0.725$ es, por interpolación lineal, igual a 0.2657. Por lo tanto, el nivel de confianza es igual al doble del área anterior, es decir, 2(0.2657) = 0.5314 (o 53.14%), tal como se muestra en la fig 9.1.

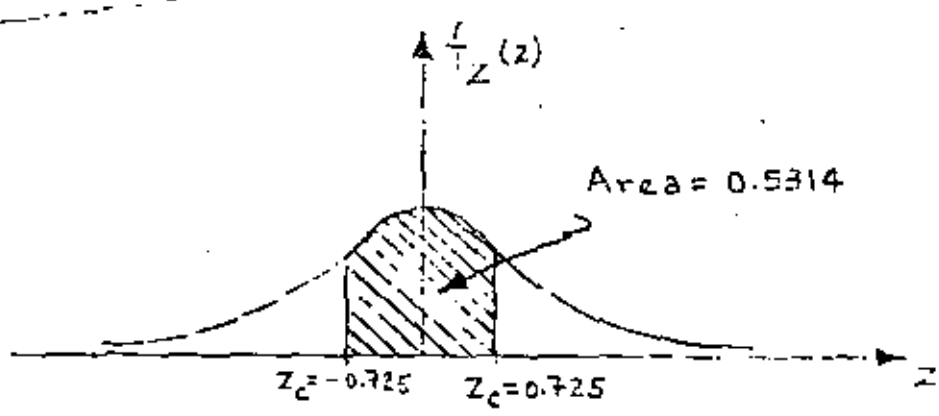


Fig 9.1

10. Intervalos de confianza para diferencias de medias

Los límites de confianza para la diferencia de las medias cuando las poblaciones X y Y son infinitas, o cuando el muestreo se realiza con remplazo de poblaciones finitas, se encuentran dados por

$$\bar{X} - \bar{Y} \pm z_c \sigma_{\bar{X} - \bar{Y}} = \bar{X} - \bar{Y} \pm z_c \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$$

en donde $\bar{X}$, n_X y $\bar{Y}$, n_Y son los respectivos promedios aritméticos y tamaños de las dos muestras extraídas de las poblaciones, y σ_X y σ_Y las desviaciones estándar de estas últimas.

En el caso de que las poblaciones X y Y sean finitas y el muestreo sin remplazo, los límites de confianza son

$$\bar{X} - \bar{Y} \pm Z_c \sigma_{\bar{X}-\bar{Y}} = \bar{X} - \bar{Y} \pm Z_c \sqrt{\frac{\sigma_X^2}{n_X}\,\frac{N_X - n_X}{N_X - 1} + \frac{\sigma_Y^2}{n_Y}\,\frac{N_Y - n_Y}{N_Y - 1}}$$

en donde N_X y N_Y son los tamaños de las poblaciones X y Y, respectivamente.

Las dos ecuaciones anteriores son válidas únicamente si las muestras aleatorias seleccionadas son independientes.

Ejemplo 10.1

Para el ejemplo de las varillas tratado anteriormente (5.2), encontrar el intervalo de confianza del 95.45% para las diferencias de las medias de las poblaciones.

Siendo $\bar{X}_A = \mu_A = 6.5$ kg, $\sigma_A = 0.4$ kg, $\bar{X}_B = \mu_B = 6.3$ kg, $\sigma_B = 0.3$ kg y $n_A = n_B = 100$, los límites de confianza para la diferencia de las medias son, empleando la tabla B.1

$$\bar{X}_A - \bar{X}_B \pm z_c \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} = 6.5 - 6.3 \pm 2 \sqrt{\frac{(0.4)^2}{100} + \frac{(0.3)^2}{100}} =$$

$$= 0.2 \pm 0.1$$

Por lo tanto, el intervalo de confianza respectivo es (0.1, 0.3).

Ejemplo 10.2

Se tienen en una bodega 3000 focos de marca X, y 5000 de marca Y. Se extrae una muestra aleatoria de 150 focos de la marca X, y se obtiene una duración promedio de 1400 horas, con desviación estándar igual a 120 horas. Otra muestra aleatoria de 200 focos de la marca Y tuvo una duración promedio de 1200 horas, con desviación estándar igual a 80 horas. Obtener intervalos de confianza de

a. 95%

b. 99%

para la diferencia de los tiempos medios de duración de los focos de ambas marcas.

a: Puesto que se trata de poblaciones finitas y $\bar{X} = 1400$ h, $S_X = 120$ h, $N_X = 3000$, $n_X = 150$, $\bar{Y} = 1200$ h, $S_Y = 80$ h, $N_Y = 5000$ y $n_Y = 200$, se obtiene, estimando a σ_X y σ_Y con S_X y S_Y, respectivamente

$$1400 - 1200 \pm 1.96 \sqrt{\frac{(120)^2}{150}\,\frac{3000 - 150}{3000 - 1} + \frac{(80)^2}{200}\,\frac{5000 - 200}{5000 - 1}}$$

$$200 \pm 1.96\ (11.04)$$

$$200 \pm 21.638$$

o sea, (178.362, 221.638), puesto que de la tabla 8.1, para un nivel de confianza de 95%, $Z_c = 1.96$.

b. En este caso, al emplear la tabla 8.1 se obtiene

$Z_c = 2.58$ para un nivel de confianza de 99%, por lo cual los límites son

$$1400 - 1200 \pm 2.58 \sqrt{\frac{(120)^2}{150}\frac{3000 - 150}{3000 - 1} + \frac{(80)^2}{200}\frac{5000 - 2000}{5000 - 1}}$$

$$200 \pm 2.58\ (11.04)$$

$$200 \pm 28.483$$

y el intervalo de confianza es

$$(171.517,\ 228.483)$$

11. Pruebas de hipótesis

Supóngase que una empresa armadora de automóviles está en la disyuntiva de emplear una nueva marca de bujías en sus unidades o la que regularmente utiliza, y que su departamento de control de calidad debe decidir, con base en la información de las muestras de las dos marcas distintas. Las decisiones de este tipo, es decir, que se basan en estudios estadísticos, reciben el nombre de *decisiones estadísticas*, y a los procedimientos que permiten decidir si se acepta o rechaza una hipótesis se les llama *pruebas de hipótesis, pruebas de significancia* o *reglas de decisión.*

Al tomar decisiones estadísticas, es necesario postular las diversas alternativas o cursos de acción que pueden adoptarse.

En el caso particular de una *prueba de hipótesis* solamente se tienen dos cursos de acción posibles, los que se denotarán como H_0 y H_1. A la acción H_0 se le llama *hipótesis nula*, y a la H_1, *hipótesis alternativa*. Por ejemplo, si la hipótesis nula establece que $\mu_1 = \mu_2$, la hipótesis alternativa puede ser una de las siguientes:

$$\mu_1 > \mu_2, \mu_1 < \mu_2 \text{ o } \mu_1 \neq \mu_2$$

Al realizar una prueba de hipótesis, se prueba siempre la verdad de la hipótesis nula H_0, aun cuando de antemano se desee rechazarla.

12. Errores de los tipos I y II. Nivel de significancia

En muchas ocasiones se presenta el caso de que se rechaza una hipótesis nula cuando en realidad debería ser aceptada; cuando esto sucede se dice que se ha cometido un *error de tipo I*. En otras ocasiones se acepta una hipótesis nula siendo en realidad falsa; en este caso se dice que se ha cometido un *error de tipo II*.

Al probar una hipótesis nula, a la máxima probabilidad con la que se está dispuesto a cometer un error del tipo I se le llama *nivel de significancia*, α, de la prueba, el cual dentro de la práctica se acostumbra establecer de 5 por ciento (0.05) o 10 por ciento (0.1). El complemento del nivel de significancia, $1 - \alpha$, se conoce como *nivel de confianza*.

Si, por ejemplo, al realizar una prueba de hipótesis se escoge un nivel de significancia de 10 por ciento, significa que existen 10 posibilidades en 100 de que se rechace ésta cuando debería ser aceptada; es decir, que se rechaza a un nivel de significancia del 10 por ciento, y que la probabilidad de que la decisión haya sido errónea es de 0.1.

13. Comportamiento de los errores tipos I y II

Supóngase que se trata de probar la hipótesis nula de que la media, μ_S, de la distribución muestral de la estadística S es μ_1, en contra de la hipótesis alternativa que establece que $\mu_S = \mu_2$, donde $\mu_2 > \mu_1$, es decir

$$H_0 : \mu_S = \mu_1$$

$$H_1 : \mu_S = \mu_2$$

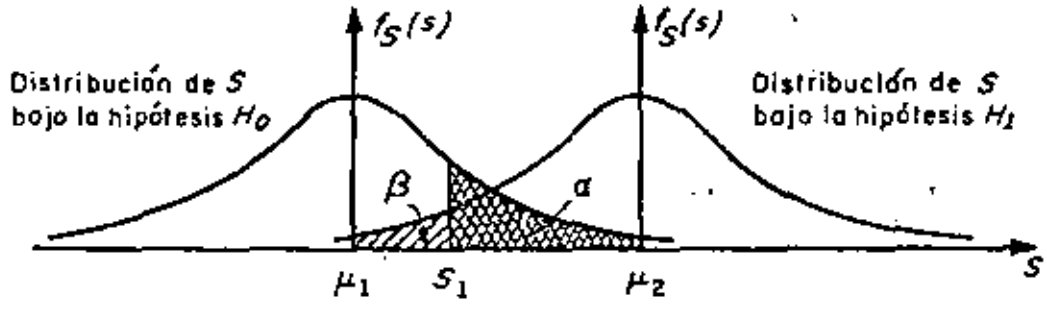
En la fig 13.1 se muestra en forma gráfica la relación entre los errores tipos I y II en el caso en el que la regla de decisión para aceptar o rechazar H_0 es la siguiente:

> *Si el valor de la estadística S obtenido de una muestra excede de cierto valor crítico S_1, rechácese H_0; en caso contrario, acéptese.*

Es evidente que si H_0 es verdadera, entonces α (área con rayado doble) es la probabilidad de que $S > S_1$, o sea la de rechazar a H_0 siendo verdadera (error tipo I). Por otro lado, si H_1 es verdadera, entonces β (área con rayado sencillo) es la probabilidad

de que $S < S_1$, o sea la de aceptar H_0 siendo falsa (error tipo II).

Obsérvese que si se aumenta el valor de S_1 se reduce la probabilidad α, pero se incrementa la β; lo contrario sucede si se disminuye el valor de S_1.



$$P[S>S_1] = \alpha \text{ (error tipo I)}$$
$$P[S<S_1] = \beta \text{ (error tipo II)}$$

Fig 13.1 Probabilidades de los errores tipos I y II en pruebas de hipótesis.

En realidad, la única forma posible en la cual se pueden minimizar simultáneamente los errores de tipos I y II es aumentando el tamaño de la muestra, para hacer más "picudas" las distribuciones muestrales de la estadística bajo las hipótesis H_0 y H_1.

Al observar la fig 13.2 siguiente, es posible concluir

que el tamaño de los errores I y II es menor para un tamaño de muestra igual a 100 que para un tamaño igual a 50, considerando la misma regla de decisión anterior.

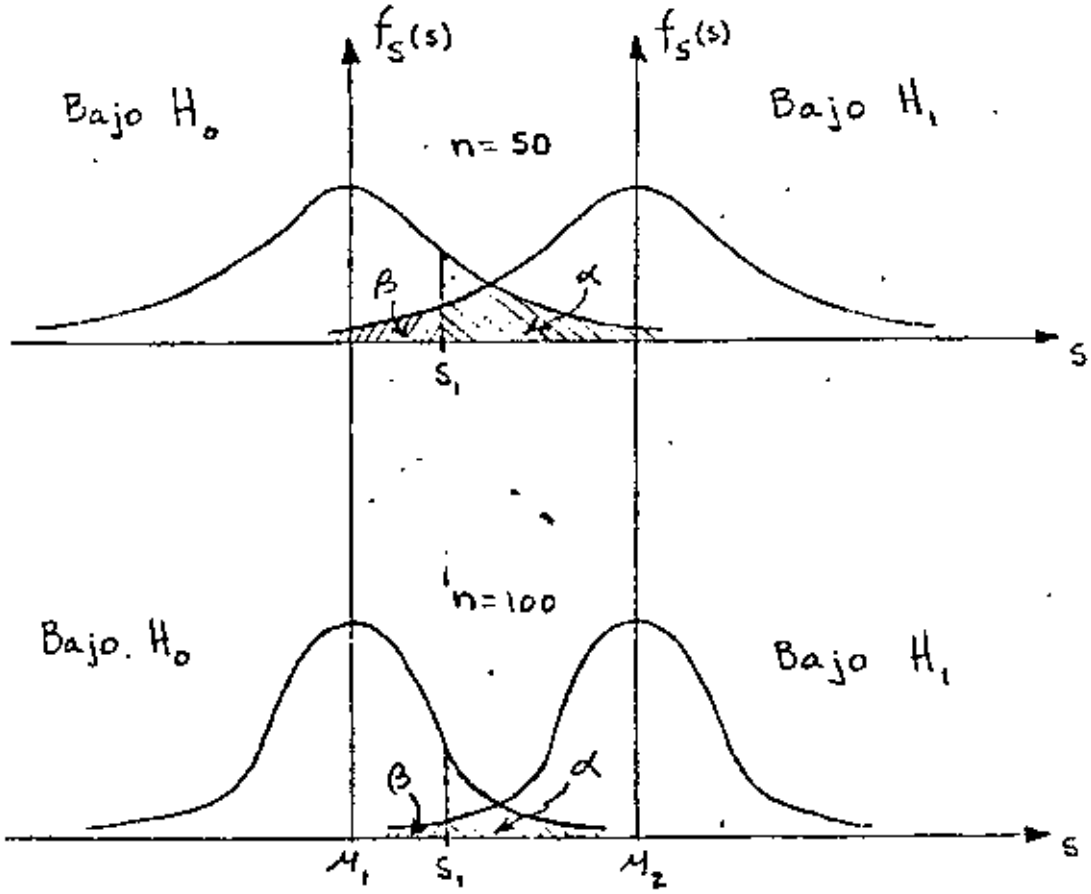


Fig 13.2

Sin embargo, esta técnica de reducción simultánea de ámbos tipos de errores no siempre puede ponerse en práctica, debid a razones de costo, tiempo, etc.

14. Regiones críticas, de rechazo ó de significancia. Regiones de aceptación.

Cuando una hipótesis nula no se acepta se dice que se *rechaza a un nivel de significancia del* α *por ciento*, o que el valor estandarizado de la estadística involucrada es *significativo a un nivel de significancia* α.

Al conjunto de los valores de la estadística en el que se rechaza la hipótesis nula se le denomina *región crítica, de rechazo, o de significancia*. Por el contrario, al conjunto de los valores de la estadística en que se acepta la hipótesis, se le llama *región de aceptación*.

Considérese que la distribución muestral de la estadística S es normal con desviación estándar σ_S, que la variable Z resulta de estandarizar a S, que la hipótesis nula, H_0, es que la media de S vale μ_S, y que la hipótesis alternativa H_1 es que dicha media es diferente de μ_S, es decir, que

$$Z = \frac{S - \mu_S}{\sigma_S}$$

H_0: media de la distribución muestral de $S = \mu_S$

H_1: media de la distribución muestral de $S \neq \mu_S$

Si se adopta la regla de decisión de aceptar la hipótesis H_0, si el valor de Z caé dentro del intervalo central que encierra al 99 por ciento del área de la distribución de probabilidades, entonces H_0 se aceptará en el caso en que

$$-2.58 \leq Z \leq 2.58$$

empleando la tabla de áreas bajo la curva normal estándar. Pero si el valor estandarizado de la estadística se encuentra fuera de dicho intervalo, se concluye que el evento puede ocurrir con probabilidad de 0.01 si la hipótesis H_0 es verdadera (área rayada total de la fig 14.1). En tal caso, el valor Z de la variable estándar difiere *significativamente* del que se podría esperar de acuerdo con la hipótesis nula, lo cual inclina a rechazarla a un nivel de confianza del 99 por ciento.

De lo anterior de deduce que el área total rayada de la fig 14.1 es el nivel de significancia α de la prueba, y representa la probabilidad de cometer un error del tipo I. Por ello, la región de aceptación de H_0 es $-2.58 \leq Z \leq 2.58$, y la de rechazo es $Z > 2.58$ y $Z < -2.58$.

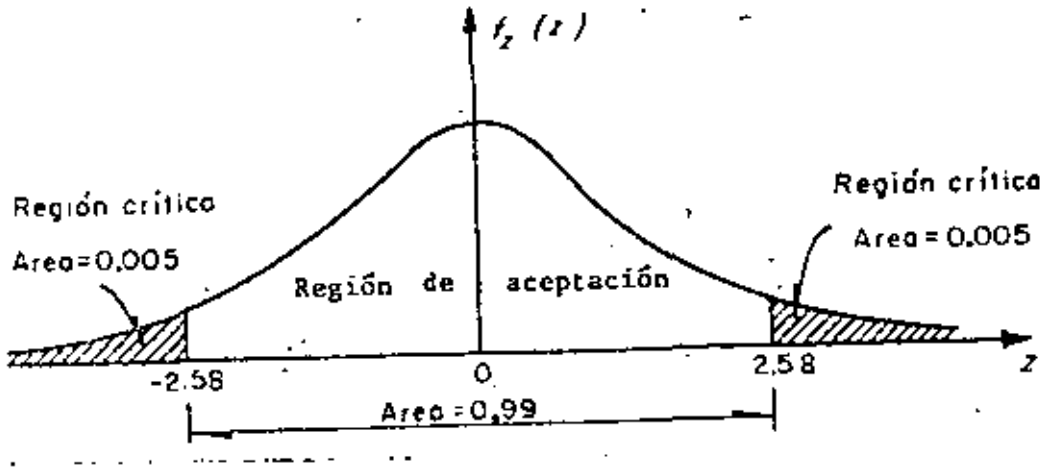


Fig 14.1 Región de significancia

En la tabla 14.1 se presentan los valores de la variable estandarizada, z, que limitan las regiones de aceptación y de rechazo para el caso en el que la estadística involucrada en la prueba tenga distribución muestral normal. Cuando en alguna prueba de hipótesis se consideren niveles de significancia diferentes a los que aparecen en la tabla mencionada, resulta necesario emplear la de áreas bajo la curva normal estándar.

TABLA 14.1 VALORES CRITICOS DE z

Nivel de significancia, α	Valores de z para pruebas de una cola	Valores de z para pruebas de dos colas
0.1	−1.281 o 1.281	−1.645 y 1.645
0.05	−1.645 o 1.645	−1.960 y 1.960
0.01	−2.326 o 2.326	−2.575 y 2.575
0.005	−2.575 o 2.575	−2.810 y 2.810

15. Pruebas de una y de dos colas

En la prueba de hipótesis del ejemplo anterior, la región de rechazo de la hipótesis nula quedó en ambos extremos (colas) de la distribución muestral de la estadística involucrada en la prueba; a las pruebas de este tipo se les denomina *pruebas de dos colas*. Cuando la región de rechazo se encuentra solamente en un extremo de la distribución muestral en cuestión, se les llama *pruebas de una cola*.

Las pruebas de dos colas se presentan cuando en la hipótesis alternativa aparece el signo $\neq$ (diferente de), como en el siguiente caso

$$H_0 : \mu_S = \mu_1 ,$$

$$H_1 : \mu_S \neq \mu_1$$

en donde μ_S es la media de la estadística S, y μ_1 es un valor fijo.

En los casos

$$H_0 : \mu_S = \mu_1$$

$$H_1 : \mu_S < \mu_1$$

y

$$H_0 : \mu_S = \mu_1$$

$$H_1 : \mu_S > \mu_1$$

las pruebas resultan de una cola.

16. Pruebas de hipótesis para la media

Para el caso de una población infinita (o finita en que se muestree con remplazo), cuya desviación estándar σ se conoce o se puede estimar adecuadamente, si se tiene que la estadística S obtenida de la muestra es el promedio aritmético, entonces la media de su distribución muestral es $\mu_S = \mu_{\bar{X}} = \mu$, y su desviación estándar es $\sigma_S = \sigma_{\bar{X}} = \sigma/\sqrt{n}$, en donde μ y σ son, respectivamente, la media y la desviación estándar de la variable aleatoria X asociada a la población, y n es el tamaño de la muestra. En tal caso, si $\bar{X}$ tiene distribución normal, la variable estandarizada correspondiente será

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

Para el caso de muestreo sin remplazo de población finita, se tiene que $\sigma_S^2 = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N_p - n}{N_p - 1}}$, en donde N_p es el tamaño de la población, por lo que la variable estandarizada será

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N_p - n}{N_p - 1}}}.$$

En los dos casos anteriores, el valor de Z correspondiente al de $\bar{X}$ de la muestra es el que se debe comparar con el valor crítico correspondiente al nivel de significancia fijado, para así aceptar o no la hipótesis nula (prueba de una cola). Si se trata de una prueba de dos colas, el valor de Z se debe comparar con los dos valores críticos que corresponden al valor de α seleccionado. En cualquiera de los casos anteriores, el valor o valores críticos se pueden obtener de la tabla 14.1, para valores comunes de α.

Ejemplo 16.1

Se sabe que el promedio de calificaciones de una muestra aleatoria de tamaño 100 de los estudiantes de tercer año de ingeniería civil es de 7.6, con una desviación estándar de 0.2. Si μ denota la media de la población de esas calificaciones, X, y si se supone que $\bar{X}$ tiene distribución normal, probar la hipótesis

$\mu = 7.65$ en contra de la hipótesis alternativa $\mu \neq 7.65$, usando un nivel de significancia de

a. 0.05

b. 0.01

Para la solución se deben considerar las hipótesis

$$H_0 : \mu = 7.65$$

$$H_1 : \mu \neq 7.65$$

Puesto que $\mu \neq 7.65$ incluye valores menores y mayores de 7.65, se trata de una prueba de dos colas.

La estadística bajo consideración es el promedio aritmético, $\bar{X}$, de la muestra, que se supone extraída de una población infinita. La distribución muestral de $\bar{X}$ tiene media $\mu_{\bar{X}} = \mu$, y desviación estándar $\sigma_{\bar{X}} = \sigma/\sqrt{n}$, en donde μ y σ denotan, respectivamente, la media y la desviación estándar de la población de calificaciones.

Bajo la hipótesis H_0 (considerándola verdadera), se tiene que

$$\mu_{\bar{X}} = 7.65 = \mu$$

y utilizando la desviación estándar de la muestra como una estimación de σ, lo cual se supone razonable por tratarse de una muestra grande,

$$\sigma_{\bar{X}} = \sigma/\sqrt{n} = 0.2/\sqrt{100} = 0.2/10 = 0.02$$

a. Para la prueba de dos colas a un nivel de significancia de 0.05 se establece la siguiente regla de decisión

Aceptar H_0 si el valor Z correspondiente al valor del promedio de la muestra se encuentra dentro del intervalo de -1.96 a 1.96 (tabla 14.1). En caso contrario, rechazar H_0.

Puesto que

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{7.6 - 7.65}{0.02} = -2.5$$

se encuentra fuera del rango de -1.96 a 1.96, se rechaza la hipótesis H_0 a un nivel de significancia de 0.05.

b. Si el nivel de significancia es 0.01, el intervalo de -1.96 a 1.96 de la regla de decisión del inciso a se reemplaza por el de -2.58 a 2.58 tabla (14.1). Entonces, puesto que el valor muestral $Z = -2.5$ se encuentra dentro de este intervalo, se acepta la hipótesis H_0 a un nivel de significancia de 0.01.

Ejemplo 16.2

La resistencia media a la ruptura de cables de acero fabricados por la empresa X es de 905 kg. Una empresa consultora sugiere a X que cambie su proceso de manufactura, con lo cual incrementará la resistencia de sus cables. Se prueba el nuevo proceso, y se extrae una muestra aleatoria de 50 cables, obteniéndose para ellos una resistencia promedio de 926 kg, con des-

-viación estándar igual a 42 kg. ¿Se puede considerar que el nuevo proceso realmente incrementa la resistencia, con un nivel de confianza de 99%?

En este caso, se debe plantear una prueba de hipótesis de una cola, para la cual

$$H_0 : \mu = 905 \text{ kg}$$

$$H_1 : \mu > 905 \text{ kg}$$

Puesto que el tamaño de la muestra es suficientemente grande, se puede aproximar la distribución muestral de la resistencia promedio mediante una normal, y estimar el valor de σ de la población mediante S_X de la muestra.

Considerando a la población infinita, y suponiendo como verdadera a H_0, se tiene que

$$\mu_{\bar{X}} = \mu = 905 \text{ kg}$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{42}{\sqrt{50}} = 5.94$$

Para la prueba de una cola a un nivel de significancia de $\alpha = 1 - (1 - \alpha) = 1 - 0.99 = 0.01$, la regla de decisión es

Aceptar H_0 si el valor estandarizado de $\bar{X}$ de la muestra es menor o igual a $Z_c = 2.326$ [tabla 14.1]; en caso contrario, rechazar H_0.

En virtud de que

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{926 - 905}{5.94} = 3.535$$

es mayor de 2.326, se rechaza H_0 a un nivel de significancia de 1%, concluyéndose que en realidad el nuevo proceso sí incrementa la resistencia de los cables.

17. Pruebas de diferencias de medias.

Sean $\bar{X}$ y $\bar{Y}$ los promedios aritméticos obtenidos de dos muestras de tamaños n_X y n_Y, extraídas respectivamente de dos poblaciones con medias μ_X y μ_Y, y desviaciones estándar σ_X y σ_Y. Se trata de probar la hipótesis nula, H_0, de que no existe diferencia entre las medias, es decir, que $\mu_X = \mu_Y$. Si n_X y n_Y son suficientemente grandes (>30), la distribución muestral de las diferencias de los promedios es aproximadamente normal. Dicha distribución muestral es rigurosamente normal si las variables aleatorias X y Y asociadas a la población tienen distribución normal, aunque n_X y n_Y sean menores de 30. Para esta distribución muestral, la variable estandarizada Z, que se compara con los valores críticos correspondientes, se encuentra dada por

$$Z = \frac{X - Y - \mu_{\bar{X}-\bar{Y}}}{\sigma_{\bar{X}-\bar{Y}}} = \frac{X - Y - 0}{\sigma_{\bar{X}-\bar{Y}}} = \frac{\bar{X} - \bar{Y}}{\sigma_{\bar{X}-\bar{Y}}}$$

con la cual se puede probar la hipótesis nula H_0 en contra de otras hipótesis alternativas, H_1, a un nivel apropiado de significancia.

Ejemplo 17.1

En el laboratorio de pruebas de una empresa fabricante de aparatos electrónicos se ensayaron dos marcas de transistores, A y B, de características similares, con objeto de comprobar su ganancia de voltaje. Se tomaron muestras aleatorias de 100 transistores de cada marca, arrojando una ganancia promedio de 31 decibeles, con desviación estándar de 0.3 decibeles para la marca A, y 30.9 decibeles de ganancia promedio, con desviación estándar de 0.4 decibeles para la otra. ¿Existe una diferencia significativa entre las ganancias en voltaje de los transistores a un nivel de significancia de

a. 0.05

b. 0.01?

Si μ_A y μ_B son las medias respectivas de las dos poblaciones infinitas a las que corresponden las muestras, la prueba de hipótesis adopta la forma siguiente:

$$H_0 : \mu_A = \mu_B$$

$$H_1 : \mu_A \neq \mu_B$$

Entonces, el valor de Z es, bajo la hipótesis H_0:

$$Z = \frac{\bar{X}_A - \bar{X}_B}{\sigma_{\bar{X}_A - \bar{X}_B}} = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} = \frac{31 - 30.9}{\sqrt{\frac{(0.3)^2}{100} + \frac{(0.4)^2}{100}}} = 2$$

a. Puesto que se trata de una prueba de dos colas a un nivel de significancia de 0.05, la diferencia es significativa si el valor de Z se encuentra fuera del intervalo de -1.96 a 1.96. Como este es el caso, puede concluirse que efectivamente existe diferencia significativa en la ganancia en voltaje de los transistores.

b. Si la prueba es a un nivel de significancia de 0.01, la diferencia es significativa si Z se encuentra fuera del rango de -2.58 a 2.58. Partiendo del hecho de que Z = 2, la diferencia entre las ganancias es producto del azar, y se acepta la hipótesis de que ambos tipos de transistores tienen igual ganancia media en voltaje a un nivel de confianza de 99 por ciento.

Ejemplo 17.2

La estatura promedio de 50 estudiantes varones tomados al azar que participan en actividades deportivas es de 173 cm, con desviación estándar de 6.3 cm. Otra muestra aleatoria de 50 estudiantes varones que no participan en ese tipo de actividades tiene promedio de estatura igual a 171 cm, con desviación estándar igual a 7.1 cm. Probar la hipótesis de que los estudiantes varones que practican deportes son más altos que los que no lo hacen, a un nivel de significancia de 0.05.

Se debe decidir entre las hipótesis

$$H_0 : \mu_X = \mu_Y$$

$$H_1 \quad \mu_X > \mu_Y$$

siendo X la variable aleatoria asociada a la población infinita de estaturas de alumnos que practican deportes, y Y la asociada a la de estudiantes que no lo hacen, que también es infinita.

Bajo la hipótesis H_0 , se tiene que

$$\mu_{\bar{X}-\bar{Y}} = 0$$

$$\sigma_{\bar{X}-\bar{Y}} = \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} = \sqrt{\frac{(6.3)^2}{50} + \frac{(7.1)^2}{50}} = 1.3424$$

Entonces, el valor de Z es

$$Z = \frac{\bar{X} - \bar{Y}}{\sigma_{\bar{X}-\bar{Y}}} = \frac{173 - 171}{1.3424} = \frac{2}{1.3424} = 1.489$$

Puesto que se trata de una prueba de hipótesis de una cola, a un nivel $\alpha = 0.05$, se rechazaría H_0 si el valor de Z muestral fuera mayor del valor crítico para dicho nivel, el cual es $Z_c = 1.645$. Puesto que $Z < Z_c$, en este caso se concluye que la diferencia en las estaturas de ambos grupos de estudiantes se debe únicamente al azar.

tiene ordenadas mayores de cero en el lado de las abscisas negativas. De hecho, la estadística S_X^2 se puede estudiar si se consideran muestras aleatorias de tamaño n extraídas de una población normal con desviación estándar σ_X y si para cada muestra se calcula el valor de la estadística.

$$\chi^2 = \frac{n\, S_X^2}{\sigma^2} \tag{3.14}$$

donde S_X^2 es la variancia de la muestra.

El número de grados de libertad, ν, de una estadística se define como

$$\nu = n - k$$

siendo n el tamaño de la muestra y k el número de parámetros de la población que deben estimarse a partir de ella.

La distribución muestral de la estadística χ^2 está dada por la ecuación

$$f(\chi^2) = U\, \chi^{\nu-2}\, e^{-\frac{1}{2}\chi^2}$$

en la que U es una constante que hace que el área total bajo la curva resulte igual a uno, y $\nu = n - 1$ es el número de grados de libertad. Esta distribución se llama *Ji cuadrada*, misma que se presenta en la fig 21 para distintos valores de ν.

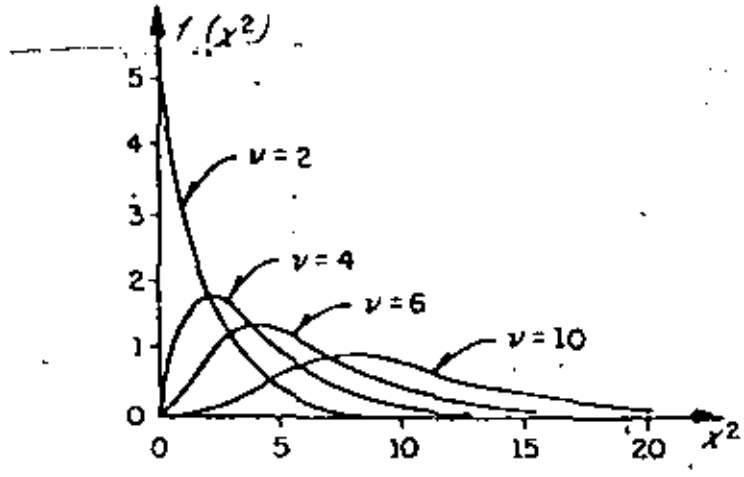


Fig 21. Distribución Ji cuadrada para distintos valores de ν

3.4 Muestras pequeñas

Como ya se indicó, para muestras grandes ($n \geq 30$) las distribuciones muestrales de muchas estadísticas son aproximadamente normales, siendo tanto mejor la aproximación cuanto mayor es el tamaño de n. Sin embargo, cuando se trata de muestras en las que $n < 30$, llamadas *muestras pequeñas*, la aproximación no es suficientemente buena, por lo que resulta necesario introducir una teoría apropiada para su estudio.

Al estudio de las distribuciones muestrales de las estadísticas para muestras pequeñas se le llama *teoría estadística de las muestras pequeñas*. Existen al respecto tres distribuciones importantes: *Ji cuadrada*, *F* y *t de Student*.

3.4.1 Distribución *Ji* cuadrada (χ^2)

Hasta ahora solo se ha tratado la distribución muestral de la media. En esta sección se verá lo concerniente a la distribución muestral de la variancia, S_X^2, para muestras aleatorias extraídas de poblaciones normales. Puesto que S_X no puede ser negativa, es de esperarse que su distribución muestral no sea una curva normal, ya que esta

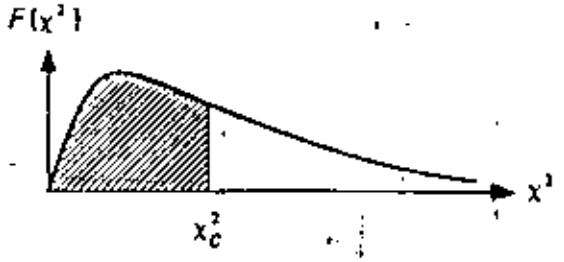


TABLA-B: VALORES CRITICOS χ^2_c

ν	$\chi^2_{.995}$	$\chi^2_{.99}$	$\chi^2_{.975}$	$\chi^2_{.95}$	$\chi^2_{.90}$	$\chi^2_{.75}$	$\chi^2_{.50}$	$\chi^2_{.25}$	$\chi^2_{.10}$	$\chi^2_{.05}$	$\chi^2_{.025}$	$\chi^2_{.01}$	$\chi^2_{.005}$
1	7.88	6.63	5.02	3.84	2.71	1.32	.455	.102	.016	.0039	.0010	.0002	.0000
2	10.6	9.21	7.38	5.99	4.61	2.77	1.39	.575	.211	.103	.0506	.0201	.0100
3	12.8	11.3	9.35	7.81	6.25	4.11	2.37	1.21	.584	.352	.216	.115	.072
4	14.9	13.3	11.1	9.49	7.76	5.39	3.36	1.92	1.06	.711	.483	.297	.207
5	16.7	15.2	12.8	11.15	9.2.	6.63	4.35	2.67	1.61	1.15	.831	.554	.413
6	18.5	16.8	14.4	12.6	10.6	7.84	5.35	3.45	2.20	1.64	1.24	.872	.676
7	20.3	18.5	16.0	14.1	12.0	9.04	6.35	4.25	2.83	2.18	1.69	1.24	.989
8	22.0	20.1	17.5	15.5	13.4	10.2	7.34	5.07	3.49	2.73	2.18	1.65	1.34
9	23.6	21.7	19.0	16.9	14.7	11.4	8.34	5.90	4.17	3.33	2.70	2.09	1.73
10	25.2	23.2	20.5	18.3	16.0	12.5	9.34	6.74	4.87	3.94	3.25	2.56	2.16
11	26.8	24.7	21.9	19.7	17.3	13.7	10.35	7.57	5.58	4.57	3.82	3.05	2.60
12	28.3	26.2	23.2	21.0	18.5	14.8	11.3	8.44	6.30	5.23	4.40	3.57	3.07
13	29.8	27.7	24.7	22.4	19.8	16.0	12.3	9.30	7.04	5.89	5.01	4.11	3.57
14	31.3	29.1	26.1	23.7	21.1	17.2	13.3	10.2	7.79	6.57	5.63	4.66	4.07
15	32.7	30.6	27.5	25.1	22.3	18.2	14.3	11.0	8.55	7.26	6.25	5.22	4.60
16	34.3	32.0	28.8	26.3	23.5	19.4	15.3	11.9	9.31	7.96	6.91	5.81	5.14
17	35.7	33.4	30.2	27.6	24.8	20.5	16.3	12.8	10.1	8.67	7.56	6.41	5.70
18	37.3	34.8	31.5	28.9	26.0	21.6	17.3	13.7	10.9	9.39	8.23	7.01	6.26
19	38.6	36.2	32.9	30.1	27.2	22.7	18.3	14.6	11.73	10.1	8.91	7.63	6.84
20	40.0	37.6	34.2	31.4	28.45	23.8	19.3	15.5	12.4	10.9	9.59	8.26	7.43
21	41.4	38.8	35.6	32.7	29.6	24.9	20.3	16.3	13.2	11.6	10.3	8.90	8.02
22	42.8	40.3	36.8	33.9	30.8	26.0	21.3	17.2	14.0	12.3	11.0	9.54	8.64
23	44.2	41.6	38.1	35.2	32.0	27.1	22.3	18.1	14.8	13.1	11.7	10.2	9.26
24	45.6	43.0	39.4	36.4	33.2	28.2	23.3	19.0	15.7	13.8	12.4	10.9	9.89
25	46.9	44.3	40.6	37.7	34.4	29.3	24.3	19.9	16.5	14.5	13.15	11.5	10.5
26	48.3	45.6	41.9	38.9	35.6	30.4	25.3	20.8	17.3	15.4	13.8	12.2	11.2
27	49.6	47.0	43.2	40.1	36.7	31.5	26.3	21.7	18.1	16.2	14.6	12.9	11.8
28	51.0	48.3	44.5	41.3	37.9	32.6	27.3	22.7	18.9	16.9	15.3	13.6	12.5
29	52.3	49.6	45.7	42.5	39.1	33.7	28.3	23.6	19.8	17.7	16.0	14.3	13.1
30	53.7	50.9	47.0	43.8	40.3	34.8	29.3	24.5	20.6	18.5	16.8	15.0	13.8
40	66.8	63.7	59.3	55.8	51.8	45.7	39.3	33.7	29.1	26.5	24.4	22.2	20.7
50	79.5	76.2	71.4	67.5	63.2	56.3	49.3	43.0	37.7	34.8	32.4	29.7	28.0
60	92.0	88.4	83.3	79.1	74.4	67.0	59.3	52.3	46.5	43.2	40.5	37.5	35.5
70	104.2	100.4	95.0	90.5	85.5	77.6	69.3	61.7	55.3	51.7	48.8	45.4	43.3
80	116.3	112.3	106.6	101.9	96.6	88.1	79.3	71.1	64.3	60.4	57.2	53.5	51.2
90	128.3	124.1	118.1	113.1	107.6	98.6	89.3	80.6	73.3	69.1	65.6	61.8	59.2
100	140.2	135.8	129.6	124.3	118.5	109.1	99.3	90.12	82.4	77.9	74.2	70.1	67.3

No obstante que la distribución ji cuadrada solo se ha presentado en el estudio de las muestras pequeñas, cabe aclarar que es válida para aquellas mayores de 30 si la variable aleatoria involucrada tiene distribución normal.

3.4.1.1 Intervalo de confianza para la varianza

Tal como se hizo para la distribución normal, se pueden establecer intervalos de confianza para la varianza de la población en términos de la varianza de una muestra extraída de ella, a un nivel de confianza dado $1 - \alpha$, si se hace uso de los valores críticos χ^2_c de la tabla 8. Por lo tanto, un intervalo de confianza para la estadística χ^2, estaría dado por

$$\chi^2_{c'} < \frac{n S_X^2}{\sigma^2} < \chi^2_c$$

donde $\chi^2_{c'}$ y χ^2_c son los valores críticos para los cuales el $(1 - \alpha)/2$ por ciento del área se encuentra en los extremos izquierdo y derecho de la distribución, respectivamente.

Con base en lo anterior, se concluye que

$$\frac{n S_X^2}{\chi^2_c} < \sigma^2 < \frac{n S_X^2}{\chi^2_{c'}}$$

es un intervalo de confianza para estimar a σ^2 a un nivel de confianza $1 - \alpha$.

3.4.1.2 Prueba de hipótesis para la varianza

La prueba de hipótesis para la varianza de una población normal se efectúa calculando el valor de la estadística χ^2 y estableciendo las hipótesis H_0 y H_1 apropiadas, es decir, se adoptan reglas de decisión similares a las usadas para la estadística Z.

Ejemplo

La variancia del tiempo de elaboración de cierto producto es igual a 40 min; sin embargo, su proceso de manufactura se modifica y se toma una muestra de

veinte tiempos, para la cual la variancia resulta ser igual a 62 min. ¿Es significativo el aumento del tiempo de elaboración a un nivel de significancia de

a) 0.05

b) 0.01?

Se debe decidir de entre las hipótesis

$$H_0 : \sigma^2 = 40 \text{ min}$$

$$H_1 : \sigma^2 > 40 \text{ min}$$

Suponiendo que la hipótesis nula es correcta, el valor de la estadística χ^2 para la muestra considerada es

$$\chi^2 = \frac{n S_X^2}{\sigma^2} = \frac{(20)\,(62)}{40} = 31$$

a) Como se trata de una prueba de una cola, la hipótesis H_0 se rechazaría si el valor de la estadística χ^2 fuera mayor que el de χ^2 para un nivel de significancia igual a 0.05, el cual, para $\nu = 20 - 1 = 19$ grados de libertad resulta ser 30.1 (tabla 8). Como $31 > 30.1$, H_0 se rechaza a un nivel de significancia de 0.05.

b) En este caso, el valor de χ^2 para un nivel de significancia de 0.01 y 19 grados de libertad es igual a 36.2. Puesto que $31 < 36.2$, se acepta H_0 a un nivel de significancia de 0.01.

3.4.2 Distribución F

Al efectuar la prueba de hipótesis de igualdad de medias para muestras pequeñas, en la siguiente sección se supondrá que las variancias de las poblaciones a las que corresponden tales muestras son iguales. Por lo tanto, es necesario probar antes si tal suposición es correcta. Para ello, debe considerarse que si S_X^2, n_X y S_Y^2, n_Y son respectivamente la variancia y el tamaño de dos muestras extraídas de poblaciones normales que tienen igual variancia, entonces

$$F = \frac{S_X^2}{S_Y^2} \tag{3.15}$$

TABLA 9. VALORES F_c PARA $\alpha = 0{,}01$

v_2 Grados de libertad del denominador	v_1 Grados de libertad del numerador																	
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4 052	5.000	5.403	5 625	5.764	5.859	5.928	5.982	6.023	6.056	6.106	6.157	6.209	6.235	6.261	6.287	6.313	6.339	6.366
2	98,50	99.00	99.20	99.20	99.30	99.30	99.40	99.40	99.40	99.40	99.40	99.40	99.40	99.50	99.50	99.50	99.50	99.50	99.50
3	34,10	30.80	29.50	28.70	28.20	27.90	27.70	27.50	27.30	27.20	27.10	26.90	26.70	26.60	26.50	26.40	26.30	26.20	26.10
4	21.20	18.00	16.70	16.00	15.50	15.50	15.00	14.80	14.70	14.50	14.40	14.20	14.00	13.90	13.80	13.70	13.70	13.60	13.50
5	16.30	13.30	12.10	11.40	11.00	10.70	10.50	10.30	10.20	10.10	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.13	9.02
6	13.70	10.90	9.77	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.87
7	12.20	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6 62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.30	8.65	7.59	7.01	6.63	6.37	6.17	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.83
9	10 60	8.02	6.99	6.42	6.06	5.81	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.00	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.66	7.22	6.22	5.68	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.93	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4 50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6 36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.43	4.20	4.03	3.89	3.78	3.69	3.55	3.40	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.50	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.53	2.45	2.36	2.27	2.17
30	7.56	5.39	4.51	4.02	3.70	3.46	3.30	3.17	3.07	2.98	2.84	2.71	2.54	2.47	2.39	2.29	2.20	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

resulta ser el valor de una variable aleatoria (estadística) que tiene distribución F, con parámetros $\nu_X = n_X - 1$ y $\nu_Y = n_Y - 1$. Esta distribución (fig 22) cuenta con dos parámetros, ν_X y ν_Y, que son los grados de libertad que corresponden a la variancia del numerador y del denominador de la ec 3.15, respectivamente. Cuando se hace referencia a una distribución F en particular, siempre se dan primero los grados de libertad para la variancia del numerador; es decir, $F(\nu_X, \nu_Y)$. En la tabla 9 se presentan los valores críticos F_c para distintos valores de ν_X y ν_Y y un nivel de significancia de 0.01. Cuando los grados de libertad ν_X o ν_Y no se encuentren en dicha tabla, el valor de F se puede obtener mediante interpolación lineal. Si se desea probar la hipótesis a otros niveles de significancia, es factible emplear las tablas de la distribución F (refs 9 y 11).

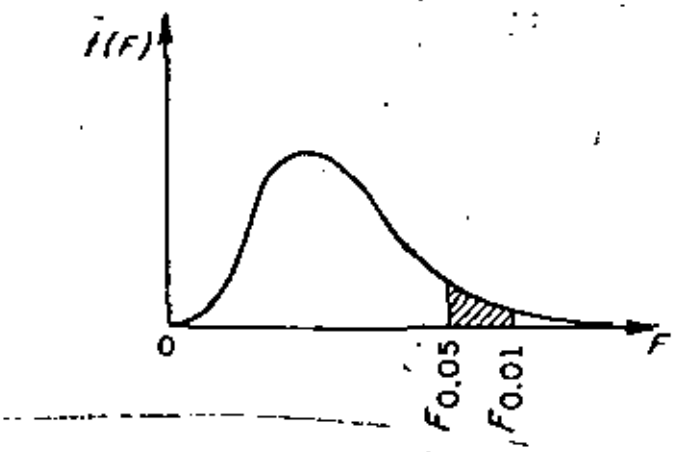


Fig 22. Distribución F.

De acuerdo con lo anterior, se puede probar la hipótesis nula

$$H_0: \sigma_X^2 = \sigma_Y^2$$

en contra de alguna hipótesis alternativa adecuada haciendo uso del hecho de que el cociente S_X^2/S_Y^2 es una estadística que tiene distribución F.

Ejemplo

Una empresa manufacturera de cartón prensado va a decidir acerca del empleo de una prensadora A o una B a fin de obtener un grosor determinado en su producto. El problema estriba en que ambas prensadoras proporcionan grosores muy similares, es decir, que la variancia de los grosores para las dos máquinas es la misma. Para decidir acertadamente, se tomó una muestra aleatoria de 31 cartones prensados por la máquina A y otra de 41 por la B. Como las variancias del grosor para los cartones de las muestras resul-

tan ser de 12 y de 5 micras, respectivamente, se establecen las hipótesis

$$H_0: \sigma_A^2 = \sigma_B^2$$

$$H_1: \sigma_A^2 > \sigma_B^2$$

con objeto de probarlas a un nivel de significancia de 0.01.

El valor de la estadística F resulta

$$F = \frac{S_A^2}{S_B^2} = \frac{12}{5} = 2.4$$

Puesto que $\nu_A = 31 - 1 = 30$ y $\nu_B = 41 - 1 = 40$, en la tabla 9 se puede ver que para un nivel de significancia de 0.01 el valor, F_c, de $F(30, 40)$ es 2.11. De acuerdo con estos valores, la hipótesis H_0 se rechazaría si el valor de F fuera mayor que F_c (30, 40).

Puesto que lo anterior resulta ser cierto, se rechaza H_0, concluyéndose que la prensadora B sería la mejor elección.

3.4.3 Distribución *t* de Student

Si se consideran muestras de tamaño n extraídas de una población normal con media μ y variancia desconocida, para cada muestra se puede calcular la estadística T definida mediante la fórmula

$$T = \frac{\bar{X} - \mu}{S_X} \sqrt{n - 1} \qquad (3.16)$$

donde $\bar{X}$ es el promedio y S_X la desviación estándar de la muestra.

La distribución muestral de T (fig 23) está dada por la ecuación

$$f(t) = \frac{U}{\left(1 + \frac{t^2}{\nu}\right)^{(\nu+1)/2}}$$

en la que U es una constante que hace que el área bajo la curva sea igual a uno, y $\nu = n - 1$ es el número de grados de libertad.

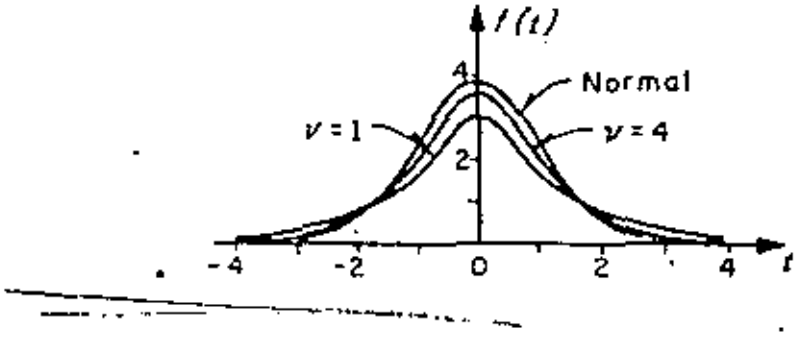


Fig 23. Distribución t de Student para distintos valores de ν

En la fig 23 se aprecia que conforme ν (o n, el tamaño de la muestra) aumenta, la distribución de $f(t)$ se aproxima a la distribución normal.

3.4.3.1 Límites e intervalos de confianza

De manera similar a como se hizo con la distribución normal, es posible estimar los límites de confianza de la media, μ, de una población mediante los *valores críticos*, t_c, de la distribución t, que dependen del tamaño de la muestra y del nivel de confianza deseado, encontrándose dichos valores en la tabla 10.

Así pues,

$$-t_c \leq \frac{\bar{X} - \mu}{S_X}\sqrt{n-1} \leq t_c$$

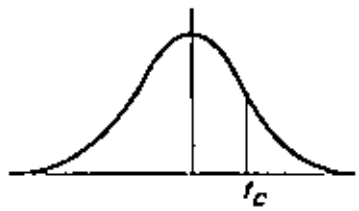
representa un intervalo de confianza para t, a partir del cual se puede estimar que μ se encuentra dentro del intervalo

$$\bar{X} - t_c \frac{\sigma_X}{\sqrt{n-1}} < \mu < \bar{X} + t_c \frac{\sigma_X}{\sqrt{n-1}}$$

En términos generales, los límites de confianza para la media de la población se representan como

$$\bar{X} \pm t_c \frac{\sigma_X}{\sqrt{n-1}}$$

TABLA 10. VALORES t_c PARA LA DISTRIBUCION t DE STUDENT



ν	$t_{.995}$	$t_{.99}$	$t_{.975}$	$t_{.95}$	$t_{.90}$	$t_{.80}$	$t_{.75}$	$t_{.70}$	$t_{.60}$	$t_{.55}$
1	63.66	31.82	12.71	6.31	3.07	1.376	1.000	.727	.325	.158
2	9.92	6.96	4.30	2.92	1.89	1.061	.816	.617	.289	.142
3	5.84	4.54	3.18	2.35	1.64	.978	.765	.584	.275	.138
4	4.60	3.75	2.78	2.13	1.53	.941	.741	.569	.271	.134
5	4.04	3.36	2.58	2.02	1.48	.920	.727	.560	.267	.132
6	3.71	3.14	2.45	1.94	1.44	.906	.718	.553	.265	.131
7	3.50	3.00	2.36	1.91	1.43	.896	.711	.549	.263	.130
8	3.36	2.90	2.31	1.86	1.40	.889	.706	.546	.262	.130
9	3.25	2.82	2.26	1.83	1.38	.883	.703	.543	.261	.129
10	3.17	2.76	2.23	1.81	1.37	.879	.700	.542	.260	.129
11	3.11	2.72	2.20	1.80	1.36	.876	.697	.540	.260	.129
12	3.06	2.68	2.18	1.78	1.36	.873	.695	.539	.259	.128
13	3.01	2.65	2.16	1.77	1.36	.871	.694	.538	.259	.128
14	2.98	2.62	2.14	1.76	1.34	.868	.693	.537	.258	.128
15	2.95	2.61	2.13	1.75	1.34	.866	.691	.536	.258	.128
16	2.92	2.58	2.12	1.75	1.34	.865	.690	.535	.258	.128
17	2.90	2.57	2.11	1.74	1.33	.863	.689	.534	.257	.128
18	2.88	2.55	2.10	1.73	1.33	.862	.688	.534	.257	.128
19	2.87	2.54	2.09	1.73	1.33	.861	.688	.533	.257	.127
20	2.84	2.53	2.09	1.72	1.32	.860	.687	.533	.257	.127
21	2.83	2.52	2.08	1.72	1.32	.859	.686	.532	.256	.127
22	2.82	2.51	2.07	1.72	1.32	.858	.686	.532	.256	.127
23	2.81	2.50	2.07	1.71	1.32	.858	.685	.532	.256	.127
24	2.80	2.49	2.06	1.71	1.32	.857	.685	.531	.256	.127
25	2.79	2.48	2.06	1.71	1.32	.856	.684	.531	.256	.127
26	2.78	2.48	2.05	1.71	1.32	.856	.684	.531	.256	.127
27	2.77	2.47	2.05	1.71	1.31	.855	.683	.531	.256	.127
28	2.76	2.47	2.05	1.70	1.31	.855	.683	.530	.256	.127
29	2.76	2.46	2.04	1.70	1.31	.854	.683	.530	.256	.127
30	2.75	2.46	2.04	1.70	1.30	.853	.683	.530	.256	.127
40	2.70	2.43	2.02	1.68	1.30	.851	.681	.529	.255	.126
60	2.66	2.39	2.00	1.67	1.30	.848	.679	.528	.254	.126
120	2.62	2.36	1.98	1.66	1.29	.845	.677	.526	.254	.126
∞	2.58	2.33	1.96	1.645	1.28	.842	.674	.524	.253	.126

3.4.3.2 Pruebas de hipótesis

La prueba de hipótesis para la media de una población se puede efectuar con muestras pequeñas en forma análoga a la de muestras de tamaño mayor de 30 si en lugar de utilizar a la estadística Z se emplea la T. Entonces, si se consideran dos muestras aleatorias cuyos tamaños, desviaciones estándar y promedios son n_X, S_X, $\overline{X}$ y n_Y, S_Y, $\overline{Y}$, respectivamente, extraídas de poblaciones normales de igual variancia ($\sigma_X^2 = \sigma_Y^2$), se puede probar la hipótesis, H_0, de que las muestras provienen de una misma población, es decir, de que también sus medias son iguales, utilizando la estadística T definida por

$$T = \frac{\overline{X} - \overline{Y}}{\epsilon \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \qquad (3.17)$$

donde

$$\epsilon = \sqrt{\frac{n_X S_X^2 + n_Y S_Y^2}{n_X + n_Y - 2}} \qquad (3.18)$$

cuya distribución es la t de Student, con $\nu = n_X + n_Y - 2$ grados de libertad.

Ejemplo

Conforme al plan de desarrollo agrícola de una región, se probó un nuevo fertilizante para maíz. Para ello se escogieron 24 ha de terreno, aplicándose dicho producto a la mitad de ellas. El promedio de producción de maíz en la zona que se usó fertilizante fue de 5.3 ton, con una desviación estándar de 0.40 ton, en tanto que en la otra zona el promedio fue de 5.0 ton, con desviación estándar de 0.36 ton.

De acuerdo con los resultados, ¿se puede concluir que existe un aumento significativo en la producción de maíz al usar fertilizante, si se utiliza un nivel de significancia de

a) 0.01

b) 0.05?

Solución

Para probar la hipótesis de igualdad de medias es indispensable saber primero si las muestras provienen de dos poblaciones normales de igual variancia. En ese caso, si σ_X^2 y σ_Y^2 denotan a las variancias de la producción de maíz en la zona tratada y en la no tratada, respectivamente, se debe probar la hipótesis nula H_0: $\sigma_X^2 = \sigma_Y^2$ en contra de la hipótesis alternativa H_1: $\sigma_X^2 > \sigma_Y^2$ a los dos niveles de significancia establecidos.

El valor de la estadística F es, de la ec 3.15,

$$F = \frac{S_X^2}{S_Y^2} = \frac{(0.40)^2}{(0.36)^2} = 1.27$$

y el valor crítico de F (11, 11), obtenido de la tabla 9 mediante interpolación lineal, resulta 4.47. Por lo tanto, como $1.27 < 4.47$, se acepta la hipótesis nula a un nivel de significancia de 0.01.

El valor crítico de F (11, 11) a un nivel de significancia de 0.05 (ref 9) es 2.82, de ahí que como $1.27 < 2.82$, también se acepta la hipótesis H_0.

Con base en lo anterior, se debe decidir entre las hipótesis

H_0: $\mu_X = \mu_Y$ (la diferencia en los promedios se debe al azar)

H_1: $\mu_X > \mu_Y$ (el fertilizante mejora la producción)

Bajo la hipótesis H_0, se tiene que

$$\epsilon = \sqrt{\frac{n_X S_X^2 + n_Y S_Y^2}{n_X + n_Y - 2}} = \sqrt{\frac{12(0.40)^2 + 12(0.36)^2}{12 + 12 - 2}} = 0.397$$

por lo cual

$$t = \frac{5.3 - 5.0}{0.397\sqrt{\frac{1}{12} + \frac{1}{12}}} = 1.85$$

a) Puesto que se trata de una prueba de una cola a un nivel de significancia de 0.01, se rechaza la hipótesis H_0 si t es mayor que el valor crítico, t_c, correspondiente a dicho nivel, el cual para $\nu = n_X + n_Y - 2 = 12 + 12 - 2 = 22$ grados de libertad, se obtiene de la tabla 8 como $t_c = 2.51$. Como $t < t_c$, la hipótesis H_0 no se puede rechazar a un nivel de significancia de 0.01.

b) Si el nivel de significancia de la prueba es de 0.05, se rechaza H_0 si t es mayor que el valor t_c respectivo que para 22 grados de libertad es $t_c = 1.72$, por lo que de acuerdo con lo anterior, H_0 se rechaza a un nivel de significancia de 0.05.

MUESTREO ALEATORIO SIMPLE PARA RAZONES

1. En una pequeña comunidad se realiza una investigación para determinar qué proporción del gasto familiar es dedicado a la alimentación y qué proporción es dedicado a la atención médica y medicamentos. Se selecciona una muestra aleatoria simple de 40 familias de un total de 2000 familias que forman la comunidad.

Estimar: a. Proporción del gasto familiar dedicado a la alimentación.

b. La varianza del inciso a.

c. El error estándar del inciso a.

d. Intervalos de confianza del 95% para la proporción del gasto familiar dedicado a la alimentación.

e. La proporción del gasto familiar dedicado en atención médica y medicamentos.

f. La varianza del inciso e.

g. Intervalos de confianza del 90% para la proporción del gasto familiar dedicado a la atención médica y medicamentos.

Los datos que se presentan corresponden a los gastos de un mes y están expresados en miles de pesos.

Familia	Gasto Familiar	Gasto Alimentación	Gasto en Médicos y Medicamentos
1	10	4	1
2	11	4.2	0.8
3	8	4	0.5
4	9	3.5	1
5	8.5	3	0.3
6	5	2.5	0
7	10	3.5	0.4
8	6	2	0.4
9	4	1.5	0
10	12	4.5	0.85
11	9	3	0.5
12	3.5	1.5	0.5
13	5	2	0
14	2	1.5	0
15	8	3	0.9

Familia	Gasto Familiar	Gasto Alimentación	Gasto en Médicos y Medicamentos
16	7	3.5	0.8
17	9.5	3	1.5
18	6	2.5	0.5
19	5	2	0.5
20	4.5	2.4	0.6
21	7.8	3	0
22	8	2.5	1.2
23	3	1.8	0
24	5.5	2	0.4
25	4.5	2.5	0
26	7	3.5	1
27	9	3	1.2
28	8.5	3.8	0.5
29	12	4.5	1.6
30	6.5	2.5	0.5
31	3.8	1.8	0
32	4	1.8	0.4
33	2.9	2	0
34	3.5	1.5	0.2
35	5	2	0.25
36	6.8	3	0.3
37	4	1.8	0.4
38	4.5	2	0.2
39	5.8	2.5	0.4
40	6.5	3	0.2
Totales	261.6	107.6	19.8

2. En una granja se está experimentando con una nueva alimentación para pollos. Se trabaja con una población de 600 pollos a los que se les pesa al iniciar el experimento, el peso total inicial resultó de 780 kilos . Después de un mes de desea conocer el peso medio por pollo y el peso total de los pollos , para lo cual se seleccionó una muestra aleatoria simple de 30 pollos que proporcionaron la siguiente información; considerando x el peso inicial del pollo , y el peso al mes del experimento

$\sum_{i=1}^{30} x_i = 37.5$, $\sum_{i=1}^{30} y_i = 116.7$, $\sum x_i^2 = 47.07$, $\sum y_i^2 = 254.09$, $\sum xy = 107.3$

Estimar:
a. Peso medio por pollo al mes de iniciado el experimento.
b. Peso total de los pollos.
c. Varianzas de los incisos a y b.
d. Intervalos de confianza del 95% para los incisos a y b.

3. Una institución bancaria cuenta con 1000 clientes con cuentas de cheques y para un estudio económico desea conocer qué proporción del ingreso mensual gastan sus clientes en promedio. Para estimar esta proporción deciden llevar a cabo un muestreo aleatorio simple de 30 clientes. De los estados de cuentas mensuales obtienen los gastos del cliente y de los datos confidenciales obtienen el ingreso mensual. Las observaciones obtenidas de la muestra fueron las siguientes:

Cliente	Gasto (en miles de pesos)	Ingreso(miles de pesos)
1	20	25
2	20	22
3	30	30
4	35	40
5	30	29
6	30	35
7	25	25
8	20	20
9	40	50
10	30	30
11	50	60
12	28	30
13	39	40
14	23	25
15	28	30
16	20	20
17	38	40
18	30	30
19	32	30
20	20	20
21	29	30
22	25	25
23	38	40
24	30	30
25	20	20
26	29	30
27	25	28
28	20	20
29	27	30
30	35	40
Totales	866	904

Estimar:

a. La proporción del ingreso que los clientes gastan mensualmente e intervalos del 95% de confianza.
b. La cantidad media del gasto por cliente y su varianza.
c. El total de gastos en la población y su error estándar.

Adela Abad Carrillo.

Junio de 1980

centro de educación continua
división de estudios de posgrado
facultad de ingeniería unam

FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO

TAMAÑO DE LA MUESTRA

M. EN I. AUGUSTO VILLARREAL ARANDA

JUNIO, 1980

4. TAMAÑO DE LA MUESTRA

Por: M en I Augusto Villarreal Aranda

INTRODUCCION

Dentro de un plan de muestreo, cuando ya se ha establecido la característica (o características) a estimar, así como el nivel de confianza y el grado de precisión requeridos, se debe decidir cuál debe ser el tamaño de la muestra o número de elementos a seleccionar por el procedimiento de muestreo que vaya a emplearse, en forma tal que los resultados que se obtengan no sean en exceso costosos o imprecisos.

Una vez que se ha fijado el error máximo admisible, que representa la precisión mínima que se exige tengan los resultados, así como el nivel de confianza $P_K = 1 - \alpha$, se requiere conocer además, en la forma más precisa posible, la variabilidad de la población,

ya que cuanto más dispersos estén los valores de la variable asociada a ella más arriesgado será el utilizar una muestra de tamaño pequeño.

A continuación se expondrá el procedimiento para seleccionar el tamaño de muestra más adecuado en el caso del muestreo aleatorio simple o irrestrictamente aleatorio (sin remplazo). Más adelante se estudiarán los métodos para cálcular el tamaño de la muestra para otros procedimientos de muestreo.

4.1 Tamaño de una muestra aleatoria simple (Medias)

En este caso se trata de estimar la media μ de una población con variable aleatoria asociada X mediante el empleo del promedio aritmético $\overline{X}$, obtenido de una muestra aleatoria de tamaño n con un error máximo admisible absoluto e y un nivel de confianza P_K. Es natural que a la probabilidad P_K le corresponderá un cierto valor de desviación K, obtenido a partir de la desigualdad de Chebyshev, o bien considerando a K como el número de desviaciones estándar para una distribución normal o para una t de Student.

El procedimiento para obtener el tamaño de la muestra se fundamenta en el hecho de que

$$P\left\{ \overline{X} - K\sigma_{\overline{X}} \leq \mu \leq \overline{X} + K\sigma_{\overline{X}} \right\} = P_K = 1 - \alpha$$

o sea que con probabilidad o nivel de confianza P_K se puede asegurar que el valor de μ de una población se encuentra dentro del

$(1-\alpha)$ % de los intervalos formados a partir de muestras de tamaño n, de la forma siguiente

$$(\bar{X} - K\sigma_{\bar{X}} \ , \ \bar{X} + K\sigma_{\bar{X}})$$

Lo anterior implica que los límites de confianza del P_K % para estimar a μ son

$$\bar{X} \pm K\sigma_{\bar{X}}$$

es decir, que el error en la estimación del valor de μ es, en valor absoluto,

$$|\text{error en la estimación de } \mu| = K\sigma_{\bar{X}} \qquad (4.1)$$

Por lo tanto, es posible escribir

$$|\text{error máximo admisible}| = |\text{error en la estimación de } \mu| = e$$

4.1.1 Muestreo de una población finita

De la inferencia estadística, el valor de $\sigma_{\bar{X}}$, <u>la desviación estándar de la distribución muestral de $\bar{X}$ (o error estándar de $\bar{X}$)</u> cuando la población es finita es

$$\sigma_{\bar{X}} = \sqrt{\frac{N_p - n}{N_p - 1} \frac{\sigma_X^2}{n}}$$

pudiéndose escribir entonces

$$e = K\sigma_{\bar{X}} = K \sqrt{\frac{N_p - n}{N_p - 1} \frac{\sigma_X^2}{n}}$$

siendo K la desviación correspondiente al nivel de confianza P_k, N_p el tamaño de la población, σ_x^2 la variancia de esta última y n el tamaño de la muestra.

Puesto que se desea conocer el tamaño de la muestra, éste se puede obtener despejando de la ecuación anterior el valor de n. Para ello, se requiere elevar al cuadrado ambos miembros, es decir

$$e^2 = K^2 \frac{N_p - n}{N_p - 1} \frac{\sigma_x^2}{n}$$

$$e^2 = \frac{K^2 \sigma_x^2 N_p - K^2 \sigma_x^2 n}{(N_p - 1) n}$$

despejando a n:

$$ne^2 (N_p - 1) = K^2\sigma_x^2N_p - K^2\sigma_x^2n$$

$$ne^2N_p - ne^2 = K^2\sigma_x^2N_p - K^2\sigma_x^2n$$

$$ne^2N_p - ne^2 + K^2\sigma_x^2n = k^2\sigma_x^2N_p$$

$$n(e^2N_p - e^2 + K^2\sigma_x^2) = k^2\sigma_x^2N_p$$

$$\therefore n = \frac{K^2\sigma_x^2N_p}{e^2N_p - e^2 + K^2\sigma_x^2} \qquad (4.2)$$

La fórmula anterior permite obtener el tamaño de la muestra considerando conocidos K, e, N_p y σ_X^2 . Puesto que el valor de σ_X^2 de la población usualmente se desconoce, se debe estimar previamente en forma adecuada considerando la información disponible de poblaciones semejantes a la que deberá muestrearse, o tomando una muestra preliminar suficientemente grande de dicha población.

Puesto que el tamaño de la muestra debe corresponder a un número entero positivo, se deberá asignar a n el valor entero más próximo por exceso al obtenido mediante la fórmula 4.2.

4.1.2 Muestreo de una población infinita

Cuando el muestreo se realiza a partir de una población infinita, el valor de $\sigma_{\bar{X}}$, la desviación estándar de la distribución muestral de $\bar{X}$, es

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$$

en donde σ_X es la desviación estándar de la población y n el tamaño de la muestra.

considerando la ecuación 4.1, se puede escribir en este caso

$$|\text{error en la estimación de } \mu| = e = K\sigma_{\bar{X}} = K\frac{\sigma_X}{\sqrt{n}}$$

Para obtener el valor de n, se elevan al cuadrado ambos miembros de la expresión anterior, es decir,

$$e^2 = \frac{K^2\ \sigma_X^2}{n}$$

Por lo cual

$$n = \frac{K^2 \sigma_X^2}{e^2}$$

Para resaltar el hecho de que en este caso el tamaño de la muestra se obtiene a partir de una población infinita, en lugar de emplear n se puede emplear n_∞, es decir

$$n_\infty = \frac{K^2 \sigma_X^2}{e^2} \qquad (4.3)$$

Al igual que en el caso de una población finita, el tamaño de la muestra dado por la ec 4.3 debe corresponder a un número natural, por lo cual se debe aproximar por exceso al valor entero más cercano.

4.1.3 Comparación entre n y n_∞

Si se divide entre $N_p e^2$ el numerador y el denominador del miembro izquierdo de la ecuación 4.2, se obtiene

$$n = \frac{\frac{K^2\sigma_X^2 N_p}{N_p e^2}}{\frac{e^2 N_p - e^2 + K^2\sigma_X^2}{N_p e^2}} = \frac{\frac{K^2\sigma_X^2}{e^2}}{1 - \frac{1}{N_p} + \frac{K^2\sigma_X^2}{N_p e^2}}$$

$$n = \frac{\frac{K^2\sigma_X^2}{e^2}}{1 + \frac{1}{N_p}\left(\frac{K^2\sigma_X^2}{e^2} - 1\right)}$$

y, considerando el valor de n_∞ dado por la ec 4.3, se obtiene finalmente

$$n = \frac{n_\infty}{1 + \frac{1}{N_p}(n_\infty - 1)} \qquad (4.4)$$

Como se puede apreciar de la ec 4.4, el valor de n es menor que el de n_∞, a menos que $N_p = \infty$.

4.1.4 Empleo adecuado de n y n_∞

Para una población finita, se definirá la <u>fracción de muestreo</u> como

$$\text{fracción de muestreo} = fm = \frac{n_\infty}{N_p}$$

siendo n_∞ el tamaño de la muestra calculada con la ec 4.3, y N_p el tamaño de la población.

Al obtener el tamaño de la muestra cuando se trata de una población finita, usualmente se acostumbra emplear la fórmula 4.3, que proporciona dicho tamaño para población infinita, y considerar como bueno dicho valor siempre que se cumpla la condición

$$fm \leq 0.05$$

Lo anterior quiere decir que en la práctica se calcula el valor de n_∞, y si n_∞ / N_p cumple con la condición mencionada, entonces se considera que n_∞ es una aproximación satisfactoria de n. Si la

condición no se cumple, entonces se emplea la ec 4.4 para obtener el valor de n.

Es claro que tomando como tamaño de la muestra a n_∞ siempre se estará del lado más prudente, en el sentido de que se toma una muestra igual o mayor que la necesaria. Sin embargo, la eficiencia del diseño exige que el gasto y el tiempo de muestreo no sean superiores a los que haya que efectuar.

Ejemplo 4.1

Sea una población normal finita con variancia aproximadamente igual a 500. Se desea obtener una muestra aleatoria para estimar mediante $\bar{X}$ a la media poblacional μ_X, con error en la estimación no mayor de 10 y nivel de confianza igual a 90%. Obténgase el valor de n considerando que el tamaño de la población es igual a

a. 1000

b. 100

Solución

a. Puesto que $\sigma_X^2 = 500$, $e = 10$ y $1 - \alpha = 0.90$, tratándose de una población normal se tiene que

$$K = Z_{0.45} = 1.645$$

por lo cual

$$n_\infty = \frac{K^2\sigma_X^2}{e^2} = \frac{(1.645)^2\ (500)}{10^2}$$

$$= (2.706)\ (5) = 13.53$$

$$\therefore n_\infty = 14$$

En virtud de que en este caso

$$fm = \frac{n_\infty}{N_p} = \frac{14}{1000} = 0.014 < 0.05$$

se considera que n = 14.

b. En este caso

$$fm = \frac{14}{100} = 0.14 > 0.05$$

por lo cual se emplea la ec 4.4 para obtener el valor de n, es decir,

$$n = \frac{n_\infty}{1 + \frac{1}{N_p}\ (n_\infty - 1)} = \frac{14}{1 + \frac{1}{100}\ (14 - 1)}$$

$$= \frac{14}{1 + \frac{13}{100}} = \frac{14}{1.13} = 12.389$$

$$\therefore n = 13$$

Ejemplo 4.2

Cierta universidad cuenta con 4726 estudiantes, y se desea conocer el rendimiento académico medio de todos ellos, en términos de una escala de calificación que va de cero a cien puntos. En estudios semejantes en otras universidades, se obtuvo que la desviación estándar de las calificaciones es aproximadamente igual a 7 puntos. Si el error en la estimación de la media de calificaciones no debe ser mayor de un punto en valor absoluto, y el nivel de confianza es igual a 99%, ¿cuál debe ser el tamaño de la muestra para realizar la estimación?

Solución

En este caso, aproximando la distribución muestral de $\bar{X}$ mediante la distribución normal, se debe considerar que

$$P_K = 1 - \alpha = 0.99 \quad \therefore \quad K = Z_{0.495} = 2.58$$

$$\sigma_X^2 = (7)^2 = 49 \quad ; \quad e = 1 \text{ punto}$$

Por lo tanto,

$$n_\infty = \frac{Z_C^2 \, \sigma_X^2}{e^2} = \frac{(2.58)^2 \ (49)}{(1)^2}$$

$$= \frac{(6.656) \ (49)}{1} = 326.144$$

O sea $n_\infty = 327$

Puesto que

$$fm = \frac{n_\infty}{N_p} = \frac{327}{4726} = 0.0692 > 0.05$$

se procede a calcular n, es decir,

$$n = \frac{n_\infty}{1 + \frac{1}{N_p}(n_\infty - 1)} = \frac{327}{1 + \frac{1}{4726}(327 - 1)}$$

$$= \frac{327}{1 + \frac{326}{4726}} = \frac{327}{1.069} = 305.89$$

$$\therefore n = 306$$

Ejemplo 4.3

Una muestra aleatoria de 14 observaciones de la altura alcanzada por cierto tipo de planta arrojó los siguientes datos:

N° de elemento	Altura, X, en pulgadas
1	52.3
2	48.1
3	55.7
4	56.8
5	50.1
6	49.2
7	47.7
8	50.8
9	57.9
10	52.5
11	54.7
12	49.6
13	53.9
14	56.0

Obténgase el tamaño de muestra necesario para asegurar, con una probabilidad igual a 0.95, que el error en la estimación de la media de alturas de esta variedad de planta no sea mayor del 2.86%.

Solución

Se deben obtener primero los valores de $\bar{X}$ y S_X^2 de la muestra, con los cuales se estimarán los de μ_X y σ_X^2 de la población. Para ello, se dispone la información en la forma siguiente :

	X_i	X_i^2
	52.3	2735.3
	48.1	2313.6
	55.7	3102.5
	56.8	3226.2
	50.1	2510.0
	49.2	2420.6
	47.7	2275.3
	50.8	2580.6
	57.9	3352.4
	52.5	2756.2
	54.7	2992.1
	49.6	2460.2
	53.9	2905.2
	56.0	3136.0
Σ	735.3	38766.2

Por lo tanto,

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{n} X_i = \frac{1}{14}(735.3) = 52.52 \text{ pulgadas}$$

$$S_X^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{14}(38766.2) - (52.52)^2$$

$$= 2769.01 - 2758.35 = 10,66 \text{ pulgadas}$$

Puesto que el error en la estimación de la media no debe ser mayor del 2.86%, y el estimador de u_X es $\bar{X} = 52.52$, se tiene que

$$e = 52.52\ (0.0286) = 1.5 \text{ pulgadas}$$

Por otra parte, se desconoce el valor real de σ_X^2 de la población, además de que S_X^2, su estimador, se ha obtenido de una muestra menor de 30 elementos. Por lo tanto, la distribución teórica a la cual se debe aproximar la muestral debe ser la t de Student, siendo en este caso $K = t_c$. Sin embargo, puesto que en este caso se estima σ_X^2 mediante S_X^2 de la muestra, se debe tener presente que el error en la estimación de μ_X es

$$e = K\,\sigma_{\bar{X}} = t_c\,\sigma_{\bar{X}} = t_c\,\frac{S_X}{\sqrt{n-1}}$$

O sea, elevando al cuadrado

$$e^2 = t_c^2\,\frac{S_X^2}{n-1}$$

y, despejando a n,

$$n - 1 = \frac{t_c^2\ S_X^2}{e^2}$$

$$n = \frac{t_c^2\ S_X^2}{e^2} + 1$$

Por ser muestreo de población infinita, se puede escribir finalmente

$$n_\infty = \frac{t_c^2 \, S_X^2}{e^2} + 1 \qquad (4.5)$$

Ya que el valor de t_c depende del número de grados de libertad de la muestra υ, y este último depende del tamaño de la muestra (ya que $\upsilon = n - 1$), la fórmula anterior para obtener el valor de n_∞ contiene dos incógnitas. Por ello, se sigue el siguiente proceso iterativo para obtener el valor de n_∞ :

1. Se hace $t_{0.025} = Z_{0.475}$, es decir

$$t_{0.025} = 1.96$$

Con dicho valor de t_c se obtiene

$$n_\infty = \frac{(1.96)^2 \ (10.66)}{(1.5)^2} + 1 = 18.2 + 1 = 19.3 \Rightarrow 20$$

De la tabla de la distribución t, se obtiene $t_{0.025} = 2.09$, para $\upsilon = 20-1 = 19$ grados de libertad.

2. Se toma ahora $t_{0.025} = 2.09$, y se obtiene

$$n_\infty = \frac{(2.09)^2 \ (10.66)}{(1.5)^2} + 1 = 20.7 + 1 = 21.7 \Rightarrow 22$$

De la tabla de la distribución t, se obtiene $t_{0,025} = 2.08$, para $\upsilon = 22-1 = 21$ grados de libertad.

3. Se toma ahora $t_{0,025} = 2.08$, y se obtiene

$$n_\infty = \frac{(2.08)^2\ (10.66)}{(1.5)^2} + 1 = 20.5 + 1 = 21.5 \Rightarrow 22$$

En este paso se obtiene un valor de n_∞ igual al del paso anterior, por lo que se puede considerar que el tamaño de muestra adecuado es igual a 22 plantas.

En este caso la población es infinita, por lo cual no se requiere hacer la corrección para población finita con la ec 4.4. Sin embargo, debe aclararse que es posible emplear la ec 4.5 para obtener n_∞ primero y, si la población de la que se muestrea es finita, usar después la ec 4.4 para obtener el valor de n corregido.

4.2 Tamaño de una muestra aleatoria simple (Totales)

Una característica o parámetro poblacional de gran interés es el <u>total</u>, que corresponde a la suma de todos los valores y_i que constituyen la población, es decir,

$$Y = \sum_{i=1}^{N_p} y_i$$

en donde Y denota al total, y N_p es el número de elementos de la misma.

Si se multiplica y divide por N_p el 2° miembro de la ecuación ante

rior, se obtiene

$$Y = \frac{N_p}{N_p} \sum_{i=1}^{N_p} Y_i = N_p \mu_Y$$

Es decir, el total de una población es igual al tamaño de la misma multiplicado por la media correspondiente.

Como estimador puntual del total de la población se puede tomar el de la estadística

$$\hat{Y} = N_p \bar{Y}$$

en donde $\bar{Y}$ es el promedio aritmético de la muestra, y $\hat{Y}$ un estimador insesgado en virtud de que

$$E\{\hat{Y}\} = E\{N_p \bar{Y}\} = N_p E\{\bar{Y}\} = N_p \mu_Y = Y$$

Por otra parte, la varianza de la distribución muestral de $\hat{Y}$ es

$$\sigma^2_{\hat{Y}} = \sigma^2_{N_p \bar{Y}} = \text{Var}\{N_p \bar{Y}\} = N_p^2 \text{Var}\{\bar{Y}\} = N_p^2 \sigma^2_{\bar{Y}}$$

y la desviación estándar es

$$\sigma_{\hat{Y}} = \sigma_{N_p \bar{Y}} = N_p \sigma_{\bar{Y}} = N_p \frac{\sigma_Y}{\sqrt{n}} \sqrt{\frac{N_p - n}{N_p - 1}}$$

De igual manera a como se hizo para las medias, el valor del tamaño de muestra para estimar el total con un nivel de confianza y un error absoluto dados, se obtiene en la forma siguiente

$$e = K\,\sigma_{\hat{Y}} = K\,N_p\,\frac{\sigma_Y}{\sqrt{n}}\sqrt{\frac{N_p - n}{N_p - 1}}$$

Elevando al cuadrado y realizando operaciones algebraicas,

$$e^2 = K^2\,N_p^2\,\frac{\sigma_Y^2}{n}\,\frac{N_p - n}{N_p - 1}$$

$$e^2 = \frac{K^2\,N_p^3\,\sigma_Y^2 - K^2\,N_p^2\,\sigma_Y^2\,n}{n(N_p - 1)}$$

$$n\left(1 + \frac{K^2\,N_p^2\,\sigma_Y^2}{e^2(N_p - 1)}\right) = \frac{K^2\,N_p^3\,\sigma_Y^2}{e^2(N_p - 1)}$$

O sea

$$n = \frac{K^2\,N_p^3\,\sigma_Y^2}{e^2(N_p - 1) + K^2\,N_p^2\,\sigma_Y^2}$$

Dividiendo el numerador y denominador de la expresión anterior entre $N_p e^2$, se obtiene

$$n = \frac{\dfrac{K^2 N_p^3\,\sigma_Y^2}{N_p e^2}}{\dfrac{e^2 N_p - e^2 + K^2 N_p^2\,\sigma_Y^2}{N_p e^2}}$$

$$= \frac{N_p^2\,\dfrac{K^2\sigma_Y^2}{e^2}}{1 - \dfrac{1}{N_p} + \dfrac{N_p^2}{N_p}\dfrac{K^2\,\sigma_Y^2}{e^2}}$$

Considerando la ec 4.3, queda finalmente

$$n = \frac{N_p^2\,n_\infty}{1 + \frac{1}{N}\,(N_p^2\,n_\infty - 1)} \qquad (4.6)$$

Ejemplo 4.4

Con el fin de hacer una solicitud al Gobierno, se recogieron firmas de habitantes de una ciudad en 676 hojas. Cada hoja tenía espacio suficiente para 42 firmas, pero en varias hojas se recolectó un número menor de ellas. Para obtener una estimación del total de firmas, se contó el número de firmas por hoja en una muestra aleatoria de 50 hojas, obteniéndose los datos que aparecen en la tabla siguiente:

Número de firmas, y_i	Número de hojas, f_i
42	23
41	4
36	1
32	1
29	1
27	2
23	1
19	1
16	2
15	2
14	1
11	1
10	1
9	1
7	1
6	3
5	2
4	1
3	1

Obtener el tamaño de muestra necesario para estimar el valor del total de firmas con un error absoluto igual al 5%, considerando un nivel de confianza igual a 95%.

Solución : Por conveniencia para realizar los cálculos, se dispone la información en la forma siguiente:

Y_i	f_i	y_i^2	$f_i Y_i$	$f_i Y_i^2$
42	23	1764	966	40572
41	4	1681	164	6724
36	1	1296	36	1296
32	1	1024	32	1024
29	1	841	29	841
27	2	729	54	1458
23	1	529	23	529
19	1	361	19	361
16	2	256	32	512
15	2	225	30	450
14	1	196	14	196
11	1	121	11	121
10	1	100	10	100
9	1	81	9	81
7	1	49	7	49
6	3	36	18	108
5	2	25	10	50
4	1	16	4	16
3	1	9	3	9
Σ	50		1471	54497

$$\bar{Y} = \frac{1}{50} \sum_{i=1}^{19} f_i\, y_i = \frac{1471}{50} = 29.42$$

$$S_Y^2 = \frac{1}{50} \sum_{i=1}^{19} f_i\, y_i^2 - (\bar{Y})^2 = \frac{54497}{50} - (29.42)^2 =$$

$$= 1089.94 - 865.44 = 224.5$$

Entonces

$$\hat{Y} = N_p\, \bar{Y} = 676 \times 29.42 = 19888 \text{ firmas}$$

y, puesto que el error absoluto debe ser igual al 5%, se tendría

$$e = (0.05)\ (19888) = 995$$

Por otra parte, el tamaño inicial de muestra igual a 50 permite suponer que la estimación de σ_Y^2 de la población es suficientemente buena con S_Y^2, y que la distribución muestral de totales puede aproximarse mediante la normal. Por lo tanto,

$$K = Z_{0.475} = 1.96$$

$$N_p = 676$$

$$\sigma_Y^2 \doteq S_Y^2 = 224,5$$

$$N_p^2 n_\infty = N_p^2 \frac{K^2 \sigma_Y^2}{e^2} = \frac{(676)^2\ (1,96)^2\ (224,5)}{(995)^2} = 397,9$$

$$n = \frac{N_p^2 n_\infty}{1 + \frac{1}{N_p}(N_p^2 n_\infty - 1)} = \frac{397.9}{1 + \frac{1}{676}(397.9 - 1)}$$

$$= \frac{397.9}{1 + 0.58} = \frac{397.9}{1.58} = 251.83$$

$$\therefore \quad n = 252 \text{ hojas}$$

4.3 Tamaño de una muestra aleatoria simple (Proporciones)

4.3.1 Antecedentes

Supóngase una población binomial de tamaño N_p tal que cada uno de sus elementos únicamente puede estar en una de dos clases: A c B (buenos o malos, negros o blancos, grandes o chicos, etc). La proporción de elementos de la población que están en la clase A es

$$P = \frac{A}{N_p}$$

y la proporción de elemtos que están en B es

$$Q = \frac{B}{N_p}$$

por lo cual

$$P + Q = \frac{A}{N_p} + \frac{B}{N_p} = 1 \quad ; \quad (A + B = N_p)$$

Si a todos los elementos X_i de la población que están en A se les asigna el valor 1 y a los de B el 0, se obtiene

$$P = \frac{A}{N_p} = \frac{\sum_{i=1}^{N_p} X_i}{N_p} = \mu_X$$

Es decir, la proporción puede considerarse un caso particular de la media cuando los elementos de la población son unos y ceros.

La variancia es

$$\sigma_X^2 = \frac{1}{N_p} \sum_{i=1}^{N_p} (X_i - P)^2$$

o sea

$$\sigma_X^2 = \frac{1}{N_p} \sum_{i=1}^{N_p} X_i^2 - P^2$$

Sin embargo, como X_i sólo puede ser igual a uno o cero, se tiene que $X_i = X_i^2$, por lo cual

$$\sigma_X^2 = \frac{1}{N_p} \sum_{i=1}^{N_p} X_i - P^2 = P - P^2 = P(1 - P) = PQ$$

En virtud de lo anterior, si se muestrea sin remplazo y con tamaño n de una población binomial finita, para estimar la proporción de elementos con cierta característica, se obtienen, considerando que la proporción se puede calcular como una media, los siguientes parámetros de la distribución muestral de proporciones

$$\mu_p = P$$

$$\sigma_p = \frac{\sigma_X}{\sqrt{n}} \sqrt{\frac{N_p - n}{N_p - 1}} = \sqrt{\frac{PQ}{n}} \cdot \sqrt{\frac{N_p - n}{N_p - 1}}$$

Si la población es infinita, se obtiene

$$\mu_p = P$$

$$\sigma_p = \frac{\sigma_x}{\sqrt{n}} = \sqrt{\frac{PQ}{n}}$$

estimándose $\underline{P}$ en ambos casos con el valor de $\underline{p}$ de la muestra, si se desconoce P de la población.

En la práctica se considera que la distribución muestral de proporciones es aproximadamente igual a la normal para tamaños de muestra mayores o iguales a 30 elementos.

4.3.2 Obtención del tamaño de la muestra

Aprovechando el hecho de que la proporción se puede calcular como una media simple, las ecs 4.3 y 4.4 se pueden emplear en este caso para obtener el tamaño de la muestra haciendo $\sigma_x^2 = PQ$. Entonces,

$$n_\infty = \frac{K^2\, PQ}{e^2} \qquad (4.7)$$

para muestreo de población infinita, y

$$n = \frac{n_\infty}{1 + \frac{1}{N_p}(n_\infty - 1)}$$

para muestreo de población finita con tamaño N_p.

Usualmente se calcula primero el valor de n_∞, y si la fracción de muestreo es mayor de 0.05, se calcula a continuación el valor de n.

Ejemplo 4.5

En una colonia con 4000 casas se desea estimar el porcentaje de inquilinos que son a la vez propietarios de su casa, con un error estándar en la estimación no mayor del 1%. Se supone, de estudios semejantes, que el porcentaje real de inquilinos-propietarios se acerca al 10%. ¿Cuántas casas se deben muestrear para que se satisfaga la condición establecida?

Solución

El error estándar en la estimación de P de la población es

$$\sigma_p = \sqrt{\frac{PQ}{n}} \sqrt{\frac{N_p - n}{N_p - 1}}$$

y no debe ser mayor en este caso del 1%. Por lo tanto, siendo $N_p = 4000$, $P = 0.1$ y $Q = 1 - P = 0.9$, se obtiene

$$0.01 = \sqrt{\frac{(0.1)\ (0.9)}{n}} \sqrt{\frac{4000 - n}{4000 - 1}}$$

Elevando al cuadrado y realizando operaciones algebraicas

$$0.0001 = \frac{0.09}{n} \frac{4000 - n}{3999}$$

$$0,0001 = \frac{360 - 0.09\ n}{3999\ n}$$

$$0.3999\ n = 360 - 0.09\ n$$

$$n(0.3999 + 0.09) = 360$$

$$n = \frac{360}{0.4899} = 734.84$$

$$\therefore\ n = 735 \text{ casas}$$

Ejemplo 4.6

En un estudio antropológico para estimar el porcentaje de habitantes de una isla con sangre del grupo O, se obtuvo una muestra aleatoria de 50 isleños, en la cual 22 de ellos pertenecen al grupo sanguíneo mencionado. Si en la isla habitan 3208 gentes, ¿cuál debe ser el tamaño de muestra mínimo para estimar con un error absoluto del 5% el valor real de P, suponiendo que el nivel de confianza es del 95%?

Solución

En este caso la proporción de la muestra es

$$p = \frac{22}{50} = 0.44$$

$$= \frac{339}{1.105} = 306.787$$

$$\therefore\ n = 307 \text{ habitantes}$$

o sea

$$q = 1 - p = 1 - 0.44 = 0.56$$

Considerando que la muestra inicial es suficientemente grande, se aproxima mediante la distribución normal, obteniéndose

$$K = Z_{0.475} = 1.96$$

por lo cual

$$n_\infty = \frac{K^2\,PQ}{e^2} = \frac{K^2\,pq}{e^2} = \frac{(1.96)^2\ (0.44)\ (0.50)}{(0.05)^2}$$

$$= \frac{0.84515}{0.0025} = 338.06$$

$$\therefore\ n_\infty = 339$$

Como

$$fm = \frac{n_\infty}{N_p} = \frac{339}{3208} = 0.106 > 0.05$$

se corrige el valor anterior, obteniéndose finalmente

$$n = \frac{n_\infty}{1 + \frac{1}{N_p}\ (n_\infty - 1)} = \frac{339}{1 + \frac{1}{3208}\ (339-1)}$$

$$= \frac{339}{1.105} = 306,787$$

$$\therefore\ n = 307 \text{ habitantes}$$

MEXICO, D.F. MAYO 29 DE 1980.

MUESTREO ESTRATIFICADO,
MUESTREO POR CONGLOMERADOS Y
SUBMUESTREO. -

EJERCICIOS ADICIONALES.

NOTA.- ESTOS EJERCICIOS FORMAN PARTE DE LA SEGUNDA EDICION DEL LIBRO: INTRODUCCION AL MUESTREO. ABAD Y SERVIN. LIMUSA, LA CUAL APARECERA PROXIMAMENTE.

Ejercicio a.

Se desea hacer un estudio sobre el personal que labora en una fábrica la cual cuenta con edificios en quince estados del país. El estudio se refiere a opiniones y actitudes de los empleados y obreros. En la muestra se desea tener representados a 1 de cada 30 empleados y existen en total 42 090 de ellos. Administrativamente, el personal de cada estado es independiente de la oficina central en cuanto a su nómina, de tal manera que, las listas de obreros y empleados se tienen para cada uno de ellos.

La distribución del personal en cada entidad aparece en la tabla No. 1.

TABLA NO. 1

ENTIDAD	NUMERO DE EMPLEADOS	NUMERO DE HOJAS
Guanajuato	19 043	635
Hidalgo	429	15
Jalisco	5 010	167
Michoacán	1 114	38
Morelos	721	25
Nayarit	474	16
Nuevo León	4 415	148
Oaxaca	450	15

ENTIDAD	NUMERO DE EMPLEADOS	NUMERO DE HOJAS
Puebla	2 750	92
Querétaro	487	17
Quintana Roo	150	5
S.Luis Potosí	925	31
Sinaloa	2 800	94
Sonora	2 900	97
Tabasco	422	15
	42 090	1 410

Para obtener a uno de cada treinta empleados en la muestra se requiere una muestra total de n=1 403 (¿Por qué?) los cuales serán sorteados a partir de algún esquema de selección apropiado.

Ejercicio b.(Continuación del ejercicio a)
Entonces, debemos elejir a 1 403 empleados de entre los 42 090 esparcidos en las quince entidades federativas. Disponemos de 15 listados de empleados,uno de cada entidad federativa. Si deseamos obtener una selección aleatoria simple, pues, habría necesidad de unir a esos listados de tal manera de asegurar una identificación única para cada empleado, y posteriormente efectuar la selección. Esta tarea resulta engorrosa (Intente el método). Si esto fuera necesario, posiblemente fuera mas adecuado recurrir a una selección sistemática (Capítulo 7) con intervalo igual a 30, de tal manera que seleccionaríamos a un número aleatorio entre 1 y 30, el cual tomaríamos como arranque (Supongamos que fue el 15) y, las instrucciones para la selección serían las siguientes:

i) Ordene los listados, digamos alfabéticamente.

ii) Encuentre al renglón número 15 del primer listado, este empleado se encuentra en la muestra.

iii) A partir del empleado número 15, vuelva a contar del 1 al 30. El empleado número 30 está en la muestra.

iv) Recomience la cuenta del 1 al 30 hasta agotar a todas las listas.

La selección anterior puede ser superada mediante una estratificación en la cual cada estrato es definido como un estado. Tendríamos 15 estratos, tales que, usando afijación proporcional, sus tamaños de muestra serían:

$n_1=(19\ 043/42\ 090)1\ 403\doteq 635$

$n_2=(429/42\ 090)1\ 403\doteq 14$

$n_3=(5\ 010/42\ 090)1\ 403=167$

$n_4=(1\ 114/42\ 090)1\ 403\doteq 37$

$n_5=(721/42\ 090)1\ 403\doteq 24$

$n_6=(474/42\ 090)1\ 403=16$

$n_7=(4\ 415/42\ 090)1\ 403\doteq 147$

$n_8=(450/42\ 090)1\ 403=15$

$n_9=(2\ 750/42\ 090)1\ 403=92$

$n_{10}=(487/42\ 090)1\ 403\doteq 16$

$n_{11}=(150/42\ 090)1\ 403=5$

$n_{12}=(925/42\ 090)1\ 403=31$

$n_{13}=(2\ 800/42\ 090)1\ 403\doteq 93$

$n_{14}=(2\ 900/42\ 090)1\ 403\doteq 97$

$n_{15}=(422/42\ 090)1\ 403=14$

Ejercicio c. (Continuación del ejercicio b)

Para efectuar la selección anterior (Ejercicio b) en cada uno de los estratos y en el caso concreto del primero de ellos, podemos obtener una muestra aleatoria simple de tamaño 635 de entre los 19 043 empleados en Guanajuato. Esto equivale a obtener 635 números aleatorios diferentes entre 1 y 19 043 (si estuvieran numerados del 1 al 19 043). Para continuar con Hidalgo, habría que elejir a 14 números aleatorios diferentes entre 1 y 429, y así sucesivamente.
Como en el ejercicio b, la selección aleatoria anterior pudo haberse efectuado mediante una selección sistemática con fracción de muestreo 1 de cada 30 (¿Realmente es la misma fracción de muestreo, 1/30, para cada estrato?, ¿por qué?).

Ejercicio d. (Continuación del ejercicio c).

Supongamos que las listas del personal en cada estado aparecen mecanografiadas en hojas tamaño carta y que el número de hojas por delegación es el registrado en la tabla No. 1. Ahí se ha supuesto que en promedio cada hoja tiene 30 nombres, debido a ello, y pensando en un esquema de muestreo por conglomerados, se requieren 47 hojas en la muestra (¿Por qué?). Lo mas posible, es que en estas condiciones, el tamaño de muestra resultante no coincida con el deseado, pero se parecerá a él. Esto ocurre en la práctica y de hecho constituye una manera de diseño; es decir, se diseña para terminar con un tamaño de muestra esperado igual a algún valor en particular y realmente, al final se terminará con una cantidad mayor o menor que la deseada. Esto ocurre con frecuencia cuando uno fija la fracción de muestreo

general para obtener estimadores autoponderados o como también se dice, diseños con igual probabilidad, aunque también existen métodos para controlar esta variación. (Ver por ejemplo el capítulo 7 del libro de Kish).

Las 47 hojas en la muestra pueden ser seleccionadas con f=47/1 410 ó 1 de cada 30 y esto puede ser hecho, ya sea con muestreo aleatorio simple o mediante una selección sistemática como en el ejercicio b.

Ejercicio e. (Continuación del ejercicio d).

Como continuación del ejercicio anterior y avanzando en la complejidad del diseño, consideramos ahora un submuestreo en el cual, la unidad primaria es la hoja con los nombres y la unidad secundaria es el nombre en sí. Nuestro objetivo es terminar con n=1 403 nombres. Aunque pudiéramos proponer un esquema tal que procurara terminar con exactamente 1 403, nuestro objetivo actual es proponer esquemas autoponderados en los cuales pudiera fluctuar el tamaño de muestra final, pero que mantendría las probabilidades de selección constantes y por lo tanto la sencillez en los estimadores. En el ejercicio anterior d, al seleccionar 47 hojas en la muestra podemos decir que las probabilidades de selección fueron las siguientes:

$$f_1 = 47/1\ 410$$
$$f_2 = 1/1$$

Es decir, la fracción de muestreo general fue $f = f_1 \cdot f_2 = 47/1\ 410(1/1) = 47/1\ 410 = 1/30$ y requerimos un censo en las primarias seleccionadas. Si f=94/1 410 y, $f_2 = 1/2$ entonces, $f=(94/1\ 410)(1/2)=1/30$
Aquí, estamos seleccionando 94 hojas en la muestra y

dentro de cada una de ellas entramos con fracción de muestreo 1 de cada 2. El número de primarias en la muestra se ha duplicado. Otra opción es usar 188 primarias en la muestra y dentro de cada una de ellas seleccionar a 1 de cada 4 nombres, es decir, f=(188/1 410)(1/4)=1/30. A medida que aumentamos el número de primarias en la muestra, reducimos el número de nombres en la muestra dentro de cada hoja. Si las listas de nombres siguen al orden de adscripción del personal a cada uno de los departamentos u oficinas en la empresa, lo que estamos haciendo al aumentar el número de primarias en la muestra, es aumentar el número de departamentos u oficinas en la muestra, aumentando, por así decirlo, la representatividad de la muestra, y el precio que se está pagando por ello, es el tener que recorrer mas edificios o ciudades(digamos) buscando a las distintas oficinas seleccionadas.

Ing. Luis Servín.

centro de educación continua
división de estudios de posgrado
facultad de ingeniería unam

FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO

PRACTICA DE MUESTREO

M. EN C. ADELA ABAD DE SERVIN

JUNIO, 1980

PRACTICA DE MUESTREO.

ESTIMADORES DE RAZON

*

1. De una lista de 468 academias de 2 años de estudio fue sacada una muestra aleatoria simple de 100. La muestra contenía 54 instalaciones públicas y 46 privadas. Los datos para el número de estudiantes (y) y el número de profesores (x) se muestran a continuación.

	n	$\sum(y)$	$\sum(x)$
Pública	54	31 281	2 024
Privada	46	13 707	1 075

	$\sum(y^2)$	$\sum(yx)$	$\sum(x^2)$
Pública	29 881 219	1 729 349	111 090
Privada	6 366 785	431 041	33 119

   a. Para cada tipo de instalación en la población, estimar la proporción (número de estudiantes/número de profesores)
   b. Calcular los errores estándar de los estimadores.
   c. Para las instalaciones públicas encontrar los límites de confianza del 95% para la proporción de estudiante/maestro en la población.

2. En un estudio realizado en una zona formada por 70 manzanas, se listaron las 3000 familias que la componían y se eligieron aleatoriamente 30 familias. A cada familia se le preguntó el número de miembros y el número de autos que tenían. Se obtuvieron los siguientes resultados, donde la y indica número de miembros y la x número de coches.

$$\sum y_i = 236 \qquad \sum x_i = 115 \qquad \sum y_i x_i = 685$$

$$\sum y_i^2 = 1494 \qquad \sum x_i^2 = 401$$

   a. Estimar el número de miembros por auto en la población.
   b. Encontrar el error estándar de su estimador.
   c. Calcular intervalos de confianza del 95% para el número de miembros por auto en la población.

3. En una zona formada por 2450 manzanas se desea estimar el número total de niños en edad pre-escolar en 1977. De acuerdo con el Censo de Población de 1970, este total fue de 25,000 niños. Se selecciona una muestra aleatoria simple de 10 manzanas y se obtienen los siguieates datos;

i	1	2	3	4	5	6	7	8	9	10
y_i	15	10	15	14	16	10	44	10	5	5
x_i	20	15	12	13	10	5	6	4	0	5

   Estimar el número total de niños en esta zona en 1977, su varianza e intervalos de confianza del 95%.

4. Una compañía desea estimar el monto promedio en dinero pagado a sus empleados en gastos médicos, durante los 3 primeros meses del presente año. Los promedios trimestrales están disponibles en los reportes fiscales del año anterior. De la población de 1000 empleados fue seleccionada una muestra aleatoria de 100 registros de empleados. Los resultados de la muestra son totalizados y se encuentra que el total para el presente trimestre es de $\sum_{i=1}^{100} y_i = 1750$ y el total correspondiente al trimestre del año anterior es de $\sum_{i=1}^{100} x_i = 1200$. Los gastos médicos de la población total, correspondientes al mismo trimestre en el año anterior fue de 12500. Estime el monto promedio en dinero pagado por la compañía en el primer trimestre del presente año y su error estándar.
Se proporcionan los siguientes datos:
$\sum_{i=1}^{100} y_i^2 = 30650$, $\sum_{i=1}^{100} x_i^2 = 15620$, $\sum_{i=1}^{100} y_i x_i = 21859.35$

*

5. En un campo de cebada se pesaron el grano, y_i, y el grano más la paja, x_i, en cada una de un gran número de unidades de muestreo localizadas al azar por todo el campo.

También se pesó la cosecha total (grano más paja) del campo completo. Se obtuvieron los siguientes datos:
$c_{yy} = 1.13$ $c_{yx} = 0.78$ $c_{xx} = 1.11$

Calcule la ganancia en precisión obtenida estimando el rendimiento en grano del campo, de la razón grano a cosecha total en lugar de usar el rendimiento medio de grano por unidad.

6. Se desea estimar el número total de niños en edad pre escolar en el Estado de México en 1979. De acuerdo a información obtenido del censo de Población de 1970, se sabe que este total fue de 25,000 niños. Se selecciona una muestra aleatoria simple de 10 manzanas y se investiga el número de niños en edad pre escolar en el año de 1979 por manzana. El mismo dato se obtiene del censo para 1970.

MANZANA	1	2	3	4	5	6	7	8	9	10
1979	15	10	15	14	16	10	4	10	5	5
1970	20	15	12	13	10	5	6	4	0	5

a. Estimar el número total de niños en edad pre escolar en 1979 en el Estado de México.

b. Calcular intervalos del 95% de confianza.

La Primera Cirujana del Ejército de E. U. Privada de su Presea

★ Fue Sufragista y Usaba Traje Masculino.
★ Una Vieja Historia de la Guerra Civil
★ Iniciativa Para Restituirle la Medalla

WASHINGTON, 24 de enero. (NYT)—Hace sesenta años la doctora Mary Edwards Walker, cirujana con nombramiento durante la Guerra Civil de Estados Unidos, fue privada de su Medalla de Honor por una junta revisora del gobierno. La doctora Walker es la única mujer que jamas haya recibido la presea, lo cual también puede ser la razón para que la perdiera.

No es la primera vez que se envia una petición al Comité de los Servicios Armados del Senado para que devuelva póstumamente a la doctora Walker la medalla que es la más alta distinción del pais por valor en combate.

El mes pasado la Junta del Ejército para la Corrección de los Expedientes Militares, junta revisora que actúa en nombre del secretario del Ejército, Clifford L. Alexander Jr., celebró una audiencia para examinar el caso de la doctora Walker. El grupo envió su recomendación al secretario Alexander, si bien los portavoces del ejército se niegan a decir qué se recomendó ni cuál será la decisión del secretario.

La medalla fue conferida a la doctora Walker por el Presidente Andrew Johnson el 11 de noviembre de 1865. Los generales William T. Sherman y George H. Tomas habian recomendado la presea y el Presidente Lincoln habia firmado el testimonio poco antes de su muerte.

La doctora Walker fue citada por su papel como la primera cirujana del Ejército de Estados Unidos. La cita original se ha perdido y no se sabe que existan copias.

En la Junta de Acción Adversa de la Medalla de Honor revocó la medalla, argumentando que había encontrado ambiguedades en la situación de la doctora Walker como miembro del ejército. El caso fue expuesto a Brooke por la señora Anne Walker, de Mt. Vernon, Virginia, quien se dice "sobrina lejana" de la doctora Walker y cuya campaña para la restitución de la medalla le lleva casi todo su tiempo.

"La doctora Mary perdió la medalla —dijo recientemente la señora Walker— solo porque se habia adelantado cien años a su época y eso nadie lo podia aceptar".

La señora Walker quizás tenga razón. La doctora Walker fue sufragista toda su vida y partidaria de la reforma del vestido de la mujer. Desde la época de la Guerra Civil usó pantalones de hombre y sacos de levita. Daba conferencias feministas a la viuda con traje de gala masculino, con la Medalla de Honor pendiente de las anchas solapas.

Durante el decenio de... 1870, trabajó en la sede de las sufragistas en Washington, al lado de Susan B. Anthony, Lucy Stone, Mary Livermore y Belva Lockwood. Las mujeres se convirtieron en blanco favorito de los impugnadores. "Esa curiosa antropoide", llamó un reportero de The New York Times a la doctora

Números aleatorios					
04433	80674	24520	18222	10610	05794
60298	47829	72648	37414	75755	04717
67884	59651	67533	68123	17730	95862
89512	32155	51906	61662	64130	16688
32653	01895	12506	88535	36553	23757

1. Estimar el número medio de letras por renglón en el artículo de periódico que se adjunta, utilizando muestreo estratificado (2 estratos) y m.a.s. dentro de cada estrato. Una muestra de 10 renglones debe ser repartida con afijación proporcional.

   Calcular intervalos del 95% de confianza para la media en consideración.

   De manera muy breve vaya narrando en cada paso el procedimiento utilizado.

2. Utilizando la misma muestra del ejercicio 1, estimar:

   a. La proporción de vocales en el artículo de periódico que se adjunta e intervalos de confianza del 95%.

   b. El número total de vocales en el artículo y su error estándar.

2. Una compañía desea estimar el número total de horas hombres perdidas, para un mes dado, a causa de accidentes entre sus empleados. Pero obreros, técnicos y administradores tienen distintas tasas de accidentes, por lo que se decide usar muestreo aleatorio estratificado considerando cada grupo un estrato separado. Información de años anteriores proporciona las varianzas para número de horas hombres perdidas por empleado en los 3 grupos y además se proporciona el tamaño de los estratos.

I(obreros)	II(técnicos)	III(administradores)
$s_1^2 = 36$	$s_2^2 = 25$	$s_3^2 = 9$
$N_1 = 132$	$N_2 = 92$	$N_3 = 27$

a. Repartir en los estratos una muestra de 30 empleados con afijación proporcional

b. Estimar el número total de horas hombres perdidas durante el mes dado y establezca intervalos de confianza del 95% para este total. Utilice la siguiente información obtenida de una muestra de 18 obreros, 10 técnicos y 2 administradores.

I(Obreros)			II(Técnicos)		III(Administradores)
8	24	0	4	5	1
0	16	32	0	24	8
6	0	16	8	12	
7	4	4	3	2	
9	5	8	1	8	
18	2	0			

3.

3. *Un muestrista tiene una población de N=5 elementos y lo máximo que puede tomar es una muestra de tamaño 2.*
A continuación se indican los valores de los característicos para esas cinco unidades:

Unidad i	y_i
1	*0*
2	*1*
3	*0*
4	*1*
5	*1*

El muestrista tiene completa libertad para utilizar muestreo aleatorio simple, muestreo estratificado, etc.

a. *Qué plan le dará una varianza mínima para el el estimador de Ȳ?*
b. *Qué formato toma la varianza de $\bar{y}_{est}$ en el caso de afijación proporcional si se supone que la variabilidad por unidad en cada estrato es constante?*
c. *Proponga un estimador insesgado de la variabilidad por unidad en el caso b.*
d. *Proponga las fórmulas para la estimación de una proporción y su varianza en el caso de muestreo estratificado con muestreo aleatorio simple en cada estrato.*

4. En una estratificación con dos estratos, los valores de W_h y S_h son como sigue:

Estrato	W_h	S_h
1	0.8	2
2	0.2	4

Calcule los tamaños de muestra n_1 y n_2 en los dos estratos necesarios para satisfacer las siguientes condiciones. Cada caso requiere un cálculo separado(ignore el cpf).

a. El error estándar del estimador de la media de la población $\bar{y}_{est}$ debe ser 0.1 y el tamaño de la muestra total $n = n_1 + n_2$ debe ser minimizado.

b. El error estándar de la media estimada de cada estrato debe ser de 0.1.

5. Cuatro recipientes, que contienen un número igual (y muy grande) de repuestos representan cuatro turnos de producción de una fábrica. El número de unidades correspondientes a muestras tomadas al azar de cada recipiente y el número de defectuosos encontrados se incluyen a continuación:

Recipiente	h	1	2	3	4
	n_h	200	200	200	200
	a_h	4	2	10	8

a. Compute una estimación insesgada de la proporción de defectuosos en la población total (4 recipientes).

b. Compute una estimación de la varianza de su estimación en (a).

6. Dada una población de N = 500 recipiebtes con 200 unidades cada uno y una muestra de n = 5 recipientes., se obtiene la siguiente información con p_i, proporción de defectuosos en el recipiente i de la muestra:

i	1	2	3	4	5
p_i	$\frac{2}{200}$	$\frac{3}{200}$	$\frac{8}{200}$	$\frac{1}{200}$	$\frac{2}{200}$

a. Estime el porcentaje de defectuosos para ese período de producción y la varianza de su estimación.

b. Si en la situación anterior se toma una muestra al azar de tamaño 2 en cada uno de los 500 recipientes y a_i indicase el número de defectuosos encontrados en cada recipiente (i = 1,..,500), proponga Ud. un estimador para el porcentaje de defectuosos y una expresión para la varianza de seu estimador.

c. ¿Cuál de los dos esquemas estimaría más adecuado?

1- En una ciudad pequeña, para efecto de estimación del ingreso medio por varón adulto, se define cada manzana como un conglomerado. Al numerar las manzanas se encuentra que en total hay 415. De ellas se extrae una m.a.s. de 25 manzanas. Los resultados obtenidos se presentan en el cuadro siguiente.
   a. Estime el ingreso medio por varón adulto en la ciudad y su e.e.
   b. Estime el ingreso total de los varones adultos en la ciudad y su e.e.
   c. Sabiendo que hay 2500 varones adultos en la ciudad, estime el ingreso total de los varones adultos en la ciudad y su e.e.

2. En adición a la información sobre ingreso de varones adultos, se les pregunta si viven en casa propia o rentada. Los resultados se presentan en el mismo cuadro. Estime la proporción de varones adultos en la ciudad que viven en casa rentada y establezca su e.e.

Conglomerado i	Nº de varones adultos m_i	ingreso total x congl. y_i	No. de varones adultos en casas rentadas (a_i)
1	8	96,000	4
2	12	121,000	7
3	4	42,000	1
4	5	65,000	3
5	6	52,000	3
6	6	40,000	4
7	7	75,000	4
8	5	65,000	2
9	8	45,000	3
10	3	50,000	2
11	2	85,000	1
12	6	43,000	3
13	5	54,000	2
14	10	49,000	5
15	9	53,000	4
16	3	50,000	1
17	6	32,000	4
18	5	22,000	2
19	5	45,000	3
20	4	37,000	1
221	6	51,000	3
22	8	30,000	3
23	7	39,000	4
24	3	47,000	0
25	8	41,000	3
	$\sum m_i = 151$	$\sum y_i = 1,329,000$	

3. El gerente de una editorial periodística desea estimar el número promedio de periódicos comprados por casa en una comunidad dada. Los costos de transportes de casa a casa son considerables, por lo que las 4000 casas en la comunidad se listan en 400 conglomerados geográficos de 10 casas cada uno, y se selecciona una m.a.s. de 4 conglomerados. Los resultados de la entrevistas son:

Conglom.	No. de periódicos										Total
1	1	2	1	3	3	2	1	4	1	1	19
2	1	3	2	2	3	1	4	1	1	2	20
3	2	1	1	1	1	3	2	1	3	1	16
4	1	1	3	2	1	5	1	2	3	1	20

Estime el número medio de periódico por casa en la comunidad y

4. En una ciudad pequeña se desea estimar el número medio de alumnos, por salón de clase, en las escuelas primarias. No se dispone de un listado de salones, pero se cuenta con un listado de 500 escuelas. La investigación se realiza considerando cada escuela un conglomerado y se selecciona una muestra de 25 escuelas y se obtienen los siguientes resultados:

ESCUELA (i)	No. DE SALONES POR ESCUELA (m_i)	No. TOTAL DE ALUMNOS por escuela (y_i).
1	12	520
2	6	310
•	•	•
•	•	•
25	8	380
	275	10500

Además $\sum y_i^2 = 2,100,000$ $\sum m_i^2 = 3050$ $\sum y_i m_i = 115,500$

Estimar:

a. El número medio de alumnos por salón de clase.
b. La varianza del número medio de alumnos por salón de clase.
c. Intervalos de confianza del 95% para el inciso a.
d. Si se sabe que el número total de salones es de 6000 en toda la población, estimar el número total de alumnos en la ciudad.
e. La varianza del inciso d.
f. El número medio de alumnos por escuela y su desviación estándar.
g. Si no se conoce el número total de salones en la comunidad, estimar el número total de alumnos en la ciudad y su error estándar.

* Extraído del libro " Técnicas de Muestreo", de William Cochran, Ed. CECSA.

centro de educación continua
división de estudios de posgrado
facultad de ingeniería unam

FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO

C O N G L O M E R A D O S
(MICAS)

M. EN C. ADELA ABAD DE SERVIN

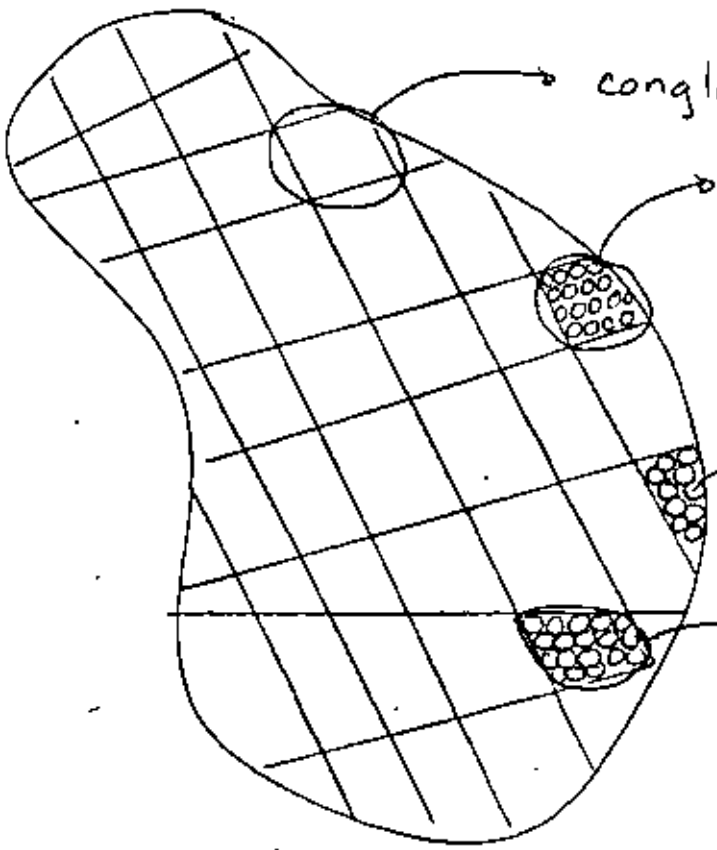
JUNIO, 1980

Muestreo por Conglomerados

Conceptos
- elemento
- unidad muestral
- conglomerado
- marco de referencia
- ventajas

Notación

Población : N conglomerados



conglomerado

M_i nº de elementos del conglomerado i-ésimo

$M = \sum_{i=1}^{N} M_i$ total de elem. en la pobl.

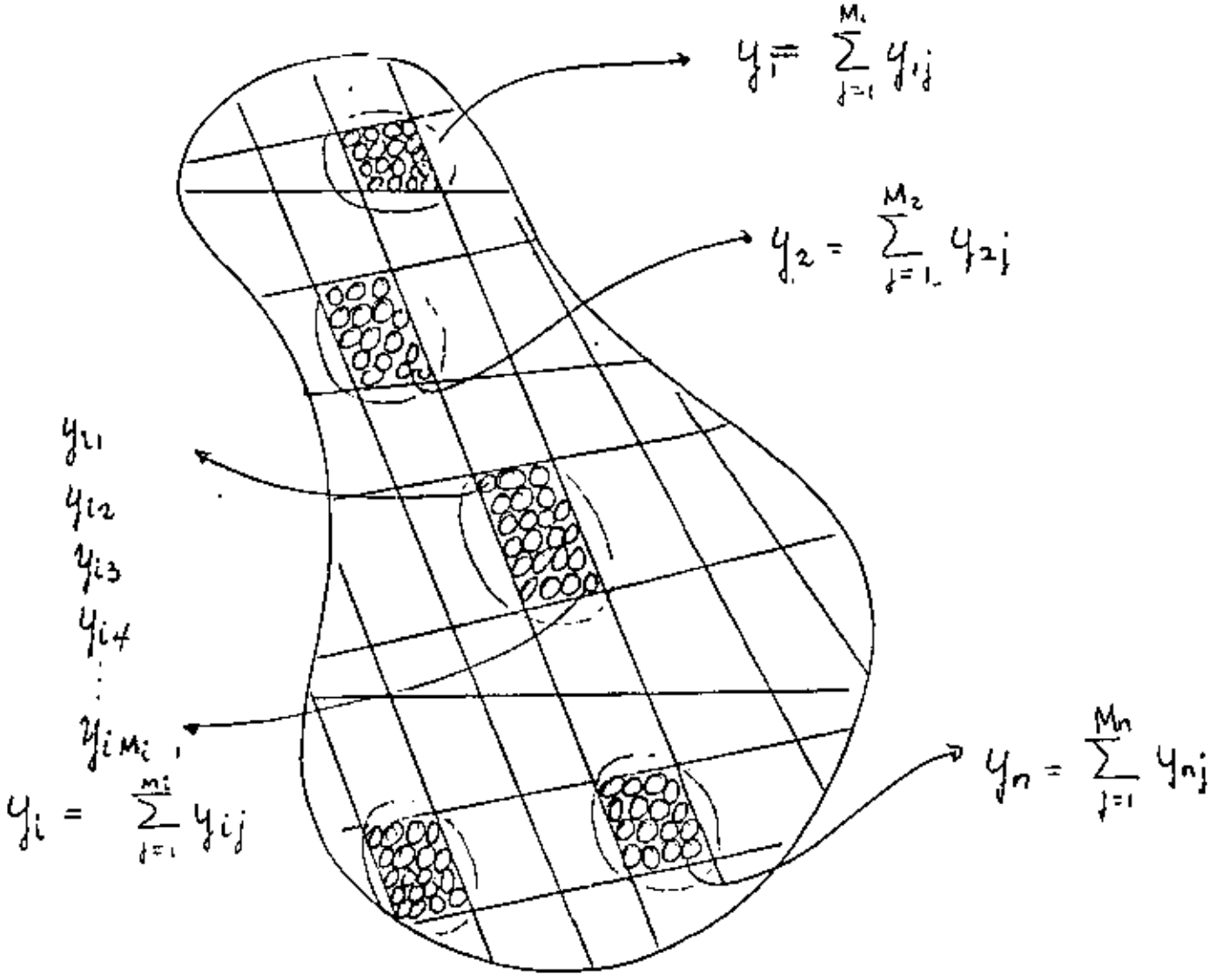
y_{ij} característica del elem. j-ésimo del conglomerado i-ésimo.

$y_i = \sum_{j=1}^{M_i} y_{ij}$ característica del congl. i-ésimo

$\bar{M} = \frac{M}{N}$ tamaño medio de los conglomerados en la población.

Método de Selección en Conglomerados.

- Marco de referencia de conglomerados.
- Se selecciona una muestra de n conglomerados.



Los conglomerados que se seleccionan en la muestra se investigan totalmente.

- Parámetros a estimar:

  - Total: de la característica de los elementos en la población. Y
  - Media por unidad. $\bar{Y}$
  - Media por elemento $\bar{\bar{Y}}$

C. 2 A: Conglomerados de tamaños desiguales.

- ESTIMADORES INSESGADOS y LAS ESTIMACIONES DE SUS VARIANZAS.

• MEDIA POR UNIDAS (conglomerado) $\hat{\bar{Y}} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ donde $y_i = \sum_{j=1}^{M_i} y_{ij}$

$$v(\hat{\bar{Y}}) = v(\bar{y}) = \frac{1-f}{n}\,\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}$$

• TOTAL $\hat{Y} = N\bar{y}$ $\qquad v(\hat{Y}) = N^2 v(\bar{y})$

• MEDIA por elemento $\hat{\bar{\bar{Y}}} = \bar{\bar{y}} = \frac{\hat{Y}}{M} = \frac{1}{\bar{M}n}\sum_{i=1}^{n} y_i$

$$v(\hat{\bar{\bar{Y}}}) = v(\bar{\bar{y}}) = \frac{1-f}{\bar{M}^2 n}\,\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}$$

- ESTIMADORES DE RAZÓN: La variable auxiliar es M_i tamaño del conglomerado.

• Media por elemento $\hat{\bar{\bar{Y}}}_R = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i}$, $v(\hat{\bar{\bar{Y}}}_R) \doteq \frac{1-f}{n\bar{M}^2}\,\frac{\sum_{i=1}^{n}(y_i - \bar{\bar{y}}_R M_i)^2}{n-1}$

• Total $\hat{Y}_R = M\,\hat{\bar{\bar{Y}}}_R$, $v(\hat{Y}_R) \doteq M^2 v(\hat{\bar{\bar{Y}}}_R)$

• Proporción $\hat{P}_R = p_R = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} M_i}$, a_i: nº de elementos del conglom i-ésimo que pertenecen a la clase de interés.

$$v(p_R) \doteq \frac{1-f}{n\bar{M}^2}\,\frac{\sum a_i^2 - 2p_R\sum a_i M_i + p_R^2 \sum M_i^2}{n-1}, \quad \hat{\bar{M}} = \frac{\sum M_i}{n}$$

CASO B: Conglomerados de tamaños iguales

Población : N conglomerados
Muestra : n conglomerados

$M' = \bar{M}$ tamaños de los conglomerados
$M = NM'$ total de elem. en la población.

$$y_i = \sum_{j=1}^{M'} y_{ij}$$

Estimadores:

- Media por unidad (conglomerado)

$$\hat{\bar{Y}} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \quad , \quad v(\bar{y}) = \frac{1-f}{n} \frac{\sum (y_i - \bar{y})^2}{n-1}$$

- Total

$$\hat{Y} = N\bar{y} \quad , \quad v(\hat{Y}) = N^2 v(\bar{y})$$

- Media por elemento

$$\hat{\bar{\bar{Y}}} = \frac{\sum y_i}{nM'} \quad , \quad v(\hat{\bar{\bar{Y}}}) = \frac{1}{(M')^2} v(\bar{y})$$

- Proporción

$$\hat{P} = p = \frac{\sum a_i}{nM'}$$

centro de educación continua
división de estudios de posgrado
facultad de ingeniería unam

FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO

B I B L I O G R A F I A

JUNIO, 1980

BIBLIOGRAFIA

1. Mendenhall, W. y Scheaffer, R. L., "Mathematical statistics with applications", Duxbury Press (1973).

2. Marascuilo, L. A. y Mc Sweeney, M., "Nonparametric and distribution-free methods for the social sciences", Brooks/Cole Publ. Co. (1977)

3. Blake, I. F., "An introduction to applied probability" John Wiley (1979)

4. Ott, L., "An introduction to statistical methods and data analysis", Duxbury Press (1977)

5. Afifi, A. A. y Azen, S. P., "Statistical analysis", Academic Press (1979)

6. Cassel, C. M., Särndal, C. E. y Wretman, J. H., "Foundations of inference in survey sampling", John Wiley (1977)

7. Davies, O. L., "The design and analysis of industrial experiments", Longman Group Limited (1979)

8. Timm, N. H., "Multivariate analysis with applications in education and psychology", Brook/Cole Publ. Co.

9. Spatz, Ch. y Johnston, J. O., "Basic statistics: tales of distributions" Brooks/Cole Publ. Co.

10. Kreyszig, E., "Introducción a la estadística matemática", Limusa-Wiley (1973)

11. Larson, H. J., "Introducción a la teoría de probabilidades e inferencia estadística", Limusa-Wiley (1978)

12. Rascón, O. A., "Introducción a la Estadística Descriptiva", Vols. I y II, Ed. UNAM

13. Rascón, O. A., "Introducción a la Teoría de Probabilidades", Ed. UNAM

14. Bair, D., "Experimentation: an introduction to measurement theory and experiment design", Prentice Hall (1962)

15. Benjamin, J., "Probability, statistics, and decision for civil engineers", McGraw-Hill (1970)

16. Bruning, J. and B. Kintz, "Computational handbook of statistics", Scott, Foreman and Co. (1968)

17. Cochran, W., "Experimental designs", Wiley (1957)

18. Dubes, R., "the theory of applied probability", Prentice-Hall (1968)

19. Feller, W., "Introducción a la teoría de probabilidades y sus aplicaciones", Limusa-Wiley (1973)

20. Freund, J., "Mathematical statistics", Prentice Hall (1971)

21. Hays, W., "Statistics, probability, inference and decisions", Holt-Rinehart and Winston (1970)

22. Kish, L.,"Muestreo de encuestas", Trillas (1972)

23. Lindgren, B., "Statistical theory", Macmillan (1968)

24. Van der Gerr, J., "Introduction to multivariate analysis for the social sciences", Freeman (1971)

Directorio de Alumnos del curso: "FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO, 1980 .

1. David A. Jarquín Aguado
COVITUR
Av. Juárez 42 B
México 1, D.F.
585.10.11 Ext. 212

Edif. Edo. de Chiapas B-8
México 3, D.F.
529.22.40

2. Martha Angeles Alvarez
C F E
Ejército Nal. 1112-2°
México 10, D.F.
557.51 .18

Universidad 1900 -5-303
México 21 DF
548 93 49

3. Teócrito Arteaga N.
Esc. Sup. de Educ. Física
Pto. No. 4 de la
Cda. Deportiva
México 8, DF
519 50 60

Ote 116 # 1203
Col. Héroes de Churubusco
México 13, DF.
581 38 60

4. Sabino Becerril Sánchez
S A H O P
Xola y Ave. Universidad
México 12, DF.

Pestalozzi 422
México 12 D F
519. 91.48

5. Antonio Becerril Téllez Girón
SAHOP
Lago Pte 16
Col. Lago
México, D.F.
519 78.15

Clzada. Sta. Anita 152-2
México 13, DF.

6. Omar Fernando Bon Rojo
S A R H
P. de la Reforma 107-4°
México 4, D.F.
546.16.71

General Cano 8-16
México 18, DF.
516.58.25

7. Josefina Brizuela González
I M S S
Toledo 10
México, D.F.
525.83.31

Horacio Nelson 42-1
México 13, D.F.
590.16.42

8. Joaquín Cardoso Frías
CIECCA
Av. Sn. Bernabe 549
México 20, D.F.
595.23.22

Oscar Wilde 29-203
México 5, D.F.
545.76.52

9. Javier Alma Engracia Cortés
SARH
Tonalá 104-3°.
México 7; DF.
574.17.86

Retorno 2 # 33
Jardín Balbuena
ZP 9
552.18.29

10. José Manuel Flores Ramos
I M S S
Toledo 10-1°
México , D.F.
511.71.65

Pte. Carranza 101
México 21 DF
554.25.89

11. Ana Bertha Fregoso Ríos
C F E
Río Ródano 14.
México 5, DF.
286.33.17

Anaxágoras 321
México 12, DF
536.10.79

12. Patricia González Juárez
S A R H
Av. Sn. Bernabe 549
México21, DF
595.3450

Acolotanco 49-133
Azcapotzalco
ZP 16
352.46.03

13. Justo Gutiérrez Moyado
Colegio de Bachilleres
Cuauhtémoc 1236-7°
México 13, DF
559.55.22 Ext.182

Virgina 178 Depto. 45
México 13 DF
674.01.98

14. Roberto Gutiérrez Pinal
S A H O P
Lago Pte. 16
Col. Lago
ZP 13
674.17.27

Matías Romero 1603
México 13, DF

15. René Hernández Carreón
S A H O P
Jalapa 147-2°
México 7 DF
574.82.37

Odesa 821
México 13 DF
539.67.63

16. Oscar Huerta Martínez
C F E
Ejército Nal. 1112
México, D.F.
557.57.44

17. Miguel Enrique Jaras Tabares
I D D E C S A
Insurgentes Sur 576 -1°
México, D.F.
543.96.50

E . Zapata Manz 1 Lote 5
México 19, D.F.

18. Armando Daniel Jasso Arías
S A R H
Reforma 107-8°.
México 4, D.F.
592.10.62

Calle 25 No. 94
México 18, DF.
55 80.44

19. Jorge Ledesma Aguillón
Inst. Tecnológico Regional de Querétaro
Av. Tecnológico y Gral. Mariano Escobedo
Querétaro, Qro.
2 22.81

La Piedad II
Querétaro, Qro.

20. Manuel Llano O.
Asbestos de México, S.A.
Gustavo Baz Km. 12.5
Tlalnepantla, Edo. de México
565.01.00

Libertad 26
Col. las arboledas
Edo. de México
379.77.93

21. María del Pilar López Marco
S A R H
Reforma 107-4°
México 4, DF.
546.16.71

Guerrero 380 B -1008
México 3, DF
583.88.95

22. Federico Mac Gregor
Dirección Gral. de Ana. de Inversiones
SAHOP
Lago Pte. 16
México 13, DF.
674.17.43

Frontera 6 Bis
San Angel
ZP 20
548.41.38

23. Miguel F. Martínez Gómez
Sur 73 No. 4245
Col. V. Piedad
ZP.13
519.18.86

24. Daniel Meza Pérez
Comisión Nal. Coordinadora de Puertos
Cuernavaca 5
México 11 DF
553.87.11

Av. Fuentes 15
Ampliación Vista Hermosa
Tlalnepantla, Edo. de Méx.
572.01.27

25. Francisco Miranda G.
Industrias Resistol, SA
Camino Lago de Guadalupe 59

Sierra Nevada 112
Parque Residencial Coacalco
Edo. de Méx.

26. Eric Moreno Mejía
C F E
Melchor Ocampo 436-5°
México 5, DF
286.12.61

Chiapas 202
México 7 DF
584.82.65

27. Fernando Páez Bolio
Depto. de Catastro Edo. de Méx.
Allende 10
Sn. Cristobal Ecatepec

Tapicería 55
México 2, DF.
789.25.23

28. Vicente Pineda Vera
Invases Generales Continental
de México SA
Oriente 107 # 114 Col. Bondojito
México , D.F.
759.16.88

Guanajuato 53-18
México 7 DF
584.30.14

29. Lisandro Ramírez Mercado
S A H O P
Jalapa 147-2°
México 7 DF
574.82.37

Asperón 6327
Col. Tres Estrellas
Z.P.14

30. Alfonso Requena Fernández

31. Ofelia Rodríguez Gómez
Tesorería del D F
Niños Héroes y Dr. Lavista
México 7, DF.
761.29.21

Zempoala 394-12
México 12, DF
696.18.11

32. Albino Rodríguez Araiza
Banco de México SA
5 de Mayo No.2
México 1, D.F.
585.42.99 Ext.154

F. Mex. de Fut Ball 26
V. Lázaro Cárdenas
México 22 DF
594.59.02

33. Miguel Angel Salcedo González
FOVISSSTE
López 15-7°
México I DF
585.11.66

Izamal 11
Pedregal Ajusco
ZP.20
568.75.14

34. Eduardo Salgado Valladares
Tesorería del D D F
Niños Héroes y Dr. Lavista
México 7, DF
588.11.06

Silvestre A. Vargas 159
Colonial Iztapalapa
México 13, DF.
558.47.10

35. Luis Angel Serrano Sandoval
CABLEVISION SA
Dr. Río de la Loza 182
México 7, DF.
588.15.39

Ind. Editorial 161
Tlalnepantla, Edo. de Méx.

36. Eduardo Wolf Fuentes
IMSS
Toledo 10-1°
México, D.F.

Oasis 67
Claveria Z.P.16
399.23.80