

DIRECTORIO, PROFESORES DEL CURSO:
FUNDAMENTOS DE LAS TÉCNICAS DE MUESTREO ESTADÍSTICO

1981

1. M en C Adela Abad Carrillo
Profesora e Investigadora
Escuela Nacional de Estudios Profesionales, ACATLAN
Acatlán, Edo de México
373 23 18
373 23 99 ext 170
2. M en C Edmundo Berumen Torres
Gerente de Investigación de Mercados
Fomento Industrial SOMEX
Hamburgo 18, Planta Baja
Col Juárez
México 6, D F
592 58 33
3. Dr. Octavio A Rascón Chávez
Subdirector del Instituto de Ingeniería
UNAM
Ciudad Universitaria
México 20, D F
548 54 79
4. M en C Luis Alejandro Servín Andrade
Asesor de la Jefatura de Servicios de
Planeación
Paseo de la Reforma 476, piso 2.
México 6, D F
5. M en I Augusto Villarreal Aranda (COORDINADOR)
Profesor
División de Estudios de Posgrado
Facultad de Ingeniería
UNAM
Ciudad Universitaria
México 20, D F
554 45 31

FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO

Fecha	Tema	Horario	Profesor
Junio 8	Introducción	18 a 21 h	M. en C. Edmundo Berumen Torres
Junio 10,12 y 15	Muestreo Aleatorio Simple	18 a 21 h c/día	Dr. Octavio A. Rascón Chávez
Junio 17	Muestreo Aleatorio Simple	18 a 19:30 h	" "
	Tamaño de la Muestra	19:30 a 21 h	M. en I. Augusto Villarreal Aranda
Junio 19 a 22	Tamaño de la Muestra	18 a 21 h c/día	" "
Junio 24 y 26	Muestreo Aleatorio Simple para Razones	18 a 21 h c/día	M.en C. Adela Abad Carrillo
Junio 29	Muestreo Estratificado	18 a 21 h c/día	M. en C. Luis Alejandro Servín Andrade
Julio 1º y 3			
Julio 6 y 8	Muestreo por Por Conglomerados	18 a 21 h c/día	M. en C. Adela Abad Carrillo
Julio 10 y 13	Muestreo Sistemático	18 a 21 h c/día	M. en C. Edmundo Berumen Torres

'cdcs.

EVALUACION DEL PERSONAL DOCENTE

①

CURSO: Fundamentos de las Técnicas de Muestreo Estadístico

FECHA: junio 8 al 13 de julio, 1981.

	DOMINIO DEL TEMA	EFICIENCIA EN EL USO DE AYUDAS AUDIOVISUALES	MANTENIMIENTO DEL INTERES. (COMUNICACION CON LOS ASISTENTES, AMENIDAD, FACILIDAD DE EXPRESION).	PUNTUALIDAD
--	------------------	--	---	-------------

CONFERENCISTA					
1.	M. EN C. EDUARDO BERUMEN TORRES				
2.	OCTAVIO A. PASCON CHAVEZ				
3.	M. EN I. AUGUSTO VILLARREAL ARANDA				
4.	M. EN C. ADELA ABAD CARRILLO				
5.	M. EN C. LUIS ALEJANDRO SERVIN ANDRADE				
6.					
7.					
8.					
9.					

ESCALA DE EVALUACION : 1 a 10

SU EVALUACION SINCERA NOS AYUDARA A MEJORAR LOS PROGRAMAS POSTERIORES QUE DISEÑAREMOS PARA USTED.

TEMA	ORGANIZACION Y DESARROLLO DEL TEMA	GRADO DE PROFUNDIDAD LOGRADO EN EL TEMA	GRADO DE ACTUALIZACION LOGRADO EN EL TEMA	UTILIDAD PRACTICA DEL TEMA	
Introducción					
Muestreo Aleatorio Simple					
Tamaño de la Muestra					
Muestreo aleatorio simple para razones					
Muestreo estratificado					
Muestreo por conglomerados					
Muestreo sistemático.					
ESCALA DE EVALUACION: 1 a 10					

EVALUACION DEL CURSO

3

CONCEPTO		EVALUACION
1.	APLICACION INMEDIATA DE LOS CONCEPTOS EXPUESTOS	
2.	CLARIDAD CON QUE SE EXPUSIERON LOS TEMAS	
3.	GRADO DE ACTUALIZACION LOGRADO CON EL CURSO	
4.	CUMPLIMIENTO DE LOS OBJETIVOS DEL CURSO	
5.	CONTINUIDAD EN LOS TEMAS DEL CURSO	
6.	CALIDAD DE LAS NOTAS DEL CURSO	
7.	GRADO DE MOTIVACION LOGRADO CON EL CURSO	

ESCALA DE EVALUACION DE 1 A 10

1. ¿Qué le pareció el ambiente en la División de Educación Continua?

MUY AGRADABLE	AGRADABLE	DESAGRADABLE

2. Medio de comunicación por el que se enteró del curso:

PERIODICO EXCELSIOR ANUNCIO TITULADO DE VISION DE EDUCACION CONTINUA	PERIODICO NOVEDADES ANUNCIO TITULADO DE VISION DE EDUCACION CONTINUA	FOLLETO DEL CURSO

CARTEL MENSUAL	RADIO UNIVERSIDAD	COMUNICACION CARTA, TELEFONO, VERBAL, ETC.

REVISTAS TECNICAS	FOLLETO ANUAL	CARTELERA UNAM "LOS UNIVERSITARIOS HOY"	GACETA UNAM

3. Medio de transporte utilizado para venir al Palacio de Minería:

AUTOMOVIL PARTICULAR	METRO	OTRO MEDIO

4. ¿Qué cambios haría usted en el programa para tratar de perfeccionar el curso?

5. ¿Recomendaría el curso a otras personas?

SI	NO

6. ¿Qué cursos le gustaría que ofreciera la División de Educación Continua?

7. La coordinación académica fue:

EXCELENTE	BUENA	REGULAR	MALA

8. Si está interesado en tomar algún curso intensivo ¿Cuál es el horario más conveniente para usted?

LUNES A VIERNES DE 9 A 13 H. Y DE 14 A 18 H. (CON COMIDAS)	LUNES A VIERNES DE 17 A 21 H.	LUNES, MIÉRCOLES Y VIERNES DE 18 A 21 H.	MARTES Y JUEVES DE 18 A 21 H.

VIERNES DE 17 A 21 H. SABADOS DE 9 A 14 H.	VIERNES DE 17 A 21 H. SABADOS DE 9 A 13 Y DE 14 A 18 H.	O T R O

9. ¿Qué servicios adicionales desearía que tuviese la División de Educación Continua, para los asistentes?


10. Otras sugerencias:



**DIVISION DE EDUCACION CONTINUA
FACULTAD DE INGENIERIA U.N.A.M.**

FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO

TAMAÑO DE LA MUESTRA



M. EN I. AUGUSTO VILLARREAL ARANDA

JUNIO, 1981

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 calcular 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 \bar{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

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

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-α) % 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,

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

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 $\sigma_{\bar{X}}$, la desviación estándar de la distribución muestral de \bar{X} (o error estándar de \bar{X}) 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, ésta 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:

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

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

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

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

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

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

$$n_{\infty} = \frac{K^2 \sigma_x^2}{e^2} \quad (4.3)$$

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_u dado por la ec 4.3, se obtiene finalmente

$$n = \frac{n_u}{1 + \frac{1}{N_p} (n_u - 1)} \quad (4.4)$$

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

4.1.4 Empleo adecuado de n y n_u

Para una población finita, se definirá la fracción de muestreo como

$$\text{fracción de muestreo} = f_m = \frac{n}{N_p}$$

siendo n_u 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

$$f_m \leq 0.05$$

Lo anterior quiere decir que en la práctica se calcula el valor de n_u , y si n_u/N_p cumple con la condición mencionada, entonces se considera que n_u 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_u 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

- 1000
- 100

Solución

- 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

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

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

$$\therefore n = 14$$

En virtud de que en este caso

$$f_m = \frac{n_e}{N_p} = \frac{14}{1000} = 0.014 < 0.05$$

se considera que $n = 14$.

b. En este caso

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

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

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

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

$$\therefore n = 13$$

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 calificación 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

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

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

Por lo tanto,

$$n = \frac{Z^2 \sigma_X^2}{e^2} = \frac{(2.58)^2 (49)}{(1)^2}$$

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

O sea $n_0 = 327$

Puesto que

$$f_m = \frac{n_0}{N_p} = \frac{327}{4726} = 0.0692 > 0.05$$

se procede a calcular n , es decir,

$$n = \frac{n_0}{1 + \frac{1}{N_p} (n_0 - 1)} = \frac{327}{1 + \frac{1}{4726} (327 - 1)}$$

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

$$n = 306$$

Ejemplo 4.3

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

Nº 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,

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

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

$$= 2769.01 - 2750.35 = 10.66 \text{ pulgadas}$$

Puesto que el error en la estimación de la media no debe ser mayor del 2.86%, y el estimador de μ_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

$$n_2 = \frac{t_c^2 \frac{s^2}{n} + 1}{1.5^2} \quad (4.5)$$

Ya que el valor de t_c depende del número de grados de libertad de la muestra n , y este último depende del tamaño de la muestra (ya que $v = n - 1$), la fórmula anterior para obtener el valor de n_2 contiene dos incógnitas. Por ello, se sigue el siguiente proceso iterativo para obtener el valor de n_2 :

1. Se hace $t_{0.025} = 2.075$, es decir

$$t_{0.025} = 1.96$$

Con dicho valor de t_c se obtiene

$$n_2 = \frac{(1.96)^2 (10.66)}{(1.5)^2} + 1 = 18.2 + 1 = 19.2 \rightarrow 20$$

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

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

$$n_2 = \frac{(2.09)^2 (10.66)}{(1.5)^2} + 1 = 20.7 + 1 = 21.7 \rightarrow 22$$

De la tabla de la distribución t , se obtiene $t_{0.025} = 2.08$, para $v = 22 - 1 = 21$ grados de libertad.

1. Se toma ahora $t_{0.025} = 2.08$, y se obtiene

$$n_2 = \frac{(2.08)^2 (10.66)}{(1.5)^2} + 1 = 20.5 + 1 = 21.5 \rightarrow 22$$

En este paso se obtiene un valor de n_2 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_2 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 total, 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 el 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

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

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 variancia de la distribución muestral de \hat{Y} es

$$\sigma_{\hat{Y}}^2 = \sigma_{N_p \bar{y}}^2 = \text{Var}(N_p \bar{y}) = N_p^2 \text{Var}(\bar{y}) = N_p^2 \sigma_{\bar{y}}^2$$

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 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 tre $N_p e^2$, se obtiene

$$n = \frac{\frac{K^2 N_p^3 \sigma_y^2}{N_p e^2}}{\frac{e^2 N_p - e^2 + K^2 N_p^2 \sigma_y^2}{N_p e^2}} = \frac{N_p^2 \frac{K^2 \sigma_y^2}{e^2}}{1 - \frac{1}{N_p} + \frac{N_p^2 K^2 \sigma_y^2}{N_p e^2}}$$

Considerando la ec 4.3, queda finalmente

$$n = \frac{N_p^2 n_w}{1 + \frac{1}{N_p} (N_p^2 n_w - 1)}$$

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

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

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 pueda aproximarse mediante la normal. Por lo tanto,

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

$$N_p = 676$$

$$\sigma_y^2 \approx S_y^2 = 224.5$$

$$N_p^2 n_p = N_p^2 \frac{K^2 \sigma_y^2}{e^2} = \frac{(676)^2 (1.96)^2 (224.5)}{(995)^2} = 397.9$$

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

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

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

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 pueda estar en una de dos clases: A o 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 elementos que están en B es

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

por lo cual

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

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

$$P = \frac{A}{N_p} = \frac{\sum_{i=1}^{N_p} X_i}{N_p} = \bar{u}_x$$

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

$$\mu_p = P$$

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

Si la población es infinita, se obtiene

$$\sigma_p = P$$

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

estimándose P en ambos casos con el valor de 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,

$$n_m = \frac{K^2 PQ}{e^2} \tag{4.7}$$

para muestreo de población infinita, y

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

para muestreo de población finita con tamaño N_p .

Usualmente se calcula primero el valor de n_m , y si la fracción de muestreo es mayor de 0.05, se calcula a continuación el valor de

$$0.0001 = \frac{360 - 0.09 n}{3999 n}$$

$$0.3999 n = 360 - 0.09 n$$

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

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

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

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

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

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

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

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

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

$$K = z_{0.475} = 1.96$$

por lo cual

$$n = \frac{K^2 pq}{e^2} = \frac{(1.96)^2 (0.44)(0.56)}{(0.05)^2}$$

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

$$\therefore n = 339$$

Como

$$fn = \frac{n}{N} = \frac{339}{3208} = 0.106 > 0.05$$

se corrige el valor anterior, obteniéndose finalmente

$$n = \frac{n}{1 + \frac{1}{N} (n - 1)} = \frac{339}{1 + \frac{1}{3208} (339 - 1)}$$

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

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



DIVISION DE EDUCACION CONTINUA
FACULTAD DE INGENIERIA U.N.A.M.

FUNDAMENTOS DE LAS TECNICAS DE MUESTRO ESTADISTICO

INFERENCIA ESTADISTICA

M. EN I. AUGUSTO VILLARREAL ARANDA

JUNIO, 1981

INFERENCIA ESTADÍSTICA

Por: M. en I. Augusto V. L. 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

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 reposición.

4. Distribución muestral del promedio aritmético

Supóngase que se extraen sin reposición 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 al muestreo es con reposición), los resultados anteriores se reducen a

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

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

puesto que

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

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 reposición.

Primer procedimiento.

Siendo la población finita y el muestreo sin reposición, 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{y}_1		\bar{x}_1
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}_1	6/3	7/3	8/3	8/3	9/3	9/3	10/3	10/3	11/3	12/3
\bar{x}_1^2	36/9	49/9	64/9	64/9	81/9	81/9	100/9	100/9	121/9	144/9

$$\sum_{i=1}^{10} \bar{x}_1 = 90/3$$

$$\sum_{i=1}^{10} \bar{x}_1^2 = 840/9$$

$$\mu_{\bar{x}} = \bar{x} = \frac{1}{10} \sum_{i=1}^{10} \bar{x}_1 = \frac{1}{10} \cdot \frac{90}{3} = 3$$

$$\sigma_{\bar{x}}^2 = \frac{1}{10} \sum_{i=1}^{10} \bar{x}_1^2 - \bar{x}^2 = \frac{1}{10} \cdot \frac{840}{9} - (3)^2 =$$

$$= 9.333 - 9.000 = 0.333 \rightarrow \sigma_{\bar{x}} = \sqrt{0.333} = 0.577$$

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

Segundo procedimiento.

Por tratarse de una población finita, se verifica que

$$\mu_{\bar{x}} = \mu \quad \text{y} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-p}{N_p-1}}$$

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

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

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

- entre 496 y 500 kg
- 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 10 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

$$\begin{aligned} P[496 < X < 500] &= P[-2.22 < Z < -0.74] = \\ &= P[-2.22 < Z < 0] - P[-0.74 < Z < 0] \end{aligned}$$

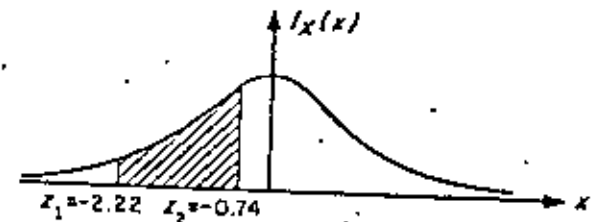


Fig 4.1 Distribución normal correspondiente al ejemplo

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

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

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

$$\begin{aligned} P[X > 510] &= P[Z > 2.96] = P[Z > 0] - P[0 < Z < 2.96] = \\ &= 0.5 - 0.4985 = 0.0015 \end{aligned}$$

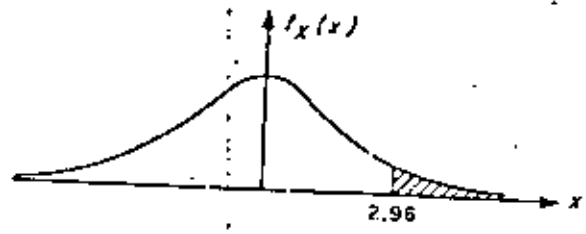


Fig 4.2 Distribución normal correspondiente al ejemplo

5. Distribución muestral de diferencias de promedios aritmético

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:

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

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

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

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

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

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

$$\begin{array}{r} 3 - 2 \quad 7 - 2 \quad 8 - 2 \\ 3 - 4 \quad 7 - 4 \quad 8 - 4 \end{array} \rightarrow \begin{array}{r} 1 \quad 5 \quad 6 \\ -1 \quad 3 \quad 4 \end{array}$$

Es decir,

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

$$\begin{aligned} \sigma_{\bar{X}-\bar{Y}}^2 &= \frac{(-1-3)^2 + (1-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2 + (6-3)^2}{6} \\ &= \frac{34}{6} = \frac{17}{3} \end{aligned}$$

Segundo procedimiento

Se sabe que

$$\mu_{\bar{X}-\bar{Y}} = \mu_{\bar{X}} - \mu_{\bar{Y}} \quad ; \quad \sigma_{\bar{X}-\bar{Y}}^2 = \sigma_{\bar{X}}^2 + \sigma_{\bar{Y}}^2$$

Por ello,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

resulta que la estadística S_n es un estimador consistente. Por ejemplo, el promedio aritmético es un estimador insesgado y consistente de la media, 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 S 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

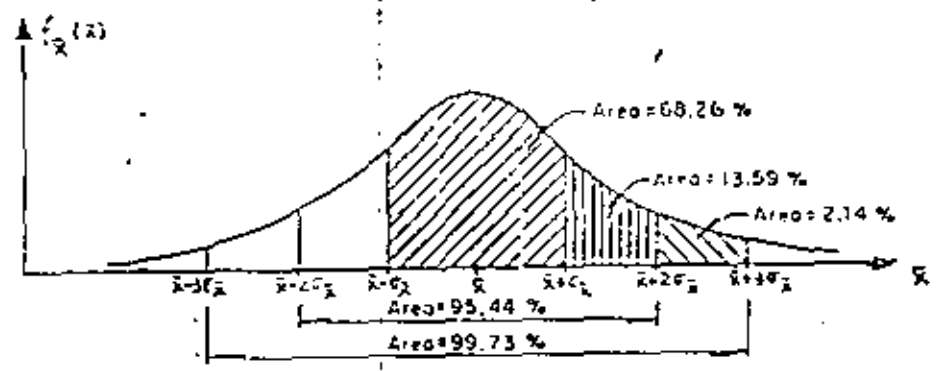


Fig. 8.1

Ejemplo 8.1

Sea el promedio aritmético \bar{x} una estadística con distribución normal. Las probabilidades o niveles de confianza de que μ_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 μ_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.

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

- 95 por ciento
- 97 por ciento

para el diámetro medio de todos los tubos.

- De la tabla 9.1, 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

- El intervalo de confianza del 95% para la media del total de calificaciones.
- 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.
- El nivel de confianza para el cual la media de la población sea 72 ± 1 puntos.

Elevando al cuadrado la desigualdad, queda

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

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

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

a. Si se estima 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%

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

$$\frac{19.6}{\sqrt{n}} \sqrt{\frac{1018 - n}{1018 - 1}} < 2$$

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.

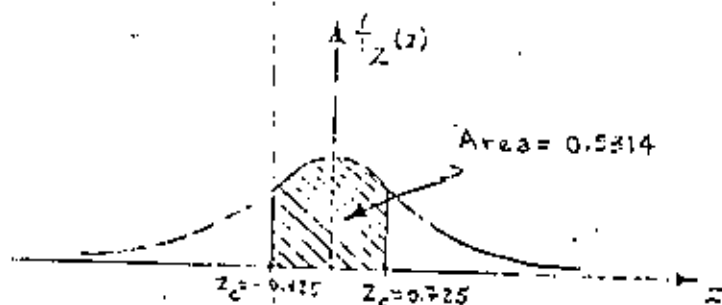


Fig 9.1

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 9.1

$$\begin{aligned} \bar{X}_A - \bar{X}_B \pm z_c \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} &= 6.5 - 6.3 \pm 2 \sqrt{\frac{(0.4)^2}{100} + \frac{(0.3)^2}{100}} \\ &= 0.2 \pm 0.1 \end{aligned}$$

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

- 95%
- 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 B.1, para un nivel de confianza de 95%, $Z_c = 1.96$.

b. En este caso, al emplear la tabla B.1 se obtiene

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

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

$$200 \pm 2.58 (11.04)$$

$$200 \pm 28.483$$

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 las 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 desea 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 esta 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

$$H_0 : \mu_S = \mu_1$$

$$H_1 : \mu_S = \mu_2$$

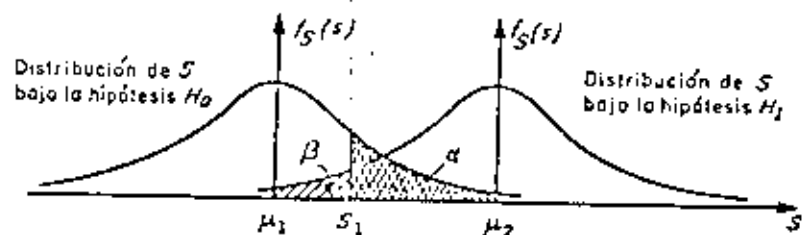
En la fig 13.1 se muestra en forma gráfica la relación entre los errores tipos I y II en el caso en el que la regla de decisión para aceptar o rechazar H_0 es la siguiente:

Si el valor de la estadística S obtenido de una muestra excede de cierto valor crítico S_1 , recházese H_0 ; en caso contrario, aceptese.

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.

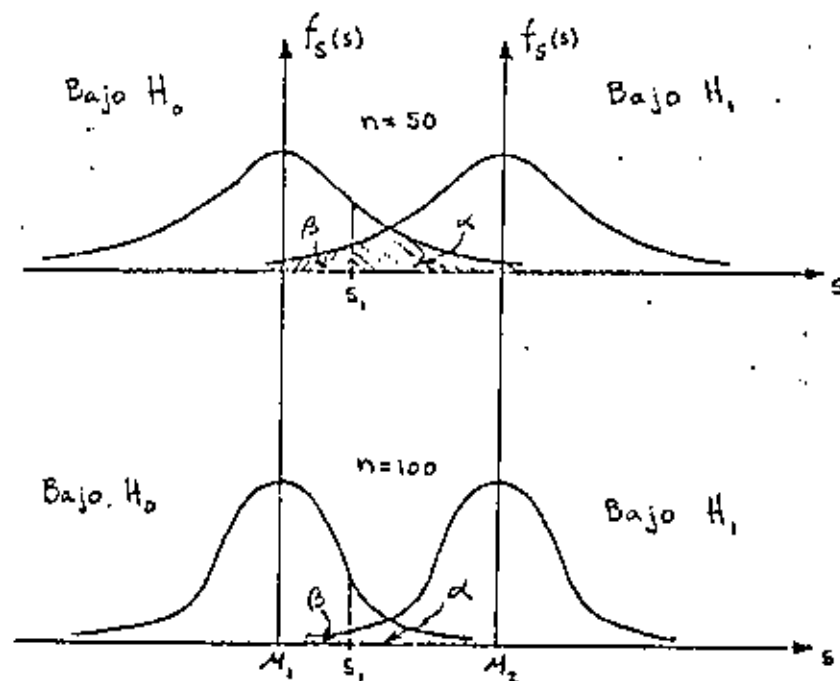


Fig 13.2

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

14. Regiones críticas, de rechazo o 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 lo 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 cae 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 se 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 $Z < -2.58$.

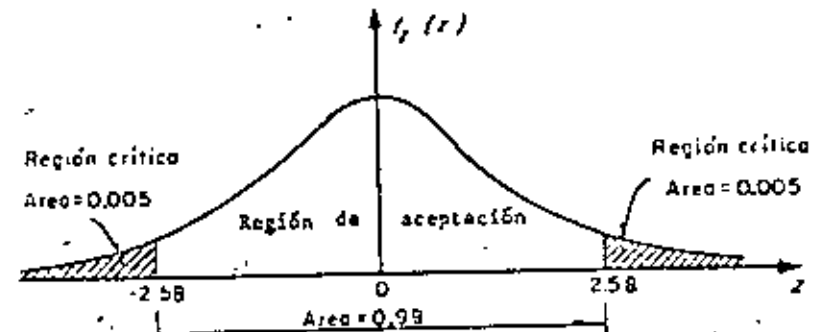


Fig 14.1 Región de significancia

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

$$H_0 : \mu_S = \mu_1$$

$$H_1 : \mu_S < \mu_1$$

$$H_0 : \mu_S = \mu_1$$

$$H_1 : \mu_S > \mu_1$$

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

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Para el caso de muestreo sin remplazo de población finita,

se tiene que $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N_p - n}{N_p - 1}}$, 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

- 0.05
- 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

Acepta 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, rechaza 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 un promedio de resistencia de 912 kg y una desviación estándar igual a 42 kg. ¿Se puede considerar que el nuevo proceso realmente incrementa la resistencia, con un nivel de confianza de 99%?

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

Acepta 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, rechaza 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

$$Z = \frac{X - Y - (\mu_X - \mu_Y)}{\sigma_{\bar{X} - \bar{Y}}} = \frac{X - Y - 0}{\sigma_{\bar{X} - \bar{Y}}} = \frac{\bar{X} - \bar{Y}}{\sigma_{\bar{X} - \bar{Y}}}$$

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. 0.017

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

$$H_0 : \mu_X = \mu_Y$$

$$H_1 : \mu_X > \mu_Y$$

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.

$$\chi^2 = \frac{n S_x^2}{\sigma^2} \quad (3.14)$$

donde S_x^2 es la variancia de la muestra.

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

$$v = n - k$$

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

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

en la que U es una constante que hace que el área total bajo la curva resulte igual a uno, y $v = n - 1$ 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 v .

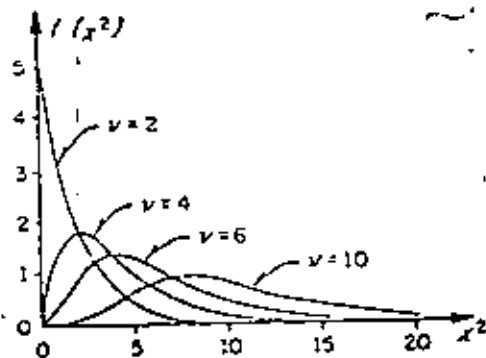


Fig 21. Distribución Ji cuadrada para distintos valores de v

3.4 Muestras pequeñas

Como ya se indicó, para muestras grandes ($n > 30$) 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

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 eq 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).

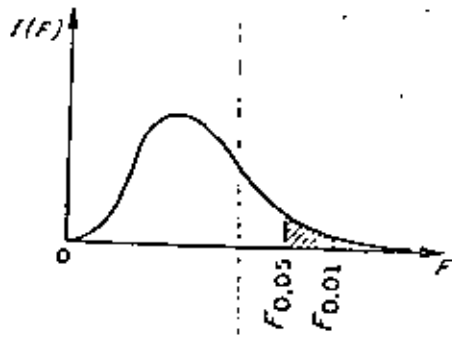


Fig 22. Distribución F .

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 proporcionen grosores muy similares, es decir, que la variancia de los grosores para las dos máquinas es la misma. Para decidir acertadamente, se toma 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

$$T = \frac{\bar{X} - \mu}{S_x} \sqrt{n - 1} \tag{3.16}$$

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

$$f(t) = \frac{U}{1 + \frac{t^2}{v} \frac{\Gamma(v + 1/2)}{\Gamma(v)}} \tag{3.17}$$

U = 1 de aproximadamente

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

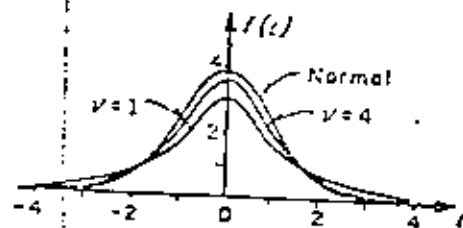


Fig 23. Distribución t de Student para distintos valores de v

En la fig 23 se aprecia que conforme v (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,

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

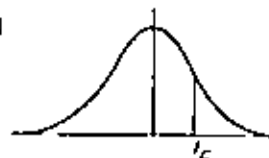
representa un intervalo de confianza para μ , a partir del cual se puede estimar que μ se encuentra dentro del intervalo

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

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



v	$t_{.995}$	$t_{.99}$	$t_{.975}$	$t_{.95}$	$t_{.90}$	$t_{.80}$	$t_{.75}$	$t_{.70}$	$t_{.60}$	$t_{.50}$
1	63.80	31.82	17.71	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	.133
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	1.66	1.30	.851	.681	.529	.255	.126
60	2.66	2.39	2.00	1.67	1.30	.848	.679	.528	.254	.126
80	2.62	2.36	1.98	1.66	1.29	.845	.677	.526	.254	.126
100	2.58	2.33	1.96	1.64	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, \bar{X} y n_Y, S_Y, \bar{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

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

donde

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

cuya distribución es la t de Student, con $v = n_X + n_Y - 2$ 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

- 0.01
- 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,

$$F = \frac{S_X^2}{S_Y^2} = \frac{10.40^2}{(0.36)^2} = 1.27$$

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.



**DIVISION DE EDUCACION CONTINUA
FACULTAD DE INGENIERIA U.N.A.M.**

FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO

PRACTICA DE MUESTREO

M. EN C. ADELA ABAD CARRILLO

JUNIO, 1981

PRACTICA DE MUESTREO.

ESTIMADORES DE RAZON

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	\bar{y} (y)	\bar{x} (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	111 890
Privada	6 366 785	431 041	33 119

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

En un estudio realizado en una zona formada por 70 manzanas, se llevaron 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.

$$\bar{y}_1 = 236 \quad \bar{x}_1 = 115 \quad \sum y_1 x_1 = 685$$

$$\sum y_1^2 = 1494 \quad \sum x_1^2 = 401$$

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

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 siguientes datos:

	1	2	3	4	5	6	7	8	9	10
y_i	15	10	15	14	16	10	14	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%.

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 $\sum y_i = 1750$ y el total correspondiente al trimestre del año anterior es de $\sum x_i = 1200$. 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:
 $\sum y_i^2 = 30650$, $\sum x_i^2 = 15620$, $\sum \frac{1}{n} y_i x_i = 2185.35$

- 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:

$$c_{yy} = 1.13 \quad c_{yx} = 0.78 \quad c_{xx} = 1.11$$

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.

- Se desea estimar el número total de niños en edad pre escolar en el Estado de México en 1979. De acuerdo a información obtenida 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

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

Calcular intervalos del 95% de confianza.

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

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

WASHINGTON, 21 de mayo. (NYT)—Una veterana de la guerra Mary Edwina Walker, capitana durante la guerra Civil de Estados Unidos, fue privada de su Medalla de Honor por el presidente del gobierno, la señora Walker se le negó el honor que otras mujeres habían ganado. La capitana Walker fue la única mujer que se le concedió la Medalla de Honor de la Secretaría Armada. Ella es la única de esta distinción y el único honor de guerra que se le concedió a una mujer.

El mes pasado la dama del Ejército fue la única mujer que se le concedió la Medalla de Honor por el presidente del gobierno, la señora Walker se le negó el honor que otras mujeres habían ganado. La capitana Walker fue la única mujer que se le concedió la Medalla de Honor de la Secretaría Armada. Ella es la única de esta distinción y el único honor de guerra que se le concedió a una mujer.

La señora Walker fue la única mujer que se le concedió la Medalla de Honor por el presidente del gobierno, la señora Walker se le negó el honor que otras mujeres habían ganado. La capitana Walker fue la única mujer que se le concedió la Medalla de Honor de la Secretaría Armada. Ella es la única de esta distinción y el único honor de guerra que se le concedió a una mujer.

La señora Walker fue la única mujer que se le concedió la Medalla de Honor por el presidente del gobierno, la señora Walker se le negó el honor que otras mujeres habían ganado. La capitana Walker fue la única mujer que se le concedió la Medalla de Honor de la Secretaría Armada. Ella es la única de esta distinción y el único honor de guerra que se le concedió a una mujer.

Números aleatorios					
04433	80674	34520	18222	10610	05794
80298	47829	72648	37414	75755	04717
67884	59551	67533	68123	17730	95882
89512	32155	51906	61662	64130	16563
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.a. dentro de cada estrato. Una muestra de 10 renglones debe ser repartida con asignación proporcional.

Calcular intervalos del 95% de confianza para la media en consideración.
De manera muy breve voy 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 hombre 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 variantes para número de horas hombre 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 asignación proporcional
- b. Estimar el número total de horas hombre perdidas durante el mes dado y establecer 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	1	5	1
0	16	32	0	24	8
6	0	16	8	12	
7	4	4	3	2	
9	5	8	1	6	
18	2	0			

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 nivel de error se dará una varianza mínima para el estimador de \bar{y} ?
 - b. ¿Qué fórmula para la varianza de \bar{y} es en el caso de asignación proporcional si se supone que la variabilidad por unidades 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.

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

Estrato	n_h	S_h
1	0.6	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 spj).

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

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

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

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

Dada una población de $N = 500$ recipientes 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	$\frac{2}{200}$	$\frac{3}{200}$	$\frac{6}{200}$	$\frac{1}{200}$	$\frac{2}{200}$

- Estime el porcentaje de defectuosos para ese período de producción y la varianza de su estimación.
- 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 ($i = 1, \dots, 500$), proponga Ud. un estimador para el porcentaje de defectuosos y una expresión para la varianza de su estimador.
- ¿Cuál de los dos esquemas estimaría más adecuado?

Una ciudad pequeña, para efecto de estimación del ingreso medio de 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.e.s. de 25 manzanas. Los resultados obtenidos se presentan en el cuadro siguiente.

- Estime el ingreso medio por varón adulto en la ciudad y su e.e.
- Estime el ingreso total de los varones adultos en la ciudad y su e.e.
- Sabiendo que hay 2500 varones adultos en la ciudad, estime el ingreso total de los varones adultos en la ciudad y su e.e.

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	Nº de varones adultos	Ingreso total x cong.	No. de varones adultos en casa rentada (a_i)
1	6	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	63,000	2
9	8	45,000	3
10	3	50,000	2
11	2	65,000	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
21	6	51,000	3
22	8	30,000	3
23	7	39,000	4
24	3	47,000	0
25	8	41,000	3

$\Sigma x_i = 151$ $\Sigma y_i = 1,329,000$

3. El gerente de una editorial periodística desea estimar el número promedio de periódicos comprados por casa en una comunidad dada. Los costos de transporte 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.e.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 periódicos por casa en la comunidad y obtenga su error estándar.

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:

- El número medio de alumnos por salón de clase.
- La varianza del número medio de alumnos por salón de clase.
- Intervalos de confianza del 95% para el inciso a.
- 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.
- La varianza del inciso d.
- El número medio de alumnos por escuela y su desviación estándar.
- 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.



**DIVISION DE EDUCACION CONTINUA
FACULTAD DE INGENIERIA U.N.A.M.**

FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO

Muestro Estratificado

Muestreo por Conglomerados

Submuestreo

M en C Luis Alejandro Servín

Junio, 1981

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
Guerrero	19 043	635
Hidalgo	429	15
Jalisco	5 010	167
Michoacán	1 114	38
Morales	721	25
Nayarit	474	16
Nuevo León	4 415	148
Oaxaca	450	15

ENTIDAD	NUMERO DE EMPLEADOS	NUMERO DE HOJAS
Puebla	2 750	92
Queretaro	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 elegir a 1 403 empleados de entre los 42 090 agrupados 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 más 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) Encuentra 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) Reconozca 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:

- $$n_1 = (19\ 043/42\ 090)1\ 403 = 635$$
- $$n_2 = (429/42\ 090)1\ 403 = 14$$
- $$n_3 = (5\ 010/42\ 090)1\ 403 = 167$$
- $$n_4 = (1\ 114/42\ 090)1\ 403 = 37$$
- $$n_5 = (721/42\ 090)1\ 403 = 24$$
- $$n_6 = (474/42\ 090)1\ 403 = 16$$
- $$n_7 = (4\ 415/42\ 090)1\ 403 = 147$$
- $$n_8 = (450/42\ 090)1\ 403 = 15$$
- $$n_9 = (2\ 750/42\ 090)1\ 403 = 92$$
- $$n_{10} = (487/42\ 090)1\ 403 = 16$$
- $$n_{11} = (150/42\ 090)1\ 403 = 5$$
- $$n_{12} = (925/42\ 090)1\ 403 = 31$$
- $$n_{13} = (2\ 800/42\ 090)1\ 403 = 93$$
- $$n_{14} = (2\ 900/42\ 090)1\ 403 = 97$$
- $$n_{15} = (422/42\ 090)1\ 403 = 14$$

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 Cuernavaca. 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 elegir 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 más 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 $f=47/1\ 410$ 5 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

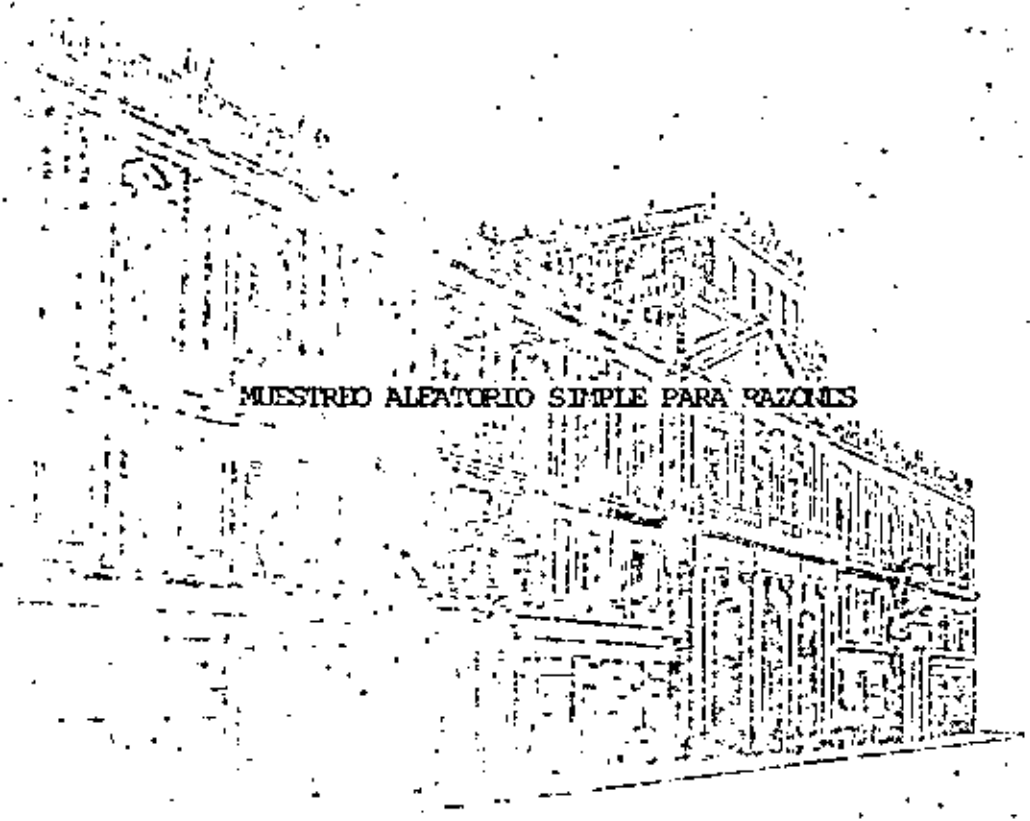
dentro de cada una de ellas entramos con fracción de muestreo 1 de cada 2. El número de primarias en la muestra se ha duplicado. Otra opción es usar 188 primarias en la muestra y dentro de cada una de ellas seleccionar a 1 de cada 4 nombres, es decir, $f=(188/1\ 410)(1/4)=1/30$. A medida que aumentamos 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 más edificios o ciudades (digamos) buscando a las distintas oficinas seleccionadas.

Ing. Luis Servín.



**DIVISION DE EDUCACION CONTINUA
FACULTAD DE INGENIERIA U.N.A.M.**

FUNDAMENTOS DE LAS TÉCNICAS DE MUESTREO ESTADÍSTICO



MUESTREO ALEATORIO SIMPLE PARA RAZONES

M. EN C. ADELA ABAD CARRILLO

JUNIO, 1991

MUNICIPIO ALFONSO GARCÍA RIVERA

1. En una pequeña comunidad se realizó una investigación para determinar qué proporción del gasto familiar es dedicada a la alimentación y qué proporción es dedicada a la atención médica y medicamentos. Se seleccionó una muestra aleatoria simple de 40 familias de la 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	Médicos	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		1
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	6.5	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	1.6		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 760 kilos. Después de un mes de dejó 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 el peso inicial del pollo, y el peso al mes del experimento

$$\sum_{i=1}^{30} x_i = 37.5, \quad \sum_{i=1}^{30} y_i = 16.7, \quad \sum x_i^2 = 17.67, \quad \sum y_i^2 = 25.67, \quad \sum x_i y_i = 10.7$$

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 cuantos 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	25	25
15	28	30
16	20	20
17	38	40
18	30	30
19	32	30
20	20	20
21	25	30
22	25	25
23	32	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:

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

Adela Abad Carrillo.

Junio de 1980



**DIVISION DE EDUCACION CONTINUA
FACULTAD DE INGENIERIA U.N.A.M.**

FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO

CONGLOMERADOS

M en C Adela Abad Carillo

Junio, 1981

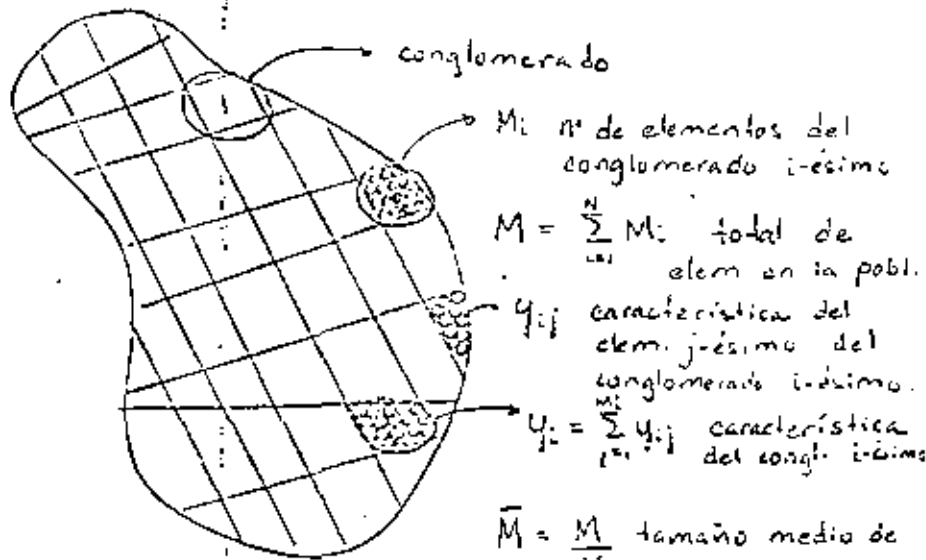
Muestras por Conglomerados

Conceptos:

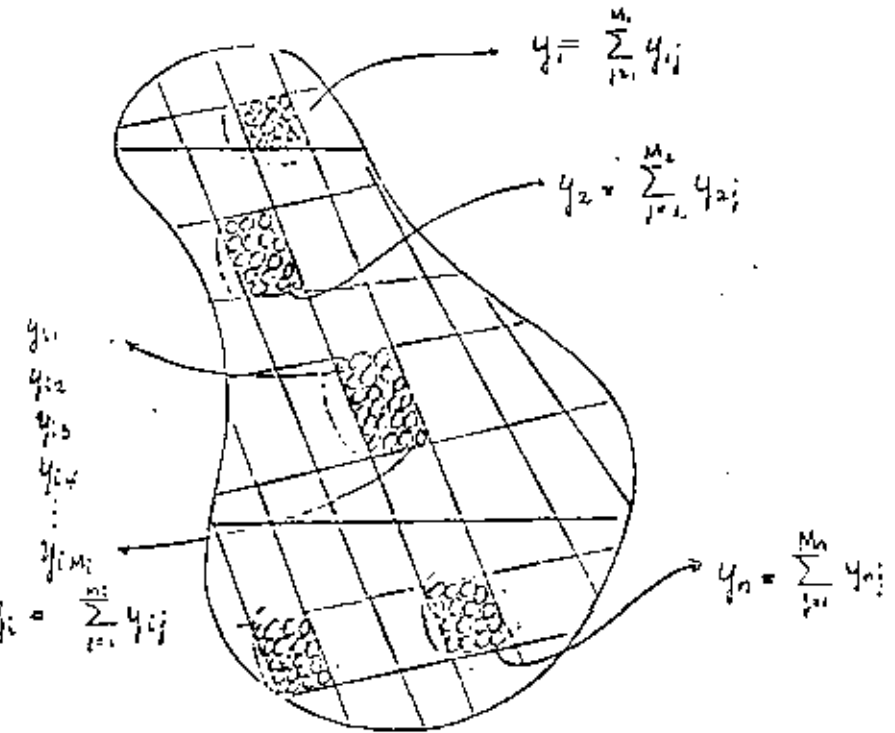
- elemento
- unidad muestral
- conglomerado
- marco de referencia
- ventajas

Notación

Población : N conglomerados



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



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{y}
- Proporción. P

Caso A: Conglomerados de tamaños desiguales.

ESTIMADORES INSESADOS Y LAS ESTIMACIONES DE SUS VARIANZAS.

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

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

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

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

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

• ESTIMADORES DE RAZÓN: la variable auxiliar es M_i tamaño del conglomerado.

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

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

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.

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

Caso B: Conglomerados de tamaños iguales

Población : N conglomerados

$M' = \bar{M}$ tamaños de los conglomerados

Muestra : n conglomerados

$M = NM'$ total de elem. en la población.

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

Estimadores:

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

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

• Media por elemento $\hat{\bar{Y}} = \frac{\sum_{i=1}^n y_i}{nM'}$, $v(\hat{\bar{Y}}) = \frac{1}{(M')^2} v(\bar{y})$

• Proporción $\hat{P} = p = \frac{\sum_{i=1}^n a_i}{nM'}$



**DIVISION DE EDUCACION CONTINUA
FACULTAD DE INGENIERIA U.N.A.M.**

FUNDAMENTOS DE LAS TÉCNICAS DE MUESTREO ESTADÍSTICO

BIBLIOGRAFIA



JUNIO, 1981

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. P., "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 Azon, 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. Fround, 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 Cerr, J., "Introduction to multivariate analysis for the social sciences", Freeman (1971)



**DIVISION DE EDUCACION CONTINUA
FACULTAD DE INGENIERIA U.N.A.M.**

FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO

WHAT IS A SURVEY?

Junio, 1981

What Is A Survey?

By Robert Ferber, Chair
Paul Sheatsley
Anthony Turner
Joseph Waksberg

Subcommittee of the Section on Survey Research Methods

American Statistical Association
Washington, D.C.

This publication is available free upon request, in single copies. For multiple copy orders a small per-copy charge is made. For further information write to the American Statistical Association, 806 Fifteenth Street, N.W., Washington, D.C. 20005. "What Is A Survey" is not copyrighted; users may reproduce portions without formal permission.

Preface

People are accustomed to seeing the results of surveys reported in the daily press, incorporated in advertising claims, and mentioned on numerous occasions by political analysts, social critics, and economic forecasters. Much less frequent, however, is any discussion of the reliability of these surveys or what is involved in carrying them out. The wealth of reported information may easily lull the user into assuming that surveys are easy to undertake, and to overlook the many steps involved in a properly-conducted survey. If technical issues are recognized, there is a frequent tendency to assume that they should be safely left to the survey expert.

In fact, many of the surveys that appear in the daily press are conducted under great time pressure and with insufficient allowance for the many different aspects of the process that need to be controlled. Yet, unless the reader of these survey results is aware of what is involved in a survey, and what quality controls are needed, s(he) is unable to form any opinion of the confidence to be placed in the results, and usually is not even in a position to know what questions to ask about such surveys.

In an effort to fill this gap, the Section on Survey Research Methods of the American Statistical Association appointed a committee to prepare a brochure that would describe survey operations without using technical terminology, and be understandable to interested persons not trained in statistics. The result is the present brochure which, it is hoped, will promote a better understanding of what is involved in carrying out a sample survey, and aspects that have to be taken into account in evaluating the results of surveys.

The American Statistical Association is pleased to publish the report in the hope and expectation that it will prove useful to a wide readership. The association is fortunate to have had four such able statisticians as Robert Ferber, Professor of Economics and Business, Survey Research Laboratory, University of Illinois; Paul Sheatsley, Survey Director, National Opinion Research Center, University of Chicago; Anthony Turner, Chief of International Mathematical-Statistical Staff, Statistical Methods Division, U.S. Bureau of the Census; and Joseph Waksberg, Vice President, Westat, Inc., form the committee that undertook the work. It thanks them for their efforts.

Margaret E. Martin, President
American Statistical Association
1980

Contents

Introduction	1
Characteristics of Surveys	
The Need	3
Who Does Surveys?	4
Types of Surveys	6
What Sort of People Work on Surveys?	7
Are Responses Confidential?	8
How a Survey Is Carried Out	
Designing a Survey	9
Sampling Aspects	11
Conducting a Survey	13
Shortcuts to Avoid	14
Using the Results of a Survey	
How Good is the Survey?	16
Sources of Errors	17
Budgeting a Survey	20
Where to Get More Information	24

Introduction

The growing popularity of surveys for throwing light on different problems has led to a tendency to overlook the fact that surveys involve many technical problems. Too many surveys seem to be conducted more or less on an ad hoc basis, with the result that the GIGO (garbage in, garbage out) principle is brought into play. This brochure seeks to help the non-statistician to avoid this danger, by providing a nontechnical introduction to sample surveys of human populations and the many different ways in which such surveys are used.

The principal focus is on the design of a survey and on the collection of survey data—two areas in which the many intricacies involved are frequently overlooked. However, attention is also given to the need for proper evaluation of survey data, an essential prerequisite for assessing the value of a survey as well as a basis for proper analysis of the data. (Analysis of survey data is a major topic in itself, and is not covered here.)

This brochure can be used in a variety of ways, such as:

- By statisticians and survey agencies, to give prospective clients some appreciation of what is involved in a sample survey.
- By research executives, to help their nonresearch counterparts understand how surveys are conducted.
- By instructors in introductory social science and other courses, to give students a brief introduction to sample surveys.

* By national agencies and others advising in other countries, to give government officials in these other countries an understanding of the various steps of a sample survey.

It should be stressed that this brochure is *not* intended to provide students of statistics or prospective specialists in the field with a comprehensive understanding of survey methods. For this purpose, the books listed at the end of the brochure need to be used, plus many of the specialized sources dealing with the techniques of survey design and data collection. This brochure is meant for nonspecialists, for the users of survey data. If it leads them to have a better appreciation of what is involved in a sample survey, its purpose will have been served.

Characteristics of Surveys

The Need

Any observation or investigation of the facts about a situation may be called a survey. But today the word is most often used to describe a method of gathering information from a number of individuals, a "sample," in order to learn something about the larger population from which the sample has been drawn. Thus, a sample of voters is surveyed in advance of an election to determine how the public perceives the candidates and the issues. A manufacturer makes a survey of the potential market before introducing a new product. A government agency commissions a survey to gather the factual information it needs in order to evaluate existing legislation or draft new legislation. For example, what medical care do people receive, and how is it paid for? Who uses food stamps? How many people are unemployed?

It has been said that the United States is no longer an industrial society but an "information society." That is, our major problems and tasks no longer focus merely on the production of the goods and services necessary to our survival and comfort. Rather, our major problems and tasks today are those of organizing and managing the incredibly complex efforts required to meet the needs and wishes of nearly 220 million Americans. To do this requires a prompt and accurate flow of information on preferences, needs and behavior. It is in response to this critical need for information on the part of the government, business and social institutions that so much reliance is placed upon surveys.

ys come in many different forms and have a wide variety of purposes, but they do have certain characteristics in common. Unlike a census, they gather information from only a small sample of people (or farms, businesses or other units, depending on the purpose of the study). In a bonafide survey, the sample is not selected haphazardly or only from persons who volunteer to participate. It is scientifically chosen so that each individual in the population has a known chance of selection. In this way, the results can be reliably projected to the larger public.

Information is collected by means of standardized questions so that every individual surveyed responds to exactly the same question. The survey's intent is not to describe the particular individuals who by chance are part of the sample, but to obtain a statistical profile of the population. Individual respondents are never identified and the survey's results are presented in the form of summaries, such as statistical tables and charts.

The sample size required for a survey will depend on the reliability needed which, in turn, depends on how the results will be used. Consequently, there is no simple rule for sample size that can be used for all surveys. However, analysts usually find that a moderate sample size is sufficient for most needs. For example, the well-known national polls generally use samples of about 1,500 persons to reflect national attitudes and opinions. A sample of this size produces accurate estimates even for a country as large as the United States with a population of over 200 million.

When it is realized that a properly selected sample of only 1,500 individuals can reflect various characteristics of the total population within a very small margin of error, it is easy to understand the value of surveys in a complex society such as ours. They provide a speedy and economical means of determining facts about our economy and people's knowledge, attitudes, beliefs, expectations, and behavior.

Who Does Surveys?

We all know of the public opinion polls which are reported in the press and broadcast media. The Gallup Poll and the Harris Survey issue reports periodically, describing national public opinion on a wide range of current issues. State polls and metropolitan area polls, often supported by a local newspaper or TV station, are reported regularly in many localities. The major broadcasting networks and national news magazines also conduct polls and report their findings.

But the great majority of surveys are not exposed to public

view. The reason is that, unlike the public opinion polls, most surveys are directed to a specific administrative or commercial purpose. The wide variety of issues with which surveys deal is illustrated by the following listing of actual uses:

1. The U.S. Department of Agriculture conducted a survey to find out how poor people use food stamps.
2. Major TV networks rely on surveys to tell them how many and what types of people are watching their programs.
3. Auto manufacturers use surveys to find out how satisfied people are with their cars.
4. The U.S. Bureau of the Census conducts a survey every month to obtain information on employment and unemployment in the nation.
5. The National Center for Health Statistics sponsors a survey every year to determine how much money people are spending for different types of medical care.
6. Local housing authorities make surveys to ascertain satisfaction of people in public housing with their living accommodations.
7. The Illinois Board of Higher Education surveys the interest of Illinois residents in adult education.
8. Local transportation authorities conduct surveys to acquire information on people's commuting and travel habits.
9. Magazine and trade journals utilize surveys to find out what their subscribers are reading.
10. Surveys are used to ascertain what sort of people use our national parks and other recreation facilities.

Surveys of human populations also provide an important source of basic social science knowledge. Economists, psychologists, political scientists and sociologists obtain foundation or government grants to study such matters as income and expenditure patterns among households, the roots of ethnic or racial prejudice, comparative voting behavior, or the effects of employment of women on family life. (Surveys are also made of nonhuman populations, such as of animals, soils and housing; they are not discussed here, although many of the principles are the same.)

Moreover, once collected, survey data can be analyzed and reanalyzed in many different ways. Data tapes with identification

of individuals removed can be made available for analysis by community groups, scientific researchers and others.

Types of Surveys

Surveys can be classified in a number of ways. One dimension is by size and type of sample. Many surveys study the total adult population, but others might focus on special population groups: physicians, community leaders, the unemployed, or users of a particular product or service. Surveys may be conducted on a national, state or local basis, and may seek to obtain data from a few hundred or many thousand people.

Surveys can also be classified by their method of data collection. Thus, there are mail surveys, telephone surveys, and personal interview surveys. There are also newer methods of data collection by which information is recorded directly into computers. This includes measurement of TV audiences carried out by devices attached to a sample of TV sets which automatically record in a computer the channels being watched. Mail surveys are seldom used to collect information from the general public because names and addresses are not often available and the response rate tends to be low, but the method may be highly effective with members of particular groups; for example, subscribers to a specialized magazine or members of a professional association. Telephone interviewing is an efficient method of collecting some types of data and is being increasingly used. A personal interview in a respondent's home or office is much more expensive than a telephone survey but is necessary when complex information is to be collected.

Some surveys combine various methods. Survey workers may use the telephone to "screen" for eligible respondents (say, women of a particular age group) and then make appointments for a personal interview. Some information, such as the characteristics of the respondent's home, may be obtained by observation rather than questioning. Survey data are also sometimes obtained by self-administered questionnaires filled out by respondents in groups, e.g., a class of school children or a group of shoppers in a central location.

One can further classify surveys by their content. Some surveys focus on opinions and attitudes (such as a pre-election survey of voters), while others are concerned with factual characteristics or behavior (such as a survey of people's health, housing or transportation habits). Many surveys combine questions of both types. Thus, a respondent will be asked if s(he) has heard or read about an

issue, what s(he) knows about it, his (her) opinion, how strongly s(he) feels and why, interest in the issue, past experience with it, and also certain factual information which will help the survey analyst classify the responses (such as age, sex, marital status, occupation, and place of residence).

The questions may be open-ended ("Why do you feel that way?") or closed ("Do you approve or disapprove?"); they may ask the respondent to rate a political candidate or a product on some kind of scale; they may ask for a ranking of various alternatives. The questionnaire may be very brief—a few questions taking five minutes or less, or it can be quite long—requiring an hour or more of the respondent's time. Since it is inefficient to identify and approach a large national sample for only a few items of information, there are "omnibus" surveys which combine the interests of several clients in a single interview. In such surveys, the respondent will be asked a dozen questions on one subject, half a dozen more on another subject, and so on.

Because changes in attitude or behavior cannot be reliably ascertained from a single interview, some surveys employ a "panel design," in which the same respondents are interviewed two or more times. Such surveys are often used during election campaigns, or to chart a family's health or purchasing pattern over a period of time. They are also used to trace changes in behavior over time, as with the social experiments that study changes by low-income families in work behavior in response to an income maintenance plan.

What Sort of People Work on Surveys?

The survey worker best known to the public is the interviewer who calls on the phone, appears at the door, or stops people at a shopping center. Though survey interviewing may occasionally require long days in the field, it is normally part-time occasional work and is thus well suited for individuals who do not seek full-time employment or who wish to supplement their regular income. Previous experience is not usually required for an interviewing job. Most research companies will provide their own basic training for the task. The main requirements are an ability to approach strangers, to persuade them to participate in the survey, and to conduct the interview in exact accordance with instructions.

Behind the interviewers are the in-house research staff who design the survey, determine the sample design, develop the questionnaire, supervise the data collection, carry out the clerical and computer operations necessary to process the completed inter-

view, analyze the data, and write the reports. In most survey research agencies, the senior people will have taken courses in survey methods at the graduate level and will hold advanced degrees in sociology, statistics, marketing, or psychology, or they will have the equivalent in business experience. Middle-level supervisors and research associates frequently have similar academic backgrounds, or they have advanced out of the ranks of clerks, interviewers or coders on the basis of their competence and experience.

Are Responses Confidential?

The privacy of the information supplied by survey respondents is of prime concern to all reputable survey organizations. At the U.S. Bureau of the Census, for example, the confidentiality of the data collected is protected by law (Title 13 of the U.S. Code). In Canada, the Statistics Act guarantees the confidentiality of data collected by Statistics Canada, and other countries have similar safeguards. Also, a number of professional organizations that rely on survey methods have codes of ethics that prescribe rules for keeping survey responses confidential. The recommended policy for survey organizations to safeguard such confidentiality includes:

1. Using only code numbers for the identity of a respondent on a questionnaire, and keeping the code separate from that of the questionnaires.
2. Refusing to give names and addresses of survey respondents to anybody outside of the survey organization, including clients.
3. Destroying questionnaires and identifying information about respondents after the responses have been put onto computer tape.
4. Omitting the names and addresses of survey respondents from computer tapes used for analysis.
5. Presenting statistical tabulations by broad enough categories that individual respondents cannot be singled out.

How a Survey Is Carried Out

As noted earlier, a survey usually has its beginnings when an individual or institution is confronted with an information need and there are no existing data which suffice. A politician may wish to tap prevailing voter opinions in his district about a proposal to build a superhighway through the county. A government agency may wish to assess the impact on the primary recipients and their families of one of its social welfare programs. A university researcher may wish to examine the relationship between actual voting behavior and expressed opinion on some political issue or social concern.

Designing a Survey

Once the information need has been identified and a determination made that existing data are inadequate, the first step in planning a survey is to lay out the objectives of the investigation. This is generally the function of the sponsor of the inquiry. The objectives should be as specific, clear-cut and unambiguous as possible. The required accuracy level of the data has a direct bearing on the overall survey design. For example, in a sample survey whose main purpose is to estimate the unemployment rate for a city, the approximate number of persons to be sampled can be estimated mathematically when one knows the amount of sampling error that can be tolerated in the survey results.

Given the objectives, the methodology for carrying out the survey is developed. A number of interrelated activities are involved. Rules must be formulated for defining and locating eligible respondents, the method of collecting the data must be decided

upon. Questionnaire must be designed and pretested, procedures must be developed for minimizing or controlling response errors, appropriate samples must be designed and selected, interviewers must be hired and trained (except for surveys involving self-administered questionnaires), plans must be made for handling nonresponse cases, and tabulation and analysis must be performed.

Designing the questionnaire represents one of the most critical stages in the survey development process, and social scientists have given a great deal of thought to issues involved in questionnaire design. The questionnaire links the information need to the realized measurement.

Unless the concepts are clearly defined and the questions unambiguously phrased, the resulting data are apt to contain serious biases. In a survey to estimate the incidence of robbery victimization, for example, one might want to ask, "Were you robbed during the last six months?" Though apparently straightforward and clearcut, the question does present an ambiguous stimulus. Many respondents are unaware of the legal distinction between robbery (involving personal confrontation of the victim by the offender) and burglary (involving breaking and entering but no confrontation), and confuse the two in a survey. In the National Crime Survey, conducted by the Bureau of the Census, the questions on robbery victimization do not mention "robbery." Instead, several questions are used which, together, seek to capture the desired responses by using more universally understood phrases that are consistent with the operational definition of robbery.

Designing a suitable questionnaire entails more than well-defined concepts and distinct phraseology. Attention must also be given to its length, for unduly long questionnaires are burdensome to the respondent, are apt to induce respondent fatigue and hence response errors, refusals, and incomplete questionnaires, and may contribute to higher nonresponse rates in subsequent surveys involving the same respondents. Several other factors must be taken into account when designing a questionnaire to minimize or prevent biasing the results and to facilitate its use both in the field and in the processing center. They include such diverse considerations as the sequencing of sections or individual questions in the document, the inclusion of check boxes or precoded answer categories versus open-ended questions, the questionnaire's physical size and format, and instructions to the respondent or to the interviewer on whether certain questions are to be skipped depending on response patterns to prior questions.

Selecting the proper respondent in a sample unit is a key ele-

ment in survey planning. For surveys where the inquiries are basically factual in nature, any knowledgeable person associated with the sample unit may be asked to supply the needed information. This procedure is used in the Current Population Survey, where the sample units are households and any responsible adult in a household is expected to be able to provide accurate answers on the employment-unemployment status of the eligible household members.

In other surveys, a so-called "household" respondent will produce erroneous and/or invalid information. For example, in attitude surveys it is generally accepted that a randomly chosen respondent from among the eligible household members produces a more valid cross section of opinion than does the nonrandomly selected household respondent. This is because a nonrandomly selected individual acting as household respondent is more likely to be someone who is at home during the day, and the working public and their attitudes would be underrepresented.

Another important feature of the survey planning process is devising ways to keep response errors and biases to a minimum. These considerations depend heavily on the subject matter of the survey. For example, memory plays an important role in surveys dealing with past events that the respondent is expected to report accurately, such as in a consumer expenditure survey. In such retrospective surveys, therefore, an appropriate choice of reference period must be made so that the respondent is not forced to report events that may have happened too long ago to remember accurately. In general, attention must be given to whether the questions are too sensitive, whether they may prejudice the respondent, whether they unduly invade the respondent's privacy, and whether the information sought is too difficult even for a willing respondent to provide. Each of these concerns has an important bearing on the overall validity of the survey results.

Sampling Aspects

Virtually all surveys that are taken seriously by social scientists and policy makers use some form of scientific sampling. Even the decennial Censuses of Population and Housing use sampling techniques for gathering the bulk of the data items, although 100 percent enumeration is used for the basic population counts. Methods of sampling are well-grounded in statistical theory and in the theory of probability. Hence, reliable and efficient estimates of a needed statistic can be made by surveying a carefully constructed sample

of a population, as opposed to the entire population, provided of course that a large proportion of the sample members give the requested information.

The particular type of sample used depends on the objectives and scope of the survey, including the overall survey budget, the method of data collection, the subject matter and the kind of respondent needed. A first step, however, in deciding on an appropriate sampling method is to define the relevant population. This target population can be all the people in the entire nation or all the people in a certain city, or it can be a subset such as all teenagers in a given location. The population of interest need not be people; it may be wholesale businesses or institutions for the handicapped or government agencies, and so on.

The types of samples range from simple random selection of the population units to highly complex samples involving multiple stages or levels of selection with stratification and/or clustering of the units into various groupings. Whether simple or complex, the distinguishing characteristics of a properly designed sample are that all the units in the target population have a known, nonzero chance of being included in the sample, and the sample design is described in sufficient detail to permit reasonably accurate calculation of sampling errors. It is these features that make it scientifically valid to draw inferences from the sample results about the entire population which the sample represents.

Ideally, the sample size chosen for a survey should be based on how reliable the final estimates must be. In practice, usually a trade-off is made between the ideal sample size and the expected cost of the survey. The complexity of a sample plan often depends on the availability of auxiliary information that can be used to introduce efficiencies into the overall design. For example, in a recent Federal Government survey on characteristics of health-care institutions, existing information about the type of care provided and the number of beds in each institution was useful in sorting the institutions into "strata," or groups by type and size, in advance of selecting the sample. The procedure permitted more reliable survey estimates than would have been possible if a simple random selection of institutions had been made without regard to size or type.

A critical element in sample design and selection is defining the source of materials from which a sample can be chosen. This source, termed the sampling frame, generally is a list of some kind, such as a list of housing units in a city, a list of retail establishments in a county or a list of students in a university. The sampling frame

can also consist of geographic areas with well-defined natural or artificial boundaries, when no suitable list of the target population exists. In the latter instance, a sample of geographic areas (referred to as segments) is selected and an interviewer canvasses the sample "area segments" and lists the appropriate units—households, retail stores or whatever—so that some or all of them can be designated for inclusion in the final sample.

The sampling frame can also consist of less concrete things, such as all possible permutations of integers that make up banks of telephone numbers, in the case of telephone surveys that seek to include unlisted numbers. The quality of the sampling frame—whether it is up-to-date and how complete—is probably the dominant feature for ensuring adequate coverage of the desired population.

Conducting a Survey

Though a survey design may be well conceived, the preparatory work would be futile if the survey were executed improperly. For personal or telephone interview surveys, interviewers must be carefully trained in the survey's concepts, definitions, and procedures. This may take the form of classroom training, self-study, or both. The training stresses good interviewer techniques on such points as how to make initial contacts, how to conduct interviews in a professional manner and how to avoid influencing or biasing responses. The training generally involves practice interviews to familiarize the interviewers with the variety of situations they are likely to encounter. Survey materials must be prepared and issued to each interviewer, including ample copies of the questionnaire, a reference manual, information about the identification and location of the sample units, and any cards or pictures to be shown to the respondent.

Before conducting the interview, survey organizations frequently send an advance letter to the sample member explaining the survey's purpose and the fact that an interviewer will be calling soon. In many surveys, especially those sponsored by the Federal Government, information must be given to the respondent regarding the voluntary or mandatory nature of the survey, and how the answers are to be used.

Visits to sample units are scheduled with attention to such considerations as the best time of day to call or visit and the number of allowable callbacks for no-one-at-home situations. Controlling the quality of the field work is an essential aspect of good

surveillance practice. This is done in a number of ways, most often through observation or rechecking of a subsample of interviews by supervisory or senior personnel, and through office editing procedures to check for omissions or obvious mistakes in the data.

When the interviews have been completed and the questionnaires filled out, they must be processed in a form so that aggregated totals, averages or other statistics can be computed. This will involve clerical coding of questionnaire items which are not already precoded. Occupation and industry categorizations are typical examples of fairly complex questionnaire coding that is usually done clerically. Also procedures must be developed for coding open-ended questions and for handling items that must be transcribed from one part of the questionnaire to another.

Coded questionnaires are keypunched, entered directly onto tape so that a computer file can be created, or entered directly into the computer. Decisions may then be needed on how to treat missing data and "not answered" items.

Coding, keypunching and transcription operations are subject to human error and must be rigorously controlled through verification processes, either on a sample basis or 100 percent basis. Once a computer file has been generated, additional computer editing, as distinct from clerical editing of the data, can be accomplished to alter inconsistent or impossible entries, e.g., a six-year-old grandfather.

When a "clean" file has been produced, the survey data are in a form where analysts can specify to a computer programmer the frequency counts, cross-tabulations or more sophisticated methods of data presentation or computation that are needed to help answer the concerns outlined when the survey was initially conceived.

The results of the survey are usually communicated in publications and in verbal presentations at staff briefings or more formal meetings. Secondary analysis is also often possible to those other than the survey staff by making available computer data files at nominal cost.

Shortcuts to Avoid

As we have seen, conducting a creditable survey entails scores of activities, each of which must be carefully planned and controlled. Taking shortcuts can invalidate the results and badly mislead the user. Four types of shortcuts that crop up often are failure to use a proper sampling procedure, no pretest of the field procedures, failure to follow up nonrespondents and inadequate quality control.

One way to ruin an otherwise well-conceived survey is to use a convenience sample rather than one which is based on a probability design. It may be simple and cheap, for example, to select a sample of names from a telephone directory to find out which candidate people intend to vote for. However, this sampling procedure could give incorrect results since persons without telephones or with unlisted numbers would have no chance to be reflected in the sample, and their voting preferences could be quite different from persons who have listed telephones. This is what happened with the *Literary Digest* presidential poll of 1936 when use of lists of telephone owners, magazine subscribers and car owners led to a prediction that President Roosevelt would lose the election.

A pretest of the questionnaire and field procedures is the only way of finding out if everything "works," especially if a survey employs a new procedure or a new set of questions. Since it is rarely possible to foresee all the possible misunderstandings or biasing effects of different questions and procedures, it is vital for a well-designed survey plan to include provision for a pretest. This is usually a small-scale pilot study to test the feasibility of the intended techniques or to perfect the questionnaire concepts and wording.

Failure to follow up nonrespondents can ruin an otherwise well-designed survey, for it is not uncommon for the initial response rate to most surveys to be under 50 percent. Plans must include returning to sample households where no one was home, attempting to persuade persons who are inclined to refuse and, in the case of mail surveys, contacting all or a subsample of the nonrespondents by telephone or personal visit to obtain a completed questionnaire. A low response rate does more damage in rendering a survey's results questionable than a small sample, since there is no valid way of scientifically inferring the characteristics of the population represented by the nonrespondents.

Quality control, in the sense of checking the different facets of a survey, enters in at all stages—checking sample selection, verifying interviews and checking the editing and coding of the responses, among other things. In particular, sloppy execution of the survey in the field can seriously damage the results. Without proper quality control, errors can occur with disastrous results, such as selecting or visiting the wrong household, failing to ask questions properly, or recording the incorrect answer. Insisting on proper standards in recruitment and training of interviewers helps a great deal, but equally important is proper review, verification and other quality control measures to ensure that the execution of a survey corresponds to its design.

Using the Results of a Survey

How Good is the Survey?

The statistics derived from a survey will rarely correspond exactly with the unknown truth. (Whether "true" values always exist is not important in the present context. For fairly simple measurements—the average age of the population, the amount of livestock on farms, etc.—the concept of a true value is fairly straightforward. Whether true values exist for measurements of such items as attitudes toward political candidates, I.Q.'s, etc., is a more complex matter.)

Fortunately, the value of a statistic does not depend on its being exactly true. To be useful, a statistic need not be exact, but it does need to be sufficiently reliable to serve the particular needs. No overall criterion of reliability applies to all surveys since the margin of error that can be tolerated in a study depends on the actions or recommendations that will be influenced by the data. For example, economists examining unemployment rates consider a change of 0.2 percent as having an important bearing on the United States economy. Consequently, in the official United States surveys used to estimate unemployment, an attempt is made to keep the margin of error below 0.2 percent. Conversely, there are occasions when a high error rate is acceptable. Sometimes a city will conduct a survey to measure housing vacancies to determine if there is a tight housing supply. If the true vacancy rate is very low, say one percent, survey results that show double this percentage will not do any harm; any results in the range of zero to two or

three percent will lead to the same conclusion—a tight housing market.

In many situations the tolerable error will depend on the kind of result expected. For example, during presidential elections the major television networks obtain data on election night from a sample of election precincts, in order to predict the election results early in the evening. In a state in which a large difference is expected (pre-election polls may indicate that one candidate leads by a substantial majority and is likely to receive 60 percent of the vote), even with an error of five or six percent it would still be possible to predict the winner with a high probability of being correct. A relatively small sample size may be adequate in such a state. However, much more precise estimates are required in states where the two candidates are fairly evenly matched and where, say, a 52-48 percent vote is expected.

Thus, no general rule can be laid down to determine the reliability that would apply to all surveys. It is necessary to consider the purpose of the particular study, how the data will be used, and the effect of errors of various sizes on the action taken based on the survey results. These factors will affect the sample size, the design of the questionnaire, the effort put into training and supervising the interview staff, and so on. Estimates of error also need to be considered in analyzing and interpreting the results of the survey.

Sources of Errors

In evaluating the accuracy of a survey, it is convenient to distinguish two sources of errors: 1. sampling errors, and 2. nonsampling errors, including the effect of refusals and not-at-homes, respondents providing incorrect information, coding or other processing errors, and clerical errors in sampling.

Sampling errors

Good survey practice includes calculation of sampling errors, which is possible if probability methods are used in selecting the sample. Furthermore, information on sampling errors should be made readily available to all users of the statistics. If the survey results are published, data on sampling errors should be included in the publication. If information is disseminated in other ways, other means of informing the public are necessary. Thus, it is not uncommon to hear television newscasters report on the size of sampling errors as

part of the results of some polling activity.

There are a number of ways of describing and presenting data on sampling errors so that users can take them into account. For example, in a survey designed to produce only a few statistics (such as the votes that the candidates for a particular office are expected to get), the results could be stated that Candidate A's votes are estimated at 57 percent with the error unlikely to be more than 3 percent, so that this candidate's votes are expected to fall in the range of 54-60 percent. Other examples can be found in most publications of the principal statistical agencies of the United States Government, such as the Bureau of the Census.

Nonsampling errors

Unfortunately, unlike sampling errors, there is no simple and direct method of estimating the size of nonsampling errors. In most surveys, it is not practical to measure the possible effect on the statistics of the various potential sources of error. However, in the past 30 or 40 years, there has been a considerable amount of research on the kinds of errors that are likely to arise in different kinds of surveys. By examining the procedures and operations of a specific survey, experienced survey statisticians will frequently be able to assess its quality. Rarely will this produce actual error ranges, as for sampling errors. In most cases, the analyst can only state that, for example, the errors are probably relatively small and will not affect most conclusions drawn from the survey, or that the errors may be fairly large and inferences are to be made with caution.

Nonsampling errors can be classified into two groups—random types or errors whose effects approximately cancel out if fairly large samples are used, and biases which tend to create errors in the same direction and thus cumulate over the entire sample. With large samples, the possible biases are the principal causes for concern about the quality of a survey.

Biases can arise from any aspect of the survey operation. Some of the main contributing causes of bias are:

1. *Sampling operations.* There may be errors in sample selection, or part of the population may be omitted from the sampling frame, or weights to compensate for disproportionate sampling rates may be omitted.

2. *Noninterviews.* Information is generally obtained for only part of the sample. Frequently there are differences between the non-

interview population and those interviewed.

3. *Adequacy of respondent.* Sometimes respondents cannot be interviewed and information is obtained about them from others, but the "proxy" respondent is not always as knowledgeable about the facts.

4. *Understanding the concepts.* Some respondents may not understand what is wanted.

5. *Lack of knowledge.* Respondents in some cases do not know the information requested, or do not try to obtain the correct information.

6. *Concealment of the truth.* Out of fear or suspicion of the survey, respondents may conceal the truth. In some instances, this concealment may reflect a respondent's desire to answer in a way that is socially acceptable, such as indicating that s(he) is carrying out an energy conservation program when this is not actually so.

7. *Loaded questions.* The question may be worded to influence the respondents to answer in a specific (not necessarily correct) way.

8. *Processing errors.* These can include coding errors, data keying, computer programming errors, etc.

9. *Conceptual problems.* There may be differences between what is desired and what the survey actually covers. For example, the population or the time period may not be the one for which information is needed, but had to be used to meet a deadline.

10. *Interviewer errors.* Interviewers may misread the question or twist the answers in their own words and thereby introduce bias.

Obviously, each survey is not necessarily subject to all these sources of error. However, a good survey statistician will explore all of these possibilities. It is considered good practice to report on the percent of the sample that could not be interviewed, and as many of the other factors listed as practicable.

ance is made for the expense involved in the different tasks, as well as for quality checks at all stages of the work. Thus, among the factors that enter into an expense budget are the following:

1. Staff time for planning the study and steering it through the various stages.
2. Labor and material costs for pretesting the questionnaire and field procedures.
3. Supervisory costs for interviewer hiring, training and supervision.
4. Interviewer labor costs and travel expense (and meals and lodging, if out-of-town).
5. Labor and expense costs of checking a certain percentage of the interviews (by reinterviews).
6. Cost of preparing codes for transferring information from the questionnaire.
7. Labor and material costs for editing, coding and keypunching the information from the questionnaire onto computer tape.
8. Cost of spot-checking to assure the quality of the editing, coding and keypunching.
9. Cost of "cleaning" the final data tapes, that is, checking the tapes for inconsistent or impossible answers.
10. Programming costs for preparing tabulations and special analyses of the data.
11. Computer time for the various tabulations and analyses.
12. Labor time and material costs for analysis of the data and report preparation.
13. Telephone charges, postage, reproduction and printing costs.

An integral part of a well-designed survey, both in terms of time and of costs is allowance for quality checks all along the way. For example, checks have to be made that the sample was selected according to specifications, that the interviewers did their work properly, that the information from the questionnaires was coded accurately, that the keypunching was done correctly, and that the computer programs used for data analysis work properly. For these reasons, a good survey does not come cheap, although some are more economical than others. As a rule, surveys made by personal

interview are more expensive than by mail or by telephone; and costs will increase with the complexity of the questionnaire and the amount of analysis to be carried out. Also, surveys that involve more interviews tend to be cheaper on a per interview basis than surveys with fewer interviews. This is particularly so where the sample size is less than about a thousand because "tooling up" is involved for just about any survey, except one that is to be repeated on the same group.

Where to Get More Information

Several professional organizations have memberships heavily involved in survey research. They also frequently have workshops or sessions on surveys as parts of their regional and annual meetings. The principal organizations are the following:

1. The *American Statistical Association* is concerned with survey techniques and with general application of survey data. It has a separate Section on Survey Research Methods which sponsors sessions on surveys at the annual meetings of the association. The many chapters of the association in the various parts of the country also periodically have meetings and workshops on survey methods, and its publications, the *Journal of the American Statistical Association* and the *American Statistician*, carry numerous articles about surveys.

2. The *American Marketing Association* is concerned, among other things, with the application of survey methods to marketing problems. Like the American Statistical Association, it sponsors sessions on survey methods at its annual meetings, and still other sessions are sponsored by its local chapters. Its publications, the *Journal of Marketing* and the *Journal of Marketing Research*, frequently contain articles on surveys.

3. The *American Association for Public Opinion Research* focuses on survey methods as applied to social and media problems. Its journal, the *Public Opinion Quarterly*, regularly carries articles on

survey techniques and on the application of survey methods to political and social problems.

A number of other professional associations in North America place emphasis periodically on survey methods, for example, the Statistical Society of Canada, the American Sociological Association, the American Political Science Association, the Association for Consumer Research, the American Public Health Association, the American Psychological Association, and the Canadian Psychological Association. There are also various business oriented associations such as the Advertising Research Foundation and the American Association of Advertising Agencies, that give attention to survey methods as applied to business. These and other associations publish a number of journals that carry a great deal of material on survey methods.

There are many good books on survey methods written for nontechnical readers. A few of these are:

1. Tanur, Judith, et al., *Statistics: A Guide to the Unknown*. San Francisco: Holden-Day Pub. Co., 1972.
2. Hauser, Philip, *Social Statistics in Use*. New York: Russell Sage Foundation, 1975.
3. Williams, William H., *A Sampler on Sampling*. New York: John Wiley & Sons, 1978.

For further information, contact . . .

Executive Director
American Statistical Association
806 15th Street, N.W.
Washington, D.C. 20005



**DIVISION DE EDUCACION CONTINUA
FACULTAD DE INGENIERIA U.N.A.M.**

FUNDAMENTO DE LAS TECNICAS DE MUESTREO ESTADISTICO

ANEXO DE APUNTES

DR. OCTAVIO A. RASCON CHAVEZ

JUNIO 1981

EJEMPLO

EN UN ESTUDIO DE MERCADOTECNIA SE TOMO 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 DESVIACION ESTANDAR $S = \$8.00$. ENCONTRAR EL 95% DE NIVEL DE CONFIANZA DE LA VARIANCIA

$$\text{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 DESVIACION ESTANDAR ES

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

SI LA MUESTRA ES PEQUEÑA, EN VEZ DE USAR LA DISTRIBUCION NORMAL SE USA LA t DE STUDENT Y, EN TAL CASO, EL INTERVALO DE CONFIANZA ES

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

EN DONDE $S(X)$ ES LA DESVIACION ESTANDAR DE LA MUESTRA

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

Y t_c ES EL VALOR CRITICO CORRESPONDIENTE A UN NIVEL DE SIGNIFICANCIA α , QUE SE OBTIENE DE LA DISTRIBUCION t DE STUDENT.

EN LA ECUACION (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 RAZONES O COCIENTES

SEAN W Y X DOS VARIABLES ALEATORIAS. LA RAZON DE LA PRIMERA RESPECTO A LA SEGUNDA SE ESTIMA COMO:

$$\hat{R} = \frac{\bar{W}}{\bar{X}}$$

SE PUEDE DEMOSTRAR QUE LA VARIANCA DE ESTE ESTIMADOR ES

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

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

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

$$\text{INTERVALO DE CONFIANZA} = \bar{y} \pm t_c \sqrt{\hat{V}(\bar{y})} = 4.9 \pm 2.26 \times 0.359$$

(t_c SE OBTIENE PARA $10 - 1 = 9$ GRADOS DE LIBERTAD)

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

- b. ESTIMAR EL NUMERO TOTAL DE LETRAS EN LOS 100 NOMBRES, U,
TANTO EN FORMA PUNTUAL COMO CON UN INTERVALO DE 90%.

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

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

- c. ESTIMAR EL PORCENTAJE DE PERSONAS CON SEXO FEMENINO Y EL
ERROR ESTANDAR DE LA ESTIMACION

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

$$\hat{G}(\hat{p}) = \sqrt{\frac{N-n}{(n-1)N} \hat{p}\hat{q}} = \sqrt{\frac{100-10}{(10-1)100} 0.4 \times 0.6} = 0.155$$

EL INTERVALO DE CONFIANZA DEL 95% ES

$$I. \text{ C.} = 0.155 \pm 2.26 \times 0.155 = \{1.91, 2.61\}$$

- d. ESTIMAR EL NUMERO DE VOCALES POR CONSONANTE

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

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



**DIVISION DE EDUCACION CONTINUA
FACULTAD DE INGENIERIA U.N.A.M.**

FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO



TABLAS DE CONTINGENCIA

Dr. Octavio A. Rascón Chávez

Junio, 1981

TABLAS DE CONTINGENCIA

CON FRECUENCIA SE DESEA DETERMINAR SI LA CLASIFICACION DE UNA MUESTRA EN TERMINOS DE 2 O MAS CRITERIOS ES TAL QUE PERMITA INFERIR SI ESOS CRITERIOS SON INDEPENDIENTES ENTRE SI.

POR EJEMPLO, UNA MUESTRA DE PERSONAS QUE HAN FALLECIDO SE PUEDE CLASIFICAR DE LA SIGUIENTE MANERA:

	MUERTE POR CANCER DEL PULMON	MUERTE POR OTRAS CAUSAS
FUMADORES	348	3152
NO FUMADORES	82	1418

EN UN CASO COMO ESTE SE PRETENDERIA PROBAR LA HIPOTESIS DE QUE FUMAR Y MORIR POR CANCER DEL PULMON SON CARACTERISTICAS INDEPENDIENTES.

CUANDO LOS DATOS SE CATEGORIZAN DE ESTA MANERA, SE DICE QUE SE FORMA UNA TABLA DE CONTINGENCIA.

SEA UNA MUESTRA DE TAMAÑO n Y QUE EL EXPERIMENTO SE HA DISEÑADO PARA CLASIFICARLA EN DOS CATEGORIAS, UNA CON r NIVELES Y LA OTRA CON c NIVELES.

SEA x_{ij} EL NUMERO (LA FRECUENCIA) DE ELEMENTOS DE LA MUESTRA QUE QUEDAN EN LA CELDA (i,j) .

LA TABLA DE CONTINGENCIA SERÍA

CLASIFICACION 1	CLASIFICACION 2					TOTAL
	1	2	3	c	
1	x ₁₁	x ₁₂	x ₁₃		x _{1c}	x _{1.}
2	x ₂₁	x ₂₂	x ₂₃		x _{2c}	x _{2.}
3	x ₃₁	x ₃₂	x ₃₃		x _{3c}	x _{3.}
.						
.						
r	x _{r1}	x _{r2}	x _{r3}		x _{rc}	x _{r.}
TOTAL	x _{.1}	x _{.2}	x _{.3}		x _{.c}	n

LOS TOTALES POR RENGLON SE DENOTAN CON $x_{i.}$, ES DECIR

$$x_{i.} = \sum_{j=1}^c x_{ij}$$

LOS TOTALES POR COLUMNA SE DENOTAN

$$x_{.j} = \sum_{i=1}^r x_{ij}$$

LA SUMA DE LOS TOTALES POR COLUMNA O POR RENGLON DEBE SER EL TAMAÑO DE LA MUESTRA, ES DECIR

$$\sum_{i=1}^r x_{i.} = \sum_{j=1}^c x_{.j} = n = \sum_{i=1}^r \sum_{j=1}^c x_{ij}$$

EL PROBLEMA DE VERIFICAR SI LAS CATEGORIAS SON INDEPENDIENTES EQUIVALE AL DE VERIFICAR SI LA PROBABILIDAD DE QUE EL ESPÉCIMEN CUMPLA CON ALGÓN NIVEL DE LA CATEGORIA 1 DEPENDE DE EN QUÉ NIVEL DE LA CATEGORIA 2 SE ENCUENTRA. ASÍ, EN EL

EJEMPLO ANTERIOR SE TRATARIA DE VERIFICAR SI LA MUERTE POR CANCER PULMONAR DEPENDE O NO DE SI LA PERSONA ES O NO FUMADORA.

EN INFERENCIA ESTADISTICA SE DEMUESTRA QUE LA ESTADISTICA

$$V = \sum_{i=1}^k \frac{(x_i - nP_i)^2}{nP_i} \quad (1)$$

TIENDE A UNA VARIABLE ALEATORIA CON DISTRIBUCION DE PROBABILIDADES χ^2 CON $k-r-1$ GRADOS DE LIBERTAD CONFORME CRECE n , EN DONDE n ES EL TAMAÑO DE LA MUESTRA, x_i ES LA FRECUENCIA CON QUE SE OBSERVO EL EVENTO i Y P_i ES LA PROBABILIDAD DE OBSERVARLO EN UNA REALIZACION DEL EXPERIMENTO.

EN NUESTRO CASO, SI P_{ij} ES LA PROBABILIDAD DE QUE UN RESULTADO TENGA EL VALOR i DE LA CARACTERISTICA 1 Y EL VALOR j DE LA 2, Y SI LOS DOS METODOS DE CLASIFICACION SON REALMENTE INDEPENDIENTES, ENTONCES:

$$P_{ij} = \omega_i S_j, \quad i = 1, 2, \dots, r; \quad j = 1, 2, \dots, c$$

DONDE ω_i ES LA PROBABILIDAD DE QUE EL ELEMENTO OBSERVADO CAIGA EN EL i -ESIMO NIVEL DE LA CLASIFICACION 1, Y S_j ES LA PROBABILIDAD DE QUE CAIGA EN EL j -ESIMO NIVEL DE LA CLASIFICACION 2.

POR OTRA PARTE, LOS ESTIMADORES DE MAXIMA VEROSIMILITUD DE ω_i Y S_i SON

$$\hat{\omega}_i = \frac{x_{i.}}{n}, \quad \hat{S}_j = \frac{x_{.j}}{n}$$

POR LO TANTO, CON LA EC. (1) SE OBTIENE QUE

$$V = \sum_{i=1}^r \sum_{j=1}^c \frac{(x_{ij} - n \hat{\omega}_i \hat{S}_j)^2}{n \hat{\omega}_i \hat{S}_j} \quad (2)$$

TIENE DISTRIBUCION χ^2 CON $(r-1)(c-1)$ GRADOS DE LIBERTAD PARA n GRANDE. ESTE NUMERO DE GRADOS DE LIBERTAD SE JUSTIFICA DE LA SIGUIENTE MANERA: SE TIENEN $K = rc$ CLASES Y PARA ESTIMAR LAS P_{ij} SE REQUIERE ESTIMAR $r-1$ VALORES DE ω Y $c-1$ VALORES DE S , ES DECIR, SE ESTIMAN $(r-1) + (c-1)$ PARAMETROS; POR LO TANTO LOS GRADOS DE LIBERTAD SON

$$rc - (r-1) - (c-1) - 1 = (r-1)(c-1)$$

EJEMPLO

CON EL FIN DE VERIFICAR SI EL FUMAR Y EL MORIR POR CANCER PULMONAR SON INDEPENDIENTES, SE DISEÑO UN EXPERIMENTO ESTADISTICO QUE CONSISTIO EN SACAR UNA MUESTRA ALEATORIA DE 5000 EXPEDIENTES CLINICOS DE PERSONAS FALLECIDAS EN UNA CADENA DE HOSPITALES, Y CLASIFICARLA EN UNA TABLA DE CONTINGENCIA. EL RESULTADO FUE EL SIGUIENTE:

	MUERTE POR CANCER PULMONAR	MUERTE POR OTRAS CAUSAS	TOTAL $X_{i.}$	$\hat{\omega}_i$
FUMADORES	348	3152	3500	0.7
NO FUMADORES	82	1418	1500	0.3
TOTAL : $X_{.j}$	430	4570	5000	1.0
\hat{S}_j	0.086	0.914	1.000	

PARA REALIZAR LA PRUEBA DE INDEPENDENCIA SE UTILIZA LA EC (2), Y SE DETERMINA EL VALOR CRÍTICO DE χ^2 QUE CORRESPONDA A UN NIVEL DE CONFIANZA PRESTABLECIDO, $1-\alpha$, USANDO $(2-1) \times (2-1) = 1$ GRADOS DE LIBERTAD, YA QUE $r = c = 2$.

$$\hat{w}_1 = \frac{3500}{5000} = 0.7, \hat{w}_2 = \frac{1500}{5000} = 0.3$$

$$\hat{s}_1 = \frac{430}{5000} = 0.086, \hat{s}_2 = \frac{4570}{5000} = 0.914$$

$$\begin{aligned} v = & \frac{[348-5000(0.7)(0.086)]^2}{5000(0.7)(0.086)} + \frac{[3152-5000(0.7)(0.914)]^2}{5000(0.7)(0.914)} + \\ & + \frac{[82-5000(0.3)(0.086)]^2}{5000(0.086)(0.3)} + \frac{[1418-5000(0.3)(0.914)]^2}{5000(0.3)(0.914)} \end{aligned}$$

$$\begin{aligned} v = & \frac{2209}{301.00} + \frac{2209}{3199.00} + \frac{2209}{129.00} + \frac{2209}{1371.00} = \\ & = 7.34 + 0.69 + 17.12 + 1.61 = 26.76 \end{aligned}$$

SI $1-\alpha = 0.99$, ENTONCES

$$(\chi^2)_{0.99,1} = 6.63 < 26.76$$

POR LO QUE SE RECHAZA LA HIPOTESIS DE INDEPENDENCIA.

EJEMPLO

UN CLUB DE PESCA DEPORTIVA ESTA INTERESADO EN SABER SI SE PESCA CADA TIPO DE PESCADO CON LA MISMA FRECUENCIA EN LOS MESES DE JUNIO A SEPTIEMBRE. PARA ELLO SE DISEÑÓ UN EXPERIMENTO CONSISTENTE EN REGISTRAR LA PESCA MENSUAL EN UNO DE LOS BARCOS DE LOS TRES TIPOS DE PECES DE LA ZONA: ABADEJO, PEZ AZUL Y COLA AMARILLA.

LA TABLA DE CONTINGENCIA QUE SE FORMULO FUE LA SIGUIENTE:

	ABADEJOS.	PECES AZULES	COLAS AMARILLAS	TOTAL	$\hat{\omega}$
JUNIO	315	1347	620	2282	0.2611
JULIO	270	1250	514	2034	0.2327
AGOSTO	295	1480	710	2485	0.2843
SEPTIEM BRE	246	1200	494	1940	0.2219
TOTAL	1126	5277	2338	8741	1.0000
\hat{S}	0.1288	0.6037	0.2675	1.0000	

PARA LA PRUEBA DE INDEPENDENCIA CON CONFIABILIDAD $1-\alpha = 0.95$
 Y $(4-1)(3-1) = 6$ GRADOS DE LIBERTAD, SE TIENE QUE $(\chi_c^2)_{0.95,6} =$
 $= 12.6.$

EL VALOR DE LA ESTADISTICA V SE OBTIENE EN LA EC (2)

$$V = \sum_{i=1}^4 \sum_{j=1}^3 \frac{(\chi_{ij} - 8741 \hat{\omega}_i \hat{S}_j)^2}{8741 \hat{\omega}_i \hat{S}_j}$$

$$\text{CON } \hat{\omega}_1 = \frac{2782}{8741} = 0.2611, \quad \hat{\omega}_2 = \frac{2034}{8741} = 0.2327$$

$$\hat{\omega}_3 = \frac{2485}{8741} = 0.2843, \quad \hat{\omega}_4 = \frac{1940}{8741} = 0.2219$$

$$\hat{S}_1 = \frac{1126}{8741} = 0.1288, \quad \hat{S}_2 = \frac{5277}{8741} = 0.6037$$

$$\hat{S}_3 = \frac{2338}{8741} = 0.2675$$

8741	$\hat{\omega}_1$	\hat{S}_1	=	8741	(0.2611)	(0.1288)	=	293.957
8741	$\hat{\omega}_1$	\hat{S}_2	=	8741	(0.2611)	(0.6037)	=	1377.809
8741	$\hat{\omega}_1$	\hat{S}_3	=	8741	(0.2611)	(0.2675)	=	610.509
8741	$\hat{\omega}_2$	\hat{S}_1	=	8741	(0.2327)	(0.1288)	=	261.983
8741	$\hat{\omega}_2$	\hat{S}_2	=	8741	(0.2327)	(0.6037)	=	1227.944
8741	$\hat{\omega}_2$	\hat{S}_3	=	8741	(0.2327)	(0.2675)	=	544.103
8741	$\hat{\omega}_3$	\hat{S}_1	=	8741	(0.2843)	(0.1288)	=	320.077
8741	$\hat{\omega}_3$	\hat{S}_2	=	8741	(0.2843)	(0.6037)	=	1500.235
8741	$\hat{\omega}_3$	\hat{S}_3	=	8741	(0.2843)	(0.2675)	=	664.755
8741	$\hat{\omega}_4$	\hat{S}_1	=	8741	(0.2219)	(0.1288)	=	249.824
8741	$\hat{\omega}_4$	\hat{S}_2	=	8741	(0.2219)	(0.6037)	=	1170.953
8741	$\hat{\omega}_4$	\hat{S}_3	=	8741	(0.2219)	(0.2675)	=	518.850

$$v = \frac{(315-293.957)^2}{293.957} + \frac{(1347-1377.809)^2}{1377.809} + \frac{(620-610.509)^2}{610.509} +$$

$$\frac{(270-261.983)^2}{261.983} + \frac{(1250-1227.944)^2}{1227.944} + \frac{(514-544.103)^2}{544.103} +$$

$$\frac{(295-320.077)^2}{320.077} + \frac{(1480-1500.235)^2}{1500.235} + \frac{(710-664.755)^2}{664.755} +$$

$$\frac{(246-249.824)^2}{249.824} + \frac{(1200-1170.953)^2}{1170.953} + \frac{(494-518.850)^2}{518.850}$$

$$v = 1.506+0.689+0.148+0.245+0.396+1.665+1.965+0.273+3.079$$

$$0.059+0.721+1.190$$

$$v = 11.936 < 12.6$$

POR LO TANTO SE ACEPTA LA HIPOTESIS DE QUE LA CANTIDAD DE PECES ES INDEPENDIENTE DEL MES EN EL PERIODO DE JUNIO A SEPTIEMBRE, CON UN 95% DE NIVEL DE CONFIANZA.

EJEMPLO

CON EL FIN DE VERIFICAR SI LAS VARIABLES REGION GEOGRAFICA, PARTIDO DE AFILIACION Y SEXO SON INDEPENDIENTES, SE DISEÑO UN EXPERIMENTO ESTADISTICO QUE CONSISTIO EN SACAR UNA MUESTRA ALEATORIA DE 1500 PERSONAS Y CLASIFICAR A CADA UNA DE ACUERDO CON ESAS VARIABLES; CON ESTO SE OBTUVO LA SIGUIENTE TABLA DE CONTINGENCIA:

PARTIDO	ESTE		OESTE		TOTAL	\hat{w}_i
	MASCULINO	FEMENINO	MASCULINO	FEMENINO		
DEMOCRATA	183	217	223	227	850	0.5667
REPUBLICANO	196	154	137	113	600	0.4000
OTRO	$\frac{12}{391}$	$\frac{8}{379}$	$\frac{14}{374}$	$\frac{16}{356}$	50	0.0333
TOTAL	=770		=730		1500	
\hat{S}_j	770/1500 = 0.5133		730/1500 = 0.4867			

$$391 + 374 = 765, \quad 379 + 356 = 735, \quad \hat{r}_1 = 765/1500 = 0.51$$

$$\hat{r}_2 = 735/1500 = 0.49, \quad r = 3, \quad c = 2, \quad m = 2$$

$$\text{LA ESTADISTICA } V = \sum_{i=1}^3 \sum_{j=1}^2 \sum_{k=1}^2 \frac{(x_{ijk} - 1500 \hat{w}_i \hat{S}_j \hat{r}_k)^2}{1500 \hat{w}_i \hat{S}_j \hat{r}_k} = 30.88$$

TIENE DISTRIBUCION χ^2 CON $rcm - (r+c+m) + 2 = 7$ GRADOS DE LIBERTAD.

SI $1-\alpha = 95\%$, ENTONCES

$$\chi_{0.95, 7}^2 = 14.1 < 30.88$$

POR LO QUE SE RECHAZA LA HIPOTESIS DE QUE LAS TRES VARIABLES SON INDEPENDIENTES.

FORMULA CORTA PARA TABLAS DE CONTINGENCIA DE 2 x 2

SI DENOTAMOS A LAS FRECUENCIAS DE LA TABLA CON a, b, c Y d,
 O SEA $x_{11} = a$, $x_{12} = b$, $x_{21} = c$ Y $x_{22} = d$, SE PUEDE DEMOSTRAR
 QUE EL VALOR DE LA ESTADISTICA V SE CALCULA CON LA FORMULA

$$V = \frac{(ad - bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

EJEMPLO

CON EL FIN DE VERIFICAR SI DOS FABRICANTES DE TELEVISORES DE COLOR TIENEN IGUAL NIVEL DE CALIDAD SE DISEÑO UN EXPERIMENTO CONSISTENTE EN PREGUNTAR A 412 COMPRADORES DE LAS MISMAS SI SE REQUIRIO DE SERVICIO DE GARANTIA EN LOS DOS PRIMEROS AÑOS DE FUNCIONAMIENTO, CON LO CUAL SE INTEGRO LA SIGUIENTE TABLA DE CONTINGENCIA:

	REQUIRIO SERVICIO	NO REQUIRIO SERVICIO	
FABRICA			TOTAL
A	111 = a	152 = b	273
B	85 = c	54 = d	139
TOTAL	196	216	412

$$v = \frac{[(111)(54) - (152)(85)]^2 412}{(273)(139)(196)(216)} = 15.51$$

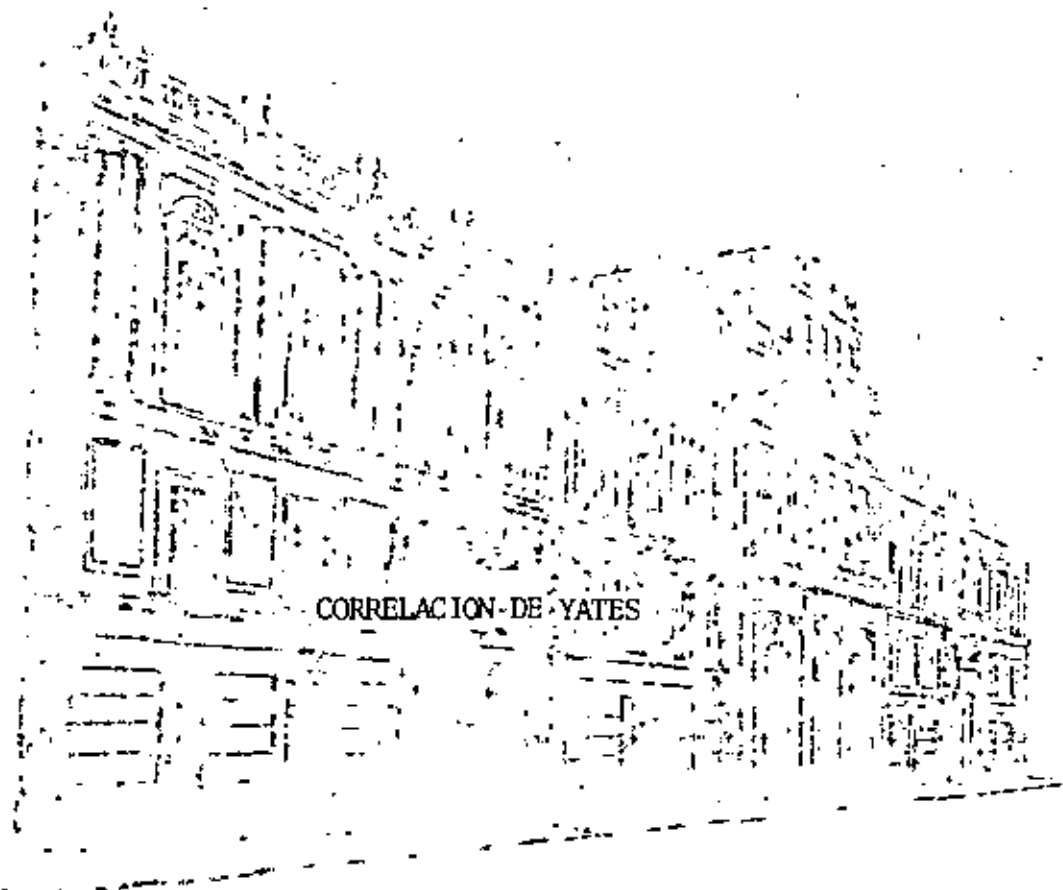
$$x_{0.95,1}^2 = 3.84 < 15.51$$

POR LO QUE SE RECHAZA LA HIPOTESIS DE IGUALDAD DE CALIDAD, A UN 95% DE NIVEL DE CONFIANZA.



**DIVISION DE EDUCACION CONTINUA
FACULTAD DE INGENIERIA U.N.A.M.**

FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO



Dr. Ocatvio A. Rascón Chávez

Junio, 1981

CORRECCION DE YATES

CON EL FIN DE MEJORAR LA APROXIMACION DE LA DISTRIBUCION χ^2 COMO DENSIDAD DE PROBABILIDADES DE LA ESTADISTICA V, CUANDO SE TIENEN POCAS CELDAS EN LA TABLA DE CONTINGENCIA, SE HA PROPUESTO INTRODUCIR UNA CORRECCION A LAS DIFERENCIAS DE LAS FRECUENCIAS OBSERVADAS MENOS LAS ESPERADAS, CONSISTENTE EN SUSTRARLE 0.5 AL VALOR ABSOLUTO DE CADA DIFERENCIA, ES DECIR,

$$V = \sum_{i=1}^r \sum_{j=1}^c \frac{(|x_{ij} - n\hat{w}_i\hat{s}_j| - 0.5)^2}{n\hat{w}_i\hat{s}_j}$$

CON ESTA CORRECCION LA FORMULA CORTA PARA TABLAS DE 2×2 QUEDA EN LA FORMA

$$V = \frac{(|ad - bc| - 0.5n)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

EN EL EJEMPLO INMEDIATO ANTERIOR, AL APLICAR ESTA CORRECCION SE OBTIENE:

$$V = \frac{[|(111)(54) - (162)(85)| - 0.5(412)]^2 412}{(273)(139)(196)(216)} = \frac{(7570)^2 412}{(273)(139)(196)(216)} = 14.69$$

COMPARACION DE DOS TRATAMIENTOS

CUANDO INTERESA VERIFICAR SI DOS PROCEDIMIENTOS DISTINTOS PARA LOGRAR UN MISMO OBJETIVO CONDUCE A RESULTADOS IGUALES, SE DISEÑA UN EXPERIMENTO QUE CONSISTE EN OBTENER UNA MUESTRA ALEATORIA DE LOS RESULTADOS LOGRADOS CON CADA TRATAMIENTO, Y COMPARAR ENTRE SI LAS MEDIAS Y VARIANCIAS CORRESPONDIENTES.

CUANDO LAS OBSERVACIONES SON INDEPENDIENTES, ESTO SE LOGRA MEDIANTE LAS PRUEBAS DE HIPOTESIS DE IGUALDAD DE MEDIAS Y DE VARIANCIAS.

CUANDO NO LO SON, LA COMPARACION SE HACE EN TERMINOS DE LAS DIFERENCIAS INDIVIDUALES DE CADA PAREJA DE RESULTADOS.

AL DISEÑAR EL EXPERIMENTO SE DEBEN CONSIDERAR DOS ALTERNATIVAS:

- a. ASIGNAR AL AZAR A CADA ESPECIMEN EL TRATAMIENTO QUE LE SERA APLICADO; A ESTE PROCESO SE LE LLAMA DE ALEATORIZACION.

EJEMPLO

POR EJEMPLO, SI SE TRATARA DE VERIFICAR SI UN FERTILIZANTE ES MAS EFICIENTE QUE OTRO, UNA VEZ DEFINIDOS LOS LOTES PARA SIEMBRA NOMINALMENTE IGUALES, HABRIA QUE ASIGNAR AL AZAR CADA LOTE A CADA FERTILIZANTE. SUPONGAMOS QUE SE DISPONE DE 11 LOTES Y QUE 5 SE TRATARAN CON EL FERTILIZANTE A Y 6 CON EL B. EL EXPERIMENTO ALEATORIZADO SERIA

LOTE	1	2	3	4	5	6	7	8	9	10	11
FERTILIZANTE	A	A	B	B	A	B	B	B	A	A	B
COSECHA DE TOMATE	29.9	11.4	26.6	23.7	25.3	28.5	14.2	17.9	16.5	21.1	24.3

PÁGINA

Formulas o
REGLAS

DICE

DEBE DECIR -

18

UN PRODUCTO TERMINADO { estado civil
.....
.....
.....

ELIMINARLO.

44

2

$N\bar{y}$

$N\bar{y}$

47

16

$$= \frac{1}{n^2} \left(\frac{n}{N} \sum_{j=1}^N y_j^2 + n(n-1) \sum_{k \neq l} y_k^2 y_l^2 \frac{1}{N} \frac{1}{N-1} \right) - \bar{y}^2$$

$$= \frac{1}{n^2} \left(\frac{n}{N} \sum_{j=1}^N y_j^2 + n(n-1) \sum_{k \neq l} y_k y_l \frac{1}{N} \frac{1}{N-1} \right) - \bar{y}^2$$

47

17

$$+ \frac{n-1}{nN(N-1)} \sum_{k \neq l} y_k y_l - \bar{y}^2$$

$$+ \frac{n-1}{nN(N-1)} \sum_{k \neq l} y_k y_l - \bar{y}^2$$

48

1

$$+ \frac{n-1}{nN(N-1)} \sum_{j=1}^N y_j^2 +$$

$$+ \frac{n-1}{nN(N-1)} \left(\sum_{j=1}^N y_j \right)^2 =$$

48

2

$$+ \frac{n-1}{nN(N-1)} \sum_{k \neq l} y_k y_l - \bar{y}^2$$

ELIMINARLO.

48

9

$$\frac{S^2}{n}$$

$$\frac{S^2}{n}$$

48

10

reucida

reducida.

48

10

factor $(1 - \frac{n}{N})$

factor $(1 - \frac{n}{N})$

50

2

$N\bar{y}$

$N\bar{y}$

50

19

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

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

51

TABLA 3.9

$$V(p) = \frac{NPQ(1-f)}{N-1}$$

$$V(p) = \frac{NPQ(1-f)}{(N-1)n}$$

51

"

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

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

PÁGINA	FORMULA o NÚMERO	DICE	DEBE DECIR
61	10	$V(y) = (1 - \frac{n}{N}) \frac{s^2}{n}$	$V(\bar{y}) = (1 - \frac{n}{N}) \frac{s^2}{n}$
61	12	$V(y)$	$V(\bar{y})$
66	1	precisión	error
67	5	$V = (\frac{d}{t})^2$	$V = (\frac{d}{t})^2$
87	5.7	$V(\hat{Y}_R) = \frac{1-f}{n} \frac{\sum (y_i - R x_i)^2}{n-1}$	$V(\hat{Y}_R) = \frac{1-f}{n} \frac{\sum (y_i - R x_i)^2}{N-1}$
89	5.14	$V(\hat{Y}_d) = \frac{1 - \frac{n_d}{N_d}}{n_d} \frac{\sum (y_{di} - Y_d)^2}{N_d - 1}$	$V(\hat{Y}_d) = \frac{1 - \frac{n_d}{N_d}}{n_d} \frac{\sum (y_{di} - \bar{Y}_d)^2}{N_d - 1}$
111	21	$\bar{y}_2 = \frac{\sum_{i=1}^{n_2} y_{2i}}{n_2}$	$\bar{y}_2 = \frac{\sum_{i=1}^{n_2} y_{2i}}{n_2}$
114	17	$= (\frac{N_1}{N})^2 \hat{V}(\bar{y}_1) + (\frac{N_2}{N})^2 x^2 \hat{V}(\bar{y}_2) + (\frac{N_3}{N})^2 \hat{V}(\bar{y}_3)$	$= (\frac{N_1}{N})^2 \hat{V}(\bar{y}_1) + (\frac{N_2}{N})^2 \hat{V}(\bar{y}_2) + (\frac{N_3}{N})^2 \hat{V}(\bar{y}_3)$
117	6.14	$\hat{V}(p_{est}) = \sum_h \frac{N_h^2 (N_h - n_h) (p_h q_h)}{N^2 (N_h - 1) n_h}$	$V(p_{est}) = \sum_h \frac{N_h^2 (N_h - n_h) (P_h Q_h)}{N^2 (N_h - 1) n_h}$
128	6	$\frac{n_h}{N_h}$	$\frac{n_h}{N_h}$
131	6.26	$\hat{Y}_{RC} = \frac{y_{est}}{x_{est}} X$	$\hat{Y}_{RC} = \frac{\bar{y}_{est}}{\bar{x}_{est}} X$
132	6.29	$\hat{Y}_d = \sum_h \hat{Y}_{hd} = \sum_h \frac{N_h}{n_h} \sum_{i=1}^{n_{hd}} y_{hdi}$	$\hat{Y}_d = \sum_h \hat{Y}_{hd} = \sum_h \frac{N_h}{n_h} \sum_{i=1}^{n_{hd}} y_{hdi}$
133	6.32	$[\sum (y_{hdi} - \bar{y}_{hd})^2 + n_{hd} (1 - \frac{n_{hd}}{n_h}) (\bar{y}_{hd} - Y_d)^2]$	$[\sum (y_{hdi} - \bar{y}_{hd})^2 + n_{hd} (1 - \frac{n_{hd}}{n_h}) (\bar{y}_{hd} - \hat{Y}_d)^2]$

PAGINA	Fórmula o EJEMPLO	DICE	DEBE DECIR
155	7.8	$Y_{ppt} = M \bar{y}_{ppt}$	$\hat{Y}_{ppt} = M \bar{y}_{ppt}$
195	ejer 1.2	iv) $\frac{14}{118} = 0.0153$	iv) $\frac{14}{118} = 0.153$
195	ejer 3.5	error estándar = 0.537	error estándar = 0.398
196	ejer 5.1	el error estándar vale 0.659	el error estándar vale 0.598

FUNDAMENTOS DE LAS TECNICAS DE MUESTREO
ESTADISTICO
1981

Directorio de Asistentes

1. Carlos Gpe. Anaya Sánchez
Alfa Industrias
Sector Bienes de Capital
J.L. Lagrange 103 10º piso
Col. Morales Chapultepec
México 20, D. F.
395 29 44 ext. 262
Moras 125
Col. Florida
México 20, D. F.
524 31 38
2. Ma. Magdalena Bahena Jaimes
SARH
Av. San Bernabé 549
Col. Sn. Jerónimo México 20, D. F.
595 44 53
Calle Cuauhtémoc 148
Col. Coyoacán
México, 21, D. F.
3. Leonel Florencio Bojorquez Landa
Departamento de Pesca
Alvaro Obregón 269
Col. Roma
México 7, D. F.
511 04 84
Sevilla 113-9
Col. Portales
México 13, D. F.
532 53 26
4. José Miguel Casanova Garcés
Instituto Nacional de Cardiología
Juan Badiano 1
Col. Tlalpan
México 22, D. F.
573 29 11
Petén Norte 8-5
Col. Narvarte
México 12, D. F.
530 38 28
5. Jesús Javier Castillo Aveitia
SAHOP
Xola y Av. Universidad
Centro Exop, cuerpo C
México 12, D. F.
530 46 77
Eugenia 1164 - 304
Col. del Valle
México 12, D.
6. Marina Castillo Garduño
SARH
Reforma 107-7º piso
Col. San Rafael
México 4, D. F.
535 0252
Doctor Durán 3
Col. Doctores
México 7, D. F.
7. Guillermo Contreras Ordaz
Teléfonos de México S. A.
Antonio Caso 179 4º piso
México 4, D. F.
546 46 97
And. 32 Ent. 34-1
Col. Acueducto de Guadalupe
México 14, D. F.
392 69 53

- | | | |
|-----|--|---|
| 7. | Lorenzo Eduardo Cuevas Velasco
Proyecto y Estudios Sociales y Asesoría
564 51 44 | Calz. de Tlalpan 2247 Bis-2
Cd. Jardín
México 21, D. F. |
| 8. | Carlos Alfonso del Rio Echeverría
Departamento de Pesca
Alvaro Obregón 269
México 7, D. F.
514 02 65 | Av. Sta. Lucia Manz. 15-7
Col. Olivar del Conde
México 19, D. F.
651 42 83 |
| 9. | Socorro Eugenia Del Villar Barrera | Miguel Laurent 1202-4
Col. Letrán Valle
México 12, D. F.
575 02 00 |
| 10. | Lilia H. Durén Salguero
Departamento de Pesca
Av. Alvaro Obregón 286 5º piso
México 7, D. F.
286 29 54 | Rumanía 305 - 8
Col. Potaless
México 13, D. F.
532 25 35 |
| 11. | Juan Guillero Figueroa Perea
Jef. de Servicios de Planificación Familiar
I M S S
Paseo de la Reforma 506 - 12º piso
México 6, D. F.
285 33 37 | Sabino 66
Sta. Ma. la Rivera
México 4, D. F.
541 07 12 |
| 12. | Arturo Godínez Marínes
UAP
4 Sur 104
Puebla, Pue. | Hacienda de Balvanera 5558
Col. Hacienda
Puebla, Pue. 45 30 74 |
| 13. | Joel González Cabral | |
| 14. | Rafael González Ezeta
IMSS
Río Blanco 6
Col. Magdalena de las Salinas
México 16, D. F.
754 40 99 | Pallares Porcillo 83-15
Col. Coyoacán
México 21, D. F.
544 74 17 |
| 15. | José Manuel Grande Vidal
Departamento de Pesca
Alvaro Obregón 269 10º piso
México 7, D. F.
514 04 00 | Dr. Lucio 102 - 602
Col. Doctores
México 7, D. F.
588 38 87 |
| 16. | Antonio Guerrero Cruz
Dir. Gral. de Geografía del Territorio Nal.
San Antonio Abad 124
México 8, D. F.
578 62 00 ext. 173 | Calle Elena 6 - 2
Col. Nativitas
México 13, D. F.
579 76 71 |

17. Patricia Guzmán Amaya
Departamento de Pesca
Alvaro Obregón 286
México 7, D. F.
286 29 54
Canarias 1203
Col. Portales
México 9, D. F.
18. José Jorge Herrera Meléndez
DGGTN
San Antonio Abad 124 3er. piso
Col. Tránsito
México 5, D. F.
761 22 85
Rio Lerma 303-102
Col. Cuahémoc
México 5, D. F.
19. José Ramiro Huacuz Díaz
IMSS
Rio Blanco # 6
Col. Magdalena de las Salinas
754 40 94
Oriente 108-3152
Col. r. Millán
México 14, D. F.
657 56 61
20. Ernesto Alejandro Assais Gómez
SARH
Fco. Villa Carretera A Torreón Km 5
Durango, Dgo.
1 56 91
Patoni 100C Sur Int. 5
Durango, Dgo.
2 18 76
21. Ma. del Pilar Leal Hernández
SARH
Av. San Bernabé 549
Col. San Jerónimo L.
México 20, D. F.
595 23 22
Calle 1521 - 25
Col. El Parque
México 8, D. F.
552 08 49
22. Regulo López Guzman
Depto. de Pesca
Alvaro Obregón 269 10º piso
México 7, D. F.
514 04 00
Tehuantepec 134 - 1
Col. Roma Sur
México 7, D. F.
23. Fco. Javier Magallón Barajas
Departamento de Pesca
Alvaro Obregón 269 10º piso
México 7, D. F.
525 49 60 ext. 357
Planta Xia 30
Tlalnepantla, Edo. de México
397 60 85
24. José Luis Martínez Medina
SARH
Plaza de la República 31 - 8º piso
México 1, D. F.
546 06 55
Valladolid 22 - 1
México 7, D. F.
25. José Luis Medina Calvo
IMSS
Toledo 21 3er. piso
Col. Juárez
México 6, D. F.
525 46 80
Tuxpan 80-201
México 7, D. F.
584 67 82

26. Martín Minor Hernández
UAP
Calle 4 Sur No. 104
Puebla, Pue.
Rio Papagayo 5906
Col. Hacienda
Puebla, Pue.
45 61 06
27. Rafael Angel Morales Martínez
SARH
Av. San Bernabé 549
Col. San Jerónimo
México 20, D. F.
594 44 53
Poemas Rústicos 5 - 2
U. Independencia
México 20, D. F.
595 22 75
28. Gloria Irene Parra Salazar
SAHOP
Xola y Av. Universidad
Centro SCOP cuerpo D Méx. 12, D.F.
530 46 77
Calle 8-10
Eje Satélite Viveros del Valle
Edo. de México
29. Salvador Preciat Lambarri
SARH
Reforma, 107 - 2º piso Méx 4, D. F.
566 95 59
Palenque 673
Col. Vertiz Narvarte
México 13, D. F.
559 06 08
30. Norberto Dardo Rocha Panizza
Centro Nal. de Investigaciones Agrarias
Azores 527
Col. Portales
México 13, D. F.
539 76 61 672 73.03
Av. Universidad 1900 Ed. 33-301
Col. Copilco Universidad
México 20, D. F.
548 38 91
31. Ma. Concepción Rodríguez Ramírez
Departamento de Pesca
Av. Alvaro Obregón 268 10º piso
México 7, D. F.
514 02 44
Amsterdan 148 - B
Col. Condesa
México 11, D. F.
286 13 57
32. Indiana Salamanca Castillo
Ministerio de Salud Pública
Managua, Nicaragua
3a. privada de Amores No. 21
Col. Del Valle
México 12, D. F.
523 24 45
33. Amira Solano Azar
SARH
Av. Paseo de la Reforma 107 - 2º piso
Col. Juárez
México 21, D. F.
566 95 59
Centenario 174
Col. Coyoacán
México 21, D. F.
554 59 27
34. Gerardo Sampayo Luna
Cia. de Luz y Fuerza del Centro S. A.
Melchor Ocampo 171
Col. Anahuac México 17, D. F.
539 96 51
Oaxaca 34
Col. Valle Ceylán
Tlalnepantla, Edo. de México
390 09 48

35. Ernesto Soto Morán
SAHOP
Xola y Av. Universidad
México 12, D. F.
519 73 46
Av. Universidad 2074 (9-2)
Col. Copilco
México 20, D. F.
548 40 54
36. Abner Uribe Jiménez
Centro de Información
Dir. Gral. de Operación Hidráulica
San Antonio Abad 231
Col. Obrera México 8, D. F.
588 23 70
588 23 80
Valle del Nilo 26 - 3
Col. Valle de Aragón
795 21 67
37. Rubén Emiliano Viveros Vergara
IMSS
Toledo 21 3er. piso
Col. Juarez
México 6, D. F.
514 84 45
Tuxpan 80 - 201
Col. Roma
México 7, D. F.
584 67 82
38. Myrna Ivonne Wong Ríos
Departamento de Pesca
Alvaro Obregón 286 5º piso
México 7 D. F.
286 29 54
Dr. Lucio 102-602 Edif. -A 2
Col. Doctores
México 7, D. F.
588 38 87