FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO

del 23 de agosto al 20 de octubre, 1977.

Fecha	Duración			Profesor
Agosto 23	18 a 21 h	I	INTRODUCCION	M. en C. Alejandro Servín Andrade
Ago. 25 y 30; Septiembre 6	18 a 21 h c/d	II	MUESTREO ALEATORIO SIMPLE	Dr. Octavio A. Rascón Chávez
Sep. 8, 13 y 20	18 a 21 h c/d	III	TAMAÑO DE LA MUESTRA	M. en I. Augusto Villarreal Aranda
Sep. 22, 27 y 29	18 a 21 h c/d	IV	MUESTREO ALEATORIO SIMPLE PARA RAZONES O COCIENTES	M. en C. Adela Abad de Servín
Oct. 4, 6, 11	18 a 21 h c/d	V	MUESTREO ESTRATIFICADO	M. en C. Alejandro Servín Andrade
Oct. 13 y 18	18 a 21 h c/d	VI	MUESTREO POR CONGLOMERADOS	M. en C. Alejandro Servín Andrade
Oct. 20	18 a 21 h c/d	VII	MUESTREO SISTEMATICO	M. en C. Adela Abad de Servín

18, VIII. 77.

'edcs.

DIRECTORIO DE PROFESORES DEL CURSO

FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO

M. EN C. ADELA ABADE DE SERVIN
Profesora
Escuela Nacional de Estudios Profesionales
Acatlán, Edo. de México.
Tel.:

DR. OCTAVIO A. RASCON CHAVEZ
Jefe de la División de Estudios Superiores
Facultad de Ingeniería, UNAM
México 20, d.f.
Tel.: 548.09.50

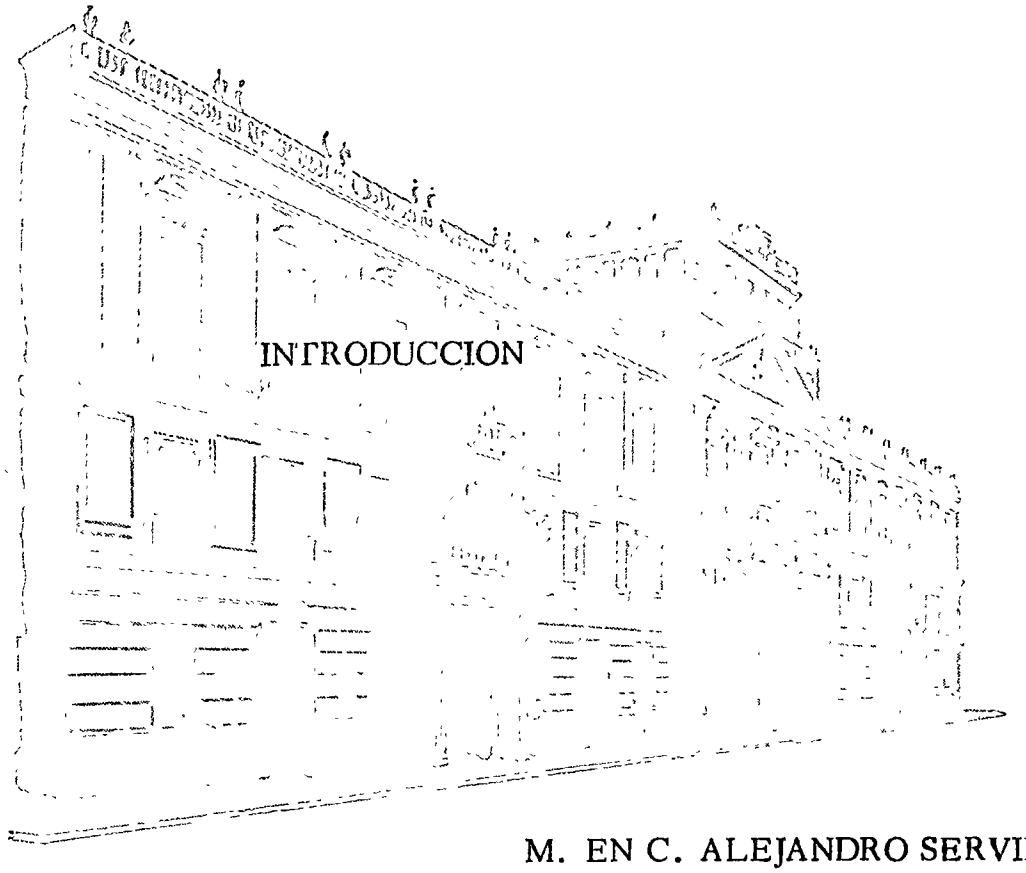
M. EN C. ALEJANDRO SERVIN ANDRADE
Asesor de la Jefatura de Sistematización
Toledo 21-6°
México 6, D.F.
Tel.: 525.46.80 y 525.03.64 E.194 y 109

M. EN I. AGUSTO VILLARREAL ARANDA
Secretario Académico de la División de
Estudios Superiores de la Facultad de Ingeniería
UNAM
México 20, D.F.
Tel.: 548.09.50

'edcs. 19, VIII.77.

centro de educación continua
división de estudios superiores
facultad de ingeniería, unam

FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO



INTRODUCCION

M. EN C. ALEJANDRO SERVIN
M. EN C. ADELA ABADE DE S.

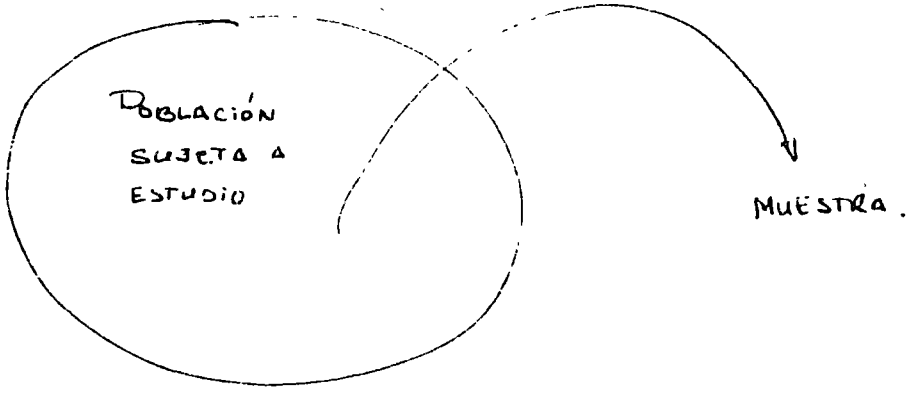
AGOSTO-OCTUBRE, 1977.

TECNICAS BASICAS DE MUESTREO

INTRODUCCION

Los métodos existentes para derivar una conclusión para toda una población, a partir del conocimiento que se tiene sobre una fracción de ella, denominada muestra, se clasifican en métodos o muestreos probabilísticos y no probabilísticos. En estos últimos, los nó probabilísticos, no puede conocerse o enunciarse ningún juicio sobre el error cometido en base al conocimiento que se tiene sobre la muestra. En oposición a esto, en el muestreo probabilístico pueden enunciarse juicios estadísticos sobre el error cometido. Este tipo de muestreo trajaba de la manera siguiente:

A partir del investigador o usuario de la información y con la consideración de la información disponible, así como de los recursos existentes se conforma la población que estará sujeta a estudio y de ella se elige una fracción denominada muestra, en base al conocimiento que se tiene sobre la muestra, se derivan inferencias o conclusiones sobre la población muestreada, las cuales son estadísticamente válidas.



Una encusta siempre queda estructurada a partir del propósito del usuario de la información; es decir se hace un diseño, se elaboran cuestionarios, se colecta información y se procesa según un propósito específico. El del usuario de la información.

El desarrollo de una encuesta probabilística requiere que el técnico cumpla o siga determinadas instrucciones o reglas. En particular, requiere que las unidades que son seleccionadas para la muestra hayan sido elegidas por el azar y no por la persona. En este sentido son selecciones inválidas aquellas como las siguientes: me cubrí los ojos y dirigí la mano a la lista para elegir a cinco personas; abrí el archivero y tomé aquellos expedientes que supuse adolecían del defecto buscado.

Para asegurarse que la muestra seleccionada es aleatoria o al azar, el técnico usa las tablas de números aleatorios. Estas son construídas de tal manera que su utilización nos asegura aleatoriedad en la selección, la cual es un requerimiento del muestreo probabilístico: la selección debe ser aleatoria.

Hacer la selección para obtener la muestra, requiere necesariamente de la existencia de una lista o algo equivalente que nos permita, digamos, numerar los elementos o unidades de la población (marco de muestreo). Esto significa que las técnicas de muestreo probabilístico se enuncian y se aplican a poblaciones finitas, es decir, que constan de un número determinado de personas, cosas o animales. Algunos ejemplos de poblaciones sujetas a estudio son las siguientes:

Conjunto de empleados en una compañía.

Conjunto de estudiantes en una escuela.

Conjunto de enfermos en un hospital.

Conjunto de expedientes médicos en un hospital.

Conjunto de ganado vacuno en una región geográfica.

etc

Por ende el muestreo probabilístico es usado ampliamente en las diferentes actividades económicas de los países. Se les usa en agricultura, en ganadería, en silvicultura, caza y pesca; en industrias, comercios y en servicios. Se les usa para estimar porcentajes como son: porcentaje de familias con más de cinco hijos y el porcentaje de empleados en una institución que hacen uso de los servicios de guardería; para estimar valores medios como el número medio de hijos por familia o el ingreso medio por empleado; para estimar totales como el total de ganado caballar, existente en una región o el total de artículos defectuosos en un lote producido en una empresa; y por último se le usa para estimar cocientes o razones del tipo: número de mujeres a número total de habitantes en una cierta ciudad.

En el desarrollo de una encuesta probabilística se presentan diferentes etapas o actividades como son:

Descripción de objetivos

Diseño muestral

Selección de la muestra

Trabajo de campo

Procesamiento de la información

Interpretación de los resultados

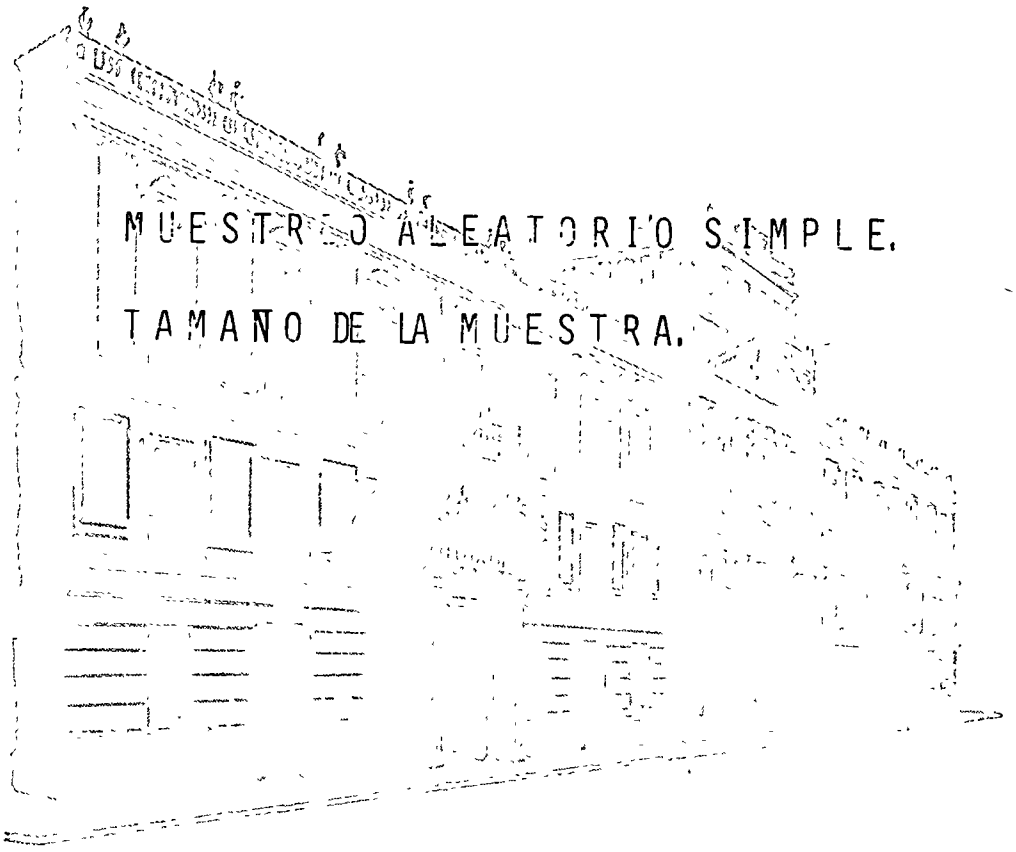
Cada una de estas actividades son importantes en si mismas, para el buen éxito de una encuesta, sin embargo, dada la amplitud del tema, en este curso nos restringimos al diseño muestral, es decir, a la especificación de maneras para efectuar la selección y de maneras para efectuar la estimación.

CONCEPTO DE VARIABILIDAD: En muestreo es de gran importancia conocer o tener idea de la variabilidad de la característica en estudio, por ejemplo: número de hijos por familia; algunas familias tienen cero hijos, otras 1 ó más de uno, digamos hasta 14. En este ejemplo podemos decir que el número de hijos en las familias que tienen a lo más 5 años constituidas como unidad familiar es menos variable que el número de hijos en cualquier familia.

Usualmente para referirnos a la variabilidad de una característica lo haremos a través de la varianza o del error estándar.

centro de educación continua
división de estudios superiores
facultad de ingeniería, unam

FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO



MUESTREO ALEATORIO SIMPLE.

TAMAÑO DE LA MUESTRA.

DR. OCTAVIO A. RASCON.

M EN I AUGUSTO VILLARREAL ARANDA.

AGOSTO DE 1977.

EXPERIMENTO. PARA FINES DE ESTE CURSO, SE ENTENDERA POR EXPERIMENTO A TODO PROCESO DE OBSERVACION. ASI UN EXPERIMENTO PUEDE SER PLANEADO Y REALIZADO POR EL HOMBRE, O PUEDE SER EFECTUADO POR LA NATURALEZA EN CASO DE UN FENOMENO NATURAL. POR EJEMPLO, EL LANZAR UNA MONEDA O UN DADO Y OBSERVAR LA CARA QUE QUEDA HACIA ARRIBA, ES UN EXPERIMENTO PLANEADO Y REALIZADO POR EL HOMBRE. EL OBSERVAR LA CANTIDAD DE AGUA QUE LLUEVE ANUALMENTE EN UNA CIUDAD, ES UN EXPERIMENTO ASOCIADO A UN FENOMENO NATURAL.

A LOS RESULTADOS DE UN EXPERIMENTO SE LES DENOMINA *DATOS*. A UN GRUPO DE DATOS SE LE LLAMA *MUESTRA*.

PROBABILIDAD: ES UNA MEDIDA DE LA CERTIDUMBRE QUE SE LE ASOCIA A LA OCURRENCIA U OBSERVACION DE UN RESULTADO DETERMINADO, AL REALIZARSE EL EXPERIMENTO CORRESPONDIENTE.

LA *TEORIA DE PROBABILIDADES* ES UNA RAMA DE LAS MATEMATICAS APLICADAS QUE TRATA LO CONCERNIENTE A LA ASIGNACION Y MANEJO DE PROBABILIDADES.

ESTADISTICA: ES LA RAMA DE LAS MATEMATICAS QUE SE ENCARGA DE ENSEÑAR LAS REGLAS PARA COLECTAR, PRESENTAR Y PROCESAR LOS DATOS OBTENIDOS AL REALIZAR VARIAS VECES EL EXPERIMENTO ASOCIADO A UN FENOMENO DE INTERES. PROPORCIONA, ADEMAS, LOS METODOS PARA EL DISEÑO DE EXPERIMENTOS Y PARA TOMAR DECISIONES CUANDO APARECEN SITUACIONES DE INCERTIDUMBRE.

ESTADISTICA

* *DESCRIPTIVA*.- TRATA LO CONCERNIENTE A LA OBTENCION, ORGANIZACION, PROCESAMIENTO Y PRESENTACION DE LOS DATOS.

* *INFERENCIAL*.- TRATA LO CONCERNIENTE A LOS METODOS PARA INFERIR CONCLUSIONES ACERCA DEL FENOMENO DEL CUAL PROVIENEN LOS DATOS

SIMBOLOS O DESIGUALDADES:

$<$ menor que

$\leq$ menor o igual que

$>$ mayor que

$\geq$ mayor o igual que

$\neq$ diferente de

TEORIA DE CONJUNTOS

UN CONJUNTO ES UNA COLECCION BIEN DEFINIDA DE OBJETOS

NOTACION. LOS CONJUNTOS SE DENOTAN USUALMENTE CON LETRAS MAYUSCULAS, Y SUS ELEMENTOS SE ANOTAN DENTRO DE UN PAR DE LLAVES.

EJEMPLOS

A) EL CONJUNTO DE NUMEROS ANOTADOS EN UN DADO ES

$S = \{1, 2, 3, 4, 5, 6\}$

B) EL CONJUNTO DE LOS NUMEROS ENTEROS MENORES QUE 5 ES

$S = \{-\infty, \ldots, -3, -2, -1, 0, 1, 2, 3, 4\}$

o $S = \{x: x \text{ ES ENTERO Y } x \leq 4\}$

C) EL CONJUNTO DE LOS NUMEROS ENTEROS POSITIVOS MENORES QUE 5 ES

$E = \{0, 1, 2, 3, 4\}$

$E = \{x: \text{ES ENTERO Y } 0 \leq x \leq 4\}$

D) EL CONJUNTO DE LOS CONTINENTES ES

$C = \{\text{ASIA, EUROPA, AMERICA, AFRICA, OCEANIA}\}$

E) EL CONJUNTO DE MANERAS QUE TIENE UNA MONEDA ES

$M = \{\text{CARA, CRUZ}\}$

F) EL CONJUNTO DE NUMEROS MAYORES DE 5 PERO MENORES O IGUALES QUE 10

$S_1 = \{x: 5 < x \leq 10\}$

CONJUNTOS

- *FINITOS.-* CUANDO TIENEN UN NUMERO FINITO DE ELEMENTOS
- *INFINITOS.-* CUANDO TIENEN UN NUMERO INFINITO DE ELEMENTOS

PARA EXPRESAR QUE UN ELEMENTO *PERTENECE* A UN CONJUNTO SE USA EL SIMBOLO ε. PARA EXPRESAR QUE *NO PERTENECE* SE USA EL SIMBOLO $\notin$.

EJEMPLO

SI $S_1 = \{X: 5<X\leq 10\}$, ENTONCES:

$3 \notin S_1$; $5 \notin S_1$; $8 \varepsilon S_1$; $10 \varepsilon S_1$.

PARA EXPRESAR QUE UN CONJUNTO *ESTA CONTENIDO* EN OTRO SE USA EL SIMBOLO C; SI *NO ESTA CONTENIDO* SE USA EL SIMBOLO $\not C$.

PARA QUE UN CONJUNTO ESTE CONTENIDO EN OTRO SE REQUIERE QUE *TODOS* SUS ELEMENTOS LO ESTEN, ES DECIR, QUE TODOS SUS ELEMENTOS PERTENEZCAN A AMBOS CONJUNTOS.

EJEMPLO

SEAN $E=\{3,5\}$; $F=\{3,8\}$; $G=\{7,9\}$. $E \not C S_1$; $F \not C S_1$; $G C S_1$

SI UN CONJUNTO, B, ESTA CONTENIDO EN OTRO, S, SE DICE QUE B ES *SUBCONJUNTO* DE S.

EJEMPLO

$B = \{X: 3\leq X\leq 8\}$ Y $S_1=\{X: 5<X\leq 10\}$.

EN ESTE CASO:

$A C S_1$ ⇒A ES SUBCONJUNTO DE S_1

$B \not C S_1$ ⇒B NO ES SUBCONJUNTO DE S_1

SE DICE QUE DOS CONJUNTOS SON IGUALES CUANDO CONTIENEN LOS MISMOS ELEMENTOS (NO IMPORTA EL ORDEN EN QUE ESTOS SE ESCRIBAN)

EJEMPLO

SEAN A={1,3,5,7}, B={7,5,1,3} Y C={7,5,1}

EN TAL CASO, A = B≠C

DE LA MISMA MANERA QUE EXISTE EL CERO EN LOS NUMEROS, EN LA TEORIA DE CONJUNTOS EXISTE EL *CONJUNTO VACIO*, EL CUAL NO TIENE ELEMENTOS. USUALMENTE SE DENOTA ∅.

EJEMPLO

¿CUAL ES EL CONJUNTO DE ELEMENTOS, X, TALES QUE 2X=7 Y X ES ENTERO?

SOLUCION :- ES EL CONJUNTO VACIO, ∅.

A ∅ SE LE CONSIDERA COMO SUBCONJUNTO DE CUALQUIER CONJUNTO. ASI, POR EJEM, TODOS LOS SUBCONJUNTOS DEL CONJUNTO

S ={2,5,10} SON: {2};{5};{10};{2,5} ;{2,10};{5,10};{2,5,10} Y ∅.

ESPACIO DE EVENTOS

ASOCIADO A UN EXPERIMENTO SIEMPRE HAY UN CONJUNTO DE RESULTADOS POSIBLES; A DICHO CONJUNTO SE LE LLAMA *ESPACIO DE EVENTOS.*

EJEMPLOS

EL ESPACIO DE EVENTOS ASOCIADO AL EXPERIMENTO DE LANZAR UN DADO Y ANOTAR LA CARA QUE QUEDA HACIA ARRIBA ES

S ={1,2,3,4,5,6}

EL ESPACIO DE EVENTOS CORRESPONDIENTE AL EXPERIMENTO DE LANZAR DOS DADOS Y ANOTAR LOS NUMEROS QUE QUEDAN HACIA ARRIBA ES

$$S=\left\{\begin{array}{l}(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)\\(2,1),(2,2),(2,3),(2,4),(2,5),(2,6)\\(3,1),(3,2),(3,3),(3,4),(3,5),(3,6)\\(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)\\(5,1),(5,2),(5,3),(5,4),(5,5),(5,6)\\(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\end{array}\right\}$$

SI EN ESTE EXPERIMENTO LA OBSERVACION DE INTERES FUESE LA SUMA DE LOS DOS NUMEROS OBSERVADOS, ENTONCES EL ESPACIO DE EVENTOS DEL EXPERIMENTO SERIA

$S= \{2,3,4,5,6,7,8,9,10,11,12\}$

. TODO SUBCONJUNTO DE UN ESPACIO DE EVENTOS SE LE LLAMA *EVENTO*. A LOS EVENTOS QUE TIENEN UN SOLO ELEMENTO DEL ESPACIO SE LES LLAMA *EVENTOS SIMPLES*.

SI AL REALIZAR UN EXPERIMENTO SE OBSERVA UN ELEMENTO DEL EVENTO A, ENTONCES SE DICE QUE *OCURRIO* O SE *VERIFICO* EL EVENTO A. POR EJEMPLO, SI A={2,4} Y AL LANZAR UN DADO SE OBSERVA EL 2 O 4, SE DICE QUE OCURRIO EL EVENTO A; SI SE OBSERVA CUALQUIER OTRO NUMERO, ENTONCES SE DICE QUE *NO* OCURRIO A.

ESPACIOS DE EVENTOS

- *DISCRETOS*.- SI SUS ELEMENTOS PUEDEN NUMERARSE O CONTARSE. TIENEN UN NUMERO FINITO O INFINITO NUMERABLE DE ELEMENTOS.
- *CONTINUOS*.- SI SUS ELEMENTOS *NO* PUEDEN ENUMERARSE. TIENEN UN NUMERO INFINITO NO NUMERABLE DE ELEMENTOS

EJEMPLO

LOS ESPACIOS DE EVENTOS S_1={CARA, CRUZ}; S_2={1,2,3,4,5,6,}; S_3={VERDE, ROJO} SON DISCRETOS. LOS ESPACIOS DE EVENTOS S_4={X: $-\infty<X\leq 0$}; S_5={Z: $Z\geq 3$}; S_6={Y: $3\leq Y\leq 80$} SON CONTINUOS.

EJEMPLO

¿QUE TIPOS DE ESPACIOS DE EVENTOS CORRESPONDEN A LOS SIGUIENTES EXPERIMENTOS?

A) CONTEO DEL NUMERO DE GRANOS DE UNA MAZORCA DE MAIZ

S={0,1,2,3,...,∞}, ES DISCRETO E INFINITO

B) MEDICION DE LA LONGITUD DE UNA ESPIGA DE TRIGO

S={X: $0<X<\infty$},X EN CM, ES CONTINUO E INFINITO

C) MEDICION DEL EFECTO DE UNA VACUNA, EN TERMINOS DE "EXITO" O "FRACASO"

S={EXITO, FRACASO} ES DISCRETO Y FINITO.

D) MEDICION DEL NUMERO DE MILIGRAMOS DE UN ANTIBIOTICO CONTENIDO EN UNA CAPSULA

S={Y: $0\leq Y<\infty$} X en mg, ES CONTINUO E INFINITO.

COMPLEMENTO DE UN EVENTO

EL *COMPLEMENTO* DE UN EVENTO A ES OTRO EVENTO QUE CONTIENE TODOS LOS ELEMENTOS DEL ESPACIO DE EVENTOS CORRESPONDIENTE QUE NO ESTAN EN A. USUALMENTE SE DENOTA CON UNA TILDE SOBRE EL SIMBOLO QUE CORRESPONDE AL EVENTO QUE COMPLEMENTA.

EJEMPLOS

I S={1,2,3,4,5,6} Y A={1,3,5} ENTONCES $\tilde{A}$={2,4,6}.

S={X: $0\leq X\leq 58$} Y A={X: $3\leq X\leq 17$}, ENTONCES $\tilde{A}$={X: $0\leq X\leq 3$, $17<X\leq 58$}

EVENTOS MUTUAMENTE EXCLUSIVOS

CUANDO DOS O MAS EVENTOS NO PUEDEN OCURRIR SIMULTANEAMENTE AL REALIZAR UNA SOLA VEZ UN EXPERIMENTO, SE DICE QUE ESTOS SON *MUTUAMENTE EXCLUSIVOS*. ES DECIR, DOS EVENTOS SON MUTUAMENTE EXCLUSIVOS CUANDO NO TIENEN NI UN SOLO ELEMENTO EN COMUN.

EJEMPLO

A) CUALQUIER EVENTO Y SU COMPLEMENTO SON MUTUAMENTE EXCLUSIVOS.

B) ¿SON $E=\{Y: 0\leq Y\leq 25\}$ Y $A=\{2,50,100\}$ MUTUAMENTE EXCLUSIVOS?
NO, PORQUE TIENEN EL ELEMENTO 2 EN COMUN.

OPERACIONES CON EVENTOS

LA UNION DE DOS EVENTOS ES OTRO EVENTO CUYOS ELEMENTOS SON TODOS LOS DE AMBOS. LA OPERACION DE UNION SE DENOTA CON EL SIMBOLO U.

EJEMPLOS

A) SI $A=\{2,4,6\}$ Y $B=\{1,6,12\}$, ENTONCES
$C=AUB=\{1,4,6,12,2\}$

B) ¿SON A Y B MUTUAMENTE EXCLUSIVOS? NO PORQUE TIENEN EL 6 EN COMUN.

C) SI $D=\{Y: 0\leq Y\leq 13\}$ Y $E=\{Y: 20\leq Y\leq 50\}$,
ENTONCES
$DUE=\{Y: 0\leq Y\leq 13,\ 20\leq Y\leq 50\}$

D) SI $F=\{Y: 8\leq Y\leq 20\}$, ENTONCES
$DUF=\{Y;\ 0\leq Y\leq 20\}$.

E) SI $G=\{Y: 3\leq Y\leq 10\}$, ENTONCES
$DUG=\{Y: 0\leq Y\leq 13\}=D$; OBSERVESE QUE EN ESTE CASO GCD. EN GENERAL, SI ACB, ENTONCES AUB=B.

EN GENERAL, LA UNION DE VARIOS EVENTOS ES OTRO EVENTO CUYOS ELEMENTOS SON TODOS LOS DE LOS EVENTOS QUE SE UNEN.

EJEMPLO

AUBUF = K = {1,2,4,6, y: 8≤ y ≤20}

LA *INTERSECCION* DE DOS EVENTOS ES EL CONJUNTO DE ELEMENTOS QUE PERTENECEN SIMULTANEAMENTE A AMBOS. PARA DENOTAR LA OPERACION DE INTERSECCION SE USA EL SIMBOLO ∩ .

EJEMPLOS

A) A = {2,3,6} Y B = {2,6,10} ENTONCES A∩B = C = {2,6}

B) D = {Y: 4≤Y≤5}, ENTONCES A∩D = Ø.

OBSERVESE QUE EN ESTE EJEMPLO A Y D SON MUTUAMENTE, EXCLUSIVOS, YA QUE *NO* TIENEN NINGUN ELEMENTO EN COMUN. SIEMPRE QUE DOS EVENTOS SON MUTUAMENTE EXCLUSIVOS, SU INTERSECCION ES EL CONJUNTO VACIO.

EN GENERAL, LA *INTERSECCION* DE VARIOS EVENTOS ES EL CONJUNTO DE ELEMENTOS QUE TODOS ELLOS TIENEN EN COMUN.

EJEMPLO

SI A = {2,3,6,8}; B={2,3,10,100}; C={Y: 0≤Y≤5} Y D={Y: 2≤Y≤4}, ENTONCES

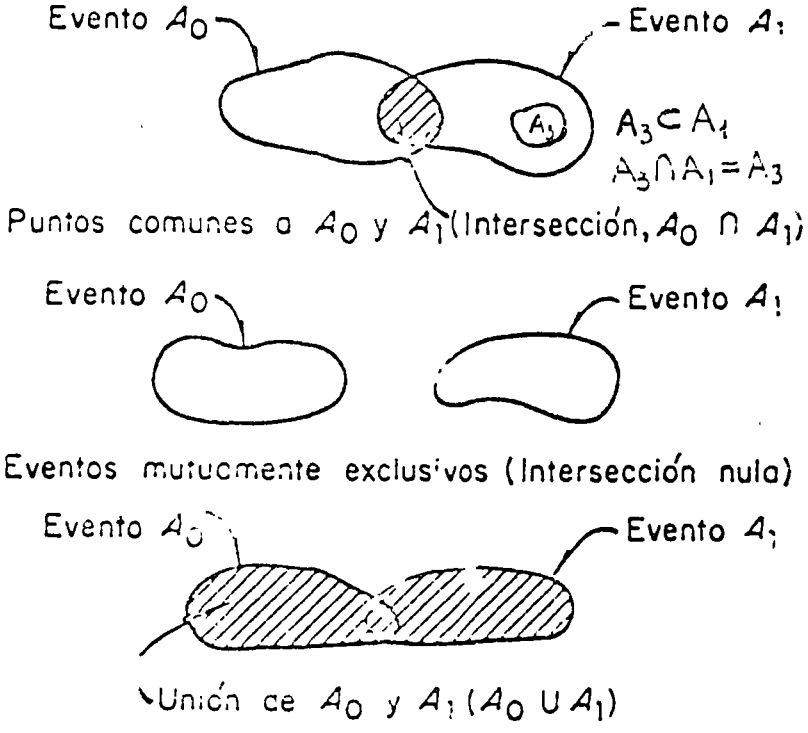
A∩B∩C∩D = E ={2,3}

AUBUCUD = F ={Y: 0≤Y≤5, 6,8,10,100}

LA OCURRENCIA DE UN EVENTO "Y" OTRO IMPLICA LA OCURRENCIA DE AMBOS A LA VEZ, ES DECIR, QUE SE VERIFIQUE LA *INTERSECCION*. LA OCURRENCIA DE UN EVENTO "O" ALGUN OTRO, IMPLICA LA OCURRENCIA DE CUALQUIERA DE ELLOS, ES DECIR DE LA *UNION*.

DIAGRAMAS DE VENN

UNA MANERA DE ILUSTRAR GRAFICAMENTE LAS OPERACIONES CON CONJUNTOS ES MEDIANTE LOS *DIAGRAMAS DE VENN*. EN ESTOS, CADA CONJUNTO SE REPRESENTA POR UNA CURVA CERRADA QUE ENCIERRA LOS ELEMENTOS QUE LE CORRESPONDEN.



Diagramas de Venn (unión e intersección de eventos)

TEORIA DE PROBABILIDADES

AL LANZAR UNA MONEDA NO PODEMOS PREDECIR CON CERTEZA CUAL CARA QUEDARA HACIA ARRIBA. LO UNICO QUE SE PUEDE ASEGURAR, SI LA MONEDA NO ESTA CARGADA, ES QUE AMBAS CARAS TIENEN LA MISMA OPORTUNIDAD DE SALIR, ES DECIR, QUE LOS EVENTOS SIMPLES {CARA} Y {CRUZ} TIENEN LA MISMA *PROBABILIDAD* DE OCURRIR.

COMO YA SE DIJO, LA *PROBABILIDAD* DE QUE OCURRA UN EVENTO ES UNA MEDIDA DEL GRADO DE CONFIANZA QUE SE TIENE DE QUE ESTE OCURRA AL REALIZAR EL EXPERIMENTO CORRESPONDIENTE.

EXISTEN POR LO MENOS TRES MANERAS DE ASIGNARLE UNA PROBABILIDAD A UN EVENTO:

1. EN TERMINOS DE LOS RESULTADOS DE REPETIR VARIAS VECES UN EXPERIMENTO (METODO FRECUENCIAL).
2. APLICANDO LA DEFINICIÓN CLASICA DE PROBABILIDADES.
3. CON BASE EN UN MODELO MATEMATICO PROBABILISTICO DEL FENOMENO DE QUE SE TRATE.

METODO FRECUENCIAL

SI N(A) ES EL NUMERO DE VECES QUE SE OBSERVA EL EVENTO A AL REALIZAR N VECES UN EXPERIMENTO, LA *FRECUENCIA RELATIVA* DE A, DEFINIDA COMO N(A)/N, SE CONSIDERA COMO ESTIMACION DE LA PROBABILIDAD DE A,

$$P(A) = \lim_{N\to\infty} \frac{N(A)}{N}$$

EJEMPLO

DE UNA URNA QUE CONTIENE BOLAS ROJAS, BLANCAS Y AZULES, SE SACO UNA BOLA, SE ANOTO SU COLOR Y SE REGRESO A LA URNA. SI ESTE EXPERIMENTO SE REPITE 20 VECES Y LOS RESULTADOS SON

b b,a,r,r,r,a,b,r,a,b,b,a,r,b,r,r,a,r,a, DONDE

r = ROJA, b = BLANCA, a = AZUL

¿QUE PROBABILIDADES LE ASIGNARIA A LOS EVENTOS B={b}, A={a}, y R={r}, DE ACUERDO CON EL METODO FRECUENCIAL?

EN ESTA MUESTRA SE TIENE QUE N(B)=6, N(A)=6, N(R)=8, N=20

POR LO QUE $P(B)=\frac{6}{20} = \frac{3}{10}$; $P(A) = \frac{6}{20} = \frac{3}{10}$; $P(R) = \frac{8}{20} = \frac{4}{10}$

NOTESE QUE LOS EVENTOS B, A Y R SON MUTUAMENTE EXCLUSIVOS, YA QUE SON EVENTOS SIMPLES, Y QUE

$$P(B) + P(A) + P(R) = \frac{3}{10} + \frac{3}{10} + \frac{4}{10} = 1$$

DEFINICION CLASICA DE PROBABILIDADES

SI N(A) ES EL NUMERO DE MANERAS *IGUALMENTE PROBABLES* EN QUE PUEDE OCURRIR EL EVENTO A Y N ES EL NUMERO TOTAL DE ELEMENTOS DEL ESPACIO DE EVENTOS CORRESPONDIENTE, ENTONCES LA PROBABILIDAD DE A ES

$$P(A) = \frac{N(A)}{N}$$

EJEMPLOS

A) SI EN UNA URNA SE TIENEN 5 BOLAS BLANCAS Y 15 NEGRAS, Y SE VA A SELECCIONAR UNA AL AZAR, ¿CUAL ES LA PROBABILIDAD DE QUE SEA ROJA (A={ROJA})?:

$N= 5+15=20;\ N(A)=5 \Rightarrow P(A)=\frac{5}{20}= \frac{1}{4}$

B) SI SE LANZAN DOS DADOS, ¿CUAL ES LA PROBABILIDAD DE QUE

1. SALGA UN 2 Y UN 5 (EVENTO B)?
2. LA SUMA SEA 7 (EVENTO A)

PARA EL INCISO 1 EL ESPACIO DE EVENTOS ES:

$$S = \begin{Bmatrix} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2.1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{Bmatrix}$$

SI EL DADO NO ESTA CARGADO, CADA PAREJA DE NUMEROS ES IGUALMENTE PROBABLE. EN TAL CASO, N=36 y N(B)=2 (APARECE (2,5) O (5,2)) $\Rightarrow$ P(B)=2/36=1/18.

PARA EL INCISO 2 EL ESPACIO DE EVENTOS ES

$S_1=\{2,3,4,5,6,7,8,9,10,11,12\}$

PERO NO TODOS LOS ELEMENTOS (EVENTOS SIMPLES) SON IGUALMENTE PROBA-

BLES, YA QUE, POR EJEMPLO, EL 2 SOLO APARECERA SI SE OBSERVA LA PAREJA (1,1), EN CAMBIO EL 3 APARECERA SI OCURREN LAS PAREJAS (1,2) O (2.1), ES DECIR, EL 3 TIENE EL DOBLE DE PROBABILIDAD QUE EL 2. POR ESTO, PARA CALCULAR LA PROBABILIDAD DE QUE LA SUMA SEA 7 ES NECESARIO TRABAJAR CON EL ESPACIO S Y CONTAR LAS MANERAS POSIBLES DE QUE LA SUMA SEA 7, LO CUAL OCURRE SI SE OBSERVA CUALQUIERA DE LAS PAREJAS (6,1), (5,2), (4,3), (3,4), (2,5) o (1,6), ES DECIR, HAY 6 MANERAS IGUALMENTE PROBABLES DE QUE OCURRA EL EVENTO A. POR LO TANTO

$$P(A) = \frac{N(A)}{N} = \frac{6}{36} = \frac{1}{6}$$

PROCEDIENDO DE ESTA MANERA SE PUEDEN CALCULAR LAS PROBABILIDADES DE QUE LA SUMA SEA 2,3,4, ETC. LOS RESULTADOS SON:

$$P(\{2\}) = \frac{1}{36};\ P(\{3\}) = \frac{2}{36};\ P(\{4\}) = \frac{3}{36};\ P(\{5\}) = \frac{4}{36};$$

$$P(\{6\}) = \frac{5}{36};\ P(\{7\}) = \frac{6}{36};\ P(\{8\}) = \frac{5}{36};\ P(\{9\}) = \frac{4}{36};$$

$$P(\{10\}) = \frac{3}{36}\ ;\ P(\{11\}) = \frac{2}{36} \text{ y } P(\{12\}) = \frac{1}{36}$$

(OBSERVESE QUE $\sum_{i=2}^{12} P(\{i\}) = 1$

ASIGNACION DE PROBABILIDADES MEDIANTE UN MODELO MATEMATICO

MEDIANTE ESTE METIDO LAS PROBABILIDADES SE ASIGNAN A PARTIR DE UN MODELO MATEMATICO QUE INVOLUCRE TODOS LOS FACTORES POSIBLES QUE INTERVIENEN EN LA ALEATORIEDAD DEL FENOMENO.

AXIOMAS DE LA TEORIA DE PROBABILIDADES

LAS PROBABILIDADES QUE SE ASIGNAN A LOS DIFERENTES EVENTOS RELACIONADOS CON UN FENOMENO ALEATORIO DEBEN CUMPLIR CON LOS SIGUIENTES TRES AXIOMAS:

AXIOMA 1: LA PROBABILIDAD DE OCURRENCIA DE UN EVENTO A ES UN NUMERO, P(A), QUE SE LE ASIGNA A DICHO EVENTO, CUYO VALOR QUEDA EN EL INTERVALO

$$0 \leq P(A) \leq 1$$

AXIOMA 2: SI S ES UN ESPACIO DE EVENTOS, ENTONCES

$$P(S)=1$$

AXIOMA 3: LA PROBABILIDAD, P(C), DE LA UNION, C, DE DOS EVENTOS MUTUAMENTE EXCLUSIVOS, A Y B, ES IGUAL A LA SUMA DE LAS PROBABILIDADES DE ESTOS, ES DECIR,

$$P(A \cup B) = P(C) = P(A) + P(B)$$

EJEMPLOS

A) EN EL PROBLEMA DEL LANZAMIENTO DE UN DADO QUE NO ESTA CARGADO SE PUEDE ASIGNAR A CADA NUMERO (A CADA EVENTO SIMPLE) UNA PROBABILIDAD DE 1/6, SI A={2,4} Y B={5.6}, ENTONCES, PUESTO QUE A={2}U{4} Y B={5}U{6}, Y QUE LOS EVENTOS ELEMENTALES SON MUTUAMENTE EXCLUSIVOS ENTRE SI, APLICANDO EL AXIOMA 3 SE OBTIENEN:

$$P(A)=P(\{2\}) + P(\{4\}) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

$$P(B)=P(\{5\}) + P(\{6\}) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

SI C=AUB, Y DADO QUE A Y B SON EVENTOS MUTUAMENTE EXCLUSIVOS:

$$P(C) = P(A) + P(B) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$$

ADEMAS, OBSERVESE QUE SE CUMPLE CON LOS AXIOMAS 1 Y 2, YA QUE

$$P(S) = P(\{1\}) + P(\{2\}) + P(\{3\}) + P(\{4\}) + P(\{5\}) + P(\{6\})$$

$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{6}{6} = 1$$

EJEMPLO

EN EL PROBLEMA DEL LANZAMIENTO DE DOS DADOS, ¿CUAL ES LA PROBABILIDAD QUE AL REALIZAR UNA VEZ EL EXPERIMENTO LA SUMA DE LOS DOS NUMEROS QUE QUEDEN HACIA ARRIBA SEA 7 U 11? ESTO ES EQUIVALENTE A PREGUNTAR POR LA PROBABILIDAD DE QUE OCURRA EL EVENTO $C = \{7\} \cup \{11\}$. PUESTO QUE {7} Y {11} SON EVENTOS MUTUAMENTE EXCLUSIVOS:

$$P(C) = P(\{7\}) + P(\{11\}) = \frac{6}{36} + \frac{2}{36} = \frac{8}{36} = \frac{2}{9}$$

EJEMPLO

EN UN LABORATORIO SE PROBARON 100 VIGAS DE CONCRETO REFORZADO NOMINALMENTE IDENTICAS, Y SE ANOTARON LAS CARGAS CON LAS CUALES FALLO CADA UNA. DE ESTA SUCESION DE EXPERIMENTOS SE ASIGNARON, EN TERMINOS DE LAS FRECUENCIAS RELATIVAS CORRESPONDIENTES, LAS SIGUIENTES PROBABILIDADES:

SI $A = \{X: 0 \leq X \leq 20 \text{ ton}\}$; $P(A) = 0.17$

SI $B = \{X: 20 < X \leq 40 \text{ ton}\}$; $P(B) = 0.24$

SI $C = \{X: 40 < X \leq 60 \text{ ton}\}$; $P(C) = 0.27$

SI $D = \{X: 60 < X \leq 80\}$; $P(D) = 0.13$

SI $E = \{X: 80 < X \leq 100\}$; $P(E) = 0.11$

SI $F = \{X: 100 < X\}$; $P(F) = 0.08$

$\Sigma P(.) = 1.00$

SI SE REALIZA UNA VEZ MAS EL EXPERIMENTO, CALCULEMOS LAS SIGUIENTES PROBABILIDADES:

A) QUE LA RESISTENCIA SEA MENOR O IGUAL QUE 80 TON. PUESTO QUE $G = \{X: 0 \leq X \leq 80 \text{ ton}\}$ SE TIENE QUE $G = AUBUCUD$, POR LO QUE

$P(G)=P(A) + P(B) + P(C) + P(D) = 0.17 + 0.24 + 0.27 + 0.13 = 0.81$

B) LA PROBABILIDAD QUE RESISTA MAS DE 60 TONS. PUESTO QUE $H = \{X: 60 < X \leq \infty\}$ O $H=\{X: X>60\}$ SE TIENE QUE $h = DUEUF$

POR LO QUE $P(H) = P(D) + P(E) + P(F) = 0.13 + 0.11 + 0.08 = 32$

C) LA PROBABILIDAD QUE RESISTA MAS DE 40 TON, PERO CUANDO MUCHO 100 TON.

PUESTO QUE $I=\{X: 40<X\leq 100\}$ SE TIENE QUE $I = CUDUE$

POR LO QUE $P(I) = P(C) + P(D) + P(E) = 0.27 + 0.13 + 0.11 = 0.51$

TEOREMAS

DOS TEOREMAS IMPORTANTES QUE SE DEDUCEN A PARTIR DE LOS AXIOMAS SON:

TEOREMA 1.

SI A ES UN EVENTO DEL ESPACIO S, ENTONCES $P(\bar{A})=1-P(A)$

DEMOSTRACION

PUESTO QUE A Y $\bar{A}$ SON MUTUAMENTE EXCLUSIVOS Y ADEMAS $A \cup \bar{A}=S$, ENTONCES, $P(S)=P(A)+P(\bar{A})=1$

$\Rightarrow P(\bar{A})=1-P(A)$

CASO PARTICULAR: PUESTO QUE $P(\bar{S})=1-P(S)=0$ Y $\bar{S}=\emptyset$, SE TIENE QUE $P(\emptyset)=0$

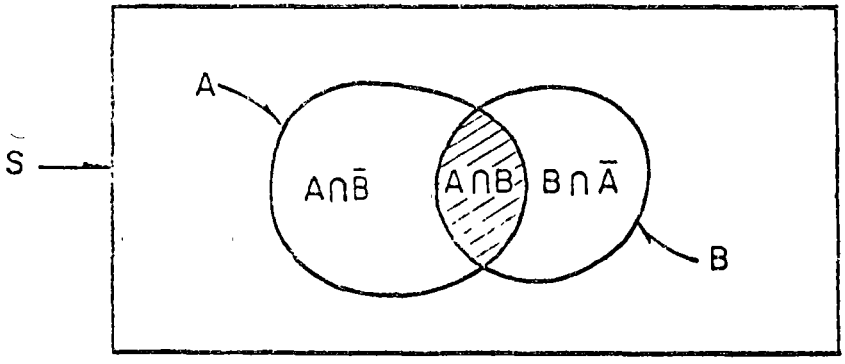
TEOREMA 2.

SI A Y B SON DOS EVENTOS CUALQUIERA DE DE S, ENTONCES

$$P(A\cup B)=P(A)+P(B)-P(A\cup B)$$

DEMOSTRACION

SEA EL DIAGRAMA DE VENN:



$A\cup B=(A\cap B)\cup(A\cap\bar{B})\cup(B\cap\bar{A})$. PUESTO QUE $A\cap B$, $A\cap\bar{B}$ Y $B\cap\bar{A}$ SON MUTUAMENTE EXCLUSIVOS, SE TIENE QUE $P(A\cup B)=P(A\cap B)+(A\cap\bar{B})+P(B\cap\bar{A})$.

SUMANDO Y RESTANDO $P(A\cap B)$ Y AGRUPANDO TERMINOS SE OBTIENE

$$P(A\cup B) = [P(A\cap B)+P(A\cap\bar{B})]+[P(A\cap B)+P(B\cap\bar{A})]-P(A\cap B)$$

PERO $A=(A\cap\bar{B})\cup(A\cap B)\Rightarrow P(A\cap B)+P(A\cap\bar{B})=P(A)$

Y $B=(A\cap B)\cup(B\cap\bar{A})\Rightarrow P(A\cap B)+P(B\cap\bar{A})=P(B)$, POR LO QUE

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)$$

EJEMPLO

EN UNA URNA SE TIENEN 28 TIRAS DE PAPEL Y EN CADA UNA SE ENCUENTRA ANOTADA UNA LETRA DISTINTA DEL ALFABETO. CALCULE LA PROBABILIDAD DE QUE AL EXTRAER AL AZAR UNA TIRA:

A) SE OBTENGA UNA VOCAL

B) SE OBTENGA a O z

C) OCURRAN C Y D, DONDE C={x,y,z} y
D={b,c,y,z}

D) OCURRA C O D

A) $A=\{a,e,i,o,u\} \Rightarrow P(A)=\frac{5}{28}$

B) $B=\{a,z\} \Rightarrow P(B)=\frac{2}{28}$

C) $F=C\cap D=\{y,z\} \Rightarrow P(F)=\frac{2}{28}$

D) $E=C\cup D=\{b,c,x,y,z\} \Rightarrow P(E)=\frac{5}{28}$

o $P(E)=P(C)+P(D)-P(C\cap D)$

$P(C\cap D)=P(F)=\frac{2}{28} \Rightarrow P(E)=\frac{3}{28}+\frac{4}{28}-\frac{2}{28}=\frac{5}{28}$

PROBABILIDAD CONDICIONAL

LA PROBABILIDAD CONDICIONAL, $P(A|B)$, DEL EVENTO A, DADO QUE EL B HA OCURRIDO SE CALCULA CON LA FORMULA

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad ; \; P(B)>0 \qquad (1)$$

SI DOS EVENTOS A Y B, SON INDEPENDIENTES, LA PROBABILIDAD DE A NO SE ALTERA SI OCURRE EL EVENTO B; ES DECIR, DOS EVENTOS SON *INDEPENDIENTES* SI

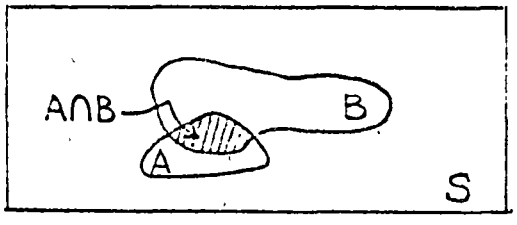
$$P(A|B) = P(A)$$

EN TAL CASO, DE LA ECUACION 1 :

$$P(A \cap B) = P(A) \times P(B)$$

PUESTO QUE $P(A \cap B) = N(A \cap B)/N(S)$ Y $P(B) = N(B)/N(S)$ LA ECUACION 1 SE PUEDE ESCRIBIR COMO

$$P(A|B) = \frac{\frac{N(A \cap B)}{N(S)}}{\frac{N(B)}{N(S)}} = \frac{N(A \cap B)}{N(B)} \qquad (2)$$



EL TRABAJAR CON LA ECUACION 2 EQUIVALE A EMPLEAR UN ESPACIO DE EVENTOS REDUCIDO DE S A B.

EJEMPLO

EN UNA URNA HAY 10 TRANSISTORES BUENOS Y 10 DEFECTUOSOS. ¿CUAL ES LA PROBABILIDAD DE SACAR UNO BUENO Y UNO DEFECTUOSO (EN CUALQUIER ORDEN) AL REALIZAR DOS EXTRACCIONES AL AZAR, SI HAY REEMPLAZO DEL PRIMER TRANSISTOR OBSERVADO?

HAY VARIAS FORMAS DE RESOLVER ESTE PROBLEMA:

1. PUESTO QUE EL NUMERO DE DEFECTUOSOS ES IGUAL AL DE BUENOS, SE PUEDE FORMULAR EL SIGUIENTE ESPACIO DE EVENTOS, EN EL QUE TODOS LOS ELEMENTOS SON IGUALMENTE PROBABLES:

   $S = \{(D,D), (D,B), (B,B), (B,D)\}$

   EL EVENTO DE INTERES ES:

   $A = \{(D,B), (B,D)\}$

   POR LO QUE $N(S) = 4$, $N(A) = 2$

   Y $P(A) = 2/4 = 1/2$

2. HAY 10×10 MANERAS DISTINTAS DE QUE SALGA PRIMERO EL BUENO Y LUEGO EL DEFECTUOSO, Y OTRAS TANTAS DE QUE OCURRA DE MANERA INVERSA. POR LO TANTO:

   $N(A) = (10 \times 10) \times 2 = 200$

   $N(S) = 20 \times 20 = 400$

   $P(A) = 200/400 = 1/2$

3. SEAN LOS EVENTOS

   B = {SALE PRIMERO EL BUENO Y LUEGO EL MALO}

   C = {SALE PRIMERO EL MALO Y LUEGO EL BUENO}

   D = {SALE PRIMERO EL BUENO}

   E = {SALE SEGUNDO EL MALO}

   O = {SALE PRIMERO EL MALO}

   R = {SALE SEGUNDO EL BUENO}

   POR LO TANTO

   $B = D \cap E$ Y $F = O \cap R$

SI A = {SALE UNO BUENO Y UNO MALO} = BUF

SE TIENE QUE $P(A) = P(B)+P(F)$

YA QUE B Y F SON MUTUAMENTE EXCLUSIVOS, Y

$$P(B) = P(D \cap E) = \frac{10}{20} \times \frac{10}{20} = \frac{100}{400} = \frac{1}{4}$$

$$P(F) = P(O \cap R) = \frac{10}{20} \times \frac{10}{20} = \frac{1}{4}$$

YA QUE D Y E, Y O Y R SON INDEPENDIENTES. ESTO CONDUCE A

$$P(A) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

RESOLVAMOS AHORA ESTE PROBLEMA SI NO HAY REEMPLAZO:

$P(D \cap E) = P(E|D)P(D)$

$P(D) = 10/20$, $P(E|D) = 10/19$

POR LO QUE $P(A) = \frac{10}{38} + \frac{10}{38} = \frac{10}{19}$

EN GENERAL, LOS EVENTOS $A_1, A_2, \ldots, A_M$

SON INDEPENDIENTES SI, Y SOLO SI,

$$P(A_{K_1} \cap A_{K_2} \cap \ldots \cap A_{K_R}) = P(A_{K_1}) \times P(A_{K_2}) \times \ldots \times P(A_{K_R})$$

PARA *CUALQUIER* GRUPO DE ENTEROS $K_1, K_2, \ldots, K_R$, CON $K_R \leq M$ (TODAS LAS PAREJAS, TERCIAS, ETC, DE EVENTOS POSIBLES DE FORMARSE DEBEN SER INDEPENDIENTES).

EJEMPLO

EN UN ESTUDIO SOCIOLOGICO SE INTERROGARON 1200 PERSONAS DE UNA COLONIA RESIDENCIAL, Y SE OBTUVIERON LOS SIGUIENTES DATOS:

GUSTO POR LA MUSICA CLASICA	TITULO UNIVERSITARIO		SIN TITULO UNIVERSITARIO		
	VARONES	DAMAS	VARONES	DAMAS	Σ
ALTO	100	50	200	250	600
BAJO	150	100	150	200	600
Σ	250	150	350	450	1200

SI A = {VARON} , B = {CON TITULO}

C = {GUSTO ALTO}

¿CUAL ES LA PROBABILIDAD DE QUE SI SE SELECCIONA UN CIUDADANO AL AZAR DE LA MISMA COLONIA, ESTE SEA VARON, TENGA TITULO Y GUSTO ALTO POR LA MUSICA?

POR EL METODO FRECUENCIAL:

NUMERO DE VARONES = 250 + 350 = 600

NUMERO DE PERSONAS CON TITULO = 250 + 150 = 400

NUMERO DE PERSONAS CON ALTO GUSTO POR LA MUSICA CLASICA = 600

POR LO TANTO

$P(A) = 600/1200 = 1/2$, $P(B) = 400/1200 = \frac{1}{3}$

Y $P(C) = 600/1200 = 1/2$. PUESTO QUE

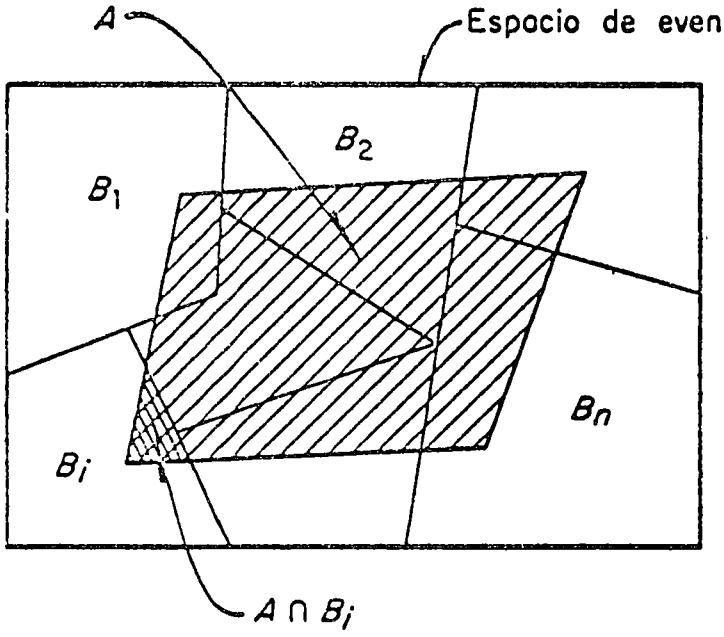
$D = A \cap B \cap C$ Y A, B Y C SON INDEPENDIENTES, SE TIENE QUE

$$P(D) = \frac{1}{2} \times \frac{1}{3} \times \frac{1}{2} = \frac{1}{12}$$

DE OTRA MANERA: $P(D) = 100/1200 = 1/12$

TEOREMA DE LA PROBABILIDAD TOTAL

SE DICE QUE UN GRUPO DE EVENTOS ES *COLECTIVAMENTE EXHAUSTIVO* SI LA UNION DE TODOS ELLOS ES EL ESPACIO DE EVENTOS CORRESPONDIENTE.



EN UN GRUPO DE EVENTOS COLECTIVAMENTE EXHAUSTIVOS Y MUTUAMENTE EXCLUSIVOS, B_1, B_2,...,B_n, SI A ES UN EVENTO CUALQUIERA DEFINIDO EN EL MISMO ESPACIO, ENTONCES, APLICANDO EL AXIOMA 3, RESULTA

$$P(A) = P(A\cap B_1) + P(A\cap B_2) +\ldots+ P(A\cap B_n) = \sum_{i=1}^{i=n} P(A\cap B_i)$$

YA QUE LOS EVENTOS $A\cap B_i$ SON MUTUAMENTE EXCLUVIVOS.

TOMANDO EN CUENTA QUE $P(A\cap B_i) = P(B_i)P(A|B_i)$, SE OBTIENE FINALMENTE LA ECUACION

$$P(A) = \sum_{i=1}^{i=n} P(B_i)P(A|B_i)$$

CON LA CUAL SE DEFINE EL LLAMADO *TEOREMA DE LA PROBABILIDAD TOTAL.*

EJEMPLO

EN UNA FABRICA SE RECIBEN REGULADORES DE VOLTAJE DE DOS PROVEEDORES, B_1 Y B_2, EN PROPORCION DE 3 A 1; ES DECIR, LA PROBABILIDAD DE QUE UN REGULADOR TOMADO AL AZAR PROVENGA DEL PROVEEDOR B_1 ES $P(B_1)=3/4$, Y DEL B_2 ES $P(B_2)=1/4$.

SUPONGAMOS ADEMAS QUE EL CONTROL DE CALIDAD DEL PROVEEDOR B_1 ES MEJOR QUE EL DE B_2, DE MANERA QUE EL 95% DE LOS REGULADORES DE B_1 TRABAJAN BIEN, Y SOLO EL 80% DE LOS DE B_2 FUNCIONAN CORRECTAMENTE. CALCULEMOS LA PROBABILIDAD DE QUE UN REGULADOR TOMADO AL AZAR FUNCIONE BIEN (EVENTO A).

$$P(A|B_1) = 0.95;\ P(A|B_2) = 0.80$$

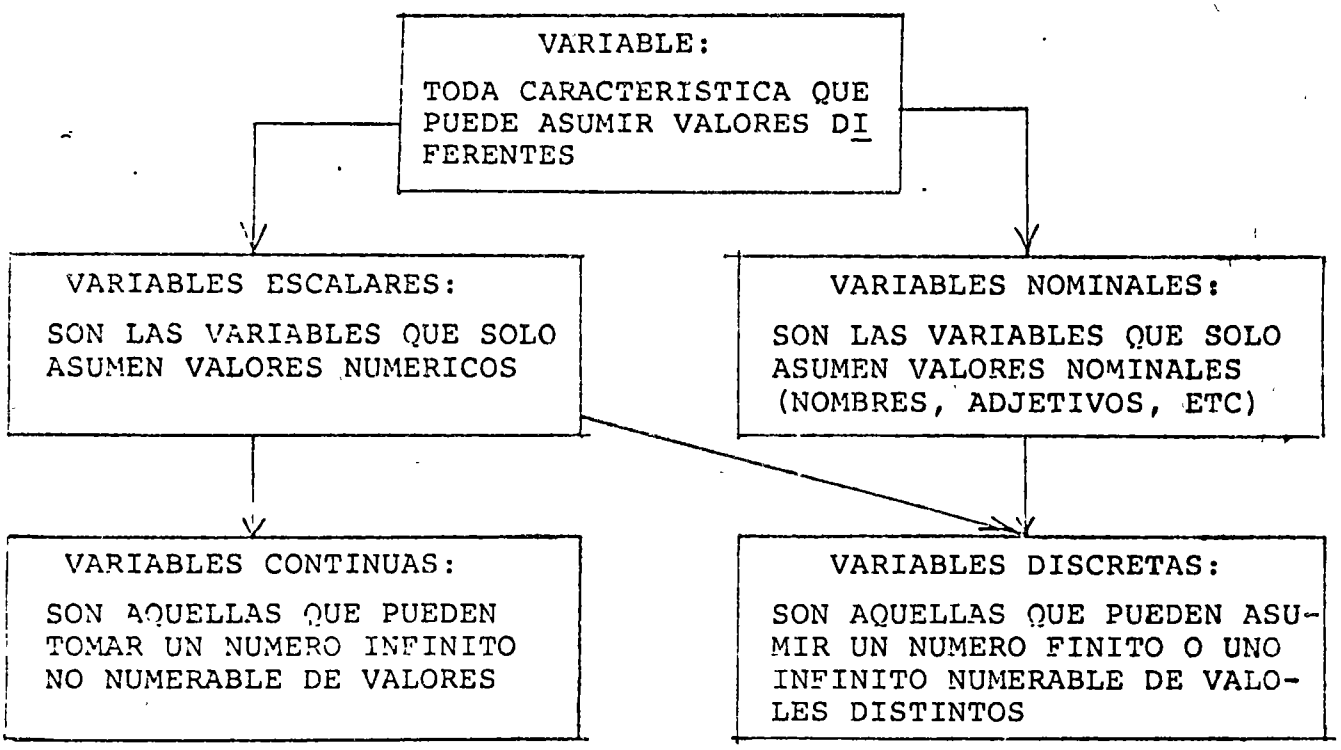
DEL TEOREMA DE LA PROBABILIDAD TOTAL:

$$P(A) = P(A|B_1)P(B_1)+P(A|B_2)P(B_2)$$

$$= 0.95 \times \frac{3}{4} + 0.80 \times \frac{1}{4} = 0.9125$$

VARIABLES ALEATORIAS

CLASIFICACION DE VARIABLES



UNA *VARIABLE ALEATORIA* ES UNA VARIABLE TAL QUE NO PUEDE PREDECIRSE CON CERTEZA EL VALOR QUE ASUMIRA ANTES DE REALIZAR UN EXPERIMENTO. POR EJEMPLO, LA RESISTENCIA O CARGA DE FALLA DE UNAS VIGAS ES UNA VARIABLE ALEATORIA, YA QUE ANTES DE ROMPER UNA VIGA TOMADA AL AZAR NO SE PUEDE PRECISAR CUAL SERA SU RESISTENCIA. EN LA SIGUIENTE TABLA SE PRESENTAN LOS RESULTADOS EXPERIMENTALES CON 15 VIGAS DE CONCRETO REFORZADO, OBSERVANDOSE QUE ESTOS VARIAN DE UNAS A OTRAS DE MANERA ALEATORIA.

TABLA 2. PRUEBAS DE VIGAS DE CONCRETO REFORZADO

Número de la viga	Carga de agrietamiento, en kg	Carga de falla, en kg
1	4 700	4 700
2	3 840	4 220
3	3 270	4 360
4	2 310	4 680
5	2 950	4 270
6	4 810	4 810
7	2 720	4 590
8	2 720	4 490
9	4 310	4 310
10	2 950	4 630
11	4 220	4 220
12	2 720	4 340
13	2 720	4 340
14	2 630	4 770
15	2 950	4 630

A TODO EXPERIMENTO SE LE PUEDE ASOCIAR AL MENOS UNA VARIABLE ALEATORIA, DEPENDIENDO ESTA DEL PROBLEMA QUE SE TENGA PLANTEADO. POR EJEMPLO, EN EL CASO DE LA RESISTENCIA DE LAS VIGAS DE VARIABLE ALEATORIA PUEDE SER DIRECTAMENTE [illegible] DICHA RESISTENCIA, EN CUYO CASO SU ESPACIO DE EVENTOS SERIA

$$S_1 = \{X: 0<X<\infty\}$$

LA VARIABLE TAMBIEN PUDO HABER SIDO UNA CUYO ESPACIO DE EVENTOS FUERA

$$S_2 = \{\text{EXITO, FRACASO}\}$$

EN DONDE EL EXITO OCURRIRIA SI LA VIGA RESISTIERA MAS DE CIERTA CANTIDAD, POR EJEMPLO 4600 KG, Y EL FRACASO OCURRIRIA SI RESISTIERA MENOS, ES DECIR:

EXITO: SI $X \geq 4600$ KG

FRACASO: SI $X < 4600$ KG

LEYES DE PROBABILIDADES

EL COMPORTAMIENTO DE UNA VARIABLE ALEATORIA SE DESCRIBE MEDIANTE SU *LEY DE PROBABILIDADES*, LA CUAL PUEDE ESPECIFICARSE DE DIFERENTES FORMAS. LA MANERA MAS COMUN DE HACERLO ES MEDIANTE SU *DISTRIBUCION O DENSIDAD DE PROBABILIDADES*.

A FIN DE EVITAR CONFUSION, SE EMPLEARA UNA LETRA MAYUSCULA PARA DENOTAR UNA VARIABLE ALEATORIA, Y LA MINUSCULA CORRESPONDIENTE PARA LOS VALORES QUE PUEDE ASUMIR. SI LA VARIABLE ALEATORIA X ES DISCRETA Y PUEDE ASUMIR LOS VALORES x_i, SU DENSIDAD DE PROBABILIDADES, $f_X(x)$ SERA EL CONJUNTO DE LAS PROBABILIDADES

$$P_X(x_i) = P(X = x_i)$$

LA CUAL SE LEE "PROBABILIDAD DE QUE $X = x_i$". ESTO ES

$$f_X(x) = \{P_X(x_i)\}$$

PARA QUE UNA DENSIDAD DE PROBABILIDADES SATISFAGA LOS TRES AXIOMAS DE LA TEORIA DE PROBABILIDADES, SE DEBEN CUMPLIR LOS SIGUIENTES REQUISITOS

A) $0 \leq P_X(x_i) \leq 1$ PARA TODA x_i

B) $\sum_{i=1}^{n} P_X(x_i) = 1$, DONDE n ES EL NUMERO TOTAL DE VALORES QUE PUEDE ASUMIR X

C) $P(X_m \leq X \leq X_r) = \sum_{i=m}^{i=r} P_X(x_i)$

LA DISTRIBUCION DE PROBABILIDADES ACUMULADAS O FUNCION DE DISTRIBUCION

OTRA FORMA DE ESPECIFICAR LA LEY DE PROBABILIDADES DE UNA VARIABLE ALEATORIA ES MEDIANTE LA DISTRIBUCION DE PROBABILIDADES ACUMULADAS, $F_X(x)$, QUE SE DEFINE COMO EL CONJUNTO DE LAS SUMAS PARCIALES DE LAS PROBABILIDADES, $P_X(x_i)$, CORRESPONDIENTES A TODOS LOS VALORES DE X MENORES O IGUALES QUE x_i. POR LO TANTO, ESTA FUNCION DA LAS PROBABILIDADES DE QUE LA VARIABLE ALEATORIA TOME VALORES MENORES O IGUALES QUE x_m PARA CUALQUIER m, ES DECIR

$$F_X(x) = \{F_X(x_m)\}$$

EN DONDE

$$F_X(x_m) = \sum_{i=1}^{i=m} P_X(x_i) = P(X \leq x_m)$$

EJEMPLO

SEA X LA VARIABLE ALEATORIA DISCRETA "NUMERO TOTAL DE CARROS QUE SE DETIENEN EN UNA ESQUINA DEBIDO A LA LUZ ROJA DE UN SEMAFORO". SI LAS PROBABILIDADES ASOCIADAS A CADA VALOR, DETERMINADAS POR EL METODO FRECUENCIAL, SON

$$P_X(x) = \begin{cases} 0.1 & SI \quad x = 0 \\ 0.2 & SI \quad x = 1 \\ 0.3 & SI \quad x = 2 \\ 0.2 & SI \quad x = 3 \\ 0.1 & SI \quad x = 4 \\ 0.1 & SI \quad x = 5 \\ 0 & SI \quad x \geq 6 \end{cases}$$

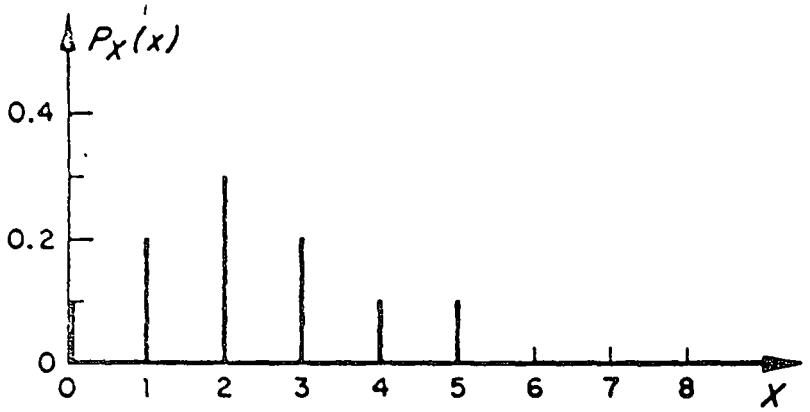
LA DISTRIBUCION DE PROBABILIDADES Y LA DE PROBABILIDADES ACUMULADAS CORRESPONDIENTES SERAN

x	$f_X(x)$	$F_X(x)$
<0	0	0
0	0.1	0.1
1	0.2	0.3
2	0.3	0.6
3	0.2	0.8
4	0.1	0.9
5	0.1	1.0
≥ 6	0	1.0

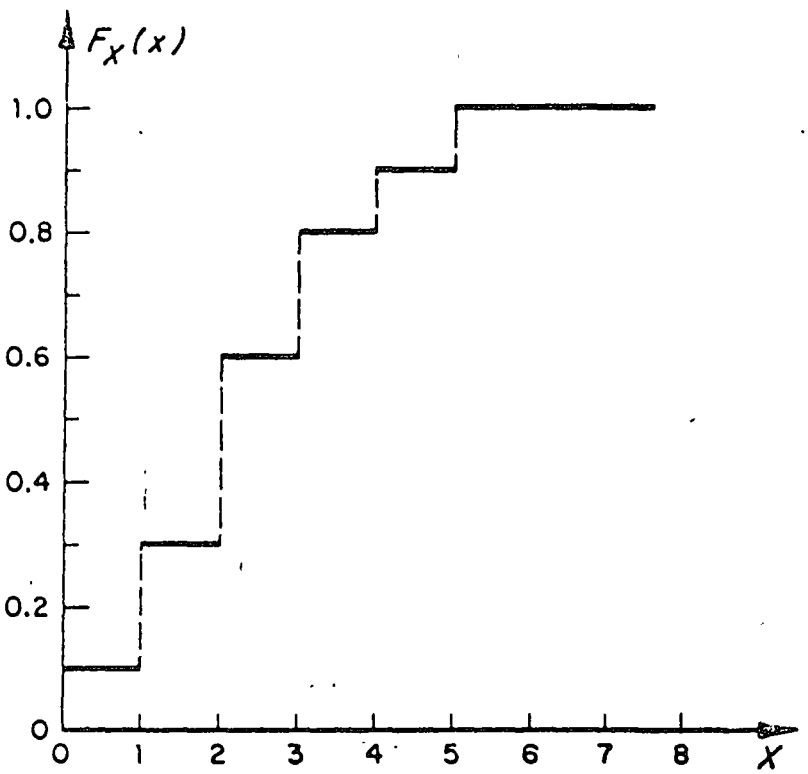
O SEA

$$F_X(x) = \begin{cases} 0 & SI \quad x < 0 \\ 0.1 & SI \quad 0 \leq x < 1 \\ 0.3 & SI \quad 1 \leq x < 2 \\ 0.6 & SI \quad 2 \leq x < 3 \\ 0.8 & SI \quad 3 \leq x < 4 \\ 0.9 & SI \quad 4 \leq x < 5 \\ 1.0 & SI \quad 5 \leq x \end{cases}$$

LAS GRAFICAS DE ESTAS DISTRIBUCIONES SE PRESENTAN EN LA FIGURA DE LA SIGUIENTE HOJA.



a) Distribución de probabilidades



b) Función de distribución

Ley de probabilidades del ejemplo del tráfico

EJEMPLO

SEA LA VARIABLE ALEATORIA X DEFINIDA POR LA SUMA DE LOS DOS NUMEROS QUE QUEDEN HACIA ARRIBA AL LANZAR DOS DADOS. EN ESTE CASO EL ESPACIO DE EVENTOS ES

$$S = \{2,3,4,5,6,7,8,9,10,11,12\}$$

Y LA DENSIDAD DE PROBABILIDADES ES

$$f_X(x) = \{\frac{1}{36}, \frac{2}{36}, \frac{3}{36}, \frac{4}{36}, \frac{5}{36}, \frac{6}{36}, \frac{5}{36}, \frac{4}{36}, \frac{3}{36}, \frac{2}{36}, \frac{1}{36}\}$$

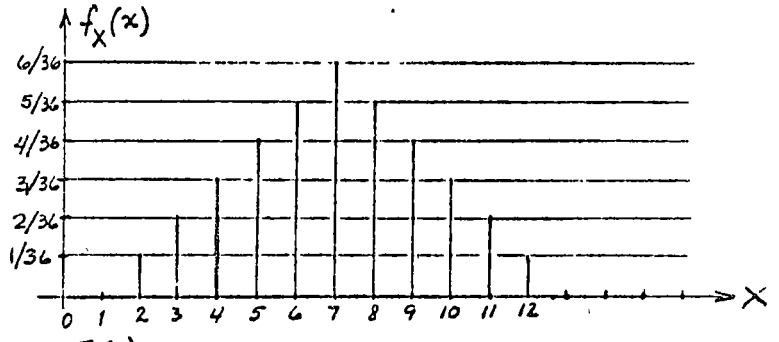
EN ESTE CASO $X_1=2$, $X_2=3,\ldots,X_{11}=12$

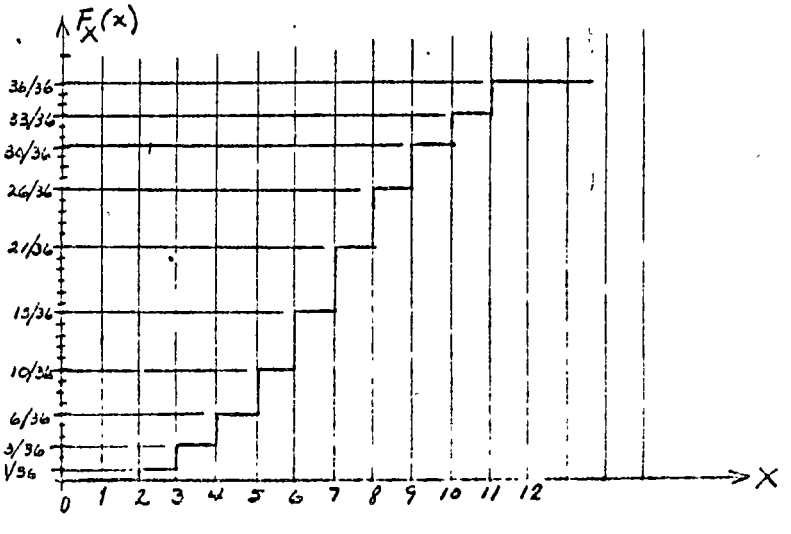
$$\text{Y} \quad f_X(2) = \frac{1}{36},\ f_X(3) = \frac{2}{36},\ldots,f_X(12) = \frac{1}{36}$$

ESTAS PROBABILIDADES FUERON CALCULADAS EN UN EJEMPLO PREVIO SOBRE PROBABILIDADES DE EVENTOS .

CON ESTAS PROBABILIDADES SE PUEDE OBTENER LA FUNCION DE DISTRIBUCION O DE PROBABILIDADES ACUMULADAS, DE LA SIGUIENTE MANERA:

x	$f_X(x)$	$F_X(x)$
<2	0	0
2	1/36	1/36
3	2/36	3/36
4	3/36	6/36
5	4/36	10/36
6	5/36	15/36
7	6/36	21/36
8	5/36	26/36
9	4/36	30/36
10	3/36	33/36
11	2/36	35/36
12	1/36	36/36=1
>12	Σ=36/36=1	1





EN EL CASO DE UNA VARIABLE ALEATORIA CONTINUA, X, LA PROBABILIDAD DE QUE ESTA TOME UN VALOR COMPRENDIDO ENTRE x Y x + dx ESTA DADA POR $f_X(x)dx$, DONDE $f_X(x)$ ES LA DENSIDAD DE PROBABILIDADES DE X. POR LO TANTO, LA PROBABILIDAD DE QUE X ASUMA VALORES COMPRENDIDOS EN EL INTERVALO $x_1 \leq X \leq x_2$ ES

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f_X(x)\, dx$$

LA INTERPRETACION GRAFICA DE ESTA PROBABILIDAD ES QUE CORRESPONDE AL AREA BAJO LA CURVA DE $f_X(x)$ COMPRENDIDA ENTRE x_1 Y x_2.

PUESTO QUE $F_X(x) = P(X \leq x) = P(-\infty \leq X \leq x)$, Y EN VIRTUD DE LA ECUACION ANTERIOR SE TIENE QUE

$$F_X(x) = \int_{-\infty}^{x} f_X(U)dU$$

DONDE U ES SOLO UNA VARIABLE MUDA DE INTEGRACION. EL VALOR DE ESTA INTEGRAL ES IGUAL AL AREA BAJO LA CURVA DE $F_X(x)$ A LA IZQUIERDA DE x. DE ESTA ECUACION SE CONCLUYE QUE

$$\frac{dF_X(x)}{dx} = \frac{d}{dx}\left(\int_{-\infty}^{x} f_X(U)dU\right) = f_X(x)$$

ALGUNAS PROPIEDADES DE $F_X(x)$ SON:

$$0 \leq F_X(x) \leq 1$$

$$F_X(-\infty) = 0$$

$$F_X(\infty) = 1$$

$$F_X(x + \varepsilon) \geq F_X(x), \text{ SI } \varepsilon \geq 0$$

$$F_X(x_2) - F_X(x_1) = P(x_1 \leq X \leq x_2)$$

PARA SATISFACER LOS AXIOMAS DE LA TEORIA DE PROBABILIDADES SE NECESITA QUE

$$F_X(x) \geq 0 \text{ PARA TODA } x$$

$$\int_{-\infty}^{\infty} f_X(x)\, dx = 1$$

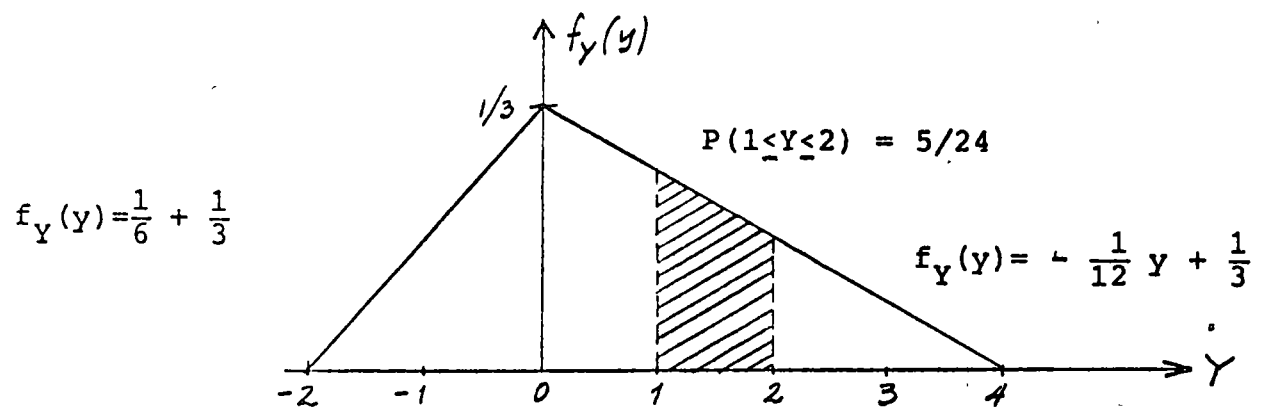
EJEMPLO

SEA UNA VARIABLE ALEATORIA CONTINUA CUYA DENSIDAD DE PROBABILIDADES ES DE FORMA TRIANGULAR DADA POR LAS SIGUIENTES ECUACIONES:

$$f_Y(y) = \frac{1}{6}\,y + \frac{1}{3}\,, \text{ SI } -2 \leq Y \leq 0$$

$$f_Y(y) = -\frac{1}{12}\,y + \frac{1}{3}\,, \text{ SI } 0 \leq Y \leq 4$$

$$f_Y(y) = 0 \qquad \text{SI } Y \leq 2 \text{ O } Y \geq 4$$



LA DISTRIBUCION DE PROBABILIDADES ACUMULADAS ES, ENTONCES:

SI $-2 \leq Y \leq 0$

$$F_Y(y) = \int_{-\infty}^{Y} f_Y(U)\,dU = \int_{-2}^{Y} \left(\frac{1}{6}\,U + \frac{1}{3}\right) dU$$

$$= \left[\frac{U^2}{12} + \frac{U}{3}\right]_{-2}^{Y} = \frac{Y^2}{12} + \frac{Y}{3} + \frac{1}{3}$$

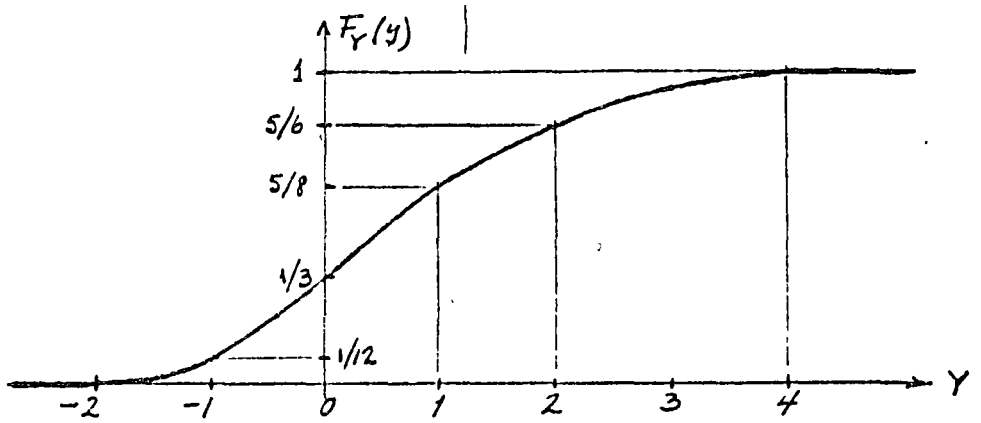
SI $0 \leq Y \leq 4$

$$F_Y(Y) = F_Y(0) + \int_0^Y \left(-\frac{1}{12}\,U + \frac{1}{3}\right) dU = \frac{1}{3} + \left[-\frac{U^2}{24} + \frac{U}{3}\right]_0^Y =$$

$$= \frac{1}{3} - \frac{Y^2}{24} + \frac{Y}{3} \qquad \text{SI } 0 \leq y \leq 4$$

$F_Y(y) = 0$ SI $y \leq -2$

$F_Y(y) = 1$ SI $y \geq 4$



SI SE DESEA CALCULAR LA PROBABILIDAD DE QUE AL REALIZAR UNA VEZ EL EXPERIMENTO QUE INVOLUCRA A DICHA VARIABLE, EL VALOR QUE SE OBSERVE CAIGA EN EL INTERVALO $1 \leq Y \leq 2$, ENTONCES

$$P[1 \leq Y \leq 2] = \int_1^2 \left(-\frac{1}{12}y + \frac{1}{3}\right) dy = \left[\frac{-y^2}{24} + \frac{y}{3}\right]_1^2 = \frac{5}{24}$$

O

$$P[1 \leq Y \leq 2] = F_Y(2) - F_Y(1) = \frac{5}{6} - \frac{5}{8} = \frac{5}{24}$$

ESPERANZAS

LA ESPERANZA DE UNA FUNCION g(X), DE UNA VARIABLE ALEATORIA DISCRETA, X, ES, POR DEFINICION

$$E(g(X)) = \sum_{i=1}^{i=n} g(x_i) P_X(x_i)$$

O PARA UNA VARIABLE CONTINUA

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx$$

EJEMPLOS

1. SI $g(X)$ = CONSTANTE = c

$$E(c) = c \int_{-\infty}^{\infty} \delta_X(x)\,dx = c$$

2. SI $g(X) = cx$

$$E[cx] = c\int_{-\infty}^{\infty} x\delta_X(x)\,dx = cE[X].$$

3. SI $g(X) = a + bx$

$$E[a + bx] = a \int_{-\infty}^{\infty} \delta_X(x)\,dx + b \int_{-\infty}^{\infty} x\delta_X(x)\,dx = a + bE[X]$$

4. SI $g(X) = g_1(X) + g_2(X)$

$$E[g_1(X) + g_2(X)] = \int_{-\infty}^{\infty} g_1(x)\delta_X(x)\,dx + \int_{-\infty}^{\infty} g_2(x)\delta_X(x)\,dx$$

$$= E[g_1(X)] + E[g_2(X)]$$

EJEMPLO

SI X ES UNA VARIABLE ALEATORIA CON DENSIDAD DE PROBABILIDADES EXPONENCIAL, CALCULAR LA ESPERANZA DE LA FUNCION

$$g(X) = X^2$$

EN ESTE CASO SE TIENE QUE

$$f_X(x) = \lambda e^{-\lambda x}$$

POR LO QUE

$$E(X^2)=E[g(X)]= \int_{-\infty}^{\infty} g(x) f_X(x)\,dx = \lambda \int_{-\infty}^{\infty} x^2 e^{-\lambda x}\,dx$$

$$=\lambda\left[\frac{-x^2 e^{-\lambda x}}{\lambda}\right]_0^{\infty} + \frac{2\lambda}{\lambda}\int_0^{\infty} x e^{-\lambda x}\,dx = \frac{-2}{\lambda^2}\left[e^{-\lambda x}(1+\lambda x)\right]_0^{\infty} = \frac{2}{\lambda^2}.$$

EN GENERAL, A LA ESPERANZA DE X^2 SE LE DENOMINA *VALOR MEDIO CUADRATICO.*

MEDIDAS DE TENDENCIA CENTRAL

LA *MEDIA O ESPERANZA* , $E[X]$, DE UNA VARIABLE ALEATORIA, X, SE CALCULA CON LAS ECUACIONES ANTERIORES PARA EL CASO EN QUE $g(X)=X$. DE ESTA MANERA, SI LA VARIABLE ES DISCRETA, SU ESPERANZA QUEDA DADA POR

$$E(X) = \sum_{i=1}^{i=n} x_i \, P_X(x_i)$$

DONDE n ES EL TOTAL DE VALORES QUE X PUEDE ASUMIR.

PARA EL CASO DE UNA VARIABLE ALEATORIA CONTINUA, LA MEDIA ES

$$m_X = E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

OTRAS MEDIDAS USUALES DE TENDENCIA CENTRAL DE UNA VARIABLE ALEATORIA SON LA *MEDIANA* Y EL *MODO*. LA PRIMERA SE DEFINE COMO EL VALOR DE LA VARIABLE AL CUAL CORRESPONDE UNA PROBABILIDAD ACUMULADA DE 50%, Y LA SEGUNDA, COMO EL VALOR DE LA VARIABLE AL CUAL CORRESPONDE LA MAYOR PROBABILIDAD.

EJEMPLO

SI LA DENSIDAD DE PROBABILIDADES DE LA VARIABLE ALEATORIA X CORRESPONDE A LOS ERRORES EN UNA NIVELACIÓN, ES LA DE LA SEGUNDA COLUMNA DE LA SIGUIENTE TABLA, LA MEDIA DE DICHA VARIABLE RESULTA SER 4 167 LA MEDIANA 4000 Y EL MODO 4000 MICRAS. LOS CALCULOS CORRESPONDIENTES SE LOCALIZAN EN LA TERCERA COLUMNA.

x_i, EN MICRAS	$P_X(x_i)$	$x_i P_X(x_i)$, EN MICRAS	$F_X(x_i)$
0	6/60	0	6/60
1 000	2/60	2 000/60	8/60
2 000	4/60	8 000/60	12/60
3 000	8/60	24 000/60	20/60
4 000	13/60	52 000/60	33/60≥0.5
5 000	12/60	60 000/60	45/60
6 000	7/60	42 000/60	52/60
7 000	4/60	28 000/60	56/60
8 000	2/60	16 000/60	58/60
9 000	2/60	18 000/60	60/60
TOTAL: $E[X]$ = 250 000/60=4 167 MICRAS			

EJEMPLO

CALCULAR LA ESPERANZA DE UNA VARIABLE ALEATORIA CUYA DENSIDAD DE PROBABILIDADES ES TRIANGULAR DADA POR

$$F_Y(y) = \frac{1}{6}y + \frac{1}{3} \qquad \text{SI } -2 \leq y \leq 0$$

$$F_Y(y) = \frac{-1}{12}y + \frac{1}{3} \qquad \text{SI } 0 \leq y \leq 4$$

$$F_Y(y) = 0 \qquad \text{SI } y \leq -2 \text{ O } y \geq 4$$

$$E(Y) = \int_{-\infty}^{\infty} y f_Y(y)dy = \int_{-2}^{0} y\left(\frac{y}{6} + \frac{1}{3}\right)dy + \int_{0}^{4} y\left(\frac{-y}{12} + \frac{1}{3}\right)dy$$

$$= \left[\frac{y^3}{18} + \frac{y^2}{6}\right]_{-2}^{0} + \left[\frac{-y^3}{36} + \frac{y^2}{6}\right]_{0}^{4} = \frac{2}{3}$$

EJEMPLO

CALCULAR LA ESPERANZA DE UNA VARIABLE ALEATORIA CON DENSIDAD DE PROBABILIDADES EXPONENCIAL

$$f_X(x) = \lambda e^{-\lambda x}$$

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx = \lambda \int_0^{\infty} x e^{-\lambda x}\,dx = \lambda \left[\frac{-e^{-\lambda x}}{\lambda^2}(1+\lambda x)\right]_0^{\infty} = \frac{1}{\lambda}$$

MEDIDAS DE DISPERSION

UNA MEDIDA MUY COMUY DE LA DISPERSION O VARIABILIDAD DE LOS VALORES QUE PUEDE ASUMIR UNA VARIABLE ALEATORIA ES LA *VARIANCIA*, LA CUAL SE DENOTA COMO $\sigma^2(X)$ O VAR (X), LA CUAL SE DEFINE COMO LA ESPERANZA DE LA FUNCION $g(X) = [X-E(X)]^2$. ASI, PARA UNA VARIABLE ALEATORIA DISCRETA

$$\sigma^2(X) = VAR(X) = \sum_{i=1}^{i=n} (x_i - E(X))^2 \, P_X(x_i)$$

Y PARA UNA CONTINUA

$$\sigma^2(X) = VAR(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f_X(x)\,dx$$

DESARROLLANDO EL INTEGRANDO DE ESTA ULTIMA ECUACION:

$$\sigma^2(X) = \int_{-\infty}^{\infty} (x^2 - 2xE(X) + E^2(X)) f_X(x)\,dx$$

$$= \int_{-\infty}^{\infty} x^2 f_X(x)\,dx - 2E(X)\int_{-\infty}^{\infty} x f_X(x)\,dx + E^2(X)\int_{-\infty}^{\infty} f_X(x)\,dx = E[X^2] - E^2[X]$$

ES DECIR, LA VARIANCIA SE PUEDE CALCULAR COMO LA DIFERENCIA DEL VALOR MEDIO CUADRATICO Y EL CUADRADO DE LA MEDIA DE X.

OTRAS MEDIDAS DE DISPERSION DE LA VARIABLE ALEATORIA X SON LA *DESVIACION ESTANDAR*, $\sigma(X)$, LA CUAL ES IGUAL A LA RAIZ CUADRADA DE LA VARIANCIA, Y EL *COEFICIENTE DE VARIACION* QUE SE DEFINE COMO

$$\nu(X) = \sigma(X)/E(X)$$

EJEMPLO

EN LA SIGUIENTE TABLA SE CALCULA LA VARIANCIA DE LA VARIABLE ALEATORIA CUYA DENSIDAD DE PROBABILIDADES SE PRESENTO EN EL EJEMPLO ANTERIOR

x_i - E(X) EN MICRAS	$P_X(x_i)$	$(x_i - E(X))^2 P_X(x_i)$, EN MICRAS
-4 167	6/60	1 740 000
-3 167	2/60	333 000
-2 167	4/60	313 000
-1 167	8/60	181 000
- 167	13/60	6 000
833	12/60	139 000
1 833	7/60	390 000
2 833	4/60	531 000
3 833	2/60	487 000
4 833	2/60	687 000
	TOTAL:	4 798 000 MICRAS^2 $=\sigma^2(X)$

LA DESVIACION ESTANDAR Y EL COEFICIENTE DE VARIACION DE ESTA VARIABLE ALEATORIA SON, RESPECTIVAMENTE,

$$\sigma(X) = \sqrt{4\ 798\ 000} \approx 2\ 200 \text{ MICRAS, Y } v(X) = \sigma(X)/E(X) = \frac{2\ 200}{4\ 167} = 0.528$$

EJEMPLO

SI X ES UNA VARIABLE ALEATORIA CON DISTRIBUCION DE PROBABILIDADES EXPONENCIAL, CALCULAR SU VARIANCIA, DESVIACION ESTANDAR Y COEFICIENTE DE VARIACION:

$$\sigma^2(X) = E(X-E[X])^2 = \int_{-\infty}^{\infty}(x-E[X])^2\lambda e^{-\lambda x}dx = \lambda \int_0^{\infty}(x^2-2xE[X]+E^2[X])e^{-\lambda x}dx$$

$$= \lambda \int_0^{\infty}x^2e^{-\lambda x}dx - 2E[X]\lambda\int_0^{\infty}xe^{-\lambda x}dx + E^2[X]\int_0^{\infty}\lambda e^{-\lambda x}dx$$

$$= \frac{2}{\lambda^2} - 2\,\frac{1}{\lambda}\,\frac{1}{\lambda} + \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

YA QUE $E(X) = 1/\lambda$.

USANDO LA FORMULA $\sigma^2(X) = E[X^2] - E^2[X]$, Y TOMANDO EN CUENTA QUE $E[X^2] = 2/\lambda^2$ SE OBTIENE:

$$\sigma^2(X) = 2/\lambda^2 - 1/\lambda^2 = 1/\lambda^2$$

EN CONSECUENCIA, LA DESVIACION ESTANDAR ES

$$\sigma(X) = \sqrt{1/\lambda^2} = 1/\lambda$$

Y EL COEFICIENTE DE VARIACION

$$\nu(X) = \sigma(X)/E(X) = \frac{\frac{1}{\lambda}}{\frac{1}{\lambda}} = 1$$

EJEMPLO

SEA Y UNA VARIABLE ALEATORIA CON DENSIDAD DE PROBABILIDADES TRIANGULAR DADA POR

$$f_Y(y) = \frac{1}{6}y + \frac{1}{3} \qquad \text{SI} \quad -2 \leq y \leq 0$$

$$f_Y(y) = \frac{-1}{12}y + \frac{1}{3} \qquad \text{SI} \quad 0 \leq y \leq 4$$

$$f_Y(y) = 0 \qquad \text{SI} \quad y \leq -2 \text{ O } y \geq 4$$

CALCULAR LA VARIANCIA, LA DESVIACION ESTANDAR Y EL COEFICIENTE DE VARIACION.

CALCULAREMOS PRIMERO EL VALOR MEDIO CUADRATICO PARA LUEGO APLICAR LA ECUACION $\sigma^2(Y) = E(Y^2) - E^2(Y)$

$$E[Y^2] = \int_{-2}^{0} y^2\left(\frac{1}{6}y + \frac{1}{3}\right)dy + \int_{0}^{4} y^2\left(-\frac{y}{12} + \frac{1}{3}\right)dy = \left[\frac{y^4}{24} + \frac{y^3}{9}\right]_{-2}^{0} + \left[\frac{-y^4}{48} + \frac{y^3}{9}\right]_{0}^{4} = 2$$

$$\sigma^2(Y) = 2-(2/3)^2 = 14/9$$

$$\sigma(Y) = 1.25$$

$$v(Y) = 1.25/(2/3) = 1,88$$

DISTRIBUCIONES PARTICULARES

DISTRIBUCION BINOMIAL O DE BERNOULLI

LA DISTRIBUCION BINOMIAL O DE BERNOULLI SE EMPLEA COMO DENSIDAD DE PROBABILIDADES DE VARIABLES ALEATORIAS DISCRETAS ASOCIADOS A EXPERIMENTOS EN LOS QUE SOLO HAY (O SOLO IMPORTAN) DOS RESULTADOS POSIBLES, UNO DE LOS CUALES USUALMENTE SE DENOMINA "EXITO" Y, EL OTRO, "FRACASO".

SEAN p= PROBABILIDAD DE OBSERVAR "EXITO" AL REALIZAR UNA VEZ EL EXPERIMENTO

q= PROBABILIDAD DE "FRACASO" = 1-p

X= VARIABLE ALEATORIA "NUMERO DE EXITOS OBSERVADOS AL REPETIR n VECES EL EXPERIMENTO "CON REEMPLAZO"

LA DISTRIBUCION DE PROBABILIDADES BINOMIAL ES

$$f(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x} \quad : \quad x = 0, 1, \ldots, n$$

SE PUEDE DEMOSTRAR QUE LOS PARAMETROS DE ESTA DISTRIBUCION SON

$$E(X) = np \ , \quad \sigma^2(X) = npq$$

EJEMPLO

SI SE LANZA AL AIRE SEIS VECES UNA MONEDA HOMOGENEA,

A) ¿CUAL ES LA PROBABILIDAD DE OBTENER DOS "CARAS"?

B) ¿CUAL ES LA PROBABILIDAD DE OBTENER POR LO MENOS CUATRO "CARAS" $(X \geq 4)$?

C) ¿CUANTO VALEN LA ESPERANZA Y LA DESVIACION ESTANDAR?

A) PUESTO QUE LA MONEDA ES HOMOGENEA SE TIENE $p=1/2$ Y $q=1-1/2=1/2$, DONDE p ES LA PROBABILIDAD DE OBSERVAR "CARA" (CARA = EXITO) EN UN LANZAMIENTO. POR TANTO

$$P[X = 2] = f_X(2) = \frac{6!}{2!\,(6-2)!}\underbrace{(1/2)^{6-2}}_{(\frac{1}{2})^2} = \frac{6!}{2!\;4!}(1/2)^6 = \frac{15}{64}$$

B) PARA QUE SE CUMPLA $X \geq 4$ EN SEIS LANZAMIENTOS, SE NECESITA QUE SE OBSERVEN 4, 5 o 6 CARAS. PUESTO QUE ESTOS TRES EVENTOS SON MUTUAMENTE EXCLUSIVOS, SE TIENE

$$P[X \geq 4] = f_X(4) + f_X(5) + f_X(6)$$

CALCULANDO LOS TRES SUMANDOS COMO EN LA PREGUNTA ANTERIOR, RESULTA

$$P[X \geq 4] = \frac{6!}{4!\;2!}(1/2)^4(1/2)^{6-4} + \frac{6!}{5!\;1!}(1/2)^5(1/2)^{6-5} + \frac{6!}{6!\;0!}(1/2)^6(1/2)^{6-6}$$

$$= \frac{15}{64} + \frac{6}{64} + \frac{1}{64} = \frac{11}{32}$$

C) $E[X] = np = 6(1/2) = 3$

$\sigma^2[X] = npq = 6(1/2)(1/2) = 3/2$, $\sigma(X) = \sqrt{3/2} = 1.22$

DISTRIBUCION HIPERGEOMETRICA

CUANDO SE TIENE UNA VARIABLE ALEATORIA DISCRETA CUYO ESPACIO DE EVENTOS TIENE SOLO DOS ELEMENTOS, DIGAMOS S={EXITO, FRACASO}, Y SE REALIZA UN MUESTREO *SIN REEMPLAZO*, ENTONCES LOS RESULTADOS DE CADA EXPERIMENTO *NO SON INDEPENDIENTES NI LA PROBABILIDAD DE EXITO PERMANECE CONSTANTE*, COMO EN LA DISTRIBUCION BINOMIAL, POR LO QUE ESTA ULTIMA NO ES APLICABLE.

SEA X LA VARIABLE ALEATORIA NUMERO DE EXITOS OBSERVADOS AL REPETIR n VECES EL EXPERIMENTO CONSISTENTE EN EXTRAER, SIN REEMPLAZO, ELEMENTOS DE UN LOTE QUE TIENE N OBJETOS DE LOS CUALES M SON "EXITOS". EL NUMERO DE ELEMENTOS QUE TIENE EL ESPACIO DE EVENTOS DEL EXPERIMENTO ES

$$N(S) = {}_{N}C_{n}$$

EL NUMERO, N({X=x}), DE MANERAS POSIBLES E IGUALMENTE PROBABLES DE OBTENER x EXITOS ES

EN EL LOTE

N ELEMENTOS

$\underbrace{\odot\odot\odot\cdots\odot\odot}_{\text{M EXITOS}}\,\underbrace{\odot\odot\odot\cdots\odot\odot}_{\text{(N-M) FRACASOS}}$

EN LA MUESTRA

n ELEMENTOS

$\underbrace{\odot\odot\odot\cdots\odot\odot}_{\text{x EXITOS QUE SE PUEDEN ELEGIR DE } {}_{M}C_{x} \text{ MANERAS}}\,\underbrace{\odot\odot\cdots\odot}_{\text{(n-x) FRACASOS QUE SE PUEDEN ELEGIR DE } {}_{N-M}C_{n-x} \text{ MANERAS}}$

CADA ELECCION POSIBLE DE x EXITOS SE COMBINA CON CADA ELECCION POSIBLE DE (n-x) FRACASOS; POR LO TANTO, EL NUMERO TOTAL DE MANERAS DE OBTENER x EXITOS EN n EXTRACCIONES SIN REEMPLAZO ES

$$N(\{X=x\}) = ({}_{M}C_{x})({}_{N-M}C_{n-x})$$

POR LO TANTO

$$P(\{X=x\}) = f_X(x) = \frac{({}_MC_x)\,({}_{N-M}C_{n-x})}{{}_NC_n} = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \quad x=0,1,\ldots,n$$

EN DONDE $\binom{M}{x} = \frac{M!}{x!\,(M-x)!}$, $\binom{N-M}{n-x} = \frac{(N-M)!}{(n-x)!\,(N-M-n+x)!}$

Y $\binom{N}{n} = \frac{N!}{n!\,(N-n)!}$

QUE SE CONOCE COMO *DISTRIBUCION HIPERGEOMETRICA*, LA MEDIA Y LA VARIANCIA DE ESTA DISTRIBUCION SON

$$E(X) = \sum_{x=0}^{n} x\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = nM/N$$

Y

$$\sigma^2(X) = \sum_{x=0}^{n} \left(x-\frac{nM}{N}\right)^2 \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = \frac{Mn\,(N-M)\,(N-n)}{N^2\,(N-1)}$$

RESPECTIVAMENTE.

EJEMPLO

EN UN PROBLEMA DE CONTROL ESTADISTICO DE CALIDAD, SE TIENE UN LOTE DE 100 TRANSFORMADORES DE CORRIENTE ELECTRICA, DE LOS CUALES 40 SON DEFECTUOSOS (NO CUMPLEN LAS NORMAS DE FABRICACION). ¿CUAL ES LA PROBABILIDAD DE OBTENER UNO DEFECTUOSO DE TRES SELECCIONADOS AL AZAR SIN REEMPLAZO?

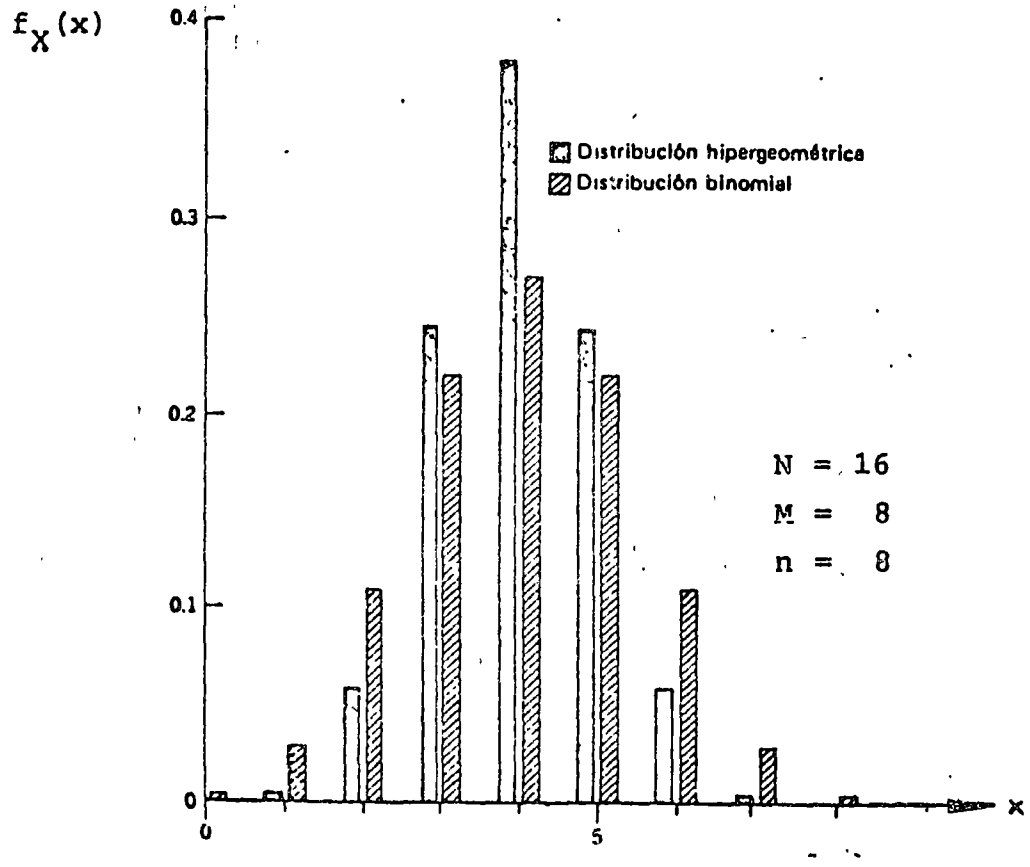
$$P[X=1] = \frac{\binom{40}{1}\binom{100-40}{3-1}}{\binom{100}{3}} = \frac{\binom{40}{1}\binom{60}{2}}{\binom{100}{3}}$$

$$= \frac{\frac{40!}{39!\times 1!} \times \frac{60!}{58!\times 2!}}{\frac{100!}{97!\times 3!}} = 0.438$$

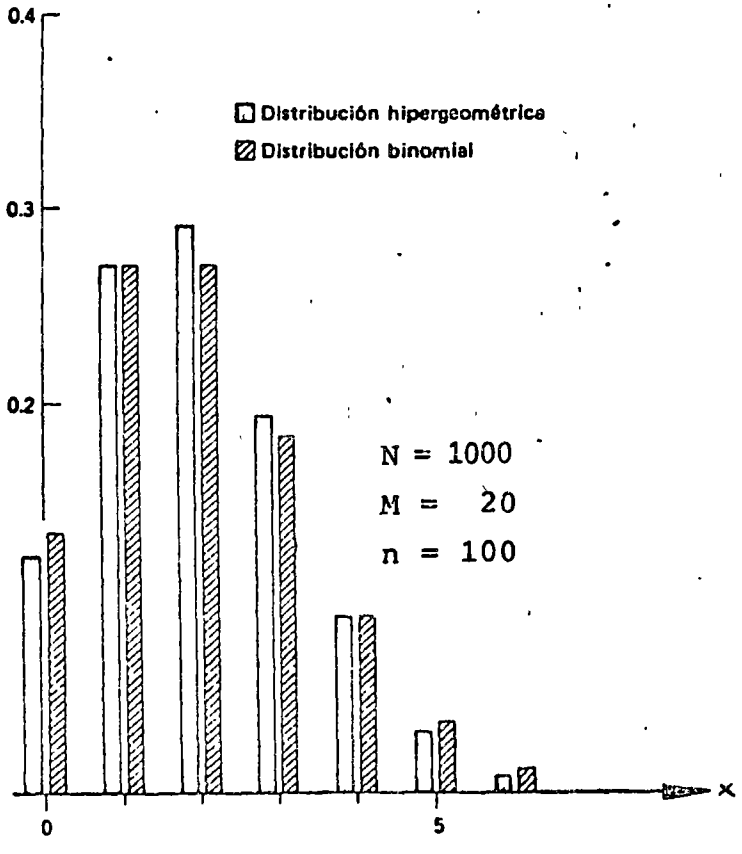
CUANDO *N ES GRANDE* Y *n PEQUEÑO*, LA DISTRIBUCION BINOMIAL SE PUEDE USAR *COMO APROXIMACION* DE LA HIPERGEOMETRICA. DE ESTA APROXIMACION SE HECHA MANO CUANDO LOS CALCULOS CON ESTA ULTIMA RESULTAN TEDIOSOS.

EN EL CASO DEL EJEMPLO ANTERIOR, SI SE USA LA DENSIDAD BINOMIAL SE OBTIENE, CON p=40/100 = 0.40 Y n=3

$$P[X=1] = \frac{3!}{1!\ 2!}(0.40)^1(0.60)^2 = 0.432$$



COMPARACION DE LAS DISTRIBUCIONES HIPERGEOMETRICA Y BINOMIAL

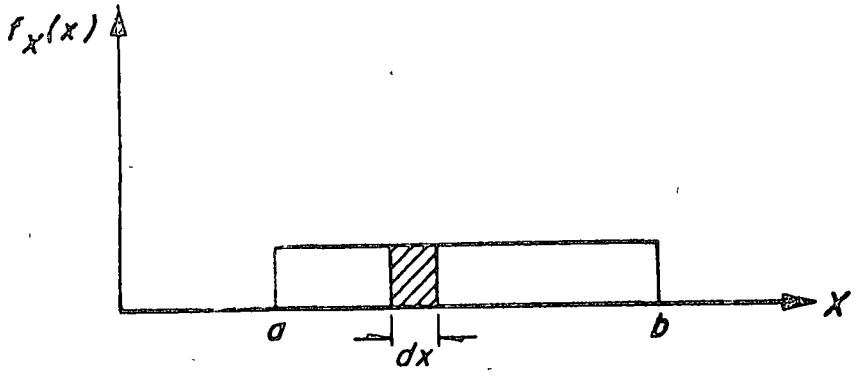


DISTRIBUCION UNIFORME

SE DICE QUE UNA VARIABLE ALEATORIA CONTINUA, X, TIENE *DISTRIBUCION UNIFORME* ENTRE X = a Y X = b(b>a) SI

$$f_X(x) = \text{CONSTANTE} = \frac{1}{b - a}$$

LO QUE SIGNIFICA QUE LA PROBABILIDAD DE OBTENER UN VALOR ENTRE x Y x + dx ES LA MISMA PARA CUALQUIER x COMPRENDIDA ENTRE a Y b. LA GRAFICA DE DICHA DISTRIBUCION ES



Distribución uniforme de una variable aleatoria continua

LA ESPERANZA Y LA VARIANCIA DE LA DISTRIBUCION UNIFORME SE CALCULAN DE LA SIGUIENTE MANERA:

$$E[X] = \int_a^b x \frac{1}{b-a}\,dx = \left[\frac{x^2}{2(b-a)}\right]_a^b = \frac{b^2 - a^2}{2(b-a)} = (b + a)/2$$

$$\sigma^2(X) = \int_a^b (x - E[X])^2 \frac{1}{b-a}\,dx = \int_a^b \frac{x^2}{b-a}\,dx + \int_a^b \frac{(E[X])^2}{b-a}\,dx -$$

$$- \int_a^b \frac{2xE[X]}{a-b}\,dx$$

$$= \left[\frac{x^3}{3(b-a)}\right]_a^b + \left[\frac{(E[X])^2}{b-a}\,x\right]_a^b - \left[\frac{2\,E[X]}{b-a}\,\frac{x^2}{2}\right]_a^b =$$

$$= \frac{b^3 - a^3}{3(b-a)} + (E[X])^2 - E[X](b + a) = \frac{(b-a)^2}{12}$$

DISTRIBUCION NORMAL

UNA DE LAS DISTRIBUCIONES DE VARIABLES ALEATORIAS CONTINUAS MAS UTIL ES LA *DISTRIBUCION NORMAL* O DE *GAUSS*, DEFINIDA POR LA ECUACION

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

DONDE μ ES LA MEDIA Y σ LA DESVIACION ESTANDAR DE X.

SI SE HACE LA TRANSFORMACION

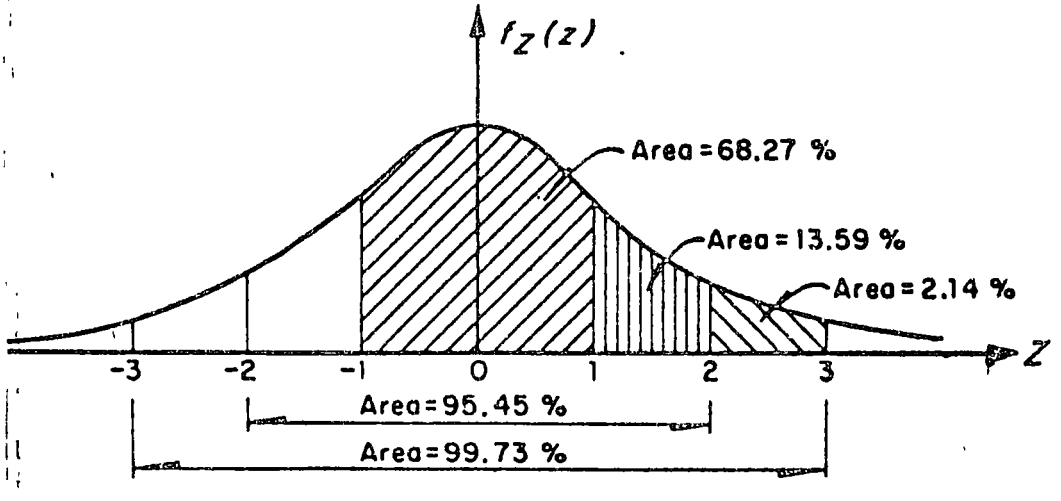
$$Z = (X-\mu)/\sigma$$

ENTONCES LA ECUACION ANTERIOR SE REDUCE A LA LLAMADA *FORMA ESTANDAR*, CUYA ECUACION ES

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-Z2/2}$$

EN ESTE CASO LA VARIABLE ALEATORIA Z TIENE DISTRIBUCION NORMAL CON *MEDIA IGUAL A CERO Y VARIANCIA IGUAL A UNO.*

EXISTEN TABLAS PARA CALCULAR LAS PROBABILIDADES DE UNA VARIABLE ASOCIADA A UNA DISTRIBUCION NORMAL ESTANDAR. EN LA SIGUIENTE FIGURA SE MUESTRA LA FORMA DE CAMPANA DE ESTA DISTRIBUCION, OBSERVANDOSE LA SIMETRIA RESPECTO A Z=E(Z)=0.



Distribución normal de una variable aleatoria contin

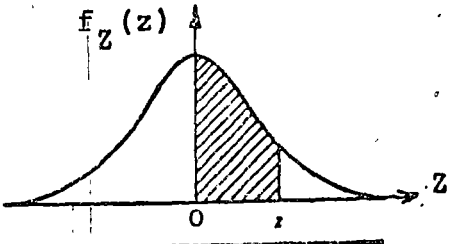
LA UTILIDAD DE LA DISTRIBUCION NORMAL ESTANDAR RADICA EN QUE

$$P[x_1 \leq X \leq x_2] = \int_{x_1}^{x_2} f_X(x)dx = P[z_1 \leq Z \leq z_2] = \int_{z_1}^{z_2} f_Z(z)dz$$

DONDE

$$z_1 = \frac{z_1 - \mu}{\sigma} \quad Y \quad z_2 = \frac{x_2 - \mu}{\sigma}$$

TABLA 5. AREAS BAJO LA CURVA NORMAL ESTANDAR ENTRE 0 Y z



z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0754
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2258	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2996	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3189	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	46.33
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	4999	.4999
3.7	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000	.5000	.5000	5000	.5000	.5000	.5000	.5000	.5000	.5000

EJEMPLO

COMO RESULTADO DE UNA LARGA SERIE DE EXPERIMENTOS PROBANDO A COMPRESION SIMPLE CILINDROS DE CONCRETO, SE HA ESTIMADO QUE LA ESPERANZA DE LA RESISTENCIA ES DE 240 KG/CM^2 Y LA DESVIACION ESTANDAR DE 30 KG/CM^2.

A) ¿CUAL ES LA PROBABILIDAD DE QUE OTRO CILINDRO TOMADO AL AZAR RESISTA MENOS DE 240 KG/CM^2?

B. ¿CUAL ES LA PROBABILIDAD DE QUE RESISTA MAS DE 330 KG/CM^2?

C) ¿CUAL ES LA PROBABILIDAD DE QUE SU RESISTENCIA ESTE EN EL INTERVALO DE 210 A 240 KG/CM^2?

SUPONGASE QUE LA DISTRIBUCION DE PROBABILIDADES ES NORMAL.

A) PARA EMPLEAR LAS TABLAS DE DISTRIBUCION NORMAL ES NECESARIO ESTANDARIZAR LA VARIABLE X, EMPLEANDO $\mu=240$ Y $\sigma=30$, CON $x_1=240$:

$$z_1 = \frac{240 - 240}{30} = 0$$

RECURRIENDO A LA TABLA DE LA DISTRIBUCION NORMAL SE OBTIENE

$$P[X \leq 240] = P[Z \leq 0] = 0.5$$

O SEA, LA PROBABILIDAD QUE CORRESPONDE AL AREA SOMBREADA DE LA SIGUIENTE FIGURA:

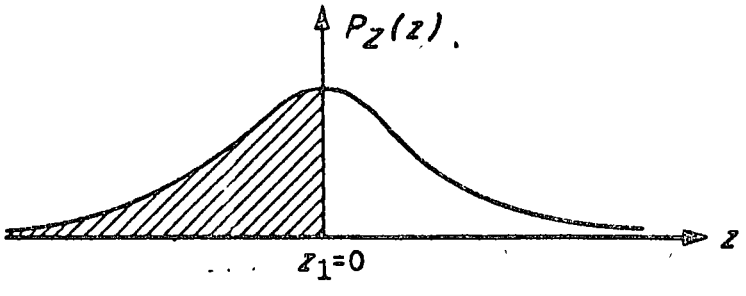


Fig 16. Distribución normal correspondiente al inciso c del ejemplo

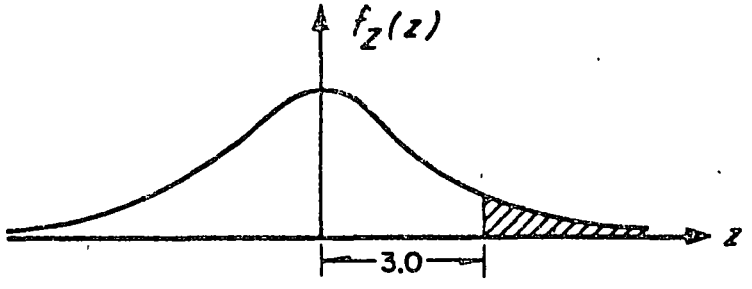
B) EL VALOR ESTANDARIZADO DE LA VARIABLE, PARA x_1=330 KG/CM^2, ES

$$z_1 = \frac{330 - 240}{30} = 3$$

POR LO QUE

$$P[X \geq 330] = P[Z \geq 3] = 1-0.9987 = 0.0013$$

QUE ES EL AREA SOMBREADA DE LA SIGUIENTE FIGURA:



Distribución normal correspondiente al inciso b del ejemplo

C) LOS VALORES ESTANDARIZADOS DE LA VARIABLE, PARA x_1=210 Y x_2=240 SON:

$$z_1 = \frac{210 - 240}{30} = -1$$

$$z_2 = \frac{240 - 240}{30} = 0$$

POR LO QUE

$$P[210 \leq X \leq 240] = P[-1 \leq Z \leq 0] = 0.3413$$

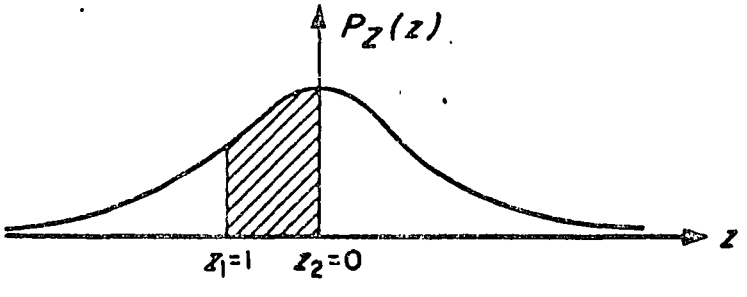


Fig 16. Distribución normal correspondiente al inciso c del ejemplo

TEOREMA CENTRAL DEL LIMITE

SEAN LAS VARIABLES ALEATORIAS $X_1, X_2, \ldots, X_k$, CON DENSIDADES DE PROBABILIDADES ARBITRARIAS, CUYA SUMA SE DENOTARA COMO W, ES DECIR

$$W = X_1 + X_2 + \ldots + X_k$$

ES POSIBLE DEMOSTRAR EL TEOREMA DENOMINADO *TEOREMA CENTRAL DEL LIMITE*, CUYO ENUNCIADO INDICA QUE CONFORME AUMENTA EL NUMERO DE VARIABLES INVOLUCRADAS EN LA SUMA ANTERIOR (AL AUMENTAR k), LA DENSIDAD DE PROBABILIDADES DE W TIENDE A SER LA DISTRIBUCION NORMAL. ADEMAS SE PUEDE DEMOSTRAR QUE SI TODAS LAS VARIABLES $X_1, X_2, \ldots, X_k$ TIENEN DISTRIBUCION NORMAL, ENTONCES, RIGUROSAMENTE, W TAMBIEN LA TIENE, INDEPENDIENTEMENTE DEL NUMERO DE VARIABLES QUE APAREZCAN EN LA SUMA.

A PARTIR DEL TEOREMA DEL LIMITE CENTRAL SE DEMUESTRA QUE LA DISTRIBUCION DE BERNOULLI SE PUEDE APROXIMAR MEDIANTE LA NORMAL CUANDO EL NUMERO DE REPETICIONES DEL EXPERIMENTO ES GRANDE (30 O MAS), CON LO CUAL SE LOGRA UN AHORRO CONSIDERABLE DE LABOR NUMERICA EN LA SOLUCION DE ALGUNOS PROBLEMAS. PARA MEJORAR ESTA APROXIMACION, CONVIENE EFECTUAR UNA CORRECCION POR CONTINUIDAD, LA CUAL SE JUSTIFICA POR USAR UNA DISTRIBUCION CONTINUA EN VEZ DE UNA DISCRETA, SUMANDO O RESTANDO, SEGUN SEA EL CASO, 0.5 AL VALOR DE X QUE SE USE. POR EJEMPLO, SI SE DESEA CUANTIFICAR LA PROBABILIDAD DE QUE DE 2000 ENSAYES SE LOGREN DE 3 A 6 EXITOS, LOS LIMITES REALES QUE SE DEBEN USAR AL APLICAR LA DISTRIBUCION CONTINUA SON $x_1=2.5$ Y $x_2=6.5$.

ESTADISTICA DESCRIPTIVA

DATO Y OBSERVACION: ES EL RESULTADO DE REALIZAR UN EXPERIMENTO.

MUESTRA: ES UNA COLECCION DE DATOS

MUESTREO: PROCESO DE ADQUISICION DE UNA MUESTRA

MUESTREO
- *CON REEMPLAZO*:- CUANDO CADA ELEMENTO OBSERVADO SE REINTEGRA AL LOTE DEL CUAL FUE EXTRAIDO ANTES DE EXTRAER EL SIGUIENTE.
- *SIN REEMPLAZO*:- CUANDO CADA ELEMENTO OBSERVADO NO SE REINTEGRA AL LOTE.

POBLACION: TOTAL DE DATOS QUE SE PUEDEN OBTENER AL REALIZAR UNA SECUENCIA EXHAUSTIVA DE EXPERIMENTOS

POBLACION
- *DISCRETA*:- TIENE UN NUMERO FINITO O UN NUMERO INFINITO NUMERABLE DE DATOS POSIBLES
- *CONTINUA*:- TIENE UN NUMERO INFINITO NO NUMERABLE DE DATOS POSIBLES

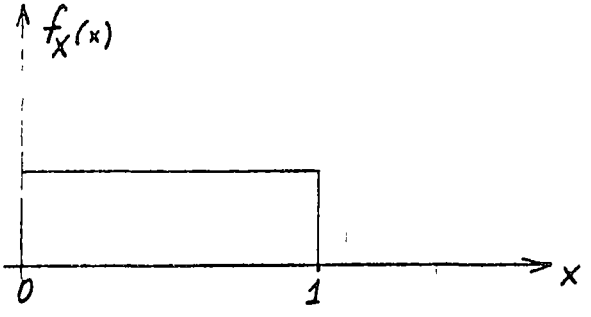
EJEMPLOS

1. EXPERIMENTO: LANZAMIENTO DE UNA MONEDA DIEZ VECES
   POBLACION: SUCESION INFINITA NUMERABLE DE "CARAS" Y "CRUCES" (DISCRETA)
   MUESTRA: GRUPO DE 10 OBSERVACIONES

2. EXPERIMENTO: MEDICION DE LA PRECIPITACION PLUVIAL MAXIMA DIARIA EN LA CIUDAD DE MEXICO DURANTE DIEZ AÑOS
   POBLACION: SUCESION INFINITA NO NUMERABLE DE VALORES (CONTINUA)
   MUESTRA: GRUPO DE 3652 OBSERVACIONES (TOMANDO DOS AÑOS BISIESTOS DE 29 DIAS EN FEBRERO)

MUESTRA ALEATORIA: ES UNA MUESTRA OBTENIDA DE TAL MANERA QUE *TODOS* LOS ELEMENTOS DE LA POBLACION TIENEN LA *MISMA* PROBABILIDAD DE SER OBSERVADOS Y, ADEMAS, LA OBSERVACION DE UN ELEMENTO *NO AFECTA* LA PROBABILIDAD DE OBSERVAR CUALQUIER OTRO, ES DECIR, SI SON INDEPENDIENTES.

TABLA DE NUMEROS ALEATORIOS: ES UNA TABLA QUE CONTIENE NUMEROS QUE CONSTITUYEN UNA MUESTRA ALEATORIA OBTENIDA DE UNA DISTRIBUCION DE PROBABILIDADES UNIFORME, QUE GENERALMENTE CORRESPONDE A UNA VARIABLE ALEATORIA QUE PUEDE ASUMIR VALORES ENTRE 0 Y 1, MULTIPLICADOS POR 10^r, EN DONDE r ES EL NUMERO DE DIGITOS QUE SE DESEA TENGAN LOS NUMEROS.



LAS TABLAS QUE SE USEN PARA OBTENER UNA MUESTRA ALEATORIA DEBEN CONTENER NUMEROS CON MAYOR NUMERO DE DIGITOS QUE LOS QUE TIENE EL TOTAL DE ELEMENTOS DE LA POBLACION QUE SE VA A MUESTREAR. POR EJEMPLO, SI SE VA A OBTENER UNA MUESTRA ALEATORIA DE UN LOTE DE LENTES PARA MICROSCOPIO QUE TIENE 10,000 ELEMENTOS, LA TABLA QUE SE USE DEBERA TENER NUMEROS ALEATORIOS CON 5 O MAS DIGITOS.

METODO DE MUESTREO ALEATORIO

1. SE ENUMERAN LOS ELEMENTOS DE LA POBLACION
2. SE FIJA EL CRITERIO DE SELECCION DE LOS NUMEROS ALEATORIOS (POR EJEMPLO, SE DEFINE QUE RENGLONES Y QUE COLUMNAS SE VAN A LEER)
3. SE INDICA QUE DIGITOS SE VAN A ELIMINAR EN CASO DE QUE LOS NUMEROS DE LA TABLA TENGAN MAS DIGITOS QUE LOS NECESARIOS
4. SE LEEN LOS NUMEROS, DE ACUERDO CON LO FIJADO EN LOS PUNTOS 2 Y 3, Y SE EXTRAEN DEL LOTE LOS ELEMENTOS QUE TIENEN LOS NUMEROS LEIDOS. ESTOS CONSTITUYEN LA MUESTRA FISICA CON LA CUAL REALIZAR LOS EXPERIMENTOS. LAS OBSERVACIONES CONSTITUIRAN LA MUESTRA ALEATORIA DESEADA.

NOTA: TODOS LOS NUMEROS QUE SE REPITAN SE CONSIDERAN SOLO UNA VEZ. TAMBIEN SE ELIMINAN LOS NUMEROS MAYORES DEL TAMAÑO DEL LOTE.

EJEMPLO

SE TIENE UN LOTE DE 1,000 TRANSISTORES NUMERADOS DEL UNO AL M.L, CUYA CALIDAD SE VA A VERIFICAR ESTADISTICAMENTE, PARA LO CUAL SE DECIDE TOMAR UNA MUESTRA DE 40 ELEMENTOS Y MEDIR SU AMPLIFICACION USANDO LA TABLA DE NUMEROS ALEATORIOS ANEXA, CON EL CRITERIO DE TOMAR TODOS LOS RENGLONES IMPARES ELIMINANDO EL ULTIMO DIGITO, LA MUESTRA FISICA SERIAN LOS TRANSISTORES CORRESPONDIENTES A LOS NUMEROS 0415, 0006, 0394, 0998, 0530, 0160, ETC.

TABLA DE NUMEROS ALEATORIOS

Renglón \ Columna	1	2	3	4	5	6	7	8	9	10	11
1	16408	81899	04153	53381	79401	21438	83035	92350	36693	31238	59649
2	18629	81953	05520	91962	04739	13092	37662	94822	94730	06496	35090
3	73115	47498	47498	87637	99016	00060	88824	71013	18735	20286	23153
4	57491	16703	23167	49323	45021	33132	12544	41035	80780	45393	44812
5	30405	03946	23792	14422	15059	45799	22716	19792	09983	74353	68668
6	16631	35006	85900	32388	52390	52390	16815	69298	38732	38480	73817
7	96773	20206	42559	78985	05300	22164	24369	54224	35083	19687	11052
8	38935	64202	14349	82674	66523	44133	00697	35552	35970	19124	63318
9	31624	76384	17403	03941	44167	64486	64758	75366	76554	01601	12614
10	78919	19474	23632	27889	47914	02584	37680	20801	72152	39339	34806

AGRUPAMIENTO DE DATOS

FRECUENCIA DE UN EVENTO: ES EL NUMERO DE VECES QUE OCURRE EL EVENTO AL OBTENER UNA MUESTRA DE LA POBLACION CORRESPONDIENTE.

FRECUENCIA RELATIVA DE UN EVENTO: ES EL COCIENTE DE SU FRECUENCIA ENTRE EL TOTAL DE ELEMENTOS (TAMAÑO) DE LA MUESTRA.

FRECUENCIA RELATIVA ACUMULADA: ES LA ACUMULACION (SUMA) DE LAS FRECUENCIAS RELATIVAS HASTA UN VALOR DADO, PARTIENDO DEL VALOR (O DEL INTERVALO) MAS PEQUEÑO. EN OTRAS PALABRAS, ES LA FRECUENCIA DE VALORES MENORES O IGUALES QUE UN VALOR DADO.

FRECUENCIA COMPLEMENTARIA: ES LA FRECUENCIA DE VALORES MAYORES QUE UN VALOR DADO = NUMERO DE DATOS - FRECUENCIA ACUMULADA.

DISTRIBUCION DE FRECUENCIAS

CON OBJETO DE FACILITAR LA INTERPRETACION DE LOS DATOS QUE SE TIENEN EN UNA MUESTRA, ES CONVENIENTE AGRUPARLOS POR VALORES O POR INTERVALOS DE VALORES, FORMANDO ASI UNA TABLA DE DISTRIBUCION DE FRECUENCIAS.

PARA FACILITAR EL CALCULO DE LAS FRECUENCIAS ES UTIL ORDENAR LOS DATOS EN FORMA CRECIENTE O DECRECIENTE DE VALORES, FORMANDO ASI UNA *TABLA DE DATOS ORDENADOS*.

EJEMPLO

EN UNA ESCUELA SECUNDARIA SE LES APLICO A 30 PROFESORES UN EXAMEN SOBRE PEDAGOGIA. LAS CALIFICACIONES (DATOS) QUE SE OBTUVIERON FUERON (YA ESTAN ORDENADOS EN FORMA CRECIENTE)

57, 59, 65, 67, 67, 67, 69, 72, 73, 73, 77, 78, 78,

A B C

81, 81, 83, 83, 83, 84, 84, 87, 88, 89, 89, 91, 91, 93,

D E

95, 97, 99

E

AGRUPAMIENTO DE VALORES

CALIFICACION	FRECUENCIA	FRECUENCIA RELATIVA	FRECUENCIA RELATIVA ACUMULADA
57	1	1/30	1/30
59	1	1/30	2/30
65	1	1/30	3/30
67	3	3/30	6/30
69	1	1/30	7/30
72	1	1/30	8/30
73	2	2/30	10/30
77	1	1/30	11/30
78	2	2/30	13/30
81	2	2/30	15/30
83	3	3/30	18/30
84	2	2/30	20/30
87	1	1/30	21/30
88	1	1/30	22/30
89	2	2/30	24/30
91	2	2/30	26/30
93	1	1/30	27/30
95	1	1/30	28/30
97	1	1/30	29/30
99	1	1/30	30/30=1
	Σ=30	Σ=30/30=1	

AGRUPAMIENTO POR INTERVALOS

LIMITES DE CLASES: SON LOS VALORES MINIMO Y MAXIMO DE CADA INTERVALO

MARCAS DE CLASE: SON LOS VALORES MEDIOS DE CADA INTERVALO DE CLASE

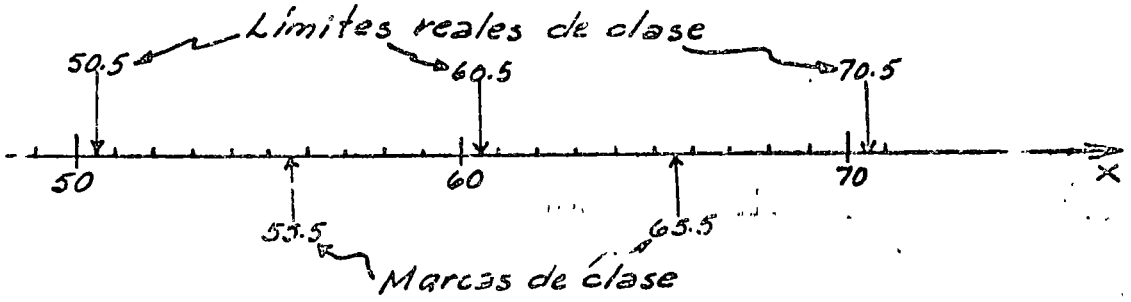
LIMITES REALES DE CLASE: SON LOS VALORES MINIMO Y MAXIMO QUE SON FRONTERA ENTRE LOS INTERVALOS. ESTOS DEBEN TENER UNA CIFRA DECIMAL MAS QUE LOS DATOS.

EVENTO (INTERVALO DE CALIFICACIONES)	ELEMENTOS OBSERVADOS	FRECUENCIA	FRECUENCIA RELATIVA
A = {51-60}	57,59	2	2/30
B = {61-70}	65,67,67,67,69	5	5/30
C = {71-80}	72,73,73,77,78,78	6	6/30
D = {81-90}	81,81,83,83,83,84, 84,87,88,89,89	11	11/30
E = {91-100}	91,91,93,95,97,99	6	6/30
		Σ=30	30/30=1

LIMITES INFERIORES DE CLASE | LIMITES SUPERIORES DE CLASE

EVENTO	LIMITES DE CLASE		LIMITES REALES DE CLASE		MARCAS DE CLASE
	INFERIOR	SUPERIOR	INFERIOR	SUPERIOR	
A	51	60	50.5	60.5	55.5
B	61	70	60.6	70.5	65.5
C	71	80	70.5	80.5	75.5
D	81	90	80.5	90.5	85.5
E	91	100	90.5	100.5	95.5

Evento	Elementos corresp. a los intervalos	Frecuencia	Frecuencia relativa	Frecuencia acumulada	Frecuencia relativa acumulada
A: 51-60	59,57	2	2/30=0.067(6.7%)	2	0.067
B: 61-70	67,65,69,67,67	5	5/30=0.166(16.6%)	2+5=7	0.067+0.166=0.233
C: 71-80	72,73,73,77,78,78,	6	6/30=0.200(20%)	7+6=13	0.233+0.200=0.433
D: 81-90	83,88,84,89,83,84, 89,87,81,83,81	11	11/30=0.367(36.7%)	13+11=24	0.433+0.367=0.800
E: 91-100	99,91,97,95,91,93	6	6/30=0.200(20%)	24+6=30	0.800+0.200=1.000
		30	1.000		



$A = \{X: 50.5 < X \leq 60.5\}$

$B = \{X: 60.5 < X \leq 70.5\}$

$C = \{X: 70.5 < X \leq 80.5\}$

$D = \{X: 80.5 < X \leq 90.5\}$

$E = \{X: 90.5 < X \leq 100.5\}$

LIMITES REALES INFERIORES DE CLASE

LIMITES REALES SUPERIORES DE CLASE

A MAYOR NUMERO DE DATOS SE REQUIERE MAYOR NUMERO DE INTERVALOS, PERO SE RECOMIENDA QUE ESTE NUMERO ESTE ENTRE 5 Y 20, SUPONIENDO QUE EN PROMEDIO CAIGAN 5 O MAS ELEMENTOS EN CADA INTERVALO. ASI, SI SE TIENEN 30 DATOS, SE RECOMIENDA USAR 30/5=6 INTERVALOS.

EL PROCESO DE AGRUPAMIENTO SE INDICARA AL MISMO TIEMPO QUE SE REALIZA EL SIGUIENTE EJEMPLO.

EJEMPLO

EN UN ESTUDIO ANTROPOLOGICO SE OBTUVO UNA MUESTRA DE 30 ESTATURAS

DE LOS VARONES ADULTOS RESIDENTES EN UNA REGION. LOS DATOS, ORDENADOS EN FORMA CRECIENTE DE VALORES, FUERON LOS SIGUIENTES:

160,161,163,163,163,167,167,167,167,168,168,168,169,169,170,
171,171,173,174,175,175,175,178,179,181,181,183,184,187,191 CM.

OBTENER LA TABLA DE DISTRIBUCION DE FRECUENCIAS.

SOLUCION:

1. DETERMINACION DEL RANGO DE LA MUESTRA

   RANGO = VALOR MAXIMO - VALOR MINIMO = 191-160=31 CM

2. DETERMINACION DEL NUMERO DE INTERVALOS

   NUMERO DE INTERVALOS = $\frac{30}{5} = 6$

3. DETERMINACION DE LOS LIMITES DE CLASE

   ANCHO DE LOS INTERVALOS = $\frac{\text{RANGO}}{\text{NUMERO}} = \frac{31}{6} = 5.1$

   TOMAREMOS UN ANCHO DE 6 CM, CON LO CUAL EL RANGO DEL AGRUPAMIENTO ES 6 x 6 =36 CM. LA DIFERENCIA DE RANGOS ES 36-31=5, QUE SE REPARTE EN LOS DOS INTERVALOS EQUITATIVAMENTE. POR LO TANTO, LOS INTERVALOS RESULTAN SER:

   157-162, 163-168, 169-174, 175-180, 181-186, 187-192

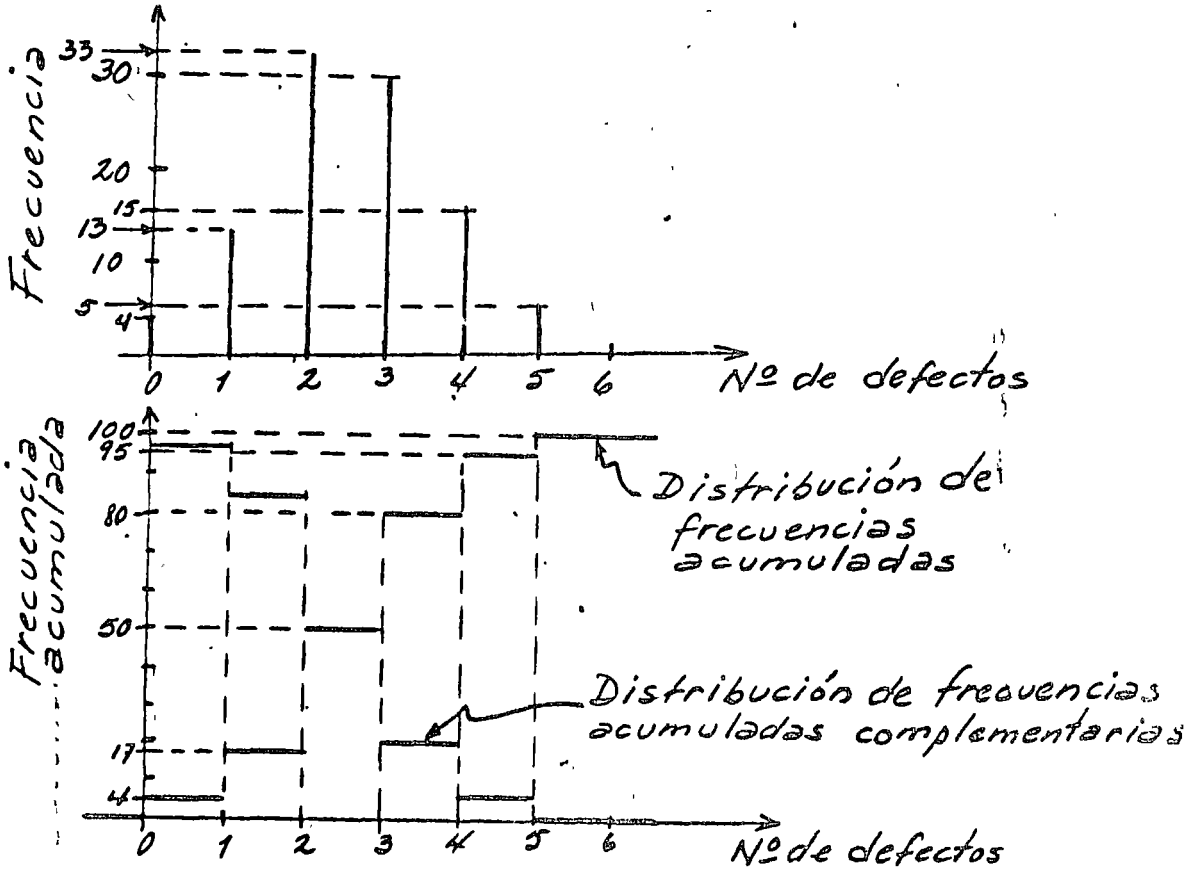
4. INTEGRACION DE LA TABLA:

INTERVALO	LIMITES REALES		FREC.	FREC. REL.	FREC. ACUM.	FREC. REL. ACUM.
	INF.	SUP.				
157-162	156.5	162.5	2	$\frac{2}{30} = 0.067$	2	0.067
163-168	162.5	168.5	10	$\frac{10}{30} = 0.333$	12	0.400
169-174	168.5	174.5	7	$\frac{7}{30} = 0.233$	19	0.633
175-180	174.5	180.5	5	$\frac{5}{30} = 0.167$	24	0.800
181-186	180.5	186.5	4	$\frac{4}{30} = 0.133$	28	0.933
187-192	186.5	192.5	2	$\frac{2}{30} = 0.067$	30	1.000
			$\Sigma=30$	$\Sigma = 1.000$		

EJEMPLO

EN UN ESTUDIO SOBRE LA CALIDAD DE LOS MONOBLOCKS PRODUCIDOS POR UNA FABRICA, SE OBTUVO UNA MUESTRA ALEATORIA DE 100 ELEMENTOS, A LOS CUALES SE LES CONTO EL NUMERO DE DEFECTOS DE FABRICACION. LA DISTRIBUCION DE FRECUENCIAS QUE SE OBTUVO ES LA SIGUIENTE:

NUMERO DE DEFECTOS	FRECUENCIA	FRECUENCIA ACUMULADA	FRECUENCIA ACUMULADA COMPLEMENTARIA
0	4	4	96
1	13	17	83
2	33	50	50
3	30	80	20
4	15	95	5
5	5	100	0
	100		



MEDIDAS REPRESENTATIVAS DE LOS DATOS

MEDIDAS DE TENDENCIA CENTRAL

VALOR MEDIO O PROMEDIO ARITMETICO

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

DONDE x_i SON LOS VALORES DE LOS DATOS Y n ES EL TAMAÑO DE LA MUESTRA.

SI LOS DATOS ESTAN AGRUPADOS Y f_j ES LA FRECUENCIA DEL j-ESIMO INTERVALO Y x_j ES LA MARCA DE CLASE CORRESPONDIENTE, ENTONCES

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{K} f_j x_j \quad ; \quad K = \text{NUMERO DE INTERVALOS}$$

EJEMPLO

SEA EL EJEMPLO ENUNCIADO ANTERIORMENTE DE LOS DEFECTOS EN MONOBLOCKS. SE TENIA:

j	No. DE DEFECTOS x	FRECUENCIA f	fx
1	0	4	4 x 0 = 0
2	1	13	13 x 1 = 13
3	2	33	33 x 2 = 66
4	3	30	30 x 3 = 90
5	4	15	15 x 4 = 60
K=6	5	5	5 x 5 = 25
		Σ=100	$\sum_{j=1}^{6}$ 254

$\bar{x} = \frac{254}{100}$

$\bar{x}$ = 2.54 DEFECTOS POR MONOBLOCK

MODO.- ES EL VALOR DE LA VARIABLE QUE APARECE CON MAYOR FRECUENCIA EN UNA MUESTRA. SI LOS DATOS ESTAN AGRUPADOS, EL MODO ES LA MARCA DE CLASE DEL INTERVALO QUE TIENE LA MAYOR FRECUENCIA.

EJEMPLO

EN EL PROBLEMA DE LOS MONOBLOCKS EL MODO ES 2. EN EL PROBLEMA DE LAS ESTATURAS DE LOS VARONES ADULTOS DE UNA CIUDAD EL MODO ES 165.5 CM.

MEDIANA: ES EL VALOR DE LA VARIABLE QUE CORRESPONDE AL 50% DE LA FRECUENCIA RELATIVA ACUMULADA.

SI LOS DATOS ESTAN AGRUPADOS POR INTERVALOS, LA MEDIANA SE PUEDE CALCULAR CON LA FORMULA (ADEMAS DE GRAFICAMENTE, COMO YA SE VIO):

$$\text{MEDIANA} = M = L_M + \frac{\frac{n}{2} - F_M}{f_M} d_M$$

DONDE L_M = LIMITE INFERIOR REAL DEL INTERVALO QUE CONTIENE A LA MEDIANA

f_M Y d_M = RESPECTIVAMENTE, A LA FRECUENCIA Y ANCHO DEL INTERVALO QUE CONTIENE A LA MEDIANA

F_M = FRECUENCIA ACUMULADA HASTA EL INTERVALO QUE CONTIENE A LA MEDIANA EXCLUSIVE

n = TAMAÑO DE LA MUESTRA

EJEMPLO

EN UN ESTUDIO PARA DETERMINAR LOS TIEMPOS EN QUE UNA MUESTRA ALEATORIA DE INDIVIDUOS REACCIONABA A CIERTOS ESTIMULOS PSICOLOGICOS SE OBTUVO LO SIGUIENTE:

j	MARCA DE CLASE x, EN SEG	LIMITES REALES	FRECUENCIA f	FRECUENCIA ACUMULADA, F	fx, SEG
1	0.10	0.075-0.125	2	2	0.20
2	0.15	0.125-0.175	7	9	1.05
3	0.20	0.175-0.225	14	23	2.80
4	0.25	0.225-0.275	4	27	1.00
K=5	0.30	0.275-0.325	3	30	0.90
			Σ=30		$\sum_{j=1}^{5} f_j x_j$=5.95

$$\bar{x} = \frac{5.95}{30} = 0.198 \text{ SEG}$$

MODO = 0.20 SEG

$$d_M = 0.05,\ L_M = 0.20 - \frac{0.05}{2} = 0.175,\ F_M = 9$$

$$n/2 = 30/2 = 15,\ f_M = 14$$

$$\text{MEDIANA} = M = 0.175 = \frac{15 - 9}{14}\ 0.05$$

$$M = 0.175 + \frac{0.30}{14} = 0.175 + 0.021 = 0.196 \text{ SEG}$$

MEDIDAS DE DISPERSION

RANGO = MAXIMO VALOR OBSERVADO - MINIMO VALOR OBSERVADO

VARIANCIA : SI LOS DATOS *NO* ESTAN AGRUPADOS;

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

SI LOS DATOS ESTAN AGRUPADOS:

$$S_X^2 = \frac{1}{n} \sum_{j=1}^{K} (x_j - \bar{x})^2$$

DONDE LAS x_j SON LOS VALORES DE LAS MARCAS DE CLASE DE LOS INTERVALOS.

DESVIACION ESTANDAR

$$S_X = \sqrt{S_X^2}$$

COEFICIENTE DE VARIACION

$$v_X = S_X/\bar{x}$$

EJEMPLO

EN UN ESTUDIO SOBRE LA TEMPERATURA MAXIMA DIARIA EN UNA CIUDAD SE OBTUVO LO SIGUIENTE DURANTE UNA PRIMAVERA:

j	INTERVALOS DE TEMPERATURA, °F	MARCA DE CLASE, °F	FRECUENCIA f	xf	$x-\bar{x}$	$(x-\bar{x})^2$	$(x-\bar{x})^2 f$
1	55 - 63	59	2	118	-21.3	453.7	907.4
2	64 - 72	68	6	408	-12.3	151.3	907.8
3	73 - 81	77	7	539	- 3.3	10.9	76.3
4	82 - 90	86	9	774	5.7	32.5	292.5
5	91 - 99	95	6	570	14.7	216.1	1296.6
			30	2409			3480.6

$$\bar{x} = \frac{2409}{30} = 80.3 \text{ °F}$$

$$s_X^2 = \frac{3480.6}{30} = 116 \text{ °F}^2$$

$$s_X = \sqrt{116} = 10.8 \text{ °F}$$

$$v_X = \frac{10.8}{80.3} = \underset{(13.4\%)}{0.134}$$

MODO = 86

$d_M=9$, $L_M=72.5$, $f_M=7$, $F_M=8$, $\frac{n}{2} = \frac{30}{2} = 15$

$$\text{MEDIANA} = M = 72.5 + \frac{15 - 8}{7} 9 = 72.5 + 9 = 81.5 \text{ °F}$$

TRANSFORMACION DE VARIABLES

SEA X UNA VARIABLE ALEATORIA CON DENSIDAD DE PROBABILIDADES $f_X(x)$, Y SEA LA TRANSFORMACION

$$Y = g(X) = a + bX$$

PARA EL CASO EN QUE X ES UNA VARIABLE CONTINUA, EN TEORIA DE PROBABILIDADES SE VIO QUE

$$E(Y) = \int_{-\infty}^{\infty} (a+bx) f_X(x)\,dx = a \int_{-\infty}^{\infty} f_X(x)\,dx + b \int_{-\infty}^{\infty} x f_X(x)\,dx$$

$$= a + bE(X)$$

Y, ANALOGAMENTE QUE

$$\sigma^2(Y) = b^2\sigma^2(X)$$

(ESTOS RESULTADOS SON VALIDOS TAMBIEN PARA VARIABLES ALEATORIAS DISCRETAS.)

AHORA, SI SE TIENE UNA MUESTRA DE TAMAÑO n DE LA VARIABLE X, A CADA VALOR, x_i, DE DICHA MUESTRA LE CORRESPONDE UN VALOR, y_i, DE LA MUESTRA DE Y DADO POR

$$y_i = a + bx_i$$

POR LO TANTO, EL PROMEDIO ARITMETICO DE LAS y_i ES

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} (a + bx_i) = \frac{1}{n}\sum_{i=1}^{n} a + \frac{b}{n}\sum_{i=1}^{n} = a + b\bar{x}$$

ANALOGAMENTE, EL VALOR MEDIO CUADRATICO RESULTA SER

$$\overline{y^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2 = \frac{1}{n}\sum_{i=1}^{n} (a+bx_i)^2 = \frac{1}{n}\sum_{i=1}^{n} a^2 + \frac{1}{n}\sum_{i=1}^{n} 2abx_i + \frac{1}{n}\sum_{i=1}^{n} b^2x_i^2$$

$$= a^2 + 2ab\bar{x} + b^2\overline{x^2}$$

Y, LA VARIANCIA,

$$S^2(y) = \overline{y^2} - \bar{y}^2 = a^2 + 2ab\bar{x} + b^2\overline{x^2} - (a+b\bar{x})^2 = b^2\overline{x^2} - b^2\bar{x}^2 = b^2 S^2(x)$$

ESTAS TRANSFORMACIONES SE PUEDEN EMPLEAR PARA CALCULAR EL PROMEDIO $\bar{y}$, Y LA VARIANCIA $S^2(y)$ DE LA MUESTRA DE UNA VARIABLE QUE RESULTA DE UNA TRANSFORMACION Y, CON BASE EN ELLOS, CALCULAR $\bar{x}$ Y $S^2(x)$ DE LA MUESTRA ORIGINAL, MEDIANTE LAS ECUACIONES

$$\bar{x} = (\bar{y} - a)/b$$
$$S^2(x) = S^2(y)/b^2$$

ESTE PROCEDIMIENTO AHORRA BASTANTE TIEMPO DE CALCULOS CUANDO LOS DATOS ESTAN AGRUPADOS, EN CUYO CASO LOS x_i SON LAS MARCAS DE CLASE.

EJEMPLO

EN EL PROBLEMA DE LOS RESULTADOS, x_i, DE UN EXAMEN SOBRE PEDAGOGIA SE OBTUVO LA DISTRIBUCION DE FRECUENCIAS INDICADAS EN LAS DOS PRIMERAS COLUMNAS DE LA SIGUIENTE TABLA:

MARCAS DE CLASE x_i	FRECUENCIAS f_i	MARCAS DE CLASE TRANSFORMADA, y_i	$y_i f_i$	y_i^2	$y_i^2 f_i$
55.5	2	-2	- 4	4	8
65.5	5	-1	- 5	1	5
75.5	6	0	0	0	0
85.5	11	1	11	1	11
95.5	6	2	12	4	24
	Σ=30		Σ=14		Σ=48

$\bar{y} = 14/30 = 0.467$, $\overline{y^2} = 48/30 = 1.6$, $S^2(y) = 1.6 - (0.467)^2 = 1.382$

$\bar{x} = [0.467 - (-7.55)]/(1/10) = 80.17$, $S^2(x) = 1.382/(0.1)^2 = 138.2$

CALCULAREMOS EL PROMEDIO Y LA VARIANCIA DE ESTA MUESTRA, CALCULANDO PRIMERO $\bar{y}$ Y $S^2(y)$ DE LA TRANSFORMACION

$$y = a + bx = \frac{x - C_1}{C_2}$$

(a = $\frac{-C_1}{C_2}$, b = $\frac{1}{C_2}$ CON C_1 = MARCA DE CLASE CENTRAL Y C_2 = ANCHO DE CLASE)

TOMANDO a = -75.5/10 y b=1/10 (C_1=75.5 y C_2=10), SE TIENE

$$y_i = (-75.5 + y_i)/10$$

POR LO QUE

$$y_1 = (-75.5 + 55.5)/10 = 2$$

$$y_2 = (-75.5 + 65.5)/10 = -1$$

$$y_3 = (-75.5 + 75.5)/10 = 0$$

$$y_4 = (-75.5 + 85.5)/10 = 1$$

$$y_5 = (-75.5 + 95.5)/10 = 2$$

OBSERVESE QUE SE OBTIENE $y = 0$ PARA EL INTERVALO CORRESPONDIENTE A $X_i = C_1$, Y PARA LOS INTERVALOS CON VALORES MAYORES DE X BASTA CON IRLE SUMANDO UNA UNIDAD, MIENTRAS QUE A LOS DE VALORES MENORES, IRLE RESTANDO UNA UNIDAD.

centro de educación continua
división de estudios superiores
facultad de ingeniería, unam

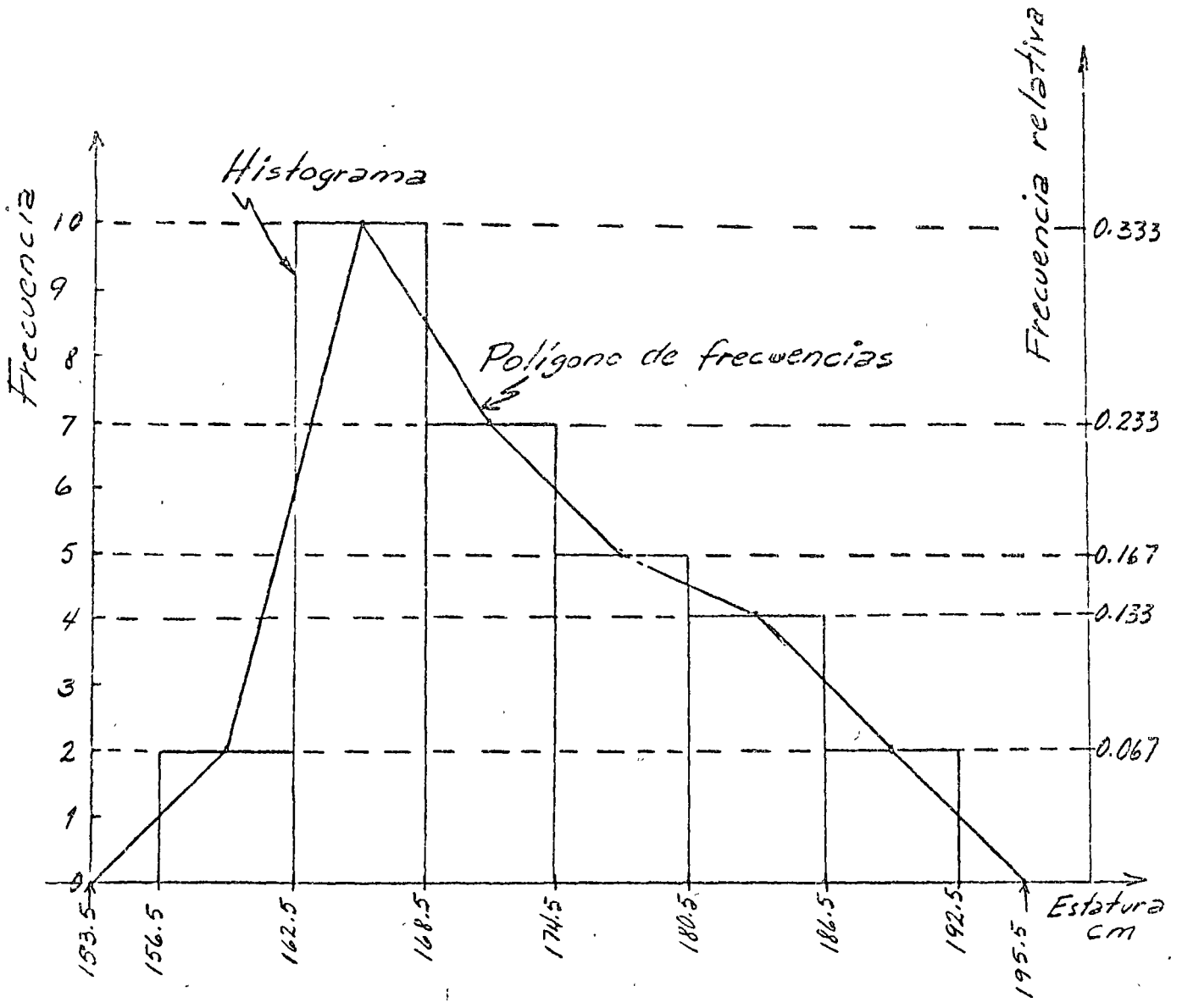
FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO

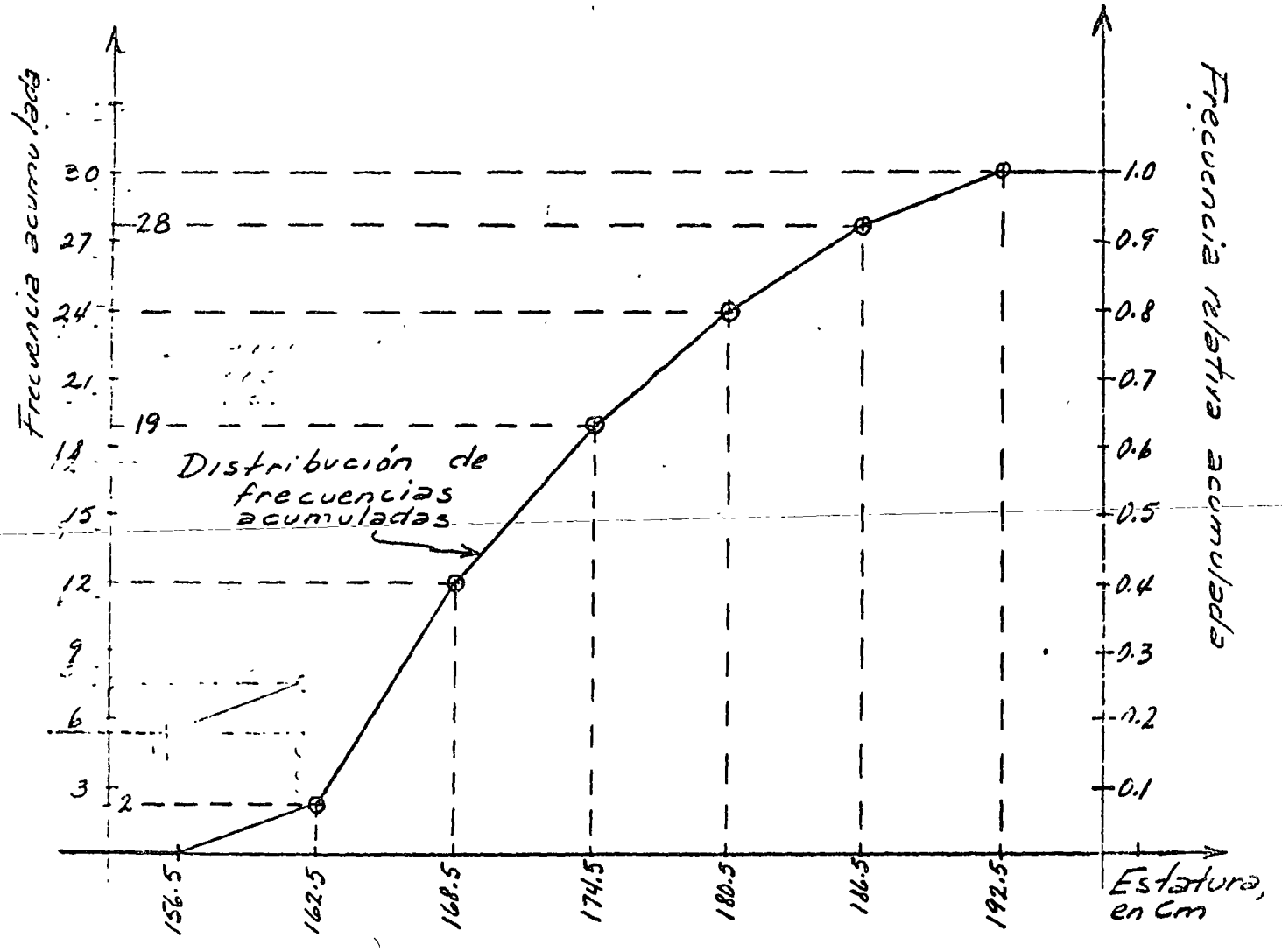
MUESTREO ALEATORIO SIMPLE (Continuación)

DR. OCTAVIO A. RASCON CHAVEZ

PALACIO DE MINERIA
Tacuba 5, primer piso, México 1, D. F.

PRESENTACION GRAFICA DE LAS DISTRIBUCIONES DE FRECUENCIAS





¿CUAL ES LA FRECUENCIA DE VALORES MAYORES QUE 180.5?: 30-24=6

LA FRECUENCIA RELATIVA ACUMULADA COMPLEMENTARIA ES: 1-0.800=0.200 (20%)

EJEMPLO

SI LA PROBABILIDAD DE QUE FALLE UNA VARILLA DE ACERO AL APLICARLE CIERTA CARGA ES DE 0.001, DETERMINAR LA PROBABILIDAD DE QUE EN 2000 VARILLAS PROBADAS FALLEN MAS DE DOS.

USANDO LA DISTRIBUCION DE BERNOULLI SE OBTIENE

$$P[X > 2] = 1 - P[X \leq 2] = 1 - (P[X = 0] + P[X = 1] + P[X = 2]) =$$

$$= 1 - \left(\frac{2\,000!}{2\,000!\ 0!}(0.001)^0(0.999)^{2\,000} + \frac{2\,000!}{1\,999!\ 1!}(0.001)^1(0.999)^{1\,999} + \right.$$

$$\left. + \frac{2\,000!}{1\,998!\ 2!}(0.001)^2(0.999)^{1\,998}\right) = 0.3255$$

LOS CALCULOS NECESARIOS PARA OBTENER LA SOLUCION SON BASTANTE MAS TEDIOSOS QUE LOS QUE DEBEN EFECTUARSE APROVECHANDO QUE EL NUMERO DE REPETICIONES DEL EXPERIMENTO ES GRANDE, A FIN DE UTILIZAR LA DISTRIBUCION NORMAL. EN ESTAS CIRCUNSTANCIAS, LA PROBABILIDAD DE QUE $X \leq 2$ EN EL CASO DISCRETO, EQUIVALE A LA DE QUE $X \leq 2.5$ EN EL CONTINUO; ASI

$$\mu = np = 2\,000 \times 0.001 = 2$$

$$\sigma = \sqrt{npq} = \sqrt{2\,000 \times 0.001 \times 0.999} = 1.41$$

$$P[X \leq 2.5] = P\left|Z \leq \frac{2.5 - 2}{1.41}\right| = P|Z \leq 0.355| = 0.6387$$

DE DONDE

$$P[x > 2.5] = 1 - P|X \leq 2.5| = 1 - 0.6387 = 0.3613$$

EJEMPLO

EN UNA SERIE DE 462 EXPERIMENTOS CON FINES ANTROPOLOGICOS, CONSISTENTES EN MEDIR EL TAMAÑO DE LA CABEZA DE LOS INDIGENAS RESIDENTES EN UNA ZONA TROPICAL, SE OBTUVIERON LOS RESULTADOS ANOTADOS EN LAS DOS PRIMERAS COLUMNAS DE LA SIGUIENTE TABLA. SI LA VARIABLE ALEATORIA "TAMAÑO DE LA CABEZA" SE CONSIDERA QUE TIENE DISTRIBUCION NORMAL, ¿QUE CANTIDAD DE RESULTADOS SE ESPERARIA OBTENER ANTES DE HACER LAS MEDICIONES, SI SE CONSIDERA QUE $\mu = \bar{x} = 191.8$MM y $\sigma = s = 6.48$MM, DONDE $\bar{x}$=PROMEDIO ARITMETICO Y s=DESVIACION ESTANDAR DE LOS DATOS?

$$z_1 = \frac{171.5 - 191.8}{6.48} = -3.13;\quad z_2 = \frac{175.5 - 191.8}{6.48} = -2.51;\quad x_3 = \frac{179.5 - 191.8}{6.48}$$

$$= -190, \text{ ETC.}$$

$P(-3.13 \leq Z \leq -2.51) = 0.0051$; $P(-2.51 \leq Z \leq -1.90) = 0.0227$;

$P(-1.90 \leq Z \leq -1.28) = 0.0716$, ETC.

$462 \times 0.0051 = 2.4$; $462 \times 0.0227 = 10.5$; $462 \times 0.0716 = 31.1$, ETC.

INTERVALO DE VALORES DE X, EN MM	NUMERO DE OBSERVACIONES (frecuencia, f)	INTERVALO DE $Z = \frac{X-\mu}{\sigma}$	PROBABILIDAD $P(z_1 \leq z \leq z_2) = P$	FRECUENCIA ESPERADA = 462 P
171.5-171.5	3	(-3.13)-(-2.51)	0.0051	2.4
171.5-179.5	9	(-2.51)-(-1.90)	0.0227	10.5
179.5-183.5	29	(-1.90)-(-1.28)	0.0716	33.1
183.5-187.5	76	(-1.28)-(-0.66)	0.1543	71.3
187.5-191.5	104	(-0.66)-(-0.05)	0.2255	104.2
191.5-195.5	110	(-0.05)- 0.57	0.2356	108.8
195.5-199.5	88	0.57 - 1.19	0.1673	77.3
199.5-203.5	30	1.19 - 1.80	0.0811	37.5
203.5-207.5	6	1.80 - 2.42	0.0281	13.0
207.5-211.5	4	2.42 - 3.04	0.0066	3.0
211.5-215.5	2	3.04 - 3.66	0.0011	0.5
215.5-219.5	1	3.66 - 4.27	0.0001	0.0
	TOTAL: 462			TOTAL: 461.6

centro de educación continua
división de estudios superiores
facultad de ingeniería, unam

FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO

TEMA: TECNICAS DE MUESTREO

DR. OCTAVIO A. RASCON CHAVEZ

AGOSTO-OCTUBRE 1977

PALACIO DE MINERIA
Tacuba 5, primer piso. México 1, D. F.
Teléfonos: 521-30-95 y 513-27-95

DISTRIBUCIONES MUESTRALES

DISTRIBUCION t DE STUDENT

SI SE CONSIDERAN MUESTRAS DE TAMAÑO n ESTRAIDAS DE UNA POBLACION CON MEDIA μ Y VARIANCIA DESCONOCIDA, PARA CADA MUESTRA SE PUEDE CALCULAR LA ESTADISTICA

$$T = \frac{\bar{x} - \mu}{S_x}\sqrt{n-1}$$

DONDE $\bar{x}$ Y S_x SON, RESPECTIVAMENTE, EL PROMEDIO Y LA DESVIACION ESTANDAR DE LA MUESTRA.

LA DISTRIBUCION MUESTRAL DE ESTA ESTADISTICA ES LA DISTRIBUCION "t DE STUDENT", CUYA ECUACION

$$f(t) = \frac{\Gamma(\frac{1}{2}[\nu+1])}{\sqrt{\nu\pi}\ \Gamma(\frac{1}{2}n)}\ (1 + \frac{t^2}{\nu})^{-(\nu+1)/2}$$

DONDE $\nu=n-1$ ES EL "NUMERO DE GRADOS DE LIBERTAD"

Entonces, el valor de Z es

$$z = \frac{\bar{x}_A - \bar{x}_B}{\sigma_{\bar{X}_1 - \bar{X}_2}} = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} = \frac{31 - 30.9}{\sqrt{\frac{(0.3)^2}{100} + \frac{(0.4)^2}{100}}} = 2$$

a) Puesto que se trata de una prueba de dos colas a un nivel de significancia de 0.05, la diferencia es significativa si 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.

3.4 Muestras pequeñas

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

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

3.4.1 Distribucion *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

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

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

donde S_X^2 es la variancia de la muestra.

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

$$\nu = n - k$$

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

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

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

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

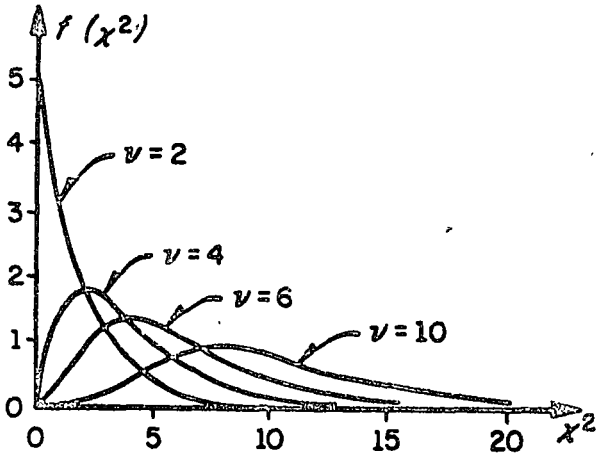
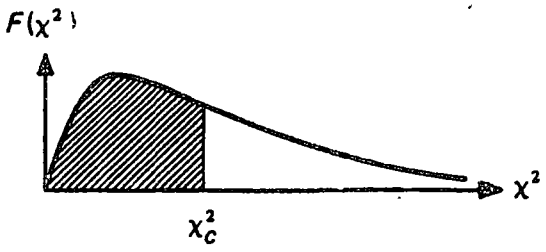


Fig 21. Distribución Ji cuadrada para distintos valores de ν

TABLA 8. VALORES CRITICOS χ^2_c



ν	$\chi^2_{.995}$	$\chi^2_{.99}$	$\chi^2_{.975}$	$\chi^2_{.95}$	$\chi^2_{.90}$	$\chi^2_{.75}$	$\chi^2_{.50}$	$\chi^2_{.25}$	$\chi^2_{.10}$	$\chi^2_{.05}$	$\chi^2_{.025}$	$\chi^2_{.01}$	$\chi^2_{.005}$
1	7.88	6.63	5.02	3.84	2.71	1.32	.455	.102	.016	.0039	.0010	.0002	.0000
2	10.6	9.21	7.38	5.99	4.61	2.77	1.39	.575	.211	.103	.0506	.0201	.0100
3	12.8	11.3	9.35	7.81	6.25	4.11	2.37	1.21	.584	.352	.216	.115	.072
4	14.9	13.3	11.1	9.49	7.76	5.39	3.36	1.92	1.06	.711	.483	.297	.207
5	16.7	15.2	12.8	11.15	9.2	6.63	4.35	2.67	1.61	1.15	.831	.554	.413
6	18.5	16.8	14.4	12.6	10.6	7.84	5.35	3.45	2.20	1.64	1.24	.872	.676
7	20.3	18.5	16.0	14.1	12.0	9.04	6.35	4.25	2.83	2.18	1.69	1.24	.989
8	22.0	20.1	17.5	15.5	13.4	10.2	7.34	5.07	3.49	2.73	2.18	1.65	1.34
9	23.6	21.7	19.0	16.9	14.7	11.4	8.34	5.90	4.17	3.33	2.70	2.09	1.73
10	25.2	23.2	20.5	18.3	16.0	12.5	9.34	6.74	4.87	3.94	3.25	2.56	2.16
11	26.8	24.7	21.9	19.7	17.3	13.7	10.35	7.57	5.58	4.57	3.82	3.05	2.60
12	28.3	26.2	23.2	21.0	18.5	14.8	11.3	8.44	6.30	5.23	4.40	3.57	3.07
13	29.8	27.7	24.7	22.4	19.8	16.0	12.3	9.30	7.04	5.89	5.01	4.11	3.57
14	31.3	29.1	26.1	23.7	21.1	17.2	13.3	10.2	7.79	6.57	5.63	4.66	4.07
15	32.7	30.6	27.5	25.1	22.3	18.2	14.3	11.0	8.55	7.26	6.25	5.22	4.60
16	34.3	32.0	28.8	26.3	23.5	19.4	15.3	11.9	9.31	7.96	6.91	5.81	5.14
17	35.7	33.4	30.2	27.6	24.8	20.5	16.3	12.8	10.1	8.67	7.56	6.41	5.70
18	37.2	34.8	31.5	28.9	26.0	21.6	17.3	13.7	10.9	9.39	8.23	7.01	6.26
19	38.6	36.2	32.9	30.1	27.2	22.7	18.3	14.6	11.73	10.1	8.91	7.63	6.84
20	40.0	37.6	34.2	31.4	28.45	23.8	19.3	15.5	12.4	10.9	9.59	8.26	7.43
21	41.4	38.8	35.6	32.7	29.6	24.9	20.3	16.3	13.2	11.6	10.3	8.90	8.02
22	42.8	40.3	36.8	33.9	30.8	26.0	21.3	17.2	14.0	12.3	11.0	9.54	8.64
23	44.2	41.6	38.1	35.2	32.0	27.1	22.3	18.1	14.8	13.1	11.7	10.2	9.26
24	45.6	43.0	39.4	36.4	33.2	28.2	23.3	19.0	15.7	13.8	12.4	10.9	9.89
25	46.9	44.3	40.6	37.7	34.4	29.3	24.3	19.9	16.5	14.5	13.15	11.5	10.5
26	48.3	45.6	41.9	38.9	35.6	30.4	25.3	20.8	17.3	15.4	13.8	12.2	11.2
27	49.6	47.0	43.2	40.1	36.7	31.5	26.3	21.7	18.1	16.2	14.6	12.9	11.8
28	51.0	48.3	44.5	41.3	37.9	32.6	27.3	22.7	18.9	16.9	15.3	13.6	12.5
29	52.3	49.6	45.7	42.5	39.1	33.7	28.3	23.6	19.8	17.7	16.0	14.3	13.1
30	53.7	50.9	47.0	43.8	40.3	34.8	29.3	24.5	20.6	18.5	16.8	15.0	13.8
40	66.8	63.7	59.3	55.8	51.8	45.7	39.3	33.7	29.1	26.5	24.4	22.2	20.7
50	79.5	76.2	71.4	67.5	63.2	56.3	49.3	43.0	37.7	34.8	32.4	29.7	28.0
60	92.0	88.4	83.3	79.1	74.4	67.0	59.3	52.3	46.5	43.2	40.5	37.5	35.5
70	104.2	100.4	95.0	90.5	85.5	77.6	69.3	61.7	55.3	51.7	48.8	45.4	43.3
80	116.3	112.3	106.6	101.9	96.6	88.1	79.3	71.1	64.3	60.4	57.2	53.5	51.2
90	128.3	124.1	118.1	113.1	107.6	98.6	89.3	80.6	73.3	69.1	65.6	61.8	59.2
100	140.2	135.8	129.6	124.3	118.5	109.1	99.3	90.12	82.4	77.9	74.2	70.1	67.3

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

3.4.1.1 Intervalo de confianza para la variancia

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

$$\chi_{c'}^2 < \frac{n S_X^2}{\sigma^2} < \chi_c^2$$

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

Con base en lo anterior, se concluye que

$$\frac{n S_X^2}{\chi_c^2} < \sigma^2 < \frac{n S_X^2}{\chi_{c'}^2}$$

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

3.4.1.2 Prueba de hipótesis para la variancia

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

Ejemplo

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

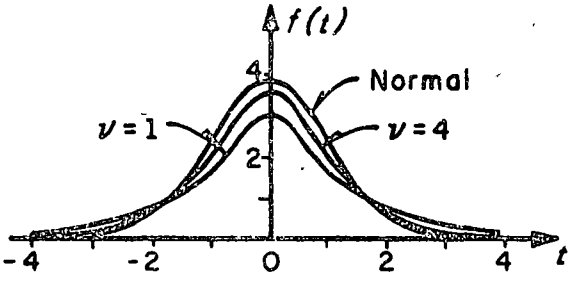


Fig 23. Distribución t de Student para distintos valores de ν

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

3.4.3.1 Límites e intervalos de confianza

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

Así pues

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

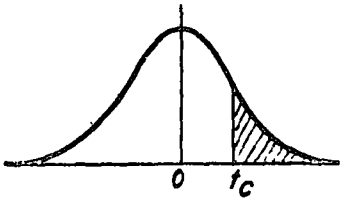
representa un intervalo de confianza para t, a partir del cual se puede estimar que μ se encuentra dentro del intervalo

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

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

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

TABLA 10. VALORES t_c PARA LA DISTRIBUCION t DE STUDENT



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

centro de educación continua
división de estudios superiores
facultad de ingeniería, unam

FUNDAMENTOS DE LAS TECNICAS DEL MUESTREO ESTADISTICO

TEMA: TAMAÑO DE MUESTRA.

PROF. M. en I. AUGUSTO VILLARREAL ARANDA.

SEPTIEMBRE DE 1977.

4. TAMAÑO DE LA MUESTRA

Por: M en I Augusto Villarreal Aranda*

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,

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

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

intervalo

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

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

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

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

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

Por lo tanto, es posible escribir

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

4.1.1 Muestreo de una población finita

De la inferencia estadística, el valor de $\sigma_{\bar{X}}$, 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, éste se puede obtener despejando de la ecuación anterior el valor de n. Para ello, se requiere elevar al cuadrado ambos miembros, es decir

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

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

despejando a n:

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

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

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

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

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

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

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

4.1.2 Muestreo de una población infinita

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

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

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

Considerando la ecuación 4.1, se puede escribir en este caso

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

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

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

Por lo cual

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

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

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

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

4.1.3 Comparación entre n y n_∞

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

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

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

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

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

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

4.1.4 Empleo adecuado de n y n_∞

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

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

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

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

$$fm \leq 0.05$$

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

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

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

Ejemplo 4.1

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

a. 1000

b. 100

Solución

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

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

por lo cual

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

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

$$\therefore n_\infty = 14$$

En virtud de que en este caso

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

se considera que n = 14.

b. En este caso

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

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

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

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

$$\therefore n = 13$$

Ejemplo 4.2

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

Solución

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

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

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

Por lo tanto,

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

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

O sea $n_\infty = 327$

Puesto que

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

se procede a calcular n, es decir,

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

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

$$\therefore n = 306$$

Ejemplo 4.3

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

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,

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

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

$$= 2769.01 - 2758.35 = 10.66 \text{ pulgadas}$$

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

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

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

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

$$t_{0.025} = 1.96$$

Con dicho valor de t_C se obtiene

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

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

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

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

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

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

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

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

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

4.2 Tamaño de una muestra aleatoria simple (Totales)

Una característica o parámetro poblacional de gran interés es el 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 al total, y N_p es el número de elementos de la misma.

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

rior, se obtiene

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

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

Cono 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

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

y la desviación estándar es

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

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

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

Elevando al cuadrado y realizando operaciones algebraicas

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

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

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

O sea

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

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

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

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

Considerando la ec 4.3, queda finalmente

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

Ejemplo 4.4

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

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

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

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

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

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

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

$$= 1089.94 - 865.44 = 224.5$$

Entonces

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

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

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

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

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

$$N_p = 676$$

$$\sigma_Y^2 \doteq S_Y^2 = 224.5$$

$$N_p^2 n_\infty = N_p^2 \frac{K^2 \sigma_Y^2}{e^2} = \frac{(676)^2\ (1,96)^2\ (224.5)}{(995)^2} = 397.9$$

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

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

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

4.3 Tamaño de una muestra aleatoria simple (Proporciones)

4.3.1 Antecedentes

Supóngase una población binomial de tamaño N_p tal que cada uno de sus elementos únicamente puede estar en una de dos clases: A 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 elemtos que están en B es

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

por lo cual

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

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

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

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

La variancia es

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

o sea

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

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

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

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

$$\mu_p = P$$

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

Si la población es infinita, se obtiene

$$\mu_p = P$$

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

estimándose 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,

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

para muestreo de población infinita, y

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

para muestreo de población finita con tamaño N_p.

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

Ejemplo 4.5

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

Solución

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

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

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

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

Elevando al cuadrado y realizando operaciones algebraicas

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

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

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

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

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

$\therefore$ n = 735 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

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

o sea

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

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

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

por lo cual

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

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

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

Como

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

se corrige el valor anterior, obteniéndose finalmente

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

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

$\therefore$ n = 307 habitantes

INFERENCIA ESTADISTICA

Por: M en I Augusto Villarreal Aranda*

1. Introducción

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

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

2. Distribuciones muestrales

Si se consideran todas las muestras posibles de tamaño

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

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

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

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

3. Muestreo con y sin remplazo

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

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

4. Distribucion muestral del promedio aritmético

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

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

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

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

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

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

puesto que

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

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

Ejemplo 4.1

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

Primer procedimiento.

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

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

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

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

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

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

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

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

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

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

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

Segundo procedimiento.

Por tratarse de una población finita, se verifica que

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

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

El valor de σ^2 de la población es

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

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

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

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

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

Ejemplo 4.2

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

a. entre 496 y 500 kg

b. de más de 510 kg.

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

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

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

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

es decir,

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

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

En la fig 4.1 se puede apreciar que

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

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

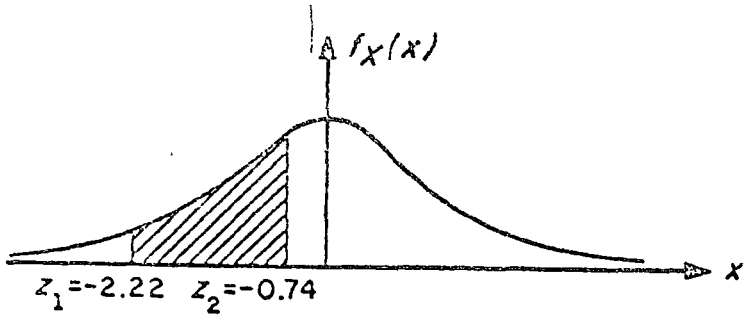


Fig 4.1 Distribución normal correspondiente al ejemplo

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

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

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

Estandarizando dicho valor, queda

$$z_3 = \frac{5.10 - 5.02}{0.027} = 2.96$$

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

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

$$= 0.5 - 0.4985 = 0.0015$$

6. Teoría estadística de la estimación

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

7. Estimadores puntuales. Clasificación

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

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

$$\mu_S = \theta$$

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

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

Empleando símbolos, si

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

resulta que la estadística S_n es un estimador consistente. Por ejemplo, el promedio aritmético es un estimador insesgado y consistente de la 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 θ se localice en el intervalo de $S - z_c \sigma_S$ a $S + z_c \sigma_S$, donde z_c es una constante, se escribe en la forma

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

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

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

TABLA 8.1 VALORES DE z_c PARA DISTINTOS NIVELES DE CONFIANZA

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

Ejemplo 8.1

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

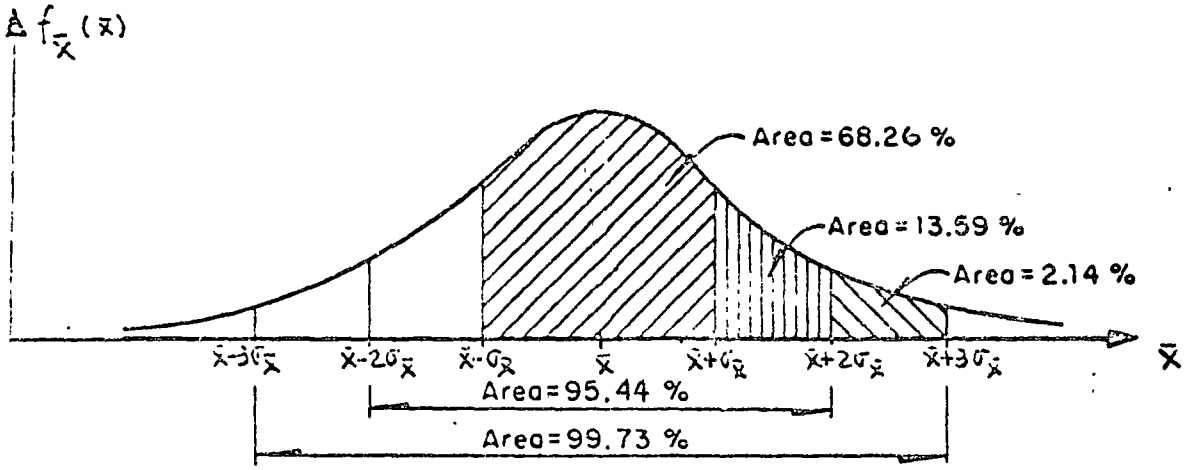


Fig 8.1

9. Estimación de intervalos de confianza para la media

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

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

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

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

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

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

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

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

Ejemplo 9.1

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

a. 95 por ciento
b. 97 por ciento

para el diámetro medio de todos los tubos.

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

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

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

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

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

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

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

Ejemplo 9.2

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

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

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

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

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

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

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

$$72 \pm 2.704$$

y el intervalo de confianza respectivo es

$$(69.296, \ 74.704)$$

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

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

en este caso se tendría

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

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

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

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

Elevando al cuadrado la desigualdad, queda

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

o sea

$$87.85 < n$$

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

c. Los límites de confianza son, en este caso

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

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

o sea

$$72 \pm 1.3795 \, Z_c$$

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

$$1 = 1.3795 \, Z_c$$

Es decir

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

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

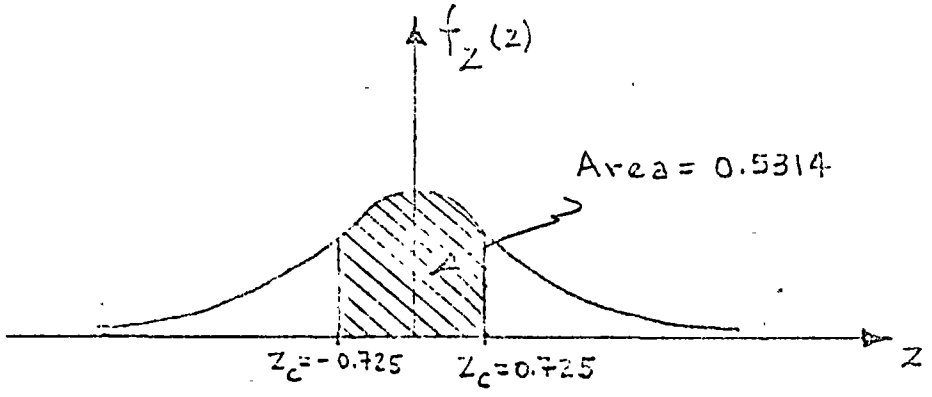


Fig 9.1

11. Pruebas de hipótesis

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

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

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

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

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

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

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

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

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

13. Comportamiento de los errores tipos I y II

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

$$H_0 : \mu_S = \mu_1$$

$$H_1 : \mu_S = \mu_2$$

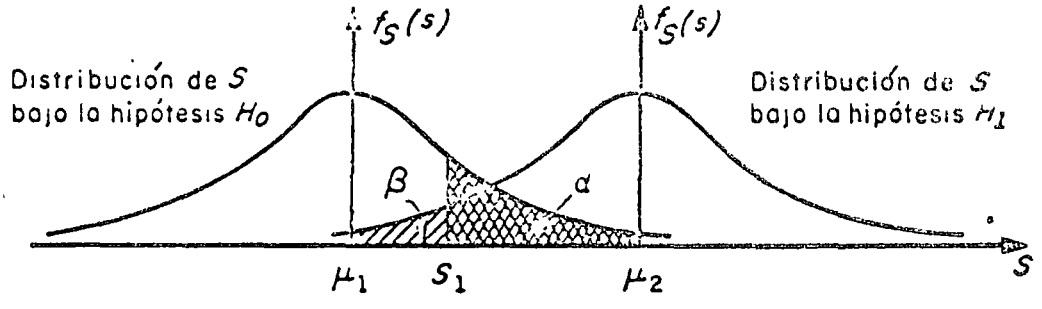
En la fig 13.1 se muestra en forma gráfica la relación entre los errores tipos I y II en el caso en el que la regla de decisión para aceptar o rechazar H_0 es la siguiente:

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

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

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

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



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

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

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

Al observar la fig 13.2 siguiente, es posible concluir

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

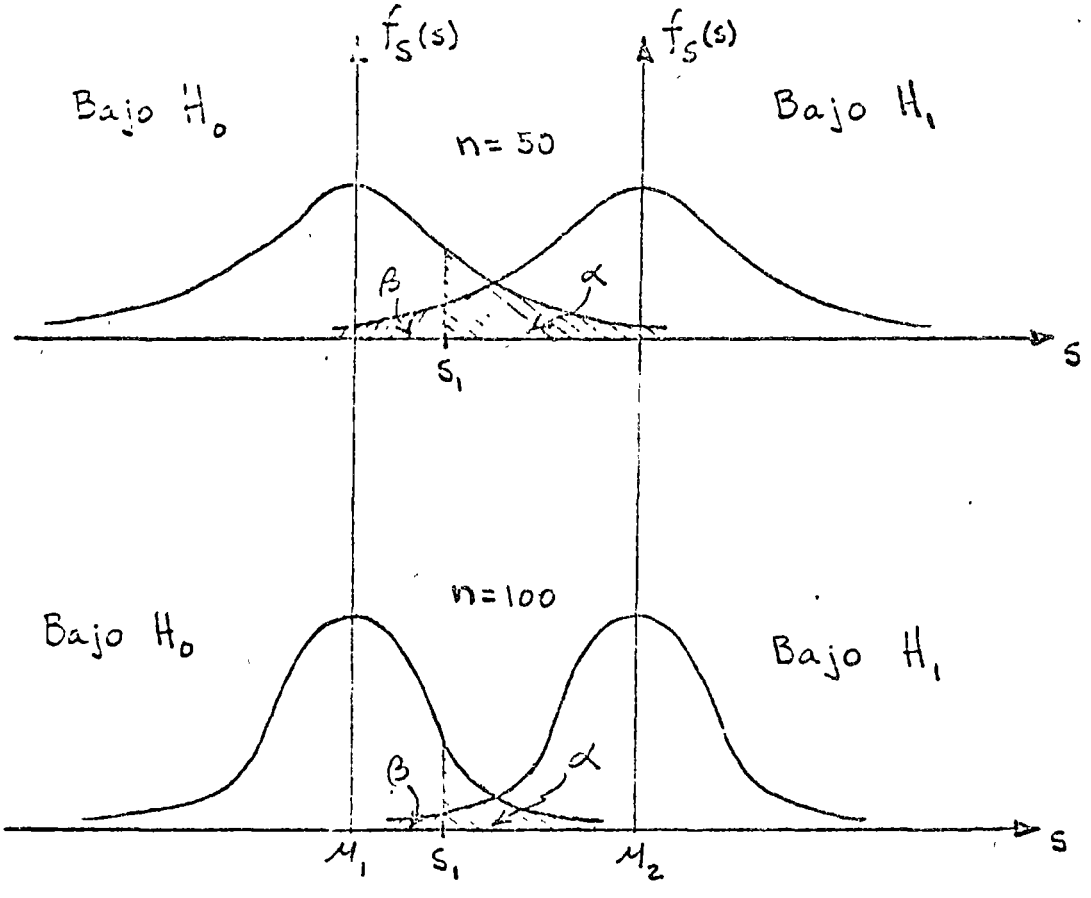


Fig 13.2

Sin embargo, esta técnica de reducción simultánea de 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 le denomina *región crítica, de rechazo, o de significancia*. Por el contrario, al conjunto de los valores de la estadística en que se acepta la hipótesis, se le llama *región de aceptación*.

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

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

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

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

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

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

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

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

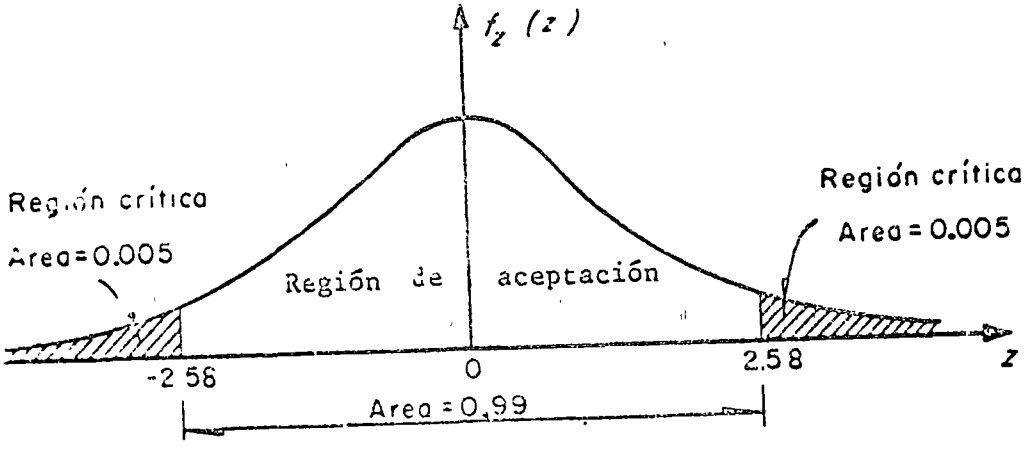


Fig 14.1 Región de significancia

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

TABLA 14.1 VALORES CRITICOS DE z

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

15. Pruebas de una y de dos colas

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

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

$$H_0 : \mu_S = \mu_1$$

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

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

En los casos

$$H_0 : \mu_S = \mu_1$$

$$H_1 : \mu_S < \mu_1$$

y

$$H_0 : \mu_S = \mu_1$$

$$H_1 : \mu_S > \mu_1$$

las pruebas resultan de una cola.

16. Pruebas de hipótesis para la media

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

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

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

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

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

Ejemplo 16.1

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

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

a. 0.05

b. 0.01

Para la solución se deben considerar las hipótesis

$$H_0 : \mu = 7.65$$

$$H_1 : \mu \neq 7.65$$

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

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

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

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

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

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

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

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

Puesto que

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

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

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

Ejemplo 16.2

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

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

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

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

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

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

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

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

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

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

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

En virtud de que

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

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

centro de educación continua
división de estudios superiores
facultad de ingeniería, unam

FUNDAMENTOS DE LAS TECNICAS DE MUESTREO ESTADISTICO

MUESTREO ALEATORIO SIMPLE PARA RAZONES O COCIENTES

M. EN C. ADELA ABAD DE SERVIN

M. EN C. ALEJANDRO SERVIN ANDRADE

AGOSTO-OCTUBRE, 1977.

4. EL MUESTREO ALEATORIO SIMPLE EN EL CASO DE LA ESTIMACION DE RAZONES O COCIENTES

4.1 ESTIMACION DE RAZONES O COCIENTES:

El objetivo de este tema es cómo estimar la razón de dos variables aleatorias. Es decir en muchas ocasiones, la estimación necesaria no es la de un total, ni la de una media, ni la de una proporción. El interés radica muchas veces en conocer, por ejemplo, la razón de préstamos dedicados a maquinaria agrícola a préstamos totales dedicados a agricultura, la razón de gastos en alimentación a gastos totales en las familias, la razón de salarios pagados a los obreros a salarios totales de la empresa. Estas estimaciones muchas veces reciben el nombre de proporciones, ya que se entienden como : la proporción de los préstamos agrícolas que fueron dedicados a maquinaria agrícola, la proporción del gasto familiar que fue dedicado a gastos de alimentación y la proporción de salarios de la empresa que se utiliza para obreros.

Debemos tener presente que la población que vamos a investigar consta de N unidades muestrales, de la cual vamos a obtener una muestra aleatoria simple de n unidades muestrales y a las unidades que cayeron en la muestra las vamos a investigar con dos características, que le vamos a llamar y_i e x_i (i= 1,...,n) el parámetro que queremos estimar es:

$$R = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i}$$

y el estimador que vamos a utilizar o que se propone es:

$$\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}}$$

Este estimador es <u>consistente</u>, es decir a medida que el tamaño de muestra aumenta el estimador se acerca más al valor poblacional. Dicho de otra manera, cuando n = N el estimador se convierte en el parámetro poblacional.

La distribución del estimador $\hat{R}$ no es sencilla, ya que se trata de un estimador que es razón de variables aleatorias y el numerador y denominador varían de muestra a muestra. $\hat{R}$ es un estimador sesgado de R, pero el sesgo se vuelve despreciable cuando la muestra es grande* y la distribución de $\hat{R}$ tiende a ser normal.

Si n es grande y es extraída aleatoriamente de una población de tamaño N, se demuestra que la varianza aproximada de $\hat{R}$ es

$$V(\hat{R}) \doteq \frac{1-f}{n\bar{X}^2}\left[\frac{\sum_{i=1}^{N}(y_i - Rx_i)^2}{N-1}\right]$$

un estimador de esta varianza es

$$\hat{V}(R) \doteq \frac{1-f}{n\bar{x}^2}\left[\frac{\sum_{i=1}^{n}(y_i - \hat{R}x_i)^2}{n-1}\right]$$

En estas expresiones $f = \frac{n}{N}$ y se conoce como fracción de muestreo, $\bar{X}$ es la media de la población (parámetro) para la característica x_i, en el caso de que se desconozca se sustituye $\bar{X}$ por su estimador $\bar{x}$, la media muestral para la característica x_i.

La expresión : $\Sigma (y_i - \hat{R} x_i)^2$ que aparece en el estimador de la varianza se puede calcular a través de su expresión equivalente

---

* Se habla de muestra grande cuando n > 30.

que resulta más cómoda:

$$\sum_{i=1}^{n}(y_i - \hat{R}x_i)^2 = \sum_{i=1}^{n} y_i^2 - 2\hat{R}\sum_{i=1}^{n} x_i y_i + \hat{R}^2 \sum_{i=1}^{n} x_i^2$$

Para realizar la estimación por intervalos para el parámetro de razón, suponemos que la distribución del estimador es Normal y que estamos trabajando con una muestra grande. Bajo estas consideraciones los límites de confianza son:

$$\left(\hat{R} - t\sqrt{\hat{V}(\hat{R})} \;,\; \hat{R} + t\sqrt{\hat{V}(\hat{R})}\right)$$

donde t es el desvío normal correspondiente a la probabilidad de confianza escogida.

La raíz cuadrada de la varianza es por definición el error estándar o la desviación estándar, de manera que los límites de confianza o intervalos de confianza se pueden expresar también como:

$$\left(\hat{R} - t\,\sigma(\hat{R}) \;,\; \hat{R} + t\,\sigma(\hat{R})\right)$$

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

a) Se pide estimar la proporción del gasto familiar dedicado a la alimentación, la estimación de la varianza y del error estándar correspondiente a esta proporción, así como intervalos de confianza del 95%.

b) Estimar la proporción del gasto familiar empleado en atención médica y medicamentos, su varianza, su error estándar e intervalos de confianza del 90%.

Los datos obtenidos en la encuesta son los siguientes:

x_i: representa gasto familiar

y_i: representa gasto correspondiente a alimentación

z_i: representa gasto correspondiente a atención médica y medicamentos.

LOS DATOS CORRESPONDEN A UN MES Y ESTAN EXPRESADOS EN PESOS

Familia (i)	Gastos familiares x_i	Gastos en alimentación y_i	Gastos en médicos y medicamentos z_i
1	10000	4000	1000
2	11000	4200	800
3	8000	4000	500
4	9000	3500	1000
5	8500	3000	300
6	5000	2500	0
7	10000	3500	400
8	6000	2000	400
9	4000	1500	0
10	12000	4500	850
11	9000	3000	500
12	3500	1500	500
13	5000	2000	0
14	2000	1500	0
15	8000	3000	900
16	7000	3500	800
17	9500	3000	1500
18	6000	2500	500
19	5000	2000	500
20	4500	2400	600
21	7800	3000	0
22	8000	2500	1200
23	3000	1800	0
24	5500	2000	400
25	4500	2500	0
26	7000	3500	1000
27	9000	3000	1200
28	8500	3800	500
29	12000	4500	1600
30	6500	2500	500
31	3800	1800	0
32	4000	1800	400
33	2900	2000	0
34	3500	1500	200
35	5000	2000	250
36	6800	3000	300
37	4000	1800	400
38	4500	2000	200
39	5800	2500	400
40	6500	3000	200
TOTALES	261,600	107,600	19,800

Para obtener la varianza necesitamos calcular x_i^2 , y_i^2 y $x_i y_i$

ESTOS DATOS SE PRESENTAN EN LA SIGUENTE TABLA:

Familia (i)	x_i^2	y_i^2	$x_i y_i$	z_i^2	$x_i z_i$
1	100000000	16000000	40000000		
2	121000000	17640000	46200000		
3	64000000	16000000	32000000		
4	81000000	12250000	31500000		
5	72250000	9000000	25500000		
6	25000000	6250000	12500000		
7	100000000	12250000	35000000		
8	36000000	4000000	12000000		
9	16000000	2250000	6000000		
10	144000000	20250000	54000000		
11	81000000	9000000	27000000		
12	12250000	2250000	5250000		
13	25000000	4000000	10000000		
14	4000000	2250000	3000000		
15	64000000	9000000	24000000		
16	49000000	12250000	24500000		
17	90250000	9000000	28500000		
18	36000000	6250000	15000000		
19	25000000	4000000	10000000		
20	20250000	5760000	10800000		
21	60840000	9000000	23400000		
22	64000000	6250000	20000000		
23	9000000	3240000	5400000		
24	30250000	4000000	11000000		
25	20250000	6250000	11250000		
26	49000000	12250000	24500000		
27	81000000	9000000	27000000		
28	72250000	14440000	32300000		
29	144000000	20250000	54000000		
30	42250000	42250000	16250000		
31	14440000	3240000	6480000		
32	16000000	3240000	7200000		
33	8410000	4000000	5800000		
34	12250000	2250000	5250000		
35	25000000	4000000	10000000		
36	46240000	9000000	13600000		
37	16000000	3240000	7200000		
38	20250000	4000000	9000000		
39	33640000	6250000	14500000		
40	42250000	9000000	19500000		
	1,973,320,000	318,800,000	776,740,000		

El estimador de R, o sea de la proporción del gasto familiar dedicado a la alimentación, utilizando los datos de la muestra quedan dados por:

$$\hat{R} = \frac{\sum_{i=1}^{40} y_i}{\sum_{i=1}^{40} x_i} = \frac{107,600}{261,000} = 0.4099 \simeq 0.41$$

significa que el 41% del gasto familiar se dedica a la alimentación.

Calculemos el estimador de varianza

$$\hat{V}(\hat{R}) = \frac{1-f}{n\bar{x}^2}\left[\frac{\sum_{i=1}^{40} y_i^2 - 2\hat{R}\sum_{i=1}^{40} x_i y_i + R^2 \sum_{i=1}^{40} x_i^2}{n-1}\right]$$

$$= \frac{1 - \frac{40}{2000}}{(40)(6525)^2}\left[\frac{318,800,000 + (0.41)^2(1973320000) - 2(0.41)(776740000)}{40-1}\right]$$

En este caso el parámetro $\bar{X}$ es desconocido, de manera que utilizamos su estimador:

$$\bar{x} = \frac{\sum_{i=1}^{40} x_i}{40} = \frac{261,600}{40} = 6525$$

continuando con la varianza tenemos:

$$\hat{V}(\hat{R}) = 0.002 \quad , \quad \sigma(\hat{R}) = \sqrt{0.002} = 0.044$$

95% de confianza nos da una $t = 1.96 \simeq 2$

Los intervalos de confianza son:

$$[0.41 - 2(0.044) \ , \ 0.41 + 2(0.44)] = (0.322 \ , \ 0.498)$$

quiere decir que la proporción del gasto familiar que se dedica a la alimentación se encuentra entre 0.322 y 0.498 con 95% de confianza. Dicho de otra manera significa que el porcentaje del gasto familiar empleado en alimentación se encuentra entre 32.2% y 49.8% con 95% de confianza.

La parte b de este problema la dejamos a cargo del lector.

4.2 ESTIMACION DE VALORES MEDIOS Y DE TOTALES CON ESTIMADORES DE RAZON.

En muchas ocasiones podemos mejorar la precisión de nuestras estimaciones de medias y totales utilizando estimadores de razón. Para utilizar estos estimadores es necesario utilizar una variable auxiliar, esta variable auxiliar es el valor de una característica que debe estar correlacionada positivamente con la característica que estamos investigando.

Generalmente se pide que el coeficiente de correlación entre la característica de interés (y_i) y la característica auxiliar (x_i), tome un valor mayor que 0.5. El coeficiente de correlación se calcula con la siguiente expresión.

$$\rho = \frac{cov(x,y)}{\sigma(x)\,\sigma(y)}$$

donde cov (x , y) es la covarianza entre las variables x e y. Y σ (x) y σ (y) son respectivamente las desviaciones estándar de x e y.

El coeficiente de correlación ρ toma valores entre -1 y 1, es decir $-1 \leq \rho \leq 1$

Cuando ρ toma valores positivos indica que a medida que una variable aumenta la otra también aumenta, valores negativos de ρ indican que cuando una de las variables decrece, la otra crece, y cuando $\rho = 0$ indica que no existe dependencia lineal entre las variables.

El valor de la variable auxiliar puede ser un valor anterior de la misma característica o puede ser un dato que ya se tenga de la población. De otra forma la variable auxiliar se puede obtener con la encuesta.

En estos casos los estimadores que se proponen son:

Para la media $$\hat{\bar{Y}}_R = \bar{y}_R = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \bar{X} = \hat{R}\bar{X}$$

Para el total $$\hat{Y}_R = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} X = \hat{R}X$$

En estos casos los valores $\bar{X}$ y X son los verdaderos en la población, es decir son parámetros para la característica x_i. y es necesario conocerlos para utilizar estos estimadores.

Las expresiones de las varianzas y sus estimadores son:

$$V(\hat{\bar{Y}}_R) \doteq \frac{1-f}{n} \frac{\sum_{i=1}^{N} (y_i - Rx_i)^2}{N-1}$$

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

$$V(\hat{Y}_R) \doteq \frac{N^2(1-f)}{n} \frac{\sum^{N} (y_i - R x_i)^2}{N-1}$$

$$\hat{V}(\hat{Y}_R) \doteq \frac{N^2(1-f)}{n} \frac{\sum^{n} (y_i - \hat{R} x_i)^2}{n-1}$$

Los estimadores de razón si se utilizan adecuadamente son más precisos que los estimadores obtenidos con M.A.S., es decir que $\bar{y}$ y $N\bar{y}$.

Ejemplo 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 fue de 780 kilos. Después de un mes se desea conocer el peso medio por pollo y el peso total. Se selecciona una muestra aleatoria simple de 30 pollos que nos proporciona la siguiente información:

Pollo (i)	Peso inicial x_i	Peso al mes y_i	x_i^2	y_i^2	$x_i y_i$
1	1	2.5	1	6.25	2.5
2	1.1	2.7	1.21	7.29	2.97
3	1	2.6	1	6.76	2.6
4	1.2	2.7	1.44	7.29	3.24
5	1.4	3.2	1.96	10.24	4.48
6	1.4	3.4	1.96	11.56	4.76
7	1.2	2.9	1.44	8.41	3.48
8	1.5	3.4	2.25	11.56	5.1
9	1.4	3.4	1.96	11.56	4.76
10	1	2.5	1	6.25	2.5
11	1.3	2.6	1.69	6.76	3.38
12	1.1	2.6	1.21	6.76	2.86
13	1.5	3.2	2.25	10.24	4.8
14	1.4	3.2	1.96	10.24	4.48
15	1.1	2.9	1.4	8.41	3.19
16	1.2	2.6	1.44	6.76	3.12
17	1.4	3.2	1.96	10.24	4.48
18	1.1	2.6	1.21	6.76	2.86
19	1.3	2.9	1.69	8.41	3.77
20	1.2	2.7	1.44	7.29	3.24
21	1.3	2.9	1.69	8.41	3.77
22	1	2.5	1	6.25	2.5
23	1.4	3.2	1.96	10.24	4.48
24	1.5	3.2	2.25	10.24	4.8
25	1.3	2.7	1.69	7.29	3.51
26	1.1	2.7	1.21	7.29	2.97
27	1.2	3.4	1.44	11.56	4.08
28	1.3	2.6	1.69	6.76	3.38
29	1.5	3.2	2.25	10.24	4.8
30	1.1	2.6	1.21	6.76	2.86
TOTALES	37.5	86.8	47.67	254.08	109.65

Estimación del peso medio por pollo:

$$\hat{\bar{Y}}_R = \frac{86.8}{37.5} \times 1.3 = 2.31 \times 1.3 = \underline{3.003 \text{ kilos}}$$

donde $\hat{R} = 2.31$ y $\bar{X} = 1.3$

Estimación del peso total:

$$\hat{Y}_R = 2.31 \times 780 = \underline{1801.8 \text{ kilos}}$$

Estimación de las varianzas.

$$\hat{V}(\hat{\bar{Y}}_R) \doteq \frac{1-f}{n} \frac{(\sum y_i^2 - 2\hat{R}\sum x_i y_i + \hat{R}^2 \sum x_i^2)}{n-1}$$

$$\doteq \frac{(1 - \frac{30}{600})(254.08 - 2 \times 2.31 \times 109.65 + (2.31)^2 47.67)}{30 \cdot 29}$$

$$\doteq \frac{\frac{19}{20}(254.08 - 506.583 + 254.37)}{870} = .00204$$

$$\hat{V}(\hat{\bar{Y}}_R) \doteq 0.00204$$

$$\hat{V}(\hat{Y}_R) \doteq N^2 \hat{V}(\hat{\bar{Y}}_R) = 600^2(0.00204) = 734.4$$

Estimación de los errores estándar:

$$\sigma(\hat{\bar{Y}}_R) = \sqrt{0.00204} = .045 \text{ kilos}$$

$$\sigma(\hat{Y}_R) = \sqrt{734.4} = 27.1 \text{ kilos}$$

Estimación de intervalos de confianza del 95%

Para la media $\left(\hat{\bar{Y}}_R \mp t\,\sigma(\hat{\bar{Y}}_R)\right)$

Para el total $\left(\hat{Y}_R \mp t\,\sigma(\hat{Y}_R)\right)$

Donde t corresponde al desvío normal correspondiente a la probabilidad de confianza deseada.

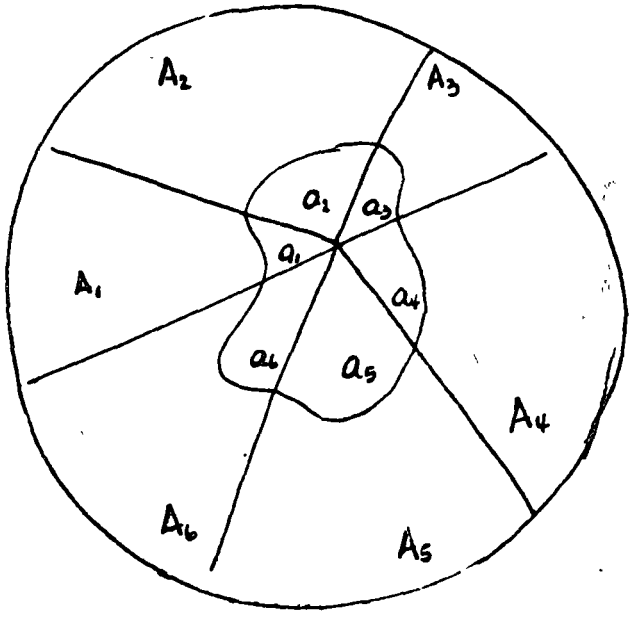
Para el 95% de confianza corresponde un t = 1.96. Realizando los cálculos para los intervalos de confianza tenemos:

media: (3.003 ± 1.96 x 0.045) = (2.9148, 3.0912)

total: (1801.8 ± 1.96 x 27.1) = (1748.7 , 1854.9)

4.3 ESTIMACION DE PORCENTAJES CUANDO SE ESTUDIAN VARIAS "CLASES" EN LA MISMA POBLACION.

Las clases determinan una partición dentro de la población, es decir una unidad pertenece a una y sólo una clase.



Supongamos una población dividida en 6 clases, de manera que el número de unidades que caen en la clase 1, la denominamos A_1,...., las que caen en la clase 6, A_6. Así el número de unidades en la población es

$$N = \sum_{i=1}^{6} A_i$$

La proporción de unidades que pertenece a la clase i en la población está dada por:

$$P_i = \frac{A_i}{N}$$

El porcentaje de unidades en la población que pertenecen a la clase i es

$$\mathcal{P}_i = \frac{A_i}{N}(100)$$

Al obtener una muestra de la población vamos a tener a_1 unidades que pertenecen a la clase 1, ..., a_6 unidades de la clase 6. Y el tamaño de la muestra n es:

$$n = \sum_{i=1}^{6} a_i$$

El estimador de la proporción es

$$\hat{P}_i = p_i = \frac{a_i}{n}$$

El estimador del porcentaje es

$$\hat{\mathcal{P}}_i = \frac{a_i}{n}(100)$$

Si estamos interesados en la proporción de unidades que pertenecen a la clase i y a la clase j entonces su estimador es

$$\hat{P}_{i+j} = \frac{a_i + a_j}{n}$$

El estimador del porcentaje es

$$\hat{\mathcal{P}}_{i+j} = \frac{a_i + a_j}{n}(100)$$

En cualquiera de los casos la varianza y su estimador se calculan con las siguientes expresiones:

$$V(p) = \frac{NPQ}{N-1}\frac{(1-f)}{n}, \qquad Q = 1-P.$$

$$\hat{V}(p) = \frac{N-n}{N}\frac{pq}{n-1} = \frac{1-f}{n-1}pq, \qquad q = 1-p.$$

Ejemplo: En un centro universitario se elige una muestra aleatoria simple de 50 estudiantes entre 2000. Se desea saber que porcentaje de estudiantes tienen menos de 18 años, tienen entre 18 y 24 años y que porcentaje de estudiantes son mayores de 24 años. Los datos de la muestra se presentan en el siguiente cuadro:

Clase	a_i
Menores de 18 años	5
Entre 18 y 24 años	32
Mayores de 24 años	13
	n=50

Estimación de los porcentajes:

$\hat{P}_1 = \frac{5}{50} \times 100 = 10\%$ de estudiantes menores de 18 años

$\hat{P}_2 = \frac{32}{50} \times 100 = 64\%$ de estudiantes entre 18 y 24 años

$\hat{P}_3 = \frac{13}{50} \times 100 = 26\%$ de estudiantes mayores de 24 años.

Si nos interesara conocer el porcentaje de estudiantes menores de 24 años sería:

$$\hat{P}_{1+2} = \frac{5+32}{50}(100) = \frac{37}{50}(100) = 74\%.$$

ESTIMACION DE MEDIAS, PROPORCIONES Y TOTALES EN SUBPOBLACIONES

Una subpoblación o dominio de estudio es una parte de la población o subconjunto de la población sobre la que se tiene algún interés en especial.

Se selecciona la muestra de tamaño n en la población, pero de esta muestra una parte de ella que vamos a llamar n_d va a corresponder al dominio o subpoblación que nos interesa, que llamamos domínio d-ésimo, así el estimador de la media en el dominio d-ésimo es

$$\hat{\bar{Y}}_d = \frac{\sum_{i=1}^{n_d} y_{di}}{n_d}$$

es decir tomamos en cuenta sólo las características que corresponden al dominio de interés, la sumamos y la dividimos entre el número de ellas, es decir n_d.

Formalmente el estimador que se acaba de presentar es un estimador de razón. Pero en la mayoría de los casos se utiliza como estimador de la varianza la siguiente expresión

$$\hat{V}(\hat{\bar{Y}}_d) = \frac{1-f_d}{n_d} \frac{\sum_{i=1}^{n_d} (y_{di} - \bar{y}_d)^2}{n_d - 1} = \frac{1-f_d}{n_d} s_d^2$$

En esta expresión $\bar{y}_d$ es el estimador de $\bar{Y}_d$, $f_d = \frac{n_d}{N_d}$ y en el caso de no conocer N_d, que es bastante frecuente, entonces se sustituye f_d por $f = \frac{n}{N}$.

En una población se pueden considerar varios dominios de estudios y obtener estimaciones para cada uno de ellos, siempre que en la muestra caigan unidades correspondiente a los dominios de inte-

rés.

Para el caso de proporciones, se considera que la característica Yd_i vale uno o cero, según que la unidad se encuentre o no en la clase de interés. Así si en el dominio d-ésimo cayeron n_d unidades muestrales, se observan cuántas unidades de las n_d pertenecen a la clase de interés y al número de ellas le llamanos n_{dc} y el estimador de la proporción de unidades que pertenecen a la clase c en el dominio d-ésimo queda dado por

$$\hat{P}_d = p_d = \frac{n_{dc}}{n_d}$$

El estimador de la varianza de p_d es el mismo que el de la media considerando que y_{di} vale 1 ó 0 y que $\bar{y}_d$ se sustituye por p_d*.

Cuando nos interesa estimar totales en dominios de estudios, lo primero que tenemos que preguntarnos es si conocemos el tamaño del dominio, cuando tenemos el N_d, entonces los estimadores del total y de la varianza son:

$$\hat{Y}_d = N_d \bar{y}_d$$

$$\hat{V}(\hat{Y}_d) = N_d^2 \frac{1-f_d}{n_d} \frac{\sum^{n_d} (y_{di} - \bar{y}_d)^2}{n_d - 1}$$

Cuando no conocemos el tamaño del dominio N_d, entonces se proponen los siguientes estimadores para el total y para su varianza:

$$\hat{Y}_d = \frac{N}{n} \sum_{i=1}^{n_d} y_{di}$$

$$\hat{V}(\hat{Y}_d) = \frac{N^2(1-\frac{n}{N})}{n} (s_d')^2$$

---

* Si lo que se desea estimar es un porcentaje, entonces p_d se multiplica por 100.

donde $$(D_d')^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n_d} y_{di}^2 - \frac{(\sum_{i=1}^{n_d} y_{di})^2}{n}\right]$$

Veamos en un ejemplo como se utilizan los estimadores propuestos:

Una empresa desea instalar el servicio de guardería para sus empleados. Para ello desea estimar que proporción de empleados de base se beneficiarían con el servicio, el número medio de hijos en edad pre-escolar por empleado de base y el número total de niños entre los empleados de base. Para realizar las estimaciones se selecciona una muestra aleatoria simple de 20 empleados, de un total de 250. La información obtenida aparece en el siguiente cuadro.

Además para cada estimación se solicita intervalos de confianza del 90%.

En el cuadro, bajo la columna de "tipo" aparece una c cuando se trata de empleado de confianza y una b para el empleado de base, en la columna de guardería aparece, si cuando el empleado piensa utilizar el servicio de guardería y no en caso contrario.

EMPLEADO	TIPO	GUARDERIA	No DE HIJOS	y_{di}^2
1	c	no	2	
2	b	si	3	9
3	b	si	4	16
4	c	no	0	
5	b	si	5	25
6	c	no	3	
7	b	si	4	16
8	b	si	2	4
9	b	no	1	1
10	c	si	3	
11	b	no	0	0
12	c	no	1	
13	b	si	4	16
14	b	si	3	9
15	c	si	2	
16	b	si	3	9
17	b	si	4	16
18	c	si	5	
19	b	si	4	16
20	c	no	2	
				137

Para estimar la proporción de empleados de base que se beneficiarán con el servicio, necesitamos conocer el número de empleados de base en la muestra, es decir n_d, $n_d = 12$ y ahora veamos cuántos de ellos van a utilizar la guardería es decir n_{dc}, $n_{dc} = 10$, así nuestro estimador es:

$$p_d = \frac{n_{dc}}{n_d} = \frac{10}{12} = 0.833$$

El 83% de los empleados de base se beneficiarán con el servicio de guardería.

El estimador de la varianza

$$\hat{V}(p_d) = \frac{\left(1 - \frac{20}{250}\right)}{12} \quad \frac{\left[10\,(1-0.833)^2 + 2\,(0-0.833)^2\right]}{11} =$$

$$\hat{V}(p_d) = 0.0116 \quad , \quad \sigma(p_d) = 0.1077$$

Así el intervalo de confianza del 90% es

$$(p_d \pm t\,\sigma(p_d)) = (0.833 \pm 1.64\,(0.1077))$$

$$= (0.657 \ , \ 1.009)$$

Lo que significa que la proporción se encuentra entre 0.657 y 1 con 90% de confianza.

Para el estimador del número medio de hijos por empleado de base tenemos

$$\hat{\bar{y}}_d = \bar{y}_d = \frac{\sum^{12} y_{di}}{n_d} = \frac{3+4+5+4+2+1+0+4+3+3+4+4}{12} = \frac{37}{12}$$

$$\bar{y}_d = 3.08 \quad \text{hijos por empleado de base}$$

El estimador de la varianza:

$$\hat{V}(\bar{y}_d) = \frac{(1 - \frac{20}{250})}{12} \frac{\left[137 - \frac{(37)^2}{12}\right]}{11} = 0.16$$

El estimador del error estándar

$$\sigma(\bar{y}_d) = \sqrt{0.16} = 0.4$$

El intervalo de confianza del 90% es

$$(\bar{y}_d \mp \sigma(\bar{y}_d)) = (3.08 \mp 1.64\ (0.4)) = (2.454, 3.736)$$

El estimador del número total de niños para los empleados de base: En este caso no conocemos el número total de empleados de base, de manera que vamos a utilizar para el total:

$$\hat{Y}_d = \frac{N}{n} \sum_{i=1}^{12} y_{di} = \frac{250}{20}(37) = 12.5 \cdot 37 = 462.5 \text{ niños.}$$

El estimador nos dice que entre los empleados de base existen aproximadamente 462.5 niños.

$$\hat{V}(\hat{Y}_d) = \frac{250^2 (1 - \frac{20}{250})}{20} \frac{\left[137 - \frac{(37)^2}{20}\right]}{19} = 11055.6$$

$$\sigma(\hat{Y}_d) = 105.14 \text{ niños}$$

El intervalo de confianza del 90% es $(462.5 \pm 1.64(105.14)) =$

$= (290.07, 634.93)$

OTROS USOS DE LAS SUBPOBLACIONES

En la práctica se recomienda utilizar subpoblaciones o dominios de estudio, para cuando al realizar el trabajo de campo no se obtienen las n entrevistas que se planearon, o por algunas otras causas como pueden ser: las respuestas aparecen borrosas en los cuestionarios, la persona entrevistada se rehusa a contestar, el entrevistador se saltó accidentalmente la pregunta, etc.

DIFICULTAD PRACTICA EN LA APLICACION DEL M.A.S.

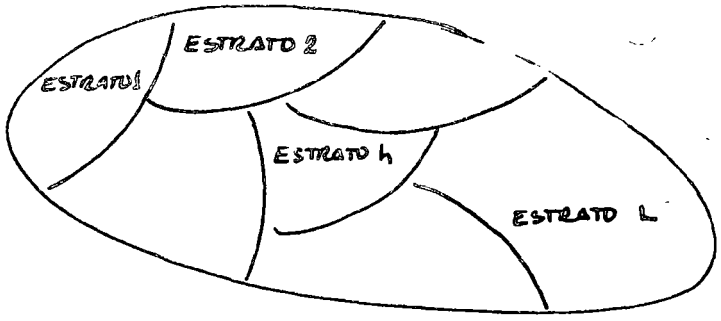
1. Es necesario un listado de la población para poder seleccionar la muestra, y cada unidad en la población debe tener asociado un número.

2. La muestra se encuentra dispersa en toda la población, lo que conduce a los entrevistadores a viajar por toda la población.

3. En la práctica se recomienda en las poblaciones pequeñas.

EL MUESTREO ESTRATIFICADO

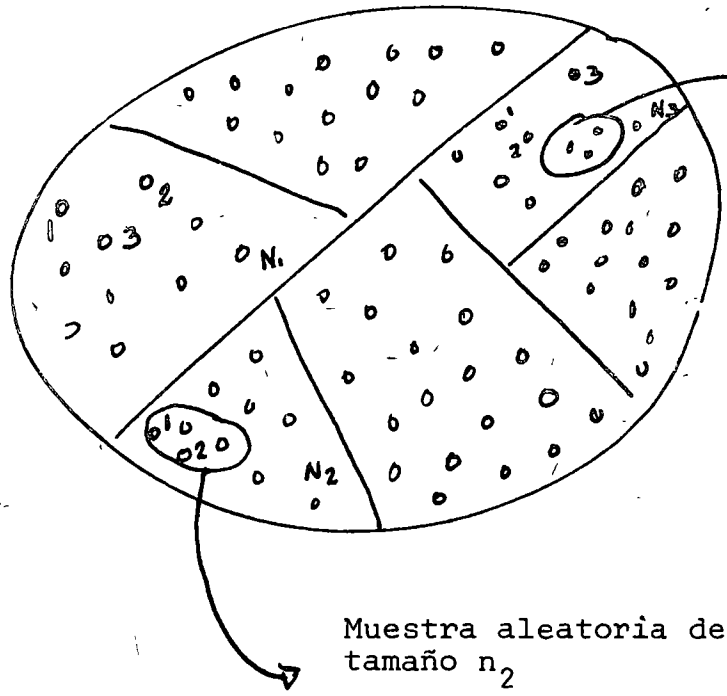
DESCRIPCION DEL METODO: La población sujeta a estudio consta de N unidades las cuales se encuentran repartidas en L subdivisiones de la población denominadas estratos. Cada unidad se localiza en uno y sólo en un estrato y la unión de todos los estratos conforma a la población original



Una vez que la población ha quedado dividida de esa manera, la selección se efectúa de la manera siguiente:

a) Elíjanse n_1 unidades en el estrato número 1
b) Elíjanse n_2 unidades en el estrato número 2 de manera independiente a la selección ya hecha en el estrato 1
c) Continue en la misma forma con el resto de los estratos.

NOTACION: En el estrato N° 1 existen N_1 unidades, en el estrato 2 existen N_2, ... y en el último estrato existen N_L unidades. De entre las N_1 unidades que conforman al estrato 1 se eligen a n_1 de ellas; de entre las N_2 unidades que conforman al estrato número 2 se eligen a n_2 de ellas y así sucesivamente



La media poblaciónal se define igual que antes, a saber:

$$\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{Y_1 + Y_2 + \dots + Y_L}{N} = \frac{N_1\bar{Y}_1 + N_2\bar{Y}_2 + \dots + N_L\bar{Y}_L}{N}$$

$$\bar{Y} = \frac{\sum_{h=1}^{h=L}\sum_{i=1}^{N_h} y_{hi}}{N} \qquad 5.1$$

donce Y_1 es el total de la característica en el estrato 1, Y_2 es el total de la característica en el estrato 2, y así sucesivamente, entonces $\bar{Y}_1$, $\bar{Y}_2$, ... se refieren a los valores medios en los estratos 1,2,... respectivamente. En la ecuación 5.1 la sumatoria sobre el índice i se desarrolla sobre el estrato h-ésimo. De acuerdo con la ecuación 5.1, el total poblacional es:

$$Y = N\bar{Y} = \sum_{h=1}^{h=L} \sum_{i=1}^{i=N_h} y_{hi}$$

En muestreo aleatorio simple se habla de la varianza S^2, en muestreo estratificado se habla de la varianza S^2 dentro de cada estrato, por lo cual también es necesario emplear un subíndice para ella, a saber: S_1^2 , S_2^2,..., S_L^2, donde:

$$S_h^2 = \frac{\sum_{i=1}^{i=N_h} (y_{hi} - \bar{y}_h)^2}{N_h - 1}$$

Su Uso: Al usar muestreo estratificado es posible dar estimaciones para cada estrato ya que se hacen selecciones independientes para cada uno de ellos; y por la misma razón esas estimaciones por estrato pueden efectuarse con la misma precisión solicitada para cada una de ellas; ya que bajo muestreo aleatorio simple; el tamaño de muestra requerido en el estrato h-ésimo sería:

$$n_{oh} = \frac{S_h^2}{V_n} \qquad \text{para estimar una media}$$

$$n_{oh} = \frac{N_h^2 S_h^2}{V_h} \qquad \text{para estimar un total}$$

$$n_{oh} = \frac{P_h Q_h}{V_h} \qquad \text{para estimar un porcentaje}$$

y en los tres casos

$$n_h = \frac{n_{oh}}{1 + \frac{n_{oh}}{N_h}}$$

y el tamaño total de la muestra será:

$$n = n_1 + n_2 + n_3 + \ldots + n_L$$

En ocasiones, aunque no se requieran estimaciones por separado para cada estrato, este esquema es usado para facilitar la selección y el trabajo de campo.

Ejemplo 5.1: Los empleados de una institución parecen listados en la nómina general, pero esta se presenta fragmentada por departamentos de la manera siguiente:

Depto 1	Depto 2	Depto 6
1. _____	1. _____	1. _____
2. _____	2. _____	2. _____
3. _____	3. _____	3. _____
.	.	.
.	.	.
.	.	.
300 _____	500 _____	400 _____

TOTAL DE EMPLEADOS = 300 + 500 + . . . + 400 = 5.000

Para llevar a cabo un estudio sobre estos empleados se decide emplear estas listas y como cada una de ellas presenta numeración independiente, para facilitar la selección se define a cada departamento o a cada lista como un estrato. En estas condiciones existen 6 estratos (L = 6) y sus tamaños son $N_1 = 300$, $N_2 = 500$, ..., $N_6 = 400$. El tamaño de esta población es de

$N = 300 + 500 + \ldots + 400 = 5000$ empleados

ESTIMACION DE VALORES MEDIOS: Para estimar al valor medio $\bar{Y}$ se usa la media estratificada $\bar{y}_{est}$ definida de la manera siguiente:

$$\bar{y}_{est} = \frac{N_1\bar{y}_1 + N_2\bar{y}_2 + \cdots + N_L\bar{y}_L}{N} = \frac{\sum_{h=1}^{L} N_h \bar{y}_h}{N}$$

donde $\bar{y}_h$ es el estimador del valor medio en el estrato h-ésimo. Si la selección dentro de cada estrato se ha hecho con muestreo aleatorio simple:

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}$$

y además

$$\hat{V}(\bar{y}_h) = \left(1 - \frac{n_h}{N_h}\right)\frac{s_h^2}{n_h} = \frac{\left(1 - \frac{n_h}{N_h}\right)}{n_h(n_h - 1)}\left[\sum y_{hi}^2 - \frac{\left(\sum^{n_h} y_{hi}\right)^2}{n_h}\right]$$

entonces

$$\hat{V}(\bar{y}_{est}) = \sum_h \left(\frac{N_h}{N}\right)^2 \hat{V}(\bar{y}_h)$$

Intervalos confidenciales del 95% para la media poblacional se calculan como sigue:

$$L_i = \bar{y}_{est} - 2\left(\sqrt{\hat{V}(\bar{y}_{est})}\right)$$

$$L_s = \bar{y}_{est} - 2\left(\sqrt{\hat{V}(\bar{y}_{est})}\right).$$

AFIJACION PROPORCIONAL: En muchas ocasiones se tiene como dato el tamaño total n de la muestra que debe emplearse; y el problema consiste en distribuir o afijar, esta muestra entre los diferentes estratos. Un método ampliamente usado para hacaerlo es la "afijación proporcional" y ella es tal que los tamaños de muestra para cada

estrato se calculan como sigue;

$$n_1 = \frac{N_1}{N} n$$

$$n_2 = \frac{N_2}{N} n$$

$$\vdots$$

$$n_L = \frac{N_L}{N} n$$

Ejemplo 5.2: En referencia al ejemplo 5.1 anterior el cual en el departamento N° 1 existen 300 empleados, en el N° 2 existen 500, ..., y en el N° 6 400 empleados, y siendo el total de ellos N=5000; podemos encontrar el tamaño de la muestra que le corresponde a cada estrato si el total de la muestra debe ser de 50 y se utiliza afijación proporcional;

$$n_1 = \frac{300}{5000} 50 = \frac{15000}{5000} = 3$$

$$n_2 = \frac{500}{5000} 50 = \frac{25000}{5000} = 5$$

$$\vdots$$

$$n_6 = \frac{400}{5000} 50 = \frac{20000}{5000} = 4$$

Ejemplo 5.3: Supongamos que en el ejemplo de los empleados en 6 departamentos, la característica en estudio es el número de hijos por empleados y lo que se desea estimar es el número medio de hijos por empleado. En el estrato 1 existen 300 empleados y la muestra es de tamaño 3 (ejemplo 5.2). Para obtener la muestra en este es-

trato se eligen 3 números aleatorios entre 1 y 300 y resultan seleccionados los empleados cuyos números son : 57, 288 y 17. Se localizó a los empleados y se les preguntó por el número de hijos que tiene cada uno de ellos; las repuestas fueron: 1, 1, 4. Entonces para este estrato, el número medio estimado de hijos por empleado, está dado por la media muestral:

$$\bar{y}_1 = \frac{y_1 + y_2 + y_3}{3} = \frac{1 + 1 + 4}{3} = 2$$

y para efectos de la obtención de intervalos de confianza del 95% calculamos s_1^2 y posteriormente $\hat{V}(\bar{y}_1)$

$$s_1^2 = \frac{\sum^3 y_{ni}^2 - \frac{(\Sigma y_{ni})^2}{n_1}}{n_1 - 1} = \frac{18 - \frac{36}{3}}{3 - 1} = \frac{18 - 12}{2} = \frac{6}{2} = 3$$

$$\hat{V}(\bar{y}_1) = \left(1 - \frac{n_1}{N_1}\right)\frac{s_1^2}{n_1} = \left(1 - \frac{3}{300}\right)\frac{3}{3} = \frac{297}{300} \doteq 1$$

En los demás estratos se trabajó de manera similar obteniéndose:

$\bar{y}_2 = 3.1$	$\hat{V}(\bar{y}_2) = 3$
$\bar{y}_3 = 2$	$\hat{V}(\bar{y}_3) = 2.5$
$\bar{y}_4 = 2.7$	$\hat{V}(\bar{y}_4) = 2.1$
$\bar{y}_5 = 6.0$	$\hat{V}(\bar{y}_5) = 7.2$
$\bar{y}_6 = 2$	$\hat{V}(\bar{y}_6) = 1$

Resumamos toda la información:

Estrato	Tamaño N_h	Muestra n_h	$\bar{y}_h$	$\hat{V}(y_h)$
1	300	3	2	1
2	500	5	3.1	3
3	1000	10	2	2.5
4	800	8	2.7	2.1
5	2000	20	6.0	7.2
6	400	4	2	1
TOTALES	5000	50		

Entonces la media estratificada vale

$$\bar{y}_{est} = \frac{300(2) + 500(3.1) + 1000(2) + 800(2.7) + 2000(6.0) + 400(2)}{5000}$$

$$= \frac{600 + 1550 + 2000 + 2160 + 12000 + 800}{5000} = \frac{19110}{5000} = 3.822$$

$\bar{y}$ est = 3.822 hijos por empleado en toda la institución. 3.822 es el número medio estimado de hijos por empleado en toda la institución. Ahora calculemos $\hat{V}(\bar{y}_{est})$

$$\hat{V}(\bar{y}_{est}) = \left(\frac{300}{5000}\right)^2(1) + \left(\frac{500}{5000}\right)^2(3) + \left(\frac{1000}{5000}\right)^2(2.5) + \left(\frac{800}{5000}\right)^2(2.1) + \left(\frac{2000}{5000}\right)^2(7.2) + \left(\frac{400}{5000}\right)^2(1)$$

$$= 1.346$$

y el error estándar vale $\sqrt{1.346} = 1.16$

Luego el intervalo de confianza del 95% vale

$$L_i = 3.822 - 2(1.16) = 1.502$$

$$L_s = 3.822 + 2(1.16) = 6.142$$

ESTIMACION DE TOTALES Y DE PORCENTAJES:

Para estimar a un total solo es necesario multiplicar el tamaño N de la poblaci on por la media estratificada:

$$\hat{Y} = N\bar{y}_{est} = N_1\bar{y}_1 + N_2\bar{y}_2 + \cdots + N_L\bar{y}_L$$

$$\hat{V}(\hat{Y}) = \hat{V}(N\bar{y}_{est}) = N_1^2\hat{V}(\bar{y}_1) + N_2^2\hat{V}(\bar{y}_2) + \cdots + N_L^2\hat{V}(\bar{y}_L)$$

y los intervalos del 95% de confianza se calculan como:

$$L_i = N\bar{y}_{est} - 2\left(\hat{V}(N\bar{y}_{est})\right)^{1/2}$$

$$L_s = N\bar{y}_{est} + 2\left(\hat{V}(N\bar{y}_{est})\right)^{1/2}$$

Ejemplo 5.4 : Continuando con el ejercicio anterior podemos estimar el número total de hijos que tienen los 5,000 empleados en la institución. En base a la estratificación efectuada y considerando la media estratificada ya calculada, así como su varianza tenemos.

$$\hat{Y} = N\bar{y}_{est} = 5000\ (3.822) = 19,110 \text{ HIJOS}$$

Esté es el valor estimado del total solicitado. Para obtener intervalos confidenciales para este total calculamos

$$\hat{V}(N\bar{y}_{est}) = N^2\hat{V}(\bar{y}_{est}) = (5000)^2(1.346) = 33,650,000$$

El error estandar vale $\sqrt{33,650,000} = 5800.86$ Por lo cual el intervalo es [7508.28; 30711.72] esta estimación es poco precisa ya que el intervalo está muy abierto.

Para estimar porcentaje usamos la expresión siguiente:

$$p_{est} = \frac{\sum_{h=1}^{L} N_h p_h}{N} = \frac{N_1 p_1 + N_2 p_2 + \ldots + N_L p_L}{N}$$

donde p_h es el estimador del porcentaje en el estrato h, con muestreo aleatorio simple dentro de cada estrato, y este estimador vale:

$$p_h = \frac{a_h}{n_h} \times 100 = \frac{\text{CASOS FAVORABLES EN EL ESTRATO } h}{\text{TAMAÑO DE MUESTRA EN EL ESTRATO } h}(100)$$

donde la varianza estimada de p_{est} es

$$\hat{V}(p_{est}) = \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2(N_h - n_h)}{N_h}\frac{p_h q_h}{n_h - 1} = \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h(N_h - n_h)}{n_h - 1}p_h q_h$$

y $q_h = 1 - p_h$.

entonces, para calcular intervalos confidenciales del 95% usamos

$$L_i = p_{est} - 2\left(\hat{V}(p_{est})\right)^{1/2}$$

$$L_s = p_{est} + 2\left(\hat{V}(p_{est})\right)^{1/2}$$

Ejemplo 5.5 Continuando con el ejemplo de los empleados que están repartidos en 6 departamentos, se quiere conocer que porcentaje de ellos son del sexo femenino. Para ello utilizamos la misma muestra y se obtiene la información siguiente:

Estrato	Tamaño N_h	Muestra n_h	Personas del sexo fem. a_h	p_h	q_h
1	300	3	1	1/3	2/3
2	500	5	2	2/5	3/5
3	1000	10	4	4/10	6/10
4	800	8	3	3/18	5/8
5	2000	20	8	8/20	12/20
6	400	4	2	2/4	2/4
TOTALES	5000	50			

La proporción estimada de personas del sexo femenino:

$$p_{est} = \frac{300(33.3) + 500(40) + 1000(40) + 800(37.5) + 2000(40) + 400(50)}{5000}$$

$$p_{est} = \frac{10000 + 20000 + 40000 + 30000 + 80000 + 20000}{5000} = \frac{200000}{5000} = 40\%$$

La estimación del porcentaje de personas del sexo femenino es: 40%.

El estimador de la varianza con afijación proporcional

$$\hat{V}(p_{est}) = \frac{\left(1 - \frac{50}{5000}\right)}{5000}\left[\frac{300^2(33.3 \times 66.6)}{50(300) - 5000} + \frac{500^2(40 \times 60)}{50(500) - 5000} + \frac{1000^2(40 \times 60)}{50(1000) - 5000} + \right.$$

$$\left. + \frac{800^2(37.5 \times 62.5)}{50(800) - 5000} + \frac{2000^2(40 \times 60)}{50(2000) - 5000} + \frac{400^2(50 \times 50)}{50(400) - 5000}\right] =$$

$$\hat{V}(p_{est}) = 64.1$$

$$\sigma(p_{est}) = \sqrt{54.1} = 7.36\ \%$$

el estimador de la varianza del porcentaje es 54.1 y su error estandar 7.36

Intervalo de confianza del 95% para el porcentaje:

$$L_i = 40 - 2(7.36) = 25.28$$

$$L_s = 40 + 2(7.36) = 54.72$$

FORMULAS PARA EL CALCULO DE LOS TAMAÑOS DE LA MUESTRA

Si se desea estimar una media poblacional y la muestra se afija proporcionalmente

Primero calculamos $$n_o = \frac{1}{NV} \sum_{h=1}^{L} N_h S_h^2$$

Donde V se refiere a la varianza de la media estratificada.
Si $\frac{n_o}{N}$ es despreciable tomamos n_o como el tamaño de muestra indicado, en caso contrario corregimos el tamaño de muestra con

$$n = \frac{n_o}{1 + \frac{n_o}{N}}$$

Después que se ha obtenido el tamaño de muestra para la población total se reparte o afija proporcionalmente con la siguiente expresión:

$$n_h = \left(\frac{N_h}{N}\right) n \qquad \text{para} \quad h = 1, 2, \ldots, L.$$

Para estimar el total poblacional se sigue el mismo procedimiento lo único que las expresiones con las que se calculan los tamaños de muestra son las siguientes:

$$n_0 = \left(\frac{N}{V}\right) \sum_{h=1}^{L} N_h S_h^2$$

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

y la muestra se afija de la misma forma.

Para cuando se desea estimar un porcentaje entonces las expresiones para calcular el tamaño de muestra con afijación proporcional son:

$$n_0 = \frac{1}{NV} \sum_{h=1}^{L} N_h P_h Q_h$$

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

En este caso la varianza V se expresa como $(\frac{d}{t})^2$ en donde d , P_h, Q_h están expresadas en porcentajes.

ESTIMACION DE VALORES MEDIOS Y DE TOTALES EN SUBPOBLACIONES DE TAMAÑO CONOCIDO.

Si queremos conocer una estimación global para una subpoblación o un dominio de estudio cuando se ha realizado una estratificación en la población, debemos tener presente que el dominio se va encontrar esparcido en todos los estratos.

Vamos a suponer que estamos trabajando con el dominio d-ésimo de tamaño N_h, existen N_{hd} unidades que pertenecen al dominio d-ésimo, del estrato h seleccionamos n_h unidades muestrales de las cuales n_{hd} pertenecen al dominio d, esto ocurre para cada uno de los estratos de la población.

De manera que para estimar la media de una característica del dominio utilizamos la siguiente expresión:

$$\hat{\bar{Y}}_d = \frac{\sum_h N_{hd} \sum_{i=1}^{n_{hd}} \frac{y_{hdi}}{n_{hd}}}{\sum_h N_{hd}} = \frac{\sum_h N_{hd} \bar{y}_{hd}}{\sum N_{hd}}$$

Y como estimador de su varianza:

$$\hat{V}(\hat{\bar{Y}}_d) = \frac{1}{(\sum_h N_{hd})^2} \sum_h \frac{N_{hd}^2 S_{hd}^2}{n_{hd}} \left(1 - \frac{n_{hd}}{N_{hd}}\right)$$

Para estimar el total del dominio:

$$\hat{Y}_d = \sum_h N_{hd} \sum_{i=1}^{n_{hd}} \frac{y_{hdi}}{n_{hd}} = \sum_h N_{hd} \bar{y}_{hd}$$

y como estimador de su varianza:

$$\hat{V}(\hat{Y}_d) = \sum_h \frac{N_{hd}^2 S_{hd}^2}{n_{hd}} \left(1 - \frac{n_{hd}}{N_{hd}}\right)$$

donde $S_{hd}^2 = \frac{1}{n_{hd}-1} \left[\sum_{i=1}^{n_{hd}} y_{hdi}^2 - \frac{(\sum y_{hdi})^2}{n_{hd}} \right]$

Es la varianza entre las unidades del domino d dentro del estrato h.

EL MUESTREO POR CONGLOMERADOS.

Descripción del método:

El muestreo por conglomerado se recomienda cuando no se dispone de un marco muestral de los elementos que forman la población y el costo de elaborar dicho marco es muy alto o cuando el costo de transporte entre los elementos de la población a investigar son muy altos. En estos casos lo que se sugiere es formar conjuntos o grupos de elementos a los que se les llama conglomerado y el marco muestral está formado por un listado de conglomerados, se selecciona una muestra de conglomerados y los conglomerados que forman la muestra se censan, es decir se investigan todos los elementos de los conglomerados que cayeron en la muestra. En el muestreo por conglomerados la característica que se investiga es una característica propia de los elementos que forman los conglomerados.

Veamos un ejemplo, se desea investigar que tipo de programas de televisión les agrada más a las personas de una ciudad, no se dispone de un listado de personas por lo que se hace un listado de viviendas, es decir que una vivienda es considerada un conglomerado formado por todas las personas que viven en ella.

En este caso la muestra está formada por n unidades muestrales, donde cada unidad muestral es un conglomerado, hay que tener pre

sente que en este esquema de muestreo cada unidad muestral nos proporciona un conjunto de observaciones que corresponden a los elementos que forman el conglomerado.

Las observaciones se van a representar por y_{ij}, donde el índice i representa el conglomerado y el índice j elemento dentro del conglomerado. Así la observación y_{23} representa al elemento 3 del conglomerado 2.

Notación: El método de selección de la muestra consiste en elegir n conglomerados de un total de N conglomerados que forman la población total. Cada conglomerado está formado por M_i elementos donde i representa el conglomerado. $M = \sum_{i=1}^{N} M_i$ representa el número total de elementos en la población $\bar{M} = \frac{M}{N}$ es el tamaño promedio de los conglomerados en la población de la misma forma $\bar{m} = \frac{\sum_{i=1}^{n} M_i}{n}$ es el tamaño promedio de los conglomerados en la muestra. $y_i = \sum_{j=1}^{M_i} y_{ij}$ representa el total de las observaciones del conglomerado i.

Con esta notación vamos a presentar los estimadores para el caso de muestreo por conglomerados.

ESTIMACION DE VALORES MEDIOS, DE TOTALES Y DE PORCENTAJES

En muestreo por conglomerados tenemos dos valores medios que son la media por unidad y la media por elemento.

La media por unidad se refiere al valor medio por conglomerado de alguna característica y el estimador que se utiliza es.

$$\hat{\bar{Y}} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{M_i} y_{ij}$$

El estimador de su varianza es:

$$\hat{V}(\bar{y}) = \frac{1-f}{n}\,\frac{\sum^{n}(y_i-\bar{y})^2}{n-1}$$

El estimador del total en la población es

$$\hat{Y} = N\bar{y} = \frac{N}{n}\sum_{i=1}^{n} y_i = \frac{N}{n}\sum_{i=1}^{n}\sum_{j=1}^{M_i} y_{ij}$$

como estimador de su varianza tenemos la siguiente expresión.

$$\hat{V}(\hat{Y}) = \frac{N^2(1-f)}{n}\,\frac{\sum^{n}(y_i-\bar{y})^2}{n-1}$$

Para la media por elemento se propone la siguiente expresión:

$$\hat{\bar{\bar{Y}}} = \bar{\bar{y}} = \frac{\hat{Y}}{M} = \frac{\sum_{i=1}^{n} y_i}{\bar{M}\,n}$$

que se puede analizar de la siguiente manera: para el valor medio de una característica debemos tener el total de la característica entre el número de elementos, por eso el estimador tiene en el numerador, un estimador del total y en el denominador tenemos el total de elementos. Este estimador se presenta con una doble barra para distinguirlo de la media por unidad.

El estimador de la varianza de la media por elemento es:

$$\hat{V}(\bar{\bar{y}}) = \frac{1-f}{n\bar{M}^2}\,\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1} = \frac{1-f}{n\bar{M}^2}\,\frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum^{n} y_i\right)^2\right]$$

Además de los estimadores presentados que son insesgados, existen

otros estimadores para conglomerados que se conocen como estimadores de razón, que en muchas ocasiones son más poderosos.

Para estimar la media por elemento se presenta el siguiente estimador:

$$\hat{\bar{\bar{Y}}}_R = \bar{\bar{y}}_e = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i}$$

como estimador de su varianza tenemos:

$$\hat{V}(\bar{\bar{y}}_e) \doteq \frac{1-f}{n\bar{M}^2} \frac{\sum_{i=1}^{n} (y_i - \bar{\bar{y}}_e M_i)^2}{n-1}$$

$$\doteq \frac{1-f}{n\bar{M}^2(n-1)} \left[\sum_{i=1}^{n} y_i^2 - 2\bar{\bar{y}}_e \sum_{i=1}^{n} M_i y_i + \bar{\bar{y}}_e^2 \Sigma M_i^2 \right]$$

Para estimar el total poblacional tenemos

$$\hat{Y}_e = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i} M$$

y para estimar la varianza del total

$$\hat{V}(\hat{Y}_e) \doteq \frac{N^2(1-f)}{n} \frac{\sum_{i=1}^{n} (y_i - \bar{\bar{y}}_e M_i)^2}{n-1} = M^2 \hat{V}(\bar{\bar{y}}_e)$$

Si el total de elementos en la población , M es desconocido sustituimos $\bar{M}$ por su estimador que es $\bar{m}$.

Para estimar porcentajes en una pobalción considerando muestreo por conglomerados, debemos tener presente que dentro de cada conglomerado los elementos pueden pertenecer a la clase C de interés o a su complemento. Así dentro del conglomerado i, el número de elementos que pertenecen a la clase de interés C lo representamos por a_i.

Así tenemos los siguientes estimadores:

$$\hat{\mathcal{P}}_R = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} M_i} \times 100$$

$$\hat{V}(\hat{\mathcal{P}}_R) \doteq \frac{1-f}{n\bar{M}^2} \frac{\left[\sum a_i^2 - 2\hat{\mathcal{P}}_R \sum a_i M_i + \hat{\mathcal{P}}_R^2 \sum M_i^2\right]}{n-1}$$

Ejemplo: En una ciudad que cuenta con 600 escuelas primarias y 9000 maestros, se desea realizar una investigación para estimar:

a) El número medio de alumnos por escuela que se retiraron antes de finalizar el año escolar.

b) El número total de alumnos que se retiraron de sus estudios primarios en la ciudad.

c) El número de alumnos por maestro que se retiraron de sus estudios

d) Porcentaje de maestros que están de acuerdo con una modificación del plan de estudios.

Para ello se considera cada escuela un conglomerado y se seleccionan 20 escuelas con muestreo aleatorio simple para formar la muestra.

De la muestra se obtienen los siguientes datos:

(i) Escuela	(M_i) No. de maestros por escuela	(y_i) No. de alumnos retirados	(a_i) Maestros de acuerdo con la modif.	a_iM_i	M_iy_i
1	12	38	8	96	546
2	15	40	9	135	600
3	10	28	6	60	280
4	6	30	4	24	180
5	8	32	4	32	256
6	10	35	9	90	350
7	9	40	9	81	360
8	12	45	10	120	540
9	14	43	10	140	602
10	11	38	9	99	418
11	18	52	12	216	936
12	20	65	20	400	1300
13	22	70	20	440	1540
14	12	29	11	132	348
15	24	75	23	552	1800
16	18	49	15	270	882
17	6	39	6	36	234
18	10	25	5	50	350
19	20	82	14	280	1640
20	15	64	9	135	960
TOTALES	272	919	213	3356	13932

$$\sum_{i=1}^{20} y_i^2 = 46533 \quad , \quad \sum_{i=1}^{20} M_i^2 = 4224 \quad , \quad \sum_{i=1}^{20} a_i^2 = 2813 \; .$$

Para caso se piden estimadores de varianza y de error estandar

a) $\bar{y} = \frac{\sum y_i}{n} = \frac{919}{20} = 45.95$ alumnos retirados por escuela

$$\hat{V}(\bar{y}) = \frac{(1 - \frac{20}{600})}{(20)(19)} \left[46533 - \frac{(919)^2}{20} \right] = 11.193$$

$$\sigma(\bar{y}) = \sqrt{11.193} = 3.36 \text{ alumnos.}$$

b) $\hat{Y} = N\bar{y} = 600 \cdot 45.95 = 27570$ alumnos retirados en la población

$$\hat{V}(\hat{Y}) = 600^2 (11.193) = 4,029.480$$

$$\sigma(\hat{Y}) = 2007.36 \text{ alumnos}$$

c) $\bar{\bar{y}} = \frac{\hat{Y}}{M} = \frac{27\,570}{9\,000} = 3.06$ alumos retirados por maestro.

$$\hat{V}(\bar{\bar{y}}) = \frac{(1 - \frac{20}{600})}{20 \left(\frac{9000}{600}\right)^2} \frac{[46533 - 42228]}{19} = 0.049$$

$$\sigma(\bar{\bar{y}}) = \sqrt{0.049} = 0.22 \text{ alumnos.}$$

d) $\hat{\rho}_R = \frac{\sum a_i}{\sum M_i} \times 100 = \frac{213}{272} \times 100 = 78.3\%$.

$$\hat{V}(\hat{P}_R) = \frac{\left(1-\frac{20}{600}\right)}{20(15)^2}\left[42813 - 2(783)(3356) + (783)^2(422K)\right]}{19}$$

$$= 267$$

$$\sigma(\hat{P}_R) = 16.3\%$$

Utilizando estimadores de razón estimar a) el número medio de alumnos por maestro que se retiraron y b) el número total de alumnos retirados en la población. Compare estos resultados con los anteriores.

a) $\bar{\bar{y}}_R = \frac{\Sigma y_i}{\Sigma M_i} = \frac{919}{272} = 3.38$ alumnos retirados por maestro

$$\hat{V}(\bar{\bar{y}}_R) = \frac{\left(1-\frac{20}{600}\right)}{20(15^2)}\left[46533 - 2(3.38)(13932) + (3.38)^2(422K)\right]}{19} = 0.0061$$

$$\sigma(\bar{\bar{y}}_R) = 0.078$$

b) $\hat{Y}_R = 3.38\,(9000) = 30420$ alumnos retirados

$$\hat{V}(\hat{Y}_R) = M^2\hat{V}(\bar{\bar{y}}_R) = (9000^2)\,0.0061 = 494,100$$

$$\sigma(\hat{Y}_R) = 702.92 \text{ alumnos}$$

COMENTARIOS SOBRE LOS TAMAÑOS DE LOS CONGLOMERADOS Y SOBRE SU CONSTRUCCION

En el muestreo por conglomerados la precisión de los estimadores depende mucho del tamaño de los conglomerados y de la estructura interna de los mismos.

En cuanto al tamaño de los conglomerados, puede ocurrir que en la población todos sean de igual tamaño, M_c, en cuyo caso los estimadores se simplifican notablemente, o puede ocurrir que los conglomerados sean de tamaños distintos, en cuyo caso los conglomerados que forman la muestra pueden ser elegidos con probabilidad proporcional al tamaño.

En muchas poblaciones al tratar de aplicar este diseño muestral, los conglomerados formados resultan muy grandes de tal manera que al tratar de investigar los conglomerados que cayeron en la muestra el censo de estos conglomerados, resulta impracticable, por lo que se recomienda cambiar de diseño muestra y probablemente un submuestreo sería adecuado.

Al construir los conglomerados debemos tratar que ellos sean lo más heterogéneos posible internamente, es lo que generalmente se expresa pidiendo que el coeficiente de correlación intraclásico sea negativo o muy cercano a cero. De esta manera se gana mucho en precisión con el esquema de muestreo por conglomerados

SELECCION DE CONGLOMERADOS CON PROBABILIDAD PROPORCIONAL AL TAMAÑO.

En este caso los conglomerados de mayor tamaño, tiene asignada una probabilidad mayor de aparecer en la muestra, y los conglomerados más chicos tienen una menor oportunidad de aparecer o de caer en la muestra. El estimador del tamaño del conglomerado debe estar correlacionado positivamente con la característica que se está investigando, cuando como medida del tamaño del conglomerado se conside-

ra el número de elementos que forman el conglomerado, entonces la probabilidad de selección del conglomerado i-ésimo está dada por

$Z_i = \frac{M_i}{\sum_{i}^{N} M_i}$ y se dice que la muestra es seleccionada con probabilidad proporcional al tamaño del conglomerado (ppt)

MANERA DE REALIZAR LA SELECCION DE LA MUESTRA CON ppt

Veamos una forma de seleccionar la muestra con probabilidad proporcional al tamaño: En una cadena de tiendas de autoservicio se desea realizar una encuesta entre sus empleados. La empresa cuenta con 8 tiendas con 60, 150, 80, 110, 76, 130, 95, 99 empleados respectivamente. Las probabilidades que se deben asignar a cada tienda son:

$Z_1 = \frac{60}{800}$, $Z_2 = \frac{150}{800}$, $Z_3 = \frac{80}{800}$, $Z_4 = \frac{110}{800}$, $Z_5 = \frac{76}{800}$

$Z_6 = \frac{130}{800}$, $Z_7 = \frac{95}{800}$ y $Z_8 = \frac{99}{800}$ DE TAL MANERA QUE

$$Z_1 + Z_2 + Z_3 + Z_4 + Z_5 + Z_6 + Z_7 + Z_8 = 1$$

Formamos el siguiente cuaddro:

Tienda	N° de empleados	N° acumulado de empleados	intervalo de selección
1	60	60	1 a 20
2	150	210	21 a 210
3	80	290	211 a 290
4	110	400	291 a 400
5	76	476	401 a 476
6	130	606	477 a 606
7	95	701	607 a 701
8	99	800	702 a 800

Vamos a elegir una muestra de n = 3, de manera que seleccionamos tres números aleatorios entre 1 y 800 buscamos los números en una tabla de números y resultan ser: 83, 695 y 420. Los conglomerados que forman la muestra son: la tienda 2 porque 83 está en el intervalo de 21 a 210, la tienda 5 porque 420 está en el intervalo de 401 a 476 y la tienda 7 porque 695 se encuentra en el intervalo 607 a 701.

Si la muestra fuera de 4 buscamos otro número en la tabla, esta resulta ser 199, que se encuentra entre 21 y 210 de manera que el conglomerado 2 está 2 veces en la muestra, o sea sus observaciones se repetirán 2 veces.

Por lo general con este tipo de selección se permite el reemplazo y un conglomerado puede aparecer varias veces en la muestra.

Los estimadores para cuando la muestra se selecciona con probabilidad proporcional al tamaño y con reemplazo son los siguientes:

Estimadores de la media por elemento y de su varianza:

$$\hat{\bar{Y}}_{ppt} = \bar{\bar{y}}_{ppt} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{M_i}$$

$$\hat{V}(\bar{\bar{y}}_{ppt}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{M_i} - \bar{\bar{y}}_{ppt}\right)^2$$

Estimadores del total y de su varianza:

$$\hat{Y}_{ppt} = M\,\bar{\bar{y}}_{ppt}$$

$$\hat{V}(\hat{Y}_{ppt}) = M^2\left[\hat{V}(\bar{\bar{y}}_{ppt})\right]$$

Estos estimadores son insesgados y se utilizan cuando se seleccio na una muestra de tamaño n de una población formada por N conglo merados donde el i-ésimo conglomerado es de tamaño M_i y es selec cionado con probabilidad proporcional al tamaño .

$$Z_i = \frac{M_i}{\Sigma M_i} .$$

SUBMUESTREO

El proceso de selección en submuestreo consiste en seleccionar una muestra aleatoria simple de conglomerados y después seleccionar una muestra aleatoria simple de elementos dentro de cada conglomerado que cayó en la muestra.

Para realizar esta selección debemos contar con un listado de los conglomerados que forman la población y después de elegir los conglomerados muestra, contar con un listado de los elementos de los conglomerados que forman la muestra, para poder realizar la selec ción de los elementos dentro de los conglomerados.

La población está formada por N conglomerados de los cuales se elige una muestra aleatoria simple de n conglomerados, donde el número de elementos del conglomerado i-ésimo lo representamos por M_i, m_i es el tamaño de la muestra dentro del conglomerado i. Así $M = \sum_{i=1}^{N} M_i$ es el número de elementos en la población, $\bar{M} = \frac{M}{N}$ es el tamaño medio de los conglomerados en la población, y_{ij} representa la característica j-ésima del i-ésimo conglomerado, a la media muestral del conglomerado i-ésimo la representamos por $\bar{y_i} = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$

Bajo estas condiciones los estimadores que se proponen son:

El estimador de la media poblacional $\bar{Y}$

$$\hat{\bar{Y}} = \bar{y} = \frac{N}{M} \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{n}$$

El estimador de la varianza de $\hat{\bar{Y}}$ es

$$\hat{V}(\hat{\bar{Y}}) = \frac{1-f_1}{n\bar{M}^2} s_1^2 + \frac{1}{nN\bar{M}^2} \sum_{i=1}^{n} M_i^2 (1-f_{2i}) \frac{s_{2i}^2}{m_i}$$

donde $$s_1^2 = \frac{\sum_{i=1}^{n} (M_i \bar{y}_i - \bar{M}\bar{y})^2}{n-1}$$

$$s_{2i}^2 = \frac{\sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2}{m_i - 1}$$

$$f_1 = \frac{n}{N} \quad , \quad f_{2i} = \frac{m_i}{M_i} \quad , \quad i = 1, 2, \ldots n$$

el estimador del total poblacional Y es:

$$\hat{Y} = M\bar{y} = N \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{n}$$

El estimador de la varianza de $\hat{Y}$

$$\hat{V}(\hat{Y}) = M^2 \hat{V}(\bar{y}) = \frac{1-f_1}{n} s_1^2 + \frac{N}{n} \sum_{i=1}^{n} M_i^2 (1-f_{2i}) \frac{s_{2i}^2}{m_i}$$

Estimadores de razón para la media y total poblacional

$$\hat{\bar{Y}}_R = \bar{y}_R = \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i}$$

La varianza estimada de $\hat{\bar{Y}}_R$ es:

$$\hat{V}(\bar{y}_R) = \frac{1 - f_1}{n \bar{M}^2} s_1^2 + \frac{1}{n N \bar{M}^2} \sum_{i=1}^{n} M_i^2 (1 - f_{2i}) \frac{s_{2i}^2}{m_i}$$

donde

$$s_1^2 = \frac{\sum_{i=1}^{n} M_i^2 (\bar{y}_i - \bar{y}_R)^2}{n - 1}$$

$$s_{2i}^2 = \frac{\sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2}{m_i - 1}$$

El estimador del total:

$$\hat{Y}_R = \bar{y}_R \sum_{i=1}^{N} M_i$$

El estimador de la varianza de $\hat{Y}_R$ es

$$\hat{V}(\hat{Y}_R) = \left(\sum_{i=1}^{N} M_i\right)^2 \hat{V}(\bar{y}_R)$$

El estimador de razón para la media se recomienda cuando M no se conoce, es decir cuando es desconocido el número total de elementos que forman la población.

Estimador de la proporción poblacional

Es estimar que parte de la población pertenece a una clase dada, para ello vamos a considerar a p_i la proporción de elementos del conglomerado i, que caen en la clase de interés.

El estimador de la proporción poblacional

$$\hat{P} = p = \frac{\sum_{i=1}^{n} M_i p_i}{\sum_{i=1}^{n} M_i}$$

La varianza estimada de p es:

$$\hat{V}(p) = \frac{1-f}{n\bar{M}^2} A_1^2 + \frac{1}{nN\bar{M}^2} \sum_{i=1}^{n} M_i^2 (1-f_{2i}) \frac{p_i q_i}{m_i - 1}$$

donde

$$A_1^2 = \frac{\sum_{i=1}^{n} M_i^2 (p_i - p)^2}{n-1} \qquad y \qquad q_i = 1 - p_i$$

MUESTREO SISTEMATICO

La población está formada por N unidades de las cuales se debe elegir una mustra de tamaño n. Para elegir esta muestra se procede de la siguiente forma:

Se calcula $k = \frac{N}{n}$, si k resulta ser un número entero entonces se elige un número aleatorio entre 1 y k, el que representaremos por r, de esta manera la muestra queda dada por las unidades:

$$r,\ r+k,\ r+2k,\ r+3k,\ \ldots$$

Si k no es entero, entonces se selecciona un número aleatorio entre 1 y N, al que se representa con r y las unidades que forman la muestra son:

$$\ldots,\ r-2k,\ r-k,\ r,\ r+k,\ r+2k,\ \ldots$$

esta muestra recibe el nombre de muestra sistemática cíclica, porque va hacia adelante y hacia atrás, hasta cubrir toda la población.

En el muestreo sistemático se encuentra la primera unidad de la muestra aleatoriamente y las n - 1 unidades restantes quedan automáticamente seleccionadas.

Este esquema de muestreo es muy utilizado en la práctica, debido a la facilidad de aplicación pero es necesario tener conocimiento sobre la población a la que se le va a aplicar, por eso se pide que la población se encuentre aleatorizada con respecto a la característica que vamos a estudiar en ella. Esto no significa que el muestreo sistemático tenga esta restricción, ya que si la población se encuentra ordenada o existe alguna tendencia, se hace un análisis y se utilizan los estimadores convenientes a la situación.

En el caso en que la población se encuentre aleatorizada con respecto a la característica en estudio, los estimadores que se proponen son los siguientes:

Un estimador insesgado de la media poblacional es la media sistemática:

$$\bar{y}_{sist} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Estimador de la varianza de la media sistemática:

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

Estimador del total poblacional:

$$\hat{Y}_{sist} = N \bar{y}_{sist}$$

Estimador de la varianza de $\hat{Y}_{sist}$ es

$$\hat{V}(\hat{Y}_{sist}) = N^2 \hat{V}(\bar{y}_{sist})$$

Estimador de una proporción poblacional

$$\hat{P}_{s,sT} = p_{s,sT} = \frac{a}{n}$$

donde a es el número de unidades en al muestra que pertenecen a la clase de interés.

Estimador de la varianza de $\hat{P}_{sist}$

$$\hat{V}(\hat{P}_{s,sT}) = \frac{p_{s,sT}\, q_{s,sT}}{n-1}(1-f)$$

donde

$$q_{s,sT} = 1 - p_{s,sT}.$$

El muestreo sistemático se usa principalmente en archivos de tarjetas, expedientes, hojas para elegir una muestra en algún proceso de producción, etc.

Ejemplo: Una empresa quiere investigar el número medio de días al año que sus empleados faltan por motivos de enfermedad y que proporción de sus empleados son mayores de 50 años. La empresa cuenta con 150 empleados y los expedientes de ellos se encuentran ordenados alfabéticamente en un archivo. Se decide elegir una muestra sistemática de 1 de cada 15 empleados. Para ello se elige en número aleatorio entre 1 y 15 y resulta ser 10, de manera que los empleados que forman la muestra son los empleados que les corresponden los siguientes números 10, 25, 40, 55, 70, 85, 100, 115, 130, 145.

Empleado	Nº días ausente por enfermedad (y_i)	Edad	y_i^2
1	8	52	64
2	5	45	25
3	0	25	0
4	10	56	100
5	3	21	9
6	6	40	36
7	2	28	4
8	1	19	1
9	4	35	16
10	7	58	49
TOTALES	46		304

$\bar{y}_{sist} = \frac{46}{10} = 4.6$ días al año por empleado

$\hat{V}(\bar{y}_{sist}) = \frac{(1 - \frac{10}{150})}{(10)9} \left(304 - \frac{(46)^2}{10}\right) = 0.924$

$\sigma(\bar{y}_{sist}) = 0.96$ días

El número total de días que han faltado los empleados durante el año:

$\hat{y}_{sist} = 150\,(4.6) = 690$ días

$$\hat{V}(\bar{y}_{s.s.r}) = 150^2 (0.924) = 20790$$

$$\sigma(\bar{y}_{s.s.r}) = 144.19 \text{ días}$$

La proporción de empleados mayores de 50 años

$$p_{s.s.r} = \frac{3}{10} = 0.33$$

Es decir el 33% de los empleados tienen más de 50 años.

$$\hat{V}(p_{s.s.r}) = \frac{0.33\,(0.67)}{9}\left(1 - \frac{10}{150}\right)$$

$$\hat{V}(p_{s.s.r}) = 0.023$$

$$\sigma(p_{s.s.r}) = 0.15$$

centro de educación continua
división de estudios superiores
facultad de ingeniería, unam

TECNICAS DE MUESTREO ESTADISTICO

SUB-MUESTREO

ING. LUIS ALEJANDRO SERVIN
OCTUBRE, 1977

Palacio de Minería Calle de Tacuba 5, primer piso. México 1, D. F.

Wages and Industrial Relations

Chapter 18. Occupational Pay and Supplementary Benefits

Background

The Bureau of Labor Statistics, for many decades, has conducted studies of wages by occupation and industry, based upon employer records. The Bureau's first such study, growing out of a study by the U.S. Senate in 1891, resulted in a wage rate record extending back continuously to 1860. Systematic collection of wage data by occupation and industry has continued since the turn of the century; changes in coverage have been dictated mainly by government requirements. A large survey program undertaken for the War Industries Board in 1919 produced occupational pay rates by industry and State, and (for some industries) by city. Between 1934 and 1940, the selection of industries studied was determined largely by administrative needs under the National Recovery Act, Public Contracts Act, and the Fair Labor Standards Act, with emphasis on nationwide data for relatively low-wage industries.

Survey activity shifted in the 1940–41 defense period to heavy industries essential to war production. Implementation of wage stabilization policy during the war required a large-scale program of occupational wage studies by industry and locality. The emphasis on data by locality has continued since 1945 within the framework of industry studies generally designed to yield national and regional estimates. In addition, the Bureau developed three new types of surveys.

Area wage surveys, initiated in the late 1940's, were designed to meet the growing demand for pay data related to office clerical and manual jobs that are common to a wide variety of manufacturing and nonmanufacturing industries within metropolitan areas. This survey program was firmly established and temporarily expanded for use in the wage stabilization effort during the Korean emergency. The need for nationwide estimates of white-collar pay in private industry for use in appraising the Federal white-collar salary structure resulted in a survey design that would produce national averages, based on an area sample. Data for individual areas studied also serve the wage administration needs for other government agencies.

Prior to 1960, studies in a very few professions provided salary data. Beginning in that year, salary surveys have been made on a nationwide basis covering professional, administrative, technical, and clerical jobs in a broad spectrum of industries. Averages for these jobs are used by the administrative agencies directly concerned with Federal pay matters.

Recognizing the increased interest in governmental pay, and the fact that government employees represented a large and growing segment of the total work force, the Bureau began a series of wage and benefit studies in eight city governments during 1970. The series was later expanded to include all cities having 500,000 inhabitants or more; this group included 26 cities in 1975.

Description of Surveys

Although differing in industrial, geographic, and occupational coverage, the four types of surveys described form an integrated program of occupational wage surveys based upon a common set of administrative forms, manual of procedures, and common concepts and definitions. Employer cooperation in surveys is on a voluntary basis. Confidential individual establishment data compiled by the Bureau's field economists are grouped in published reports in a manner that will avoid possible disclosure of an establishment's rates. Establishments included in all surveys are classified by industry as defined in the 1967 edition of the *Standard Industrial Classification Manual* prepared by the U.S. Office of Management and Budget.[1] Survey reports identify the minimum size of establishment (measured by total employment) studied. Definitions for *Standard Metropolitan Statistical Areas* are employed in all programs.[2]

Industry wage surveys provide data for occupations selected to represent the full range of activities performed by workers. Consideration also is given, in their selection, to the prevalence in the industry, definiteness and clarity of duties, and importance as reference points in collective bargaining.

In addition to collecting straight-time first-shift rates (or hours and earnings for incentive workers) for individual workers in the selected occupations, surveys in most industries also establish the wage frequency distribution for broad employment groups, i.e., pro-

[1] See app. B.
[2] See app. C.

Reprint from
Handbook of Methods
Bulletin 1910

U.S. DEPARTMENT OF LABOR
Bureau of Labor Statistics

duction and related workers or nonsupervisory workers. Weekly work schedules; shift operations and differentials, paid holiday and vacation practices; and health, insurance, and retirement benefits are included in the information collected, along with the provisions made for other items, applicable to certain industries. The studies also provide estimates of labor-management agreement coverage, proportions employed under incentive pay plans, and the extent to which establishments provide a single rate or range of rates for individual job categories.

Fifty manufacturing and 20 nonmanufacturing industries, accounting for about 22.5 million employees, are surveyed on a regularly recurring basis. A majority are studied on a 5-year cycle, but a number of comparatively low-wage industries are on a 3-year cycle. In addition, special wage surveys also are undertaken at the request of others.

Nearly all of the manufacturing, utilities, and mining industries are studied on a nationwide basis and estimates are provided also for regions and major areas of concentration. Surveys in trade, finance, and service industries usually are limited to a number of metropolitan areas. Nationwide surveys generally develop separate estimates by size of establishment, size of community, labor-management agreement coverage, and type of product or plant group.

Area wage surveys provide data for occupations common to a wide variety of industries in the areas surveyed. The 76 occupational categories studied include 29 office clerical; 17 electronic data processing, drafting, and industrial nurses, and 30 maintenance, toolroom, powerplant, and custodial and material movement jobs. Thus, they provide representation of the range of duties and responsibilities associated with white-collar, skilled maintenance trades, and other "indirect" manual jobs. Weekly salaries reported for individuals in white-collar jobs relate to regular straight-time salaries that are paid for standard workweeks. Average hourly earnings for maintenance and other manual jobs relate to first-shift hourly rates.

Industry divisions included are (1) manufacturing; (2) transportation, communication, and other public utilities; (3) wholesale trade, (4) retail trade; (5) finance, insurance, and real estate, and (6) selected service industries. Establishments employing fewer than 50 workers are excluded—with a minimum of 100 applying to manufacturing, transportation, communication,and other public utilities; and to retail trade in the 13 largest communities.

In addition to the all-industry averages and distributions of workers by earnings classes, separate data are provided for manufacturing and nonmanufacturing in each area and, wherever possible, for individual industry divisions in the nonmanufacturing sector. Among the 70 Standard Metropolitan Statistical Areas in this annual survey program as of 1976, separate data are provided for transportation, communication, and other public utilities in 68 areas; for retail trade in 32 areas; for wholesale trade and finance, insurance, and real estate in 18 areas; and for the selected service industries in 20 large areas. In 31 of the larger areas, wage data are presented separately for establishments that have 500 workers or more.

Data on weekly work schedules; paid holiday and vacation practices; and health, insurance, and retirement benefits are recorded separately for nonsupervisory office workers and plant workers (nonoffice). Shift operations and differentials are collected for plant workers in manufacturing. Data on minimum entrance rates for inexperienced office workers are collected in all industries. These items are studied every 3 years in all areas. This survey program also has developed information on profit-sharing plans, characteristics of sick leave plans, wage payment systems, and other items related to employee compensation.

Special area wage surveys have been conducted annually since 1967 at the request of the Employment Standards Administration for use in administering the Service Contract Act of 1965. Cross-industry surveys provide information on hourly earnings for 14 office occupations, 10 professional and technical jobs, and 20 maintenance, toolroom, powerplant, and custodial and material movement jobs. The industrial scope includes manufacturing; transportation, communication, and other public utilities; wholesale trade, retail trade; finance, insurance, and real estate, and selected service industries. Establishments with fewer than 50 employees are excluded from the scope of these special area wage surveys.

In addition to the cross-industry surveys, special industry studies are conducted for the Employment Standards Administration. These studies provide information on hourly earnings for 10 moving and storage jobs; 6 refuse hauling jobs; 24 contract construction jobs; 7 laundry jobs; and 6 food service jobs. For both the cross-industry surveys and special industry studies, data on incidence of paid holidays and vacation practices, and health, insurance, and retirement benefits are provided every 3 years.

The National Survey of Professional, Administrative, Technical, and Clerical Pay provides a fund of broadly based information on salary levels and distributions in private employment. The 72 occupation-work levels studied in 1975 were selected from the following fields: Accounting, legal services, personnel management, engineering and chemistry, buying, clerical supervisory, drafting, and clerical. Definitions for these occupations provide for classification of employees according to appropriate work levels (or classes). Although reflecting duties and responsibilities in industry, the definitions were designed to be translatable to specific pay grades in the General Schedule applying to Federal Classification Act employees. This survey, thus, provides information in a form suitable for use in comparing the compensation of salaried employees in

the Federal civil service with pay in private industry.

Monthly and annual average salaries are reported for all occupations. Data relate to the standard salaries that were paid for standard work schedules i.e., to the straight-time salary corresponding to the employee's normal work schedule, excluding overtime hours. Nationwide salary distributions and averages are presented for men and women combined. Averages also are presented for establishments in metropolitan areas combined and for establishments employing 2,500 workers or more.

Industry divisions included are: (1) manufacturing, (2) transportation, communication, electric, gas and sanitary services, (3) wholesale trade, (4) retail trade, (5) finance, insurance, and real estate, and (6) engineering and architectural services, and commercially operated research, development, and testing laboratories.

Limited to the Nation's metropolitan areas for the years 1960 through 1964, the annual survey was expanded in 1965 to include nonmetropolitan counties. The minimum establishment size included in the survey is 250 workers in manufacturing and retail trade and 100 in the other industries studied. The minimum establishment size has been adjusted at various times since 1961. Since the survey scope is subject to change, users are directed to the Scope and Method of Survey appendix in the bulletins for a description of current practice.[3]

Municipal government wage surveys provide data for occupations common to many municipal governments. The 50 occupations studied include 10 office clerical; 5 data processing; 13 maintenance, custodial, and trades and labor; 6 public safety and correction; 2 sanitation; and 14 professional, administrative, and technical jobs. To facilitate comparisons, the surveys are designed to be as comparable as possible to the Bureau's area wage surveys of private industry and to other related studies. Average salaries relate to base salaries for a standard workweek, plus longevity pay, reported on a monthly basis. In addition to wage data, comprehensive information is provided on city pay plans and their administration, work practices, unionization, and health, insurance, and retirement benefits of municipal employees. To assist in making inter-city comparisons and comparisons with private industry and unions, the principal features of the benefit plans are described in standard formats. These formats are almost identical to those used in the Bureau's *Digest of Health and Insurance Plans*, 1974 Edition, and the *Digest of Selected Pension Plans, 1973 Edition.*

Concepts. The Bureau's occupational wage surveys summarize a highly specific wage measure—the rate of pay, excluding premium pay for overtime and for work on weekends, holidays, and late shifts, for individual workers. In the case of workers paid under piecework or other types of production incentive pay plans, an earned rate is computed by dividing straight-time earnings for a time period by corresponding hours worked. Production bonuses, commissions, and cost-of-living bonuses are counted as earnings. In general, bonuses that depend on factors other than the output of the individual worker or group of workers are excluded; examples of such nonproduction payments are safety, attendance, year-end or Christmas bonuses, and cash distributions under profit-sharing plans.

Unless stated otherwise, rates do not include tips or allowances for the value of meals, room, uniform, etc. The earnings figures, thus, represent cash wages (prior to deductions for social security, taxes, savings bonds, premium payments for group insurance, meals, room or uniforms) after the exclusion of premium pay for overtime, weekend, holiday, or late shift work.

Hours shown for salaried occupations relate to standard weekly hours for which the employee receives his regular straight-time salary.

Occupational classifications are defined in advance of the survey. Because of the emphasis on interestablishment and interarea comparability of occupational content, the Bureau's job descriptions may differ significantly from those in use in individual establishments or those prepared for other purposes. The job descriptions used for wage survey purposes are typically brief and usually more generalized than those used for other purposes. The primary objective of the descriptions is to identify the essential elements of skill, difficulty, and responsibility that establish the basic concept of the job.[4]

Although work arrangements in any one establishment may not correspond precisely to those described, those workers meeting the basic requirements established for the job are included.[5]

In applying these job descriptions, the Bureau's field representatives exclude working supervisors, apprentices, learners, beginners, trainees, handicapped workers, part-time or temporary workers, probationary workers unless provision for their inclusion is specifically stated in the job description.

[3] The terms "in scope" or "within scope" are used throughout this chapter to refer to the coverage of the particular survey being described.

[4] An example of a job description:

MACHINIST, MAINTENANCE

Produces replacement parts and new parts in making repairs of metal parts of mechanical equipment operated in an establishment. Work involves *most of the following*: interpreting written instructions and specifications, planning and laying out of work, using a variety of machinist's handtools and precision measuring instruments, setting up and operating standard machine tools, shaping of metal parts to close tolerances, making standard shop computations relating to dimensions of work, tooling, feeds, and speeds of machining, knowledge of the working properties of the common metals, selecting standard materials, parts, and equipment required for this work, and fitting and assembling parts into mechanical equipment. In general, the machinist's work normally requires a rounded training in machine-shop practice usually acquired through a formal apprenticeship or equivalent training and experience.

Paid holidays, paid vacations, and health, insurance, and retirement plans are treated statistically on the basis that these are applicable to all nonsupervisory plant or office workers if a majority of such workers are eligible or can expect eventually to qualify for the practices listed. Data for health, insurance, and retirement plans are limited to those plans for which at least a part of the cost is borne by the employer. This limitation does not apply, however, to data for health, insurance, and retirement plans reported in the municipal government surveys. Informal provisions are excluded.

Survey Methods

Planning. Consultations are held with appropriate management, labor, and Government representatives to obtain views and recommendations related to scope, timing, selection, and definitions of survey items, and types of tabulations. Particularly in planning surveys in specific industries, these discussions importantly supplement comments and suggestions received from the regional offices at the conclusion of the previous study. Reflecting its use in evaluation of Federal white-collar pay, the design of the National Survey of Professional, Administrative, Technical, and Clerical Pay was developed in conjunction with the Office of Management and Budget and the Civil Service Commission. Changes in the survey scope, item coverage, and job definitions are initiated by these agencies.

The industrial scope of each survey is identified in terms of the classification system provided in the *Standard Industrial Classification Manual*. The scope may range from part of a 4-digit code for an industry study to a uniform combination of broad industry divisions and specific industries for the area wage surveys or the salary survey of professional, administrative, technical, and clerical jobs. The needs of major users are a prime consideration in designing the multipurpose occupational studies.

The minimum size of establishment included in a survey is set at a point where the possible contribution of the excluded establishments is regarded as negligible for most of the occupations surveyed. Another practical reason for the adoption of size limitations is the difficulty encountered in classifying workers in small establishments where they do not perform the specialized duties indicated in the job definitions.

[5] In general, workers are included in a classification if the duties as described are performed a major part of the time and the remainder is spent on related duties requiring similar or lesser skill and responsibility. However, in some jobs, particularly office and skilled production-worker categories, workers may regularly perform a combination of duties involving more than one occupation. Unless indicated otherwise in the description, in these situations consideration for classification purposes is given to those elements of the job which are most important in determining its level for pay purposes. Thus, a worker meets the basic concept of the stenographer classification if taking of dictation is a regular requirement of the job even though a majority of time is spent on routine typing.

Considerations in timing of industry surveys include date of expiration of major labor-management agreements, deferred wage adjustments, seasonality of production (e.g., garments), and interests of users. Wherever possible, area wage surveys are timed to follow major wage settlements as well as to meet the needs of government agencies engaged in wage administration as required by law.

The types of occupations studied and criteria used in their selection were identified in the description of the various types of surveys. The job list for each survey is selected to represent a reasonably complete range of rates in the wage structure for the employment categories involved, i.e., production and related workers in a specific manufacturing industry or nonsupervisory office, maintenance, material handling, and custodial workers in a metropolitan area. The established hierarchy of job rates to be found within establishments and industries permits the use of pay data for such key or benchmark jobs for interpolating rates for other jobs. Technological developments or user interests may dictate changes in the job lists and definitions. New definitions for jobs usually are pretested in a variety of establishments prior to their use in a full-scale survey.

Questionnaires. Two basic schedules are used in obtaining data in all surveys. The first (BLS 2751A) includes items relating to products or services, employment, shift operations and differentials, work schedule, overtime premiums, paid holidays and vacations, insurance and retirement plans, union contract coverage, and other items applicable to the establishment. The second (BLS 2753G) is used in recording occupation, sex, method of wage payment, hours (where needed), and pay rate or earnings for each worker studied. Supplementary forms are used to meet particular needs.

Collection. Bureau field economists collect data by personal visit to each of the sample establishments. Job functions and factors in the establishment are carefully compared with those included in the Bureau job definitions. The job matching may involve review of records such as pay structure plans and organizational charts, company position descriptions, interviews with appropriate officials, and, on occasion, observation of jobs within plants. A satisfactory completion of job matching permits acceptance of company-prepared reports where this procedure is preferred by the respondent. Generally, however, the field economist secures wage or salary rates (or hours and earnings, when needed) from payroll or other records and data on the selected employer practices and supplementary benefits from company officials, company booklets, and labor-management agreements.

Area wage surveys in all areas involve personal visits every third year with partial collection by mail or telephone in the intervening years. Establishments par-

BLS 2751A
(Rev. June 1973)

U.S. DEPARTMENT OF LABOR
Bureau of Labor Statistics

Budget Bureau No. 44-R0338
Approval expires November 30, 1972

WAGE SURVEY
GENERAL ESTABLISHMENT INFORMATION

Your report will be held in confidence

1. ESTABLISHMENT IDENTIFICATION

A Survey	Payroll Period 19__	Payroll Period 19__
ESTABLISHMENT NAME		
STREET ADDRESS	CITY	
COUNTY, STATE AND ZIP CODE	AREA CODE - TELEPHONE	
NAME AND TITLE OF AUTHORIZING OFFICIAL 19__	NAME AND TITLE OF OFFICIAL SUPPLYING DATA	
19__		
ADDRESS OF OFFICE FROM WHICH DATA WAS OBTAINED IF DIFFERENT FROM ABOVE		
B Central Office (Complete if clearance ☐ and/or data ☐ obtained from this source)		TELEPHONE
NAME OF COMPANY		NAME OF AUTHORIZING OFFICIAL
STREET ADDRESS	CITY, STATE AND ZIP CODE	
		TITLE

2. CURRENT PRODUCTS OR SERVICES AND PROCESSES

A. PRODUCT OR SERVICE	APPROXIMATE % ANNUAL VALUE	PRODUCT OR SERVICE	APPROXIMATE % ANNUAL VALUE	B. SCOPE OF OPERATIONS
19__		19__		

3. OFFICE USE ONLY

YEAR	SCHEDULE NO	IDENT	AREA	REGION	STATE	CITY SIZE	BIC CODE	EST. SIZE	UNION	WEIGHT	SPECIAL CHARACTERISTICS (1)	SPECIAL CHARACTERISTICS (2)
	1-5	6-8	9-11	12	13-14	15	16-19	20	21	22-24	25-27	28-30
19__												
19__												

OCCUPATIONAL PAY AND SUPPLEMENTARY BENEFITS

BLS 2751A—Continued
(Rev. June 1973)

U.S. DEPARTMENT OF LABOR
Bureau of Labor Statistics

SURVEY	PAYROLL PERIOD	SCHEDULE NO.
ESTABLISHMENT		

5. UNION CONTRACT COVERAGE

		No	Yes	
A. Are a majority of your production workers covered by union agreements?	31	0	1	
B. Are a majority of your office workers covered by union agreements?	32	0	1	

C. With what unions does this establishment have contracts? *(Give name and affiliation below.)*	D. What occupational groups are covered by the contract? *(List groups below opposite the appropriate union.)*
Production Workers:	
Office Workers:	

6. ESTABLISHMENT EMPLOYMENT (APPROXIMATE)

A. What is the approximate total employment* in this establishment?	33-38	
B. How many are nonsupervisory production (plant) workers?	39-43	
Men	44-48	
Women	49-53	
C. Nonsupervisory office workers?	54-58	
Men	59-63	
Women	64-68	
D. Other employees (executive, professional, supervisory, etc.)?	69-73	
E.	74-78	
F.		
	End of card—⇒80	1

*Includes salaried officers of corporations but does not include proprietors, members of unincorporated firms, pensioners, members of the armed forces carried on the payroll, or unpaid family workers.

7. Remarks

BLS HANDBOOK OF METHODS

BLS 275 30
Rev 65

OCCUPATIONAL RATES

Payroll period_____

Schedule No Est Name Page

OCCUPATION AND GRADE	Occupational code (1)	Sex (2)	Method of pay (3)	Number of workers (4)	Hours (5)	Salary, rate, or earnings (6)	Line No (7)	(8)
							1	
							2	
							3	
							4	
							5	
							6	
							7	
							8	
							9	
							10	
							11	
							12	
							13	
							14	
							15	
							16	
							17	
							18	
							19	
							20	
							21	
							22	
							23	
							24	
							25	
							26	
							27	
							28	

ticipating in the mail collection receive a transcript of the job matching and wage data obtained previously, together with the job definitions. The up-dated returns are scrutinized and questionable entries are checked with the respondent. Personal visits are made to establishments not responding to the mail or telephone request and to those reporting unusual changes from year-earlier data.

The work of all field economists is checked for quality of reporting, with particular attention directed to accuracy in job matching. The revisits are made by supervisory and senior economists. Systematic technical audits of the validity of survey definitions, made by staff with specialized training, also are maintained for the technically complex nationwide white-collar salary survey.

Sampling

Before the sample is selected, a suitable sampling "frame" must be located or developed. A sampling frame is a list of establishments which fall within the designated scope of the survey. The frame is as close to a universe as possible but is often incomplete. BLS uses frames primarily compiled from lists provided by regulatory governmental agencies (primarily State unemployment insurance agencies). Because these are sometimes incomplete, they are supplemented by data from trade directories, trade associations, labor unions, and other sources.

The survey design employs a high degree of stratification. Each geographic-industry unit for which a separate analysis is to be presented is sampled independently. Within these broad groupings, a finer stratification by product (or other pertinent attributes) and size of establishment is made. Stratification may be carried still further in certain industries: Textile mills, for instance, are classified on the basis of integration, i.e., whether they spin only, weave only, or do both. Such stratification is highly important if the occupational structure of the various industry segments differs widely.

The sample for each industry-area group is a probability sample, each establishment having a predetermined chance of selection. However, in order to secure maximum accuracy at a fixed level of cost (or a fixed level of accuracy at minimum cost), the sampling fraction used in the various strata ranges downward from all large establishments through progressively declining proportions of the establishments in each smaller size group. This procedure follows the principles of optimum allocation where the standard deviation of the characteristic being estimated is proportional to the average employment in the stratum. Thus, each sampled stratum will be represented in the sample by a number of establishments roughly proportionate to its share of the total employment. Though this procedure may appear at first to yield a sample biased by the over-representation of large firms, the method of estimation employed yields unbiased estimates by the assignment of proper weights to the sampled establishments.

In the event a sample establishment within scope is uncooperative in supplying usable data, a substitute is assigned in the same industry-location-size class. (Since no close relation exists between failure to participate in these surveys and the items being studied, little bias is introduced by this procedure.)

The size of the sample in a particular survey depends on the size of the universe, the diversity of occupations, and their distribution, the relative dispersion of earnings among establishments, the distribution of the establishments by size, and the degree of accuracy required. Estimates of variance based on data from previous surveys are used in determining the size of the sample needed.

As indicated earlier, area wage surveys are limited to selected metropolitan areas. These areas, however, form a sample of all such areas, and, when properly combined (weighted), yield estimates of the national and regional levels. The sample of areas is based on the selection of one area from a stratum of similar areas. The criteria of stratification are region, type of industrial activity as measured by percent of manufacturing employment, and major industries. Each area is selected with its probability of selection proportionate to its nonagricultural employment. The largest metropolitan areas are self-representing, i.e., each one forms a stratum by itself and is certain of inclusion in the area sample. The present area sample contained about 70 percent of all nonagricultural employment of the metropolitan area complex of the entire country in 1973.

Estimating Procedures

Estimated average earnings (hourly, weekly, monthly, or annual) for an industry or an occupation are computed as the arithmetic mean of the individual employee's earnings. They are not estimated by dividing total payrolls by the total time worked, since such information almost never is available on an occupational basis.

All estimates are derived from the sample data. The averages for occupations, as well as for industries, are weighted averages of individual earnings and not computed on an establishment basis. The proportion of employees affected by any fringe provision likewise is estimated from the sample; all plant and office workers in each establishment are considered to be covered by the predominant benefit policy in effect, and the entire plant and office employment of the establishment is separately classified accordingly.

As mentioned previously, the use of a variable sampling ratio in different strata of the population would result in biased estimates if straight addition of the data for the various establishments were made. Therefore, each establishment is assigned a weight that is the inverse of the sampling rate for the stratum from which it was selected—e.g., if a third of the establishments in one stratum are selected, each of the sampled establishments is given a weight of 3.

To illustrate the use of weights, suppose the universe were 7 establishments, from which a sample of 3 was selected. Assume that establishment A was drawn from a cell, or stratum, in which one of the two establishments was used in the sample. It therefore is given a weight of 2. Establishment B, on the other hand, was taken with certainty (or a probability of 1) and is thus given a weight of 1. Establishment C was taken from the remaining group where one of the four establishments was used in the sample, and hence is given a weight of 4. The following calculations are made in estimating average earnings for a given occupation.

		Workers in occupation in sample establishments at specified rate			
Establishment	Weight	Total number	Average hourly earnings	Estimates of total in stratum	
				Workers	earnings
A	2	40	$2 60	2×40	2×40×$2 60
B	1	30	2 70	1×30	1×30× 2 70
		20	2 95	1×20	1×20× 2 95
C	4	10	2 65	4×10	4×10× 2 65
Estimated universe				170	$454 00

The estimated average hourly earning is thus $\frac{\$454\ 00}{170}$ or $2.67.

A similar method applies to any characteristic estimated from the sample. To estimate the proportion of employees in establishments granting paid vacations of 2 weeks after 2 years of service, for instance, the establishments are classified according to the length of vacation granted after 2 years' service, establishment weights are applied to employment, as in the previous example, and the proportion of the estimated employment in the 2-week category of the estimated total employment then is computed. Using the same three establishments as in the previous example, this can be illustrated as follows:

Establishment	Weight	Actual total establishment employment	Weighted employment	Vacation provisions after 2 years
A	2	100	200	1 week.
B	1	500	500	2 weeks
C	4	75	300	1 week
Estimated universe			1,000	

Thus, the estimated percentage of workers in establishments granting 2 weeks' vacation after 2 years of service is $\frac{500}{1,000}$ or 50 percent.

When a large establishment within survey scope, for which no substitute exists, is unable to supply data, the deficiency is alleviated by increasing the weight of the most nearly similar units. Should any segment be affected by a substantial amount of such noncooperation, the publication of materials will be diminished by omitting separate presentation of sectors seriously affected.

Where a sample of selected metropolitan areas is used to represent the totality of such areas, a second stage of weighting is used to expand the individual area totals to region and/or national estimates. Since, as indicated in the description of the sampling method, each area represents a stratum of similar areas, the total from each area is weighted to the estimated stratum totals by multiplying by the inverse of the chance of selection. This procedure provides the ratio of nonagricultural employment in the stratum to that in the sample area (one in the case of the large self-representing areas). Summing all such estimated stratum totals yields the earnings and employment totals for the region and the country as a whole.

Analysis and Presentation

Where an industry survey is designed to yield estimates for selected States or areas, these are published separately as information becomes available from all sample firms in the State or area unit. Industry surveys limited to selected areas do not provide a basis for the examinations of pay levels by size of community, size of establishment, product, or labor-management agreement coverage that generally are included in bulletin reports on nationwide surveys. Regardless of geographic scope, industry survey reports record the incidence of incentive pay plans and, to the extent possible, average pay levels separately for time and incentive workers.

Individual bulletin reports on individual area wage surveys are supplemented by two summary bulletins. The first compiles the results of individual area surveys made during a year. The second contains information on occupational earnings, employer practices, and supplementary wage benefits for all metropolitan areas combined and by industry division within the four broad census regions.

Percent increases, adjusted for changes in employment, are computed for broad occupational groups, e.g., office clerical, electronic data processing, skilled maintenance, and unskilled plant. These increases are computed annually, separately for all industries, manufacturing, and nonmanufacturing, for each metropolitan area studied, for all metropolitan areas combined, and for four broad census regions. Area pay relatives for the four occupational categories are published annually, permitting ready comparisons of average pay levels among areas. Estimates of labor-management

agreement coverage are also presented annually. Occupational pay relationships within individual establishments are summarized periodically.

Bulletins on the National Survey of Professional, Administrative, Technical, and Clerical Pay present occupational averages and distributions on an all-industry basis, nationwide and separately for all metropolitan areas combined, and for establishments employing 2,500 workers or more. Average pay levels for industry divisions are shown as percentages of the all-industry averages. Year-to-year percent changes for occupation-work levels and trend estimates for occupations are reported.

Industry wage, area wage, and municipal government wage survey reports are issued throughout the year as the surveys are completed. The bulletin on the National Survey of Professional, Administrative, Technical, and Clerical Pay is available in December.

Summaries of the data in the bulletins and special analyses appear also in the *Monthly Labor Review*.

Uses and Limitations

Occupational wage data developed in these surveys have a variety of uses. They are used by Federal, State, and local agencies in wage and salary administration and in the formulation of public policy on wages, as in minimum wage legislation. They are of value to Federal and State mediation and conciliation services and to State unemployment compensation agencies in judging the suitability of job offers. Knowledge of levels and trends of pay rates by occupation, industry, locality, and region is required in the analysis of current economic developments and in studies relating to wage dispersion and differentials.

Bureau data are used in connection with private wage or salary determinations by employers or through the collective bargaining process. To the extent that wages are a factor, survey data also are considered by employers in the selection of location for new facilities and in cost estimating related to contract work.

Occupational wage survey programs are not designed to supply mechanical answers to questions of pay policy. As suggested earlier, limitations are imposed in the selection and definition of industries, of geographic units for which estimates are developed, of occupations and associated items studied, and in determination of periodicity and timing of particular surveys. Depending upon his needs, the user may find it necessary to interpolate for occupations or areas missing from the survey on the basis of knowledge of pay relationships.

Because of interestablishment variation in the proportion of workers in the jobs studied and in the general level of pay, the survey averages do not necessarily reflect either the absolute or relative relationships found in the majority of establishments. To illustrate, employment in the specialized maintenance crafts tends to be concentrated in the larger establishments, whereas employment is custodial and material movement jobs is distributed more widely within an industry or area. Thus, to the extent that pay rates in the larger establishments vary from the average level, the skill differential measure based on the survey averages will differ to some degree from that obtainable within each of the larger establishments.

The incidence of incentive methods of payment may vary greatly among the occupations and establishments studied. Since hourly averages for incentive workers generally exceed those for hourly-rated workers in the same job, averages for some incentive-paid jobs may equal or exceed averages for jobs positioned higher on a job evaluation basis but normally paid on a time basis. Wherever possible, data are shown separately for time workers and incentive workers in the industry surveys. Incentive plans (generally plant-wide in application) apply to only a very small proportion of the workers in the indirect plant jobs studied in the area wage program.

Although year-to-year changes in averages for a job or job group primarily reflect general wage and salary changes or merit increases received by individuals, these averages also may be affected by changes in the labor force resulting from labor turnover, labor force expansions and reductions for other reasons, as well as changes in the proportion of workers employed in establishments with different pay levels. A labor force expansion might increase the proportion of lower paid workers and thereby lower the average, or the closing of a relatively high-paying establishment could cause average earnings in the area to drop.

This problem has been overcome for area wage surveys by holding establishment employments constant while computing percent increases in earnings. That is, the previous and current year earnings of each establishment are weighted by that establishment's previous year's employment. An establishment which does not have workers or has not been sampled in the previous year is not included in the calculation.

Reliability of surveys. Results of the surveys generally will be subject to sampling error. This error will not be uniform, since, for most occupations, the dispersion of earnings among establishments and frequency of occurrence of the occupation differ. In general, the sample is designed so that the chances are 9 out of 10 that the published average does not differ by more than 5 percent from the average that would be obtained by enumeration of all establishments in the universe.

The sampling error of the percentage of workers receiving any given supplementary benefit differs with the size of the percentage. However, the error is such that rankings of predominant practices almost always will appear in their true position. Small percentages may be subject to considerable error, but will always remain in the same scale of magnitude. For instance,

the proportion of employees in establishments providing more than 5 weeks' paid vacation to long-service employees may be given as 2 percent, when the true percentage for *all* establishments might be only 1 percent. Such a sampling error, while considerable, does not affect the essential inference that the practice is a rare one.

Estimates of the number of workers in a given occupation are subject to considerable sampling error, due to the wide variation among establishments in the proportion of workers found in individual occupations. (It is not unusual to find these estimates subject to sampling error of as much as 20 percent.) Hence, the estimated number of workers can be interpreted only as a rough measure of the relative importance of various occupations. The greatest degree of accuracy in these employment counts is for those occupations found principally in large establishments. This sampling error, however, does not materially affect the accuracy of the average earnings shown for the occupations. The estimate of average earnings is technically known as a "ratio estimate," i.e., it is the ratio of total earnings (*not* payrolls) to total employment in the occupation. Since these two variables are highly correlated (i.e., the errors tend to be in the same direction), the sampling error of the estimate (average hourly earnings) is considerably smaller than the sampling error of either total earnings or total employment.

Since completely current and accurate information regarding establishment products and the creation of new establishments is not available, the universe from which the sample is drawn may be incomplete. Sample firms incorrectly classified are accounted for in the actual field work, and the universe estimates are revised accordingly. Those firms which should have been included but were classified erroneously in other industries cannot be accounted for.

Since some measure of subjective judgment enters into the classification of occupations and other characteristics, there is some reporting variability in the results. A repetition of the survey in any establishment with different interviewers and respondents would undoubtedly produce slightly different results. However, when spread over a large number of establishments the differences, being random, would tend to balance out. Hence, analyses based on a small number of respondents must be used with care, even when all eligible establishments are included. No evidence of any consistent error has been uncovered.

Technical References

Number

1. Cohen, Samuel E., "Studies of Occupational Wages and Supplementary Benefits," *Monthly Labor Review*, March 1954 (pp. 292–297)
   An earlier description of the methods of wage surveys, similar to the present article
2. Douty, H. M., "Survey Methods and Wage Comparisons," *Labor Law Journal*, April 1964 (pp. 222–230).
   A discussion of the uses of wage survey results, and the pitfalls to be avoided. A short discussion of the factors affecting survey methods is also included.
3. Houff, James N., "Improving Area Wage Survey Indexes," *Monthly Labor Review*, January 1973 (pp. 52–57).
4. Kanninen, Toivo P., "New Dimensions in BLS Wage Survey Work." *Monthly Labor Review*, October 1959 (pp. 1081–1084)
   An outline of the occupational wage survey programs, as expanded in fiscal 1960. Lists the type of survey and cycle for each of 70 industries studied separately, and identifies the area sample as originally determined for the labor market survey program.
5. Talbot, Deborah B., "Improved Area Wage Survey Indexes," *Monthly Labor Review*, May 1975 (pp. 30–34).
   A discussion of differences in computing Area Wage Survey pay increases by the matched and unmatched sample techniques
6. Ward, Virginia L., "Area Sample Changes in the Area Wage Survey Program." *Monthly Labor Review*, May 1975 (pp. 49–50)
   A description of the Area Wage Survey program and changes in the program's area sample