Capitulo 10

Variable aleatoria.

Distribuciones de probabilidad

Discreta y continua.

"*La teoría de la probabilidad como disciplina matemática puede y debe ser desarrollada a partir de axiomas, de la misma manera que la geometría o el álgebra*".
A. N. Kolmogorov[1]

10.1 Variable aleatoria

Un experimento aleatorio es el proceso mediante el cual se obtiene una observación. Si bien una observación no siempre es numérica, estas observaciones pueden recodificarse por números; por ejemplo, el resultado de un nacimiento, si asignamos al resultado femenino con "0" y masculino "1". Así, el interés se centra en experimentos que producen resultados numéricos. Generalmente asignamos una letra a la variable medida en un experimento, por ejemplo x.

Una variable es un símbolo que actúa en las funciones, fórmulas, algoritmos y proposiciones de la matemática y la estadística. La relación que se asigna a los resultados de un experimento aleatorio se le llama *variable aleatoria* (o estocástica). Las variables aleatorias (v. a.) son ampliamente utilizadas en la teoría de la probabilidad y en la estadística. En las aplicaciones estadísticas, las variables aleatorias se utilizan para modelar el resultado de un experimento no determinista que genera un resultado aleatorio. La variable aleatoria, nos mide entonces la probabilidad de que la variable tome ciertos valores.

Por ejemplo, considere un muestreo de opinión de 40 personas sobre sus preferencias políticas. El número de personas que prefieren un partido político puede considerarse una variable aleatoria que puede tomar cualquiera de los valores $0, 1, 2 \ldots 40$. Cada uno de estos valores corresponde a un resultado posible del experimento. Otro posible experimento podría ser la precipitación de agua en una zona agrícola. En este caso la variable aleatoria es la 'cantidad de agua' que no es un valor entero, toma valores no numerables, valores reales.

Los ejemplos anteriores nos sugieren que las variables aleatorias pueden ser de dos tipos.

a) Llamaremos ***variables aleatorias discretas***, aquellas que solamente pueden tomar valores numerables enteros positivos. Estos valores pueden ser finitos o infinitos, como podría suponerse.

[1] **Andréi Nikoláyevich Kolmogórov**, nació en Tambov Rusia, el 25 de abril de 1903 y murió en Moscú el 20 de octubre de 1987. Matemático ruso que hizo grandes aportaciones a los campos de la teoría de probabilidad y de la topología. En particular, desarrolló una base axiomática que supone el pilar básico de la teoría de las probabilidades a partir de la teoría de conjuntos. Trabajó al inicio de su carrera en lógica constructivista y en la serie de Fourier. Fue el fundador de la teoría de la complejidad algorítmica.

Ejemplos de variables aleatorias discretas;

a. El número de ventas de una empresa en un mes.
b. El número de votos de un candidato en una elección.
c. Número de accidentes mensuales en una empresa.

b) Si la variable aleatoria toma valores reales, cualquier valor en un intervalo de una recta, es una ***variable aleatoria continua.*** Los valores de una variable continua también pueden ser finitos o infinitos. Por ejemplo;

a. El consumo de gasolina de un auto en 100 km.
b. El valor de una acción en la bolsa de valores un día determinado.
c. El salario mensual de un trabajador.

Es importante la distinción entre variables aleatorias discretas y continuas ya que se requieren modelos probabilísticos distintos para cada una de ellas.

En lo que sigue utilizaremos letras mayúsculas para las variables aleatorias y minúsculas para los resultados de las variable.

10.2 Distribución de probabilidades

Recordemos que una distribución probabilística es la enumeración de todos los resultados de un experimento junto con las probabilidades asociadas. Es una distribución de frecuencia relativa respecto a resultados de un espacio muestral. Las leyes básicas de la probabilidad siguen siendo válidas.

1) La probabilidad de un determinado evento no puede ser menor que cero ni mayor que uno.
2) La suma de las probabilidades de todos los eventos es igual a uno.

No importa si la variable aleatoria es discreta o continua; la distribución de probabilidad la podemos describir como:

a. Un listado teórico de resultados y probabilidades que se pueden obtener con un modelo matemático o función que representen algún fenómeno de interés.
b. Un listado empírico de resultados y sus frecuencias relativas observadas.
c. Un listado subjetivo de resultados asociados con sus probabilidades subjetivas o “inventadas” que representan el grado de convicción del tomador de decisiones en cuanto a la viabilidad de posibles resultados.

10.2.1 Distribuciones de variable discreta

Se denomina distribución de variable discreta a aquella cuya función de probabilidad sólo toma valores positivos numerables en un conjunto de valores de *X.* A esta función se le llama función de distribución de probabilidad. Se deben cumplir las siguientes condiciones;

a) El conjunto de todos los valores que toma la variable aleatoria es finito y numerable
b) Para todo resultado de un experimento de una variable aleatoria se dice que, "la variable aleatoria X toma el valor de x_i; es decir, $[X = x_i]$ y

$$P(X = x_i)$$

Se llama distribución de probabilidad, o *Ley de probabilidad,* de la v. a. X

Considere el experimento de lanzar una moneda. Sea $'x'$ el número de caras obtenidas al lanzar tres veces una moneda. En este caso, los valores que puede tomar la variable aleatoria X son $x = 0, 1, 2, 3$. Los puntos muestrales para este experimento y las probabilidades asociadas se dan en la tabla siguiente. Encuentre la distribución de probabilidad de x.

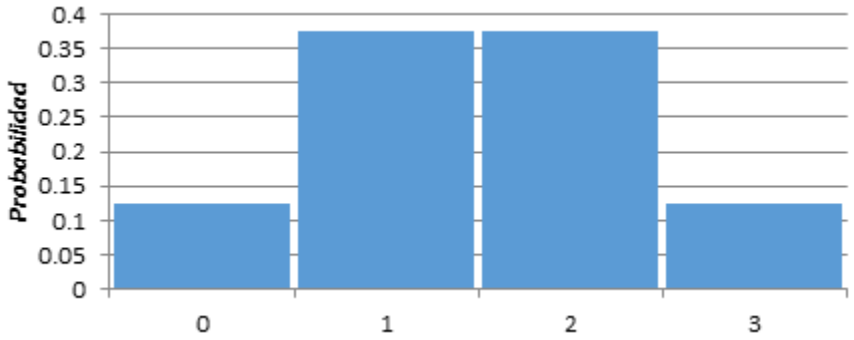
Punto muestral	*Evento*	$P(E_i)$	x
E_1	*XXX*	$1/8$	*0*
E_2	*XXC*	$1/8$	*1*
E_3	*XCX*	$1/8$	*1*
E_4	*CXX*	$1/8$	*1*
E_5	*CCX*	$1/8$	*2*
E_6	*CXC*	$1/8$	*2*
E_7	*XCC*	$1/8$	*2*
E_8	*CCC*	$1/8$	*3*
Total		*1*	

Los valores que toma la variable aleatoria son, de $x = 0$ al evento E_1, $x = 1$ al evento E_2, $x = 2$ al evento E_3 y $x = 3$ al evento E_4. La distribución de probabilidad es la siguiente,

x	Punto muestral	$P(x)$
0	E_1	$1/8$
1	E_2, E_3, E_4	$3/8$
2	E_5, E_6, E_7	$3/8$
3	E_8	$1/8$
	Total	1

En forma gráfica

Tabla 10.1 Distribución de probabilidad.
Número de caras que se obtienen al lanzar tres veces una moneda.



Para evitar confusiones entre una distribución de probabilidad y la distribución de frecuencias, en esta última los datos se agrupan en categorías que muestran el número de observaciones en cada una de ellas. Por otro lado, la distribución de probabilidad nos muestra las probabilidades asociadas a diferentes valores que puede tomar una variable aleatoria.

¿Cuál es la probabilidad de obtener por lo menos dos caras en tres lanzamientos de una moneda?

$$P(por\ lo\ menos\ dos\ caras) = P(dos\ caras) + P(tres\ caras) = \frac{3}{8} + \frac{1}{8} = \frac{1}{2}$$

En resumen, la probabilidad de una variable aleatoria discreta está enteramente determinada por las probabilidades p_i de sus eventos, es decir por el par ordenado (x_i, p_i). Para simplificar la escritura, podemos rescribir el resultado anterior,

$$P(X \geq 2) = P(X = 2) + P(X = 3) = \frac{1}{2}$$

Debemos destacar la analogía de la distribución de probabilidad con la frecuencia relativa. De hecho, podemos considerar que la distribución de probabilidad es una forma ideal, o teórica, cuando el número de observaciones es alto. Por esta razón *las distribuciones de probabilidad se aplican más en poblaciones; mientras que las distribuciones de frecuencia son más utilizadas cuando trabajamos con muestras de una población.*

Si los valores de una variable aleatoria x toman valores enteros $x = 0,1,2,3,4..$ como en el caso anterior, podemos definir la función de distribución acumulada como,

$$F(x) = \sum_{\forall x} p(x)$$

Donde $F(x)$ es la probabilidad de que la variable aleatoria X tome un valor inferior o igual a x.

La tabla del ejercicio anterior se presenta entonces,

x	$P(X \leq x_i)$	$F(x)$
0	$1/8$	$1/8$
1	$3/8$	$4/8$
2	$3/8$	$7/8$
3	$1/8$	1
	1	

Esperanza o Valor esperado de una variable aleatoria discreta.

Si X es una variable aleatoria discreta, llamaremos Esperanza o valor esperado, media, en un espacio de eventos elementales donde,

$$E(X) = \mu = \sum_{i=1}^{n} x_i p_i$$

Una variable aleatoria discreta toma los siguientes valores con las siguientes probabilidades asociadas,

X	*0*	*1*	*2*	*3*	*4*
$P(X = x)$	*0.20*	*.35*	*.25*	*.1*	*.1*

$$E(X) = 0(.2) + 1(.35) + 2(.25) + 3(.1) + 4(.1) = \frac{31}{20} = 1.55$$

Si se realiza la prueba un número suficiente de veces el experimento, en promedio obtendremos como resultado 1.55

Varianza de una variable aleatoria discreta.

La varianza de una variable aleatoria es la desviación al cuadrado de la variable aleatoria con respecto a la media. Es un parámetro de dispersión que es equivalente a la varianza observada. Para calcular la varianza se evalúan las desviaciones cuadradas y se pondera por su probabilidad.

$$Var(X) = \sigma^2 = \sum (x - \mu)^2 P(x)$$

$$= E[(X - \mu)^2] = E[X^2] - (E[X])^2$$

Para el ejercicio anterior,

X	$X-\mu$	$(X-\mu)^2$	$P(X)$	$(X-\mu)^2P(X)$
0	$0-1.55=-1.55$	2.402	0.2	0.48
1	$1-1.55=-.55$	0.302	0.35	0.105
2	$2-1.55=0.45.$	0.202	0.25	0.05
3	$3-1.55=1.45$	2.102	0.1	0.210
4	$4-1.55=2.45$	6.002	0.1	0.6

Así, $Var(X)=\sigma^2=1.445$

Los siguientes teoremas son útiles para el cálculo del valor esperado de una función. Sean X y Y dos variables aleatorias con distribución de probabilidad $P(X)$, $P(Y)$ y c una constante.

a. $E[c]=c$
b. $E[cX]=cE[X] \quad \forall c\in\mathbb{R}$
c. $E[g_1(X)+g_2(X)+\cdots+g_k(X)]=E[g_1(X)]+E[g_2(X)]+\cdots+E[g_k(X)]$
d. $Var(X)=E[(X-\mu)^2]$
   Vamos a demostrar que $\sigma^2=E[(X-\mu)^2]=E[X^2]-(E[X])^2$

$$Var(X)=E[(X-\mu)^2]=E[X^2-2X\mu+\mu^2]$$
$$=E[X^2]-2E[X\mu]+E[\mu^2]$$
$$=E[X^2]-2\mu E[X]+\mu^2 \qquad si\ \mu=E(X)$$
$$Var(X)=E[X^2]-2E[X]^2+E(X)^2=E[X^2]-(E[X])^2$$

e. $Var(X+c)=Var(X)=\sigma^2$

$$Var(X+c)=E[((X+c)-E(X+c))^2]$$
$$=E[(X+c-E(X)-c)^2]$$
$$=E[(X-E(X))^2]=E[(X-\mu)^2]$$

f. $Var(cX)=c^2\sigma^2$

$$Var(cx)=E[(cX)^2]-[E(cX)]^2=c^2E[X^2]-c^2(E[X])^2$$
$$=c^2(E[X^2]-(E[X])^2)=c^2\sigma^2$$

La desviación estándar de la variable aleatoria X será la raíz cuadrada de la varianza.

$$\sigma(X)=\sqrt{Var(X)}$$

Con los datos anteriores tenemos,

$$\sigma(X) = \sqrt{1.445} = 1.202$$

Esta desviación estándar tiene la ventaja de que se expresa en las mismas unidades de la variable aleatoria.

Ejemplo. La demanda de un producto de una empresa varía de mes a mes. La distribución de probabilidad que se presenta en la tabla siguiente, basada en los datos de los dos últimos años. Muestra la demanda mensual de la empresa.

Demanda unitaria	300	400	500	600
Probabilidad	0.2	0.30	0.35	0.15

a) Si la empresa basa las órdenes mensuales en el valor esperado de la demanda mensual, ¿cuál debería ser la cantidad ordenada mensualmente por la empresa para este producto?
b) Suponga que cada unidad demandada genera $70 pesos de ganancia y que cada unidad ordenada cuesta $50 pesos. ¿Cuánto ganará o perderá la empresa en un mes si coloca una orden con base en su respuesta anterior y la demanda real de este artículo es de 300 unidades?

Solución,

X	$P(x)$	$xP(x)$	$X-\mu$	$(X-\mu)^2$	$(X-\mu)^2P(X)$
300	0.2	60	-145	21025	4205
400	0.3	120	-45	2025	607.5
500	0.35	175	55	3025	1058.75
600	0.15	90	155	24025	3603.75
		445			9475

a) El valor esperado de la demanda mensual

$$E(X) = \sum_{i=1}^{n} x_i p_i = 445$$

b) $Demanda\ estimada – demanda\ real = 445 - 300 = 145$. De esta manera pagó un excedente de 145 piezas a un costo de $\$50$ pesos cada una $145 * 50 = \$7{,}250$. Entonces, si la demanda real es de 300 piezas, las ventas son de $300 * 70 = 21{,}000$ menos los costos de $300 * 50 = 15{,}000$. Finalmente, si restamos $ventas\ menos\ los\ costos = 21{,}000 - 15{,}000 = \$6000\ pesos$; por otro lado, como pagamos un excedente de $7,250 por 145 piezas, el resultado final es de una perdida de $1,250 pesos.

10.2.2 Ensayos Bernoulli

Es aquel modelo que sigue un experimento que se realiza una sola vez y que puede tener dos resultados mutuamente excluyentes, por lo regular se denotan como acierto o fracaso:

Los ensayos Bernoulli tienen las siguientes características:

1. La prueba tiene dos resultados posibles ***éxito*** (E) y ***fracaso*** (F).
2. Los ensayos son mutuamente exclusivos; no ocurren a la vez, éxito o fracaso.
3. La probabilidad de éxito es $P(E) = p$ y la probabilidad de fracaso es $P(F) = q$. Nótese que $q = 1 - p$.
4. Una variable aleatoria Bernoulli toma los valores $x = 1$, si hay éxito y $x = 0$ si es fracaso.
5. Los eventos son independientes.

Algunos ejemplos Bernoulli;

- La probabilidad de salir cara al lanzar una moneda al aire (sale cara o no sale);
- Probabilidad de ser admitido en una universidad (o te admiten o no te admiten)
- Probabilidad de acertar una quiniela (o aciertas o no aciertas).

Media y varianza de un ensayo Bernoulli.

Sea X una variable aleatoria tipo Bernoulli con parámetros, p éxito y q fracaso, si sabemos que $p + q = 1$, de esta manera la variable aleatoria toma los valores,

$$X = \begin{cases} 1 & si\ éxito\ \ (p) \\ 0 & si\ es\ fracaso\ (q) \end{cases}$$

La variable X es igual al número de éxitos, o el número de fracasos, según sea el caso luego,

$$\begin{aligned} E[X] &= p * 1 + (1-p) * 0 = p \quad el\ valor\ esperado\ \ E[X] = p \\ E[X^2] &= p * 1^2 + (1-p) * 0^2 = p \quad entonces \\ V[X] &= E[X^2] - E[X]^2 = p - p^2 \quad así, la\ varianza\ es\ \ V[X] = p(1-p) \end{aligned}$$

Y la desviación estándar $\sigma = \sqrt{V[X]} = \sqrt{p(1-p)}$ $o\ bien$ $\sigma = \sqrt{E[(X-\mu)^2]}$

Ejemplo. Se sabe que el 3% de la cuenta de crédito de una institución bancaria están en cartera vencida. Elegimos una cuenta al azar para conocer si no está en cartera vencida ¿cómo se distribuye la variable aleatoria X, si vale 1 cuando la cuenta no está en vencida y 0 si se encuentra en cartera vencida? ¿Cuáles son su media y varianza?

$$\begin{aligned} E[X] &= 0.97 \\ V[X] &= (0.97)(0.03) = 0.0291 \end{aligned}$$

10.2.3 Distribución binomial

La *distribución binomial* se aplica cuando se realizan n ensayos Bernoulli, siendo cada ensayo independiente del anterior. La variable aleatoria puede tomar valores de:

0 Si todos los experimentos han sido fracaso
n Si todos los experimentos han sido éxitos

La variable aleatoria binomial y su distribución están basadas en un experimento que satisface las siguientes condiciones:

- El experimento consiste en una secuencia de *n* ensayos, donde *n* se fija antes del experimento.
- Los ensayos se realizan bajo idénticas condiciones, y cada uno de ellos tiene únicamente dos posibles resultados, que se denotan a conveniencia por $éxito\ (E)$ o $fracaso\ (F)$. De tal manera que, $p(E) + p(F) = 1$.
- Los ensayos son independientes, por lo que el resultado de cualquier ensayo en particular no influye sobre el resultado de cualquier otro intento.
- La probabilidad de éxito es idéntica para todos los ensayos.

Siguiendo estas premisas, la variable aleatoria binomial X está definida como X el número de éxitos en *n* intentos.

De esta manera, si un evento tiene una probabilidad p de que ocurra y por consecuencia una probabilidad $q = 1 - p$ de que no ocurra, decimos que la probabilidad de tener r éxitos en n intentos está dada por el siguiente modelo,

$$b(x = r|n, p) = \binom{n}{r} p^r (1-p)^{n-r}$$

O bien,

$$b(x = r|n, p) = \frac{n!}{r!\,(n-r)!} p^r (1-p)^{n-r}$$

Ejemplo ¿Cuál es la probabilidad de obtener 6 caras al lanzar una moneda 10 veces?

r Es el número de aciertos. En este ejemplo r es igual a 6. En cada acierto decíamos que la variable toma el valor 1; como son 6 aciertos, entonces $r = 6$
n Es el número de ensayos. En nuestro ejemplo son 10
p Es la probabilidad de éxito, obtener **"cara"** al lanzar la moneda. Por lo tanto $p = 0{,}5$

$$b(r = 6|n = 10, p = 0.5) = \binom{10}{6} 0.5^6 (1 - 0.5)^{10-6} = 0.205$$

Es decir, el 20.5% de las veces se obtendrá 6 caras al lanzar 10 veces una moneda.

Ejemplo ¿Cuál es la probabilidad de obtener cuatro veces el número 3 al lanzar un dado ocho veces?

$\boldsymbol{r}$ Número de aciertos, toma el valor 4
$\boldsymbol{n}$ Toma el valor 8
$\boldsymbol{p}$ Probabilidad de que salga un 3 al tirar el dado) es $\frac{1}{6}$ (=0.1666)

$$b(r=4|n=8,p=0.166)=\binom{8}{4}0.166^4(1-0.166)^{8-4}=0.026$$

Es decir, él 2.6% se obtendrán cuatro veces el número 3 al tirar un dado 8 veces.

Ejemplo. Una organización de productores decide invertir en el desmonte de 10 parcelas agrícolas de 5 has cada una. La probabilidad de que una parcela sea improductiva es de 0.3. Las parcelas no son contiguas y podemos suponer independencia. a) ¿Cuál es la probabilidad de que exactamente tres parcelas sean improductivas? b) ¿Cuál es la probabilidad de que al menos 5 sean improductivas? c) ¿Más de 7 de las 10 sean improductivas?

Solución

a) $b(r=3|n=10,p=0.3)=\binom{10}{3}0.3^3(1-0.3)^{10-3}$

$$=\frac{10!}{7!\,3!}(.3)^3(.7)^7=\frac{10*9*8}{3*2}(.3)^3(.7)^7=0.2668$$

b) $b(r\geq 5|n=10,p=0.3)=b(5)+b(6)+\cdots+b(10)$

$$=1-b(0)-b(1)-b(2)-b(3)-b(4)$$

$$b(r=0|10,0.3)=\binom{10}{0}0.3^0(0.7)^{10}=0.0282$$

$$b(r=1|10,0.3)=\binom{10}{1}0.3^1(0.7)^9=0.1211$$

$$b(r=2|10,0.3)=\binom{10}{2}0.3^2(0.7)^8=0.2335$$

$$b(r=3|10,0.3)=\binom{10}{3}0.3^3(0.7)^7=0.2668$$

$$b(r=4|10,0.3)=\binom{10}{4}0.3^4(0.7)^6=0.2001$$

$$b(r\geq 5|n=10,p=0.3)=1-(0.0282+0.1211+0.2335+0.2668+0.2001)$$
$$=0.1503$$

c) $b(r>7|n=10,p=0.3)=b(8)+b(9)+b(10)=0.0014+0.0001+0.0000059$
$$=0.0016$$

Para obtener la probabilidad binomial con un programa de cómputo como Excel.

$$=DISTR.BINOM.N(r,n,p,FALSO)$$

El valor esperado de la distribución binomial, considerando que es un proceso Bernoulli,

$$\mu=El\ valor\ esperado\ es\ E[y]=np\quad y$$
$$\sigma^2=la\ varianza\ Var[y]=npq$$

Ejemplo. El 75% de las viviendas en una alcaldía de la ciudad de México no tienen agua. Una colonia, de esta alcaldía, cuenta con 2500 viviendas. Si X es la variable aleatoria que indica el número de viviendas que no tienen agua, a) ¿Cuál es el número esperado de viviendas sin agua? b) Encuentre la varianza y la desviación estándar de las viviendas sin agua y c) Utilice el teorema de Tchebyscheff para encontrar entre que limites esperaría que estuviera el valor de x?

Solución,

a) $\mu = E(X) = np = 2500(.75) = 1875$
b) $\sigma^2 = npq = 2500(.75)(.25) = 468.75$ y $\sigma = \sqrt{npq} = \sqrt{468.75} = 21.65$
c) Por la regla empírica de Tchebyscheff, sabemos que,

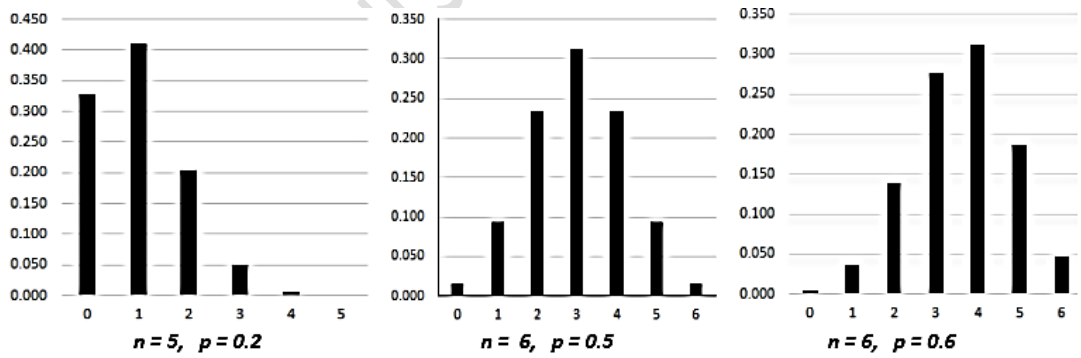
$(\mu \pm 2\sigma)$ Se encuentran el 75% de las observaciones
$(\mu \pm 3\sigma)$ Se encuentran el 89% de las observaciones
$(\mu \pm 4\sigma)$ Se encuentran el 94% de las observaciones

De esta manera para nuestro ejercicio,

$[1875 \pm 2(21.65)]$ Entre 1831.7 y 1918.3 están el 75% de las observaciones
$[1875 \pm 3(21.65)]$ Entre 1810.05 y 1939.95 están el 89% de las observaciones
$[1875 \pm 4(21.65)]$ Entre 1788.4 y 1961.6 están el 94% de las observaciones

La representación, en un gráfico de barras, de la distribución binomial se muestra a continuación. La forma de la distribución depende de los valores de p y de q, como se muestra en los siguientes gráficos.

Grafica 10.1 *Comparación de la distribución binomial para diferentes valor de* p



Se pueden hacer varias observaciones sobre estos diagramas.

a) La forma de la distribución es simétrica si $p = 0.5$, sin importar el valor de n.
b) Como consecuencia, es asimétrica cuando $p \neq 0.5$. Si p es menor que 0.50, las probabilidades son más altas en el lado izquierdo de la distribución que en el lado derecho (asimetría positiva). Si p es mayor que 0.5, es lo opuesto (asimetría negativa).

c) La distribución tiende a ser simétrica cuando n es grande. Además, si p no está demasiado cerca de 0 o 1, se aproximará a la distribución de la distribución normal como lo veremos más tarde.

Suma de dos variables binomiales

Si X_1 y X_2 se distribuyen como una binomial y son variables independientes, entonces

$$X_1 + X_2 \sim b(x = k|n_1 + n_2, p)$$

Es decir, si X_1 y X_2 son el número de éxitos en n_1 y n_2 ensayos Bernoulli respectivamente, con la misma probabilidad de éxito; entonces, $X_1 + X_2$ representa el número de éxitos en $n_1 + n_2$ pruebas idénticas e independientes

10.2.4 Distribución Pascal o binomial negativa (b^-)

La distribución Pascal, también llamada *Distribución binomial negativa,* se caracteriza porque si se realizan n pruebas Bernoulli estamos interesados en el número de éxitos r que pueden ocurrir. Es decir, nos interesa conocer el número de ensayos $'n'$ que se deben efectuar para obtener $'r'$ éxitos. Así, la última prueba siempre será un éxito. Los ensayos son Bernoulli con probabilidad de éxito p; así, la probabilidad de efectuar exactamente n pruebas hasta obtener r éxitos es:

$$b^-(N = n \mid r, p) = \binom{n-1}{r-1} p^r q^{n-r} \ \ para\ 0 < r \leq n$$

Donde,

p Es la probabilidad de éxito en una prueba tipo Bernoulli
N Es el número de pruebas hasta que se observan $r - 1$ éxitos

La media y la varianza de una distribución Pascal, o Binomial negativa, es

$$E(N) = \frac{r}{p} \quad y \quad Var(N) = \left(\frac{r}{p}\right)\left(\frac{1}{p} - 1\right) = \frac{r(1-p)}{p^2}$$

Ejemplo. Una empresa decide invertir en un grupo de acciones. De acuerdo con estudios de una casa de bolsa, sabe que el 5% de estas acciones no aumentarán su valor. Encontrar los siguiente.

a) Si se seleccionan 20 acciones al azar, ¿cuál es la probabilidad de que al menos 2 de ellas no aumenten su valor? La probabilidad de que una acción no aumente su valor es $p = 0.05$.

Es una binomial, por lo tanto, la probabilidad que nos solicitan es:

$$b(r \geq 2 \,|N = 20\,, 0.05) = 1 - \binom{20}{0} p^0 q^{20} - \binom{20}{1} p^1 q^{19}$$
$$= 1 - 0.359 - 0.3774 = 0.264$$

b) Si la empresa requiere que 10 acciones aumenten su valor, ¿cuál es la probabilidad de que tengan que seleccionar exactamente 12 para obtener 10 acciones redituables?

$$b^-(N = 12|\ 10, .95) = \binom{12-1}{10-1} 0.95^{10} 0.05^2 = 0.083$$

c) ¿Cuál es el número esperado de acciones que debería seleccionar para obtener 10 redituables? En este caso usamos como probabilidad de éxito, cuando la acción aumenta su valor; entonces $p = 1 - 0.05 = 0.95$

$$E(N) = \frac{10}{0.95} = 10.5 \quad y \quad Var(N) = \frac{10(1-.95)}{0.95^2} = 0.55$$

Ejemplo. Si la probabilidad de que una persona con gripe contagie a otros trabajadores en una empresa es de 0,40, ¿cuál es la probabilidad de que la décima persona expuesta a la enfermedad sea la tercera en contraerla? En este caso, N es el número de personas expuestas a la enfermedad y, $n = 10, r = 3\ y\ p = 0.4$

$$b^-(N = 10\,|3{,}0.4) = \binom{10-1}{3-1} 0.4^3 (1-0.4)^{10-3} = 0.0645$$

10.2.5 Distribución geométrica- (G).

Un caso especial de la distribución Pascal es la distribución geométrica. En la distribución Pascal los valores de n y r pueden ser cualquier valor entero donde $n \geq r$. Si el valor de $r = 1$; es decir, el proceso se repite hasta obtener el primer éxito, entonces la variable aleatoria N se distribuye como una distribución geométrica,

$$G(N = n|p) = pq^{n-1}$$

Donde N es el número de ensayos hasta observar el primer éxito. La media y la varianza de una distribución geométrica es,

$$E(N) = \frac{1}{p} \quad y \quad Var(N) = \left(\frac{1}{p}\right)\left(\frac{1}{p} - 1\right) = \frac{(1-p)}{p^2}$$

Demostración.

$$E(X) = \sum_{N \geq 1} npq^{n-1} = p \sum_{N \geq 0} nq^{n-1} = \frac{p}{(1-q)^2} = \frac{1}{p}$$

Ejemplo. Una máquina despachadora de refrescos falla en la entrega del producto el 5% de las veces, calcular la probabilidad de que la maquina falle en la entrega de producto con la octava persona que adquiera el producto.

Solución. Si llamamos *'éxito'* que la máquina no entregue el producto. Entonces $N = 8$ y la probabilidad p de éxito es $p = 0.05$.

De esta manera la probabilidad q de fracaso es $q = 1 - 0.05 = 0.95$. Asi la probabilidad de que la máquina falle con la octava persona,

$$G(N = 8) = (0.05)(0.95)^7 = 0.0349$$

El número esperado de entregas hasta que la máquina falle es de,

$$E(N) = \frac{1}{0.05} = 20 \quad y \quad Var(N) = \frac{(1 - 0.05)}{0.05^2} = 380$$

10.2.6 Distribución Hipergeométrica

Al igual que la binomial, se trata de encontrar el número de éxitos en una muestra que tiene n observaciones. Lo que los distingue es la forma en que se obtienen los datos. En el caso binomial la distribución es adecuada si la probabilidad de un éxito permanece constante para cada intento. Si la población es pequeña y no hay remplazo, la probabilidad de éxito será diferente en cada ensayo. En el modelo Hipergeométrica los datos de la muestra se extraen sin remplazo de una población finita, Es decir, el valor de la probabilidad de éxito p cambia.

Una variable aleatoria tiene una distribución Hipergeométrica si,

a) El experimento consiste en extraer al azar y sin remplazo n elementos de una población N, k de los cuales son éxitos.
b) El tamaño de la muestra n es grande, pero no mayor que la población N, y se relacionan de acuerdo con la expresión $\frac{n}{N} > 0.05$
c) La variable aleatoria es el número r de éxitos en la muestra de n elementos.

$$h(X = r|N, n, k) = \frac{\binom{k}{r}\binom{N-k}{n-r}}{\binom{N}{n}}$$

Dónde:

$N =$ Tamaño de la población
$n =$ Tamaño de la muestra.
$k =$ Número de éxitos en la población.
$N - k =$ Número de fracasos en la población
$r =$ Número de éxitos en la muestra

La distribución hipergeométrica tiene algunas limitantes, como el tamaño de la muestra o el número de éxitos en la población. Está claro que el número de éxitos en la muestra r no debe exceder el número de éxitos en la población.

Si la probabilidad p es la proporción de éxitos en la población,

$$p = \frac{éxitos\ en\ la\ población}{Población} = \frac{k}{N}$$

Y por consiguiente la proporción de fracasos es,

$$(1-p) = \frac{número\ de\ fracasos\ en\ la\ población}{tamaño\ de\ la\ población} = \frac{N-k}{N}$$

La media y la varianza de una distribución hipergeométrica es,

$$\mu = E(X) = np = \frac{nk}{N}$$

$$\sigma^2 = Var(X) = np(1-p) = \frac{nk(N-k)(N-n)}{N^2(N-1)}$$

Ejemplo Una serie de 8 lámparas se conectan de tal forma que si una falla, el sistema no funcionará. Se sabe que 2 lámparas no funcionan.

a) ¿Cuál es la probabilidad de que la primera que se inspeccione sea una de las que falló?

$$N = 8;\quad k = 2;\ \ n = 1;\ r = 1$$

$$h(r = 1|N = 8, n = 1, k = 2) = \frac{\binom{2}{1}\binom{6}{0}}{\binom{8}{1}} = 0.25$$

b) ¿Cuál es la probabilidad de encontrar las dos que fallan si se inspeccionaron 4 de ellas?

$$N = 8;\quad k = 2;\ \ n = 4;\ r = 2$$

$$h(r = 2|N = 8, n = 4, k = 2) = \frac{\binom{2}{2}\binom{6}{2}}{\binom{8}{4}} = \frac{15}{70} = 0.2143$$

c) ¿Cuántas lámparas se pueden inspeccionar para tener un 70% de probabilidades de encontrar las dos lámparas defectuosas?

$$N = 8;\quad r = 2;\ \ n = ?\,;\ k = 2$$

$$h(r=2|N=8,n=?,k=2)=\frac{\binom{2}{2}\binom{6}{n-2}}{\binom{8}{n}}=0.70$$

$$\frac{\frac{6!}{(n-2)!\,(8-n)!}}{\frac{8!}{(8-n)!\,n!}}=0.70 \Rightarrow \frac{6!\,n!}{8!\,(n-2)!}=0.70 \Rightarrow \frac{n(n-1)}{8*7}=0.70$$

$$n^2-n=39.2 \quad \Rightarrow \quad n=6.78 \quad n=7\ discreta$$

10.2.6.1 Aproximación de binomial a la hipergeométrica

Como regla empírica, cuando se muestrea sin remplazo, siempre que el tamaño de la muestra es menor de 5% del tamaño de la población, se puede utilizar la distribución binomial para aproximar la hipergeométrica $(n<0.05N)$.

Para el ejemplo anterior, si nos interesa la probabilidad de que 0,1 o 2 focos no sirven en una muestra de 4. En la siguiente tabla comparamos los dos cálculos.

En primer lugar, la regla empírica nos dice que, podemos utilizar la aproximación de la binomial.

$$n<0.05N \qquad 4<0.05(100)$$

La tabla siguiente nos ayuda con la comparación de ambos modelos probabilísticos.

$Binomial, con\ remplazo$	$Hipergeométrica, sin\ remplazo$
$b\left(x=0\middle\|4,p=\frac{2}{100}=0.02\right)=\ 0.922$	$h(x=0\|N=100,n=4,k=2)=\ 0.921$
$b(x=1\|4,p=0.02)=\ 0.075$	$h(x=1\|N=100,n=4,k=2)=\ 0.077$
$b(x=2\|4,p=0.02)=\ 0.002$	$h(x=2\|N=100,n=4,k=2)=\ 0.0012$

La probabilidad de que 2 o menos focos no funcionan, que es la suma de las tres probabilidades es,

$b(x\leq 2\|4,p=0.02)=0.99$	$h(x\leq 2\|N=100,n=4,k=2)=\ 0.99$

Supongamos ahora que tenemos una población que contiene un número finito n de elementos divididos en R clases mutuamente exclusivas donde $k_1$es el grupo 1, k_2 es el grupo 2,…. k_i. Se selecciona una muestra de n observaciones al azar y sin remplazo y obtenemos r_1 elementos de la clase 1, r_2 elementos de la clase 2, etc. La probabilidad de ocurrencia de esta muestra es,

$$P(r = r_1, r_2, .., r_R | N, n, k = k_1, k_2, .., k_R) = \frac{\binom{k_1}{k_1}\binom{k_2}{r_2}...\binom{k_R}{r_R}}{\binom{N}{n}}$$

Donde

$$k_1 + k_2 + \cdots + k_R = N \quad \text{y} \quad r_1 + r_2 + \cdots + r_R = r$$

Ejemplo. Supongamos que una organización de productores cafetaleros produce tres tipos de café, arábica, robusta y orgánico. Para su comercialización cuentan con una cartera pequeña, población, de 20 compradores. 10 de estos prefieren el café arábico, 6 prefieren el robusta y 4 prefieren café orgánico. Se toma una muestra de 10 compradores, sin remplazo. ¿Cuál es la probabilidad de que 7 prefieran el café arábico, 2 robusta y solo 1 prefiera el café orgánico?

$$P(r = 7,2,1 | N = 20, n = 10, k = 10,6,4) = \frac{\binom{10}{7}\binom{6}{2}\binom{4}{1}}{\binom{20}{10}} = \frac{(120)(15)(4)}{(184756)} == .0389$$

Ejemplo. En un estudio biológico se emplea un grupo de 10 individuos. El grupo consiste en 3 personas con sangre tipo O, 4 con tipo A y 3 con tipo B. ¿Cuál es la probabilidad de que una muestra aleatoria de 5 contenga una persona con sangres tipo O, 2 con tipo A y dos con tipo B?

En este ejemplo: $N = 10; \quad k_1 = 3; \; k_2 = 4; \; k_3 = 3$

$$n = 5; \quad r_1 = 1; \; r_2 = 2; \; r_3 = 3$$

$$P(r = 1,2,2 | N = 10, n = 5, k = 3,4,3) = \frac{\binom{3}{1}\binom{4}{2}\binom{3}{2}}{\binom{10}{5}} = 0.2143$$

10.2.7 Distribución de Poisson

La distribución de Poisson es una distribución de probabilidad discreta que expresa la probabilidad de que un número r de eventos ocurren en un tiempo fijo y si estos eventos ocurren con una tasa media conocida, y son independientes del tiempo desde el último evento. Para eventos de este tipo nos interesa solo el número de ocurrencias del evento, no su falta de ocurrencia.

La distribución Poisson, Se aplica a varios fenómenos discretos de la naturaleza (esto es, aquellos fenómenos que ocurren 0, 1, 2, 3, ... veces durante un periodo definido de tiempo o en un área determinada) cuando la probabilidad de ocurrencia del fenómeno es constante en el tiempo o el espacio.

Eventos que pueden ser modelados por la distribución Poisson pueden ser:

- El número de autos que pasan a través de un cierto punto en una ruta (suficientemente distantes de los semáforos) durante un periodo definido de tiempo.
- El número de errores en las páginas de un libro
- El número de personas que adquieren una enfermedad, en una estación, por año, etc.
- El número de accidentes en un tramo de una carretera.
- El número de cheques emitidos sin fondos.
- El número de accidentes que ocurren por asegurado en un intervalo de tiempo, por día, mes año
- El número de accidentes que ocurren a los trabajadores de una empresa en una jornada de trabajo, por semana, por mes, por año, etc.

Se dice que una variable aleatoria sigue una distribución tipo Poisson, si al considerar un experimento binomial, donde el tamaño de la muestra es grande y la probabilidad de éxito es pequeña, en el que se observa la aparición de sucesos puntuales en un intervalo continúo de tiempo, longitud, área, etc., en cualquier intervalo suficientemente pequeño, se verifica que:

1. El experimento consiste en contar el número x de veces que ocurre un evento en particular, durante una unidad de tiempo, o en un área o volumen.
2. La probabilidad de que un evento ocurra en una unidad de tiempo, área, volumen, etc., es la misma para todas las unidades.
3. El número de eventos que ocurren en una unidad de tiempo es independiente de otros.
4. El número medio (o esperado de eventos en cada unidad se denota por λ.

La distribución de Poisson sigue el siguiente modelo:

$$P(x = r) = e^{-\lambda}\ \frac{\lambda^r}{r!}$$

Donde;

r Es el número de éxitos

λ Es la media de la distribución de probabilidad. También es conocida como la "tasa entre arribos". Se calcula como $\boldsymbol{\lambda = np}$; donde n son las veces que se repite el experimento y p es la probabilidad de éxito.

Los parámetros estadísticos de la distribución Poisson son:

El valor esperado $E(X) = \sum_{r=0}^{\infty} r\,\frac{e^{-\lambda}\lambda^r}{r!} = \lambda$

La varianza $V(X) = E(X^2) - [E(X)]^2 = \lambda$

Demostración[2],

$$E(X) = \sum_{r\geq 0} r\frac{\lambda^r e^{-\lambda}}{r!} = e^{-\lambda}\sum_{r\geq 1}\frac{\lambda^r}{(r-1)!} = e^{-\lambda}\lambda e^{\lambda} = \lambda$$

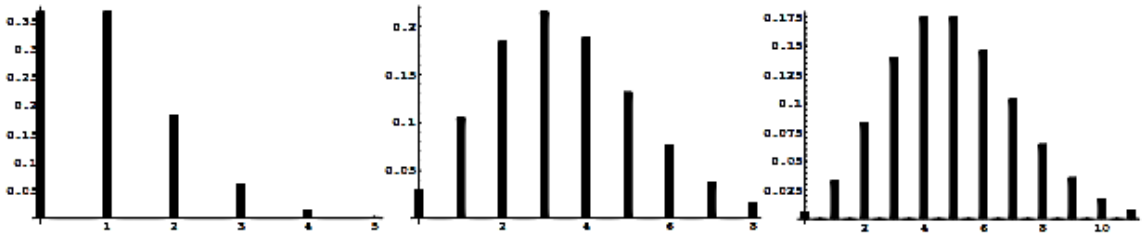
$$Var(X) = E(X^2) - \big(E(X)\big)^2 = E\big(X(X-1)\big) + E(X) - \big(E(X)\big)^2$$

$$E\big(X(X-1)\big) = \sum_{k\geq 0} r\,(r-1)\frac{\lambda^r e^{-\lambda}}{r!} = e^{-\lambda}\sum_{k\geq 0} r\,(r-1)\frac{\lambda^r}{r!}$$

$$= e^{-\lambda}\lambda^2 e^{\lambda} = \lambda^2$$

$$Var(X) = \lambda^2 + \lambda - \lambda^2 = \lambda$$

La forma de la distribución Poisson depende del parámetro λ



$\lambda = 1$ $\lambda = 3.5$ $\lambda = 5$

En general, el diagrama es asimétrico con relación a λ, y a medida que el valor de λ aumenta, la distribución tiende a la simetría, se aproxima a una distribución normal, que veremos más adelante.

Ejemplo. La probabilidad de tener un accidente de tráfico es de 0.02 cada vez que se viaja. Si se realizan 300 viajes, ¿cuál es la probabilidad de tener 3 accidentes?

$$n = 300 \quad p = 0.02 \quad \lambda = np = 300(0.02) = 6$$

$$P(x = 3) = e^{-6} * \frac{6^3}{3!} = 0.0892$$

Por lo tanto, la probabilidad de tener 3 accidentes de tráfico en 300 viajes es del 8.9%

Ejemplo. Una empresa se dedica a sembrar maíz y experimentan una plaga llamada "gusano elotero". La inspección de 5000 mazorcas seleccionadas al azar revelo que se encontraron

[2] Recordemos que la serie $1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \cdots .. = \sum_{n=0}^{\infty}\frac{x^n}{n!} = e^x$

3500 gusanos. ¿Cuál es la probabilidad de que una mazorca al azar no tenga gusanos?, ¿Cuál es la probabilidad de que 5 mazorcas tengan gusanos?

$$n = 5000 \quad \lambda = \frac{3500}{5000} = 0.7 \; número\ de\ gusanos\ por\ mazorca$$

a) $P(x = 0) = \frac{(0.7)^0 e^{-0.7}}{0!} = 0.497$

b) 5 mazorcas tengan gusanos $P(x = 5) = 0.7^5 * \frac{e^{-0.7}}{5!} = 0.0007$. Por lo tanto, la probabilidad de que 5 mazorcas tengan gusanos es de 0.07%.

Ejemplo. Una empresa fabrica frascos de frutas en almíbar, 1.2% de estos frascos no cumplen con las normas de calidad y entonces son rechazados. Se recibe un pedido de 750 frascos ¿Cuál es la probabilidad de que haya 5 de mala calidad?

$$n = 750, \;\; p = 0.012 \;\; \lambda = np = 750(0.012) = 9$$

$$P\ (x = 5) = e^{-9}\frac{(9^5)}{5!} = 0.0607$$

Podríamos intentar resolver el problema por medio de la binomial. En este caso tendríamos;

$$b(x = 5|750,0.012) = \binom{750}{5} 0.012^5(1 - 0.012)^{795}$$

Estos valores no se encuentran en tablas. Por lo tanto, tenemos que resolver en forma manual.

$$\binom{750}{5} 0.012^5(1 - 0.012)^{745} = \frac{750 * 749 * 748 * 747 * 746}{120} 0.012^5(1 - 0.012)^{745} = 0.0602$$

Lo mismo que lo visto para la distribución binomial, para calcular las probabilidades de Poisson se dispone de tablas estadísticas tabuladas para distintos valores (ver apéndice 2).

Ejemplo. Suponga que estamos investigando la seguridad de un cruce de avenidas peligroso. Los archivos de la policía indican que ocurren en promedio 5 accidentes por mes. ¿Cuál es la probabilidad de tener exactamente 0, 1, 2, 3 y 4 accidentes en un mes?

$\lambda = 5$ La media es igual a la varianza.

$$P(0) = \frac{5^0 e^{-5}}{0!} = \frac{(1)(0.00674)}{1} = 0.00674$$
$$P(1) = \frac{5^1 e^{-5}}{1!} = 0.0404 - 0.0067 = 0.3370 \;\; (obtenido\ de\ tablas)$$
$$P(2) = \frac{5^2 e^{-5}}{2!} = 0.1247 - 0.0404 = 0.0843 \;\; (obtenido\ de\ tablas)$$
$$P(3) = \frac{5^3 e^{-5}}{3!} = 0.2650 - 0.1247 = 0.1403 \;\; (obtenido\ de\ tablas)$$

$$P(4) = \frac{5^4 e^{-5}}{4!} = 0.4405 - 0.2650 = 0.1755 \ \ (obtenido\ de\ tablas)$$

10.2.8.1 Aproximación de la binomial por la distribución Poisson.

En Economía se presentan muchas aplicaciones de experimentos donde la muestra es grande, por lo que se recomienda el uso de la distribución Poisson. En virtud de que es difícil encontrar tablas de la distribución binomial para valores grandes de n, es necesario utilizar métodos de cálculo simples que nos den una solución aproximada de estas probabilidades.

La regla empírica que utilizan con más frecuencia los estadísticos es que la distribución de Poisson es una buena aproximación de la distribución binomial cuando $n \geq 20$ y la probabilidad de éxito p es igual o menor que 0.05. O bien, cuando $np \leq 5$ o bien $n(1-p) \leq 5$. En los casos en que se cumplen estas condiciones, podemos sustituir la media de la distribución binomial np por λ que es la media de la distribución de Poisson.

Ejemplo. Supongamos que una organización de 5000 productores agrícolas desea que sus productores no estén en cartera vencida. Si la probabilidad de que un productor este en cartera vencida es de 0.001. ¿Cuál es la probabilidad de tener exactamente 4 productores en cartera vencida?
Solución. Por el modelo binomial

$$b(x = 4|5000,0.001p) = \binom{5000}{4}(0.001)^4(1-0.001)^{4996} = 0.1755$$

Las tablas de la distribución binomial son insuficientes, y el cálculo manual es muy laborioso. La alternativa de solución fácil es por medio de la aproximación de la binomial a la Poisson. Si hacemos,

$$n = 5000 \quad p = 0.001 \quad \lambda = np = 5000(0.001) = 5$$

La probabilidad por medio de la distribución Poisson es entonces;

$$P(4) = \frac{5^4 e^{-5}}{4!} = 0.4405 - 0.2650 = 0.1755 \ \ (obtenido\ de\ tablas)$$

Tabla distribución Poisson

X	*3.50*	*4.00*	*4.50*	*5.00*	*5.50*	*6.00*
0	.0302	.0183	.0111	.0067	.0041	.0025
1	.1359	.0916	.0611	.0404	.0266	.0174
2	.3208	.2381	.1736	.1247	.0884	.0620
3	.5366	.4335	.3423	.2650	.2017	.1512
4	.7254	.6288	.5321	.4405	.3575	.2851
5	.8576	.7851	.7029	.6160	.5289	.4457

En algunas tablas, como la que se muestra, los valores que se obtienen son acumulados; es decir para el caso, como en nuestro ejemplo, de obtener 4 éxitos, la tabla nos da el resultado

de obtener “hasta 4 éxitos”. Por esta razón es que restamos el valor de Poisson para cuatro éxitos de el de tres.

Ejercicios

1. A un muelle de carga llegan camiones en forma aleatoria con un promedio de 1 por hora. ¿Cuál es la probabilidad de que en una hora no llegue ningún camión al muelle?
2. Una casa de créditos recibe en promedio 2.2 solicitudes de préstamos para mejoramiento de la vivienda por semana. Calcule la probabilidad de que en una semana:
   a) lleguen tres solicitudes de crédito
   b) llegue por lo menos una solicitud
   c) lleguen entre 2 y 5 solicitudes (incluidos estos valores)
   d) lleguen a lo sumo tres solicitudes
3. La llegada de vehículos a un puesto de peaje sigue un proceso de Poisson con promedio 4 llegadas por minuto. Calcule la probabilidad de que en dos minutos lleguen por lo menos tres vehículos al puesto de peaje.
4. El número de averías semanales de una computadora es una variable aleatoria que tiene distribución de Poisson con promedio l = 0.4. ¿Cuál es la probabilidad de que la computadora trabaje sin averías durante dos semanas consecutivas?

Resumen distribución de probabilidad discretas

Distribución	*Características*	*Fórmula*	*Medidas*
Binomial	Obtener r, éxitos en n intentos	$b(x=r\|n,p)=\binom{n}{r}p^r(1-p)^{n-r}$	$E[y]=np\ \ Var[y]=npq$
Pascal o Binomial negativa	Probabilidad de efectuar exactamente n pruebas hasta obtener r éxitos	$b^-(N=n\mid r,p)=\binom{n-1}{r-1}p^rq^{n-r}$ $para\ 0<r\leq n$	$E(N)=\frac{k}{p}$ $Var(N)=\frac{r(1-p)}{p^2}$
Geométrica	Probabilidad de efectuar exactamente n pruebas hasta obtener el primer éxito	$P(N=n\|p)=pq^{n-1}$	$E(N)=\frac{1}{p}$ $Var(N)=\frac{(1-p)}{p^2}$
Hipergeométrica	Si el número de elementos en la población es pequeño en relación con la muestra. La probabilidad de acierto en un ensayo depende de los resultados precedentes. La muestra $n<0.05N$	$P(X=r\|N,n,k)=\frac{\binom{k}{r}\binom{N-k}{n-r}}{\binom{N}{n}}$	$E(K)=\frac{nk}{N}$ $Var(K)=\frac{nk(N-k)(N-n)}{N^2(N-1)}$
Poisson	Expresa la probabilidad de que un número r de eventos ocurren en un tiempo fijo y si estos eventos ocurren con una tasa media conocida	$P(x=k)=e^{-\lambda}\ \frac{\lambda^r}{r!}$	$E(X)=\lambda$ $Var(X)=\lambda$

10.3 Variables aleatorias continuas

En las distribuciones de probabilidad presentadas hasta ahora, las variables aleatorias toman sólo valores aislados, por lo tanto, se hacía referencia a distribuciones de probabilidad de variables aleatorias discretas. Si x es una variable aleatoria continua, la probabilidad de que tome un valor particular cualquiera es nulo. Para definir una distribución de probabilidad de una variable continua deben de establecerse con claridad que su valor debe caer en un intervalo.

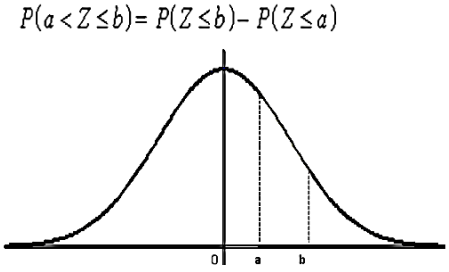
Una variable aleatoria es **continua** cuando el rango o recorrido de esta es un conjunto infinito no numerable.

Ejemplos de variables aleatorias continuas pueden ser,

a) Tiempo de demora en el otorgamiento de un crédito,
b) Años de antigüedad de los empleados de una empresa,
c) Salario de los trabajadores de una empresa,
d) Duración de componentes electrónicas, etc.

Función de densidad

La representación gráfica de una variable continua se realiza a través de un histograma con su correspondiente polígono de frecuencias relativas. La siguiente figura muestra que a medida que aumenta el tamaño de la muestra y disminuye la amplitud del intervalo de clase, el polígono de frecuencias tiende a una curva llamada **curva de densidad**



Por *DENSIDAD* entendemos la concentración de probabilidad dentro de un intervalo de valores de la variable aleatoria x.

La curva de densidad es la representación de una función, llamada **función de densidad,** que se simboliza con $f(x)$ y que verifica:

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$
3. $P(a \leq X \leq b) = \int_{a}^{b} f(x)\, dx$

La probabilidad de que un valor de X sea igual a cualquier valor específico es cero, en consecuencia, el signo de igualdad puede o no incluirse en la especificación del intervalo.

$$P\ (a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b)$$

Donde $\boldsymbol{F(x) = P(X < x) = \int_{-\infty}^{x} f(s)ds}$ es la función de distribución de probabilidad acumulada de X.

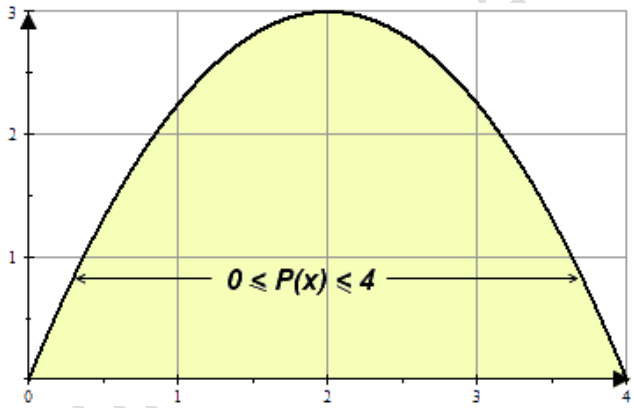
Se deduce: $f\ (x)\ = \frac{d}{dx}\ F(x)$

Ejemplo; Supongamos que x es una variable aleatoria continua con función de densidad,

$$f(x) = \begin{cases} k(-2x^2 + 8x) & si \quad 0 < x < 4 \\ 0 & si\ no \end{cases}$$

¿Cuál es el valor de la constante k, y cual es la probabilidad de que $P(x > 2)$

Grafica 10.x función de densidad de probabilidades



Solución, el hecho de que $f(x)$ sea una función de densidad, nos permite suponer que

$$\int_{-\infty}^{\infty} f(x)\ dx = 1 \quad entonces \qquad \int_{0}^{4} k(-2x^2 + 8x)\ dx = 1$$

Resolvemos esta integral,

$$1 = k\int_{0}^{4}(-2x^2 + 8x)\ dx = k\left[-\frac{2x^3}{3} + 4x^2\Big|_0^4\right] = k\left(-\frac{2(4)^3}{3} + 4(4)^2\right) - 0$$

$$= k\left(-\frac{128}{3} + 64\right) \rightarrow \frac{64}{3}k = 1 \quad k = \frac{3}{64}$$

Para encontrar la probabilidad de $p(x > 2)$

$$P(x > 2) = \frac{3}{64}\int_{2}^{4}(-2x^2 + 8x)\ dx = \frac{3}{64}\left[-\frac{2x^3}{3} + 4x^2\Big|_2^4 = \frac{3}{64}\left(\frac{64}{3} - \frac{32}{3}\right) = \frac{1}{2}\right]$$

Parámetros estadísticos

Si x es una variable aleatoria con función de densidad $f(x)$, entonces

- *Esperanza matemática o promedio poblacional* $E(X) = \int_{-\infty}^{\infty} xf(x)dx$
- *Varianza poblacional* $V(X) = \sigma^2 = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx = E(X^2) - [E(X)]^2$
- *Desviación estándar poblacional* $\sigma = +\sqrt{\sigma^2}$

La propiedad de linealidad del valor esperado es válida también para las variables aleatorias continuas; así como las propiedades de la varianza y de la desviación estándar.

Ejemplo. Considere una función del tipo $f(x) = kx^2$

a) Encontrar el valor de k que hace que $f(x)$ sea una función de densidad de probabilidades en el intervalo $0 \leq x \leq 4$

$$\int_0^4 kx^2\,dx = 1$$

$$\int_0^4 kx^2\,dx = \frac{1}{3}kx^3\Big]_0^4 = \frac{1}{3}k(0) - 0 \to 1 = \frac{64}{3}k \quad \therefore \quad k = \frac{3}{64}$$

b) Sea x una variable aleatoria continua con función de densidad $f(x)$ calcular la probabilidad de $P(1 \leq x \leq 2)$

$$P(1 \leq x \leq 2) = \int_1^2 \frac{3}{64}x^2dx = \frac{1}{64}x^3]_1^2 = \frac{8}{64} - \frac{1}{64} = \frac{7}{64}$$

c) Calcular el valor esperado, varianza y la desviación estándar, cuando $0 \leq x \leq 4$

$$E(X) = \int_0^4 \frac{3}{64}xx^2dx = \frac{3}{64}\int_0^4 x^3dx = \frac{3}{256}x^4]_0^4 = 3$$

$$\sigma^2 = E(X^2) - [E(X)]^2 = \frac{3}{64}\int_0^4 x^2x^2dx - 3^2$$

$$= \frac{3}{64}\int_0^4 x^4dx - 3^2 = \frac{3}{320}x^5\Big|_0^4 - 9 = \frac{48}{5} - 9 = \frac{3}{5}$$

$$\sigma = \sqrt{\frac{3}{6}} = 0.774$$

10.3.1 Teorema de CHEBYSHEV

El teorema de Chebyshev muestra una propiedad muy útil de la desviación estándar. En particular dice que, independientemente de cómo se distribuyan los valores de la variable aleatoria X, al menos $(1 - 1/k^2)$ % de dichos valores estarán a menos de k desviaciones estándar del promedio.

Es decir: $$P[|X-\mu| \geq k\sigma] \leq \frac{1}{k^2} \qquad y \qquad P[|X-\mu| \leq k\sigma] \geq \frac{1}{k^2}$$

Como lo indicamos anteriormente, esta ecuación no tiene sentido para valores de $k = 1$. Si el valor de $k = 2$, tendremos el 75% $\left(1 - \frac{1}{4}\right)$ de las observaciones, o la probabilidad de 0.75. Igualmente para $k = 3\ y\ k = 4$, tendremos el 89% y el 94% con probabilidades de 0.89 y de 0.94 respectivamente.

La aplicación de esta regla es útil cuando no se conoce la función de probabilidad de la variable y nos permite obtener valores aproximados de las probabilidades asociadas a dicha variable.

Ejemplo. En una fábrica la edad promedio de los operarios es de 44.8 años con una desviación estándar de 9.7 años. Al usar la desigualdad de Chebyshev para $k = 2$ se obtiene:

$$P[|X-\mu| \leq k\sigma] = \mu - k\sigma \leq X \leq \mu + k\sigma = \\ = 44.8 - 2(9.7) \leq X \leq 44.8 + 2(9.7) = \mathbf{25.4 \leq X \leq 64.2}$$

Es decir que al menos el 75 % de las edades de los operarios de la fábrica están entre 25.4 y 64.2 años, o lo que es lo mismo, la probabilidad de que un operario de la fábrica (elegido al azar) tenga entre 25.4 y 64.2 años es al menos de 0,75. Si se hubiese conocido la función de densidad correspondiente a los años de los operarios de la fábrica se hubiese podido calcular con exactitud el porcentaje de edades de los operarios que se encuentra a menos de dos desviaciones estándar del promedio.

Ejercicios

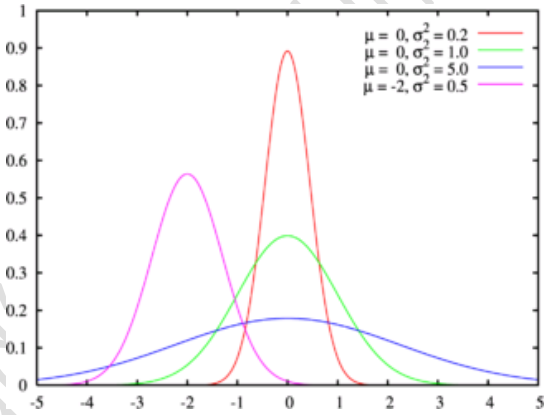
1. Una variable aleatoria continua X tiene un promedio de 9.2 ¿Qué valor máximo puede admitirse para el desvío estándar σ, si se desea que la variable se encuentre en el intervalo $(9.0\ ;\ 9.4)$ con una probabilidad de al menos $0{,}889$?
2. Sea X una variable aleatoria continua que representa el gasto familiar mensual familiar en una comunidad. Si el gasto promedio es de 5,800 pesos, con una desviación estándar de 360 pesos. ¿Cuál es el porcentaje de familias que tienen un gasto promedio mensual entre 5,080 y 6,520 pesos?

La distribución de probabilidad normal (o campana de Gauss)

Es una distribución de variable aleatoria continua. Este tipo de variables son aquellas que pueden tomar cualquier valor de entre todos los contenidos en el intervalo de una recta. El modelo probabilístico de este tipo de funciones es una variable aleatoria que corresponde a la llamada *función de densidad de probabilidades.* A pesar de que este tipo de funciones pueden ser de muy diferentes tipos, la forma más común es la acampanada.

La **distribución normal**, también llamada distribución de Gauss o distribución gaussiana, es la distribución de probabilidad que con más frecuencia aparece en estadística y teoría de probabilidades. Esto se debe a dos razones fundamentalmente:

- Su *función de densidad* es simétrica y con forma de campana, lo que favorece su aplicación como modelo a gran número de variables estadísticas.
- Es, además, límite de otras distribuciones y aparece relacionada con multitud de resultados ligados a la teoría de las probabilidades gracias a sus propiedades matemáticas.



La curva normal es una curva perfectamente simétrica, y unimodal basada en un número infinito de casos, por lo que sólo puede ser tratada de forma aproximada cuando se opera con datos reales. Por esta simétria, coinciden la media, la moda y la mediana.

La *función de densidad normal* es de la forma:

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

Donde μ es la media; es el centro de la distribución y el valor máximo y σ es la desviación estándar; entre mayor sea este parámetro la función será más aplastada, platicurtica.

Muchas variables aleatorias continuas presentan una función de densidad cuya gráfica tiene forma de campana. La importancia de la distribución normal se debe principalmente a que hay muchas variables asociadas a fenómenos naturales que siguen el modelo de la normal, como, por ejemplo

- Caracteres morfológicos de individuos como estatura, peso
- Caracteres sociológicos como consumo, ingreso, etc.
- Caracteres sociológicos como el consumo de un producto por un grupo de individuos
- Caracteres psicológicos como el cociente intelectual
- Valores estadísticos muestrales como la media y aproximaciones de la normal a otras distribuciones como la binomial y la Poisson

Teorema del límite central.

El teorema del límite central es uno de los resultados fundamentales de la estadística. Este teorema nos dice que, si una muestra es lo bastante grande, sea cual sea la distribución de la media muestral, seguirá aproximadamente una distribución normal. Es decir, dada cualquier variable aleatoria, si extraemos muestras de tamaño n $(n > 30)$ de una población con media finita μ y desviación estándar σ, entonces, si n es grande.

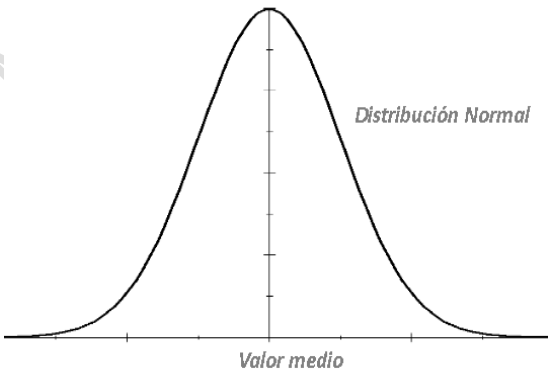
Para muestras, la media muestral $\bar{x}$ tiene una distribución aproximadamente normal con media μ y desviación estándar de la muestra de la distribución es $\frac{\sigma}{\sqrt{n}}$. La aproximación es mejor a medida que n crece.

La importancia del teorema del límite central explica el por qué algunas mediciones tienen una distribución aproximadamente normal y su más importante contribución; en inferencia estadística, muchos de los estimadores que se usan para hacer inferencias acerca de los parámetros de la población son sumas o promedios de mediciones muestrales. Cuando esto ocurre y el tamaño de la muestra es suficientemente grande, se espera que el estimador tenga una distribución aproximadamente normal.

La distribución de la media muestral de una población normal es una distribución normal con la misma media poblacional y con desviación estándar, sin importar el tamaño de la muestra. Si la distribución de la población es simétrica, la distribución de la media también se puede aproximar como una distribución normal, incluso para muestras pequeñas. Este hecho nos permite calcular probabilidades cuando tenemos una muestra de una variable con distribución normal y desviación estándar conocida.

Esta distribución de caracteriza porque los valores se distribuyen formando una **campana de Gauss**, en torno a un valor central que coincide con el valor medio de la distribución:

Un 50% de los valores están a la derecha de este valor central y otro 50% a la izquierda Esta distribución viene definida por **dos parámetros**:



$$\boldsymbol{X}: \boldsymbol{N}\ (\boldsymbol{\mu}, \boldsymbol{\sigma^2})$$

$\boldsymbol{\mu}$ es el valor medio de la distribución y es precisamente donde se sitúa el centro de la curva (de la campana de Gauss). $\boldsymbol{\sigma^2}$ Es la varianza. Indica si los valores están más o menos alejados del valor central: si la varianza es baja los valores están próximos a la media; si es alta, entonces los valores están muy dispersos.

Distribución normal estandarizada

Cuando la media de la distribución normal es 0 y la varianza 1, se denomina ***normal estandarizada***, y su ventaja reside en que hay tablas donde se recoge la probabilidad acumulada para cada punto de la curva de esta distribución. Además, toda distribución normal se puede transformar en una normal estandarizada:

Ejemplo: una variable aleatoria sigue el modelo de una distribución normal con media 10 y varianza 4. Transformarla en una normal estandarizada.

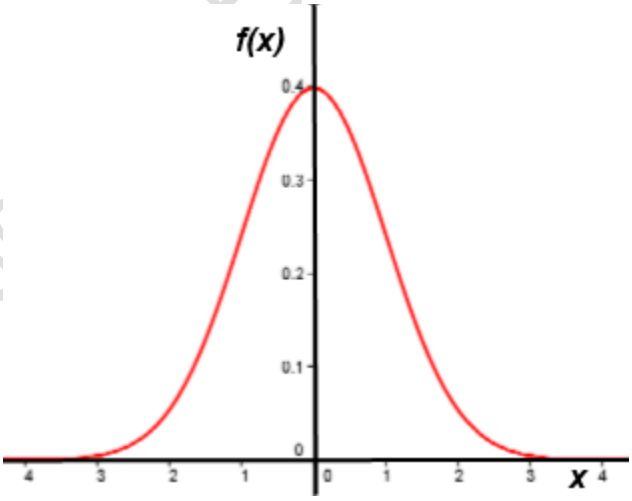
$$\boldsymbol{X \sim N\ (10, 4)}$$

Para transformarla en una normal estandarizada se crea una nueva variable Z que será igual a la anterior X menos su media y dividida por su desviación estándar (que es la raíz cuadrada de la varianza)

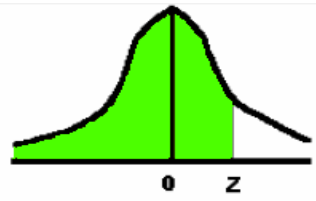
$$Z = \frac{X - \mu}{\sigma}, \text{ en la nueva variable es } Z = \frac{X - 10}{2}$$

Esta nueva variable se distribuye como una normal estandarizada, permitiéndonos, por tanto, conocer la probabilidad acumulada en cada valor.

$$\boldsymbol{Z \sim N\ (0, 1)}$$



La ***distribución normal estandarizada*** tiene la ventaja, como ya hemos indicado, de que las probabilidades para cada valor de la curva se obtienen de tablas, como la siguiente



Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54395	.54776	.55172	.55567	.55962	.56356	.56750	.57124	.57534
0.2	.57926	.58617	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61781	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79373	.79955	.80234	.80510	.80785	.81057	.81327
0.9	.81594	.81859	.82124	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87923	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89616	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91308	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96079	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96637	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98299	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99001	.99036	.99061	.99086	.99110	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99491	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99597	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99830	.99836	.99841	.99846	.99851	.99856	.99860
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99897	.99900
3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
3.4	.99956	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99996	.99997	.99997

La columna de la izquierda indica el valor cuya probabilidad acumulada queremos conocer. La primera fila nos indica el segundo decimal del valor que estamos consultando.

Ejemplo: Para obtener la probabilidad acumulada en el valor 2.75. Entonces buscamos en la columna de la izquierda el valor 2.7 y en la primera fila el valor 0.05. La casilla en la que se cruzan es su probabilidad acumulada; 0.99702, o bien 99.7%.

El valor que obtenemos de la tabla es la probabilidad acumulada; es decir, la que va desde el inicio de la curva por la izquierda hasta dicho valor. No nos da la probabilidad en ese punto. En una distribución continua en el que la variable puede tomar infinitos valores, la probabilidad en un punto concreto es prácticamente despreciable.

***Ejemplos**:*

a) Probabilidad acumulada en el valor 0.67: la respuesta es 0.7486
b) Probabilidad acumulada en el valor 1.35: la respuesta es 0.9115
c) Probabilidad acumulada en el valor 2.19: la respuesta es 0.98574

d) El salario medio de los empleados de una empresa se distribuye según una distribución normal, con promedio de 5500 pesos y desviación estándar 1250. Calcular el porcentaje de empleados con un sueldo inferior a 8000 pesos.

Lo primero que haremos es transformar esa distribución en una normal estandarizada, para ello se crea una nueva variable Z que será igual a la anterior X menos su media y dividida por la desviación estándar

$$P(X < 8000) = P\left(Z < \frac{X-\mu}{\sigma}\right) = P\left(\frac{X-5500}{1250}\right)$$

Esta nueva variable se distribuye como una normal estandarizada. La variable Z que corresponde a una variable X de valor 8000 es:

$$P\left(Z < \frac{8000-5500}{1250}\right) = P\left(Z < \frac{2500}{1250}\right) = P(Z < 2)$$

Ya podemos consultar en la tabla la probabilidad acumulada para el valor 2 (equivalente a la probabilidad de sueldos inferiores a 8000 pesos). Esta probabilidad es 0.97725

Por lo tanto, el porcentaje de empleados con salarios inferiores a 8000 pesos es del 97.725%.

e) La renta media de los habitantes de un país es de 6000 pesos/mes, con una desviación estándar de 3545.7. Se supone que se distribuye según una distribución normal. Calcular:

i. Porcentaje de la población con una renta inferior a 3000 pesos.
ii. Renta a partir de la cual se sitúa el 10% de la población con mayores ingresos.
iii. Ingresos mínimo y máximo que engloba al 60% de la población con renta media.

***i.** Porcentaje de la población con una renta inferior a 3000 pesos por mes.*

Lo primero es calcular la normal estandarizada:

$$P\left(Z < \frac{3000-6000}{3545.7}\right) = P(Z < -0.8461)$$

El valor de Z equivalente a 3000 pesos es -0.8461.

$$P(X < 3000) = P(Z < -0.8461)$$

Ahora tenemos que ver cuál es la probabilidad acumulada hasta ese valor. En algunas tablas no se incluyen los valores negativos, ya que la distribución normal es simétrica respecto al valor medio. Si este fuera el caso, tendríamos que proceder así,

$$P\,(Z\ <\ -0.8461)\ =\ P\,(Z\ >\ 0.8461)$$

Por otra parte, la probabilidad que hay a partir de un valor es igual a 1 (100%) menos la probabilidad acumulada hasta dicho valor:

$$P(Z > 0.8461) = 1 - P(Z < 0.8461) = 1 - 0.8012\ (aprox.) = 0.1988$$

Luego, el 19.88% de la población tiene una renta inferior a 3000 pesos mensuales[3].

ii. *Nivel de ingresos a partir del cual se sitúa el 10% de la población con renta más elevada.*

Vemos en la tabla el valor de la variable estandarizada cuya probabilidad acumulada es el 0.9 (90%), lo que quiere decir que por encima se sitúa el 10% superior.

Ese valor corresponde a $Z = 1.281$ (aprox.). Ahora calculamos la variable normal X equivalente a ese valor de la normal tipificada:

$$1.281 = \frac{X - 6000}{3545.7}$$

Despejando X, su valor es $10{,}538.5$. Por lo tanto, aquellas personas con ingresos superiores a $\$\ 10{,}538.5$ pesos mensuales son el 10% de la población con renta más elevada.

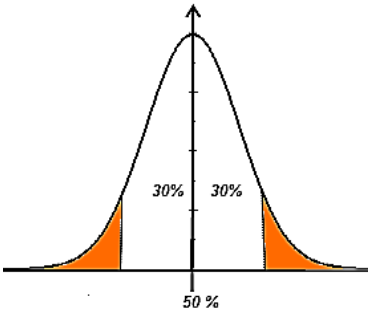
iii. *Nivel de ingresos mínimo y máximo que engloba al 60% de la población alrededor de la media*

Como sabemos que, hasta la media, la probabilidad acumulada es del 50%, quiere decir que entre la media y este valor de Z hay un 30% de probabilidad adicional. De esta manera, el valor en la tabla que nos interesa es de la variable normalizada Z cuya probabilidad acumulada es el $0.5 + .30$ (80%) del lado derecho de la distribución.

[3] Para encontrar el valor de una variable, esta entre dos valores de la tabla, normalmente cuando tenemos más de dos cifras decimales, podemos utilizar la fórmula de interpolación, $y = \frac{x - x_1}{x_2 - x_1}(y_2 - y_1) + y_1$
En el ejemplo anterior $y = \frac{0.8461 - 0.84}{0.85 - 0.84}(0.80234 - 0.79955) + 0.79955 = 0.8012$

Por otra parte, al ser la distribución normal simétrica, entre $-Z$ y la media hay otro 30% de probabilidad del lado izquierdo. En definitiva, el segmento $(-Z, Z)$ engloba al 60% de población con renta media.

El valor de Z que acumula el 80% de la probabilidad es 0.8416 (aprox.), por lo que el segmento viene definido por (-0.8416, +0.8416). Ahora calculamos los valores de la variable X correspondientes a estos valores de Z.



$$0.8416 = \frac{X - 6000}{3545.7} \quad y \quad -0.8416 = \frac{X - 6000}{3545.7}$$

Los valores de X son 3016.0 y 8984.1. Por lo tanto, las personas con ingresos superiores a \$3,016.0 e inferiores a \$8,984.1 pesos constituyen el 60% de la población con un nivel medio de renta.

f) La esperanza de vida de los habitantes de un país es de 68 años, con una varianza de 25. Se hace un estudio en una pequeña ciudad de 10,000 habitantes. ¿Cuántas personas superarán previsiblemente los 75 años? ¿Cuántos vivirán menos de 60 años?

Solución. Las personas que vivirán (previsiblemente) más de 75 años. Calculamos el valor de la normal tipificada equivalente a 75 años. Por lo tanto

$$Z = \frac{75 - 68}{\sqrt{25}} = 1.4 \quad P(X > 75) = P(Z > 1.4) = 1 - P(Z < 1.4) = 1 - 0.9192 = 0.0808$$

Luego, el 8.08% de la población (808 habitantes) vivirán más de 75 años.

Para obtener el número de personas que vivirán (previsiblemente) menos de 60 años. Calculamos el valor de la normal estándar equivalente a 60 años.

$$Z = \frac{60-68}{\sqrt{25}} = -1.6 \quad \text{Por lo tanto} \quad P(X < 60) = P(Z < -1{,}6) = 0.0548$$

Si utilizamos tablas que no incluyen los valores negativos tendríamos que calcular esta probabilidad así,

$$P(Z > 1.6) = 1 - P(Y < 1.6) = 0{,}0548$$

Luego, el 5.48% de la población (548 habitantes) no llegarán probablemente a esta edad.

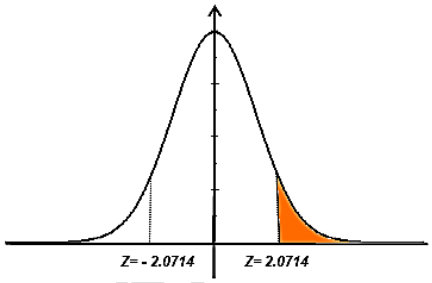
g) El rendimiento promedio al vencimiento de los bonos industriales emitidos durante un trimestre fue de 8.55 con una desviación estándar $\sigma = 0.70$. Suponiendo que el rendimiento de los bonos de una empresa fue de 7.10, obtenga el porcentaje correspondiente a este nivel.

Solución. Primero encontramos el valor de la variable estandarizada

$$Z = \frac{7.10 - 8.55}{0.7} = -2.0714$$

Así, entonces la probabilidad de Z es igual a

$$P(Z < -2.0714) = 1 - P(Z > 2.0714)$$
$$= 1 - 0.98077 = 0.01923$$



Es decir que los bonos de la empresa tuvieron un rendimiento de 1.92%, lo que nos indica que la situación financiera de la empresa es dudosa.

h) El ingreso anual promedio de los habitantes de una comunidad es de $36,200 con una desviación estándar de $4,000 (en miles de pesos).

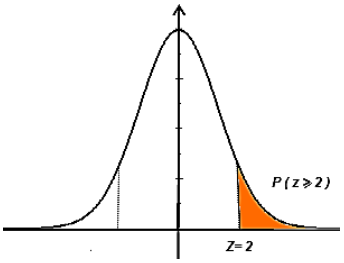
I. Para una muestra de 64 personas, calcule la probabilidad de que el ingreso anual promedio en la muestra exceda a $37,200.

Por el teorema del límite central tenemos:

$$\mu = 36200 \ y \ \sigma = 4000 \ \ entonces \ \ \overline{y} = 36200$$

En la muestra la desviación estándar $s = \frac{\sigma}{\sqrt{n}}$

$$s = {}^{4000}/_{\sqrt{64}} = \frac{4000}{8} = 500$$



Por lo tanto

$$Z = \frac{37200 - 36200}{500} = \frac{1000}{500} = 2$$

$$P(Z \geq 2) = 1 - P(Z \leq 2) = 1 - .97725 = 0.02275$$

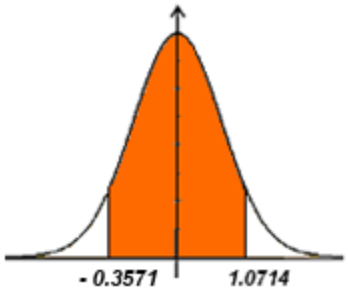
II. Si se toma otra muestra independiente de 64 personas, encuentre la probabilidad de que las dos medias muestrales excedan a $ 37,200

$$0.02275 * \ 0.02275 = 0.000517$$

Ejercicio. Una empresa de renta de autos posee un parque de 500 autos de alquiler en la ciudad de México. Estudios de la empresa indican que el recorrido que realizan los autos cada día se distribuye como una variable aleatoria con media $\mu = 300$ y desviación estándar $\sigma = 28$. ¿Cuál es la probabilidad de que un auto cualquiera recorra? 1) entre 290 y 330 km diarios 2) no más de 310 km por día y 3) al menos 349 km diarios.

Solución,

1) En el primer caso, probabilidad que nos solicitan en forma gráfica es la siguiente;



$$\mathrm{P}(290 \le \mathrm{x} \le 330) = \mathrm{P}\left(\frac{290-300}{28} \le \mathrm{z} \le \frac{330-300}{28}\right)$$

$$= \mathrm{P}(-0.3571 \le \mathrm{z} \le 1.0714)$$

Para encontrar esta área, tendríamos

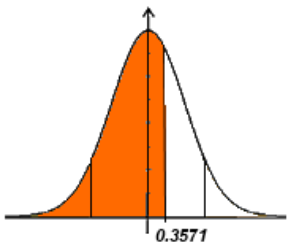
$$= P(z \le 1.0714) - P(z \le -0.3571) = P(z \le 1.0714) - (1 - P(z \le 0.3571))$$

$$= 0.8577 - (1 - 0.6395) = 0.4972$$

La tabla de la distribución normal que utilizamos proporciona áreas del lado izquierdo al valor de z, por esta razón el cálculo obliga a restar uno para obtener el valor que nos interesa. El valor de la probabilidad de $P(z \le 0.3571) = 0.6394$ se obtuvo por interpolación ($z \le 0.3571$).

Es decir, $\left(\frac{0.3571-0.35}{0.36-0.35}\right)(0.64058 - 0.6368) + 0.63683 = 0.6395.$

2) No más de 310 km por día



$$\mathrm{P}(\mathrm{x} \le 310) = \mathrm{P}\left(\mathrm{z} \le \frac{310-300}{28}\right)$$

$$= \mathrm{P}(\mathrm{z} \le 0.3571) = 0.6395 \ \text{(por interpolación)}$$

3) Al menos 349 km al día.

$$P(x \geq 349) = P\left(z \geq \frac{349 - 300}{28}\right) = P(z \geq 1.75)$$
$$= 1 - P(z \leq 1.75)$$
$$= 1 - 0.9599 = 0.0401$$

Aproximación de la Normal a la binomial.

Cuando "n" es grande, la distribución binomial resulta laboriosa y complicada, por lo que el matemático Abraham de Moivre (1667-1754), demostró que cuando se dan ciertas condiciones una distribución Binomial se puede aproximar a una distribución Normal de parámetros,

$$\boldsymbol{\mu = np} \text{ y desviación estándar } \boldsymbol{\sigma = \sqrt{npq}}$$

Sabemos que la variable aleatoria binomial es el número de éxitos que tienen lugar cuando se realizan **n** repeticiones independientes de un experimento con pruebas Bernoulli. La variable aleatoria x puede escribirse como la suma de n variables aleatorias de Bernoulli:

$$x = x_1 + x_2 + \cdots + x_n \qquad E(x) = np \quad Var(x) = npq$$

Si x es una variable aleatoria binomial, $b(n, p)$ con media $E(x) = np$ y desviación estándar $\sqrt{npq}$, entonces, cuando $n \to \infty$ la variable aleatoria binomial tiende a una distribución normal estandarizada donde:

$$Z = \frac{x - np}{\sqrt{npq}} \sim N(0,1)$$

En la práctica, para aproximar la binomial con una normal requiere se recomienda que:

$$np > 15 \quad y \quad p \leq \frac{1}{2}$$

$$nq > 20 \quad y \quad p \leq \frac{1}{2}$$

Ejemplo. Se lanza una moneda al aire 100 veces, si sale cara le damos el valor 1 y si sale cruz el valor 0. Cada lanzamiento es una variable independiente que se distribuye según el modelo de Bernoulli, con media 0,5 y varianza 0,25.

Calcular la probabilidad de que en estos 100 lanzamientos obtener más de 60 caras. La variable suma de estas 100 variables independientes se distribuye, por tanto, según una distribución normal.

$$Media = \mu = np = 100 * 0.5 = 50$$
$$Varianza = \sigma^2 = npq = 100 * .5 * .5 = 25$$
$$\sigma = \sqrt{25} = 5$$

Para ver la probabilidad de que salgan más de 60 caras calculamos la variable normal tipificada equivalente:

$$Z = \frac{60 - 50}{5} = 2$$

$$P(X > 60) = P(Z > 2.0) = 1 - P(Z < 2.0) = 1 - 0.9772 = 0.0228$$

Es decir, la probabilidad de que al tirar 100 veces la moneda conseguir más de 60 caras es tan sólo del 2.28%

Ejemplo. Un investigador agrícola estima que la probabilidad de que un grano de maíz no germine después de tres años de conservación es del 70%. Se toma una muestra de 100 granos almacenados por tres años. ¿Cuál es la probabilidad de que menos de 25 germinen?

Solución.
Definimos a X como la v.a, *número de granos que germinan en la muestra de 100 granos.*
La probabilidad de que un grano germine es $p = 0.3$ si la muestra es independiente.

Solución binomial, decimos que la variable aleatoria sigue una distribución, $B(25{,}100{,}0.3)$ y buscamos la probabilidad de

$$P(X < 25) = P(X = 0) + P(X = 1) + \cdots + P(X = 24) = 0.1136$$

Obtenido en Excel, utilizando la función DISTR.BINOM.N(24,100,0.3,VERDADERO)

El cálculo es laborioso, quizá por tablas de función binomial acumulada; o bien ajustar una distribución normal. Esta opción será posible si los productos np y nq son suficientemente grandes, en nuestro caso toman los valores de

$$np = 100(.3) = 30 \quad y \quad nq = 100(.7) = 70$$

Así

$$\mu = np = 30 \quad y \quad \sigma = \sqrt{npq} = \sqrt{100 * 0.3 * 0.7} = 4.58$$

Entonces, el resultado que buscamos deberá calcularse para $P(X < 25) = P(X \leq 24)$.

$$P(X < 25) = P(X \leq 24) = P\left(z \leq \frac{24 - 30}{4.582}\right)$$
$$P(z \leq -1.31) = 1 - P(z \leq 1.31) = 1 - 0.90490 = 0.0951$$

Corrección de continuidad

Ya hemos visto que, para valores específicos de una variable continua, la probabilidad es cero; sabemos que el área bajo la curva en un punto es cero. Así, Cuando aproximamos una distribución binomial mediante una normal, estamos convirtiendo una variable discreta en una distribución continua; y por tanto la probabilidad sería cero. Para evitar este problema de la aproximación de los valores fijos, se corrigen por medio de la corrección de Yates[4]. Básicamente, sustituimos los puntos extremos por intervalo centrado en el punto y de amplitud unitaria.

De esta manera, si la variable aleatoria $X \sim B(n,p)$, entonces x es una variable discreta y la aproximación de la normal, con la corrección de continuidad sería,

$$P(X \leq x) = P(X < x + 0.5)$$
$$P(X \geq x) = P(X > x - 0.5)$$

Finalmente

$$P(x_1 \leq X \leq x_2) = P(x_1 - 0.5 < X < x_2 + 0.5)$$

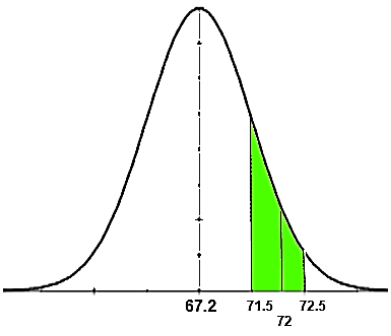
Ejemplo. Para el ejercicio anterior sabemos que la variable aleatoria se distribuye como una binomial $b(100, 0.3)$. Obtener la probabilidad de $P(X < 25)$.

Con el factor de corrección,

$$P(X \leq 24.5) = P\left(X \leq \frac{24.5 - 100 * 0.3}{\sqrt{100 * .3 * .7}}\right) = P\left(Z \leq \frac{24.5 - 30}{4.582}\right)$$

$$P(Z \leq 1.2003) = 1 - P(Z \leq 1.2003) = 1 - 0.88493 = 0.115$$

Ejercicio. El 45% de la cámara de diputados y senadores tiene un grado académico superior al de la licenciatura. ¿Cuál es la probabilidad de que, si seleccionamos a 150 de ellos, 72 tengan un grado superior al de licenciatura?



Solución.

Probamos si es posible aproximar con la normal.

$$np = 150(.45) = 67.5 \qquad nq = 150(.55) = 82.5$$

Aplica utilizar la normal como una aproximación de la binomial.

$$\mu = np = 67.5 \quad y \quad \sigma = \sqrt{npq} = \sqrt{150 * 0.45 * 0.55} = 6.09$$

[4] Frank Yates, estadístico inglés 1902-1994, nació en Manchester el 12 de mayo de 1902.

Resolvemos con la corrección de continuidad

$$P(71.5 \leq X \leq 72.5) = P\left(\frac{71.5 - 67.5}{6.09} < Z < \frac{72.5 - 67.5}{6.09}\right)$$

$$P(0.66 < Z < 0.82) = 0.2930 - 0.2454 = 0.0485$$

Si utilizamos la función de Excel, citada antes tendríamos,

$$DISTR.BINOM.N(72{,}150{,}0.45, FALSO) = 0.0496$$

Aproximación de la normal a la distribución Poisson

En el caso de la distribución de Poisson, la variable aleatoria nos establece el número de veces que ocurre un suceso en un determinado intervalo de tiempo, sabemos que la media y la varianza de esta distribución coincide con el parámetro λ.

Si el número de ocurrencias esperadas λ es grande y el intervalo de tiempo se divide en subintervalos de idéntica longitud. En ese caso, el número total de ocurrencias es la suma de las ocurrencias de cada subintervalos, y puede verse como la suma de un número moderadamente grande de variables aleatorias, cada una de las cuales representa el número de ocurrencias en un subintervalos del periodo de tiempo, puede utilizarse la distribución normal como una aproximación a la distribución de Poisson. En la práctica la aproximación es aceptable sí $\lambda \geq 10$.

El procedimiento práctico es análogo al caso de la binomial, así pues, si tenemos una variable aleatoria x que se distribuye según una distribución de Poisson de parámetro λ, entonces cuando $\lambda \geq 10$ la variable aleatoria:

$$Z = \frac{x - E(x)}{\sqrt{Var(x)}} = \frac{x - \lambda}{\sqrt{\lambda}} \sim N(0,1)$$

Cuando se utiliza la aproximación de la Normal a la Poisson, es importante aplicar la corrección de continuidad

Ejemplo. **Si la variable aleatoria $X \sim Poisson(36)$. Estimar la probabilidad de de $P(31 \leq X \leq 39)$.**

$$P(31 \leq X \leq 39) = P(30.5 < X < 39.5)$$

$$P\left(\frac{30.5 - 36}{\sqrt{36}} < \frac{X - 36}{\sqrt{36}} < \frac{39.5 - 36}{\sqrt{36}}\right) = P(-0.916 < Z < 0.583)$$

$$= P(Z < 0.583) - \big(1 - P(Z < .916)\big) = 0.719 - (1 - .8176)$$

$$0.719 - 0.1823 = 0.5366$$

Si utilizamos la distribución Poisson directamente, la probabilidad buscada sería,

$$P(31 \leq X \leq 39) = \sum_{x=31}^{x=39} \frac{36^x e^{-36}}{x!} = 0.5456$$

Obtenida por cálculos en Excel.

Ejercicio. Un banco recibe en promedio 6 cheques falsos al día, suponiendo que el número de cheques falsos sigue una distribución de Poisson, hallar:
a) Probabilidad de que se reciban cuatro cheques falsos en un día.
b) Probabilidad de que se reciban más de 30 cheques falsos en una semana

Solución.
La variable aleatoria x se distribuye como una Poisson, con media $\mu = 6$. La $v.a.X$ representan los cheques falsos por día.

$a)$ $P(X = 4) = \dfrac{e^{-6}6^4}{4!} = 0.1339$

b) Vamos a utilizar la aproximación de la Normal a la Poisson. Como el intervalo de tiempo es de 7 días $\mu = 6 * 7 = 42$

$$P(x > 30) = P(z > \frac{30 - 42}{\sqrt{42}} = P(z > -1.85)$$
$$= P(z < 1.85) = 0.96784$$

Si realizamos este último cálculo, utilizando Excel tendríamos,

$$= 1 - POISSON.DIST(30{,}42, VERDADERO) = 0.9670$$

Ejercicio. El 6% de los cítricos que produce una organización agrícola son de baja calidad; después de revisar el tamaño de la fruta, el espesor de la cascara, etc. Se empaquetan 100 productos en una caja para sus distribución o venta. ¿Cuál es la probabilidad de que en una caja haya entre 8 y 10 cítricos de baja calidad?

Solución:
Sea $X =$ "número de cítricos de baja calidad en una caja"

Esta variable aleatoria X se distribuye como una binomial ya que un cítrico puede ser de baja calidad o no, con parámetros $b(r = 8{,}10|\, n = 100, p = 0.06)$, además

$$\mu = np = 100(0.06) = 6 \quad y \quad \sigma = \sqrt{npq} = \sqrt{100(.06).0.95)} = 5.64$$

Podemos aproximar la normal a la binomial ya que, $p \leq 0.5 \;\; y \;\; np = 6 > 5$

Por lo tanto $b(r = 8{,}10|\, n = 100, p = 0.06) \backsim N(\mu = 6, \sigma = 5.64)$

$P(8 \leq X \leq 10)$ y por continuidad transformamos a $P(7.5 \leq X \leq 10.5)$, Así,

$$P\left[\frac{7.5-6}{5.68} \leq \frac{x-6}{5.68} \leq \frac{10.5-6}{5.68}\right] = P[0.26 \leq z \leq 0.79]$$

Ejercicios

1. *Un proceso de fabricación produce 10% de artículos defectuosos. Si se toma muestra de 100 unidades al azar y x es la variable aleatoria para el número de artículos defectuosos de la muestra. Utilizar la aproximación de la normal a la binomial para calcular las probabilidades*
   a. $P(x \leq 20)$ $b)\, P(x = 20)$ $c)\, P(10 \leq x \leq 15)$ $d(x > 20)$
2.

Capitulo V.

Muestreo.

Muestreo

Como ya hemos establecido con anterioridad, el principal objetivo de la estadística es hacer inferencias de una población a partir de los datos de una muestra. El muestreo, es el proceso por el cual se define la muestra de una población en estudio, y es el que nos proporciona las consideraciones a tomar de dicha muestra, es decir, cuál será el tamaño de la muestra y como se genera el conjunto de elementos o datos muestrales.

Partiendo de que objeto de nuestro interés es la población. Normalmente, trabajar con la población es muy costoso y el tiempo que requiere recoger los datos puede resultar muy largo. Quizá por esta razón los censos de población se realizan cada 10 años.

Supongamos que nos interesa conocer los hábitos de lectura de un grupo de 30 estudiantes. En este caso no tenemos problemas para trabajar con la población, ya que contamos solamente con 30 personas. Sin embargo, las inferencias sobre sus hábitos estarán solo limitadas a este grupo de estudiantes. Si requerimos un estudio más amplio, por ejemplo, para unos 300,000 estudiantes, entonces realizar el estudio en una población tan grande va a requerir que tomemos más tiempo y el costo se eleva considerablemente. En estos casos es que el muestreo puede ser más práctico y económico.

Por otro lado, si al obtener la muestra de la población solo la tomamos de los estudiantes del área de matemáticas, o si la muestra solo incluye a estudiantes de escuelas particulares, etc. En estos casos, los resultados no van a ser muy fiables ya que no incluirán los hábitos de lectura de estudiantes de otras disciplinas, condiciones económicas, sexo, etc.

Así, resulta que no todas las muestras que se pueden extraer de una población son útiles desde el punto de vista estadístico. Ahora comenzaremos el estudio, muy descriptivo y resumido, de diferentes técnicas de muestreo que permiten evitar (o al menos reducir) el impacto sobre el resultado final de los errores que acabamos de mencionar.

Pasos para seleccionar una muestra:

1. Definir el objetivo del estudio.
2. Definir la población objetivo.
3. Seleccionar un procedimiento.
4. Definir el tamaño de la muestra.
5. Seleccionar las unidades muéstrales.

Existen varias clases de muestreo, y se usan ó no de acuerdo con el tipo de estudio, objetivo del estudio, la forma de las variables, características de la población, incluso hasta la cantidad de recursos disponibles. Sin embargo, solo tenemos dos métodos para seleccionar muestras de poblaciones: el muestreo determinista, no aleatorio, y el muestreo aleatorio (que incorpora el azar como recurso en el proceso de selección). Cuando este último cumple con la condición de que todos los elementos de la población tienen alguna oportunidad de

ser escogidos en la muestra, si la probabilidad correspondiente a cada sujeto de la población es conocida de antemano, recibe el nombre de *muestreo probabilístico*. Una muestra obtenida por muestreo determinista puede basarse en la experiencia de alguien con la población. Algunas veces una muestra determinista se usa como guía o muestra tentativa para decidir cómo tomar una muestra aleatoria más adelante.

Muestreo probabilístico

Forman parte de este tipo de muestreo todos aquellos métodos para los que puede calcularse la probabilidad de extracción de cualquiera de las muestras posibles. Este conjunto de técnicas de muestreo es el más aconsejable, aunque en ocasiones no es posible optar por él. En este caso se habla de muestras probabilísticas, pues no es en rigor correcto hablar de *muestras representativas* dado que, al no conocer las características de la población, no es posible tener certeza de que tal característica se haya conseguido.

La herramienta del muestreo más importante para recoger los datos de la muestra es la encuesta. Algunas definiciones son necesarias,

a) El diseño de muestreo o *diseño de encuesta* especifica el método de obtención de la muestra.
b) Un *elemento* es un objeto del cual se toma una medición. Por ejemplo, en el caso de la encuesta de ingreso-gasto de los hogares, el elemento es la familia,
c) Las *unidades muestrales* son grupos de elementos de una población.

Para obtener una muestra aleatoria, de unidades muestrales, es necesario contar con el ***marco muestral***, que es una lista de unidades muestrales.

Entonces, la primera acción para realizar una encuesta por muestreo consiste en la identificación de las unidades de muestreo y la lista que contiene estas unidades.

Errores en el muestreo.

Las fuentes de error más comunes en el muestreo son:

a) Variación aleatoria. Por ejemplo, para determinar el ingreso de los hogares de una comunidad, podemos seleccionar solamente a los estratos más altos, o más aún los niveles de ingreso medios que pueden pasar desapercibido y producir inferencias erróneas.
b) Especificación errónea o deficiente de la población. Este tipo de error lo produce generalmente un marco muestral erróneo. Por ejemplo, en una encuesta de opinión sobre las elecciones podemos incluir en el marco muestral a personas que no van a votar el día de la elección, o como es una moda actual realizar las encuestas por

teléfono, lo cual impide que participen electores que no tienen teléfono, o no lo pagaron, etc.

c) La no respuesta. Es común suponer que los elementos de la muestra que responden tienen comportamientos similares a los que no responden. Es común en las encuestas de opinión que los que no responden son quienes prefieren que las cosas se queden como están.

Cuando se trabaja con una muestra aleatoria se deben tener en cuenta dos aspectos principales:

- El método de selección
- El tamaño de la muestra

Muestreo simple aleatorio

Es la extracción de una muestra de una población finita, en el que el proceso de extracción es tal que garantiza que cada uno de los elementos de la población tiene la misma oportunidad de ser elegido en la muestra. Nótese que esto no es lo mismo que decir que todos los elementos de la población tienen igual probabilidad de ser elegidos. Esta condición garantiza sin embargo la equiprobabilidad. Muchas de las estimaciones resultantes de un muestreo de este tipo se dicen *insesgadas*. Esto significa en el caso particular (por ejemplo) de un porcentaje, lo siguiente: si en la población un determinado porcentaje de individuos presenta la característica A, la extracción aleatoria garantiza matemáticamente que, por término medio, se obtendrá el mismo porcentaje de datos muestrales con esa característica.

El muestreo aleatorio simple puede ser de tres tipos:

Sin reposición de los elementos**,** cada elemento extraído se descarta para la subsiguiente extracción. Por ejemplo, si se extrae una muestra de una "población" de bombillas para estimar la vida media de las bombillas que la integran, no será posible medir más que una vez la bombilla seleccionada.

Con reposición de los elementos las observaciones se realizan con reemplazamiento de los individuos, de forma que la población es idéntica en todas las extracciones. En poblaciones muy grandes, la probabilidad de repetir una extracción es tan pequeña que el muestreo puede considerarse sin reposición, aunque, realmente, no lo sea.

Con reposición múltiple En poblaciones muy grandes, la probabilidad de repetir una extracción es tan pequeña que el muestreo puede considerarse sin reposición. Cada elemento extraído se descarta para la subsiguiente extracción.

El primer paso en el muestreo simple aleatorio es encontrar el tamaño de la muestra y posteriormente enumerar todos los individuos de la población. Una vez cubiertos estos pasos procedemos a seleccionar la muestra. Hay varias maneras de hacerlo, escribir números en papelitos y meterlos en una urna, posteriormente seleccionamos el número de papelitos que correspondan al tamaño de la muestra y al relacionar estos con la lista nos permitirá obtener la muestra correspondiente. El método más utilizado es utilizar las llamadas tablas de números aleatorios, que representan el procedimiento más habitual para obtener una muestra aleatoria simple dentro de una población finita, estos números pueden ser generados también por medios electrónicos, incluso si no tenemos algo más con los números del directorio telefónico.

Procedimiento de selección de la muestra.

1) Definir la población de estudio.
2) Asignar un número a cada individuo de la población
3) Determinar el tamaño de muestra óptimo o para el estudio.
4) Seleccionar la(s) muestra(s) de manera sistemática por medio de algún medio mecánico (Tablas de números aleatorios, bolas dentro de una bolsa, números aleatorios generados con una calculadora, etc.)
5) Y se eligen tantos individuos como sea necesario para completar el tamaño de muestra que necesitamos.

Para calcular la media poblacional μ y el total poblacional τ, basados en un muestreo aleatorio simple, utilizamos las siguientes formulas;

El cuadro siguiente ilustra una tabla de números aleatorios. La tabla completa se muestra el final.

TABLA DE NUMEROS ALEATORIOS

17841	**49597**	**92623**	**80005**	**11177**	**15145**	**46379**
84970	**47043**	**64048**	**06993**	**17369**	**70932**	**47950**
30524	**27250**	**73072**	**52654**	**33653**	**30422**	**22347**
56211	**27219**	**44652**	**09467**	**62848**	**82479**	**35068**
66110	**69181**	**13200**	**93239**	**25591**	**21248**	**06881**
…..	**…..**	**…..**	**…..**	**……**	**……**	**…….**
…..	**……**	**……**	**……**	**…….**	**……**	**…….**
72208	**67425**	**77273**	**35454**	**43798**	**89958**	**98485**
62663	**32726**	**14266**	**48467**	**36706**	**90411**	**84898**
99530	**11547**	**35629**	**86192**	**25909**	**97084**	**30951**
36626	**80491**	**21369**	**48285**	**59708**	**44408**	**75096**

Fuente. Elaboración propia

Ejemplo. En una compañía con 750 trabajadores se quiere obtener una muestra aleatoria de 15 elementos para un chequeo médico. Los trabajadores fueron numerados del 1 al 750

y mediante una tabla de números aleatorios se procedió a seleccionarlos. El punto de arranque en la tabla se fijó mediante la hora en ese momento, 2:04, por lo tanto, se inició en la columna 3, renglón 4. Como los números de los trabajadores van desde 1 hasta 750 solo se toman en cuenta las primeras 3 cifras de cada número que se encuentren en ese rango. En seguida se muestra una parte de la tabla, con el primer y segundo seleccionado:

TABLA DE NUMEROS ALEATORIOS

17841	49597	92623	80005	11177	15145	46379
84970	47043	64048	06993	17369	70932	47950
30524	27250	73072	52654	33653	30422	22347
56211	**272**19	**446**52	**094**67	**628**48	82479	**350**68
66110	**691**81	**132**00	93239	**255**91	**212**48	**068**81
28710	**524**14	**558**93	**256**32	64856	51745	46855
38939	15777	66270	53052	05160	94786	81987
31297	00722	88300	21109	13124	96742	64968
34043	19959	77949	24510	93510	40492	81113
74996	32698	29430	58603	43879	7861	15870
36626	80491	21369	48285	59708	44408	75096

Fuente. Elaboración propia

Es decir, la muestra estaría integrada con los trabajadores, 272, 446, 094, 628, 350, 661, 691, 132, 255, 212, 068, 287, 524, 558 y 256.

Estimador de la media poblacional μ, para un muestreo aleatorio simple

$$\mu = \overline{y} = \sum_{i=1}^{n} \frac{y_i}{n}$$

Varianza estimada del estimador $\overline{y}$

$$\hat{\sigma}^2_{\overline{y}} = \left(\frac{s^2}{n}\right)\left(\frac{N-n}{N}\right) \quad \text{donde} \quad s^2 = \sum_{i=1}^{n} \frac{(y_i - \overline{y})^2}{n-1} = \frac{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}{n-1}$$

Límite para el error de estimación. $\overline{y} \pm 2\hat{\sigma}_{\overline{y}}$

A. Estimador del total poblacional τ, para un muestreo aleatorio simple

$$\hat{\tau} = N\overline{y} = \frac{N\sum_{i=1}^{n} y_i}{n}$$

Varianza estimada de $\hat{\tau}$

$$\hat{\sigma}^2_{\hat{\tau}} = N^2\hat{\sigma}^2_{\overline{y}} = N^2\left(\frac{s^2}{n}\right)\left(\frac{N-n}{N}\right)$$

Límite para el error de estimación

$$N\overline{y} \pm 2\hat{\sigma}_{\hat{\tau}}$$

Ejemplo. En la tabla aparecen los saldos correspondientes a las cuentas de una muestra de un laboratorio de análisis clínico de un hospital de tamaño $n = 8$. El laboratorio tiene un total de 1000 cuentas por cobrar

a) Estime el saldo promedio para las N cuentas y establezca una cota para el error de estimación.
b) Estime el total τ de los saldos de todas las cuentas y establezca una cota para el error de estimación.

No. Elemento	Adeudo y_i	y_i^2
1	30.2	912.04
2	14.5	210.25
3	33.5	1122.25
4	32	1024
5	17.5	306.25
6	10	100
7	23.4	547.56
8	27.5	756.25
Suma	188.6	4978.6

a) $\overline{y} = \frac{188.6}{8} = \$\ 23.58$

La cota del error $s^2 = \frac{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}{n-1} = \frac{\left(4978.6 - \frac{188.6^2}{8}\right)}{7} = \frac{532.36}{7} = \$\ 76.05$

La varianza estimada es entonces $\hat{\sigma}_{\overline{y}}^2 = \left(\frac{s^2}{n}\right)\left(\frac{N-n}{N}\right) = \left(\frac{76.05}{8}\right)\left(\frac{1000-8}{1000}\right) = 9.43$

Estimación del saldo promedio μ, y una cota de error; $23.58 \pm 2\sqrt{9.43} = \$23.58 \pm 6.14$

b) El total τ de los saldos de todas las cuentas $\hat{\tau} = N\overline{y} = 1000(23.575) = 23575$

Dado que la varianza estimada de $\hat{\tau}$ es $\hat{\sigma}_{\hat{\tau}}^2 = N^2\hat{\sigma}_{\overline{y}}^2$ una estimación del total de los saldos. La cota de error que le corresponde es;

$$\hat{\tau} \pm 2\hat{\sigma}_{\hat{\tau}} = N\overline{y} \pm 2N\hat{\sigma}_{\overline{\hat{y}}} = \$23{,}575 \pm 2(1000)\sqrt{9.43} = \$23{,}575 \pm 9{,}430$$

B. Estimación de la proporción poblacional para una muestra aleatoria simple.

Este análisis se realiza cuando estamos interesados en estimar la proporción (porcentaje) de la población que tiene una característica específica. Tiene un comportamiento binomial ya que una observación pertenece o no a la categoría estudiada.

Estimador de la proporción poblacional $\hat{\mathbf{p}} = \frac{\mathbf{y}}{\mathbf{n}}$ *donde* y *es en este caso el número total de los elementos de la muestra que tienen determinada característica.*

Varianza estimada del estimador $\hat{\sigma}^2_{\hat{p}} = \left(\frac{\hat{p}\hat{q}}{n-1}\right)\left(\frac{N-n}{N}\right)$ con $\hat{p} + \hat{q} = 1$

Cota para el error de estimación. $\hat{\mathrm{p}} \pm 2\hat{\sigma}_{\hat{p}}$

Ejemplo:
Para el ejercicio anterior, supongamos que 2 de las 8 cuentas de la muestra tienen saldos vencidos. Estime el total de cuentas atrasadas y establezca una cota para el error de la estimación.

La proporción muestral está dada por: $\hat{\mathrm{p}} = \frac{2}{8} = 0.25$

La cota del error de estimación es $\hat{\sigma}^2_{\hat{p}} = \left(\frac{0.25(0.75)}{8-1}\right)\left(\frac{1000-8}{1000}\right) = 0.026$ con $\hat{q} = 1 - 0.25 = 0.75$

Por lo tanto, se estima que el 25% (0.25) de las cuentas tienen saldo vencido, con una cota de error de

$$0.25 \pm 2\sqrt{0.026} = 0.25 \pm 0.32$$

Tamaño de la muestra

El número de observaciones necesarias para estimar una media y/o total poblacional con un límite para el error de estimación de magnitud B se obtiene de la siguiente manera, con un límite para el error de la estimación,

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2} \qquad donde \quad D = \frac{B^2}{4}$$

Ejemplo. La cantidad promedio de cuentas por cobrar de un hospital debe ser estimada, aunque no se cuenta con mucha información, se sabe que las cuentas tienen una varianza de $\sigma^2 = 625$. Existen $N = 1000$ cuentas abiertas, encuentre el tamaño de muestra necesario para estimar μ con un límite de para el error de estimación de $B = \$3$.
Se calcula entonces D:

$$D = \frac{B^2}{4} = \frac{3^2}{4} = 2.25$$

De esta manera el tamaño de la muestra es,

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2} = \frac{1000 * 625}{(1000-1) * 2.25 + 625} = 217.56$$

Se necesitan tomar 218 observaciones para estimar la media de las cuentas por cobrar μ, con un límite para el error de estimación de $3 pesos

De manera similar para determinar el tamaño de la muestra para estimar el total poblacional τ

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2} \qquad donde \quad D = \frac{B^2}{4N^2}$$

Ejemplo. Un investigador está interesado en estimar la ganancia en peso total en un periodo de 4 semanas de $N = 1000$ pollitos alimentados con cierto alimento. Obviamente, pesar cada uno de los pollitos sería una tarea tediosa y demorada. Se debe determinar el número de pollitos que serán seleccionados para estimar τ con un límite para el error de estimación igual a 1000 gramos. Estudios previos dicen que la varianza poblacional σ^2 es de aproximadamente 36 gramos. Determine el tamaño de muestra requerido.

$$D = \frac{B^2}{4N^2} = \frac{1000^2}{4(1000)^2} = 0.25$$

De esta manera el tamaño de la muestra es,

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2} = \frac{1000 * 36}{(1000-1) * 0.25 + 36} = 125.98$$

Se requiere pesar a 126 pollitos para estimar el total poblacional τ***, la ganancia en peso*** en las 4 semanas de N=1000 pollitos, con un límite para el error de estimación igual a 1000 gramos.

Muestreo sistemático

Es el método donde se selecciona un punto aleatorio de inicio y posteriormente se elige cada $k\_ésimo$ miembro de la población. Es similar al muestreo simple aleatorio; sin embargo, no requiere contar con el marco muestral. Pudiera ser más económico y rápido, no obstante, tiene la desventaja de la periodicidad, es decir, al obtener las unidades o elementos muestrales de manera sistemática, se pueden realizar mediciones que obtienen estimaciones sesgadas.

El muestreo aleatorio sistemático, parte de encontrar una muestra a partir de un factor determinista. A partir de una selección al azar de un elemento de la población, seleccionamos los restantes al aplicar este factor de selección, el cual es proporcional al tamaño de la población.

$$k = \frac{N}{n}$$

Por ejemplo una empresa que esté interesada en estudiar el comportamiento de sus ventas, puede ordenar los totales de ventas diarias en una lista, si el factor es por ejemplo de 7, seleccionaría siempre el mismo día de la semana, lo cual podría ser un inconveniente ya que siempre será seleccionado el mismo día de la semana, por ejemplo, el lunes, presentando estimadores de ventas muy por debajo de lo real, o por el contrario seleccionar todos los viernes, donde la estimación queda sobre valorada.

El resultado de k se redondea al entero más cercano. Este procedimiento se hace más sencillo porque en lugar de extraer n números aleatorios sólo se extrae uno. Y porque es fácil si al igual que el muestreo aleatorio simple, se tienen enumerados todos los elementos de la población, o si de lo contrario no se tienen enumerados de todos modos se puede realizar, pero se debe observar el orden físico de los elementos de la población. Cuando el orden físico de la población se relaciona con la característica de la población no se debe aplicar el muestreo aleatorio sistemático. El riesgo de este tipo de muestreo está en los casos en que se dan periodicidades en la población ya que al elegir a los miembros de la muestra con una periodicidad constante podemos introducir una homogeneidad que no se da en la población.

Procedimiento:

1) Definir la población de estudio.
2) Determinar el tamaño de muestra requerido.
3) Se calcula la muestra sistemática dividiendo la población entre el tamaño de la muestra.
4) El valor de *k* es el intervalo de selección que indica cada *k* de veces que un elemento de la población se integrará a la muestra (en el caso de no estar enumerados los elementos). Y también es el intervalo de selección del cual se escogerá un número aleatoriamente dentro de este intervalo (en caso de que los elementos estén enumerados), y de ahí se parte para seleccionar las muestras en los demás grupos o intervalos de selección.

Ejemplo:

Cuando los elementos no están enumerados. Una empresa armadora de autos va a probar el sistema de frenos, el departamento de ingeniera considera que una muestra de 50 autos de una población de 500 será suficiente para considerar si el lote se regresa a fabrica, $k = \frac{500}{50}$, $k = 10$. Ya los autos se encuentran estacionados en forma ordenada en un almacén, este intervalo de selección indica que cada 10 autos que contemos se integrarán a la muestra. El primero en la muestra es el décimo auto estacionado, el segundo en la muestra es el vigésimo, el tercero el trigésimo y así de diez en diez hasta completar los 50 autos de la muestra.

Cuando los elementos están numerados. Si la población se compone de una cartera de clientes pre numerados; por ejemplo, $N = 800$ y se quiere extraer una muestra sistemática de *n=40,* se aplica la formula,

$$k = \frac{N}{n} = \frac{800}{40} \cong 20$$

De este intervalo selecciona un número aleatorio entre 1 y 20, y se incluye cada vigésimo elemento tras la primera selección de la muestra. Supongamos que el primer número seleccionado es 8. Los elementos que serían seleccionados para integrar la muestra,

28, 48, 68, 88, 108, 128, 148, 168, 188, 208, 228, 248, 268, 288, 308, 328, 348, 368, 388, 408, 428, 448, 468, 488, 508, 528, 548, 568, 588, 608, 628, 648, 668, 688, 708, 728, 748, 768 y 788.

Muestreo estratificado

Consiste en la división previa de la población de estudio en grupos o clases que se suponen homogéneos con respecto a alguna característica de las que se van a estudiar y tomar una muestra de cada uno de estos grupos. Posteriormente, si utilizamos muestreo simple aleatorio para extraer los elementos de la muestra de cada estrato, tendremos un ***Muestreo aleatorio estratificado***. Normalmente si no se especifica lo contrario se llamará solamente ***muestreo estratificado***.

Un estrato se define mediante algunas características comunes como son el sexo, la población, la edad, la profesión, etc. A diferencia del muestreo simple, el muestreo estratificado es útil porque garantiza que cada grupo esté representado en la muestra. El procedimiento de muestreo será más útil mientras más homogéneos sean los diferentes estratos.

Hay algunas ventajas al utilizar el muestreo estratificado,

- Además de estimar los parámetros de la población, nos permite realizar estimaciones de cada uno de los estratos. Por ejemplo, es posible estimar el comportamiento de los consumidores en cada estrato para después estimar el comportamiento de la población.
- El muestreo estratificado nos puede garantizar una muestra más representativa. La estratificación reduce la probabilidad de obtener una muestra no representativa al asegurar que un número de elementos sean seleccionados de cada estrato.
- No es necesario disponer de la lista de toda la población sino de las subpoblaciones de orden superior extraídas.
- Existe una considerable reducción de costos. Es más fácil tomar una muestra de por ejemplo zonas o colonias específicas de una ciudad que muestrear toda la ciudad

La distribución de la muestra en función de los diferentes estratos, puede ser de diferentes tipos:

a) Afijación simple. A cada estrato le corresponde igual número de elementos muestrales. La desventaja es que favorece a los estratos más pequeños
b) Proporcional. Cada estrato se encuentra representado en la muestra en proporción exacta al tamaño de la población total.
c) Óptima. Se tiene en cuenta la previsible dispersión de los resultados, de modo que se considera la proporción y la desviación típica. Tiene poca aplicación ya que no se suele conocer la desviación.

A cada uno de estos estratos se le asignaría una cuota que determinaría el número de miembros del mismo que compondrán la muestra. Dentro de cada estrato se suele usar, como ya lo indicamos antes, la técnica de muestreo aleatorio o sistemático, una de las técnicas de selección más usadas en la práctica.

Si N es el tamaño de la población y n el de la muestra. La población será dividida en L estratos de tamaño $l_1, l_2, l_3, \ldots l_L$

$$N = l_1 + l_2 + l_3 + \ldots l_L = \sum_{i=1}^{L} l_i$$

Y la muestra

$$n = n_1 + n_2 + n_3 + \ldots n_L = \sum_{i=1}^{L} n_i$$

Por ejemplo, para un estudio de opinión, puede resultar interesante estudiar por separado las opiniones de hombres y mujeres pues se estima que, dentro de cada uno de estos grupos, puede haber cierta homogeneidad. Así, si la población está compuesta de un 55% de mujeres y un 45% de hombres, se tomaría una muestra que contenga también esos mismos porcentajes de hombres y mujeres.

Una vez especificados los estratos, usamos el procedimiento mostrado en el muestreo simple aleatorio para seleccionar una muestra aleatoria para cada estrato. Usando el método de afijación (o asignación) proporcional, el tamaño de la muestra n se divide en cada uno de los estratos, de acuerdo a la siguiente regla,

$$n_i = n\left(\frac{N_i}{N}\right) \quad i = 1,2,3,\ldots,L \quad \textit{donde } N_i \textit{ es el número de elementos de estrato } i$$

$$Por\ lo\ tanto\quad N=\sum_{i=1}^{L}N_i\quad es\ el\ tamaño\ de\ la\ población.$$

En resumen, para seleccionar una muestra estratificada por afijación proporcional realizamos las siguientes acciones:

1) Definir la población en estudio
2) Determinar el tamaño de la muestra requerido
3) Establecer los estratos
4) Determinar la frecuencia relativa del muestreo de cada estrato por el tamaño de la muestra total, para obtener de cada estrato la cantidad de individuos que integrarán la muestra total.
5) Seleccionar y extraer de cada estrato la cantidad de elementos que formarán parte de la muestra, por el método de muestreo aleatorio simple.

Ejemplo. En un club de tenis, los 500 socios se reparten por edades en cuatro categorías: la 1ª con 200 socios, la 2ª con 175, la 3ª con 75 y la 4ª con 50. Se quiere seleccionar una muestra de 40 socios.

a) ¿Qué tipo de muestreo deberíamos realizar si queremos que estén representadas todas las edades?
b) ¿Cuántos individuos elegiríamos de cada categoría, si atendiéramos a razones de proporcionalidad?

Solución:

a) Deberíamos realizar un muestreo aleatorio estratificado.

b) Llamamos n1, n2, n3, n4 al número de individuos que tendríamos que seleccionar en cada categoría (1a, 2a, 3ª y 4ª , respectivamente). Entonces:

$$\frac{n_1}{200}=\frac{n_2}{175}=\frac{n_3}{75}=\frac{n_4}{50}=\frac{40}{500}$$

Así, debemos seleccionar $\frac{n_1}{200}=\frac{40}{500}\ \rightarrow\ n_1=200\left(\frac{4}{50}\right)=16$ de la misma manera calculamos las demás categorías.

$$n_1=16;\quad n_2=175\left(\frac{4}{50}\right)=14;\quad n_2=75\left(\frac{4}{50}\right)=6;\ n_2=50\left(\frac{4}{50}\right)=4$$

A. Estimación de la media poblacional para una muestra aleatoria estratificada.

Ahora podemos encontrar la *media y la varianza para cada estrato*, de la siguiente forma,

$$Media\ \ \overline{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} \quad y\ la\ varianza\ de\ cada\ estrato\ \ s_i^2 = \frac{\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2}{n_i - 1} \quad i = 1,2,3, \ldots\ L$$

$$El\ término\ y_{ij}\ es\ la\ \ j-ésima\ observación\ del\ estrato\ i$$

Para estimar la *media poblacional* para una muestra aleatoria estratificada.

$$Estimador\ \ \overline{y}_{est} = \frac{1}{N}\sum_{i=1}^{L} N_i \overline{y}_i$$

Y la varianza estimada del estimador es

$$\hat{\sigma}^2_{\overline{y}_{est}} = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \left(\frac{s_i^2}{n_i}\right)\left(\frac{N_i - n_i}{N_i}\right)$$

Finalmente, la cota para el error de la estimación es $\overline{y}_{est} \pm 2\hat{\sigma}_{\overline{y}_{est}}$

Ejemplo: Si se tiene que seleccionar una muestra de 20 personas, de una comunidad de 500 habitantes, con el fin de hacerles una encuesta sobre la inversión en ahorros diaria. Los habitantes están repartidos en 5 comunidades, en donde el tamaño de cada estrato es:

Estrato	Colonia	Tamaño N_i		No de muestras por estrato	s_i^2
1	Mochitlán	100	$n_1 = 40\left(\frac{100}{500}\right)$	8	$ 16.81
2	Quechultenango	150	$n_2 = 40\left(\frac{150}{500}\right)$	12	$ 22.09
3	Juan R Escudero	50	$n_3 = 40\left(\frac{50}{500}\right)$	4	$ 125.44
4	San Marcos	125	$n_4 = 40\left(\frac{125}{500}\right)$	10	$ 45.2
5	Ayutla	75	$n_5 = 40\left(\frac{75}{500}\right)$	6	$ 130.2
Total		500		40	

Para fines prácticos se considera como dato la varianza de cada estrato s_i^2

58	144	147	94	40	26	135	9	2	16	129	42	5	150	22	126	149	69	109	19	51	3	4	39	11
114	116	79	50	146	104	87	33	83	126	71	68	53	41	122	62	6	144	8	149	111	98	31	146	2
70	5	36	55	148	141	81	144	112	99	36	107	104	145	95	43	95	73	39	52	30	131	140	88	60
52	118	110	33	144	15	25	58	76	29	49	108	67	34	88	38	129	4	101	72	105	144	59	132	51
137	106	41	113	39	139	128	55	17	16	105	116	96	45	86	71	96	129	94	118	40	68	9	9	16
131	35	68	69	61	42	35	9	116	108	2	145	80	27	121	13	116	94	49	121	11	47	62	64	103

Los habitantes de cada comunidad están registrados y se les asignará un número, por ejemplo, en el estrato 1 hay 100 habitantes entonces se numerará de 001 a 100, en el estrato 2 hay 150 y se numerará de 001 a 150 y así sucesivamente se hará con los demás estratos. Y del tamaño de cada estrato se sacaran el número de muestras que se obtuvieron, por medio del método de muestreo aleatorio simple con la tabla de números aleatorios siguiente.

Los datos de cada comunidad se anexan al final.

Del estrato 1 (1 a 100) se tomarán las 8 muestras de la fila 1 de izquierda a derecha. Las muestras son: 58, 94, 40, 26, 9, 2, 16 y 42

Del estrato 2 (1 a 150) se tomarán las 12 muestras de la fila 2 de izquierda a derecha. Las muestras son: 114, 116, 79,50, 146, 104, 87, 33, 83, 126, 71 y 68

Del estrato 3 (1 a 50) se tomarán las 4 muestras de la fila 3 de izquierda a derecha. Las muestras son: 5, 36, 43 y 39

Del estrato 4 (1 a 125) se tomarán las 10 muestras de la fila 4 de izquierda a derecha. Las muestras son: 52, 118, 110, 33, 15, 25, 58, 76, 29 y 49

Del estrato 5 (1 a 75) se tomarán las 6 muestras de la fila 5 de izquierda a derecha. Las muestras son: 41, 39, 55, 17, 16 y 45

Estrato	Colonia	Tamaño N_i	Promedio de la muestra $\bar{y}$	No de muestras por estrato	s_i^2
1	Mochitlán	100	10.05	8	$ 16.81
2	Quechultenango	150	24.8	12	$ 22.09
3	Juan R Escudero	50	24.3	4	$ 125.44
4	San Marcos	125	34.6	10	$ 45.2
5	Ayutla	75	53.4	6	$ 130.2
Total		500		40	

A partir de la tabla anterior, se estima la inversión en ahorros promedio.

$$\tilde{y}_{est} = \frac{1}{N}\sum_{i=1}^{L} N_i \overline{y}_i$$

$$= \frac{1}{500}[100(10.05) + 150(24.8) + 50(24.3) + 125(34.6) + 75(53.4)]$$

$$= \frac{1}{500}(15270) \approx 29.0$$

Por lo tanto el ahorro promedio estimado que realizó la comunidad es $ 28.54 y la varianza estimada es,

$\hat{\sigma}^2_{\overline{y}_{est}} = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \left(\frac{s_i^2}{n_i}\right)\left(\frac{N_i - n_i}{N_i}\right) = \frac{1}{(500)^2}\left[(100)^2\left(\frac{100-8}{100}\right)\left(\frac{16.81}{8}\right) + (150)^2\left(\frac{150-12}{150}\right)\left(\frac{22.09}{12}\right) + (50)^2\left(\frac{50-4}{50}\right)\left(\frac{125.44}{4}\right) + (125)^2\left(\frac{125-10}{125}\right)\left(\frac{45.2}{10}\right) + (75)^2\left(\frac{75-6}{75}\right)\left(\frac{130.2}{6}\right)\right]$

$\hat{\sigma}^2_{\overline{y}_{est}} == \frac{1}{(500)^2}\ (19331.5 + 38105.25 + 72128 + 64975 + 112297.5) = 1.227$

Los ahorros promedio con una cota de error de la estimación son entonces,

$$\overline{y}_{est} \pm 2\hat{\sigma}_{\overline{y}_{est}} = \$\ 29 \pm 2\sqrt{1.227} = \$29 \pm 1.1077$$

B. Estimador del total poblacional para una muestra aleatoria estratificada

Si el objetivo de la encuesta es utilizar el muestreo estratificado para estimar el total poblacional τ procedemos de la siguiente forma,

Estimador del total poblacional $\hat{\tau} = N\overline{y}_{est}$ y la varianza estimada del estimador $\hat{\sigma}^2_{\hat{\tau}} = N^2\hat{\sigma}^2_{\overline{y}_{est}}$

La cota de error de la estimación $\hat{\tau} \pm 2\hat{\sigma}_{\hat{\tau}}$

Regresando a nuestro ejemplo, para estimar el ahorro total poblacional.

$$\hat{\tau} = 500(29) = 14500$$

Las cotas del error son,

$\hat{\sigma}^2_{\hat{\tau}} = N^2\hat{\sigma}^2_{\overline{y}_{est}} = (500)^2(1.227) = 306925$

y la cota $\hat{\tau} \pm 2\hat{\sigma}_{\hat{\tau}} = 14500 \pm 2\sqrt{306925} = \$14{,}500 \pm \$1{,}108$

C. Estimación de la proporción poblacional para una muestra aleatoria estratificada.

Estimador de la proporción poblacional $\hat{p}_{est} = \frac{1}{N}\sum_{i=1}^{L} N_i \hat{p}_i$

Varianza estimada del estimador $\widehat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\hat{\boldsymbol{p}}_{\boldsymbol{est}}} = \frac{\mathbf{1}}{\boldsymbol{N}^{\mathbf{2}}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{L}} \boldsymbol{N}_{\boldsymbol{i}}^{\mathbf{2}} \left(\frac{\hat{\boldsymbol{p}}_{\boldsymbol{i}}\,\hat{\boldsymbol{q}}_{\boldsymbol{i}}}{\boldsymbol{n}_{\boldsymbol{i}}-\mathbf{1}}\right)\left(\frac{\boldsymbol{N}_{\boldsymbol{i}}-\boldsymbol{n}_{\boldsymbol{i}}}{\boldsymbol{N}_{\boldsymbol{i}}}\right)$ con $\widehat{\boldsymbol{p}}_{\boldsymbol{i}} + \widehat{\boldsymbol{q}}_{\boldsymbol{i}} = \mathbf{1}$

Y las cotas para el error de la estimación $\hat{p}_{est} \pm 2\hat{\sigma}_{\hat{p}_{est}}$

Tamaño de muestra para estimar el promedio con asignación proporcional.

$$n = \frac{\sum_{i=1}^{k} n_i \tilde{s}_i^2}{\frac{NB^2}{k^2} + \frac{1}{N}\sum_{i=1}^{k} n_i \tilde{s}_i^2}$$

Dónde: B = error de estimación

k = Percentil que se halla en la tabla de la distribución normal y depende del nivel de confianza

EJEMPLO. Se desea estimar la nota promedio de los estudiantes de administración de empresas diurna y nocturna en una universidad. En la carrera diurna (estrato 1) hay 280 estudiantes y en la nocturna (estrato 2) hay 200 estudiantes. Determine el tamaño de muestra necesario para cumplir el objetivo con un error máximo de 0.15 y una confiabilidad del 95 por ciento.

Por un estudio realizado tiempo atrás se conocen las varianzas de las notas de administración diurna y nocturna, las que respectivamente son: 0.31 y 0.28.

Solución

Considerando que las varianzas son similares, se trabaja con muestreo estratificado con asignación proporcional. El error (B) es 0.15 y para una confiabilidad del 95 por ciento el valor correspondiente en la distribución normal es 1.96, entonces, k = 1.96:

$$n_1 = 280; \quad n_2 = 200; \quad N = 480; \quad \tilde{s}_1^2 = 0.31; \quad \tilde{s}_2^2 = 0.28$$

Para hallar el tamaño de muestra se utiliza la ecuación anterior

$$n = \frac{280(0.31) + 200(0.28)}{480\,\frac{0.15^2}{1.96^2} + \frac{1}{480}[280(0.31) + 200(0.28)]} = 45.93$$

El tamaño de la muestra es de 46 estudiantes. Esta muestra se reparte proporcionalmente al tamaño de los estratos, como se realizó con anterioridad

$$n_1 = 46\frac{280}{480} = 26.83; \qquad n_2 = 46\frac{200}{480} = 19.17$$

Se deben seleccionar 27 estudiantes de administración de empresas diurna y 19 de la nocturna.

Tamaño de muestra para estimar el total con asignación proporcional

$$n = \frac{\sum_{i=1}^{k} n_i \tilde{s}_i^2}{\frac{B^2}{k^2 N} + \frac{1}{N}\sum_{i=1}^{k} n_i \tilde{s}_i^2}$$

***Ejemplo*.** Se desea hacer un estudio para estimar el consumo total de gasolina en una ciudad, halle el tamaño de muestra necesario para cumplir este objetivo. Los vehículos se clasificaron en tres grupos o estratos, particulares (1), públicos (2) y oficiales (3). En la oficina de circulación y tránsito se obtuvo la siguiente información sobre los vehículos matriculados en la ciudad; vehículos particulares 7,627, públicos 2,392 y oficiales 534.

Solución

Como no se dispone de estudios similares, se toma una muestra piloto, con la cual se obtienen las siguientes varianzas sobre el consumo semanal en galones:

$$\tilde{s}_1^2 = 137.6; \quad \tilde{s}_2^2 = 138; \quad \tilde{s}_3^2 = 135.86$$

Asumiendo un error de estimación máximo de 15,000 galones, (B = 15,000), y una confiabilidad del 95 por ciento, el valor de k en la distribución normal es 1.96.

Considerando que las varianzas en los tres estratos son similares, se trabaja con muestreo estratificado con asignación proporcional. Para calcular el tamaño de la muestra se utiliza la ecuación anterior.

$$N_1 = 7{,}627 \quad N_2 = 2{,}392 \quad N_3 = 534 \quad N = 10{,}553$$

$$n = \frac{7627(137.6) + 2392(138) + 534(135.86)}{\frac{15{,}000^2}{1.96^2(10{,}553)} + \frac{1}{10553}[7627(137.6) + 2392(138) + 534(135.86]} = 255$$

$$n_1 = 255\frac{7{,}627}{10{,}553} = 184; \qquad n_2 = 255\frac{2{,}392}{10{,}553} = 58; \qquad n_3 = 255\frac{534}{10{,}553} = 13$$

Para estimar el consumo total de gasolina con un error máximo de 15.000 galones/semana, se debe seleccionar una muestra de 255 autos repartida así: 184 autos particulares, 58 públicos y 13 oficiales.

Recuerde que, si se desea, se puede disminuir el error máximo admisible, pero esto conlleva a un aumento en el tamaño de la muestra.

Tamaño de muestra para estimar la proporción con asignación proporcional

$$n = \frac{\sum_{k=1}^{L} n_k p_k q_k}{N\frac{B^2}{k^2} + \frac{1}{N}\sum_{k=1}^{L} n_k p_k q_k}$$

***Ejemplo*.** En vista de la recesión económica existente, una empresa textil pretende reducir el número de días laborables por semana a cuatro. Otra alternativa consiste en clausurar una de sus tres plantas y despedir a los trabajadores. Para tener una idea de la opinión de los trabajadores, el gerente de personal de la empresa desea seleccionar una muestra de empleados de las tres plantas para estimar la proporción de trabajadores que prefieren la reducción de la semana de trabajo, con un error de estimación máximo de 0.1.

La empresa emplea 150 personas en la planta 1.65 en la planta 2 y 40 en la 3. Se estima que cerca del 75 por ciento de los de la planta tres están a favor de la reducción de la semana de trabajo, mientras que en las otras plantas este porcentaje parece corresponder al 50 por ciento. Encuentre el tamaño de muestra y la asignación necesaria en cada estrato.

Solución

$$N_1 = 150 \quad N_2 = 65 \quad N_3 = 40 \quad N = 255$$
$$p_1 = 0.5 \quad p_2 = 0.5 \quad p_3 = 0.75 \quad B = 0.1$$

Por la diferencia en el tamaño de las plantas, se utiliza el muestreo estratificado con asignación proporcional.

Asumiendo un nivel de confianza del 95 por ciento, el valor correspondiente en la distribución normal es 1.96 (k=1.96).

Para determinar el tamaño de la muestra sustituimos en la ecuación,

$$n = \frac{150(0.5) + 65(0.5)(0.5) + 40(0.75)(0.25)}{255\frac{0.1^2}{1.96^2} + \frac{1}{255}[150(0.25) + 65(0.25) + 40(0.1875)]} = 67.76$$

$$n_1 = 68\frac{150}{255} = 40; \quad n_2 = 68\frac{65}{255} = 17.33; \quad n_3 = 68\frac{40}{255} = 10.67$$

Muestreo por conglomerados

Técnica similar al muestreo por estadios múltiples, se utiliza cuando la población se encuentra dividida, de manera natural, en grupos que se supone que contienen toda la variabilidad de la población, es decir, la representan fielmente respecto a la característica a elegir, pueden seleccionarse sólo algunos de estos grupos o *conglomerados* para la realización del estudio.

En una muestra de conglomerados, se divide *N* elementos de la población en varios grupos de tal manera que cada uno sea representativo de toda la población. Este procedimiento tiende a proporcionar mejores resultados cuando los elementos dentro de los conglomerados no son semejantes. Lo ideal es que cada conglomerado sea una representación, a pequeña escala, de la población. Se aplica en el muestreo de áreas, en la que los conglomerados son manzanas, ciudades, distritos electorales, países, etc. En este tipo de muestreo es imprescindible diferenciar entre unidad de análisis entendida como quiénes va a ser medidos y unidad muestral que se refiere al conglomerado a través del cual se logra el acceso a la unidad de análisis.

Procedimiento:

1) Dividir la población en conglomerados.
2) Crear una lista de todos ellos
3) Seleccionar al azar el número de conglomerados que desee.
4) Tomar una muestra aleatoria simple de uno de los elementos de cada conglomerado.

Dentro de los conglomerados seleccionados se ubicarán las unidades elementales, por ejemplo, las personas a encuestar, y podría aplicársele el instrumento de medición a todas las unidades, es decir, los miembros del conglomerado, o sólo se podría aplicar a algunos de ellos, seleccionados al azar. Este método tiene la ventaja de simplificar la recogida de información muestral.

Cuando, dentro de cada conglomerado seleccionado, se extraen algunos individuos para integrar la muestra, el diseño se llama *muestreo bietápico*.

Las ideas de estratos y conglomerados son, en cierto sentido, opuestas. El primer método funciona mejor cuanto más homogénea es la población respecto del estrato, aunque más

diferentes son éstos entre sí. En el segundo, ocurre lo contrario. Los conglomerados deben presentar toda la variabilidad, aunque deben ser muy parecidos entre sí.

Ejemplo:

Si se va a realizar una encuesta sobre las políticas y leyes del municipio, se podría dividir el municipio en distritos, por ejemplo en 13 distritos, de esos tres se toma al azar el 4, 5, 9 y 11, y solo concentrándonos en estos distritos, tomamos una muestra aleatoria de habitantes de cada uno de esos distritos, para entrevistarlos.

A. *Estimación de la media poblacional en un muestreo por conglomerados*

Al usar muestreo por conglomerados, la media poblacional, se estima de la siguiente forma. En estos cálculos, n_i es el número de elementos del i $-$ ésimo conglomerado y t_i es el total de las mediciones del conglomerado.

$$\hat{\mu} = \frac{\sum_{i=1}^{m} t_i}{\sum_{i=1}^{m} n_i}$$

Varianza estimada del estimador

$$\hat{\sigma}^2_{\overline{y}_c} = \left(\frac{M-m}{Mm\overline{m}^2}\right)\left(\frac{\sum_{i=1}^{m}(t_i - \overline{y}_c n_i)^2}{m-1}\right)$$

Cotas para el error de la estimación $\overline{y}_c \pm 2\hat{\sigma}_{\overline{y}_c}$ donde $\overline{n} = \frac{1}{m}\sum_{i=1}^{m} n_i \quad \overline{t} = \frac{1}{m}\sum_{i=1}^{m} t_i$

M es el número de conglomerados en la población y m es el número de conglomerados en la muestra.

B. *Estimación del total poblacional en un muestreo por conglomerados*

$$\hat{\tau} = \frac{M}{m}\sum_{i=1}^{m} t_i$$

Varianza estimada del estimador

$$\hat{\sigma}^2_{\hat{\tau}} = M^2\left(\frac{M-m}{Mm}\right)\left(\frac{\sum_{i=1}^{m}(t_i - \overline{t})^2}{m-1}\right)$$

Cotas para el error de la estimación $\hat{\tau} \pm 2\hat{\sigma}_{\hat{\tau}}$

C. *Estimación de la proporción poblacional para un muestreo por conglomerados.*

Estimador $\hat{p}_c = \frac{\sum_{i=1}^{m} a_i}{\sum_{i=1}^{m} n_i}$

Varianza estimada del estimador $\widehat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\hat{\boldsymbol{p}}_{\boldsymbol{c}}} = \left(\frac{M-m}{Mm\overline{m}^2}\right)\left(\frac{\sum_{i=1}^{m}(a_i-\overline{p}_c n_i)^2}{m-1}\right)$

Y las cotas para el error de la estimación $\hat{p}_c \pm 2\hat{\sigma}_{\hat{p}_c}$

ANEXO

TABLA DE NUMEROS ALEATORIOS

17841	49597	92623	80005	11177	15145	46379
84970	47043	64048	06993	17369	70932	47950
30524	27250	73072	52654	33653	30422	22347
56211	27219	44652	09467	62848	82479	35068
66110	69181	13200	93239	25591	21248	06881
28710	52414	55893	25632	64856	51745	46855
38939	15777	66270	53052	05160	94786	81987
31297	00722	88300	21109	13124	96742	64968
34043	19959	77949	24510	93510	40492	81113
74996	32698	29430	58603	43879	7861	15870
09224	49628	26353	25592	78113	27589	87512
16925	04512	74150	64475	04497	64977	30847
95370	15305	08474	58306	65393	86919	16478
74611	68568	45153	72541	14812	64511	20253
24791	18151	68084	01936	53838	48954	25322
22283	30815	82384	19084	41248	77855	22366
71898	26726	89650	31162	86245	00370	31000
95413	08931	96334	31263	45687	89601	25395
50790	90191	37070	59230	21080	86042	46441
86175	96384	63337	73013	34939	59945	62412
51485	90027	98827	43212	93302	64337	34026
24958	56475	29207	62272	41011	56041	37735
47249	28708	17767	20087	43020	20963	59504
62152	80266	99282	22863	81820	09317	74915
09135	46518	71377	14410	69712	65884	14366
43770	34210	35225	08830	65793	43288	22567
49358	18612	11688	52443	39456	65328	44806
67452	60795	63023	21400	02021	20485	09224
39578	43182	40366	02955	47485	54797	16874
49630	42256	95206	52914	15086	01292	24360
86158	26615	20228	14854	00161	64983	59471
81648	63523	82624	81928	54646	62114	36529
72208	67425	77273	35454	43798	89958	98485
62663	32726	14266	48467	36706	90411	84898
99530	11547	35629	86192	25909	97084	30951
36626	80491	21369	48285	59708	44408	75096

Fuente. Elaboración propia

Anexo 2. Datos ejercicio muestreo estratificado

	Datos muestreo estratificado comunidad Mochitlán							
Número	Ahorro	Número	Ahorro	Número	Ahorro	Número	Ahorro	
1	8.9	26	8.0	51	12.7	76	7.6	
2	11.4	27	8.0	52	15.8	77	6.0	
3	9.8	28	9.7	53	17.7	78	7.6	
4	11.9	29	8.9	54	8.5	79	10.7	
5	6.8	30	13.7	55	12.7	80	6.7	
6	9.4	31	8.9	56	16.9	81	6.0	
7	10.1	32	8.3	57	12.9	82	4.3	
8	4.7	33	10.5	58	12.7	83	11.4	
9	7.2	34	16.6	59	11.1	84	9.9	
10	11.1	35	15.9	60	11.9	85	5.8	
11	11.5	36	9.7	61	13.5	86	6.1	
12	9.1	37	8.9	62	7.9	87	10.0	
13	9.2	38	12.1	63	9.0	88	11.4	
14	8.8	39	16.9	64	6.0	89	4.8	
15	17.7	40	6.5	65	7.4	90	6.0	
16	12.1	41	11.3	66	7.2	91	9.2	
17	15.1	42	15.2	67	14.8	92	8.3	
18	8.6	43	17.7	68	7.6	93	9.1	
19	7.9	44	20.0	69	8.4	94	7.3	
20	10.3	45	13.5	70	6.8	95	6.2	
21	15.2	46	13.5	71	8.2	96	5.1	
22	7.2	47	15.8	72	6.8	97	6.0	
23	8.9	48	9.6	73	8.4	98	8.1	
24	9.7	49	11.5	74	7.6	99	4.9	
25	10.1	50	20.1	75	14.3	100	6.8	
Promedio	10.16							

	Datos muestreo estratificado comunidad Juan R Escudero								
Número	Ahorro	Número	Ahorro	Número	Ahorro	Número	Ahorro	Número	Ahorro
1	13.0	11	20.8	21	14.4	31	28.6	41	20.1
2	31.1	12	24.4	22	16.0	32	27.3	42	17.0
3	18.9	13	14.3	23	19.5	33	27.3	43	15.3
4	15.9	14	17.3	24	18.9	34	21.8	44	17.5
5	15.6	15	18.6	25	16.0	35	18.2	45	16.0
6	16.0	16	23.2	26	29.2	36	38.7	46	23.3
7	21.8	17	14.8	27	38.7	37	23.3	47	21.5
8	15.7	18	31.9	28	21.8	38	22.9	48	13.0
9	25.7	19	17.2	29	16.8	39	27.7	49	23.0
10	27.3	20	17.2	30	47.6	40	11.4	50	36.1
Promedio	21.80								

	Datos muestreo estratificado comunidad Quechultenango						
Número	Ahorro	Número	Ahorro	Número	Ahorro	Número	Ahorro
1	29.2	39	32.3	77	20.7	115	16.9
2	20.4	40	24.1	78	43.0	116	17.7
3	24.6	41	18.3	79	37.8	117	15.4
4	33.9	42	23.8	80	22.4	118	16.9
5	25.8	43	20.7	81	15.4	119	37.8
6	20.0	44	20.7	82	27.2	120	17.1
7	28.9	45	19.0	83	39.5	121	37.8
8	42.8	46	23.8	84	29.1	122	33.7
9	25.8	47	33.9	85	18.5	123	30.5
10	22.4	48	18.5	86	19.0	124	16.9
11	25.8	49	17.1	87	28.9	125	19.0
12	18.7	50	19.0	88	19.0	126	20.7
13	39.5	51	22.4	89	20.4	127	15.4
14	56.3	52	18.8	90	22.0	128	23.1
15	28.9	53	37.8	91	22.4	129	25.5
16	18.6	54	28.9	92	36.1	130	24.1
17	36.1	55	28.9	93	18.6	131	19.0
18	32.9	56	15.4	94	22.3	132	21.6
19	24.1	57	36.1	95	32.3	133	30.5
20	33.9	58	32.4	96	34.6	134	24.1
21	27.6	59	22.4	97	20.7	135	20.3
22	24.6	60	23.8	98	32.3	136	27.6
23	13.8	61	21.6	99	19.0	137	24.1
24	32.3	62	18.1	100	22.2	138	23.9
25	20.4	63	32.5	101	27.2	139	22.0
26	28.9	64	29.2	102	35.6	140	20.4
27	30.6	65	27.6	103	19.0	141	39.7
28	22.4	66	25.8	104	22.1	142	24.2
29	32.3	67	22.4	105	45.9	143	21.1
30	25.8	68	36.0	106	27.6	144	18.5
31	36.8	69	18.5	107	43.0	145	22.3
32	17.0	70	32.5	108	17.1	146	24.6
33	13.5	71	15.4	109	30.8	147	25.1
34	22.4	72	19.3	110	27.2	148	25.4
35	27.2	73	19.0	111	33.9	149	21.9
36	27.5	74	20.5	112	25.5	150	19.3
37	22.3	75	18.7	113	19.0		
38	19.0	76	32.3	114	22.4		
Promedio	25.50						

	Datos muestreo estratificado comunidad San Marcos

Número	Ahorro	Número	Ahorro	Número	Ahorro	Número	Ahorro	Número	Ahorro
1	40.7	26	64.5	51	26.7	76	47.4	101	38.7
2	31.5	27	29.1	52	32.5	77	31.0	102	28.7
3	45.4	28	29.1	53	26.7	78	36.3	103	36.3
4	26.0	29	26.0	54	28.7	79	43.0	104	43.3
5	31.5	30	31.0	55	21.6	80	28.8	105	33.5
6	36.3	31	26.2	56	21.6	81	24.0	106	40.6
7	33.7	32	30.3	57	53.1	82	24.9	107	31.0
8	29.1	33	31.4	58	48.6	83	31.5	108	38.7
9	21.7	34	27.1	59	55.5	84	40.9	109	41.1
10	19.0	35	31.5	60	41.0	85	38.7	110	33.9
11	28.1	36	60.2	61	50.6	86	45.4	111	33.1
12	26.7	37	36.3	62	50.7	87	23.8	112	40.6
13	45.4	38	45.4	63	31.5	88	21.7	113	31.5
14	21.7	39	45.4	64	38.6	89	19.5	114	31.5
15	42.8	40	33.4	65	47.7	90	42.8	115	50.7
16	28.7	41	26.7	66	26.6	91	25.7	116	31.2
17	38.3	42	50.0	67	31.4	92	26.0	117	38.2
18	45.4	43	47.6	68	53.1	93	19.4	118	35.8
19	29.1	44	24.0	69	26.7	94	34.6	119	33.9
20	45.7	45	26.7	70	40.6	95	47.6	120	40.6
21	28.6	46	45.5	71	29.1	96	36.3	121	64.5
22	79.2	47	26.7	72	26.2	97	31.5	122	33.5
23	60.4	48	33.9	73	26.7	98	50.7	123	26.3
24	24.0	49	26.2	74	32.5	99	26.7	124	33.9
25	21.7	50	29.1	75	31.5	100	38.7	125	33.4
Promedio	35.4								

	Datos muestreo estratificado comunidad Ayutla								
Número	Ahorro	Número	Ahorro	Número	Ahorro	Número	Ahorro	Número	Ahorro
1	45.0	16	40.5	31	85.0	46	53.4	61	78.0
2	40.5	17	56.9	32	102.3	47	58.9	62	70.0
3	47.5	18	36.9	33	44.6	48	53.6	63	102.6
4	62.0	19	69.8	34	49.5	49	80.6	64	69.1
5	50.0	20	52.7	35	58.9	50	45.0	65	79.0
6	45.3	21	44.2	36	90.4	51	57.7	66	134.6
7	36.9	22	45.4	37	49.5	52	48.7	67	54.0
8	88.0	23	48.7	38	45.4	53	65.7	68	36.8
9	44.9	24	77.3	39	53.6	54	60.5	69	53.6
10	61.8	25	65.8	40	65.0	55	46.1	70	44.6
11	90.4	26	65.8	41	36.8	56	65.0	71	86.1
12	45.4	27	50.0	42	81.0	57	53.6	72	69.0
13	78.0	28	80.5	43	65.8	58	69.0	73	47.0
14	44.2	29	45.4	44	61.8	59	110.0	74	61.0
15	49.0	30	90.3	45	86.3	60	77.3	75	52.0
Promedio	62.1								

Bibliografía

Anderson, D. R., D. J. Sweeney y T. A. Williams. (2008). *Estadística para la administración y la economía.* (10ª ed). México: CENGAGE Learning. 260-262.

Levine, D. M., T. C. Krehbiel y M. L. Berenson. (2006). *Estadística para la administración.* (4ta ed). México: Pearson Prentice Hall. 221.

Lind, D. A., W. G. Marchal, y S. A. Wathen. (2008). *Estadística aplicada a los negocios y a la economía.* (13ª ed). México: McGraw-Hill. 262, 265, 266