Daena: International Journal of Good Conscience. 4(1): 149-178. Marzo 2009. ISSN 1870-557X.

Distribuciones probabilísticas de uso común

(Probabilistic distributions of common use)

Badii, M. H. y J. Castillo*

Resumen. Se discutan las características de las distribuciones probabilísticos de uso más común. Se presentan explicaciones para el uso correcto de diferentes distribuciones tales como la distribución binomial, geométrica, hipergeométrica, Poisson y normal. Para cada una de estas distribuciones proporcionan sus ecuaciones y ejemplos prácticos. Se contrastan las diferencias entre la aplicación adecuadas de cada una de estas distribuciones probabilísticas.

Palabras claves. Distribución binomial, geométrica, hipergeométrica, normal, Poisson

Abstract. Features of probabilistic distributions of common usage are discussed. Explications for correct usage of distinct distributions such as binomial, geometric, hipergeometric, normal, Poisson are given. Equations and practical examples for each of these distributions are provided. Distinctions for adequate application of these distributions are noted.

Keywords. Binomial, geometric, hipergeometric, normal, Poisson distributions

Introducción

Los valores de una variable sirven para describir o clasificar individuos o distinguir entre ellos. La mayoría de nosotros hacemos algo más que simplemente describir, clasificar o distinguir, porque tenemos ideas respecto a las *frecuencias relativas* de los valores de una variable. En estadística decimos que la variable tiene una *función* de *probabilidad,* una *función* de *densidad* de *probabilidad* o simplemente una *función de distribución* (Badii & Castillo, 2007).

Las distribuciones de probabilidad están relacionadas con la distribución de frecuencias. De hecho, podemos pensar en la distribución de probabilidad como una distribución de frecuencias teórica. Una distribución de frecuencias teórica es una distribución de probabilidades que describe la forma en que se espera que varíen los resultados. Debido a que estas distribuciones tratan sobre expectativas de que algo suceda, resultan ser modelos útiles para hacer inferencias y tomar decisiones de incertidumbre (Badii et al., 2007a, 2007b).

Los objetivos de distribuciones de probabilidad son:

- **a)** Introducir las distribuciones de probabilidad que más se utilizan en la toma de decisiones.
- **b)** Utilizar el concepto de valor esperado para tomar decisiones.
- **c)** Mostrar qué distribución de probabilidad utilizar, y cómo encontrar sus valores.
- **d)** Entender las limitaciones de cada una de las distribuciones de probabilidad que utilice.

Distribuciones muestrales

Consideremos todas las posibles muestras de tamaño *n* en una población dada (con o sin reposición). Para cada muestra, podemos calcular un estadístico (tal como la media o la

desviación estándar o típica), dicho estadístico varía de una muestra a otra. De esta manera obtenemos una distribución de la estadística que se llama *distribución de muestreo*. Si, por ejemplo, la estadística utilizada es la media muestral, entonces la distribución se llama la distribución de muestreo de medias, o distribución de muestreo de la media. Análogamente, podríamos tener distribución de muestreo de la desviación típica, de la varianza, de la mediana, de las proporciones, etcétera. Supongamos que se toman todas las posibles muestras de tamaño *n*, sin reposición, de una población finita de tamaño *N*>*n*. Si denotamos la media y la desviación típica de la distribución de muestreo de medias por $\mu_{\bar{x}}$, $\sigma_{\bar{x}}$ y las de la población por μ y σ, respectivamente, entonces:

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

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \qquad (1)$$

Donde:

$\mu_{\bar{x}}$ = la media de la distribución de muestreo de las medias.
μ = la media de la población.

$\sigma_{\bar{x}}$ = error estándar de la media de muestreo.
σ = desviación típica de la población.
N = tamaño de la población.
n = tamaño de la muestra.

Este nuevo factor que aparece del lado derecho de la ecuación y que multiplica a nuestro error estándar original, se conoce como *multiplicador de población finita:*

Multiplicador de población finita = $\sqrt{\frac{N-n}{N-1}}$.

Ejemplo 1. Supongamos que estamos interesados en una población de 20 compañías textiles del mismo tamaño, todas estas fábricas experimentan una producción excesiva de trabajo. Nuestro estudio indica que la desviación estándar de la distribución de la producción anual es igual a 75 empleados. Si muestreamos 5 de estas compañías (sin reemplazo), y deseamos calcular el error estándar de la media para la población finita, usaríamos la ecuación:

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

$$= \frac{75}{\sqrt{5}} X \sqrt{\frac{20-5}{20-1}} = 29.8$$

En casos en los que la población es muy grande en relación con el tamaño de la muestra, este multiplicador de población finita adquiere un valor cercano a 1 y tiene poco efecto sobre el

cálculo del error estándar. Este último pone de manifiesto que cuando muestreamos una pequeña fracción de la población entera, la fracción *n/N* se define como la *fracción de muestreo.*
Cuando la fracción de muestreo es pequeña, el error estándar de la media para poblaciones finitas es tan cercano a la media para poblaciones infinitas que bien podríamos utilizar la misma fórmula para ambas desviaciones. *Si la fracción de muestreo es menor a 0.05, no se necesita usar el multiplicador de población finita.* Si la población es infinita los resultados anteriores se reducen a:

$$\mu_{\bar{x}} = \mu$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad (2)$$

Para valores grandes de *n* (*n*≥30), la distribución de muestreo de medias es aproximadamente normal con media $\mu_{\bar{x}}$ y desviación típica $\sigma_{\bar{x}}$ independientemente de la población. Este resultado para una población infinita es un caso especial del teorema del límite central de la teoría avanzada de probabilidades, que afirma que la precisión de la aproximación mejora al crecer *n*. En estas ocasiones se dice que la distribución de muestreo es *asintóticamente normal*.

Distribución de muestreo de proporciones

Supongamos que una población es infinita y que la probabilidad de ocurrencia de un suceso (su éxito) es *p*, mientras la probabilidad de que no ocurra es *q* = 1 - *p*. Por ejemplo, la población puede ser la de todas las posibles tiradas de una moneda, en la que la probabilidad del suceso "cara" es *p* = ½. Consideremos todas las posibles muestras de tamaño *n* de tal población, y para cada una de ellas determinemos la proporción de éxitos (*P)*. En el caso de una moneda, *P* sería la proporción de caras de *n* tiradas. Obtenemos así una *distribución de muestreo de proporciones* cuya media μ_p y su desviación típica σ_p vienen dadas por:

$$\mu_p = p$$
$$\sigma_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}} \qquad (3)$$

donde:

μ_p = media de muestreo de proporciones.
P = media de la población de proporciones.
p = probabilidad de éxito.
q = probabilidad de fracaso.

Para muestras grandes (*n*≥30), la distribución de muestreo está, muy aproximadamente, normalmente distribuida. Nótese que la población está *binomialmente distribuida*. La ecuación (3) es válida también para una población finita en la que se hace muestreo con reposición. Para poblaciones finitas en que se haga muestreo sin reposición, la ecuación (3) queda sustituida por la ecuación:

$$\mu = p \text{ y } \sigma = \sqrt{pq}$$

Distribución de muestreo con diferencias y sumas

Sean dadas dos poblaciones. Para cada muestra de tamaño n_1 de la primera, calculamos una estadística S_1; eso da una distribución de muestreo para S_1, cuya media y desviación típica denotaremos por μ_1 y σ_1. Del mismo modo, para la segunda muestra de tamaño n_2 de la segunda población, calcula una estadística S_2. De todas las posibles combinaciones de estas muestras de las dos poblaciones podemos obtener una distribución de las diferencias, S_1 - S_2, que se llama *distribución de muestreo de las diferencias de los estadísticos*. La media y la desviación típica de esta distribución de muestreo, denotadas respectivamente por las siguientes: $\mu_{s1\text{-}s2}$ y $\sigma_{s1\text{-}s2}$ vienen dadas por:

$$\mu_{s1-s2} = \mu_{s1} - \mu_{s2}$$
$$\sigma_{s1-s2} = \sqrt{\sigma_{s1}^2 + \sigma_{s2}^2} \qquad (4)$$

en lo que:

$\mu_{s_1} - \mu_{s_2}$ = media de muestreo de las diferencias de los estadísticos.

$\sigma^2{}_{s_1}$ = la varianza de la estadística 1.

$\sigma^2{}_{s_2}$ = la varianza de la estadística 2.

Si S_1 y S_2 son las medias muestrales de ambas poblaciones, cuyas medias denotaremos por x_1 y x_2, respectivamente, entonces la distribución de muestreo de las diferencias de medias viene dada para poblaciones infinitas como:

$$\mu_{\bar{x}1-\bar{x}2} = \mu_{\bar{x}1} - \mu_{\bar{x}2} = \mu_1 - \mu_2$$

y

$$\sigma_{x1-\bar{x}2} = \sqrt{\sigma_{\bar{x}1}^2 + \sigma_{\bar{x}2}^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad (5)$$

donde:

$\mu_{\bar{x}_1-\bar{x}_2}$ = la media de la diferencia muestral.

$\sigma_{x_1-x_2}$ = el error estándar de la diferencia entre dos medias.

σ_1^2 = la varianza de la muestra una.

σ_2^2 = la varianza de la muestra dos.

Resultados correspondientes se pueden obtener para las distribuciones de muestreo de diferencias de proporciones de dos poblaciones binomialmente distribuidas con parámetros (p_1, q_1) y (p_2, q_2), respectivamente. En este caso, S_1 y S_2 corresponden a la proporción de éxitos P_1 y P_2, y las ecuaciones (4 y 5) llevan a:

$$\mu_{p1-p2} = \mu_{p1} - \mu_{p2} = p_1 - p_2$$

y

$$\sigma_{p1-p2} = \sqrt{\sigma_{p1}^2 + \sigma_{p2}^2} = \sqrt{\frac{p_1 q_1}{n} + \frac{p_2 q_2}{n_2}} \qquad (6)$$

Si n_1 y n_2 son grandes (n_1, $n_2 \geq 30$), la distribución de muestreo de diferencias de medias o proporciones están casi normalmente distribuidas. A veces es útil hablar de la distribución de muestreo de la suma de estadísticos. La media y la desviación típica de tal distribución son (para muestras independientes):

$$\mu_{s1+s2} = \mu_{s1} + \mu_{s2}$$

y

$$\sigma_{s1+s2} = \sqrt{\sigma_{s1}^2 + \sigma_{s2}^2} \qquad (7)$$

Combinación de probabilidades y valores monetarios

Veamos el caso de un vendedor al mayoreo de frutas y legumbres que comercia fresas. Este producto tiene una vida útil muy limitada. Si no se vende el día de su entrega, ya no tiene valor. Una cajita de fresas cuesta $20 y el vendedor recibe $50 por ella. Este no puede especificar el número de cajitas que un cliente pedirá en cualquier día dado, pero su análisis de registros pasados ha reducido la información que presentamos en la Tabla 1.

Tabla 1. Ventas durante 100 días (ejemplo 1).

Venta diarias	Número de Días de venta	Probabilidad de venta de cada cantidad
10	15	0.15
11	20	0.20
12	40	0.40
13	25	0.25
Total	**100**	**1.00**

Definición de los tipos de pérdidas

El vendedor al mayoreo ha sufrido dos tipos de pérdidas: **1)** pérdidas de abundancia, ocasionadas por tener en existencia demasiada fruta en un día y tener que tirarla al día siguiente; y **2)** pérdidas de oportunidad, ocasionadas por no tener en existencia el producto al tiempo que un cliente lo solicita.

Cada valor de pérdida está condicionado a un número específico de cajas que se encuentran en existencia y a un número específico de solicitudes. Los valores que se tienen en la Tabla 2 no solamente incluyen pérdidas por frutas echadas a perder, sino que también aquellas que son resultados de la pérdida de recuperación cuando el vendedor no es capaz de suministrar los pedidos que se le hacen.

Ninguno de estos tipos de pérdidas se obtiene cuando en existencia en un día cualquiera es el mismo que el número de cajas solicitadas. Cuando sucede lo anterior, el vendedor vende todo lo que tiene almacenado y no obtiene pérdidas. Esta situación se indica con el cero en negrita que aparece en la columna correspondiente. Las cifras que se encuentren por encima de un cero cualquiera representan las pérdidas obtenidas por tener que tirar la fruta. En todo caso, en este ejemplo, el número de cajas almacenadas es mayor que el número de cajas solicitadas.

Los valores por debajo representan las pérdidas de oportunidad que resultan de pedidos que no se pueden cumplir. Si sólo tiene en existencia 10 cajas de fresas en un cierto día y se solicitan 11, el vendedor sufre una pérdida de oportunidad de $30 por la caja que no pudo vender por no tenerla ($50 de la entrada por caja menos $20 de su costo, igual a $30).

Tabla 2. Pérdidas condicionales.

Posibles peticiones	**Posibles opciones de existencia**			
	10	**11**	**12**	**13**
10	$0	$20	$40	$60
11	30	0	20	40
12	60	30	0	20
13	90	60	30	0

Cálculo de pérdidas esperadas

Al examinar cada acto de almacenamiento posible, podemos calcular la pérdida esperada, como indica la Tabla 3.

Tabla 3. Pérdida esperada al tener en existencia 10 cajas.

Posibles solicitudes	**Pérdida ($)**	**Probabilidad**	**Esperada**
10	0	0.15	$0
11	30	0.20	6.00
12	60	0.40	24.00
13	90	0.25	22.50
Total		**1.00**	**52.50**

Las pérdidas esperadas se resumen en las Tablas 4 a 6 que resulta de tomar la decisión de tener en existencia 11, 12 y 13 cajas de fresas, respectivamente. La acción de almacenamiento óptima es aquella que minimiza las pérdidas esperadas. Esta acción corresponde al hecho de tener en existencia 12 cajas diarias, en cuyo caso las pérdidas esperadas toman el valor mínimo de $17.50. Con la misma facilidad pudimos haber resuelto este problema tomando un camino alternativo, es decir, maximizando la ganancia esperada ($50-20 del costo de cada caja), en lugar de minimizar la pérdida esperada.
En el cálculo de pérdidas esperadas hemos supuesto que la demanda del producto puede tomar únicamente cuatro valores y que las fresas ya no valen nada al día siguiente. Estas dos suposiciones reducen el valor de la respuesta que hemos obtenido.

Tabla 4. Pérdida esperada al tener en existencia 11 cajas.

Posibles solicitudes	Pérdida ($)	Probabilidad	Esperada
10	20	0.15	$3.0
11	0	0.20	0
12	30	0.40	12.00
13	60	0.25	15.50
Total		**1.00**	**30.00**

Tabla 5. Pérdida esperada al tener en existencia 12 cajas (pérdida mínima esperada).

Posibles solicitudes	Pérdida ($)	Probabilidad	Esperada ($)
10	40	0.15	6.00
11	20	0.20	4.00
12	0	0.40	0.00
13	30	0.25	7.50
Total		**1.00**	**17.50**

Tabla 6. Pérdida esperada al tener en existencia 13 cajas.

Posibles solicitudes	Pérdida condicional	Probabilidad de que se tengan estas solicitudes	Pérdida esperada
10	60	0.15	9.00
11	40	0.20	8.00
12	20	0.40	8.00
13	0	0.25	0.00
Total		**1.00**	**25.00**

El valor esperado en una situación de toma de decisiones es un valor teórico que puede no presentarse nunca. En la mejor solución para la situación óptima de existencias que tenemos en las tablas anteriores, la menor pérdida esperada es de $17.50, pero la pérdida real que puede sufrir el vendedor en cualquier día puede ser de $0, $20, $30 ó $40. Recuerde que el valor esperado es un promedio de todos los resultados posibles, pesados con la probabilidad de que cada resultado se presente.

La distribución binomial

Una distribución de probabilidad de una variable aleatoria discreta utilizada ampliamente es la distribución binomial. Esta distribución es apropiada par una variedad de procesos que describe datos discretos, que son resultado de un experimento (Badii et al., 2007d) conocido como proceso de *Bernoulli* en honor al matemático Suizo Jacob Bernoulli (1654-1705), el cual nos llevará a uno de sólo dos resultados posibles que son mutuamente exclusivos, tales como muerto o vivo, enfermo o saludable, etc., en donde la obtención del resultado deseado se considera como éxito "p" y el resultado no deseado como fracaso "q", donde, $q = 1 - p$ (Badii et al., 2007c).

Características del proceso de Bernoulli

Podemos utilizar el resultado del lanzamiento de una moneda no alterada un cierto número de veces como ejemplo de proceso de Bernoulli. Podemos describir el proceso de la manera siguiente:

1. Cada ensayo conduce a uno de dos resultados posibles, mutuamente exclusivos, uno denominado éxito y el otro fracaso.
2. La probabilidad del resultado de cualquier intento permanece fijo con respecto al tiempo.
3. Los ensayos son estadísticamente independientes, es decir, el resultado de un ensayo en particular no es afectado por el resultado de cualquier otro ensayo.

Si *p* y *1-p* son las probabilidades de éxito y fracaso respectivamente en cada ensayo, entonces, la probabilidad de obtener *x* éxitos y *n-x* fracasos en algún orden específico se da por la siguiente ecuación: $p^x(1\text{-}p)^{n\text{-}x}$; entonces el número de formas en que podemos obtener *x* éxitos en *n* ensayos es el número de combinaciones de *x* objetos seleccionados de un conjunto de *n* objetos (*n/x*) así llegamos al siguiente resultado:

$$P(x) = \binom{n}{x} P^x q^{n-x}, \text{ Para } x = 0, 1, 2, 3, ..., n \qquad (8)$$

Donde,

n = número de ensayos realizados.
p = probabilidad de éxito.
q = (1-*p*) = probabilidad de fracaso.
n - *x* = número de fracasos deseados.
p^x = probabilidad favorable.

Recordando la fórmula de combinaciones, la ecuación 8 se transformará a la siguiente:

$$P(x) = \frac{n!}{x!(n-x)!} p^x q^{n\text{-}x} \qquad (9)$$

Aunque esta fórmula pueda parecer un tanto complicada, se le puede utilizar con bastante facilidad. El símbolo ! significa *factorial.* Por ejemplo factorial cinco (5! = 5 * 4 * 3 * 2 *1 = 120), 0! = 1. Utilizando la fórmula binomial para resolver nuestro problema descubrimos que la ecuación 9 puede desarrollarse como:

$$(q+p)^n = q^n + \binom{n}{1} q^{n-1} p + \binom{n}{2} q^{n-2} p^2 + \ldots + p^n \qquad (10)$$

Donde: $1, \binom{n}{1}, \binom{n}{2}$ .... , son *coeficientes binomiales*; *x* = número de éxitos deseados.

Ejemplo 2. Se lanza una moneda corriente 6 veces, donde llamamos cara a un éxito. Por consiguiente $n = 6$ y $p = q = 1/2$. Solo pueden ocurrir dos cosas (p ó q) por lo tanto la probabilidad de que ocurra 1 de ellas es la mitad, es decir ½.

a) La probabilidad de que suceda 2 caras exactamente (o sea $x = 2$) es:

$$P(2) = \frac{6!}{2!(6-2)!}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^4 = \frac{6!}{2!4!}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^4 = \frac{6x5x4x3x2x1}{(2x1)(4x3x2x1)}\left(\frac{1}{4}\right)\left(\frac{1}{6}\right) = \frac{30}{2}\left(\frac{1}{64}\right) = \frac{15}{64}$$

b) La probabilidad de conseguir por lo menos cuatro caras (o sea $x = 4$, 5 ó 6) es:

$$\frac{6!}{4!2!}\left(\frac{1}{16}\right)\left(\frac{1}{4}\right) + \frac{6!}{5!4!}\left(\frac{1}{32}\right)\left(\frac{1}{2}\right) + \frac{6!}{6!}\left(\frac{1}{64}\right) = \frac{6x5x4}{4!2x1}\left(\frac{1}{64}\right) + \frac{6*5!}{5!*4}\left(\frac{1}{64}\right) + \left(\frac{1}{64}\right) = \frac{15}{64} + \frac{6}{64} + \frac{1}{64} = \frac{22}{64} = \frac{11}{32}$$

Propiedades de la distribución binomial

- **a)** la media: $\mu = np$
- **b)** la varianza: $\sigma^2 = npq$
- **c)** el coeficiente de sesgo: $\alpha_3 = \frac{q-p}{\sqrt{npq}}$
- **d)** desviación típica: $\sigma = \sqrt{npq}$
- **e)** cuando p es menor que 0.5, la distribución binomial está sesgada hacia la derecha.
- **f)** conforme p aumenta, el sesgo es menos notable.
- **g)** cuando $p = 0.5$, la distribución binomial es simétrica.
- **h)** cuando p es mayor que 0.5, la distribución esta sesgada hacia la izquierda.

Ejemplo 3. En 100 tiradas de una moneda el # promedio de caras es $\mu = np = (100)(½) = 50$, este es el número esperado de caras en 100 lanzamientos. La desviación típica es:

$$\sigma = \sqrt{npq} = \sqrt{(100)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)} = 5$$

Distribución de probabilidad hipergeométrica

Recuérdese que si se selecciona una muestra aleatoria de n consumidores de una población de N consumidores, el número x de usuarios que favorecen un producto específico tendría una distribución binomial cuando el tamaño muestra n es pequeño respecto al número de N de consumidores en la población, el número x a favor del producto tiene una *distribución de probabilidad hipergeométrica*, cuya fórmula es:

$$P(x) = \frac{C_x^r C_{N-x}^{N-r}}{C_n^N} \qquad (11)$$

Donde:

N = número de elementos en la población.
r = número de elementos que tienen una característica especifica, por ejemplo el número de personas a favor un producto particular.
n = número de elementos en el muestra.

En base a la ecuación 11 podemos realizar las operaciones factoriales siguientes:

$$C_x^r = \frac{r!}{x!(r-x)!}$$

$$C_{n-x}^{N-r} = \frac{(N-r)!}{(n-x)!\{(N-r)-(n-x)\}!}$$

$$C_n^N = \frac{N!}{n!(N-n)!}$$

Medidas de tendencia central y de dispersión para la distribución hipergeométrica

La distribución hipergeométrica al igual que otras distribuciones de probabilidades tiene un valor esperado o media (μ) y una desviación estándar (σ), y vamos a ver la forma en que ambas medidas estadísticas se pueden calcular. Simbólicamente, podemos representar la media de una distribución hipergeométrica como:

Media aritmética $$\mu = \frac{nr}{N} \qquad (12)$$

En la que:

n = número de muestras.
r = número de elementos de la muestra con ciertas características.
N = tamaño de la población.

Y podemos calcular la variancia y la desviación estándar de una distribución hipergeométrica haciendo uso de la fórmula:

Variancia: $$\sigma^2 = \frac{r(N-r)n(N-n)}{N^2(N-1)} \qquad (13)$$

Desviación estándar:
$$\sigma = \sqrt{\frac{r(N-r)n(N-n)}{N^2(N-1)}} \qquad (14)$$

En la que:

σ^2 = la variancia.
σ = la desviación estándar.

Ejemplo 4. Un furgón contenía 20 computadoras electrónicas grandes, 2 de las cuales estaban defectuosas. Si se seleccionan al azar tres computadoras del furgón ¿cuál será la probabilidad de que dos de ellas tengan desperfectos?

Solución:

$N = 20$
$n = 3$
$r =$ 2(computadoras defectuosas)
$x =$ número de computadoras con averías en la muestra

Entonces,

$$P(x) = \frac{C_x^r C_{n-x}^{N-r}}{C_n^N}$$

$$P(2) = \frac{C_2^2 C_{3-2}^{20-2}}{C_3^{20}}$$

Donde,

$$\frac{1}{50} = 0.02$$

$$C_{3-2}^{20-2} = C_1^{18} = \frac{18!}{1!17!} = 18$$

$$C_3^{20} = \frac{20!}{3!17!} = 1140$$

Entonces la probabilidad de sacar $x = 2$ computadoras defectuosas en una muestra de $n = 3$ es:

$$P(2) = \frac{(1)(8)}{1140} = 0.016$$

La distribución de probabilidad geométrica

Se recordará que en un experimento binomial se tiene una serie de eventos idénticos e independientes, y que cada uno origina un "éxito" E o un "fracaso" F, con $p(E) = p$ y $p(F) = 1- p = q$. Si se interesa el número x de pruebas hasta la observación del primer éxito, entonces x posee una *distribución de probabilidad geométrica*. Nótese que el número de pruebas podría seguir indefinidamente y que x es un ejemplo de variable aleatoria discreta que puede tomar un número infinito (pero contable) de valores. Las fórmulas para la distribución geométrica son:

$$p(x) = pq^{x-1} \qquad x = 1,2,\ldots\ldots,8 \qquad (15)$$

Donde,
x = número de pruebas independientes hasta la ocurrencia del primer éxito.
p = probabilidad de éxito en una sola prueba, $q = 1 - p$.

Media: $$\mu = \frac{1}{p} \qquad (16)$$

Variancia: $$\sigma^2 = \frac{1-p}{p^2} \qquad (17)$$

Desviación estándar: $$\sqrt{\frac{1-p}{p^2}} \qquad (18)$$

La distribución de probabilidad geométrica es un modelo para el intervalo de tiempo que un jugador (o inversionista) tiene que esperar hasta ganar. Por ejemplo, si la ganancia media en una serie de apuestas idénticas en la ruleta (o en alguna otra serie de pruebas idénticas), no es una buena medida para su prospectiva de ganar, podría tener una racha de mala suerte y quedarse sin dinero antes de tener la posibilidad de recuperar sus pérdidas.

La distribución de probabilidad geométrica también proporciona un modelo discreto para el lapso, digamos el número x de minutos, antes de que un consumidor en una fila o línea de espera (en un supermercado, servicio de reparaciones, hospital, etc.) reciba la atención [nótese que el lapso o intervalo de tiempo es una variable aleatoria continua. La distribución de probabilidad geométrica es una analogía discreta de (una aproximación para) una distribución de probabilidad continua particular, conocida como distribución exponencial]. Este modelo discreto para la distribución de probabilidad de tiempo de espera x se basa en la suposición de que la probabilidad de recibir el servicio durante cualquier

minuto es idéntica e independiente del resultado durante cualquier otro minuto y que x se mide en minutos "enteros", es decir, $x = 1, 2, 3$.

Ejemplo 5. Los registros indican que una cierta vendedora tiene éxito en formular venta en 30% de sus entrevistas. Supóngase que una venta en una entrevista es independiente de una venta cualquier otro momento.

a) ¿Cuál es la probabilidad de que esta vendedora tenga que tratar con 10 personas antes de hacer su primera venta?
b) ¿Cuál es la probabilidad que la primera venta se realice antes o en la décima oportunidad?

Solución:

a) La probabilidad de realizar una venta en un solo contacto o entrevista es $p = 0.3$ Entonces la probabilidad de que necesita exactamente $x = 10$ contactos antes de hacer su primera venta es:

$$p(x) = pq^{x-1}$$

o bien $$p(10) = (0.3)(0.7)^9 = 0.012$$

b) La probabilidad de que se realice la primera venta antes de o en la décima entrevista, es

$$P(x = 10) = p(0) + p(1) + p(2) + \ldots\ldots + p(10)$$

La manera más sencilla de hallar esta probabilidad es expresarla como el complemento de una serie infinita, es decir:

$$P(x = 10) = 1 - P(x > 10)$$

Donde,

$$P(x > 10) = p(11) + p\,(12) + p\,(13) + \\ = pq^{10} + pq^{11} + \ldots$$

La suma de una serie geométrica infinita es igual a $a\,/\,(1 - r)$, donde a es el primer término de la serie, r es la razón, común, y $r^2 < 1$. Utilizando este resultado, tenemos.

$$P(x = 10) = 1 - p(x \succ 10) = 1 - \frac{pq^{10}}{1-q} = 1 - q^{10} = 1 - (0.7)^{10} = 0.972$$

Por lo tanto existe una alta probabilidad (0.972) de que la vendedora realice primera venta antes de o en el décimo contacto.

La distribución de probabilidad normal

Hasta ahora hemos presentado diversas variables aleatorias discretas y sus distribuciones de probabilidad. Ahora fijaremos nuestra atención a los casos en que la variable puede tomar cualquier valor que esté en un intervalo de valores dado, y en los cuales la distribución de probabilidad es continua. El objetivo de la distribución de probabilidad normal es conducir la variable aleatoria normal, una de las variables aleatorias continuas más importantes y que se utiliza con mayor frecuencia. Se da su distribución de probabilidad y se muestra cómo puede emplearse la distribución de probabilidad (Badii et al., 2007a).

Varios matemáticos han contribuido al desarrollo de la distribución normal, entre los que podemos contar al astrónomo-matemático del siglo XIX, Karl Gauss. En honor a su trabajo, la distribución de probabilidad normal a menudo también se le llama distribución Gaussiana.

Existen dos razones básicas por las cuales la distribución normal ocupa un lugar tan prominente en la estadística. Primero, tiene algunas propiedades que la hacen aplicable a un gran número de situaciones en las que es necesario hacer inferencias mediante la toma de muestras. Segundo, la distribución normal casi se ajusta a las distribuciones de frecuencias reales observadas en muchos fenómenos, incluyendo características humanas (peso, altura, IQ), resultados de procesos físicos y muchas otras medidas de interés para los investigadores, tanto en el sector público como en el privado.

Variables aleatorias continuas

Una variable aleatoria continua es la que puede tomar un número infinitamente grande de valores que corresponden a los puntos en un intervalo de una recta. Las estaturas y los pesos de las personas, el tiempo entre dos eventos o la vida útil de un equipo de oficina, son ejemplos típicos de variables aleatorias continuas.

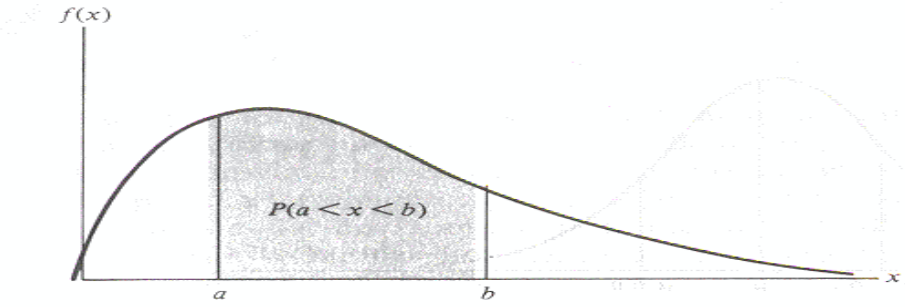
El modelo probabilístico para la distribución de frecuencias de una variable aleatoria continua implica la selección de una curva, generalmente regular o aislada, a la que se llama *distribución de probabilidad o función de densidad de probabilidad* de una variable aleatoria. Si la ecuación de esta distribución de probabilidad continua es *f(x)*, entonces la probabilidad de que x esté en el intervalo $a < x < b$ es el área bajo la curva de distribución para *f(x)* entre los dos puntos a y b (Figura 1).

Esto concuerda con la interacción de un histograma de frecuencias relativas, donde las áreas sobre un intervalo, bajo el histograma, correspondieron a la proporción de observación que cae en dicho intervalo. Ya que el número de valores que puede tomar x es infinitamente grande y no se puede contar, la probabilidad de que x sea igual a un valor específico, por ejemplo a, es 0.

Entonces las afirmaciones probabilísticas acerca de las variables aleatorias continuas siempre corresponden a áreas bajo la distribución de probabilidad sobre un intervalo por ejemplo, de a a b, y se expresan como $P(a < x < b)$. Nótese que la probabilidad en $a < x < b$ es igual a la probabilidad de que $a < x = b$, pues $P(x = a) = P(x = b) = 0$. Hay muchas distribuciones de probabilidad continuas y cada una se representa mediante una ecuación

f(x), que se escoge de la manera que el área total bajo la curva de distribución de probabilidad sea igual a 1.

Figura 1. Distribución de probabilidad para una variable aleatoria continua.



Una vez que conocemos la ecuación *f(x)* de una distribución de probabilidad particular se pueden encontrar probabilidades específicas, por ejemplo, la probabilidad de que *x* esté en el intervalo $a < x < b$, de dos maneras. Podemos graficar la ecuación y utilizar métodos numéricos para aproximar el área sobre el intervalo $a < x < b$. Este cálculo puede realizarse utilizando métodos muy aproximados o una computadora para obtener cualquier grado de precisión. O bien, si *f(x)* tiene una forma particular, podemos usar el cálculo integral para encontrar $P(a < x < b)$. Afortunadamente, no hay que utilizar en la práctica, ninguno de estos métodos, porque se han calculado y tabulado las áreas bajo la mayoría de las distribuciones de probabilidades continuas más empleadas.

Estudiaremos en las secciones siguientes, una de las distribuciones de probabilidad continuas de mayor uso la distribución de probabilidad normal (de campana). La distribución de probabilidad normal se trata de enseñar a encontrar la probabilidad de un suceso por medio de la curva normal y la tabla de las áreas bajo la curva normal. La distribución normal se utiliza cuando existe una variable aleatoria continua, donde dicha variable puede asumir cualquier valor de una gama de ellos y por tanto la distribución de probabilidad es continua. La distribución normal representa las siguientes propiedades:

1. La curva es simétrica, tiene un solo pico, por consiguiente es unimodal, presenta una forma de campana.
2. La media de una población distribuida normalmente se encuentra en el centro de su curva normal.
3. A causa de la simetría de la distribución normal de probabilidad, la media, la moda y la mediana de la distribución se encuentran también en el centro; en consecuencia, para una curva normal, la media, la mediana y la moda tienen el mismo valor.
4. Teóricamente, la curva se extiende en ambas direcciones, y tiende gradualmente a unirse con el eje horizontal. Sin embargo, se extiende al infinito, sin tocar nunca el eje de la abscisa.

Por consiguiente, uno de los más importantes ejemplos de una distribución de probabilidad continua es la *distribución normal, curva normal o distribución Gaussiana*, definida por la ecuación:

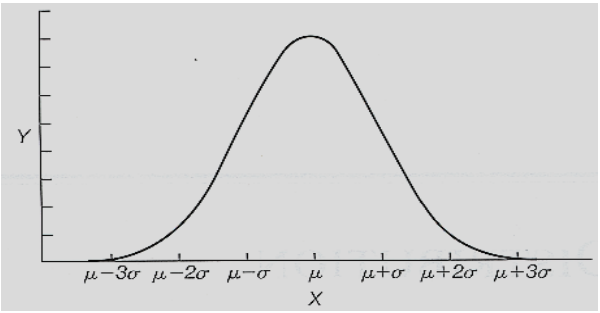
$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(X-\mu)^2/\sigma^2} \qquad (19)$$

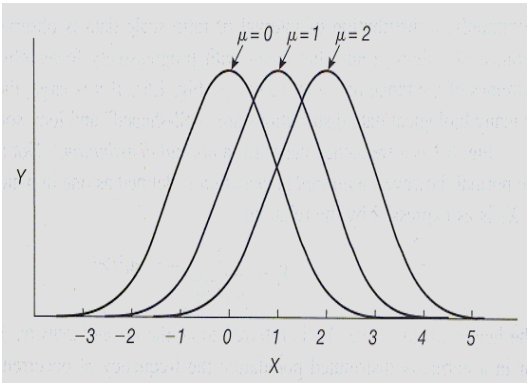
Donde:
μ = la media.
σ² = la varianza.
Σ = la desviación típica.
π = constante (3.14159).
e (exponencial) = 2.71828.

La ecuación de una distribución normal con μ = 0 y σ = 1 (una distribución normal estandarizada) es igual a:

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \qquad (20)$$

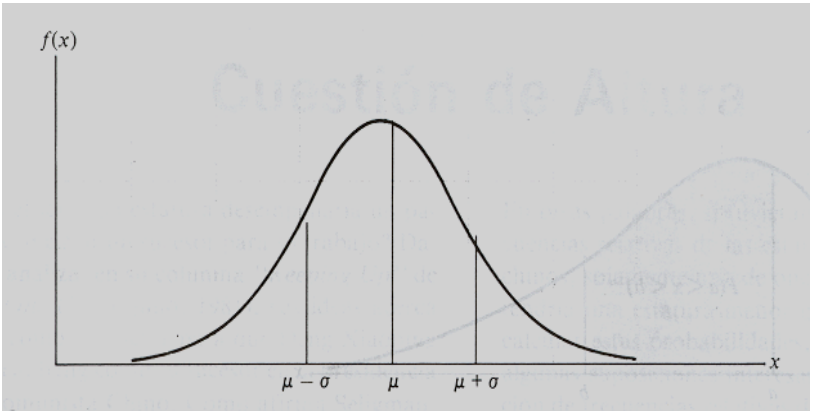
Figura 2. a) Una distribución normal, b) Una distribución normal con varianzas iguales y medias aritméticas





desiguales.

Figura 3. Función de densidad de probabilidad normal.



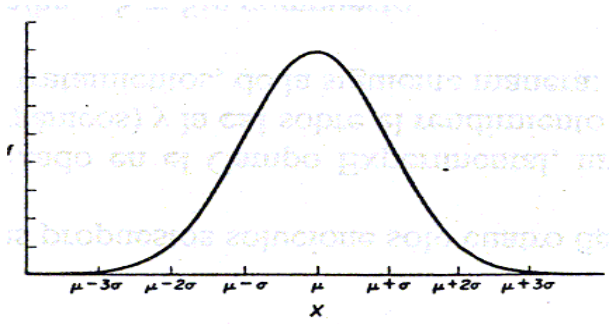
La distribución de probabilidad normal, mostrada en la Figura 3 (una desviación estándar de σ), es simétrica respecto a la media μ. En la practica se encuentran pocas veces variables que cambien de "menos infinito" a "más infinito", cualquier que sea el significado que se debe atribuir a estas expresiones. Ciertamente, la estatura de personas o la vida útil de un equipo de oficina no satisfacen estos requisitos, sin embrego, el histograma de frecuencia relativa para muchas de las mediciones tiene forma acampanada se puede aproximar con la función mostrada en la Figura 3.

El área total limitada por la curva y el eje x es 1; por tanto, el área bajo la curva entre $x = a$ y $x = b$, con $a < b$, representa la probabilidad de que x esté entre a y b. Esta probabilidad se denota por $P\{a < x < b\}$, y se calcula mediante la fórmula:

$$\left[z = \left(X - \mu\right) / \sigma\right] \qquad (21)$$

Donde: z = la desviación normal con la media igual a cero y desviación típica σ = 1.

Figura 4. Una distribución canónica.



Cuando se expresa la variable x en unidades estándares, la ecuación (20) es reemplazada por la llamada forma *canónica* (21). La Figura 4 es un gráfico de esta forma *canónica,* y muestra que las áreas comprendidas entre $z = \pm 1$, $z = \pm 2$, y $z = \pm 3$ son iguales, respectivamente, a 68.27%, 95.45% y 99.73% del área total, que es 1. La tabla de Z en cualquier libro de estadística muestra las áreas bajo esta curva acotadas por las ordenadas $z = 0$ y cualquier valor positivo de z. De esta tabla se puede deducir el área entre todo par de coordenadas usando la simetría de la curva respecto de $z = 0$.

Tabulaciones de las áreas de la distribución de la probabilidad normal

No importa cual sean los valores de μ y σ para una distribución de probabilidad normal, el área total bajo la curva es 1, de manera que podemos pensar en áreas bajo la curva como si fuera probabilidades. Matemáticamente es verdad que:

1. Aproximadamente 68% de todos los valores de una población normalmente distribuida se encuentra dentro de ± 1 desviación estándar de la media.

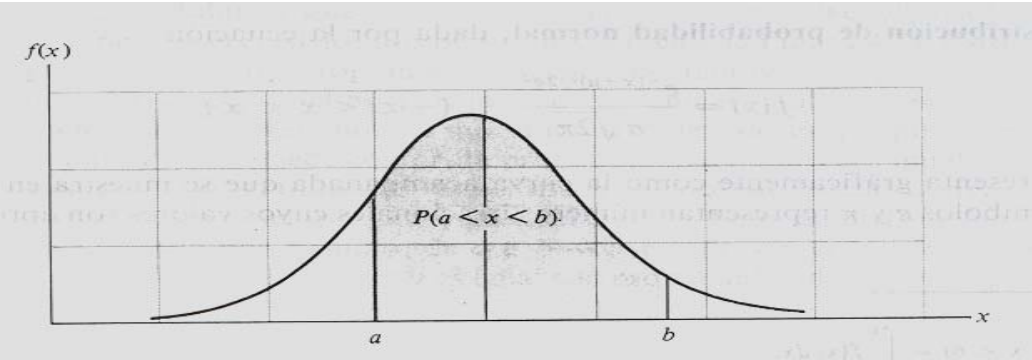
2. Aproximadamente 95.5% de todos los valores de una población normalmente distribuida se encuentran dentro de ± 2 desviaciones estándar de la media.
3. Aproximadamente 99.7% de todos los valores de una población normalmente distribuida se encuentran dentro de ± 3 desviaciones estándar de la media.

Recuérdese que la probabilidad de una variable aleatoria continua toma un valor en el intervalo de *a* hasta *b*, es el área bajo la función de la densidad de probabilidad, entre los puntos *a* y *b* **(**Figura 5) a fin de evaluar las áreas bajo la curva normal. En base con los valores numéricos de μ.y σ podemos generar un número infinitamente grande de distribuciones normales dando diversos valores a estos parámetros. Obviamente, no es práctico tener en tablas separadas las áreas de estas curvas, sino conviene tener una tabla de áreas aplicable a todas las curvas.

La manera fácil de utilizar una sola tabla, es trabajar con áreas situadas dentro de un número específico de desviaciones estándares respecto a la media como se hizo en la caso de la Regla Empírica. Por ejemplo, sabemos que aproximadamente 0.68 de esta área estará dentro de una desviación estándar de la media, 0.95 dentro de dos y casi la totalidad, dentro de tres. ¿Qué fracción del área total caerá dentro de 0.7 desviaciones estándares? A esta pregunta y a otras se les dará respuesta con la Tabla *Z*.

Como la curva normal es simétrica respecto a su media, la mitad del área bajo la curva se encuentra a la izquierda de la media y la otra mitad a la derecha. También, debido a la simetría podemos simplificar la tabla de las áreas listándolas entre la media y un número especificado z de desviaciones estándares a la derecha de la media (μ). Las áreas a la izquierda de la media se pueden calcular utilizando el área correspondiente e igual a la de la derecha de la medida. La distancia de un valor de x a la medida es (x - μ). Al expresar esta distancia en unidades de desviaciones estándar (σ), obtenemos: $z = (x - \mu) / \sigma$.

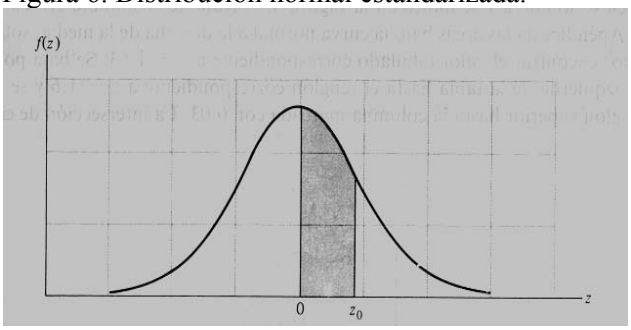
Figura 5. Probabilidad $p(a < x < b)$ para una variable aleatoria distribuida normalmente.



Nótese que hay una correspondencia de uno a uno entre z y x, en particular, cuando $x = \mu$, $z = 0$,. El valor de z será positivo cuando x esté por arriba de la media, y negativo cuando x sea menor que dicha media. La distribución de probabilidad de z muchas veces se designa por *distribución normal estandarizada* pues su media es igual a cero y su desviación

estándar es igual a uno. El valor bajo la curva normal entre la media $z = 0$ y un valor especificado de $z>o$, por ejemplo z_o es la probabilidad $P(o \leq z \leq z_0)$. Esta área se registra en la Tabla Z y se identifica como el área sombreada en la Figura 6.

Figura 6. Distribución normal estandarizada.



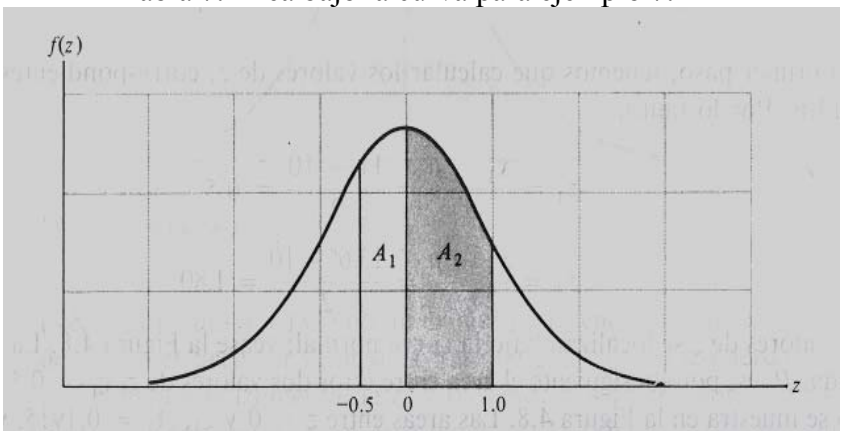
Ejemplo 6. Obtenga $p(0 \leq z \leq 1.63)$. Esta probabilidad corresponde al área entre la media ($z = 0$) y un punto $z = 1.63$ desviaciones estándares a la derecha de la media.

Solución: De la tabla de z de las áreas bajo la curva normal a la derecha de la media, solamente necesita encontrar el valor tabulado correspondiente a $z = 1.63$. Se baja por la columna de la izquierda de la tabla hasta el renglón correspondiente a $z = 1.6$ y se va luego por el renglón superior hasta la columna marcada con 0.03. La intersección de esta combinación de renglón da el área A = 0.4484. Por lo tanto, $P(0 < z < 1.63) = 0.4484$.

Ejemplo 7. Calcular $P(-0.5 \leq z \leq 1.0)$, que corresponde al área entre $z = -0.5$ y $z = 1.0$.

Solución: El área requerida es igual a la suma de A_1 y A_2 mostrada en la Figura 7. De la Tabla obtenemos $A_2 = 0.3413$. El área A_1 es igual al área correspondiente entre $z = 0$ y $z = 0.5$ o bien $A_1 = 0.1915$. Por lo tanto: $A = A_1 + A_2 = 0.1915 + 0.3413 = 0.5328$.

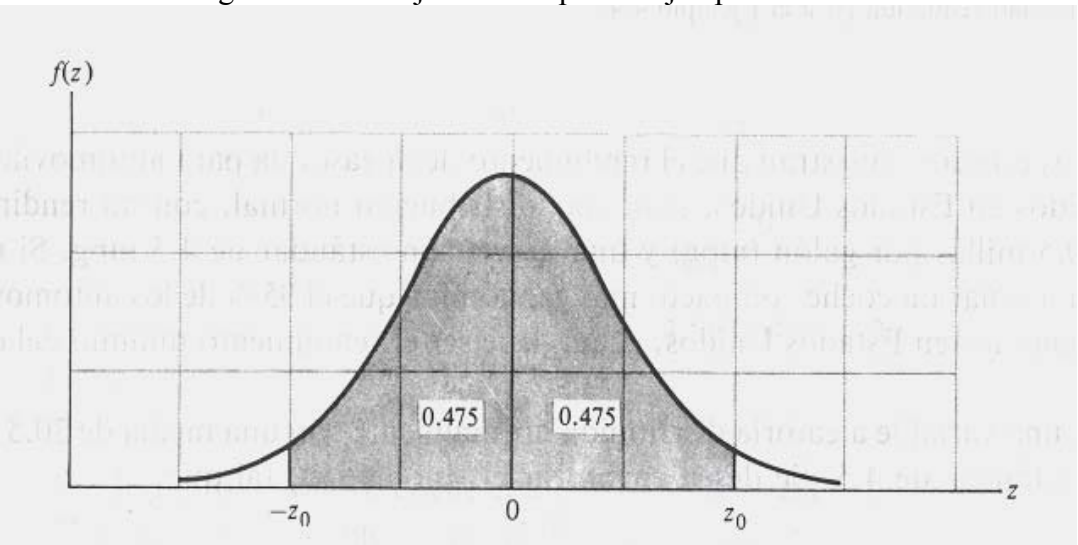
Tabla 7. Área bajo la curva para ejemplo 7.



Ejemplo 8. Hallar los valores de z, por ejemplo z_o tales que exactamente (hasta cuatro cifras decimales) 0.95 del área quede dentro de $\pm z_0$ desviaciones estándares de la media.

Solución: La mitad del área de 0.95 se encontrara a la izquierda de la media y la otra mitad a la derecha, porq la simetría de la distribución normal, por lo tanto, se desea encontrar el valor z_o que corresponde a un área igual a 0.475. Esta área esta sombreada en la Figura 8. Al referirnos a la Tabla Z se ve que el área 0.475 está en el renglón correspondiente a z = 1.9 y en la columna 0.06. Por lo tanto z_o = 1.96. Nótese que este resultado está muy cerca del valor aproximado z = 2 que se utiliza en la regla empírica.

Figura 8. Área bajo la curva para el ejemplo 8.



…

Ejemplo 9. Sea x una variable aleatoria distribuida normalmente con una media igual a 10 y una desviación estándar igual a 2. Encuentre la probabilidad de que x esté entre 11 y 13.6.

Solución: Como primer paso, tiene que calcular los valores de z, correspondientes a x = 11 y x = 13.6. Por lo tanto:

$$z_1 = (x_1-\mu)/\sigma = (11-10)/2 = 0.5$$

$$z_2 = (x_2-\mu)/\sigma = (13.6-10)/2 = 1.80$$

La probabilidad deseada P, por consiguiente, es el área entre estos dos valores de z, z_1 = 0.5 y z_2 = 1.80, Las áreas entre z = 0 y z_1, A_1 = 0.1915 y entre z = 0 y z_2, A_2 = 0.4641, se obtiene d la Tabla Z. La probabilidad P es igual a la diferencia entre las dos áreas A_1 y A_2; es decir: $P = A_2 - A_1 = 0.4641 - 0.1915 = 0.2726$.

Ejemplo 10. Ciertos estudios muestran que el rendimiento de la gasolina para automóviles compactos vendidos en Estados Unidos, tienen distribución normal, con un rendimiento

medio de 30.5 millas por galón (mpg) y una desviación estándar de 4.5 mpg. Si un fabricante desea diseñar un coche compacto más económico que el 95 % de los automóviles compactos vendidos en Estados Unidos, ¿cuál debe ser el rendimiento mínimo del coche nuevo?

Solución: Sea x una variables aleatoria distribuida normalmente con una media de 30.5 y una desviación estándar de 4.5. Se desea encontrar el valor de x_o tal que

$$P\,(x< x_o) = 0.95$$

Solución: Como un primer paso encuentra el valor de z_o tal que el área a la izquierda sea igual a 0.95. Puesto que el área a la izquierda de $z = 0$ es 0.5, z_0 será el valor de z en la Tabla que corresponde a una área igual a 0.45. Este valor es $z_0 = 1.645$. El paso final es encontrar el valor x_0 correspondiente a $z_0 = 1.645$. Se obtiene utilizando la ecuación que relaciona x y z a saber:

$$z = \frac{x - \mu}{\sigma}$$

Donde, $\mu = 30.5$ y $\sigma = 4.5$. Al sustituir los valores de μ, σ y z_o en esta ecuación y despejando x_o resulta:

$$1.645 = \frac{x_0 - 30.5}{4.5}$$

$$x_0 = (4.5)(1.645) + 30.5 = 37.9$$

Por lo tanto, el nuevo coche compacto del fabricante debe desarrollar un rendimiento de 37.9 mpg, para ser mejor que el 95 % de los coches compactos que actualmente se venden en Estados Unidos.

Ejemplo 11.

a) ¿Cuál es la proporción de reclutas que tienen un C.I. entre 100 y 105.7? Sea $\mu = 100$ y $\sigma = 10$.

$$z = (x - \mu)/\sigma = (105.7 - 100) / 10 = 0.57$$

En la tabla normal de área hallamos 0.2842 de la siguiente forma:

$$P(x \geq 105.7) = P\left(\frac{105.7 - 100}{10}\right) \geq P(z \geq 0.57) = 0.284$$

Así la proporción que deseamos es 0.5000 - 0.2843 = 0.2157

b) ¿Cuál es la proporción de reclutas entre 103 y 105.7?

$$z = (x - \mu)/\sigma = (103 - 100) / 10 = 0.3$$

$$P(103 \leq x \leq 105.7)$$

$$p\left(\frac{103-100}{10} \leq \frac{105.7-100}{10}\right)$$

$$= P(0.3 \leq z \leq 0.57) = 0.3821 - 0.2843 = 0.0978.$$

Con la tabla normal de áreas, se deduce que la proporción del área a partir del extremo es 0.3821. Se conoce que la proporción que corresponde a 105.7 es 0.2483, entonces el área rayada que se busca será: 0.3821 - 0.2843 = 0.0978.

c) ¿Qué proporción de reclutas tienen un C.I. inferior a 83.6?

$$z = (x - \mu)/\sigma = (83.6 - 100) / 10 = -1.64$$

En la tabla normal se encuentran que es 0.0505

d) ¿Cuál es la proporción superior a 120?

$$z = (x - \mu)/\sigma = (120 - 100) / 10 = 2$$

$$P(X \geq 130) = P(Z \geq 2) = 0.0228$$

Desventajas de la distribución normal

Hemos notado que los extremos de la distribución normal se acercan al eje horizontal, pero nunca llegan a tocarlo. Esto implica que existe algo de probabilidad (aunque puede ser muy pequeña) de que la variable aleatoria pueda tomar valores demasiado grandes. Debido a la forma del extremo derecho de la curva, es posible que la curva de la distribución normal asigne una probabilidad minúscula a la existencia de una persona que pese dos toneladas. Desde luego, nadie creería en la existencia de tal persona. Un peso de una tonelada o más estaría a aproximadamente 50 desviaciones estándar a la derecha de la media y tendría una probabilidad con 250 ceros justo después del punto decimal. No perdemos mucha precisión al ignorar valores tan alejados de la media. Pero a cambio de la conveniencia del uso de este modelo teórico, debemos aceptar el hecho de que puede asignar valores empíricos imposibles.

La distribución normal como una aproximación de la distribución binomial

Aunque la distribución normal es continua, resulta interesante hacer notar que algunas veces puede utilizarse para aproximar la distribución binomial, suponga que nos gustaría saber la probabilidad de obtener 5, 6, 7 u 8 caras en diez lanzamientos de una moneda no alterada:

$$P(5, 6, 7 \text{ u } 8) = P(5) + P(6) + P(7) + P(8)$$
$$= 0.2461 + 0.2051 + 0.1172 + 0.0439 = 0.6123$$

Para $n = 10$ y $p = ½$ se puede calcular la media ($\mu = np = 10\ (½) = 5$) y desviación estándar = $(\sigma = \sqrt{npq} = \sqrt{10x1/2x1/2} = 1.58)$.

Observe el área bajo la curva normal entre 5 ± ½. Nos damos cuenta de que esta área es de *aproximadamente* el mismo tamaño que el área de la barra que representa la probabilidad binomial de obtener 5 caras. Los dos ½ que agregamos y restamos a cinco se conocen como *factores de corrección de continuidad* y se utilizan para mejorar la precisión de la aproximación.

Al usar los factores de corrección de continuidad, vemos que la probabilidad binomial de obtener 5, 6, 7 u 8 caras puede ser aproximada por el área bajo la curva normal entre 4.5 y 8.5. Determine esta probabilidad mediante el cálculo de los valores de z correspondientes a 4.5 y 8.5.

$$z_1 = \frac{x - \mu}{\sigma} = \frac{4.5 - 5}{1.581} = -0.32$$ desviación estándar

$$z_2 = \frac{8.5 - 5}{1.581} = 2.21$$ desviación estándar

$p(z_1 \leq -0.32) = p(z_1 \geq 0.32) = 0.1255$ de que x esté entre 4.5 y 5.

$p(z_2 \geq 2.21) = 0.4864$ correspondiente de que x esté entre 5 y 8.5.

La probabilidad de que x esté entre 4.5 y 8.5:

$$A = 0.1255 + 0.4864 = 0.6119$$

Comparando la probabilidad binomial de 0.6123 (Tabla Z) con la aproximación normal de 0.6119, vemos que el error en la aproximación es menor a 1/10 (1%).

La aproximación normal a la distribución binomial resulta muy conveniente, pues nos permite resolver el problema sin tener que consultar grandes tablas de la distribución binomial. Debemos hacer notar que se necesita tener algo de cuidado al utilizar esta aproximación, que es bastante buena siempre y cuando np y nq sean de al menos cinco.

La distribución de probabilidad de Poisson

La *distribución de probabilidad de Poisson* debe su nombre a Siméon Denis Poisson (1781-1840), un francés que desarrollo la distribución en el año 1834 a partir de los estudios sobre esta distribución (Badii et al., 2000). La distribución de Poisson es un buen modelo para la distribución de frecuencias relativas del número de eventos raros que ocurren en una unidad de tiempo, de distancia, de espacio, etcétera. Por esta razón se utiliza mucho en área de investigación científica tanto en administración de empresas como en las actividades biológicas para modelar la distribución de frecuencias relativas del número de accidentes industriales por unidad de tiempo (como el accidente en la planta nuclear de Three Mile Island) o por administradores de personal, para modelar la distribución de frecuencias relativas del número de accidentes de los empleados o el número de reclamaciones de seguros, por unidad de tiempo, o la frecuencia de enfermedades raras que ocurre una población dada. La distribución de probabilidad de Poisson puede proporcionar, en algunos casos, un buen modelo para la distribución de frecuencias relativas del número de llegadas por unidad de tiempo a una unidad de servicio (por ejemplo, el número de pedidos recibidos en una planta manufacturera o el número de clientes que llegan a una instalación de servicio, a una caja registradora en un supermercado, etc.).

Características de la distribución de Poisson

1. Las consecuencias de los eventos son independientes. La ocurrencia de un evento en un intervalo de espacio o tiempo no tiene efecto sobre la probabilidad de una segunda ocurrencia del evento en el mismo, o cualquier otro intervalo.
2. Teóricamente, debe ser posible un número infinito de ocurrencias del evento en el intervalo.
3. La probabilidad de la ocurrencia única del evento en un intervalo dado es proporcional a la longitud del intervalo.

Una Particularidad de la distribución de Poisson es el hecho de que la media y la varianza son iguales.

Cálculo de la probabilidad de Poisson

La distribución de probabilidad de Poisson, tiene que ver con ciertos procesos que pueden ser descritos por una variable aleatoria discreta. La letra *X* por lo general representa a esta variable discreta y puede tomar valores enteros (0, 1, 2, 3, 4, etc.). Utilizamos la letra mayúscula *X* para representar a la variable aleatoria y la letra minúscula *x* para señalar un valor específico que dicha variable puede tomar. La probabilidad de tener exactamente *x* presentaciones en una distribución de Poisson se calcula con la fórmula:

$$P(x) = f(x) = \frac{e^{-\lambda}\ \lambda^{x}}{x!} \qquad (22)$$

Donde:

P(x) = probabilidad de tener exactamente *x* presentaciones.

$e^{-\lambda}$ = Exponencial = 2.71828 (base de los logaritmos naturales), elevada a la lamda potencia negativa.

$\lambda^x = \lambda$ (el número medio de presentaciones por intervalo de tiempo) elevada a la x potencia.

$x! = x$ factorial.

λ = parámetro de distribución o la media donde $\lambda = p(x)$, es el número promedio de ocurrencias del evento aleatorio por intervalo de tiempo.

X = número de eventos raros por unidad de tiempo de distancia de espacio.

Ejemplo 12. Un administrador de un hospital ha estado estudiando las admisiones diarias de emergencia durante un periodo de varios años, los estudios revelan que en dicho periodo en promedio se presentaron 3 emergencias por día: encuentre la probabilidad de que:

a) En un día dado ocurran sólo dos admisiones de emergencia.

Solución: En este ejemplo $\lambda = 3$ que es igual al valor promedio de ocurrencia en la población y $x = 2$ como una variable aleatoria discreta. La probabilidad de ocurrencia se calcula como:

$$p(x=2) = f(2) = \frac{e^{-3}3^2}{2!} = \frac{(0.05*9)}{2*1} = 0.225$$

b) ¿Cuál es la probabilidad de que en un día particular no ocurra ni una sola admisión de emergencia?

$$f(x) = \frac{e^{-3}\ 3^0}{0!} = \frac{0.050*1}{1} = 0.05$$

c) En un día particular sean admitidos 3 ó 4 casos de emergencia.

Dado que los dos eventos son mutuamente exclusivos se usa la regla de adición:

$$f(3) + f(4) = \frac{e^{-3}3^3}{3!} + \frac{e^{-3}3^4}{4!} = \frac{(0.05*27)}{3*2*1} + \frac{(0.05*81)}{4*3*2*1} = 0.225 + 0.16875 = 0.39$$

Para los casos en que se desea obtener la probabilidad de ocurrencia de x o menos frecuencias del evento en cuestión que tiene el número promedio de ocurrencias igual a λ se debe utilizar la tabla de distribución de Poisson hecha exclusivamente para estos casos.

Ejemplo 13. En el estudio de cierto organismo acuático se tomaron gran número de muestras de un estanque y se contó el número promedio de organismos igual a dos. Encuentre la probabilidad de que:

a) La siguiente muestra que se tome contenga uno o más organismos.

Solución: La sumaria de todas las posibles situaciones que se puedan presentar es igual a uno; por lo tanto si se desea obtener la probabilidad de que la muestra tenga uno ó más organismos, solo necesito restarle a uno la probabilidad de que obtenga 0 organismos. Es decir, $P(x \geq 1) = 1 - P(x = 0)$. En la tabla de distribución acumulada de Poisson se ve que cuando $\lambda = 2$ la probabilidad de que $x = 0$ es 0.135, por lo tanto $P(x \geq 1) = 1 - 0.135 = 0.865$.

b) La siguiente muestra que se tome contenga exactamente 3 organismos. Este caso se puede resolver la probabilidad de la otra manera [con $p(x = 3)$], pero para explicar el uso de esta tabla lo resolveremos de la siguiente manera:

$$P(x = 3) = P(x \leq 3) - P(x \leq 2) = 0.857 - 0.677 = 0.180$$

Si la probabilidad de que ocurra 3 es igual a la probabilidad de $x \leq 3$ menos la probabilidad de que ocurra $x \leq 2$.

c) La siguiente muestra que se tome contenga menos de cinco organismos. Ya que el conjunto de menos de cinco organismos no incluye a 5 se está pidiendo la probabilidad acumulada desde 0 hasta 4, llegando a: $P(x \leq 4) = 0.947$

Nótese que x es normalmente pequeño en la práctica; teóricamente podría ser muy grande, sin límite. Por lo tanto, la variable aleatoria de Poisson es un ejemplo de variable aleatoria discreta que puede tomar un número infinito (pero contable) de valores.

La distribución de probabilidad binomial, cuando n es grande y p es pequeño, y cuando la media $\mu = np$ de la distribución de probabilidad binomial es aproximadamente menor que 7. Esta aproximación elimina el cálculo tedioso necesario para determinar las probabilidades binomiales cuando n es grande.

Ilustramos estos dos tipos de aplicaciones en los ejemplos siguientes.

Ejemplo 14. Las lesiones laborales graves que ocurren en una planta siderúrgica, tienen una media anual de 2.7. Dado que las condiciones de seguridad serán iguales en la planta durante el próximo año, ¿cuál es la probabilidad de que el número de lesiones graves sea menor que dos?

Solución: El evento de que ocurrirán menos de dos lesiones graves, es el evento que $x = 0$ o bien $x = 1$, por lo tanto,

$$P(x<2) = p(1) \text{ en donde, p (x)} = \frac{(2.7)^x e^{-2.7}}{x}$$

Sustituyendo en la fórmula para $p(x)$, obtenemos:

$$P(x < 2) = P(0) + P(1) = \frac{(2.7)^0 0.067206}{0} + \frac{(2.7)^1 0.067206}{1} = 0.249$$

Recuérdese que 0! = 1, por lo tanto, la probabilidad de que haya menos de dos lesiones laborales graves el próximo año en la planta fabril de acero, es 0.249.

Por conveniencia se proporciona en la tabla de Poisson**,** las sumas parciales, para la distribución de probabilidad de Poisson según valores de μ desde 0.25, hasta 5.0, con incrementos de 0.25. El ejemplo siguiente ilustrara el uso de la distribución de Poisson para aproximar la distribución de probabilidad binomial.

Ejemplo 15. Supóngase que se tiene un experimento binomial con $n = 25$ y $p = 0.1$. Hallar el valor exacto de $P(x = 3)$, utilizando la una tabla de Poisson, de las sumas parciales para la distribución de probabilidad de Poisson. Compare el valor aproximado con el valor exacto para $P(x = 3)$.

Solución: De la Tabla de Poisson, el valor exacto $P(x = 3)$ es la suma de $P(x) = 0.764$. El valor correspondiente con parcial de Poisson donde $\mu = np = (25)(0.1) = 2.5$ se da en la Tabla Poisson: $P(x = 3) = P(x) = 0.758$. Al comparar ambos resultados, vemos que la aproximación es bastante buena. Solamente difiere en 0.006 del valor exacto.

Búsqueda de probabilidades de Poisson utilizando la tabla de Poisson

En la tabla de Poisson se tienen los mismos resultados que si hiciéramos los cálculos, pero nos evitamos el trabajo tedioso. Por ejemplo, los registros indican el número promedio de accidentes en un crucero es igual a 5 accidentes mensuales. Si deseamos calcular la probabilidad de que cualquier mes ocurran 4 accidentes. Podemos utilizar la tabla de Poisson para evitar el tener que calcular *e* elevadas a potencias negativas. Aplicando la fórmula:

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

$$P(4) = \frac{(5)^4 e^{-5}}{4!} = 0.17552$$

Para utilizar esta tabla, todo lo que necesitamos saber son los valores de x y de λ (lamda), en este ejemplo 4 y 5, respectivamente. Ahora busque en la tabla, primero encuentre la columna cuyo encabezado es 5; luego recórrala hacia abajo hasta que esté a la altura del 4 y lea la respuesta directamente, 0.1755.

Relaciones entre la distribución normal, binomial y Poisson

Si el tamaño de la muestra es grande y si ni p ni q son muy próximos a cero, la distribución binomial puede aproximarse estrechamente por una distribución normal con variable canónica dada por:

$$z = \frac{x - np}{\sqrt{npq}} \qquad (23)$$

La aproximación mejora al aumentar la *n*, y en el límite exacto; esto se muestra en las propiedades de ambas distribuciones, donde es claro que al crecer *n*, el sesgo y la curtosis de la distribución binomial se aproximan a los de la distribución normal. En la práctica, la aproximación es muy buena si tanto *np* como *nq* son número mayores.

En la distribución binomial, si *n* es grande y la probabilidad *p* de ocurrencia de un suceso es muy pequeña, de modo que *q* = 1- *p* es casi 1, el suceso se llama un suceso raro. En la práctica, un suceso se considera raro si el número de ensayos es al menos 50 (*n* = 50) mientras *np* es menor que 5. En tal caso, la distribución binomial queda aproximada muy estrechamente por la distribución de Poisson con μ= *np*. Esto se comprueba comparando las propiedades, pues al poner μ = *np*, *q* ≈1 y *p* ≈ 0 de las propiedades binomial obtenemos las propiedades de poisson.

Como hay una relación entre la distribución binomial y la distribución normal, se sigue que también están relacionadas la distribución de Poisson y la distribución normal. De hecho, puede probarse que la distribución de Poisson tiende a una distribución normal con variable canónica (Tabla 7).

Tabla 7. Relación entre distribuciones normales, binomial y de Poisson.

Parámetro	Normal	Binomial	Poisson
Media	μ	$\mu = np$	$\mu = \lambda$
Varianza	σ^2	$\sigma^2 = npq$	$\sigma^2 = \lambda$
Desviación típica	σ	$\sigma = \sqrt{npq}$	$\sigma = \sqrt{\lambda}$
Coeficiente de sesgo	$\alpha_3 = 0$	$\alpha_3 = (q - p)/\sqrt{npq}$	$\sigma_3 = \frac{1}{\sqrt{\lambda}}$
Coeficiente de curtosis	$\alpha_4 = 3$	$\alpha_4 = 3 + [(!-6pq\ )/npq\ ]$	$\sigma_4 = 3 + \frac{1}{\lambda}$
Desviación media	$\sigma\sqrt{2/\pi} = 0.7979$		

Algunas veces si se desea evitar el tedioso trabajo de calcular las distribuciones binomiales, se puede usar en cambio la de Poisson. Esta última es una aproximación razonable de la distribución binomial, pero sólo en determinadas circunstancias. Estas condiciones se cumplen cuando *n* es grande y *p* es pequeña; es decir, cuando el número de ensayos es extenso y la probabilidad binomial es pequeña. La regla de mayor uso entre los estadísticos establece que una distribución de Poisson es una buena aproximación de la distribución binomial cuando *n* es igual o mayor que 20 y cuando *p* es igual o menor que 0.05.
En los casos en que se satisfacen tales condiciones, podemos sustituir la media de la distribución binomial *np* en lugar de la media de la distribución de Poisson λ de modo que la formula será:

$$P(x)= (np)^n \, e^{-np}/x! \qquad (24)$$

Ejemplo 19. Supongamos que tenemos un hospital con 20 máquinas de diálisis renal y que la probabilidad de que una de ellas no funcione bien durante un día cualquiera es de 0.02 ¿Cuál es la probabilidad de que exactamente 3 queden fuera de servicio en un mismo día?

En la Tabla 8, se muestra las respuestas a esta pregunta. Como se aprecia en ella, la diferencia entre las dos distribuciones de probabilidad es ligera (apenas cerca de 10% de error en el ejemplo):

$$n = 20 \qquad P = 0.02 \qquad x = 3 \qquad q = 1 - p = 0.98$$

Tabla 8. Solución de problema bajo dos enfoques de distribución.

Enfoque de Poisson	Enfoque binomial
$P(x) = (np)^n \, e^{-np}/x!$ $P(x) = (20*0.02)^3 \, e^{-(20*0.02)}/3!$ $P(x) = 0.00715$	$P(x) = [(n!)/ (n - x)!] (p^x \, q^{n-x})$ $P(x) = [(20!)/ (20 - 3)!] (0.02^3 * 0.98^{(20-3)})$ $P(x) = 0.0065$

Conclusiones

La estadística es la ciencia que se trata de cuantificar la probabilidad de la ocurrencia o el efecto de cualquier evento, sujeto, proceso, fenómeno o interacciones resultantes. Hay que recalcar que la estadística es solamente un medio y no el fin. Sin embargo, algunos investigadores se involucran tanto en los detalles de la estadística que parece que hasta se trata de ajustar la realidad a los métodos estadísticos, es decir, para estos investigadores, si la estadística no define una realidad, uno debe deshacerse de la realidad. En otras palabras, hemos sido testigo de abuso, mal uso y sobre uso de esta herramienta en las investigaciones en diferentes disciplinas de la búsqueda de patrones repetitivas, que forman el propósito de una franquicia muy sería denominada la ciencia (Badii & Castillo, 2007). Los ejemplos de este mal uso de la estadística abundan en las mejores revistar científicas del mundo. Es con este objetivo que debemos utilizar de forma adecuada las diferentes distribuciones probabilísticas de uso actual en nuestras investigaciones.

Referencias

Anscombe, F.J. & W.W. Glynn, 1983. Distributions of the kurtosis statistic for normal statistics. Biometrika: 70: 227-234.

Badii, M.H. & J. Castillo (eds.). 2007. Técnicas Cuantitativas en la Investigación. 348 pp. UANL, Monterrey. ISBN: 970-694-377-3.

Badii, M.H., J. Castillo, J. Landeros & K. Cortez. 2007a. Papel de la estadística en la investigación científica. Innovaciones de Negocios. 4(1): 107-145.

Badii, M.H., J. Castillo, A. Wong & J. Landeros. 2007b. Precisión de los índices estadísticos: técnicas de jacknife & bootstrap. Innovaciones de Negocios. 4(1): 63-78.

Badii, M.H., J. Castillos, R. Foroughbakhch & K. Cortez. 2007c. Probability and scientific research. Daena, 2(2): 358-369.

Badii, M.H., J. Castillo, K. Cortez, A. Guillen & P. Villalpando. 2007d. Diseños experimentales e investigación científica. Innovaciones de Negocios, 4(2): 283-330.

Best, D.J. 1975. The difference between two Poisson expectations. Austral. J. Statatist.
Burstein, H. 1981. Binomial test for independent samples with independent proportions. Comimnic. Statist.-Theor , Meth.10:11-29.
Dale, A.I. 1989. An early occurrence of the Poisson distribution. Statist. Prob. Lett. 7:21-22.
Groeneveld, R.A. & G.Meeden, 1984. Measuring skewness and kurtosis. Statistician, 33:391-399.
Little, R.J.A. 1989. Testing the equality of two independent binomial proportions. J. Amer. Statist.
Moors, J.J.A., 1986. The meaning of kurtosis: Darlington revisted. Amer. Statist. 40:283-284.
Moors, J.J.A., 1988. A quantile alternative for kurtosis. Statistician 37:25-32.

***Acerca de los autores**

El Dr. Mohammad Badii es Profesor e Investigador de la Facultad de Administración y Contaduría Pública de la U. A. N. L. San Nicolás, N. L., México, 66450. mhbadii@yahoo.com.mx

El Dr. Jorge Castillo es Profesor e Investigador de la Facultad de Administración y Contaduría Pública de la U. A. N. L. San Nicolás, N. L., México, 66450. daena@spentamexico.org