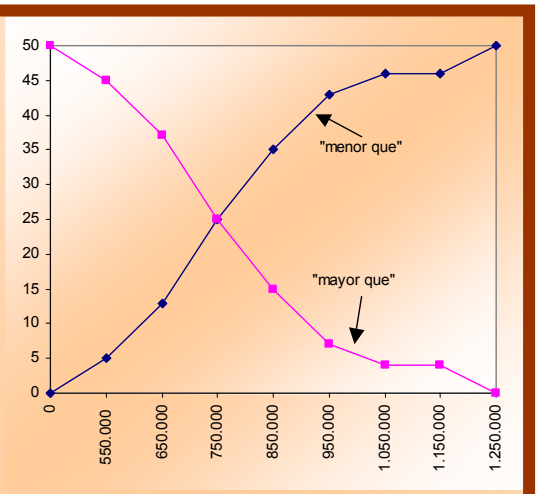


ESTADÍSTICA I

Guía teórico – práctica

Dr. Genaro Mosquera C.
Ing. Luis A. Martínez R.

ÍNDICE

Introducción

El propósito de esta guía didáctica es presentar a los estudiantes de administración y economía una exposición concisa y ordenada de los aspectos más importantes de la estadística elemental, buscando compatibilizar el tratamiento matemático de los tópicos tratados con el conocimiento básico de matemática de los primeros años de estudios universitarios, sin que ello significara pérdida de rigurosidad dentro del proceso de aprendizaje.

La guía fue diseñada ajustándose en lo posible al contenido programático del curso de Estadística I que se imparte en las carreras de Administración y Economía. En ella se consideran aspectos tales como la naturaleza general de la Estadística y sus propósitos, en relación con problemas y situaciones que usualmente se habrán de enfrentar durante el ejercicio profesional de estas y otras disciplinas.

El material didáctico abarca los conceptos básicos de la estadística, usos y notación; la naturaleza, tipología y manejo de datos estadísticos; las herramientas básicas de la estadística descriptiva, para luego concluir con el tema de las probabilidades y las diferentes formas de distribución de las variables aleatorias, como base de la estadística inductiva.

A título de complemento, se presenta aparte una serie de láminas en formato electrónico donde se recogen los aspectos más significativos del curso, con vínculos automáticos a un grupo de problemas resueltos mediante el uso de la hoja de cálculo de Microsoft Excel, además de un extracto de ejercicios seleccionados, mediante los cuales se aspira contribuir a una mayor comprensión y reafirmación de los temas tratados.

ESTADÍSTICA I

1. Estadística. Definición y aplicaciones.

La estadística es una disciplina científica que forma parte de las matemáticas aplicadas y reúne en sí un conjunto de métodos y herramientas que encuentran utilización práctica en una variada gama de campos, incluyendo, entre otras, las ciencias sociales, físicas, biológicas y de la salud, demostrándose de particular utilidad como medio de apoyo en los procesos de investigación y de toma de decisiones.

La estadística es la ciencia de la **inducción lógica**, vale a decir que permite extraer conclusiones de carácter general a partir de un reducido número de observaciones; en otras palabras, la inducción permite formular generalizaciones acerca de la naturaleza o característica de una determinada clase de objetos, sobre la base de observaciones realizadas sobre una cantidad limitada de tales objetos. La **deducción**, al contrario de la inducción, conduce a establecer, sobre la base de premisas generales acerca de las propiedades de una clase de objetos, si todos los objetos considerados poseen tales características. Es posible entonces decir que la deducción es un razonamiento *a priori*, mientras de la inducción es un razonamiento en base a la *evidencia empírica*.

Para un investigador, la estadística proporciona los medios por los cuales se hace posible utilizar una cantidad limitada o incompleta de información, para formular conclusiones acerca de causas y efectos de algún fenómeno estudiado, para comprobar teorías o determinar las relaciones existentes entre datos iniciales y resultados finales. De tal forma que es posible, por ejemplo, decidir cuál entre varias formulaciones químicas de un fertilizante agrícola es la más conveniente para obtener las cosechas más abundantes de un determinado rubro.

En el área de las ciencias sociales, las técnicas estadísticas pueden ser usadas para predecir el resultado de unas elecciones o la probabilidad de éxito o fracaso de un cierto negocio. Como ejemplo, en asuntos de tipo económico, es factible usar la estadística para elegir, entre varias alternativas de posibles funciones teóricas de consumo, aquella que mejor explica el comportamiento de los datos reales observados.

En el campo médico, los métodos estadísticos son aplicables al resultado de pruebas experimentales realizadas con un nuevo fármaco, con la finalidad de evaluar su efectividad en el tratamiento de algún tipo de enfermedad o, en otro caso, si realmente existe una relación causal directa o indirecta entre el consumo de cigarrillos y el cáncer pulmonar.

Aun cuando se hayan mencionado solamente algunos pocos ejemplos de la gran variedad de situaciones en las cuales es factible utilizar los métodos estadísticos para encontrar soluciones a problemas de diversa naturaleza, concluyendo se puede afirmar que la teoría y las herramientas estadísticas pueden contribuir efectivamente, con un alto grado de confianza en la validez de los resultados, a resolver el dilema de tener que tomar un curso de acción, en escenarios donde solamente se dispone de un conocimiento parcial o limitado de los elementos para decidir.

2. Conceptos y notación

Se denomina *población* o *universo* a la totalidad de un grupo de objetos de interés para el estudio estadístico, mientras que por *muestra* se entiende una parte relativamente pequeña de la población, seleccionada con el propósito de extraer conclusiones acerca de algunas características o parámetros propios del universo.

El tratamiento estadístico considera dos categorías principales de datos: *variables* y *constantes*. Las variables se denotan usualmente por medio de las últimas letras del alfabeto, tales como ***x,*** *y,* o ***z.*** Dentro de esta categoría se incluyen también los *estadísticos* y los *parámetros*. Un estadístico es una característica medida u observada en una muestra y usualmente se denota mediante una letra específica. Por ejemplo, la media aritmética de una muestra se denota mediante la letra equis mayúscula con un guión superpuesto $\overline{X}$. La característica o parámetro correspondiente a la población o universo al cual pertenece la muestra seleccionada se denota mediante **μ** (la letra griega mu minúscula). Generalmente los estadísticos de una muestra se denotan mediante letras del alfabeto latino, mientras los parámetros de la población correspondiente se denotan mediante letras griegas.

En relación a las constantes, existen dos tipos: ordinarias y "naturales". Las constantes ordinarias se denotan mediante las primeras letras minúsculas del alfabeto, es decir *a*, *b*, o *c*. También en este caso, es necesario distinguir entre constantes basadas en la observación de una muestra y los parámetros de la población correspondiente. Supóngse por ejemplo la siguiente relación entre dos variables X y Y observada en una muestra: $Y = a + bX$. La ecuación correspondiente a la población o universo sería: $Y = \alpha + \beta X$. En el ejemplo, a y b son estadísticos de la muestra mientras que α y β son los parámetros de la respectiva población.

Por constantes "naturales" se entienden ciertas constantes específicas utilizadas frecuentemente en todas las ramas de las matemáticas, siendo el número irracional *e* = 2,71828..... y $\pi = 3{,}14159...$ dos de las constantes naturales más frecuentemente utilizadas.

En los procesos de cómputo se utilizan los llamados *operadores* es decir un símbolo para indicar que una determinada operación debe ser ejecutada. Los dos *operadores* utilizados con mayor frecuencia son los de *sumatoria* y *productoria*. El primero se denota por medio de **Σ** (la letra griega sigma mayúscula) y se interpreta como las suma de una serie de números. El segundo se denota mediante **Π** (la letra griega pi mayúscula) y se interpreta como el producto de una serie de términos dados. A título de ejemplo, si se escribe:

$$\sum_{i=1}^{n} x_i \quad \text{esto significa} \quad \sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + ... + x_n$$

Similarmente, si se escribe:

$$\prod_{i=1}^{n} x_i \quad \text{esto significa} \quad \prod_{i=1}^{n} x_i = x_1 \cdot x_2 \cdot x_3 \cdot ... \cdot x_n$$

3. Los datos: materia prima de la estadística.

Si se quisiera distinguir entre *conocimiento* e *información*, se podría decir que el conocimiento implica la formulación de conclusiones con un cierto grado de certidumbre, derivada de información obtenida a partir de un número limitado de individuos, objetos o experiencias pasadas, mediante la aplicación de técnicas estadísticas. La información de partida, en este caso, se conoce como *datos estadísticos* y se asemejan a la materia prima de un proceso de manufactura. Los datos se obtienen de alguna fuente, se aplican las operaciones estadísticas apropiadas y se obtiene como resultado un producto terminado bajo la forma de estimaciones o conclusiones que sirven de base para la toma de decisiones, permitiendo elegir uno entre varios cursos de acción, con el propósito de alcanzar con la mayor efectividad posible un objetivo predeterminado.

3.1. Tipos de datos

Los datos estadísticos pueden ser de dos tipos: cualitativos y cuantitativos. Datos como el color del cabello, la ciudad de nacimiento y las preferencias políticas o religiosas de las personas son datos cualitativos. Por otra parte, el nivel de ingresos, las ventas, el salario, peso o estatura son datos de tipo cuantitativo.

En cuanto a su origen, los datos pueden provenir de fuentes internas o externas. Las fuentes internas incluyen los registros o archivos de la empresa u organización que los utiliza. Los datos internos en el caso de una empresa están conformados por los registros de ventas, producción, nómina, personal y los asientos contables. Si los datos provienen de fuentes que se encuentran más allá de las fronteras de la organización que los utiliza, se habla entonces de fuentes externas.

Los datos obtenidos de fuentes externas a través de publicaciones pueden ser clasificados a su vez en dos categorías: fuentes primarias, donde los datos son publicados por quienes los recolectaron y/o produjeron originariamente (censos de población, estadísticas de ingresos, reportes anuales de actividad industrial, etc.) y fuentes secundarias, como los diarios, revistas especializadas, libros y otros medios de comunicación, que reproducen la información elaborada por las fuentes primarias.

Mientras exista la posibilidad, resulta preferible obtener la información de las fuentes primarias, debido a las imprecisiones que se pueden encontrar en los datos provenientes de fuentes secundarias (errores de trascripción, datos incompletos, ausencia de notas aclaratorias, etc.). Sin embargo, cuando no se logra satisfacer las necesidades de la investigación con la información extraída de las fuentes internas y externas, es necesario recurrir a la recolección directa de los datos, ya sea mediante encuestas, entrevistas, observación directa, mediciones, o cualquier otra técnica apropiada, para obtener la información deseada. En cualquier caso, deberá cuidarse que la muestra seleccionada para este propósito sea efectivamente representativa de la población o universo al cual pertenece; de lo contrario, el grado de imprecisión o incertidumbre podría ser de tal magnitud como para invalidar los resultados de la observación o de la investigación.

3.2. Algunos aspectos del manejo de los datos.

En toda investigación científica debe tomarse en cuenta algunos aspectos relacionados con la exactitud en el manejo y procesamiento de los datos numéricos. Específicamente, se habla de la aplicación de aproximaciones, redondeos y el uso de la cantidad apropiada de dígitos significativos.

Números aproximados

Generalmente, los datos estadísticos cuantitativos representan mediciones aproximadas y no números exactos. Por ejemplo, cuando se afirma que la temperatura ambiente en un determinado lugar y en un momento dado es de 28° C, es necesario tener claro que la medición, por varias razones, no es totalmente exacta y el verdadero valor se halla, por ejemplo, en el intervalo entre 27,5 y 28,5° C. A pesar de ello, el valor de 28° C es suficientemente preciso para muchos propósitos y en la mayor parte de los casos resulta más conveniente usar datos aproximados en lugar de valores exactos.

Reglas de redondeo

Cuando se dispone de datos numéricos exactos, es siempre posible reducirlos a cantidades aproximadas recurriendo a las reglas de redondeo. En el caso de la temperatura citado anteriormente, se redondeó a la unidad entera más próxima. La regla básica es redondear al valor más cercano al último dígito que se desea destacar. Por ejemplo, si la vida útil de un lote de neumáticos es de 35.374,53 kilómetros, se puede aproximar al millar expresando que la vida útil es de 35.000 kilómetros, o de 35.400 kilómetros si se redondea a la centena más cercana. Sin embargo, en el caso de valores como 22.500, que se encuentra exactamente a mitad de camino entre 22.000 y 23.000, se utiliza una regla arbitraria, pero prácticamente de uso universal, que establece lo siguiente: si el último dígito que se desea conservar es par, entonces se redondea por defecto o se descartan los demás dígitos a la derecha. Si el último dígito a conservar es impar, se redondeará por exceso al dígito par más cercano.

Obviamente, cuando se redondea, se introducen errores debido al uso de números aproximados. Sin embargo, cabe esperar que en promedio un dígito par aparezca tantas veces cuanto aparece un dígito impar, y considerando que un número entro 0 y 5 podrá aparecer en tantas oportunidades cuanto un número entre 5 y 10, es de esperar que en una gran cantidad de números redondeados haya tantas cifras aproximadas por defecto cuanto por exceso. Por consiguiente, el error de redondeo tendería a anularse.

Dígitos significativos.

Cuando se usan números redondeados, se denominan dígitos significativos a la cantidad de dígitos del número original que haya sido conservada. En el número redondeado a 22.000 del caso arriba señalado, aparecen dos dígitos significativos. Sin embargo, en el caso de números con decimales, por ejemplo 3.420,0 las cifras significativas son cinco. Para números menores que 1, se cuentan los dígitos que aparezcan a la derecha del último cero decimal; así 0,000190 tiene tres dígitos significativos, lo mismo que 0,0320 o 0,00255.

4. Distribuciones de frecuencias

Una vez recolectados los datos, estos se presentan usualmente como una abundante masa de cifras que, a primera vista, parecen carecer de significado. Es por ello que antes someter los datos a un proceso de análisis, estos deben ser recompuestos o arreglados de manera ordenada o compendiados de forma tal que adquieran una cierta coherencia y sea posible obtener una visión global acerca de su distribución. Precisamente, las distribuciones de frecuencias son las herramientas que hacen posible cumplir con tales propósitos.

Supóngase, a título de ejemplo, que la nómina mensual los trabajadores de una cierta fábrica se presenta como muestra el Cuadro N° 1

Cuadro N° 1 - Salario mensual (en Bs.)

952.500	834.800	999.900	899.000	1.010.100
913.700	693.300	600.500	781.000	823.700
895.900	984.700	672.000	651.000	786.000
807.300	989.900	632.500	933.500	676.700
936.500	751.000	800.000	889.900	803.200
1.003.300	839.200	767.000	998.800	687.500
994.800	654.000	854.300	643.200	600.100
753.600	1.253.300	1.114.500	1.253.300	1.126.400
634.900	1.100.000	714.800	1.291.000	856.000
1.303.500	938.000	913.200	667.700	799.000

Tal como se presentan los datos, resultaría difícil aseverar algo acerca de la forma en que se distribuyen los salarios, más aun si se tratara de una empresa que empleara varios cientos o miles de trabajadores. Si se construyera una tabla donde se establecieran intervalos de salarios y la cantidad de trabajadores cuya paga está comprendida en cada uno de los intervalos definidos, se tendría entonces una panorámica más comprensible de la manera en la cual se distribuyen los salarios en la nómina mensual de la empresa. Este tipo de arreglo se denomina *tabla de frecuencias* o *distribución de frecuencias*.

4.1 Construcción de una distribución de frecuencias

Utilizando los datos del ejemplo anterior, se procederá a construir una tabla de frecuencias que ilustre la distribución de los datos. El primer paso consiste en determinar los *intervalos de clase* apropiados, estableciendo sus *límites* y *fronteras* (los límites de las clases son límites nominales e indican los valores de las frecuencias contenidas en cada clase después del redondeo; las fronteras de las clases indican los valores reales de la frecuencias contenidas en el intervalo antes del redondeo). Es de notar que a pesar de ser en cierta medida arbitraria, la definición de los intervalos de clase debe cumplir con ciertas reglas o lineamientos generales, como se describe a continuación:

a) No debe haber solapamiento entre intervalos consecutivos.
b) Los intervalos definidos deben contener en ellos la totalidad de los valores o datos observados.

c) Los intervalos deben tener todos la misma amplitud y es deseable, para facilitar la lectura e interpretación, que sean en múltiplos de cinco.
d) No se debe incluir intervalos de clase semiabiertos (ej.: "menores que XX" o "YY y más").
e) La cantidad de intervalos de clase no debe ser excesiva, pero tampoco exageradamente limitada. Se recomienda entre cinco y quince intervalos, si bien el número exacto dependerá de la cantidad de datos y la diferencia entre los valores máximo y mínimo de la distribución; cuanto mayores son ambos factores, más intervalos de clase deben existir.

Para construir la tabla de frecuencias (Cuadro N° 2) se definen los intervalos de clases y se realiza manualmente el conteo del número de empleados cuyo salario está comprendido entre los valores establecidos para cada intervalo.

Límites de clase			Frecuencias
Bs. 550.000	hasta Bs.	649.900	5
650.000		749.900	8
750.000		849.900	12
850.000		949.900	10
950.000		1.049.900	8
1.050.000		1.149.900	3
1.150.000		1.249.900	0
1.250.000		1.349.900	4

Finalmente, la tabla de frecuencias se presentará de la forma siguiente:

Cuadro N° 2 - Tabla de Frecuencia de Salarios mensuales

Salarios mensuales			N° de trabajadores
Bs. 550.000	hasta Bs.	649.900	5
650.000		749.900	8
750.000		849.900	12
850.000		949.900	10
950.000		1.049.900	8
1.050.000		1.149.900	3
1.150.000		1.249.900	0
1.250.000		1.349.900	4
Total			50

4.2 Frecuencias relativas y acumulativas.

Para la interpretación de los datos, a menudo resulta conveniente disponer de la distribución en términos de frecuencia relativa; es decir, expresando las frecuencias incluidas en una determinada clase como un porcentaje del total de frecuencias de la distribución. El cuadro N° 3 que se presenta a continuación, ilustra el resultado del cálculo de las frecuencias relativas.

Cuadro N° 3 - Tabla de Frecuencia de Salarios mensuales (absolutas y relativas)

Salarios mensuales		N° de trabajadores	Frecuencias relativas (%)
Bs. 550.000	hasta Bs. 649.900	5	10
650.000	749.900	8	16
750.000	849.900	12	24
850.000	949.900	10	20
950.000	1.049.900	8	16
1.050.000	1.149.900	3	6
1.150.000	1.249.900	0	0
1.250.000	1.349.900	4	8
Total		50	100

De igual manera, resulta útil, a los fines del análisis de los datos, contar con las frecuencias expresadas en forma acumulativa. En otras palabras, se podría determinar la cantidad o porcentaje de datos que son "menores", "iguales" o "superior" a un determinado límite (superior o inferior) de las frecuencias de clase consideradas. El cuadro N° 4 contiene los valores acumulados "menores que" y "superiores a" las clases de frecuencia de la distribución, tanto en términos absolutos como relativos. Se puede observar rápidamente que la mitad de las frecuencias corresponden a valores por debajo de Bs. 850.000 y que el 30% de las frecuencias corresponden a valores de al menos Bs. 950.000.

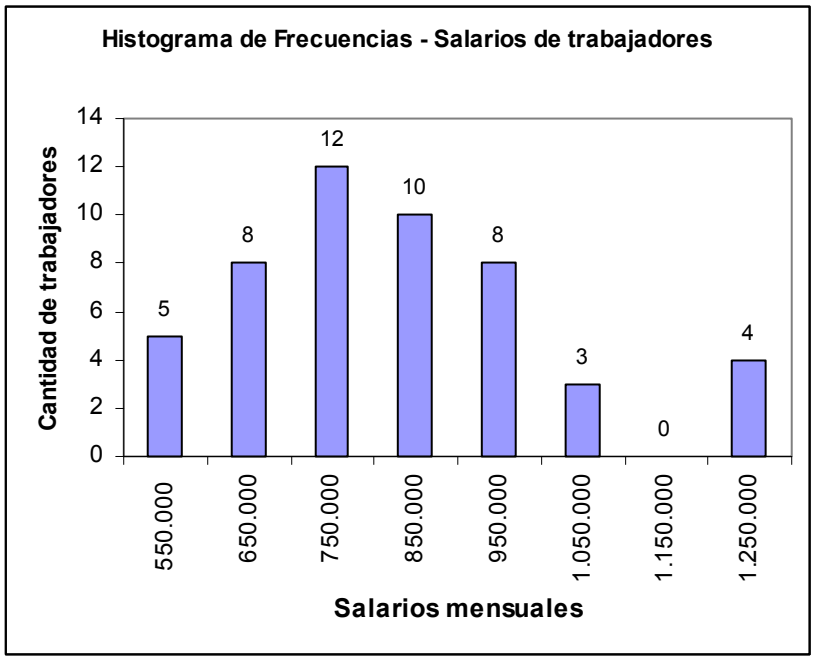
Cuadro N° 4- Distribución de frecuencias acumuladas "menores que" o "superiores a"

		Menores que		Superiores a	
Salarios mensuales		Absoluto	Relativo %	Absoluto	Relativo %
Bs. 550.000	hasta Bs. 649.900	5	10	50	100
650.000	749.900	13	26	45	90
750.000	849.900	25	50	37	74
850.000	949.900	35	70	25	50
950.000	1.049.900	43	86	15	30
1.050.000	1.149.900	46	92	7	14
1.150.000	1.249.900	46	92	4	8
1.250.000	1.349.900	50	100	4	8

4.3. Gráficos de frecuencias

El gráfico de una distribución de frecuencia es un medio que permite al analista captar de manera inmediata la naturaleza de la distribución de los datos y facilita la presentación y comprensión de los resultados de una investigación estadística. Existen tres tipos básicos de gráficos de frecuencia: el histograma, el polígono de frecuencias y la función acumulativa de frecuencias u ojiva. Para el trazado de los gráficos se pueden utilizar indistintamente los valores absolutos o relativos de las frecuencias. En el eje de las abscisas se dispondrán los límites o los puntos medios de cada clase, mientras que los valores de las frecuencias se ubicarán sobre el eje de las ordenadas.

Un histograma es un gráfico en el cual las frecuencias correspondientes a cada clase de la distribución se representa mediante una serie de rectángulos. Los lados de los rectángulos se trazan verticalmente a partir de los puntos que indican los límites de las frecuencias de clase y sus alturas representan el número absoluto o relativo de frecuencias correspondientes a cada clase. Un histograma basado en el cuadro de distribución de frecuencias absolutas y relativas del salario mensual de los trabajadores referido anteriormente, se presenta en el gráfico a continuación:

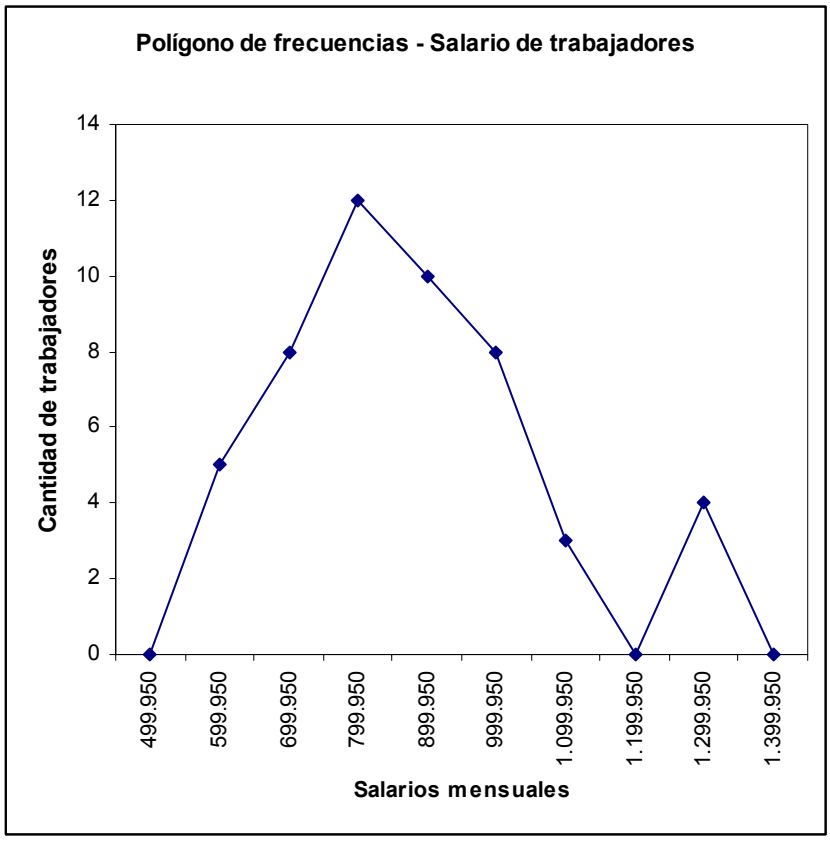


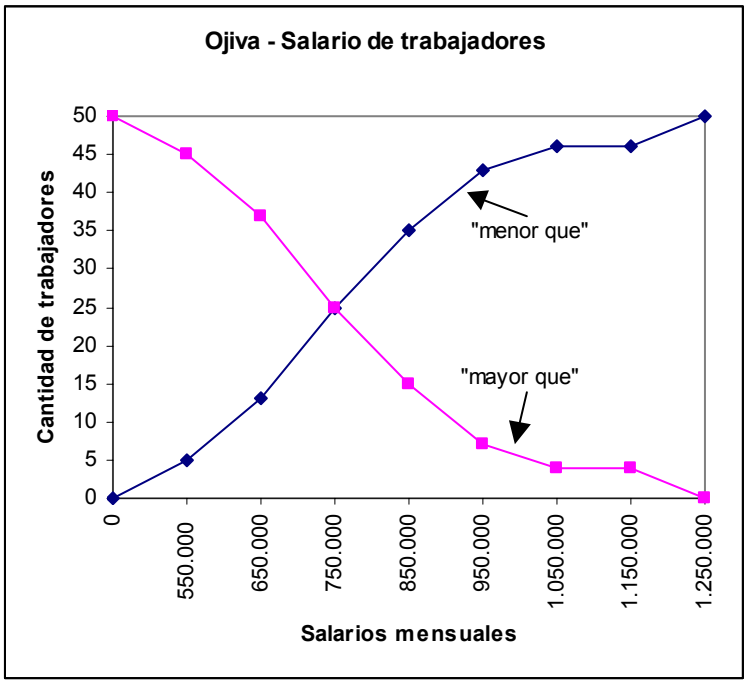
Los gráficos expuestos en la página siguiente representan respectivamente un polígono de frecuencias y una ojiva o gráfico de frecuencias acumuladas.

En el caso del polígono de frecuencias, en el eje de las abscisas se disponen los valores medios de los intervalos de clase y en el eje de las ordenadas el número absoluto o relativo de las frecuencias. El punto medio del intervalo se considera representativo de la clase y las frecuencias se trazan en correspondencia de dichos puntos.

La ojiva puede ser construida a partir de los valores absolutos o relativos de las frecuencias y en ella se pueden representar los valores "menor que" o "mayor que" o ambos. De forma similar a los casos anteriores, en el eje de las abscisas aparecerán los límites de los intervalos de clase y en las ordenadas las cantidades acumuladas de trabajadores correspondiente a cada clase.

Se puede notar que en ambos casos fueron agregados intervalos de clase no incluidos en los datos originales, solamente con el propósito de obtener figuras poligonales cerradas de mejor apariencia.


Polígono de frecuencias - Salario de trabajadores
Cantidad de trabajadores
14
12
10
8
6
4
2
0
499.950
599.950
699.950
799.950
899.950
999.950
1.099.950
1.199.950
1.299.950
1.399.950
Salarios mensuales


Ojiva - Salario de trabajadores
Cantidad de trabajadores
50
45
40
35
30
25
20
15
10
5
0
"menor que"
"mayor que"
0
550.000
650.000
750.000
850.000
950.000
1.050.000
1.150.000
1.250.000
Salarios mensuales

5. Medidas de tendencia central

El propósito de las medidas de tendencia central es describir, mediante una sola cifra, un valor representativo de un determinado conjunto de datos. **D**e todos los valores posibles de una cierta distribución, se podría desear conocer en qué punto de la escala se encuentra el valor alrededor del cual se centra la distribución. Por ejemplo, los economistas a menudo tienen interés en conocer el valor medio del ingreso de las personas habitantes en una determinada área geográfica. Igualmente, podría ser apropiado conocer el valor medio de vida de un componente mecánico, o saber cuál es el color de calzado que mayormente prefiere un cierto grupo de hombres y mujeres de edad adulta. Mediante las medidas de tendencia central es posible encontrar ese valor medio, típico o representativo de los datos analizados en cada caso.

5.1 La moda.

La moda se define sencillamente como el valor que se repite o aparece con mayor frecuencia en una cierta distribución de valores. Esta medida es aplicable a todos los ejemplos antes señalados, tanto en el caso de datos de tipo cualitativo (ej.: color del cabello, color de los ojos, sexo, raza, lugar de origen, ocupación, etc.) como para datos cuantitativos (ej.: estatura, peso, ingresos, ventas, etc.). Sin embargo, se debe acotar que si los datos cuantitativos se encuentran agrupados bajo la forma de una distribución de frecuencias, no será posible encontrar el valor exacto de la moda, aunque sí es factible determinar la clase modal, es decir la clase que posee el mayor número de frecuencias.

La principal ventaja de la moda como valor "típico" es su sencillez, sea desde el punto de vista conceptual como de su determinación. Además, en el caso de valores cualitativos, es la única medida que tiene significado. Las mayores desventajas de la moda se refieren a su unicidad y a su misma existencia, pues hay numerosos casos de distribuciones que no la poseen, mientras que otras pueden tener dos o más valores modales.

5.2 La Mediana

La mediana, al igual que la moda, es un tipo simple de promedio. Se define sencillamente como aquel valor que divide una distribución de frecuencias (ordenadas numéricamente) en dos partes iguales. A diferencia de la moda, la mediana adquiere significado solamente en el tratamiento de datos cuantitativos. De manera que se podrá hablar de la mediana en el caso de edades, ingresos, ventas de una tienda por departamentos y otros casos similares.

Para calcular el valor de la mediana, se toman los datos ordenados de manera ascendente o descendente y se ubica el valor que corresponde a la siguiente expresión :

$$m_d = a_{(n+1)/2}$$

donde (n + 1) / 2 es el subíndice del valor medio de la variable A, siempre que se tenga una cantidad impar de valores. Si la cantidad de valores es par, entonces la mediana será dada mediante la expresión:

$$m_d = \frac{a_{n/2} + a_{(n/2)+1}}{2}$$

Es decir que, en este caso, la mediana se define como el promedio de las dos frecuencias centrales de la distribución.

Si los datos se encuentran agrupados en frecuencias de clase, será posible calcular un valor estimado de la mediana, siempre y cuando las frecuencias dentro de la clase que contiene la mediana estén distribuidas uniformemente. En tal caso, la expresión a utilizar es la siguiente:

$$m_d = L + c \cdot \left(\frac{n/2 - \sum^{L} f}{f_m} \right)$$

donde:

L = límite inferior de la clase que contiene la mediana

c = amplitud de la clase que contiene la mediana y

$\sum^{L} f$ = sumatoria de todas las clases por encima de la que contiene la mediana.

Para ilustrar el proceso de cálculo, considérese la siguiente tabla de retribuciones por hora de un grupo de obreros (Cuadro N° 5). Nótese que siendo n/2 = 740 / 2 = 370, la mediana se encuentra en el intervalo de clase "*2.250 y menor que 2.750*". L = 2250, c = 500, f_m = 224 y la suma de frecuencias por encima de la clase que contiene la mediana es $\Sigma f = 288$.

Cuadro N° 5 – Retribuciones por hora de trabajo

Salario en Bs./hora	N° de obreros f_i	Frec. acum. Σf_i
1.250 y menor que 1.750	105	105
1.750 y menor que 2.250	183	288
2.250 y menor que 2.750	224	512
2.750 y menor que 3.250	148	660
3.250 y menor que 3.750	75	735
3.750 y menor que 4.250	2	740
Total	740	-

Aplicando la expresión para el cálculo de la mediana se tiene:

$$\begin{aligned} m_d &= 2250 + 500\left(\frac{370 - 228}{224} \right) \\ &= 2250 + 316{,}96 \\ &= \text{Bs. } 2.566{,}96 \end{aligned}$$

Respecto a la moda y la media aritmética, la mediana presenta algunas ventajas importantes. A diferencia de la moda, la mediana siempre existe y es única para cada distribución, aun si los intervalos de clase son desiguales o abiertos. Su significado es fácilmente comprensible y generalmente es sencillo calcularla. Al no depender de los valores como tales, sino de la posición relativa de las frecuencias en la distribución, la mediana no resulta afectada por la existencia de valores extremos, revelándose sumamente útil como valor representativo de distribuciones muy sesgadas (con gran cantidad de valores concentrados a la derecha o izquierda del valor medio).

5.3 La Media Aritmética

Media aritmética es el nombre técnico del promedio ordinario, ampliamente utilizado, inclusive desde los primeros niveles de la escuela elemental. Respectivamente para una muestra de *n* elementos y una población de tamaño *N*, la media se determina mediante las siguientes expresiones:

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

Para cada distribución, la media constituye un valor único, pero a diferencia de la moda y la mediana, resulta sensiblemente afectada cuando en la distribución existen valores extremadamente grandes o extremadamente pequeños.

Cálculo de la media de una distribución de frecuencias

Considerando que en una distribución de frecuencias los elementos individuales pierden su identidad, el cálculo de la media será posible solamente si se conoce la media de los elementos pertenecientes a cada clase. Se podrá entonces calcular una media ponderada de cada clase mediante la expresión siguiente:

$$\mu = \frac{\sum_{i=1}^{N} m_i f_i}{N}$$

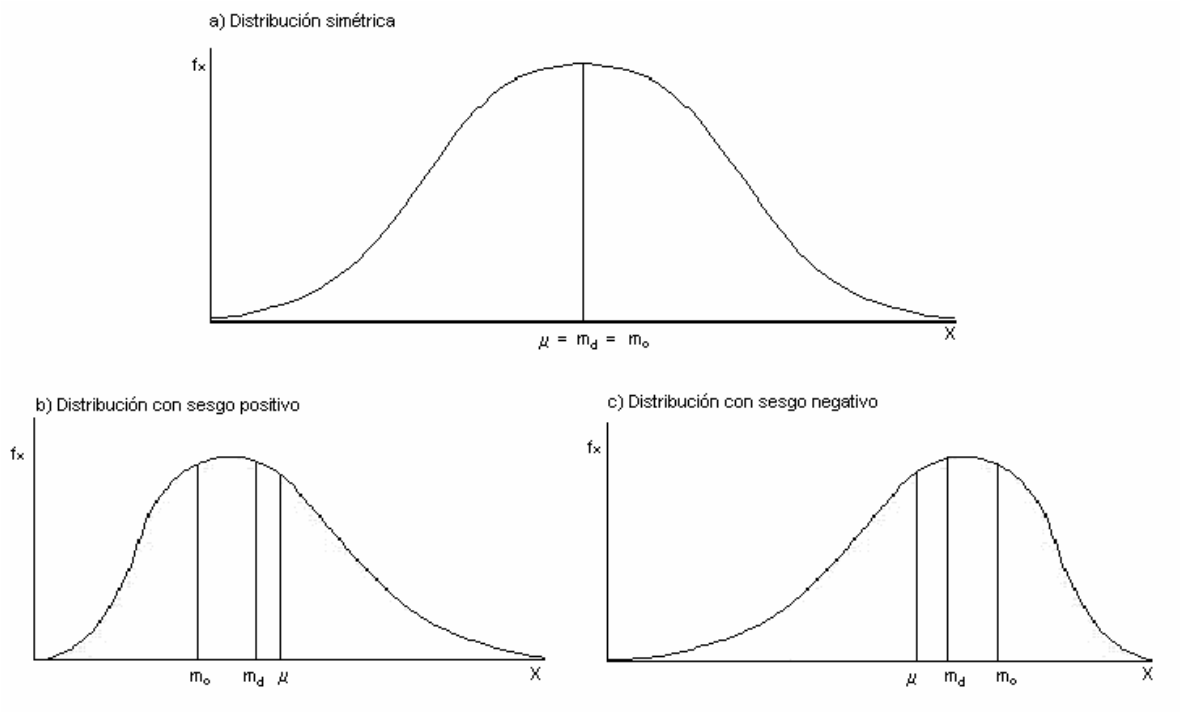
donde N es la cantidad de clases presente en la distribución y f_i representa el número de frecuencias dentro de cada clase. De manera similar, si los datos constituyen una muestra de valores, se calculará la media a partir de:

$$\bar{x} = \frac{\sum_{i=1}^{N} m_i f_i}{n}$$

Posición relativa de la media, mediana y moda en una distribución

Una distribución de frecuencias puede ser simétrica respecto a la media, presentar un insesgamiento positivo (a la derecha) o negativo (a la izquierda), según que los valores de

la variable se encuentren simétricamente ubicados respecto a la media, o tiendan a acumularse a uno u otro lado del valor medio. En el primer caso, media, mediana y moda coinciden, localizándose en el punto medio de la distribución. En el segundo caso, la media se ubica a la derecha de la moda y la mediana, mientras que en el último se coloca en un punto a la izquierda de la mediana y la moda.



5.4. La Media Geométrica

Es la medida de tendencia central de una distribución, cuyos valores son elementos de una serie geométrica. Los tipos de datos a los cuales se aplica la media geométrica son valores tales como porcentajes y razones de incremento de los datos respecto al tiempo, la relación entre ventas de un período a otro, la proporción de un componente respecto a la suma total, etc. La media geométrica se calcula mediante la fórmula:

$$m_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

El cálculo de la media geométrica se facilita convirtiendo los valores de X a logaritmos y computando la media aritmética de los valores obtenidos.

$$\log m_g = \frac{\sum_{i=1}^{n} \log x_i}{n}$$

Es decir que el logaritmo de la media aritmética es igual a la media aritmética de los logaritmos de los datos. Es de notar que la media geométrica tiene sentido si y solo si todos los valores de la serie geométrica son diferentes de cero.

5.5 La Media Armónica

La media armónica es otro clase de medida que se utiliza para promediar ciertos tipos de relaciones o proporciones cuando las cantidades a calcular se expresan en diferentes unidades de medida. Por ejemplo, si se quisiera promediar la velocidad por hora de dos automóviles que recorren una cierta distancia, se debe recurrir a la media armónica. El cálculo se obtiene a partir de la siguiente fórmula:

$$m_a = \frac{n}{\sum_{i=1}^{n}(1/x_i)}$$

lo cual equivale a decir que la media armónica es el recíproco de la media aritmética de los recíprocos de los valores observados.

Sin embargo, es oportuno resaltar que la media armónica es usada raras veces ya que todos los problemas a los cuales se aplica, pueden ser igualmente resueltos mediante el cálculo de un promedio aritmético ponderado.

5.6 Cuartiles, Deciles y Percentiles

Un conjunto de datos observados puede ser dividido en grupos de manera tal que cada uno de ellos contenga valores menores que la medida calculada para el correspondiente cuartil, decil o percentil. En el caso de los cuartiles, el primer cuartil es aquel valor por debajo del cual se encuentra el 25 por ciento de las frecuencias observadas. El segundo cuartil (que corresponde a la mediana de la distribución) es el valor por debajo del cual se coloca 50 por ciento de los valores de frecuencia. El tercero y último cuartil es el valor por debajo del cual se ubica el 75 por ciento de las frecuencias observadas.

Las fórmulas para el cálculo de los cuartiles son las siguientes:

Cuando n es par

$$Q_1 = a_{(n+1/4)} \;;\; Q_2 = a_{(n+1/2)} \;;\; Q_3 = a_{3(n+1)/4}$$

Cuando n es impar

$$Q_1 = [a_{(n/4)} + a_{(n/4+1)}]/2 \;;\; Q_2 = [a_{(n/2)} + a_{(n/2+1)}]/2 \;;\; Q_3 = [a_{3n/4} + a_{(3n/4+1)}]/2 \;;$$

Los deciles son nueve y el primero es el valor por debajo del cual se coloca el 10 por ciento de las frecuencias de la distribución. Por debajo del segundo decil se ubica el 20% de las frecuencias, y así sucesivamente hasta el noveno decil, por debajo del cual se encuentra el

90 por ciento de las frecuencias observadas. Para calcular los deciles, se emplea la fórmula siguiente:

$$D_i = L_i + c\left(\frac{\frac{i \cdot n}{10} - \sum^{L} f}{f_i}\right)$$

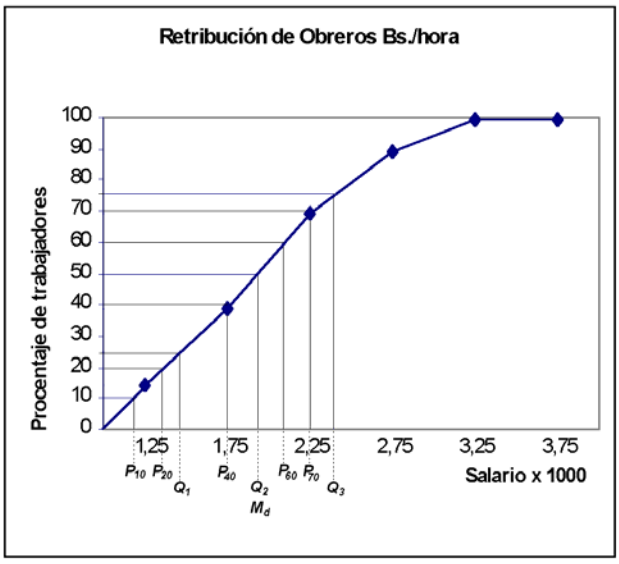
donde L_i es el límite inferior de la clase que contiene el i-ésimo decil y f_i es la cantidad de frecuencias presentes en la clase que contiene dicho decil.

Los percentiles se interpretan de igual forma que los cuartiles y deciles. Existen 99 percentiles y el décimo percentil equivale al primer decil, o sea el valor por debajo del cual se halla el 10 por ciento de las frecuencias. Así mismo, el percentil 25 equivale al primer cuartil; el percentil 50 equivale al segundo cuartil y al quinto decil, y así sucesivamente. La fórmula general para el cálculo de los percentiles es la siguiente:

$$P_i = L_i + c\left(\frac{\frac{i \cdot n}{100} - \sum^{L} f}{f_i}\right)$$

donde L_i es el límite inferior de la clase que contiene el i-ésimo percentil y f_i es la cantidad de frecuencias presentes en la clase que contiene dicho percentil.

El gráfico ilustra la relación existente entre cuartiles y percentiles y la ubicación de la mediana, cuyo trazado se realizó utilizando los datos del ejemplo de retribución de un grupo de obreros (Cuadro N° 5).



6. Medidas de dispersión

A pesar de que las medidas de tendencia central o de posición resultan sumamente valiosas por cuanto representan valores típicos o representativos de una distribución, su utilidad resulta algo restringida si no se acompañan de información adicional acerca de la naturaleza de la distribución misma. Cualquiera que sea la población considerada, siempre existirá alguna variación en las características de los elementos que la integran. Por lo tanto, no es suficiente determinar alguna de las medidas de posición, sino que se debe establecer así mismo la magnitud de la variación respecto a esa medida de tendencia central.

6.1 La desviación media.

Desde el punto de vista matemático, la desviación media representa el promedio de las diferencias en valor absoluto entre los datos observados y su media aritmética, calculada mediante la siguiente ecuación:

$$DM = \frac{\sum_{i=1}^{N} \left| x_i - \mu \right|}{N}$$

El cálculo de la desviación media del conjunto de datos mostrado a continuación (Cuadro N° 6), cuya media es igual a 14,6 se expone seguidamente:

Cuadro N° 6

x_i	$\lvert x_i - \mu \rvert$
24	9.4
20	5.4
18	3.4
18	3.4
15	0.4
12	2.6
12	2.6
10	4.6
9	5.6
8	6.6
μ = 146/10 = 14,6	DM = 44/10 = 4,4

Es decir, que la diferencia promedio entre los valores individuales de X y su media es igual a 4,4.

6.2 La desviación estándar y la varianza

Si en lugar del promedio de las diferencias en valor absoluto entre los valores de X y su media aritmética, se calcula el promedio de los cuadrados de dichas diferencias, se obtiene la varianza de una población, que se denota mediante el símbolo σ^2 cuyo valor se obtiene mediante la ecuación:

$$\sigma^2 = \frac{\sum_{i=1}^{N} \left(x_i - \mu \right)^2}{N}$$

Si se extrae la raíz cuadrada de la varianza, se obtiene una medida de dispersión expresada en las mismas unidades que la media o cualquiera de las otras medidas de tendencia central. Este valor es denominado desviación estándar o típica y su expresión matemática es:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

La media aritmética y la desviación estándar son respectivamente las medidas de posición y dispersión que mejor caracterizan a una determinada distribución. La desviación estándar mide, en un cierto sentido, el grado de representatividad de la media. Las distribuciones que muestran valores pequeños de desviación estándar, se alejan poco de su media. Si los valores de la desviación estándar son elevados, significa que la distribución presenta un alto grado de dispersión respecto a la media.

Cuando se trata de muestras, las expresiones para el cálculo de la varianza y la desviación estándar son las siguientes:

$$s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{n-1} \qquad y \qquad s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{n-1}}$$

Las formulas anteriores difieren de las expresiones usadas para una población por emplear la media de la muestra para calcular los desvíos y la cantidad de datos menos 1 en el denominador. Cabe señalar que si las muestras son suficientemente grandes, es decir de 100 o más datos, es prácticamente indiferente usar n o n-1. De hecho, a medida que crece el número de datos, la desviación estándar *s* de la muestra, se aproxima cada vez más a la desviación estándar σ de la población de la cual se extrajo.

Cálculo de la varianza y la desviación estándar para datos no agrupados.

El proceso de cálculo, se presenta continuación:

Cuadro N° 7

x_i	$(x_i - \mu)$	$(x_i - \mu)^2$
7.88	-0.074	0.005476
8.56	0.606	0.367236
8.43	0.476	0.226576
6.67	-1.284	1.648656
8.23	0.276	0.076176
6.98	-0.974	0.948676
7.55	-0.404	0.163216
7.90	-0.054	0.002916
8.22	0.233	0.070756
9.12	1.166	1.359556
μ =79.54/10=7.954	-	σ^2 = 4.86924 /10

Por consiguiente, la varianza es igual a 0,49 y la desviación estándar a 0,70. Si los datos se refirieran al peso de un grupo de niños de dos años de edad, significaría que el peso promedio del grupo es de 7,954 kilogramos y que los valores individuales difieren de la media en 0,7 kilogramos.

Cálculo de la varianza y la desviación estándar para datos agrupados.

Si los datos se presentan en forma de distribuciones de frecuencias, se utilizan la siguientes expresiones para el cálculo de la varianza, respectivamente para una población y una muestra:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(m_i - \mu)^2 \cdot f_i}{N} \qquad s^2 = \frac{n \cdot \sum_{i=1}^{k} m_i^2 \cdot f_i - \left(\sum_{i=1}^{k} m_i^2 \cdot f_i\right)^2}{n \cdot (n-1)}$$

Retomando los datos del ejemplo de retribuciones por hora contenidas en el cuadro N° 5, el proceso de cálculo para una población se realiza como sigue:

Salario en Bs./hora	Punto medio m_i	N° de obreros f_i	$m_i \cdot f_i$	$m_i^2 \cdot f_i$
1.250 y menor que 1.750	1.500	105	157.500	23.625 E4
1.750 y menor que 2.250	2.000	183	366.000	73.200 E4
2.250 y menor que 2.750	2.500	224	560.000	140.000 E4
2.750 y menor que 3.250	3.000	148	444.000	133.200 E4
3.250 y menor que 3.750	3.500	75	262.500	91.875 E4
3.750 y menor que 4.250	4.000	5	20.000	8.000 E4
Totales		740	1.810.000	469.900 E4

Por consiguiente, aplicando la fórmula de cálculo de la varianza para datos agrupados, se obtiene:

$$\sigma^2 = \frac{(740 \cdot 469.900 \cdot 10^4) - (181 \cdot 10^4)^2}{(740)^2} = 367.348{,}43$$

Extrayendo la raíz cuadrada de la varianza resulta que $\sigma = 606{,}10$. Finalmente, en relación a la población analizada, es posible afirmar que la retribución media de los trabajadores es de Bs. 2.450 por hora, y que en promedio los valores individuales difieren en Bs. 606,10 de la media aritmética de la distribución.

7. Probabilidades

7.1. Probabilidad. Conceptos y definiciones

Variable aleatoria: Una variable se dice aleatoria cuando cada uno los valores específicos que puede asumir está asociado a una probabilidad.

Sucesos aleatorios: la ocurrencia de sucesos donde intervienen variables aleatorias se denominan sucesos aleatorios.

Sucesos elementales: se dice que un evento o suceso es elemental cuando no puede ser expresado en términos de otros eventos más sencillos

Probabilidad: desde un punto de vista práctico, se puede asumir que la probabilidad es una medida de la verosimilitud o expectativa de ocurrencia de un determinado suceso aleatorio, cuya representación matemática es un número real comprendido entre cero y uno. Si el suceso es totalmente imposible, se le asigna probabilidad cero y si el suceso es absolutamente cierto que ocurra, se le asigna probabilidad uno. En el resto de los casos, la probabilidad de ocurrencia del suceso estará comprendida entre estos dos límites, y mayor será la certeza de que el suceso se produzca cuanto más cercano se encuentre su valor de probabilidad al número *uno*.

Concepto clásico de probabilidad: si un experimento aleatorio puede tener *n* posibles resultados (eventos elementales) que tienen todos la misma probabilidad de ocurrir, y *r* de ellos poseen una característica predefinida, entonces, de acuerdo al concepto clásico, se define la probabilidad de ocurrencia del suceso (S) como el cociente de *r* (casos favorables) entre *n* (casos posibles) es decir:

$$P(S) = r / n$$

La aplicación del concepto clásico de probabilidad requiere del conocimiento previo de número total de posibles resultados y de los resultados favorables, así como asumir que todos los resultados son igualmente probables.

Por ejemplo, considérese un lote de producción 5000 calculadoras electrónicas de bolsillo. Si se conoce que 25 unidades del lote presenta defectos de fabricación, determínese la probabilidad, al escoger una al azar, que sea una calculadora defectuosa.

$$P(S) = 25/5000 = 1/200 = 0{,}005$$

Este enfoque resulta difícilmente aplicable a situaciones reales en el campo económico-financiero, puesto que no es posible conocer anticipadamente la totalidad de los posibles resultados de un determinado negocio y cuáles de ellos son favorables, como tampoco es razonable pensar que todos los posibles resultados sean igualmente probables.

Concepto de probabilidad en base a frecuencias relativas: Cuando los experimentos aleatorios son en teoría indefinidamente repetibles, en la práctica se pueden repetir gran cantidad de veces, y los resultados se aproximan en buena medida a los que se obtendrían repitiendo el experimento indefinidamente. Bajo estas condiciones, y considerando N repeticiones, en las que un suceso A haya ocurrido N_A veces (frecuencia de A) como producto de la repetición de uno cualquiera de los sucesos elementales que pertenecen a A, entonces es experimentalmente observable que a medida que N aumenta, el cociente N_A / N, denominado frecuencia relativa del suceso A, tiende a aproximarse a un cierto valor, considerado como el límite al cual tiende la frecuencia relativa cuando N es suficientemente grande, lo que permite utilizar el concepto de frecuencia relativa como aproximación numérica al concepto de probabilidad.

En este caso, considérese una situación en la cual se requiere conocer qué cantidad de un producto altamente perecedero se debe adquirir en un determinado momento. Evidentemente, el concepto clásico de probabilidad resulta inaplicable puesto que no es posible saber con antelación la magnitud ni la frecuencia de la demanda. Sin embargo, es factible analizar cuál fue su comportamiento a lo largo de un período de tiempo suficientemente extenso y construir, a partir de los datos recopilados, una distribución de frecuencias relativas, basada en las cantidades de producto demandadas. Estas frecuencias relativas pueden ser interpretadas como las probabilidades de ocurrencia de la demanda de 0, 1, 2, ...unidades del producto.

Distribución de la demanda

Unidades	0	1	2	3	4	5	más de 5
Días	260	150	90	70	24	10	6

Para estimar la probabilidad de que en un día cualquiera la demanda sea de 2 unidades, se calcula $P(S) = A_N / N$, donde $A_N = 90$ y $N = 610$, por lo tanto

$$P(S) = 90 / 610 = 0{,}1475$$

Para conocer cuál será la probabilidad de que la demanda sea superior a 3 unidades, es decir cuatro o más unidades, el cálculo se realiza considerando las frecuencias acumuladas a partir de 4 unidades, o sea: $A_N = 24 + 10 + 6 = 40$ y por consiguiente

$$P(S) = 400 / 610 = 0{,}0655$$

Concepto subjetivo de probabilidad: Es factible encontrarse frente a problemas prácticos en los cuales no es posible aplicar ni el concepto clásico, ni el de frecuencias relativas. Es el caso, por ejemplo, de considerar la oportunidad de desarrollar y lanzar al mercado un nuevo producto. Se trata, en síntesis, de decidir si deberá emprenderse el desarrollo del producto y determinar si la ganancias potenciales podrán compensar el esfuerzo y los costos de la iniciativa. Siendo un producto totalmente nuevo, no se cuenta con datos en base a los cuales calcular la frecuencia relativa de los sucesos de éxito y falla; sin embargo, el emprendedor se encuentra ante la disyuntiva de tener que decidir al respecto.

El concepto subjetivo de probabilidad sostiene que una persona experimentada está en condiciones de enfrentar el problema sobre la base de experiencias similares, y capacitado para formular estimaciones subjetivas, razonablemente precisas, acerca de las probabilidades asociadas a las alternativas del caso. En esta oportunidad, las probabilidades no representan cantidades objetivamente mensurables, sino expresiones puramente subjetivas determinadas por el "peso de la experiencia". En este sentido, se puede considerar la probabilidad como una medida de la fortaleza del convencimiento personal acerca de la ocurrencia de un determinado suceso. En otras palabras, es el pensamiento subjetivo de que un evento o conjunto de eventos ocurrirá o no, sujeto al grado de fortaleza (ponderación) de las propias convicciones personales.

Naturaleza de la probabilidad: En cada uno de los conceptos presentados, quedó claramente establecido que el valor asignado o calculado de la probabilidad de ocurrencia de uno o más sucesos está comprendido en un rango entre 0 y 1, ambos inclusive.

Ahora bien, ¿cómo se interpreta que la probabilidad de un suceso sea igual a 0 o a 1? La respuesta depende del concepto de probabilidad que se esté manejando en cada caso. Si se trata del concepto clásico, P(S) = 0 significa que el suceso es imposible, que no puede ocurrir. Sin embargo, bajo el concepto de frecuencia relativa P(S) = 0 no necesariamente implica que el suceso sea imposible, sino simplemente que no ocurrió al realizar el experimento. De manera similar, de acuerdo al concepto clásico, P(S) = 1 significa que el suceso es absolutamente seguro que ocurra. En cambio, para el concepto de frecuencia relativa, significa sencillamente que el suceso ocurrió en todas las pruebas realizadas.

La probabilidad de ocurrencia se puede interpretar como la proporción de veces que se espera ocurra el suceso, al repetirse un experimento u observación una gran cantidad de veces. Si el suceso no se produce al realizar una cualquiera de las pruebas, esto no significa que no sea probable o posible que suceda.

Espacio muestral: se denomina espacio muestral al conjunto de todos los resultados posibles de un experimento aleatorio.

Sucesos mutuamente excluyentes: Se dice que dos sucesos o eventos A y B son mutuamente excluyentes si no pueden ocurrir simultáneamente. Es decir que al ocurrir el suceso A, no puede ocurrir B y viceversa. Además, si conjuntamente los eventos elementales de A y de B incluyen todos los eventos posibles del espacio muestral, y uno de ellos debe forzosamente producirse, resulta obvio que $P(A) = 1 – P(B)$ razón por lo cual se dice que B es el complemento del evento A. Si los sucesos A y B son mutuamente excluyentes, pero no agotan colectivamente todos los eventos presentes en el espacio muestral, esto significa que existe la posibilidad de que no ocurran ni A ni B.

La probabilidad de que A o B ocurran, será equivalente a la unión de los subconjuntos constituidos por los elementos pertenecientes a A y B. Si n es el total de elementos del espacio muestral, n_a el total de elementos del suceso A y n_b los elementos del suceso B, entonces $P(A) = n_a / n$ y $P(B) = n_b / n$, y la suma de las probabilidades de ocurrencia de ambos sucesos será: $P(A \cup B) = P(A) + P(B)$.

Cuando existan elementos comunes entre los subconjuntos A y B, en el cálculo de la probabilidad de ocurrencia de uno u otro de los sucesos (o de ambos) se deberá descontar el valor equivalente a la intersección de los subconjuntos A y B, a fin de no contabilizar dos veces el aporte a la probabilidad de los elementos comunes a los dos subconjuntos. En este caso, la probabilidad será: $P(A \cup B) = P(A) + P(B) – P(A \cap B)$.

Sucesos Independientes: Dos sucesos A y B definidos en un mismo espacio muestral, se dicen independientes, si la ocurrencia de uno de ellos, no altera la probabilidad de ocurrencia del otro. La probabilidad resultante es equivalente al producto de las probabilidades de ocurrencia de los sucesos considerados o, en otras palabras, de la intersección de ambos subconjuntos. Así que $P(A \cap B) = P(A) . P(B)$.

Sucesos dependientes y probabilidad condicionada: Considerados dos sucesos A y B, se dice que la probabilidad de ocurrencia del suceso A es condicionada por el suceso B, cuando su resultado depende de la ocurrencia del suceso B. La expresión para calcular la probabilidad de A condicionada por B es : $P(A/B) = P(A \cap B) / P(B)$. Igualmente, la probabilidad de B condicionada por A será : $P(B/A) = P(A \cap B) / P(A)$.

Teorema de Bayes: Es denominado a menudo como regla de Bayes o de la probabilidad causal de los sucesos, ya que el interés no se centra en determinar la probabilidad de ocurrencia de un suceso B, sino sabiendo que B ocurrió, hallar la probabilidad de que una causa específica A_k haya sido la que provocó su ocurrencia. Este teorema establece que si un suceso B puede ocurrir asociado a los sucesos A_1, A_2....A_k, los cuales son mutuamente excluyentes dos a dos, entonces la probabilidad de que haya ocurrido el suceso A_k sabiendo que ocurrió el suceso B, viene dada por:

$$P(A_k / B) = \frac{P(B / A_k) \cdot P(A_k)}{\sum_{J=1}^{K} P(B / A_J) \cdot P(A_J)}$$

Para ilustrar la aplicación del teorema de Bayes, supóngase que en un proceso de acabado de rodamientos para cojinetes, la máquina utilizada para este propósito tiene cuatro dispositivos de ajuste que deben ser calibrados correctamente. Si el reglaje de alguno de ellos queda fuera de la norma, se producirá en los rodamientos un cierto tipo de defecto (suceso B). Supóngase además que las posibles causas del defecto son A_1, A_2, A_3 y A_4 cuyas probabilidades de ocurrencia son respectivamente 0,01; 0,03; 0,02 y 0,02, y que las probabilidades de ocurrencia de B al ocurrir A_1, A_2, A_3 y A_4 son respectivamente 0,45; 0,30, 0,25 y 0,50. Luego, las probabilidades de A_k dado que B ha ocurrido son como se indica en la tabla siguiente:

Sucesos A_k	$P(A_k)$	$P(B/A_k)$	$P(A_k)$. $P(B/A_k)$	$P(A_k/B)$
A_1	0,01	0,45	0,0045	0,16
A_2	0,03	0,30	0,0090	0,32
A_3	0,02	0,25	0,0050	0,17
A_4	0,02	0,50	0.0100	0,35
			0.0285	1,00

De los resultados obtenidos se concluye que la causa más probable del defecto (suceso B) está en el ajuste A_4 y que la segunda causa más probables es el ajuste A_2, lo que permitirá tomar decisiones y actuar de consecuencia respecto a los puntos de regulación.

Aun cuando en la vida real es improbable que se conozcan con precisión las probabilidades de los sucesos causales y los sucesos condicionados, especialmente en problemas ligados a procesos administrativos o de tipo económico-financiero, será siempre posible asignarlas sobre la base de experiencias pasadas, de acuerdo con el concepto subjetivo de probabilidad referido en párrafos anteriores, teniendo presente que gran parte de la teoría de decisiones sobre bases estadísticas se fundamenta en estas nociones básicas.

8. Distribución de variables aleatorias

8.1 Conceptos básicos.

Una variable aleatoria X puede ser discreta o continua, dependiendo de los valores específicos x que asuma. Las variables aleatorias discretas, toman solamente valores numéricos de tipo entero de la escala de posibles valores. Las variables aleatorias continuas pueden asumir cualquier valor numérico real dentro de un rango específico.

Al considerar las variables aleatorias de tipo discreto, la probabilidad de ocurrencia para un cierto valor x se denota como f(x) o, en otros términos, que $P(X = x) = f(x)$, donde f(x) puede asumir distintas formas dependiendo de la naturaleza de la variable aleatoria.

Supóngase por ejemplo que la demanda diaria (número de unidades demandadas) de un cierto artículo en una tienda tenga la siguiente distribución:

$$P(X=0) = f(0) = 0{,}3679$$
$$P(X=1) = f(1) = 0{,}3679$$
$$P(X=2) = f(2) = 0{,}1839$$
$$P(X=3) = f(3) = 0{,}0613$$
$$P(X=4) = f(4) = 0{,}0153$$
$$P(X\geq 5) = f(5) = 0{,}0037$$

Por cada valor de x existe su correspondiente f(x) que equivale a la probabilidad que la variable X asuma ese valor en particular. La expresión f(x) se denomina función de densidad de probabilidades, o simplemente función de densidad de X. Es de notar que las x son mutuamente excluyentes, de manera que la sumatoria de sus probabilidades es igual a la unidad.

De igual manera, para las variables aleatorias existe una función de distribución F(a) que se define como

$$F(a) = \sum_{\min X}^{a} f(x) = P(X \leq a)$$

Es decir que la función de distribución es una función acumulativa de probabilidades de X, desde el valor más pequeño hasta un cierto valor específico de X. En efecto, F(a) es igual a $P(X \leq a)$ lo cual significa que el valor de la variable aleatoria X es menor o igual a cierto valor específico *a*.

8.2 Esperanza matemática de una variable discreta.

La esperanza matemática o valor esperado de una variable aleatoria discreta, equivale a la sumatoria de los productos de todos los posibles valores de *x* por su respectiva probabilidad y representa el valor medio de la variable al realizar un número grande de experimentos.

Retomando el ejemplo del consumo diario de un determinado artículo y asumiendo que la demanda máxima es de 5 unidades, la demanda diaria esperada en un largo período de tiempo está dada por:

x	f(x)	x.f(x)
0	0,3679	0,0000
1	0,3679	0,3679
2	0,1839	0,3678
3	0,0613	0,1839
4	0,0153	0,0612
5	0,0037	0,0185
		E(X) = 0.9994 ≈ 1,000

donde $E(X) = \Sigma(x \ . \ f(x))$

8.3. Varianza de una variable aleatoria discreta.

En el parágrafo anterior se expresó que el valor esperado de una variable aleatoria puede ser considerado como el valor medio de la variable luego de realizar una larga serie de experimentos, vale a decir que la media μ de una variable aleatoria se expresa como:

$$\mu = \sum_X x \cdot f(x) = E(X)$$

La varianza representará entonces la medida en la cual la variable aleatoria difiere de su media o valor esperado. La expresión para calcular su valor es:

$$\sigma^2 = \sum_X (x-\mu)^2 \cdot f(x) \ = E(X-\mu)^2 \ = \sum_X x^2 f(x) - \mu^2$$

Extrayendo la raíz cuadrada de la varianza, se obtendrá la desviación estándar de la variable aleatoria, es decir:

$$\sigma = \sqrt{E(X^2) - [E(X)]^2}$$

8.4 Distribuciones de variables aleatorias discretas

Seguidamente se describen las características más importantes de algunas funciones de distribución asociadas a variables aleatorias discreta..

Distribución Hipergeométrica

Si se tiene un conjunto finito S de N elementos, compuesto por dos subconjuntos: *A* que contiene n elementos y el complementario de *A* que contiene los restantes N - *n* elementos, la distribución hipergeométrica es el modelo que representa la probabilidad de extraer una selección donde x elementos provengan de un subconjunto *B* perteneciente a *A* y n – x sean elementos del subconjunto complemento de B. La función de densidad de la distribución es:

$$f(x) = \frac{\binom{A}{x}\binom{N-a}{n-x}}{\binom{N}{n}}$$

donde $A = N \cdot P$; la función de distribución se obtiene mediante:

$$f(x) = \sum_{x=0}^{c} \frac{\binom{A}{x}\binom{N-a}{n-x}}{\binom{N}{n}}$$

y la variable x_i toma valores de la manera siguiente: $x_i : 1, 2, \ldots . c$

La Esperanza Matemática de la distribución es:

$$E(x) = E(n\,p) = n \cdot \frac{A}{N}$$

Los valores de A se estimarán a partir de una muestra donde p es la proporción muestral, esto es: $p = \frac{a}{n}$ donde:

$$A = N \cdot p$$

y la desviación típica de la distribución:

$$\sigma = \sqrt{\frac{N-n}{n} \cdot \frac{p.q}{n}}$$

Distribución Binomial

En este caso se considera una variable discreta x_i cuyo campo de variación es x_i: 1, 2, 3, . . . c y se desea calcular la probabilidad de x éxitos en n intentos. Si se define como P la proporción de elementos con una característica 1 en una población N, se tiene:
$P = \frac{A}{N}$, siendo A la suma de los elementos que presentan una característica que se ha llamado 1; la ausencia de esa característica se identifica por (0). La proporción de elementos con característica (0) es:

$$Q = 1 - P$$

Al considerar una población N, el número de veces o intentos que se realiza en el experimento o prueba corresponde a una muestra con reemplazamiento de n elementos. En tal caso, la función de densidad es:

$$f(x) = \binom{n}{x} P^x Q^{n-x}$$

la función de distribución es:

$$F(x) = \sum_{x=0}^{c} \binom{n}{x} P^x Q^{n-x}$$

la Esperanza Matemática de esta distribución es:

$$E(x) = n P$$

y su varianza

$$\sigma^2 = n \cdot P \cdot Q$$

Distribución de Poisson

La distribución de Poisson es aquella en la cual la variable discreta X puede tomar valores enteros positivos entre 0 y +♄ cuando X corresponde al número de ocurrencias de un suceso por unidad de tiempo o espacio, cuya función de densidad viene dada por la expresión:

$$f(x) = e^{-\lambda} \cdot \frac{\lambda}{x!}$$

donde $\lambda > 0$ es el parámetro de la distribución.

La función de distribución en este caso es:

$$f(x) = \sum_{x=0}^{c} e^{-\lambda} \cdot \frac{\lambda}{x!}$$

la Esperanza Matemática correspondiente es:

$$E(x) = \lambda$$

y la varianza:

$$\sigma^2 = \lambda$$

Cuando la muestra es suficientemente grande, es decir $n \equiv$ ♄, y $p \equiv 0$, la distribución binomial b(x) se aproxima a la distribución de Poisson p(x). Si se fija $\lambda = 2$, la distribución de Poisson se aproximará a la distribución binomial.

$$b(x) \rightarrow e^{-\lambda} . \frac{\lambda}{x!} = p(x)$$

donde $\lambda = n \, . \, p.$

8.5 Variables aleatorias continuas

Cuando se consideraron las nociones fundamentales acerca de las variables aleatorias, se introdujo el concepto de función de distribución de probabilidades F(a) para variables discretas como una probabilidad acumulativa para valores de X a partir de un mínimo hasta un valor *a* determinado. De forma similar, es posible definir una función de distribución para las variables continuas donde $F(a) = P(X \leq a)$. Es decir, para una variable continua el área bajo la curva de la función de densidad hasta $X = a$ es:

$$P(X \leq a) = \int_{\min X}^{a} f(x) \; dx$$

Por lo tanto, el área bajo la curva de densidad en un intervalo determinado es equivalente a la probabilidad de que X se encuentre dentro de dicho intervalo.

El valor esperado para una variable continua X se define a través de la integración, en todo el rango de valores posibles, del producto de x por f(x). Es decir:

$$E(X) = \int_{\min X}^{\max x} x \cdot f(x) dx$$

y la varianza se obtiene mediante la expresión:

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

8.5 Distribuciones de variables aleatorias continuas

Desde el punto de vista de las aplicaciones en estadística, la distribución normal es sin lugar a dudas la más importante de todas, a pesar de que existen otras distribuciones cuyo empleo en el análisis estadístico tiene particular significación. En esta oportunidad, el estudio se limitará a considerar las características más importantes de la distribución normal.

Distribución Normal o de Gauss

Esta distribución es de particular importancia para la comprensión y aplicación de los diferentes métodos estadísticos. Su rango de variación para la variable aleatoria correspondiente es de carácter continuo y su función de densidad se expresa mediante:

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

la función de distribución correspondiente viene dada mediante la siguiente expresión:

$$F(X) = \int_{-\infty}^{z} f(x)\ dx \quad \text{donde } z = \frac{x_i - \mu}{\sigma}$$

la Esperanza Matemática de la distribución corresponde a la media aritmética $E(X) = \mu$ y su varianza es σ^2

Distribución Normal Standard

Aunque generalmente se habla de *la distribución normal* en forma excluyente, es oportuno resaltar que no existe una distribución normal única, sino un número infinito de ellas. Para cada par diferente de valores de μ y σ, denominados *parámetros de la distribución*, existe una distribución específica, ya que una vez fijado el valor de la media y la desviación estándar, la distribución queda completamente definida.

Ahora bien, teniendo presente que una transformación lineal de la variable X no introduce alguna modificación en la naturaleza de la distribución sino en sus parámetros, se puede definir una nueva variable Z, tal que:

$$Z = \frac{X - \mu}{\sigma}$$

La importancia de Z, por la forma en que ha sido definida, reside en el hecho que no solamente su distribución es normal, sino que tiene por media $\mu = 0$ y varianza $\sigma^2 = 1$, cuya función de densidad se conoce como normal unitaria o normal standard y que tiene por expresión:

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

siendo su función de distribución

$$F(x) = \int_{-\infty}^{z} f(x)\ dx$$

Es conveniente recordar que las distribuciones discretas a las cuales se hizo mención anteriormente, tienen aproximaciones a la distribución normal. Por ejemplo, la distribución binomial se podrá aproximar a una distribución normal para valores de *p* cercanos a 0,5, teniendo por media $\mu = n.p$ y por varianza $\sigma^2 = n \cdot p \cdot q$

A partir de estos elementos, es posible hacer una transformación a los valores de Z mediante la expresión siguiente:

$$Z = \frac{X - np}{\sqrt{npq}}$$

Así, la probabilidad de que X sea igual a un valor *a*, cuya distribución binomial se expresa mediante $b(a) = \binom{n}{a} p^a q^{n-a}$ es aproximadamente igual a la distribución normal definida por: $f(x) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$

BIBLIOGRAFÍA

- **Burford, Roger.** Basic Statistics for Business and Economics. C. Merrill Publiching Company, Columbus Ohio. USA. 1970.

- **Meyer, Stuart L**. Data Analysis for Scientists and Engineers. John Wiley and Sons.USA.1975.

- **Mosquera C. Genaro**. Hipótesis Estadística. Ediciones Sobre Visión C.A. Caracas, Venezuela. 1974.

- **Salinas, José F**. Introducción al Cálculo de Probabilidades. Ediciones Sobre Visión C.A. Caracas, Venezuela. 1977.

Si se desea profundizar algunos de los temas tratados en esta guía, se recomienda consultar las siguientes publicaciones:

- Rivas González, E. Estadística General. Editorial Cbvc.
- Spiegel. Estadística. Editorial McGraw Hill.
- Soto Negrín, Armando. Iniciación a la Estadística. Editorial José Martí.
- Gómez Rondón. Estadística Aplicada. Editorial Fragor.
- Johnson y Kelly. Estadística Elemental. Editorial Thomson.
- Gómez Rondón. Estadística Metodológica. Editorial Fragor.
- Hamett Murphy. Introducción al Análisis Estadístico. Editorial Addison Wesley.
- Soto y Galarraga. Iniciación Práctica a la Estadística. Editorial Experhente.
- Soto Negrín. Principios de Estadística. Editorial Panapo.