Capítulo 5. DISTRIBUCIONES EN EL MUESTREO

En los capítulos anteriores se estableció que el principal "material" del proceso de investigación econométrica son los datos y que, para obtenerlos, el investigador debe plantear una tabla de datos de unidades de observación por variables. También se introdujo el concepto de que estas variables tienen características que las distinguen a unas de otras (medidas de posición y dispersión) y poseen distribuciones de probabilidad. En este capítulo se estudiará cómo aquellas medidas de posición y dispersión poblacionales de una variable –denominadas parámetros (media, varianza, etc.)- son estimados a partir de medidas de posición y dispersión muestral -denominadas estadísticos- obtenidos en la muestra aleatoria de las unidades de observación. También, las condiciones estadísticas que se analizarán deben cumplir para que los estimadores sean considerados como tal, dejando para los próximos capítulos la forma de realizar estimaciones puntuales o por intervalo de dichos parámetros y cómo contrastar si los valores estimados se corresponden con los naturales de los parámetros o lo supuestos por el investigador.

5.1 ESTADÍSTICOS Y PARÁMETROS

Hasta aquí se ha desarrollado el concepto de probabilidad y sus distribuciones asociadas, ahora se estudia la forma en que los *estadísticos* de la muestra se pueden utilizar para hacer inferencias sobre los *parámetros* de la población.

En la práctica, una muestra de unidades de observación de tamaño determinado se selecciona aleatoriamente entre los **N** elementos de la población. Supóngase que se propone seleccionar **n** unidades y que se observarán las características X, Y y Z. La tabla de datos se puede representar según la figura 5.1, donde:

- Los valores x_i, y_i, z_i, para todo i =1,…,n representan *los datos* que se van a obtener cuando se lleve a cabo la investigación;
- Las medidas de posición y dispersión que se obtendrán, últimas dos filas de la tabla, son estadísticos, por ejemplo el estadístico:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

correspondiente a la *media aritmética* de la variable X en la muestra y se utilizará para estimar al parámetro

$$\bar{\mu}_X = \frac{\sum_{i=1}^{n} X_i}{N}$$

correspondiente a la media de la variable X en la población;

- **N** representa el tamaño de la población (son los elementos de la población) y **n** el de la muestra (son las unidades de observación seleccionadas en la población).

T	X	Y	Z
1	x_1	y_1	z_1
2	x_2	y_2	z_2
3	x_3	y_3	z_3
⋮	⋮	⋮	⋮
i	⋮	⋮	⋮
⋮	⋮	⋮	⋮
n	x_n	y_n	z_n
Medida de posición	$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$	$\frac{\bar{Y} = \sum_{i=1}^{n} y_i}{n}$	$\frac{\bar{Z} = \sum_{i=1}^{n} z_i}{n}$
Medida de dispersión	$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$	$S^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{n-1}$	$S^2 = \frac{\sum_{i=1}^{n}(z_i - \bar{Z})^2}{n-1}$

Figura 5.1 Tabla de datos experimental

Las **n** unidades de observación se van a obtener aplicando un diseño de muestreo. A fin de poder utilizar la media de la muestra para estimar la media de la población, se deberían examinar todas las muestras posibles de la población y calcular una media para cada muestra. Esto es, se debería repetir el experimento anterior tantas veces como muestras con reposición se podrían obtener de los N elementos de la población. La distribución de estos resultados se denomina *distribución en el muestreo*. En la práctica con la selección de una muestra, y utilizando la teoría de la probabilidad, se pueden hacer inferencias a la población.

Esta parte del proceso se denomina *inferencia estadística* y es uno de los objetivos de la investigación econométrica. La inferencia se hace utilizando estadísticos de la muestra; los estadísticos son "mediciones" que se realizan en la muestra para estimar las mismas "mediciones" en la población que se denominan parámetros.

El investigador, a través de la tabla de datos planteada en la Figura 5.1, podrá obtener las medidas de posición y dispersión de la muestra (estadísticos), supongamos para la variable X y buscará estimar valores probables de las medidas de posición o dispersión poblacionales (parámetros). Las conclusiones a las que arribará serán respecto a la población y no respecto de la muestra.

Ejemplo. Un encuestador político se interesa en los resultados de la muestra sólo como medio para estimar la *proporción* de votos que recibirá cada candidato entre la población de votantes.

En la Figura 5.2 se observa la relación entre estadísticos y parámetros

Nombre	Parámetros	Estimador
Media	μ_x	$\overline{X}$
Diferencia entre las medias de dos poblaciones	$\mu_{x_1} - \mu_{x_2}$	$\overline{X}_1 - \overline{X}_2$
Proporción	π	p
Varianza	σ_X^2	s_X^2
Desviación estándar	σ_X	s_X

Por ejemplo, el estadístico $\overline{X}$, que surge de una muestra, es estimador del parámetro μ_x en la población. El estadístico $\overline{X}_1 - \overline{X}_2$, diferencia de las medias provenientes de dos muestras, en distintas poblaciones, es estimador del parámetro $\mu_{X_1} - \mu_{X_2}$ que indica la diferencia en las medias proveniente de dos poblaciones.

Figura 5.2: Parámetros poblacionales y estadísticos muestrales

En la práctica, se selecciona solo una muestra; por lo tanto, se debe examinar el concepto de distribución en el muestreo, con el propósito de poder utilizar la teoría de la probabilidad para hacer inferencias en cuanto a los parámetros de la población.

En este capítulo se estudiarán las distribuciones en el muestreo sin abordar específicamente el problema de cómo se obtiene una muestra aleatoria, tema que será abordado más adelante, ya que corresponde a la tercera etapa del proceso de investigación econométrica planteado como objetivo académico en este manual.

5.2 DISTRIBUCIÓN EN EL MUESTREO DE LA MEDIA

Existen diversas medidas con que se pueden caracterizar los datos de la muestra. Dentro de las medidas de posición, tanto el *modo* como la *mediana* y la *media*, calculadas de los datos muestrales de una variable, podrían ser considerados estadísticos útiles para estimar la media de esa variable en la población.

No obstante, el mejor estimador para inferir la media poblacional es la *media aritmética* obtenida sobre una determinada variable, a partir de los datos correspondientes a las unidades de observación de la muestra. Esto es así pues, este estadístico, tiene tres propiedades deseables y muy importantes en todo proceso de estimación: es *insesgado*, *eficiente* y *consistente*.

La media aritmética, obtenida de los datos de una muestra aleatoria, es una *variable aleatoria*, ya que se ha calculado a partir de un proceso de estimación aleatorio. Por lo tanto, al ser una

variable aleatoria tiene una distribución de probabilidad, una esperanza matemática y una varianza. Esta propiedad es propia de todo estadístico calculado con datos provenientes de una muestra aleatoria.

Esto es así ya que si se repite sucesivamente un experimento muestral, en un espacio y tiempo determinado, se obtendrán sucesivos valores diferentes para las variables consideradas, lo que producirá diferentes estadísticos para estimar los parámetros. Es decir, se tendrán tantas medias aritméticas como veces se repita el experimento, por lo que la media es una variable y como proviene de un proceso de selección aleatorio, será una variable aleatoria. Por lo tanto, tendrá media, varianza y distribución de probabilidad asociada.

Para el caso particular de la media aritmética de los datos de una muestra aleatoria, se demostrará empíricamente que:

$$E(\overline{X})=\mu_X \tag{5.1}$$

Nótese que se está diciendo que el promedio de las medias aritméticas, obtenidas de las sucesivas muestras, es igual al parámetro poblacional que estima y no que la media aritmética de la muestra es igual al parámetro que estima.

Esto es, la esperanza matemática de la media aritmética de los datos obtenidos a partir de una muestra aleatoria, es igual al parámetro poblacional que se está estimando (la media de la población). Al cumplir con esta propiedad, se dice que la media aritmética es un *estimador insesgado* de la media poblacional.

Por lo tanto, si se toma una cantidad determinada de muestras aleatorias de una población, **n** muestras, y de cada una se calcula, a partir de los datos observados de una variable, una media aritmética, se obtendrán **n** medias aritméticas muestrales. La esperanza matemática de todas estas medias aritméticas muestrales es la media de la población,

$$\sum_{i=1}^{n}\frac{\overline{X}_i}{n}=\mu_X \tag{5.2}$$

Experimento. Bajo el supuesto de que la variable X representa el número de visitas al dentista en un año determinado por individuos de una cierta población, y que es una variable aleatoria discreta con distribución uniforme. Esto es, X puede asumir solo valores enteros y cada uno de estos valores se observa con la misma probabilidad. El Cuadro 5.1 y el Gráfico 5.1 muestran la frecuencia relativa en la población.

Identificando con *a* el valor mínimo y con *b* el valor máximo de *X*, la media de una distribución uniforme discreta es

$$E(X)=\mu_X=\frac{a+b}{2}=\frac{0+9}{2}=4.5$$

y la desviación estándar:

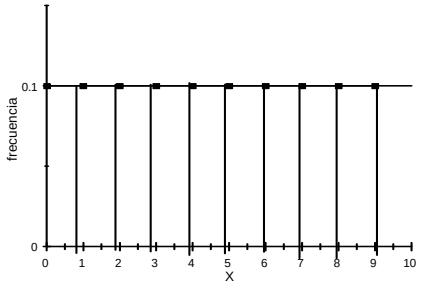
$$\sigma_X = \sqrt{\frac{[(b-a)+1]^2 - 1}{12}} = \sqrt{\frac{[(9-0)+1]^2 - 1}{12}} = 2.87$$

De modo que la media y la desviación estándar de la población bajo estudio asumen los valores 4.5 y 2.87.

Cuadro 5.1

Valores de X	Frecuencia relativa en la población
0	0.1
1	0.1
2	0.1
3	0.1
4	0.1
5	0.1
6	0.1
7	0.1
8	0.1
9	0.1
	1.0

Gráfico 5.1



De esta población se extraen 100 muestras posibles de tamaño 5. Es decir, se seleccionan 5 personas al azar de la población y se le pregunta cuántas veces visitó al dentista en el año. En una de las muestras las respuestas de las personas fueron:

PERSONA	NUMERO DE VISITAS
Primera	2
Segunda	0
Tercera	4
Cuarta	1
Quinta	1

Sobre esta muestra se calcula la media aritmética:

$$\frac{(2+0+4+1+1)}{5} = 1.6$$

Este experimento se repitió 100 veces y se calcularon 100 medias aritméticas. Los resultados del experimento se muestran en el Cuadro 5.2.

Aunque la variable (y, por lo tanto, los datos individuales) solo pueden tomar valores enteros entre 0 y 9, las medias de las muestras no serán, generalmente, enteros y la distribución muestral será una *distribución de frecuencias* cuyas clases estarán definidas por intervalos y no por puntos. Se sabe que para representar cada clase se puede elegir un valor, como, por ejemplo, el centro de cada intervalo.

Cuadro 5.2			
Valores de la Media de la muestra X		Frecuencia	
Intervalo	Punto Medio	Absoluta	Relativa
0.5 a 1.499	1	1	0.01
1.5 a 2.499	2	5	0.05
2.5 a 3.499	3	12	0.12
3.5 a 4.499	4	31	0.31
4.5 a 5.499	5	28	0.28
5.5 a 6.499	6	15	0.15
6.5 a 7.499	7	5	0.05
7.5 a 8.499	8	3	0.03
8.5 a 9.499	9	0	0.00
		100	1.00

Las principales características de esta distribución son

$$\text{Media} = \sum_{i=1}^{9} \overline{X}_i f_i = 1\,(0.01) + 2\,(0.05) + 3\,(0.12) + 4\,(0.31) + 5\,(0.28)$$

$$+ 6\,(0.15) + 7\,(0.05) + 8\,(0.03) + 9\,(0.00) = 4.60$$

$$\text{Desviación estándar} = \sqrt{\sum_{i=1}^{9} (\overline{X}_i - 4.60)^2 f_i} = 1.3638$$

Los resultados indican que un 59% de los valores estimados (0.31+0.28) están comprendidos en el intervalo ± 1 alrededor del valor real 4.5 y que el 86% de las estimaciones están comprendidas en el intervalo ± 2 alrededor de dicho valor.

Realizado el mismo experimento con la variante de tomar 100 muestras de tamaño 10 se obtuvo la siguiente característica

$$\text{Media } \sum_{i=1}^{9} \overline{X}_i f_i = 4.57\text{, Desviación estándar} = \sqrt{\sum_{i=1}^{9} (\overline{X}_i - 4.57)^2 f_i} = 1.0416$$

Los resultados del experimento permiten realizar las siguientes generalizaciones

a) la media de la muestra, $\overline{X}$, es un estimador *insesgado* de la media de la población, μ_X

b) a medida que aumenta el tamaño de la muestra, la distribución muestral de $\overline{X}$ se va concentrando cada vez más en torno a la media de la población y, por tanto, $\overline{X}$ es un estimador *consistente* de μ_X

c) el estimador $\overline{X}$ es más estable de una muestra a otra que la mediana o el modo, esto lo convierte en un estimador *eficiente*.

Se ha trabajado con distribuciones muestrales experimentales. Si se lo hubiese hecho con las distribuciones teóricas se hubieran obtenido idénticos resultados pero con mucho menos trabajo ya que sólo se tomaría una muestra.

5.3 FUNCIÓN GENERATRIZ DE MOMENTOS DE LA MEDIA

La media aritmética de la muestra es un estimador insesgado de la media de la población; como tal, es un estadístico y se vio que los mismos son variables aleatorias. Por lo tanto, como cualquier variable tienen ciertas características que la distingue de otras variables, como son: media, varianza y distribución de probabilidad. En esta sección se estudia la forma de derivar esas características.

La media de una variable aleatoria se conoce como el momento natural de orden 1. Partiendo de la fórmula de cálculo de la media aritmética, se puede derivar ese momento usando la función generatriz de momentos.

Se sabe que:

$$\overline{X}=\sum_{i=1}^{n}\frac{x_i}{n} \qquad (5.3)$$

Entonces si se toma el numerador de esta fórmula, cabe preguntarse:

¿A que es igual la función generatriz de momentos de la suma de variables aleatorias independientes[1] $X_1, X_2,\ldots X_n$?

Para responder a esta pregunta será necesario recordar, en primer lugar, qué es una función generatriz de momentos y cuáles son sus propiedades.

Función generatriz de momentos

La función generatriz de momentos de una *variable aleatoria continua* es:

$$M_X = E\left(e^{\theta x}\right)=\int_{-\infty}^{\infty} e^{\theta x} f(x)dx \qquad (1)$$

Si la *variable aleatoria es discreta*, se expresa

$$M_x = E\left(e^{\theta x}\right)=\sum_{X} e^{\theta x} f(x)$$

[1] Variable aleatoria independiente: la probabilidad de ocurrencia de una variable no depende de la probabilidad de ocurrencia de la otra.

M_X , o $M_X(\theta)$ donde θ es el parámetro, simboliza la función generatriz de momentos de una variable aleatoria.

Propiedades

a) La función generatriz de momentos de la sumatoria de variables aleatorias independientes en el parámetro θ es:

$$M_{X_1+X_2+\ldots+X_n}(\theta) = E\left\{ e^{\theta(X_1+X_2+\ldots+X_n)} \right\} \qquad (2)$$

Aplicando la propiedad distributiva

$$E\left\{ e^{\theta X_1} e^{\theta X_2} e^{\theta X_3} \ldots . e^{\theta X_n} \right\} \qquad (3)$$

Dado que las variables son independientes, lo anterior puede expresarse como:

$$E\left(e^{\theta X_1}\right)\ E\left(e^{\theta X_2}\right)\ldots\ldots E\left(e^{\theta Xn}\right) \qquad (4)$$

Por (1) se sabe que

$$E\left(e^{\theta x}\right) = M_x$$

Teniendo en cuenta (1) en (4)

$$M_{X_1}(\theta)\ M_{X_2}(\theta)\ldots\ldots\ldots\ldots M_{X_n}(\theta) \qquad (5)$$

Pero, en muestreo aleatorio, las características de la población se obtienen a partir de las características de la muestra, entonces:

$$M_{X_i}(\theta) = M_X^n(\theta) \qquad (6)$$

Observación: X_1, X_2,….X_n son los datos para las unidades de observación correspondientes a la variable aleatoria X. Estos datos provienen de una muestra aleatoria de unidades de observación y, por lo tanto, son también variables aleatorias, ya que si se toma otra muestra, las unidades de observación serán otras y los datos de la variable X (X_1, X_2,….X_n) serán otros. Como se dijo anteriormente, en la práctica se toma una sola muestra aleatoria; es decir, estos datos se corresponden con una sola muestra, representando características no aleatorias para las unidades de observación seleccionadas. Entonces, como las características de la población se obtienen a partir de las características de la muestra, cada función generatriz de momentos de los datos coincide con la de la variable observada.

Por ende $M_X^n(\theta)$ es la función generatriz de momento elevado a la potencia n.

En síntesis, la función generatriz de momentos de una suma de variables independientes

$M_{x_1+x_2+\ldots+x_n}(\theta)$

es el producto, de las n funciones generatrices de momentos del parámetro θ, para las n variables

$M_x(\theta)\,M_x(\theta)\ldots\ldots\ldots\ldots M_x(\theta)$

y esto es igual a la función generatriz de momentos del parámetro θ elevado a la n

$M_X^n(\theta)$

$$M_{x_1+x_2+\ldots+x_n}(\theta)=M_x(\theta)\;\;M_x(\theta)\ldots\ldots\ldots\ldots M_x(\theta)\;=\;M_X^n(\theta) \qquad (7)$$

b) La *función generatriz de momentos del producto de una constante por una variable* para el parámetro θ, $M_{cx}(\theta)$, es la función generatriz de momentos del producto del parámetro (θ) por la constante (c) para la variable X

$$M_{CX}(\theta) = M_X(c\,\theta) \qquad (8)$$

Con las propiedades (7) y (8) se puede hallar la función generatriz de momentos de la media aritmética de la muestra en el parámetro θ $\left(M_{\bar{x}}(\theta)\right)$

$$M_{\bar{X}}(\theta) = M_{\frac{1}{n}(x_1+x_2+\ldots+x_n)}(\theta) = M_{(x_1+x_2+\ldots+x_n)}\left(\frac{\theta}{n}\right)$$

Con este último resultado y teniendo en cuenta (7)

$$M_{\bar{X}}(\theta) = M_X^n\left(\frac{\theta}{n}\right) \qquad (9)$$

La función generatriz de momentos de la media $\overline{X}$ en el parámetro (θ), es la función generatriz de momentos de la variable X elevado a la n en el parámetro θ/n.

La expresión (9) relaciona la función generatriz de momentos de la media muestral con la variable observada en la población, a partir de los datos correspondientes a las unidades de observación seleccionadas por medio de una muestra, en esa población.

5.4. DISTRIBUCIONES EN EL MUESTREO DE LA MEDIA

Con la distribución de probabilidad de X (variable aleatoria poblacional) y muestras aleatorias de tamaño n, de donde se obtiene $\overline{X}$ (variable aleatoria muestral), se analizará cómo se distribuye $\overline{X}$. La respuesta no es única, pues depende de cuál sea la distribución de probabilidad de la variable X. Atendiendo a esto, se tendrán dos alternativas, según se esté en presencia de:

a) La variable X con distribución de probabilidad normal
b) La variable X con cualquier tipo de distribución de probabilidad

Distribución de la media muestral, $\overline{X}$, cuando X~N (0,1)

Si la variable aleatoria X~N (μ_X , σ_X), su función de densidad es

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]};\quad -\infty<X<\infty$$

Si se define

$$Z=\frac{X-\mu_X}{\sigma_X} \qquad (10)$$

donde Z ~ N(0,1), entonces Z es una variable aleatoria normal centrada y reducida[2], denominada variable normal estándar. La función de densidad de Z es

$$f(Z)=\frac{1}{\sqrt{2\pi}}e^{\left(-\frac{1}{2}Z^2\right)};\ 0<Z<1$$

La función generatriz de momentos de la variable Z, usando (1), es

$$M_z(\theta)=E\left\{e^{\theta\cdot Z}\right\}=\int_{-\infty}^{\infty}e^{\theta\cdot Z}f(z)dz \qquad (11)$$

[2] Una variable centrada es aquella a la que se le ha restado, a cada valor, el de su media; y es reducida cuando se la divide por su desvío.

Reemplazando $f(z)$ por su igual

$$M_z(\theta) = \int_{-\infty}^{\infty} e^{\theta \cdot Z} \frac{1}{\sqrt{2\pi}} e^{\left(-\frac{1}{2}Z^2\right)} dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\theta \cdot Z} e^{\left(-\frac{Z^2}{2}\right)} dz \qquad (12)$$

En la última expresión se tiene un producto de exponentes de igual base

$$e^{\theta Z} \cdot e^{-\frac{Z^2}{2}}$$

el resultado de este producto es conservar la base y sumar los exponentes

$$e^{\theta Z - \frac{Z^2}{2}} \qquad (13)$$

A fin de encontrar una expresión menos compleja, se suma y resta $\frac{1}{2}\theta^2$ en el exponente de e

$$\theta Z - \frac{Z^2}{2} + \frac{1}{2}\theta^2 - \frac{1}{2}\theta^2$$

Al ordenar convenientemente y sacar factor común $-\frac{1}{2}$

$$-\frac{1}{2}\left(\theta^2 - 2\theta Z + Z^2\right) + \frac{1}{2}\theta^2$$

Se tiene el desarrollo del cuadrado[3]

$$-\frac{1}{2}(Z-\theta)^2 + \frac{1}{2}\theta^2$$

[3] Definir el binomio $(\theta - Z)^2$ es equivalente a $(Z-\theta)^2$. En esta oportunidad es conveniente adoptar la segunda expresión para que, en la sustitución de variables que se realiza en (16), dz sea positivo.

En síntesis:

$$\theta Z - \frac{Z^2}{2} = -\frac{1}{2}(Z-\theta)^2 + \frac{1}{2}\theta^2 \qquad (14)$$

Reemplazando el resultado de (14) en (15)

$$e^{\theta Z - \frac{Z^2}{2}} = e^{\left[-\frac{1}{2}(Z-\theta)^2 + \frac{1}{2}\theta^2\right]} = e^{-\frac{1}{2}(Z-\theta)^2} e^{\frac{1}{2}\theta^2} \qquad (15)$$

El resultado de (15) en (12) da lugar a una nueva expresión de la función generatriz de momentos de la variable Z en el parámetro θ

$$M_Z(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(Z-\theta)^2} e^{\frac{1}{2}\theta^2} dz = \frac{e^{\left(\frac{1}{2}\theta^2\right)}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(Z-\theta)^2} dz \qquad (16)$$

La diferencia entre (16) y (12) radica en la expresión a integrar, (16) es de sencilla resolución a través de sustitución de variables. Para lo cual se define $Z - \theta = t$

Entonces si $t = z - \theta$

$dt = dz$ (porque θ es constante)

Reemplazando en (16)

$$M_Z(\theta) = \frac{e^{\left(\frac{1}{2}\theta^2\right)}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}t^2} dt$$

Esta última expresión tiene la ventaja de contener una integral notable cuyo resultado es $\sqrt{2\pi}$

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}t^2}\,dt = \sqrt{2\pi} \qquad (17)$$

Entonces

$$M_Z(\theta) = \frac{e^{\left(\frac{1}{2}\theta^2\right)}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}t^2}\,dt = \frac{e^{\left(\frac{1}{2}\theta^2\right)}}{\sqrt{2\pi}} \sqrt{2\pi} = e^{\left(\frac{1}{2}\theta^2\right)}$$

En síntesis, si $X \sim \mathrm{N}(0,1)$ se está trabajando con una variable aleatoria estandarizada *Z.* La función generatriz de momentos de la variable Z, en el parámetro θ, es una función exponencial con exponente $\theta^2/2$

$$M_Z(\theta) = e^{\left(\frac{\theta^2}{2}\right)} \qquad (18)$$

Al derivar $M_z(\theta)$, se encuentran los momentos de orden 1 y 2 que permiten hallar la media y la varianza de $\overline{X}$ cuando $X \sim \mathrm{N}(0,1)$

$$\frac{\partial M_Z(\theta)}{\partial \theta} = \frac{\partial\left(e^{\frac{\theta^2}{2}}\right)}{\partial \theta} = e^{\frac{\theta^2}{2}}\,\theta\Big|_{\theta=0} = 0 \;\rightarrow E(X) = 0$$

$$\frac{\partial^2 M_Z(\theta)}{\partial \theta^2} = \frac{\partial\left(e^{\frac{\theta^2}{2}}\right)}{\partial \theta} = e^{\frac{\theta^2}{2}}\,\theta + e^{\frac{\theta^2}{2}}\,1\Big|_{\theta=0} = 1 \;\rightarrow V(X) = m_2 - m_1^2 = 1 - 0 = 1$$

Distribución de la media muestral, $\overline{X}$, cuando X~N(μ_X , σ_X)

Retomando la variable normal estándar definida en (10), si se multiplica en ambos miembros por σ_X , la diferencia entre la observación de la variable y la media es la medida estándar por el desvío:

$$X - \mu_x = Z\sigma_X \qquad (19)$$

La función generatriz de la variable aleatoria normal centrada $x - \mu_X$, en el parámetro θ, es la función generatriz de momentos de la medida estándar por el desvío en el parámetro θ

$$M_{X-\mu_x}(\theta) = M_{Z\sigma_x}(\theta) \text{ (donde } \sigma_x \text{ es constante)}$$

Si se tiene en cuenta lo visto en (8), donde $M_{cx}(\theta) = M_X(c\theta)$, la función generatriz de momentos de la variable centrada $x - \mu_X$ en el parámetro θ, es la función generatriz de momentos de la medida estándar por el desvío en el parámetro θ, que es igual a la función generatriz de momentos de la variable estándar Z en el parámetro $\sigma_X\theta$

$$M_{X-\mu_x}(\theta) = M_{Z\sigma_x}(\theta) = M_Z(\sigma_X\theta)$$

Por (18) se sabe que la función generatriz de momentos de la variable Z en el parámetro θ es $e^{\theta^2/2}$ $\left(M_Z(\theta) = e^{\frac{\theta^2}{2}}\right)$

Entonces

$$M_Z(\theta\sigma_X) = e^{\left(\frac{(\theta\sigma_x)^2}{2}\right)} = e^{\left[\frac{1}{2}(\theta\sigma_x)^2\right]}$$

Es decir, la función generatriz de momentos de la variable normal aleatoria centrada $x - \mu_X$ en el parámetro θ es

$$M_{X-\mu_x}(\theta) = e^{\left[\frac{1}{2}(\theta\sigma_x)^2\right]} \qquad (20)$$

La variable aleatoria Z fue reexpresada de acuerdo a (19), si se suma μ_x en ambos miembros se tiene la expresión de la variable aleatoria X

$$X = Z\sigma_x + \mu_x \qquad (21)$$

La función generatriz de momentos de una variable aleatoria en el parámetro θ, teniendo en cuenta (21), es la función generatriz de momentos de la suma de un producto (la variable Z ponderada por el desvío) y una constante en el parámetro θ

$$M_X(\theta) = M_{Z\sigma_x + \mu_x}(\theta) \qquad (22)$$

Pero, teniendo en cuenta una propiedad de función generatriz de momentos

$$M_{c+X}(\theta) = e^{c\theta} M_X(\theta) \qquad (23)$$

y la ya expresada en (8), $M_{cX}(\theta) = M_X(c\theta)$, la expresión (22) es igual a

$$M_X(\theta) = M_{Z\sigma_x + \mu_x}(\theta) = e^{\mu_x \theta} M_Z(\sigma_X \theta) \qquad (24)$$

La función generatriz de momentos de la variable X en el parámetro θ es el producto de una constante, $e^{\mu_x \theta}$, y la función generatriz de momentos de una variable Z en el parámetro $\sigma_x \theta$.

Con el resultado alcanzado en (18), se reexpresa $M_z(\sigma_x \theta)$ de (24)

$$\mathrm{M_z}(\sigma_\mathrm{x}\theta) = \mathrm{e}^{\frac{(\sigma_\mathrm{x}\theta)^2}{2}}$$

y se tiene

$$\mathrm{M_X}(\theta) = \mathrm{e}^{\left(\mu_\mathrm{x}\theta + \frac{1}{2}\sigma_\mathrm{x}^2\theta^2\right)} \qquad (25)$$

(25) es la función generatriz de momentos de la variable aleatoria X normal de media μ_X y desviación estándar σ_X

En (9) se había hallado la función generatriz de momentos de la media muestral en el parámetro θ

$$M_{\bar{X}}(\theta) = M_X^n\left(\frac{\theta}{n}\right)$$

sustituyendo por (25), queda la función generatriz de momentos de la variable aleatoria *X* elevada a la *n*

$$M_X^n\left(\frac{\theta}{n}\right) = \left[e^{\mu_x \frac{\theta}{n} + \frac{1}{2}\sigma_x^2 \frac{\theta^2}{n^2}}\right]^n$$

Por lo tanto, la función generatriz de momentos de la media muestral, $\overline{X}$, cuando X se distribuye normal es

$$M_{\bar{x}}(\theta) = \left[e^{\mu_x \frac{\theta}{n} + \frac{1}{2}\sigma_x^2 \frac{\theta^2}{n^2}}\right]^n = \left[e^{\mu_x \frac{\theta}{n}} e^{\frac{1}{2}\sigma_x^2 \frac{\theta^2}{n^2}}\right]^n = \left(e^{\mu_x \frac{\theta}{n}}\right)^n \left(e^{\frac{1}{2}\sigma_x^2 \frac{\theta^2}{n^2}}\right)^n = e^{\mu_x \frac{\theta}{n} n} e^{\frac{1}{2}\sigma_x^2 \frac{\theta^2}{n^2} n}$$

$$M_{\bar{x}}(\theta) = e^{\mu_x \theta + \frac{1}{2}\sigma_x^2 \frac{\theta^2}{n}} \qquad (26)$$

Al comparar las funciones generatrices de momentos de las variables X y $\overline{X}$ dadas en (25) y (26)

$$M_X(\theta) = e^{\theta \mu_X + \frac{1}{2}\theta^2 \sigma_X^2}$$

$$M_{\bar{X}}(\theta) = e^{\theta \mu_X + \frac{1}{2}\theta^2 \frac{\sigma_X^2}{n}}$$

se acepta el *teorema del límite central* que dice: "Si la variable aleatoria X asociada a una población tiene distribución normal de media μ_x y desviación estándar σ_x, entonces la variable aleatoria $\overline{X}$ asociada a la distribución de medias de las muestras también es normal, de media:

$$E(\overline{X}) = \mu_{\overline{X}} = \mu_X$$

y desvío:

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

Para demostrar el teorema se calcula el momento de orden 1 y el momento de orden 2. Para esto se deriva la función generatriz de momentos respecto de ϑ, y se evalúa cuando ϑ en 0. Entonces,

$$M_1 = \frac{\partial M_{\overline{X}}(\theta)}{\partial \theta} = e^{\mu_X \theta + \frac{1}{2}\sigma_X^2 \frac{\theta^2}{n}} \left(\mu_X + \frac{\sigma_X^2 \theta}{n} \right) \qquad (27)$$

Evaluando en $\vartheta = 0$

$$M_1 = e^{\mu_X 0 + \frac{1}{2}\sigma_X^2 \frac{0^2}{n}} \left(\mu_X + \frac{\sigma_X^2}{n} 0 \right) = \mu_X$$

De modo que el momento de orden 1 es

$$M_1 = E(\overline{X}) = \mu_X \qquad (28)$$

Con lo que se demuestra la propiedad enunciada más arriba, esto es que la media aritmética de la muestra es un estimador insesgado de la media poblacional.

Para hallar el momento de orden 2, se deriva la derivada que dio origen al momento de orden 1:

$$M_2 = \frac{\partial^2 M_{\overline{X}}(\theta)}{\partial \theta^2} = e^{\mu_X \theta + \frac{1}{2}\theta^2 \frac{\sigma_X^2}{n}} \left(\mu_X + \frac{\sigma_X^2}{n} \theta \right)^2 + e^{\mu_X \theta + \frac{1}{2}\theta^2 \frac{\sigma_X^2}{n}} \frac{\sigma_X^2}{n}$$

Evaluando en $\vartheta = 0$

$$M_2 = e^0\left(\mu_X + \frac{\sigma_X^2}{n}0\right)^2 + e^0\,\frac{\sigma_X^2}{n}$$

Entonces el momento centrado de orden 2 es

$$M_2 = \mu_X^2 + \frac{\sigma_X^2}{n}$$

La varianza de $\overline{X}$ es,

$$V(\overline{X}) = \sigma_{\overline{X}}^2 = M_2 - M_1^2 = \mu_X^2 + \frac{\sigma_X^2}{n} - \mu_X^2 = \frac{\sigma_X^2}{n} \qquad (29)$$

y el desvío de $\overline{X}$ será

$$\sigma_{\overline{X}} = \sqrt{\frac{\sigma_X^2}{n}} = \frac{\sigma_X}{\sqrt{n}} \qquad (30)$$

Los resultados alcanzados en (28) y (30) confirman la aseveración del teorema.

En síntesis, si $X \sim N(\mu_X, \sigma_X)$ la función generatriz de momentos de la variable aleatoria X en el parámetro θ es $M_X(\theta) = e^{\theta\mu_X + \frac{1}{2}\theta^2\sigma_X^2}$ y la función generatriz de momentos de la variable aleatoria $\overline{X}$ es $M_{\overline{X}}(\theta) = e^{\theta\mu_X + \frac{1}{2}\theta^2\frac{\sigma_X^2}{n}}$ siendo $\overline{X} \sim N(\mu_X, \frac{\sigma_X}{\sqrt{n}})$.

Distribución de la media muestral , $\overline{X}$, cuando X tiene distribución diferente a la normal

Es decir, qué sucede si X tiene alguna distribución distinta a la normal pero tiene una media μ_X y un desvío σ_X.

Para demostrarlo se define una variable aleatoria estandarizada correspondiente a la distribución de medias de las muestras:

$$t = \frac{\overline{X} - \mu_x}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu_x}{\sigma_X / \sqrt{n}}$$

La función generatriz de momentos de *t* en el parámetro θ

$$M_t(\theta) = M_{\frac{\overline{X} - \mu_x}{\sigma_X / \sqrt{n}}}(\theta)$$ donde $\sigma_X / \sqrt{n}$ es constante

Aplicando lo visto en (8) que $M_{CX}(\theta) = M_X(C\theta)$

$$M_t(\theta) = M_{\overline{X} - \mu_X}\left(\frac{\theta}{\sigma_X / \sqrt{n}}\right)$$ donde μ_X es constante

En (23) se vio que $M_{C+X}(\theta) = e^{C\theta} M_X(\theta)$, con lo cual

$$M_t(\theta) = e^{\frac{-\mu_X \theta}{\sigma_X / \sqrt{n}}} M_X\left(\frac{\theta}{\sigma_X / \sqrt{n}}\right)$$

Y en (9) que $M_{\overline{X}}(\theta) = M_X^n\left(\frac{\theta}{n}\right)$

Entonces,

$$M_t(\theta) = e^{\frac{-\mu_x \theta \sqrt{n}}{\sigma_X}} M_X^n\left(\frac{\theta / n}{\sigma_X / \sqrt{n}}\right) = e^{\frac{-\mu_x \theta \sqrt{n}}{\sigma_X}} M_X^n\left(\frac{\theta}{\sigma_X \sqrt{n}}\right) \quad (31)$$

Esto es así porque

$$\frac{\theta/n}{\sigma_X/\sqrt{n}} = \frac{\theta\sqrt{n}}{\sigma_X n} = \frac{\theta n^{1/2}}{\sigma_X n} = \frac{\theta n^{1/2-1}}{\sigma_X} = \frac{\theta n^{-1/2}}{\sigma_X} = \frac{\theta}{\sigma_X\sqrt{n}}$$

Al tomar logaritmo natural en (31)

$$In[M_t(\theta)] = \frac{-\mu_X\theta\sqrt{n}}{\sigma_X} + n\cdot In\left[M_X\left(\frac{\theta}{\sigma_X\cdot\sqrt{n}}\right)\right] \qquad (32)$$

Hay que tener en cuenta que

$$\mathrm{M_X(\theta) = E\{e^{\theta X}\}}$$

y que $e^{\theta X}$ puede desarrollarse en series de potencias, esto significa que el número *e* elevado a cualquier potencia se puede expresar de la siguiente manera:

$$\mathrm{e^w = 1 + w + \frac{w^2}{2!} + \frac{w^3}{3!} + \ldots}$$

Entonces, haciendo $w = \theta X$

$$M_X(\theta) = E\left\{1 + \theta X + \frac{(\theta X)^2}{2!} + \frac{(\theta X)^3}{3!} + \ldots\right\}$$

La esperanza es distributiva respecto de la suma

$$M_X(\theta) = E\{1\} + E\{\theta X\} + E\left\{\frac{(\theta X)^2}{2!}\right\} + E\left\{\frac{(\theta X)^3}{3!}\right\} + \ldots$$

Pero luego $E\{CX\} = C\cdot E\{X\}$

$$M_X(\theta) = E\{1\} + \theta E\{X\} + \frac{\theta^2}{2!}E\{X^2\} + \frac{\theta^3}{3!}E\{X^3\} + \ldots$$

De modo que la función generatriz de la variable X es

$$M_X(\theta) = 1 + \theta\mu_1^{\prime} + \frac{\theta^2}{2!}\mu_2^{\prime} + \frac{\theta^3}{3!}\mu_3^{\prime} + \ldots \qquad (33)$$

Donde se ha tenido en cuenta que $E\{1\} = E\{C\} = C = 1$; y que, $E\{X\} = \mu_1^{\prime}$ es el momento con respecto al origen de la variable aleatoria X de órdenes crecientes.

Derivando sucesivamente la expresión (33)

$$\frac{d}{d\theta} M_X(\theta) = \frac{d}{d\theta}\left[1 + \theta\mu_1^{\prime} + \frac{\theta^2}{2!}\mu_2^{\prime} + \frac{\theta^3}{3!}\mu_3^{\prime} + \frac{\theta^4}{4!}\mu_4^{\prime} + \ldots\right]$$

$$= 0 + \mu_1^{\prime} + \frac{2}{2}\theta\mu_2^{\prime} + \frac{3}{3\cdot 2}\theta^2\mu_3^{\prime} + \frac{4}{4\cdot 3\cdot 2}\theta^3\mu_4^{\prime} + \ldots$$

Si $\theta = 0$ entonces $\frac{d}{d\theta} M_X(\theta) = \mu_1^{\prime}$

A continuación se busca la segunda derivada

$$\frac{d^2}{d\theta^2} M_X(\theta) = \frac{d}{d\theta}\left[\mu_1^{\prime} + \frac{2}{2}\theta\mu_2^{\prime} + \frac{3}{3\cdot 2}\theta^2\mu_3^{\prime} + \frac{4}{4\cdot 3\cdot 2}\theta^3\mu_4^{\prime} + \ldots\right]$$

$$= 0 + \mu_2^{\prime} + \frac{2}{2}\theta\mu_3^{\prime} + \frac{3}{3\cdot 2}\theta^2\mu_4^{\prime} + \ldots$$

Si $\theta = 0$ entonces $\frac{d^2}{d\theta^2} M_X(\theta) = \mu_2^{\prime}$

Luego la tercer derivada

$$\frac{d^3}{d\theta^3} M_X(\theta) = \frac{d}{d\theta}\left[\mu_2^{\prime} + \frac{2}{2}\theta\mu_3^{\prime} + \frac{3}{3\cdot 2}\theta^2\mu_4^{\prime} + \ldots\right]$$

$$= 0 + \mu_3^{\prime} + \theta\mu_4^{\prime} + \ldots$$

Si $\theta = 0$ entonces $\frac{d^3}{d\theta^3} M_X(\theta) = \mu_3^{\prime}$

Las derivaciones anteriores permiten generalizar el resultado en la siguiente expresión

$$\frac{d^k}{d\theta^k} M_X(\theta)\Big|_{\theta=0} = \mu_k^l$$

La derivada k-ésima respecto al parámetro θ, k veces, evaluada en $\theta = 0$, es el k-ésimo momento de la variable aleatoria con respecto al origen.

En (32) se tenía que

$$In[M_t(\theta)] = \frac{-\mu_X \theta \sqrt{n}}{\sigma_X} + n \cdot In\left[M_X\left(\frac{\theta}{\sigma_X \cdot \sqrt{n}}\right)\right]$$

Remplazando aquí la expresión encontrada en (33) para $M_X\left(\frac{\theta}{\sigma_X \cdot \sqrt{n}}\right)$, se tiene

$$\begin{aligned} \mathrm{In}[\mathrm{M_t}(\theta)] = \frac{-\mu_x \theta \sqrt{n}}{\sigma_X} + n \cdot \ln\Bigg[1 + \frac{\theta}{\sigma_X \sqrt{n}} \mu_1^l + \frac{1}{2!} \frac{\theta^2}{\sigma_X^2 n} \mu_2^l + \\ + \frac{1}{3!} \frac{\theta^3}{\sigma_X^3 n^{3/2}} \mu_3^l + \frac{1}{4!} \frac{\theta^4}{\sigma_X^4 n^{4/2}} \mu_4^l + \ldots \Bigg] \end{aligned} \qquad (34)$$

Si n>30 se lo considera lo suficientemente grande como para que lo contenido en el corchete sea una suma de términos que tiende a 1 + algo cercano a 0. Esto se puede desarrollar usando una expansión en serie de potencia donde

$$In(1+Z) = Z - \frac{Z^2}{2} + \frac{Z^3}{3} - \frac{Z^4}{4} + \ldots \quad |Z| < 1$$

Z será igual a

$$Z = \frac{\theta}{\sqrt{n}\,\sigma_X}\mu_1^l + \frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\mu_2^l + \frac{1}{3!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_3^l + \frac{1}{4!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_4^l + \ldots$$

que es parte de la expresión que se encuentra entre corchetes en (34).

Si se desarrolla la serie en potencia

$$\begin{aligned} Ln(1+Z) = &\left[\frac{\theta}{\sqrt{n}\,\sigma_X}\mu_1^l + \frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\mu_2^l + \frac{1}{3!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_3^l + \frac{1}{4!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_4^l + \ldots\right] - \\ &-\frac{1}{2}\left[\frac{\theta}{\sqrt{n}\,\sigma_X}\mu_1^l + \frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\mu_2^l + \frac{1}{3!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_3^l + \frac{1}{4!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_4^l + \ldots\right]^2 + \\ &+\frac{1}{3}\left[\frac{\theta}{\sqrt{n}\,\sigma_X}\mu_1^l + \frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\mu_2^l + \frac{1}{3!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_3^l + \frac{1}{4!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_4^l + \ldots\right]^3 - \\ &-\frac{1}{4}\left[\frac{\theta}{\sqrt{n}\,\sigma_X}\mu_1^l + \frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\mu_2^l + \frac{1}{3!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_3^l + \frac{1}{4!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_4^l + \ldots\right]^4 + \ldots \end{aligned}$$

Luego se tienen que desarrollar los cuadrados, los cubos, las cuartas potencias… Para resolver el cuadrado puede considerarse que se está en presencia de un binomio (A+B)

$$\underbrace{\left[\frac{\theta}{\sqrt{n}\,\sigma_X}\mu_1^l + \frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\mu_2^l + \frac{1}{3!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_3^l + \frac{1}{4!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_4^l + \ldots\right]^2}_{\text{A}\qquad\qquad\qquad\qquad\text{B}}$$

El desarrollo del cuadrado da lugar a la expresión

$$\frac{\theta^2}{n\sigma_X^2}\left(\mu_1^{\prime}\right)^2+2\frac{\theta}{\sqrt{n}\,\sigma_X}\mu_1^{\prime}\left[\frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\mu_2^{\prime}+\frac{1}{3!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_3^{\prime}+\frac{1}{4!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_4^{\prime}+\ldots\right]+$$

$$+\left[\frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\mu_2^{\prime}+\frac{1}{3!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_3^{\prime}+\frac{1}{4!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_4^{\prime}+\ldots\right]^2$$

De igual manera hay que trabajar para resolver las potencias restantes. Entonces, se tiene que

$$\begin{aligned}\ln(1+Z)=&\frac{\theta}{\sqrt{n}\,\sigma_X}\mu_1^{|}+\frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\mu_2^{|}+\frac{1}{3!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_3^{|}+\frac{1}{4!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_4^{|}+\frac{1}{5!}\frac{\theta^5}{n^{5/2}\sigma_X^5}\mu_5^{|}+\cdots\\&-\frac{1}{2}\left\{\frac{\theta^2}{n\sigma_X^2}\left(\mu_1^{|}\right)^2+\frac{2}{2!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_1^{|}\,\mu_2^{|}+\frac{2}{3!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_1^{|}\,\mu_3^{|}+\frac{2}{4!}\frac{\theta^5}{n^{5/2}\sigma_X^5}\mu_1^{|}\,\mu_4^{|}\right.\\&\left.+\frac{1}{6!}\frac{\theta^6}{n^{6/2}\sigma_X^6}\mu_1^{|}\,\mu_5^{|}+\cdots\left[\frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\mu_2^{|}+\frac{1}{3!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_3^{|}+\frac{1}{4!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_4^{|}+\cdots\right]^2\right\}\\&+\frac{1}{3}\left[\frac{\theta}{\sqrt{n}\,\sigma_X}\mu_1^{|}+\frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\mu_2^{|}+\frac{1}{3!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_3^{|}+\frac{1}{4!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_4^{|}+\cdots\right]^3+\cdots\end{aligned}$$

Agrupando los términos por factor común

$$\begin{aligned}\ln(1+Z)=&\frac{\theta}{\sqrt{n}\,\sigma_X}\mu_1^{|}+\frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\left[\mu_2^{|}-\left(\mu_1^{|}\right)^2\right]+\frac{\theta^3}{n^{3/2}\sigma_X^3}\left[\frac{\mu_3^{|}}{3!}-\frac{\mu_1^{|}\,\mu_2^{|}}{2!}\right]\\&+\frac{\theta^4}{n^{4/2}\sigma_X^4}\left[\frac{\mu_4^{|}}{4!}-\frac{\mu_1^{|}\,\mu_3^{|}}{3!}\right]+\frac{\theta^5}{n^{5/2}\sigma_X^5}\left[\frac{\mu_5^{|}}{5!}-\frac{\mu_1^{|}\,\mu_4^{|}}{4!}\right]+\frac{\theta^6}{n^{6/2}\sigma_X^6}\left[\frac{\mu_6^{|}}{6!}-\frac{\mu_1^{|}\,\mu_5^{|}}{5!}\right]\\&+\cdots\left[\frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\mu_2^{|}+\frac{1}{3!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_3^{|}+\frac{1}{4!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_4^{|}+\cdots\right]^2\\&+\frac{1}{3}\left[\frac{\theta}{\sqrt{n}\,\sigma_X}\mu_1^{|}+\frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\mu_2^{|}+\frac{1}{3!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_3^{|}+\frac{1}{4!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_4^{|}+\cdots\right]^3+\cdots\end{aligned}$$

Reemplazando el resultado del ln(1+Z) en (34)

$$\ln[M_t(\theta)] = \frac{-\mu_X\theta\sqrt{n}}{\sigma_X} + n\left\{ \frac{\theta}{\sqrt{n}\,\sigma_X}\mu_1^{|} + \frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\left[\mu_2^{|} - \left(\mu_1^{|}\right)^2\right] + \frac{\theta^3}{n^{3/2}\sigma_X^3}\left[\frac{\mu_3^{|}}{3!} - \frac{\mu_1^{|}\,\mu_2^{|}}{2!}\right] + \frac{\theta^4}{n^{4/2}\sigma_X^4}\left[\frac{\mu_4^{|}}{4!} - \frac{\mu_1^{|}\,\mu_3^{|}}{3!}\right] + \frac{\theta^5}{n^{5/2}\sigma_X^5}\left[\frac{\mu_5^{|}}{5!} - \frac{\mu_1^{|}\,\mu_4^{|}}{4!}\right] + \frac{\theta^6}{n^{6/2}\sigma_X^6}\left[\frac{\mu_6^{|}}{6!} - \frac{\mu_1^{|}\,\mu_5^{|}}{5!}\right] + \cdots\left[\frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\mu_2^{|} + \frac{1}{3!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_3^{|} + \frac{1}{4!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_4^{|} + \cdots\right]^2 + \frac{1}{3}\left[\frac{\theta}{\sqrt{n}\,\sigma_X}\mu_1^{|} + \frac{1}{2!}\frac{\theta^2}{n\sigma_X^2}\mu_2^{|} + \frac{1}{3!}\frac{\theta^3}{n^{3/2}\sigma_X^3}\mu_3^{|} + \frac{1}{4!}\frac{\theta^4}{n^{4/2}\sigma_X^4}\mu_4^{|} + \cdots\right]^3 + \cdots\right\}$$

Al reordenar y simplificar la expresión:

$$\ln[M_t(\theta)] = \frac{-\mu_X\sqrt{n}}{\sigma_X}\theta + \frac{\sqrt{n}}{\sigma_X}\theta\mu_1^{|} + \frac{\mu_2^{|} - \left(\mu_1^{|}\right)^2}{2!\,\sigma_X^2}\theta^2 + \frac{1}{n^{1/2}\sigma_X^3}\left[\frac{\mu_3^{|}}{3!} - \frac{\mu_1^{|}\,\mu_2^{|}}{2!}\right]\theta^3 + \frac{1}{n^{2/2}\sigma_X^4}\left[\frac{\mu_4^{|}}{4!} - \frac{\mu_1^{|}\,\mu_3^{|}}{3!}\right]\theta^4 + \frac{1}{n^{3/2}\sigma_X^5}\left[\frac{\mu_5^{|}}{5!} - \frac{\mu_1^{|}\,\mu_4^{|}}{4!}\right]\theta^5 + \cdots términos\ en\ \theta^k\ para\ k > 5$$

Si se tiene en cuenta que $\mu_X = \mu_1^{|}$ y que $\sigma_X^2 = \mu_2^{|} - \left(\mu_1^{|}\right)^2$, entonces

$$\ln[M_t(\theta)] = \underbrace{\left[-\frac{\mu_X\sqrt{n}}{\sigma_X} + \frac{\mu_X\sqrt{n}}{\sigma_X}\right]}_{1}\theta + \frac{\sigma_X^2}{2!\,\sigma_X^2}\theta^2 + \frac{1}{n^{1/2}\sigma_X^3}\left[\frac{\mu_3^{|}}{3!} - \frac{\mu_1^{|}\,\mu_2^{|}}{2!}\right]\theta^3 + \frac{1}{n^{2/2}\sigma_X^4}\left[\frac{\mu_4^{|}}{4!} - \frac{\mu_1^{|}\,\mu_3^{|}}{3!}\right]\theta^4 + \frac{1}{n^{3/2}\sigma_X^5}\left[\frac{\mu_5^{|}}{5!} - \frac{\mu_1^{|}\,\mu_4^{|}}{4!}\right]\theta^5 + \cdots términos\ en\ \theta^k\ para\ k > 5$$

Simplificando la expresión

$$\mathrm{In}\left[\mathrm{M}_t(\theta)\right] = \frac{\theta^2}{2} + \text{términos en } \theta^k \text{ con k>2}$$

Pero los términos en θ^k cuando k>2 tienden a 0 cuando $n \to \infty$, por lo que se puede afirmar que

$$\lim_{n\to\infty} In[M_t(\theta)] = \frac{\theta^2}{2}$$

Es decir que $\lim_{n\to\infty} M_t(\theta) = e^{\frac{\theta^2}{2}}$ porque $\ln \cdot e^{\frac{\theta^2}{2}} = \frac{\theta^2}{2}$ (35)

Si se compara (35) con (18) se observa que la función generatriz de momentos de la variable aleatoria estándar *t* tiende a la función generatriz de momentos de la distribución normal estándar a medida que el tamaño de la muestra aumenta.

Este se puede expresar en el siguiente teorema: "Si X es una variable aleatoria con cualquier distribución de probabilidad de media μ_X y desviación estándar σ_X, para la cual existe su función generatriz de momentos, entonces la variable aleatoria $\overline{X}$ tiende asintóticamente a la distribución normal de media μ_X y desviación estándar $\sigma_X / \sqrt{n}$".

En el Teorema Central del Límite radica la importancia de la distribución normal, ya que basta con que el tamaño de la muestra sea suficientemente grande, mayor a 30, para poder asegurar la buena aproximación de la distribución de medias de la muestra a la distribución normal, independientemente de la propia distribución que tenga la población de donde se extraigan las muestras.

Con los subincisos anteriores queda resuelto el problema de la distribución de probabilidad de las medias de las muestras, cuando estas provienen de una población normal o de una población con cualquier distribución de probabilidad y el tamaño de las muestras es grande. En ambos casos se recurre a la distribución normal.

Para Gomez Villegas (2005), la aplicación del teorema central del límite como justificación de que los errores de medida se distribuyen aproximadamente según una normal es considerada omo una de las más importantes aportaciones científicas. En realidad, en los siglos XVII y XVIII, el teorema central del límite se conoció como la ley de frecuencias de los errores. Muchos científicos pensaron que la ley de frecuencias de los errores suponía un gran avance. Observe al respecto el comentario de Francis Galton (*La herencia natural* publicado en 1889) "difícilmente conozco algo que alimente tanto mi imaginación como el maravilloso orden cósmico que se deriva de la "Ley de frecuencias de los errores" Si los griegos hubieran conocido esta ley, seguro que la habrían endiosado. Reina con serenidad y en completa auto-modestia entre la confusión más salvaje. Cuanto más vigentes están la ley de la calle y la aparente anarquía, más perfecto es su balanceo. Es la ley suprema de la sinrazón".

5.5. FACTOR DE CORRECCIÓN PARA POBLACIONES FINITAS

Las poblaciones pueden ser finitas o infinitas. Si de una población finita se obtienen muestras con reemplazo, puede considerarse teóricamente que la población es infinita, ya que cualquier número de muestras pueden obtenerse de la población sin agotarla. Además, para casi cualquier propósito práctico, el muestreo de poblaciones finitas de tamaño muy grande puede considerase como muestreo de poblaciones infinitas.

De acuerdo al concepto de muestreo aleatorio, el muestreo sin reemplazo de poblaciones finitas no conduce a muestras aleatorias, debido a que al dejar permanentemente fuera el elemento ya extraído, influye en la probabilidad de extracción del siguiente; o sea, que las pruebas sucesivas para obtener una muestra no son independientes. Esto no ocurre si la población es infinita.

Resumiendo lo anterior puede decirse que, para obtener muestras aleatorias, el muestreo debe ser de cualquier tipo en poblaciones infinitas y sólo con reemplazo en poblaciones finitas, a excepción del caso de poder considerar a la población infinita por ser de tamaño muy grande.

Con respecto al teorema, cabe indicar que cuando la población es finita y el muestreo se realiza sin reemplazo, hay que incluir -en la determinación de σ_X- el valor del tamaño de la población, es decir,

$$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

donde $\frac{N-n}{N-1}$ recibe el nombre de factor de corrección para poblaciones finitas en muestreo sin reemplazo. Obsérvese que este factor de corrección tiende a cero cuando $n \to N$.

Ejemplo. La vida útil de los focos fabricados por una empresa tiene distribución normal de media 200 horas y desviación estándar 25 horas. Calcular la probabilidad de que la duración media de 25 focos escogidos al azar sea superior a 208 horas.

De acuerdo al enunciado del problema, la población de las duraciones de los focos de la empresa tiene distribución normal de parámetros.

$\mu_x = 200$ horas y $\sigma_x = 25$ horas

y puede considerarse que es de tamaño infinito. Entonces, la distribución de las medias de las muestras también tendrá distribución normal de parámetros:

$n = 25$ focos

$\mu_{\bar{x}} = \mu_x = 200$ horas

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{25}{\sqrt{25}} = 5 \quad \text{horas}$$

Estandarizando la variable aleatoria $\overline{X}$, se tiene

$$Z = \frac{\overline{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\overline{X} - 200}{5}$$

Si $\overline{X} = 208, \quad z = \frac{208 - 200}{5} = 1.6$

Por lo tanto, la probabilidad de que la media de los 25 focos de la muestra tenga un valor mayor a 208 horas valdrá:

$$P(\overline{X} > 208) = P(z > 1.6)$$
$$= 1 - P(z \le 1.6)$$
$$= 1 - 0.9452$$
$$= 0.0548$$

Sólo el 5.48% de los focos de la muestra tendrán una vida superior a las 208 horas, es decir, un solo foco.

Ejemplo. Los pesos de 600 ejes producidos en un torno tienen distribución normal de media 53 kg y desviación estándar de 2.5 kg. Estos se empacan en cajas de 10, las que soportan hasta 540 kg de peso. Si se envían 35 de estas cajas, calcular cuántas cajas cabe esperar se rompan por exceso de peso.

La población de pesos de los ejes tiene distribución normal de acuerdo al enunciado explícito del problema. Sus parámetros son:

$$N = 600, \quad \mu_x = 53\text{Kg}, \quad \sigma_x = 2.5\text{Kg}$$

Como la población es normal, la distribución de medias de las muestras también es normal y sus parámetros valen;

$$n = 10 \quad \mu_{\bar{x}} = \mu_x = 53\text{Kg}$$

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{2.5}{\sqrt{10}}\sqrt{\frac{600-10}{600-1}} = 0.785\text{Kg}$$

Una caja de ejes se romperá por exceso de peso, si el peso conjunto de los 10 ejes es mayor de 540 kg; y el peso conjunto de los 10 ejes es mayor de 540 kg si el peso medio de los 10 ejes es superior a 54kg (540/10).

Entonces, se trata de calcular cuántas de las 35 muestras tienen media mayor de 54 kg, lo que se hará por medio del cálculo de la probabilidad $P(\overline{X} > 54)$, en donde $\overline{X}$ tiene distribución normal de media 53 y desviación estándar 0.785.

Estandarizando $\overline{X} = 54$, se tiene

$$z = \frac{\overline{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{54 - 53}{0.785} = 1.274$$

Por lo tanto,

$$P\left(\overline{X} > 54\right) = P(z > 1.274)$$
$$= 1 - P(z \le 1.274)$$
$$= 1 - 0.8987$$
$$= 0.1013$$

El 10.13% de las 35 cajas se romperán, o sea, de 3 a 4 cajas.

Ejemplo. Los tabiques comprimidos que se usan en una construcción tienen un peso medio de 5.50 kg y una desviación estándar de 0.85 kg. Estos se elevan en lotes al lugar en donde se emplean por medio de una grúa cuyo límite de seguridad es de 200 kg. Calcular el tamaño máximo de los lotes de manera de que la probabilidad de exceder el límite de seguridad de la grúa sea menor de 5%.

Como en el ejemplo anterior, se vuelve a tener una población de pesos cuyos parámetros son

$\mu_X = 5.50Kg$ y $\sigma_X = 0.85Kg$

con distribución de probabilidad desconocida.

La simple comparación entre el peso de un tabique y el límite de seguridad de la grúa, hace que los lotes de tabiques que se cargan den muestras de tamaño grande. Esto permite asegurar que la distribución de medias de las muestras de pesos de los tabiques es aproximadamente normal de parámetros

$$\mu_{\bar{x}} = \mu_x = 5.50\text{Kg} \text{ y } \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{0.85}{\sqrt{n}}$$

en donde sólo se sabe que n es grande (mayor de 30).

El peso de la muestra de n tabiques excede el límite de seguridad de la grúa de 200 kg, si la media de la muestra es superior a 200/n. Se busca saber cuál es el tamaño de la muestra del lote (cantidad de tabiques) que hace que el límite de seguridad sea superado. El problema se plantea como:

$$P\left(\overline{X} > \frac{200}{n}\right) < 0.05$$

Estandarizando *x* resulta $z = \dfrac{\overline{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \dfrac{\overline{X} - 5.50}{0.85/\sqrt{n}}$

Si $\overline{X} = \dfrac{200}{n}$;

Entonces $z = \dfrac{\frac{200}{n} - 5.50}{\frac{0.85}{\sqrt{n}}} = \dfrac{200 - 5.50n}{0.85\sqrt{n}}$

De modo que

$$P\left(\overline{X} > \frac{200}{n}\right) = P\left(z > \frac{200 - 5.50n}{0.85\sqrt{n}}\right) < 0.05$$

lo que es lo mismo $P\left(z \le \frac{200-5.50n}{0.85\sqrt{n}}\right) \ge 0.95$

De la tabla de probabilidades de la normal se obtiene que ésta es efectivamente mayor de 0.95 si el valor crítico de *z* es mayor o igual a 1.654.

Luego se obtiene

$$\frac{200-5.50n}{0.85\sqrt{n}} \ge 1.645$$

$$200 - 5.50n \ge 1.398\sqrt{n}$$

Elevando ambos miembros al cuadrado

$$(200 - 5.50n)^2 \ge (1.398\sqrt{n})^2$$

$$40000 - 2200n + 30.25n^2 \ge 1.955n$$

$$30.25n^2 - 2201.955n + 40000 \ge 0$$

Resolviendo la ecuación se obtienen las raíces 34.86 y 37.93. Cada una de estas raíces representa dos tamaños de muestras diferentes.

Si $z \le \frac{200-5.50n}{0.85\sqrt{n}}$, cuando $n = 34.86 \Rightarrow z \le -1.645$ y $n = 37.93 \Rightarrow z \le 1.645$

Luego

$$P(z < -1.645) = 0.05$$

$$P(z < 1.645) = 0.95$$

Se busca que el tamaño del lote no supere el límite de seguridad de la grúa fijado en el 5%. Por esta razón, el tamaño de muestra requerido es

$$n \le 34.86$$

CASOS DE ESTUDIO, PREGUNTAS Y PROBLEMAS

Caso 5.1: Edades y estaturas del curso Inferencia Estadística.

El tema de estudio es la distribución de las edades y las estaturas en el curso de Inferencia Estadística del presente ciclo lectivo.

El objetivo general es conocer las medidas antropométricas de la población; específicamente se busca hallar la media de estatura y la media de edad, adicionalmente utilizar la teoría de distribuciones en el muestreo para inferir datos a la población.

Se parte de las siguientes premisas:

- la existencia de datos extremos en la muestra aumenta la dispersión de la media
- a medida que aumenta el número de muestras, la media de las muestras tiende a la media de la población

Para realizar este estudio se debe:

a) Generar una tabla de datos de n observaciones por 2 variables cuantitativas; donde. la unidad de observación sean los alumnos de Inferencia Estadística, y las variables su respectivas edad y estatura.
b) Cada alumno debe

1.seleccionar cinco muestras de tamaño cinco.
2.calcular la media de cada muestra
3.calcular la media de las cinco muestras

c) Generar una tabla con las medias de cada una de las cinco muestras generadas por cada alumno (el número de observaciones de esta tabla será nx5) y calcular la media de las medias.
d) Generar una tabla con la media de cada alumno (el número de observaciones de esta tabla será n) y calcular la media de las medias.
e) En la tabla generada en a), calcular la media de la población.
f) Comparar las medias obtenidas en c), d) y e).

Problemas

5.1. Las calificaciones de examen final, de un curso de introducción a la estadística, tiene distribución normal con una media de 73 y una desviación estándar de 8.

1) ¿cuál es la probabilidad de obtener, como máximo, una calificación de 91 en este examen?
2) ¿qué porcentaje de estudiantes calificó entre 65 y 89?
3) ¿qué porcentaje de estudiantes calificó entre 81 y 89?
4) ¿cuál debe ser la calificación final del examen, si solo el 5% de los estudiantes examinados tuvieron la calificación más alta?

El profesor decide aprobar al 10% de los alumnos que tengan la mayor nota, sin que importe su calificación.

5) El alumno que haya obtenido una calificación de 81 en este examen, ¿aprueba?.

Con una calificación de 68 en un examen en que la media es 62 y la desviación es 3,

6) el alumno ¿aprueba?.

7) ¿Cuál de las dos situaciones es más conveniente para el alumno? Múestrelo estadísticamente y explíquelo.

5.2 El diámetro de las pelotas de ping-pong fabricada por una industria ubicada en la Provincia de Córdoba, tienen una distribución aproximadamente normal con una media de 1.30 pulgadas y una desviación estándar de 0.04 pulgadas.

Cuál es la probabilidad de que una pelota de ping pong seleccionada aleatoriamente tenga un diámetro de

1) entre 1.28 y 1.30 pulgadas?
2) entre qué dos valores simétricamente distribuidos alrededor de la media caerá el 60% de las pelotas de ping-pong (en términos del diámetro)?

Si se seleccionaron muchas muestras de tamaño 16

3) cuáles se esperaría que fueran la media y el desvío estándar de la media?
4) qué distribución seguirían las medias de la muestra?
5) qué proporción de las medias de muestras estarían entre 1.28 y 1.30 pulgadas?
6) entre qué valores se encontraría el 60% de las medias de las muestras?

Qué es más probable que ocurra

7) una pelota individual mayor a 1.34 pulgadas?
8) una media muestral por arriba de 1.32 pulgada en una muestra de tamaño 4?
9) una media muestral por arriba de 1.31 pulgadas en una muestra de tamaño 16?

Bibliografía

- **Berenson, M y Levine, D.** *Estadística Básica En Administración*. México: Prentice Hall, 1996.
- **Daniel, W.** *Bioestadística, Base Para El Análisis De Las Ciencias De La Salud.* México: Editorial Limussa, 1999.
- **Dixon, W.J. y Massey, F.J.** *Introduction to Statistical Analysis*. Nueva York: Mc Graw Hill, 1957.

- **Gomez Villegas, Miguel Angel.** *Inferencia Estadística*. España: Ediciones Díaz de Santo, 2005.
- **Hernández Sampieri, R.; Fernández Collado, C. y Baptista Lucio, P.** *Metodología De La Investigación*. México: McGraw Hill, 2010.
- **Kazmier, L y Diaz Mata, A.** *Estaística Aplicada a La Administración Y a La Economía*. México: McGraw Hill, 1993.
- **Kinnear, T.y Taylor, J.** *Investigación De Mercado. Un Enfoque Aplicado*. McGraw Hill, 1993.
- **Mao, J.C.T.** *Análisis Financiero*. Buenos Aires: El Ateneo, 1980.
- **Mendenhall, W. Wackerly, D. Scheaffer, R. .** *Estadística Matemática Con Aplicaciones*. México: Grupo Editorial Iberoamérica, 1990.
- **Meyer.** *Probabilidad Y Aplicaciones Estadísticas*. México: Fondo Educativo Interamericano SA, 1973.
- **Padua, J.** *Técnicas De Investigación Aplicadas a Las Ciencias Sociales*. México: Fondo de Cultura Económica, 1996.
- **Tramutola, C.D.** *Modelos Probabilísticos Y Decisiones Financieras*. Capital Federal: E.C.Moderan, 1971.