

METODOLOGIA ESTADISTICA

ANTONIO CAMPAÑA

Este documento ha sido elaborado sólo con fines pedagógicos y su objetivo es presentar algunos de los temas esenciales que se ofrecen en el Curso 403: Metodología estadística.

CELADE - SISTEMA DOCPAL
DOCUMENTACION
SOBRE POBLACION EN
AMERICA LATINA

El objetivo de este documento es presentar, en forma sucinta, los principales conceptos y técnicas estadísticas para la investigación en temas de población. Asimismo, se hará ^hincapié en establecer la relación entre la estadística y las ciencias sociales, por un lado, y entre las variables de la población y las variables económicas y sociales, por el otro.

ESTADISTICA DESCRIPTIVA

1. Escalas de Intervalo: medidas de tendencia central

El promedio es un valor típico de un conjunto de datos. Los promedios también reciben el nombre de medidas de tendencia central o centralización, puesto que sus valores tienen la tendencia a ubicarse en el centro de un conjunto de datos que, a la vez, han sido ordenados de acuerdo a su magnitud.

En las ciencias sociales destacan, principalmente, dos tipos de medidas de tendencia central que se usan en la investigación: la media aritmética y la mediana. También se incluyen dentro de estas tendencias el modo, la media geométrica y la media armónica.

LA MEDIA ARITMETICA

La media aritmética -o, simplemente, la media- de un conjunto de N números $X_1, X_2, X_3, \dots, X_N$, se indica con el símbolo \bar{X} y se define como:

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_N}{N} = \frac{\sum_{i=1}^N X_i}{N}$$

en la que X_1 representa el primer valor de la variable, X_2 el segundo valor de la variable y así sucesivamente, siendo X_n el último valor.

Propiedades de la media aritmética:

- 1.- La suma de las desviaciones de un conjunto de números respecto de la media será siempre cero.

Por ejemplo, la media de los números 72, 81, 86, 69 y 57 se obtiene al sumar todos los números y luego dividir por cinco: $\bar{X}=73$. Al restar la media de cada uno de las cifras y luego al sumar las cifras restantes, se verifica que la sumatoria es igual a cero. Supongamos, en cambio, que hubiésemos obtenido una media de 70. Al restar ésta de cada una de las cifras en cuestión nos daríamos cuenta que la media de 70 es errónea, puesto que la suma resultante no es igual a cero.

X	X-73	X-70
72	-1	2
81	8	11
86	13	16
69	-4	-1
57	-16	-13
	0	15

2.- La suma de las desviaciones cuadradas de cada cifra dentro de un conjunto de números con respecto a la media es menor que la suma de las desviaciones cuadradas con respecto a cualquier otro número.

Es decir: $\sum_{i=1}^N (X_i - \bar{X})^2 = \text{mínimo.}$

X	$(X-73)^2$	$(X-70)^2$
72	1	4
81	64	121
86	169	256
69	16	1
57	256	169
	506	551

Tomando los números usados en el ejemplo anterior, se obtiene que la suma de las desviaciones cuadradas respecto de la media es menor (506) que la de cualquier otro número (551).

Cálculo de la media aritmética de datos agrupados:

Cuando el número de datos es lo suficientemente grande, es más conveniente agrupar los datos en categorías o grupos y calcular la media a partir de la distribución de frecuencia resultante. Al efectuar los cálculos se procede a tomar en consideración ciertos supuestos que facilitan o simplifican la labor estadística. Por lo tanto, en el caso de la media se toman todos los casos como si estuvieran concentrados en los puntos medios de sus intervalos respectivos. Estas simplificaciones conducen a ciertos grados de inexactitud. Sin embargo, en la medida en que el número de datos sea más grande las distorsiones introducidas serán menores y menos insignificantes. El cálculo de la media de datos agrupados se puede hacer a partir de dos métodos: el método largo y el método corto.

La fórmula del método largo es la siguiente:

$$\bar{X} = \frac{\sum_{i=1}^k f_i m_i}{N}$$

donde f_i = número de casos de la categoría i

m_i = punto medio de la categoría i

k = número de las categorías

Ejemplo 1:

Límites fijados	Puntos medios (m_i)	f_i	$f_i m_i$
1950 - 2950	2450	17	41650
2950 - 3950	3450	26	89700
3950 - 4950	4450	38	169100
4950 - 5950	5450	51	277950
5950 - 6950	6450	36	232200
6950 - 7950	7450	21	156450
		189	967050

$$\bar{X} = \frac{967050}{189} = 5117$$

El método corto se calcula de la siguiente manera:

$$\bar{X} = \bar{X}' + \frac{\sum_{i=1}^k f_i d_i}{N}$$

donde \bar{X}' = media anticipada (corresponde al punto medio de uno de los intervalos)

f_i = número de casos de la categoría i

$d_i = X_i - \bar{X}'$

Escogemos 5450 como punto medio de un intervalo, puesto que la media debería ser un poco menor.

Ejemplo 2:

Límites fijados	Puntos medios (m_i)	f_i	d_i	$f_i d_i$
1950 - 2950	2450	17	-3000	-51000
2950 - 3950	3450	26	-2000	-52000
3950 - 4950	4450	38	-1000	-38000
4950 - 5950	5450	51	0	0
5950 - 6950	6450	36	1000	36000
7950 - 8950	7450	21	2000	42000
		189		-63000

$$\bar{X} = 5450 + \frac{(-63000)}{189} = 5450 - 333 = 5117$$

LA MEDIANA

La mediana de un conjunto de números ordenados en relación a su magnitud es el valor medio o la media aritmética de los dos valores medios.

Por ejemplo, para el conjunto de números 3, 4, 4, 5, 6, 8, 8, 8, 10 la mediana es 6. En cambio, para el conjunto de números 5, 5, 7, 9, 11, 12, 15, 18 la mediana se obtiene en la siguiente forma: $(9+11)/2=10$.

Cálculo de la mediana de datos agrupados:

El procedimiento para calcular la mediana de datos agrupados se obtiene mediante interpolación y se resume en la siguiente fórmula:

$$Md = L + \frac{N/2 - F}{f} \times i$$

donde Md = intervalo en que se encuentra la mediana

L = límite inferior del intervalo que contiene la mediana

N = numero de casos

F = frecuencia acumulativa correspondiente al límite inferior

f = número de casos del intervalo que contiene la mediana

i = amplitud del intervalo que contiene la mediana

Para obtener el cálculo de la media aritmética de datos agrupados se requiere, en primer lugar, localizar el intervalo que contiene el caso medio. En el presente ejemplo el número total de frecuencias es 189, por lo que el caso medio equivale a $189/2 = 94.5$. En segundo lugar se busca el intervalo que contenga el dato medio. Dado que hay 81 casos por debajo de \$4950 y 132 casos por debajo de \$5950, la mediana ha de quedar en algún lugar del intervalo que va de \$4950 a \$5950.

Límites fijados	f_i	F
1950 - 2950	17	17
2950 - 3950	26	43
3950 - 4950	38	81
4950 - 5950	51	132
5950 - 6950	36	168
7950 - 8950	21	189
	189	

$$Md = 4950 + \frac{94.5 - 81}{51} \times 1000$$

$$\begin{aligned}
 &= 4950 + 13.5 \frac{1000}{51} \\
 &= 4950 + 265 \\
 &= \$ 5215
 \end{aligned}$$

Asimismo, hay que tener presente que la mediana de datos agrupados se puede obtener restando cierta cantidad al límite superior U. Para este caso la fórmula es la siguiente:

$$Md = L + \frac{F - N/2}{f} \times i$$

en donde F representa la frecuencia acumulativa correspondiente al límite superior del intervalo. Por lo tanto,

$$\begin{aligned}
 Md &= 5950 + \frac{132 - 94.5}{51} \times 1000 \\
 &= \$5215
 \end{aligned}$$

EL MODO

El modo de un conjunto de números se define como aquel valor que ocurre más frecuentemente; es decir, el valor más común. Puede ocurrir que un conjunto de números no tenga modo o que éste no sea único.

Por ejemplo,

- (i) 21, 27, 63, 27, 65, 69
- (ii) 21, 27, 63, 15, 65, 69
- (iii) 21, 27, 63, 27, 63, 69

La primera serie de números tiene un modo de 27, en cambio la segunda serie no tiene modo. La tercera serie cuenta con dos modos: el 27 y 63.

En caso de referirnos a una distribución de frecuencias, el modo se representará por el punto más alto de la curva. En cambio, en una distribución simétrica -que cuenta con un sólo modo ubicado en el centro de la curva- la media, la mediana y el modo serán idénticos.

Aquellas series de números que sólo cuenten con un modo recibirán el nombre de distribuciones unimodales; en cambio cuando existan dos modos en una serie, las distribuciones se denominarán bimodales.

LA MEDIA GEOMETRICA

La media geométrica, G , de un conjunto de N números $X_1, X_2, X_3, \dots, X_N$ es la raíz N del producto de los números:

$$G = \sqrt[N]{X_1 X_2 X_3 X_4 \dots X_N}$$

Por ejemplo, la media geométrica de los números 2, 4, 8 es;

$$G = \sqrt[3]{(2)(4)(8)} = \sqrt[3]{64} = 4$$

LA MEDIA ARMONICA

La media armónica, H , de un conjunto de N números $X_1, X_2, X_3, \dots, X_N$ es el recíproco de la media aritmética del recíproco de los números.

$$H = \frac{N}{\sum_{i=1}^N \frac{1}{X}}$$

Por ejemplo, la media armónica de los números 2, 4, 8 es:

$$H = \frac{3}{\frac{1}{2} + \frac{1}{4} + \frac{1}{8}} = \frac{3}{\frac{7}{8}} = 3.43.$$

2. Escalas de intervalo: medidas de dispersión

La dispersión o variación de la información muestra el grado en que los datos numéricos tienden a esparcirse en relación a un valor medio. Existen varias medidas de dispersión, entre las cuales podemos mencionar el rango, la desviación media y la desviación estándar.

EL RANGO

El rango de un conjunto de datos se define como la diferencia entre el número mayor y el menor. En el caso que los números se hayan agrupado, se toma como recorrido la diferencia entre los puntos medios de las categorías extremas.

Por ejemplo:

(i) 21, 27, 63, 27, 65, 69

(ii) 500 9500

El primer conjunto de datos presenta un rango equivalente a 48. En cambio el rango correspondiente al segundo conjunto de datos equivale a 9000.

Como punto de referencia es conveniente precisar que el rango se basa única y exclusivamente en dos casos, los cuales, además, son casos extremos. Suele suceder en problemas empíricos que los casos extremos no sean representativos del conjunto total de datos, por lo que se crea una situación de extrema delicadeza, pues el rango no sería un fiel representante de una medida de dispersión.

LA DESVIACION MEDIA

La desviación media de un conjunto de datos se define como la media aritmética de las diferencias absolutas de cada valor de la variable con respecto a la media.

Es decir: Desviación Media =
$$\frac{\sum_{i=1}^N |X_i - \bar{X}|}{N}$$

Por ejemplo:

La media de los números 72, 81, 86, 69 y 57 es 73. Para obtener la desviación media sustraemos la media -73- de cada uno de los números, se ignoran los signos, se suman los resultados y se divide por el número de datos del conjunto.

$$\text{Desviación Media} = \frac{|72-73| + |81-73| + |86-73| + |69-73| + |57-73|}{5}$$

$$= \frac{1 + 8 + 13 + 4 + 16}{5} = \frac{42}{5} = 8.4$$

Podemos, por consiguiente, decir que en promedio los datos difieren de la media en 8.4 unidades.

LA DESVIACION ESTANDAR

La desviación estándar es la más útil y frecuente medida de dispersión. Se define como la raíz cuadrada de la media aritmética de las desviaciones cuadradas con respecto a la media.

$$\text{Es decir: } s = \sqrt{\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}} = \sqrt{\frac{x^2}{N}}$$

$$\text{donde: } (X_i - \bar{X})^2 = x^2$$

s = desviación estándar

Por ejemplo:

X_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
72	-1	1
81	8	64
86	13	169
69	-4	16
57	-16	256
$X=73$	0	506

$$s = \sqrt{506/5} = \sqrt{101.2} = 10.06$$

Una formula de cálculo alternativa de la desviación estándar que tiene el beneficio de cometer menos errores de redondeo, por lo cual se recomienda, es la siguiente:

$$s = \frac{1}{N} \sqrt{N \sum_{i=1}^N X_i^2 - (\sum_{i=1}^N X_i)^2}$$

Por ejemplo:

X_i	X_i^2
72	5184
81	6561
86	7396
69	4761
57	3249
365	27151

$$\begin{aligned}
 s &= 1/5 \sqrt{5(27151) - (365)^2} \\
 &= 1/5 \sqrt{135755 - 133225} \\
 &= 10.06
 \end{aligned}$$

Otra medida de uso frecuente en las ciencias sociales es la varianza, que se define como el cuadrado de la desviación estándar. Es decir:

$$\text{Varianza} = s^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}$$

En algunos casos es necesario distinguir entre la desviación estándar de una población $-\sigma-$ y la desviación estándar de una muestra de la población $-s-$. Por lo tanto, la varianza de una muestra y de una población presentan la misma terminología; s^2 y σ^2 , respectivamente. Asimismo, es importante mencionar que en algunos casos el denominador de la fórmula de la desviación estándar de una muestra se define como $(N-1)$ en vez de N . Esto se debe, principalmente, a que el valor obtenido representa mejor la estimación de la desviación estándar de la población de la cual se obtuvo la muestra. Para valores grandes de N ($N > 30$) prácticamente no existe diferencia entre ambas definiciones. En el caso que dispongamos de la población total, se usa la siguiente fórmula:

$$\sigma^2 = (X - \mu)^2 / N.$$

Desviación estándar de una población: σ

Desviación estándar de una muestra: s

Cálculo de la desviación estándar de datos agrupados:

Por ejemplo:

Límites verdaderos	Puntos medios	f_i	d'_i	$f_i d'_i$	$f_i d'^2_i$
1950 - 2950	2450	17	-3	-51	153
2950 - 3950	3450	26	-2	-52	104
3950 - 4950	4450	38	-1	-38	38
4950 - 5950	5450	51	0	0	0
5950 - 6950	6450	36	1	36	36
6950 - 7950	7450	21	2	42	84
		189		-63	415

$$s = \frac{i}{N} \sqrt{N \sum_{i=1}^k f_i d'^2_i - (\sum_{i=1}^k f_i d'_i)^2}$$

$$s = \frac{1000}{189} \sqrt{189(415) - (-63)^2}$$

$$s = 5.291 \sqrt{78435 - 3969}$$

$$s = 5.291 \sqrt{74466}$$

$$s = 5.291 (272.885)$$

$$s = 1443.84 = 1444$$

3.- La Distribución Normal

La noción de distribución de frecuencias pasa a ser parte fundamental de la estadística descriptiva, dado que, en primer lugar, la curva normal se emplea generalmente para interpretar la desviación estándar y, en segundo lugar, por su significado teórico en la comprensión del mismo.

Areas bajo la curva normal:

La curva normal posee la propiedad de que, independiente de la media o de la desviación estándar que una curva presente, habrá un área constante (o proporción de casos) entre la media y una ordenada, que es una distancia determinada a partir de la media en términos de unidades estándar.

Al colocarse en una desviación estándar a la derecha (o izquierda) de la media, encontraremos siempre .3413 del área incluida entre la media y la ordenada en dicho punto. Por lo tanto, dos veces dicha área, ó .6826, estará incluida entre las dos ordenadas situadas a una desviación estándar a ambos lados de la media. Asimismo, el área comprendida entre la media y la ordenada a dos desviaciones estándar de aquella será siempre .4773 y, por tanto, el área entre las dos ordenadas a dos desviaciones estándar a ambos lados de la media será .9546. Para fines prácticos se puede decir que todos los casos estarán dentro del área que va desde la media hasta tres desviaciones estándar a ambos lados, aunque la curva normal se extienda teóricamente al infinito en ambas direcciones. Es importante tener presente que aunque la curva normal proporciona una interpretación de la desviación estándar, esta propiedad no puede emplearse para definir lo que se entiende por desviación estándar.

Tal vez el aspecto de mayor importancia de la distribución normal es que resulta posible tomar cualquier curva normal y transformar sus valores numéricos de tal forma que pueda utilizarse un simple cuadro para evaluar la proporción de casos al interior de cualquier intervalo deseado. Es decir:

$$Z = \frac{X - \bar{X}}{s}$$

donde Z: representa la desviación con respecto a la media en unidades de desviación estándar; y

X: es el valor de la ordenada.

Por ejemplo:

Supongamos que tenemos una curva normal con una media de 50 y una desviación estándar de 10. Se desea obtener la proporción de casos que se encuentran en el intervalo 50 a 65. Para ello se requiere precisar a cuántas desviaciones estándar se halla 65 de la media 50.

$$Z = \frac{65 - 50}{10} = 1.5$$

Este procedimiento estipula que en tanto la distribución de la variable X es normal con una media de \bar{X} y una desviación estándar de s, la nueva variable, en cambio, es normal con una media de cero y una desviación estándar de uno.

ESTADISTICA INDUCTIVA

1.- Introducción a la estadística inductiva

Uno de los aspectos que reviste gran importancia en la comprensión de la estadística inductiva es aquel que distingue entre las características propias de una población y de una muestra, ésta última obtenida de dicha población o universo. De esta manera, las características de la población se designarán como parámetros; en cambio, las características de la muestra, como estadísticos. Así, en adelante se designará la media de la población con μ y la de la muestra con \bar{X} ; la desviación estándar de la primera con σ y la de la muestra con s .

Es importante tener presente que el objetivo primordial que se persigue es obtener información acerca de la población y no de una muestra cualquiera. La muestra se debe comprender como una herramienta de conveniencia sin importancia en sí misma. Las conclusiones que se obtengan -utilizando muestras escogidas- deben estar basadas en una serie de parámetros de la población. Como lo ha expuesto Blalock en su libro Estadística Social: "En las verificaciones de hipótesis formulamos supuestos a propósito de los parámetros desconocidos, y preguntamos a continuación cómo serían nuestras estadísticas específicas si dichos supuestos fueran correctos. Al proceder así, tratamos de decidir racionalmente si los valores supuestos de dichos parámetros son o no razonables a la vista de la evidencia de que disponemos".

Característica de la población:	parámetros
Característica de la muestra:	estadísticos
Media de la población:	μ
Media de la muestra:	\bar{X}
Desviación estándar de la población:	σ
Desviación estándar de la muestra:	s

2.- Pruebas de muestras simples

El teorema del límite central:

Si de una población normal con una media de μ y una varianza de σ^2 se extraen reiteradas muestras al azar, la distribución de selección de las medias de las muestras será normal, con la media μ y la varianza σ^2/N .

En otras palabras, se obtienen varias muestras con sus respectivas medias \bar{X} . Cada una de estas medias de las muestras variará con respecto al resto, pero en general se agruparán alrededor de la verdadera media μ de la población. El teorema, entonces, dice que un gráfico de la distribución de estas muestras será una curva normal.

Al referirnos a las pruebas estadísticas, es más bien la distribución de las muestras y no la población original la que se utiliza directamente en las pruebas de significación. En resumen, las medias y las desviaciones estándar de las tres clases de distribución son como sigue:

	Media	Desviación estándar
Población	μ	σ
Muestra	\bar{X}	s
Distribución de las muestras	μ	σ / \sqrt{N}

El teorema del límite central pone de manifiesto que, suponiendo que se hayan evitado distorsiones, puede tenerse más confianza en la apreciación de la media de una muestra grande que de una pequeña.

La ley de los grandes números:

Si se extraen al azar diversas muestras de magnitud N de una población cualquiera (de la forma que sea) con una media de μ y una varianza de σ^2 entonces, a medida que N crece, la distribución de las muestras (que corresponden a las medias de las muestras) se aproxima a la normalidad, con la media μ y la varianza σ^2/N .

En otras palabras, por muy notable que sea la distribución de la que partimos, a condición que N sea lo bastante grande, podemos contar con una distribución de la muestra aproximadamente normal. Para comprender integralmente el teorema del límite central y para convencerse que el error estándar es realmente σ/\sqrt{N} , se extrae un número de muestras de una población cuya media y desviación estándar son conocidas, luego se procede a calcular las medias de las muestras y, finalmente, se compara el resultado obtenido con σ/\sqrt{N} .

3.- Estimación de intervalo

El procedimiento efectivo empleado para obtener una estimación de intervalo o, lo que comúnmente se designa como intervalo de confianza, es el siguiente:

Primero se decide acerca del riesgo de error que se está dispuesto a asumir al afirmar que el parámetro se sitúa en algún punto al interior del intervalo si en realidad no es así. En el caso de intervalos de confianza nos referimos a la unidad menos la probabilidad de error. Esto significa que tenemos confianza de estar en lo cierto, por ejemplo, el 95 por ciento de las veces. El intervalo se obtiene apartándose en ambas direcciones de la estimación del punto cierto múltiplo de errores estándar correspondiente al nivel de confianza elegido. Así, por ejemplo, para apreciar la media μ de la población obtenemos un intervalo como sigue (tomando un nivel de confianza del 95 por ciento):

$$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{N}} = \bar{X} \pm 1.96 \frac{\sigma}{\sqrt{N}}$$

en donde 1.96 corresponde a la región crítica de la curva normal, usando el nivel de confianza equivalente a 95% y una prueba de dos colas. Si $\bar{X}=15$, $\sigma=5$ y $N=100$, el intervalo de confianza sería:

$$15 \pm 1.96 \frac{5}{\sqrt{100}} = 15 \pm 0.98$$

en otros términos el intervalo iría de 14.02 a 15.98. Por lo tanto, sabemos que sólo un 5% de las veces obtendremos con este procedimiento intervalos que no comprendan el parámetro. El 95% restante de las veces el procedimiento nos dará medias de una muestra lo suficientemente cercanas al parámetro para que los intervalos de confianza obtenidos comprendan efectivamente a éste. Por último, cabe recordar que el parámetro es un valor fijo y que son los intervalos los que varían de una muestra a otra. Si se escoge un nivel de confianza mayor, existe más certeza que el intervalo contiene μ ; pero, por otro lado, necesitamos un intervalo mayor para tener un nivel mayor de confianza.

Tamaño de la muestra:

Para tener un intervalo corto y que al mismo tiempo tenga un nivel alto de confianza se tendrá que aumentar el tamaño de la muestra. Tomando un ejemplo sencillo para facilitar la comprensión, tenemos que se desea calcular la longitud de un intervalo de confianza de 95%. Dado que el intervalo es $\bar{X} \pm 1.96 \sigma/\sqrt{n}$, la longitud total del intervalo es $2(1.96) \sigma/\sqrt{n}$. Si el investigador desea que la longitud sea igual a 60, la ecuación se puede resolver de la siguiente manera:

$$60 = \frac{2(1.96)120}{\sqrt{n}}$$

$$\sqrt{n} = \frac{2(1.96)120}{60} = 7.84$$

$$n = 61.47 = 62.$$

Es importante tener presente que este procedimiento se puede llevar a cabo debido a que el investigador conoce el valor de la varianza de la muestra que decidió utilizar.

La distribución t:

En la mayor parte de las investigaciones el valor de la varianza se desconoce, lo que significa que debe ser estimada con base en los datos. En general el investigador desconoce el valor de σ , por lo tanto obtiene una estimación de ésta mediante el cálculo de s . Acto seguido se forma la cantidad $t = (\bar{X} - \mu) / (s / \sqrt{n-1})$, la que será utilizada en vez de $z = (\bar{X} - \mu) / (\sigma / \sqrt{n})$. El valor o la cantidad que adquiere t no tiene una distribución normal. La distribución de t es diferente para distintos valores de n , el tamaño de la muestra. Las áreas bajo las curvas de distribución para la cantidad t han sido obtenidas y puestas en forma de una tabla. La primera columna de la tabla en cuestión da a conocer un número que se denomina grados de libertad. Este es el número que se usó en el denominador al calcular s^2 ó $n-1$ en este caso.

Intervalo de confianza para la media usando la distribución t:

Cuando la desviación estándar se estima de la muestra, un intervalo de confianza para μ , la media de la población, se forma de la misma manera que cuando σ se conoce, con la excepción de que s reemplaza a σ y las tablas de distribución t reemplazan a las tablas normales. Al conocer σ , el intervalo de confianza equivalente a 95% para μ es $\bar{X} \pm z_{.95}\sigma/\sqrt{n}$, donde $z_{.95}$ señala el valor donde se ubica el 95% de las zetas ($z_{.95}=1.96$). Usando s , que ha sido calculada de la muestra da $\bar{X} \pm t_{.95}s/\sqrt{n-1}$, donde $t_{.95}$ señala el valor donde el 95% de las t se ubican en la distribución t con $n-1$ grados de libertad.

Por ejemplo:

$X=311.9$ gramos

$s^2=20,392$

$s=142.8$

$n=17$

Por lo tanto, el intervalo de confianza equivalente a 95% es:

$$311.9 \pm t_{.95} \frac{142.8}{\sqrt{17-1}}, \quad \text{d.f.} = 16$$

$$311.9 \pm 2.120 \frac{142.8}{\sqrt{16}}$$

$$311.9 \pm 2.120(35.7)$$

$$311.9 \pm 75.7$$

$$236.2 \text{ a } 387.6 \text{ gramos}$$

Este intervalo de confianza debe interpretarse de la siguiente manera: el investigador tiene un 95% de confianza de que μ se encuentra entre 236.2 y 387.6 gramos, puesto que si el experimento se repitiera -con una muestra de tamaño 17- usando siempre la fórmula $\bar{X} \pm t_{.95}s/\sqrt{n-1}$ para formar un intervalo de confianza, el 95% de los intervalos formados incluirían μ .

De la misma forma que se efectuó con las distribuciones normales, el investigador puede definir intervalos de confianza equivalentes a 95% ó 99% mediante el uso de la tabla de la distribución t. De esta manera podemos estimar la media de la población μ dentro de límites específicos de confianza.

Por ejemplo, si $-t_{.95}$ y $t_{.95}$ son los valores de t para los cuales 2.5% del área se ubica en cada cola de la distribución t (5% para ambos), entonces un intervalo de confianza equivalente a 95% para t es:

$$-t_{.95} < \frac{\bar{X} - \mu}{s} \sqrt{n-1} < t_{.95}$$

por lo que se puede estimar que μ se ubicara en el siguiente intervalo:

$$\bar{X} - t_{.95} \frac{s}{\sqrt{n-1}} < \mu < \bar{X} + t_{.95} \frac{s}{\sqrt{n-1}}$$

con una confianza equivalente a 95%. Es importante tener presente que $t_{.95}$ representa el valor equivalente al 95 percentil.

En general, podemos representar límites de confianza para medias de la siguiente manera:

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

donde los valores t_c , definidos como valores críticos o coeficientes de confianza, dependen del nivel de confianza deseado y del tamaño de la muestra.

El número de grados de libertad de una estadística se define como el número N de observaciones independientes en la muestra menos el número k de parámetros de la población que se deben estimar basándose en las observaciones de la muestra.

Es decir: $V = N - K$

BIBLIOGRAFIA

Blalock Jr., Hubert M., Estadística social, Fondo de Cultura Económica, México, D.F., 1966.

Dixon, Wilfred J. Y Massey Frank J., Introducción al análisis estadístico, 2a edición, Ediciones Castilla, Madrid, 1966.

Spiegel, Murray R., Teoría y problemas de estadística, McGraw Hill, Bogotá, Colombia, 1977.