TEMAS DE MATEMÁTICAS
(Oposiciones de Secundaria)

TEMA 58

POBLACION Y MUESTRA. CONDICIONES DE REPRESENTATIVIDAD DE UNA MUESTRA. TIPOS DE MUESTREO. TAMAÑO DE UNA MUESTRA.

TEMA 58

POBLACION Y MUESTRA. CONDICIONES DE REPRESENTATIVIDAD DE UNA MUESTRA. TIPOS DE MUESTREO. TAMAÑO DE UNA MUESTRA.

1. INTRODUCCIÓN.

El objetivo de la Estadística es hacer una inferencia con respecto a la población basándose en la información contenida en una muestra. Como las poblaciones se describen mediante medias numéricas denominadas parámetros, el objetivo de la mayoría de las investigaciones estadísticas es hacer una inferencia con respecto a uno o más parámetros de la población. La generalidad de los procedimientos de la inferencia estadística involucran ya sea la estimación o bien la prueba de hipótesis

2. TIPO DE MUESTREO.

Llamaremos muestra a la parte de la población que utilizamos para conocer a toda la población, aunque sea de un modo aproximado.

Las muestras deben cumplir las siguientes condiciones:

a) Ser Representativa.

Esta condición está asociada al tamaño de la muestra n, ya que cuanto más grande sea n es evidente que más información proporcionará, y por lo tanto, más representativa. A su vez, el tamaño de la muestra también depende de la dispersión ya que, si la población esta muy dispersa, tendremos que coger una muestra de gran tamaño, para no perder mucha información.

b) Ser Aleatoria.

Todos los análisis estadísticos se basan en que la muestra sea aleatoria, es decir, que todos los elementos de la población tienen la misma probabilidad de formar parte de la muestra. En caso contrario, corremos el peligro de coger una subpoblación, que es un subconjunto de la población que cumple una determinada condición, con lo cual perdemos el principio de representatividad.

En el caso de que se pierda la aleatoriedad y, por lo tanto la muestra no sea del todo representativa, se dice que se han cometido errores de Sesgo.

Otro fallo a la hora de elegir una muestra es que una variable condicione a otra, ya que las variables deben ser independientes y no condicionadas.

Ejemplo. Un ejemplo de una muestra representativa y aleatoria es coger cinco números con dos decimales utilizando la función de Randomize, donde la población es de N=99 y la muestra es n=5.

0’61, 0’39, 0’30, 0’44, 0’12, 0’57

Para aumentar la representatividad sin necesidad de aumentar el tamaño de la muestra se recurre al muestreo o técnicas de muestreo. En la práctica resuelven el problema de la representatividad. Hay varios tipos de muestreo:

a) Muestreo Aleatorio Simple.

Se realiza en poblaciones en las que los datos son homogéneos. Es decir, no existen factores que produzcan variabilidad sistemática. en este tipo de muestreo los elementos de la población homogénea se eligen al azar.

b) Muestreo Aleatorio Estratificado.

Si en la población existe variabilidad, este muestre consiste en descomponer la población en partes que se llaman estratos, de manera que dentro de cada estrato los elementos sean homogéneos, siendo diferentes los elementos de estratos distintos. Posteriormente se realiza un muestreo aleatorio simple en cada estrato, obteniéndose así la muestra.

Criterios de Estratificación.

Se deben coger como estratos aquellos factores que producen variabilidad de los datos. Por ejemplo, en las alturas de los españoles, los criterios de estratificación son:

- Edad: E_1, E_2, E_3
- Sexo; H, M
- Zona: Rural, Urbana

	E_1		E_2		E_3	
	Rural	Urbana	Rural	Urbana	Rural	Urbana
H						
M						

Los cuadros representan a los estratos. Si llamamos N_i al tamaño del estrato y n_i al tamaño de la muestra del estrato, la proporción que existe entre el tamaño de la muestra del estrato y el del estrato coincide con la proporción que existe entre el tamaño total de la muestra y el de la población.

$$\frac{n_i}{N_i} = \frac{n}{N}$$

c) Muestreo por conglomerados

Se aplica cuando la población presenta heterogeneidad y se actúa de la siguiente manera:

Paso 1: Se descompone la población en clases llamadas conglomerados, de forma que dentro de cada conglomerado haya la máxima dispersión o heterogeneidad (es decir, que haya de todo), de tal forma que los conglomerados se parezcan entre sí.

Paso 2: Para elegir la muestra se realiza un muestreo aleatorio de conglomerados. Cuando se elige un conglomerado, todos los elementos del mismo forman parte de la muestra.

d) Muestreo Sistemático.

Se realiza cuando los elementos se encuentran en una lista. Una vez que se elige un número, el resto ya está condicionado. Para introducir la aleatoriedad, se dice por donde se empiezan a coger los elementos.

Ejemplo. Si en una lista hay 100 números y deseamos coger 25 de ellos, elegiremos uno al azar entre los cuatro primeros y, a partir de ese momento, tomamos los elementos cada cuatro.

e) Muestreos Polietápicos.

Los muestreos estratificados se realizan de la siguiente manera: se forman estratos y después se hace un muestreo (de los tipos anteriores), luego se hace otro, etc.

Además, los tipos de muestreo pueden ser con reemplazamiento o sin reemplazamiento.

- Muestreo sin reemplazamiento. Una vez tomado un elemento no se devuelve a la población.
- Muestreo con reemplazamiento. El elemento elegido sí se devuelve a la población, pudiendo ser seleccionado de nuevo.

3. ESTIMACIÓN.

Iniciaremos el estudio estadístico de colectivos (población) mediante la elección de unos pocos (muestra), de los que inferiremos las características de toda la población. La Estadística Inferencial tratará de obtener información acerca de los parámetros poblacionales a partir de la muestra.

Los Estimadores son variables aleatorias utilizadas `para estimar parámetros de la población. Los estimadores que proporcionan un único valor para el parámetro poblacional se denominan Estimadores Puntuales, mientras que los que especifican un intervalo de valores se denominan Estimadores por Intervalos.

Para llevar a cabo tales estimaciones es necesario que la muestra sea representativa de la población, para lo cual ha de ser aleatoria en su selección y poseer un adecuado tamaño.

3.1. Propiedades de un Buen Estimador.

Para que un estadístico sea considerado un buen estimador de un parámetro dado, conviene que reúna las siguientes propiedades:

- Ser Insesgado (centrado)
- Ser Consistente.
- Ser Eficiente.
- Ser Suficiente.

Sea un estadístico (’ del que nos vamos a servir para estimar un parámetro (.

3.1.1. Estimadores Centrados.

Diremos que θ’ es un Estimador Centrado o Insesgado de θ si se verifica que $E(\theta')=\theta$. Por el contrario, diremos que el estimador es Sesgado si $E(\theta')=\theta+b(\theta)$, y se denomina Sesgo del estimador a la cantidad $b(\theta)=E(\theta') - \theta$.

3.1.2. Consistencia.

Un estadístico θ’ utilizado para estimar un parámetro θ es consistente si para n tendiendo a infinito, se verifica que $\theta' \rightarrow \theta$ en probabilidad, para lo cual es suficiente que se cumplan las dos condiciones siguientes:

i) Que sea Asintóticamente Centrado: $E(\theta') \rightarrow \theta$
ii) Que la Varianza tienda a Cero: $Var(\theta') \rightarrow 0$

3.1.3. Eficiencia.

Si para estimar un mismo parámetro θ tenemos dos estimadores asintóticamente centrados θ’ y θ’’, decimos que θ’ es más eficiente que θ’’ si la varianza del primero es menor que la del segundo. Esto es, si $Var(\theta') \leq Var(\theta'')$.

La Eficiencia Relativa de θ’’ respecto de θ’ se define como el cociente entre ambas varianzas:

$$efr(\theta'|\theta'') = \frac{Var(\theta')}{Var(\theta'')}$$

3.1.4. Suficiencia.

Un estimador θ’ del parámetro θ es Suficiente si contiene tanta información como la contenida en la propia muestra. Dicho de otra forma: Diremos que un estadístico $T=T(X_1,X_2,...,X_n)$ es Suficiente para θ si la distribución de X_1, X_2, ..., X_n dado T es independiente del valor del parámetro.

3.2. Métodos de Estimación Puntual.

La Estimación Puntual constituye el método más elemental de asignar los valores obtenidos de la muestra (estadísticos) a toda la población (parámetros). En los métodos de estimación puntual se busca un estimador, con base en los datos muestrales, que proporcione un único valor del valor del parámetro. Estimar un parámetro θ no es más que dar una función de las observaciones que no dependa, por tanto, del parámetro desconocido:

$$\theta' = \theta'(X_1, X_2, ..., X_n)$$

así pues, para cada valor de la muestra asigna un valor al parámetro θ. A esta función se la denomina Estimador y a sus valores Estimaciones del Parámetro.

3.2.1. Estimación por Máxima Verosimilitud.

Este método selecciona como estimación aquel valor del parámetro que tiene la propiedad de maximizar el valor de la probabilidad de la muestra aleatoria observada. En otras palabras, el método de máxima verosimilitud consiste en encontrar el valor del parámetro que maximiza el valor de la función de verosimilitud.

Si denotamos a la función de verosimilitud por L, es decir, la función de probabilidad de la muestra (caso discreto) o la de densidad de probabilidad (caso continuo), en ambos casos la función depende del parámetro desconocido (o parámetros) θ, con lo que tenemos:

$$L(\boldsymbol{q}; X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} f(X_i; \boldsymbol{q})$$

para una muestra aleatoria simple X_1, X_2, ..., X_n de una distribución con función de probabilidad o densidad de probabilidad $f(x;\theta)$. La función $L(\theta; x_1, x_2, ..., x_n)$, considerada como función del parámetro θ, recibe el nombre de Función de Verosimilitud de la Muestra. Si $t=g(x_1, x_2, ..., x_n)$ es el valor de θ para el cual el valor de la función de verosimilitud es máxima, entonces $T=g(X_1, X_2, ..., X_n)$ es el estimador de máxima verosimilitud de θ. Así pues, el estimador máximo verosímil debe satisfacer la ecuación:

$$L(\theta'; X_1, X_2, ..., X_n) = \max_{\theta \in \Theta}{}^{o} L(\theta; X_1, X_2, ..., X_n)$$

donde Θ es el Espacio Paramétrico (conjunto de posibles valores del parámetro θ). El método de máxima verosimilitud tiene la propiedad de proporcionar estimadores que son funciones de estadísticos suficientes si y sólo si el estimador de máxima verosimilitud es único. Debido a la naturaleza de la función de verosimilitud resulta a menudo mucho más fácil de obtener el estimador máximo verosímil del logaritmo neperiano de dicha función, $Ln[L(\theta; X_1, X_2, ..., X_n)]$, que de la propia función $L(\theta; X_1, X_2, ..., X_n)$.

3.2.2. Método de los Momentos.

Este método es quizás el más antiguo para la estimación de parámetros. Consiste en igualar un determinado número de momentos teóricos de la distribución de población con los correspondientes momentos muestrales para obtener una o varias ecuaciones que, una vez resueltas, permiten estimar los parámetros desconocidos de la distribución poblacional.

3.2.3. Método de Mínimos Cuadrados.

El método de Mínimos Cuadrados, introducido por Legendre y Gauss, consiste en buscar los valores de los parámetros que minimizan una cierta función cuadrática de los mismos (la suma de los cuadrados de los errores). Estos métodos son interesantes debido a sus propiedades asintóticas, pues para muestras grandes:

Suelen dar estimadores asintóticamente centrados, es decir, con un sesgo despreciable.

Por lo general, son asintóticamente normales, es decir, que su distribución de probabilidad es aproximadamente normal.

3.3. Estimación Por Intervalos.

Aún el estimador centrado más eficiente es improbable que estime con exactitud el valor del parámetro de la población. De aquí nace la necesidad de obtener un intervalo dentro del cual se espera hallar el valor del parámetro, lo que nos lleva a la Estimación por Intervalos. Una estimación por intervalos de un parámetro θ de la población es un intervalo de la forma

$$\theta_L < \theta < \theta_U$$

donde θ_L y θ_U dependen de las observaciones muestrales. Puesto que muestras diferentes generalmente proporcionarán valores diferentes de θ_L y θ_U, estos puntos extremos del intervalo son valores aleatorios y se buscan de modo que fijado γ entre 0 y 1 se verifique que $P(\theta_L < \theta < \theta_U)=\gamma$

El Intervalo $\theta_L < \theta < \theta_U$ obtenido a partir de la muestra seleccionada recibe el nombre de Intervalo de Confianza. El valor γ se denomina Coeficiente de Confianza. Los valores θ_L y θ_U son respectivamente Límite de confianza Inferior y Límite de Confianza Superior.

4. ESTADÍSTICOS.

Definimos como Estadístico cualquier función que dependa de los valores de la muestra.

Distintos tipos de Estadísticos:

1) Media Muestral. $$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Se diferencia de la media poblacional μ en que $\overline{X}$ es una función y μ es un parámetro.

2) Varianza Muestral. $$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^2}{n}$$

Se diferencia de la varianza poblacional, ya que σ^2 es un número.

3) Cuasivarianza Muestral. $$S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^2}{n-1}$$

4) Proporción Muestral. Se define para valores cualitativos donde x=nº de casos con cualidad y n=tamaño de la muestra. Se define por: $p = \frac{x}{n}$

5) Proporción Poblacional. Se define por Π

La media muestral puede tomar un conjunto de valores que dependen de las variables X_1, X_2, ..., X_n. Los valores que puede tomar un Estadístico t_1, t_2, ..., t_n forman la población del estadístico. La distribución de frecuencias o de probabilidad se llamará Distribución del Estadístico.

Ejemplo. Dada la población {1,3,5,7} con μ=4 y σ^2=5, calcular la distribución del estadístico de $\overline{X}$ de tamaño 2:

$$\overline{X} = \frac{X_1 + X_2}{2}$$

Valores que puede tomar el Estadístico: $$\begin{cases} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \end{cases}$$

La media de las medias muestrales es:

$$m_{\overline{X}} = \frac{1 + 2\cdot 2 + 3\cdot 3 + 4\cdot 4 + 3\cdot 5 + 2\cdot 6 + 7}{16} = 4$$

que coincide con la media de la población.

La varianza de las medias muestrales $s_{\overline{X}}^2 = \frac{5}{2}$, que coincide con la varianza de la población dividida por n. Entonces:

$$s_{\overline{X}}^2 = \frac{s^2}{n}$$

Por lo tanto, la distribución del Estadístico de $\overline{X}$ se ajusta a una distribución normal de media μ y varianza $\frac{s^2}{n}$.Así pues:

$$\overline{X} \to N\left(m, \frac{s^2}{n}\right)$$

TEOREMA. Teorema Central del Límite.

Cuando la población es Normal, la distribución del estadístico es Normal, con media μ y varianza $\frac{s^2}{n}$.

5. ERRORES DE MUESTREO.

Calculemos la varianza de la media muestral para establecer la precisión de las estimaciones:

$$s^2(\overline{x}) = E\left(\frac{\sum_{i=1}^{n} x_i}{n} - E(\overline{x})\right)^2 = E\left(\frac{\sum_{i=1}^{n} x_i}{n} - \overline{X}\right)^2 =$$

$$= \frac{1}{n^2} E\left(\sum_{i=1}^{n}(x_i - \overline{X})\right)^2 = \frac{1}{n^2} E\left(\sum_{i=1}^{n}(x_i - \overline{X}) + 2\sum_{i\neq j}^{n}(x_i - \overline{X})(x_j - \overline{X})\right)$$

Ahora bien:

$$E\left(\sum_{i=1}^{n}(x_i - \overline{X})^2\right) = \sum_{i=1}^{n} E(x_i - \overline{X})^2 = n \cdot V = n \cdot \frac{N-1}{N} S^2$$

$$E\left(\sum_{i\neq j}^{n}(x_i - \overline{X})(x_j - \overline{X})\right) = \sum_{i\neq j}^{n} E\left((x_i - \overline{X})(x_j - \overline{X})\right) = \sum_{i\neq j}^{n} \frac{\sum_{i\neq j}^{N} E\left((X_i - \overline{X})(X_j - \overline{X})\right)}{\binom{n}{2}}$$

Por otra parte:

$$0 = \left(\sum_{i=1}^{N}(X_i - \overline{X})\right)^2 = \sum_{i=1}^{N}(X_i - \overline{X})^2 + 2\sum_{i\neq j}^{N}(X_i - \overline{X})(X_j - \overline{X})$$

Luego:

$$\sum_{i\neq j}^{n} E\left(x_i-\overline{X}\right)\left(x_j-\overline{X}\right)=\sum_{i\neq j}^{n}\left(-\frac{2}{N(N-1)}\sum_{i=2}^{N}\left(X_i-\overline{X}\right)^2\right)=\sum_{i\neq j}^{n}-\frac{1}{N-1}\cdot V$$

Sustituyendo tenemos:

$$s^2(\overline{x})=\frac{1}{n^2}\left(n\cdot\frac{N-1}{N}S^2-\binom{n}{2}\cdot 2\cdot\frac{1}{N-1}\cdot V\right)=\frac{1}{n}\left(\frac{N-1}{N}\cdot S^2-\frac{n-1}{N-1}\cdot\frac{N-1}{N}S^2\right)=$$

$$=\frac{1}{n}S^2\left(\frac{N-n}{N}\right)=\frac{N-n}{Nn}S^2=\frac{N-n}{Nn}\cdot\frac{V\cdot N}{N-1}=\frac{N-n}{N-1}\cdot\frac{V}{n}$$

Así pues:

$$s^2(\overline{x})=\begin{cases}\dfrac{N-n}{Nn}S^2 & \text{en función de la cuasivarianza poblacional } S\\ \dfrac{N-n}{N-1}\cdot\dfrac{V}{n} & \text{en función de la varianza poblacional } V\end{cases}$$

PROP La cuasivarianza muestral es un estimador insesgado de la varianza poblacional.

Dem.

Probaremos que $E(s^2)=S^2$

$$s^2=\frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}{n-1}\qquad E\left(s^2\right)=\frac{1}{n-1}E\left(\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2\right)$$

Ahora bien:

$$\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2=\sum_{i=1}^{n}\left[\left(x_i-\overline{X}\right)-\left(\overline{x}-\overline{X}\right)\right]^2=$$

$$=\sum_{i=1}^{n}\left[\left(x_i-\overline{X}\right)^2-2\left(x_i-\overline{X}\right)\cdot\left(\overline{x}-\overline{X}\right)+\left(\overline{x}-\overline{X}\right)^2\right]=$$

$$=\sum_{i=1}^{n}\left[\left(x_i-\overline{X}\right)^2-2\left(\overline{x}-\overline{X}\right)\sum_{i=1}^{n}\left(x_i-\overline{X}\right)+\left(\overline{x}-\overline{X}\right)^2\right]=$$

$$=\sum_{i=1}^{n}\left(x_i-\overline{X}\right)^2-2\left(\overline{x}-\overline{X}\right)\left(n\overline{x}-n\overline{X}\right)+n\left(\overline{x}-\overline{X}\right)^2=$$

$$=\sum_{i=1}^{n}\left(\left(x_i-\overline{X}\right)^2-n\left(\overline{x}-\overline{X}\right)^2\right)$$

Entonces:

$$E(s^2)=\frac{1}{n-1}\sum_{i=1}^{n}\left[E\left(x_i-\overline{X}\right)^2-nE\left(\overline{x}-\overline{X}\right)^2\right]=$$

$$\frac{1}{n-1}\left[n\cdot V-n\cdot\frac{N-n}{N-1}\cdot\frac{V}{n}\right]=\frac{1}{n-1}\cdot V\cdot\frac{n-1}{N-1}\cdot N=S^2$$

A la raíz cuadrada de la varianza de la media muestral se le llama error de muestreo, y su valor es:

$$\sigma_{\overline{X}}=\sqrt{\frac{N-n}{N-1}}\cdot\sqrt{\frac{V}{n}} \qquad \text{o también} \qquad \sigma_{\overline{X}}=\sqrt{\frac{N-n}{N}}\cdot\frac{S}{\sqrt{n}}$$

Llamaremos Factor de Corrección para Poblaciones Finitas a $\sqrt{\frac{N-n}{N}}$.

Si la población es infinita ($N\to\infty$) se obtiene que el error de muestreo es el cociente de la derivación típica poblacional y la raíz cuadrada del tamaño muestral n.

En el caso de una proporción, al tratarse de una media, resulta que la proporción muestral es un estimador insesgado de la proporción poblacional y

$$\sigma_r^2(P)=\frac{N-n}{N}\cdot\frac{S^2}{n}$$

pero en este caso S^2 se puede simplificar más

$$S^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(x_i-\overline{X}\right)^2=\frac{N}{N-1}\cdot\sum_{i=1}^{N}\frac{\left(x_i-\overline{X}\right)^2}{N}=\frac{N}{N-1}\left(\frac{\sum_{i=1}^{N}x_i^2}{N}-\overline{X}^2\right)=\frac{N}{N-1}(P-P^2)=\frac{N}{N-1}PQ$$

(puesto que $\sum_{i=1}^{N}x_i^2=\sum_{i=1}^{N}x_i$ al tomar x_i los valores 0 y 1, y Q=1 – P) con lo que el error de muestreo de la proporción es

$$\sigma_p=\sqrt{\frac{N-n}{N}\cdot\frac{N}{N-1}\cdot\frac{PQ}{n}}=\sqrt{\frac{N-n}{N-1}\cdot\frac{PQ}{n}}$$

6. ESTIMACIÓN PUNTUAL. TAMAÑO DE UNA MUESTRA.

Consiste en asignar valores a un parámetro de la población a partir de los valores de la muestra, donde la media tiene un valor determinado que es estimado, no calculado.

Para ello se usan los estimadores, que son los estadísticos que dan valores aproximados del parámetro que se quiere estimar.

Estimador	Parámetro
$\overline{X}$	μ
S^2	σ^2
P	Π

La media μ de la población es aproximadamente la que obtenemos como media de la muestra, y lo mismo ocurre con la varianza. Pero para saber como es de "aproximado" debe conocerse el grado de aproximación que viene dado por la llamada Cota de Error e.

$$e = K \cdot \frac{s}{\sqrt{n}}$$

La cota de error la utilizamos para estimar la media de la población mediante el estimador media muestral.

Debemos saber que la cota de error depende de la dispersión del estadístico.

Tenemos que el estadístico es:

$$N \to N\left(m, \frac{s^2}{n}\right) \qquad con \quad (X_1, X_2, \ldots, X_n)$$

Se pueden tomar los valores de X, pues todos tienen 0 o se ajustan a la misma distribución normal. En una muestra n-dimensional $(X_1, X_2, \ldots, X_n)$ se tiene que:

$$\frac{X_i}{n} \to N\left(\frac{m}{n}, \frac{s^2}{n^2}\right)$$

donde

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \sum_{i=1}^{n} \frac{X_i}{n}$$

por lo tanto podemos ver que la suma de variables normales es también normal y son idénticamente distribuidas e independientes. Se tiene que:

$$N\left(n \cdot \frac{m}{n}, n \cdot \frac{s^2}{n^2}\right) \to N\left(m, \frac{s^2}{n}\right)$$

Para saber cuanto vale la media de una población se obtiene una muestra y se toma la media de la misma como estimador.

El problema está en que la cota de error no es segura, ya que todavía no sabemos el valor que debe tomar K. Para ello debemos plantearnos que nivel de confianza queremos en nuestra estimación.

6.1. Nivel de Confianza.

Se mide en términos probabilísticos, donde K se determina en función del nivel de confianza:

$$P\left(\left|\overline{X}-\mu\right|<e\right)=1-\alpha \qquad (1)$$

Para que esto suceda con una probabilidad elevada, se necesita conocer el nivel de confianza, es decir, $1-\alpha$ (donde para que el nivel de confianza sea alto debe ocurrir que α sea muy pequeño).

El nivel de confianza se define como "la probabilidad de que la diferencia entre el estimador y el parámetro que se quiere estimar sea menor que la cota de error", o sea (1).

El nivel de confianza suele establecerse entre 0'95 y 0'99, es decir

$$P\left(\left|\overline{X}-\mu\right|<K\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

Tipificando se tiene que:

$$P\left(\frac{\left|\overline{X}-\mu\right|}{\frac{\sigma}{\sqrt{n}}}<K\right)=1-\alpha \quad \Rightarrow \quad P\left(\left|Z\right|<K\right)=1-\alpha \quad \text{donde } Z\equiv N(0,1).$$

Por lo tanto tenemos que $K=Z_{\alpha/2}$

Y de aquí deducimos que el error es $e=Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}$

Debemos saber que la cota de error depende de:

- De la dispersión de la población.
- Del tamaño de la muestra.
- Del Nivel de Confianza.

Ejemplo. Notas de Bioquímica {8, 8'33, 8, 8'1, 7, 6, 8'9, 7'2, 7'8, 8'5}

$$\begin{matrix}\sigma=2\\ \overline{X}=7'78\end{matrix} \qquad 1-\alpha=0'95 \quad \Rightarrow \quad \textit{Podemos tomar} \quad \begin{matrix}Z_{\alpha/2}=1'96\\ Z_{1-\frac{\alpha}{2}}=0'975\end{matrix}$$

$$\Rightarrow \quad e=Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}=1'96\cdot\frac{2}{\sqrt{10}}=1'24$$

y obtenemos que $\mu=7'78\pm1'24$

6.2. Cálculo del Tamaño de una Muestra que Corresponde a un Error.

Si definimos "e" como el error máximo admisible, y partimos de la fórmula del error $e = Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ se tiene que:

$$e = Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow \sqrt{n}e = Z_{\alpha/2} \cdot \sigma \Rightarrow \sqrt{n} = \frac{Z_{\alpha/2} \cdot \sigma}{e} \Rightarrow n = \frac{Z_{\alpha/2}^2 \cdot \sigma^2}{e^2}$$

Esta fórmula la aplicaremos porque conocemos $Z_{\alpha/2}$ y porque "e" lo damos nosotros.

Sin embargo, esta fórmula sólo nos sirve para muestreos sobre poblaciones infinitas o finitas con reemplazamiento.

Por lo tanto, para poblaciones finitas utilizaremos esta otra fórmula:

$$e_N = Z_{\alpha/2} \cdot \sqrt{\frac{N-n}{N-1}} \cdot \frac{\sigma}{\sqrt{n}}$$

Esta fórmula del error para poblaciones finitas aparece a partir de la de poblaciones infinitas, pero añadiéndole un término llamado Factor de Corrección, que es $\sqrt{\frac{N-n}{N-1}}$, siendo N el tamaño de la población y n es el tamaño de la muestra. Además, podemos decir que este factor de corrección disminuye el error, ya que $\sqrt{\frac{N-n}{N-1}} < 1$, por razones obvias.

En el caso, ahora, de que queramos calcular el tamaño de una muestra para poblaciones finitas, utilizaremos la fórmula:

$$n_N = \frac{n_\infty}{1 + \frac{n_\infty - 1}{N}}$$

Por lo tanto, recapitulando tenemos que:

	Error	Tamaño de la Muestra
Población Infinita	$e = Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$	$n = \frac{Z_{\alpha/2}^2 \cdot \sigma^2}{e^2}$
Población Finita	$e_N = Z_{\alpha/2} \cdot \sqrt{\frac{N-n}{N-1}} \cdot \frac{\sigma}{\sqrt{n}}$	$n_N = \frac{n_\infty}{1 + \frac{n_\infty - 1}{N}}$

6.3. Estimación de la Varianza.

También se puede estimar la varianza mediante la cuasivarianza

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2$$

pero como $Z_{a/2}$ procede de una distribución normal y la cuasivarianza no, tenemos que establecer una nueva fórmula del error, que es:

$$e_\infty = t_{a,n-1}\cdot\frac{S}{\sqrt{n}}$$

Como vemos, aparece un término nuevo que es la distribución continua t-Student, que depende del nivel de confianza α que queramos para la cota del error y de los grados de libertad n-1, donde n es el tamaño de la muestra.

Análogamente a lo hecho antes, podemos establecer una fórmula para la cota del error pero en poblaciones finitas, que dependen de la cuasivarianza:

$$E_N = t_{a,n-1}\cdot\frac{S}{\sqrt{n}}\cdot\sqrt{\frac{N-n}{N-1}}$$

OBS Vamos a establecer una relación entre la varianza y la cuasivarianza, de manera que podamos conocer la cota de error, conocida una o conocida la otra.

$$\left.\begin{array}{ll} S^2 = \dfrac{\sum_{i=1}^{n}\left(x_i-\overline{X}\right)^2}{n-1} & \Rightarrow \quad \sum_{i=1}^{n}\left(x_i-\overline{X}\right)^2 = S^2\cdot(n-1) \\ s^2 = \dfrac{\sum_{i=1}^{n}\left(x_i-\overline{X}\right)^2}{n} & \Rightarrow \quad \sum_{i=1}^{n}\left(x_i-\overline{X}\right)^2 = S^2\cdot n \end{array}\right\} \Rightarrow \quad \frac{S^2}{n} = \frac{s^2}{n-1} \quad \Rightarrow \quad \frac{S}{\sqrt{n}} = \frac{s}{\sqrt{n-1}}$$

o lo que es lo mismo $\sqrt{n-1}\cdot S = \sqrt{n}\cdot s$

Por lo tanto se tiene que:

	Error
Población Infinita	$e_\infty = t_{a,n-1}\cdot\frac{S}{\sqrt{n}}$ ó $e_\infty = t_{a,n-1}\cdot\frac{S}{\sqrt{n-1}}$
Población Finita	$E_N = t_{a,n-1}\cdot\frac{S}{\sqrt{n}}\cdot\sqrt{\frac{N-n}{N-1}}$ ó $E_N = t_{a,n-1}\cdot\frac{S}{\sqrt{n-1}}\cdot\sqrt{\frac{N-n}{N-1}}$

Ejemplo. Se quiere medir la efectividad de un somnífero. Para ello se aplica a 17 personas que duermen una media de 8’2 h. Con una desviación de 0’25 h. Estimar la efectividad del somnífero con una confianza del 95%.

Resolución.

Como la conclusión que queremos sacar es general, supondremos una población infinita.

$$\left.\begin{array}{r} n=17 \\ \overline{X}=8'2 \\ s=0'25 \\ 1-\alpha=0'95 \end{array}\right\}$$

La efectividad del somnífero es el estimador $\overline{X}$. Tenemos que calcular el error, ya que $\overline{X}$ =8’2 horas.

$$e_{\infty}=t_{a,n-1}\cdot\frac{S}{\sqrt{n-1}}$$

como α=0’05 y n-1=16 entonces $t_{0'05,16}$=2’12

Entonces $e_{\infty}=2'12\cdot\frac{0'25}{\sqrt{16}}=0'13$ horas.

Luego $\mu=\overline{X}\pm e_{\infty}$=8’2±0’13 horas $\Rightarrow$ $\mu\in(8'07,8'33)$

6.4. Tamaño de la Muestra de la Varianza.

Existe ahora un problema, porque para conocer el tamaño de la muestra necesitamos más datos, ya que hasta ahora hemos partido de la muestra ya extraída.

Para ver el tamaño de la muestra, hemos de tener en cuenta la **Paradoja de Friedman** que dice que cuando no conocemos la varianza y queremos saber el tamaño de la muestra, debemos usar una de estas tres opciones:

1) La varianza de un estudio similar.
2) La varianza de una muestra piloto.
3) Una muestra lo más grande posible.

Ejemplo. Queremos estimar la efectividad de un fármaco con un error de 1 minuto (máximo). ¿Cuántos pacientes tendrían que recibir el fármaco?

$$\left.\begin{array}{r} e=\frac{1}{60} \\ 1-\alpha=0'95 \\ \alpha=0'05 \\ Z_{\alpha/2}=1'96=t_{a,\infty} \end{array}\right\} \Rightarrow \quad n_{\infty}=\frac{Z_{\alpha/2}^2\cdot s^2}{e^2}$$

La expresión anterior no podemos resolverla ya que nos falta la varianza y al no tener una muestra similar, podemos calcularla a partir de la muestra del ejemplo anterior, aunque la varianza de la muestra piloto no es la varianza que buscamos, pero se aproxima.

$$n_{\infty} = \frac{1'96^2 \cdot (0'25)^2}{\left(\frac{1}{60}\right)^2} = 864'36 \approx 865$$

Para no perder precisión aproximamos por arriba.

Cuando queremos reducir el error a la mitad necesitamos aumentar el tamaño de la muestra cuatro veces.

$$e' = Z_{\alpha/2} \cdot \frac{s}{\sqrt{4n}} = Z_{\alpha/2} \cdot \frac{s}{2\sqrt{n}} = \frac{1}{2} Z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} = \frac{1}{2} e$$

Para poblaciones pequeñas, si ésta se duplica, el tamaño de la muestra también se duplica. Si ésta vuelve a duplicarse, da un resultado mayor, pero llega un momento en el que por mucho que aumente la población, su tamaño no influye en el de la muestra.

$$\lim_{N \to \infty} \frac{n_{\infty}}{1 + \frac{n_{\infty} - 1}{N}} = \frac{n_{\infty}}{1 + 0} = n_{\infty}$$

6.5. Estimación de la Proporción.

Si definimos Π como la proporción poblacional y definimos p como la proporción muestral, la cual se ajusta a una distribución $N\left(\Pi, \frac{p(1-p)}{n}\right)$ donde n es el tamaño muestral, se tiene que las cotas de error para poblaciones finitas e infinitas son:

$e_{\infty} = Z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}$ Cota de error para poblaciones infinitas.

Y añadiéndole el factor de corrección: $E_N = Z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$

Cuando queramos estimar el Intervalo de Confianza de la Proporción Poblacional, éste nos vendrá dado por la proporción muestral y por la cota de error. Es decir:

$$\Pi \approx p^* \pm e^*$$

donde p^*=p*100 y e^*=e*100, que se dan en tanto por ciento.

6.6. Tamaño de una Muestra de la Proporción.

Para una población infinita se tiene que, despejando n de la fórmula de la cota de error:

$$n_\infty = \frac{Z^2_{a/2} \cdot p \cdot (1-p)}{e^2}$$

Pero esta fórmula no nos sirve, pues no tiene sentido que si queremos hallar el tamaño de la muestra, estemos utilizando la proporción muestral. Por tanto, para poder solucionar este problema, sustituimos en la fórmula la proporción muestral por la proporción poblacional.

$$n_\infty = \frac{Z^2_{a/2} \cdot \Pi \cdot (1-\Pi)}{e^2}$$

El problema se plantea cuando tampoco conocemos Π. Entonces la pregunta sería ¿Cómo estudiar Π sin conocerlo? La solución a este problema consiste en elegir una de las siguientes opciones:

1) Tomar Π de un estudio similar.
2) Tomar Π de una muestra piloto.
3) Darle a Π el valor que haga máximo a n. Es decir, nos planteamos el caso más desfavorable de modo que así no podamos errar, y dicho valor es Π=0’5.

El tamaño de una muestra para poblaciones finitas se obtiene mediante la siguiente fórmula:

$$n_N = \frac{n_\infty}{1 + \frac{n_\infty - 1}{N}}$$

OBS. Podríamos establecer la relación $e = \frac{\Pi}{h}$ donde h es el índice de proporción de error. Es decir, si h=2 es porque se quiere un error que sea la mitad de la proporción poblacional. Entonces, utilizando esta relación se tiene que:

$$n_\infty = \frac{Z^2_{a/2} \cdot \Pi \cdot (1-\Pi)}{\frac{\Pi^2}{h^2}} \quad \Rightarrow \quad n_\infty = Z^2_{a/2} \cdot h^2 \cdot \left(\frac{1}{\Pi} - 1 \right)$$

Ejemplo. ¿Qué tamaño de muestra se necesitaría para estimar el porcentaje de fumadores de la Región de Murcia, sabiendo que en un estudio anterior se conoció que fumaban el 30%, y que podemos permitirnos un error máximo del 5%?

Sabemos que:

$$1 - \alpha = 0'95 \quad \Rightarrow \quad Z_{a/2} = 1'96$$
$$\text{e} = 5\% \quad \Rightarrow \quad \text{e} = 0'05$$
$$\Pi \approx 30\% \quad \Rightarrow \quad \Pi \approx 0'3$$

$$n_\infty = \frac{Z^2_{\alpha/2} \cdot \Pi \cdot (1-\Pi)}{e^2} = \frac{(1'96)^2 \cdot 0'3 \cdot 0'7}{0'05^2} = 322'7 \approx 323$$

Ejemplo. Si quisiéramos hacer un estudio para saber a cuántos de los 14 que estamos en este aula habría que entrevistar para calcular el porcentaje de fumadores bajo las condiciones del ejemplo anterior, se tiene que:

$$n_N = \frac{n_\infty}{1+\frac{n_\infty - 1}{N}} = \frac{323}{1+\frac{323-1}{14}} = 13'458 \approx 14$$

BIBLIOGRAFÍA RECOMENDADA.

Introducción a la Teoría de la Estadística. Aut.: Mood/Graybill. Ed. Aguilar.

Introducción a la Probabilidad y la Medida. Aut. Procopio Zoroa. Ed. PPU