AULA ESTADÍSTICA

INFERENCIA ESTADÍSTICA
ESTIMACIÓN DE PARÁMETROS

JORGE R. LORENZO

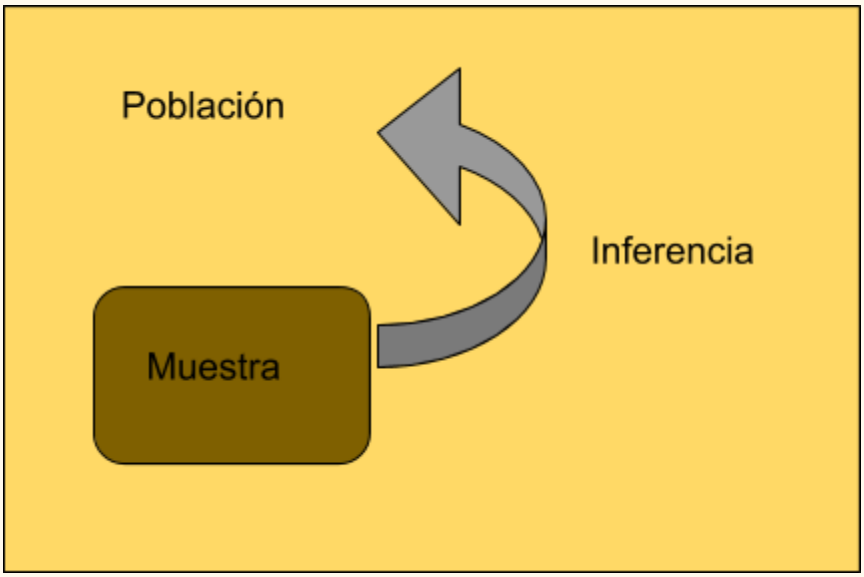
INTRODUCCIÓN

Los modelos estadísticos son ampliamente utilizados en las las ciencias experimentales en la medida en que facilitan obtener conocimientos generales de lo particular (razonamiento inductivo). Cualquier generalización sobre una población a partir de una muestra de la misma, solo está garantizada en la medida en que tal muestra es representativa. La inferencia estadística es aquella rama de la estadística mediante la cual se trata de sacar conclusiones de una población, a partir de la información que proporciona una muestra representativa de la misma. También es denominada estadística inductiva.

La necesidad de utilizar muestras de la población se comprende fácilmente si tenemos en cuenta los costes económicos de la experimentación o el hecho de que muchos de los métodos de medida impiden que las unidades de muestreo se repongan a la población. Por ejemplo, si estamos trabajando sobre una muestra de estudiantes y aplicamos una prueba de logro académico, no podremos recolectar una nueva muestra con esos estudiantes dado que ya conocen las preguntas, y por tanto, les será más sencillo responderlas. Diremos en este caso que estaríamos introduciendo un sesgo.

En toda inferencia inductiva existe un término de error, pero también es posible medirlo si el experimento se ha realizado adecuadamente. Las técnicas estadísticas para hacer inferencias inductivas permiten medir el grado de incertidumbre de tales inferencias. La incertidumbre se expresa en términos de probabilidad. Si miramos el esquema de la página que sigue, es posible ver el proceso completo: de toda población bien definida es posible extraer una muestra. Si la muestra es representativa de la población, es posible calcular diversos valores de las variables de interés en la muestra y estimar los parámetros poblacionales. Este proceso es el que recibe el nombre de inferencia estadística.

Esquema de la inferencia estadística



Definimos como población al conjunto de individuos sobre los que se desea información. La población debe estar perfectamente definida a la hora de comenzar el estudio. Por ejemplo, si estamos realizando un ensayo clínico de un medicamento se deben especificar puntualmente la población objetivo para la que se estima que la droga produce un alivio. Dicho en otras palabras se debe especificar los criterios de inclusión de un paciente en la población para conformar la muestra. Otro tanto ocurre cuando la población de interés es muy grande, por ejemplo jóvenes entre 18 y 25 años. Sin criterios que la delimiten, esta población puede abarcar a cualquier jóven alrededor del mundo. En tal caso será mejor decir, jóvenes entre 18 y 25 años que viven en la Ciudad Autónoma de Buenos Aires. Como se dijo, la muestra debe ser representativa de la población, en el sentido de que debe tener una composición similar en cuanto a la proporción de las distintas características poblacionales. La representatividad de la muestra queda garantizada con la elección correcta del método de

muestreo. Desde ya destacamos que no todas las muestras obtenidas de una población resultan representativas de la misma.

Si se tiene una muestra representativa de una población, es factible medir en cada una de las unidades de muestreo las variables de interés. Una población puede caracterizarse en función de una o varias variables. De este modo, a cada elemento de la población le corresponde una variable aleatoria cuya notación estadística por convención es *x*. Los subíndices de *x* indicarán la variable de que se trate. Por ejemplo, si hemos tomado registro de tres variables diferentes, estas se designarán como. x_1; x_2; x_3. De esa manera quedan identificadas población y variable aleatoria asociada. Así en teoría de la inferencia población será el conjunto de individuos que nos interesa estudiar, y variable aleatoria asociada será la característica que medimos en los individuos.

Para estudiar la distribución de la variable en la población supondremos un modelo de distribución de probabilidad aleatorio que resuma las características de la misma; aunque desconocemos los parámetros, intentaremos estimarlos a partir de una muestra. Por ejemplo, la variable aleatoria *x* en la población se distribuye como: $N(\mu; \sigma)$ donde los dos parámetros son media poblacional y varianza. La estimación consistirá en capturar uno o ambos mediante la inferencia estadística. En algunos casos no es necesario especificar tales distribuciones y las inferencias se hacen sobre características de la distribución que no son necesariamente parámetros.

La inferencia Estadística puede dividirse en dos tipos de acuerdo con el conocimiento sobre la distribución de la variable en la población.

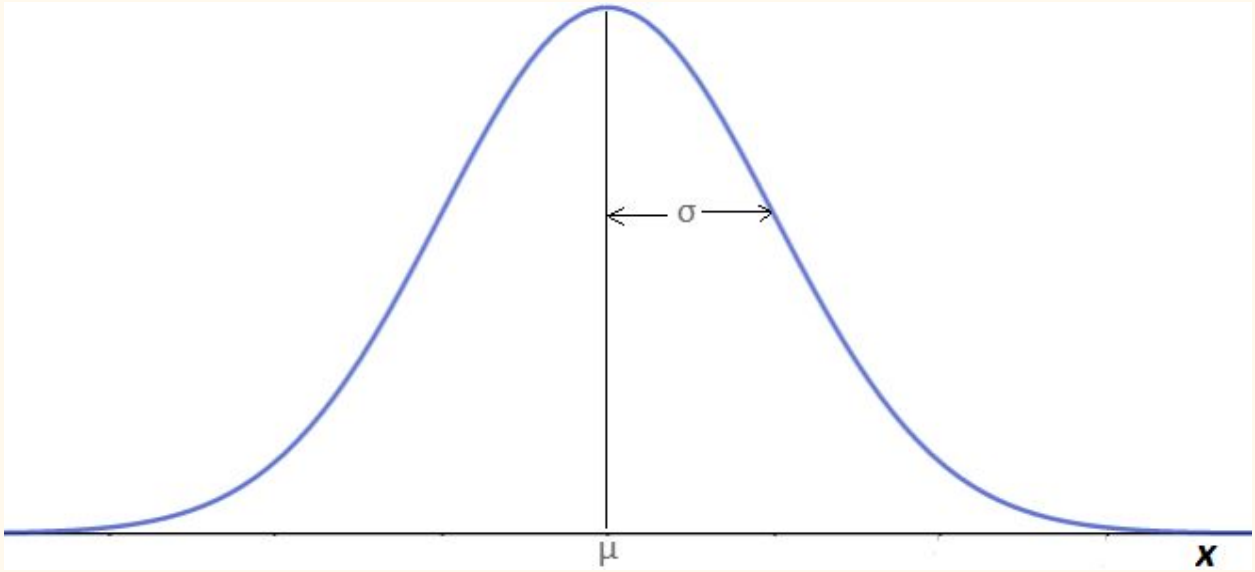
Inferencia Paramétrica: la forma de la distribución es conocida (Normal, Binomial, Poisson, otra) pero se desconocen sus parámetros. Se realizan inferencias sobre los parámetros desconocidos utilizando la distribución de probabilidad que la modela.

Inferencia No Paramétrica: forma y parámetros desconocidos. Se realizan inferencias sobre características que se ajustan a parámetros (estadísticos de Orden, posición, porcentuales, otros).

De acuerdo con la forma en que se estudian los parámetros o características desconocidas, la inferencia puede dividirse en dos apartados:

Estimación: se estiman los parámetros desconocidos sin hipótesis previas sobre posibles valores de los mismos. Así tenemos:
Estimación puntual: Un único valor para cada parámetro.
Estimación por intervalos: Intervalo de valores probables para el parámetro.

Contraste de Hipótesis: se plantean hipótesis sobre los parámetros desconocidos y se desarrolla un procedimiento para comprobar la verosimilitud de la hipótesis planteada. El primer paso de cualquier investigación necesita de una definición clara de la población en estudio, por ejemplo: estudiantes que se encuentren cursando sexto año de escuela secundaria en la Ciudad Autónoma de Buenos Aires. Nuestra variable de interés será el rendimiento académico, que se expresará mediante puntuaciones estandarizadas de los operativos nacionales de evaluación. Por tratarse de una variable aleatoria, se asume que su distribución en la población sigue el modelo normal.



Distribución poblacional del rendimiento académico en estudiantes de sexto año CABA.

El modelo predice que la distribución aleatoria estará centrada en el parámetro μ (media paramétrica) y su dispersión será σ (sigma). Por definición el parámetro resulta desconocido y tendremos que estimarlo a partir de una muestra. En los apartados siguientes, veremos de qué manera podemos obtener una muestra representativa de la población de interés para realizar la estimación deseada.

MUESTREO

Recordemos que estamos interesados en obtener una muestra que sea representativa de la población que hemos delimitado. Por lo tanto, conviene repasar algunas propiedades de las técnicas de muestreo. Los pasos a seguir para la recolección de una muestra son los siguientes:

1) Definir la población en estudio especificando las unidades que la componen, el área geográfica donde se realiza el estudio y el periodo de tiempo en el que se realizará el mismo.

2) Definir el marco de muestreo que consta del listado o descripción de los elementos que forman la población.

3) Definir la unidad de muestreo, que para nuestro ejemplo se trata de personas, pero pueden ser otras unidades distintas.

4) Definir las variables a medir, para el ejemplo planteado se trata del rendimiento académico según resultados de una prueba estandarizada. Las variables pueden cambiar tanto como el tipo de población que se defina.

5) Seleccionar el método de muestreo: para poder estimar un parámetro necesitamos un tipo de muestreo probabilístico. Existen otros tipo de muestreo a menudo denominados no-probabilísticos, pero éstos últimos nos permiten la estimación correcta de parámetros.

6) Calcular el tamaño muestral necesario para una precisión determinada en la estimación.

En cuanto al tipo de muestreo, algunas de las características más importantes de los muestreos probabilísticos se describen a continuación, aunque no se detallan en profundidad. Sobre este punto sugerimos remitirse a bibliografía especializada.

Muestreo Aleatorio Simple

Se trata de un procedimiento de muestreo (sin reemplazo), en el que se seleccionan *n* unidades de las en la población (N), de forma que cualquier posible muestra del mismo tamaño tiene la misma probabilidad de ser elegidas. Se realizan selecciones independientes de las unidades de muestreo de forma que en cada selección todos los individuos tengan la misma probabilidad de ser elegidos. El procedimiento habitual consiste en numerar todos los elementos de la población y seleccionar muestras del tamaño deseado utilizando una secuencia de números aleatorios. Entre las ventajas de este procedimiento esta

la compensación de valores altos y bajos con lo que la muestra tiene una composición similar a la de la población, es además un procedimiento sencillo y produce estimadores de los parámetros desconocidos próximos a los valores reales de los mismos. El principal inconveniente de este tipo de muestreo es que necesita un marco adecuado y amplio que no siempre es fácil de conseguir y que no contiene información a priori sobre la población que podría ser útil en la descripción de la misma.

Muestreo aleatorio sistemático

Es una variante del muestreo aleatorio simple, pero en este caso se ordenan los individuos de la población y se enumeran (se debe cuidar que en tal ordenamiento no se introduzcan sesgos que pudieran viciar la muestra). Se calcula una fracción de muestreo dividiendo la población en tantos grupos como individuos se quieren tener en la muestra: N/*n*. Se selecciona valor de arranque al azar en el primer grupo y se elige el que ocupa el mismo lugar en todos los grupos. La ventaja principal es que es más sencillo y más barato que el muestreo aleatorio simple, además, se comporta igual si no hay patrones o periodicidades en los datos. Se debe cuidar que no existan patrones desconocidos ya que puede llevar a importantes errores en la estimación de los parámetros. Este tipo de muestreo puede utilizarse toda vez que la población se encuentre listada en una base de datos. Existen software específico para realizar este tipo de muestreos cuando los datos se encuentran en bases estándares.

Muestreo por conglomerados

En este tipo de muestreo se divide la población en grupos de acuerdo con su proximidad geográfica (conglomerados). Cada grupo ha de ser heterogéneo y tener representados todos las características de la población. Un ejemplo de conglomerados en zonas rurales pueden ser los municipios de la zona. En zonas urbanas los conglomerados se pueden delimitar a partir de los lindes barriales.

Se selecciona una muestra de conglomerados al azar y se toma el conglomerado completo o una muestra del mismo. Necesitan menos información previa sobre los individuos particulares y soluciona el problema de los patrones en los datos.

Si el número de bloques no es muy grande se puede incurrir en errores de estimación si se han incluido conglomerados atípicos. Los conglomerados que se realizan teniendo en cuenta proximidad geográfica pueden no tener un significado importante en la población pues no responden a una característica real. Este tipo de muestreo se utiliza fundamentalmente para reducir los costes de muestreo al tomar grupos de individuos completos, o bien cuando la población no se encuentra listada o en una base de datos completa.

Muestreo estratificado

Si es posible identificar y definir una característica que permita parcelar la población, ésta puede utilizarse para estratificarla. Cada estrato divide la población en grupos homogéneos de acuerdo con las características a estudiar. En nuestro ejemplo, el tipo de gestión de la escuela (Publica, Privada u Otra), presentar características diferenciales que pueden operar para delimitar el tamaño del estrato. Además, cada unidad de muestreo se cuenta solo en un tipo de escuela. Una vez definido el estrato, se selecciona una muestra aleatoria de cada uno, tratando de que todos los estratos de la población queden representados. Esta técnica permite utilizar información a priori sobre la estructura de la población en relación con las variables a estudiar. Dependiendo del tamaño de cada estrato, se procederá a ponderarlos para obtener una muestra proporcional de cada uno de ellos.

ESTADÍSTICOS Y DISTRIBUCIONES MUESTRALES

El tema que sigue se basa en el supuesto de cualquier población es infinita. En nuestro ejemplo este supuesto debe entenderse en el sentido de que la población es lo suficientemente grandes como para poder abarcala en su totalidad. Otro supuesto subyacente es que el procedimiento de muestreo responde a lo que hemos mostrado como aleatorio simple, es decir se garantiza una muestra representativa de la población y la obtención de observaciones independientes.

Dada una población de interés, el proceso de muestreo consiste en obtener, al azar, un valor de la variable X. Sea x_1 el valor obtenido aleatoriamente, puede considerarse como una realización particular de la variable aleatoria X bajo la misma distribución. El siguiente valor de X, sea x_2 debe obtenerse independientemente de la primera observación, que también será considerado como una realización particular de una variable aleatoria X con la misma distribución e independiente de x_1. Obsérvese que la población no se modifica al extraer uno de sus individuos ya que es infinita. En caso de una población finita podría utilizarse un muestreo con reemplazo.

El proceso continúa hasta obtener una muestra de tamaño *n*, de observaciones x_1, x_2, ... , x_n de la variable aleatoria X, independientes e idénticamente distribuidas. Las letras minúsculas refieren a las observaciones particulares de una muestra, mientras que X mayúscula denota la variable aleatoria de las que se han tomado las observaciones.

Otra forma de ver la muestra es como una variable aleatoria multivariante con función de densidad de probabilidad es el producto de las funciones de densidad de cada una de las componentes, que son independientes, donde las funciones de densidad son iguales a:

$$f(X_1, X_2, \dots, X_n) = f(X_1)\, f(X_2) \dots f(X_n)$$

Una vez obtenida la muestra la describimos en términos de algunas de sus características fundamentales como la media o la desviación estándar, a la que denominaremos estadísticos. Entonces, un estadístico es una función de los valores muestrales que no depende de ningún parámetro poblacional desconocido. Un estadístico es una variable aleatoria ya que es una función de variables aleatorias. Por ejemplo la media muestral es una variable aleatoria de la que tenemos una sola observación.

Supongamos que disponemos de una población finita compuesta por números naturales que toman los siguientes valores: {1, 2, 3, 4}. Al obtener diferentes muestras sin reemplazamiento de tamaño *n*=2, las distintas posibilidades se realizan de la siguiente manera:

{1, 2} {1, 3} {1, 4} {2, 3} {2, 4} {3, 4}

Siguiendo el orden en que se obtuvieron las muestras, los promedios obtenidos para cada una de ellas son:

1.5 2 2.5 2.5 3 3.5

El promedio se obtuvo aplicando la siguiente ecuación:

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

Vemos que la media muestral no es un valor fijo sino que puede considerarse también como una variable aleatoria de la que tenemos una sola observación: la media de la muestra concreta seleccionada. Dicha variable tendrá una distribución de probabilidad asociada. En este caso una distribución discreta que toma los valores 1.5 (p=1/6) 2, (p=1/6), 2.5 (p=2/6), 3 (p=1/6) y 3.5 (p=1/6).

A la distribución de un estadístico calculado a partir de los valores tomados de una muestra se la denomina distribución muestral del estadístico. En la mayor parte de los casos supondremos que nuestra población tiene distribución normal y que los estadísticos que vamos a utilizar son la media y la desviación estándar.

DISTRIBUCIONES MUESTRALES DE LA MEDIA Y LA DESVIACIÓN ESTÁNDAR

Sea $X_1, X_2, \ldots, X_n$, una muestra aleatoria de una población X en la que:
E(X)= μ y Var (X)= σ^2
Entonces el valor esperado de la media y la varianza son:

$$E(\bar{x}) = \mu$$

$$var(\bar{x}) = \frac{\sigma^2}{n}$$

$$Desv\ Est\ (\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

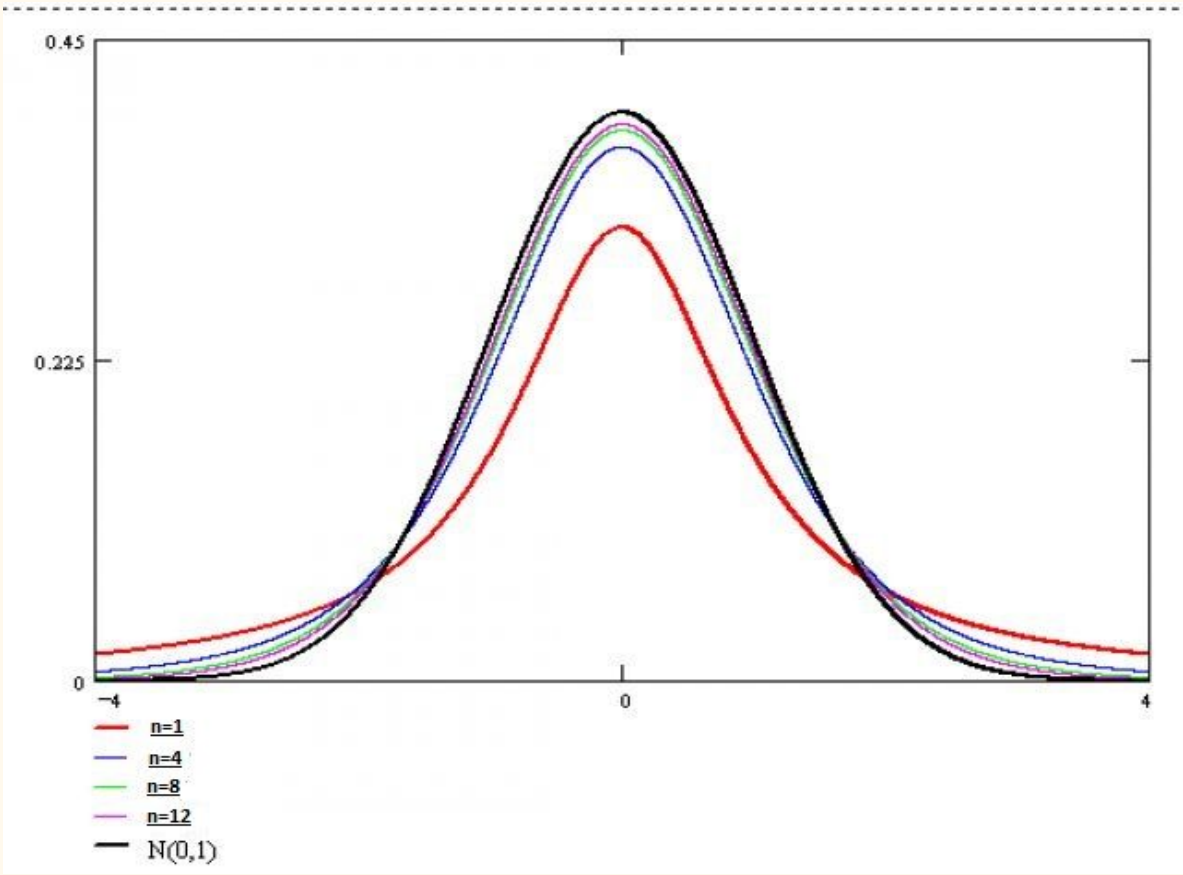
La comprobación del resultado es obvia si tenemos en cuenta que la esperanza de la suma de varias variables aleatorias independientes es la suma de las esperanzas, y que la varianza es la suma de las varianzas, y además que si multiplicamos una variable por una constante, la varianza queda multiplicada por la constante al cuadrado. Por lo dicho:

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\, n\mu = \mu$$

$$var\ (\bar{X}) = var\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = \sum_{i=1}^{n} \frac{1}{n^2}\, var\,(X_i) = n\frac{\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Si además, la población es normal, es decir, $X \equiv N(\mu, \sigma)$ entonces la media muestral es también normal $x \equiv N(\mu, \sigma)$.

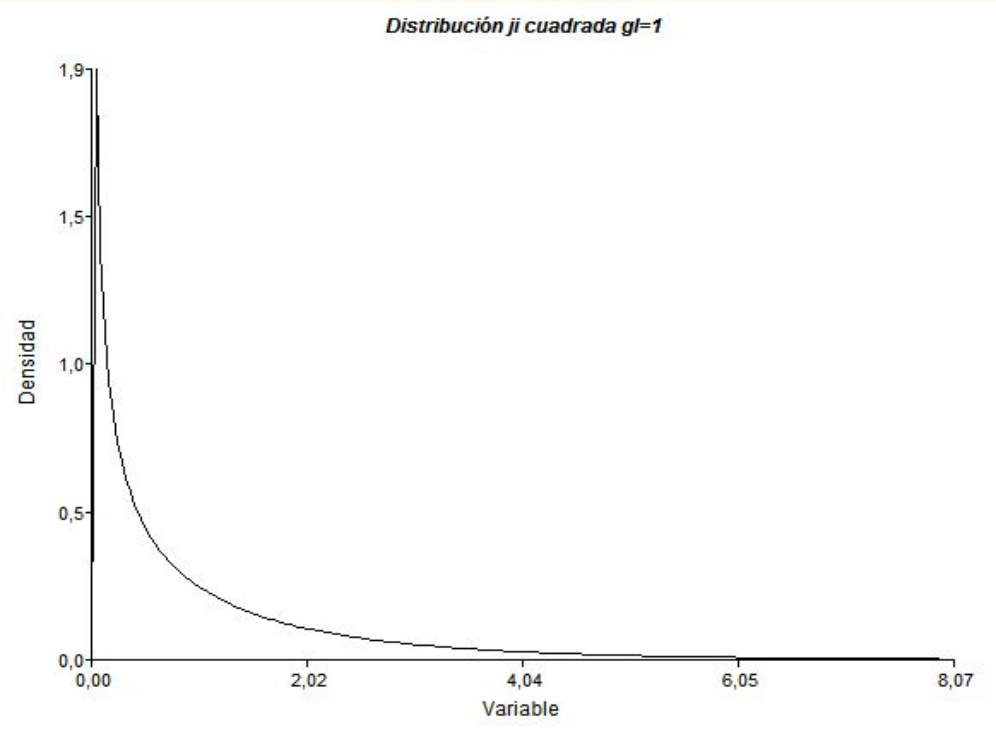
El resultado es importante en estimación ya que, aunque la media poblacional y la media muestral no coincidan, los posibles valores de la media muestral se concentran de forma simétrica alrededor de la media poblacional, además, la dispersión es menor a medida que aumenta el tamaño muestral. En la siguiente figura se muestran distintas distribuciones muestrales de tamaño, 1; 4; 8 y 12, donde se verifica lo dicho anteriormente.



La distribución muestral asociada a varianzas se obtiene con la siguiente ecuación: Sea X_1, X_2, ... , X_n, una muestra aleatoria simple de una población $X \equiv N(\mu, \sigma)$, entonces la variable aleatoria

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

sigue una ji-cuadrado con n-1 grados de libertad.



Del resultado anterior se deduce que las variables

$$\frac{n\,S^2}{\sigma^2} \quad y \quad \frac{(n-1)\,\widehat{S}^2}{\sigma^2}$$

siguen ambas una distribución ji-cuadrado con n-1 grados de libertad.

EL TEOREMA DEL LÍMITE CENTRAL

Existen muchos casos en los que no es posible suponer distribución normal en la población, o saber que dicha distribución no es normal. Sin embargo, tal como se demostrará el proceso de extraer sucesivas muestras aleatorias de la población, permite trabajar con el modelo de la distribución para una distribución de medias aunque la población no lo sea. Esto se deriva de los postulados del Teorema del Límite Central.

Sea X_1, X_2, ..., X_n , una muestra aleatoria de una población X con una distribución de probabilidad no especificada para la que la media es E(X) = μ y la varianza Var(X)= s^2 finita. La media muestral tiene una distribución con media μ y varianza s^2 /*n* que tiende a una distribución normal cuando *n* tiende a infinito. La aproximación a la distribución normal es mejor para n grande ya que se trata de una aproximación y no de una distribución exacta como en el caso de poblaciones normales. Consideramos n grande cuando es mayor de 30.

Una consecuencia directa del teorema es que la suma de los valores muestrales sigue una distribución normal de media nμ, y varianza n σ^2.

PROPIEDADES DE LOS ESTIMADORES

Supongamos ahora que disponemos de una población en la que se mide una variable X con distribución de forma conocida y parámetros desconocidos, por ejemplo una normal con media y varianzas desconocidas. De la población se extrae una muestra aleatoria simple de tamaño *n*, X_1, X_2, ..., X_n. Se trata de calcular, a partir de los valores muestrales, una función de los mismos que proporcione un valor $\theta = u(X_1, X_2 \ldots X_n)$ que sustituya al parámetro desconocido de la población θ, de forma que ambos sean parecidos. A tal valor obtenido de la muestra se le denomina estimador. Un estimador es también una variable aleatoria. Se trata básicamente de buscar estimadores centrados alrededor del verdadero valor del parámetro y con la menor varianza posible.

La distancia entre el estimador y el parámetro a estimar puede medirse mediante el error cuadrático medio, que se define como el valor esperado de la diferencia entre el estimador y el verdadero parámetro.

$$ECM\ (\widehat{\theta}) = E(\widehat{\theta} - \theta)$$

Otra manera en que puede escribirse esta fórmula es la siguiente:

$$ECM\ (\widehat{\theta}) = var\ (\widehat{\theta}) + \left[\theta - E\ (\widehat{\theta})\right]^2$$

la varianza del estimador es igual al cuadrado del sesgo.

Las propiedades deseables que ha de tener un estimador para considerarse adecuado son las siguientes:

Ausencia de sesgo

Se dice que un estimador es insesgado (o centrado) si la esperanza del estimador coincide con el parámetro a estimar:

$$E(\widehat{\theta}) = \theta.$$

En caso contrario se dice que es sesgado por tanto

$$b(\theta) = \left[q - E(\widehat{\theta})\right]$$

se la denomina sesgo.

La propiedad es importante ya que los posibles valores del estimador fluctúan alrededor del verdadero parámetro. Si utilizamos la media muestral como estimador de la media poblacional en una distribución normal, se trata de un estimador insesgado ya que la esperanza de su distribución muestral es la media poblacional μ. Que tenga distribución normal, es importante en la práctica, ya que aunque la media muestral y la poblacional no coinciden exactamente, los valores de aquella fluctúan de forma simétrica alrededor de esta, son valores próximos con probabilidad alta y la dispersión disminuye cuando aumenta el tamaño muestral, tal como se mostró en la anterior figura.

Consistencia

Se dice que un estimador es consistente si se aproxima cada vez más al verdadero valor del parámetro a medida que se aumenta el tamaño muestral. El estimador es consistente si se cumple que:

$$Pr\,[|\widehat{\theta} - \theta| > \varepsilon] \to 0 \; cuando \; n \to \infty \; , \; para \; \varepsilon > 0$$

Esto predice que la distribución del estimador se concentra alrededor del verdadero parámetro cuando el tamaño muestral aumenta.
La media muestral es un estimador consistente de la media poblacional en una distribución normal, ya que, la varianza de la misma σ^2/n tiende a cero toda vez que n $\to \infty$, de forma que la distribución se concentra alrededor del verdadero valor μ cuando *n* crece.

Eficiencia

Un estimador será tanto mejor cuanto menor sea su varianza, ya que se concentra más alrededor del verdadero valor del parámetro. Se dice que un estimador insesgado es eficiente si tiene varianza mínima. Una cota inferior para la varianza viene dada por la denominada cota de Cramer-Rao.

Sea X_1, X_2, ..., X_n. una muestra aleatoria simple de una distribución con densidad f(x; θ). Sujeto a ciertas condiciones de regularidad en la función de densidad, cualquier estimador insesgado verifica que:

$$var\,(\widehat{\theta}) \geq \frac{1}{nE\left[\left(\frac{\partial \ln f\,(X;\theta)}{\partial \theta}\right)^2\right]}$$

$$\ln\,(\theta) = nE\left[\left(\frac{\partial \ln f\,(X;\theta)}{\partial \theta}\right)^2\right]$$

A este último se los denomina cantidad de información de Fisher asociada a una muestra aleatoria simple de tamaño *n*.

METODOS DE ESTIMACION

Método de los Mínimos Cuadrados

Consiste en minimizar la suma de cuadrados de los errores. Éstos se definen como las diferencias entre valores observados y esperados bajo el supuesto que las observaciones se obtienen como la suma de una parte sistemática o controlada y una parte aleatoria no controlada.

En la estimación de la media de una población normal, cada observación x_i puede suponerse como la suma de una constante (la media μ) y un error experimental aleatorio (ε_i):

$$x_i = \mu + \varepsilon_i$$

donde el error experimental aleatorio se obtiene de la sustracción entre cada observación y la media paramétrica bajo una distribución normal N(0, σ).

El método de los mínimos cuadrados consiste en minimizar la suma de cuadrados de los errores o diferencias entre valores observados y esperados.

$$D = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(x_i - \mu \right)^2$$

Derivando con respecto a μ e igualando la derivada a cero obtenemos la media muestral como estimador de la poblacional.

$$\frac{\partial D}{\partial \mu} = \sum_{i=1}^{n} 2\left(x_i - \mu \right)(-1) = 0$$

$$\sum_{i=1}^{n} \left(x_i - \mu \right) = 0$$

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

Método de la Máxima Verosimilitud

Consiste en sustituir los parámetros por aquellos valores que maximizan el logaritmo de la función de verosimilitud de la muestra, esto es, la función de densidad conjunta de todos los valores muestrales en el supuesto de que son independientes. Para la media y varianza de una población normal, los valores muestrales X_1, X_2,... , X_n, se supone que son variables aleatorias independientes y todas con distribución N(μ, σ). La función de densidad conjunta será el producto de las funciones de densidad de cada una de ellas.

$$L\left(\frac{x_1,\ldots x_n}{\mu,\sigma}\right)=\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\frac{\left(x_i-\mu\right)^2}{\sigma^2}}=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{\sum_{i=1}^{n}-\frac{1}{2}\frac{\left(x_i-\mu\right)^2}{\sigma^2}}$$

Tomando logaritmos y derivando con respecto a μ y σ y resolviendo se obtienen como estimadores para la media y la varianza:

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

$$\widehat{\sigma}^2=s^2=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{n}$$

Propiedades de los estimadores Máximo-verosímiles

Los estimadores máximo-verosímiles juegan un papel importante en estadística debido a que se obtienen mediante un método simple y tienen son robustos en cuanto a sesgo, eficiencia y consistencia.

Bajo ciertas condiciones de regularidad se verifica:

a) Si existe un estimador insesgado y de varianza mínima, cuya varianza alcance la cota de Cramer-Rao, este estimador es máximo verosímil y es la única solución de la ecuación de verosimilitud.

b) Si el estimador es sesgado, su sesgo tiende a cero al aumentar el tamaño de la muestra, además es asintóticamente eficiente (cuando *n* grande).

ESTIMADORES PUNTUALES DE LOS PARÁMETROS DE UNA POBLACIÓN NORMAL

Sea una muestra aleatoria simple, X_1, X_2,..., X_n de una población con distribución N(μ, σ), los estimadores de media y varianza son:

Estimador de la media

$$\widehat{\mu} = \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Como se mostró, se trata de un estimador eficiente, insesgado y de varianza mínima. La distribución muestral de la media está dada por:

$$\overline{X} \equiv N\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$$

El error estándar de la media está dado por:

$$EE = \frac{S}{\sqrt{n}}$$

y estima la desviación estándar de la media.

$$\frac{\sigma}{\sqrt{n}}$$

El error estándar de la media mide la variabilidad de la media en el proceso de muestreo.

Estimador de la Varianza

Bajo las condiciones mencionadas anteriormente, el estimador de la varianza muestral es:

$$\widehat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2}{n}$$

este es el estimador sesgado, la cuasi-varianza o estimador insesgado lleva en su denominador *n*-1.

ESTIMADORES DE LOS PARÁMETROS PARA DISTRIBUCIONES DISCRETAS

Así como existen estimadores para distribuciones continuas, los hay para el caso de las distribuciones discretas. Expondremos en este caso dos de tales distribuciones. Si se dispone de una muestra de tamaño n en la que el resultado de la observación es una variable dicotómica (v.g. éxito - fracaso), se trata de estimar la probabilidad *p* de éxito en la población. La variable X= número de éxitos en las *n* pruebas, puede tener distintas distribuciones dependiendo de las condiciones en las que se toma la muestra. En este caso revisaremos la distribución Binomial e Hipergeométrica.

Distribución Binomial

Puede utilizarse toda vez que se toman muestras de poblaciones infinitas o se realiza un muestreo con reemplazamiento de una población finita, donde se realizan *n* pruebas y se contabiliza el número de éxitos en las mismas. El estimador de la proporción de éxito es:

$$\hat{p} = \frac{x}{n}$$

donde *x* es el número de éxitos sobre el total de ensayos.
Si se aproxima *x* mediante una distribución normal, la distribución muestral del estimador de la probabilidad de éxito para muestras grandes es:

$$\hat{p} = \frac{x}{n} \equiv N\left(p, \sqrt{\frac{pq}{n}}\right)$$

q en este caso, es el complemento de la probabilidad de éxitos.

Distribución Hipergeométrica

Si se toman muestras sin reemplazamiento de una población finita de tamaño *N* conocido.El estimador de la proporción de éxito es:

$$\hat{p} = \frac{x}{n}$$

donde *x* es el número de éxitos sobre el total de ensayos.

Aproximando *x* mediante una distribución normal, la distribución muestral del estimador de la probabilidad de éxito para muestras grandes es:

$$\hat{p} = \frac{x}{n} \equiv N\left(p, \sqrt{\frac{pq}{n}\,\frac{N-n}{N-1}}\right)$$

ESTIMACIÓN POR INTERVALOS

Dada una muestra aleatoria X_1, X_2, ..., X_n, de una población con función de densidad $f(x;\ \theta)$, un intervalo de confianza, de extremos L_1 y L_2, para el parámetro θ de la población, es un par ordenado de funciones reales de las *n* medidas de la muestra: $I_\theta = \left[L_1(X_1,..,X_n);L_1(X_1,..,X_n)\right]$ construidas de forma que la probabilidad de que los extremos contenga al verdadero valor del parámetro

es un valor prefijado 1 - α. Al número que resulta de 1 - α se le denomina nivel de confianza.

Por convención, el nivel de confianza suele ser α=0,95 (95%), o bien α=0,99 (99%). La interpretación práctica de este valor es una predicción que determina que si un experimento se repite indefinidamente, el 95% de las veces el intervalo de confianza calculado contendría al verdadero valor del parámetro y en el 5% restante el intervalo no lo contendría (una interpretación similar vale cuando α=0,99).

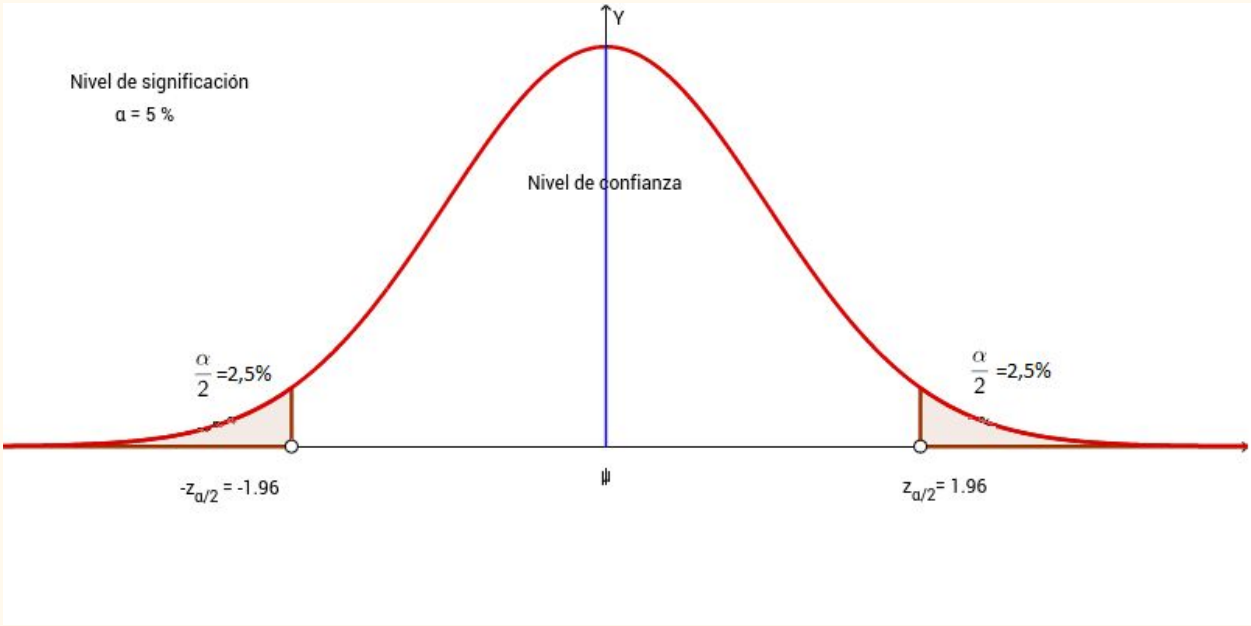
Una vez que el intervalo de confianza ha sido calculado para una muestra concreta, el intervalo obtenido contiene (o no, según el caso), al verdadero valor del parámetro con probabilidad igual a 1. Cuando se han establecido los límites se habla de confianza y no de probabilidad. El valor porcentual confirma cuánta confianza tenemos en que el intervalo que hemos calculado contenga el verdadero valor del parámetro.

INTERVALO DE CONFIANZA PARA LA MEDIA DE UNA POBLACIÓN NORMAL DE VARIANZA CONOCIDA

Supongamos que disponemos de una población en la que sabemos que la variable bajo estudio sigue una distribución N(μ, σ) con σ conocida. Obtenemos una muestra de tamaño *n* y deseamos estimar la media μ de la población. Tal como vimos, el estimador puntual de la misma es la media muestral cuya distribución es conocida. Sobre una distribución N(0,1 - normal estándar) podremos seleccionar dos puntos simétricos -$Z_{\alpha/2}$ y $Z_{\alpha/2}$ tal que:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$p\left(-z_{\frac{\alpha}{2}} \leq z \leq z_{\frac{\alpha}{2}}\right)$$

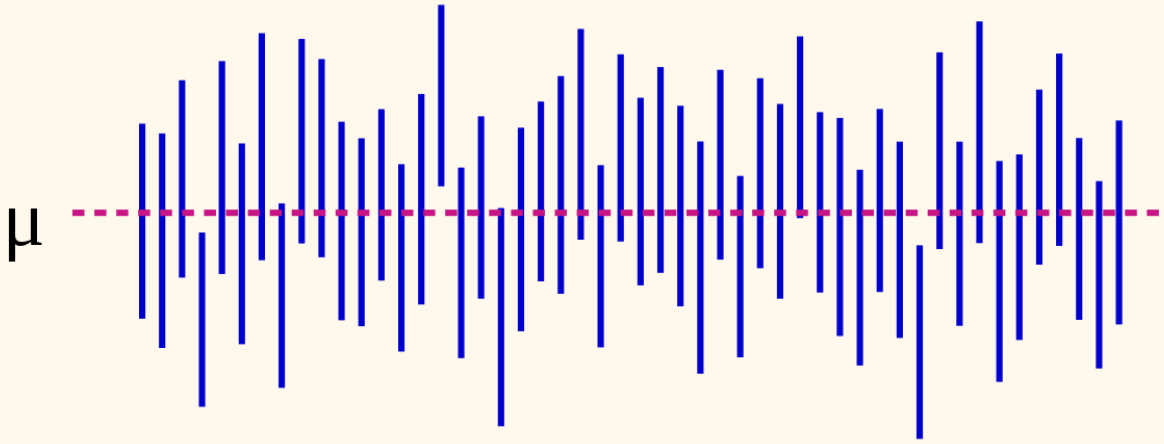


En la práctica, de todos los posibles valores de la media muestral, tenemos uno sólo y por tanto un único intervalo de todos los posibles para distintas muestras:

$$I_{\mu}^{1-\alpha} = \left[\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

La importancia del intervalo de confianza para la estimación está en el hecho de que el intervalo contiene información sobre el estimador puntual (valor central del intervalo) y sobre el posible error en la estimación a través de la dispersión y de la distribución muestral del estimador. El error en la estimación está directamente relacionado con la distribución muestral del estimador y con la varianza poblacional, e inversamente relacionado con el tamaño muestral. El gráfico siguiente ilustra la interpretación del nivel de confianza para un intervalo dado, de una media muestral obtenida de una distribución normal con varianza conocida. Para los distintos posibles valores de la media, representados mediante su distribución muestral, obtenemos distintos intervalos de confianza. La mayor parte incluye al verdadero valor del parámetro. En caso de que el

intervalo se haya calculado en 95% de confianza, debemos esperar un 5% de resultados que no logren capturar el verdadero parámetro. En la práctica disponemos de una única repetición del experimento, y por tanto de un único intervalo de confianza. Por lo tanto, se espera que el intervalo obtenido sea uno de los que contiene el verdadero valor del parámetro.



Si observamos la siguiente figura, tenemos que el conjunto de franjas que se muestran corresponden a diferentes intervalos de confianza. El punto donde intersectan la línea que representa el parámetro poblacional indica que dicho intervalo lo contiene. Como puede verse, algunos intervalos no intersectan al parámetro porque están más abajo (subestiman el verdadero valor), o bien están por encima (sobreestiman el verdadero valor).

AMPLITUD DEL INTERVALO Y ERROR EN LA ESTIMACIÓN

En la práctica hemos de tratar de que la amplitud del intervalo sea lo más acotada posible, es decir, que el error en la estimación sea lo más pequeño posible. Se denomina longitud del intervalo a la distancia que separa ambos límites, y es una medida de su amplitud.

$$long = 2\, z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Esto puede conseguirse modificando las distintas magnitudes que aparecen en la fórmula: a) el valor crítico para el nivel de confianza, b) la variabilidad y c) el tamaño muestral.

La longitud del intervalo de confianza aumenta al aumentar el nivel de confianza ya que el valor crítico de la distribución es mayor. Si consideramos un nivel de confianza del 100%, el intervalo de confianza será (-∞; +∞) y por tanto contendrá al parámetro por defecto. Puesto que contamos solo con una muestra, se hace necesario ajustar el nivel de confianza a valores razonables, que, como se mostró, suelen ser del 95% o del 99%.

La longitud del intervalo de confianza disminuye con la varianza, por lo tanto, la estimación será más precisa cuanto menor sea la variabilidad en la población. Vale decir que para aquellas variables donde la población es más homogénea, la longitud del intervalo será menor.

Finalmente, la longitud del intervalo de confianza disminuye al aumentar el tamaño muestral, lo que significa que se obtienen estimaciones más precisas cuanto mayor sea el tamaño de la muestra. Debido a consideraciones prácticas de coste y tiempo, en general no es posible aumentar indefinidamente el tamaño muestral para obtener estimaciones más precisas, es por ello que en la práctica se selecciona el tamaño muestral necesario para obtener una determinada precisión establecida antes del estudio o experimento. Cabe destacar en este punto que el método por el cual se selecciona la muestra es tanto o más importante que su tamaño.

Si se pretende estimar la media de una población (cuya distribución en la variable de interés se asimila a una normal), de modo que la diferencia existente entre la media muestral obtenida y la media poblacional, esté por debajo de un error prefijado:

$$\left|\bar{x} - \mu\right| \leq E$$

entonces, se debe considerar el intervalo de confianza:

$$P=\left(\bar{x}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\leq\mu\leq+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

el error estaría dado por:

$$E=z_{\frac{\sigma}{2}}\frac{\sigma}{\sqrt{n}}$$

despejando *n* de la igualdad:

$$E^2=z^2_{\frac{\alpha}{2}}\frac{\sigma^2}{n}$$

$$n=\frac{z^2_{\frac{\alpha}{2}}\sigma^2}{E^2}$$

obtenemos la expresión deseada para el tamaño muestral.

En este caso, *n* ha sido calculado contando con variabilidad conocida. Si no tiene este dato, puede aproximarse la variabilidad a partir de trabajos o experimentos previos, o bien, realizando un experimento piloto. El cálculo del tamaño muestral se han igualado el error fijado a priori con el error en la estimación obtenido del intervalo de confianza y que este último incluye el nivel de confianza.

Muchas veces no se conoce la varianza de la población, que habrá que estimar a partir de los datos muestrales. En tal caso se utiliza la cuasi-varianza muestral como estimador. La distribución muestral asociada a la cuasi-varianza es la siguiente:

$$\frac{(n-1)\ s^2}{\sigma^2}\equiv\chi^2_{n-1}$$

Teniendo en cuenta la distribución normal asociada a las medias y combinándola con la ji- cuadrado, obtenemos una distribución t de Student:

$$t = \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}} = \frac{\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\frac{(n-1)\ \hat{s}^2}{\sigma^2}}{n-1}}} = \frac{\bar{x} - \mu}{\frac{\hat{s}}{\sqrt{n}}} \equiv t_{n-1}$$

Siguiendo los mismos pasos para el cálculo del intervalo de confianza en el caso de una distribución normal, resulta que:

$$I_{\mu}^{1-\alpha} = \left[\bar{x} \pm t_{n-1,\alpha} \frac{\hat{s}}{\sqrt{n}} \right]$$

Como se aprecia, el intervalo calculado para este caso es el mismo que para la distribución normal, salvo en el valor crítico y en que la varianza ha sido estimada a partir de la muestra. Esto implica que los valores críticos son un poco más grandes y, por tanto el intervalo tiene mayor longitud. Cuando el tamaño muestral es grande, la distribución *t* es muy similar a la normal, de forma que pueden intercambiarse los valores críticos correspondientes. El intervalo de confianza para la media en muestras grandes se puede escribir como ya lo hemos descrito.

$$I_{\mu}^{1-\alpha} = \left[\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\hat{s}}{\sqrt{n}} \right]$$

TAMAÑO DE MUESTRA EN ESTUDIOS CLÍNICOS

Un aspecto fundamental en el diseño de estudios clínicos es la determinación del tamaño de muestra apropiado. Cuando la muestra es grande, el mayor problema son los costos asociados a su recolección, pero si la muestra es de tamaño menor al requerido, el estudio tendrá baja potencia estadística y en consecuencia, las estimaciones serán menos precisas y la probabilidad de encontrar diferencias significativas entre tratamientos o grupos será menor. El problema entonces es encontrar una muestra equilibrada.

Entre los factores que se deben considerar para el cálculo del tamaño muestral prima el objetivo del estudio, ya que éste determina el parámetro que se desea estimar. También es importante contar con datos sobre el tamaño de la población, la varianza de la variable de interés, el error máximo que se está dispuesto a aceptar, el nivel de confianza deseado, la magnitud del efecto que se pretende encontrar y la potencia estadística deseada. La estimación de parámetros, casi siempre recae en un promedio o una proporción.

Para determinar el tamaño de muestra es preciso definir el nivel de confianza deseado (1-α: 95%, por ejemplo) y el error máximo aceptable. Este último valor dependerá del tipo de estudio clínico que se realice. Normalmente se deducen de los estudios previos realizados.

La fórmula para calcular el tamaño de la muestra en el caso de que la población sea infinita (que en términos prácticos significa que ésta es muy grande), es la siguiente:

$$n_o = \frac{z_\alpha s^2}{d^2}$$

En la fórmula, el tamaño de la muestra viene determinado por el producto del nivel de confianza escogido para el estudio y la desviación estándar de la

variable en estudio (este dato debe obtenerse de investigaciones previas), dividido por error máximo aceptable.

La varianza (y consecuentemente la desviación estándar) de la variable bajo investigación resulta un dato que puede obtenerse de la recolección bibliográfica previa al experimento. De ello resulta que el tamaño de la población también puede resultar conocido. En tal caso, si se puede dar cuenta del tamaño de la población, la fórmula anterior puede reemplazarse por la siguiente:

$$n = \frac{n_o}{1 + \frac{n_o}{N}}$$

Esto permite ajustar el tamaño muestral calculado, al tamaño real de la población.

Si nos interesa estimar una proporción, el método para calcular el tamaño de muestra necesita información previa de la proporción que se espera encontrar en la población. Nuevamente, debemos consultar lo que se haya reportado previamente en la literatura al respecto. También se debe definir el nivel de confianza deseado y el error máximo permitido. Todos los términos de la fórmula han sido previamente definidos; en el numerador se reemplaza por el complemento de la proporción que esperamos encontrar.

$$n_o = \frac{z_\alpha^2 (1 - p)}{d^2}$$

En la situación en que estemos interesados en comparar promedios o proporciones de dos grupos, bajo una hipótesis que establece que existe una diferencia entre ellos (en este caso, trabajamos bajo la hipótesis nula de no diferenciación en la variable criterio), el cálculo del tamaño muestral varía. El cálculo del tamaño de muestra en éstos casos requiere tener en cuenta la

magnitud de la diferencia que se considera importante clínicamente, el nivel de significación o probabilidad de error tipo I máximo que se desea, la potencia estadística , la varianza de la variable bajo estudio, si la prueba de hipótesis es de una o dos colas y el tamaño de la población. Considerando cada una de estas cuestiones, la fórmula para el cálculo del tamaño muestral es la siguiente:

$$n_o = \frac{\left(z_\alpha + z_\beta\right)\, s^2}{d^2}$$

Los valores Z del numerador corresponden a la confianza deseada y la potencia estadística respectivamente. En este caso, se supone que la población es infinita, es decir, desconocemos su límite. En caso de que contemos con información precisa sobre el tamaño de la población, el cálculo del tamaño muestral se puede ajustar con la fórmula que hemos visto anteriormente.

$$n = \frac{n_o}{1 + \frac{n_o}{N}}$$

Para el caso de pruebas de hipótesis sobre dos proporciones, el cálculo del tamaño de muestra es el siguiente.

$$p = \frac{\left|p_1 - p_2\right|}{2}$$

$$n_o = \frac{\left(z_\alpha\sqrt{2p(1-p)} + z_\beta\sqrt{p_1\left(1-p_1\right) + p_2\left(1-p_2\right)}\right)^2}{\left(p_1 - p_2\right)^2}$$

En toda investigación clínica existe la posibilidad de perder pacientes de la muestra, esta es una situación que es necesario prever incrementando el tamaño de muestra para compensar las posibles pérdidas. Si se estima la proporción de

pérdidas, y sea R la expresión de tal proporción, el tamaño de muestra ajustado será:

$$n_{ajustado} = n\left(\frac{1}{1-R}\right)$$

TAMAÑO MUESTRAL EN ESTUDIOS DE MERCADO

A diferencia de los estudios clínicos, en las encuestas de mercado el instrumento más utilizado suelen ser cuestionarios donde se recogen diferentes tipos de datos. Contienen además preguntas estructuradas (dicotómicas o politómicas) sobre diferentes variables de las que se intenta inferir el comportamiento de la población. En el diseño del estudio y los instrumentos hay diversas cuestiones a tener en cuenta: a) cuántas variables deben manejarse en el estudio, b) qué tipo de variables contiene el cuestionario (ordinales, nominales, métricas), c) qué características de la población permitirían reducir el tamaño muestral sin perder fiabilidad en los datos, d) cuál es el error de estimación tolerado. Todas estas cuestiones hacen al diseño e implementación del estudio de mercado, pero para calcular el tamaño muestral apropiado las consideraciones fundamentales son la siguientes: a) los estadísticos que se pretenden estimar (E), b) máximo error permitido en la estimación del estadístico (ε), y c) nivel de confianza con el que se trabaja ($1-\alpha$). Como ya se ha mostrado, estos elementos se relacionan de la siguiente manera:

$$p\left(\left|E_o - E\right| \leq \varepsilon\right) = 1-\alpha$$

La expresión especifica que la probabilidad que la diferencia entre el estadístico que se estime y el valor real del mismo, sea menor o igual que el error prefijado. Tal probabilidad se viene dada por $1-\alpha$. Existen valores convenidos para este

tipo de error, usualmente del 5% y del 1%, pero en estudios de estas características pueden utilizarse otros valores.

El ejemplo más sencillo (que puede generalizarse a diferentes distribuciones de respuestas medidas en escalas politómicas), es el caso de variables dicotómicas. Una variable dicotómica admite dos tipos de respuestas, que al modelizarse bajo una distribución conocida se denominan éxito y fracaso. Las verdaderas respuestas pueden ser de otro tipo, tales como verdadero-falso, acuerdo-desacuerdo, sí-no, etc. A los fines de mostrar el cálculo del tamaño muestral para este tipo de mediciones, una de las categorías de respuesta será tratada como éxito.

La tasa de éxitos en la población se denomina P, y su complemento Q. Estos parámetros deberán estimarse a partir de una pregunta realizada sobre una muestra de la población. Tratándose de variables dicotómicas, la estimación será la proporción de éxitos. Al calcularse sobre la muestra desde donde se hará la inferencia, los estadísticos se denominan p y q respectivamente. El valor estimado de p y q se acercara al verdadero valor de P y Q, cuando el número de encuestados aproxime a N. Esta dependencia se modeliza estadísticamente considerando que el número de respuestas positivas sigue una distribución Binomial de parámetros N y P: $x \approx Bi(N,P)$ donde el valor de P es desconocido y se estima como $p=x_0/n_0$, donde el numerador es igual al número de respuestas afirmativas registradas en la muestra.

Dado que el valor estimado p depende del tamaño de la muestra, se necesita calcularlo de manera tal que el valor estimado p y el valor real P se ajusten a ε. En términos estadísticos, necesitamos saber cuál es el valor de n_0 para que $|P - p| < \varepsilon$. Siendo p una variable aleatoria, esta diferencia sólo se puede asegurar si $n_0 = N$, y por tanto p=P. Pero siendo P un valor estimado $\widehat{P}$, se debe asegurar que el tamaño muestral ajuste la diferencia a $Pr(|P- p| \leq \varepsilon)=1-\alpha$.

Bajo el supuesto que el número de respuestas afirmativas x, se distribuye como una variable Binomial, se tiene que:

$$n_o = \frac{\dfrac{z^2_{1-\alpha} PQ}{\varepsilon^2}}{1 + \dfrac{1}{N}\left(\dfrac{z^2_{1-\alpha} PQ}{\varepsilon^2} - 1\right)}$$

En este caso $Z_{1-\alpha}$ corresponde al cuantil $1-\alpha$ para la distribución normal estándar. Esta expresión se utiliza cuando el tamaño de la población es grande.

De forma general el tamaño muestral es una función decreciente respecto del error permitido. Cuanto mayor es éste menor será el número de encuestas necesarias para alcanzar la estimación.

Para una pregunta en la que son tres las posibles respuestas. Denominaremos P, Q y R a la proporción verdadera o poblacional correspondiente a cada uno de las tres posibles respuestas; *p*, *q* y *r* serán los valores estimados a través de una muestra. Si en el estudio se acepta una diferencia máxima entre ambos valores de un 5% y la probabilidad de que el error supere el valor de α=0.05 (o equivalentemente $1-\alpha$ = 1 - 0.05 = 0.95 = 95%). El tamaño muestral necesario para que la diferencia entre el porcentaje estimado para cada una de las tres respuestas y el real, no supere el 5%, con una probabilidad del 95%, queda definido como: Pr(|P - *p*|<5 y |Q - *q*|<5 y |R - *r*|<5) = 0.95. Modelizada a través de la distribución normal, el tamaño de la muestra se calcula de la siguiente manera:

$$n_o = 1 + \left(\frac{N z^2_{\frac{1-\alpha}{6}}}{\varepsilon^2} - z^2_{\frac{1-\alpha}{6}}\right)$$

Material de Consulta

Villardón, J. L. (2004). Introducción a la inferencia estadística: muestreo y estimación puntual y por intervalos. Salamanca, España: Departamento de Estadística–Universidad de Salamanca.

Camacho-Sandoval, J. (2008). Tamaño de muestra en estudios clínicos. Acta Médica Costarricense, 50(1), 20-21.

Fernández, P. (1996). Determinación del tamaño muestral. Cad Aten Primaria, 3, 138-141.

Se pueden consultar otro títulos del autor sobre este tema:

Lorenzo, J. (2013). Nociones Básicas de Muestreo. Working Paper. Repositorio Ansenuza, U.N.C. http://hdl.handle.net/11086.1/746