Introducción a la Estadística y Probabilidad

Tema 8. Teoría de Muestras

MANUEL MONGE, Ph.D.

Departamento de Economía Aplicada y Métodos Cuantitativos
Facultad de Derecho, Economía y Gobierno
Universidad Francisco de Vitoria

Contenido

1. Introducción

1. Introducción

- A partir de ahora tratamos de conocer la distribución y características de la población a partir de la información que proporciona una **muestra**, o subconjunto de elementos de una población.
- Uno de los objetivos fundamentales de la estadística es obtener conclusiones, o inferir valores sobre características poblacionales (parámetros) a partir de la información que proporciona una muestra. A este proceso se le conoce como **inferencia estadística**.
- La **Estadística Inferencial** o estadística inductiva engloba una serie de **estrategias que permiten generalizar** las propiedades de un conjunto de datos empíricos (**muestra**) al conjunto total de datos (**población**) a los que representan. Para poder efectuar estas generalizaciones (inferencia) es imprescindible que la **muestra sea representativa de la población**.

2. Muestra Aleatoria

Muestra Aleatoria

- Para que se pueda realizar una inferencia correcta, desde el punto de vista de la estadística, la muestra deber ser **representativa** (de la población) y **aleatoria**.
- Trabajaremos con **muestras aleatorias simples**
    1. Cada componente de la muestra sigue la misma distribución de probabilidad que la población de donde ha sido extraída.
    2. Las componentes muestrales son independientes.

3. Estadísticos Muestrales

3. Estadísticos Muestrales

- **Media muestral** $(\bar{X})$

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

- **Media de la distribución en el muestreo** $E(\bar{X})$

  $E(\bar{X}) = E(\frac{X_1+X_2+\ldots+X_n}{n}) = \frac{E(X_1)+E(X_2)+\ldots+E(X_n)}{n} = \frac{\mu+\mu+\ldots+\mu}{n} = \frac{n\mu}{n} = \mu$

  Por lo tanto la media, $E(\bar{X})$, de la distribución en el muestreo de la media muestral, $\bar{X}$, es la media poblacional, μ.

3. Estadísticos Muestrales

- **Varianza de la distribución en el muestreo** ($\sigma^2_{\bar{X}}$)

  $\sigma^2_{\bar{X}} = Var(\bar{X}) = Var(\frac{X_1+X_2+\ldots+X_n}{n}) = (\frac{Var(X_1)+Var(X_2)+\ldots+Var(X_n)}{n^2}) = \frac{\sigma^2+\sigma^2+\ldots+\sigma^2}{n} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$

  Nótese que la varianza de la distribución en el muestreo de $\bar{X}$ disminuye a medida que aumenta el tamaño de la muestra n. Es decir, que cuanto mayor es el tamaño de la muestra, más concentrada está la distribución en el muestreo entorno a la media poblacional.

  Si el tamaño de la muestra (n) es considerablemente grande con respecto al tamaño de la población (N), los individuos de la muestra no están distribuidos independientemente, ya que las observaciones no se seleccionan independientemente porque la probabilidad de que un individuo sea la segunda observación depende del individuo elegido como primera observación. En este caso, la varianza de la media muestral queda modificada por **el factor de corrección** ($\frac{N-n}{N-1}$)

  $(\sigma^2_{\bar{X}}) = Var(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$

- **Desviación estándar de la distribución muestral de la media** llamada también **error típico de** $\bar{X}$

  $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

4. Teorema Central del Límite

4. Teorema Central del Límite

- El Teorema Central del Límite es la base de la teoría de muestras.
- Esta permite utilizar la distribución de probabilidad normal para crear intervalos de confianza de la media poblacional y llevar a cabo pruebas de hipótesis.
- El **Teorema Central del Límite** hace hincapié en que, en el caso de muestras aleatorias grandes, la forma de la distribución muestral de la media se aproxima a la distribución de probabilidad normal.
- En otras palabras, el teorema central del límite se cumple en el caso de todas las distribuciones de la población.
- Si la población obedece a una distribución normal, entonces, en el caso de cualquier tamaño de la muestra, la distribución muestral de las medias también será de narturaleza normal.
- La aproximación es más exacta en el caso de muestras grandes.

4. Teorema Central del Límite

Si una población tiene una media (μ) y desviación típica (σ), tomamos muestras de tamaño n con $n \geq 30$ (o cualquier tamaño si la población es 'normal'), la media de estas muestras siguen aproximadamente una distribución normal dada por

$$\mathrm{N}(\mu, \frac{\sigma}{\sqrt{n}})$$

donde,

$$Z = \frac{\bar{X}-\mu}{\sigma_{\bar{X}}} = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

¿Para qué sirve el teorema central del límite?

- Permite averiguar la probabilidad de que la media de una muestra concreta esté en un cierto intervalo.
- Permite calcular la probabilidad de que la suma de los elementos de una muestra esté, a priori, en un cierto intervalo.

$$\sum_{i=1}^{n} \rightarrow N(n\mu, n\sigma^2)$$

- Inferir la media de la población a partir de una muestra.

Cuando la población es normal, el tamaño de la muestra es inferior a 30 y desconocemos la varianza de la población utilizaremos el estadístico

$$\frac{\bar{X}-\mu}{\frac{S_c}{\sqrt{n}}}$$

que se distribuye como una t de Student de n-1 grados de libertad.

Ejemplo

Supongamos que las subidas salariales porcentuales anuales de los directores generales de todas las empresas de tamaño medio siguen una distribución normal que tiene una media del 12,2 % y una desviación típica de 3,6 %. Se extrae una muestra aleatoria de nueve observaciones de esta población y se calcula la media muestral.. ¿Cuál es la probabilidad de que la media muestral sea inferior a un 10 %?.

- Sabemos que $\mu = 12,2$; $\sigma = 3,5$ y $n = 9$.
- Sea $\bar{X}$ la media muestral y calculamos su error típico

  $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3,5}{9} = 1,2$

- Por tanto, la probabilidad de que la media muestral sea inferior a un 10 % es

  $P(\bar{X} < 10) = (\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} < \frac{10-12,2}{1,2}) = P(Z < -1,83) = 1 - P(Z < 1,83) = 0,0336$

Ejemplo

Un fabricante de bujías sostiene que la duración de sus bujías sigue una distribución normal que tiene una media de 36.000 kilómetros y una desviación típica de 4.000 kilómetros. Una muestra de 16 bujías tenía una duración media de 34.500 kilómetros. Si la afirmación del fabricante es correcta, ¿cuál es la probabilidad de obtener una media muestral de 34.500 o menos?

$\bar{X} \sim N(\mu, \sigma^2_{\bar{X}})$, donde $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{4{,}000}{\sqrt{16}} = 1{,}000$

$P(\bar{X} \leq 34{,}500) = P(\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} \leq \frac{34{,}500-36{,}000}{1000}) = P(Z \leq -1{,}50) = 1 - P(Z \leq 1{,}50) = 0,0668$

Esta probabilidad sugiere que, si las afirmaciones del fabricante ($\mu = 36{,}000; \sigma = 4{,}000$) son ciertas, una media muestral de 34.500 o menos tiene una pequeña probabilidad. Por lo tanto, dudamos de las afirmaciones del fabricante.

4. Teorema Central del Límite

Ejemplo

Cafés La Mexicana, S.A., está considerando la posibilidad de abrir una tienda de cafés en Villalegre, ciudad de 50.000 habitantes. Según algunos estudios de mercado realizados anteriormente, sus tiendas tendrán éxito en las ciudades de ese tamaño si la renta anual per cápita es de más de 60.000 euros con una desviación típica de 5.000 euros. Se ha obtenido una muestra aleatoria de 36 personas y la renta media es de 62.300 euros. ¿Constituye esta muestra una prueba para concluir que debe abrirse una tienda?

El teorema central del límite nos permite asumir que la media muestral sigue aproximadamente una distribución normal. Para responder a la pregunta necesitamos hallar la probabilidad de obtener una media muestral de al menos $\bar{X} = 62{,}300$, considerando que la media poblacional es de $\mu = 60{,}000$.

$$P(\bar{X} \geq 62{,}300) = 1 - P(\bar{X} \leq 62{,}300) \simeq 1 - P\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{62{,}300-60{,}000}{\frac{5{,}000}{\sqrt{36}}}\right) = 1 - P(Z \leq 2,76) = 0,0029$$

Como esta probabilidad es muy baja, podemos concluir que es probable que la renta media de la población no sea 60.000 euros sino mayor. Este resultado es una poderosa prueba de que la renta media de la población es de más de 60.000 euros y de que la tienda de café probablemente será un éxito. En este ejemplo, podemos ver la importancia de las distribuciones en el muestreo y del teorema central del límite para resolver problemas.