INTRODUCCIÓN A LA INFERENCIA ESTADÍSTICA DISTRIBUCIÓN EN EL MUESTREO

Objetivos generales del tema

En este tema se introducirá el concepto de estadístico como medio para extraer información acerca de la ley de probabilidad del fenómeno en estudio. Se verá que, al ser función de v.a. también es una v.a. y se verá cómo obtener su distribución. Por último, se estudiarán algunos casos concretos de estadístico.

Principales contenidos

- Introducción
- Estimación de la media de una población
- Estimación de la varianza de una población
- Estimación de una proporción
- Estadísticos ordenados

1 Introducción

El muestreo estadístico es la herramienta que la Matemática utiliza para el estudio de las características de una población a través de una determinada parte de la misma.

La muestra de estudio debe ser lo más pequeña posible ya que del hecho de que una muestra sea más grande, no se desprende necesariamente que la información sea más fiable.

Además, la muestra elegida debe serlo por un proceso aleatorio para que sea lo más representativa posible.

Términos usuales en un estudio estadístico

- Población: conjunto de todos los individuos que son objeto del estudio.
- Muestra: parte de la población en la que miden las características estudiadas.
- Muestreo: proceso seguido para la extracción de una muestra.
- Encuesta: proceso de obtener información de la muestra.

Los resultados obtenidos del estudio de **una muestra** no son del todo fiables, pero sí en buena medida. Los parámetros que obtienen de una muestra (estimadores estadísticos) nos permitirán arriesgarnos a predecir una serie de resultados para toda la población. De estas predicciones y del riesgo que conllevan se ocupa la Inferencia Estadística.

Las observaciones de una muestra se denotan por $x_1, \ldots, x_n$.

Sin embargo, antes de hacer un muestreo o de experimentar, cualquier observación en particular estará sujeta a incertidumbre (por ejemplo, antes de saber cuál es el gasto medio de una familia de la muestra en alimentación, ésta podría ser $x_1 = 125.000$ ó $x_1 = 87.000$ ó muchos otros valores posibles).

Debido a esta incertidumbre, antes de que se disponga de datos numéricos concretos, consideramos las observaciones como variables aleatorias y las denotamos por letras mayúsculas $X_1, \ldots, X_n$.

Esto a su vez implica que hasta que se hayan obtenido los datos, cualquier función de las observaciones (media muestral, varianza de la muestra, etc.) son funciones de variables aleatorias, y por tanto variables aleatorias con distribución de probabilidad propia, llamada *distribución en el muestreo.*

Definition 1 *Se llama* ***espacio muestral*** *al conjunto de muestras posibles que pueden obtenerse al seleccionar una muestra aleatoria, de un determinado tamaño, de una cierta población.*

Definition 2 *Se llama* ***estadístico*** *a cualquier función $T(X_1, \ldots, X_n)$ de la muestra $(X_1, \ldots, X_n)$. El estadístico $T(X_1, \ldots, X_n)$, como función de variables aleatorias $(X_1, \ldots, X_n)$, es también una variable aleatoria, y tendrá por tanto una distribución de probabilidad, llamada* ***distribución en el muestreo****.*

Algunos estadísticos importantes:

$T(X_1, \ldots, X_n) = \overline{X} = \dfrac{X_1 + \ldots + X_n}{n}$ *Media Muestral*

$T(X_1, \ldots, X_n) = Min(X_1, \ldots, X_n)$

$T(X_1, \ldots, X_n) = Max(X_1, \ldots, X_n)$

$T(X_1, \ldots, X_n) = S_X^2 = \dfrac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n}$ *Varianza Muestral*

$T(X_1, \ldots, X_n) = \widehat{S}_X^2 = \dfrac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}$ *Cuasivarianza Muestral*

Example 3 *Se está interesado en conocer la probabilidad θ de obtener cara con una moneda, es decir, se trata de estudiar la variable aleatoria*

$$X = \begin{cases} 1 & \textit{si se obtiene cara} \\ 0 & \textit{si se obtiene cruz} \end{cases}$$

cuya distribución está caracterizada por

X_i	*Probabilidad*
1	θ
0	$1-\theta$

que depende del parámetro θ que varía en el espacio paramétrico $\Theta = [0,1]$. Se realizan tres lanzamientos, con lo que se dispone de una m.a.s X_1, X_2, X_3. Puesto que la muestra es aleatoria simple, se verifica que

$$\mathbb{P}\left(X_1 = x_1, X_2 = x_2, X_3 = x_3\right) = \mathbb{P}\left(X_1 = x_1\right) \cdot \mathbb{P}(X_2 = x_2) \cdot \mathbb{P}(X_3 = x_3)$$

La probabilidad de todas las muestras posibles es, por tanto, la siguiente:

X_1	X_2	X_3	*Probabilidad*
1	1	1	θ^3
1	1	0	$\theta^2(1-\theta)$
1	0	1	$\theta^2(1-\theta)$
0	1	1	$\theta^2(1-\theta)$
1	0	0	$\theta(1-\theta)^2$
0	1	0	$\theta(1-\theta)^2$
0	0	1	$\theta(1-\theta)^2$
0	0	0	$(1-\theta)^3$

Es decir, tenemos la distribución de la muestra a través de su función de

probabilidad. Si nos interesa el estadístico "media muestral"

$$\overline{X} = \frac{X_1 + X_2 + X_3}{3}$$

a partir de la distribución en el muestreo de la muestra, se tiene la siguiente distribución en el muestreo para el estadístico:

$\overline{X}$	*Probabilidad*
1	θ^3
2/3	$3\theta^2(1-\theta)$
1/3	$3\theta(1-\theta)^2$
0	$(1-\theta)^3$

Example 4 ***De una población*** X ***con distribución de Poisson de parámetro*** λ, $Poisson(\lambda)$, ***se obtiene una m.a.s. de tamaño*** n. ***Determinar la distribución en el muestreo de la media muestral*** $\overline{X}$.

Si X ~ $Poisson(\lambda)$, *entonces* $\mathbb{P}(X=k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$

Dada una m.a.s. $(X_1, \ldots X_n)$, $\quad T(X_1, \ldots X_n) = \frac{1}{n}\sum_{i=1}^n X_i = \overline{X}$. *Su distribución es:*

$$\mathbb{P}\left(\overline{X}=k\right) = \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i = k\right) = \mathbb{P}\left(\sum_{i=1}^n X_i = nk\right) \qquad = e^{-n\lambda}\frac{(n\lambda)^{nk}}{(nk)!}$$

$$\uparrow$$

$$\sum_{i=1}^n X_i \text{ ~ } Poisson(n\lambda)$$

Definition 5 ***Estimador*** *Dada una muestra* $(X_1, \ldots X_n)$, *un estimador del parámetro* θ *es una función de la muestra* $T(X_1, \ldots X_n)$ *que aproxima el valor de* θ. *Se suele denotar por* $T(X_1, \ldots X_n) = \widehat{\theta}$

Definition 6 ***Estimador Insesgado****: Sea* $T(X_1, \ldots X_n)$ *un estadístico que estima un parámetro* θ, *denotémoslo por* $T(X_1, \ldots X_n) = \widehat{\theta}$. *El* ***sesgo*** *del estimador de* θ *es*

$$Sesgo\left[\widehat{\theta}\right] = E\left[\widehat{\theta}\right] - \theta = E\left[T(X_1, \ldots X_n)\right] - \theta$$

El estimador es insesgado si $Sesgo\left[\widehat{\theta}\right] = 0$, *es decir* $E\left[\widehat{\theta}\right] = \theta$.

Definition 7 *Es un estimador* ***asintóticamente insesgado*** *si*

$$Sesgo\left[\widehat{\theta}\right] \to 0 \text{ cuando } n \to \infty$$

Definition 8 ***Estimador Consistente****: Sea* $T(X_1, \ldots X_n)$ *un estadístico que estima un parámetro* θ, *denotémoslo por* $T(X_1, \ldots X_n) = \widehat{\theta}$. *El estimador es consistente si*

$$Var\left[\widehat{\theta}\right] = VAr\left[T(X_1, \ldots X_n)\right] \to 0 \text{ cuando } n \to \infty$$

2 Estimación de la media de una población

Sea $X_1, \ldots, X_n$ una m.a.s. de una variable aleatoria X con media $\mathbb{E}(X) = \mu$ y varianza $Var(X) = \sigma^2$. El estimador más razonable de la media poblacional μ es la **media muestral**

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

que verifica las siguientes propiedades:

1. Es un **estimador insesgado** de μ.

$$\mathbb{E}\left(\overline{X}\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

2. Es un estimador **consistente en media cuadrática** de μ, puesto que es insesgado y su varianza tiende a 0 cuando $n \to \infty$.

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

3. La distribución de exacta de $\overline{X}$ depende de la distribución de la población X. Por ejemplo, si X es normal la distribución de $\overline{X}$ también lo será. Para muestras grandes, por el Teorema Central del Límite, la distribución de $\overline{X}$ puede aproximarse por una normal $(n \geq 30)$.

Theorem 9 *Sea $X_1, \ldots, X_n$ una m.a.s. de una variable aleatoria $X \text{ ~ } N(\mu, \sigma)$, entonces*

$$\boxed{\overline{X} \text{ ~ } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)}$$

Example 10 ***Sea X una población con distribución*** $N(90, \sigma = 20)$.

a) Si se obtiene una m.a.s. de tamaño 16, ¿cuál es la probabilidad de que la media muestral $\overline{X}$ sea mayor o igual que 92?

b) Determinar el tamaño muestral para que la probabilidad de que la media muestral sea menor o igual que 98 sea $\mathbb{P}\left(\overline{X} \leq 98\right) = 0,99$.

$X \text{ ~ } N(\mu = 90, \sigma = 20)$

$$\left.\begin{array}{c} \mathbb{E}\left(\overline{X}\right) = \mu = 90 \\ Var\left(\overline{X}\right) = \dfrac{\sigma^2}{n} = \dfrac{20^2}{16} \Rightarrow \sigma_{\overline{X}} = \sqrt{\dfrac{400}{16}} = \dfrac{20}{4} = 5 \end{array}\right\} \Rightarrow \overline{X} \text{ ~ } N(90, 5)$$

a) $\mathbb{P}\left(\overline{X} \geq 92\right) = \mathbb{P}\left(N(90, 5) \geq 92\right) = \mathbb{P}\left(N(0, 1) \geq \dfrac{92 - 90}{5}\right) =$

$= \mathbb{P}\left(Z \geq \frac{2}{5}\right) = \mathbb{P}\left(Z \geq 0.4\right) = \boxed{0,3446}$

b) $0.99 = \mathbb{P}\left(\overline{X} \leq 98\right) = \mathbb{P}\left(N\left(90, \dfrac{20}{\sqrt{n}}\right) \leq 98\right) = \mathbb{P}\left(N(0, 1) \leq \dfrac{98 - 90}{\frac{20}{\sqrt{n}}}\right) =$

$= \mathbb{P}\left(Z \leq \sqrt{n}\frac{2}{5}\right) \Longrightarrow \sqrt{n}\dfrac{2}{5} = 2,33 \Longrightarrow n = \left(\dfrac{2,33 \cdot 5}{2}\right)^2 = 33,9 \Longrightarrow \boxed{n \geq 34}$

3 Estimación de la varianza de una población

Sea $X_1, \ldots, X_n$ una m.a.s. de una variable aleatoria X con media $\mathbb{E}(X) = \mu$ y varianza $Var(X) = \sigma^2$. El estimador más razonable de la varianza poblacional σ^2 es la **varianza muestral**

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

que verifica las siguientes propiedades:

1. Es un **estimador consistente** de σ^2.

2. Es un estimador **asintóticamente insesgado** de σ^2 puesto que

$$\boxed{\mathbb{E}(S_X^2) = \frac{\sigma^2}{n} + \sigma^2}$$

   Puesto que la varianza muestral S^2 es un estimador sesgado (aunque asintóticamente insesgado), para estimar la varianza σ^2 se usa la **cuasivarianza muestral:**

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

   que es un estimador **insesgado** y **consistente** de σ^2.

$$\mathbb{E}\left(\widehat{S^2}\right) = \mathbb{E}\left(\frac{n}{n-1}S^2\right) = \frac{n}{n-1}\mathbb{E}\left(S^2\right) = \frac{n}{n-1}\left(\sigma^2 - \frac{\sigma^2}{n}\right) = \sigma^2$$

3. (**Teorema de Fisher**) Bajo hipótesis de normalidad ($X \sim N(\mu, \sigma)$) se

verifica que

$$\boxed{\frac{nS^2}{\sigma^2} = \frac{(n-1)\,\widehat{S}^2}{\sigma^2} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\sigma^2} \sim \chi^2_{n-1}}$$

4. Si conocemos la varianza poblacional μ, se verifica

$$\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2} \sim \chi_n$$

5. Bajo hipótesis de normalidad $(X \sim N(\mu, \sigma))$, las variables aleatorias $\overline{X}$ y $\widehat{S}^2$ **son independientes**.

Example 11 ***Dada una población*** $X \sim N(6, \sigma = 2.5)$***, y tomando una muestra aleatoria simple*** $X_1, \ldots, X_n$ ***de tamaño*** $n = 12$***, calcular la probabilidad de que la varianza muestral sea mayor que*** 4.9.

$$\mathbb{P}(S^2 > 4.9) = 1 - \mathbb{P}(S^2 \leq 4.9) = 1 - \mathbb{P}\left(\frac{nS^2}{\sigma^2} \leq \frac{n \cdot 4.9}{\sigma^2}\right) =$$

$$= 1 - \mathbb{P}\left(\chi^2_{n-1} \leq \frac{12 \cdot 4.9}{2.5^2}\right) = 1 - \mathbb{P}(\chi^2_{11} \leq 9.41)$$

En la tabla de la distribución de la χ^2_n *aparecen los siguientes resultados*

$$\mathbb{P}(\chi^2_{11} \leq 7.58) = 0.25$$

$$\mathbb{P}(\chi^2_{11} \leq 10.03) = 0.5$$

Mediante una interpolación numérica se calcula la probabilidad que in-

teresa

x	y
7.58	0.25
9.41	p
10.03	0.5

$$y = y_1 + \frac{y_2 - y_1}{x_2 - x_1}(x - x_1) = 0.25 + \frac{0.5 - 0.25}{10.03 - 7.58}(x - 7.58)$$

$$y = 0.102x - 0.52347 \Longrightarrow p = 0.102 \cdot 9.41 - 0.52347 = 0.43635$$

$$\mathbb{P}(S^2 > 4.9) = 1 - \mathbb{P}(\chi^2_{11} \leq 9.41) = 1 - 0.43635 = \boxed{0.56365}$$

3.1 Estadístico t de Student. Estimación de la media poblacional cuando σ^2 es desconocida.

Bajo hipótesis de normalidad ($X \sim N(\mu, \sigma)$) se verifica que $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, o equi-valentemente

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Este resultado puede ser de poca utilidad si la varianza poblacional σ^2 es desconocida, ya que entonces no se podrá usar esta conclusión para hacer previsiones acerca de $\overline{X}$. Cabe pensar que el resultado no será muy distinto si se sustituye σ por la cuasidesviación típica muestral $\widehat{S}$, puesto que, al menos para muestras grandes, σ^2 y $\widehat{S}^2$ tendrán valores semejantes. Tal idea llevó a Student (pseudónimo de W. Gosset) a considerar el estadístico

$$\boxed{\frac{\overline{X} - \mu}{\frac{\widehat{S}}{\sqrt{n}}} = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n-1}}} \sim t_{n-1}}$$

Example 12 ***Dada una población*** $X \text{ - } N(-1, \sigma)$***, se extrae una m.a.s. de tamaño*** $n = 10$ ***con los siguientes resultados:***

$$1.08,\ -1.79,\ -2.54,\ 0.37,\ -0.6,\ 0.53,\ 0.28,\ -2.21,\ -2.66,\ 1.45$$

Calcular la probabilidad de que la media muestral $\overline{X}$ ***sea mayor que*** -1.2.

$$\mathbb{P}\left(\overline{X} > -1.2\right) = 1 - \mathbb{P}\left(\overline{X} \leq -1.2\right) = 1 - \mathbb{P}\left(\frac{\overline{X} - \mu}{\frac{\widehat{S}}{\sqrt{n}}} \leq \frac{-1.2 - \mu}{\frac{\widehat{S}}{\sqrt{n}}}\right)$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = \frac{1}{10}(1.08^2 + 1.79^2 + 2.54^2 + 0.37^2 + 0.6^2 + 0.53^2 + 0.28^2$$

$$+2.21^2 + 2.66^2 + 1.45^2) - [\frac{1}{10}(1.08 + 1.79 + 2.54 + 0.37 + 0.6 + 0.53 +$$

$$+0.28 + 2.21 + 2.66 + 1.45)]^2 = \frac{1}{10} \cdot 25.7405 - 0.609^2 = 2.20317$$

$$S = \sqrt{2.20317} = 1.4843 \Longrightarrow \widehat{S} = \sqrt{\frac{n}{n-1}S^2} = \sqrt{\frac{10}{9}2.20317} = 1.5646$$

$$\mathbb{P}\left(\overline{X} > -1.2\right) = 1 - \mathbb{P}\left(\frac{\overline{X} - \mu}{\frac{\widehat{S}}{\sqrt{n}}} \leq \frac{-1.2 - \mu}{\frac{\widehat{S}}{\sqrt{n}}}\right) = 1 - \mathbb{P}\left(t_{n-1} \leq \frac{-1.2 - (-1)}{\frac{1.5646}{\sqrt{10}}}\right) =$$

$$= 1 - \mathbb{P}\left(t_9 \leq \frac{-1.2 - (-1)}{\frac{1.5646}{\sqrt{10}}}\right) = 1 - \mathbb{P}\left(t_9 \leq -0.4042\right) = \mathbb{P}\left(t_9 \leq 0.4042\right)$$

Se calcula la probabilidad anterior por interpolación numérica

x	y
0.261	0.6
0.4042	p
0.543	0.7

$$y = y_1 + \frac{y_2 - y_1}{x_2 - x_1}(x - x_1) = 0.6 + \frac{0.7 - 0.6}{0.543 - 0.261}(x - 0.4042)$$

$$y = 0.35462x + 0.456 \Longrightarrow p = 0.35462 \cdot 0.4042 + 0.456 = 0.6$$

$$\boxed{\mathbb{P}\left(\overline{X} > -1.2\right) = 0.6}$$

4 Estimación de una proporción

Se desea estimar la proporción p de individuos de una población que tiene una determinada característica. Para ello se toma una m.a.s. de elementos de la población, anotando un 1 si dicho elemento tiene la característica, y 0 en otro caso, es decir, se tiene una m.a.s. $X_1, \ldots, X_n$ de una $B(1, p)$

$$X = \begin{cases} 1 & \text{(tiene la caract.) con probab. } p \\ 0 & \text{(no la tiene) con probab. } 1 - p \end{cases}, \ \mathbb{E}(X) = p, \ Var(X) = p(1 - p)$$

Un estimador razonable de p es la proporción de elementos de la muestra que tiene dicha característica, es decir

$$\boxed{\widehat{p} = \frac{X_1 + \cdots + X_n}{n}}$$

Propiedades

1. $\mathbb{E}(\widehat{p}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}np = p$

2. $Var(\widehat{p}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} p(1-p) = \frac{p(1-p)}{n}$

3. La distribución de $\widehat{p}$ depende de la distribución de la población X, pero cuando n es grande

$$\boxed{\widehat{p} \text{ ~ } N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)}$$

Example 13 ***En el proceso de producción de una empresa, el 1% de los productos sale defectuoso. Para corroborarlo se obtiene una m.a.s.de tamaño*** $n = 25$ ***y se estima la proporción de productos defectuosos. Estimar la probabilidad de que la proporción estimada sea mayor que el*** 2%.

$$\mathbb{P}(\widehat{p} > 0.02) = \mathbb{P}\left(\frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} > \frac{0.02 - p}{\sqrt{\frac{p(1-p)}{n}}}\right) \simeq \mathbb{P}\left(Z > \frac{0.02 - \widehat{p}}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}}\right) =$$

$$= \mathbb{P}\left(Z > \frac{0.02 - 0.01}{\sqrt{\frac{0.01(1-0.01)}{25}}}\right) = \mathbb{P}(Z > 0.50) = 0.3085$$

$$\boxed{\mathbb{P}(\widehat{p} > 0.02) = 0.3085}$$

5 Estadísticos ordenados

Sea $X_1, \ldots, X_n$ una m.a.s de una población X con función de distribución $F(x)$ y densidad $f(x)$. Es importante estudiar entre qué valores podrían estar los valores muestrales; se consideran entonces $X_{(1)}, \ldots, X_{(n)}$ los estadísticos de orden (los valores muestrales ordenados de menor a mayor $\left(X_{(1)} \leq \ldots \leq X_{(n)}\right)$. Aunque $X_1, \ldots, X_n$ son independientes idénticamente distribuidos (i.i.d.) por tratarse de una m.a.s., $X_{(1)}, \ldots, X_{(n)}$ no lo son.

Ejemplos

$X_{(1)} = \min(X_1, \ldots, X_n)$

$X_{(n)} = \max(X_1, \ldots, X_n)$