

Estadística I (notas breves)

<http://www.est.uc3m.es/svegas>

1 Distribuciones en el muestreo

1.1 Muestras aleatorias

- Es necesario seleccionar un subconjunto representativo de la población, al que se denominará muestra.
- Si la muestra está bien escogida es posible inferir característica de la población a partir de los datos.
- Los estimadores son valores que se obtienen a partir de los datos y aproximan el verdadero valor del parámetro que nos interesa estudiar, siempre y cuando la muestra escogida sea representativa de la población.
- Existen diversos procedimientos de muestreo que nos ayudan a conseguir que la muestra sea representativa:
 - muestreo aleatorio simple: una muestra se denomina muestra aleatoria simple si cada elemento de la población tiene la misma probabilidad de ser elegido y los elementos se seleccionan de forma independiente.
 - muestreo estratificado: una muestra se denomina muestra estratificada si los elementos de la población se dividen en clases o estratos y cada uno de éstos es homogéneo (p.ej. edades).
 - muestra por conglomerados: una muestra se denomina muestra por conglomerados si los elementos de la población se dividen en clases, siendo cada clases tan heterogéneas como la población y las clases homogéneas entre sí (p.ej. provincias).
- Nos centraremos en las muestras aleatorias simples o m.a.s. Así, una muestra aleatoria simple es una colección de n variables aleatorias v.a. $(X_1, X_2, \dots, X_n)$ tal que $(X_1, X_2, \dots, X_n)$ son independientes y cada X_i tiene una función de distribución $F_\theta(x)$, siendo θ la colección de parámetros de la distribución que nos interesa conocer.
- Suponemos que la forma de la distribución $F_\theta(x)$ es conocida salvo por θ . Por ejemplo, si X sigue una distribución Normal, $X \sim \mathcal{N}(\mu, \sigma^2)$, los parámetros de interés son $\theta = (\mu, \sigma^2)$. Así, para estimar el parámetro μ escogeremos una m.a.s y calcularemos su media $\bar{x}$.
- Al número n le denominaremos tamaño muestral.
- La función de distribución conjunta de la m.a.s $(X_1, X_2, \dots, X_n)$ es

$$F_\theta(x_1, x_2, \dots, x_n) = F_\theta(x_1) \cdot F_\theta(x_1) \cdots F_\theta(x_n)$$

- La función de densidad conjunta de la m.a.s $(X_1, X_2, \dots, X_n)$ es

$$f_\theta(x_1, x_2, \dots, x_n) = f_\theta(x_1) \cdot f_\theta(x_1) \cdots f_\theta(x_n)$$

- A una realización concreta de una m.a.s $(X_1, X_2, \dots, X_n)$ la denominaremos $(x_1, x_2, \dots, x_n)$

1.2 Estadísticos

- Un estadístico es cualquier función de la muestra, luego también es una variable aleatoria: su valor va cambiando a medida que se obtienen diferentes muestras.
- Denotaremos a los estadísticos por $\hat{\theta} = g(X_1, X_2, \dots, X_n)$
- Estadísticos habituales:

- media muestral: $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ (su correspondiente característica poblacional es la media poblacional o esperanza: $\mu = E(X) = \int x f(x) dx$)

- varianza muestral: $\hat{S}_X^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ (su correspondiente característica poblacional es la varianza poblacional $\sigma^2 = Var(X) = \int (x - \mu)^2 f(x) dx$)
- cuasi-varianza muestral: $S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$
- mínimo: $\hat{\theta} = \min\{X_1, X_2, \dots, X_n\}$
- máximo: $\hat{\theta} = \max\{X_1, X_2, \dots, X_n\}$
- logaritmo: $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n \log X_i$

1.3 Distribución muestral

- La distribución de probabilidad correspondiente a un estadístico se denomina distribución muestral.
- La esperanza, la varianza y la desviación típica de la v.a. media muestral $\bar{X}$ de una m.a.s con $E(X_i) = \mu$ y $Var(X_i) = \sigma^2$ toman la forma siguiente,

$$\begin{aligned} E(\bar{X}) &= \frac{1}{n} \sum_{i=1}^n E(X_i) = \mu \\ Var(\bar{X}) &= \frac{1}{n^2} Var(\sum_{i=1}^n X_i) = \frac{\sigma^2}{n} \\ sd(\bar{X}) &= \frac{\sigma}{\sqrt{n}} \end{aligned}$$

- La esperanza de la v.a. varianza muestral $\hat{S}_X^2$ de una m.a.s con $E(X_i) = \mu$ y $Var(X_i) = \sigma^2$ es

$$E(\hat{S}_X^2) = E\left(\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right) = \frac{n-1}{n} \sigma^2$$

- A medida que aumenta el tamaño de la muestra el error estándar (o desviación típica) disminuye, por lo que cuanto mayor sea el tamaño muestral mejor será la inferencia sobre la media poblacional.
- La distribución Chi-cuadrado con n grados de libertad, $\mathcal{X}_n^2$, se define como la suma de n normales independientes $\mathcal{N}(0, 1)$ al cuadrado,

$$\begin{aligned} \mathcal{X}_n^2 &= \mathcal{N}(0, 1)^2 + \dots + \mathcal{N}(0, 1)^2 \\ E(\mathcal{X}_n^2) &= n \\ Var(\mathcal{X}_n^2) &= 2n \end{aligned}$$

- La distribución t de Student con n grados de libertad, t_n , se define como el cociente de dos distribuciones independientes, una de ellas $\mathcal{N}(0, 1)$ y la otra $\mathcal{X}_n^2$,

$$\begin{aligned} t_n &= \frac{\mathcal{N}(0, 1)}{\sqrt{\frac{1}{n} \mathcal{X}_n^2}} \\ E(t_n) &= 0 \\ Var(t_n) &= \frac{n}{n-2} \\ t_n &\xrightarrow{n \rightarrow \infty} \mathcal{N}(0, 1) \end{aligned}$$

- **Teorema Central del Límite (TCL):** Sean $(X_1, X_2, \dots, X_n)$ v.a.i. con $E(X_i) = \mu_i$ y $Var(X_i) = \sigma_i^2$, entonces,

$$\text{i) } \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \xrightarrow{n \rightarrow \infty} \mathcal{N}(0, 1)$$

$$\text{ii) si las variables } (X_1, X_2, \dots, X_n) \text{ son además idénticamente distribuidas (v.a.i.d.) con } \mu_i = \mu \text{ y } \sigma_i^2 = \sigma^2, \text{ entonces } \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{n \rightarrow \infty} \mathcal{N}(0, 1)$$

- **Lema de Fisher:** Sean $(X_1, X_2, \dots, X_n)$ v.a.i.i.d $\mathcal{N}(\mu, \sigma^2)$, entonces

i) $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ o equivalentemente $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$

ii) $\frac{(n-1)}{\sigma^2} S_X^2 \sim \chi_{n-1}^2$

iii) $\bar{X}$ e S_X^2 son independientes

- **Lema de Student:** Sean $(X_1, X_2, \dots, X_n)$ v.a.i.i.d $\mathcal{N}(\mu, \sigma^2)$, entonces

$$\frac{\bar{X}-\mu}{S_X/\sqrt{n-1}} = \frac{\bar{X}-\mu}{S_X/\sqrt{n}} \sim t_{n-1}$$

- Se suele considerar que n es suficientemente grande si $n > 30$
- La esperanza y la varianza de la v.a. cuasi-varianza muestral S_X^2 de una m.a.s con $E(X_i) = \mu$ y $Var(X_i) = \sigma^2$ toman la forma siguiente

$$E(S_X^2) = \frac{\sigma^2}{n-1} E(\chi_{n-1}^2) = \sigma^2$$

$$Var(S_X^2) = \frac{\sigma^4}{(n-1)^2} Var(\chi_{n-1}^2) = \frac{2\sigma^4}{n-1}$$

- Sea α cualquier número entre 0 y 1, definimos

· z_α como el valor de la normal estándar $Z \sim \mathcal{N}(0, 1)$ tal que $P(Z > z_\alpha) = 1 - \Phi(z_\alpha) = \alpha$, siendo Φ la función de distribución de la normal estándar.

· $t_{n,\alpha}$ como el valor de la t de Student t_n tal que $P(t_n > t_{n,\alpha}) = \alpha$

· $\chi_{n,\alpha}^2$ como el valor de la χ_n^2 tal que $P(\chi_n^2 > \chi_{n,\alpha}^2) = \alpha$