Análisis y Visualización de Datos

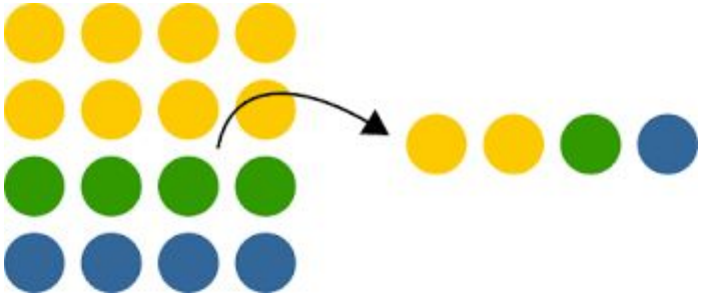
Diplomatura CDAAyA
2024

Teoría para aplicar

Muestreo aleatorio

Cuando recogemos los datos muchas veces es imposible relevar la característica de interés de todo el grupo o universo, se examina una pequeña parte, llamada muestra.

Población   Muestra



La muestra debe ser representativa de la población

Muestra aleatoria

Al medir una característica en una muestra, se consideran los datos $x_1, x_2, \dots, x_n$,
como realizaciones de $X_1, X_2, \dots, X_n$ muestra aleatoria (m.a.)

Notar la diferencia entre minúscula y mayúscula

Muestra aleatoria

Una sucesión de v.a. $X_1, X_2,..., X_n$ se dice **muestra aleatoria (m.a.)** si son v. a.

independientes e idénticamente distribuidas (i.i.d.). “Clones” de una misma X

Todas las medidas antes mencionadas para una muestra de datos podemos pensarlas a

partir de una muestra aleatoria, también serán variable aleatorias, llamadas estadísticos.

Como por ejemplo el **estadístico Media Muestral:**

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

Algunas propiedades teóricas: Muestra aleatoria

- Si $X_1, X_2, \ldots, X_n$ m.a. (v.a.i.i.d.) tal que $X_i \sim N(\mu,\sigma^2)$, entonces:

  - $X_1 + X_2 + \ldots + X_n \sim N(n\mu, n\sigma^2)$
  - $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu,\sigma^2/n)$

- Si $Z_1, Z_2, \ldots, Z_n$ m.a. tal que $Z_i \sim N(0,1)$, entonces:

$$V = \sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$$

LGN: Ley de los Grandes Números

- Dada. X_1, ...X_n m.a. c/u con media μ ("clones" de la misma variable con distribución cualquiera pero con esperanza $E(X_1)= \mu$, media poblacional)

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu \qquad\qquad P\left(\left|\overline{X} - \mu\right| \leq \epsilon\right) \underset{n\to\infty}{\longrightarrow} 1$$

TCL: Teorema Central del Límite

- Sea X_1, ...X_n m.a. c/u con media μ y varianza σ^2. ("clones" de la misma variable con distribución cualquiera pero con media μ y varianza σ^2)

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \xrightarrow{\mathfrak{D}} Z \sim N(0,1))$$

$$\overline{X} \approx N(\mu, \sigma^2/n)$$

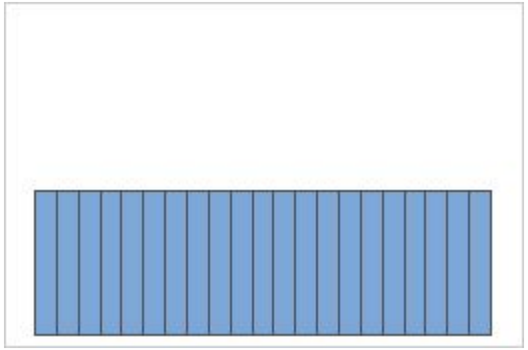
TCL: Teorema Central del Límite

- La densidad de **la v.a. media muestral** parece acampanada p/ **n grande**, **aprox. normal**, cualquiera sea la distribución en la población;

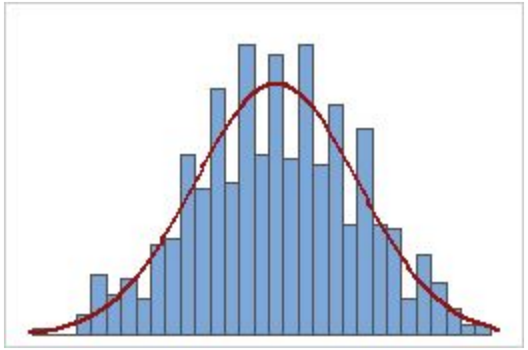
- La densidad de **la media muestral** crece en altura y decrece en dispersión si n crece.

- La media de la distribución del promedio muestral es igual a la media de la población
- La varianza de la distribución de la media muestral es menor que la varianza de la población;

TCL: Teorema Central del Límite

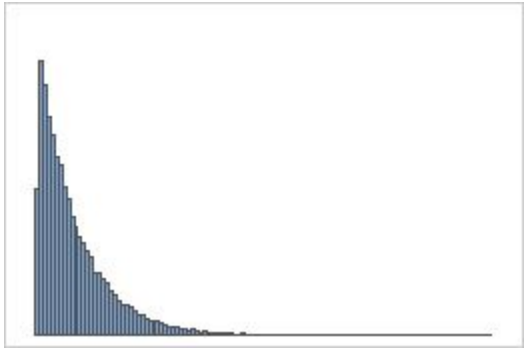


X_i ~ Uniforme

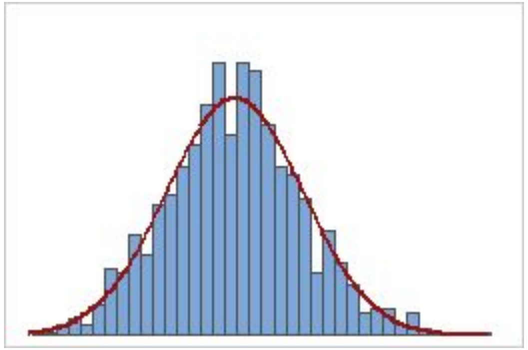


$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

TCL: Teorema Central del Límite



X_i ~ exponencial



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

Estadística Inferencial

Inferencia Estadística

Métodos utilizados para **tomar decisiones** o para **obtener conclusiones** sobre una población (usa modelos y parámetros generalmente).

Estos métodos utilizan la información contenida en una muestra de la población.

Permiten **inferir** el comportamiento de la población con un riesgo medible en términos de **probabilidad de error**.

Inferencia Estadística

Si nos ubicamos dentro de la estadística paramétrica:
Se considera una característica de interés de la **población (Ω).** Se supone que la característica está modelada por una **variable aleatoria X** con distribución "conocida" y paramétrica $f_\theta(x)=f(x,\theta)$. (ej $\theta=(\mu,\sigma^2)$ en Normal)

Se considera una **muestra aleatoria (m.a.) X_1, ...X_n**, con la misma distribución (paramétrica) que X.

Inferencia Estadística

Incluye dos grandes áreas:

- estimación de parámetros (p/estadística paramétrica)
- pruebas de hipótesis

Demo con Notebook

04 Estadísticos.ipynb