Inferencia Estadística

1. INTRODUCCIÓN

ITAM - Primavera 2020

Juan B. Martínez Parente

Probabilidad *vs.* estadística

Son dos conceptos muy diferentes pero estrechamente relacionados.

La **probabilidad** cuantifica la incertidumbre asociada a fenómenos que presentan variabilidad.
Se preocupa por predecir los **posibles** resultados de un fenómeno.
Es una rama **teórica** de las matemáticas.

La **estadística** utiliza datos **observados** a partir de los cuales infiere o generaliza respecto a la naturaleza de un fenómeno.
Es una rama **aplicada** de las matemáticas.

Otra forma de verlo: la probabilidad nos ayuda a determinar las consecuencias de un fenómeno ideal, mientras que la estadística nos permite medir hasta qué punto el fenómeno es ideal.

Inferencia estadística

inferir Conjugar

Del lat. *inferre* 'llevar a'.

Conjug. actual c. *sentir.*

1. tr. Deducir algo o sacarlo como conclusión de otra cosa. *Se infiere* DE *su rostro que está contento.*

Fuente: RAE

En nuestro caso, «inferencia» se refiere al proceso de deducir propiedades de una **población** a partir de una **muestra** de la misma.

Ramas de la estadística

Dentro de la inferencia estadística se pueden identificar dos ramas.

Estadística descriptiva: consiste en recopilar, depurar, describir, analizar y presentar los datos. Se basa únicamente en los datos observados de la muestra.

Estadística inferencial: consiste en hacer estimaciones y validar las mismas con el propósito de generalizar el comportamiento de los datos. Se basa en la información observada y en supuestos estadísticos.

$X=$ ingreso promedio familiar al mes

$\underline{X \sim N(\mu, \sigma^2)}$, $\mu \sim Exp(\lambda)$, $\sigma^2 \sim ga(\alpha, \beta)$

$\lambda \sim Exp(\lambda_0)$ $\alpha, \beta \sim N(\mu_2, \sigma_2^2)$

Enfoques y tipos de inferencia

Adicional a lo anterior, se pueden identificar dos **enfoques** diferentes en el proceso inferencial:

Clásico o **frecuentista**: hace uso únicamente de la información observada.

Bayesiano: combina la información observada con la información (subjetiva) del tomador de decisiones.

Además de los enfoques de inferencia, se pueden identificar diferentes **tipos** de inferencia dependiendo de los supuestos que se hagan sobre el comportamiento probabilístico de los datos, los cuales pueden ser:

Paramétrico: se especifica un modelo de probabilidad en particular a los datos.

No paramétrico: no se especifica un modelo de probabilidad, en su lugar dejamos que los datos «hablen por sí mismos», es decir, determinen su propio comportamiento probabilístico.

Enfoques y tipos de inferencia

	Clásico	Bayesiano
Paramétrico	$\boxed{\text{I}}$	III
No paramétrico	II	IV

Población y muestra

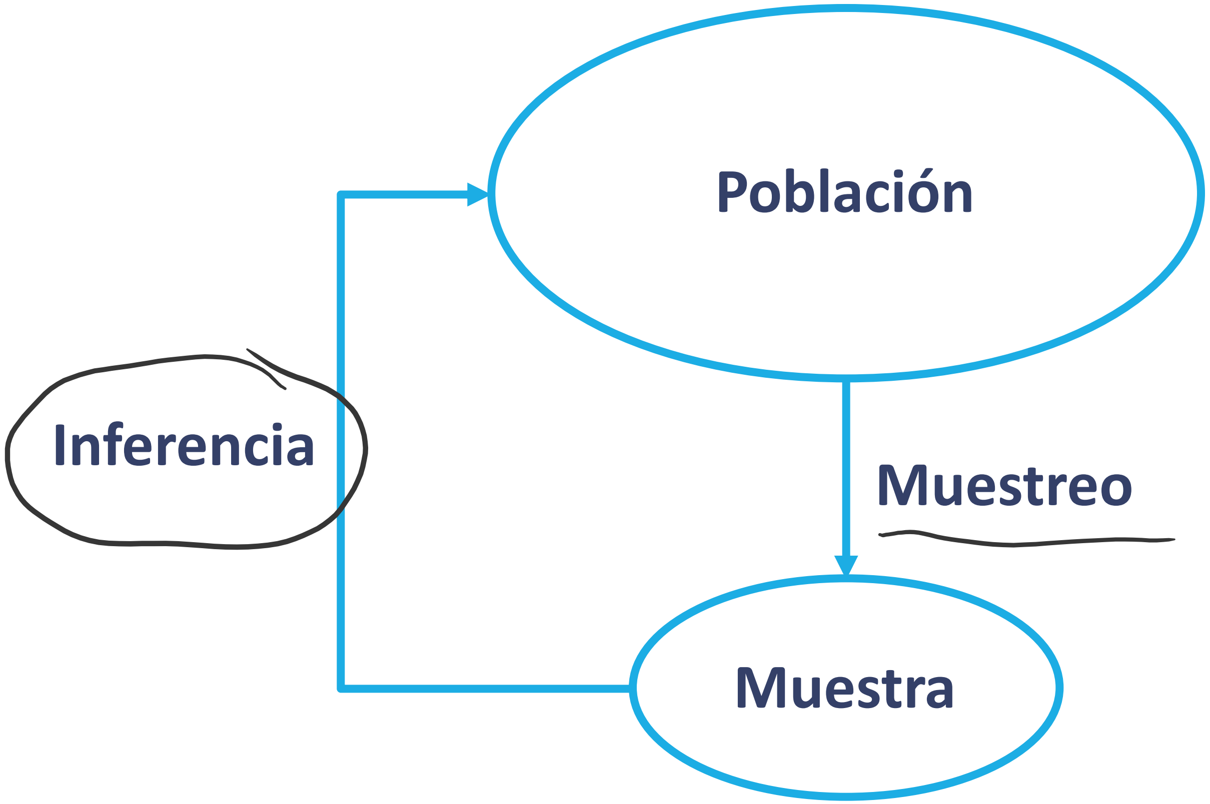
La **población** es el conjunto de elementos de interés; representa a la totalidad de elementos bajo estudio.

La **muestra** es un subconjunto de la población.

Una muestra debe tener la característica de representar lo mejor posible la heterogeneidad de la población, de lo contrario se dice que la muestra es sesgada o poco representativa.

Es vital que la muestra sea una buena / insesgada / **representativa**: de lo contrario la inferencia será incorrecta, aun si usamos correctamente otras técnicas estadísticas.

Población y muestra



Población y muestra

Una forma de resumir la información de todos los elementos de la población es mediante **parámetros**. A partir de ellos se puede caracterizar a los elementos de una población.

Durante el curso, estudiaremos cómo a partir de muestras es posible inferir los parámetros que definen a una población.

$$\underline{\mu, \sigma^2, \lambda, \alpha, \beta, \eta, \theta}$$

$$X \sim N(\underbrace{\mu, \sigma^2}_{?})$$

Estimadores: aproximaciones de los parámetros

$$\hat{\sigma}^2 = s^2$$

$$\mu = \bar{X}$$

Medidas que describen a una población

Los parámetros de una población generalmente se asocian a medidas de centralidad, dispersión, posición y asociación de los elementos que la conforman.

1. Medidas de tendencia
2. Medidas de posición
3. Medidas de dispersión

Medidas de tendencia central

Las medidas de tendencia central son útiles para tener una idea del **comportamiento típico** de los datos.

Generalmente se emplea a la media muestral como métrica para evaluar dicho comportamiento, aunque con frecuencia ocurre que la existencia de datos atípicos o valores extremos subestimen o sobrestimen la centralidad de los datos.

Cuando esto ocurre, lo más recomendable es emplear otras medidas de tendencia central, como la mediana, que resulta ser más robusta en estos escenarios pues no sólo considera el valor de los datos, si no la cantidad de estos.

Ejemplos de medidas de tendencia central:

1) Media muestral: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$\lfloor 3.5 \rfloor = 3$

2) Mediana (muestral): $med(\underline{X}) = \frac{1}{2}\left(X_{\lfloor \frac{n}{2} \rfloor} + X_{\lfloor \frac{n+1}{2} \rfloor}\right)$

$X_1 \leq X_2 \leq X_3 \leq \cdots \leq X_n$

1, 2, 3, 4, 5

$\rightarrow$ 1, 2, $\boxed{3, 4}$, 5, 6

$\frac{1}{2}(3+4) = 3.5$

$\underline{X} = \{1, 2, \boxed{3, 4}, 5, 50\}$

$\bar{X} = 10.83$

$med(\underline{X}) = \underline{3.5}$

$\underline{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$

$\underline{X}^T = (X_1, X_2, \ldots, X_n)$

Medidas de tendencia central

Media muestral:

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

Mediana: si la muestra está ordenada ascendentemente,

$$mediana(\bar{X}) = \frac{X_{\left\lfloor \frac{n+1}{2} \right\rfloor} + X_{\left\lceil \frac{n+1}{2} \right\rceil}}{2}$$

Medidas de posición

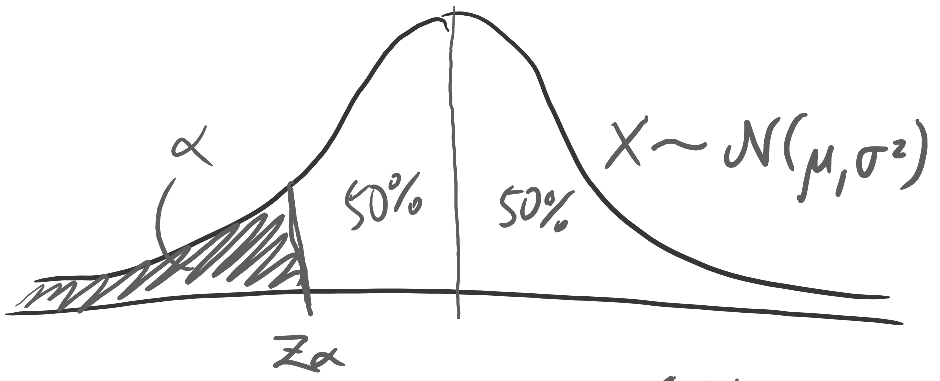
Las medidas de posición (llamados también **percentiles** o **cuantiles**) dividen a una distribución ordenada en partes iguales, los cuales corresponden al valor de la variable por debajo del cual se encuentra un porcentaje dado de observaciones, una vez que los datos han sido ordenados de menor a mayor.

Cuando el intervalo de porcentaje se divide en cuatro partes iguales, los percentiles se denominan **cuartiles** y ésta es la forma más común de representar medidas de posición de los datos.

Medidas de posición:

x es el k-ésimo percentil (o cuantil) si

$$P(X \leq x) = k$$



$$P(X \leq z_\alpha) = \alpha$$

$\underbrace{1,\ 2,}\ 3,\ \underbrace{4,\ 5}$

$\underbrace{1,\ 2,\ 3}\ |\ \underbrace{4,\ 5,\ 6}$

3.5

Medidas de posición

x es el k-ésimo percentil si:

$$\mathrm{P}(X \leq x) = k$$

Medidas de dispersión

Las medidas de dispersión son métricas que nos ayudan a entender el comportamiento de los datos con respecto a su centroide (o punto de referencia central).

Es decir, miden la **variabilidad** de una muestra en torno a su centro.

Ejemplos de medidas de dispersión:

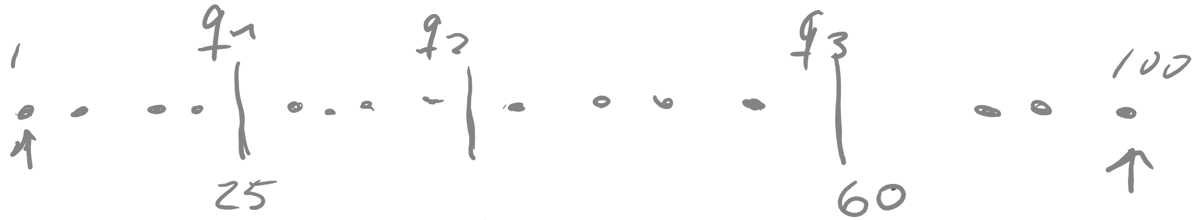
1) Varianza muestral: $\underline{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (\underline{x}_i - \underline{\bar{x}})^2$

$x_i - \bar{x}$ grande,
$(x_i - \bar{x})^2$ muy grande

$x_i - \bar{x}$ chico,
$(x_i - \bar{x})^2$ muy chico

2) Rango: $R(\underline{X}) = x_n - x_1$

3) Rango intercuartílico: $q_3 - q_1$

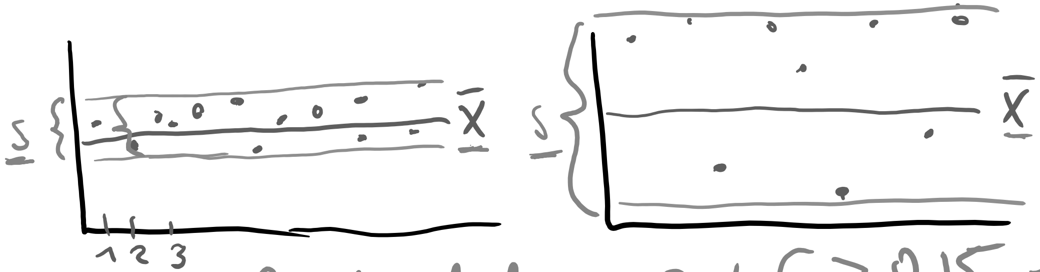


$R(\underline{X}) = 100 - 1 = 99$

$RI(\underline{X}) = 60 - 25 = 35$

4) Coeficiente de variación

$$CV(\underline{X}) = \frac{s}{\bar{X}} \longrightarrow s = \sqrt{s^2}$$



Regla de dedo: CV: $\begin{cases} > 0.15 & \text{variabilidad grande} \\ < 0.15 & \text{poca variabilidad} \end{cases}$

Medidas de dispersión

Rango de los datos:

$$R(\bar{X}) = X_n - X_1$$

Varianza muestral:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Rango intercuartílico:

$$RI(\bar{X}) = q_3 - q_1$$

Coeficiente de variación:

$$CV(\bar{X}) = \frac{s}{\bar{\bar{X}}}$$