Distribuciones muestrales

Astelarra Mateo, Ferrari Analia, Kuna Paula

Niveladora de estadística

Población, muestra y parámetros

Definiciones y notación

Una **población** es un conjunto o universo de items (objetos, hogares, personas, etc) de interés.

- A los items se los denomina comúnmente **unidades**.
- **N** designa la cantidad de unidades (tamaño) de la población.
- Una **muestra** es un subconjunto de items de la población.
- n designa la cantidad de unidades (tamaño) de la muestra.
- Un **parámetro** es una característica específica de la población.

Poblaciones y parámetros: ejemplos

Ejemplos

- **Población:** Todos los votantes registrados de la ciudad de Buenos Aires.
  Parámetro: Proporción de votantes decididos a votar al candidato A.
- **Población:** Todos los hogares de la ciudad de Buenos Aires.
  Parámetro: Ingreso medio de los hogares en Buenos Aires.
- **Población:** Todos los estudiantes ingresantes en escuelas secundarias de la ciudad de Buenos Aires.
  Parámetro: Proporción de ingresantes que finalizan los estudios.
- **Población:** Todos los empleados de una hamburgueseria.
  Parámetro: Media del tiempo de contrato.
- **Población:** Todos las ventas de televisores en noviembre de una cadena de electrodomésticos.
  Parámetro: Total del valor monetario de las ventas.

Inferencia estadística

El objetivo de la inferencia estadística es sacar conclusiones y/o tomar decisiones concernientes a algun parámetro desconocido de una población basándose solo en los datos de una muestra. A grandes rasgos hay tres tipos de inferencias :

Estimación puntual	valiéndose del ingreso de c/persona en la muestra, **adivinar** el valor de la media del ingreso mensual de la población.
Estimación por intervalos	valiéndose del ingreso de c/persona en la muestra, proveer un intervalo que con alta probabilidad cubra a la media del ingreso mensual de la población
Test de hipótesis	valiéndose del ingreso de c/persona en la muestra, evaluar la evidencia a favor o en contra de la hipotesis de que la proporción de hogares con ingreso menor a 50000 pesos es mayor o igual que 0,9.

¿Por qué muestrear?

- Consume menos tiempo y recursos que un censo.
- Muchas veces es prácticamente imposible evaluar todos los items de una población.
- Muchas veces es posible obtener resultados con mucha precisión basándose simplemente en los datos de una muestra.

Datos de la muestra

- En cualquier estrategia de muestreo al azar, las unidades que serán seleccionadas son desconocidas y, por lo tanto, **aleatorias**.
- Los datos numéricos de los atributos que se medirán en cada unidad i seleccionada es **una variable** X_i.
- La totalidad de la informacion numérica a recabar en la muestra esta conformada por $X_1, \ldots, X_n$.
- Una vez que se han registrado los datos numéricos de la muestra, designamos genéricamente con letra minuscula a los datos registrados: $x_1, \ldots, x_n$.

O sea

- X_i : los datos a recoger en la unidad i de la muestra (antes de recogerlos).
- x_i : los datos registrados en la unidad i de la muestra (una vez que los datos han sido recogidos).

Datos de la muestra: ejemplo

De una ciudad se toma una muestra aleatoria de hogares:

- Si en cada hogar se mide solo el ingreso mensual, entonces

$$X_i = \text{ ingreso mensual del hogar } i$$

- Si en cada hogar se mide el ingreso mensual y el número de habitantes, entonces

$$X_i = (X_{1i}, X_{2i})$$

donde X_{1i} es el ingreso del hogar i y X_{2i} es el número de habitantes del hogar i.

Histograma de una lista de valores numéricos

El **histograma de una lista de valores numéricos** $x_1, \ldots, x_n$ (en los que pueden aparecer valores repetidos)

- solo puede tomar valores en el conjunto $\{x_1, \ldots, x_n\}$
- su probabilidad de que tome el valor x_i es igual a la frecuencia con la que aparece x_i en la lista.

Ejemplo

El histograma de la lista $x_1 = 7, x_2 = 9, x_3 = 7$

Media, varianza y desvío de una lista de valores numéricos

Definición

La media (o promedio), varianza y desvío estándar de una lista de valores numéricos $x_1, \ldots, x_n$ es igual a la media, varianza y desvío estándar de la distribución asociada con el histograma de la lista.

- Media

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Varianza

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Desvío estándar

$$\sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Estadístico

Definición

Un **estadístico** es una función de los valores numéricos de las variables que se registrarán en la muestra y, tal vez, de los parámetros poblacionales.

Ejemplos de estadísticos

- promedio entre las unidades de la muestra de los valores de un atributo (media poblacional del atributo).
- varianza entre las unidades de la muestra de los valores de un atributo.

Estimador

Un **estimador de un parámetro poblacional desconocido** es la función de los datos de la muestra que se empleará para estimar al parámetro poblacional.

- Un estimador es un estadistico. Pero es un estadistico que **no** puede depender de los parámetros poblacionales desconocidos.
- Un estimador es una variable o vector aleatorio porque es una función de las v.a. $X_1, \ldots, X_n$

El **valor estimado de un parámetro poblacional desconocido** es el valor que toma el estimador del parámetro una vez que se han recogido los datos de la muestra.

- El valor estimado no es una variable aleatoria, porque es una funcion de $x_1, \ldots, x_n$.

Muestreo aleatorio simple

Definición

Un **muestreo aleatorio simple** es una estrategia de muestreo al azar en la que todos los subconjuntos de igual tamaño la población tienen la misma probabilidad de ser la muestra seleccionada.

Muestreo aleatorio simple

Theorem

En un muestreo aleatorio simple, las X_i son igualmente distribuidas, es decir:

$$f_{X_1}(x) = f_{X_2}(x) = \cdots = f_{X_n}(x) \text{ si cada } X_i \text{ es continua}$$

o

$$p_{X_1}(x) = p_{X_2}(x) = \cdots = p_{X_n}(x) \text{ si cada } X_i \text{ es discreta}$$

Theorem

En un muestreo aleatorio simple

$$P(a < X_i < b) = \text{proporcion de unidades de la población con valores numéricos del atributo } X \text{ mayores que } a \text{ y menores que } b.$$

Propiedades de la muestra en un muestreo aleatorio simple

Theorem

En un muestreo aleatorio simple, si el tamano de la muestra n es muy inferior al tamano N de la población, entonces

$$X_1, \ldots, X_n$$

son aproximadamente mutuamente independientes. Si $N = \infty$ entonces $X_1, \ldots, X_n$ son mutuamente independientes.

De ahora en más supondremos que los datos provendrán de un muestreo aleatorio simple con N mucho mayor que n, y abusando de la aproximación asumiremos que $X_1, \ldots, X_n$ son i.i.d.

Parámetros poblacionales

Supongamos que el atributo es escalar. Designemos al conjunto de los valores del atributo en la población como $\mathcal{U} = \{x_1, \ldots, x_N\}$.

- **Media Poblacional**

$$m = \frac{1}{N} \sum_{j=1}^{N} x_j$$

- **Varianza poblacional**

$$v = \frac{1}{N} \sum_{j=1}^{N} (x_j - m)^2$$

- Parámetros de la distribución del valor X_i a registrar en la unidad i:

$$\mu_X = E(X_i) \quad \text{y} \quad \sigma_X^2 = var(X_i)$$

Theorem

En un muestreo aleatorio simple, se verifica que $\mu_X = m$ y $\sigma_X^2 = v$.

Media muestral

Definición

Sean $X_1, \ldots, X_n$ las variables o vectores aleatorios que representan los datos numéricos a recoger en una muestra.

La **media muestral** se define como el promedio de los datos numéricos que se registrarán en la muestra, esto es

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

A la distribución de $p_{\bar{X}}(x)$ o $f_{\bar{X}}(x)$ de $\bar{X}$ se la llama **distribución muestral de la media muestral**.

Media muestral

Ejemplos

- Media de los ingresos de los hogares de la muestra. (X_i = ingreso del hogar i)
- Proporción de votantes de la muestra decididos a votar al candidato A ($X_i = 1$ si el votante i votará al candidato A, $X_i = 0$ si no lo votará)

OJO: no confundir los siguientes conceptos!

- **la media muestral** $\bar{X}$ (una variable aleatoria)
- **la media poblacional** m (un valor fijo)
- **la media** $\bar{x}$ **de los valores del atributo observados en la muestra** ($\bar{x}$ es un valor observado, y por lo tanto, no aleatorio)

Ejemplo

- Población de tamaño $N = 4,000,000$
- Muestra aleatoria simple de tamaño $n = 2$
- X_i = edad del individuo $i, i = 1, 2$.
- Posibles edades en los miembros de la población: 18, 20, 22, 24 (años)
- Igual número de personas de cada edad en la población

Ejemplo

1. **Distribución conjunta de (X_1, X_2)**

$$p_{X_1 X_2}(x_1, x_2) = p_{X_1}(x_1)\, p_{X_2}(x_2) = \frac{1}{4} \times \frac{1}{4} = \frac{1}{16} \text{ para todo } (x_1, x_2)$$

2. **Enumeración de los posibles valores de la media muestral**

		edad de la 2da unidad			
		18	**20**	**22**	**24**
edad	**18**	18	19	20	21
de	**20**	19	20	21	22
la	**22**	20	21	22	23
1ra unidad	**24**	21	22	23	24

3. **Distribución muestral de la media muestral**

x	**18**	**19**	**20**	**21**	**22**	**23**	**24**
$p_{\bar{X}}(x)$	1/16	2/16	3/16	4/16	3/16	2/16	1/16

Propiedades de la media muestral

De ahora en mas usaremos la siguiente notación:

$$\mu_{\bar{X}} = E\left(\bar{X}\right) \text{ y } \sigma^2_{\bar{X}} = var\left(\bar{X}\right)$$

De las propiedades sobre la esperanza y la varianza de sumas de variables aleatorias que vimos en clases anteriores podemos deducir los siguientes resultados para $X_1, \ldots, X_n$ v.a. iid.

Theorem

La esperanza de $\bar{X}$ es igual a la media poblacional, es decir:

$$\mu_{\bar{X}} = \mu_X$$

Theorem

La varianza de $\bar{X}$ es igual a la varianza poblacional dividido n,

$$\sigma^2_{\bar{X}} = \frac{1}{n}\sigma^2_X$$

Demostraciones

Esperanza.

$$\mu_{\bar{X}} = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right)$$
$$= \frac{1}{n} nE\left(X_1\right) = E\left(X_1\right) = E\left(X_i\right)$$

□

Varianza.

$$\sigma^2_{\bar{X}} = var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \left(\frac{1}{n}\right)^2 var\left(\sum_{i=1}^{n} X_i\right) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} var\left(X_i\right)$$
$$= \left(\frac{1}{n}\right)^2 \cdot n \cdot \ var\left(X_1\right) = \frac{1}{n} var\left(X_1\right) = \frac{1}{n} var\left(X_i\right)$$

□

Distribución asintótica de la media muestral

Theorem

Si $X_1, \ldots, X_n$ son v.a. i.i.d., el histograma de la distribución de $\bar{X}$

1. *está centrado en la media poblacional μ_X para cualquier n*
2. *tiene varianza que tiende a 0 a medida que $n \to \infty$*

Theorem (Ley de los Grandes Numeros (LGN))

Si $X_1, \ldots, X_n$ son v.a.i.i.d, cada una con esperanza μ_X, para cualquier $\varepsilon > 0$ vale que

$$P\left(\left|\bar{X} - \mu_X\right| < \varepsilon\right) \underset{n \to \infty}{\to} 1$$

Interpretacion de la LGN

Si una muestra aleatoria simple es rande, entonces con muy alta probabilidad la media muestral caerá cerca de la media poblacional.

- Si tomamos una muestra aleatoria simple grande, puede que la media muestral no caiga cerca de la media poblacional.
- Sin embargo si repetimos muchas veces, digamos $M = 100,000$ veces, el experimento de tomar una muestra aleatoria, en la mayoria de los M experimentos la media muestral caera cerca de la media poblacioal, porque la probabilidad de que $\bar{X} \approx \mu_X$ es muy alta.