Lección 1: Distribuciones muestrales

Módulo 3: Inferencia Estadística

Magdalena Cornejo

Estimadores puntuales

- Queremos hacer inferencia respecto de una población, basados en la información contenida en una muestra aleatoria.

Estimadores puntuales

- Queremos hacer inferencia respecto de una población, basados en la información contenida en una muestra aleatoria.
- Nos centraremos en características específicas, **parámetros**, de la población. Por ejemplo:
  - El nivel medio de consumo de cierto producto.
  - La proporción de mujeres con cargos jerárquicos en una empresa.

Estimadores puntuales

- Queremos hacer inferencia respecto de una población, basados en la información contenida en una muestra aleatoria.
- Nos centraremos en características específicas, **parámetros**, de la población. Por ejemplo:
  - El nivel medio de consumo de cierto producto.
  - La proporción de mujeres con cargos jerárquicos en una empresa.
- Cualquier inferencia sobre la población estará basada en **estadísticos** muestrales.

Definición

Un **estadístico** es una función de la información muestral. La elección del estadístico adecuado dependerá en cuál es el parámetro poblacional de interés.

Estimadores puntuales

Por sus propiedades, ciertos estadísticos son generalmente preferidos como estimadores de la media poblacional (μ) o la proporción poblacional (p)

Parámetro	**Estimador**
μ	$\overline{X} = \frac{\sum_i X_i}{n}$
p	$\hat{p} = \frac{\text{nro éxitos}}{n}$

Muestral vs Poblacional

Todo concepto de que vimos en Estadística Descriptiva tiene su correlato poblacional reemplazando frecuencias observadas por probabilidades:

Muestral	**Poblacional**
Frecuencia	Probabilidad
Histograma	Distribución de probabilidades
Media muestral ($\overline{X}$)	Media poblacional (μ)
Varianza muestral (S^2)	Varianza poblacional (σ^2)
Proporción muestral ($\hat{p}$)	Proporción poblacional (p)
Covarianza muestral (S_{XY})	Covarianza (COV(X,Y) o σ_{XY})
Coeficiente de correlación (r_{XY})	Coeficiente de correlación (ρ_{XY})

Distribuciones muestrales

Definición

Una **distribución muestral** es la distribución de probabilidad de un estadístico.

- ¡Sí! ¡Los estadísticos (como $\overline{X}$) son también variables aleatorias!
- Entonces, podemos estar interesados en conocer su distribución.
- Pensemos en la media muestral ($\overline{X}$)...

Media muestral ($\overline{X}$)

- Depende de la muestra, ¡que es aleatoria!
- Por ser una variable aleatoria tiene una media y una varianza.
- Si se extrae una muestra aleatoria de una población con media μ y varianza σ^2. La media y la varianza de $\overline{X}$ serán:

$$E(\overline{X}) = E\left(\frac{\sum_i X_i}{n}\right) = \frac{1}{n}E\left(\sum_i X_i\right) = \frac{1}{n}n\mu = \mu$$

$$V(\overline{X}) = V\left(\frac{\sum_i X_i}{n}\right) = \frac{1}{n^2}V\left(\sum_i X_i\right) = \frac{1}{n^2}n\sigma^2 = \sigma^2/n$$

- **Nota**: para el cálculo de la varianza tengan en cuenta que $X_1, X_2, ..., X_n$ son independientes porque vienen de una muestra aleatoria.

Distribución muestral de la media

Entonces, $\overline{X}$ tiene:

$$E(\overline{X}) = \mu$$
$$V(\overline{X}) = \frac{\sigma^2}{n}$$

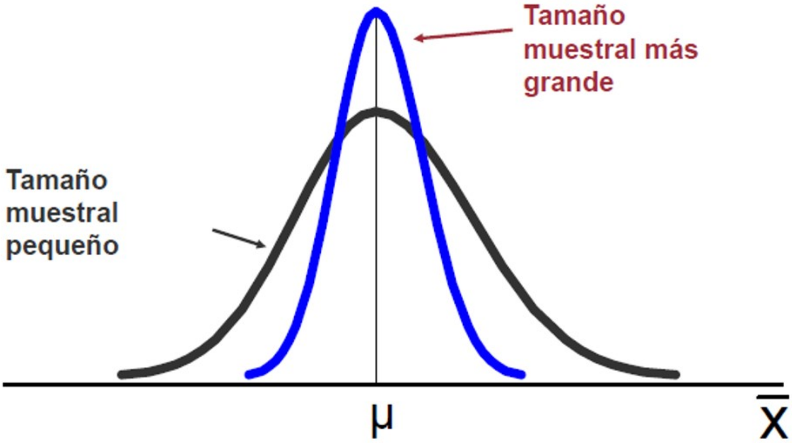
El desvío estándar será:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

que en algunas ocasiones recibe el nombre de **error estándar** de la media (porque es un estimador).

El "n" importa

- A medida que n aumenta, $\sigma_{\overline{X}}$ disminuye:



Distribución muestral de la media

Ya conocemos la media (μ) y la varianza (σ^2/n) de la media muestral ($\overline{X}$). Ahora queremos conocer su distribución.

- Si la población tiene distribución normal, entonces ¡$\overline{X}$ también!
- Ya que la suma (finita) de variables aleatorias normales también es normal.
- Si la estandarizamos:
$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$
- tiene una **distribución normal estándar**, $N(0, 1)$.

Teorema Central del Límite (TCL)

Un famoso teorema matemático (el **teorema central del límite**) mostró que si el tamaño de la muestra es considerablemente grande, entonces la distribución de la media muestral (de n variables aleatorias IID) se aproxima a una distribución normal, aunque la población no fuera normal.

A medida que n se incrementa, la distribución muestral de la media muestral estandarizada se aproxima a una distribución normal con media 0 y varianza 1.

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

Poniendo a prueba el TCL

- En Excel genere 5 variables aleatorias con distribución uniforme (0,1) con 1000 datos (para cada una) usando la herramienta de análisis de datos (`Analysis ToolPak`).
- Haga el histograma para una de estas variables.
- Genere una nueva variable aleatoria que sea el promedio de las anteriores.
- Haga el histograma para esta nueva variable. ¿Qué observa?

- Repita el experimento generando 10000 variables y 100000 datos. ¿Qué forma va tomando el histograma?

La ley de los grandes números

Esta ley afirma que el promedio de variables aleatorias independientes con una distribución común va a converger a la media de la distribución a medida que el tamaño crece.

Es decir, de acuerdo a esta ley, el promedio de los resultados obtenidos de una larga serie de repeticiones se acerca al valor esperado.

Ley de los Grandes Números: Si $X_1, X_2, ...$ son variables aleatorias independientes idénticamente distribuidas (IID), cada una con media $E(X) = \mu$. Entonces,

$$\overline{X} = \frac{\sum_i X_i}{n} \to \mu \text{ cuando } n \to \infty$$

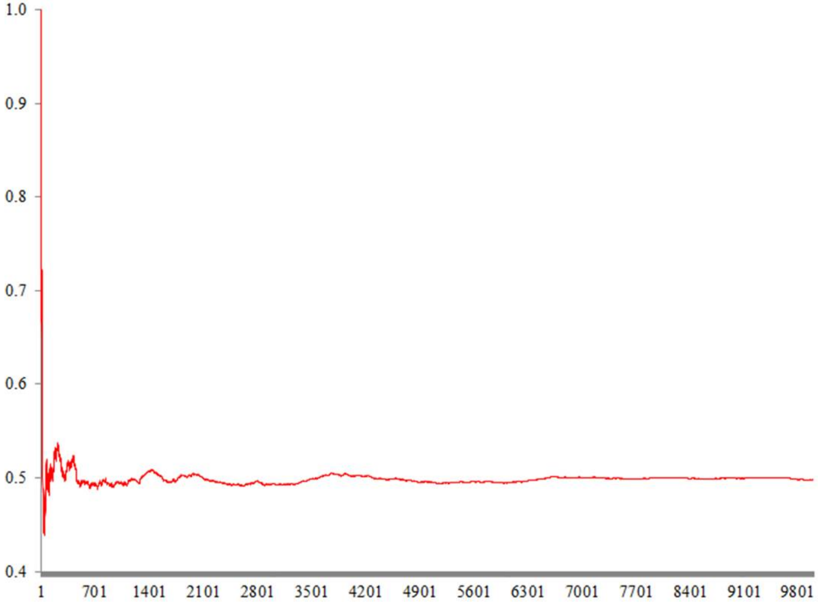
La ley de los grandes números

Ejemplo

Simulamos un experimento en el que arrojamos una moneda.
Sabemos que la probabilidad de que salga cara o ceca es $1/2$, la siguiente tabla muestra cómo la probabilidad se acerca a 0.5 a medida que aumentamos la cantidad de repeticiones del experimento:

N	**Cara**	**Ceca**	**P(Cara)**	**P(Ceca)**
10	7	3	0.70	0.30
50	23	27	0.46	0.54
100	49	51	0.49	0.51
500	247	253	0.494	0.506
1000	495	505	0.495	0.505
10000	4980	5020	0.498	0.502

La ley de los grandes números



Proporción muestral

- Sea X es el número de éxitos en una muestra de n observaciones (independientes entre sí), donde p es la probabilidad de éxito.
- Vimos que $\hat{p}$ (la **proporción muestral**) es:

$$\hat{p} = \frac{X}{n}$$

- Al igual que con la media muestral, la proporción muestral es un estadístico y por lo tanto tendrá asociado una distribución.

Distribución muestral de la proporción muestral

- La media y la varianza de la distribución muestral de la proporción muestral pueden deducirse fácilmente de la media y la varianza del número de éxitos (X) que vimos que eran:

$$E(X) = np \qquad V(X) = np(1-p)$$

- Entonces:

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = p \qquad V(\hat{p}) = V\left(\frac{X}{n}\right) = \frac{p(1-p)}{n}$$

- El desvío estándar de la proporción muestral se conoce como **error estándar**:

$$\sqrt{\frac{p(1-p)}{n}}$$

Distribución muestral de la proporción muestral

- Si el tamaño de la muestra (n) es lo suficientemente grande entonces:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

- Es decir, ¡la distribución binomial se aproxima a una distribución normal!

Video: La Máquina de Galton

A modo de conclusión

- $\overline{X}$ y $\hat{p}$ son estadísticos y, por lo tanto, variables aleatorias.
- Entonces, tienen una distribución asociada (**distribución muestral**).
- Necesitamos conocer dichas distribuciones para poder hacer **inferencia estadística**.