Inferencia Estadística

Juan Sosa, PhD

Universidad Externado de Colombia

I - 2018

Objetivo

- **Caracterizar una población** con base en la información de una **muestra representativa** de la población.
- **Recolectar datos** para **decrecer** la **incertidumbre** acerca de cantidades poblacionales desconocidas (**parámetros**).

Introducción

- Una **población** consiste de todos los **entes de interés**. ¿Por qué la mayoría de las veces no es posible hacer un **censo**?
- Escoger la muestra de manera que los elementos escogidos sean **similares** a los de la población en todas aquellas **características que son relevantes** para el estudio.
- La generalización se hace sobre aquellos elementos que satisfacen tales características de similaridad (e.g., **estudios observacionales**).
- El enemigo en la recolección de datos es el **sesgo** (tendencias sistemáticas hacia algo en particular).
- La inferencia estadística **no es comprobable directamente**, a diferencia los **modelos predictivos**.

Inferencia frecuentista basada en el modelo

Interpretación frecuentista de la probabilidad (J. Venn, R. von Mises)

La **probabilidad** de un evento A se define como la **frecuencia relativa** de su ocurrencia en **repeticiones sucesivas** (repeticiones hipotéticas o reales) del fenómeno de interés:

$$\mathbb{P}\text{r}\left[A\right] = \lim_{n\to\infty} \frac{\#\text{ de veces } A \text{ ocurre}}{n}$$

Modelo probabilístico/estocástico (R. Fisher)

Los datos son producto de un **mecanismo aleatorio** indexado por un conjunto de **parámetros**. Tal mecanismo caracteriza **cómo surgen los datos**.
Ej. Distribución Bernoulli, Distribución Normal, Distribución Exponencial, etc.

"All models are wrong, but some are useful" (G. Box)

Objetivo

Hacer inferencia sobre los parámetros del modelo:

- Estimación puntual.
- Intervalos de confianza.
- Pruebas de hipótesis.

Proporción poblacional

Modelo

El enfoque frecuentista se basa en la idea de **repeticiones hipotéticas o reales** del proceso que se estudia, bajo condiciones que son lo más cercanas posibles al muestro **Independiente e Idénticamente Distribuido** (IID):

$$X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim} \text{Ber}(\theta)$$

- Resultados de un experimento aleatorio.
- Muestra de una población "infinita" sin reemplazo.

Estimador

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

Propiedades del promedio muestral $\bar{X}$

$$\mathbb{E}\left[\bar{X}\right] = \mu \qquad \mathbb{V}\text{ar}\left[\bar{X}\right] = \frac{\sigma^2}{n}$$

donde $\mu = \mathbb{E}\left[X\right]$ es el **promedio poblacional** y $\sigma^2 = \mathbb{V}\text{ar}\left[X\right]$ es la **varianza poblacional**.

Ley de los grandes números (D. Bernoulli, 1713; S. D. Poisson, 1837)

Si $X_1, X_2, \ldots$ es una sucesión infinita de variables aleatorias independientes que tienen el mismo valor esperado μ y varianza σ^2, entonces el promedio $\boldsymbol{\bar{X}}$ **converge en probabilidad a** $\boldsymbol{\mu}$. En otras palabras, para cualquier número positivo ϵ se tiene que

$$\lim_{n\to\infty} \mathbb{P}\mathsf{r}\left[|\bar{X} - \mu| < \epsilon\right] = 1.$$

Teorema del Límite Central (TLC, Lindeberg–Lévy)

Sea $X_1, X_2, \ldots, X_n$ un conjunto de variables aleatorias independientes e idénticamente distribuidas de una distribución con media μ y varianza $\sigma^2 < \infty$. **Para** $\boldsymbol{n}$ **"grande", se tiene que el promedio muestral** $\boldsymbol{\bar{X}}$ **aproximadamente (asintóticamente) sigue una distribución Normal.** Esto es,

$$\bar{X} \overset{\mathsf{A}}{\sim} \mathsf{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

- El TLC solo dice algo sobre la **distribución muestral** de $\bar{X}$, no sobre la distribución de X en sí.
- El TLC dice que la distribución muestral de $\bar{X}$ es aproximadamente normal cuando n es "grande". ¿Cuándo es n grande? ¿Qué tan buena es esta aproximación?

Inferencia sobre la proporción poblacional (Neyman)

Modelo probabilístico

$$X_1, X_2, \ldots, X_n \stackrel{\textsf{IID}}{\sim} \textsf{Ber}(\theta)$$

Intervalo de confianza

Dado que

$$Z = \frac{\hat{\theta} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \stackrel{\textsf{A}}{\sim} \textsf{N}(0,1)$$

de acuerdo con el **Teorema del Límite Central**, se obtiene que un **intervalo de confianza para θ usando una confiabilidad del $100(1-\alpha)\%$** es

$$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

donde $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$ es la estimación puntual de θ y $z_{1-\alpha/2}$ es el percentíl $100(1-\alpha/2)$ de la distribución Normal estándar.

Nota: Empíricamente se ha visto que este intervalo suele ser “apropiado” cuando $n \geq 30$, $n\hat{\theta} \geq 5$ y $n(1-\hat{\theta}) \geq 5$.

Modelo probabilístico

$$X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim} F$$

donde F es **cualquier distribución** con media μ y varianza $\sigma^2 < \infty$.

Intervalo de confianza

Dado que

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \overset{\text{A}}{\sim} \mathsf{N}(0,1)$$

de acuerdo con el **Teorema del Límite Central**, se obtiene que un **intervalo de confianza para μ usando una confiabilidad del $100(1-\alpha)\%$** es

$$\bar{x} \pm z_{1-\alpha/2} \, \frac{s}{\sqrt{n}}$$

donde $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ y $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ son las estimaciones puntuales de μ y σ, respectivamente, y $z_{1-\alpha/2}$ es el percentíl $100(1-\alpha/2)$ de la distribución Normal estándar.