Inferencia

Alejandro Zubiri

Fri Nov 08 2024

Contents

La inferencia es el estudio de lo poco que conocemos para sacar conclusiones de lo mucho que desconocemos.

1 Estimación

Sirve para hacer prediciones sobre parámetros de la población.
El estimador de un parámetro θ es una función de los valores de la muestra para estimar θ. Un estimador se denota por $\hat{\theta}$

$$\hat{\theta} = g(x_1, \ldots, x_n) \tag{1}$$

Cuanto más aproximado es un estimador al próximo parámetro, decimos que es **insesgado**. Si un estimador lo es se verifica que

$$E(\hat{\theta}) = \theta \tag{2}$$

Donde $E(\hat{\theta})$ es la esperanza (media) del estimador.
Una de las propiedades es que:

$$E(\hat{\theta}) = \theta + b(\theta) \tag{3}$$

Donde $b(\theta)$ es el sesgo ("error").
También tenemos la **varianza mínima** o **precisión**, que es la proximidad entre muestras repetidas. Calculamos el error de muestreo como:

$$e(\hat{\mu}) = \frac{\sigma}{\sqrt{n}} \tag{4}$$

Y finalmente la **eficiencia**, que es un estimador insesgado de varianza mínima:

$$efic(\hat{\theta}) = \frac{1}{var(\hat{\theta})} \tag{5}$$

Un ejemplo de estimador eficiente y lineal es el **ELIO**.
El criterior para un estimador es el **error cuadrático medio** (ECM):

$$ECM(\hat{\theta}) = E((\hat{\theta} - \theta)^2) = var(\hat{\theta}) + b^2(\hat{\theta}) \tag{6}$$

Distribución muestral. Distribución de probabilidad de un estadístico □

1.1 Varianza conocida

Sea una variable aleatoria:

$$E(\bar{X}) = \frac{E(X_1) + E(X_2) + \cdots + E(X_n)}{n} = \frac{\mu + \mu + \cdots + \mu}{n} = \frac{n\mu}{n} = \mu \tag{7}$$

$$var(\bar{X}) = \frac{var(X_1 + X_2 + \cdots + X_n)}{n^2} = \frac{\sigma^2 + \cdots + \sigma^2}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \tag{8}$$

Teorema central del límite. Sean $x_1, x_2, \ldots, x_n$ variables independientes idénticamente distribuidas, con media μ_i y con desviación típica σ_i:

$$\frac{\sum x_i - \sum \mu_i}{\sqrt{\sum \sigma^2}} \to N(0,1) \tag{9}$$

Si n es grande ($n > 30$), es asintóticamente normal:

$$\bar{X} \to N(\mu, \frac{\sigma}{\sqrt{n}}) \tag{10}$$

Si la población ya es normal, se cumple siempre □

1.2 Varianza desconocida

Si la varianza es desconocida, usamos la varianza muestral $\hat{S}^2$:

$$\frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}} \to t_{n-1} \tag{11}$$

Donde t es la t de Student.

1.3 Distribución muestral de proporción

Si p es la proporción de individuos que cumplen una determinada característica:

$$\hat{p} = \frac{\sum x_i}{n} = \bar{X} \tag{12}$$

Donde x_i es aleatoria y toma 0 o 1.
Si $n > 30$ y $np(1-p) > 5$, aplicamos el teorema central del límite:

$$\frac{p - \hat{p}}{\sqrt{\hat{p}(1-\hat{p})/n}} \approx N(0,1) \tag{13}$$

1.4 Distribución muestral de la varianza

$$\hat{S}^2 = \frac{\sum x_i - \bar{x}}{n-1} \tag{14}$$

Es el estimador insesgado de varianza poblacional. Cuando la muestra viene de una distribución normal:

$$\frac{S^2}{\sigma^2} \approx \frac{\chi^2_{n-1}}{n-1} \qquad \text{Lema de Fisher-Cochran}$$

1.5 Intervalos de confianza

Queremos la media μ de una población, **pero no podemos**. Como alternativa, calculamos la media con incertidumbre (probabilidad):
Para el parámetro θ con nivel de confianza $1-\alpha$, es el intervalo $\theta_1(X), \theta_2(X)$ tal que:

$$P(\theta_1(X) \leq \theta \leq \theta_2(x)) = 1 - \alpha \tag{15}$$

Donde α es el nivel de significación.

1.5.1 Cálculo de los intervalos - media con σ conocida

Necesitamos un pivote:

- Que incluya lo que queremos calcular: μ
- Que incluya lo que conocemos: $\bar{x}$
- Una distribución conocida

$$\bar{x} \rightarrow N(\mu, \frac{\sigma}{\sqrt{n}}) \tag{16}$$

1.5.2 Cálculo de intervalos - media con σ desconocida

Si n es grande:

$$\bar{x} \pm Z_{1-\frac{\alpha}{2}} \frac{S_x}{\sqrt{n}} \tag{17}$$

Si n es pequeña pero dada por una distribución normal:

$$IC(\mu_x) = (\bar{x} \pm t_{n-1,1-\frac{\alpha}{2}}) \cdot \frac{S_x}{\sqrt{n}} \tag{18}$$

Para definir $Z_{1-\frac{\alpha}{2}}$:

$$p(Z > Z_{1-\frac{\alpha}{2}}) = \frac{\alpha}{2} \tag{19}$$

$$p(Z < Z_{1-\frac{\alpha}{2}}) = \frac{\alpha}{2} \tag{20}$$

$$p(-Z_{1-\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu_x}{\sigma_x/\sqrt{n}} \leq Z_{1-\frac{\alpha}{2}}) = 1 - \alpha \tag{21}$$

$$IC_{1-\alpha}(\mu_x) = (\bar{x} \pm Z_{1-\alpha} \frac{\sigma_x}{\sqrt{n}}) \tag{22}$$

1.6 Para una proporción

Nuestra media muestral será nuestra **media muestral**: $\bar{x} = \hat{p}$. Y

$$\hat{\sigma}^2 = p(1-p) \tag{23}$$

Entonces, en caso de contar con una muestra grande,

$$IC(p_x) = (\hat{p}_x \pm Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) \tag{24}$$

2 Varianza en poblaciones normales

Contando con el teorema de Fisher-Cochran:

$$\frac{(n-1)S^2}{\sigma^2} \approx \chi^2_{n-1} \tag{25}$$

Llamamos a $\chi^2_{(n-1,1-\alpha/2)}$ y $\chi^2_{(n-1,\alpha/2)}$ los valores que dejan una probabilidad de $1-\alpha$:

$$p(\frac{(n-1)S^2}{\chi^2_{(n-1,1-\alpha/2)}} \geq \sigma^2 \geq \frac{(n-1)S^2}{\chi^2_{(n-1,\alpha/2)}}) = 1-\alpha \tag{26}$$

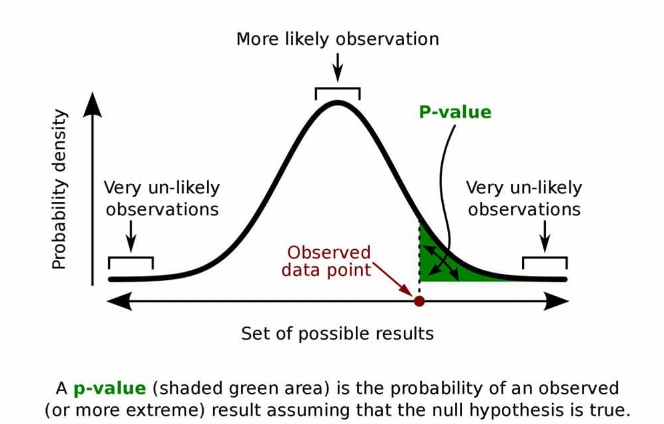
3 Contraste de hipótesis

Tenemos dos alternativas:

- H_0: nula o neutra. Se contraste, y la mantenemos a menos que se demuestre lo contrario, sino se acepta.
- H_1: alternativa. Se acepta si se deniega la neutra.

3.1 Método

La hipótesis nula es verdadera hasta que se demuestra lo contrario.
Una discrepancia grande tiene poca probabilidad de ocurrir si la hipótesis es nula:



Hay dos tipos de errores:

- Rechazar hipótesis verdadera: tipo I (α)
- Aceptar hipótesis falsa: Tipo II

Y definimos α como:

$$\alpha = p\left(\frac{\text{Rechazar } H_0}{H_0 \text{ cierta}}\right) \tag{27}$$

3.2 Valor p

Es la región de rechazo, el **nivel crítico** (p-valor).

p-valor. Probablidad de obtener discrepancia mayor o igual que la observada si H_0 es cierta.
$p < 0.05$: poca evidencia, la rechazamos. □

3.3 Cálculo de contrastes

3.3.1 Media con σ conocida

Contrastamos la hipótesis, sabiendo que la media de la distribución es μ_0 **Primer caso**:

$$\begin{aligned} H_0 : &\mu = \mu_0 \\ H_1 : &\mu \neq \mu_0 \end{aligned} \tag{28}$$

Sabemos que:

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \approx N(0,1) \tag{29}$$

Por tanto, rechazamos H_0 si:

$$\begin{aligned} Z &> Z_{1-\alpha/2} \\ Z &< -Z_{1-\alpha/2} \end{aligned} \tag{30}$$

Segundo caso:

$$\begin{aligned} H_0 &: \mu \leq \mu_0 \\ H_1 &: \mu > \mu_0 \end{aligned} \tag{31}$$

En este caso, rechazaremos la hipótesis si:

$$\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} > Z_{1-\alpha} \tag{32}$$

Tercer caso:

$$\begin{aligned} H_0 &: \mu \geq \mu_0 \\ H_1 &: \mu < \mu_0 \end{aligned} \tag{33}$$

La rechazaremos si:

$$\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} < -Z_{1-\alpha} \tag{34}$$

3.3.2 Media con σ desconocida

Partiendo de una población que debe ser normal:

$$\frac{\bar{x}-\mu}{\hat{S}/\sqrt{n}} \approx t_{n-1} \tag{35}$$

Rechazamos la hipótesis nula H_0 si:

$$\begin{aligned} \frac{\bar{x}-\mu}{\hat{S}/\sqrt{n}} &> t_{n-1,1-\alpha/2} \\ \frac{\bar{x}-\mu}{\hat{S}/\sqrt{n}} &< t_{n-1,\alpha/2} \end{aligned} \tag{36}$$

3.3.3 Contraste para una proporción

Ya sabemos que para una proporción:

$$\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/\sqrt{n}}} \approx N(0,1) \tag{37}$$

si la muestra es suficientemente grande. En estos casos, rechazaremos la hipótesis nula si:

$$\begin{aligned} \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/\sqrt{n}}} &> Z_{1-\alpha/2} \\ \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/\sqrt{n}}} &< -Z_{1-\alpha/2} \end{aligned} \tag{38}$$

3.3.4 Contraste para la varianza σ^2

Ya sabemos que para la varianza:

$$\frac{(n-1)S^2}{\sigma_0^2} \approx \chi^2_{n-1} \tag{39}$$

En caso de partir de una población normal. En este caso, rechazamos la hipótesis si:

$$\begin{aligned} \frac{(n-1)S^2}{\sigma_0^2} &> \chi^2_{n-1,1-\alpha/2} \\ \frac{(n-1)S^2}{\sigma_0^2} &< \chi^2_{n-1,\alpha/2} \end{aligned} \tag{40}$$

3.3.5 Contraste para el tamaño muestral

El tamaño muestral es el número de elementos del universo que se seleccionan para cada muestra.
Para aproximar la media, tenemos que:

$$\frac{rango}{4} \approx \sigma \tag{41}$$

Aunque la σ real siempre será menor. Para el caso de proporciones, siempre se asume el caso más desfavorable: $p = q = 0.5$.
Se considera **universo grande** cuando:

$$\begin{aligned} N &\geq 100n \\ N &\geq 20n \end{aligned} \tag{42}$$

Para calcular el tamaño de la muestra, este será:

$$\boxed{n = \frac{Z^2_{1-\frac{\alpha}{2}}\sigma^2}{e^2}} \tag{43}$$

Esto se aplica también para una proporción, donde

$$\sigma^2 = \hat{p}(1-\hat{p}) \tag{44}$$

Para el caso de **universo pequeño**, debemos aplicar la corrección:

$$\frac{N-n}{N-1} \tag{45}$$

Donde N es el tamaño de la población y n el de la muestra. Con esta corrección, el tamaño sería

$$\boxed{n = \frac{NZ^2_{1-\alpha/2}\sigma^2}{(N-1)e^2 + Z^2_{1-\alpha/2}\sigma^2}} \tag{46}$$

4 Resumen de intervalos

- Media con varianza: $Z_{1-\alpha/2}$
- Media sin varianza:
  - Con DNORM: t_{n-1}
  - Con n grande: $Z_{1-\alpha/2}$
- Proporción con n grande: $Z_{1-\alpha/2}$
- Varianza con DNORM: $\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}$

5 Resumen de contrastes

- Media con varianza: $Z_{1-\alpha/2}$
- Media sin varianza con DNORM: $t_{n-1,1-\alpha/2}$
- Proporción con n grande: $Z_{1-\alpha/2}$
- Varianza con DNORM: $\chi^2_{n-1,1-\frac{\alpha}{2}}$

6 Afijación

Es la distribución del número de individuos por muestra:

6.0.1 Simple

Dividimos el total entre el número de estratos:

$$n_i = \frac{n}{L} \tag{47}$$

donde L es el número de estratos.

6.0.2 Proporcional

Lo ponderamos con lo proporcional que sea cada estrato:

$$n_i = n\frac{N_i}{N} \tag{48}$$

Donde N_i es la población del estrato i.

6.0.3 Óptima o de Neyman

Ponderamos cada estrato por la varianza:

$$n_i = n\frac{N_i\sigma_i}{\sum N_i\sigma_i} \tag{49}$$

7 Error de muestreo para la estimación

- Universo grande:

$$e_x = \frac{\sigma}{\sqrt{n}} \tag{50}$$

- Universo pequeño:

$$e_x = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \tag{51}$$

Esto funcionará igual para la proporción, donde $\sigma^2 = p(1-p)$.