Tema 6: Inferencia estadística y Contraste de Hipótesis

Estadística inferencial

Se trata de predecir, inferir, generalizar parámetros de una población determinada a partir de los datos de una muestra.

Ejemplo: Sondeo de antes de unas elecciones. No se pregunta a toda la población sobre su intención de voto (medios técnicos costosos) En lugar de eso, se realiza un sondeo. se toma una muestra representativa de la sociedad.

Se miden aquellos parámetros en los que estoy interesado. y se intenta estimar o inferir los parámetros poblacionales (media, varianza, proporciones, ...) a partir de los parámetros muestrales.

- En este curso estudiaremos muestreos aleatorios simples (no confundir con muestreo al. con rep)

Parámetros:

Población

Muestra

→ μ = media ⟶ $\bar{x}$ = media muestral

→ σ = desv. típica ⟶ s = cuasidesviación típica

$$s = \frac{n}{n-1} \cdot \sigma_{muestra} \qquad n: \text{tamaño muestra}$$

p = proporción poblacional ⟶ $\hat{p}$ = proporción muestral

$$p = \frac{n_A}{N} \qquad \hat{p} = \frac{\hat{n}A}{n}$$

$\hat{n}A$ = nº de individuos en la muestra que verifica una condición A.

nA = nº de individuos en la población que verifica una condición A

$\bar{x}, s, \hat{p}$ → Estadísticos

Dada una muestra, si quiero inferir parámetros de población, puedo hacer 2 tipos de estimaciones.

- Estimaciones puntuales (por ejemplo, aquí diré que la media μ tomará un valor fijo)
- Estimaciones por intervalos. (por ejemplo aquí diré que la media μ está en un intervalo I. Aquí diré cuál es la probabilidad de que esa media μ se encuentre en I. $p(\mu \in I)$

Estimación puntual

- Dada una población que sigue $N(\mu, \sigma)$ $\quad \bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$
- Dada una población que sigue $N(\mu, \sigma)$ si desconozco σ
  - $\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{n-1}$ ≡ T-Student
    $p(t_{n-1} \geq \alpha)$
  - Si $n \geq 30$
    $$\bar{x} \sim N\left(\mu, \frac{s}{\sqrt{n}}\right)$$

Ejemplo: La altura de una población sigue una distribución normal $N(\mu, \sigma)$

Se tomará una muestra, $s = 20$ cm
$N = 40$ individuos

-) Un estimador puntual para la media μ (poblacional)

$$\bar{x} = 170$$

- Estimador para σ

$$s = 20$$

- Calculo la probabilidad de que en esa muestra, encuentre a alguien que mida más de 180cm.

Sé que $n \geq 30$ y también desconozco σ

luego $\bar{x} \sim N\left(\mu = 170, \frac{s}{\sqrt{n}} = \frac{20}{\sqrt{40}}\right) =$

$$P(\bar{x} > 180)$$

$$P\left(\frac{\bar{x} - 170}{3'162} > \frac{180 - 170}{3'162}\right) = P(Z > 3'162) \simeq 0$$

Estimación por intervalos

La idea será inferir un parámetro θ de la población, dando un intervalo I, al que pertenezca.
Además, podremos dar la probabilidad $P(\theta \in I)$

A esa probabilidad le daré el nombre de nivel de confianza y se denota como:

$$P(\theta \in I) = \underbrace{1 - \alpha}_{\text{nivel de confianza}}$$

A α se le llama nivel de significación

Intervalo de confianza para la media de una normal $N(\mu, \sigma)$

Si conozco σ, el intervalo para la media sería $I = \left[\bar{x} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$

$Z_{\alpha/2}$ es un número tal que

$$P(Z \geq Z_{\alpha/2}) = \frac{\alpha}{2}, \quad Z \sim N(0, 1)$$

Ejercicio:

Un servicio técnico debe atender las averías producidas en un modelo de lavadoras. Se observa que el tiempo empleado para realizar las reparaciones sigue una distribución normal. Se miden los tiempos al azar de 200 reparaciones que resultan ser los siguientes:

$\sum_{i=1}^{200} t_i = 6000$ min. $\quad \sum t_i^2 = 600.000$

Calcular un intervalo de confianza al 90% para el tiempo empleado en la reparación:

μ no la conozco: en realidad es la que quiero estimar.

$\sigma \rightarrow$ no la conozco.

Voy a tratar de inferir μ a partir de los datos:

$$\bar{x} = \frac{\sum t_i}{n} = \frac{6000}{200} = 30$$

$$S = \frac{n}{n-1} \cdot \sigma_{muestral} \qquad \sigma_{muestral} = \sqrt{\frac{\sum_{i=1}^{200} t_i^2}{n} - 30^2} = \sqrt{\frac{600.000}{200} - 900} = \sqrt{2100}$$

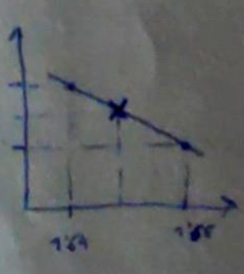
$$\bar{x} = 30 \qquad S = \frac{200}{199} \cdot \sqrt{2100} = 46'056$$

$$I = \left[30 - z_{\alpha/2} \cdot \frac{46'056}{\sqrt{200}} \,,\, 30 + z_{\alpha/2} \cdot \frac{46'056}{\sqrt{200}} \right]$$

$z_{\alpha/2}$: $P(Z \geq z_{\alpha/2}) = \alpha/2$

¿Quién es α? Confianza es del 90% $\quad \alpha = 1 - 0'9 = 0'1 \quad P(Z \geq z_{\alpha/2}) = 0'05$

$P(Z \geq 1'64) = 0'0505 \qquad P(Z \geq 1'65) = 0'0495$



$$z_{\alpha/2} = \frac{1'64 + 1'65}{2} = 1'645$$

$$I = \left[30 - 1'645 \cdot \frac{46'056}{\sqrt{200}} \,,\, 30 + 1'645 \cdot \frac{46'056}{\sqrt{200}} \right]$$

Interpretación del resultado

La media μ de la población se encuentra en $I = [\ldots]$ y lo puedo asegurar al 90% de confianza

Hoy nos vamos a centrar en 3 cuestiones.

1. Hasta ahora, suponíamos que una población sigue una determinada distribución. Ahora lo que haremos será ajustar la distribución a la población, y mediante un contraste de hipótesis, determinar la bondad del ajuste.

2. Si quisiera ver si 2 determinadas muestras pertenecen a una misma población, puedo realizar también un contraste de hipótesis. (Contraste de homogeneidad)

3. Dadas 2 muestras de una población, puedo determinar si son independientes mediante un contraste de hipótesis (Contraste de dependencia o independencia)

Gemplo Bondad del ajuste.

Prob. dada por la distribución $P_o(\lambda)$

x_i	0	1	2	3	4	5	6 o más
O_i	200	220	150	68	25	~~10~~ 14	~~x~~
$P_i = O_i/N$	200/677	220/677	150/677	68/677	25/677	~~10/677~~ 14/677	~~−~~
$\hat{P}_i$	(*) 0'2654	(**) 0'3521	0'2335	0'1032	0'0342		(***) ~~0'0025~~
$N \cdot \hat{P}_i = \hat{O}_i$	179'67	238'34	158'07	69'87	23'17	$N \cdot P(x \geq 5)$	

$\sum O_i = N = 677$

$P_i \simeq \hat{P}_i \rightarrow$ Probabilidad de una distribución (Poisson)

$X \sim P_o(\lambda) \rightarrow$ Depende de un único parámetro λ

El problema de ajustar los datos a una $P_o(\lambda)$ se reduce a ajustar el parámetro λ.

$E(x) = \lambda \Rightarrow$ la mejor forma de estimar λ será tomando la media muestral.

$$\bar{x} = \frac{0 \cdot 200 + 1 \cdot 220 + 2 \cdot 150 + 3 \cdot 68 + 4 \cdot 25 + 5 \cdot 10 + 6 \cdot 5 \cdot 4}{677} = 1'3264 \Rightarrow \lambda = 1'3264$$

(*) $P(x=0) = e^{-\lambda} \frac{\lambda^0}{0!} = e^{-\lambda} \simeq 0'2654$

(**) $P(x=1) = e^{-\lambda} \frac{\lambda^1}{1!} = e^{-\lambda} \lambda^1 \simeq 0'3521$

(***) $P(x \geq 6) = 1 - P(x < 6) = 1 - (P(x=0) + \dots + P(x=5)) \simeq 0'0025$.