Inferencia estadística. Estimación puntual y estimación por intervalos de confianza.

1- **Inferencia estadística** ← Conocer el método estadístico que se usa para obtener conclusiones sobre una variable aleatoria X.

2- **Estimación puntual** ← Estimar los parámetros desconocidos del modelo de probabilidad más adecuado para explicar X. ← Método de los momentos

3- **Estimación por intervalos de confianza** ← Al estimar los parámetros se comete un error. Se trata de que sea el menor posible y acotarlo.

Inferencia estadística

- Trata de obtener conclusiones sobre una variable aleatoria X medida en una población de individuos. Por eso a X se la suele llamar población.

- Para realizar un estudio sobre una variable aleatoria X, se sigue el método estadístico:

Elección de la muestra de X, $x_1, x_2, \dots, x_n$

↓

Estudio descriptivo de la muestra

↓

Proposición modelo para X

↓

Estimación de parámetros del modelo propuesto ← Método de los momentos

↓

Error de estimación de parámetros: intervalos de confianza.

↓

Verificar resultados obtenidos

↙ ↘

Contrastes paramétricos → (vuelve a: Estimación de parámetros del modelo propuesto)

Contrastes no paramétricos hipótesis sobre el modelo para X. → (vuelve a: Proposición modelo para X)

↘ ↙

Toma de decisiones sobre X (si todo es correcto)

Definiciones

Para la variable aleatoria X que se estudia se supone que tiene distribución F_θ, donde:

- F es conocida
- θ es el parámetro(s) desconocido(s)

Muestra aleatoria simple

Sea X una v.a. con distribución F_θ. Llamaremos muestra aleatoria simple de tamaño n a n variables $X_1, X_2, \ldots X_n$ que verifican:

- Son independientes.
- X_i tiene la misma distribución que la población $\boxed{X, \; i = 1, 2, \ldots, n.}$

Estadístico o estimador

Función de la muestra aleatoria simple, $T(X_1, X_2, \ldots, X_n)$ que no dependa del parámetro θ.

$E[X] = \mu(x)$
$V(X) = \sigma(x)$

- Permite estimar el valor del parámetro desconocido θ.
- Un estadístico $T(X_1, X_2, \ldots, X_n) = T$ es una v.a T porque es una función de v.a.

Estadísticos frecuentes

Media muestral $\bar{X} = \frac{1}{n} \sum_1^n X_i$

Varianza muestral $V = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$

Cuasivarianza muestral $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \rightarrow$ Cuasidesviación típica $S = \sqrt{S^2}$

Máximo y mínimo $\begin{cases} \max\{X_1, X_2, \ldots, X_n\} \\ \min\{X_1, X_2, \ldots, X_n\} \end{cases}$

Teorema

Sea X una variable aleatoria con media $E[X]$ y varianza $V(X)$. Sea $X_1, X_2, \dots, X_n$ una muestra aleatoria simple de X. Entonces siempre verifica que:

- Para el estadístico media muestral, $\overline{X} = \frac{1}{n}\sum_{1}^{n} X_i$

  1.- $E[X] = E[\overline{X}]$

  2.- $\frac{V(X)}{n} = V(\overline{X})$

- Para el estadístico cuasivarianza muestral $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

$$V(X) = E[s^2]$$

Estimación puntual

Sea X una v.a. con distribución F_θ

- Datos de X: $x_1, x_2, \dots x_n$.
- Consideramos la muestra aleatoria simple de X asociada a estos datos $X_1, X_2, \dots, X_n$.
- Se considera que $X_1 = x_1$; $X_2 = x_2$; ... ; $X_n = x_n$.

Para estimar θ se usa un buen estimador $T(\overbrace{X_1, X_2, \dots, X_n}^{\text{variables aleatorias}})$ ← Con el método de los momentos

y se asigna a θ el valor numérico $\hat{\theta}$ obtenido $\hat{\theta} = T(\underbrace{x_1, x_2, \dots, x_n}_{\text{datos}})$

este valor numérico se llama estimación de θ

Método de los momentos

Caso 1: El modelo de X depende de 1 solo parámetro desconocido ← $P(\lambda), \dots$

Planteamos una sola ecuación

$$E[X] = \overline{x} \quad ó \quad V(X) = v$$

Caso 2: El modelo de X depende de 2 parámetros desconocidos ← $N(\mu, \sigma), \dots$

Planteamos sistema de dos ecuaciones

$$E[X] = \overline{x} \quad y \quad V(X) = v$$

$\overline{x}$ y v son la media y varianza de los datos.

Distribuciones de probabilidad de algunas variables aleatorias para intervalos de confianza.

Distribución Chi Cuadrado

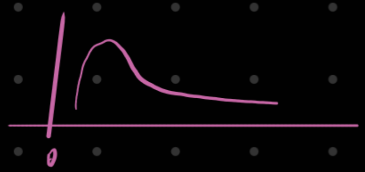
Si $X_1, \dots, X_n$ son v.a. independientes $N(0,1)$, entonces la variable aleatoria

$$Y = \sum_{i=1}^{n} X_i^2 \sim \chi_n^2$$

1) El parámetro n es el número de sumandos que intervienen en la definición.

2) La variable chi-cuadrado es una variable de tipo continuo. Se puede demostrar que $\boxed{\chi_n^2 = \gamma\left(\alpha = \frac{n}{2}, \beta = \frac{1}{2}\right)}$. Por tanto, el gráfico de su función de densidad es:

Una variable aleatoria chi cuadrado con n grados de libertad sigue también una distribución Gamma de forma $\alpha = \frac{n}{2}$ y tasa $\beta = \frac{1}{2}$



Teorema de Fisher

Si $X_1, \dots, X_n$ es una m.a.s (muestra aleatoria simple) de una v.a. $X \sim N(\mu, \sigma)$

1) Los estadísticos o estimadores

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{y} \quad S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{X})^2$$ son variables aleatorias independientes.

2) La distribución de la variable aleatoria U es

$$U = \frac{(n-1) S^2}{\sigma^2} \sim \chi_{n-1}^2$$

Distribución t de Student

Si $X \sim N(0,1)$, $Y \sim \chi_n^2$ son variables independientes, entonces la variable aleatoria Z definida como

$$Z = \frac{X}{\sqrt{\frac{Y}{n}}}$$

sigue una distribución t de student de parámetro n $Z \sim t_n$.

El parámetro de la t_n es el parámetro de la χ_n^2 que interviene en la definición. La distribución t de student es continua, toma valores en todo $\mathbb{R}$, es simétrica respecto del 0 y su gráfica tiene la misma forma que la $N(0,1)$.

$$t_n \approx N(0,1), \; n \geq 120$$

Teorema

Si $X_1, \ldots, X_n$ es una muestra aleatoria simple de una v.a. X con distribución $N(\mu, \sigma)$, entonces, la distribución de la variable aleatoria U es:

$$U = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad y \quad S = \sqrt{S^2} \quad \text{siendo } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

Cuasidesviación típica

Estimación por intervalos de confianza.

- Sea X una v.a. con distribución F_θ conocida donde θ es el parámetro desconocido. Sea $\{X_1, X_2, \ldots, X_n\}$ una muestra simple aleatoria de X.
- Disponemos de $\hat{\theta} = T(X_1, X_2, \ldots, X_n)$ un estimador puntual para θ, obtenido por el método de los momentos.

Objetivo en intervalos

Encontrar una cota del error, K, cometido al estimar θ mediante $\hat{\theta}$. Esta cota K vendrá dada en términos de probabilidad; queremos encontrar K tal que:

$$|\theta - \hat{\theta}| < K \quad \text{tal que } P(|\theta - \hat{\theta}| < K) \text{ sea alta.}$$

Obtener cota de error, K, es equivalente a obtener un intervalo que contenga al parámetro θ.

$$P(|\theta - \hat{\theta}| < K) = P(\theta \in (\hat{\theta} - K, \hat{\theta} + K))$$

Definición

Sea una v.a. $X \sim F_\theta$. Sea $X_1, \dots, X_n$ una muestra aleatoria simple de X. Llamaremos intervalo de confianza para θ con nivel de confianza $1-\alpha$ a un intervalo (a,b) tal que:

$$P(a < \theta < b) = P(\theta \in (a,b)) = 1-\alpha$$

- Los extremos del intervalo a y b son variables aleatorias porque dependen de la muestra aleatoria simple tomada para X.

		Mejor estimador para el parámetro
$\mu = E[X]$	$X \sim N(\mu, \sigma)$ ó X cualquiera y $n \geq 100$ (σ^2 desconocida)	$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$
$\sigma^2 = V(X)$ (ó σ)	Sólo si $X \sim N(\mu, \sigma)$	$\hat{\sigma}^2 = S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X})^2$
p	Sólo si $X \sim B(1,p)$ y $n \geq 100$	$\hat{p} = \frac{\text{nº elementos con la propiedad}}{\text{tamaño muestra}}$

Método de la construcción de intervalos de confianza

Método de la cantidad pivotal

1.- Sea $\hat{\theta} = T(X_1, X_2, \dots, X_n) = T$ es el estimador obtenido para θ.

2.- A partir de T se define otra v.a. U que llamaremos pivote.
U tiene que cumplir:
- La distribución de U tiene que ser conocida y no depender de θ.
- En la expresión de U tiene que aparecer el parámetro θ.

3.- Probabilidad alta, $1-\alpha$. Usando la distribución de U encontrar h_1 y h_2:

$$P(h_1 < U < h_2) = 1-\alpha$$

4.- Conocidos h_1 y h_2, despejar θ usando $P(h_1 < U = f(\theta) < h_2) = 1-\alpha$

$$P(\theta \in (a,b)) = P(a(X_1, \dots, X_n) < \theta < b(X_1, \dots, X_n)) = 1-\alpha$$

- X
  - X distribución normal
    - intervalo para la media μ
    - intervalo para la varianza σ^2
  - X distribución cualquiera, $n \geq 100$
    - intervalo aproximado para la media μ
    - intervalo aproximado para la proporción p

Contrastes de hipótesis

Técnica estadística para decidir si aceptamos o rechazamos la veracidad de cierta hipótesis, en términos probabilísticos.

Se plantea en términos de dos hipótesis (nula y alternativa).

Tipos de contrastes

Paramétricos ← Hipótesis sobre los valores que puede tomar el parámetro

- Una característica X: la media μ, varianza σ^2 (desviación típica σ) o la proporción p.
- Dos características X e Y: comparar medias μ_X con μ_Y o varianzas σ_X^2 con σ_Y^2. ← Prácticas solo

Solo prácticas → **No paramétricos** ← Hipótesis sobre el modelo de distribución

- Contraste Chi-Cuadrado
- Contraste de Kolgomorov-Smirnov

Elementos de un contraste

- Hipótesis nula (H_0) e hipótesis alternativa (H_1)

  se considera cierta mientras los datos no muestren lo contrario.

- Nivel de significación α: Máximo error que se permite al tomar la decisión.
- Estadístico del contraste: Sobre el que se toma la decisión
- Región crítica o de rechazo: intervalo o región tal que si el valor $T(x_1, \dots, x_n)$ pertenece a R, se rechazará H_0.

α = Error de tipo 1 ← Rechazar H_0 siendo cierta = $P(\text{rechazar } H_0 / H_0 \text{ cierta})$

β = Error de tipo 2 ← Aceptar H_0 siendo falsa = $P(\text{aceptar } H_0 / H_0 \text{ falsa})$

Estadístico de un contraste

$\mu = E[X]$
$X \sim N(\mu, \sigma)$ ó
X cualquiera, $n \geq 100$

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

$\sigma^2 = V(X)$
$X \sim$ Normal

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

p
$n \geq 100$

$$\hat{p} = \frac{\text{nº de 1s}}{n}$$

Pasos para realizar un contraste de hipótesis.

1.- Planteamiento hipótesis, términos del parámetro (μ, σ^2, p).

2.- Región crítica del contraste, se usa el estadístico del contraste y tiene la misma forma de la hipótesis alternativa.

3.- $\alpha = P(\text{error tipo 1})$ y usando el pivote, obtener la región crítica.
Método equivalente contraste con p-valor.

4.- Toma de decisiones, cae o no en la región crítica.
En p-valor $< \alpha$ se rechaza H_0.

p-valor

Cuantifica el grado de "seguridad" al tomar la decisión.

Se toma la misma decisión usando región crítica que con el p-valor.

p-valor $\in [0, 1]$ si es mayor a α se acepta H_0 y cuanto más cerca a 1 con mayor confianza.

Planteamientos posibles en contrastes paramétricos

Caso 1

$H_0: \theta = \theta_0$ ó $\theta \geq \theta_0$

$H_1: \theta < \theta_0$

Caso 2

$H_0: \theta = \theta_0$ ó $\theta \leq \theta_0$

$H_1: \theta > \theta_0$

Caso 3

Bilateral

$H_0: \theta = \theta_0$

$H_1: \theta \neq \theta_0$

A mano mediante intervalos de confianza $1-\alpha$.

Relación entre los p-valores

$$p\text{-valor}_{caso\,1} = 1 - (p\text{-valor}_{caso\,2})$$

Si $\bar{x} > \mu_0, \sigma_0^2, p_0$ $\quad p\text{-valor}_{caso\,3} = 2\,(p\text{-valor}_{caso\,2})$

Si $\bar{x} < \mu_0, \sigma_0^2, p_0$ $\quad p\text{-valor}_{caso\,3} = 2\,(p\text{-valor}_{caso\,1})$