

TEMA 2

CONTRASTE DE SIGNOS PARA LA MEDIANA

TAMBIÉN PUEDE APLICARSE PARA LA MEDIANA DE LA DIFERENCIA DE DOS MUESTRAS APAREADAS

Objetivo: contrastar $\begin{cases} H_0: m = m_0 \\ H_1: m \neq m_0 \end{cases}$, $m_0 \in \mathbb{R}$

Supongamos: F continua y estrictamente creciente $\Rightarrow H_0: \theta = 0.5$, $\theta = P(X \geq m_0)$

Si H_0 es cierta cabría esperar que la mitad de los valores de la muestra estén por encima de m_0 y la otra mitad por debajo \equiv mitad de los signos de $X_i - m_0$ fueran positivos.

Estatístico de contraste: $S = \# \{i: X_i > m_0\} = \# \text{ signos positivos}$

Como F es abs. continua: $S \sim B(n, \theta)$ si H_0 es cierta $\rightarrow S \sim B(n, 0.5)$
 \hookrightarrow no depende de F .

$p\text{-valor} = 2 \cdot \min \{ P(S \geq s_{obs}), P(S \leq s_{obs}) \} \rightarrow$ Rechazamos H_0 si $p\text{-valor} \leq \alpha$.

Distribución asintótica

Como $S \sim B(n, 0.5) \Rightarrow \frac{S - E(S)}{\sqrt{\text{Var}(S)}} = \frac{S - 0.5n}{\sqrt{0.25n}} = Z \sim N(0,1) \Rightarrow K_\alpha = 0.5\sqrt{n} Z_\alpha + 0.5n + 0.5$
corrección por continuidad

CONTRASTE DE RANGOS ASIGNADOS DE WILCOXON

Objetivo: contrastar $\begin{cases} H_0: m = m_0 \\ H_1: m \neq m_0 \end{cases}$, $m_0 \in \mathbb{R}$

Supondremos: F absolutamente continua y simétrica respecto a su mediana.

Tendremos en cuenta los signos y las magnitudes $A_i = |X_i - m_0| \Rightarrow$ Definimos $\begin{cases} T^+: \text{suma largos de } A_i \text{ positivos} \\ T^-: \text{suma largos de } A_i \text{ negativos} \end{cases}$

Si H_0 es cierta cabría esperar que T^+ y T^- fueran similares.

Estadístico de contraste: $T^+ = \sum_{i=1}^n R(i) \cdot \underset{\substack{\downarrow \text{ si es } + \\ 0 \text{ si es } -}}{Z_i}$ $\xrightarrow{R(i)=j \Rightarrow i=R^{-1}(j)}$ $T^+ = \sum_{j=1}^n j \cdot \underset{\sim \text{Bernoulli}}{2R^{-1}(j)}$ $\begin{matrix} i: \text{índices sin ordenar} \\ j: \text{índices ordenados} \end{matrix}$

Como F es continua: no hay que preocuparse por los empates. $\xrightarrow{\text{si } H_0 \text{ cierta}}$ $T^+ \sim B = \sum_{j=1}^n j \cdot B_j$: $B_j \sim \text{Be}(0.5)$
la variable B es discreta con valores en $\{0, 1, \dots, \frac{n(n+1)}{2}\}$ y $P(B=t) = \frac{\text{fav}}{\text{pos}} = \frac{n(t)}{2^n}$ $\xrightarrow{\text{no depende de F.}}$
 $n(t)$: formas que hay 0 y 1 a B_1, \dots, B_n
t.q. $B = t$

Distribución asintótica

Como $T^+ \sim \sum_{j=1}^n j \cdot B_j \Rightarrow \frac{T^+ - E(T^+)}{\sqrt{\text{Var}(T^+)}} = \frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1) \cdot (2n+1)}{24}}} = Z \sim N(0,1) \Rightarrow K_\alpha = \text{Va}(T^+) Z_\alpha + E(T^+) + 0.5$
 $\xrightarrow{\text{corrección por continuidad}}$
tomando el menor entero se mejora la aprox

CONTRASTE DE DESPLAZAMIENTO DE MEDIANAS - MANN-WHITNEY

Para dos muestras independientes $\left\{ \begin{array}{l} X_1, \dots, X_m ; X \sim F_X \\ Y_1, \dots, Y_n ; Y \sim F_Y \end{array} \right.$

Objetivo: contrastar $\left\{ \begin{array}{l} H_0 : m_X = m_Y \\ H_1 : m_X \neq m_Y \end{array} \right.$

Supondremos que las distribuciones de X e Y son la misma salvo por un desplazamiento de su mediana.

Si H_0 es cierta: las $m+n$ observaciones provienen de la misma distribución \Rightarrow si las ordenamos y vemos sus rangos, la suma T_X de los rangos de datos X será similar a T_Y

Estadísticos de contraste:

$$\textcircled{1} T_X = \sum_{j=1}^{m+n} j \cdot I_j \quad \rightarrow \quad I_j = \begin{cases} 1 & \text{si dato con rango } j \text{ viene de } X \\ 0 & \text{si no} \end{cases}$$

$$T_X \text{ es discreta con valores } \left\{ \frac{m \cdot (m+1)}{2}, \dots, \left[\frac{(m+n)(m+n+1)}{2} - \frac{n \cdot (n+1)}{2} \right] \right\} \Rightarrow P(T_X = t) = \frac{n(t)}{\binom{m+n}{n}}$$

\hookrightarrow X todas al principio
 \hookrightarrow X todas al final

manera de ordenar las X y las Y para que sume t

Distribución asintótica

$$Z = \frac{T_X - E(T_X)}{\sqrt{V(T_X)}} = \frac{T_X - \frac{m \cdot (m+n+1)}{2}}{\sqrt{\frac{m \cdot n \cdot (m+n+1)}{12}}} \sim N(0,1)$$

$$\textcircled{2} \text{ Mann-Whitney: } U = \sum_{i=1}^m \sum_{j=1}^n U_{ij} ; U_{ij} = I\{X_i > Y_j\} \rightarrow \text{para cada } X, \text{ cuántas } Y\text{'s tiene por debajo}$$

$$\frac{1}{2} \Rightarrow T_X = U + \frac{m \cdot (m+1)}{2}$$

CONTRASTE PARA MÁS DE DOS MUESTRAS INDEPENDIENTES - KRUSKAL WALLIS

$$\text{Para } K \text{ muestras independientes, } \begin{cases} X_{11}, \dots, X_{1n_1} \sim X_1 \\ X_{21}, \dots, X_{2n_2} \sim X_2 \\ \vdots \\ X_{K1}, \dots, X_{Kn_K} \sim X_K \end{cases} \Rightarrow n_1 + n_2 + \dots + n_K = n$$

Objetivo: contrastar $H_0: m_1 = m_2 = \dots = m_K$

Supondremos que las distribuciones de X_1, \dots, X_K son la misma salvo por un desplazamiento de su mediana.

Si H_0 es cierta las n observaciones forman una muestra de una misma distribución \Rightarrow si asignamos rangos a cada muestra, la media de rangos de cada una de las K muestras deberá de estar cerca de la media de todos los rangos $\frac{n \cdot (n+1)}{2} / n = \frac{n+1}{2}$. \Rightarrow Si R_j es la suma de rangos en $X_j \Rightarrow \bar{R}_j$ deberá estar cerca de $\frac{(n+1)}{2}$.

Estadístico de Kruskal-Wallis $Q = \frac{\sum_{j=1}^K n_j \cdot \left(\bar{R}_j - \frac{n+1}{2} \right)^2}{\frac{n \cdot (n+1)}{12}} \rightarrow$ Distribución de Chi-cuadrado

Distribución asintótica

$$Q \sim \chi^2_{K-1} \text{ cuando } \min_j n_j \rightarrow +\infty$$

Comparaciones múltiples

Cuando se rechaza H_0 decimos que hay al menos una mediana diferente al resto \Rightarrow realizamos comparaciones 2 a 2; hay $\binom{K}{2}$ comparaciones y contrastamos $H_0: x_j \cdot m_i = m_j \rightarrow$ Utilizaremos el contraste de Cramer-Iman.

CONTRASTE PARA MÁS DE DOS MUESTRAS RELACIONADAS - FRIEDMAN

Para K poblaciones (o tratamientos) relacionados, organizamos las observaciones en b bloques de tal manera que se observen variables X_{ij} donde $i = 1, \dots, b$ representa el bloque y $j = 1, \dots, K$ el tratamiento. \Rightarrow tenemos b vectores aleatorios X_{i1}, \dots, X_{iK} que suponemos independientes; tenemos $n = b \cdot K$ observaciones.

Objetivo: contrastar $H_0: m_1 = m_2 = \dots = m_K$

Supondremos que existe una distribución F_0 con mediana 0 t.q. $X_{ij} \sim F_0(x - \theta - \beta_i - m_j)$

Si H_0 es cierta las n observaciones forman una muestra de una misma distribución \Rightarrow asignamos rangos en cada bloque; R_{ij} es rango de X_{ij} dentro del bloque i y $R_j = \sum_{i=1}^b R_{ij}$ denota la suma de rangos debida al trat. j .

Estadístico de Friedman: $S = \frac{\sum_{j=1}^K \left(R_j - \frac{b(K+1)}{2} \right)^2}{\frac{bK(K+1)}{12}} = \frac{12}{bK(K+1)} \cdot \sum_{j=1}^K R_j^2 - 3b(K+1) \rightarrow$ Distribución discreta

Distribución asintótica

$S \sim \chi^2_{K-1}$ cuando $b \rightarrow +\infty$

Comparaciones múltiples

Cuando se rechaza H_0 decimos que hay al menos una mediana diferente al resto \Rightarrow realizamos comparaciones 2 a 2; hay $\binom{K}{2}$ comparaciones y contrastamos $H_{0,j}: m_i = m_j \rightarrow$ Utilicemos el contraste exacto de FHPG.

CONTRASTES DE BONDAD DE AJUSTE

Queremos saber si una determinada variable aleatoria $X = (X_1, \dots, X_n) \sim F$ sigue o no una distribución determinada.

Tenemos dos tipos de hipótesis a contrastar; $H_0: F = F_0$ (simple) y $H_0: F \in \{F_\theta: \theta \in \Theta\}$ (compuesta)

A) CONTRASTE DE BONDAD DE AJUSTE - HIPÓTESIS NULA SIMPLE

A1 - FUNCIÓN DE DISTRIBUCIÓN EMPÍRICA

Disponemos de una muestra $X = (X_1, \dots, X_n) \sim F$

Objetivo: contrastar $H_0: F = F_0$, donde F_0 totalmente especificada y absolutamente continua.

Estimador: $F_n(x) = \frac{\#\{i: X_i \leq x\}}{n} \rightarrow$ toma valores en $\{0, \dots, 1\}$

\rightarrow vale 0 en $-\infty$
vale 1 en $+\infty$
es creciente

\rightarrow pone igual masa de probabilidad en cada uno de los n datos observados

• si $X_1 = x_1, \dots, X_n = x_n \Rightarrow F_n \sim U\{x_1, \dots, x_n\}$

• $n \cdot F_n(x) = \#\{i: X_i \leq x\}$, $i = 1, \dots, n \Rightarrow n \cdot F_n(x)$ cuenta el número de éxitos en n eventos independientes, luego $n \cdot F_n(x) \sim B(n, p)$, donde $p = P(X_i \leq x) = F(x) \rightarrow$ por lo tanto, $E(F_n(x)) = F(x)$ y $V(F_n(x)) = \frac{p \cdot F(x) \cdot (1 - F(x))}{n} \rightarrow$ y como $F_n(x)$ es ~~ensogado~~ y $U(F_n(x)) \xrightarrow{n \rightarrow \infty} 0$, $F_n(x) \xrightarrow{P} F(x)$ \rightarrow ~~ensogado~~

A2 - CONTRASTE DE BONDAD DE AJUSTE DE KOLMOGOROV-SMIRNOV

Objetivo: contrastar $H_0: F = F_0$ para una cierta F_0 continua $\Rightarrow F(x) = F_0(x) \quad \forall x \in \mathbb{R}$

$H_1: F(x) \neq F_0(x)$ para algún $x \in \mathbb{R}$

Estadística de contraste de Kolmogorov-Smirnov: $D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)| = \max_{x \in \mathbb{R}} \{D_n^+, D_n^-\}$ donde

$$\begin{cases} D_n^+ = \sup_{x \in \mathbb{R}} \{F_n(x) - F_0(x)\} \\ D_n^- = \sup_{x \in \mathbb{R}} \{-F_n(x) + F_0(x)\} \end{cases}$$

↪ Distancia uniforme

Si H_0 es cierta D_n es un contraste de distribución libre: como F_0 es continua, $F_0(x) \sim U(0,1) \Rightarrow X_i = F_0^{-1}(U_i)$,

con U_1, \dots, U_n muestra $U(0,1)$ y por lo tanto $F_n(x) = \frac{\#\{i: X_i \leq x\}}{n} = \frac{1}{n} \cdot \sum_{i=1}^n \mathbb{I}\{X_i \leq x\} = \frac{1}{n} \cdot \sum_{i=1}^n \mathbb{I}\{F_0^{-1}(U_i) \leq x\}$
 $= \frac{1}{n} \cdot \sum_{i=1}^n \mathbb{I}\{U_i \leq F_0(x)\} = G_n(F_0(x))$ donde G_n es la función de distribución empírica de U_1, \dots, U_n .

Con esto, tenemos que $D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)| = \sup_{x \in \mathbb{R}} |G_n(F_0(x)) - F_0(x)| = \sup_{u \in [0,1]} |G_n(u) - u|$ ↪ se calcula a partir de una muestra $U(0,1) \Rightarrow$ no depende de F

Distribución exacta

$D_n^+ = \sup_{x \in \mathbb{R}} \{F_n(x) - F_0(x)\} \longrightarrow$ si ordenamos las n observaciones $X_{(1)} < \dots < X_{(n)}$, $D_n^+ = \max_{1 \leq i \leq n} \{F_n(X_{(i)}) - F_0(X_{(i)})\}$ y como $F_n(X_{(i)}) = \frac{\#\{j: X_{(j)} \leq X_{(i)}\}}{n} = \frac{i}{n}$ tenemos entonces que $D_n^+ = \max_{1 \leq i \leq n} \left\{ \frac{i}{n} - F_0(X_{(i)}) \right\}$ y $D_n^- = \max_{1 \leq i \leq n} \left\{ F_0(X_{(i)}) - \frac{i-1}{n} \right\}$ de igual forma.

Si H_0 es cierta la distribución de D_n^+ y D_n^- es la misma: $G(x) = 1 - x \cdot \sum_{j=0}^{n(x)} \binom{n}{j} (1-x - \frac{j}{n})^{n-j} (x + \frac{j}{n})^{j-1} \Rightarrow P(D_n \leq x) = (G(x))^2$.

Distribución asintótica

Para D_n la aproximación más usada es: $P(\sqrt{n} D_n \leq n) \longrightarrow 1 - 2 \cdot \sum_{k=0}^{\infty} (-1)^k \cdot e^{-2k^2 x^2}$

El cuantil $1-\alpha$ $D_{n,\alpha}$ de la distribución se aproxima: $\begin{cases} D_{n,0.05} = 1.36/\sqrt{n} \\ D_{n,0.01} = 1.63/\sqrt{n} \end{cases}$

A3 - CONTRASTE DE BONDAD DE AJUSTE DE CRAMÉR-VON MISES

Objetivo: contrastar $H_0: F = F_0$ para una cierta F_0 continua $\Rightarrow F(x) = F_0(x) \forall x \in \mathbb{R}$

$H_1: F(x) \neq F_0(x)$ para algún $x \in \mathbb{R}$

Estadístico de contraste de Cramér-von Mises: $W_n^2 = E_X(\{F_n(x) - F_0(x)\}^2 | x_1, \dots, x_n) = \int_{-\infty}^{\infty} \{F_n(x) - F_0(x)\}^2 dF_0(x)$

↪ Distancia cuadrática

En la práctica se utiliza $T_n = n W_n^2 = \frac{1}{12n} + \sum_{i=0}^n \left\{ \frac{2i-1}{2n} - F_0(X_{(i)}) \right\}^2$

Distribución exacta

Complicada pero está tabulada

Distribución asintótica

$P(T_n \leq x) \rightarrow 1 - \frac{1}{\pi} \sum_{j=1}^{\infty} (-1)^{j-1} w_j(x)$ ↪ integral complicada que depende de x

A4 - CONTRASTE DE BONDAD DE AJUSTE DE ANDERSON-DARLING

Objetivo: contrastar $H_0: F = F_0$ para una cierta F_0 continua $\Rightarrow F(x) = F_0(x) \forall x \in \mathbb{R}$

$H_1: F(x) \neq F_0(x)$ para algún $x \in \mathbb{R}$

Estadístico de contraste de Anderson-Darling: $A_n^2 = n \int_{-\infty}^{+\infty} \frac{\{F_n(x) - F_0(x)\}^2}{F_0(x) \cdot \{1 - F_0(x)\}} dF_0(x)$
↳ Distancia cuadrática ponderada

Cuando H_0 es cierta la varianza de $F_n(x)$ es $\frac{1}{n} \cdot F_0(x) \cdot \{1 - F_0(x)\} \rightarrow$ de ahí el estadístico de contraste
En la práctica se utiliza $A_n^2 = -n - \frac{1}{n} \sum_{i=1}^n \left\{ (2i-1) \log F_0(X_{(i)}) + (2n+1-2i) \log (1 - F_0(X_{(i)})) \right\}$

Distribución exacta

Complicada pero está tabulada

Distribución asintótica


La distribución de A_n^2 se puede aproximar mediante la va $\sum_{j=1}^{\infty} \frac{Y_j}{j \cdot (j+1)}$, donde $Y_j \sim \chi_1^2$, $j \geq 1$, son i.i.d.

A5 - CONTRASTE DE BONDAD DE AJUSTE DE CHI-CUADRADO DE PEARSON

Objetivo: contrastar $H_0: F = F_0$ para una cuenta F_0 continua $\Rightarrow F(x) = F_0(x) \forall x \in \mathbb{R}$

$H_1: F(x) \neq F_0(x)$ para algún $x \in \mathbb{R}$

- 1) Dividimos el rango de posibles valores de X en k celdas A_1, \dots, A_k t.q. formen una partición de \mathbb{R}
- 2) Calculamos frecuencias esperadas de cada celda $E_j = n \cdot p_j$, $p_j = P(A_j | H_0 \text{ cuenta})$
- 3) Calculamos frecuencias observadas de cada celda $O_j = \#\{i: X_i \in A_j\}$

Estadístico de contraste de Pearson: $T = \sum_{j=1}^k \frac{(O_j - E_j)^2}{E_j}$  en sus cuadrados relativos

Distribución asintótica

$T \rightsquigarrow \chi_{k-1}^2$ (al menos 80% $E_j \geq 5$)

B) CONTRASTE DE BONDAD DE AJUSTE - HIPÓTESIS NULA COMPUESTA

Objetivo: contrastar $H_0: F \in \{F_\theta : \theta \in \Theta\}$

Obtenemos estimador $\hat{\theta}$ de máxima verosimilitud de θ

Estadístico de contraste: $\hat{D}_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_{\hat{\theta}}(x)| \Rightarrow$ la distribución de \hat{D}_n no coincide con la distribución del estadístico de contraste en el caso de H_0 simple debido a que \hat{D}_n depende de la familia paramétrica considerada.

\hat{D}_n tabulada para algunas familias paramétricas de interés; Lilliefors tabuló la clave $F = \{\Phi_{\mu, \sigma^2} : \mu \in \mathbb{R}, \sigma > 0\}$ de distribuciones $N(\mu, \sigma^2)$.

B1- CONTRASTE DE BONDAD DE AJUSTE DE NORMALIDAD DE SHAPIRO-FRANCA

Basado en el QQ-plot: gráfico que representa los cuantiles muestrales en función de los teóricos de $N(0,1)$.

Si $z_{\alpha; \mu, \sigma^2}$ es el cuantil α de $N(\mu, \sigma^2) \Rightarrow z_{\alpha; \mu, \sigma^2} = \mu + \sigma \cdot z_{\alpha; 0, 1}$.

Si $X_1, \dots, X_n \sim N(\mu, \sigma^2) \Rightarrow$ el QQ-plot debe ajustarse a una recta.

Estadístico de contraste de Shapiro-Francia: $|W_p|$ es el cuadrado del coef. correlación de los puntos del QQ.

\rightarrow Existe una versión más elaborada, test Shapiro-Wilk.

B2 - CONTRASTE DE BONDAD DE AJUSTE DE CHI-CUADRADO DE PEARSON

Objetivo: contrastar $H_0: F \in \{F_\theta: \theta \in \Theta \subseteq \mathbb{R}^r\}$

- 1) Calculamos el estimador de máxima verosimilitud $\hat{\theta}$ de los r parámetros $\theta = (\theta_1, \dots, \theta_r)$
- 2) Realizamos el contraste de bondad de ajuste de χ^2 de Pearson para $H_0: F = F_\theta$ mediante el estadístico
$$T = \sum_{j=1}^K \frac{(O_j - E_j)^2}{E_j}$$
, sustituyendo las frecuencias esperadas E_j por los estimados \hat{E}_j a partir de F_θ

Distribución asintótica

$$T \rightsquigarrow \chi_{K-r-1}^2 \quad (\text{al menos } 80\% E_j \geq 5)$$

CONTRASTE DE HOMOGENEIDAD

Tenemos s v.a. $X_1 \sim F_1, \dots, X_s \sim F_s$

Objetivo: contrastar $H_0: F_1 = F_2 = \dots = F_s$

→ Disponemos de s muestras independientes $\begin{cases} X_{11}, \dots, X_{1n_1} \sim X_1 \\ \vdots \\ X_{s1}, \dots, X_{sn_s} \sim X_s \end{cases}$, por lo tanto $n = n_1 + \dots + n_s$ observaciones

1) Dividimos los n datos en r clases A_1, \dots, A_r

2) Contaremos las frecuencias observadas $O_{ij} = \# \{k : X_{jk} \in A_i\}$

→ Si H_0 es cierta $O_{ij} \sim B(n_j, p_i)$ con $p_i = P(A_i | H_0 \text{ cierta})$ ↗ no se pueden calcular porque no sabemos que F es la verdadera

3) Estimamos los p_i : $\hat{p}_i = \frac{\sum_{j=1}^s O_{ij}}{n} \Rightarrow \hat{E}_{ij} = n_j \cdot \hat{p}_i$

→ Estadístico de contraste: $T = \sum_{i=1}^r \sum_{j=1}^s \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$

Distribución asintótica

$$T \rightsquigarrow \chi_{(r-1)(s-1)}^2$$

CONTRASTE DE INDEPENDENCIA

Tenemos variables X e Y

Objetivo: contrastar $H_0: X \text{ e } Y \text{ son indep.}$

→ Disponemos de una muestra $(X_1, Y_1), \dots, (X_n, Y_n)$

4) Dividimos los n datos en $\left\{ \begin{array}{l} r \text{ clases } A_1, \dots, A_r \text{ para } X \\ s \text{ clases } B_1, \dots, B_s \text{ para } Y \end{array} \right.$

2) Consideramos las frecuencias observadas: $O_{ij} = \# \{ K : X_K \in A_i, Y_K \in B_j \}$

↳ Si no es cierta: $E_{ij} = n \cdot p_{ij} = n \cdot P(X \in A_i, Y \in B_j) = n \cdot P(X \in A_i) \cdot P(Y \in B_j)$ independientes

3) Estimamos frecuencias esperadas: $\hat{E}_{ij} = n \cdot \frac{\sum_{j=1}^s a_{ij}}{n} \cdot \frac{\sum_{i=1}^r a_{ij}}{n} = \frac{O_{i \cdot} \cdot O_{\cdot j}}{n}$

El estadístico de contraste es: $T = \sum_{i=1}^r \sum_{j=1}^s \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$ \rightarrow mismo que en el contraste de homogeneidad pero frecuencias representan casos distintos

Distribución asintótica

$$T \sim \chi^2_{(r-1)(5-1)}$$

TEMA 3

