Metodos No Parametricos - Final

Gonzalo Barrera Borla

Lunes 15/08/2019

Contents

1 Notación

- $[n]$ representa el conjunto de los naturales de 1 hasta n, $1, 2, 3, \ldots, n$
- $\mathbb{I}$ representa la función indicadora, $\mathbb{I}(P) = \begin{cases} 1 & \text{si P es verdadera} \\ 0 & \text{si P es falsa} \end{cases}$
- $\Pr(X)$ es la probabilidad del evento X
- E, Var son los operadores esperanza y varianza, respectivamente

2 Ejercicios

2.1 Practica 1, ej. 4

Un empresario de la industria alimenticia asegura que menos del 10% de sus frascos de café instantáneo contiene menos café del que garantiza la etiqueta. Para probar esta afirmación se eligen al azar 15 frascos de café y se pesa su contenido. Su afirmación es aceptada si a lo sumo dos frascos contienen menos café del garantizado.

1. ¿Qué hipótesis se deben testear?

2. ¿Cuál es el nivel de la regla de decisión planteada? ¿Le parece razonable?

3. Encuentre la probabilidad de que la afirmación del empresario sea aceptada cuando el porcentaje real de frascos que contienen menos café del garantizado en la etiqueta es 5%, 10% y 20%.

4. Grafique la función de potencia del test planteado inicialmente. Muestre que es insesgado.

5. Con el tamaño de muestra dado, ¿es posible obtener un test de nivel 0.05? Hallar el tamaño de muestra mínimo para obtener un test de nivel 0.05, manteniendo la misma región de rechazo que el test anterior.

El planteo de las hipótesis a testear suele ser un poco idiosincrático. Tradicionalmente, "lo que se desea probar" se coloca en la hipótesis alternativa, y en este caso uno esperaría que lo que se desea probar, es que la afirmación del empresario. Sin embargo, la frase "su afirmacion sera *aceptada*" (id est, *no rechazada*), nos sugiere que la coloquemos en la hipótesis nula. El empresario asegura que "menos del 10% de sus frascos de café instantáneo contiene menos café del que garantiza la etiqueta", que será entonces la hiótesis nula, y su complemento será la hipotesis alternativa:

$$\begin{aligned} & H_0 : \text{Menos del 10\% de los frascos contienen menos café del garantizado} \\ vs. \quad & H_1 : \text{Mas del 10\% de los frascos contiene menos cafe del garantizado} \end{aligned}$$

Sea $X_i = 1$ si el i-ésimo frasco de cafe tiene menos cafe del que garantiza la etiqueta, y 0 en caso contrario. Supongamos además que la probabilidad p de que un frasco de café contenga menos café del garantizado, es constante e identica para todos los frascos de manera que $X_i \overset{iid}{\sim} Ber(p), i = 1, ..., n$, y $\sum_{i=1}^{n} X_i = T \sim Bi(n,p)$ (en nuestro caso, $n = 15$). Bajo estos supuestos, una forma matemáticamente tratable de las hipótesis originales es

$$H_0 : p \leq 0.1 \quad vs. H_1 : p > 0.1$$

o

$$H_0 : p \in \Theta_0 = [0, 0.1] \quad vs. H_1 : p \in \Theta_1 = [0., 1]$$

.

Sea $\Phi(T) = \mathbb{I}(T > k)$ un test que rechaza la hipótesis nula para valores altos de T. El nivel de significación α es igual al supremo de su esperanza bajo la hipótesis nula. En general,

$$\mathrm{E}(\Phi | p = p_0) = \mathrm{E}_{p_0}(\mathbb{I}(T > k)) = \mathrm{Pr}_{p_0}(T > k) = \mathrm{Pr}(Bi(n, p_0) > k)$$

Y en particular, $n = 15, k = 2$ nos dan el test planteado. Sea $f_{n,k}(p) = \mathrm{Pr}(Bi(n,p) > k)$ la probabilidad acumulada a *derecha* para una VA binomial, como función de p para $0 < k < n$ dados. f es *estrictamente creciente* en p, ya que a mayor p, menor es la probabilidad de que T realice un valor por debajo de k , de manera que $\sup_{p\in[a,b]} f_{n,k}(p) = f(b)$. Luego,

$$\alpha = \sup_{p_0\in\Theta_o} \mathrm{E}_{p_0}(\Phi) = \Pr_{0.1}(T > 2) = \sum_{i=3}^{15} \binom{15}{i} 0.1^i \ 0.9^{15-i} \approx 0.1841$$

En general, un test que rechace la hipótesis nula ("no acepte" la afirmación del empresario) casi una de cada 5 veces aún cuando éste esté diciendo la verdad, no es razonable, ya que deja demasiado "librado al azar" el resultado. Tradicionalmente, esperaríamos $\alpha \leq 0.5$.

La probabilidad de que la afirmación del empresario sea aceptada cuando la proporción real de frascos de café es p_1, es simplemente la probabilidad de que T *no* caiga en la región de rechaz

$$\Pr(\Phi = 0|p = p_1) = \Pr_{p_1}(T \leq k) = \Pr(Bi(n, p_1) \leq k) = \sum_{i=0}^{k} \binom{n}{i} p_1^i (1 - p_1)^{n-i}$$

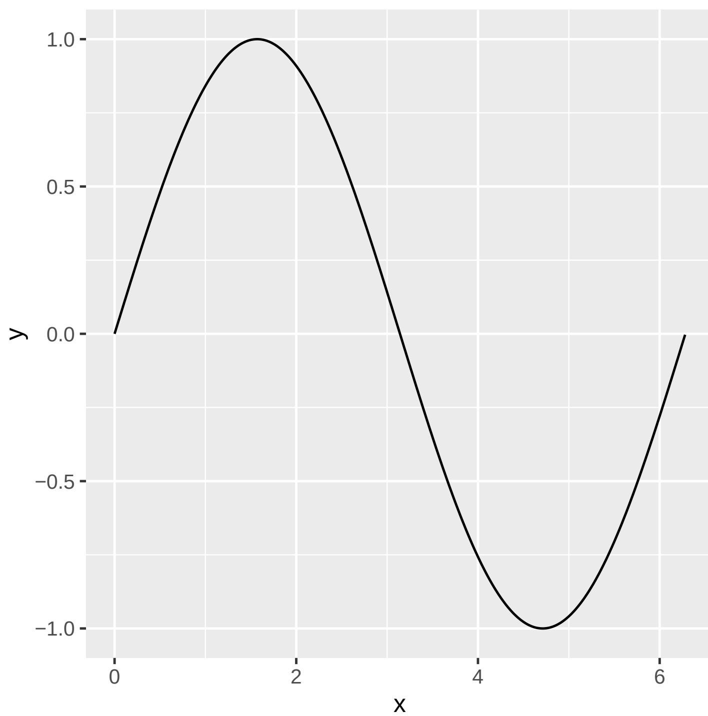
$$\Pr_{0.05}(T \leq 2) = \sum_{i=0}^{2} \binom{15}{i} 0.05^i 0.95^{15-i} \approx 0.9638$$

$$\Pr_{0.1}(T \leq 2) \approx 0.8159$$

$$\Pr_{0.2}(T \leq 2) \approx 0.3980$$

Los puntos (b) y (c) piden el nivel de significacion y la probabilidadde error de tipo II, para lo cual ya hemos generado funciones. Para el punto (c), calculamos la funcion de potencia en una grilla de valores de \$p\$, y la graficamos con ‘ggplot‘:

```
> library(tidyverse)
> ggplot(tibble(x=seq(0,2*pi, 0.01), y=sin(x)), aes(x, y)) + geom_line()
```



Se dice que el test $\Phi(T)$ es insesgado para las hipótesis H_0, H_1cuando es menos probable rechazar la hipótesis nula en caso de que esta sea verdadera, que cuando es falsa. Más formalmente, dado un test $\Phi(T) = \mathbb{I}(T \in RR)$ basado en un estadístico cuya función de distribución está unívocamente determinada por un único parámetro $\theta : T \sim F_\theta$ y un par de hipotesis $H_0 : \theta \in \Theta_0 \quad vs. \quad H_1 : \theta \in \Theta_1$

$$
\begin{aligned}
\Pr(\text{rech } H_0|H_0 \text{ Verdadera}) &< \Pr(\text{rech } H_0|H_0 \text{ Falsa}) \\
\Pr(\Phi = 1|H_0 \text{ V}) &< \Pr(\Phi = 1|H_0 \text{ F}) \\
\mathrm{E}(\Phi|H_0 \text{ V}) &< \mathrm{E}(\Phi|H_0 \text{ F}) \\
\mathrm{E}_{\theta_0}(\Phi) &< \mathrm{E}_{\theta_1}(\Phi) && \forall \quad \theta_0 \in \Theta_0, \theta_1 \in \Theta_1 \\
\sup_{\theta \in \Theta_0} \mathrm{E}(\Phi) &< \inf_{\theta \in \Theta_1} \mathrm{E}(\Phi) && \Rightarrow \Phi \text{es insesgado para } H_0, H_1
\end{aligned}
$$

Y recordemos que la funcion ‘prob_rechazo‘ es, justamente, la esperanza del test, bajo H_0 y H_1. El siguiente grafico muestra claramente la insesgadez del test, a pesar de su pesima significacion.

“‘{r grafico p1ej4d} densidad <- 0.001 p1ej4$d <- tibble(p = seq(0, 1, densidad), potencia = map_dbl(p, ~prob_rechazo(n, ., RR))) %>% ggplot(aes(p, potencia)) + geom_line() + geom_vline(xintercept = p_null, alpha = 0.3) + # Referencia: p_null geom_hline(yintercept = alfa, alpha = 0.3) # Referencia: signif. de la RR “‘

Manteniendo la RR propuesta, el punto (e) nos pide $\min n : Pr(Bi(n, 0.1) \in \{0,1,2\}) \leq 0.05$. Escribamos la funcion que lo busca: “‘{r p1e4e}

n_requerido <- function(RR, alfa, p_null, max_n = 10000) { n_req <- 0 while (n_req < max_n) { if (significacion(n_req, p_null, RR) <= alfa) { return(n_req) } else { n_req <- n_req + 1 } } return(NA) # Devuelve NA si no encuentra n antes de max_n } p1ej4$e <- n_requerido(RR, 0.05, p_null) “‘

Practica 2, ej. 2

Construya el gráfico de $S(\theta)$ para n impar y deduzca el estimador y el intervalo de confianza para θ.

Practica 3, ej. 8

Suponga que $\forall i =1,....n ,\ X_i \stackrel{iid}{\sim} F \in \Omega_s$ (distribuciones simétricas con única mediana en 0).

a. Usando que $g(X_1, ...,X_n)$ y $g(-X_1, ...,-X_n)$ tienen la misma distribución, muestre que si $g(X_1, ...,X_n) + g(-X_1, ...,-X_n) = \mu_0$, entonces $g(X_1, ...,X_n)$ está simétricamente distribuída alrededor de $\mu_0/2$, Hint: Muestre que $P(g(X_1, ...,X_n) \leq \mu_0/2 - t) = P(g(X_1, ...,X_n) \leq \mu_0/2 + t)$. b. Aplique a) al estadístico del test de rangos signados de Wilcoxon para demostrar que T^+ tiene distribución simétrica bajo la hipótesis nula. ¿Cuál es el punto de simetría?

Demostraciones

Consistencia del test de Wilcoxon

Sea $\bar{X}=(X_1, ..., X_n)$ una muestra aleatoria tal que $\forall i =1,....n ,\ X_i \stackrel{iid}{\sim} F \in \Omega_s$ (distribuciones simétricas con única mediana en 0), y definamos - la posicion o rango de X_i en la muestra ordenada segun valores absolutos: $R_i = \#\{j : |X_j| \leq |X_i|, j = 1,... n\}$, - la "funcion signo" $s(X_i) = I(X_i > 0)$, y - el estadistico $T^+ = \sum_{i=1}^{n} R_i \ s(X_i)$

Recordemos que T^+ admite la siguiente definicion equivalente en base a los promedios de Walsh, $$ T^+ = \#\left\{ i, j : \frac{X_i + X_j}{2} > 0, \ 1 \leq i \leq j \leq n \right\} = \sum_{i=1}^{j} \sum_{j=1}^{n} I(X_i + X_j > 0) = \sum_{i=1}^{j} \sum_{j=1}^{n} T_{ij} $$

donde $I(\cdot)$ es la funcion indicadora.

Utilicemos esta ultima forma para desarrollar la esperanza y varianza de T^+, dada cierta distribucion subyacente fija F_0. Gracias a la linealidad de la esperanza, y el hecho de que $E(I(Q)) = P(Q)$ $$ \begin{split} E(T^+)&= E\left(\sum_{i=1}^{j} \sum_{j=1}^{n} I(X_i + X_j > 0) \right) =\sum_{i=1}^{j} \sum_{j=1}^{n} E(I(X_i + X_j > 0)) =\sum_{i=1}^{j} \sum_{j=1}^{n} P(X_i + X_j > 0) \\ &=\sum_{i=j}P(X_i > 0) + \sum_{i<j} P(X_i + X_j > 0) \end{split} $$ Sean ahora $p_1=P(X_i > 0), \ \ p_2=P(X_i + X_j > 0 | i < j)$. Se observa que hay n sumandos donde $i=j$ (1 por cada elemento j), y $(n-1)n/2$ sumandos donde $i<j$. Luego, $E(T^+)= np_1 + (n-1)n p_2$.

Para el calculo de la varianza de T^+, usaremos el hecho de que, en general, si $Y = \sum_i X_i \Rightarrow Var(Y) = \mathbb{cov}(Y, Y) = \mathbb{cov}(\sum_i X_i, \sum_i X_i) = \sum_i Var(X_i) + 2 \sum_{i<j}

\mathbb{cov}(X_i, X_j)$. Para ello, tendremos que encontrar la expresion de la covarianza para todas las combinaciones unicas de subindices.

$$ \begin{split} Var(T^+) &= Var\left(\sum_{i=1}^{j} \sum_{j=1}^{n} T_{ij} \right) = \sum_{i=1}^{j} \sum_{j=1}^{n} Var(T_{ij}) + 2 \sum_{i=1}^{k} \sum_{j=1}^{l} \sum_{k=1}^{l} \sum_{l=1}^{n} \mathbb{cov}(T_{ij},T_{kl}) \end{split} $$

Recordemos que si $X \perp Y$ (las VA X y Y son independientes entre si, $cov(X, Y)=0$. Ademas, para funciones deterministicas arbitrarias f, g, $X \perp Y \Rightarrow f(X) \perp g(Y)$. Como $T_{ij}=f(X_i, X_j)$, y $X_i \perp X_j \ \forall i \neq j$, siempre y cuando los pares$(i,j),(k,l)$ no compartan ningun indice, $\mathbb{cov}(T_{ij},T_{kl})=0$. Consideremos a continuacion los casos en que "ambos T" comparten algun subindice: - Hay Usando que $cov(X, Y)=cov(Y,X)$ y $T_{ij}=T_{ji}$, podemos limitarnos sin perdida de generalidad a los $i \leq k \leq j \leq l$ y nos basta con limitarnos a los siguientes casos:

$$ cov(T_{ij}, T_{kl}) = \begin{cases} Var(T_{ii}) = Var(T_{11}) &\text{si } i = k = j = l \\ Var(T_{ij}) = Var(T_{12}) &\text{si } i = k < j = l \\ cov(T_{ii}, T_{il}) = cov(T_{11}, T_{12}) &\text{si } i = k = j < l \\ cov(T_{ij}, T_{il}) = cov(T_{12}, T_{13}) &\text{si } i = k < j < l \\ 0 &\text{en otro caso} \\ \end{cases} $$

Usando que $cov(X, Y) = E(XY) - E(X)E(Y)$ e $I(P) \cdot I(Q) = I(P \land Q)$ (y por ende $I(P)^2=I(P \land P)=I(P)$), calculamos

$$
\begin{aligned}
Var(T_{ii}) &= E(T_{11}^2) - E(T_{11})^2 = E(I(X_1 > 0)^2) - E(I(X_1 > 0))^2 \\
&= E(I(X_1 > 0))(1 - E(I(X_1 > 0))) \\
&= P(X_1 > 0)(1 - P(X_1 > 0)) \\
&= p_1(1 - p_1) \\
Var(T_{ij}) &= E(T_{12}^2) - E(T_{ij})^2 = E(I(X_1 + X_2 > 0)^2) - E(I(X_1 + X_2 > 0))^2 \\
&= E(I(X_1 + X_2 > 0))(1 - E(I(X_1 + X_2 > 0))) \\
&= P(X_1 + X_2 > 0)(1 - P(X_1 + X_2 > 0)) \\
&= p_2(1 - p_2) \\
cov(T_{ii}, T_{il}) &= E(T_{11}T_{12}) - E(T_{11})E(T_{12}) \\
&= E(I(X_1 > 0 \land X_1 + X_2 > 0)) - E(I(X_1 > 0)E(I(X_1 + X_2 > 0)) \\
&= P(X_1 > 0 \land X_1 + X_2 > 0) - p_1 \cdot p_2 \\
&= p_3 - p_1 \cdot p_2 \quad \text{, digamos} \\
cov(T_{ij}, T_{il}) &= E(T_{12}T_{13}) - E(T_{12})E(T_{13}) \\
&= E(I(X_1 + X_2 > 0 \land X_1 + X_3 > 0)) - E(I(X_1 + X_2 > 0)E(I(X_1 + X_3 > 0)) \\
&= P(X_1 + X_2 > 0 \land X_1 + X_3 > 0) - p_2^2 \\
&= p_4 - p_2^2 \quad \text{, digamos}
\end{aligned}
$$

Para concluir, basta ver cuantos de cada tipo de termino en la expansion de $Var(T^+)$ en sus terminos de covarianza. Cuando los 4 subindices son identicos, $cov(T_{ij}, T_{kl})=Var(T_{ii})$, y hay ${n \choose 1}=n$ de elegir un subindice entre n, asi que $Var(T_{ii})$ aparece n veces. Similarmente, hay ${n \choose 2}=n(n-1)/2$ formas de elegir 2 subindices unicos, asi que tanto $cov(T_{11}, T_{22}), cov(T_11), T_tanto $cov(T_{ii}, T_{ij})$ como $Var(T_{ij})$ aparecen ${n \choose 2}=n(n-1)/2$ veces. Por ultimo, hay ${n \choose 3}=n(n-1)(n-2)/6$ formas de elegir 3 subindices unicos, de manera que

formas de elegir 2 subindices, de manera que $Var(T_{ij})$ aparece $(n-1)n/2$ veces, y tanto $cov(T_{ii}, T_{ij})$ como $cov(T_{ij}, T_{ii})$aparece - $cov(T_{ii}, T_{il})$ aparece tambien $n(n-1)/2$ veces ($l-1$ veces por cada $l=1,...,n$), al igual que $cov(T_{il}, T_{ii})$, para un total de $n(n-1)$ veces, y -en $cov(T_{ij}, T_{il})$

$$
\begin{array}{cccccccc}
T_{11} & T_{12} & ... & T_{1i} & ... & T_{1j} & ... & T_{1n} \\
 & T_{22} & ... & T_{2i} & ... & T_{2j} & ... & T_{2n} \\
 & & ... & ... & ... & ... & ... & ... \\
 & & & T_{ii} & ... & T_{ij} & ... & T_{in} \\
 & & & & ... & ... & ... & ... \\
 & & & & & T_{jj} & ... & T_{jn} \\
 & & & & & & ... & ... \\
 & & & & & & & T_{nn}
\end{array}
$$

\begin{split} Var(T_ii) = E ## Teorema de Proyeccion

MWW: Distribución exacta de los estadistico del test de Mann-Whitney-Wilcoxon (MWW) bajo H_0

Apendice

Condiciones suficientes de consistencia

Definicion:

Sea $X_1, ..., X_n$ una muestra aleatoria tal que $\forall \ i=1,..., n \Rightarrow\ X_i \stackrel{iid}{\sim} F \in \Omega$. Se dice que la sucesion de tests $\{\Phi_n\}$ de nivel asintotico α es consistente para las hipotesis $H_0 : F \in \Omega_{0} \subset \Omega\text{ vs. } H_1 : F \in \Omega_1\subset \Omega$, si - $\alpha \geq E_{F_0}(\Phi_n) \geq \gamma > 0 \quad \forall \ F_0 \in \Omega_0$, y - $E_{F_1}(\Phi_n) \to 1 \quad \forall \ F_1 \in \Omega_1$, donde $E_G(T(X_1, ..., X_n)) = E(T | F = G)$. Es decir, que el nivel de significacion del test se mantiene acotado encima del cero para la sucesion, mientras que la potencia tiende a 1 para la alternativa, a medida que aumenta el tamano muestral.

El conjunto de todas las distribuciones para las cuales $\{\Phi_n\}$ es consistente se denomina _clase de consistencia_ $\Omega_c \subset \Omega_1$.

El siguiente teorema es util para determinar la clase de consistencia en ciertos test comunes.

Teorema Sea Φ_n un test de nivel asintotico α basado en T_n que rechaza H_0 para valores grandes, para las hipotesis previamente definidas. Si

$$ T_n \stackrel{p}{\to} \mu(F) = \begin{cases} \mu_0 &\forall F \in

\Omega_0 \\ > \mu_0 &\forall F \in \Omega_c \\ \end{cases} $$ y ademas existe σ_0 tal que

$$ \sqrt{n}\frac{T_n - \mu_0}{\sigma_0} \stackrel{D}{\to} N(0,1) \quad \forall F \in \Omega_0 $$ entonces existe una sucesion de valores k_n tal que la sucesion de tests $\Phi_n = I(T_n \geq k_n)$ es consistente para la clase Ω_c.

Demostracion

Part I
Demostraciones

Part II
Anexo