INFERENCIA ESTADÍSTICA II

Tema 5: Contrastes basados en estadísticos de rangos

Author
Victor Elvira Fernández, Tomás Ruiz Rojo, Juan Horrillo Crespo
Universidad de Valladolid

23 de diciembre de 2024

AVISO

Estos apuntes fueron creados de forma voluntaria por un grupo de estudiantes, invirtiendo tiempo, dedicación y esfuerzo para ofrecer información útil a la comunidad. Apreciamos cualquier apoyo que se nos quiera brindar, ya que nos ayuda a continuar con futuros proyectos de este tipo.

Si deseas colaborar en esta clase de proyectos puedes contactarnos y unirte o invitarnos a unas ricas patatas 5 salsas por el siguiente enlace:

Buy Me a Patatas 5 Salsas

https://www.buymeacoffee.com/ApuntesINdat

- Mail Juan Horrillo
- Mail Victor Elvira
- Mail Tomás Rojo

Si has colaborado de cualquier forma te agradecemos enormemente.

Índice

1. Test de rangos

Caso paramétrico normal.
$X_1, \dots, X_n$ i.i.d. $N(\mu_1, \sigma_1)$
$Y_1, \dots, Y_n$ i.i.d. $N(\mu_2, \sigma_2)$

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

En el caso no paramétrico:
$X_1, \dots, X_n$ i.i.d. con distribución F
$Y_1, \dots, Y_n$ i.i.d. con distribución G

$$H_0 : F = G$$

$$H_1 : F \neq G$$

En ambos casos el objetico es el mismo, comparar tratamientos o resultados.

Ejemplo:

Digamos que se quiere contrastar la eficacia de un nuevo medicamento para una enfermedad. Lo primero que se tiene que hacer es diseñar un experimento para obtener los datos.

Kowalski, opciones (para diseñar el experimento xd):

Tenemos dos opciones...

1. **Modelo de aleatorización**: Los datos vienen de un diseño controlado en el que los individuos de análisis han sido asignados aleatoriamente a diferentes grupos.

2. **Modelo poblacional:** Los datos son extraidos de una población y se asume que esa población tiene ciertas propiedades. Por ejemplo: una distribución

1.1. Modelo de aleatorización

Se dispone de N individuos. Se eligen n individuos a los que se asigna un tratamiento(una medición). Caso donde

$$X_1, \dots, X_m \quad H_0 : \text{El tratamiento no tiene efecto}$$

$$Y_1, \dots, Y_n \quad H_1 : \text{El tratamiento tiene efecto}$$

El estadístico que utilicemos, rechazará H_0 cuando los valores de la variable considerada en los tratados sean mayores a los de control. Para esto, usaremos **el estadístico basado en rangos**. El estadístico basado en rangos, no depende de unidades de medida. Tomaremos toda la muestra y asignaremos rangos.

Definición: Un rango es el lugar que ocupa la observación en la muestra ordenada. Sean $(X_1, \ldots, X_m,\ Y_1, \ldots, Y_n)$ nuestra muestra completa. Se ordenan los valores y se asignan rangos $(i_1, \ldots, i_m, i_{m_1}, i_n)$ Esto es la permutación donde i_k es el valor que ocupa la observación k-ésima en la muestra ordenada.

Ejemplo:

Tenemos $X = \{5, 8, 9\}$ e $Y = \{6, 7, 10\}$
Por tanto la muestra completa será $\{5, 8, 9, 6, 7, 10\}$
Que ordenado $\{5, 6, 7, 8, 9, 10\}$, y se le asigna un rango a cada elemento $(\{1, 2, 3, 4, 5, 6\})$

Usando notación de tuplas, (a, b) donde a es un elemento de la muestra y b el rango asociado, la anterior asignación quedaría de la siguiente manera $\{(5, 1), (6, 2), (7, 3), (8, 4), (9, 5), (10, 6)\}$

Por tanto quedarían asignados...
Rangos de X: $\{1, 4, 5\}$.
Rangos de Y: $\{2, 3, 6\}$.

Notación formal:
$R_1, \ldots, R_m \longrightarrow$ rangos correspondientes a las observaciones $X_1, \ldots, X_m$.
$S_1, \ldots, S_n \longrightarrow$ rangos correspondientes a las observaciones $Y_1, \ldots, Y_n$.
¿Cuando las Y_s son mayores que las X_s?. Cuando s_i sean más grandes, es decir, la suma de rangos sea más grande.

1.1.1. Estadístico de suma de rangos

Definición: La suma
$$W_s = s_1 + \cdots + s_n$$
es conocida como el estadístico de Wilcoxon de suma de rangos ($W_r = R_1 + \cdots + R_m$). Se rechazará H_0 para valores de W_s grandes ($W_s > c_\alpha$).

Como siempre, debemos conocer la distribución del estadístico de W_s bajo H_0. Como la distribución es discreta, y podemos calcular lo siguiente...

$$P_{H_0}(W_s = k) = \sum_{s_1+\cdots+s_n=k} P_{H_0}((S_1, \ldots, S_n) = (s_1, \ldots, s_n))$$

... encontrar la distribución de W_s bajo H_0 se reduce a encontrar la distribución de $(S_1, \ldots, S_n)$.

$$P_{H_0}((S_1, \ldots, S_n) = (s_1, \ldots, s_n)) = \frac{1}{\binom{N}{n}}$$

Cada resultado es igual de probable bajo H_0.

Ejemplo:

Dados $N = 5, m = 2, n = 3$

$$\binom{N}{n} = \binom{5}{3} = \frac{5!}{3! \cdot 2!} = 10$$

Habrá 10 posibles resultados de s_1, s_2, s_3.

Tratados	$(1,2,3)$	$(1,2,4)$	$(1,2,5)$	$\dots$	$(3,4,5)$
$P_{H_0}((S_1,\dots,S_n) = (s_1,\dots,s_n))$	0,1	0,1	0,1	$\dots$	0,1
W_s	6	7	8	$\dots$	12

Distribución de W_s bajo H_0:

k	6	7	8	9	10	11	12
$P_{H_0}(W_s = k)$	0,1	0,1	0,2	0,2	0,2	0,1	0,1

Rechazaremos H_0 cuando el valor de W_s sea poco probable.

Observaciones:

La distribución de W_s no es la misma si hubiéramos asignado n a tratamientos y n a control. Lo que es lo mismo, W_R no sigue la misma distribución que W_s bajo H_0. La distribución dependerá de n y m.

1.2. Estadístico de Mann-Whitney

En el caso anterior, el valor mínimo de W_s corresponde a la situaciñon a la que los individuos con tratamiento toman los valores más pequeños. Donde en el ejemplo anterior, $W_s = 6 \quad 6 = \frac{n\cdot(n+1)}{2}$. Si consideramos el estadístico de Mann-Whitney,

$$W_{XY} = W_S - \frac{n(n+1)}{2}$$

se nos facilitará hacer la tabla porque tomas valores $0, 1, \dots, (n \cdot m)$. Tomará el valor 0 cuando todos los valores de Y son los más pequeños y el valor n $\cdot$ m cuando todos los Y tomen los valores más grandes.

Una ventaja que tiene el estadístico es que toma los mismos valores sin importar la decisión de cuántos asignar a n y cuántos a m.

Del mismo modo:

$$W_{YX} = W_R - \frac{m(m+1)}{2}$$

W_{YX} también toma valores 0,1,$\dots$, $n \cdot m$.
W_{XY} y W_{YX} siguen la misma distribución bajo H_0.

Existen tablas para esta distribución.

Observaciones:

La distribución bajo H_0 de W_S (o W_R) es simétrica respecto a $\frac{n\cdot(N+1)}{2}$.

$$\forall k \quad P_{H_0}\left(W_S = \frac{n\cdot(N+1)}{2} + k\right) = P_{H_0}\left(W_S = \frac{n\cdot(N+1)}{2} - k\right)$$

Bajo H_0 X e Y están igualmente distribuidas, por lo que todos los elementos deberían ser indistinguibles en término de rangos. Por esto W_{XY} y W_{YX} están igualmente distribuidas bajo H_0.

Ejemplo del uso de las tablas:

Tenemos $W = 10, n = 6, m = 4$
Queremos saber $P_{H_0}(W_s \geq 35)$
Tenemos las tablas para $W_{XY} = W_S - \frac{n\cdot(N+1)}{2} = W_S - 21$
Debemos escribir $P_{H_0}(W_s \geq 35)$ como $P_{H_0}(W_{XY} \leq a)$
Sabemos que W_S es simétrico a $\frac{n\cdot(N+1)}{2} = 33$. Entonces...

$$P_{H_0}(W_S \geq 35) = P_{H_0}(W_S \geq 33+2) = P_{H_0}(W_S \leq 33-2) = P_{H_0}(W_S \leq 31)$$

$$= P_{H_0}\left(W_S - \frac{n\cdot(n+1)}{2} \leq 31 - 21\right) = P_{H_0}(W_{XY} \leq 10) = 0{,}3810 \text{(Usando las tablas)}$$

Llegaríamos al mismo resultado usando $W_S + W_R = 55$.

$$P_{H_0}(W_S \geq 35) = P_{H_0}(W_R \leq 20) = P_{H_0}\left(W_R - \frac{m\cdot(m+1)}{2} \leq 20 - 10\right)$$

$$= P_{H_0}(W_{YX} \leq 10) = 0{,}3810$$

Hay tablas hasta $n = m = 10$. A partir de así usariamos la distribución asintótica.

1.2.1. Forma alternativa del estadístico de Mann-Whitney

Una manera alternativa de ver el estadístico de Mann-Whitney que va a ser útil para calcular la potencia y también intervalos de confianza es la siguiente

Definición: Si $X_1, \ldots, X_m$ son valores para individuos sin tratamientos y $Y_1, \ldots, Y_n$ son valores para individuos con tratamiento, y $W_{XY} = W_S - \frac{n\cdot(n+1)}{2}$, W_{XY} es también el número de pares (X_i, Y_j) $i = 1, \ldots, m$, $j = 1, \ldots, n$ para los que $X_i < Y_i$

$$W_{XY} = \#[(X_i, Y_j) | X_i < Y_j]$$

Demostración 1.1. Sean $Y_{(1)}, \dots, Y_{(n)}$ valores ordenados de $Y_1, \dots, Y_n$ y sean $s_1, \dots, s_n$ los rangos correpondientes también ordenados,

- Hay $s_1 - 1$ observaciones menores que $Y_{(1)}$ y todas son X.

$$\#[(X_i, Y_{(1)})|X_i < Y_{(1)}] = S_1 - 1$$

- Hay $s_2 - 1$ observaciones menores que $Y_{(2)}$, de ellas una es $Y_{(1)}$ y el resto son X.

$$\#[(X_i, Y_{(2)})|X_i < Y_{(2)}] = S_2 - 1 - 1 = S_2 - 2$$

- Hay $s_n - 1$ observaciones menores que $Y_{(n)}$, de ellas n-1 son Y y el resto son X.

$$\#[(X_i, Y_{(n)})|X_i < Y_{(n)}] = S_n - n$$

Por lo tanto:

$$\#[(X_i, Y_j)|X_i < Y_j] = (S_1 - 1) + (S_2 - 2) + \cdots + (S_n - n)$$

$$= (S_1 + \cdots + S_n) - (1 + 2 + \cdots + n) = W_S - \frac{n \cdot (n+1)}{2} = W_{XY}$$

Distribución asintótica de W_s Si los valores n y m son grandes (mayores que 10), se considera que la distribución asintótica para W_S bajo H_0 es, por el Teorema Central del Limite...

$$\frac{W_S - E_0(W_S)}{\sqrt{Var_0(W_S)}} \xrightarrow[H_0]{L} N(0,1)$$

$$E_0(W_S) = \frac{n \cdot (N+1)}{2} \quad Var_0(W_S) = \frac{n \cdot (N-n) \cdot (N+1)}{12}$$

2. Test de rangos con observaciones coincidentes

Hemos visto contrastes para 2 muestras independientes, donde se eligen n individuos para un tratamiento y m para un grupo de control(u otro tratamiento), siendo N=n+m.

$$H_0 : \text{Tratamiento no tiene efecto}$$
$$H_1 \text{ Tratamiento tiene efecto}$$

Mediamos la variable de interés en los N individuos.

$$X_1, \dots, X_m \text{ para los individuos de control}$$
$$Y_1, \dots, Y_n \text{ para los individuos del tratamiento}$$

Obteniamos los rangos de la muestra ($R_1, \dots, R_m$ y $S_1, \dots, S_n$). Si la alternativa es mayor, se rechaza H_0 para $W_S = S_1 + \dots + S_n > C$. Si la alternativa es menor, se rechaza H_0 para $W_S = S_1 + \dots + S_n < C$

¿Que pasa si hay coincidencias?

Ejemplo:

Muestra:

$$1.2, 1.7, 1.7, 1.7, 2, 3.1, 3.1, 5$$

Los rangos serían:

$$1, 2, 2, 2, 5, 6, 6, 8$$

Pero estos rangos no serían correctos. Para aquellos valores en los que coincida el rango, se les dan distintos y el rango de todos los que coincidan se calculara como la media de sus rangos. Por tanto, los rangos serían: 1, 3, 3, 3, 5, 6.5, 6.5, 8

A esto lo vamos a llamar **semi-rangos**. Siempre que tengamos coincidencias, calcularemos los semi-rangos

2.1. Semirangos

Definición: Los semi-rangos correspondientes a observaciones coinicidentes, se calculan como la media de los rangos que les corresponderían si no tuviéramos empates.

Notación: Cuando haya coincidencias, los semi-rangos se asignan y representan como:

$$R_1^*, \dots, R_m^* \text{ semi-rango individual de control}$$
$$S_1^*, \dots, S_n^* \text{ semi-rango individual de tratamiento}$$

Para controlar H_0, se rechaza si:

$$W_S^* = S_1^* + \dots + S_n^* > C$$

Como siempre, necesitamos la distribución de W_S^* bajo H_0 que no es la misma que cuando no hay coincidencias, aunque llegaremos a la distribución de la misma forma.

Ejemplo:

$n = m = 3$
$X_1 = 5, \quad X_2 = 5, \quad X_3 = 9, \quad Y_1 = 5, \quad Y_2 = 10, \quad Y_3 = 10$
Los semi-rangos son 2,2,4 2,5.5,5.5 (Nota: si los empates estan en el mismo grupo no nos afectan en nada)

$$(S_1^*, S_2^*, S_3^*) = (2, 5{,}5, 5{,}5) \quad W_S^* = 13$$

Después de conocer los semi-rangos, calculamos la distribución de W_S^* bajo H_0 de la misma forma que anteriormente.

Bajo H_0, (hipótesis de que el tratamiento no tiene efecto), los 6 individuos recibirían los semi-rangos independientemente de que fueran asignados al grupo de tratamiento o de control. Por lo tanto, para la selección de los 3 individuos a tratamiento, hay $\binom{6}{3}$ =20 posibles ekecciones de 3 individuos a tratamiento y 3 a control. Pero no todas diferentes, porque hay repetidos.

S_1^*, S_2^*, S_3^*	$(2, 2, 2)$	$(2, 2, 4)$	$(2, 2, 5{,}5)$	$(2, 4, 5{,}5)$	$(2, 5{,}5, 5{,}5)$	$(4, 5{,}5, 5{,}5)$
W_S^*	6	8	9,5	11,5	13	15
P_{H_0}	$\frac{1}{20}$	$\frac{3}{20}$	$\frac{6}{20}$	$\frac{6}{20}$	$\frac{3}{20}$	$\frac{1}{20}$

La distribución depende de la configuración de las coincidencias. No se tienen tablas ya que habría que considerar cada caso. Para n grande, se tiene al distribución asintótica bajo H_0.

2.1.1. Configuración de las coincidencias

Definición: Configuración de las coincidencias.

- Sea N=n+m el número de individuos tal que n sea el numero de individuos de control y m el numero de individuos del tratamiento
- Sea e el número de observaciones distintas entre los tratamientos
- Sea d_1, el número de observaciones iguales a la más pequeña
- Sea d_2, el número de observaciones iguales a la siguiente más pequeña
- Sea d_e el número de observaciones iguales a la más grande

Al vector $(e, d_1, \ldots, d_e)$ se le conoce como configuración de las coincidencias.

Ejemplo

n=m=3
$X_1 = 5, \quad , X_2 = 5, \quad X_3 = 9, \quad Y_1 = 5, \quad Y_2 = 10, \quad Y_3 = 10$
Los semi-rangos son 2,2,4;2,5.5,5.5
En este caso:

$$(e, d_1, \ldots, d_e) = (3, 3, 1, 2)$$

- El semirango de las d_1:

$$\frac{1 + \cdots + d_1}{d_1} = \frac{d_1 + 1}{2}$$

- El semirango de las d_2:

$$\frac{(d_1+1)+\cdots+(d_1+d_2)}{d_2} = d_1 + \frac{d_2+1}{2}$$

- El semirango i-ésimo

$$\frac{(d_{i-1}+1)+\cdots+(d_{i-1}+d_i)}{d_i} = d_1 + \cdots + d_{i-1} + \frac{d_i+1}{2}$$

$$d_1 = \frac{3+1}{2} = 2 \quad d_2 = 3 + \frac{1+1}{2} = 4 \quad d_3 = 3+1+\frac{2+1}{2} = 5{,}5$$

Estos conteos se pueden hacer solo para n y m pequeños. En este caso, podemos relacionar W_S^* con el estadístico de Mann-Whitney análogo para el caso sin coincidencias.

2.2. Estadístico de Mann-Whitney con observaciones no distintas

El estadístico de W_S^* es una generalización de W_S cuando no todas las observaciones son distintas. Del mismo modo, se puede generalizar el estadístico de Mann-Whitney.

Sea $X_1, \ldots, X_m$ valores de la variable de interés para control y $Y_1, \ldots, Y_m$ valores de la variable de interés del tratamiento ,si todas las observaciones son distintas, definiamos el estadístico de Mann-Whitney como:

$$W_{XY} = \#[(X_i, Y_i)|X_i < Y_i]$$

($\#$ = numero de casos en que:)

En caso de tener coincidencias, se puede definir para cada par (X_i, Y_j)

$$\phi(X_i, Y_j) = \begin{cases} 1 & si \quad X_i < Y_j \\ \frac{1}{2} & si \quad X_i = Y_j \\ 0 & si \quad X_i > Y_j \end{cases}$$

Si definimios $W_{XY}^* = \sum \phi(X_i, Y_j)$, es decir:

$$W_{XY}^* = \#[(X_i, Y_i)|X_i < Y_i] + \frac{1}{2} \cdot \#[(X_i, Y_i)|X_i = Y_i]$$

Resultado: Los tests basados en W_S^* y en W_{XY}^* son equivalentes y además

$$W_{XY}^* = W_S^* - \frac{n \cdot (n+1)}{2}$$

Demostración en el campus

Nota: se puede usar para categorías

2.2.1. Distribución asintótica de W_S^*

Si n y m son grandes y la proporción máxima de observaciones coincidentes no es próxima a 1, es decir, si:

$$max_{i=1,\dots,e}\left\{\frac{d_i}{N}\right\} \ll 1$$

es decir, no hay un grupo en el que estén casi todas las observaciones.

$$\frac{W_S^* - E_\theta(W_S^*)}{\sqrt{Var_\theta(W_S^*)}} \xrightarrow{L} N(0,1)$$

$$E_\theta(W_S^*) = \frac{n\cdot(N+1)}{2}$$

$$Var_\theta(W_S^*) = \frac{n\cdot m\cdot(N-1)}{12} - n\cdot m\cdot\sum_{i=1}^{e}\frac{d_i\cdot(d_i^2-1)}{12-N\cdot(N+1)}$$

Ejercicio 9

En un estudio sobre la efectividad de los consejos psicológicos, 80 jovenes se dividen aleatoriamente en un grupo control de 40 jóvenes , a quienes se aconseja de un modo tradicional, y un grupo de 40 que recibe un tratamiento especial. El cambio en el comportamiento de los jóvenes se califica como pobre, medianamente pobre, medianamente bueno y bueno. Obtenemos los siguientes resultados:

	Pobre	Medianamente pobre	Medianamente bueno	Bueno
Tratamiento	5	7	16	12
Control	7	9	15	9

Contrastar si el efecto del tratamiento es positivo.
Nos piden contrastar:

H_0: no hay diferencias entre control y tratamiento
H_1: El tratamiento aumenta la respuesta

Hay 4 grupos, por lo tanto e=4.

- En el primer grupo hay 12 individuos

(e, d_1, d_2, d_3, d_4)=(4,12,16,31,21)

$$\text{Semi-rangos:}\begin{cases} d_1 = \frac{12+1}{2} = 6{,}5 \\ d_2 = 12 + \frac{16+1}{2} = 20{,}5 \\ d_3 = 12+16+\frac{31+1}{2} = 44 \\ d_4 = 12+16+31+\frac{21+1}{2} = 70 \end{cases}$$

$$W_S^* = \sum_{i=1}^{4} B_i(\text{semirangos}) = 5\cdot(65) + 7\cdot(205) + \cdots + 12\cdot(70) = 1720$$

Vemos si el valor es grande con su distribución asintótica
El p-valor sería:

$$E_0(W_S^*) = \frac{n\cdot(N+1)}{2} = \frac{40\cdot 81}{2} = 1620$$

$$Var(W_S^*) = 9854{,}937$$

$$P_{H_0}(W_S^* \geq 1720) = P\left(\frac{W_S^* - E(W_S^*)}{\sqrt{\mathrm{Var}(W_S^*)}} \geq \frac{1720 - E(W_S^*)}{\sqrt{\mathrm{Var}(W_S^*)}}\right)$$

$$= P\left(Z \geq \frac{1720 - 1620}{\sqrt{9854{,}937}}\right) = 1 - \Phi(1{,}01) = 0{,}16$$

3. Modelo poblacional

El precio que pagamos usando un modelo de aleatorización es que los resultados solo son válidos para los N individuos de estudio y no se pueden extrapolar a una población más amplia. Para que eso sea posible, será necesario que los N individuos representen a toda la población. Dicho de otra forma, necesitamos una **muestra aleatoria simple** de la población.

La situación es la siguiente: Tenemos $N = n + m$ individuos al azar de la población,

$$n \longrightarrow \text{elegidos al azar} \longrightarrow \text{grupo de tratamiento}$$
$$m \longrightarrow \text{restantes al grupo de control}$$

$$Y : \text{Variable respuesta de individuos que reciben el tratamiento}$$
$$X : \text{Variable respuesta de individuos que son del grupo de control}$$

X e Y son dos variables aleatorias con funciones de distribución $X \sim F$ y $Y \sim G$
Queremos contrastar la hipótesis de que el tratamiento NO es efectivo

$$H_0 : \text{El tratamiento no tiene efecto (F=G)}$$
$$H_1 : \text{El tratamiento aumenta/disminuye la respuesta (F¿G/F¡G)}$$

El modelo poblacional tiene dos ventajas fundamentales:

1. Los resultados son extrapolables
2. Podemos estudiar la potencia del test

Si tenemos un modelo poblacional sin coincidencias podemos utilizar el estadístico W_s y el test de Wilcoxon ($W_s > C_\alpha$); bajo H_0, W_s sigue la misma distribución que en el modelo de aleatorización.

$$X_1, \ldots, X_m \quad Y_1, \ldots, Y_n$$
$$R_1, \ldots, R_m \quad S_1, \ldots, S_n$$
$$W_s = S_1 + \cdots + S_n$$

Si hay coincidencias, tenemos que encontrar la distribución de W_s^* bajo H_0.
En este caso el estadístico $W_s^* = S_1^* + \cdots + S_n^*$ no es de distribución libre. La distribución bajo H_0 de los semi-rangos de los n individuos depende de F. Esto se debe (al igual que en el modelo de aleatorización) a que la distribución depende de la configuración de las coincidencias $(e, d_1, \ldots, d_e)$, que en el modelo de aleatorización son un número pero aquí son variables aleatorias cuya distribución depende de F.

Ejemplo

Supongamos F discreta de tal forma que

$$F : \begin{cases} a & \text{Con probabilidad } p \\ b & \text{Con probabilidad } 1 - p \end{cases}$$

Si a¡b, y con m=2 y n=1, entonces los posibles resultados son:

$X_1X_2Y_1$	Probabilidad	Semi-rangos
$a\ a\ a$	p^3	2 2 2
$a\ a\ b$	$p^2(1-p)$	1,5 1,5 2
$a\ b\ a$	$p^2(1-p)$	1,5 2 1,5
$b\ a\ a$	$p^2(1-p)$	2 1,5 1,5
$a\ b\ b$	$p(1-p)^2$	1 2 2
$b\ b\ a$	$p(1-p)^2$	2 2 1
$b\ a\ b$	$p(1-p)^2$	1,5 1 1,5
$b\ b\ b$	$(1-p)^3$	1 1 1

La distribución de W_s^* bajo H_0 será:

S_n^*	1	1,5	2	2,5	3
$P_0(S_1^* = s_1^*)$	$p(1-p)^2$	$2p^2(1-p)$	$p^3 + (1-p)^3$	$2p(1-p)^2$	$p^2(1-p)$

Evidentemente la distribución de W_s^* depende de p; es decir, de F.

Al igual que en el modelo de aleatorización, la distribución de W_s^* depende de la configuración de las coincidencias, solo que esta vez esas coincidencias son v.a. que dependen de F.

3.1. Potencia del test

Una ventaja del modelo poblacional es que podemos calcular la potencia del test. Para ello debemos especificar la hipótesis alternativa. Sean F y G las distribuciones de las variables respuesta en individuos de control y tratamiento respectivamente,

$H_0: \ F = G$

$H_1:$ El tratamiento aumenta la respuesta, $F > G$

¿Qué significa en términos de F y G que el tratamiento aumente la respuesta?

$$\forall z \in \mathbb{R} \qquad P(Y > z) \geq P(X > z) \iff 1 - G(z) \geq 1 - F(z) \iff F(z) \geq G(z)$$

Teorema 3.1. Sean X e Y v.a. tales que $X \sim F$ y $Y \sim G$ con F y G distribuciones de distribución. Se dice que Y (respecto a X) es estocásticamente mayor que X (respecto a Y) cuando los valores que toma la v.a. Y son mayores que los que toma la v.a. X, es decir:

$$G(z) \leq F(z) \qquad \forall z \in \mathbb{R}$$
$$H_0: \ F(x) = G(x)$$
$$H_1: \ F(x) \geq G(x)$$

El cálculo de la potencia requiere la distribución de los rangos. En el caso de F y G continuas es muy complicado, por lo que aproximaremos con la distribución asintótica y usaremos el estadístico de Mann-Whitney.

3.1.1. Potencia asintótica

$$\Pi(F,G):\ X \sim F,\ Y \sim G \text{ si n y m son suficientemente grandes}$$
$$\Pi(F,G) = \underset{\text{Bajo } H_1}{P_{F,G}}(W_{XY} \geq C_\alpha) = P_{F,G}\left(\frac{W_{XY}-E_{FG}(W_{XY})}{\sqrt{Var_{FG}(W_{XY})}} \geq \frac{C_\alpha - E_{FG}(W_{XY})}{\sqrt{Var_{FG}(W_{XY})}}\right) =$$
$$= 1 - \Phi\left(\frac{C_\alpha - E_{FG}(W_{XY})}{\sqrt{Var_{FG}(W_{XY})}}\right)$$

$$E(W_{XY}) = mnp_1$$
$$Var(W_{XY}) = mnp_1(1-p_1) + mn(n-1)(p_2 - p_1^2) + mn(m-1)(p_3 - p_1^2)$$

Siendo:

$$p_1 = P_{FG}(X < Y)$$
$$p_2 = P(X < Y, X < Y')$$
$$p_3 = P(X < Y, X' < Y)$$

El problema viene porque p_1, p_2 y p_3 son difíciles de calcular, por lo que usaremos una aproximación de la potencia.

3.2. Modelo Shift de aproximación de la potencia

Teorema 3.2. F y G se agrupan en un modelo Shift si

$$\exists \Delta > 0,\ \forall x \qquad G(x) = F(x - \Delta)$$

El modelo Shift queda

$$H_0 : F(x) = G(x) \iff \Delta = 0 \qquad H_1 : G(x) = F(x-\Delta) \iff \Delta > 0$$

La potencia se escribe como:

$$\Pi_F(\Delta) = P_\Delta(W_{XY} > C_\alpha),\ \Delta > 0$$

En particular, $\Pi_F(0) = \alpha$

Teorema 3.3. Sea F^* la función de distribución de la diferencia de las dos v.a. independientes con distribución F y sea $f^*(0)$ su densidad en el 0. Entonces

$$\Pi(\Delta) \approx \Phi\left[\sqrt{\frac{12mn}{N+1}} f^*(0)\Delta - \mu_\alpha\right]$$

Donde μ_α / $\Phi(\mu_\alpha) = 1 - \alpha$

Supongamos que N es suficientemente grande como para poder usar la aproximación normal para encontrar C_α

$$\alpha = P_0(W_{XY} > C_\alpha) = P_0\left(\frac{W_{XY} - E_0(W_{XY})}{\sqrt{Var_0(W_{XY})}} \geq \frac{C_\alpha - E_0(W_{XY})}{\sqrt{Var_0(W_{XY})}}\right)$$

$$W_{XY} = W_s - \tfrac{n(n+1)}{2}$$
$$E(W_s) = \tfrac{n(N+1)}{2}$$
$$E(W_{XY}) = \tfrac{n(N+1)}{2} - \tfrac{n(n+1)}{2} = \cdots = \tfrac{nm}{2}$$
$$Var(W_{XY}) = \tfrac{1}{12}mn(N+1)$$

$$\alpha = P_0\left(\frac{W_{XY} - \frac{1}{2}mn}{\sqrt{\frac{1}{12}mn(N+1)}} \geq \frac{C_\alpha - \frac{1}{2}mn}{\sqrt{\frac{1}{12}mn(N+1)}}\right)$$

Por lo que

$$\mu_\alpha = \frac{C_\alpha - \frac{1}{2}mn}{\sqrt{\frac{1}{12}mn(N+1)}} \Longrightarrow C_\alpha = \frac{1}{2}mn + \sqrt{\frac{1}{12}mn(N+1)}\mu_\alpha$$

Calculando la potencia:

$$\Pi_F(\Delta) = P_\Delta\left(W_{XY} \geq \frac{1}{2}mn + \sqrt{\frac{1}{12}mn(N+1)}\mu_\alpha\right) =$$

$$= P_\Delta\left(\frac{W_{XY} - E_\Delta(W_{XY})}{\sqrt{(Var_\Delta(W_{XY}))}} \geq \frac{\frac{1}{2}mn + \sqrt{\frac{1}{12}mn(N+1)}\mu_\alpha - mnp_1}{\sqrt{Var_\Delta(W_{XY})}}\right) =$$

$$= 1 - \Phi\left(\frac{\left(\frac{1}{2} - p_1\right)mn + \mu_\alpha\sqrt{\frac{1}{12}mn(N+1)}}{\sqrt{Var_\Delta(W_{XY})}}\right)$$

Sustituyendo $p_1 = P_{FG}(X < Y)$

$$p_1 = P_\Delta(X < Y) = P_0(X < Y - \Delta) = P_0(Y - X > \Delta) = P_0(\underbrace{Y - X}_{\text{Misma dist. bajo } H_0} - \Delta > 0) = 1 - F^*(\Delta)$$

$F^*(\Delta)$ será la función de distribución de la diferencia de las dos v.a. independientes con distribución F. Si desarrollamos $F^*(\Delta)$ en torno al 0 con el polinomio de Taylor, sabiendo que $(F^*(x))' = f^*(x)$ y por simetría respecto al 0:

$$F^*(\Delta) \approx F^*(0) + (\Delta - 0)f^*(0) = \frac{1}{2} + \Delta f^*(0)$$

Supongamos $D = X - Y$, $X \sim F$, $Y \sim F$ y $D \sim F^*$, entonces:

$$F^*(0) = P(D \leq 0) = P(X - Y \leq 0) = P(X \leq Y) = \frac{1}{2}$$

Por lo tanto,

$$p_1 = 1 - F^*(0) \approx \frac{1}{2} + \Delta f^*(0) \Longrightarrow p_1 - \frac{1}{2} \approx \Delta f^*(0)$$

Entonces, ya podemos hacer una primera aproximación para el cálculo de la potencia:

$$\Pi_F(\Delta) \approx \Phi\left(\frac{mn\Delta f^*(0) - \mu_\alpha\sqrt{\frac{1}{12}mn(N+1)}}{\sqrt{Var_\Delta(W_{XY})}}\right)$$

Nos faltaría calcular $Var_\Delta(W_{XY})$. Podemos hallar una aproximación cuando Δ es pequeño, ya que

$$Var_\Delta(W_{XY}) \approx Var_0(W_{XY}) = \frac{mn(N+1)}{12}$$

Por lo que la expresión para la potencia quedaría como

$$\Pi_F(\Delta) \approx \Phi\left(\frac{mn\Delta f^*(0)}{\sqrt{\frac{mn(N+1)}{12}}} - \mu_\alpha\right) = \Phi\left(\sqrt{\frac{12mn}{N+1}}\Delta f^*(0) - \mu_\alpha\right)$$

$$= \Phi\left(\sqrt{\frac{12\,m}{N+1}}\,\Phi^*(0)\Delta - \mu_\alpha\right) = \Pi$$

3.2.1. Inverso

$$\sqrt{\frac{12\,m}{N+1}}\,\Phi^*(0)\Delta - \mu_\alpha = \Phi^{-1}(\Pi) \Longrightarrow \frac{12\,m}{N+1} = \frac{\left(\Phi^{-1}(\Pi) + \mu_\alpha\right)^2}{\left(\Phi^*(0)\Delta\right)^2}$$

Aproximamos asumiendo $m \simeq n$ y asumimos también N suficientemente grande para que $N \simeq N+1$:

$$n \simeq \frac{\left(\Phi^*(0)\Delta + \mu_\alpha\right)^2}{6\Delta^2\Phi^*(0)^2}$$

3.2.2. Intervalos de confianza para pares

Calculamos diferencias de nuestros dos:

$$D_{ij} = Y_i - X_j \quad \text{para todas las pares } i = 1, \ldots, m \quad j = 1, \ldots, n$$

Tomamos como estimador $\hat{\Delta} = \text{mediana}(D_{ij})$ ya que la mediana es robusta y menos sesgada.

3.2.3. Test de signos para muestras pareadas

Antes del tratamiento:

$$X = \{2, 4, 5, 6, 8\}$$

Después del tratamiento:

$$Y = \{3, 5, 7, 4, 10\}$$

Se calculan las diferencias:

$$D = \{3-2, 5-4, \dots\} = \{1, 1, 2, -2, 2\}$$

Si no hubiera diferencias, se deberían distribuir las diferencias positivas y negativas (y viceversa). El estadístico:

$$S \sim b(n, 0{,}5)$$

En un test bilateral, el p-valor:

$$p = 2 \cdot P(5 \leq \text{mín}(S^+, S^-))$$

Referencias

[1] Juan Camilo Yepes Borrero,
Apuntes Manuscritos Tema 5.
Universidad de Valladolid 2024.

[2] Yolanda Larriba González,
Apuntes INFE2 Tema 5.
Universidad de Valladolid 2023.