Métodos No Paramétricos - Final

Gonzalo Barrera Borla

Viernes 23/08/2019

Índice

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.
- $\mathrm{E}, \mathrm{Var}, \mathrm{cov}$ son los operadores esperanza, varianza y covarianza, respectivamente.
- $\Pr_Q(X) = \Pr\left(X \middle| Q\right)$ es la probabilidad de X condicional a Q. De ella se deriva E_Q.
- $X \perp Y$ indica que las VAs X, Y son independientes entre sí.

1. Ejercicios

1.1. Práctica 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, y dependerá de dónde queramos que caiga la carga de la prueba. La frase "su afirmación sera *aceptada*" (id est, *no rechazada*), nos sugiere que coloquemos la afirmación del empresario en la hipótesis nula. Esto equivale a declararlo "inocente hasta que se pruebe lo contario", y salvo que haya clara evidencia en contra, asumiremos que la proporción de frascos defectuosos está debajo del 10 %:

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

La estrategia tradicional, sin embargo, consiste en colocar "lo que se desea probar" en la hipótesis alternativa, de manera que sólo se acepte la afirmación si la evidencia está a su favor. Por razones que quedarán claras al resolver el punto (5), ésta parece ser la estrategia correcta en el ejercicio:

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

Sea $X_i = 1$ si el i-ésimo frasco de café tiene menos café del que garantiza la etiqueta, y 0 en caso contrario. Supongamos además que la probabilidad p de que un frasco contenga menos café del garantizado es constante e idéntica para todos, de manera que $X_i \overset{iid}{\sim} Ber(p)\ \forall i \in [n]$, y $\sum_{i=1}^{n} X_i = T \sim Bi(n, p)$, $n = 15$.

Bajo estos supuestos, dos formas equivalentes y matemáticamente tratables de las hipótesis son

$$H_0 \coloneqq p \geq 0{,}1 \quad vs. \quad H_1 \coloneqq p < 0{,}1$$

$$H_0 \coloneqq p \in \Theta_0 = [0{,}1, 1] \quad vs. \quad H_1 \coloneqq p \in \Theta_1 = [0, 0{,}1)$$

.

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

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

Y en particular, $n = 15, k = 2$ nos dan el test del enunciado, con región de rechazo $RR = \{0, 1, 2\}$. Sea $f_{n,k}(p) = \Pr\left(Bi(n, p) \leq k\right)$ la probabilidad acumulada a *izquierda* para una VA binomial, como función de p para $0 < k < n$ dados. $f_{n,k}$ es *estrictamente decreciente* en p, ya que a mayor p, menor es la probabilidad de que T realice un valor menor o igual a k , de manera que $\sup_{p \in [a,b]} f_{n,k}(p) = f_{n,k}(a)$. Luego,

$$\alpha = \sup_{p_0 \in [0{,}1, 1]} \mathrm{E}_{p_0}(\Phi) = \mathrm{Pr}_{0{,}1}(T \leq 2) = \sum_{i=0}^{2} \binom{15}{i} 0{,}1^i \ 0{,}9^{15-i} \approx 0{,}8159$$

En general, un test que rechace la hipótesis nula aún cuando ésta es verdadera (*id est,* acepte la afirmación del empresario aún cuando miente) 4 de cada 5 veces, no es razonable. 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_0, es simplemente la probabilidad de que T caiga en la región de rechazo

$$\Pr\left(\Phi = 1 \middle| p = p_0\right) = \mathrm{Pr}_{p_0}(T \leq k) = \Pr\left(Bi(n, p_0) \leq k\right) = \sum_{i=0}^{k} \binom{n}{i} p_0^i (1 - p_0)^{n-i}$$

$$\mathrm{Pr}_{0{,}05}(T \leq 2) = \sum_{i=0}^{2} \binom{15}{i} 0{,}05^i 0{,}95^{15-i} \approx 0{,}9638$$

$$\mathrm{Pr}_{0{,}1}(T \leq 2) \approx 0{,}8159$$

$$\mathrm{Pr}_{0{,}2}(T \leq 2) \approx 0{,}3980$$

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 ésta 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 hipótesis $H_0 : \theta \in \Theta_0 \quad vs. \quad H_1 :$

$\theta \in \Theta_1$

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

En nuestro caso particular, $\Theta_0 = [0{,}1, 1]$, $\Theta_1 = [0, 0{,}1)$ y $\mathrm{E}(\Phi) = f_{n,k}(p)$ es una función estrictamente *decreciente*. Como $\inf \Theta_0 = 0{,}1 < \sup \Theta_1$, entonces $\sup_{\theta \in \Theta_0} \mathrm{E}_\theta(\Phi) < \inf_{\theta \in \Theta_1} \mathrm{E}_\theta(\Phi)$ y Φ es insesgado para (H_0, H_1). propiamente dicha). En la figura 1 se intuye claramente la insesgadez.

La *potencia* de un test, es justamente la probabilidad de rechazar H_0:

- cuando $\theta \in \Theta_0$, el supremo de la función de potencia nos da el nivel de signifcación α del test, y
- cuando $\theta \in \Theta_1$, obtenemos la probabilidad de detectar la hipótesis alternativa (la potencia

Finalmente, debemos encontrar el mínimo n tal que $\alpha \leq 0{,}05$, manteniendo la misma $RR = \{0, 1, 2\}$. Si al comienzo del ejercicio nos hubiésemos decidido por el *set* de hipótesis que coloca la afirmación del empresario en H_0, la región de rechazo hubiese sido $RR = \{3, ..., 15\}$, ¡que rechaza para valores "centrales" de T! Aún siendo razonables y tomando $RR = \{3, ..., n\}$, consideremos el nivel de significación en función de n, $g_{k,p}(n) = \Pr_p(T \in RR) = \Pr(Bi(n,p) > k)$ para $0 < k < n$, $0 < p < 1$ dados. $g_{k,p}$ es estrictamente creciente en n, y ¡$\alpha_n = g_{k,p}(n) \overset{n \to \infty}{\longrightarrow} 1$! Evidentemente, esta última consigna sólo tiene sentido si planteamos las hipótesis con la afirmación del empresario en H_1.

Estamos buscando $\min n \ : \ \Pr(Bi(n,p) \leq k) = 1 - g_{k,p}(n) \leq \alpha$, y si g era estrictamente creciente, $1 - g$ será estrictamente decreciente en n. Además, ya vimos que $\Pr(Bi(n,p) \leq k)$ es *estrictamente decreciente* en p, y por ende alcanza su máximo en $p = \inf \Theta_0 = 0{,}1 \; \forall \, n$. Tomando $\alpha = 0{,}05, p = 0{,}1, k = 2$, podemos buscar n con la ayuda de R, y encontramos que

$$\Pr(Bi(60, 0{,}1) \leq 2) \approx 0{,}0530 > 0{,}05 > 0{,}0491 \approx \Pr(Bi(61, 0{,}1) \leq 2)$$

de manera que el mínimo tamaño muestral que nos garantiza $\alpha \leq 0{,}05$ dada $RR = \{0, 1, 2\}$, es $n = 61$, como se observa en la figura 2.

1.2. Práctica 2, ej. 2

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

Supongamos que $\forall i \in [n]$, $X_i \overset{iid}{\sim} F(t - \theta)$, $F \in \Omega_0$ (distribuciones absolutamente continuas con única mediana en 0) de manera que las X_i tienen única

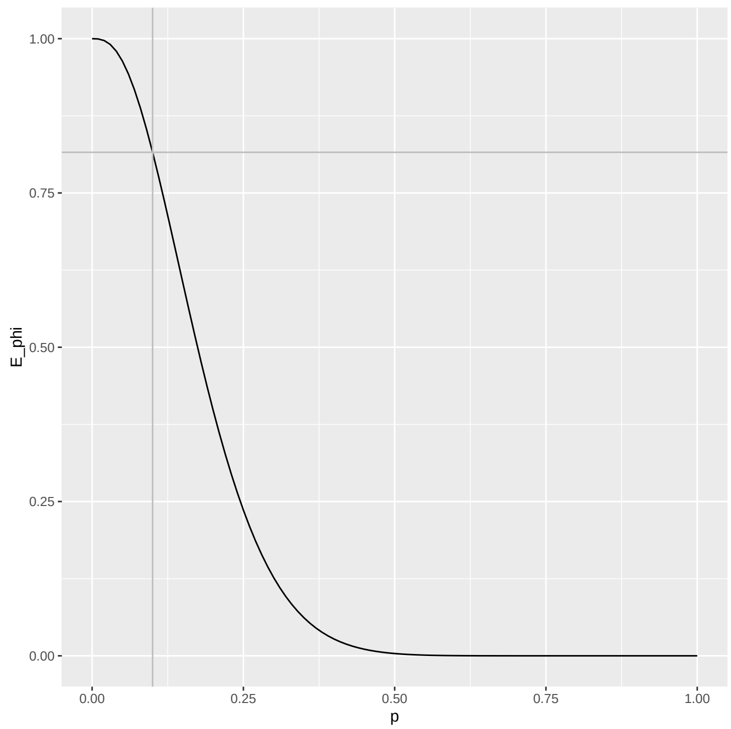

Figura 1: Potencia de Φ como función de p. La referencia vertical ($p = 0{,}1$) separa Θ_0, Θ_1 y corta la curva de potencia justo en el nivel de significación($\alpha \approx 0{,}8159$).

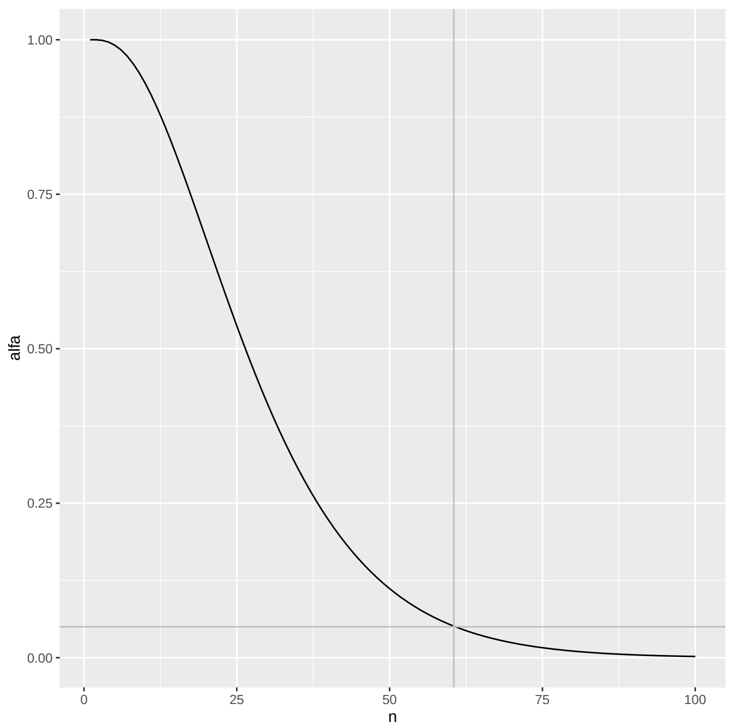

Figura 2: Nivel de significación α para Φ en función de n.

mediana en θ. Definamos el estadístico $S(\theta) = \#\{i : X_i - \theta > 0,\, i \in [n]\}$, y considerando que los conjuntos $\{X_1, \ldots, X_n\}$ y $\left\{X^{(1)}, \ldots, X^{(n)}\right\}$son idénticos, reescribámoslo como

$$S(\theta) = \#\{X_i - \theta > 0\} = \#\left\{X^{(i)} - \theta > 0\right\} = \#\left\{X^{(i)} > \theta\right\}$$

Así se ve claramente que $S(\theta)$ es una función continua a derecha, en escalera decreciente con saltos de tamaño 1 en los estadísticos de orden $X^{(i)}$. Visto de otra forma,

$$\begin{aligned} S(\theta) = n - k &\Leftrightarrow X^{(k)} \leq \theta < X^{(k+1)} \\ S(\theta) \leq k &\Leftrightarrow \theta \geq X^{(n-k)} \\ S(\theta) > k &\Leftrightarrow \theta < X^{(n-k)} \end{aligned}$$

Nótese que bajo H_0, $\forall (F, \theta, i) \in (\Omega_0 \times \mathbb{R} \times [n])$, $Y_i = \mathbb{I}(X_i - \theta > 0) \overset{iid}{\sim} Ber\left(\frac{1}{2}\right)$, y por ende $S(\theta) \sim Bi\left(n, \frac{1}{2}\right)$ y es simétrica alrededor de $\frac{n}{2}$. Luego, si elegimos $k : \Pr\left(Bi\left(n, \frac{1}{2}\right) \leq k\right) = \Pr\left(Bi\left(n, \frac{1}{2}\right) \geq n - k\right) = \frac{\alpha}{2}$,

$$\begin{aligned} 1 - \alpha =& 1 - \frac{\alpha}{2} - \frac{\alpha}{2} \\ =& \Pr(S \in \mathbb{R}) - \Pr(S \leq k) - \Pr(S \geq n - k) \\ =& \Pr(k < S < n - k) \\ \Pr(k < S < n - k) =& \Pr(S \leq n - (k+1) \cap S > k) \\ =& \Pr\left(\theta \geq X^{(n-[n-(k+1)])} \cap \theta < X^{(n-k)}\right) \\ =& \Pr\left(X^{(k+1)} \leq \theta < X^{(n-k)}\right) \end{aligned}$$

Y resulta que $\left[X^{(k+1)}, X^{(n-k)}\right)$ es un intervalo de confianza de nivel $1 - \alpha$ para θ. Como la distribución de S es simétrica alrededor de $\frac{n}{2}$, buscamos un estimador puntual $\hat{\theta} : S\left(\hat{\theta}\right) = \frac{n}{2}$. Graficando la función (por ejemplo, para $n = 9$),

vemos que cuando $\theta < X^{\left(\frac{n+1}{2}\right)} \Rightarrow S(\theta) > \frac{n-1}{2}$ y si $\theta \geq X^{\left(\frac{n+1}{2}\right)} \Rightarrow S(\theta) < \frac{n+1}{2}$. Luego, el valor más razonable del estimador puntual será la única mediana de la muestra, $\hat{\theta} = X^{\left(\frac{n+1}{2}\right)}$.

1.3. Práctica 3, ej. 8

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

1. 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 $P\left(g(X_1, ..., X_n) \leq \frac{\mu_0}{2} - t\right) = P\left(g(X_1, ..., X_n) \geq \frac{\mu_0}{2} + t\right)$.

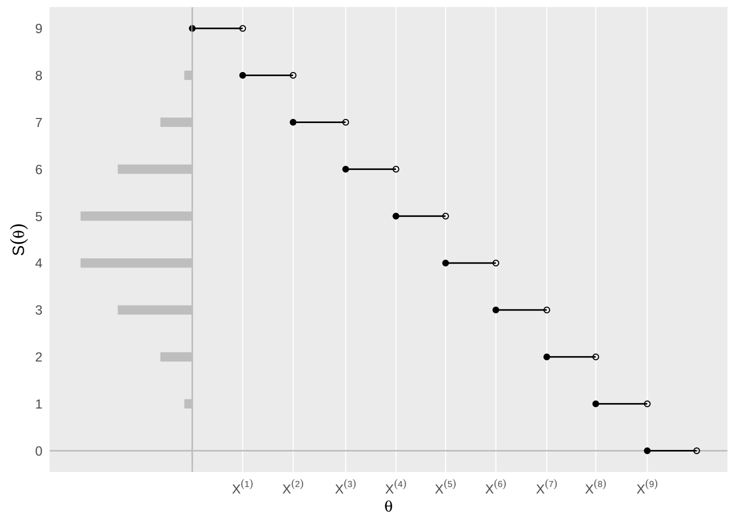

Figura 3: $S(\theta)$ para $n = 9$ (impar).

2. Aplique (1) 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?

Para simplificar la notación, consideremos el vector aleatorio $\bar{X} = (X_1, \dots, X_n)$. Además, por hipótesis $g\left(\bar{X}\right) = \mu_0 - g\left(-\bar{X}\right)$ y como $g\left(\bar{X}\right),\ g\left(-\bar{X}\right)$ tienen la misma distribución, $\Pr\left(g\left(\bar{X}\right) \leq k\right) = \Pr\left(g\left(-\bar{X}\right) \leq k\right)$.

$$\forall k \in \mathbb{R}, \Pr\left(g\left(\bar{X}\right) \leq k\right) = \Pr\left(\mu_0 - g\left(-\bar{X}\right) \leq k\right) = \Pr\left(\mu_0 - k \leq g\left(-\bar{X}\right)\right) = \Pr\left(\mu_0 - k \leq g\left(\bar{X}\right)\right)$$

Tomando $k = \frac{\mu_0}{2} - t$, $\Pr\left(g\left(\bar{X}\right) \leq \frac{\mu_0}{2} - t\right) = \Pr\left(\mu_0 - \frac{\mu_0}{2} - t \leq g\left(\bar{X}\right)\right) = \Pr\left(g\left(\bar{X}\right) \geq \frac{\mu_0}{2} + t\right)$ y por definición, g es simétrica alrededor de $\frac{\mu_0}{2}$, su mediana.

En particular, $T^+ = \sum_{i\in[n]} sg\left(X_i\right) R\left(X_i\right)$ *es función de* $\bar{X}$ y entre los supuestos del test, se encuentra que $\forall i \in [n],\ X_i \overset{iid}{\sim} F \in \Omega_s$. Luego, $T^+\left(\bar{X}\right), T^+\left(-\bar{X}\right)$ tienen la misma distribución[1]. Además, asumiendo que F es absolutamente continua, y considerando que $R_i = R\left(X_i\right) = R\left(-X_i\right)$ en la muestra ordenada por valores absolutos,

[1]Se puede probar que si $\forall i \in [n],\ X_i \overset{iid}{\sim} F \in \Omega_s,\ g\left(X_1, \dots, X_n\right) : \mathbb{R}^n \to \mathbb{R}$ tiene la misma distribución que $g\left(-X_1, \dots, -X_n\right)$ ya sea en general o al menos para el caso particular de T^+para completar la demostración, pero ya que el enunciado dice "usando...", lo tomamos como dado.

$$
\begin{aligned}
\frac{n(n+1)}{2} &= \sum_{i\in[n]} i = \sum_{i\in[n]} 1R_i \\
&= \sum_{i\in[n]} \left[\mathbb{I}\left(X_i>0\right) + \overbrace{\mathbb{I}\left(X_i=0\right)}^{=0} + \mathbb{I}\left(X_i<0\right) \right] R_i \\
&= \sum_{i\in[n]} \mathbb{I}\left(X_i>0\right) R\left(X_i\right) + \sum_{i\in[n]} \mathbb{I}\left(X_i<0\right) R\left(X_i\right) \\
&= \sum_{i\in[n]} \mathbb{I}\left(X_i>0\right) R\left(X_i\right) + \sum_{i\in[n]} \mathbb{I}\left(-X_i>0\right) R\left(-X_i\right) \\
&= \sum_{i\in[n]} sg\left(X_i\right) R\left(X_i\right) + \sum_{i\in[n]} sg\left(-X_i\right) R\left(-X_i\right) \\
\frac{n(n+1)}{2} &= T^+\left(\bar{X}\right) + T^+\left(-\bar{X}\right)
\end{aligned}
$$

Y T^+ tiene distribución simétrica alrededor de $\frac{n(n+1)}{4}$.

2. Demostraciones

2.1. Consistencia del test de Wilcoxon

Sea $\overrightarrow{X} = (X_1, ..., X_n)$ una muestra aleatoria tal que $\forall i \in [n]\,,\ X_i \overset{iid}{\sim} F\left(t-\theta\right), F \in \Omega_s$ (distribuciones simétricas con única mediana en 0), y las hipótesis a testear $H_0 \coloneqq \theta = 0 \quad vs. \quad H_1 \coloneqq \theta > 0$. Definamos

- el rango de $|X_i|$ en la muestra ordenada segun valores absolutos: $R\left(X_i\right) = R_i = \#\{j : |X_j| \leq |X_i|\,,\ j \in [n]\}$,
- el *antirrango* $D_j = i \Leftrightarrow R_i = j$, la "función inversa" de R_i, que dado el rango j de un elemento en la muestra ordenada por valores absolutos, devuelve el índice en la muestra original.
- la "función signo" $sg\left(x\right) = \mathbb{I}\left(x>0\right)$, y
- el estadistico $T^+ = \sum_{i\in[n]} R_i\, sg\left(X_i\right) = \sum_{j\in[n]} j\, W_j$, donde $W_j = sg\left(X_{D_j}\right)$

Nótese que con F absolutamente continua no hay empates, y $D_{R_i} = i,\ R_{D_j} = j,\ [n] = \{1, \ldots, n\} = \{R_1, \ldots, R_n\} = \{D_1, \ldots, D_n\}$.

Para probar la consistencia de $\Phi_n = \mathbb{I}\left(T_n^+ > k_n\right)$ para las hipótesis usando el Teorema 6 del anexo, necesitamos las siguientes condiciones:

- $T_n^+ \xrightarrow{p} \mu\left(G\right) = \begin{cases} \mu_0 & \forall\, G \in \Omega_0 \\ > \mu_0 & \forall\, G \in \Omega_c \subset \Omega_1 \end{cases}$
- Bajo $H_{0,}\ \exists \sigma_0 :\ \sqrt{n}\frac{T^+ - \mu_0}{\sigma_0} \xrightarrow{D} N\left(0,1\right)$

Para probar la normalidad asintótica, veremos que bajo H_0, T^+se puede expresar como combinación lineal de variables independientes entre sí. Para la convergencia en probabilidad de T^+, encontraremos la esperanza y varianza exactas del estadístico recurriendo a los promedios de Walsh.

2.1.1. Normalidad asintótica de T^+bajo H_0

Teorema 1. *Bajo $H_0 : \theta = 0,\ F \in \Omega_s$ las VA $sg\left(X_1\right), \ldots, sg\left(X_n\right)$y el vector $\left(R_1, \ldots, R_n\right)$ son mutuamente independientes.*

Demostración. Como $X_i \perp X_j \forall i \neq j$ por definición, y las funciones $\left|X_i\right|, sg\left(X_i\right)$ sólo dependen de X_i, los pares $\left(sg\left(X_i\right), \left|X_i\right|\right) \perp \left(sg\left(X_j\right), \left|X_j\right|\right) \forall i \neq j$. Recordemos que como $\theta = 0,\ F \in \Omega_s \Rightarrow F\left(-t-0\right) = 1 - F\left(t-0\right)$, y veamos que $sg\left(X_i\right) \perp \left|X_i\right|$:

$$
\begin{aligned}
\Pr\left(sg\left(X_i\right) = 0 \cap \left|X_i\right| \leq k\right) &= \Pr\left(X_i \leq 0 \cap -k \leq X_i \leq k\right) \\
&= \Pr\left(-k \leq X_i \leq 0\right) \\
&= F\left(0\right) - F\left(-k\right) && \Pr\left(X_i = -k\right) = 0 \\
&= \frac{1}{2}\left(1 - 2F\left(-k\right)\right) \\
&= \frac{1}{2}\left[F\left(k\right) - F\left(-k\right)\right] \\
&= \Pr\left(X_i > 0\right)\Pr\left(-k \leq X_i \leq k\right) \\
\Pr\left(sg\left(X_i\right) = 0 \cap \left|X_i\right| \leq k\right) &= \Pr\left(sg\left(X_i\right)\right)\Pr\left(\left|X_i\right| \leq k\right)
\end{aligned}
$$

Un planteo similar para $sg\left(X_i\right) = 1$ nos permite concluir que $sg\left(X_i\right) \perp \left|X_i\right|$. Como $R_k = f\left(\left|X_1\right|, \ldots, \left|X_n\right|\right),\ sg\left(X_i\right) \perp \left|X_j\right| \forall j \in [n] \Rightarrow sg\left(X_i\right) \perp R_k \forall i, k \in [n]^2$. Análogamente, $D_k = g\left(\left|X_1\right|, \ldots, \left|X_n\right|\right) \Rightarrow sg\left(X_i\right) \perp D_k \forall i, k \in [n]^2$. □

Teorema 2. *Bajo $H_0 : \theta = 0,\ F \in \Omega_s,\ sg\left(X_{D_j}\right) = W_j \overset{iid}{\sim} Ber\left(\frac{1}{2}\right)$.*

Demostración. Sea $D = \left(D_1, \ldots, D_n\right)$ el vector de antirrangos como variables aleatorias, $d = \left(d_1, \ldots, d_n\right)$ una de las $n!$ configuraciones posible de ellos y $s = \left(s_1, \ldots, s_n\right) \in \{0, 1\}^n$. Por el teorema de la probabilidad total,

$$
\begin{aligned}
\Pr\left(\bigcap_{i\in[n]} sg\left(X_{D_i}\right)=s_i\right) &= \sum_d\left[\Pr\left(\bigcap_{i\in[n]} sg\left(X_{D_i}\right)=s_i\middle|D=d\right)\Pr\left(D=d\right)\right]\\
&= \sum_d\left[\Pr\left(\bigcap_{i\in[n]} sg\left(X_{d_i}\right)=s_i\right)\Pr\left(D=d\right)\right]\\
&= \sum_d\left[\left(\prod_{i\in[n]}\Pr\left(sg\left(X_{d_i}\right)=s_i\right)\right)\Pr\left(D=d\right)\right] && sg\left(X_i\right)\perp sg\left(X_j\right)\\
&= \sum_d\left[\left[\prod_{i\in[n]}\frac{1}{2}\right]\Pr\left(D=d\right)\right]\\
&= \left(\frac{1}{2}\right)^n\overbrace{\sum_d\left[\Pr\left(D=d\right)\right]}^{=1}\\
\Pr\left(\bigcap_{i\in[n]} sg\left(X_{D_i}\right)=s_i\right) &= 2^{-n}\,\forall\left(s_1,\ldots,s_n\right)\in\{0,1\}^n
\end{aligned}
$$

Es decir, las 2^nposibles configuraciones de signos son equiprobables, y $\forall\left(s_1,\ldots,s_n\right)\in\{0,1\}^n$, $\Pr\left(\bigcap_{i\in[n]}W_i=s_i\right)=\prod_{i\in[n]}\Pr\left(W_i=s_i\right)=2^{-n}$ y $\Pr\left(W_i=1\right)=\Pr\left(W_i=0\right)=\frac{1}{2}\,\forall i\in[n]$. Luego, $W_i\overset{iid}{\sim}Ber\left(p\right)\,\forall i\in[n]$. $\square$

Hemos alcanzado una expresión de T^+como suma de VA independientes $\left(\sum_j\ j\cdot W_j\right)$, pero no idénticamente distribuidas. El siguiente teorema es la pieza faltante para encontrar la distribución asintótica.

Teorema 3. *Sean* $X_i\overset{iid}{\sim}\mathrm{E}\left(X_1\right)=0$, $\mathrm{Var}\left(X_1\right)=\sigma^2\in(0,\infty)$, *y defínase* $T=n^{-\frac{1}{2}}\sum_{i\in[n]}a_iX_i$. *Si*

$$\frac{\max|a_i|}{\sqrt{\sum_{i\in[n]}a_i^2}}\to 0$$

entonces $\frac{T}{\sqrt{\mathrm{Var}(T)}}\xrightarrow{D}N\left(0,1\right)$ *con* $\mathrm{Var}\left(T\right)=n^{-1}\sigma^2\sum_{i\in[n]}a_i^2$.

Demostración. Es una variante del Teorema Central del Límite de Lindeberg. Cf. Teorema A9 en Hettmansperger, p. 301. $\square$

Sean $X_i=W_i-\frac{1}{2}$, $a_i=i$. Luego, $\mathrm{E}\left(X_i\right)=\mathrm{E}\left(W_i\right)-\frac{1}{2}=0$ y $\mathrm{Var}\left(X_i\right)=$

$\mathrm{Var}\,(W_i) = \frac{1}{4} = \sigma^2 \in (0, \infty)$. Además, cuando $n \to \infty$

$$\frac{\text{máx}\,|a_i|}{\sqrt{\sum_{i\in[n]} a_i^2}} = \frac{\text{máx}\,|i|}{\sqrt{\sum_{i\in[n]} i^2}} = \frac{n}{\sqrt{\frac{n(n+1)(2n+1)}{6}}} = \frac{n}{\sqrt{\frac{1}{3}n^3 + \frac{1}{2}n^2 + \frac{1}{6}n}}$$

$$= \frac{\cancel{n}}{\cancel{n}\sqrt{n}\underbrace{\sqrt{\frac{1}{3} + \frac{1}{2n} + \frac{1}{6n^2}}}_{\to\sqrt{1/3}}} = \sqrt{\frac{3}{n}} \to 0$$

Luego,

$$T = n^{-\frac{1}{2}} \sum_{i\in[n]} a_i X_i = n^{-\frac{1}{2}} \sum_{i\in[n]} i\left(W_i - \frac{1}{2}\right) = \frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{n}}$$

$$\mathrm{Var}\,(T) = n^{-1}\sigma^2 \sum_{i\in[n]} a_i^2 = \frac{\sum_{i\in[n]} i^2}{4n} = \frac{(n+1)\,(2n+1)}{24}$$

Y finalmente,

$$\frac{T}{\sqrt{\mathrm{Var}\,(T)}} = \frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \xrightarrow{D} N\,(0,1)$$

2.1.2. Convergencia en Probabilidad de $\overline{T}$

Recordemos que T^+ admite la siguiente definición en base a los promedios de Walsh,

$$T^+ = \#\left\{i,j : \frac{X_i + X_j}{2} > 0,\ 1 \le i \le j \le n\right\} = \sum_{i=1}^{j}\sum_{j=1}^{n} \mathbb{I}\,(X_i + X_j > 0)$$

$$T^+ = \sum_{i=1}^{j}\sum_{j=1}^{n} T_{ij}$$

donde $T_{ij} = (X_i + X_j > 0)$. Utilicemos esta ultima forma para desarrollar la esperanza y varianza de T^+, dada cierta distribucion subyacente fija $F_0 \in \Omega_0(F_0 \in \Omega_s,\ \theta = 0)$. Gracias a la linealidad de la esperanza, y el hecho de que $\mathrm{E}\,(\mathbb{I}\,(Q)) = \Pr\,(Q)$,

$$\mathrm{E}\left(T^+\right) = \mathrm{E}\left[\sum_{i=1}^{j}\sum_{j=1}^{n} \mathbb{I}\,(X_i + X_j > 0)\right] = \sum_{i=1}^{j}\sum_{j=1}^{n} \mathrm{E}\left[\mathbb{I}\,(X_i + X_j > 0)\right] = \sum_{i=1}^{j}\sum_{j=1}^{n} \Pr\,(X + X_j > 0)$$

$$\mathrm{E}\left(T^+\right) = \sum_{i=j} \Pr\,(X_i > 0) + \sum_{i<j} \Pr\,(X_i + X_j > 0)$$

Sean ahora $p_1 = \Pr(X_i > 0)$, $p_2 = \Pr\left(X_i + X_j > 0 \middle| i \neq j\right)$ (cuando $i = j, \Pr(X_i + X_j > 0) = \Pr(2X_i > 0) = p_1$). Se observa que hay n sumandos donde $i = j$ y $\frac{(n-1)n}{2}$ sumandos donde $i < j$. Luego, $E(T^+) = np_1 + \frac{(n-1)n}{2}p_2$.

Para el cálculo de la varianza de T^+, usaremos el hecho de que, en general, si $Y = \sum_i X_i$, entonces

$$
\begin{aligned}
\operatorname{Var}(Y) = \operatorname{cov}(Y, Y) &= \operatorname{cov}\left(\sum_i X_i, \sum_i X_i\right) \\
&= \sum_i \sum_j \operatorname{cov}(X_i, X_j) \\
&= \sum_i \operatorname{Var}(X_i) + 2\sum_{i<j} \operatorname{cov}(X_i, X_j)
\end{aligned}
$$

Para ello, tendremos que encontrar la expresion de la covarianza para todas las combinaciones unicas de subindices.

$$
\operatorname{Var}\left(T^+\right) = \operatorname{Var}\left(\sum_{i\leq j} T_{ij}\right) = \sum_{i\leq j}\sum_{k\leq l} \operatorname{cov}(T_{ij}, T_{kl})
$$

Recordemos que si $X \perp Y \Rightarrow \mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y) \Rightarrow \operatorname{cov}(X, Y) = 0$. Además, para funciones deterministicas arbitrarias f, g, si $X \perp Y \Rightarrow f(X) \perp g(Y)$. 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 índice, $T_{ij} \perp T_{kl} \Rightarrow \operatorname{cov}(T_{ij}, T_{kl}) = 0$. ¿Cuántos pares (T_{ij}, T_{kl}) comparten algún índice en la expresión de $\operatorname{Var}(T^+)$? Tengamos en cuenta que cada elemento T_{ij} figura en exactamente $\frac{n(n+1)}{2}$ términos, "cruzado" una vez con cada uno de los promedios de Walsh:

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

Cuando $i < j$, T_{ij} comparte 2 índices con T_{ij}, T_{ii}, T_{jj}y 1 índice con

$$
\overbrace{(i-1)}^{col\ i} + \overbrace{(n-i-1)}^{fila\ i} + \overbrace{(j-2)}^{col\ j} + \overbrace{(n-j)}^{fila\ j} = 2(n-2)
$$

elementos. Cuando $i = j$, T_{ii} comparte 2 índices con T_{ii}, y 1 índice con

$$
\overbrace{(i-1)}^{col\ i} + \overbrace{(n-i)}^{fila\ i} = n-1
$$

elementos. Luego, sólo tenemos que preocuparnos por términos de las formas

$$\operatorname{cov}(T_{11},T_{11}) = \operatorname{Var}T_{11} \quad \operatorname{cov}(T_{12},T_{12}) = \operatorname{Var}T_{12} \quad \operatorname{cov}(T_{11},T_{12}) \quad \operatorname{cov}(T_{12},T_{13})$$

Y de lo expuesto,

$$\begin{aligned}
\forall i \neq j, \quad & \sum_{k \leq l} \operatorname{cov}(T_{ij},T_{kl}) = \operatorname{Var}(T_{12}) + 2\operatorname{cov}(T_{11},T_{12}) + 2(n-2)\operatorname{cov}(T_{12},T_{13}) \\
\forall i = j, \quad & \sum_{k \leq l} \operatorname{cov}(T_{ii},T_{kl}) = \operatorname{Var}(T_{11}) + (n-1)\operatorname{cov}(T_{11},T_{12})
\end{aligned}$$

Entonces,

$$\begin{aligned}
\operatorname{Var}\left(T^{+}\right) =& \sum_{i=j} \sum_{k \leq l} \operatorname{cov}(T_{ij},T_{kl}) + \sum_{i<j} \sum_{k \leq l} \operatorname{cov}(T_{ij},T_{kl}) \\
=& n\left(\sum_{k \leq l} \operatorname{cov}(T_{ii},T_{kl})\right) + \frac{n(n-1)}{2}\left(\sum_{k \leq l} \operatorname{cov}(T_{ij},T_{kl})\right) \\
=& \frac{n(n-1)}{2}\operatorname{Var}(T_{12}) + 2\frac{n(n-1)}{2}\operatorname{cov}(T_{11},T_{12}) + 2\frac{n(n-1)}{2}(n-2)\operatorname{cov}(T_{12},T_{13}) \\
& + n\operatorname{Var}(T_{11}) + n(n-1)\operatorname{cov}(T_{11},T_{12}) \\
\operatorname{Var}\left(T^{+}\right) =& n\operatorname{Var}(T_{11}) + \frac{n(n-1)}{2}\operatorname{Var}(T_{12}) + 2n(n-1)\operatorname{cov}(T_{11},T_{12}) \\
& + n(n-1)(n-2)\operatorname{cov}(T_{12},T_{13})
\end{aligned}$$

Recordemos las definiciones de $p_1 = \Pr(X_1 > 0)$, $p_2 = \Pr(X_1 + X_2 > 0)$. Usando que $\operatorname{cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$ y $\mathbb{I}(P) \times \mathbb{I}(Q) = \mathbb{I}(P \wedge Q)$ (y por ende $[\mathbb{I}(P)]^2 = \mathbb{I}(P \wedge P) = \mathbb{I}(P)$), calculamos

$$
\begin{aligned}
\operatorname{Var}(T_{11}) &= \mathrm{E}T_{11}^2) - \mathrm{E}(T_{11})^2 = \mathrm{E}(\mathbb{I}(X_1 > 0)^2) - \mathrm{E}(\mathbb{I}(X_1 > 0))^2 \\
&= \mathrm{E}(\mathbb{I}(X_1 > 0))(1 - \mathrm{E}(\mathbb{I}(X_1 > 0))) \\
&= \Pr(X_1 > 0)(1 - \Pr(X_1 > 0)) \\
&= p_1(1 - p_1) \\
\operatorname{Var}(T_{12}) &= \mathrm{E}(T_{12}^2) - \mathrm{E}(T_{12})^2 = \mathrm{E}(\mathbb{I}(X_1 + X_2 > 0)^2) - \mathrm{E}(\mathbb{I}(X_1 + X_2 > 0))^2 \\
&= \mathrm{E}(\mathbb{I}(X_1 + X_2 > 0))(1 - \mathrm{E}(\mathbb{I}(X_1 + X_2 > 0))) \\
&= \Pr(X_1 + X_2 > 0)(1 - \Pr(X_1 + X_2 > 0)) \\
&= p_2(1 - p_2) \\
\operatorname{cov}(T_{11}, T_{12}) &= \mathrm{E}(T_{11}T_{12}) - \mathrm{E}(T_{11})\mathrm{E}(T_{12}) \\
&= \mathrm{E}(\mathbb{I}(X_1 > 0 \wedge X_1 + X_2 > 0)) - \mathrm{E}(\mathbb{I}(X_1 > 0)\mathrm{E}(\mathbb{I}(X_1 + X_2 > 0)) \\
&= \Pr(X_1 > 0 \wedge X_1 + X_2 > 0) - p_1 \cdot p_2 \\
&= p_3 - p_1 \cdot p_2 \quad \text{, digamos} \\
\operatorname{cov}(T_{12}, T_{13}) &= \mathrm{E}(T_{12}T_{13}) - \mathrm{E}(T_{12})\mathrm{E}(T_{13}) \\
&= \mathrm{E}(\mathbb{I}(X_1 + X_2 > 0 \wedge X_1 + X_3 > 0)) - \mathrm{E}(\mathbb{I}(X_1 + X_2 > 0)\mathrm{E}(\mathbb{I}(X_1 + X_3 > 0)) \\
&= \Pr(X_1 + X_2 > 0 \wedge X_1 + X_3 > 0) - p_2^2 \\
&= p_4 - p_2^2 \quad \text{, digamos}
\end{aligned}
$$

Finalmente,

$$
\operatorname{Var}\left(T^+\right) = np_1\left(1 - p_1\right) + \frac{n\left(n-1\right)}{2} p_2\left(1 - p_2\right) + 2n\left(n-1\right)\left(p_3 - p_1 \cdot p_2\right) + n\left(n-1\right)\left(n-2\right)\left(p_4 - p_2^2\right)
$$

Consideremos la esperanza y varianza del estadístico $\overline{T} = \frac{T^+}{\frac{n(n+1)}{2}}$:

$$
\begin{aligned}
\mathrm{E}\left(\overline{T}\right) &= \mathrm{E}\left(\frac{T^+}{\frac{n(n+1)}{2}}\right) = \frac{2}{n\left(n+1\right)}\mathrm{E}\left(T^+\right) = \frac{2}{n\left(n+1\right)}\left(np_1 + \tfrac{(n-1)n}{2}p_2\right) \\
\mathrm{E}\left(\overline{T}\right) &= \tfrac{2}{n+1}p_1 + \tfrac{n-1}{n+1}p_2 \overset{n\to\infty}{\longrightarrow} p_2 \\
\operatorname{Var}\left(\overline{T}\right) &= \tfrac{4}{n^2(n+1)^2}\operatorname{Var}\left(T^+\right) \\
\operatorname{Var}\left(\overline{T}\right) &= n^{-3}k_1 + n^{-2}k_2 + n^{-1}k_3 \overset{n\to\infty}{\longrightarrow} 0
\end{aligned}
$$

donde k_1, k_2, k_3son las expresiones constantes e independientes de n que resultan al distribuir el coeficiente $\frac{4}{n^2(n+1)^2}$ entre la cornucopia de términos de $\operatorname{Var}(T^+)$. En consecuencia, $\overline{T} \xrightarrow{p} p_2$[2]. Nos resta calcular p_2 para las distribuciones $F \in \Omega_0,\ F \in \Omega_c \subset \Omega_1$.

---

[2]Resulta de una aplicación directa de Chebyshev a $Z = \overline{T} - \mathrm{E}\overline{T}$

$$p_2 = \Pr(X_1 + X_2 > 0) = \Pr(X_1 > -X_2) = \iint \mathbb{I}(x_1 > -x_2) f(x_1) f(x_2)\, dx_2 dx_1$$

$$= \iint_{-x_2}^{\infty} f(x_1) f(x_2)\, dx_1 dx_2 = \int f(x_2) \left[F(x) \Big|_{-x_2}^{\infty} \right] dx_2 = \int f(x_2) \left[1 - F(-x_2)\right] dx_2$$

Bajo $H_0, F \in \Omega_s, \theta = 0 \Rightarrow 1 - F(-x) = F(x)$, de manera que $f(x)[1 - F(-x)] = f(x) F(x)$, que tiene como primitiva a $\frac{1}{2} F(x)^2$. Luego,

$$p_2 = \int f(x_2)[1 - F(-x_2)]\, dx_2 = \tfrac{1}{2}\left(F(x)^2 \Big|_{-\infty}^{\infty} \right) = \frac{1}{2}$$

Consideremos ahora la clase de distribuciones *estocasticámente positivas* Ω_p. Diremos que $G \in \Omega_p$si $\forall k \in \mathbb{R},\ 1 - G(k) \geq G(-k)$, con desigualdad estricta en un conjunto de medida no-nula. Si $X_i \overset{iid}{\sim} G \in \Omega_p$,

$$p_2 = \int g(x_2)[1 - G(-x_2)]\, dx_2 > \int g(x_2)[G(x_2)]\, dx_2 = \tfrac{1}{2}\left(G(x)^2 \Big|_{-\infty}^{\infty} \right) = \frac{1}{2}$$

y luego

$$\overline{T} \to p_2(F, \theta) = \begin{cases} 1/2 & \text{si } F \in \Omega_0\, (F \in \Omega_s \wedge \theta = 0) \\ > 1/2 & \text{si } F \in \Omega_p \end{cases}$$

Vale notar que cuando $F \in \Omega_s,\ \theta > 0$, la desigualdad que caracteriza a Ω_p se cumple estrictamente en todo punto

$$1 - F(t) > 1 - F(t + \theta) = F(\theta - t) > F(t)$$

y si $\Omega_> = \{F \text{ con única mediana en } \theta{>}0\} \Rightarrow \Omega_1 = (\Omega_s \cap \Omega_>) \subset \Omega_p$. Luego, el test basado en $\overline{T}$ (o equivalentemente, T^+) será test consistente para las hipótesis originales.

2.2. Teorema de Proyección

Teorema 4. *(de Proyección) Sea $X_i, \ldots, X_n$ una muestra aleatoria con distribución arbitrataria $X_i \overset{iid}{\sim} F\ \forall i \in [n]$, y $V = f(X_1, \ldots, X_n)$ una variable aleatoria con* $\mathrm{E}V = 0$. *Si W es una variable aleatoria de la forma $W = \sum_i g_i(X_i)$, entonces $\mathrm{E}\left[(V - W)^2\right]$ se minimiza eligiendo las funciones $g_i = g_i^\star$:*

$$g_i^\star(x) = \mathrm{E}\left(V \middle| X_i = x\right)$$

$V_p = \sum_i g_i^\star(X_i)$ se denomina la ***proyección de V*** *sobre la clase de VAs independientes, y se cumple que*

$$\mathrm{E}\left[(V - V_p)^2\right] = \mathrm{Var}(V) - \mathrm{Var}(V_p)$$

Demostración. Sumando y restando V_p,

$$\begin{aligned}
\mathrm{E}\left[(V-W)^2\right] =& \mathrm{E}\left[((V-V_p)+(V_p-W))^2\right]\\
=& \mathrm{E}\left[(V-V_p)^2\right]+\mathrm{E}\left[(V_p-W)^2\right]+2\mathrm{E}\left(V-V_p\right)\mathrm{E}\left(V_p-W\right)
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\left(V-V_p\right) =& \mathrm{E}\left(V-\sum_i g_i^\star\left(X_i\right)\right) = \mathrm{E}\left(V-g_i^\star\left(X_i\right)-\sum_{i\neq j} g_j^\star\left(X_j\right)\right)\\
=& \mathrm{E}\left[\mathrm{E}\left(V-g_i^\star\left(X_i\right)-\sum_{i\neq j} g_j^\star\left(X_j\right)\right)\Big| X_i\right]\\
=& \mathrm{E}\left[\mathrm{E}\left(V\middle|X_i\right)-\mathrm{E}\left(g_i^\star\left(X_i\right)\right)-\sum_{i\neq j}\mathrm{E}\left(g_j^\star\left(X_j\right)\middle|X_i\right)\right]\\
=& \mathrm{E}\left[\cancel{\mathrm{E}\left(V\middle|X_i\right)}-\cancel{\mathrm{E}\left(V\middle|X_i\right)}-\sum_{i\neq j}\mathrm{E}\left(g_j^\star\left(X_j\right)\right)\right] \qquad g_j^\star\left(X_j\right)\perp X_i\\
=& \mathrm{E}\left[-\sum_{i\neq j}\mathrm{E}\left(\mathrm{E}\left(V\middle|X_j\right)\right)\right] = \mathrm{E}\left[-\sum_{i\neq j}\mathrm{E}V\right] = 0
\end{aligned}$$

Retomando la expresion anterior,

$$\begin{aligned}
\mathrm{E}\left[(V-W)^2\right] =& \mathrm{E}\left[(V-V_p)^2\right]+\mathrm{E}\left[(V_p-W)^2\right]+\cancel{2\times 0\times\mathrm{E}\left(V_p-W\right)}\\
\mathrm{E}\left[(V-W)^2\right] \geq& \mathrm{E}\left[(V-V_p)^2\right]
\end{aligned}$$

con igualdad estricta $\Leftrightarrow W = V_p$. Ademas, $\mathrm{E}V_p = \sum_i \mathrm{E}\left[\mathrm{E}\left(V\middle|X_i\right)\right] = n\mathrm{E}V = 0$, de manera que tomando $W = 0$, obtenemos la expresión de las varianzas:

$$\mathrm{E}V^2 = \mathrm{E}\left(V-V_p\right)^2+\mathrm{E}V_p^2 \Rightarrow \mathrm{E}\left(V-V_p\right)^2 = \mathrm{Var}V-\mathrm{Var}V_p$$

$\square$

En general, bajo la hipótesis nula, los estadísticos de los test no paramétricos que consideramos se pueden expresar como una suma (de funciones) de variables aleatorias independientes entre sí, y contamos con toda una artillería para probar la normalidad asintótica de expresiones bajo esas condiciones. Sin embargo, bajo la hipótesis alternativa es común perder la independencia entre dichas expresiones, y por ende necesitamos una expresión alternativa de los estadísticos sobre la que podamos dar alguna aproximación en grandes números.

Intuitivamente, el teorema de proyección nos ofrece una solucion elegante. Primero, hay que notar que $\forall i,\ g_i^\star(X_i) = \mathrm{E}\left(V \middle| X_i\right)$ es una VA que *depende únicamente de* X_i, entonces $\forall i \neq j,\ g_i^\star(X_i) \perp g_j^\star(X_j)$ y $V = \sum_i g_i^\star$es una función lineal de VA independientes con $\mathrm{E} = 0$. Luego, a medida que $n \to \infty$, podemos pensar que salvo casos patológicos, el espacio de las $g_i^\star(X_i)$ contará con suficiente información para que $V_p \approx V$ y $\mathrm{E}\left(V - V_p\right)^2 \to 0$, de manera que si $f(V_p) \xrightarrow{D} N(0,1)$, $f(V)$ también lo haga[3]. Con la distribución asintótica bajo la alternativa, podemos explorar características del estimador como su eficiencia relativa y potencia asintóticas.

2.3. Distribución exacta de los estadísticos del test de Mann-Whitney-Wilcoxon (MWW) bajo H_0

Sean $X_1 \dots, X_n$ y $Y_1, \dots, Y_m$ dos muestras aleatorias tales que $\forall i \in [n],\ X_i \overset{iid}{\sim} F(t - \theta_x),\ \forall j \in [m],\ Y_j \overset{iid}{\sim} F(t - \theta_y),\ F \in \Omega_0$. No hace falta suponer F simétrica, pero sí que la diferencia entre ambas distribuciones es únicamente de posición, no de forma. Sea $\Delta = \theta_x - \theta_y$, y se desea testear $H_0 \coloneqq \Delta = 0\ vs.\ H_1 \coloneqq \Delta > 0$. Wilcoxon (1945) plantea los siguientes estadísticos análogos a T^+ en el problema de una muestra:

$$U = \sum_{j \in [m]} R(Y_j) \qquad T = \sum_{i \in [n]} R(X_i)$$

donde $R(W_k)$ es el rango del k-ésimo elemento de $W \in \{X, Y\}$ en la muestra combinada $(X_1, \dots, X_n, Y_1, \dots, Y_m)$ de tamaño $N = m + n$. Cuando $F \in \Omega_0$ es absolutamente continua y no hay empates, por lo cual $U + T = \frac{N(N+1)}{2}$, y basta con conocer la distribución exacta de uno de los dos estadísticos para conocer la de ambos.

Mann y Whitney (1947) proponen un estadístico equivalente con forma "contable":

$$W = \sum_{j \in [m]} \sum_{i \in [n]} \mathbb{I}(Y_j > X_i)$$

Si $R_j = R(Y_j)$ el rango de Y_j en la muestra combinada, y R'_j expresa el rango de Y_j en la muestra $Y_1, \dots, Y_m$,

$$\begin{aligned}
R(Y_k) &= \#\{i : X_i < Y_k, i \in [n]\} + \#\{j : Y_j \leq Y_k, j \in [m]\} \\
R(Y_k) &= \sum_{i \in [n]} \mathbb{I}(Y_j > X_i) + R'_k \\
\sum_{k \in [m]} R(Y_k) &= \sum_{k \in [m]} \sum_{i \in [n]} \mathbb{I}(Y_j > X_i) + \sum_{k \in [m]} R'_k \\
U &= W + \tfrac{m(m+1)}{2}
\end{aligned}$$

[3]De más esta decir que se deben cuantificar formalmente estas afirmaciones, y de hecho eso lo que debemos hacer al aplicar este teorema sobre estadísticos particulares.

Nos abocamos luego a derivar la distribución exacta de U, que será idéntica a las de T, W salvo por un desplazamiento y/o rotación.

Bajo H_0, la muestra combinada $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ tiene $N = m + n$ elementos con idéntica distribución, de manera que cualquiera de las $\binom{N}{m} = \frac{(n+m)!}{n! \cdot m!}$ formas de elegir m posiciones para las Y_j entre los N rangos de la muestra combinada es equiprobable. Basta con contar qué valor toma U en cada combinación, y tenemos la distribución exacta.

Por ejemplo, si $n = m = 2 \Rightarrow N = 4$ y hay $\binom{4}{2} = 6$ formas de elegir 2 elementos $(R(Y_1), R(Y_2))$ entre 4 $(\{1,2,3,4\})$: $(1,2), (1,3), (1,4), (2,3), (2,4), (3,4)$. Cada configuración es equiprobable con $p = \frac{1}{6}$ y U vale, respectivamente, $3, 4, 5, 5, 6, 7$. Finalmente,

$$p_U\left(k \middle| n = m = 2\right) = \begin{cases} 1/6 & \text{si } k \in \{3,4,6,7\} \\ 1/3 & \text{si } k = 5 \\ 0 & \text{en otro caso} \end{cases}$$

Este trabajo "casero" se vuelve tedioso rápidamente, y como buen matemático vago uno se siente motivado a encontrar alguna definición recursiva fácilmente programable. Ya vimos que aún sin hacer suposición alguna sobre la forma de F, las $\binom{N}{m}$ combinaciones de rangos son equiprobables, de manera que calcular $p_{m,n}(k) = \Pr_{H_0}\left(U = k \middle| m, n\right)$ se reduce a contar de cúantas maneras se puede sumar k con m enteros distintos menores o iguales a $N = m + n$[4]. Llamémosle $\Delta(k, m, n)$ a esta cantidad[5]. Todos los conjuntos de m elementos distintos menores o iguales a N se pueden dividir de manera mutuamente excluyente y conjuntamente exhaustiva entre aquellos que:

- contienen a $N = m + n$ como elemento (y los $m - 1$ restantes son menores o iguales a $N - 1$), y los que
- no contienen a N (y sus m elementos son menores o iguales a $N - 1$). Luego,

$$\Delta(k, m, n) = \Delta(k - (m+n), m - 1, n) + \Delta(k, m, n - 1)$$

Como valores "base" de la recursión, tenemos que considerar que

A La suma de elementos positivos (rangos) nunca es negativa: si $k < 0 \Rightarrow \Delta(k, m, n) = 0$

B Siempre que $n = 0 \Rightarrow U = \frac{m(m+1)}{2}$: $\Delta(k, m, 0) = \begin{cases} 1 & \text{si } k = \frac{m(m+1)}{2} \\ 0 & \text{en otro caso} \end{cases}$

[4] Resulta ser que en teoría de números este tema se conoce con el nombre de particiones, y su tratamiento es bastante más áspero de lo que uno se imagina.

[5] La fórmula recursiva que sigue es equivalente a la que plantea Hettmansperger, pero la notación difiere considerablemente. En clase no la entendí pero me gustó el ejercicio y la desarrollé por mi cuenta más tarde.

C Hay una única forma de sumar 0, y es con 0 elementos. Si $(m = 0 \vee k = 0) \Rightarrow$

$$\Delta(k, m, n) = \begin{cases} 1 & \text{si } (m = 0 \wedge k = 0) \\ 0 & \text{en otro caso} \end{cases}$$

Veamos que la fórmula es consistente con el ejemplo anterior con $n = m = 2$, $\binom{2+2}{2} = 6$

$$\begin{aligned}
\Delta(1,2,2) &= \Delta(1-(2+2), 2-1, 2) + \Delta(1,2,2-1) \\
&= \cancelto{0\,(A)}{\Delta(-3,1,2)} + \Delta(1,2,1) \\
&= \cancelto{0\,(A)}{\Delta(-2,1,1)} + \cancelto{0\,(B)}{\Delta(1,2,0)} \\
&= 0 \qquad\qquad p_{2,2}(1) = \tfrac{0}{6} \\
(\ldots)& \\
\Delta(5,2,2) &= \Delta(1,1,2) + \Delta(5,2,1) \\
&= \cancelto{0\,(A)}{\Delta(-2,0,2)} + \Delta(1,1,1) + \Delta(2,1,1) + \cancelto{0\,(B)}{\Delta(5,2,0)} \\
&= \cancelto{0\,(A)}{\Delta(-1,0,1)} + \cancelto{1\,(B)}{\Delta(1,1,0)} + \cancelto{1\,(C)}{\Delta(0,0,1)} + \cancelto{0\,(B)}{\Delta(2,1,0)} \\
\Delta(5,2,2) &= 2 \qquad\qquad p_{2,2}(5) = \tfrac{2}{6} \\
(\ldots)&
\end{aligned}$$

Además, se cumple que

$$\begin{aligned}
p_{m,n}(k) =& \frac{\Delta(k,m,n)}{\binom{n+m}{m}} = \frac{n! \cdot m!}{(n+m)!}\Delta(k,m,n) \\
=& \frac{m}{(m+n)} \frac{n! \cdot (m-1)!}{(n+m-1)!}\Delta(k-(m+n), m-1, n) + \\
& \frac{n}{(m+n)} \frac{(n-1)! \cdot m!}{(n+m-1)!}\Delta(k, m, n-1) \\
=& \frac{m}{(m+n)} \frac{\Delta(k-(m+n), m-1, n)}{\binom{n+m-1}{m-1}} + \frac{n}{(m+n)} \frac{\Delta(k,m,n-1)}{\binom{n-1+m}{n-1}} \\
p_{m,n}(k) =& p_{m-1,n}(k-(m+n)) + p_{m,n-1}(k)
\end{aligned}$$

Al programar una función recursiva, se vuelve imperioso *cachear* (*i.e.*, almacenar los resultados computados previamente) cada resultado obtenido, para que la cantidad de instrucciones a ejecutar no "explote" rápidamente. A continuación, proponemos una implementación en R, y la usamos para computar la distribución exacta del ejemplo:

```
> # Defino el cache/memoria de la funcion
> max_n <- 50
```

```
> max_m <- 50
> max_k <- sum( (max_m+1):(max_n+max_m) )
> memU <- array(dim = c(max_k, max_m, max_n))
> # A falta del caracter 'Delta', usamos 'A'
> A <- function(k, m, n) {
+   # Devolver valores limites
+   if (k < 0) { return(0) } # Regla A
+   if (n == 0) { return(k==m*(m+1)/2) } # Regla B
+   if (m == 0 | k == 0) { return(k==0 & m == 0) } # Regla C
+
+   # De estar precomputado, devolver el valor
+   if (!is.na(memU[k,m,n])) { return(memU[k,m,n]) }
+
+   # Si no, calcularlo, cachearlo y devolverlo
+   valor <- A(k-(m+n), m-1, n) + A(k, m, n-1)
+   memU[k,m,n] <<- valor
+   return(valor)
+ }
> # En R, `dX` es la densidad de la distribucion X
> dU <- function(x, m, n) {
+   # Devuelve Pr(U = xi | n, m) para cada xi en x
+   posibles <- map_dbl(x, A, m=m, n=n)
+   totales <- choose(m+n, m)
+   return(posibles / totales)
+ }
> tablaU <- function(m, n) {
+   tibble(
+         # Dados n, m, el soporte de U esta dado por
+         rango = (m*(m+1)/2):(m*n + m*(m+1)/2),
+         probs = dU(rango, m, n))
+ }
> tablaU(2,2)
```

```
# A tibble: 5 x 2
  rango probs
  <int> <dbl>
1     3 0.167
2     4 0.167
3     5 0.333
4     6 0.167
5     7 0.167
```

La figura 4 muestra la distribución de U bajo H_0 para un ejemplo moderadamente grande ($m = 10, n = 15$), y al menos a mí me alcanza para convencerme (intuitivamente) de la normalidad asintótica del estadístico[6]:

[6] ¡Y con qué breve comando se puede conseguir!

```
> tablaU(10, 15) %>% ggplot(aes(rango, probs)) + geom_col(width = 0.5)
```

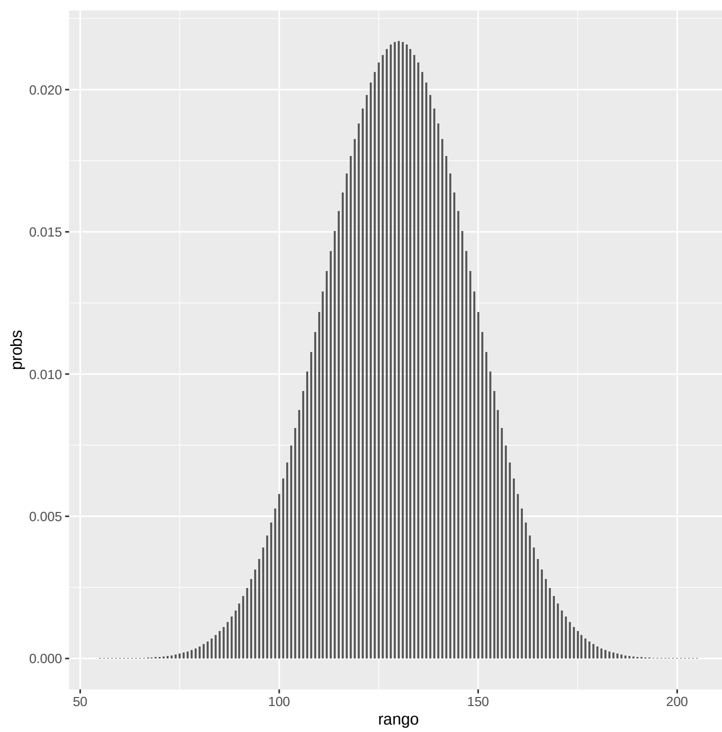

Figura 4: $p_{10,15}\left(k\right)$,o probabilidad puntual de para $m = 10, n = 15$

3. Apéndice

3.1. Condiciones suficientes de consistencia

Definición 5. Sea $X_1, ..., X_n$ una muestra aleatoria tal que $\forall\ i \in [n]\,,\ X_i \overset{iid}{\sim} F \in \Omega$, y $\{\Phi_n\}$ una sucesión de tests de nivel asintotico α para las hipótesis $H_0 : F \in \Omega_0 \subset \Omega$ vs. $H_1 : F \in \Omega_1 \subset \Omega$. Se dice que $\{\Phi_n\}$ es consistente para $F_1 \in \Omega_1$ si

- $\alpha \geq E_{F_0}(\Phi_n) \geq \gamma > 0 \quad \forall\ F_0 \in \Omega_0$, y
- $E_{F_1}(\Phi_n) \to 1$

donde $E_G\left(T\left(X_1, ..., X_n\right)\right) = E\left(T \middle| X_i \overset{iid}{\sim} G\right)$. Es decir, que el nivel de significación del test se mantiene acotado por encima de 0 para toda la sucesión, mientras que la potencia tiende a 1 para la alternativa, a medida que aumenta el tamaño 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 útil para determinar la clase de consistencia en ciertos test comunes.

Teorema 6. *Sea Φ_n un test de nivel asintótico α basado en T_n que rechaza H_0 para valores grandes, para las hipótesis previamente definidas. Si*

$$T_n \xrightarrow{p} \mu(F) = \begin{cases} \mu_0 & \forall F \in \Omega_0 \\ > \mu_0 & \forall F \in \Omega_c \subset \Omega_1 \end{cases}$$

y además existe σ_0 tal que

$$\sqrt{n}\frac{T_n - \mu_0}{\sigma_0} \xrightarrow{D} N(0,1) \quad \forall F \in \Omega_0$$

entonces existe una sucesión de valores k_n tal que la sucesión de tests $\Phi_n = \mathbb{I}\left(T_n \geq k_n\right)$ es consistente para la clase Ω_c.

Demostración. Sea $z_\alpha \coloneqq k : \Pr\left(N\left(0,1\right) > k\right) = \alpha$ el cuantil $1-\alpha$ de la distribución normal estándar, y elijamos $k_n = \mu_0 + \frac{\sigma_0}{\sqrt{n}} z_\alpha$. Luego, el nivel asintótico de Φ_n es

$$\alpha_n = \mathrm{E}_{H_0}\Phi_n = \Pr_{H_0}\left(T_n \geq \mu_0 + \tfrac{\sigma_0}{\sqrt{n}} z_\alpha\right) = \Pr_{H_0}\left(\sqrt{n}\frac{T_n - \mu_0}{\sigma_0} \geq z_\alpha\right) \to \alpha$$

Tomemos ahora una $F' \in \Omega_c$ fija, y definamos

$$\epsilon = \frac{\mu\left(F'\right) - \mu_0}{2} > 0$$

Nótese de su definición que $k_n \overset{n\to\infty}{\longrightarrow} \mu_0$, de manera que para n' suficientemente grande, $k_{n'} < \mu_0 + \epsilon$. Luego,

$$k_{n'} < \mu_0 + \epsilon = \mu\left(F'\right) - 2\epsilon + \epsilon = \mu\left(F'\right) - \epsilon$$

Si se cumple que

$$\left|T_n - \mu\left(F'\right)\right| \leq \epsilon \Rightarrow T_n - \mu\left(F'\right) \geq -\epsilon \Rightarrow T_n \geq \mu\left(F'\right) - \epsilon = k_n$$

y por ende

$$\Pr_{F'}\left(\left|T_n - \mu\left(F'\right) \leq \epsilon\right|\right) \leq \Pr_{F'}\left(T - \mu\left(F'\right) \geq -\epsilon\right) = \Pr_{F'}\left(T_n \geq k_n\right) \leq 1$$

Pero por definición de convergencia en probabilidad, $\Pr_{F'}\left(\left|T_n - \mu\left(F'\right) \leq \epsilon\right|\right) \rightarrow 1$. En consecuencia $\Pr_{F'}\left(T_n \geq k_n\right) \rightarrow 1$, y Φ_n es consistente $\forall F' \in \Omega_c$. $\square$