UNA PRUEBA CONDICIONAL DE RACHAS DE DISTRIBUCIÓN LIBRE PARA SIMETRÍA

EBER JAVIER LENES PUELLO

Monografía presentada como requisito para optar titulo de Matemático

UNIVERSIDAD DE CARTAGENA

FACULTAD DE CIENCIAS E INGENIERÍA

PROGRAMA DE MATEMÁTICA

CARTAGENA

2007

2

UNA PRUEBA CONDICIONAL DE RACHAS DE DISTRIBUCIÓN LIBRE PARA SIMETRÍA

EBER JAVIER LENES PUELLO

Monografía presentada como requisito para optar titulo de Matemático

Asesores

JIMMY ANTONIO CORZOS SALAMANCA
SANDRA GUTIERREZ

UNIVERSIDAD DE CARTAGENA

FACULTAD DE CIENCIAS E INGENIERÍA

PROGRAMA DE MATEMÁTICA

CARTAGENA

2007

AGRADECIMIENTOS

*Agradezco, ante todo a **Dios padre** por haberme acompañado en el camino que me proyecté y además por enseñarme a ver la vida de modo diferente para tener una mejor relación con las personas que más estimo .*

*A mi **Padre** que me impulsa todos los dias a seguír adelante. " TE ADORO ".*

*A mi **madre Querida** por todos sus esfuerzos realizados para poder alcanzar todas las cosas que tengo y que me has brindado con mucho aprecio " TE AMO ".*

*A mis **hermanos** que de alguna forma han contribuido en mis esfuerzos realizados .*

*Al profesor **Jimmy Corzos Salamanca** por estar dispuesto a ayudarme en lo conserniente al trabajo.*

A esas personas que de una manera u otra forma siempre estuvieron presente para ayudarme a alcanzar mi meta parcial propuesta .

*Por último dedico este trabajo a mis abuelas queridas "**CARMEN** " y " **NICOLASA** " que son una bendición de Dios.*

Índice general

Capítulo 1

INTRODUCCIÓN

En términos generales un procedimiento de inferencia estadística tiene mucho que ver con la extracción de una muestra aleatoria S de una distribución F que se quiere:

1. Verificar (rechazar) hipótesis sobre las característica de la distribución. Problema de la prueba de hipótesis, si suponemos que F depende de uno o varios parámetros.

2. Verificar (rechazar) hipótesis sobre su valor (desconocido). Problema de prueba de hipótesis.

3. Estimar puntualmente o por intervalos (su valor desconocido). Problema de estimación.

Para 3, existen métodos generales como: métodos de los momentos y método de la máxima verosimilitud. Para 2, usamos los estimadores puntuales de 3, generando prueba de hipótesis para resolver estos problemas. De 3, podemos ver que estas pruebas son generales.

Los métodos anteriormente usados son paramétricos, es conocido que una de las ventajas de los métodos no parametrizados sobre los paramétricos se da cuando los datos disponibles en la muestra se encuentran en escala ordinal o nominal. En muchos de estos casos no existen pruebas paramétricas. Por otro lado para 1, no existen pruebas generales para resolver esos problemas. Sin embargo, podemos hallar métodos que nos ayuden a solucionar alguno de los mismos, dichos métodos son llamados no paramétricos. A los procedimientos que requieren muy pocas o muy débiles supuestos acerca de la distribución muestrada, como continuidad o simetría de F, suelen llamarse métodos no paramétricos.

En este trabajo mostraremos dos pruebas de hipótesis que nos ayuda a resolver uno de los problemas que pueden presentarse en 1, verificar la simetría de una distribución. Varias estadísticas de pruebas han sido sugeridas en la literatura para resolver el mismo problema incluyendo a Butler (1969), Rothman y Woodeoofe (1972), Hill y Rao (1977) y Mc Williams (1990). Mc Williams realizó un estudio de simulación donde mostró la superioridad de su prueba sobre las otras pruebas mencionadas sobre varias alternativas asimétricas. La prueba de Mc Williams (1990) será presentada en el capítulo 3. Luego, mostraremos como la prueba de rachas de Mc Williams puede ser modificada para conseguir una mayor robustez contra la mala especificación de la mediana manteniendo su función de potencia superior, la prueba modificada será presentada en el capítulo 4. Finalmente, en el capítulo 5, mostraremos un estudio de simulación en el

6

cual se verifica que la prueba modificada (condicional) es más robusta para la mala especificación de la mediana y es más potente que la prueba de Mc Williams (1990) (incondicional).

Capítulo 2

PRELIMINARES

En este capítulo se darán algunas definiciones y teoremas que serán usados en algunos capítulos posteriores.

Definición: Sea X una variable aleatoria, sea $\mu_X = E(X)$. La varianza de X denotada por $V(X)$, es definida por:

$$V(X) = \sum_j (x_j - \mu_X)^2 p_X(x_j)$$

si X es discreta con función de masa de probabilidad $p_X(x_j)$.

$$V(X) = \int_0^\infty (x - \mu_X)^2 f_X(x) dx$$

si X es continua con función de densidad de probabilidad $f_X(x)$.

$$V(X) = \int_{-\infty}^{\infty} 2x[1 - F_X(x) + F_X(-x)]dx - \mu_X^2$$

para cualquier variable aleatoria X.

Teorema 2.0.1

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

Definición: La covarianza de dos variables aleatorias X, Y con medias μ_X y μ_Y respectivamente, entonces:

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

o

$$Cov(X, Y) = E(XY) - E(X)E(Y)$$

Teorema 2.0.2 *Si $X_1, \cdots, X_n$ son variables aleatorias, entonces* $Var(\sum_{k=1}^{n} X_k) = \sum_{k=1}^{n} Var(X_k) + \sum_{1 \leq j} \sum_{\neq k \leq n} Cov(X_k, X_j)$

prueba
Hagamos esta prueba por inducción, para n=1 Listo!. Para $n = 2$ tenemos:

$$\begin{aligned}
Var(\sum_{k=1}^{2} X_k) &= Var(X_1 + X_2) \\
&= E((X_1 + X_2 - E(X_1 + X_2))^2) \\
&= E(((X_1 - E(X_1)) + (X_2 - E(X_2)))^2) \\
&= E((X_1 - E(X_1))^2 + (X_2 - E(X_2))^2 + 2(X_1 - E(X_1))(X_2 - E(X_2))) \\
&= Var(X_1) + Var(X_2) + 2Cov(X_1, X_2) \\
&= \sum_{k=1}^{2} + \sum_{1\leq j} \sum_{\neq k \leq n} Cov(X_k, X_j)
\end{aligned}$$

Entonces, $Var(\sum_{k=1}^{2} X_k) = \sum_{k=1}^{2} + \sum_{1\leq j} \sum_{\neq k\leq n} Cov(X_k, X_j)$. Es decir, para $n = 2$ también se cumple.

Supongamos que se cumple para $n = m$, es decir, $Var(\sum_{k=1}^{m} X_k) = \sum_{k=1}^{2} + \sum_{1\leq j} \sum_{\neq k\leq m} Cov(X_k, X_j)$.
Verifiquemos que también se cumple para $n = m + 1$. En efecto.

$$\begin{aligned}
Var(\sum_{k=1}^{m+1} X_k) &= Var(X_{m+1} + \sum_{k=1}^{m} X_k) \\
&= Var(X_{m+1}) + Var(\sum_{k=1}^{m} X_k) + 2cov(X_{m+1}, \sum_{k=1}^{m} X_k) \\
&= Var(X_{m+1}) + \sum_{k=1}^{m} Var(X_k) + \sum_{1\leq j} \sum_{\neq k\leq m} Cov(X_j, X_k) + 2cov(X_{m+1}, \sum_{k=1}^{m} X_k) \\
&= \sum_{k=1}^{m+1} Var(X_k) + \sum_{1\leq j} \sum_{\neq k\leq m} Cov(X_k, X_j) + 2cov(X_{m+1}, \sum_{k=1}^{m} X_k)
\end{aligned}$$

Pero,

$$\begin{aligned}
Cov(X_{m+1}, \sum_{k=1}^{m} X_k) &= E(X_{m+1} \sum_{k=1}^{m} X_k) - E(X_{m+1})E(\sum_{k=1}^{m} X_k) \\
&= \sum_{k=1}^{m} E(X_{m+1}X_k) - \sum_{k=1}^{m} E(X_{m+1})E(X_k) \\
&= \sum_{k=1}^{m} [\, E(X_{m+1}X_k) - E(X_{m+1})E(X_k) \,] \\
&= \sum_{k=1}^{m} Cov(X_{m+1}, X_k)
\end{aligned}$$

Entonces,

$$Var(\sum_{k=1}^{m+1} X_k) = \sum_{k=1}^{m+1} Var(X_k) + \sum_{1\leq j} \sum_{\neq k\leq m} Cov(X_k, X_j) + 2\sum_{k=1}^{m} cov(X_{m+1}, X_k)$$

Así,

$$Var(\sum_{k=1}^{m+1} X_k) = \sum_{k=1}^{m+1} Var(X_k) + \sum_{1\leq j} \sum_{\neq k \leq m+1} Cov(X_k, X_j)$$

Teorema 2.0.3 *Teorema central del límite*
Sea $X_1, \cdots, X_n$ una sucesión de variables aleatorias independientes e identicamente distribuidas, cada una con media μ y varianza σ^2, entonces la distribución de $\frac{X_n - \mu}{\sigma/\sqrt{n}}$ tiende a la normal standar cuando $n \longrightarrow \infty$.

Definición: (función generadora de momentos)
Sea X una variable aletoria con densidad $f_X(x)$. El valor esperado de e^{tX} es definida como la función generadora de momentos de X si los valores esperados de X existen para cada valor de t en algún intérvalo $h < t < h$; $h > 0$. La función generadora de momentos, denotada por $m_X(t)$ o $m(t)$, es:

$$m(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tX} f_X(x)dx$$

si X es una variable aleatoria continua y es

$$m(t) = E(e^{tX}) = \sum_j e^{tX} p_X(x_j)$$

si la variable aleatoria X es discreta.

Definición: Sean $m, n \in \mathbb{Z}$, entonces:

$$\binom{m}{n} = \begin{cases} \frac{m!}{(m-n)!n!} & \text{si} \quad 0 \leq n \leq m \\ 0 & \text{en otro caso} \end{cases}$$

Definición: (Distribución libre)
Es el nombre que se da a aquellos métodos que con o sin el supuesto de sobre la forma funcional de la distribución muestrada, para la estimación de parámetros o para la prueba de hipótesis, utilizan estadístico cuyas distribuciones no dependen de la distribución muestrada.

Definición: (Racha)
Es una sucesión de uno o más símbolos idénticos que representan una propiedad común.

Definición: (Prueba de rachas)
Es una técnica muy útil para probar la hipótesis nula H_0 de que las observaciones tomadas en realidad se escogen al azar.

Teorema 2.0.4 *Sea $X_1, ..., X_n$ un conjunto de variables aleatorias independientes, cada una con función generatriz de momentos $m_{X_1}(t), ..., m_{X_n}(t)$. Si $Y = a_1X_1 + ... + a_nX_n$, donde $a_1, ..., a_n$ son constantes, entonces,*

$$m_Y(t) = m_{X_1}(t)...m_{X_n}(t)$$

Teorema 2.0.5 *Sea $X_1, ..., X_n$ un conjunto de variables aleatorias independientes, con distribución bernoulli y parametro p, entonces la variable aleatoria $Y = X_1 + ... + X_n$ tiene distribución binomial con parametros n y p.*

Capítulo 3

PRUEBA DE RACHAS PARA SIMETRÍA

En este capítulo, presentaremos una prueba basada sobre una estadística de rachas para la simetría de una distribución continua con una mediana asumida y un estudio hecho por Monte Carlo demostrando que para una gran variedad de alternativas de distribuciones asimétricas, la prueba es más potente que las propuestas por Butler (1969), Rothman (1969) y Woodroofe (1972), Hill y Rao (1977).

Sea $\{X_1, \cdots, X_n\}$ una muestra aleatoria de una distribución diferenciable continua F y asumamos que la mediana especificada es θ. Considéremos la prueba para simetría sobre F:

$$H_0 : F(X - \theta) = 1 - F(\theta - X) \quad \text{para todo} \quad X$$

versus cualquier alternativa asimétrica. Definamos la siguiente aplicación:

$$\begin{array}{rccc} d(\cdot, \theta): & S & \longrightarrow & \mathbb{R} \\ & X_i & \mapsto & d(X_i, \theta) = |X_i - \theta| \end{array}$$

Y ordenemos los X_i de la siguiente manera: $X_i \leq X_j$ para $i, j \leq n$ si $d(\ X_i\ , \theta) \leq d(\ X_j\ , \theta)$. Obteniendo asi una sucesión ordenada $S = \{X_{(1)}, \cdots, X_{(n)}\}$. Sean $S_1, \cdots, S_n$ indicadores de variables dadas por:

$$S_i = \begin{cases} 1 & \text{si} \qquad X_{(i)} \geqslant \theta \\ & \\ 0 & \text{en otro caso} \end{cases}$$

Obteniendo así la sucesión $\{S_i\}$, la estadística de prueba utilizada es R, el número de rachas en la sucesión $\{S_i\}$. R puede también ser expresado como $R = 1 + I_2 + \cdots + I_n$, donde

$$I_k = \begin{cases} 1 & \text{si} \quad S_k \neq S_{k-1} \\ & \\ 0 & \text{si} \quad S_k = S_{k-1} \end{cases}$$

$k = 1, 2, \cdots, n$

Rechazamos la hipótesis nula si R falla en la cola inferior de la distribución nula. Para hallar la región critica se procede como sigue: si puede tolerarse un error de tipo I igual a α, $0 < \alpha < 1$, entonces el punto critico es aquel valor r_0 que cumple $P(R \leq r_o : n \in \mathbb{N}) \approx \alpha$, donde n es el tamaño de la muestra.

Para comprender el razonamiento detrás de la estadística escogida y la región critica, primero consideremos la prueba de rachas tradicional por independencia, con N_1 objetos de "tipo I" y N_2 objetos de "tipo II" arreglados en una sucesión de n objetos, es decir, $n = N_1 + N_2$. Definamos:

$$I_k = \begin{cases} 1 & \text{si el k-esimo elemento} \neq \text{(k-1)-esimo elemento} \\ 0 & \text{en otro caso} \end{cases}$$

Entonces,

$$P(I_k = 1) = P(S_{k-1} \neq S_k) = \frac{N_1 N_2}{\binom{n}{2}}$$

Asi, $I_k \backsim B(P)$ $k = 1, 2, \cdots, n$ y $P = \dfrac{2N_1N_2}{n(n-1)}$. Ahora, $R = 1 + \sum\limits_{k=2}^{n} I_k$, entonces, $E(R) = 1 + \sum\limits_{k=2}^{n} E(I_k)$.

Entonces,

$$E(R) = 1 + \frac{2N_1N_2}{n}$$

$$E(R) = 1 + \frac{2N_1(n - N_1)}{n}$$

Consideremos n fijo, $f(N_1) = 1 + \dfrac{2N_1(n - N_1)}{n}$, es una función cóncava la cual es maximizada cuando $N_1 = N_2 = n/2$. En efecto, $f'(N_1) = \dfrac{2n - 4N_1}{n} = 0$ sí y sólo sí $N_1 = n/2$ si y solo si $N_1 = N_2$. $f''(N_1) = -\dfrac{4}{n} < 0$ para $N_1 = n/2$, entonces f alcanza un máximo absoluto en $N_1 = n/2$. Ahora, $E(R) = 1 + \dfrac{2N_1(n - N_1)}{n}$, entonces el valor esperado de R es máximo cuando $N_1 = N_2$ o cuando el número de objetos de tipo I es igual al número de objetos de tipo II.

Sea $\theta \in R$, consideremos el conjunto $T = \{X_i : a < |X_i - \theta| < b\}$, para a, b arbitrarios, $|\theta| \leq a < b$. Clasifiquemos los $X_i > \theta$ como objetos de tipo I y los $X_i < \theta$ como objetos de tipo II, sean $A = \{X_i : a < |X_i - \theta| < b, X_i > \theta\}$ y $B = \{X_i : a < |X_i - \theta| < b, X_i < \theta\}$, entonces el número de rachas esperadas es máxima si $\#A = \#B$. Por otro lado $\#A = \#B$ es una consecuencia de simetría, entonces el número máximo de rachas es una concecuencia de simetría. Si F es asimétrica y existen a, b tales que $\#A$ difiere sustancialmente de $\#B$, esperamos que el número de valores de tipo I difiera sustancialmente del número de valores de tipo II, resultando así pocas rachas.

La contribución hecha por T a R tiende hacer grande cuando F es simétrica, pequeña cuando F es asimétrica. De nuevo recalquemos el punto, consideremos el caso extremo en el cual F es sesgada a la derecha y existen valores a y b, tal que $\#A > 0$ pero $\#B = 0$. En este caso los valores X_i satisfacen $a < |X_i - \theta| < b$, todas están en un mismo lado de θ, así ellos tendrán una contribución de 0 a R.

DISTRIBUCIÓN DE R

Para mostrar la distribución de R en esta prueba, se usará el siguiente lema:

Lema 3.1. *Usara el siguiente hipótesis Bajo la hipótesis nula de simetría, la variable aleatoria $R-1$ tiene una distribución binomial con parámetros $n-1$ y $1/2$.*

prueba
Sea $S = \{X_1, \cdots, X_n\}$ una muestra aleatoria y θ una mediana especificada de una distribución F, la cual se va a verificar su simetría. Sea

$$D_i = \begin{cases} 1 & \text{si} \quad X_i \geq \theta \\ \\ 0 & \text{si} \quad X_i < \theta \end{cases}$$

Bajo la hipótesis de simetría $D_1, \cdots, D_n$ son todas variables aleatorias Bernoulli con parámetro $p = 1/2$. Ahora, si ordenamos las variables de S, por su distancia a θ, entonces la sucesión $\{D_i\}$ también será ordenada y por la definición de S_i en la prueba tenemos que: $S_i = D_{(i)}$, lo cual implica que S_i, $i = 1, 2, \cdots, n$ son variables aleatorias Bernoulli con parámetro $p = 1/2$, (la sucesión S_i es esencialmente una permutación aleatoria de la sucesión $\{D_i\}$). Por otro lado, para $k = 2, \cdots, n$

$$\begin{aligned} P(I_k = 1) &= P(S_{k-1} \neq S_k) \\ &= P(S_{k-1} = 0 \ \text{ y } \ S_k = 1) + P(S_{k-1} = 1 \ \text{ y } \ S_k = 0) \\ &= P(S_k = 1)P(S_{k-1} = 0) + P(S_k = 0)P(S_{k-1} = 1) \\ &= \frac{1}{4} + \frac{1}{4} \\ &= \frac{1}{2} \end{aligned}$$

Entonces, $P(I_k = 1) = \frac{1}{2}$, luego, $I_k \sim B(\frac{1}{2})$, bajo la hipótesis nula de simetría. Ahora, $R = 1 + \sum_{k=2}^{n} I_k$, entonces, $R - 1 = \sum_{k=2}^{n} I_k$. Así, bajo la hipótesis nula de simetría

$$R - 1 \sim B(n-1, 1/2)$$

Teorema 3.1. *EL valor esperado y la varianza de R estan dadas por: $E(R) = \dfrac{n+1}{2}$ y $Var(R) = \dfrac{n-1}{2} \; \dfrac{3-n}{2} + \sum_{2 \leq j} \sum_{\neq k \leq j} E(I_j I_k)$*

prueba
En efecto, $R = 1 + \sum_{k=2}^{n} I_k$, entonces $E(R) = 1 + \sum_{k=2}^{n} E(I_k)$, como $E(I_k) = \frac{1}{2}$ para $k = 2, 3, \cdots, n$, entonces, $E(R) = 1 + \dfrac{n-1}{2} = \dfrac{n+1}{2}$. Por tanto, $E(R) = \dfrac{n+1}{2}$

14

$$\begin{aligned}
Var(R) &= Var(1+\sum_{k=2}^{n} I_k) \\
&= Var(1)+Var(\sum_{k=2}^{n} I_k)+2Cov(1,\sum_{k=2}^{n} I_k) \quad \text{pues } Cov(1,\sum_{k=2}^{n} I_k)=0 \\
&= Var(\sum_{k=2}^{n} I_k)
\end{aligned}$$

Así

$$\begin{aligned}
Var(R) &= Var(\sum_{k=2}^{n} I_k) \\
&= \sum_{k=2}^{n} I_k + \sum_{2\le j}\sum_{\neq k\le n} Cov(I_j,I_k) \\
&= (n-1)Var(I_k)+\sum_{2\le j}\sum_{\neq k\le n} Cov(I_j,I_k)
\end{aligned}$$

Entonces,

$$\begin{aligned}
Var(R) &= (n-1)\,[E(I_k^2)-(E(I_k))^2]+\sum_{2\le j}\sum_{\neq k\le n}[E(I_jI_k)-E(I_j)(I_k)] \\
&= (n-1)\,[E(I_k^2)-(E(I_k))^2]+\sum_{2\le j}\sum_{\neq k\le n}E(I_jI_k)-[(n-1)^2(E(I_k))^2-(n-1)(E(I_k))^2] \\
&= (n-1)\,E(I_k^2)+\sum_{2\le j}\sum_{\neq k\le n}E(I_jI_k)-(n-1)^2(E(I_k))^2 \\
&= \frac{n-1}{2}-\frac{(n-1)^2}{4}+\sum_{2\le j}\sum_{\neq k\le n}E(I_jI_k)
\end{aligned}$$

Pues, $E(I_k^2)=E(I_k)=\frac{1}{2}$. Entonces,

$$Var(R)=(\frac{n-1}{2})(\frac{3-n}{2})+\sum_{2\le j}\sum_{\neq k\le n}E(I_jI_k)$$

DISTRIBUCIÓN ASINTÓTICA DE R

Para hallar la distribución asintótica de R, podemos utilizar el teorema de Moivre-Laplace, puesto que en el lema anterior se demostró que $R-1\sim B(n-1,1/2)$ bajo la hipótesis nula de simetría.

Teorema 3.2. *Teorema de Moivre-Laplace*
Si X_n denota el número de éxitos que ocurren cuando son desarrollados n pruebas independientes, cada una con probabilidad de éxito p, entonces para $a<b$

$$P\left(a\le \frac{X_n-np}{\sqrt{np(1-p)}}\le b\right)=\phi(b)-\phi(a)$$

cuando $n\to\infty$

Este teorema es una aproximación normal de la distribución binomial y es un caso especial del Teorema central del limite, que se demostró en los preliminares. Esta aproximación es buena cuando $np(1-p) \geq 10$

Ejemplo:
En una planta de armado se diseña una operación especifica, la cual toma un tiempo promedio de 5 minutos. El gerente tome una muestra de 16 tiempos de operación para un empleado y se obtienen los siguientes resultados (en minutos): 6.1, 4.8, 4.9, 5.3, 5.6, 4.7, 5.4, 6.2, 4.6, 5.5, 5.2, 4.5, 5.7, 4.6, 5.7, 6.4, 4.5.
Sea X la variable aleatoria que denota el tiempo de duración de un empleado, probar si la distribución de X es simétrica con un nivel de significancia de 0.05.
Solucion:
Asumamos que la mediana especificada es $\theta = 6{,}0$, al ordenar los datos de la muestra obtenemos: 6.1, 6.2, 5.7, 5.7, 5.6, 6.4, 5.5, 5.4, 5.3, 5.2, 4.9, 4.8, 4.7, 4.6, 4.6, 4.5, 4.5 y $S = (1,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0)$ entonces $R = 4$. Así para un nivel de significancia de 0.05 el valor critico de la prueba es: $r_0 = 5$ y dado que el valor $r = 4$ se encuentra dentro de la región critica de tamaño 0.05, se rechaza la hipotesis nula de simetría.

PRUEBAS REALIZADAS

El estudio de Monte Carlo ha sido realizado para evaluar la potencia de la prueba basada sobre R para detectar asimetría. Para este propósito, cuatro de las pruebas de distribución libre existentes serán consideradas.

1. La prueba del rango signado de Wilcoxon, basada sobre T^+, la suma del rango signado positivo de $X_1, X_2, \cdots, X_n$. Esta prueba que referenciada por Conover (1980) y Lehmann (1986).

2. Una prueba propuesta por Butler (1969). Esta prueba es basada sobre la estadistica $B_n = n^{-1/2} \sup_{x \leq 0} |Q_n(x)|$, donde $Q_n(x) = n[F_n(x) + F_n(-x) - 1]$ y $F_n(x)$ representa la función de distribución de la muestra $X_1, \cdots, X_n$. Si H_0 es verdadera, entonces esperamos que $F_n(x) \approx 1 - F_n(x)$, lo cual se cumple para verdaderos valores de B_n. Butler derivo la distribución asintotica exacta de B_n.

3. Un tipo de estadística propuesta por Rothman y Woodroofe (1972):

$$R_n = n \int_{-\infty}^{\infty} [F_n^*(x) + F_n^*(-x) - 1]^2 dF_n(x)$$

   Donde F_n representa una y otra vez la función de distribución de la muestra y $F_n^* = \frac{1}{2}[F_n(x+) + F_n(x_0)]$. Como el caso de la prueba de Butler, lo lógico es que $F_n(x) = 1 - F_n(x)$ bajo la hipótesis nula, lo cual se da para valores pequeños de la estadística de prueba. Rothman y Woodroofe proporcionaron percentiles de la distribución asintótica de T_n y demostraron, por un estudio de Monte Carlo que esta distribución proporciona una buena aproximación de la distribución exacta para muestras de tamaños mayores o iguales a 20.

4. Un tipo de prueba Cramer-Von Mises propuesta por Hill y Rao (1977) que combina la estadística de Rothman y Woodroofe la propuesta por Srinivasan y Gadio (1974). Esta

estadística es hallada definiendo $\delta_k = 1$ si $X_{(k)} \geq 0$, $\delta_k = -1$ en otro caso. Entonces, calculamos,

$$P_k = \frac{1}{2}\sum_{j=1}^{k}(1+\delta_j), \quad N_k = \frac{1}{2}(1-\delta_j)$$

Finalmente, calculamos la estadística de prueba

$$T_n = \frac{1}{2n^2}\left\{\sum_{k=1}^{n}(N_k - P_k)^2 + \sum_{k=1}^{n}(N_k' - P_k')^2\right\}$$

Los términos N_k' y P_k' serán calculados en la misma manera que N_k y P_k pero basados sobre $X_1^{-1}, \cdots, X_n^{-1}$ más que en los $X_1, \cdots, X_n$. Hill y Rao proporcionan una tabla de percentiles de T_n para tamaño de muestra de 10 hasta 24. Lockart y Mc Larent (1985) dan los puntos críticos para la distribución asintótica y han demostrado que sus valores asintóticos proporcionan una buena aproximación de los puntos críticos exactos cuando excede a 24. Los estudios.

Los estudios de Monte Carlo serán usados para evaluar la realización de R y competir con lo descrito previamente. Las pruebas serán evaluados para tamaños de muestras de $n = 20, 30, 50$ y 100 usando un nivel de significaría $\alpha = \cdot 05$. El resultado del estudio hecho por Monte Carlo es dado en la siguiente tabla:

PROBABILIDAD DE RECHAZO (X 1.000): $\alpha = 0{,}05$

Caso 1 (distribución simétrica)
$\lambda_1 = 0, \lambda_2 = 0{,}197454, \lambda_3 = 0{,}134915, \lambda_4 = 0{,}134915 \quad (\alpha_3 = 0.\ \alpha_4 = 3{,}0)$

n	R	T^+	B_n	R_n	T_n
20	49	42	44	41	42
30	47	47	56	48	59
50	50	52	65	53	60
100	64	43	48	42	41

Caso 2
$\lambda_1 = 0, \lambda_2 = 1{,}0, \lambda_3 = 1{,}4, \lambda_4 = 0{,}25 \quad (\alpha_3 = 0{,}5,\ \alpha_4 = 2{,}2)$

n	R	T^+	B_n	R_n	T_n
20	192	84	53	75	72
30	297	94	62	84	91
50	476	117	86	103	139
100	776	200	327	207	278

Caso 3
$\lambda_1 = 0, \lambda_2 = 1{,}0, \lambda_3 = 0{,}0007, \lambda_4 = 0{,}1 \quad (\alpha_3 = 1{,}5,\ \alpha_4 = 5{,}8)$

n	R	T^+	B_n	R_n	T_n
20	325	90	70	69	89
30	438	131	82	112	127
50	683	196	205	189	253
100	927	350	689	438	595

 17

Caso 4

$\lambda_1 = 3{,}586508, \lambda_2 = 0{,}04306, \lambda_3 = 0{,}025213, \lambda_4 = 0{,}094029 \quad (\alpha_3 = 0{,}9,\ \alpha_4 = 4{,}2)$

n	R	T^+	B_n	R_n	T_n
20	85	59	38	57	40
30	117	68	59	58	81
50	131	84	61	73	79
100	223	126	72	106	129

Caso 5

$\lambda_1 = 0, \lambda_2 = -1{,}0, \lambda_3 = -0{,}0075, \lambda_4 = 0{,}03 \quad (\alpha_3 = 1{,}5,\ \alpha_4 = 7{,}5)$

n	R	T^+	B_n	R_n	T_n
20	120	70	61	60	59
30	145	77	59	67	78
50	192	118	65	96	96
100	338	157	108	135	201

Caso 6

$\lambda_1 = -0{,}116734, \lambda_2 = -0{,}351663, \lambda_3 = -0{,}13, \lambda_4 = -0{,}16 \quad (\alpha_3 = 0{,}8,\ \alpha_4 = 11{,}4)$

n	R	T^+	B_n	R_n	T_n
20	55	49	60	51	50
30	50	45	40	43	46
50	56	42	41	41	44
100	51	52	41	43	45

Caso 7

$\lambda_1 = 0, \lambda_2 = -1{,}0, \lambda_3 = -0{,}1, \lambda_4 = -0{,}18 \quad (\alpha_3 = 2{,}0,\ \alpha_4 = 21{,}2)$

n	R	T^+	B_n	R_n	T_n
20	63	47	50	45	55
30	90	56	48	47	62
50	97	64	49	64	63
100	124	97	61	97	89

Caso 8

$\lambda_1 = 0, \lambda_2 = -1{,}0, \lambda_3 = -0{,}001, \lambda_4 = -0{,}13 \quad (\alpha_3 = 3{,}16,\ \alpha_4 = 23{,}8)$

n	R	T^+	B_n	R_n	T_n
20	387	116	76	109	111
30	534	151	102	133	148
50	744	234	291	239	337
100	972	423	841	607	776

Caso 9

$\lambda_1 = 0, \lambda_2 = -1{,}0, \lambda_3 = -0{,}0001, \lambda_4 = -0{,}17 \quad (\alpha_3 = 3{,}88,\ \alpha_4 = 40{,}7)$

18

n	R	T^+	B_n	R_n	T_n
20	398	122	73	108	87
30	560	171	127	163	155
50	816	256	318	258	344
100	976	434	883	636	794

α_3 *y* α_4 denotan el sesgo y la curtosis de las distribuciones respectivamente.

Seleccionamos nueve distribuciones de la familia lambda generalizada discutida en Ramberg y Schmeiser (1974), incluyendo una distribución simétrica y una variedad de distribuciones asimétricas. La forma de las distribuciones son mostradas en Mc Williams (1990). Esta familia es fácilmente generada y es definida en términos como:

$$X_i = \lambda_i + \frac{1}{\lambda_2}(u^{\lambda_3} - (1-u)^{\lambda_4}) \qquad i = 1, 2, \cdots, n$$

Donde u es un número aleatorio, los parámetros $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ son de la distribución lambda generalizada. En todos los casos la mediana de la distribución es asumida e investigamos la capacidad de varias pruebas para detectar la presencia de asimetría. Para cada combinación de F y n 1000 muestras aleatorias son generadas. Si la mediana de F no es cero, este valor será sustraído de los datos. Cada muestra aleatoria es usada para calcular los valores para R, T^+, B_n, R_n y T_n. Frecuentemente observamos el número de veces de 1000, que la hipótesis nula es rechazada para estadística de prueba. Las pruebas que son basadas sobre valores críticos de distribuciones discretas son aleatorizadas para tener el tamaño exacto de 0.05. Los resultados del estudio de Monte Carlo, proporcionan un caso fuerte para el uso de R para detectar la forma de distribución asimétrica.

En el peor de los casos R es competitivo en potencia, la cual en muchos casos es claramente dominante. Por ejemplo, en el caso 8 cuando $n = 30$, el resultado empírico muestra que R es 3.5 más para detectar la asimetría de distribuciones que T^+. Considerando el estudio de Monte Carlo y el hecho que R es fácil de calcular y tiene una distribución disponible, entonces, decimos que R es más recomendada sobre cualquiera de las estadísticas consideradas en este estudio. Note que la ventaja de la potencia de R es limitada a una clase especifica de hipótesis alternativa consideradas en este estudio. Supongamos que la mediana asumida de F es correctamente especificada y midamos la capacidad de varias pruebas para detectar asimetría de distribuciones. Una distribución también puede ser asimétrica con respecto al estado del valor de la mediana por virtud de un cambio de localización. En este caso una estadística más tradicional, como la estadística de Wilcoxon T^+, es recomendada, como la prueba basada sobre R no proporciona ventaja y puede en este caso ser menos potente.

Capítulo 4

PRUEBA MODIFICADA

En este capítulo se dará la segunda prueba para simetría y la distribución exacta de la estadística de prueba utilizada.
Sea $S = \{X_1, \cdots, X_n\}$ una muestra aleatoria y θ la mediana especificada de una distribución continua diferenciable G. Considere la siguiente prueba:

$$H_o: \; G(X-\theta) = 1 - G(\theta - X) \quad para \; todo \; X$$

versus cualquier alternativa asimétrica y definamos la siguiente aplicación:

$$\begin{aligned} d(\cdot, \theta) : S &\longrightarrow R \\ X_i &\longrightarrow d(X_i, \theta) = |X_i - \theta| \end{aligned}$$

Ahora, ordenemos los $X_i \in S$, $i = 1, \cdots, n$ de la siguiente manera, $X_i \leq X_j$ si $d(X_i, \theta) \leq d(X_j, \theta)$ para $i, j = 1, \cdots, n$; obteniendo así la sucesión ordenada $S = \{X_{(1)}, \cdots, X_{(n)}\}$.
Sean $S_1, \cdots, S_n$ indicadores de variables dadas por

$$S_i = \begin{cases} 1, & si \; X_{(i)} \geq \theta \quad i = 1, \cdots, n \\ 0, & en \; otro \; caso \end{cases}$$

resultando la sucesión $\{S_i\}$. La estadística de prueba R es el número total de rachas en la sucesión $\{S_i\}$. Para hallar su región crítica se procede como sigue: supongamos que puede tolerarse un error de tipo I igual a α, $0 < \alpha < 1$, si hallamos r_o de tal manera que $P(R \leq r_o : p, m) \approx \alpha$ o $\sum_{r=2}^{r_o} P(R = r) \approx \alpha$ donde P y m representan el número de ceros y unos en la sucesión $\{S_i\}$ respectivamente, entonces r_o es el punto crítico de la prueba. Ahora considerando el valor de R según la muestra observada podemos decidir bajo esta regla si se rechaza la hipótesis nula de simetría.

Los pequeños valores de la estadística R proporcionan pruebas en contra de la simetría, pues las regiones críticas en estos casos son tomados unilaterales inferiores, lo cual fue probado en el capítulo anterior.
Estas regiones tendrán alguno de los pequeños valores que puede tomar la estadística R, así que cuando R tome dichos valores la hipótesis nula de simetría será rechazada.

Consideremos una sucesión ordenada de n elementos de dos tipos, p el primer tipo y m el segundo tipo, donde $p + m = n$. Si r_1 es el número de rachas de elementos de tipo I y r_2 es el número de rachas de elementos de tipo II, el número total de rachas en la sucesión es $r = r_1 + r_2$.

DISTRIBUCIÓN DE R.

La distribución de R será hallada determinando primero la distribución conjunta de probabilidad de R_1 y R_2 y entonces la distribución de su suma. Puesto que bajo la hipótesis nula cualquier arreglo de los $p+m$ objetos es igualmente probable, la probabilidad que $R_1 = r_1$ y $R_2 = r_2$ es el número de distintos arreglos de los $p+m$ objetos con r_1 rachas de tipo I y r_2 rachas de tipo II, divididos por el número total de arreglos el cual es $\binom{n}{m}$. Para hallar la cantidad del numerador, el siguiente lema puede ser usado.

Lema 4.1. *El número de formas diferentes para distribuir n objetos en r células diferentes, no vacías es:*

$$\binom{n-1}{r-1}; \quad n \geq r.$$

Demostración. Supongamos que los n objetos son todos bolas blancas. Coloquemos las n bolas en una fila y efectuemos una división de esta en r células intersectando $r-1$ bolas negras, una por una, entre cualquier par de bolas blancas en la linea. Puesto que existen $n-1$ posiciones en la cual cada bola negra puede ser colocada, en el número de arreglos es $\binom{n-1}{r-1}$. ôo

Teorema 4.1. *Supongamos que R_1 y R_2 denotan los respectivos números de rachas de p objetos tipo I y m objetos tipo II en una muestra aleatoria de tamaño $n = p+m$, la distribución conjunta de probabilidad de R_1 y R_2 es:*

$$f_{R_1,\ R_2}(r_1, r_2) = \frac{C\binom{p-1}{r_1-1}\binom{m-1}{r_2-1}}{\binom{m+p}{p}}$$

$$r_1 = 1, \cdots, p \ y \ \ r_2 = 1, \cdots, m$$

$$r_1 = r_2 \ \ o \ \ r_1 = r_2 \pm 1.$$

Donde $C = 2$ si $r_1 = r_2$ y $C = 1$ si $r_1 = r_2 \pm 1$.

Demostración. Consideremos una sucesión $\{S_i\}$ la cual tiene n elementos, p de tipo I y m de tipo II, es decir $n = p+m$. Por 4.1, el número diferentes formas de obtener r_1 células de elementos tipo I y r_2 células de elementos tipo II es $\binom{p-1}{r_1-1}$ y $\binom{m-1}{r_2-1}$ respectivamente. Luego, por el principio básico del conteo el número de formas diferentes de obtener r_1 rachas de elementos de tipo I y r_2 rachas de elementos de tipo II en $\{S_i\}$ es:

$$\binom{p-1}{r_1-1}\binom{m-1}{r_2-1}$$

1. Empezando con una racha de tipo I.
2. Comenzando con una racha de tipo II.

Si en el primer caso la sucesión $\{S_i\}$ es terminada con una racha de tipo I, entonces $r_1 = r_2 + 1$ y si esta es terminada con una racha de tipo II, entonces $r_1 = r_2$. Si en segundo caso la sucesión $\{S_i\}$ es terminada con una racha de tipo II, entonces $r_2 = r_1 + 1$ o $r_1 = r_2 - 1$ y si es terminada con una racha de tipo I, Entonces, el número de formas diferentes de obtener r_1 rachas de tipo I y r_2 rachas de tipo II es:

$$\binom{p-1}{r_1-1}\binom{m-1}{r_2-1} \quad si \ \ r_1 = r_2 \pm 1$$

$$2\binom{p-1}{r_1-1}\binom{m-1}{r_2-1} \quad si \ \ r_1 = r_2$$

ôo

21

Corolario 4.1. *La distribución marginal de probabilidad de R_1 es:*

$$f_{R_1}(r_1) = \begin{cases} \dfrac{\binom{p-1}{r_1-1}\binom{m+1}{r_1}}{\binom{p+m}{p}}, & r_1 = 1, \cdots, p \end{cases}$$

Similarmente para R_2 con p y m intercambiados.

Demostración. (ejercicio). ♢○

Teorema 4.2. *La distribución de probabilidad de R, el número total de rachas de $n = p + m$ objetos, p de tipo I y m de tipo II, en una muestra aleatoria es:*

$$f_R(r) = \begin{cases} \dfrac{2\binom{p-1}{r/2-1}\binom{m-1}{r/2-1}}{\binom{p+m}{p}} & si\ r\ es\ par,\ r > 1 \\ \dfrac{\binom{p-1}{\frac{r-1}{2}}\binom{m-1}{\frac{r-3}{2}}+\binom{p-1}{\frac{r-3}{2}}\binom{m-1}{\frac{r-1}{2}}}{\binom{p+m}{m}} & si\ r\ es\ impar,\ r > 1 \end{cases}$$

Para $r = 2, 3, \cdots, p + m$

Demostración. En el teorema 4.1, vimos que $r_1 = r_2$ ó $r_1 = r_2 \pm 1$, además $r = r_1 + r_2$. Ahora si r es par debe cumplirse únicamente $r_1 = r_2$ de manera que $r_1 = r_2 = r/2$.
Por otro lado, si r es impar, entonces $r_1 = r_2 \pm 1$ y por tanto $r_1 = \frac{r-1}{2}$ y $r_2 = \frac{r+1}{2}$, ó $r_1 = \frac{r+1}{2}$ y $r_2 = \frac{r-1}{2}$, reemplazando este resultado en 4.1 tenemos

$$f_R(r) = \begin{cases} \dfrac{2\binom{p-1}{r/2-1}\binom{m-1}{r/2-1}}{\binom{p+m}{p}} & si\ r\ es\ par,\ r > 1 \\ \dfrac{\binom{p-1}{\frac{r-1}{2}}\binom{m-1}{\frac{r-3}{2}}+\binom{p-1}{\frac{r-3}{2}}\binom{m-1}{\frac{r-1}{2}}}{\binom{p+m}{m}} & si\ r\ es\ impar,\ r > 1 \end{cases} \tag{4.0.1}$$

♢○

MOMENTOS DE LA ESTADÍSTICA R

El k-ésimo momento de R es:

$$\sum_k r^k f_R(r) = \frac{\sum_{r\ par} 2r^k\binom{p-1}{r/2-1}\binom{m-1}{r/2-1}}{\binom{p+m}{p}} + \frac{\sum_{r\ impar} r^k\binom{p-1}{\frac{r-1}{2}}\binom{m-1}{\frac{r-3}{2}} + \binom{p-1}{\frac{r-3}{2}}\binom{m-1}{\frac{r-1}{2}}}{\binom{p+m}{p}}$$

Hallemos solo los dos primeros momentos, pero antes de ello, mostremos unos lemas que nos ayudarán para su demostración.
Veamos cual es el rango de valores de R; por definición de la estadística R determina el número de rachas en la sucesión $\{S_i\}$ proveniente de la prueba modificada (condicional), donde p y m son el número de ceros y unos en ella respectivamente, es decir la estadística R está condicionada a p y m. Por lo tanto son considerados tres casos para la determinación del rango de valores de R.

- Si $p = 0$ ó $m = 0$
- Si $p \neq 0$ y $m \neq 0$ con $p \neq m$.
- Si $p \neq 0$ y $m \neq 0$ con $p = m$

Lema 4.2. *Si $p = 0$ ó $m = 0$, entonces el rango de valores de R es $\{1\}$.*

Demostración. Sea $\{S_i\}$ una sucesión obtenida de la prueba de rachas modificadas siendo p y m el número de ceros y unos en $\{S_i\}$ respectivamente. Supongamos que $p = 0$, entonces $r = 1$ para toda permutación posible de los términos de la sucesión $\{S_i\}$, entonces $r = 1$ con la probabilidad uno. De igual forma se razona si $m = 0$. ♢○

 23

sucesión $\{S_i\}$ será terminada en cero, lo cual implica que $2\min(p,m) \in Ran(R)$.
Veamos que $r \not> 2\min(p,m)$ para cualquier permutación en la sucesión $\{S_i\}$, supongamos que no, es decir, existe una cierta permutación en la sucesión $\{S_i\}$ la cual forma r rachas, $r > 2\min(p,m)$ con p ceros y m unos, entonces $r \geq 1 + 2\min(p,m)$, entonces el número de ceros es $\geq 1 + 2\min(p,m)$ y el número de unos es $\geq 1 + \min(p,m)$ ó el número de ceros es $\geq \min(p,m)$ y el número de unos es $\geq 1 + \min(p,m)$ en la sucesión $\{S_i\}$ lo cual es absurdo y según la nota 4.1 concluimos que $Ran(R) = \{2, \cdots, 2\min(p,m)\}$. ôo

Lema 4.5. *Sean $n, m, a, b \in \mathbb{N}$, $k \in \mathbb{Z}$ con $0 \leq k \leq n$, entonces*

1. $\binom{n}{k} = \binom{n}{n-k}$.
2. $\sum_{k=0}^{n} \binom{a}{k}\binom{b}{n-k} = \binom{a+b}{n}$.
3. $\sum_{k=1}^{n} \binom{n-1}{k-1}\binom{m-1}{k-1} = \binom{n+m-2}{n-1}$.
4. $\sum_{k=1}^{n} k\binom{n-1}{k-1}\binom{m-1}{k-1} = (m-1)\binom{m+n-3}{n-2} + \binom{n+m-2}{n-1}$.
5. $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$.
6. $k\binom{n}{k} = n\binom{n-1}{k-1}$.

Demostración. 1. inmediata.

2. $(1+x)^{a+b} = (1+x)^a(1+x)^b$ para todo x; entonces

$$\sum_{i=0}^{a+b}\binom{a+b}{i}x^i = \sum_{j=0}^{b}\binom{b}{j}x^j\sum_{k=0}^{n}\binom{a}{k}x^k$$
$$= \sum_{j=0}^{b}\sum_{k=0}^{a}\binom{b}{j}\binom{a}{k}x^{k+j}$$

Sea $h = k + j$, entonces $j = h - k$

$$\sum_{i=0}^{a+b}\binom{a+b}{i}x^i = \sum_{h=k}^{b+k}\sum_{k=0}^{a}\binom{b}{h-k}\binom{a}{k}x^h$$

entonces

$$\binom{a+b}{n} = \sum_{k=0}^{n}\binom{b}{n-k}\binom{a}{k}$$

3.

$$\sum_{k=1}^{n}\binom{n-1}{k-1}\binom{m-1}{k-1} = \sum_{k=1}^{n}\binom{n-1}{n-k}\binom{m-1}{k-1} \quad por\ 1.$$

Sea $y = k - 1$, entonces $k = y + 1$ y cuando $k = 1$, $y = 0$ y cuando $k = n$, $y = n - 1$, entonces,

$$\sum_{k=1}^{n}\binom{n-1}{k-1}\binom{m-1}{k-1} = \sum_{y=0}^{n-1}\binom{n-1}{n-1-y}\binom{m-1}{y} = \binom{m+n-2}{n-1} \quad por\ 2$$

Así que

$$\sum_{k=1}^{n}\binom{n-1}{k-1}\binom{m-1}{k-1} = \binom{n+m-2}{n-1}$$

4. $$\sum_{k=1}^{n} k\binom{n-1}{k-1}\binom{m-1}{k-1} = \sum_{k=1}^{n} k\binom{n-1}{n-k}\binom{m-1}{k-1}; \quad por \ 1$$

sea $y = k-1$, entonces $k = y+1$ y cuando $k = 1$, $y = 0$ y cuando $k = n$, $y = n-1$, entonces,

$$\begin{aligned}\sum_{k=1}^{n} k\binom{n-1}{k-1}\binom{m-1}{k-1} &= \sum_{y=0}^{n-1}(y+1)\binom{n-1}{n-1-y}\binom{m-1}{y} \\ &= \sum_{y=0}^{n-1} y\binom{n-1}{n-1-y}\binom{m-1}{y} + \sum_{y=0}^{n-1}\binom{n-1}{n-1-y}\binom{m-1}{y}\end{aligned}$$

después de aplicar la parte 3 del lema tenemos:

$$\sum_{k=1}^{n} k\binom{n-1}{k-1}\binom{m-1}{k-1} = (m-1)\sum_{y=1}^{n-1}\binom{m-2}{y-1}\binom{n-1}{n-1-y} + \binom{n+m-2}{n-1}$$

sea $z = y-1$, entonces $y = z+1$ y cuando $y = 1$, $z = 0$ y cuando $y = n-1$ $z = n-2$, entonces

$$\begin{aligned}\sum_{k=1}^{n} k\binom{n-1}{k-1}\binom{m-1}{k-1} &= (m-1)\sum_{z=0}^{n-2}\binom{m-2}{z}\binom{n-1}{n-2-z} + \binom{n+m-2}{n-1} \\ &= (m-1)\binom{m+n-3}{n-2} + \binom{n+m-2}{n-1}\end{aligned}$$

es decir,

$$\sum_{k=1}^{n} k\binom{n-1}{k-1}\binom{m-1}{k-1} = (m-1)\binom{m+n-3}{n-2} + \binom{n+m-2}{n-1}$$

5. inmediata.

6. inmediata

ôo

A continuación hallaremos la media y la varianza de la estadística R.

Teorema 4.3. *La media y la varianza de la estadística R están dadas por:*

$$E(R) = 1 + \frac{pm}{p+m} \quad y \quad V(R) = \frac{2pm(2pm-p-m)}{(p+m)^2(p+m-1)}$$

Demostración. Para hallar el valor esperado y la varianza es necesario considerar los casos:

- $p < m$
- $p = m$

En efecto, supongamos que $p < m$, $p \neq 0$, entonces por (4.3) y definición de valor esperado tenemos:

$$E(R) = \sum_{r=2}^{1+2min(p,m)} rp(r) = \sum_{r=2}^{1+2p} rp(r)$$

De ahora en adelante colocaremos $r = 2k$ cuando r sea par y $r = 2k+1$ para r impar en la función de probabilidad de R, $k = 1, \cdots, p$. Así,

$$E(R) = \frac{\sum_{k=1}^{p} 4k\binom{p-1}{k-1}\binom{m-1}{k-1} + \sum_{k=1}^{p}(2k+1)[\binom{p-1}{k-1}\binom{m-1}{k} + \binom{p-1}{k}\binom{m-1}{k-1}]}{\binom{p+m}{p}}$$

$$= \frac{\sum_{k=1}^{p} 2k[\binom{p-1}{k-1}\binom{m-1}{k-1} + \binom{p-1}{k}\binom{m-1}{k-1}]}{\binom{p+m}{p}} + \frac{\sum_{k=1}^{p} 2k[\binom{p-1}{k-1}\binom{m-1}{k-1} + \binom{p-1}{k-1}\binom{m-1}{k}]}{\binom{p+m}{p}}$$

$$+ \frac{\sum_{k=1}^{p} \binom{p-1}{k-1}\binom{m-1}{k} + \binom{p-1}{k}\binom{m-1}{k-1}}{\binom{p+m}{p}}$$

Si en las dos primeras sumatorias aplicamos la ley distributiva y (4.5) incisos 5 y 6, entonces

$$E(R) = \frac{(2p+2m)\sum_{k=1}^{p} \binom{p-1}{k-1}\binom{m-1}{k-1} + \sum_{k=1}^{p}[\binom{p-1}{k-1}\binom{m-1}{k} + \binom{p-1}{k}\binom{m-1}{k-1}]}{\binom{p+m}{p}}$$

$$= \frac{\sum_{k=1}^{p} 2\binom{p-1}{k-1}\binom{m-1}{k-1} + \sum_{k=1}^{p}[\binom{p-1}{k-1}\binom{m-1}{k} + \binom{p-1}{k}\binom{m-1}{k-1}]}{\binom{p+m}{p}}$$

$$+ \frac{(2p+2m-2)\sum_{k=1}^{p} \binom{p-1}{k-1}\binom{m-1}{k-1}}{\binom{p+m}{p}}$$

Ahora por (4.5), inciso 3

$$E(R) = 1 + \frac{2(p+m-1)\binom{p+m-2}{p-1}}{\binom{p+m}{p}}$$

Luego, realizando unos cálculos matemáticos obtenemos lo pedido en el teorema.

Por otro lado, supongamos que $p = m$, $p \neq 0$, entonces por (4.4) y la definición de valor esperado tenemos

$$E(R) = \sum_{r=2}^{2p} rp(r)$$

$$= \frac{\sum_{k=1}^{p} 4k\binom{p-1}{k-1}\binom{m-1}{k-1} + \sum_{k=1}^{p-1}(2k+1)[\binom{p-1}{k-1}\binom{m-1}{k} + \binom{p-1}{k}\binom{m-1}{k-1}]}{\binom{p+m}{p}}$$

$$= \frac{4p + \sum_{k=1}^{p-1} 4k\binom{p-1}{k-1}\binom{m-1}{k-1}}{\binom{p+m}{p}} + \frac{\sum_{k=1}^{p-1}(2k+1)[\binom{p-1}{k-1}\binom{m-1}{k} + \binom{p-1}{k}\binom{m-1}{k-1}]}{\binom{p+m}{p}}$$

$$= \frac{4p + \sum_{k=1}^{p-1} 2k[\binom{p-1}{k-1}\binom{m-1}{k-1} + \binom{p-1}{k}\binom{m-1}{k-1}]}{\binom{p+m}{p}} + \frac{\sum_{k=1}^{p-1} 2k[\binom{p-1}{k-1}\binom{m-1}{k-1} + \binom{p-1}{k-1}\binom{m-1}{k}]}{\binom{p+m}{p}}$$

$$+ \frac{\sum_{k=1}^{p-1}[\binom{p-1}{k-1}\binom{m-1}{k} + \binom{p-1}{k}\binom{m-1}{k-1}]}{\binom{p+m}{p}}$$

Si en las dos primeras sumatorias aplicamos la ley distributiva y (4.5), incisos 5 y 6, entonces

$$E(R) = \frac{4p + (2p+2m)\sum_{k=1}^{p-1}\binom{p-1}{k-1}\binom{m-1}{k-1}}{\binom{p+m}{p}} + \frac{\sum_{k=1}^{p-1}[\binom{p-1}{k-1}\binom{m-1}{k} + \binom{p-1}{k}\binom{m-1}{k-1}]}{\binom{p+m}{p}}$$

$$= \frac{2(p+m-1)\sum_{k=1}^{p}\binom{p-1}{k-1}\binom{m-1}{k-1}}{\binom{p+m}{p}} + \frac{\sum_{k=1}^{p}[2\binom{p-1}{k-1}\binom{m-1}{k-1}] + \sum_{k=1}^{p-1}[\binom{p-1}{k-1}\binom{m-1}{k} + \binom{p-1}{k}\binom{m-1}{k-1}]}{\binom{p+m}{p}}$$

Así por (4.5), inciso 3

$$E(R) = \frac{2(p+m-1)\binom{p+m-2}{p-1}}{\binom{p+m}{p}} + 1$$

Por tanto

$$E(R) = 1 + \frac{2pm}{p+m}$$

Para mostrar que $V(R) = \frac{2pm(2pm-p-m)}{(p+m)^2(p+m-1)}$, primero calcularemos $E(R^2)$. Consideremos los casos:

- $p < m$
- $p = m$

Supongamos que $p < m$, con $p \neq 0$ entonces por 4.3 y definición de valor esperado

$$E(R^2) = \sum_{r=2}^{1+2p} r^2 p(r)$$

$$= \frac{\sum_{k=1}^{p} 8k^2\binom{p-1}{k-1}\binom{m-1}{k-1} + \sum_{k=1}^{p}(4k^2+4k+1)[\binom{p-1}{k-1}\binom{m-1}{k} + \binom{p-1}{k}\binom{m-1}{k-1}]}{\binom{p+m}{p}}$$

$$= \frac{\sum_{k=1}^{p-1} 4k^2[\binom{p-1}{k-1}\binom{m-1}{k-1} + \binom{p-1}{k-1}\binom{m-1}{k}]}{\binom{p+m}{p}} + \frac{\sum_{k=1}^{p} 4k^2[\binom{p-1}{k-1}\binom{m-1}{k-1} + \binom{p-1}{k}\binom{m-1}{k-1}]}{\binom{p+m}{p}}$$

$$+ \frac{\sum_{k=1}^{p}(4k+1)[\binom{p-1}{k-1}\binom{m-1}{k} + \binom{p-1}{k}\binom{m-1}{k-1}]}{\binom{p+m}{p}}$$

Si en las dos primeras sumatorias aplicamos la ley distributiva, (4.5) incisos 5 y 6 tenemos:

$$E(R^2) = \frac{4(m+p)\sum_{k=1}^{p} k\binom{p-1}{k-1}\binom{m-1}{k-1}}{\binom{p+m}{p}} + \frac{\sum_{k=1}^{p}(4k+1)[\binom{p-1}{k-1}\binom{m-1}{k} + \binom{p-1}{k}\binom{m-1}{k-1}]}{\binom{p+m}{p}}$$

$$= \frac{4(m+p-2)\sum_{k=1}^{p} k\binom{p-1}{k-1}\binom{m-1}{k-1}}{\binom{p+m}{p}} + \frac{\sum_{k=1}^{p} 2k[\binom{p-1}{k-1}\binom{m-1}{k-1} + \binom{p-1}{k-1}\binom{m-1}{k}]}{\binom{p+m}{p}}$$

$$+ \frac{\sum_{k=1}^{p} 2k[\binom{p-1}{k-1}\binom{m-1}{k-1} + \binom{p-1}{k}\binom{m-1}{k}]}{\binom{p+m}{p}}$$

$$+ \frac{\sum_{k=1}^{p}(2k+1)[\binom{p-1}{k-1}\binom{m-1}{k} + \binom{p-1}{k-1}\binom{m-1}{k}] + \sum_{k=1}^{p} 4k\binom{p-1}{k-1}\binom{m-1}{k-1}}{\binom{p+m}{p}}$$

Si aplicamos la ley distributiva en las sumatorias 2 y 3, por (4.5), incisos 5 y 6, entonces

$$E(R^2) = \frac{4(p+m-2)\sum_{k=1}^{p} k\binom{p-1}{k-1}\binom{m-1}{k-1} + 2(p+m)\sum_{k=1}^{p}\binom{p-1}{k-1}\binom{m-1}{k-1}}{\binom{p+m}{p}} + \frac{2pm+p+m}{p+m}$$

entonces por (4.5), incisos 3 y 4 tenemos

$$E(R^2) = \frac{4(p+m-2)[(m-1)\binom{p+m-3}{p-2} + \binom{p+m-2}{p-1}]}{\binom{p+m}{p}} + \frac{2(p+m)\binom{p+m-2}{p-1}}{\binom{p+m}{p}} + \frac{2pm+p+m}{p+m}$$

Al convertir cada uno de los sumandos del numerador en $\binom{p+m}{p}$ mediante multiplicaciones respectivas, tenemos

$$E(R^2) = \frac{4mp(m-1)(p-1)}{(p+m-1)(p+m)} + \frac{4mp(p+m-2)}{(p+m-1)(p+m)} + \frac{2mp}{m+p+1} + \frac{2mp+m+p}{m+p}$$

entonces

$$\begin{aligned} V(R) &= E(R^2) - (E(R))^2 \\ &= \frac{4mp(m-1)(p-1)}{(p+m-1)(p+m)} + \frac{4mp(p+m-2)}{(p+m-1)(p+m)} + \frac{2mp}{m+p-1} + \frac{2mp+m+p}{m+p} - \frac{(2mp+m+p)^2}{(m+p)^2} \end{aligned}$$

Entonces,

$$V(R) = \frac{2mp(2mp-m-p)}{(p+m)^2(m+p-1)}$$

Por otro lado, supongamos que $p = m$, $p \neq 0$, entonces por (4.4) y definición

$$\begin{aligned} E(R^2) &= \frac{\sum_{k=1}^{p} 8k^2\binom{p-1}{k-1}\binom{m-1}{k-1}}{\binom{p+m}{p}} + \frac{\sum_{k=1}^{p-1}(4k^2+4k+1)[\binom{p-1}{k-1}\binom{m-1}{k} + \binom{p-1}{k}\binom{m-1}{k-1}]}{\binom{p+m}{p}} \\ &= \frac{8p^2 + \sum_{k=1}^{p-1} 4k^2[\binom{p-1}{k-1}\binom{m-1}{k-1} + \binom{p-1}{k}\binom{m-1}{k-1}]}{\binom{p+m}{p}} + \frac{\sum_{k=1}^{p-1} 4k^2[\binom{p-1}{k-1}\binom{m-1}{k-1} + \binom{p-1}{k-1}\binom{m-1}{k}]}{\binom{p+m}{p}} \\ &+ \frac{\sum_{k=1}^{p}(4k+1)[\binom{p-1}{k-1}\binom{m-1}{k} + \binom{p-1}{k}\binom{m-1}{k-1}]}{\binom{p+m}{p}} \end{aligned}$$

Aplicando la ley distributiva en las dos primeras sumatorias, (4.5), incisos 5 y 6 tenemos que:

$$\begin{aligned} E(R^2) &= \frac{8p^2 + 4(m+p)\sum_{k=1}^{p-1} k\binom{p-1}{k-1}\binom{m-1}{k-1}}{\binom{p+m}{p}} + \frac{\sum_{k=1}^{p-1}(4k+1)[\binom{p-1}{k-1}\binom{m-1}{k} + \binom{p-1}{k}\binom{m-1}{k-1}]}{\binom{p+m}{p}} \\ &= \frac{4p + 4(m+p-2)\sum_{k=1}^{p} k\binom{p-1}{k-1}\binom{m-1}{k-1}}{\binom{p+m}{p}} + \frac{\sum_{k=1}^{p-1} 2k[\binom{p-1}{k-1}\binom{m-1}{k-1} + \binom{p-1}{k-1}\binom{m-1}{k}]}{\binom{p+m}{p}} \\ &+ \frac{\sum_{k=1}^{p-1} 2k[\binom{p-1}{k-1}\binom{m-1}{k-1} + \binom{p-1}{k}\binom{m-1}{k-1}]}{\binom{p+m}{p}} \\ &+ \frac{\sum_{k=1}^{p} 4k\binom{p-1}{k-1}\binom{m-1}{k-1} + \sum_{k=1}^{p-1}(2k+1)[\binom{p-1}{k-1}\binom{m-1}{k} + \binom{p-1}{k}\binom{m-1}{k-1}]}{\binom{p+m}{p}} \end{aligned}$$

28

Si aplicamos la ley distributiva en la segunda y tercera sumatoria, (4.5), incisos 5 y 6, concluimos que

$$E(R^2) = \frac{4p + 4(p+m-2)\sum_{k=1}^{p} k\binom{p-1}{k-1}\binom{m-1}{k-1} + 2(p+m)\sum_{k=1}^{p-1}\binom{p-1}{k-1}\binom{m-1}{k-1}}{\binom{p+m}{p}} + \frac{2pm+m+p}{m+p}$$

Entonces, por (4.5), incisos 3 y 4 tenemos:

$$E(R^2) = \frac{4(p+m-2)[(m-1)\binom{p+m-3}{p-2} + \binom{p+m-2}{p-1}] + 2(p+m)\binom{p+m-2}{p-1}}{\binom{p+m}{p}} + \frac{2pm+m+p}{p+m}$$

Finalmente concluimos que:

$$V(R) = \frac{2mp(2mp-m-p)}{(m+p)^2(m+p-1)}$$

őo

LA DISTRIBUCIÓN ASINTÓTICA DE R

La ecuación (4.0.1) puede ser usada para hallar la distribución exacta de R para valores cualesquiera de p y m, los cálculos son laboriosos aunque p y m sean ambos pequeños. Para muestras grandes, podemos usar una aproximación de la distribución de R, la cual brinda buenos resultados cuando p y m son ambos mayores que 10.

Teorema 4.4. *Sea p y m el número de ceros y unos respectivamente en una sucesión obtenida de la prueba modificada, $n = p + m$. Si $p \longrightarrow \infty$ y $m \longrightarrow \infty$ tal que $p/n \longrightarrow \lambda$ y $m/n \longrightarrow 1-\lambda$ para λ fijo, $0 < \lambda < 1$. Entonces la variable aleatoria estandarizada*

$$Z = \frac{R - 2n\lambda(1-\lambda)}{2n^{1/2}\lambda(1-\lambda)}$$

tiene como límite una distribución normal standar, cuando $n \longrightarrow \infty$

prueba

Sea p y m el número de ceros y unos en la sucesión obtenida, respectivamente. Definamos:

$$I_k = \begin{cases} 1 & \text{si} \quad S_k \neq S_{k-1} \\ 0 & \text{en otro caso} \end{cases}$$

para $k = 2, \cdots, n$. Entonces, $P(I_k = 1) = \frac{pm}{\binom{n}{2}} = \frac{2pm}{n(n-1)}$. Así, $I_k \sim B(\frac{2pm}{n(n-1)})$ para $k = 2, \cdots, n$. Entonces, $R = 1 + \sum_{k=2}^{n} I_k$, es tal que $R - 1 =\sim B(n-1, \frac{2pm}{n(n-1)})$. Ahora, por el teorema de Moivre-Laplace

$$Z = \frac{R - 1 - E(R-1)}{\sqrt{Var(R-1)}}$$

tiene como limite distribución normal standar, cuando $n \longrightarrow \infty$. Como $Var(R-1) = Var(R)$, entonces,

$$Z = \frac{R - E(R)}{\sqrt{Var(R)}} \sim N(0,1)$$

cuando $n \longrightarrow \infty$. Por otro lado, hallemos los limites de $E(R)$ y $Var(R)$ cuando $n \longrightarrow \infty$. Sabiendo que $p/n \longrightarrow \lambda$ y $m/n \longrightarrow 1-\lambda$ para $0 < \lambda < 1$, cuando $p \longrightarrow \infty$ y $m \longrightarrow \infty$.

Entonces,

$$\begin{aligned}
\lim_{n\longrightarrow\infty} Var(\frac{R}{n^{1/2}}) &= \lim_{n\longrightarrow\infty} \frac{Var(R)}{n} \\
&= \lim_{n\longrightarrow\infty} \frac{2mp(2mp-m-p)}{n(m+p)^2(m+p-1)} \\
&= 2\lambda(1-\lambda) \lim_{n\longrightarrow\infty} \frac{2mp-m-p}{n(n-1)} \\
&= 2\lambda(1-\lambda) \lim_{n\longrightarrow\infty} \frac{2mp}{n(n-1)} \\
&= 4\lambda^2(1-\lambda) \lim_{(n)\longrightarrow\infty} \frac{p}{n-1} \\
&= 4\lambda^2(1-\lambda) \lim_{(n-1)\longrightarrow\infty} \frac{p}{n-1} \\
&= 4\lambda^2(1-\lambda)^2
\end{aligned}$$

Es decir, $\lim_{n\longrightarrow\infty} Var(\frac{R}{n^{1/2}}) = 4\lambda^2(1-\lambda)^2$, lo cual implica que para n grande, se tiene

$$Z = \frac{R - 2n\lambda(1-\lambda)}{\sqrt{4\lambda^2(1-\lambda)^2}} \approx \frac{R - E(R)}{\sqrt{Var(R)}} \sim N(0,1)$$

Por tanto,

$$Z = \frac{R - 2n\lambda(1-\lambda)}{2n^{1/2}\lambda(1-\lambda)} \sim N(0,1)$$

cuando $n \longrightarrow \infty$.

Ejemplo:

En una planta de armado se diseña una operación especifica, la cual toma un tiempo promedio de 5 minutos. El gerente toma una muestra de 16 tiempos de operación para un empleado y se obtienen los siguientes resultados(en minutos): 6.1, 4.8, 4.9, 5.3, 5.6, 4.7, 5.4, 6.2, 4.6, 5.5, 5.2, 4.5, 5.7, 4.6, 5.7, 6.4, 4.5.
Sea X la variable que denota el tiempo de duración de un empleado, probar si la distribución de X es simétrica con un nivel de significancia de 0.05.

Solución:
Asumamos que la mediana especificada es $\theta = 6.0$, al ordenar los datos de la muestra obtenemos: 6.1, 6.2, 5.7, 5.7, 5.6, 6.4, 5.5, 5.4, 5.3, 5.2, 4.9, 4.8, 4.7, 4.6, 4.6, 4.5, 4.5 y $S = (1,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0)$ entonces, $R = 4$. Así para un nivel de significancia de 0.05 el valor crítico de la prueba es: $r_o = 3$ y dado que el valor $r = 4$ no se encuentra dentro de la región crítica de tamaño 0.05, no se puede rechazar la hipótesis nula de simetría.

Capítulo 5

TAMAÑO Y FUNCIÓN DE POTENCIA DE LAS ESTADÍSTICAS DE PRUEBAS

En el presente capítulo mostraremos un estudio de simulación para investigar la función de la estadística de prueba modificada y compararlo con la estadística de prueba de Mc Williams (1990).

El estudio está concernido en los siguientes criterios:

- Potencia de las pruebas.
- Robustez de los tamaños de las pruebas contra la mala especificación de la mediana.

Para el primer criterio, usamos los siguientes indices:

1. Los niveles de significancia α, el cual toma los valores 0,01, 0,05 y 0,1;
2. Los tamaños de las muestras n, la cual toma los valores 30 y 50;
3. Los parámetros $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ de la distribución λ generalizada (Ramberg y Schameiser, 1974) usado para generar las observaciones como:

$$x_i = \lambda_1 + \frac{1}{\lambda_2}(u^{\lambda_3} - (1-u)^{\lambda_4}), \quad i = 1, \cdots, n$$

Donde u es un número aleatorio uniforme usado en el (paquete matemático, versión 4.0)

El siguiente teorema demuestra que la potencia de cada prueba puede ser estimada por la fracción de rechazos de la hipótesis nula.

Teorema 5.1. *Sea H_o la hipótesis nula de simetría de la prueba modificada, P la potencia de la misma, entonces P puede ser estimada por*

$$\hat{P} = \frac{n^o \ rechazos \ H_o}{muestras \ \ generadas}$$

Demostración. Sea G una distribución asimétrica, verifiquemos usando la prueba modificada

$$H_o : \ simetría \ \ G$$

$$H_1 : \ asimetría \ \ G.$$

Sea P la potencia de la prueba, entonces: $P = P(rechazar \ \ H_o/H_1 \ \ es \ \ falso)$ ó
$P = P(rechazar \ \ H_o/ \ la \ \ distribución \ \ asimétrica)$.
POr otro lado, la prueba modificada puede dar solo dos conclusiones respecto a G.

1. No se puede rechazar H_o (fracaso).

2. Se rechaza H_o (éxito).

Definamos la variable aleatoria X dada por

$$X = \begin{cases} 0, & si, \ se \ da \ \ 1 \\ 1, & si \ se \ da \ \ 2 \end{cases}$$

Por lo tanto, $P(X = 1) = P$ y $P(X = 0) = 1 - P$, entonces $X \sim B(P)$.

Sea Y la variable aleatoria que determina el número de éxito que ocurren después de aplicar pruebas sobre n muestras aleatorias, entonces $Y \sim B(n,\ P)$ y además mediante el método de máxima verosimilitud sabemos que $\hat{P} = \frac{Y}{n}$ es un buen estimador de P. Por tanto

$$\hat{P} = P.$$

õo

La misma prueba es realizada en el teorema si usamos la prueba de *Mc Williams*(1990). Estimemos las potencias de las pruebas usando el método del teorema anterior y considerando los siguientes nueve casos [las formas de las distribuciones son descritas en Mc Williams (1990)].
Caso 1: ($\lambda_1 = 0,\ \lambda_2 = 0{,}197454,\ \lambda_3 = 0{,}134915,\ \lambda_4 = 0{,}134915$) (simetría).
Caso 2: ($\lambda_1 = 0,\ \lambda_2 = 1,\ \lambda_3 = 1{,}4,\ \lambda_4 = 0{,}25$)
Caso 3: ($\lambda_1 = 0,\ \lambda_2 = 1,\ \lambda_3 = 0{,}00007,\ \lambda_4 = 0{,}1$)
Caso 4: ($\lambda_1 = 3{,}586508,\ \lambda_2 = 0{,}04306,\ \lambda_3 = 0{,}025213,\ \lambda_4 = 0{,}094029$)
Caso 5: ($\lambda_1 = 0,\ \lambda_2 = -1,\ \lambda_3 = -0{,}0075,\ \lambda_4 = -0{,}03$)
Caso 6: ($\lambda_1 = -0{,}116734,\ \lambda_2 = -0{,}351663,\ \lambda_3 = -0{,}13,\ \lambda_4 = -0{,}16$)
Caso 7: ($\lambda_1 = 0,\ \lambda_2 = -1,\ \lambda_3 = -0{,}1,\ \lambda_4 = -0{,}18$)
Caso 8: ($\lambda_1 = 0,\ \lambda_2 = -1,\ \lambda_3 = -0{,}001,\ \lambda_4 = -0{,}13$)
Caso 9: ($\lambda_1 = 0,\ \lambda_2 = -1,\ \lambda_3 = -0{,}0001,\ \lambda_4 = -0{,}17$)

		α=0.01	α=0.01	α=0.05	α=0.05
n	casos	incond	cond	incond	cond
30	1	0.0101	0.0113	0.0460	0.0473
30	2	0.1217	0.1336	0.3138	0.3315
30	3	0.2228	0.242	0.4552	0.4794
30	4	0.0288	0.0297	0.1055	0.1122
30	5	0.0409	0.0438	0.1330	0.1418
30	6	0.0097	0.0106	0.0493	0.0499
30	7	0.0153	0.0174	0.0687	0.0722
30	8	0.2944	0.3143	0.5395	0.5632
30	9	0.3283	0.3544	0.5823	0.6056
50	1	0.0107	0.0101	0.0512	0.0498
50	2	0.2466	0.2551	0.4841	0.4991
50	3	0.4255	0.4384	0.6751	0.6850
50	4	0.0444	0.0458	0.1458	0.1521
50	5	0.0663	0.0695	0.1978	0.2010
50	6	0.0108	0.0095	0.0495	0.0514
50	7	0.0173	0.0188	0.0836	0.0848
50	8	0.5317	0.5464	0.7659	0.7735
50	9	0.5904	0.6004	0.8075	0.8186

		α=0.1	α=0.1
n	casos	incond	cond
30	1	0.0973	0.0961
30	2	0.4413	0.4585
30	3	0.5950	0.6166
30	4	0.1872	0.1901
30	5	0.2269	0.2367
30	6	0.0997	0.1004
30	7	0.1329	0.1356
30	8	0.6798	0.6997
30	9	0.7055	0.7263
50	1	0.1007	0.1011
50	2	0.6261	0.6391
50	3	0.7856	0.7975
50	4	0.2373	0.2440
50	5	0.3116	0.3204
50	6	0.1048	0.1048
50	7	0.1565	0.1540
50	8	0.8541	0.8638
50	9	0.8908	0.8971

Para cada combinación de estos índices, 10.000 muestras son generadas, además las pruebas son aleatorias en orden para conseguir el tamaño nominal. Los resultados son dados en las tablas anteriores.

Para el segundo criterio de tamaño de robustez contra la mala especificación de la mediana, usamos los siguientes índices:

1. El tamaño nominal de la prueba α, la cual toma los valores 0.01, 0.05 y 0.1;
2. El tamaño de la muestra n, la cual toma los valores de 30 y 50.
3. el parámetro λ de la distribución λ simétrica (Tukey,1962) el cual toma los valores -0.5, -0.3, -0.1, 0.1, 0.3, 0.5. Los datos son generados así:

$$x_i = \lambda + \frac{1}{\lambda}(u^\lambda - (1-u)^\lambda), \quad i = 1, \cdots, n$$

4. La mediana mal especificada m_o toma los valores 55, 60, 65, $70th$ percentil de la distribución.

(los resultados para los correspondientes percentiles inferiores serán los mismos por la simetría de la distribución de la hipótesis nula).

El siguiente teorema demuestra que el tamaño de cada prueba puede ser estimada por la fracción de rechazos de la hipótesis nula de simetría.

Teorema 5.2. *Sea H_o la hipótesis nula de simetría de la prueba modificada, α el tamaño de la misma, entonces α puede ser estimada por* $\hat{\alpha} = \frac{n^o \text{ rechazos } H_o}{\text{muestras generadas}}$.

Demostración. Sea G una distribución simétrica, verifiquemos usando la prueba modificada

$$H_o: \ simetría \ \ G$$

$$H_1 : asimetría \;\; G$$

Sea α el tamaño de la prueba, entonces: $\alpha = P(rechazar \;\; H_o/H_o \; es \; verdadero)$ ó $\alpha = P(rechazar \;\; H_o/ \;\; la \; distribuciónsimétrica)$.
Por otro lado, la prueba modificada puede dar solo dos conclusiones respecto a G.

No se puede rechazar H_o (fracaso).

Se rechaza H_o (éxito). Definamos la variable aleatoria X dada por:

$$X = \begin{cases} 0, & si \;\; se \; da \; 1 \\ 1, & si \;\; se \; da \; 2 \end{cases}$$

Por lo tanto, $P(X = 1) = \alpha$ y $P(X = 0) = 1 - \alpha$, así $X \sim B(\alpha)$.
Sea Y la variable aleatoria que determina el número de éxito que ocurren después de realizar n ensayos, entonces $Y \sim B(n, \alpha)$ y además mediante el método de máxima verosimilitud sabemos que $\hat{e} = \frac{Y}{n}$ es un buen estimador de α por tanto $\hat{e} \sim \alpha$.

La misma demostración puede ser realizada en el teorema para el caso de la prueba de Mc Williams (1990).
Para cada combinación de estos índices, 10.000 son generadas y el tamaño de cada prueba puede ser hallado como la fracción de rechazos de la hipótesis nula de simetría. Las pruebas son aleatorizadas para conseguir el mismo tamaño nominal. Los resultados son dados en las tablas siguientes:las dos primeras tablas serán hechas para $n = 30$.

		α=0.01	α=0.01	α=0.05	α=0.05
λ	m_o	incond	cond	incond	cond
-0.5	0.5	0.0113	0.0115	0.0513	0.0491
-0.5	0.55	0.0137	0.0110	0.0578	0.0485
-0.5	0.60	0.0237	0.0124	0.0882	0.0569
-0.5	0.65	0.0545	0.0151	0.1576	0.0628
-0.5	0.70	0.1247	0.0187	0.2870	0.0786
-0.3	0.5	0.0094	0.0087	0.0505	0.0478
-0.3	0.55	0.0141	0.0115	0.0598	0.0514
-0.3	0.60	0.0226	0.0098	0.0920	0.0564
-0.3	0.65	0.0505	0.0136	0.1604	0.0674
-0.3	0.70	0.1237	0.0199	0.2841	0.0817
-0.1	0.5	0.0098	0.0090	0.0497	0.0492
-0.1	0.55	0.0138	0.0119	0.0590	0.0519
-0.1	0.60	0.0236	0.0121	0.0923	0.0619
-0.1	0.65	0.0531	0.0164	0.1566	0.0702
-0.1	0.70	0.1288	0.0261	0.2917	0.0909

3(

0.1	0.5	0.0105	0.0113	0.0513	0.0506
0.1	0.55	0.0118	0.0108	0.0586	0.0515
0.1	0.60	0.0272	0.0112	0.0955	0.0591
0.1	0.65	0.0603	0.0198	0.1786	0.0791
0.1	0.70	0.1386	0.0268	0.3174	0.1024
0.3	0.5	0.0109	0.0111	0.0513	0.0511
0.3	0.55	0.0131	0.0112	0.0600	0.0540
0.3	0.60	0.0270	0.0139	0.1050	0.0729
0.3	0.65	0.0765	0.0279	0.2055	0.1007
0.3	0.70	0.1682	0.0428	0.3671	0.1418
0.5	0.5	0.0111	0.0115	0.0492	0.0493
0.5	0.55	0.0151	0.0136	0.0713	0.0650
0.5	0.60	0.0378	0.0266	0.1303	0.0935
0.5	0.65	0.0981	0.0428	0.2625	0.1454
0.5	0.70	0.2291	0.0754	0.4621	0.2150

		α=0.1	α=0.1
λ	m_o	incond	cond
-0.5	0.5	0.1002	0.0990
-0.5	0.55	0.1079	0.0976
-0.5	0.60	0.1566	0.1073
-0.5	0.65	0.2565	0.1219
-0.5	0.70	0.4053	0.1404
-0.3	0.5	0.0997	0.0970
-0.3	0.55	0.0034	0.1023
-0.3	0.60	0.1637	0.1150
-0.3	0.65	0.2520	0.1211
-0.3	0.70	0.4017	0.1438
-0.1	0.5	0.0989	0.0976
-0.1	0.55	0.1136	0.1053
-0.1	0.60	0.1629	0.1191
-0.1	0.65	0.2489	0.1254
-0.1	0.70	0.4108	0.1552
0.1	0.5	0.1032	0.1019
0.1	0.55	0.1140	0.1033
0.1	0.60	0.1686	0.1171
0.1	0.65	0.2796	0.1462
0.1	0.70	0.4381	0.1751
0.3	0.5	0.1002	0.1016
0.3	0.55	0.1199	0.1081
0.3	0.60	0.1851	0.1338
0.3	0.65	0.3106	0.1813
0.3	0.70	0.5001	0.2323
0.5	0.5	0.0975	0.0960
0.5	0.55	0.1295	0.1200

0.5	0.60	0.2155	0.1643
0.5	0.65	0.3945	0.2466
0.5	0.70	0.6024	0.3242

En la tabla siguiente se considerará $n = 50$.

		α=0.01	α=0.01	α=0.05	α=0.05
λ	m_o	incond	cond	incond	cond
-0.5	0.5	0.0080	0.0089	0.0469	0.0503
-0.5	0.55	0.0125	0.0114	0.0621	0.0530
-0.5	0.60	0.0261	0.0119	0.1006	0.0569
-0.5	0.65	0.0708	0.0157	0.1996	0.0664
-0.5	0.70	0.1830	0.0196	0.3844	0.0844
-0.3	0.5	0.0102	0.0097	0.0453	0.0454
-0.3	0.55	0.0131	0.0100	0.0610	0.0527
-0.3	0.60	0.0268	0.0121	0.1016	0.0546
-0.3	0.65	0.0718	0.0168	0.2039	0.0703
-0.3	0.70	0.1832	0.0203	0.3835	0.0832
-0.1	0.5	0.0068	0.0073	0.0443	0.0422
-0.1	0.55	0.0139	0.0102	0.0636	0.0546
-0.1	0.60	0.0283	0.0112	0.1039	0.0573
-0.1	0.65	0.0753	0.0157	0.2059	0.0706
-0.1	0.70	0.2001	0.0269	0.4047	0.0959
0.1	0.5	0.0101	0.0090	0.0455	0.0475
0.1	0.55	0.0134	0.0110	0.0673	0.0580
0.1	0.60	0.0303	0.0155	0.1170	0.0672
0.1	0.65	0.0857	0.0226	0.2296	0.0884
0.1	0.70	0.2244	0.0358	0.4361	0.1212
0.3	0.5	0.0075	0.0084	0.0471	0.0480
0.3	0.55	0.0139	0.0112	0.0635	0.0530
0.3	0.60	0.0365	0.0188	0.1289	0.0825
0.3	0.65	0.1153	0.0333	0.2797	0.1209
0.3	0.70	0.2868	0.0608	0.5336	0.1864
0.5	0.5	0.0090	0.0095	0.0477	0.0491
0.5	0.55	0.0195	0.0168	0.0812	0.0697
0.5	0.60	0.0535	0.0297	0.1765	0.1184
0.5	0.65	0.1653	0.0632	0.3723	0.1954
0.5	0.70	0.3947	0.1152	0.6495	0.2889

		α=0.1	α=0.1
λ	m_o	incond	cond
-0.5	0.5	0.0997	0.0996
-0.5	0.55	0.1228	0.1060
-0.5	0.60	0.1763	0.1067
-0.5	0.65	0.3048	0.1249
-0.5	0.70	0.5093	0.1545
-0.3	0.5	0.0948	0.0954
-0.3	0.55	0.1202	0.1044
-0.3	0.60	0.1751	0.1099
-0.3	0.65	0.3171	0.1312
-0.3	0.70	0.5120	0.1514
-0.1	0.5	0.0910	0.0893
-0.1	0.55	0.1204	0.1062
-0.1	0.60	0.1802	0.1154
-0.1	0.65	0.3179	0.1326
-0.1	0.70	0.5276	0.1748
0.1	0.5	0.0973	0.0940
0.1	0.55	0.1241	0.1075
0.1	0.60	0.1997	0.1302
0.1	0.65	0.3420	0.1564
0.1	0.70	0.5636	0.2017
0.3	0.5	0.0996	0.0976
0.3	0.55	0.1242	0.1097
0.3	0.60	0.2198	0.1471
0.3	0.65	0.4076	0.2096
0.3	0.70	0.6568	0.2902
0.5	0.5	0.0991	0.0998
0.5	0.55	0.1406	0.1278
0.5	0.60	0.2807	0.1979
0.5	0.65	0.5159	0.3066
0.5	0.70	0.7645	0.4175

La robustez de la prueba condicional de rachas sobre la incondicional se presenta por las siguientes consideraciones:

Supongamos que una distribución G, la cual se va a probar que es simétrica y la mediana θ especificada no es en realidad correcta. Sea $\{S_i\}$ la sucesión obtenida mediante la prueba condicional, entonces es muy probable que exista una desigualdad entre el número de ceros y el número de unos en la sucesión $\{S_i\}$, sean n y m el número de ceros y unos respectivamente, entonces $n << m$ ó $m << n$[1] .
Supongamos que $m << n$, entonces por el lema 4.3 el número máximo de rachas posibles en $\{S_i\}$ es bastante menor que $m+n$ en forma considerable. Ahora si $z = m+n$, es el tamaño de la muestra utilizada, entonces el valor máximo que puede tomar R es bastante menor que Z.

Ahora utilizando la prueba incondicional la cual utiliza como estadística de prueba R y $R-1 \sim B(Z-1, 1/2)$, pueden existir puntos críticos los cuales superan a $1+2m$, lo cual implica

[1] El símbolo << significa menos, pero bastante menos en consideración a los números en comparación

37

que la estadística R después de ser calculada en una muestra es muy probable que resulte un valor de la estadística que se encuentre dentro de la región crítica garantizada, lo que conduce al rechazo de la hipótesis nula de simetría.
Lo anterior demuestra que la prueba incondicional es sensible para la mala especificación de la mediana.

En la prueba condicional, la cual utiliza como estadístico de prueba R, es muy importante como están distribuidos los ceros y los unos de la sucesión $\{S_i\}$.
Si $m << n$ entonces los puntos críticos para esta prueba son: $S = \{r_o | 2 < r_o < 1 + 2m\}$, pero si la sucesión $\{S_i\}$ tiene cada cero en un par de unos, entonces no es posible el rechazo de la hipótesis nula de simetría.

Por eso en esta prueba cuando θ es mal especificada lo que nos ayuda a decidir son las acumulaciones de ceros y unos en la sucesión $\{S_i\}$ o lo que uno denomina como rachas, por eso decimos que una de las cosas mas importantes en esta prueba es su estadística y te brinda bastante confianza para que la decisión pueda ser tomada en ésta.
Lo anterior muestra que la prueba condicional es más robusta que la incondicional para la mala especificación de la mediana.

Supongamos que $m << n$, entonces $1 + 2m << m + n$, entonces usando la estadística de Mc Williams, $P(R \leq 1 + 2m)1$ y usando la estadística modificada $P(R \leq 1 + 2m) = 1$. Sean r_o y r'_o los puntos críticos en las dos pruebas respectivamente, entonces $P(r'_o \leq r_o) \approx 1$, lo cual unifica que el cero tipo I de la prueba modificada es menor que el cero tipo I de la prueba de Mc Williams. Además, los puntos críticos de prueba condicional son modificados caso por caso, dependiendo de m y n, el número de ceros y el número de unos en la sucesión $\{S_i\}$ respectivamente.
Así, la fracción de rechazos de la hipótesis nula esperada estará cercana al tamaño nominal.

Capítulo 6

APLICACIONES

Un estudio comparativo de la potencia de la prueba del signo con respecto a la potencia de la prueba t- de student ó con respecto a otra prueba de localización con una muestra, suponiendo que la distribución muestreada es diferente de la normal, mostrará las ventajas o desventajas de esta prueba frente a sus competidoras en condiciones no normales (Randles & Wolfe, 1984, Cap 4).

<u>Pruebas de localización para distribuciones simétricas</u>

La prueba del signo es uniformemente más potente para la alternativa de una cola, en condiciones generales (Hettmansperger, 1984, pp. 9 y 55).
Para desarrollar un procedimiento similarmente potente pero utilizando los rangos de las observaciones, es decir, usando más información, se introduce el supuesto de simetría de la distribución muestreada que se expresa de la manera siguiente:

$$\Omega_s = \{F : F \in \Omega_{\theta_o} \ \ y \ \ F(x - \theta_o) = 1 - F(\theta_o - x)\}$$

donde,

$$\Omega_\theta = \{F : F \ es \ continua \ y \ F(x) = 1/2 \ \ solo \ para \ x = \theta_o\}$$

Si $F \in \Omega_s$, se dice que F es una distribución simétrica alrededor de θ_o, su única mediana, que es al mismo tiempo su media , cuando esta existe

<u>Modelo de Muestreo</u>

Muestra aleatoria: $X_1, \cdots, X_n$ cada variable con función de distribución $F(x - \theta)$, donde $F \in \Omega_s$. Entonces θ es la única mediana (y la media cuando existe) y se encuentra en el centro de distribución.
Nuevamente el problema de inferencia es la prueba de la hipótesis: $H_o : \theta = \theta_o$ versus $H_1 : \theta > \theta_o$ o versus $H_2 : \theta < \theta_o$ o versus $H_3 : \theta \neq \theta_o$ con $F \in \Omega_s$.
Para desarrollar una prueba que utilice la información sobre la simetría de la distribución muestreada es necesario introducir el concepto de rango de una observación.

Rangos: Se define el rango R_1 de la cantidad X_i, en la sucesión $X_1, \cdots, X_n$ como el puesto que ocupa X_i, en la sucesión ordenada $X_{(1)}, \cdots, X_{(n)}$, esto significa que $X_i = X_{(R_i)}$.
Por ejemplo

X_1	X_2	X_3	X_4	X_5
23.5	12.6	21.4	32.8	10.4
$X_{(4)}$	$X_{(2)}$	$X_{(3)}$	$X_{(5)}$	$X_{(1)}$

3c

los rangos de $X_1, \cdots, X_5$ son:

R_1	R_2	R_3	R_4	R_5
4	2	3	5	1

Nótese, por ejemplo, que como $R_1 = 4$ entonces $X_{(R_1)} = X_{(4)} = x_1 = 23{,}5$
También se define el anti-rango D_j de la j-ésima estadística de orden $X_{(j)}$ como el subíndice que tiene la observación original de donde proviene $X_{(j)}$. Esto significa que $X_{D_j} = X_{(j)}$.
En el ejemplo anterior las estadísticas de orden son

$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$
10.4	12.6	21.4	23.5	32.8
X_5	X_2	X_3	X_1	X_4

y los anti-rangos:

D_1	D_2	D_3	D_4	D_5
5	2	3	1	4

Por ejemplo, como $D_1 = 5$ entonces $X_{D_1} = X_{(1)} = X_5 = 10{,}4$. Observese también que $D_{R_i} = i$.

Se define el rango absoluto R_i^+ de X_i como el rango de $|X_i|$ en la sucesión ordenada de los valores absolutos

$$|X|_{(1)} \leq |X_2|_{(2)} \leq \cdots \leq |X|_{(n)}$$

es decir,

$$|X_i| = |X|_{(R_i^+)}$$

El anti-rango de $|X|_{(j)}$ es el subíndice que tiene la observación donde proviene $|X|_{(j)}$, es decir, $|X_{D_j}| = |X|_{(j)}$.
Se define el rango signado de X_i al producto $S(X_i)R_i^+$, donde

$$S(X_i) = \begin{cases} 1, & X_i > 0 \\ 0, & X_i \leq 0 \end{cases}$$

es decir, el rango signado de una observación positiva es su rango absoluto, pero el rango signado de una observación no positiva es cero.

<u>La prueba del rango signado de Wilcoxon</u>

Para la prueba de la hipótesis $H_o : \theta = \theta_o$ versus $H_1 : \theta > \theta_o$, puede utilizarse la estadística del rango signado de Wilcoxon:

$$T = \sum_{i=1}^{n} R_i^+ S(X_i) = \sum_{i=1}^{n} iW_i, \quad W_i = S(X_{D_i})$$

donde la segunda igualdad se obtiene reemplazando i por D_i. La región crítica se obtiene considerando que si muchas observaciones son mayores que θ_o y están a una distancia mayor que las observaciones menores que θ_o, esto será indicador de que $\theta > \theta_o$ y por tanto, los valores de T tenderán a ser mas grandes. En consecuencia se rechaza la hipótesis nula $H_o : \theta = \theta_o$ en favor de la alternativa $H_1 : \theta > \theta_o$ cuando

$$T \geq K_\alpha$$

donde k_α se determina de tal manera que para un nivel de significancia $\alpha \in (0,1)$ dado y fijo, se cumpla

$$P(T \geq K_\alpha) \leq \alpha$$

Bibliografía

[1] Ayman Baklizi, (2003). *A conditional distribution free runs test for symmetry* . Nonparametric-Statistics Vol 15 (6). Pags 713-718.

[2] Butler, C.C.(1969) *A test for symmetry using the sample distribution function.*The Annal of mathematical Statistic, 40, 2211 - 2214.

[3] Gibbons, J. D and Chakraborti, S. (1992), *Nonparametric Statiscal Inference.* Marcel Dekker, New York.

[4] Hill, D.L and Rao, P.V(1977) *Test of symmetry based on Cramer-Von Mises statistics*, Biometrika, 64, 489-494.

[5] Lehmann, E.L. (1975), *Nonparametrics Statistical method based on ranks*, Mc Graw Hill, New York.

[6] Mc Williams, T.P. (1990). *A distribution - free for symmetry based on a runs statistic.* Journal of the American Statiscal Association, 85 (412), 1130 - 1133.

[7] Ramberg, J.S: and Schmeiser, B.W (1874) *An approximate method for generating asymmetric random variables.*Communications of the ACM, 17, 78 - 82.

[8] Rothman, E.D. and Woodroofe, M. (1972) *A Cramer-Von Mises type statistic for testing symmetry.* The Annals of Mathematical Statistics, 43, 2035 - 2038.

[9] Tukey, J.W. (1962). *The future of the Analysis*, The annals of mathematical Statistic, 33, 1-67.

[10] Alexander M. Mood, Franklin A. GrayBill y Duane C. Boes. *Introducción a la teoría de estadística*, tercera edición. Mc Graw-Hill.

[11] George C. canavos. *Probabilidad y Estadística*, Primera edición. Mc Graw-Hill.

Nonparametric Statistics
Vol. 15(6), December 2003, pp. 713–718

A CONDITIONAL DISTRIBUTION FREE RUNS TEST FOR SYMMETRY

AYMAN BAKLIZI*

Department of Statistics, Yarmouk University, Irbid, Jordan

(Received October 2001; Revised September 2003; in final form September 2003)

The problem of testing for symmetry of a continuous distribution about a specified median is considered. The runs test proposed by McWilliams (McWilliams, T. P. (1990). A distribution-free test for symmetry based on a runs statistic. *Journal of the American Statistical Association*, **85**(412), 1130–1133) is modified by conditioning on the number of observations greater than the specified median. Simulations show that the modified test slightly outperforms the unconditional one in terms of power and is quite robust for possible misspecification of the median.

Keywords: Conditional inference; Distribution free; Robust tests; Runs test; Symmetry

1 INTRODUCTION

Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution with probability density function $f(x)$. The problem we consider is that of testing the symmetry of the distribution about a specified median. Several test statistics have been suggested in the literature for this problem including Butler (1969), Rothman and Woodroofe (1972), Hill and Rao (1977), McWilliams (1990). McWilliams (1990) performed a simulation study that showed the power superiority of his test over the other mentioned tests under various asymmetric alternatives. In this article we show how the runs test of McWilliams (1990) can be modified to achieve robustness against misspecification of the median while maintaining its superior power performance. The modified test is presented in Section 2. A simulation study to investigate the powers and the sizes of the tests when the median is misspecified is presented in Section 3. Results and conclusions are given in Section 4.

2 THE MODIFIED TEST

Assume, without any loss of generality, that the specified median is zero. Let $X_{(1)}, \ldots, X_{(n)}$ denote the ordered values of the sample from smallest to largest according to the absolute value (signs are retained). Let $S_1, \ldots, S_n$ be indicator variables defined as S_i is 1 if $X_{(i)}$ is nonnegative and zero otherwise. The test statistic is the total number of runs (R^*) in the $\{S_i\}$

* E-mail: baklizi1@hotmail.com

ISSN 1048-5252 print; ISSN 1029-0311 online © 2003 Taylor & Francis Ltd
DOI: 10.1080/10485250310001634737

43

sequence. Small values of this statistic provide evidence against symmetry. The rationale behind this statistic and its critical region has been explained in McWilliams (1990). He based his test on the unconditional distribution of $R^* - 1$, which he showed to have a binomial distribution with parameters $(n - 1)$ and $1/2$. The test we propose here is based on the conditional distribution of R^* given n_1 and n_2, the numbers of zeros and ones in the $\{S_i\}$ sequence. The conditional probability distribution of R^* when n_1 and n_2 are positive integers is given by Gibbons and Chakraborti, 1992.

$$f_{R^*}(r|n_1 \cdot n_2) = \begin{cases} \dfrac{2\dbinom{n_1-1}{r/2-1}\dbinom{n_2-1}{r/2-1}}{\dbinom{n_1+n_2}{n_1}}, & r > 1 \text{ and is even.} \\[2ex] \dfrac{\dbinom{n_1-1}{(r-1)/2}\dbinom{n_2-1}{(r-3)/2} + \dbinom{n_1-1}{(r-3)/2}\dbinom{n_2-1}{(r-1)/2}}{\dbinom{n_1+n_2}{n_1}}, & r > 1 \text{ and is odd.} \end{cases}$$

If either $n_1 = 0$ or $n_2 = 0$ then $r = 1$ with probability one.

Tables of critical values for this test for some small positive values of n_1 and n_2 are given in Gibbons and Chakraborti (1992).

The mean and variance of R^* are given by

$$E(R^*) = 1 + \frac{2n_1n_2}{n_1+n_2} \quad \text{and} \quad \operatorname{var}(R^*) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1 + n_2 - 1)}.$$

If n_1 and n_2 tend to infinity in such a way that $n_1/(n_1+n_2) \to \gamma$, where $0 < \gamma < 1$, then the standardized random variable

$$Z = \frac{R^* - 2(n_1+n_2)\gamma(1-\gamma)}{2(n_1+n_2)^{1/2}\gamma(1-\gamma)}$$

has a standard normal limiting distribution.

The robustness of the conditional runs test over the unconditional one may be anticipated from the following considerations. If the distribution is symmetric and the specified median is not actually correct then we would expect an imbalance between the number of zeros and ones in the $\{S_i\}$ sequence, and consequently the maximum number of possible runs [which is $2\min(n_1, n_2) + 1$] will be reduced, hence the tendency for producing small values of R^* will be increased, which in turn results in a larger number of rejections of the null hypothesis of symmetry having used the unconditional test of McWilliams (1990), hence an inflated size of the test. On the other hand, when using the conditional test, the imbalance of ones and zeros in the $\{S_i\}$ sequence is accounted for, and the critical values of the conditional test will be modified accordingly, case by case, depending on n_1 and n_2; the numbers of zeros and ones in the $\{S_i\}$ sequence. Thus, the fraction of rejections of the null hypothesis will be expected to be closer to the nominal size.

44

3 SIZE AND POWER PERFORMANCE OF THE MODIFIED TEST STATISTIC

A simulation study is conducted to investigate the performance of the test statistic and to compare it with the runs statistic of McWilliams (1990). The study is concerned with the following two criteria

(1) power of the tests; and
(2) robustness of the sizes of the tests against misspecification of the median.

For the first criterion, we used the following indices:

(a) the significance level of the test (α), which is taken to be 0.01, 0.05, and 0.1;
(b) the total sample size (n), which is taken to be 30 and 50; and
(c) the parameters ($\lambda_1, \lambda_2, \lambda_3, \lambda_4$) of the generalized lambda distribution (Ramberg and Schmeiser, 1974) used to generate the observations as

$$x_i = \lambda_1 + \frac{1}{\lambda_2}(u^{\lambda_3} - (1-u)^{\lambda_4}), \quad i = 1, \ldots, n,$$

where u is a uniform random number generated using the mathematical package (Mathematica, version 4.0). The following nine cases are considered [the shapes of the distributions are described in McWilliams (1990)];

Case 1: ($\lambda_1 = 0, \lambda_2 = 0.197454, \lambda_3 = 0.134915, \lambda_4 = 0.134915$). (Symmetric).
Case 2: ($\lambda_1 = 0, \lambda_2 = 1, \lambda_3 = 1.4, \lambda_4 = 0.25$)
Case 3: ($\lambda_1 = 0, \lambda_2 = 1, \lambda_3 = 0.00007, \lambda_4 = 0.1$)
Case 4: ($\lambda_1 = 3.586508, \lambda_2 = 0.04306, \lambda_3 = 0.025213, \lambda_4 = 0.094029$)
Case 5: ($\lambda_1 = 0, \lambda_2 = -1, \lambda_3 = -0.0075, \lambda_4 = -0.03$)
Case 6: ($\lambda_1 = -0.116734, \lambda_2 = -0.351663, \lambda_3 = -0.13, \lambda_4 = -0.16$)
Case 7: ($\lambda_1 = 0, \lambda_2 = -1, \lambda_3 = -0.1, \lambda_4 = -0.18$)
Case 8: ($\lambda_1 = 0, \lambda_2 = -1, \lambda_3 = -0.001$, [illegible] -0.13)
Case 9: ($\lambda_1 = 0, \lambda_2 = -1, \lambda_3 = -0.000$[illegible] $= -0.17$)

TABLE 1 Powers of the Tests.

		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$	
n	*Case*	*Uncond.*	*Cond.*	*Uncond.*	*Cond.*	*Uncond.*	*Cond.*
30	1	0.0101	0.0113	0.0460	0.0473	0.0973	0.0961
30	2	0.1217	0.1336	0.3138	0.3315	0.4413	0.4585
30	3	0.2228	0.242	0.4552	0.4794	0.5950	0.6166
30	4	0.0288	0.0297	0.1055	0.1122	0.1872	0.1901
30	5	0.0409	0.0438	0.1330	0.1418	0.2269	0.2367
30	6	0.0097	0.0106	0.0493	0.0499	0.0997	0.1004
30	7	0.0153	0.0174	0.0687	0.0722	0.1329	0.1356
30	8	0.2944	0.3143	0.5395	0.5632	0.6798	0.6997
30	9	0.3283	0.3544	0.5823	0.6056	0.7055	0.7263
50	1	0.0107	0.0101	0.0512	0.0498	0.1007	0.1011
50	2	0.2466	0.2551	0.4841	0.4991	0.6261	0.6391
50	3	0.4255	0.4384	0.6751	0.6850	0.7856	0.7975
50	4	0.0444	0.0458	0.1458	0.1521	0.2373	0.2440
50	5	0.0663	0.0695	0.1978	0.2010	0.3116	0.3204
50	6	0.0108	0.0095	0.0495	0.0514	0.1048	0.1048
50	7	0.0173	0.0188	0.0836	0.0848	0.1565	0.1540
50	8	0.5317	0.5464	0.7659	0.7735	0.8541	0.8638
50	9	0.5904	0.6004	0.8075	0.8186	0.8908	0.8971

6245

45

For each combination of these indices, 10,000 samples are generated and the power of each test is estimated as the fraction of the number of rejections of the null hypothesis of symmetry. The tests are randomized in order to achieve the nominal sizes. The results are given in Table I.

For the second criterion of size robustness against an incorrectly specified median, we used the following simulation indices:

(a) the nominal size of the test (α), which is taken to be 0.01, 0.05, 0.1;
(b) the sample size (n), which is taken to be 30 and 50;
(c) the parameter (λ) of the symmetric lambda distribution (Tukey, 1962) which is taken to be $-0.5, -0.3, -0.1, 0.1, 0.3, 0.5$. The data are generated as

$$x_i = \lambda + \frac{1}{\lambda}(u^\lambda - (1-u)^\lambda), \quad i = 1, \ldots, n.$$

(d) The misspecified median m_0, which is taken to be the 55, 60, 65, and 70th percentile of the distribution. (Results for the corresponding lower percentiles will be the same due to the symmetry of the distribution under the null hypothesis.)

For each combination of these indices, 10,000 samples are generated and the size of each test is found as the fraction of rejections of the null hypothesis of symmetry. The tests are randomized to achieve the same nominal sizes. The results are given in Tables II and III.

TABLE II Sizes of the Tests ($n = 30$).

		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$	
λ	m_0	*Uncond.*	*Cond.*	*Uncond.*	*Cond.*	*Uncond.*	*Cond.*
−0.5	0.50	0.0113	0.0115	0.0513	0.0491	0.1002	0.0990
−0.5	0.55	0.0137	0.0110	0.0578	0.0485	0.1079	0.0976
−0.5	0.60	0.0237	0.0124	0.0882	0.0569	0.1566	0.1073
−0.5	0.65	0.0545	0.0151	0.1576	0.0628	0.2565	0.1219
−0.5	0.70	0.1247	0.0187	0.2870	0.0786	0.4053	0.1404
−0.3	0.50	0.0094	0.0087	0.0505	0.0478	0.0997	0.0970
−0.3	0.55	0.0141	0.0115	0.0598	0.0514	0.1134	0.1023
−0.3	0.60	0.0226	0.0098	0.0920	0.0564	0.1637	0.1150
−0.3	0.65	0.0505	0.0136	0.1604	0.0674	0.2520	0.1211
−0.3	0.70	0.1237	0.0199	0.2841	0.0817	0.4017	0.1438
−0.1	0.50	0.0098	0.0090	0.0497	0.0492	0.0989	0.0976
−0.1	0.55	0.0138	0.0119	0.0590	0.0519	0.1136	0.1053
−0.1	0.60	[illegible]	0.0121	0.0923	0.0619	0.1629	0.1191
−0.1	0.65	[illegible]	[illegible]	0.1566	0.0702	0.2489	0.1254
−0.1	0.70	0.1288	0.0261	0.2917	0.0909	0.4108	0.1552
0.1	0.50	0.0105	0.0113	0.0513	0.0506	0.1032	0.1019
0.1	0.55	0.0118	0.0108	0.0586	0.0515	0.1140	0.1033
0.1	0.60	0.0272	0.0112	0.0955	0.0591	0.1686	0.1171
0.1	0.65	0.0603	0.0198	0.1786	0.0791	0.2796	0.1462
0.1	0.70	0.1386	0.0268	0.3174	0.1024	0.4381	0.1751
0.3	0.50	0.0109	0.0111	0.0513	0.0511	0.1002	0.1016
0.3	0.55	0.0131	0.0112	0.0600	0.0540	0.1199	0.1081
0.3	0.60	0.0270	0.0139	0.1050	0.0729	0.1851	0.1338
0.3	0.65	0.0765	0.0279	0.2055	0.1007	0.3106	0.1813
0.3	0.70	0.1682	0.0428	0.3671	0.1418	0.5001	0.2323
0.5	0.50	0.0111	0.0115	0.0492	0.0493	0.0975	0.0960
0.5	0.55	0.0151	0.0136	0.0713	0.0650	0.1295	0.1200
0.5	0.60	0.0378	0.0266	0.1303	0.0935	0.2155	0.1643
0.5	0.65	0.0981	0.0428	0.2625	0.1454	0.3945	0.2466
0.5	0.70	0.2291	0.0754	0.4621	0.2150	0.6024	0.3242

TABLE III Sizes of the Tests ($n = 50$).

		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$	
λ	m_0	*Uncond.*	*Cond.*	*Uncond.*	*Cond.*	*Uncond.*	*Cond.*
−0.5	0.50	0.0080	0.0089	0.0469	0.0503	0.0997	0.0996
−0.5	0.55	0.0125	0.0114	0.0621	0.0530	0.1228	0.1060
−0.5	0.60	0.0261	0.0119	0.1006	0.0569	0.1763	0.1067
−0.5	0.65	[illegible]	0.0157	0.1996	0.0664	0.3048	0.1249
−0.5	0.70	0.1330	0.0196	0.3844	0.0844	0.5093	[illegible]
−0.3	0.50	0.0102	0.0097	0.0453	0.0454	0.0948	0.0954
−0.3	0.55	0.0131	0.0100	0.0610	0.0527	0.1202	0.1044
−0.3	0.60	0.0268	0.0121	0.1016	0.0546	0.1751	0.1099
−0.3	0.65	0.0718	0.0168	0.2039	0.0703	0.3117	0.1312
−0.3	0.70	0.1832	0.0203	0.3835	0.0832	0.5120	0.1514
−0.1	0.50	0.0068	0.0073	0.0443	0.0422	0.0910	0.0893
−0.1	0.55	0.0139	0.0102	0.0636	0.0546	0.1204	0.1062
−0.1	0.60	0.0283	0.0112	0.1039	0.0573	0.1802	0.1154
−0.1	0.65	0.0753	0.0157	0.2059	0.0706	0.3179	0.1326
−0.1	0.70	0.2001	0.0269	0.4047	0.0959	0.5276	0.1748
0.1	0.50	0.0101	0.0090	0.0455	0.0475	0.0973	0.0940
0.1	0.55	0.0134	0.0110	0.0673	0.0580	0.1241	0.1075
0.1	0.60	0.0303	0.0155	0.1170	0.0672	0.1997	0.1302
0.1	0.65	0.0857	0.0226	0.2296	0.0884	0.3420	0.1564
0.1	0.70	0.2244	0.0358	0.4361	0.1212	0.5636	0.2017
0.3	0.50	0.0075	0.0084	0.0471	0.0480	0.0996	0.0976
0.3	0.55	0.0139	0.0112	0.0635	0.0530	0.1242	0.1097
0.3	0.60	0.0365	0.0188	0.1289	0.0825	0.2198	0.1471
0.3	0.65	0.1153	0.0333	0.2797	0.1209	0.4076	0.2096
0.3	0.70	0.2868	0.0608	0.5336	0.1864	0.6568	0.2902
0.5	0.50	0.0090	0.0095	0.0477	0.0491	0.0991	0.0998
0.5	0.55	0.0195	0.0168	0.0812	0.0697	0.1406	0.1278
0.5	0.60	0.0535	0.0297	0.1762	0.1184	0.2807	0.1979
0.5	0.65	0.1653	0.0632	0.3723	0.1954	0.5159	0.3066
0.5	0.70	0.3947	0.1152	0.6495	0.2889	0.7645	0.4195

4 RESULTS AND CONCLUSIONS

Concerning the power performance it appears that the proposed conditional test slightly dominates the unconditional test of McWilliams (1990). This is true for all sample sizes and significance levels considered.

With regard to the size robustness of the tests, it is clear that the unconditional test is quite sensitive to departures from the assumed median, while the conditional test shows great robustness even for misspecified medians that are actually the 70th percentile when λ is negative and the sample size is greater than 30. For positive λ and large samples, the robustness still achieved for misspecified medians that are actually the 55th or the 60th percentile. In all cases, the conditional test shows much greater robustness than the unconditional one.

This study shows that the conditional runs test while enjoying desirable robustness properties performs even better than the unconditional test in terms of power in all the situations under consideration.

References

Butler, C. C. (1969). A test for symmetry using the sample distribution function. *The Annals of Mathematical Statistics*, 40, 2211–2214.

Gibbons, J. D. and Chakraborti, S. (1992). *Nonparametric Statistical Inference*. Marcel Dekker, New York.

Hill, D. L. and Rao, P. V. (1977). Tests of symmetry based on Cramer–Von Mises statistics. *Biometrika*, **64**, 489–494.
McWilliams, T. P. (1990). A distribution-free test for symmetry based on a runs statistic. *Journal of the American Statistical Association*, **85**(412), 1130–1133.
Ramberg, J. S. and Schmeiser, B. W. (1974). An approximate method for generating asymmetric random variables. *Communications of the ACM*, **17**, 78–82.
Rothman, E. D. and Woodroofe, M. (1972). A Cramer–Von Mises type statistic for testing symmetry. *The Annals of Mathematical Statistics*, **43**, 2035–2038.
Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, **33**, 1–67.