Capítulo 15 Estadística no paramétrica

15.1 Introducción

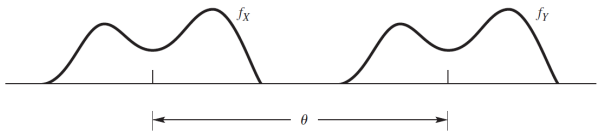
Algunos experimentos dan como resultado medidas de respuesta que desafían una cuantificación exacta. Los métodos estadísticos no paramétricos son útiles para analizar este tipo de datos. Los procedimientos estadísticos no paramétricos se aplican no sólo a observaciones que son difíciles de cuantificar, sino que también son particularmente útiles para hacer inferencias en situaciones en las que existe una seria duda acerca de las suposiciones que son la base de la metodología estándar.

Definiremos los *métodos paramétricos* como aquellos que se aplican a problemas donde la(s) distribución(es) de la(s) cual(es) se toman las muestras están especificadas excepto para los valores de un número finito de parámetros. Los métodos no paramétricos se aplican en todos los otros casos.

El empleo válido de algunos de los método paramétricos presentados en capítulos anteriores exige que se satisfagan al menos aproximadamente ciertas suposiciones de distribución. Incluso si se satisfacen todas las suposiciones, la investigación ha demostrado que las pruebas estadísticas no paramétricas son casi tan capaces de detectar diferencias entre poblaciones como los métodos paramétricos aplicables. Pueden ser y con frecuencia lo son, más potentes para detectar diferencias poblaciones cuando las suposiciones no se satisfacen.

15.2 Modelo general de desplazamiento (o cambio) de dos muestras

Sea $X_1, X_2, \ldots, X_{n_1}$ una muestra aleatoria de una población con función de distribución $F(x)$ y sea $Y_1, Y_2, \ldots, Y_{n_2}$ una muestra aleatoria de una población con función de distribución $G(y)$. Si deseamos probar que las dos poblaciones tienen la misma distribución, es decir, $H_0 : F(z) = G(z)$ contra $H_a : F(z) \neq G(z)$, con la forma real de $F(z)$ y $G(z)$ no especificada. Muchas veces se puede querer considerar la hipótesis alternativa más específica de que Y_1 tiene la misma distribución que X_1 desplazada una cantidad θ (desconocida), es decir, que las distribuciones *difieren en localización*. Un ejemplo de esto se puede ver en la siguiente figura.



Entonces se tiene, $G(y) = P(Y_1 \leq y) = P(X_1 \leq y - \theta) = F(y - \theta)$ para algún valor θ de parámetro desconocido. Observe que la forma particular de $F(x)$ continúa no especificada. Por lo tanto, para el modelo de desplazamiento de dos muestras, $H_0 : F(z) = G(z)$ es equivalente a $H_0 : \theta = 0$. Si θ es mayor (menor) que 0, entonces la distribución de los valores Y está ubicada a la derecha (izquierda) de la distribución de los valores X.

15.3 Prueba de signos para un experimento de observaciones pareadas

Suponga que tenemos n pares de observaciones de la forma (X_i, Y_i) y que deseamos probar la hipótesis de que la distribución de las X es igual a la de las Y, contra la alternativa de que las distribuciones difieren en ubicación. Hagamos $D_i = X_i - Y_i$ entonces, dada la hipótesis nula de que X_i y Y_i provienen de las mismas distribuciones continuas de probabilidad, la probabilidad de que D_i sea positiva es igual a $1/2$ (como lo es la probabilidad de que D_i sea negativa). Denotemos con M el número total de diferencias positivas (o negativas). Entonces, si las variables X_i y Y_i tienen la misma distribución, M tiene una distribución binomial con $p = 1/2$ y la región de rechazo para una prueba basada en M se puede obtener con el uso de la distribución binomial de probabilidad vista en el Capítulo 3.

Prueba de signos para un experimento de observaciones pareadas
Sea $p = P(X > Y)$.

Hipótesis nula: $H_0 : p = 1/2$.

Hipótesis alternativa: $H_a : p > 1/2$ o ($p < 1/2$ o $p \neq 1/2$).

Estadístico de prueba $M =$ número de diferencias positivas donde $D_i = X_i - Y_i$.

Región de rechazo: para $H_a : p > 1/2$, rechazar H_0 para los valores más grandes de M; para $H_a : p < 1/2$, rechazar H_0 para los valores más pequeños de M; para $H_a : p \neq 1/2$, rechazar H_0 para valores muy grandes o muy pequeños de M.

Suposiciones: los pares (X_i, Y_i) se seleccionan en forma aleatoria e independiente.

Un problema que puede surgir en relación con una prueba de signos es qeu las observaciones asociadas con uno o más pares pueden ser iguales y por lo tanto puede resultar en empate. Cuando se presente esta situación, borre los pares empatados y reduzca n, el número total de pares.

También puede haber situaciones donde n, el número de pares, es grande. Entonces los valores de α asociados con la prueba de signos pueden calcularse si se usa la aproximación normal a la distribución binomial de probabilidad que se estudio en la Sección 7.5. Estas aproximaciones serán bastante adecuadas para n de valor muy pequeño, 10 o 15. Para $n \geq 25$, la prueba Z del Capítulo 10 será suficiente, donde

$$Z = \frac{M - np}{\sqrt{npq}} = \frac{M - n/2}{(1/2)\sqrt{n}}$$

Prueba de signos para muestras grandes: n > 25

Hipótesis nula: $H_0 : p = 0.5$ (ninguno de estos tratamientos se prefiere al otro).

Hipótesis alternativa: $H_a : p \neq 0.5$ para una prueba de dos colas (*Nota*: usamos la prueba de dos colas como ejemplo. Muchos análisis requieren una prueba de una cola).

Estadístico de prueba: $Z = \dfrac{M - n/2}{(1/2)\sqrt{n}}$

Región de rechazo: rechazar H_0 si $z \geq z_{\alpha/2}$ o si $z \leq z_{\alpha/2}$, donde $z_{\alpha/2}$ se obtiene de la Tabla 3, Apéndice 3.

La prueba de signos en realidad prueba la hipótesis nula de que la *mediana* de las variables D_i es cero contra la alternativa de que es diferente de cero. [Si la mediana de las variables D_i es cero, implica que $P(D_i < 0) = P(D_i > 0)$.] Si las variables X_i y Y_i tienen la misma distribución, la mediana de las variables D_i será cero.

15.4 Prueba de rangos con signo de Wilcoxon para un experimento de observaciones pareadas

Al igual que en la sección pasada, suponga que tenemos n observaciones pareadas de la forma (X_i, Y_i) y que $D_i = X_i - Y_i$. De nuevo suponemos que estamos interesados en probar la hipótesis de que las X y las Y tienen la misma distribución contra la alternativa de que las distribuciones difieren en localización. De acuerdo con la hipótesis nula de que no hay diferencias en las distribuciones de las X y las Y, se esperaría (en promedio) que la mitad de las diferencias en los pares sean negativas y la mitad positivas. Esto es, el número esperado de diferencias negativas entre pares es $n/2$ (donde n es el número de pares).

Para llevar a cabo la prueba de Wilcoxon calculamos las diferencias (D_i) para cada uno de los n pares. Las diferencias iguales a cero se eliminan y el número de pares, n, se reduce de conformidad. Entonces clasificamos los *valores absolutos* de las diferencias asignando un 1 al más pequeño, un 2 al segundo más pequeño y así sucesivamente. Si dos o más diferencias absolutas están empatadas para el mismo rango, entonces el promedio de las clasificaciones que se hubieran asignado a estas diferencias se asigna a cada miembro del grupo empatado.

Prueba de rangos con signo de Wilcoxon para un experimento de observaciones pareadas

H_0 : las distribuciones poblacionales para las X y las Y son idénticas.

H_a : (1) las dos distribuciones poblacionales difieren en localización (dos colas), *o bien* (2) la distribución de frecuencia relativa poblacional para las X se desplaza a la derecha de la de las Y (una cola).

Estadístico de prueba:

1. Para una prueba de dos colas, use $T = \min(T^+, T^-)$, donde T^+ = suma de los rangos de las diferencias positivas y T^- = suma de los rangos de las diferencias negativas.
2. Para una prueba de una cola (para detectar la alternativa de una cola que acabamos de dar), use la suma de rango T^- de las diferencias negativas.

Región de rechazo:

1. Para una prueba de dos colas, rechace H_0 si $T \leq T_0$, donde T_0 es el valor crítico para la prueba de dos lados dada en la Tabla 9, Apéndice 3.
2. Para una prueba de una cola (como se describe líneas antes), rechace H_0 si $T^- \leq T_0$, donde T_0 es el valor crítico para la prueba unilateral.

Para detectar un desplazamiento de la distribución de las Y a la derecha de la distribución de las X, use la suma de rango T^+, la suma de los rangos de las diferencias positivas y rechace H_0 si $T^+ \leq T_0$.

Aunque la Tabla 9, Apéndice 3m es aplicable para valores de n (el número de pares de datos) de hasta $n = 50$, es conveniente observar que T^+ (o T^-) estará distribuida normalmente en forma aproximada cuando la hipótesis nula es verdadera y n sea grande (25 o más por ejemplo). Esto hace posible que construyamos una prueba Z de muestra grande, donde si $T = T^+$,

$$E(T^+) = \frac{n(n+1)}{4} \qquad V(T^+) = \frac{n(n+1)(2n+1)}{24}$$

Entonces el estadístico Z se puede usar como estadístico de prueba.

$$Z = \frac{T^+ - E(T^+)}{\sqrt{V(T^+)}} = \frac{T^+ - [n(n+1)/4]}{\sqrt{n(n+1)(2n+1)/24}}$$

Una prueba de rangos con signo de Wilcoxon con muestras grandes para un experimento de observaciones pareadas: n > 25

Hipótesis nula: H_a : las distribuciones de frecuencia relativa poblacionales para las X y Y son idénticas.

Hipótesis alternativa: (1) H_a : las dos distribuciones de frecuencia relativa poblacionales difieren en localización (una prueba de dos colas), *o bien*,

(2) la distribución de frecuencia relativa poblacional para las X está desplazada a la derecha (o izquierda) de la distribución de frecuencia relativa de las Y (pruebas de una cola).

Estadístico de prueba $Z = \dfrac{T^+ - [n(n+1)/4]}{\sqrt{n(n+1)(2n+1)/24}}$

Región de rechazo: rechazar H_0 si $z \geq z_{\alpha/2}$ o $z \leq -z_{\alpha/2}$ para una prueba de dos colas. Para detectar un desplazamiento en las distribuciones de las X a la derecha de las Y, rechazar H_0 cuando $z \geq z_\alpha$: para detectar un desplazamiento en la dirección opuesta, rechazar H_0 si $z \leq -z_\alpha$.

15.5 Uso de rangos para comparar dos distribuciones poblacionales: muestras aleatorias independientes

Una prueba estadística para comparar dos poblaciones basadas en muestras aleatorias independientes, la prueba de *suma de rangos*, fue propuesta por Frank Wilcoxon. De nuevo suponemos que estamos interesados en probar si las dos poblaciones tienen la misma distribución contra el desplazamiento (o localización) alternativo.

Supongamos que usted debe seleccionar muestras aleatorias independientes de n_1 y n_2 observaciones de las poblaciones I y II, respectivamente. Se combinan las $n_1 + n_2 = n$ observaciones y clasificarlas en orden de magnitud, de 1 (la más pequeña) a n (la más grande). Los empates se trata como en la sección pasada.

Si las observaciones se seleccionaron de poblaciones idénticas, las *sumas de rango* para las muestras deben ser más o menos proporcionales a los tamaños muestrales n_1 y n_2. En contraste, si las observaciones en una población, la I por ejemplo, tendían a ser mayores que las de la población II, las observaciones de la muestra I tenderían a recibir las clasificaciones más altas y la muestra I tendría una suma de rango mayor que lo esperado. Entonces (con los tamaños muestrales siendo iguales), si una suma de rango es muy grande (y, de modo correspondientes, la otra es muy pequeña), esto puede indicar una diferencia importante, desde el punto de vista estadístico entre las localizaciones de las dos poblaciones.

15.6 Prueba de U de Mann-Whitney: muestras aleatorias independientes

El estadístico U de Mann-Whitney se obtiene al ordenar todas las $(n_1 + n_2)$ observaciones según su magnitud y al contar el número de observaciones en la muestra I que precede a cada observación de la muestra II. El estadístico U es la suma de estas cantidades. En el resto de esta sección denotamos las observaciones en la muestra I como $x_1, x_2, \ldots, x_{n_1}$ y las observaciones en la muestra II como $y_1, y_2, \ldots, y_{n_2}$

Valores muy grandes o muy pequeños de U implican una separación de las x y las y ordenadas, en consecuencia demuestran que existe una diferencia (un desplazamiento de localización) entre las distribuciones de las poblaciones I y II.

Fórmula para el estadístico U de Mann-Whitney

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - W$$

donde

n_1 = número de observaciones en la muestra I
n_2 = número de observaciones en la muestra II
W = suma de rango para la muestra I.

Algunos resultados útiles acerca de la distribución de U:

1. Los posibles valores de U son $0, 1, 2, \ldots, n_1 n_2$.
2. La distribución de U es simétrica alrededor de $(n_1 n_2)/2$. Esto es, para cualquier $a > 0$, $P[U \leq (n_1 n_2)/2 - a] = P[U \geq (n_1 n_2)/2 + a]$.
3. El resultado en (2) implica que $P(U \leq U_0) = P(U \geq n_1 n_2 - U_0)$.

La prueba U de Mann-Whitney
La población I es la población de la cual se tomó la muestra más pequeña.

Hipótesis nula: H_0 : las distribuciones de las poblaciones I y II son idénticas.

Hipótesis alternativa: (1) H_a : las distribuciones de las poblaciones I y II tienen localizaciones diferentes (una prueba de dos colas), *o bien*

(2) la distribución de la población I se desplaza a la derecha de la distribución de la población II, *o*
(3) la distribución de la población I se desplaza a la izquierda de la distribución de la población II.

Estadístico de prueba:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - W$$

La prueba U de Mann-Whitney

Región de rechazo: (1) para la prueba de dos colas y un valor dado de α, rechazar H_0 si $U \leq U_0$ o $U \geq n_1n_2 - U_0$, donde $P(U \leq U_0) = \alpha/2$[*Nota:* observe que U_0 es el valor tal que $P(U \leq U_0)$ es igual a la mitad de α.]

(2) Para probar que la población I está desplazada a la derecha de la población II con un valor determinado de α, rechazar H_0 si $U \leq U_0$, donde $P(U \leq U_0) = \alpha$.

(3) Para probar que la población I está desplazada a la izquierda de la población II con un valor determinado de α, rechazar H_0 si $U \geq n_1n_2 - U_0$, donde $P(U \leq U_0) = \alpha$.

Suposiciones: las muestras se han seleccionado aleatoriamente y de manera independiente de sus poblaciones respectivas. Los empates en las observaciones se pueden manejar al promediar los rangos que hubieran sido asignados a las observaciones empatadas y asignar este rango promedio a cada una.

Una prueba simplificada para muestras grandes ($n_1 > 10$ y $n_2 > 10$) se puede obtener con el uso del ya conocido estadístico Z. Cuando las poblaciones son idénticas, se puede demostrar que el estadístico U tiene los siguientes valores,

$$E(U) = \frac{n_1n_2}{2} \qquad V(U) = \frac{n_1n_2(n_1 + n_2 + 1)}{12}$$

También, cuando n_1 y n_2 son grandes el estadístico Z tiene aproximadamente un distribución normal estándar.

$$Z = \frac{U - E(U)}{\sigma_U}$$

La prueba de U de Mann-Whitney para muestras grandes
$n_1 > 10$ **y** $n_2 > 10$

Hipótesis nula: H_0 : las distribuciones de frecuencia relativa para las poblaciones I y II son idénticas.

Hipótesis alternativa: (1) H_a : las distribuciones de frecuencia relativa de las dos poblaciones difieren en ubicación (una prueba de dos colas), *o bien*,

(2) la distribución de frecuencia relativa para la población I está desplazada a la derecha (o la izquierda) de la distribución de frecuencia relativa para la población II (una prueba de una cola).

Estadístico de prueba:

$$Z = \frac{U - (n_1n_2/2)}{\sqrt{n_1n_2(n_1 + n_2 + 1)/12}}.$$

Región de rechazo: rechazar H_0 si $z > z_{\alpha/2}$ o $z < -z_{\alpha/2}$ para una prueba de dos colas.

La prueba de U de Mann-Whitney para muestras grandes
$n_1 > 10$ **y** $n_2 > 10$

Para una prueba de una cola, ponga toda α en una cola de la distribución z. Para detectar un desplazamiento en la distribución de la población I a la derecha de la población II, rechace H_0 cuando $z < -z_\alpha$. Para detectar un desplazamiento en la dirección contraria, rechace H_0 cuando $z > z_\alpha$. Los valores tabulados de z se dan en la Tabla 4, Apéndice 3.

15.7 La prueba de Kruskal-Wallis para un diseño de un factor

Al igual que en la sección 13.3 suponemos que muestras aleatorias independientes han sido tomadas de k poblaciones que difieren sólo en localización, aunque no es necesario suponer que estas poblaciones poseen distribuciones normales. Para una generalización completa, permitimos que los tamaños muestrales sean desiguales y con n_i, para $i = 1, 2, \ldots, k$, representamos el tamaños de la muestra tomada de la i-ésima población. Igual que en el procedimiento de la Sección 15.5 combinamos todas las $n_1 + n_2 + \cdots + n_k = n$ observaciones y las clasificamos desde 1 (la más pequeña) hasta n (la más grande). Los empates se tratan como en las secciones pasadas.

Con R_i denotamos la suma de los rangos de las observaciones de la población i y con $\overline{R}_i = R_i/n_i$ denotamos el promedio correspondiente de los rangos. Si $\overline{R}$ es igual al promedio general de los rangos, consideramos el rango análogo SST, que se calcula usando los rangos en lugar de los valroes reales de las mediciones:

$$V = \sum_{i=1} kn_i(\overline{R}_i - \overline{R})^2$$

Si la hipótesis nula es verdadera y las poblaciones no difieren en localización, se espera que los valores de $\overline{R}_i$ sean aproximadamente iguales y el valor resultante de V sea relativamente pequeño. Si la hipótesis alternativa es verdadera esperamos valores grandes para V. Observe que

$$\overline{R} = \frac{(\text{suma de los primeros } n \text{ enteros})}{n} = \frac{[n(n+1)/2]}{n} = \frac{n+1}{2}$$

Entonces se tiene lo siguiente,

$$V = \sum_{i=1}^{k} n_i\left(\overline{R}_i - \frac{n+1}{2}\right)^2$$

En ligar de concentrarnos en V se necesita utilizar el estadístico $H = 12V/[n(n+1)]$, esto es,

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

La hipótesis nula de localizaciones iguales es rechazada a favor de la alternativa de que las poblaciones difieren en localización si el valor de H es grande. Así, la correspondiente prueba de nivel α pide el rechazo de la hipótesis nula a favor de la alternativa si $H > h(\alpha)$, donde $h(\alpha)$ es tal que, cuando H_0 es verdadera $P[H > h(\alpha)] = \alpha$

Si los n_i valores son 'grandes', la distribución nula de H se puede aproximar con una distribución χ^2 con $k - 1$ grados de libertad.

Prueba de Kruskal-Wallis basada en H para comparar k distribuciones poblacionales

Hipótesis nula: H_0 : las k distribuciones poblacionales son idénticas.

Hipótesis alternativa: H_a : al menos dos de las distribución poblacionales difieren en localización.

Estadístico de prueba:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1), \text{ donde}$$

n_i = número de mediciones en la muestra de la población i,

R_i = suma de rangos para la muestra i, donde el rango de cada medida se calcula de acuerdo con su tamaño relativo en el conjunto general de $n = n_1 + n_2 + \cdots + n_k$ observaciones formadas al combinar los datos de todas las k muestras.

Región de rechazo: rechazar H_0 si $H > \chi^2_\alpha$ con $(k-1)$ grados de libertad.

Suposiciones: las k muestras se sacan al azar y en forma independiente. Hay cinco o más mediciones en cada muestra.

15.8 La prueba de Friedman para diseños de bloques aleatorizados

La prueba de Friedman está diseñada para probar la hipótesis nula de que las distribuciones de probabilidad de los k tratamientos son idénticos contra la alternativa de que al menos dos de las distribuciones difieren en localización. Se calcula de la siguiente manera.

Después de obtener los datos de un diseño de bloques aleatorizado, *dentro de cada bloque* los valores observados de las respuestas a cada uno de los k tratamientos se clasifican de 1 (el más pequeño del bloque) a k (el más grande del bloque). Si dos o más observaciones del *mismo bloque* están empatadas se aplica el mismo procedimiento que en las secciones pasadas. No obstante, los empates necesitan resolverse de este modo sólo su se presentan dentro del mismo bloque.

Con R_i denote la suma de los rangos de las observaciones correspondientes al tratamiento i y con $\overline{R}_i = R_i/b$ denote el promedio correspondiente de los rangos (note que como es un diseño de bloques aleatorizado hay un total de bk observaciones). Debido a que los rangos de 1 a k están asignados dentro de cada bloque, la suma de los

rangos asignados en cada bloque es $1+2+\cdots+k = k(k+1)/2$. Así la suma de todos los rangos asignados en el análisis es $bk(k+1)/2$. Si $\overline{R}$ denota el promedio general de los rangos de todas las observaciones bk, se deduce que $\overline{R} = (k+1)/2$. Considere el equivalente en rangos de la SSY para un diseño de bloques aleatorizado dado por

$$W = b\sum_{i=1}^{k}(\overline{R}_i - \overline{R})^2$$

Si la hipótesis nula es verdadera y las distribuciones de probabilidad de las respuestas del tratamiento no difieren en localización, esperamos que los valores $\overline{R}_i$ sean aproximadamente iguales y el valor resultante para W sea pequeño. Si la hipótesis alternativa fuera verdadera se esperarían valores grandes de W. En lugar de W, Friedman consideró el estadístico $F_r = 12W/[k(k+1)]$, que se puede reescribir como

$$F_r = \frac{12}{bk(k+1)}\sum_{i=1}^{k}R_i^2 - 3b(k+1)$$

Como vimos antes, la hipótesis nula de localizaciones iguales es rechazada a favor de la alternativa de que las distribuciones de tratamiento difieren en localización si el valor de F_r es grande. Esto es, la prueba de nivel α correspondiente rechaza la hipótesis nula a favor de la alternativa si $F_r > f_r(\alpha)$ es tal que, cuando H_0 es verdadera, $P[F_r > f_r(\alpha)] = \alpha$. Los valores seleccionados de $f_r(\alpha)$ para varias opciones de k y b se muestran en la Tabla A.22 de Hollander y Wolfe (1999).

La distribución nula del estadístico F_r de Friedman se puede calcular con una distribución χ^2 con $k-1$ grados de libertad mientras b sea 'grande'. La evidencia empírica indica que la aproximación es adecuada si b (el número de bloques) o k (el número de tratamientos) es mayor que 5.

Prueba de Friedman basada en F_r para un diseño de bloques aleatorizado

Hipótesis nula: H_0 : las distribuciones de probabilidad para los k tratamientos son idénticas.

Hipótesis alternativa: H_a : al menos dos de las distribuciones difieren en localización.

Estadístico de prueba:

$F_r = \frac{12}{bk(k+1)}\sum_{i=1}^{k}R_i^2 - 3b(k+1)$, donde

b = número de bloques,

k = número de tratamientos,

R_i = suma de los rangos para el i-ésimo tratamiento, donde el rango de cada medición se calcula con respecto a su tamaño dentro de su propio bloque.

Región de rechazo: $F_r > \chi^2_\alpha$ con $(k-1)$ grados de libertad.

Suposiciones: los tratamientos se asignan al azar a unidades experimentales dentro de los bloques. Ya sea que el número de bloques (b) o el número de tratamientos (k) excedan de 5.

15.9 Prueba de corridas de ensayo: una prueba de aleatoriedad

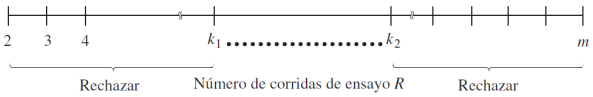
Considere un proceso de producción en le que piezas manufacturadas salen en secuencia y cada una es clasificada como defectuosa (D) o no defectuosa (N). En ocasiones pruebas para encontrar la fracción de piezas defectuosas no detectara defectos periódicos y corridos en la producción. Una forma de encontrar si hay o no aleatoriedad en la aparición de piezas defectuosas es con una *prueba de aleatoriedad*. En este capítulo se ver una de estas pruebas conocida como *prueba de corridas de ensayo*.

Como su nombre lo indica, la pruebas de corridas de ensayo se usan para estudiar una secuencia de eventos en la que cada elemento de la secuencia puede tomar uno de dos resultados, no defectuosos (S) o defectuoso (F).

Definición 15.1

Una *corrida de ensayo* es una subsecuencia máxima de elementos semejantes.

Un número muy grande o muy pequeño de corridas de ensayo en una secuencia indica no aleatoriedad. Por lo tanto, sea R (el número de corridas de ensayo en una secuencia) el estadístico de prueba y sean $R \leq k_1$ y $R \geq k_2$ la región de rechazo, como se muestra en la siguiente figura.



Supongamos que la secuencia contiene n_1 elementos S y n_2 elementos F, resultando en Y_1 corridas de ensayo de las S y Y_2 corridas de ensayo de las F, donde $(Y_1+Y_2) = R$. La distribución de probabilidad para R, $P(R=r)$ esta dada por la probabilidad de exactamente y_1 corridas de ensayo de las S y y_2 corridas de ensayo de las F, esto es

$$p(y_1,y_2) = \frac{\binom{n_1-1}{y_1-1}\binom{n_2-1}{y_2-1}}{\binom{n_1+n_2}{n_1}}$$

Entonces, $P(R=r)$ es igual a la suma de $p(y_1,y_2)$ en todos los valores de y_1 y y_2 tales que $(y_1+y_2) = r$. A continuación se muestra la demostración de esta distribución de probabilidad.

Demostración

Suponemos que la secuencia contiene n_1 elementos S y n_2 elementos F, resultando en Y_1 corridas de ensayo de las S y Y_2 corridas de ensayo de las F, donde $(Y_1+Y_2) = R$. Entonces para una Y_1 determinada, Y_2 puede ser igual a Y_1, (Y_1-1) o (Y_1+1). Con m denotamos el número máximo posible de corridas de ensayo. Observe que $m = 2n_1$ si $n_1 = n_2$ y que $m = (2n_1+1)$ si $n_1 < n_2$.

Supondremos que todo arreglo distinguible de los (n_1+n_2) elementos de la secuencia constituye un evento simple y que los puntos muestrales son igualmente probables. Como el número total de arreglos distinguibles de n_1 elementos S y n_2 elementos F es

$$\binom{n_1+n_2}{n_1} \quad \Rightarrow \quad P(\text{por punto muestral}) = \frac{1}{\binom{n_1+n_2}{n_1}}$$

El número de formas de lograr y_1 corridas de ensayo S es igual al número de arreglos identificables de n_1 elementos indistinguibles en y_1 celdas, ninguna de las cuales está vacía, como se representa en la siguiente figura,

$$|S|SSSS|SS\ldots|SS|SSS|S|$$

Esto es igual al número de formas de distribuir (y_1-1) barras interiores en los (n_1-1) espacios entre los S elementos (las dos barras exteriores permanecen fijas). EN consecuencia es igual al número de formas de seleccionar (y_1-1) espacios (para las barras) fuera de los (n_1-1) espacios disponibles, esto es

$$\binom{n_1-1}{y_1-1}$$

El número de formas de observar y_1 corridas de ensayo S y y_2 corridas de ensayo F, obtenido al aplicar la regla mn es

$$\binom{n_1-1}{y_1-1}\binom{n_2-1}{y_2-1}$$

Esto da el número de puntos muestrales en el evento 'y_1 corridas de ensayo de las S y y_2 corridas de ensayo de las F'. Entonces al multiplicar este número por la probabilidad por punto muestra, obtenemos la probabilidad de exactamente y_1 corridas de ensayo de las S y y_2 corridas de ensayo de las F:

$$p(y_1,y_2) = \frac{\binom{n_1-1}{y_1-1}\binom{n_2-1}{y_2-1}}{\binom{n_1+n_2}{n_1}}$$

Los valores de $P(R \leq \alpha)$ se dan en la Tabla 10, Apéndice 3, para todas las combinaciones de n_1 y n_2, donde n_1 y n_2 son menores o iguales a 10.

Al igual que en el caso de los otros estadísticos de prueba no paramétricos estudiados en secciones anteriores de este capítulo, la distribución de probabilidad para R tiende hacia la normalidad cuando n_1 y n_2 se hacen grandes. La aproximación es buena cuando n_1 y n_2 sean ambos mayores que 10. Por esta razón se puede usar el estadístico de prueba Z, donde

$$Z = \frac{R - R(R)}{\sqrt{V(R)}}$$

y

$$E(R) = \frac{2n_1n_2}{n_1 + n_2} + 1$$
$$V(R) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}$$

La región de rechazo para una prueba de dos colas, con $\alpha = 0.05$, es $|z| \geq 1.96$. Si α es la probabilidad deseada de un error tipo I, para una prueba de cola superior, rechazamos la hipótesis nula si $z > z_\alpha$ (para una prueba de cola inferior, rechazamos H_0 si $z < -z_\alpha$).

15.10 Coeficiente de correlación de rangos

En las secciones anteriores usamos rangos para indicar la magnitud relativa de observaciones en pruebas no paramétricas para comparación de tratamientos. Ahora empleamos las misma técnica para probar una correlación entre dos variables clasificadas. Recuerde, de la Sección 11.8, que el coeficiente de correlación muestral para observaciones $)x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ está dado por

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$
$$= \frac{\sum_{i=1}^{n} x_iy_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]\left[\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

Con $R(x_i)$ denote el rango de x_i entre $x_1, x_2, \ldots, x_n$ y con $R(y_i)$ denote el rango de y_i entre $y_1, y_2, \ldots, y_n$. El *coeficiente de correlación de rango de Spearman*, r_S, se calcula al sustituir los rangos como las mediciones pareadas en la fórmula anterior. Así, r_S tiene la siguiente forma.

$$\frac{\sum R(x_i)R(y_i) - \frac{1}{n}\left(\sum R(x_i)\right)\left(\sum R(y_i)\right)}{\sqrt{\left[\sum [R(x_i)]^2 - \frac{1}{n}\left(\sum R(x_i)\right)^2\right]\left[\sum [R(y_i)]^2 - \frac{1}{n}\left(\sum R(y_i)\right)^2\right]}}$$

Cuando no hay empates en las observaciones x ni en las observaciones y, está expresión para r_S se reduce algebraicamente a una expresión más sencilla:

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

donde $d_i = R(x_i) - R(y_i)$.

El coeficiente de correlación de rango de Spearman se puede emplear como estadístico de prueba para probar la hipótesis de que no hay asociación entre dos poblaciones. Suponemos que los n pares de observaciones (x_i, y_i) se han seleccionado al azar y, por lo tanto, la ausencia de cualquier asociación entre las poblaciones implica una asignación aleatoria de los n rangos dentro de cada muestra. Cada asignación aleatoria (para las dos muestras) representa un punto muestral asociado con el experimento y un valor r_S se puede calcular para cada uno. Es posible calcular la probabilidad de que r_S tome un valor absoluto grande debido sólo a la casualidad y, por lo tanto, sugiere una asociación entre poblaciones aun cuando existe ninguna.

La región de rechazo para una prueba de dos colas incluye valores de r_S cercanos a +1 y a -1. SI la alternativa es que la correlación entre X y Y es negativa, rechazamos H_0 para valores de r_S cercanos a -1. De manera similar, si la alternativa es que la correlación entre X y Y es positiva, rechazamos H_0 para valores positivos grandes de r_S.

Prueba de correlación de rango de Spearman

Hipótesis nula: H_0 : no hay asociación entre los pares de rangos.

Hipótesis alternativa: (1) H_a : hay asociación entre los pares de rangos (una prueba de dos colas), *o bien*,

(2) la correlación entre los pares de rangos es positiva (o negativa) (una prueba de una cola).

Estadístico de prueba:

$$r_S = \frac{n\sum R(x_i)R(y_i) - [\sum R(x_i)][\sum R(y_i)]}{\sqrt{[n\sum [R(x_i)]^2 - [\sum R(x_i)]^2][n\sum [R(y_i)]^2 - [\sum R(y_i)]^2]}}$$

donde $R(x_i)$ y $R(y_i)$ denotan el rango de x_i entre $x_1, x_2, \ldots, x_n$ y y_i entre $y_1, y_2, \ldots, y_n$, respectivamente.

Región de rechazo: Para una prueba de dos colas, rechazar H_0 si $r_S \geq r_0$ o $r_S \leq r_0$, donde r_0 se da en la Tabla 11, Apéndice 3. Duplique la probabilidad tabulada para obtener el valor α para la prueba de dos colas. Para una prueba de una cola, rechace H_0 si $r_S \geq r_0$ (para una prueba de cola superior) o $r_S \leq -r_0$ (para una prueba de cola inferior). El valor α para una prueba de una cola es el que se muestra en la tabla 11, Apéndice 3.