

Inferencia

Estadística

(Teoría y problemas)

I. Espejo Miranda
F. Fernández Palacín
M. A. López Sánchez
M. Muñoz Márquez
A. M. Rodríguez Chía
A. Sánchez Navas
C. Valero Franco

© Servicio de Publicaciones. Universidad de Cádiz
I. Espejo Miranda, F. Fernández Palacín, M. A. López Sánchez, M. Muñoz
Márquez, A. M. Rodríguez Chía, A. Sánchez Navas, C. Valero Franco

Edita: Servicio de Publicaciones de la Universidad de Cádiz
c/ Doctor Marañón, 3. 11002 Cádiz (España)
www.uca.es/publicaciones

ISBN: 978-84-9828-131-6

Se concede permiso para copiar, distribuir y/o modificar este documento bajo los términos de la Licencia de Documentación Libre de GNU, Versión 1.2 o cualquier otra versión posterior publicada por la Free Software Foundation. Una traducción de la licencia está incluida en la sección titulada "Licencia de Documentación Libre de GNU".

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation. A copy of the license is included in the section entitled "GNU Free Documentation License".

Capítulo 5

Contrastes no paramétricos

1. Introducción

En este capítulo, bajo el título común de contrastes no paramétricos, se abordan una serie de técnicas inferenciales bien diferenciadas que permiten resolver problemas de diversa índole. En primer lugar, se tratan los tests que garantizan la calidad de la información muestral. En segundo lugar, se estudia si la población se adapta a alguna estructura de probabilidad conocida, con especial atención al caso de la Normal. La situación de no normalidad obligará a la aplicación de los contrastes no paramétricos sobre valores de parámetros o sobre medidas de posición, centralización, etc.

En el capítulo anterior, se comentó que lo ideal es utilizar contrastes paramétricos siempre que sea posible, puesto que si se conoce la distribución de probabilidad, se consiguen mejores resultados incorporando esta información en la construcción del test. En el caso de que no se disponga de dicha información, se recurre a métodos alternativos basados exclusivamente en la información que proporciona la muestra, como el análisis de signos, el análisis de rangos, la función de distribución empírica, etc.

El no cumplimiento de alguna de las hipótesis a la hora de realizar

un contraste paramétrico tiene incidencias diversas. Por ejemplo, si falla el supuesto de normalidad al realizar un contraste sobre la media, pero la muestra es grande, el Teorema Central del Límite garantiza que la media muestral es asintóticamente Normal y por tanto podría aplicarse el contraste paramétrico. Sin embargo, en un contraste de igualdad de medias donde las varianzas de las poblaciones sean distintas y haya problemas de normalidad, es conveniente recurrir a la alternativa no paramétrica; con más motivo si los tamaños muestrales son muy distintos.

El concepto que subyace en el párrafo anterior se conoce en la terminología estadística como Robustez. Formalmente, un procedimiento se dice *robusto* si es “aproximadamente” válido cuando se producen desviaciones respecto a los supuestos que se exigen en su aplicación. En general los contrastes paramétricos para poblaciones Normales son robustos frente a desviaciones de normalidad y no tanto frente a desviaciones de *homocedasticidad* (igualdad de varianzas); en todo caso, es deseable trabajar con *problemas equilibrados*, esto es, con tamaños muestrales parecidos.

Como se ha venido comentando, la muestra es la materia prima que hace posible la realización del estudio inferencial. De su calidad dependerá la bondad de los resultados que se puedan ofrecer, de forma que una muestra con sesgos, falta de representatividad, dependencia entre sus valores, presencia de valores anómalos, etc., condicionará las conclusiones de la investigación, hasta el punto de distorsionar totalmente la realidad del universo que se pretende estudiar. En definitiva, el proceso inferencial es muy sensible a los desajustes provocados por la utilización de una muestra contaminada.

De los posibles problemas que puede presentar una muestra, algunos son soslayables, mientras que otros obligan a prescindir de ella y a realizar un nuevo diseño y una posterior extracción de las unidades muestrales. Entre los primeros se encuentra la presencia en la muestra de *valores anómalos*, extraños o en la terminología anglosajona “outliers”, mientras que entre los segundos, el más usual es el de la falta de aleatoriedad de los datos.

De esta forma, los test no paramétricos además de permitir estudiar una población cuando se desconoce la distribución de la misma, tiene otra utilidad como es la de comprobar que las hipótesis exigidas para llevar a cabo un contraste paramétrico realmente son satisfechas.

2. Análisis de la calidad de la muestra

2.1. Test de rachas

El test de rachas es un contraste de aleatoriedad basado exclusivamente en el análisis de los signos. A modo de introducción, se supone la siguiente situación:

Ejemplo 5.1 *En dos colas de un multicines se encuentra la siguiente ordenación de hombres (H) y mujeres (M):*

Cola 1 : H H H H H H M M M M M M

Cola 2 : H M H M H M H M H M H M

La pregunta que se hace es, ¿llegaron cada una de las personas de forma independiente a la cola, o vinieron en grupos?

La respuesta para un observador avisado parece obvia, en el primer caso da la impresión de que se trata de dos grupos, uno de cada sexo, mientras que en el segundo todo parece indicar que aparecen varias parejas.

En situaciones no tan extremas cabe preguntarse si se da o no la independencia y para contestar a esta cuestión se utiliza el test de rachas.

En primer lugar, se supone una muestra dada en el orden de extracción. Nótese que el orden de extracción es fundamental, puesto que los problemas de aleatoriedad se derivan del hecho de muestrear individuos que son vecinos en algún sentido: físico, temporal, etc. Se denomina *racha* a una sucesión de uno o más elementos de un mismo tipo, llámese A, que está seguida y precedida por un elemento de otro tipo, llámese

B. Se cuenta el número de rachas, R_{exp} , y, a continuación, el número de elementos de cada tipo: n_1 los de tipo A y n_2 los de tipo B.

La región crítica del test viene dada por

$$R_c : \{R_{exp} \leq R_{n_1, n_2, \frac{\alpha}{2}}\} \cup \{R_{exp} \geq R_{n_1, n_2, 1 - \frac{\alpha}{2}}\}$$

donde los valores críticos $R_{n_1, n_2, \frac{\alpha}{2}}$ y $R_{n_1, n_2, 1 - \frac{\alpha}{2}}$ vienen dados en la tabla A.21. Nótese que este test rechaza la hipótesis de independencia cuando el número de rachas es pequeño o cuando éste es muy grande (aparición de ciclicidad).

Ejemplo 5.2 *En el ejemplo anterior aparecen dos rachas en la Cola 1 y doce rachas en la Cola 2. Para un nivel de significación del 5 %, la tabla A.21 proporciona los valores críticos del test tanto en el caso de la Cola 1 como la Cola 2, siendo éstos 3 y 11. Por tanto, se rechaza la hipótesis de aleatoriedad en ambos casos.*

Cuando, como ocurre en la mayoría de los casos, los datos son numéricos, la forma de determinar los elementos de ambos tipos es calcular la mediana y asignar símbolos distintos en función de que el dato en cuestión, en su orden inicial, esté por encima o por debajo de dicha mediana, eliminándose los valores que coincidan con ésta.

Ejemplo 5.3 *Supóngase la muestra*

10'0 11'0 10'8 10'3 9'6 9'8 9'3
 11'2 10'7 10'5 10'4 9'9 12 10.

La mediana vale 10'35, al asignar - a los valores que se encuentran por debajo de la mediana y + a los que están por encima, resulta

-	+	+	-	-	-	-
10'1	11	10'8	10'3	9'6	9'8	9'3
+	+	+	+	-	+	-
11'2	10'7	10'5	10'4	9'9	12	10'1

de donde $n_1 = 7$, $n_2 = 7$ y $R_{exp} = 7$. Según la ta-

bla A.21, $R_{7,7,0'025} = 3$ y $R_{7,7,0'975} = 12$, por lo que no se puede rechazar la hipótesis de aleatoriedad de la muestra para el nivel dado.

2.2. Contraste de autocorrelación

Cuando los datos vienen dados en una secuencia temporal es posible que se produzca algún tipo de relación, de forma que los anteriores influyan sobre los posteriores; este concepto se conoce como *autocorrelación*. La autocorrelación puede tener diferentes grados de diferimiento, distinguiéndose entre el de una etapa, de dos etapas, etc. En el de una etapa, cada medición influye sobre la inmediatamente siguiente, el de dos etapas sobre la que se encuentra situada dos posiciones después de ella, y así sucesivamente.

Para estudiar este problema se parte de la realización muestral x_1, x_2, \dots, x_n , y se considera la distribución bidimensional

$$(x_1, x_2), (x_2, x_3), \dots, (x_{n-1}, x_n),$$

estimándose su correlación por

$$r_1 = \frac{\sum_{i=2}^n (x_i - \bar{x})(x_{i-1} - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

En general para estudiar la autocorrelación de orden k se considera la distribución

$$(x_1, x_{k+1}), (x_2, x_{k+2}), \dots, (x_{n-k}, x_n),$$

estimándose su correlación por

$$r_k = \frac{\sum_{i=k+1}^n (x_i - \bar{x})(x_{i-k} - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

La hipótesis a contrastar en el caso general es

$$\begin{cases} H_0 : & \text{Se da la independencia lineal de orden } k \\ H_1 : & \text{Existe una correlación de orden } k \end{cases}$$

Cuando la Hipótesis Nula es cierta, se sabe que

$$r_k \sim N(0, 1/\sqrt{n}), \quad \forall k,$$

viniendo dada la región crítica por

$$|r_k| \geq \frac{Z_{1-\frac{\alpha}{2}}}{\sqrt{n}}.$$

Para contrastar, de forma global, la existencia de autocorrelación en los m primeros órdenes, Ljung y Box dieron un test, cuyo estadístico de contraste viene dado por

$$Q_{exp} = n(n+2) \sum_{k=1}^m \frac{r_k^2}{n-k},$$

que bajo la hipótesis de inexistencia de autocorrelación se distribuye según una χ_{m-1}^2 , viniendo dada, en este caso, la región crítica por

$$Q_{exp} \geq \chi_{m-1, \alpha}^2.$$

Ejemplo 5.4 *El PIB (billones de ptas.) de un cierto país en los últimos años ha sido*

Año	75	76	77	78	79	80	81	82	83	84
PIB	13	14	18	21	22	19	20	23	27	30

Haciendo los cálculos oportunos los dos primeros coeficientes de autocorrelación resultan $r_1 = 0'89$ y $r_2 = 0'53$, mientras que el valor crítico para un nivel $\alpha = 0'05$ es igual a $1'96/\sqrt{10} = 0'62$, por lo que se rechaza, para el nivel fijado, la hipótesis de independencia lineal de primer orden, pero se admite la de segundo.

En cuanto a la independencia global de los dos primeros órdenes, se tiene que

$$\begin{aligned} Q_{exp} &= 10 \cdot 12 \left(\frac{0'89^2}{9} + \frac{0'53^2}{8} \right) \\ &= 14'77, \end{aligned}$$

estando situado el valor crítico para un $\alpha = 0'05$ en $\chi_{1,0'95}^2 = 3'841$, con lo que se rechaza, para ese nivel, la hipótesis de incorrelación global.

2.3. Test de valores atípicos

En la hipótesis de normalidad se puede estudiar la presencia de valores atípicos en una muestra \underline{X} atendiendo al tamaño n de la misma. Para muestras pequeñas en las que se sospecha la presencia de un valor atípico, se recurre a la desviación máxima estudentizada,

$$d_m = \max_i \left| \frac{X_i - \bar{X}}{S_c} \right|,$$

cuyos valores críticos más importantes se dan en la siguiente tabla

n	5	6	7	8	9	10	12	15	20
$\alpha = 0'05$	1'71	1'89	2'02	2'13	2'21	2'29	2'41	2'55	2'71
$\alpha = 0'01$	1'76	1'97	2'14	2'28	2'38	2'48	2'63	2'81	3'00

Se considera que existe un valor atípico o heterogéneo respecto a los demás cuando $d_m > d_{n,\alpha}$. Teniendo en cuenta la expresión del estadístico d_m , se observa que este valor coincide con el máximo (o mínimo) de la muestra.

Para estudiar la presencia simultánea de varios valores atípicos en muestras de mayor tamaño, se recurre al coeficiente de apuntamiento, viniendo éste dado por

$$g_2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^4}{nS_c^4}.$$

Los valores críticos para algunos valores de n se dan a continuación,

n	5	10	15	20	25	50	75	100	200	500
$\alpha = 0'05$	2'9	3'9	4'1	4'1	4'0	3'99	3'87	3'77	3'57	3'37
$\alpha = 0'01$	3'1	4'8	5'1	5'2	5'0	4'88	4'59	4'39	3'98	3'6

Se admite la presencia de valores atípicos cuando el apuntamiento de la distribución sea significativamente mayor que el de la Normal. La región crítica vendría dada por

$$R_c : g_2 \geq g_{n;\alpha}.$$

Cuando n es muy grande, se recurre a

$$\frac{g_2 - 3}{\sqrt{24/n}} \sim N(0, 1).$$

Ejemplo 5.5 Para comprobar si en la muestra,

2, 3, 5, 1, 7, 2, 12, 4, 6, 8, 20, 4, 5, 7, 1

existen valores atípicos, se calcula su coeficiente de apuntamiento, resultando que $g_2 = 4'45$. Para el valor de $\alpha = 0'05$ y dado el tamaño de la muestra, $n = 15$, el valor crítico es igual a 4'1, con lo que se admite la presencia de valores atípicos.

3. Análisis de la estructura de la población

En esta sección se engloban a aquellos test que justifican el ajuste de los datos a modelos teóricos o a distribuciones conocidas, lo que permitirá la utilización de métodos paramétricos en la contrastación de parámetros. Se dedicará especial atención al caso de la Normal, para la que se ofrecen test específicos.

3.1. Test de bondad de ajuste de la Chi-cuadrado

Este test, que puede denominarse de amplio espectro por la multitud de situaciones a las que se puede aplicar, está indicado cuando los

valores muestrales puedan clasificarse en clases, lo que ocurre directamente en el caso de variables discretas, o bien para variables continuas, previa agrupación de los valores en intervalos.

Supóngase un atributo o una variable que toma valores que pueden englobarse en las clases C_1, C_2, \dots, C_k , de forma que para una muestra de tamaño n las frecuencias observadas en cada clase vienen dadas por n_1, n_2, \dots, n_k , con $\sum_{i=1}^k n_i = n$. Se desea contrastar la hipótesis de que las clases A_i ocurran con probabilidad p_{i_0} , es decir

$$\begin{cases} H_0 : \forall i & p_i = p_{i_0} \\ H_1 : \exists i & p_i \neq p_{i_0} \end{cases}$$

Para cada valor observado, n_i , se calcula su correspondiente valor esperado como $E_i = np_i$. El estadístico de contraste o experimental viene dado por

$$\chi_{exp}^2 = \sum_{i=1}^k \frac{(n_i - E_i)^2}{E_i},$$

quedando la región crítica como

$$R_c : \chi_{exp}^2 \geq \chi_{k-p-1, 1-\alpha}^2,$$

donde p es el número de parámetros que se han de estimar para el cálculo de las probabilidades.

El test tiene varias restricciones, a saber:

1. $\min_i E_i \geq 3$. Cuando esto no se verifica es necesario agrupar clases para conseguir que se cumpla la restricción.
2. Cuando $k \leq 4$ se debe aplicar al estadístico la corrección de Yates

$$\chi_{exp}^2 = \sum_{i=1}^k \frac{(|n_i - E_i| - 0.5)^2}{E_i}.$$

En los ejemplos que siguen se pueden apreciar los distintos campos de aplicación de este contraste.

1. Para ajustar una distribución Uniforme discreta.

Ejemplo 5.6 *Para contrastar que un dado es honrado lo lanzamos 60 veces, obteniendo los siguientes resultados*

x_i	1	2	3	4	5	6
n_i	7	12	10	11	8	12

La hipótesis a contrastar es que $p_i = 1/6$, $\forall i$, con lo que se tiene que $E_i = 60(1/6) = 10$, $\forall i$, resultando el estadístico experimental

$$\chi_{exp}^2 = \frac{(7-10)^2}{10} + \frac{(12-10)^2}{10} + \frac{(10-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(12-10)^2}{10} = 2'2.$$

Para un nivel $\alpha = 0'05$, puesto que no se ha estimado ningún parámetro, el valor crítico para el test es igual a $\chi_{5,0'95}^2 = 11'07$, por lo que, para dicho nivel de significación, no se rechaza que el dado es honrado.

2. Para ajustar una distribución de Poisson.

Ejemplo 5.7 *Durante la Segunda guerra mundial los alemanes bombardearon en diversas ocasiones Londres. Al objeto de analizar si los bombardeos eran indiscriminados o se hacían con intención, se procedió a dividir la ciudad en cuadrículas y a contar el número de impactos en cada una de ellas. Los resultados se recogen en la siguiente tabla*

Impactos	0	1	2	3	4	5
Número cuadrículas	229	211	93	35	7	1

Las hipótesis podrían ser expresadas, en términos probabilísticos, de la siguiente manera

$$\begin{cases} H_0 : X \sim P(\lambda) \\ H_1 : X \not\sim P(\lambda) \end{cases}$$

puesto que si las bombas caen indiscriminadamente, lo hacen de forma independiente en un soporte continuo. Lo que indica que la variable que mide el número de impactos por cuadrículas debe ser Poisson.

Para estimar el parámetro λ se utiliza su estimador óptimo $\bar{x} = 0'929$, con lo que las probabilidades quedan

$$\begin{aligned} p_0 = P[X = 0] &= \frac{e^{-0'929} 0'929^0}{0!} = 0'395 \\ p_1 = P[X = 1] &= \frac{e^{-0'929} 0'929^1}{1!} = 0'367 \\ p_2 = P[X = 2] &= n \frac{e^{-0'929} 0'929^2}{2!} = 0'1705 \\ p_3 = P[X = 3] &= \frac{e^{-0'929} 0'929^3}{3!} = 0'0528 \\ p_4 = P[X = 4] &= \frac{e^{-0'929} 0'929^4}{4!} = 0'0123 \\ p_5 = P[X \geq 5] &= 1 - \sum_{i=0}^4 \frac{e^{-0'929} 0'929^i}{i!} \\ &= 0'0024 \end{aligned}$$

Aunque no se han observado más de cinco impactos en una cuadrícula, la distribución de Poisson puede, teóricamente, tomar valores hasta el infinito. Para conseguir que los valores observados y los esperados sumen lo mismo, es decir, sean comparables, $\sum_{i=0}^5 p_i$ debe ser igual a 1, con lo que

$$\sum_{i=0}^5 E_i = \sum_{i=0}^5 np_i = n \sum_{i=0}^5 p_i = n = \sum_{i=0}^5 n_i.$$

Así pues, en vez de tomar $P[X = 5]$, se toma $P[X \geq 5]$.

Los valores esperados quedan

x_i	0	1	2	3	4	5
E_i	227'5	211	98	30	7	1'5

Al no darse la restricción de que el mínimo de los E_i sea mayor o igual a tres, se debe proceder a la agrupación, así

x_i	0	1	2	3	≥ 4
n_i	229	211	93	35	8
E_i	227'5	211	98	30	8'5

Tomando $\alpha = 0'05$, el estadístico experimental vale $\chi_{exp}^2 = 1'127$, mientras que el valor crítico, teniendo en cuenta que se ha estimado el parámetro λ , vale

$$\chi_{5-1-1,0'95}^2 = \chi_{3,0'95}^2 = 7'815,$$

por lo que, para el nivel fijado, no se rechaza la Hipótesis Nula, concluyendo así que los bombardeos alemanes sobre Londres eran indistinguibles.

3. Para ajustar una distribución Normal.

Ejemplo 5.8 Se desea contrastar que los datos siguientes proceden de una distribución Normal.

$L_{i-1} - L_i$	n_i
(0, 1]	1
(1, 2]	3
(2, 3]	7
(3, 4]	12
(4, 5]	6
(5, 6]	2
(6, 7]	1

En primer lugar se calculan los estimadores de μ y de σ , al objeto de obtener las probabilidades teóricas. Tomando marcas de clase, se obtiene que $\bar{x} = 3'41$ y $S_c = 1'28$, con lo que

$$\begin{aligned} p_1 &= P[-\infty < X < 1] \\ &= P\left[-\infty < Z < \frac{1-3'41}{1'28}\right] = F_z(-1'88) \\ &= 0'0301 \\ p_2 &= P[1 \leq X < 2] \\ &= F_z(-1'10) - F_z(-1'88) = 0'1056 \\ p_3 &= P[2 \leq X < 3] \\ &= F_z(-0'32) - F_z(-1'10) = 0'2388 \end{aligned}$$

$$\begin{aligned}
p_4 &= P[3 \leq X < 4] \\
&= F_z(0'46) - F_z(-0'32) = 0'3027 \\
p_5 &= P[4 \leq X < 5] \\
&= F_z(1'24) - F_z(0'46) = 0'2153 \\
p_6 &= P[5 \leq X < 6] \\
&= F_z(2'02) - F_z(1'24) = 0'0858 \\
p_7 &= P[6 \leq X < \infty] \\
&= 1 - F_z(2'02) = 0'0217
\end{aligned}$$

En cuanto a las E_i

$L_{i-1} - L_i$	E_i
$(-\infty, 1]$	0'96
(1, 2]	3'38
(2, 3]	7'64
(3, 4]	9'69
(4, 5]	6'89
(5, 6]	2'75
(6, ∞]	0'69

Reagrupando las clases, la situación queda

$L_{i-1} - L_i$	n_i	E_i
$(-\infty, 2]$	4	4'34
(2, 3]	7	7'64
(3, 4]	12	9'69
(4, 5]	6	6'89
(5, ∞)	3	3'44

Realizando los cálculos, el estadístico experimental vale $\chi_{exp}^2 = 0'8022$, mientras que con los dos parámetros estimados el valor crítico, para $\alpha = 0'05$, vale $\chi_{5-2-1,0'95}^2 = 5'991$, con lo que, para el nivel fijado, no se rechaza la hipótesis de que los datos proceden de una $N(3'41, 1'28)$.

3.2. Test de bondad de ajuste de Kolmogorov–Smirnov

Cuando los datos proceden de una variable continua y vienen dados de forma puntual, el contraste de bondad de ajuste más adecuado

es el de Kolmogorov–Smirnov, aunque también es posible aplicarlo a variables ordinales. Genéricamente, imagínese que se desea contrastar la hipótesis $H_0 : F(x) = F_0(x)$. Dada una muestra X_1, \dots, X_n , se define la *función de distribución empírica* para una muestra concreta como

$$F_n(x) = \begin{cases} 0 & x < x_{(1)} \\ \frac{r}{n} & x_{(r)} \leq x < x_{(r+1)} \\ 1 & x \geq x_{(n)}, \end{cases}$$

Lo que se hace es calibrar las diferencias entre la función de distribución empírica y la propuesta en H_0 a través de la expresión

$$D_{exp} = \text{máx } |F_n(x) - F_0(x)|,$$

calculando dicho máximo sobre los valores de la muestra.

Cuando la variable sea continua, para el cálculo de la máxima diferencia entre la distribución empírica y la teórica hay que situarse en la peor de las circunstancias posibles, a saber

$$D_{exp}(x_k) = \text{máx } \{|F_n(x_{k-1}) - F_0(x_k)|, |F_n(x_k) - F_0(x_k)|\}.$$

Teniendo en cuenta la expresión de la función de distribución empírica, D_{exp} puede expresarse como

$$D_{exp} = \text{máx } \{D_n^+, D_n^-\},$$

siendo

$$D_n^+ = \text{máx } \left\{ 0, \text{máx}_i \left[\frac{i}{n} - F_0(x_i) \right] \right\}$$

$$D_n^- = \text{máx } \left\{ 0, \text{máx}_i \left[F_0(x_i) - \frac{i-1}{n} \right] \right\}.$$

Los gráficos que siguen ilustran las dos posibles situaciones que se pueden encontrar.

La región crítica del test viene dada por

$$R_c : D_{exp} \geq D_{n;\alpha},$$

donde los valores $D_{n;\alpha}$ vienen dados en la tabla A.27.

Para valores mayores que 100 el valor crítico se calcula como

$$D_{n;\alpha} = \sqrt{-\log(\alpha/2)/2n}.$$

Ejemplo 5.9 Se desea contrastar la hipótesis de que los datos siguientes, generados aleatoriamente mediante ordenador, procedan de una distribución Uniforme en el intervalo $[0, 1]$ con un nivel de significación $\alpha = 0'05$.

0'582 0'501 0'497 0'026 0'132 0'561
 0'642 0'994 0'948 0'081 0'179 0'619

La función de distribución de una $U(0, 1)$ es

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{si } 0 \leq x \leq 1 \\ 1 & \text{si } x > 1. \end{cases}$$

La tabla siguiente proporciona los valores de la distribución empírica, los de la distribución $U(0, 1)$ así como las diferencias entre ambos.

x	$F_n(x)$	$F_0(x)$	$D_n(x)$
0'026	0'083	0'026	0'057
0'081	0'167	0'081	0'086
0'132	0'250	0'132	0'118
0'179	0'333	0'179	0'154
0'497	0'417	0'497	0'164
0'501	0'500	0'501	0'084
0'561	0'583	0'561	0'061
0'582	0'667	0'582	0'085
0'619	0'750	0'619	0'131
0'642	0'833	0'642	0'191
0'948	0'917	0'948	0'115
0'994	1'000	0'994	0'077

Por tanto, $D_{exp} = \max |F_n(x) - F_0(x)| = 0'191$, el valor crítico, que viene dado en la tabla A.27,

vale $D_{12,0'05} = 0'375$, con lo que para ese nivel de significación no se rechazaría la hipótesis nula.

3.3. Contrastes de normalidad

En esta sección se estudian dos test específicos de normalidad, aunque el segundo de ellos es una adaptación del test de Kolmogorov–Smirnov. El test de Shapiro–Wilk se comporta mejor cuando se tienen muestras pequeñas ($n \leq 50$), mientras que en otras situaciones se emplea el test de Kolmogorov–Smirnov–Lilliefors.

3.3.1. Test de Shapiro–Wilk

Supóngase que se desea contrastar la hipótesis de normalidad de una población de la cual se ha extraído la m.a.s. X_1, X_2, \dots, X_n . Para una realización muestral x_1, x_2, \dots, x_n , se procede de la siguiente manera:

1. Se ordenan los elementos muestrales de mayor a menor

$$x_{(n)}, x_{(n-1)}, \dots, x_{(1)}.$$

2. Se calculan las diferencias entre los valores que equidistan del centro, multiplicando cada una de dichas diferencias por los coeficientes correspondientes del test, que vienen tabulados en función del tamaño muestral. La suma de los productos se denota por b ,

$x_{(n-i+1)} - x_{(i)}$	a_{n-i+1}	$(x_{(n-i+1)} - x_{(i)})a_{n-i+1}$
$x_{(n)} - x_{(1)}$	a_n	$(x_{(n)} - x_{(1)})a_n$
$x_{(n-1)} - x_{(2)}$	a_{n-1}	$(x_{(n-1)} - x_{(2)})a_{n-1}$
\dots	\dots	\dots
$x_{(k+1)} - x_{(k)}$	a_{k+1}	$(x_{(k+1)} - x_{(k)})a_{k+1}$

$$b = \sum_{i=1}^k a_{n-i+1}(x_{(n-i+1)} - x_{(i)})$$

donde $k = \frac{n}{2}$, si n es par y $k = \frac{n-1}{2}$, si n es impar.

3. El estadístico experimental se define por

$$W_{exp} = \frac{b^2}{(n-1)S_c^2},$$

debiendo comprobarse que $W_{exp} \leq 1$, dándose la igualdad sólo en el caso de que la muestra sea la réplica de una Normal.

4. Por último, la región crítica del test viene dada por

$$R_c : W_{exp} \leq W_{n,\alpha},$$

donde el valor $W_{n,\alpha}$ viene dado en la tabla A.28

Ejemplo 5.10 Se desea contrastar la hipótesis de normalidad de una población de la cual se ha extraído la siguiente muestra:

12'3 11'0 10'7 12'4 11'7 13'1
9'9 12'6 11'8 10'2 10'5.

Ordenando y haciendo las operaciones pertinentes se tiene

$x_{(n-i+1)} - x_{(i)}$	a_{n-i+1}	$(x_{(n-i+1)} - x_{(i)})a_{n-i+1}$
13'1 - 9'9	0'5601	1'7923
12'6 - 10'2	0'3315	0'7956
12'4 - 10'5	0'2260	0'4294
12'3 - 10'7	0'1429	0'2286
11'8 - 11'0	0'0695	0'0556
		$b = 3'3015$

De donde $b^2 = 10'8999$. Por otra parte,

$$(n-1)S_c^2 = 11'4818,$$

por lo que

$$W_{exp} = \frac{10'8999}{11'4818} = 0'9493.$$

En cuanto al valor crítico $W_{11;0'05} = 0'850$, por lo que para un nivel de significación de 0'05 no se puede rechazar la hipótesis de que la muestra ha sido extraída de una población Normal.

3.3.2. Test de Kolmogorov–Smirnov–Lilliefors

El test de Kolmogorov–Smirnov–Lilliefors (K–S–L) consiste en la adaptación, por parte de Lilliefors, de la tabla de valores críticos del test de Kolmogorov–Smirnov al caso de poblaciones Normales (tabla A.26). A la hora de calcular la máxima de las diferencias entre la distribución empírica y la teórica hay que tener presente que se está tratando con una variable continua.

Ejemplo 5.11 *Aplicando el test de K–S–L a los datos del ejemplo anterior*

12'3 11'0 10'7 12'4 11'7 13'1
9'9 12'6 11'8 10'2 10'5,

para los que $\bar{x} = 11'47$ y $S_c = 1'07$.

x_i	$F_n(x_i)$	$F_0(x_i)$	$D_n(x_i)$
9'9	0'0909	0'0708	0'0708
10'2	0'1818	0'1170	0'0648
10'5	0'2727	0'1814	0'0913
10'7	0'3636	0'2358	0'1278
11'0	0'4545	0'3300	0'1245
11'7	0'5454	0'5851	0'1306
11'8	0'6363	0'6217	0'0763
12'3	0'7272	0'7810	0'1447
12'4	0'8181	0'8078	0'0806
12'6	0'9090	0'8550	0'0540
13'1	1'0000	0'9350	0'0650

De donde $D_{exp} = \max D_n(x_i) = 0'1447$. Buscando en la tabla A.26 se encuentra, para $\alpha = 0'05$, que $D_{11;0'05} = 0'249$, por lo que para el nivel elegido no se rechaza la hipótesis de que la población de la que se ha extraído la muestra es Normal.

4. Contrastes de localización y escala

En general, los contrastes no paramétricos sobre parámetros de la población se hacen sobre medidas de posición y, en particular, sobre la

mediana. Estos contrastes se basan en el análisis de la situación de los elementos de la muestra respecto a este tipo de medidas. De esta forma, se estudia si los datos muestrales están por encima o por debajo de la mediana, es decir, se analiza el signo de su diferencia con la mediana; o bien, se estudia la distancia ordenada a la que se encuentra de la mediana, es decir, se considera el rango o la posición que ocupa dicho elemento en la secuencia ordenada de las diferencias.

En el caso de que la estructura de probabilidad de la población sea simétrica estos test serán aplicables al estudio de la media de la distribución. Para empezar, se considera un test que sólo utiliza la información de los signos, para luego continuar con el estudio de los test de rangos con signos que ofrecen mejores resultados, tanto en el caso de una muestra, test de Wilcoxon, como en el de dos, test de la U de Mann–Whitney. Por último, se abordan los contrastes no paramétricos sobre medidas de dispersión, como es la varianza. En concreto, se estudia el test de Siegel–Tukey, que permite contrastar si dos muestras independientes proceden de poblaciones con el mismo parámetro de escala, esto es, si tienen la misma variabilidad.

4.1. Test de los signos

Este test puede ser aplicado con la única restricción de que la distribución sea continua. Supóngase que se desea realizar el contraste

$$\begin{cases} H_0 : M_e = M_{e_0} \\ H_1 : M_e \neq M_{e_0} \end{cases}$$

para lo que se dispone de la información que ofrece una realización muestral, x_1, x_2, \dots, x_n . Se calculan las diferencias $x_i - M_{e_0}$, descartando las nulas. Sea n' el número de diferencias no descartadas y T el número de diferencias positivas. Si H_0 es cierta entonces

$$T \sim B(n'; 0'5),$$

puesto que con igual probabilidad se encuentra un valor por encima que por debajo de la mediana. De forma que, dado un nivel de significación α y tomando k tal que

$$P[T \leq k] \leq \frac{\alpha}{2},$$

llamando T_{exp} al valor que toma T en nuestra muestra, la región de rechazo del test vendrá dada por

$$R_c : \{T_{exp} \leq k\} \cup \{T_{exp} \geq n' - k\}.$$

En el caso de dos muestras apareadas, \underline{X} e \underline{Y} , el test de los signos permite resolver el contraste $H_0 : M_e(d) = 0$, donde $\underline{d} = \underline{x} - \underline{y}$.

Ejemplo 5.12 *Se quiere probar que la mediana de una población vale 5, para lo cual, se extrae una m.a.s. de tamaño 11, resultando*

$$3, 7, 8'5, 4'3, 6, 8, 7'2, 6, 4, 5, 2$$

asignando signos se tiene

$$\begin{array}{cccccccccc} 3 & 7 & 8'5 & 4'3 & 6 & 8 & 7'2 & 6 & 4 & 5 & 2 \\ - & + & + & - & + & + & + & + & - & & - \end{array}$$

Se observa que $T_{exp} = 6$, mientras que bajo H_0 , $T \sim B(10; 0'5)$, por lo que, para un nivel $\alpha = 0'05$, hay que encontrar k tal que

$$P[T \leq k] \leq 0'025.$$

Se obtiene que

$$\begin{aligned} P[T \leq 1] &= 0'0108 \\ P[T \leq 2] &= 0'0547, \end{aligned}$$

de donde

$$R_c : \{T_{exp} \leq 1\} \cup \{T_{exp} \geq 9\}.$$

Por tanto, no se rechaza que la mediana de la población de la cual se ha extraído la muestra es igual a 5, para un nivel de significación de 0'0216.

De forma más general, el test de los signos permite contrastar el valor de cualquier cuantil de la distribución. Efectivamente, si se asigna el signo $-$ a aquellos valores de la muestra que estén por encima del valor del cuantil propuesto en H_0 y el signo $+$ a aquellos que estén por

debajo; llamando T al número de signos +, bajo H_0

$$T \sim B(n'; p_0),$$

siendo n' el número de elementos que no coinciden con el cuantil y p_0 al porcentaje de elementos que se sitúan teóricamente por debajo de dicho cuantil. Así, si se habla del percentil 70, $p_0 = 0'7$ y si se trata del primer cuantil, $p_0 = 0'25$. De forma que no se rechaza H_0 si $k_1 \leq T \leq k_2$, donde

$$P[T \leq k_1] \leq \frac{\alpha}{2} \quad \text{y} \quad P[T \geq k_2] \leq \frac{\alpha}{2}.$$

Ejemplo 5.13 En el ejemplo anterior, si se desea contrastar $H_0 : Q_3 = 8$, se hace la siguiente asignación de signos a la muestra:

$$\begin{array}{cccccccccc} 3 & 7 & 8'5 & 4'3 & 6 & 8 & 7'2 & 6 & 4 & 5 & 2 \\ + & + & - & + & + & & + & + & + & + & + \end{array}$$

Por tanto $T_{exp} = 9$. Bajo H_0 , $T \sim B(10; 0'75)$, por lo que si en esta ocasión se toma $\alpha = 0'12$ ($\frac{\alpha}{2} = 0'06$), se tiene

$$\begin{aligned} P[T \leq 4] &= 0'0197 \\ P[T \leq 5] &= 0'0781 \\ P[T \geq 9] &= 0'2377 \\ P[T \geq 10] &= 0'0500, \end{aligned}$$

con lo que

$$R_c : \{T_{exp} \leq 4\} \cup \{T_{exp} \geq 10\}.$$

Por tanto, no se rechaza que el tercer cuantil de la población de la cual se ha extraído la muestra es igual a 8, para un nivel de significación de

$$P[T \leq 4] + P[T \geq 10] = 0'0697.$$

Si las probabilidades que se obtienen utilizando la distribución Binomial se alejan mucho de $\alpha/2$, se puede utilizar la aproximación a la Normal.

Ejemplo 5.14 En el último ejemplo, todas las probabilidades están lejos de 0'06, utilizando la aproximación a la Normal se tendrá que

$$T \sim N(7'5, 1'37)$$

de forma que

$$0'06 = P[T \leq k_1] = P\left[Z \leq \frac{k_1 - 7'5}{1'37}\right]$$

y buscando en las tablas

$$\frac{k_1 - 7'5}{1'37} = -1'56 \Rightarrow k_1 = 5'3628,$$

mientras que

$$0'06 = P[T \geq k_2] = P\left[Z \geq \frac{k_2 - 7'5}{1'37}\right]$$

y buscando en las tablas

$$\frac{k_2 - 7'5}{1'37} = 1'56 \Rightarrow k_2 = 9'6372,$$

con lo que, aproximando al entero más próximo, se tendría que

$$R_c : \{T_{exp} \leq 5\} \cup \{T_{exp} \geq 10\}.$$

4.2. Test de la T de Wilcoxon para una muestra o dos muestras apareadas

Cuando la distribución además de continua es simétrica, para contrastar

$$\begin{cases} H_0 : M_e = M_{e_0} \\ H_1 : M_e \neq M_{e_0}, \end{cases}$$

se recurre al test de la T de Wilcoxon que es más potente que el de los signos.

Para aplicarlo, se procede del modo siguiente: para cada elemento de la muestra se calcula $|x_i - M_{e_0}|$, reteniéndose aquellas diferencias que sean distintas de cero, cuyo número se nota con n' . A continuación, se asignan rangos desde 1 hasta n' . Si existieran grupos de diferencias coincidentes en valor absoluto, se asigna a cada elemento la media aritmética

de los rangos que correspondan al grupo.

Una vez asignados los rangos, se suman los correspondientes a diferencias negativas, por debajo de la mediana, denotados T^- y los correspondientes a diferencias positivas, por encima de la mediana, T^+ . El estadístico experimental viene dado por

$$T_{exp} = \min\{T^-, T^+\}$$

y la región crítica por

$$R_c : T_{exp} \leq T_{n', \frac{\alpha}{2}},$$

donde $T_{n', \frac{\alpha}{2}}$ viene dado en la tabla A.31.

El test admite las siguientes variaciones y aplicaciones:

1. En las hipótesis anteriores, también se puede realizar el contraste $H_0 : \mu = \mu_0$, calculándose las diferencias en valor absoluto respecto a la media propuesta.
2. Para muestras apareadas, el contraste de mediana se plantea como $M_e(z) = 0$, donde $z_i = x_i - y_i$, para cada par de valores; para las medias el test de la T permite contrastar $H_0 : \mu_d = \mu_{d_0}$, siendo μ_d la media de la población diferencia.
3. Para muestras grandes ($n \geq 25$), la distribución asintótica de $T^-(T^+)$, bajo el supuesto de que H_0 es cierta, es

$$T^- \sim N\left(\frac{n(n+1)}{4}, \sqrt{\frac{n(n+1)(2n+1)}{24}}\right).$$

4. Para un contraste unilateral del tipo $H_0 : M_e \geq M_{e_0}$, la región crítica viene dada por

$$R_c : T^+ \leq T_{n', \alpha}.$$

5. Para un contraste unilateral del tipo $H_0 : M_e \leq M_{e_0}$, la región crítica viene dada por

$$R_c : T^- \leq T_{n', \alpha}.$$

Ejemplo 5.15 Se desea contrastar que la mediana de la población de la cual se ha extraído la siguiente m.a.s. vale 5.

4, 5, 6, 5, 3, 4, 2, 7, 6, 5, 4, 3, 8, 8, 9, 4, 6, 7, 2, 5, 6

Se calculan las diferencias respecto al valor propuesto, se ordenan y se asignan los rangos:

Rangos ⁻					4'5	4'5	4'5	4'5
Diferencias	0	0	0	0	1	1	1	1
Rangos ⁺								
Rangos ⁻					10'5	10'5		
Diferencias	1	1	1	1	2	2	2	2
Rangos ⁺	4'5	4'5	4'5	4'5			10'5	10'5
Rangos ⁻	14'5	14'5						
Diferencias	3	3	3	3	4			
Rangos ⁺					14'5	14'5	17	

Resultando que $T^- = 68$, mientras que $T^+ = 85$, con lo que $T_{exp} = 68$. Buscando en las tablas del test de Wilcoxon, aparece que $T_{17;0'05} = 30$, con lo que para el nivel de significación fijado se admite que la mediana de la población vale 5.

4.3. Test de la U de Wilcoxon–Mann–Whitney para dos muestras independientes

Como requisito para la aplicación de este test no se exige la simetría de las distribuciones aunque sí que tengan una forma similar. Sean X_1, X_2, \dots, X_{n_1} e Y_1, Y_2, \dots, Y_{n_2} m.a.s. extraídas de dos poblaciones X e Y independientes. El test de la U de Wilcoxon–Mann–Whitney permite contrastar:

$$\begin{cases} H_0 : M_{e_1} = M_{e_2} \\ H_1 : M_{e_1} \neq M_{e_2} \end{cases}$$

Para ello se procede como sigue:

1. Se ordena los $n_1 + n_2$ elementos de las dos muestras, asignando a continuación rangos desde el 1 hasta el $n_1 + n_2$.

2. Se suman los rangos correspondientes a la muestra \underline{X} , R_1 y los de la muestra \underline{Y} , R_2 , calculando a continuación:

$$U_1 = n_1 \cdot n_2 + \frac{n_1(n_1 + 1)}{2} - R_1; U_2 = n_1 \cdot n_2 + \frac{n_2(n_2 + 1)}{2} - R_2.$$

Se puede comprobar que no se ha cometido error en los cálculos viendo que se verifica que $U_1 + U_2 = n_1 \cdot n_2$.

3. Se calcula el estadístico experimental como

$$U_{exp} = \min\{U_1, U_2\}.$$

4. La región crítica del test viene dada por

$$R_c : U_{exp} \leq U_{n_1, n_2, \alpha},$$

donde $U_{n_1, n_2, \alpha}$ viene dado en la tabla A.32.

El test admite las siguientes variaciones y aplicaciones:

1. En muestras grandes ($n_1, n_2 \geq 9$), Mann y Whitney demuestran que, bajo H_0 ,

$$\frac{U_{exp} - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} \xrightarrow{d} N(0, 1).$$

Si existen muchas observaciones repetidas en la muestra conjunta habría que realizar alguna corrección sobre el estadístico de Mann–Whitney.

2. Para un contraste unilateral del tipo $H_0 : F_X(x) \geq F_Y(x)$, la región crítica viene dada por

$$R_c : U_1 \leq U_{n_1, n_2, \alpha}.$$

3. Para un contraste unilateral del tipo $H_0 : F_X(x) \leq F_Y(x)$, la región crítica viene dada por

$$R_c : U_2 \leq U_{n_1, n_2, \alpha}.$$

Ejemplo 5.16 Supóngase que se quiere contrastar, con un nivel de significación $\alpha = 0'05$, que las muestras

X	7	5	8	7	9	3	5	7	9	11	2	10
Y	3	7	8	12	15	6	5	9	11	7		

proceden de la misma población.

Se ordenan los valores y se asignan rangos

Rangos X	1	2'5	5	5	10	10	10					
Orden	2	3	3	5	5	5	6	7	7	7	7	7
Rangos Y	2'5		5		7		10					
Rangos X	13'5		16		16		18		19'5			
Orden	7	8	8	9	9	9	10	11	11	12	15	
Rangos Y	10	13'5		16		19'5		21		22		

Sumando los rangos se obtiene

$$R_1 = 126'5 \quad \text{y} \quad R_2 = 126'5.$$

De donde,

$$U_1 = 12 \cdot 10 + \frac{12(13)}{2} - 126'5 = 71'5$$

$$U_2 = 12 \cdot 10 + \frac{10(11)}{2} - 126'5 = 48'5.$$

Verificándose que $71'5 + 48'5 = 120 = 12 \cdot 10$. El estadístico del contraste vale $U_{exp} = 48'5$. Puesto que los tamaños muestrales son mayores que 9, se tiene que

$$U_{exp}/H_0 \sim N(60, 15'17),$$

con lo que

$$P[U_{exp} \geq K/H_0] = 0'025,$$

es decir,

$$P \left[Z \geq \frac{K-60}{15'17} \right] = 0'025,$$

de donde

$$\frac{K-60}{15'17} = 1'96.$$

Por tanto $K = 89'73$, lo que lleva a rechazar la hipótesis de que las muestras procedan de la misma población.

4.4. Contraste de Kolmogorov–Smirnov para dos muestras

La prueba de Kolmogorov–Smirnov para dos muestras es una alternativa al test de la U de Mann–Whitney, planteándose las hipótesis en los mismos términos, es decir, la igualdad de las funciones de distribución. La filosofía que subyace es equivalente a la del test de bondad de ajuste, valorando las diferencias entre las distribuciones empíricas obtenidas a partir de muestras independientes, de tamaños n_1 y n_2 , extraídas de dos poblaciones X e Y , respectivamente.

Para aplicar el test hay que obtener, a partir de los datos muestrales, sendas distribuciones, $F_{n_1}(x)$ para la muestra \underline{X} y $G_{n_2}(y)$ para la \underline{Y} . El estadístico experimental viene dado por

$$D_{exp} = \text{máx}\{\text{máx}|F_{n_1}(x_i) - G_{n_2}(x_i)|, \text{máx}|F_{n_1}(y_j) - G_{n_2}(y_j)|\}$$

calculando dicho máximo sobre los valores de la muestra, siendo la región crítica

$$R_c : D_{exp} \geq D_{n_1, n_2, \alpha},$$

donde $D_{n_1, n_2, \alpha}$ viene dado en la tabla A.27.

4.5. Test de Siegel–Tukey

Sean \underline{X} e \underline{Y} dos m.a.s. de tamaños n_1 y n_2 , respectivamente. El test de Siegel–Tukey permite contrastar si ambas muestras proceden de poblaciones con la misma variabilidad o dispersión.

Para ello, se considera la muestra conjunta de ambas, de tamaño $n_1 + n_2 = n$ y se ordena de menor a mayor. A continuación, se asignan los rangos según el siguiente criterio:

1. Se asigna el rango 1 al menor valor, $x_{(1)}$.

2. Al mayor valor y al anterior a él, es decir, $x_{(n)}$ y $x_{(n-1)}$, se les asignan los rangos 2 y 3 respectivamente.
3. A continuación se toma el par formado por los valores $x_{(2)}$ y $x_{(3)}$ y se les asignan los rangos 4 y 5, respectivamente.
4. El proceso se repite tomando el siguiente par de valores $x_{(n-3)}$ y $x_{(n-2)}$.

La siguiente tabla refleja el proceso a seguir:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	\cdots	$x_{(n-3)}$	$x_{(n-2)}$	$x_{(n-1)}$	$x_{(n)}$
1	4	5	8	9	\cdots	7	6	3	2

Se puede observar que los mayores rangos se encuentran situados en el centro de la tabla. Cuando n es par, una de las observaciones centrales quedará sin su rango correspondiente.

El estadístico del contraste viene determinado por los rangos de la muestra de menor tamaño. Así, si se considera $n_m = \min(n_1, n_2)$, el estadístico viene dado por:

$$R_m = \sum_{i=1}^{n_m} R_i.$$

La distribución de probabilidad de este estadístico es la misma que la del estadístico U_{exp} de Wilcoxon, por lo que se puede recurrir a la tabla de valores críticos del mismo. Así, la región crítica viene dada por

$$R_c : R_m \leq U_{n_1, n_2, \alpha},$$

donde $U_{n_1, n_2, \alpha}$ viene dado en la tabla A.32. Cuando $n > 20$, la distribución asintótica de R_m , al igual que ocurría con U_{exp} , es

$$\frac{R_m - \frac{n_m(n+1)}{2}}{\sqrt{\frac{n_1 n_2 (n+1)}{12}}} \xrightarrow{d} N(0, 1).$$

Muestra Conjunta	Rangos R_i	Muestra Origen
1'8	1	x
2	4	x
5'1	5	x
10'5	8	y
11'9	9	x
15'4	12	x
19'5	13	y
23'5	16	y
24'3	17	x
24'9	20	y
28	21	x
30	19	y
30'2	18	x
34'7	15	x
40'1	14	x
40'2	11	y
41'3	10	y
45'5	7	y
51'1	6	y
54'8	3	y
60'8	2	y

Tabla 5.1: Tabla de datos

Ejemplo 5.17 Se quiere contrastar si las dos m.a.s. siguientes proceden de poblaciones con la misma variabilidad.

x	2'0	1'8	5'1	11'9	15'4	24'3	28'0	30'2	34'7	40'1	
y	10'5	23'5	30'0	54'8	19'5	24'9	40'2	41'3	51'1	45'5	60'8

La tabla 5.1 recoge las asignaciones de rangos de la muestra conjunta.

Como la muestra de menor tamaño es \underline{X} , se tiene que $n_m = \min(n_1, n_2) = 10 = n_1$. Sumando los rangos pertenecientes a los valores de la muestra \underline{X} se obtiene el estadístico

$$R_m = \sum_{i=1}^{n_m} R_i = 116$$

La distribución asintótica del estadístico es

$$N\left(\frac{10(21+1)}{2}, \sqrt{\frac{10 \cdot 11(21+1)}{12}}\right)$$

Por tanto,

$$R_m/H_0 \sim N(110, 14'2)$$

con lo que, para un nivel de confianza del 95 %,

$$\begin{aligned} P[R_m \geq K_2/H_0] &= 0'025 \\ P[R_m < K_1/H_0] &= 0'025, \end{aligned}$$

es decir,

$$\begin{aligned} P\left[Z \geq \frac{K_2-110}{14'2}\right] &= 0'025 \\ P\left[Z < \frac{K_1-110}{14'2}\right] &= 0'025, \end{aligned}$$

de donde

$$\frac{K_2 - 110}{14'2} = 1'96$$

$$\frac{K_1 - 110}{14'2} = -1'96.$$

Por tanto,

$$K_1 = 82'168 \quad y \quad K_2 = 137'832.$$

Como

$$K_1 = 82'168 < R_m < K_2 = 137'832,$$

no se puede rechazar la hipótesis de que las muestras procedan de poblaciones con la misma variabilidad.

5. Tablas de contrastes no paramétricos

Hipótesis	Estadístico	R. Crítica
Test de rachas		
$\begin{cases} H_0 : & \text{Aleatoriedad} \\ H_1 : & \text{No aleatoriedad} \end{cases}$	$R_e = \text{número de rachas}$ Si n_1 y n_2 grandes se usa:	$R_e \leq R_{n_1, n_2, \frac{\alpha}{2}}$ U $R_e \geq R_{n_1, n_2, 1 - \frac{\alpha}{2}}$
$\begin{cases} H_0 : & \text{Aleatoriedad} \\ H_1 : & \text{Agrupamiento} \end{cases}$	$Z = \frac{R - \left(\frac{2n_1 n_2}{n_1 + n_2} + 1\right)}{\sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}} \sim \mathcal{N}(0, 1)$	$R_e \leq R_{n_1, n_2, \alpha}$
$\begin{cases} H_0 : & \text{Aleatoriedad} \\ H_1 : & \text{Fluctuación} \end{cases}$		$R_e \geq R_{n_1, n_2, 1 - \alpha}$
Contraste de autocorrelación		
$\begin{cases} H_0 : & \text{Incorrelada} \\ H_1 : & \text{No incorrelada} \end{cases}$	$r_k = \frac{\sum_{i=k+1}^n (x_i - \bar{x})(x_{i-k} - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$	$ r_k \geq \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}$
	$Q_e = n(n+2) \sum_{k=1}^m \frac{r_k^2}{n-k}$	$Q_e \geq \chi_{m-1, \alpha}^2$
Test χ^2 para una muestra		
$\begin{cases} H_0 : & F_X \equiv F_0 \\ H_1 : & F_X \neq F_0 \end{cases}$	$\chi^2 = \sum_{i=1}^k \frac{(n_i - E_i)^2}{E_i}$	$\chi_e^2 \geq \chi_{k-p-1, 1-\alpha}^2$
Kolmogorov-Smirnov para una muestra		
$\begin{cases} H_0 : & F_X \equiv F_0 \\ H_1 : & F_X \neq F_0 \end{cases}$	$D_n^+ = \max \left\{ 0, \max_{i=1}^n \left[\frac{i}{n} - F_0(x_{(i)}) \right] \right\}$	$D_e \geq D_{n, \alpha}$
$\begin{cases} H_0 : & F_X \equiv F_0 \\ H_1 : & F_X > F_0 \end{cases}$	$D_n^- = \max \left\{ 0, \max_{i=1}^n \left[F_0(x_{(i)}) - \frac{i-1}{n} \right] \right\}$	$D_n^+ \geq D_{n, 2\alpha}$
$\begin{cases} H_0 : & F_X \equiv F_0 \\ H_1 : & F_X < F_0 \end{cases}$	$D_e = \max \{ D_n^+, D_n^- \}$	$D_n^- \geq D_{n, 2\alpha}$
Shapiro-Wilk		
$\begin{cases} H_0 : & X \sim \text{Normal} \\ H_1 : & X \not\sim \text{Normal} \end{cases}$	$W = \frac{b^2}{(n-1)S_c^2}$ $b = \sum_{i=1}^{n/2} a_{n-i+1} (X_{(n-i+1)} - X_{(i)})$	$W_e \leq W_{n, \alpha}$

Tabla 5.2: Contrastes no paramétricos I

Hipótesis	Estadístico	R. Crítica
Test de los signos		
$\begin{cases} H_0 : C = C_0 \\ H_1 : C \neq C_0 \end{cases}$	$T = \text{N}^\circ \text{ de valores inferiores a } C_0$ Bajo H_0 , $T \sim B(n', p_0)$ $P[T \leq k_1] \leq \frac{\alpha}{2} \quad P[T \geq k_2] \leq \frac{\alpha}{2}$	$T_e \leq k_1$ \cup $T_e \geq k_2$
$\begin{cases} H_0 : C = C_0 \\ H_1 : C < C_0 \end{cases}$	$P[T \leq k_1] \leq \alpha$	$T_e \leq k_1$
$\begin{cases} H_0 : C = C_0 \\ H_1 : C > C_0 \end{cases}$	$P[T \geq k_2] \leq \alpha$	$T_e \geq k_2$
Test de Rangos-Signos (Wilcoxon para una muestra)		
$\begin{cases} H_0 : M_e = M_{e_0} \\ H_1 : M_e \neq M_{e_0} \end{cases}$	$T_e = \min\{T^+, T^-\}$	$T_e \leq T_{n', \frac{\alpha}{2}}$
$\begin{cases} H_0 : M_e = M_{e_0} \\ H_1 : M_e < M_{e_0} \end{cases}$	Si n grande se usa:	$T^+ \leq T_{n', \alpha}$
$\begin{cases} H_0 : M_e = M_{e_0} \\ H_1 : M_e > M_{e_0} \end{cases}$	$Z = \frac{4T^+ - n(n+1)}{\sqrt{2n(n+1)(2n+1)/3}} \sim \mathcal{N}(0, 1)$	$T^- \leq T_{n', \alpha}$

Tabla 5.3: Contrastes no paramétricos II

Hipótesis	Estadístico	R. Crítica
Test U. Mann-Whitney		
$\begin{cases} H_0 : F_X \equiv F_Y \\ H_1 : F_X \neq F_Y \end{cases}$	$U_1 = n_1 \cdot n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$ $U_2 = n_1 \cdot n_2 + \frac{n_2(n_2 + 1)}{2} - R_2$	$U_e \leq U_{n_1, n_2, \alpha}$
$\begin{cases} H_0 : F_X \equiv F_Y \\ H_1 : F_X < F_Y \end{cases}$	$U_e = \min\{U_1, U_2\}$ Si $n_1, n_2 \geq 9$ se usa:	$U_1 \leq U_{n_1, n_2, \alpha}$
$\begin{cases} H_0 : F_X \equiv F_Y \\ H_1 : F_X > F_Y \end{cases}$	$U_e \sim N\left(\frac{n_1 \cdot n_2}{2}, \sqrt{\frac{n_1 \cdot n_2 (n_1 + n_2 + 1)}{12}}\right)$	$U_2 \leq U_{n_1, n_2, \alpha}$
Kolmogorov-Smirnov para dos muestras		
$\begin{cases} H_0 : F_X \equiv F_Y \\ H_1 : F_X \neq F_Y \end{cases}$	$D_e = \max_i \left \frac{F_{X_i}}{n_x} - \frac{F_{Y_i}}{n_y} \right $	$D_e \geq D_{n_1, n_2, \alpha}$
Test de los signos (2 muestras relacionadas)		
$\begin{cases} H_0 : M_e(d) = 0 \\ H_1 : M_e(d) \neq 0 \end{cases}$	$\underline{d} = \underline{x} - \underline{y}$ Se aplica el test de los signos.	
Test de Wilcoxon (2 muestras relacionadas)		
$\begin{cases} H_0 : M_e(d) = 0 \\ H_1 : M_e(d) \neq 0 \end{cases}$	$\underline{d} = \underline{x} - \underline{y}$ Se aplica el test de Wilcoxon.	

Tabla 5.4: Contrastes no paramétricos III

6. Ejercicios**6.1. Ejercicio resuelto**

5.1 Una conocida empresa está considerando su fusión con una de las dos cadenas de televisión más importante del país, la cadena A y la cadena B, por lo que decide realizar un estudio sobre el nivel de audiencia en cada una de ellas. Elige aleatoriamente distintas horas de la semana y mide el nivel de audiencia, en millones de personas, obteniendo las muestras:

Cadena A	6'4	8'9	9	2'7	4'5	10	9	4'9	3	7	15
Cadena B	8	5'9	10	15	17'5	9	3	3'2	6	8	16

a) Utilizando la información procedente de la función de distribución empírica, ¿existen diferencias significativas, a un nivel de confianza del 95 %, en las distribuciones de los niveles de audiencia de las dos cadenas?

b) Si se admite que ambas muestras proceden de poblaciones Normales, ¿se puede afirmar que la cadena A presenta más dispersión en el nivel de audiencia que la B, al mismo nivel de confianza que en el caso anterior?

Solución:

a) Sean X e Y las variables aleatorias que representan los niveles de audiencia de las cadenas A y B, respectivamente. Sean F y G sus correspondientes funciones de distribución. Para contrastar si F y G son iguales utilizando las funciones de distribución empíricas, se recurre al test de Kolmogorov–Smirnov para dos muestras.

Para el contraste:

$$\begin{cases} H_0 : F(x) = G(x) \\ H_1 : F(x) \neq G(x) \end{cases}$$

se utiliza el estadístico

$$D_{exp} = \max |F_n(x) - G_n(y)|,$$

La siguiente tabla recoge los valores de las funciones de distribución empíricas de ambas muestras:

x_i	n_i	$F_n(x_i)$	y_i	n_i	$G_n(y_i)$
2'7	1	1/11	3	1	1/11
3	1	2/11	3'2	1	2/11
4'5	1	3/11	5'9	1	3/11
4'9	1	4/11	6	1	4/11
6'4	1	5/11	8	2	6/11
7	1	6/11	9	1	7/11
8'9	1	7/11	10	1	8/11
9	2	9/11	15	1	9/11
10	1	10/11	16	1	10/11
15	1	1	17'5	1	1

Por tanto, si se consideran las diferencias de las distribuciones empíricas de ambas muestras y se calcula el máximo en los valores muestrales, se obtiene que

$$D_{exp} = \max |F_n(x) - G_n(y)| = \frac{2}{11} = 0'1818.$$

Puesto que $D_{n_1, n_2, \alpha} = D_{11, 11, 0'05} = 6/11 = 0'5455 > 0'1818 = D_{exp}$, no puede rechazarse la hipótesis de igualdad de distribuciones.

b) Suponiendo que ambas distribuciones son Normales, es decir, $X \sim N(\mu_1, \sigma_1)$ e $Y \sim N(\mu_2, \sigma_2)$, el contraste que se plantea ahora para averiguar si la cadena A tiene más dispersión es el siguiente:

$$\begin{cases} H_0 : \sigma_1^2 \geq \sigma_2^2 \\ H_1 : \sigma_1^2 < \sigma_2^2 \end{cases}$$

donde las medias μ_1 y μ_2 son desconocidas. Para ello, se recurre al estadístico

$$F = \frac{S_{c_1}^2}{S_{c_2}^2} \sim F_{n_1-1, n_2-1}.$$

Calculando las cuasivarianzas muestrales, se obtiene

$$\begin{aligned} S_{c_1}^2 &= \frac{1}{n_X - 1} \left[\sum_{i=1}^{n_X} x_i^2 - \frac{1}{n_X} \left(\sum_{i=1}^{n_X} x_i \right)^2 \right] \\ &= \frac{1}{10} \left[716'72 - \frac{80'4^2}{11} \right] = 12'9069 \end{aligned}$$

$$S_{c_2}^2 = \frac{1}{10} \left[1186'3 - \frac{101'6^2}{11} \right] = 24'7885$$

y por tanto, $F_{exp} = \frac{12'9069}{24'7885} = 0'5207$.

Para el nivel de significación 0'05, se tiene que

$$F_{n_1-1, n_2-1} = F_{10,10} = 2'98,$$

que es mayor que el F_{exp} . Por tanto, no se puede rechazar la hipótesis de que la dispersión de la cadena A sea mayor que la de la cadena B.

6.2. Ejercicios propuestos

5.1. En los últimos 200 sorteos de la lotería nacional, la última cifra del número premiado fue

Cifra	0	1	2	3	4	5	6	7	8	9
Frecuencia	12	24	18	14	28	24	26	16	8	30

Determine si se ajustan estas frecuencias empíricas a la distribución teórica que les correspondería.

5.2. Una determinada empresa está interesada en construir una gasolinera en sus terrenos, sin embargo, la realización del proyecto depende de la distribución de llegadas de vehículos. Para conocer dicha distribución se consideran 900 períodos de tiempo de 5 minutos cada uno en una gasolinera de una zona con características parecidas, contándose a su vez el número de coches que llegan. El resultado de tales averiguaciones fue el siguiente:

Valores variable	0	1	2	3	4	5	6	7	8	≥ 9
Frecuencia	195	301	210	120	45	22	5	1	1	0

Proponga la distribución que siguen los datos anteriores y estudie si se ajusta a ellos.

5.3. La siguiente tabla recoge el número de accidentes mortales, por día, contabilizados en una ciudad durante 200 días:

Accidente/día	0	1	2	3	4	5
Frecuencia	50	72	44	22	10	2

Proponga un modelo de distribución y estudie su ajuste.

5.4. Una conocida empresa de coches está realizando un estudio sobre la vida de sus motores. Para ello, se decidió llevar a cabo un estudio sobre 70 motores, observando que la vida media era de 1'6 años. Agrupando los valores observados por clases, se obtuvo lo siguiente:

Años funcionamiento	[0, 1)	[1, 2)	[2, 3)	[3, 4)	[4, ∞)
Frecuencia	30	23	6	5	6

Estudie si la vida de los motores sigue una distribución exponencial.

5.5. A partir de la siguiente realización muestral de tamaño 500,

$(-\infty, 0)$	$(0, 3)$	$(3, 5)$	$(5, 7)$	$(7, 9)$	$(9, 11)$
36	70	54	86	44	50
$(11, 13)$	$(13, 15)$	$(15, 18)$	$(18, 21)$	$(21, \infty)$	
40	24	36	24	36	

donde la primera fila corresponde a los valores de la variable y la segunda al número de valores de la muestra dentro de cada intervalo.

a) Verifique la hipótesis de que ha sido extraída de una población $N(10, 5)$ a un nivel de confianza del 90%.

b) ¿Es posible que los valores muestrales se hubieran extraído de una Normal con parámetros distintos a los anteriores?

c) Proponga valores para μ y σ , y repita el contraste.

5.6. Estudie, utilizando el contraste χ^2 de bondad de ajuste, si la siguiente muestra de tamaño 30 procede de una Normal:

107	96	91	80	103	88	101	106	112	106
93	88	101	109	102	99	93	86	100	99
104	116	87	93	106	102	89	96	104	90

5.7. Con el fin de estudiar el tiempo de vida, en horas, de las baterías de 7 voltios, se extrae aleatoriamente un muestra de 10 de ellas, obteniéndose los siguientes resultados:

28'9, 15'2, 28'7, 72'5, 48'6, 52'4, 37'6, 49'5, 62'1, 54'5.

Proponga un modelo de distribución de probabilidad y estudie su ajuste.

5.8. A partir de los datos del ejercicio anterior contraste la hipótesis de que éstos procedan de una población Normal, utilizando para ello el contraste de Shapiro–Wilks.

5.9. Utilizando el contraste de Kolmogorov–Smirnov–Lilliefors, contraste la hipótesis de normalidad de los siguientes datos:

27, 23, 17, 18, 17, 30, 22, 26, 16, 23, 20, 22, 16, 21, 17.

5.10. A partir de los datos muestrales del ejercicio anterior y suponiendo que éstos hayan sido extraídos en ese orden, contrasta la hipótesis de que la muestra sea aleatoria.

5.11. Contraste la hipótesis de que la siguiente m.a.s. proceda de una distribución $B(9, 0'3)$, para un nivel de significación del 10%.

Valores variable	0	1	2	3	4	5	6	7	8	9
Frecuencias	84	196	210	136	50	14	6	2	2	0

5.12. A partir de los siguientes datos de distribución de frecuencias, contraste la hipótesis de que tal muestra proceda de una distribución Normal, para un nivel de significación de 0'05.

$(-1, 0)$	$(0, 0'5)$	$(0'5, 1)$	$(1, 1'5)$	$(1'5, 2)$	$(2, 2'5)$	$(2'5, 2'8)$
554	228	844	584	366	256	168

5.13. Se pretende realizar un estudio sobre la influencia de un determinado fármaco en el tratamiento de la celulitis para lo cual se estudió el peso de cada paciente antes y después del tratamiento. A partir de una muestra de 32 pacientes los resultados fueron los siguientes:

Antes	73	99	75	84	102	84	65	70	78	75	78
Después	69	93	78	85	99	80	67	72	73	71	78
Antes	82	64	72	71	64	85	94	89	57	59	67
Después	80	61	74	76	63	85	94	88	54	56	65
Antes	96	97	73	58	57	63	81	84	80	67	
Después	95	99	75	56	57	62	80	85	78	68	

Decida, utilizando el test de los signos, si ha habido una alteración significativa en el peso.

5.14. Para medir la introversión se aplica a 12 individuos un test de personalidad en sus dos variantes, 1 y 2, que se supone la miden por igual. A partir de los datos de la siguiente tabla, compruebe mediante el test de rangos de Wilcoxon, con un nivel de significación del 5%, si es cierto que las formas 1 y 2 miden por igual la introversión:

Individuo	1	2	3	4	5	6	7	8	9	10	11	12
Forma 1	12	18	21	10	15	27	31	6	15	13	8	10
Forma 2	10	17	20	5	21	24	29	7	9	13	8	11

5.15. En una clase de matemáticas el profesor plantea a los alumnos dos problemas, P y Q, asignándole a cada uno un problema al azar. Con el fin de estudiar si existían diferencias de dificultad entre ambos

problemas, el profesor fue anotando el tipo de problema a medida que iban entregando, obteniendo la siguiente secuencia:

QQQPPQPQQQPQPPQPPP.

Estudie si existían diferencias de dificultad entre ambos problemas, con una probabilidad de error de tipo I menor del 5%.

5.16. Para estudiar cuál de los dos tratamientos contra la artrosis es más eficaz se eligen aleatoriamente dos muestras de 10 y 22 pacientes a los cuales se les somete a los tratamientos 1 y 2, respectivamente. Pasados tres meses se valoran ambos tratamientos de manera que el tenga mayor puntuación será más eficaz. La tabla siguiente refleja los resultados obtenidos:

Tratamiento 1	12	15	21	17	38	42	10	23	35	28	
Tratamiento 2	21	18	42	25	14	52	65	40	43	35	18
	56	29	32	44	15	68	41	37	43	58	42

Aplicando el test U de Mann–Whitney, decida si existe diferencia entre los tratamientos.

5.17. Tras un experimento donde se realizó un cruce entre gatos, se obtuvo que de 64 descendientes, 10 eran negros, 34 grises y 20 blancos. Según la genética, la proporción de estos números debería ser 3:9:4. Decida si los datos obtenidos en el experimento están de acuerdo con la genética, para un nivel de significación del 5%.