Práctica 7. Contrastes no paramétricos en dos o más poblaciones

1. Contraste de homogeneidad con dos o más muestras independientes (Kruskal-Wallis)

El procedimiento que vamos a explicar se aplica cuando la variable es cualitativa ordinal o cuantitativa, pero no Normal, y los tamaños muestrales son pequeños (en la práctica, alguno de ellos menor que 30).

Observamos una variable aleatoria cuantitativa o cualitativa ordinal en r poblaciones, y extraemos r muestras aleatorias independientes (una de cada población). El objetivo es contrastar la hipótesis nula H_0: *"Las r poblaciones son homogéneas"* (la variable aleatoria observada tiene la misma distribución en las r poblaciones) frente a la hipótesis alternativa H_1: *"Las r poblaciones no son homogéneas"*. La hipótesis nula implica que las r medias poblacionales son iguales por lo que, a veces, se sustituye aquella hipótesis nula por ésta.

En ***Minitab*** hay varios procedimientos para realizar un contraste de homogeneidad con dos o más muestras independientes, pero vamos a explicar el contraste de Kruskal-Wallis, que es una generalización del test de Mann-Whitney (que también se puede usar en ***Minitab***) para dos muestras aleatorias independientes.

Para realizar el contraste de Kruskal-Wallis hay que seleccionar **Stat** ⇒**Nonparametrics** ⇒**Kruskal-Wallis**.

Con el archivo de datos **Pulse.mtw**, veamos si se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que el nivel de actividad física de los hombres es igual al de las mujeres. Como la variable **Activity** no es Normal y no tiene sentido comparar las medias poblacionales, tenemos que realizar un contraste no paramétrico de homogeneidad con dos muestras independientes (la muestra de hombres y la muestra de mujeres). La hipótesis nula se puede enunciar como H_0 :*"El nivel de actividad física es el mismo para los hombres y para las mujeres"* o como H_0 :*"La distribución de la variable **Activity** es la misma en la población de los hombres y en la de las mujeres"*.

Para hacer este contraste seleccionamos **Stat** ⇒**Nonparametrics** ⇒**Kruskal-Wallis**. En **Response** seleccionamos, de la lista de variables de la izquierda, la columna **'Activity'**; y en **Factor** seleccionamos, de la lista de la izquierda, la columna **'Sex'**. Podemos comprobar, en la ventana de sesión, que el p-valor es $0'305$; claramente mayor que el nivel de significación, $\alpha = 0'05$, por lo que podemos aceptar la hipótesis nula; es decir, podemos aceptar que el nivel de actividad física es el mismo para los hombres y para las mujeres.

Un ejemplo con más de dos muestras independientes podría ser el siguiente: Con el archivo de datos **Pulse.mtw**, comprobemos si se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que el peso medio es el mismo para los 4 niveles de actividad física (**Activity**=0, 1, 2 y 3). Se puede comprobar que la variable **Weight** no es Normal. Además los tamaños muestrales no son grandes: $n_1 = 1$ para **Activity**=0, $n_2 = 9$ para **Activity**=1, $n_3 = 61$ para **Activity**=2, y $n_4 = 21$ para **Activity**=3. Por tanto, no podemos realizar un contraste paramétrico en el que se comparen las medias poblacionales. Tenemos que realizar un contraste no paramétrico de homogeneidad con 4 muestras independientes. La hipótesis nula es H_0 :*"La distribución de la variable* ***Weight*** *es la misma para los cuatro niveles de actividad física"*.

Para hacer este contraste seleccionamos **Stat** $\Rightarrow$**Nonparametrics** $\Rightarrow$**Kruskal-Wallis**. En **Response** seleccionamos, de la lista de variables de la izquierda, la columna '**Weight**'; y en **Factor** seleccionamos, de la lista de la izquierda, la columna '**Activity**'. Podemos comprobar, en la ventana de sesión, que el p-valor es $0'741$; claramente mayor que el nivel de significación, $\alpha = 0'05$, por lo que podemos aceptar la hipótesis nula; es decir, podemos aceptar que la distribución de la variable **Weight** es la misma para los cuatro niveles de actividad física.

Como hemos podido observar, para realizar este contraste con ***Minitab*** debemos tener una columna con todos los resultados de la variable (para todos y cada uno de los individuos de todas y cada una de las muestras) y otra columna que nos indique la muestra de la que procede cada resultado.

2. Contraste de homogeneidad con dos o más muestras apareadas (Friedman)

El contraste de Friedman es similar al de Kruskal-Wallis pero en este caso las r muestras son apareadas (están relacionadas o asociadas). El test de Friedman se aplica cuando la variable es cualitativa ordinal o cuantitativa, pero no Normal, y los tamaños muestrales son pequeños (en la práctica, alguno de ellos menor que 30)

Para realizar este contraste hay que seleccionar **Stat** $\Rightarrow$**Nonparametrics** $\Rightarrow$**Friedman**.

En general, el problema suele ser el siguiente. Supongamos que estamos interesados en comparar los efectos de r tratamientos. Se cree que hay una variable que puede interferir en nuestra capacidad para detectar diferencias reales entre los r tratamientos. Queremos controlar esta variable extraña mediante la construcción de bloques. Esto es, dividimos los individuos en n bloques, cada uno de tamaño r, siendo los individuos de un mismo bloque tan iguales como sea posible respecto de la variable extraña. Asignaremos aleatoriamente los r tratamientos a los individuos de los bloques.

Para explicar este método vamos a utilizar un ejemplo: En la Figura 1 aparece la opinión de tres expertos respecto de la calidad de 10 libros elegidos al azar (1=muy mala, 2=mala, 3=regular, 4=buena, 5=muy buena).

Vamos a comprobar si se puede aceptar, con un nivel de significación de $0'05$, que no hay diferencia significativa entre los tres expertos respecto de su opinión sobre la calidad de los libros. Notemos que, efectivamente, las tres muestras están relacionadas, pues realmente son la misma muestra (a la cual se le ha observado tres variables distintas). La hipótesis nula es, por tanto, H_0: *"No hay diferencia significativa entre los tres expertos respecto de su opinión sobre la calidad de los libros"*.

En este ejemplo tenemos 3 tratamientos (la opinión de cada uno de los 3 expertos) y 10 bloques (cada uno de los 10 libros elegidos al azar). La variable respuesta es la opinión (de 1 a 5) de cada

opinión1	opinión2	opinión3
2	1	1
5	4	5
4	5	5
2	3	4
3	3	3
1	5	3
3	3	1
1	3	2
4	2	1
2	5	3

Figura 1

experto respecto de la calidad de cada libro.

De manera análoga a lo que ocurría con el test de Kruskal-Wallis, para realizar el contraste de Friedman con ***Minitab*** debemos tener una columna con todos los resultados de la variable respuesta (para cada individuo de cada muestra); otra columna que nos indique la muestra de la que procede cada resultado (tratamiento) y otra columna que nos indique el individuo (bloque). Por tanto, para poder aplicar el contraste de Friedman no se pueden tener los datos tal y como se muestran en la Figura 1, sino que hay que tener una columna que indique el número del tratamiento (en este caso, el número del experto: de 1 a 3); otra columna que indique el número del bloque (en este caso, el número del libro: de 1 a 10) y otra columna que indique la variable respuesta (de 1 a 5) para cada combinación de resultados de las dos columnas anteriores. Los datos, por tanto, tienen que introducirse tal y como se muestra en la Figura 2. Podemos grabar estos datos en un archivo que denominaremos **Ejemplo_Friedman.mtw**.

Para realizar el contraste seleccionamos **Stat** ⇒**Nonparametrics** ⇒**Friedman**. En **Response** seleccionamos, de la lista de la izquierda, la columna **'opinión (de 1 a 5)'**; en **Treatment** seleccionamos, de la lista de la izquierda, la columna **'nº experto'**; en **Blocks** seleccionamos, de la lista de la izquierda, la columna **'nº libro'** y pulsamos en **OK**. Podemos observar, en la ventana de sesión, que el p-valor es $0'592$, mayor que el nivel de significación ($\alpha = 0'05$); por tanto, aceptamos la hipótesis nula; es decir, aceptamos que no hay diferencia significativa entre los tres expertos respecto de su opinión sobre la calidad de los libro.

3. Contraste chi-cuadrado sobre homogeneidad de dos o más poblaciones

En dos o más poblaciones distintas observamos una misma variable aleatoria, y extraemos una muestra aleatoria simple de cada población para comprobar si un determinado parámetro poblacional (μ, σ^2, ...) toma idéntico valor en las distintas poblaciones. Pero como no se cumplen las condiciones necesarias para aplicar un contraste de hipótesis paramétrico, entonces tenemos que realizar un contraste de hipótesis no paramétrico. Sin embargo, ocurre que la hipótesis nula no se puede enunciar como la igualdad de los parámetros poblacionales, sino que ahora debemos comprobar si la variable aleatoria tiene la misma distribución en las dos poblaciones. Esta hipótesis se resume diciendo que las poblaciones son homogéneas.

nº experto	nº libro	opinión (de 1 a 5)
1	1	2
1	2	5
1	3	4
1	4	2
1	5	3
1	6	1
1	7	3
1	8	1
1	9	4
1	10	2
2	1	1
2	2	4
2	3	5
2	4	3
2	5	3
2	6	5
2	7	3
2	8	3
2	9	2
2	10	5
3	1	1
3	2	5
3	3	5
3	4	4
3	5	3
3	6	3
3	7	1
3	8	2
3	9	1
3	10	3

Figura 2

El contraste chi-cuadrado de homogeneidad es el mismo que el test chi-cuadrado de independencia de variables explicado en la sección 3 de la **Práctica 4**, aunque la hipótesis nula no sea la misma.

Para realizar este tipo de contraste en ***Minitab*** se utilizan las mismas dos opciones explicadas en la sección 3 de la **Práctica 4**; es decir, si los datos están recogidos en una tabla de doble entrada, se utiliza **Stat⇒Tables⇒Chi-Square Test (Two-Way Table in Worksheet)**, y si los datos se encuentran recogidos en dos (o tres) columnas, se utiliza **Stat⇒Tables⇒Cross Tabulation and Chi-Square**.

Vamos a hacer el siguiente ejemplo: Se selecciona una muestra aleatoria simple de estudiantes de informática de universidades privadas y otra de universidades públicas, y se les somete a una prueba de rendimiento, calificada de 0 a 500. Los resultados son los expuestos en la tabla siguiente. Deseamos saber si la distribución en la prueba de rendimiento es la misma para universidades privadas que para universidades públicas.

	[0,275]	[276,350]	[351,425]	[426,500]
privadas	6	14	17	9
públicas	30	32	17	3

El objetivo es contrastar la hipótesis H_0: *"La distribución de los resultados de la prueba es la misma en las universidades públicas que en las privadas"*, frente a la hipótesis H_1: *"La distribución no es la misma"*.

Para realizar este contraste de homogeneidad con ***Minitab***, en primer lugar tenemos que introducir la tabla de doble entrada anterior. Los datos tienen que ser introducidos tal como se muestra a continuación:

privadas	públicas
6	30
14	32
17	17
9	3

Podemos guardar estos datos en un archivo denominado **Ejemplo_Homogeneidad.mtw**.

Ahora seleccionamos **Stat⇒Tables⇒Chi-Square Test (Two-Way Table in Worksheet)**; en **Columns containing the table** elegimos, de la lista de variables de la izquierda, las columnas **privadas** y **públicas** y pulsamos en **OK**. En la ventana de sesión podemos ver lo siguiente:

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

      privadas  públicas  Total
    1        6        30     36
         12,94     23,06
         3,720     2,087

    2       14        32     46
         16,53     29,47
         0,388     0,217

    3       17        17     34
         12,22     21,78
         1,871     1,050

    4        9         3     12
          4,31      7,69
         5,095     2,858

Total       46        82    128

Chi-Sq = 17,286; DF = 3; P-Value = 0,001
1 cells with expected counts less than 5.
```

Recordemos que este contraste solamente puede aplicarse si todas las frecuencias esperadas bajo la hipótesis nula son mayores o iguales que 1 y, además, todas las frecuencias esperadas bajo la hipótesis nula son mayores o iguales que 5, salvo para un 20 % como máximo. El 20 % de las casillas sería el 20 % de 8, que es $1'6$. Como solamente una de las frecuencias esperadas es menor que 5, podemos aplicar esta técnica. El resultado del p-valor es $0'001$, claramente menor que los habituales niveles de significación ($0'05$ ó $0'01$) por lo que rechazamos la hipótesis nula y, en consecuencia, aceptamos que la distribución de los resultados de la prueba no es la misma en las universidades públicas que en las privadas.

4. Ejercicios propuestos

7.1. El número de libros juveniles prestados en 15 días elegidos al azar en los meses de verano (V) e invierno (I) ha sido:

V	54	61	44	50	50	54	59	54	22	58	45	30	25	29	24
I	61	46	50	17	45	31	20	54	37	38	30	42	58	44	58

¿Hay la misma demanda de libros en verano que en invierno?

7.2. Una colección de libros ha sido incluida en un índice de dos formas distintas: a) fichero de entrada simple e índice en cadena, y b) fichero de entrada múltiple e índice simple por orden alfabético. El número de entradas en los dos ficheros para una muestra aleatoria de 12 documentos ha sido:

a)	4	3	4	4	5	4	3	3	3	5	5	2
b)	4	3	6	4	6	6	4	3	4	6	6	2

¿El número de entradas por documento depende del tipo de fichero?

7.3. Se selecciona una muestra aleatoria simple de 10 bibliotecas y se observa el número de items (libros, artículos, revistas, ...) obtenidos y el número de items pedidos por el servicio de préstamo interbibliotecario de cada una de ellas en el último año. Los resultados son los siguientes:

obtenidos	920	1.274	768	608	776	874	744	484	826	2.174
pedidos	874	489	1.175	1.034	1.752	588	670	622	747	1.793

¿En toda la población de bibliotecas, el número medio de items obtenidos es igual al número medio de items pedidos?

7.4. Se eligen aleatoria e independientemente 15 alumnos del primer curso de bachillerato y 12 alumnos del segundo curso de bachillerato, y se observa el número de libros distintos que han pedido prestados en la biblioteca de su instituto durante un curso académico determinado. Los resultados son los siguientes:

1º	2	7	5	9	7	10	8	6	4	3	1	6	9	10	11
2º	10	12	3	7	9	11	7	12	14	9	8	10			

¿Son iguales las medias del número de libros que los alumnos de 1º y 2º han pedido prestados a la biblioteca del instituto durante el curso?

7.5. En un volumen de libros para jóvenes se observa que, para edades comprendidas entre 9 y 11 años, 68 libros fueron escritos por hombres y 94 por mujeres; y para edades comprendidas entre 12 y 14 años, 116 libros fueron escritos por hombres y 28 por mujeres. ¿Hay diferencia significativa entre los dos grupos de edades respecto de la variable sexo de la persona que escribe los libros?

7.6. En un experimento se encuentra que en el año 1980 el número de citas en sociología fue 330 y el número de citas en economía fue 299. En 1990, el número de citas en sociología fue 414 y en economía fue 393. ¿Hay diferencia entre los dos años investigados respecto del número de citas en sociología y economía?

7.7. Los siguientes datos corresponden al número de libros científicos y de ficción prestados a adultos residentes en dos áreas de una determinada ciudad:

	científicos	de ficción
área A	870	745
área B	304	251

¿Hay diferencia significativa entre las dos áreas respecto del tipo de libro demandado?

7.8. Los resúmenes de *Economics Abstracts* se escriben en inglés, francés y alemán. Se extraen muestras aleatorias independientes de 8 resúmenes escritos en cada uno de los tres idiomas mencionados, observando el número de palabras por resumen, siendo los resultados los siguientes:

inglés	francés	alemán
71	111	67
118	113	75
52	84	61
47	84	99
59	84	58
65	94	107
84	90	113
111	90	95

¿La extensión de los resúmenes es la misma para los tres idiomas?

7.9. En una investigación sobre la transferencia de la información se recogieron los siguientes datos:

grupo de trabajo	en persona	por teléfono	otras
A	1.008	269	708
B	409	194	497
C	2.252	544	1.524

¿Hay diferencia entre los grupos de trabajo A, B y C en cuanto a los métodos empleados para transmitir la información?

7.10. Se pregunta a una muestra aleatoria de alumnos de 3º de una facultad de documentación, de cuatro cursos académicos distintos, si conocen los registros MARC de la *British Library*, y los resultados son los siguientes:

	No	Sí	No responde
1994–95	37	56	24
1995–96	24	44	30
1996–97	14	34	41
1997–98	28	54	15

¿Hay diferencia significativa entre los cuatro cursos académicos con respecto a la respuesta dada?

7.11. Se eligen aleatoria e independientemente 10 estantes con libros de geografía, 10 con libros de derecho, 10 con libros de matemáticas y 10 con libros de filosofía, y se cuenta el número de libros por estante. Los resultados son los siguientes:

Geografía	Derecho	Matemáticas	Filosofía
25	21	36	25
30	21	32	27
30	33	30	26
29	23	30	26
25	16	32	21
23	26	33	28
28	26	33	30
33	28	28	31
25	26	39	28
25	21	43	32

¿El número medio de libros por estante es igual para las cuatro materias?

7.12. En una muestra aleatoria simple de 12 días se observa el número de libros prestados en diferentes materias (científicos, novelas, ensayos, arte, música) siendo los resultados los siguientes:

científicos	novelas	ensayos	arte	música
24	40	19	23	21
29	39	17	15	17
33	45	15	13	20
30	38	10	19	16
36	33	12	17	14
27	30	15	20	12
24	25	20	21	11
19	38	23	9	8
16	27	25	23	21
35	39	14	17	14
37	41	21	19	12
32	47	11	14	17

¿El número medio de libros prestados diariamente es igual en las cinco materias?

7.13. En una investigación sobre el uso que los profesores de distintos departamentos hacen de las revistas científicas, se encontró que 34 de los 50 profesores del departamento A, 22 de los 40 profesores del departamento B y 15 de los 35 profesores del departamento C, utilizan las revistas como ayuda en su trabajo académico (y el resto no). ¿Hay diferencia significativa entre los tres departamentos respecto del uso que hacen de las revistas científicas?