Práctica 4. Contrastes no paramétricos en una población

1. Contraste de aleatoriedad de la muestra

El contraste de las rachas sobre aleatoriedad de una muestra se realiza mediante la opción **Stat** ⇒**Nonparametrics** ⇒**Run Test**. Esta prueba no puede utilizarse si los valores de la variable han sido ordenados en el archivo de datos.

Como ya sabemos, este contraste se basa en el concepto de racha, que es una secuencia de observaciones de un mismo tipo precedida y continuada por otro tipo de observaciones o por ninguna. Esto supone que los datos son sólo de dos tipos; es decir, que la variable está dicotomizada. Si esto no sucediera, se pueden reducir los datos a dos tipos mediante lo siguiente: asignar un símbolo (por ejemplo, "+") a los datos que son mayores que la media (o la mediana) y otro símbolo (por ejemplo, "−") a los que son menores o iguales que la media (o la mediana, respectivamente).

Con los datos del archivo **Pulse.mtw** vamos a comprobar si se puede aceptar, con un nivel de significación de $0'05$, que la muestra de resultados de la variable **Pulse1** es aleatoria. Vamos a realizar la dicotomización de los datos a través de la mediana, por lo cual la calculamos previamente. Podemos comprobar que dicha mediana es 71. Ahora seleccionamos **Stat** ⇒**Nonparametrics** ⇒**Run Test**. En el cuadro de diálogo resultante, activamos el recuadro **Variables** (haciendo *clic* dentro de él); seleccionamos (haciendo doble *clic* sobre su nombre) la columna '**Pulse1**'. Si dejamos activada la opción **Above and below the mean** la variable se dicotomizaría a través de su media. Como queremos dicotomizar a través de la mediana, activamos **Above and below** y tecleamos el valor de la mediana; es decir, **71**. Pulsando en **OK** podemos comprobar, en la ventana de sesión, que el p-valor es $0'294$, mayor que el nivel de significación elegido $(0'05)$, por lo que podemos aceptar que la muestra de resultados de dicha variable es aleatoria.

2. Contrastes de Normalidad

En ***Minitab*** hay varias formas de comprobar la Normalidad de una variable. Una de ellas es la opción **Stat**⇒**Basic Statistics**⇒**Normality Test**.

Recordemos que para poder aplicar un contraste de Normalidad es necesario comprobar previamente que la muestra de datos es aleatoria.

Con la hoja de datos **Pulse.mtw** hemos comprobado que la muestra de resultados de la columna **Pulse1** es aleatoria. Por tanto, podemos ahora realizar un contraste de Normalidad para ver si se puede aceptar, con un nivel de significación de $0'05$, que la variable **Pulse1** es Normal. Para ello, usamos **Stat**⇒**Basic Statistics**⇒**Normality Test**. En el cuadro de diálogo resultante, en **Variable** seleccionamos, de la lista de variables de la izquierda, la columna '**Pulse1**'; en **Percentile Lines** dejamos lo

que está activado por defecto, que es **None**; en **Tests for Normality** podemos activar uno de los siguientes tres contrastes: Anderson-Darling, Ryan-Joiner o Kolmogorov-Smirnov. Por ejemplo, vamos a activar el último test, **Kolmogorov-Smirnov**. El recuadro **Title** vamos a dejarlo en blanco. Por último, pulsamos en **OK**. El resultado es un gráfico probabilístico en el cual también está indicado el p-valor, que es mayor que $0'15$. Este p-valor es mayor que el nivel de significación elegido $(0'05)$ y, por tanto, podemos aceptar que la variable **Pulse1** es Normal.

3. Contraste chi-cuadrado sobre independencia de dos variables aleatorias

Hasta ahora se ha considerado una única variable cuyas observaciones en una población daban lugar a ciertas hipótesis convenientes de contrastar mediante un test. Sin embargo, es frecuente el problema de estudiar conjuntamente dos variables en los mismos individuos y preguntarse si existe o no algún tipo de relación entre ellas, es decir, si los valores que tome una de ellas van a condicionar de algún modo los valores de la otra. El método estadístico para responder a tal pregunta varía con el tipo de variables implicadas. Cuando ambas son cualitativas, la técnica oportuna es el *test chi-cuadrado de Pearson*; aunque este método también se puede emplear cuando las variables son cuantitativas.

En ***Minitab*** hay dos formas de aplicar este contraste, según tengamos recogidos los datos. Explicamos estos dos casos en los dos sub-apartados siguientes.

3.1. Datos en una tabla de doble entrada

Si los datos están recogidos en una tabla de doble entrada, se utiliza la opción **Stat⇒Tables⇒Chi-Square Test (Two-Way Table in Worksheet)**.

Vamos a hacer el siguiente ejemplo: Se desea averiguar si existe asociación entre el sexo y el uso de la biblioteca. A tal efecto, se tomó una muestra aleatoria de 30 mujeres y 30 hombres y se les clasificó de la siguiente manera:

	usuarios	no usuarios
hombres	6	24
mujeres	14	16

Para realizar este contraste con ***Minitab***, en primer lugar tenemos que introducir la tabla de doble entrada anterior en una nueva hoja de datos que podemos denominar **Ejemplo_Independencia.mtw**. Los datos tienen que ser introducidos tal y como se muestra a continuación:

↓	C1	C2
	SI	NO
1	6	24
2	14	16

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 **C1** y **C2**; es decir, **'SI'** y **'NO'** y pulsamos en **OK**. En la ventana de sesión podemos ver el resultado del p-valor, que es $0'028$.

Si consideramos un nivel de significación de $\alpha = 0'01$ entonces el p-valor es mayor que α, por lo que podríamos aceptar la hipótesis nula de independencia. Pero si consideramos un nivel de significación de $\alpha = 0'05$ (que es lo usual) entonces el p-valor es menor que α, por lo que no podríamos aceptar la hipótesis nula de independencia, aceptando entonces que existe relación entre el sexo y el uso de la biblioteca.

3.2. Datos en dos (o tres) columnas

Si los datos están recogidos en dos (o tres) columnas, se utiliza la opción **Stat⇒Tables⇒Cross Tabulation and Chi-Square**.

Ejemplo 1. Vamos a hacer el mismo ejemplo que en el apartado anterior, pero con la opción **Stat⇒Tables⇒Cross Tabulation and Chi-Square**. Para ello, en primer lugar tenemos que introducir los datos (en la hoja de datos **Ejemplo_Independencia.mtw**) tal como se muestra a continuación:

C3-T	C4-T	C5
sexo	usuario	frecuencia
H	SI	6
H	NO	24
M	SI	14
M	NO	16

Como se puede observar, hemos creado tres nuevas columnas que contienen todas las combinaciones posibles de resultados de las dos variables y sus frecuencias conjuntas: la columna **sexo** tiene por resultados **H** (hombre) y **M** (mujer); la columna **usuario** tiene por resultados **SI** (la persona sí es usuaria de la biblioteca) y **NO** (la persona no es usuaria de la biblioteca); la columna **frecuencia** contiene las frecuencias conjuntas de todas y cada una de las combinaciones posibles de los resultados de las dos variables mencionadas.

Ahora seleccionamos **Stat⇒Tables⇒Cross Tabulation and Chi-Square**. En **Categorical variables** se tienen que especificar las variables para las cuales vamos a hacer el test de independencia; en nuestro ejemplo, en **For rows** tenemos que seleccionar, de la lista de variables de la izquierda, la columna **'sexo'**; en **For columns** tenemos que seleccionar, de la lista de variables de la izquierda, la columna **'usuario'**. El recuadro **For layers** (capas) lo dejamos en blanco. En **Frequencies are in** tenemos que seleccionar, de la lista de variables de la izquierda, la columna **'frecuencia'**. Pulsamos el botón **Chi-Square** y, en el cuadro de diálogo resultante, dejamos activada la opción **Chi-Square Analysis** y pulsamos en **OK**. Dejamos lo que aparece por defecto en el cuadro de diálogo inicial y pulsamos en **OK**. En la ventana de sesión podemos comprobar que los resultados del contraste de hipótesis son los mismos que antes (p-valor=$0'028$) y, por tanto, las conclusiones, obviamente, son las mismas.

Ejemplo 2. Para utilizar la opción **Stat⇒Tables⇒Cross Tabulation and Chi-Square** no es necesario que tengamos una columna con las frecuencias de cada combinación de resultados de dos variables; también se puede utilizar dicha opción si solamente tenemos **dos columnas** que contienen los resultados de una variable bidimensional, (x_i, y_i), pero es necesario que las dos variables sean de tipo discreto, con pocos resultados distintos; de lo contrario no se puede aplicar este contraste.

Para hacer un ejemplo de este caso, vamos a activar (o abrir) la hoja de datos **Pulse.mtw**. Vamos a comprobar si existe dependencia entre las variables **Smokes** (la persona es fumadora o no) y **Sex**

(sexo). La hipótesis nula es H_0: *"No existe relación entre el sexo y ser fumador o no"*. Como vemos, en la *Worksheet* los datos están recogidos en dos columnas (no en tres). Para realizar este contraste seleccionamos **Stat⇒Tables⇒Cross Tabulation and Chi-Square**; en **For rows** seleccionamos la columna **'Smokes'**; en **For columns** seleccionamos la columna **'Sex'**; no escribimos nada en **For layers** (capas) y tampoco escribimos nada en **Frequencies are in**. Pulsamos el botón **Chi-Square** y, en el cuadro de diálogo resultante, activamos **Chi-Square Analysis** y **Expected cell counts**, y pulsamos en **OK**. Finalmente, volvemos a pulsar **OK** en el cuadro de diálogo inicial. En la ventana de sesión aparece lo siguiente:

```
Rows: Smokes   Columns: Sex

              1       2     All

1            20       8      28
          17,35   10,65   28,00

2            37      27      64
          39,65   24,35   64,00

All          57      35      92
          57,00   35,00   92,00

Cell Contents:      Count
                    Expected count


Pearson Chi-Square = 1,532; DF = 1; P-Value = 0,216
```

Como podemos observar, aparecen las frecuencias observadas y las frecuencias esperadas bajo la hipótesis nula. Podemos comprobar que estas últimas frecuencias son todas mayores o iguales que 5, por lo cual se puede aplicar esta técnica (el test chi-cuadrado de independencia). 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. Si no ocurriera esto, ***Minitab*** nos lo especificaría en la ventana de sesión, y por tanto el test quedaría invalidado. Como podemos ver, tenemos el resultado del estadístico χ^2 y el resultado del p-valor, que es $0'216$, claramente mayor que los habituales niveles de significación ($0'05$ ó $0'01$), por lo que podemos aceptar la hipótesis nula de independencia de las dos variables aleatorias; es decir, podemos aceptar que no existe relación entre el sexo y ser fumador o no.

4. Ejercicios propuestos

4.1.

a) Crea un nuevo proyecto de ***Minitab***.

b) Abre la hoja de datos **Prestamos.mtw** (datos del Ejercicio 1.1).

c) Calcula de mediana de la columna **PPU**.

d) Utilizando la mediana (para dicotomizar) en el contraste de las rachas, ¿se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la muestra de datos de la variable **PPU** (**porcentaje anual de préstamos por usuario**) es aleatoria? ¿Por qué?

e) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la variable **PPU** es Normal? ¿Por qué?

f) Graba el proyecto con el siguiente nombre: **Ejercicio4-1.mpj**

4.2.

a) Crea un nuevo proyecto de ***Minitab***.

b) Abre la hoja de datos **Transacciones.mtw** (datos del Ejercicio 1.2).

c) Utilizando la media (para dicotomizar) en el contraste de las rachas, ¿se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que las muestras de los datos de las variables **TR**, **TRF** y **Porcentaje TRF** son aleatorias? ¿Por qué?

d) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que las variables **TR**, **TRF** y **Porcentaje TRF** son Normales? ¿Por qué?

e) Graba el proyecto con el siguiente nombre: **Ejercicio4-2.mpj**

4.3. Los siguientes datos corresponden a las edades de una muestra de 10 personas que visitan una biblioteca.

19	24	83	30	17	23	33	19	68	56

a) Crea un nuevo proyecto de ***Minitab***.

b) Guarda los datos en el archivo **Edad.mtw**

c) Calcula de mediana.

d) Utilizando la mediana (para dicotomizar) en el contraste de las rachas, ¿se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la muestra es aleatoria? ¿Por qué?

e) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la variable aleatoria *edad de las personas que visitan la biblioteca* es Normal? ¿Por qué?

f) Graba el proyecto con el siguiente nombre: **Ejercicio4-3.mpj**

4.4. El rector de una universidad española desea saber la opinión del profesorado en relación con un proyecto por el cual todos los libros comprados por los departamentos se llevarían a una biblioteca general universitaria ubicada en un edificio independiente de las facultades. Para ello, selecciona una muestra aleatoria de 370 profesores de distintos rangos académicos (A.E.U.= Ayudante de Escuela Universitaria, A.F.= Ayudante de Facultad, T.E.U.=Titular de Escuela Universitaria, T.U.= Titular de Universidad, C.U.= Catedrático de Universidad). Los resultados se reflejan en la siguiente tabla:

	A.E.U.	A.F.	T.E.U.	T.U.	C.U.
en contra	30	55	95	14	12
indiferente	15	20	17	8	10
a favor	10	25	38	8	13

a) Crea un nuevo proyecto de ***Minitab***.

b) Guarda los datos en el archivo **Rango-Opinion.mtw**

c) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'01$, que existe relación entre el rango académico y la opinión de los profesores respecto del proyecto mencionado? ¿Por qué?

d) Graba el proyecto con el siguiente nombre: **Ejercicio4-4.mpj**

4.5. Un profesor de estadística de un Grado en Información y Documentación quiere estudiar la mejor forma de obtener un buen resultado en la asignatura y para ello solicita la colaboración de los alumnos durante varios cursos académicos planteándoles el siguiente esquema: al final del primer parcial califica a todos los alumnos según los resultados del examen en A (sobresaliente y notable), B (aprobado) y C (suspenso); luego les pide que contesten cuál ha sido su método de trabajo ante la signatura (I= sólo estudia teoría, II= sólo estudia problemas, III= estudia teoría y problemas). Conocidos los resultados, el profesor construye la siguiente tabla:

		Método de trabajo		
		I	II	III
Calificación	A	15	12	65
	B	58	70	85
	C	40	102	53

a) Crea un nuevo proyecto de ***Minitab***.

b) Guarda los datos en el archivo **Calificacion-Metodo.mtw**

c) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la calificación es independiente del método de trabajo empleado? ¿Por qué?

d) Graba el proyecto con el siguiente nombre: **Ejercicio4-5.mpj**

4.6. En una determinada facultad se considera una muestra de 807 alumnos y se realiza una encuesta para saber cuántas horas diarias dedica cada alumno al estudio en la biblioteca, obteniéndose la siguiente tabla de resultados:

		Curso de la licenciatura				
		1º	2º	3º	4º	5º
Nº de horas	menos de 1 hora	18	20	32	77	96
	entre 1 y 3 horas	22	35	90	83	50
	más de 3 horas	60	70	80	60	14

a) Crea un nuevo proyecto de ***Minitab***.

b) Guarda los datos en el archivo **Curso-Tiempo.mtw**

c) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que existe relación entre el curso al que pertenece el alumno y el tiempo que dedica al estudio en la biblioteca? ¿Por qué?

d) Graba el proyecto con el siguiente nombre: **Ejercicio4-6.mpj**