

Dra. Josefa Marín Fernández
Departamento de Estadística e Investigación Operativa
Universidad de Murcia

Estadística
Licenciatura en Documentación
Manual de prácticas con MINITAB 15

Curso 2010-11

Contenidos

1. Introducción a Minitab	5
1.1. Elementos de Minitab para Windows	5
1.1.1. Introducción	5
1.1.2. Barra de menús	6
1.2. Entrada, grabación y lectura de datos	7
1.2.1. Entrada de datos	7
1.2.2. Grabación de datos	8
1.2.3. Lectura de datos	9
1.3. Opciones principales de los menús <i>Data</i> y <i>Calc</i>	9
1.3.1. Desapilamiento de columnas	9
1.3.2. Apilamiento de columnas	10
1.3.3. Ordenación de datos	10
1.3.4. Codificación o clasificación de datos	11
1.3.5. Transformación de variables	11
1.3.6. Creación de datos por patrón	13
1.4. Ejercicios propuestos	13
2. Estadística descriptiva	17
2.1. Distribución de frecuencias	17
2.2. Representaciones gráficas	18
2.2.1. Gráfico de sectores o de <i>pastel</i>	18
2.2.2. Diagrama de barras	19
2.2.2.1. Diagrama de barras simple	19
2.2.2.2. Diagrama de barras agrupado (o apilado)	20
2.2.3. Histograma	21
2.3. Medidas descriptivas de los datos	22
2.4. Correlación y regresión lineal	24
2.4.1. Diagrama de dispersión o nube de puntos	24

2.4.2. Coeficiente de correlación lineal	25
2.4.3. Rectas de regresión	25
2.5. Ejercicios propuestos	26
3. Modelos de probabilidad	29
3.1. Muestras aleatorias de las distribuciones usuales	29
3.2. Función de densidad y función de probabilidad	30
3.3. Función de distribución	30
3.4. Inversa de la función de distribución (percentiles)	31
3.5. Ejercicios propuestos	31
4. Contrastes no paramétricos en una población	35
4.1. Contraste de aleatoriedad de la muestra	35
4.2. Contrastes de normalidad	36
4.3. Ejercicios propuestos	37
5. Contrastes paramétricos en una población	41
5.1. Contrastes sobre la media	41
5.1.1. Contraste sobre la media cuando la desviación típica poblacional es conocida	41
5.1.2. Contraste sobre la media cuando la desviación típica poblacional es desconocida	43
5.2. Contrastes sobre la varianza	44
5.3. Ejercicios propuestos	45
6. Contrastes paramétricos en dos poblaciones	49
6.1. Comparación de dos varianzas con muestras independientes	49
6.2. Comparación de dos medias con muestras independientes	51
6.2.1. Comparación de dos medias con muestras independientes y varianzas pobla- cionales desconocidas pero iguales	51
6.2.2. Comparación de dos medias con muestras independientes y varianzas pobla- cionales desconocidas y distintas	53
6.3. Comparación de dos medias con muestras apareadas	54
6.4. Ejercicios propuestos	55

1

Introducción a Minitab

1.1. Elementos de Minitab para Windows

1.1.1. Introducción

Al ejecutar *Minitab* 15 aparece la *ventana* de la Figura 1.

Como en cualquier otra aplicación Windows, esta *ventana* puede modificarse en cuanto al tamaño y a la disposición de sus elementos. Se trata de una *ventana* típica de una aplicación Windows que consta de los siguientes elementos:

- En la primera línea aparece la **barra de título**, que contiene el nombre de la ventana y los botones de minimizar, maximizar y cerrar.
- En la segunda línea está la **barra de menús**, que consta de los 10 menús que luego comentaremos.
- Las líneas tercera y cuarta conforman la **barra de herramientas** donde, mediante botones con iconos, se representan algunas de las operaciones más habituales. Si pasamos el puntero del ratón por cualquiera de ellos, aparecerá en la pantalla un texto indicando la función que se activa.
- Después aparece la **ventana de sesión (Session)**. Es la parte donde aparecen los resultados de los análisis realizados. También sirve para escribir instrucciones, como forma alternativa al uso de los menús.
- A continuación tenemos la **hoja de datos (Worksheet)**. Tiene el aspecto de una hoja de cálculo, con filas y columnas. Las columnas se denominan $C1, C2, \dots$, tal como está escrito, pero también se les puede dar un nombre, escribiéndolo debajo de $C1, C2, \dots$. Cada columna es una variable y cada fila corresponde a una observación o caso.
- En la parte inferior aparece (minimizada) la **ventana de proyecto (Project Manager)**. En *Minitab* un proyecto incluye la hoja de datos, el contenido de la ventana de sesión, los gráficos que se hayan realizado, los valores de las constantes y de las matrices que se hayan creado, etc.

Para activar la ventana de sesión (**Session**) podemos hacer *clic* sobre ella o podemos hacer *clic* sobre su icono en la barra de herramientas (primer icono de la Figura 2). Para activar la hoja de

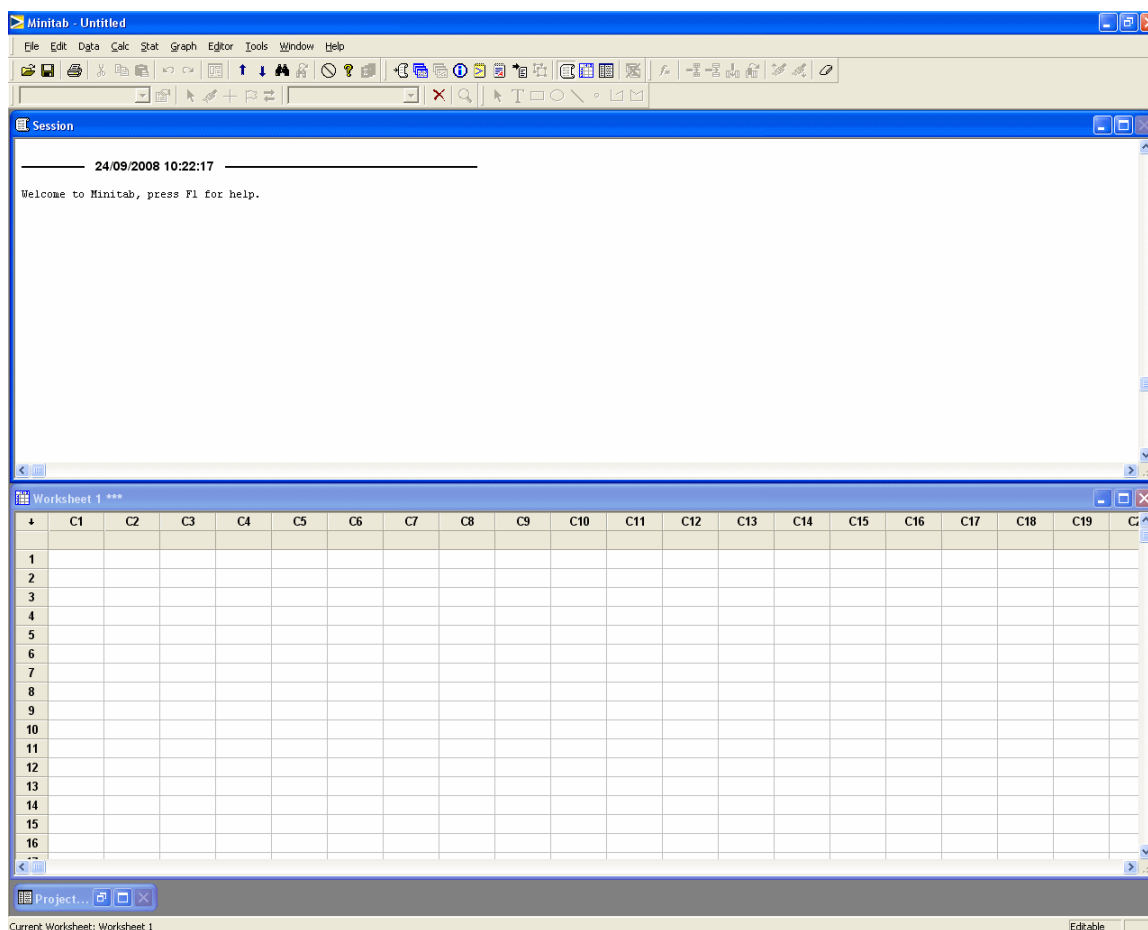


Figura 1: Ventana inicial de Minitab 15

datos (**Worksheet**) podemos hacer *clic* sobre ella o podemos hacer *clic* sobre su icono en la barra de herramientas (segundo icono de la Figura 2). Para activar la ventana de proyecto (**Project Manager**) podemos maximizarla o podemos hacer *clic* sobre su icono en la barra de herramientas (tercer icono de la Figura 2).



Figura 2: Iconos para activar las ventanas de sesión, de datos o de proyecto

Para salir del programa se selecciona la opción **File** \Rightarrow **Exit** o se pulsa el botón de la esquina superior derecha: .

1.1.2. Barra de menús

A continuación se da un resumen de lo que se puede encontrar en la **barra de menús**:

File: Mediante este menú se pueden abrir, crear o grabar los diferentes archivos que **Minitab** emplea, ya sean de datos, instrucciones, resultados o procesos. Igualmente, es posible controlar las tareas de impresión.

- Edit:** Permite realizar las tareas habituales de edición: modificar, borrar, copiar, pegar, seleccionar, etc.
- Data:** Este menú permite, entre otras cosas, efectuar modificaciones en los archivos de datos: extraer un subconjunto de datos, apilar y desapilar, ordenar, codificar, etc.
- Calc:** Aquí se encuentran todas las opciones relativas a la modificación y generación de nuevas variables, cálculo de los estadísticos, introducción de datos por patrón, cálculo de las distribuciones de probabilidad, etc.
- Stat:** Mediante este menú se accede a los diferentes análisis estadísticos que se pueden realizar con los datos.
- Graph:** Permite la creación y edición de diversos tipos de gráficos. Algunos de ellos son también accesibles a través de determinadas técnicas estadísticas.
- Editor:** Tiene distintas opciones según esté activada la ventana de sesión o la hoja de datos. Con la ventana de sesión activada permite, por ejemplo, que se pueda escribir (en dicha ventana) utilizando el *lenguaje de comandos*.
- Tools:** Entre otras cosas, permite personificar la barra de herramientas y la barra de menús.
- Windows:** Dispone de las funciones habituales para controlar las ventanas.
- Help:** Proporciona ayuda al usuario en el formato típico de Windows.

1.2. Entrada, grabación y lectura de datos

1.2.1. Entrada de datos

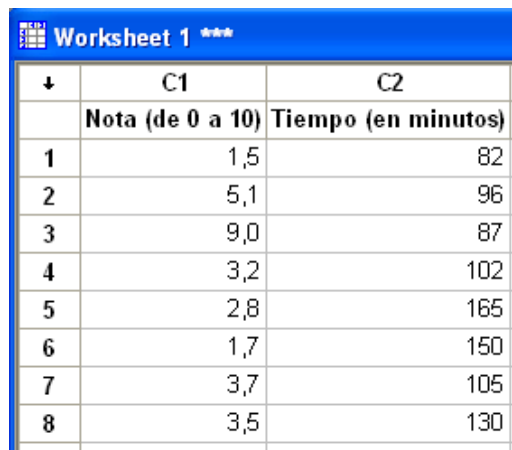
Antes de realizar ningún análisis estadístico es necesario tener un conjunto de datos en uso, para lo cual podemos proceder de cuatro formas:

- Escribirlos a través del teclado.
- Obtenerlos desde un archivo.
- Pegarlos.
- Generarlos por patrón o de forma aleatoria.

Para introducir datos a través del teclado, activamos, en primer lugar, la hoja de datos. En la parte superior aparece $C1, C2, C3, \dots$ y debajo un espacio en blanco para poner el nombre de cada variable. La flechita del extremo superior izquierdo de la hoja de datos señala hacia dónde se mueve el cursor al pulsar la tecla **Intro**. Por defecto apunta hacia abajo, \downarrow ; si se hace *clic* sobre ella, apuntará hacia la derecha, \Rightarrow . Para escribir datos por columna no hay más que situarse en la casilla del caso 1, teclear el dato y pulsar la tecla **Intro**. La casilla activa se moverá hacia abajo. Si tecleamos datos que no son numéricos podemos observar que junto a CJ aparece un guión y la letra T (es decir, $CJ - T$), lo que significa que **Minitab** reconoce que la variable es cualitativa (o de texto).

Con esta versión de **Minitab**, al introducir los resultados de una variable cuantitativa (o numérica) tenemos que recordar que la separación decimal se hace mediante una coma (en parte de abajo). Si, por ejemplo, ponemos un punto como separación decimal, entonces **Minitab** consideraría, automáticamente, que dicha la variable es cualitativa o de texto (junto a CJ aparece un guión y la letra T) y, por tanto, no podríamos hacer ningún cálculo matemático con los datos de esta variable.

Por ejemplo, podemos introducir los datos de la Figura 3, correspondientes a las calificaciones (de 0 a 10 puntos) en el examen de Estadística y el tiempo (en minutos) empleado en realizar dicho examen.



↓	C1	C2
	Nota (de 0 a 10)	Tiempo (en minutos)
1	1,5	82
2	5,1	96
3	9,0	87
4	3,2	102
5	2,8	165
6	1,7	150
7	3,7	105
8	3,5	130

Figura 3: Ejemplo para introducir datos a través del teclado

Si el nombre de la variable (columna) no es suficientemente explicativo, podemos escribir una descripción de la variable para poder consultarla en cualquier momento. Para ello, hacemos *clic* sobre el nombre de la variable (o sobre su número de columna: *CJ*); pulsamos con el botón derecho del ratón y seleccionamos **Column**⇒**Description**. Por ejemplo, podríamos escribir etiquetas descriptivas para las variables **Nota (de 0 a 10)** y **Tiempo (en minutos)**.

Para cambiar el formato de una variable (columna) numérica, hacemos *clic* sobre el nombre de la variable (o sobre su número de columna: *CJ*); pulsamos con el botón derecho del ratón y seleccionamos **Format Column**⇒**Numeric**. Una de las utilidades de esta opción es **el cambio del número de decimales** que se muestran en la hoja de datos. Por ejemplo, podríamos hacer que **Minitab** mostrase 2 decimales en la columna **Nota (de 0 a 10)**.

Una hoja de datos de **Minitab** puede contener hasta 4 000 columnas, 1 000 constantes y hasta 10 000 000 de filas, dependiendo de la memoria que tenga el ordenador.

1.2.2. Grabación de datos

Una vez introducidos los datos, éstos pueden guardarse en un archivo para poder ser utilizados en cualquier otro momento.

Para guardar únicamente la hoja de datos hay que seleccionar **File**⇒**Save Current Worksheet As** (si vamos a grabar el archivo de datos por primera vez y, por tanto, vamos a ponerle un nombre a dicho archivo) ó **File**⇒**Save Current Worksheet** (si el archivo de datos ya tiene nombre pero queremos guardar los últimos cambios realizados). Por ejemplo, podemos guardar los datos de la Figura 3 en un archivo que denominaremos **Notas_Tiempo.mtw**. Para ello, elegimos la opción **File**⇒**Save Current Worksheet As**; en **Guardar en** seleccionamos la carpeta en la que vamos a grabar esta hoja de datos; en **Nombre** escribimos **Notas_Tiempo** (**Minitab** le asigna automáticamente la extensión **.mtw**) y, por último, pulsamos en **Guardar**.

Si queremos grabar toda la información (la hoja de datos, el contenido de la ventana de sesión, los gráficos que se hayan realizado, los valores de las constantes y de las matrices que se hayan creado,

etc.) usaremos la opción **File⇒Save Project As** (si vamos a grabar el proyecto de *Minitab* por primera vez y, por tanto, vamos a ponerle un nombre a dicho archivo) ó **File⇒Save Project** (si el proyecto ya tiene nombre pero queremos guardar los últimos cambios realizados). Es muy importante diferenciar entre archivos de datos (.mtw) y archivos de proyectos (.mpj).

También se puede guardar solamente la ventana de sesión. Para ello, la activamos y seleccionamos la opción **File⇒Save Session Windows As**.

1.2.3. Lectura de datos

Un archivo sólo puede ser recuperado de la forma en que fue grabado. Si se ha grabado como hoja de datos (.mtw) se recupera con la opción **File⇒Open Worksheet**. Si se ha grabado como proyecto de *Minitab* (.mpj) se recupera con la opción **File⇒Open Project**.

Minitab 15 lleva bastantes archivos de datos como muestra. Éstos se encuentran en **C:\Archivos de programa\Minitab 15\English\Sample Data** y, como ya sabemos, llevan la extensión .mtw. En las aulas de informática de la Universidad de Murcia es posible que se encuentren en **C:\Archivos de programa\UM\Minitab 15\English\Sample Data**.

Por ejemplo, podemos abrir el archivo de datos **Pulse.mtw**. Su contenido fue recogido en una clase de 92 alumnos. De cada estudiante se observó su pulso antes de correr, **Pulse1**; su pulso después de correr, **Pulse2**; si corrió o no, **Ran** (1=Sí corrió, 2=No corrió); si es fumador o no, **Smokes** (1=Sí fuma, 2=No fuma); el sexo, **Sex** (1=Hombre, 2=Mujer); su altura en pulgadas, **Height**; su peso en libras, **Weight**; y su nivel de actividad física, **Activity** (0=Ninguna actividad física, 1=Baja, 2=Media, 3=Alta). Se puede encontrar más información de este archivo de datos con la opción **Help⇒Help⇒Indice**. Bajo la frase **Escriba la palabra clave a buscar** se teclea **Pulse.mtw** y después se hace *clic* en **Mostrar** o se hace doble *clic* sobre el nombre de dicho archivo.

Con la opción **File⇒Open Worksheet** se pueden leer otros tipos de archivos de datos, como hojas de cálculo de Excel, Lotus 1-2-3, dBase, etc. Para obtener una información más detallada sobre los tipos de archivos que *Minitab* puede leer, se selecciona **File⇒Open Worksheet** y, en el cuadro de diálogo resultante, se hace *clic* sobre **Ayuda**.

1.3. Opciones principales de los menús *Data* y *Calc*

Si queremos que en la ventana de sesión (**Session**) aparezcan los comandos que va a utilizar *Minitab* en las opciones que vamos a explicar, activamos la ventana de sesión y luego seleccionamos **Editor⇒Enable Commands**.

1.3.1. Desapilamiento de columnas

La opción **Data⇒Unstack columns** permite separar los resultados de una columna en varias columnas, según los resultados de otra variable o columna (que contiene los subíndices).

Por ejemplo, de la hoja de datos **Pulse.mtw** vamos a desapilar los resultados de la variable **Pulse2** (*pulso después de correr*) según los resultados de la variable **Ran** (*1=Sí corrió, 2=No corrió*).

En primer lugar tenemos que abrir dicha hoja de datos, si no la tenemos abierta ya. Recordemos que para abrirla elegimos la opción **Open Worksheet**; en **Buscar en** seleccionamos la carpeta donde

se encuentra la hoja de datos; activamos **Nombre**; seleccionamos el archivo **Pulse.mtw** y, por último, pulsamos en **Abrir**.

Para realizar el desapilamiento de los resultados de la variable **Pulse2** según los resultados de la variable **Ran** seleccionamos **Data⇒Unstack Columns**; activamos **Unstack the data in** (haciendo *clic* dentro del recuadro); seleccionamos (haciendo doble *clic* sobre su nombre) la variable o columna **Pulse2**; activamos el recuadro **Using subscripts in** (haciendo *clic* dentro del recuadro); y seleccionamos la columna que contiene la procedencia de cada dato, que es **Ran**; en **Store unstacked data in** activamos la opción **After last column in use**; dejamos activado **Name the columns containing the unstacked data** y pulsamos en **OK**.

En la hoja de datos **Pulse.mtw** nos aparecen dos nuevas columnas: **Pulse2_1** y **Pulse2_2**. En la columna **Pulse2_1** hay 35 datos, que son los resultados del pulso después de correr (**Pulse2**) de las personas que sí corrieron (**Ran=1**); y en la columna **Pulse2_2** hay 57 datos, que son los resultados del pulso después de correr (**Pulse2**) de las personas que no corrieron (**Ran=2**).

Debemos grabar la actual hoja de datos con un nombre distinto de **Pulse.mtw** para conservar los datos originales sin transformaciones ni nuevas columnas. Para ello, elegimos la opción **File⇒Save Current Worksheet As**; en **Guardar en** seleccionamos la carpeta en la que vamos a grabar esta hoja de datos; en **Nombre** escribimos **Pulse transformada** y, por último, pulsamos en **Guardar**.

1.3.2. Apilamiento de columnas

Con la opción **Data⇒Stack⇒Columns** se pueden apilar varias columnas en una sola. Opcionalmente se puede indicar de qué columna procede cada valor mediante una nueva variable (subíndices). Si no se hace esta indicación no se podrá identificar la procedencia de cada dato. Esta opción es la contraria de la explicada en el apartado anterior.

Para practicar esta opción podemos apilar los datos de las columnas **Pulse2_1** y **Pulse2_2** de la hoja de datos **Pulse transformada.mtw**. En primer lugar debemos asegurarnos de que la hoja de datos activa es **Pulse transformada.mtw**. Si dicha hoja de datos no está activa, debemos activarla haciendo *clic* sobre ella o seleccionando **Window⇒Pulse transformada.mtw**. A continuación, seleccionamos la opción **Data⇒Stack⇒Columns**; activamos el recuadro **Stack the following columns** y seleccionamos (haciendo doble *clic* sobre sus nombres) las dos columnas que queremos apilar: **'Pulse2_1'** y **'Pulse2_2'**; en **Store stacked data in** activamos la opción **Column of current worksheet** y tecleamos la posición de una columna que esté vacía, por ejemplo, **C11** (o escribimos un nombre para esta nueva columna). En **Store subscripts in** tecleamos la posición de la columna en la que queremos guardar la procedencia de cada dato, por ejemplo, **C12** (o escribimos un nombre para esta nueva columna). Es conveniente dejar activada la opción **Use variable names in subscript column**.

Podemos observar que la columna **Pulse2** y la columna **C11** contienen los mismos resultados, pero no en el mismo orden.

1.3.3. Ordenación de datos

La opción **Data⇒Sort** ordena los datos de una columna según los resultados de una o varias columnas. Lo normal es ordenar una columna según los resultados de dicha columna. Esto es lo que vamos a explicar.

Por ejemplo, en la hoja de datos **Pulse transformada.mtw** vamos a crear una nueva variable (columna) que contenga los resultados de la variable **Pulse1** ordenados de menor a mayor. En primer lugar, activamos dicha hoja de datos (si no la tenemos activada ya). A continuación, seleccionamos **Data⇒Sort**;

activamos el recuadro **Sort column**; seleccionamos (haciendo doble *clic* sobre su nombre) la variable **Pulse1**; activamos el primer recuadro **By column** que aparece y volvemos a seleccionar la misma columna, **Pulse1**. Dejamos desactivada la opción **Descending** para que la ordenación se realice de menor a mayor resultado.

En **Store sorted data in** activamos **Column of current worksheet** y tecleamos el nombre que queremos ponerle a dicha columna, por ejemplo, '**Pulse1 ordenado**'. En este cuadro de diálogo (en realidad, en todos los cuadros de diálogo de *Minitab*), cuando haya que escribir el nombre de una nueva variable (columna) y **el nombre contenga espacios en blanco, guiones, paréntesis, etc., entonces hay que escribirlo entre comillas simples**. La comilla simple suele estar en la misma tecla que el símbolo de cerrar interrogación.

Hay tener cuidado con la ordenación de columnas debido a que los resultados de esta nueva variable no guardan correspondencia con los casos originales. Por ejemplo, la primera persona observada tiene un pulso antes de correr (resultado de **Pulse1**) igual a 64 pulsaciones por minuto, no 48 pulsaciones por minuto, como nos ha salido en el primer lugar de la columna **Pulse1 ordenado**. Como podemos observar, el menor valor de **Pulse1** es 48 y el mayor valor es 100.

1.3.4. Codificación o clasificación de datos

La opción **Data⇒Code** permite la clasificación o codificación de los datos de una columna. Se puede codificar transformando datos numéricos en datos numéricos, datos numéricos en datos de texto, datos de texto en datos de texto, datos de texto en datos numéricos, etc.

Por ejemplo, con la hoja de datos **Pulse transformada.mtw** podemos codificar la variable **Pulse1** de la forma siguiente:

Resultados de Pulse1	Nueva categoría
comprendido entre 48, incluido, y 65, incluido	Pulso bajo
comprendido entre 65, sin incluir, y 83, incluido	Pulso medio
comprendido entre 83, sin incluir, y 100, incluido	Pulso alto

Para ello, seleccionamos **Data⇒Code⇒Numeric to Text**. En **Code data from columns** seleccionamos (haciendo doble *clic* sobre su nombre) la variable **Pulse1**. En **Store coded data in column** escribimos el nombre la nueva variable; por ejemplo, '**codificación de Pulse1**' (con comillas simples, al principio y al final, ya que el nombre tiene espacios en blanco). En la primera línea de **Original values** debemos escribir **48:65**, lo cual es interpretado por *Minitab* de la siguiente manera: todos los resultados comprendidos entre 48, incluido, y 65, incluido. En la primera línea de **New** escribimos **Pulso bajo**. En la segunda línea de **Original values** escribimos **65:83** lo cual es interpretado por *Minitab* de la siguiente manera: todos los resultados comprendidos entre 65, sin incluir, y 83, incluido. En la segunda línea de **New** escribimos **Pulso medio**. En la tercera línea de **Original values** escribimos **83:100** lo cual es interpretado por *Minitab* de la siguiente manera: todos los resultados comprendidos entre 83, sin incluir, y 100, incluido. En la tercera línea de **New** escribimos **Pulso alto**.

1.3.5. Transformación de variables

En este apartado vamos a ver el modo de generar nuevas variables mediante transformaciones efectuadas sobre los valores de las variables ya definidas. Para ello vamos a utilizar la opción **Calc⇒Calculator**

()	Paréntesis	<	Menor que	AND	Operador Y
**	Exponenciación	>	Mayor que	OR	Operador O
*	Multiplicación	<=	Menor o igual que	NOT	Operador NO
/	División	>=	Mayor o igual que		
+	Suma	=	Igual que		
-	Resta	<>	No igual que		

(a) Operadores aritméticos (b) Operadores relacionales (c) Operadores lógicos

Tabla 4: Operaciones aritméticas, relacionales y lógicas

En la Tabla 4 se encuentran recogidos los operadores aritméticos, relacionales y lógicos que están permitidos. Tanto las expresiones aritméticas como las lógicas se evalúan de izquierda a derecha. Todas las expresiones entre paréntesis se evalúan antes que las que están fuera de los paréntesis y ante varios operadores en el mismo nivel, el orden de preferencia (de mayor a menor) es el que figura en la Tabla 4 (de arriba hacia abajo).

Como ya hemos indicado, para construir una nueva variable mediante transformaciones de otras ya existentes, se tiene que elegir la opción **Calc** \Rightarrow **Calculator**, con lo que se abre una ventana que tiene cinco partes fundamentales: arriba a la derecha está el lugar para escribir el nombre de la nueva variable (**Store result in variable**), a la izquierda aparece la lista de variables y constantes existentes, a la derecha está el lugar destinado a la definición de la nueva variable (**Expression**), debajo hay una calculadora y la lista de funciones que se pueden utilizar (**Functions**).

En primer lugar se asigna un nombre a la variable que queremos generar, escribiendo el mismo en el cuadro **Store result in variable**. Normalmente se va a tratar de una variable nueva, pero también cabe la posibilidad de especificar una de las ya existentes. En tal caso la modificación consistirá en sustituir los valores antiguos de la variable con los nuevos resultantes de la transformación numérica que se efectúe.

Una vez que se ha asignado el nombre a la variable, el siguiente paso es definir la expresión que va a permitir calcular los valores de la misma. Tal expresión se escribe en el cuadro **Expression** y puede constar de los siguientes elementos: nombres de variables del archivo original, constantes, operadores y funciones. Para escribir dicha expresión, se puede teclear directamente pero **es recomendable emplear la calculadora, la lista de variables y constantes y la lista de funciones** (haciendo *clic* dentro del recuadro **Expression** y haciendo doble *clic* sobre la variable, sobre la constante o sobre la función). Una vez que hemos terminado de escribir la expresión, pulsamos en **OK**.

Por ejemplo, del archivo de datos **Pulse transformada.mtw** vamos a calcular la media geométrica de las variables **Pulse1** y **Pulse2** (raíz cuadrada del producto de ambas variables; es decir, producto de ambas variables elevado a 1/2). Para ello, seleccionamos la opción **Calc** \Rightarrow **Calculator**; en **Store result in variable** tenemos que teclear la posición de la columna que contendrá los resultados (una columna, **CJ**, que esté vacía) o el nombre que queremos darle a dicha columna. Nosotros vamos a poner a la nueva variable el siguiente nombre: '**Media geométrica Pulse1 Pulse2**' (con comillas simples, al principio y al final, ya que el nombre tiene espacios en blanco). En **Expression** tenemos que colocar la operación que se realiza para determinar la media geométrica indicada: $(\text{'Pulse1'} * \text{'Pulse2'})^{**}(1 / 2)$. Por último, pulsamos en **OK**.

1.3.6. Creación de datos por patrón

Con la opción **Calc**⇒**Make Patterned Data** se generan datos siguiendo un determinado patrón.

Por ejemplo, si queremos generar una lista de los siguientes 100 números: 0'01, 0'02, 0'03, ..., 1, seguiremos los siguientes pasos:

Como estos datos no tienen nada que ver con los datos del archivo **Pulse transformada.mtw**, creamos una nueva hoja de datos con la opción **File**⇒**New**. En el cuadro de diálogo que aparece seleccionamos **Minitab Worksheet**. A esta nueva hoja de datos **Minitab** le asignará el nombre **Worksheet J**, siendo *J* un número natural. Luego podremos cambiarle el nombre con la opción **File**⇒**Save Current Worksheet As**. Seleccionamos, a continuación, la opción **Calc**⇒**Make Patterned Data**⇒**Simple Set of Numbers**. En **Store patterned data in** podemos teclear *C1* o un nombre, por ejemplo '**Patrón entre 0 y 1**' (con comillas simples, al principio y al final, ya que el nombre tiene espacios en blanco). En **From first value** tecleamos **0,01**, en **To last value** escribimos **1** y en **In steps of** ponemos **0,01**. Tanto en **List each value** como en **List the whole sequence** dejamos lo que está puesto por defecto, que es **1**.

1.4. Ejercicios propuestos

Ejercicio 1.1 En la Tabla 5 se muestra el número anual de usuarios de una biblioteca determinada y el número anual de préstamos durante 10 años elegidos al azar.

año	usuarios	préstamos
1	296	155
2	459	275
3	602	322
4	798	582
5	915	761
6	1145	856
7	1338	1030
8	1576	1254
9	1780	1465
10	2050	1675

Tabla 5

- Crea un nuevo proyecto de **Minitab**.
- Introduce los datos (sin incluir, obviamente, la primera columna, que indica el número de caso). Pon los siguientes nombres a las dos variables: **Usuarios** y **Préstamos**. Graba la hoja de datos en un archivo denominado **Prestamos.mtw**
- Calcula, en una nueva columna, la variable que indica el **porcentaje anual de préstamos por usuario**, resultado de multiplicar por 100 el resultado de dividir el número anual de préstamos entre el número anual de usuarios. Pon a la nueva variable el siguiente nombre: **PPU**. Haz que los resultados aparezcan con tres decimales. Pon una etiqueta descriptiva a esta variable. Vuelve a grabar la hoja de datos.

- d) Ordena los datos de la variable **PPU** en orden creciente. Pon un nombre adecuado a la nueva columna. Pon una etiqueta descriptiva a esta columna. A partir de esta ordenación determina el valor mínimo y el valor máximo de **PPU**.
- e) Clasifica los datos de la variable **PPU** en 4 categorías o intervalos de la misma amplitud. Llama a la nueva variable **Intervalos PPU**. Las categorías han de denotarse como lo hacemos en las clases de teoría; es decir, $[a, b]$ o $(a, b]$ (sustituyendo, obviamente, a y b por los límites de los intervalos de clase). Vuelve a grabar la hoja de datos.
- f) Graba el proyecto con el siguiente nombre: **Ejercicio1-1.mpj**

Ejercicio 1.2 En la Tabla 6 aparece el número anual de transacciones de referencia y el número anual de transacciones de referencia finalizadas en 20 biblioteca elegidas al azar.

biblioteca	tipo de biblioteca	transacciones de referencia	transacciones de referencia finalizadas
1	1	11500	9400
2	1	8600	7200
3	1	20400	18100
4	1	5800	4600
5	1	6500	5800
6	1	13700	10900
7	1	12400	11200
8	1	5300	4700
9	1	6700	5600
10	1	15600	12500
11	2	1900	1700
12	2	9600	7800
13	2	8400	6900
14	2	6200	4900
15	2	7700	5900
16	2	5600	4200
17	2	6200	4900
18	2	4800	3500
19	2	3800	2600
20	2	2400	2200

Tabla 6

- a) Crea un nuevo proyecto de **Minitab**.
- b) Introduce los datos (sin incluir, obviamente, la primera columna, que indica el número de caso). Pon los siguientes nombres a las variables: **Tipo**, **TR** y **TRF**. Pon una etiqueta descriptiva a cada variable. En lo que respecta a la variable **Tipo** hay que dejar claro que el valor **1** significa *biblioteca pública* y el valor **2** significa *biblioteca universitaria*. Graba la hoja de datos en un archivo denominado **Transacciones.mtw**

- c) Crea una nueva variable, denominada **Tipo biblioteca**, que contenga las categorías de la variable **Tipo** designadas de la siguiente manera: *bib. pública* (en vez de **1**) y *bib. universitaria* (en vez de **2**). Vuelve a grabar la hoja de datos.
- d) Calcula, en una nueva columna, la variable que indica el **porcentaje de transacciones de referencia finalizadas**, que se determina multiplicando por cien el resultado de dividir el número anual de transacciones de referencia finalizadas entre el número anual de transacciones de referencia. Pon a la nueva variable el siguiente nombre: **Porcentaje TRF**. Haz que los resultados aparezcan con 5 decimales. Pon una etiqueta descriptiva a esta variable. Vuelve a grabar la hoja de datos.
- e) Desapila los resultados de la variable **Porcentaje TRF** según los resultados de la variable **Tipo biblioteca**.
- f) Ordena los datos de la variable **Porcentaje TRF** en orden creciente. Pon un nombre adecuado a la nueva columna. Pon una etiqueta descriptiva a esta columna. A partir de esta ordenación determina el valor mínimo y el valor máximo de **Porcentaje TRF**.
- g) Clasifica los datos de la variable **Porcentaje TRF** en 3 categorías o intervalos de la misma amplitud. Llama a la nueva variable **Intervalos Porcentaje TRF**. Las categorías han de denotarse como lo hacemos en las clases de teoría; es decir, $[a, b]$ o $(a, b]$ (sustituyendo, obviamente, a y b por los límites de los intervalos de clase). Vuelve a grabar la hoja de datos.
- h) Graba el proyecto con el siguiente nombre: **Ejercicio1-2.mpj**

2

Estadística descriptiva

2.1. Distribución de frecuencias

Con *Minitab*, para determinar la distribución de frecuencias de una (o más variables) utilizamos la opción **Stat⇒Tables⇒Tally Individual Variables**.

Para practicar esta opción, podemos utilizar el archivo de datos (Worksheet) **Pulse.mtw**. En primer lugar tenemos que abrir dicha hoja de datos, si no la tenemos abierta ya. Recordemos que su contenido fue recogido en una clase de 92 alumnos. De cada estudiante se observó su pulso antes de correr, **Pulse1**; su pulso después de correr, **Pulse2**; si corrió o no, **Ran** (1=Sí corrió, 2=No corrió); si es fumador o no, **Smokes** (1=Sí fuma, 2=No fuma); el sexo, **Sex** (1=Hombre, 2=Mujer); su altura en pulgadas, **Height**; su peso en libras, **Weight**; y su nivel de actividad física, **Activity** (0=Ninguna actividad, 1=Baja, 2=Media, 3=Alta).

De la hoja de datos **Pulse.mtw** vamos a averiguar la distribución de frecuencias de todas las variables. Para ello, seleccionamos la opción **Stat⇒Tables⇒Tally Individual Variables**; en el recuadro **Variables** seleccionamos, de la lista de variables de la izquierda, todas las columnas. En **Display** activamos los cuatro tipos de frecuencias que aparecen: **Counts** (frecuencia absoluta), **Percents** (porcentaje), **Cumulative counts** (frecuencia acumulada absoluta) y **Cumulative percents** (porcentaje acumulado). Por último, pulsamos en **OK**.

En la ventana de sesión podemos observar, por ejemplo:

- Hay 57 personas (de las 92 que componen la muestra) que no corrieron; es decir, 57 es la frecuencia absoluta de **Ran=2**.
- Hay 64 personas (de las 92 que componen la muestra) que no fuman; es decir, 64 es la frecuencia absoluta de **Smokes=2**.
- El 38'04 % del total de personas de la muestra son mujeres; es decir, 38'04 % es el porcentaje de **Sex=2**.
- 46 personas (la mitad de las personas que componen la muestra) tienen 70 pulsaciones o menos antes de correr; es decir, 46 es la frecuencia acumulada absoluta de **Pulse1=70**.

- El 75 % de las personas (las tres cuartas partes del total) tienen 84 pulsaciones o menos después de correr; es decir, 75 % es el porcentaje acumulado de **Pulse2=84**.

2.2. Representaciones gráficas

En *Minitab* la mejor opción para hacer representaciones gráficas es usar el menú **Graph**.

Una utilidad importante de todos los gráficos creados a través del menú **Graph** es que haciendo *clic* sobre ellos con el botón derecho del ratón y activando la opción **Update Graph Automatically** del menú contextual que aparece, el gráfico cambia automáticamente al modificar los datos con que se han construido (ya sea añadiendo, modificando o eliminando datos).

2.2.1. Gráfico de sectores o de *pastel*

El *gráfico de sectores* se construye de la siguiente forma: se divide el área de un círculo en sectores circulares de ángulos proporcionales a las frecuencias absolutas de las clases. Se utiliza cuando la variable es cualitativa o cuantitativa discreta con pocos resultados distintos.

En *Minitab*, este gráfico se obtiene con la opción **Graph⇒Pie Chart**.

Por ejemplo, vamos a hacer el gráfico de sectores de los datos de la columna **Activity** de la hoja de datos **Pulse.mtw**. Para ello, en el cuadro de diálogo que resulta al seleccionar **Graph⇒Pie Chart**, dejamos activada la opción **Chart counts of unique values** y seleccionamos la columna '**Activity**' en el recuadro **Categorical variables**. Podemos cambiar el aspecto que tendría el gráfico por defecto, pulsando en los botones que aparecen en este cuadro de diálogo: **Pie Options**, **Labels**, **Multiple Graphs** y **Data Options**. En principio, podríamos dejar todas las opciones por defecto a la hora de realizar este primer diagrama de sectores.

El gráfico obtenido podemos copiarlo en el portapapeles, haciendo *clic* sobre el gráfico con el botón derecho del ratón y seleccionando, del menú contextual que resulta, la opción **Copy Graph**. De esta manera, podríamos pegarlo en otro programa bajo Windows, por ejemplo, uno de edición de gráficos. También podemos almacenarlo en la ventana de proyecto, **Project Manager** (concretamente en el directorio **ReportPad**) haciendo *clic* sobre el gráfico con el botón derecho del ratón y seleccionando, del menú contextual que resulta, la opción **Append Graph to Report**. También tenemos la posibilidad de grabarlo en varios formatos (gráfico propio de Minitab, **mgf**, **jpg**, **png**, **bmp**, etc.). Para ello solo tenemos que cerrar el gráfico (botón ☐) y pulsar en **Sí** cuando *Minitab* nos pregunte si queremos guardar el gráfico en un archivo aparte.

Una vez obtenido el gráfico es posible cambiar su aspecto. Para ello, hacemos doble *clic* sobre la parte del gráfico que queremos cambiar. Aparece, entonces, una nueva ventana que nos permite hacer dicha transformación. Para practicar, vamos a cambiar el gráfico de sectores de los datos de la columna **Activity** de la siguiente manera:

- Que el título sea *Gráfico de sectores de la variable 'Actividad Física'*, en letra Verdana, cursiva, negrita, de color rojo oscuro y con un tamaño de 10 puntos.
- Que junto a los sectores circulares aparezca la frecuencia absoluta de cada categoría (*clic* sobre uno de los sectores circulares con el botón derecho del ratón; opción **Add, Slice Labels**; activamos **Frequency** y pulsamos en **OK**).

Vamos a aprender a hacer un diagrama de sectores cuando tenemos en una columna las categorías de una variable y en otra columna las frecuencias absolutas de dichas categorías. Por ejemplo, vamos a realizar el diagrama de sectores de los datos de la Figura 7, correspondientes a los idiomas en que están escritos los libros de los estantes de una determinada biblioteca.

	Idioma	Nº de estantes
1	francés	78
2	alemán	47
3	ruso	20
4	español	30

Figura 7: Idioma de los libros de una biblioteca

Como estos datos no tienen nada que ver con los datos del archivo **Pulse.mtw**, abrimos una nueva hoja de datos con la opción **File⇒New**. En el cuadro de diálogo que aparece seleccionamos **Minitab Worksheet**. A esta nueva hoja de datos **Minitab** le asignará el nombre **Worksheet J**, siendo *J* un número natural. A continuación introducimos los datos tal como se muestra en la Figura 7. Luego guardamos esta hoja de datos con el nombre **IdiomaLibros.mtw** (**File⇒Save Current Worksheet As**). Para dibujar el diagrama de sectores seleccionamos **Graph⇒Pie Chart**. En el cuadro de diálogo resultante, activamos la opción **Chart values from a table**; seleccionamos la columna '**Idioma**' en el recuadro **Categorical Variable**; seleccionamos la columna '**Nº de estantes**' en el recuadro **Summary variables** y pulsamos en **OK**. Como ya sabemos, podemos modificar este gráfico.

2.2.2. Diagrama de barras

2.2.2.1. Diagrama de barras simple

El *diagrama de barras* se construye de la siguiente manera: se sitúan en el eje horizontal las clases y sobre cada una de ellas se levanta un segmento rectilíneo (o un rectángulo) de altura igual a la frecuencia (absoluta o relativa) o al porcentaje de cada clase. Se utiliza cuando la variable es cualitativa o cuantitativa discreta con pocos resultados distintos.

En **Minitab** este gráfico se obtiene con la opción **Graph⇒Bar Chart**.

Por ejemplo, vamos a hacer el diagrama de barras de los datos de la columna **Activity** de la hoja de datos **Pulse.mtw**. En primer lugar, es necesario tener abierta y activada dicha hoja de datos. Para dibujar el diagrama de barras seleccionamos **Graph⇒Bar Chart**; dejamos activada la opción **Counts of unique values** del recuadro **Bars represent** y dejamos también activado el modelo **Simple** del diagrama de barras. En el cuadro de diálogo resultante, seleccionamos la columna '**Activity**' en el recuadro **Categorical Variables**. Como las categorías son números concretos (0, 1, 2 y 3) es más riguroso que, en vez de barras, aparezcan solamente segmentos rectilíneos; por tanto, activamos el botón **Data View** y en el cuadro de diálogo resultante activamos solo la opción **Project lines**.

Igual que ocurría con los gráficos anteriores, una vez obtenido el diagrama de barras podemos copiarlo en el portapapeles, o almacenarlo en el apartado **ReportPad** de la ventana **Project Manager**, o grabarlo en un archivo aparte. Podemos observar, además, que si hacemos *clic* sobre el gráfico (para activarlo) y luego pasamos el ratón por encima de las barras, se nos indica la frecuencia absoluta de cada categoría.

También es posible cambiar su aspecto, una vez obtenido, haciendo doble *clic* sobre la parte del gráfico que queremos cambiar. Para practicar, vamos a modificar diagrama de barras de los datos de la columna **Activity** de la siguiente manera:

- Que el título sea *Diagrama de barras de la variable ‘Actividad Física’*, en letra Comic Sans MS, cursiva, negrita, de color rojo y con un tamaño de 11 puntos.
- Que las barras (líneas) sean de color rojo y de un tamaño (grosor) de 3 puntos.
- Que en el eje vertical se muestren 13 marcas (*ticks*), en letra Arial, no negrita, de color rojo y con un tamaño de 10 puntos.
- Que el texto del eje vertical sea *Frecuencia absoluta*, en letra Arial, cursiva, no negrita, de color rojo y con un tamaño de 9 puntos.
- Que el texto del eje horizontal sea *Actividad Física (0=Ninguna, 1=Baja, 2=Media, 3=Alta)*, en letra Arial, cursiva, no negrita, de color rojo y con un tamaño de 8 puntos.
- Que en la parte superior de cada barra aparezca la frecuencia absoluta de cada categoría (*clic* sobre una de las barras con el botón derecho del ratón, opción **Add, Data Labels**, dejar activado **Use y-values labels**).

Vamos a aprender a hacer un diagrama de barras cuando tenemos en una columna las categorías de una variable y en otra columna las frecuencias absolutas de dichas categorías. Por ejemplo, vamos a realizar el diagrama de barras de los datos de la Figura 7, correspondientes a los idiomas en que están escritos los libros de los estantes de una determinada biblioteca. En primer lugar, es necesario tener abierta y activada dicha hoja de datos (**IdiomaLibros.mtw**). Para dibujar el diagrama de barras seleccionamos **Graph⇒Bar Chart**, activamos la opción **Values from a table** del apartado **Bars represent**; activamos el modelo **Simple** del apartado **One column of values** y pulsamos en **OK**. En el cuadro de diálogo resultante, seleccionamos la columna ‘**Nº de estantes**’ en el recuadro **Graph variables**; seleccionamos la columna ‘**Idioma**’ en el recuadro **Categorical Variable** y pulsamos en **OK**. Como ya sabemos, podemos modificar este gráfico.

2.2.2.2. Diagrama de barras agrupado (o apilado)

Con la opción **Graph⇒Bar Chart** existe la posibilidad de seleccionar una nueva variable para determinar las barras dentro de cada grupo; esto se realiza seleccionando **Cluster** (para un diagrama de barras agrupado según los resultados de otra variable) o **Stack** (para un diagrama de barras apilado según los resultados de otra variable).

Por ejemplo, con el archivo de datos **Pulse.mtw** vamos a hacer el diagrama de barras de los datos de la columna **Activity** en grupos definidos por la variable **Sex**. En primer lugar, es necesario tener abierta y activada dicha hoja de datos. Para dibujar el citado diagrama de barras seleccionamos **Graph⇒Bar Chart**; dejamos activada la opción **Counts of unique values** del recuadro **Bars represent**; y activamos el modelo **Cluster** del diagrama de barras. En el siguiente cuadro de diálogo seleccionamos, de la lista de variables de la izquierda, las columnas ‘**Activity**’ y ‘**Sex**’ (en este orden) para ponerlas en el recuadro **Categorical variables**. Una vez obtenido dicho diagrama de barras es conveniente modificarlo para que sea más explicativo; por ejemplo, vamos a hacer lo siguiente:

- Que el título sea *Diagrama de barras de la variable ‘Actividad Física’ en grupos definidos por la variable ‘Sexo’*, en letra Verdana, negrita, de color morado y con un tamaño de 9 puntos.

- Que las barras tengan distinto color según los resultados de la variable **Sex** y que aparezca una leyenda explicativa (doble *clic* sobre una de las barras, en el cuadro de diálogo resultante seleccionamos la carpeta **Groups**, en el recuadro **Assign attributes by categorical variables** seleccionamos la variable **Sex** y pulsamos en **OK**).
- Que en el eje vertical se muestren 10 marcas (*ticks*), en letra Verdana, no negrita, de color morado y con un tamaño de 10 puntos.
- Que el texto del eje vertical sea *Frecuencia absoluta*, en letra Verdana, no negrita, de color morado y con un tamaño de 11 puntos.
- Que en el eje horizontal todo esté escrito con la fuente Verdana, no negrita, de color morado y con un tamaño de 9 puntos. Que en dicho eje aparezcan los nombres de las variables en español: *Actividad Física* en vez de *Activity*, y *Sexo* en vez de *Sex*. Que en el mismo eje los resultados de la variable *Sex* no sean 1 y 2 sino *Hombre* y *Mujer*. Y los resultados de la variable *Activity* no sean 0, 1, 2 y 3 sino *Ninguna*, *Poca*, *Media* y *Alta*.

Vamos a aprender a hacer un diagrama de barras agrupado (o apilado) cuando tenemos los datos en una tabla de doble entrada. Por ejemplo, vamos a realizar el diagrama de barras agrupado de los datos de la Figura 8, correspondientes al número de citas en diferentes campos de investigación y en tres distintos años.

	Campo investigación	1970	1980	1990
1	sociología	330	414	547
2	economía	299	393	295
3	política	115	357	137
4	psicología	329	452	258

Figura 8: Citas anuales en distintos campos de investigación

En primer lugar, abrimos una nueva hoja de datos con la opción **File⇒New**. En el cuadro de diálogo que aparece seleccionamos **Minitab Worksheet**. A continuación introducimos los datos tal como se muestra en la Figura 8. Luego guardamos esta hoja de datos con el nombre **Citas.mtw**. Para dibujar el diagrama de barras agrupado seleccionamos **Graph⇒Bar Chart**, activamos la opción **Values from a table** del apartado **Bars represent**; activamos el modelo **Cluster** del apartado **Two-way table** y pulsamos en **OK**. En el cuadro de diálogo resultante, seleccionamos las columnas '1970', '1980' y '1990' en el recuadro **Graph variables**; seleccionamos la columna 'Campo investigación' en el recuadro **Row labels**; activamos **Rows are outermost categories and columns are innermost** y, por último, pulsamos en **OK**. Como ya sabemos, podemos modificar este gráfico.

2.2.3. Histograma

El *histograma* se construye de la siguiente manera: se sitúan en el eje horizontal los intervalos de clase y sobre cada uno se levanta un rectángulo de área igual o proporcional a la frecuencia absoluta.

En **Minitab** se puede obtener el histograma de una variable con la opción **Graph⇒Histogram**. Esta opción ofrece 4 tipos: **Simple**, **With Fit**, **With Outline and Groups** y **With Fit and Groups**.

Por ejemplo, podemos hacer el histograma de la variable **Weight** de la hoja de datos **Pulse.mtw**. En primer lugar, es necesario tener abierta y activada dicha hoja de datos. Para realizar el citado histograma seleccionamos la opción **Graph⇒Histogram**. De las cuatro opciones que aparecen seleccionamos

Simple. En el cuadro de diálogo resultante seleccionamos la variable '**Weight**' para ponerla en el recuadro **Graph variables**. Podemos cambiar el aspecto que tendría el gráfico por defecto, pulsando en los botones que aparecen en este cuadro de diálogo: **Scale**, **Labels**, **Data View**, **Multiple Graphs** y **Data Options**. En principio, podríamos dejar todas las opciones por defecto a la hora de realizar este primer histograma.

Una vez obtenido el histograma podemos copiarlo en el portapapeles, o almacenarlo en el directorio **ReportPad** de la ventana **Project Manager**, o grabarlo en un archivo aparte. También es posible cambiar su aspecto una vez obtenido. Para ello, hacemos doble *clic* sobre la parte del gráfico que queremos cambiar. Aparece, entonces, una nueva ventana que nos permite hacer dicha transformación. Los cambios más usuales son: cambio en la escala del eje horizontal, cambio en el eje vertical, aspecto de las barras, intervalos sobre los que se sitúan las barras, aspecto de la ventana del gráfico y cambio en las proporciones del gráfico. Para practicar con estas opciones, vamos a cambiar el histograma de la variable **Weight** de la siguiente manera:

- Que el título sea *Histograma de la variable 'Peso'*, en letra Arial, cursiva, negrita, de color azul oscuro y con un tamaño de 10 puntos.
- Que las barras sean de color azul claro con una trama de relleno oblicua y con los bordes de color azul oscuro.
- Que haya 7 intervalos de la misma amplitud y que en el eje horizontal aparezcan los límites de los intervalos (no los puntos medios).
- Que el texto del eje horizontal sea *Peso de los alumnos, en libras*, en letra Arial, cursiva, no negrita, de color azul oscuro y con un tamaño de 9 puntos.
- Que en el eje vertical se muestren 13 marcas (*ticks*), en letra Arial, de color azul oscuro y con un tamaño de 8 puntos.
- Que el texto del eje vertical sea *Frecuencia absoluta*, en letra Arial, cursiva, no negrita, de color azul oscuro y con un tamaño de 9 puntos.

2.3. Medidas descriptivas de los datos

La opción **Stat⇒Basic Statistics⇒Display Descriptive Statistics** de **Minitab** permite obtener los estadísticos más importantes de las columnas (variables) de la hoja de datos. También permite calcularlos separando los valores de una columna según el valor de otra. Además puede realizar una serie de gráficas que nos permiten resumir la información contenida en los datos.

Para practicar esta opción, vamos a calcular los estadísticos descriptivos más importantes de las variables **Pulse1**, **Height** y **Weight** de la hoja de datos **Pulse.mtw**. Para ello, seleccionamos **Stat⇒Basic Statistics⇒Display Descriptive Statistics** y en el recuadro **Variables** del cuadro de diálogo resultante seleccionamos, de la lista de columnas que tenemos a la izquierda, las tres variables '**Pulse1**', '**Height**' y '**Weight**'. A continuación pulsamos en **Statistics**. Nos aparece un nuevo cuadro de diálogo en el cual se pueden elegir los estadísticos que queremos determinar de las variables que hemos seleccionado en el recuadro **Variables**. Haciendo *clic* sobre el botón **Help** se obtiene información sobre el significado de cada uno de estos estadísticos. Los estadísticos que podemos seleccionar son los siguientes:

Mean	media aritmética	$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$
SE of mean	error estándar de la media	$\frac{S_x}{\sqrt{n}}$
Standard deviation	desviación típica corregida	$S_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$
Variance	varianza corregida	S_x^2
Coefficient of variation	coeficiente de variación media	$CV = \frac{S_x}{ \bar{x} } \cdot 100 \%$
First quartile	primer cuartil	Q_1
Median	mediana	$M_e = Q_2$
Third quartile	tercer cuartil	Q_3
Interquartile range	recorrido intercuartílico	$R_I = Q_3 - Q_1$
Trimmed mean	media de los datos eliminando el 5 % de los menores y el 5 % de los mayores	
Sum	suma	$\sum_{i=1}^n x_i$
Minimum	mínimo dato	x_{min}
Maximum	máximo dato	x_{max}
Range	recorrido o rango	$R = x_{max} - x_{min}$
N nonmissing	número de casos para los cuales sabemos el resultado de la variable = n	
N missing	número de casos para los cuales no sabemos el resultado de la variable	
N total	número total de casos = N nonmissing + N missing	
Cumulative N	número acumulado de casos (solo cuando se ha rellenado el recuadro By variables)	
Percent	porcentaje de casos (solo cuando se ha rellenado el recuadro By variables)	
Cumulative percent	porcentaje acumulado de casos (solo cuando se ha rellenado el recuadro By variables)	
Sum of squares	suma de cuadrados	$\sum_{i=1}^n x_i^2$
Skewness	coeficiente de asimetría	$g_1 = \frac{m_3}{s_x^3}, \text{ con } m_3 = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{n}$
Kurtosis	coeficiente de apuntamiento	$g_2 = \frac{m_4}{s_x^4} - 3, \text{ con } m_4 = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{n}$

MSSD	media de los cuadrados de las sucesivas diferencias
------	---

Siguiendo con nuestro ejemplo (cálculo de los estadísticos más importantes de las variables **Pulse1**, **Height** y **Weight**), podemos seleccionar todos los estadísticos menos **Cumulative N**, **Percent** y **Cumulative percent**. En la ventana de sesión podemos comprobar, por ejemplo, que la suma de los datos de la variable **Pulse1** es 6704 y la suma de los cuadrados de los datos de la misma variable es 499546.

Con la misma hoja de datos (**Pulse.mtw**) podemos calcular los estadísticos de la variable **Pulse2** (Pulso después de correr) separando sus resultados según los valores de la variable **Ran** (¿corrió o no corrió?). Para ello, seleccionamos **Stat**⇒**Basic Statistics**⇒**Display Descriptive Statistics**; en el recuadro **Variables** del cuadro de diálogo resultante seleccionamos la variable '**Pulse2**'; y en **By variables (Optional)** seleccionamos la variable '**Ran**'. En consecuencia, en la ventana de sesión aparecen los resultados de los mencionados estadísticos de la variable **Pulse2** separados para cada grupo de resultados de la variable **Ran**. Por ejemplo, podemos comprobar que para el grupo de personas que sí corrió (**Ran=1**) la media del pulso es 92'51 y la mediana es 88, mientras que para el grupo de personas que no corrió (**Ran=2**) la media del pulso es 72'32 y la mediana es 70.

2.4. Correlación y regresión lineal

2.4.1. Diagrama de dispersión o nube de puntos

Con **Minitab** el diagrama de dispersión se obtiene con la opción **Graph**⇒**Scatterplot**.

Por ejemplo, con la hoja de datos **Pulse.mtw** podemos dibujar el diagrama de dispersión, con la recta de regresión superpuesta, de la altura en pulgadas, **Height**, sobre el peso en libras, **Weight**. En primer lugar, es necesario tener abierta y activada dicha hoja de datos. En segundo lugar, seleccionamos la opción **Graph**⇒**Scatterplot**; en el cuadro de diálogo que aparece seleccionamos **With Regression** y pulsamos en **OK**. En el siguiente cuadro de diálogo, en el recuadro **Y Variables** seleccionamos, de la lista de variables de la izquierda, la columna '**Height**'; y en el recuadro **X Variables** seleccionamos, de la lista de variables de la izquierda, la columna '**Weight**'. Podemos cambiar el aspecto que tendría el gráfico por defecto, pulsando en los botones que aparecen en este cuadro de diálogo: **Scale**, **Labels**, **Data View**, **Multiple Graphs** y **Data Options**. En principio, podríamos dejar todas las opciones por defecto a la hora de realizar este primer diagrama de dispersión. Se puede comprobar que el diagrama de dispersión o *nube de puntos* se agrupa cerca de una línea recta, lo que significa que hay una relación lineal fuerte entre las dos variables.

Igual que ocurría con los gráficos anteriores, una vez obtenido el diagrama de dispersión se puede copiar en el portapapeles, o almacenar en el apartado **ReportPad** de la ventana **Project Manager**, o grabar en un archivo aparte. También es posible cambiar su aspecto, una vez obtenido, haciendo doble *clic* sobre la parte del gráfico que queremos modificar. Para practicar, vamos a modificar el diagrama de dispersión anterior de la siguiente manera:

- Que el título sea *Diagrama de dispersión de la 'Altura' frente al 'Peso'*, en letra Times New Roman, cursiva, negrita, de color rojo y con un tamaño de 14 puntos.
- Que los símbolos sean rombos rojos de tamaño 1.

- Que en el eje horizontal se muestren 14 marcas (*ticks*), en letra Times New Roman, no negrita, de color rojo y con un tamaño de 12 puntos.
- Que el texto del eje horizontal sea *Peso de los alumnos, en libras*, en letra Times New Roman, cursiva, no negrita, de color rojo y con un tamaño de 12 puntos.
- Que en el eje vertical se muestren 10 marcas (*ticks*), en letra Times New Roman, no negrita, de color rojo y con un tamaño de 12 puntos.
- Que el texto del eje vertical sea *Altura de los alumnos, en pulgadas*, en letra Times New Roman, cursiva, no negrita, de color rojo y con un tamaño de 12 puntos.
- Que la recta de regresión sea de color rojo y de tamaño 2.

2.4.2. Coeficiente de correlación lineal

Con *Minitab* el coeficiente de correlación lineal de Pearson se obtiene con la opción **Stat⇒Basic Statistics⇒Correlation**.

Por ejemplo, de la hoja de datos **Pulse.mtw** vamos a calcular el coeficiente de correlación lineal de Pearson entre cada par de variables de las siguientes: **Pulse1**, **Height** y **Weight**. Para ello, seleccionamos **Stat⇒Basic Statistics⇒Correlation**. En el cuadro de diálogo resultante activamos el recuadro **Variables** y seleccionamos, de la lista de variables de la izquierda, las columnas **Pulse1**, **Height** y **Weight**; desactivamos la opción **Display p-values**; dejamos desactivada la opción **Store matrix (display nothing)** y pulsamos en **OK**. Podemos comprobar, en la ventana de sesión, que el coeficiente de correlación lineal entre las variables **Pulse1** y **Height** es igual a $-0,212$ (por tanto, la fuerza de la relación lineal entre estas dos variables es muy débil); el coeficiente de correlación lineal entre las variables **Pulse1** y **Weight** es igual a $-0,202$ (por tanto, la fuerza de la relación lineal entre estas dos variables es muy débil); y el coeficiente de correlación lineal entre las variables **Height** y **Weight** es igual a $0,785$ (por tanto, la fuerza de la relación lineal entre estas dos variables es fuerte; consecuencia que ya habíamos extraído al realizar el diagrama de dispersión de **Height** sobre **Weight**).

2.4.3. Rectas de regresión

Para obtener la ecuación de la recta de regresión (mínimo cuadrática) de una variable cuantitativa Y sobre otra variable cuantitativa X , se selecciona la opción **Stat⇒Regression⇒Regression**.

Puesto que sabemos que la fuerza de la relación lineal entre las variables **Height** y **Weight** es fuerte, vamos a encontrar la ecuación de la recta de regresión de la variable **Weight** sobre la variable **Height** (de la hoja de datos **Pulse.mtw**). Para ello, seleccionamos la opción **Stat⇒Regression⇒Regression**; en el cuadro de diálogo resultante seleccionamos la variable '**Weight**' en **Response** y la variable '**Height**' en **Predictors**; pulsamos en **Results** y, en el cuadro de diálogo resultante, activamos la opción **Regression equation, table of coefficients, s, R-squared, and basic analysis of variance** y pulsamos en **OK**; en el siguiente cuadro de diálogo volvemos a pulsar en **OK**. En la ventana de sesión aparecen varios resultados, la mayoría de los cuales no pueden ser interpretados en este momento pues todavía no hemos explicado la parte de Estadística Inferencial. Lo que a nosotros nos interesa en este momento son los resultados de los coeficientes de regresión, que son: $A = -204'74$, $B = 5'0918$, siendo la ecuación de la recta de regresión $Y = A + B X$; donde $Y = \text{Weight}$ (peso) y $X = \text{Height}$ (altura). Es decir, la ecuación de la recta de regresión de la variable **Weight** sobre la variable **Height** es:

$$\text{Weight} = -204'74 + 5'0918 \cdot \text{Height}$$

2.5. Ejercicios propuestos

Ejercicio 2.1

- a) Crea un nuevo proyecto de *Minitab*.
- b) Abre la hoja de datos **Prestamos.mtw** (datos del Ejercicio 1.1).
- c) Determina la distribución de frecuencias de la variable **Intervalos PPU**.
- d) Para las variables **Usuarios**, **Préstamos** y **PPU** calcula todas las medidas descriptivas que hemos estudiado en las clases teóricas.
- e) Dibuja el diagrama de dispersión, con la recta de regresión superpuesta, de la variable **Préstamos** sobre la variable **Usuarios**. Modifícalo de la siguiente forma:
 - Que el título sea *Diagrama de dispersión del 'Nº anual de préstamos' frente al 'Nº anual de usuarios'* en letra Verdana, itálica, negrita, de color rojo y con un tamaño de 9 puntos.
 - Que los símbolos sean cuadrados rellenos, de color verde oscuro y de tamaño 2.
 - Que en el eje horizontal se muestren 20 marcas (*ticks*) y que los números sean de color azul y con un tamaño de 8 puntos.
 - Que el texto del eje horizontal sea *Número anual de usuarios*, en letra Verdana, itálica, no negrita, de color rojo y con un tamaño de 11 puntos.
 - Que en el eje vertical se muestren 18 marcas (*ticks*) y que los números sean de color azul y de un tamaño de 8 puntos.
 - Que el texto del eje vertical sea *Número anual de préstamos*, en letra Verdana, itálica, no negrita, de color rojo y con un tamaño de 11 puntos.
 - Que la recta de regresión sea de color rojo y de tamaño 2.
- f) Calcula el coeficiente de correlación lineal entre las variables **Préstamos** y **Usuarios**.
- g) Determina la ecuación de la recta de regresión de la variable **Préstamos** sobre la variable **Usuarios**.
- h) Dibuja el histograma simple de la variable **PPU**. Modifícalo de la siguiente forma:
 - Que haya 4 intervalos de la misma amplitud y que en el eje horizontal aparezcan los límites de los intervalos (no los puntos medios).
 - Que el título sea *Histograma del 'Porcentaje anual de préstamos por usuario'*, en letra Times New Roman, negrita, de color rojo oscuro y con un tamaño de 14 puntos.
 - Que las barras sean de color rojo claro con una trama de relleno horizontal y con los bordes de color rojo oscuro, de tamaño 2.
 - Que el texto del eje horizontal sea *Porcentaje anual de préstamos por usuario*, en letra Times New Roman, cursiva, no negrita, de color rojo oscuro y con un tamaño de 12 puntos.
 - Que en el eje vertical se muestren 7 marcas (*ticks*) y que los números sean de color rojo oscuro y con un tamaño de 12 puntos.
 - Que el texto del eje vertical sea *Frecuencia absoluta*, en letra Times New Roman, cursiva, no negrita, de color rojo oscuro y con un tamaño de 12 puntos.
- i) Dibuja el gráfico de sectores de la variable **Intervalos PPU**. Modifícalo de la siguiente forma:
 - Que el título sea *Gráfico de sectores de la variable 'Intervalos PPU'*, en letra Verdana, cursiva, negrita, de color azul oscuro y con un tamaño de 12 puntos.

- Que junto a los sectores circulares aparezca la frecuencia absoluta y el porcentaje de cada categoría.
 - En la leyenda, tanto la fuente de la cabecera como la fuente del cuerpo sea Verdana, de color azul oscuro y con un tamaño de 10 puntos.
- j) Graba el proyecto con el siguiente nombre: **Ejercicio2-1.mpj**

Ejercicio 2.2

- a) Crea un nuevo proyecto de *Minitab*.
- b) Abre la hoja de datos **Transacciones.mtw** (datos del Ejercicio 1.2).
- c) Determina la distribución de frecuencias de la variable **Intervalos Porcentaje TRF**.
- d) Para las variables **TR**, **TRF** y **Porcentaje TRF** calcula las medidas descriptivas siguientes: mínimo, primer cuartil, mediana, tercer cuartil, máximo, recorrido, recorrido intercuartílico, media, varianza corregida, desviación típica corregida, suma de los datos y suma de los cuadrados de los datos.
- e) Calcula la media, la mediana y la desviación típica corregida de la variable **Porcentaje TRF** separando sus resultados según los valores de la variable **Tipo Biblioteca**.
- f) Dibuja el diagrama de dispersión, con la recta de regresión superpuesta, de la variable **TRF** sobre la variable **TR**. Modifícalo de la siguiente forma:
- Que el título sea *Nube de puntos y recta de regresión* en letra Verdana, negrita, de color azul y con un tamaño de 12 puntos.
 - Que los símbolos sean triángulos rellenos, de color magenta y de tamaño 1.
 - Que en el eje horizontal se muestren 10 marcas (*ticks*) y que los números sean de color azul y de un tamaño de 9 puntos.
 - Que el texto del eje horizontal sea *Número anual de transacciones de referencia*, en letra Verdana, itálica, no negrita, de color azul y con un tamaño de 10 puntos.
 - Que en el eje vertical se muestren 10 marcas (*ticks*) y que los números sean de color azul y de un tamaño de 9 puntos.
 - Que el texto del eje vertical sea *Número anual de transacciones de referencia finalizadas*, en letra Verdana, itálica, no negrita, de color azul y con un tamaño de 9 puntos.
 - Que la recta de regresión sea de color morado y de tamaño 2.
- g) Calcula el coeficiente de correlación lineal entre las variables **TR** y **TRF**.
- h) Determina la ecuación de la recta de regresión de la variable **TRF** sobre la variable **TR**.
- i) Dibuja el diagrama de barras de la variable **Intervalos Porcentaje TRF** en grupos definidos por la variable **Tipo Biblioteca**. Modifícalo de la siguiente forma:
- Que las barras tengan distinto color según los resultados de la variable **Tipo Biblioteca** y que aparezca una leyenda explicativa.
 - Que el título sea *Diagrama de barras agrupado*, escrito con letra Arial, negrita, de color rojo oscuro y con un tamaño de 16 puntos.
 - Que el texto del eje vertical sea *Frecuencia absoluta*, escrito con letra Arial, negrita, de color rojo oscuro y con un tamaño de 12 puntos.
 - Que en el eje horizontal todo esté escrito con la fuente Arial, de color rojo oscuro y con un tamaño de 10 puntos.
- j) Graba el proyecto con el siguiente nombre: **Ejercicio2-2.mpj**

Ejercicio 2.3 El gasto de una biblioteca, en euros, durante un año determinado, es:

Gasto en personal	6570
Gasto en libros	3450
Otros gastos	2380

- Crea un nuevo proyecto de *Minitab*.
- Guarda los datos en el archivo **GastoBiblioteca.mtw**
- Haz un diagrama de barras y modifícalo a tu gusto.
- Haz un gráfico de sectores y modifícalo a tu gusto.
- Graba el proyecto con el siguiente nombre: **Ejercicio2-3.mpj**

Ejercicio 2.4 La estadística de fotocopias de 4 bibliotecas (A, B, C y D), durante un año, está recogida en la siguiente tabla:

	A	B	C	D
Reproducción de catálogos	16110	3640	0	3400
Trabajo del personal de la biblioteca	63350	11360	3080	5500
Préstamo interbibliotecario	2600	1090	560	250
Copias para usuarios de la biblioteca	43540	58040	1980	0

- Crea un nuevo proyecto de *Minitab*.
- Guarda los datos en el archivo **TipoFotocopias.mtw**
- Haz un diagrama de barras agrupado y modifícalo a tu gusto.
- Graba el proyecto con el siguiente nombre: **Ejercicio2-4.mpj**

Ejercicio 2.5 El número de descriptores (*keywords*) de 72 artículos de investigación viene dado por:

Nº de descriptores	3	4	5	6	7	8	9	10	11	12	13	14
Nº de artículos	5	8	12	7	9	9	10	5	3	2	1	1

- Crea un nuevo proyecto de *Minitab*.
- Guarda los datos en el archivo **Keywords.mtw**
- Haz un diagrama de barras en el cual las barras sean segmentos rectilíneos. Modifícalo a tu gusto.
- Graba el proyecto con el siguiente nombre: **Ejercicio2-5.mpj**

3

Modelos de probabilidad

3.1. Muestras aleatorias de las distribuciones usuales

En *Minitab* podemos generar datos de distribuciones usuales utilizando la opción **Calc⇒Random Data**. Esta opción permite generar una muestra de datos de cualquier columna de la hoja de datos actualmente abierta o de una de las distribuciones de probabilidad que aparecen listadas.

En primer lugar, creamos una nueva hoja de datos con la opción **File⇒New**. En el cuadro de diálogo que aparece seleccionamos **Minitab Worksheet**. A esta nueva hoja de datos *Minitab* le asignará el nombre **Worksheet J**, siendo *J* un número natural. Luego podremos cambiarle el nombre (por ejemplo, **Probabilidad.mtw**) con la opción **File⇒Save Current Worksheet As**. A continuación, vamos a crear una columna, en dicha hoja de datos, que lleve por nombre **‘1000 datos de N(5,2)’** y que contenga 1000 datos aleatorios procedentes de una distribución $\mathcal{N}(5, 2)$ (Normal de media 5 y desviación típica 2). Para ello, seleccionamos **Calc⇒Random Data⇒Normal**; en **Number of rows of data to generate** tecleamos **1000**; en **Store in column** tecleamos el nombre **‘1000 datos de N(5,2)’**; en **Mean** tecleamos **5** y en **Standard deviation** ponemos un **2**.

A continuación vamos a hacer el histograma, con la curva Normal superpuesta, de la muestra aleatoria obtenida en la columna **‘1000 datos de N(5,2)’**. Para ello, recordemos que hay que seleccionar la opción **Graph⇒Histogram**. En el cuadro de diálogo resultante elegimos **With Fit**. En el siguiente cuadro de diálogo, en **Graph variables** seleccionamos, de la lista de variables que tenemos a la izquierda, la columna **‘1000 datos de N(5,2)’** y pulsamos en **OK**. En la representación gráfica podemos apreciar que el histograma está cerca de la curva Normal superpuesta, lo cual es lógico puesto que hemos creado una muestra de una distribución Normal. También podemos ver, en la leyenda que aparece en la parte superior derecha del gráfico, que la media de la muestra obtenida se aproxima a 5 y la desviación típica se aproxima a 2.

3.2. Función de densidad y función de probabilidad

Minitab puede calcular el resultado de la función de densidad (cuando la distribución es continua) o de la función de probabilidad (cuando la distribución es discreta) para un valor concreto o para una lista de valores. Para ello hay que elegir la opción **Calc**⇒**Probability Distributions** y a continuación el nombre de la variable aleatoria: Chi-square (chi-cuadrado de Pearson), Normal, F (de Snedecor), t (de Student), etc.

Dentro del cuadro de diálogo que aparecerá hay que seleccionar **Probability Density** (para las distribuciones continuas) o **Probability** (para las distribuciones discretas).

Para entender mejor el interés de esta opción, vamos a determinar los resultados de la función de densidad de una distribución $\mathcal{N}(0, 1)$ (Normal Estándar) para una lista de valores que vamos a crear (todos los números comprendidos entre -4 y 4, con un incremento de 0,01). Luego haremos la representación gráfica de esta función de densidad. Para ello se procede de la siguiente manera:

- a) Mediante la opción **Calc**⇒**Make Patterned Data**⇒**Simple Set of Numbers** crearemos una nueva columna que denominaremos 'x de -4 a 4' y que contendrá todos los números comprendidos entre el -4 y el 4 con un incremento de 0,01. Podemos comprobar que en la columna 'x de -4 a 4' hay 801 números.
- b) En otra columna se calculan los resultados de la función de densidad de la variable aleatoria Normal Estándar para cada valor de la columna 'x de -4 a 4'. Para hacerlo, se selecciona **Calc**⇒**Probability Distributions**⇒**Normal**; se activa **Probability density**; en **Mean** y en **Standard deviation** se deja lo que aparece por defecto (cero y uno, respectivamente); en **Input column** se selecciona, de la lista de variables de la izquierda, la columna 'x de -4 a 4' y en **Optional storage** se teclea el nombre de la columna que contendrá los resultados de la función de densidad; por ejemplo, 'f(x) N(0,1)'.
 - c) Finalmente, para representar gráficamente la función de densidad de la variable aleatoria Normal Estándar se elige la opción **Graph**⇒**Scatterplot**, después se elige **With connect line**. En el siguiente cuadro de diálogo, en **Y variables** se selecciona, de la lista de variables de la izquierda, la columna 'f(x) N(0,1)' y en **X variables** se selecciona la columna 'x de -4 a 4'. Sería conveniente quitar los puntos del gráfico, dejando sólo la línea de conexión, para lo cual se hace doble *clic* sobre la curva, en **Attributes**⇒**Symbols** se marca la opción **Custom** y en **Type** se selecciona **None** (buscando hacia arriba). Luego se hace un *clic* dentro del gráfico, pero no sobre la curva.

3.3. Función de distribución

Para calcular el resultado de la función de distribución de una variable aleatoria X , $F(t) = P(X \leq t)$, hay que elegir la opción **Calc**⇒**Probability Distributions** y a continuación el nombre de la variable aleatoria. Dentro del cuadro de diálogo que aparece hay que seleccionar **Cumulative Probability**.

Por ejemplo, vamos a calcular la probabilidad $P(X \leq -1'36)$, siendo X una variable aleatoria Normal Estándar. Como $P(X \leq -1'36) = F(-1'36)$, para calcular su resultado seleccionamos la opción **Calc**⇒**Probability Distributions**⇒**Normal**; activamos **Cumulative Probability**; en **Mean** y en **Standard deviation** dejamos lo que aparece por defecto (cero y uno, respectivamente). No activamos la opción **Input column** sino la opción **Input constant**, en donde colocamos el valor -1,36. Podemos almacenar el resultado en una constante tecleando en el recuadro **Optional storage** una K seguida de un número o poniendo un nombre a dicho resultado. Nosotros no vamos a rellenar el recuadro **Optional storage**, por

lo que el resultado aparecerá en la ventana de sesión. Se puede comprobar que la probabilidad pedida es $P(X \leq -1'36) = F(-1'36) = 0'086915$.

Si queremos calcular probabilidades de los tipos $P(X > a)$, $P(a < X < b)$, etc., tenemos que utilizar lápiz y papel, y aplicar las propiedades de la probabilidad para llegar a expresiones en las que sólo aparezcan probabilidades del tipo $P(X \leq x)$ (función de distribución), pues éstas son las que calcula **Minitab**. No tenemos que olvidar, por ejemplo, que si X es una variable aleatoria continua, entonces $P(X = a) = 0$ para todo a , por lo que se cumplen las siguientes igualdades: $P(X \leq x) = P(X < x)$, $P(X \geq x) = P(X > x)$, \dots . Pero si X es una variable aleatoria discreta, las probabilidades $P(X \leq x)$ y $P(X < x)$ no son (en general) iguales.

3.4. Inversa de la función de distribución (percentiles)

En ocasiones, en lugar de querer calcular probabilidades de sucesos, se desea justamente lo contrario, conocer el valor t que hace que la probabilidad del suceso ($X \leq t$) sea igual a un valor determinado p ; es decir, hallar t para que se cumpla $P(X \leq t) = p$; esto no es más que calcular percentiles de variables aleatorias. Para calcular el resultado de los percentiles de una variable aleatoria hay que elegir la opción **Calc** \Rightarrow **Probability Distributions** y a continuación el nombre de la variable aleatoria. Dentro del cuadro de diálogo que aparece hay que seleccionar **Inverse cumulative probability**.

Por ejemplo, vamos a calcular el valor t que verifica $P(X \leq t) = 0'98$, cuando $X \equiv \chi_{20}^2$; es decir, X tiene una distribución chi-cuadrado de Pearson con 20 grados de libertad. Para ello seleccionamos la opción **Calc** \Rightarrow **Probability Distributions** \Rightarrow **Chi-Square**. En el cuadro de diálogo activamos **Inverse cumulative probability**. Dejamos lo que aparece por defecto (cero) en **Noncentrality parameter**. En **Degrees of freedom** tecleamos **20**. No activamos la opción **Input column** sino la opción **Input constant**, en donde colocamos el valor **0,98**. Podemos almacenar el resultado en una constante tecleando en el recuadro **Optional storage** una K seguida de un número o poniendo un nombre a dicho resultado. Nosotros no vamos a rellenar el recuadro **Optional storage**, por lo que el resultado aparecerá en la ventana de sesión. Se puede comprobar que el valor t que verifica $P(X \leq t) = 0'98$ es 35'0196; es decir, $P(X \leq 35'0196) = 0'98$, siendo $X \equiv \chi_{20}^2$.

3.5. Ejercicios propuestos

Ejercicio 3.1 Si Z es una variable Normal Estándar, determina:

- a) $P(Z \leq 2'21)$.
- b) $P(Z < 3'47)$.
- c) $P(Z \leq -1'75)$.
- d) $P(Z > 2'46)$.
- e) $P(Z \geq 3'24)$.
- f) $P(Z > -3'08)$.
- g) $P(1'12 \leq Z \leq 2'68)$.
- h) $P(-0'85 < Z < 1'27)$.
- i) $P(-2'97 < Z \leq -1'33)$.

Ejercicio 3.2 Si X es una variable Normal con media 8'46 y desviación típica 1'14, halla:

- a) $P(X \leq 9'11)$.
- b) $P(X < 12'33)$.
- c) $P(X \leq 6'41)$.
- d) $P(X > 10'52)$.
- e) $P(X \geq 12'61)$.
- f) $P(X > 4'01)$.
- g) $P(6'11 \leq X \leq 11'91)$.
- h) $P(7'53 < X < 10'33)$.
- i) $P(5'05 \leq X < 6'83)$.

Ejercicio 3.3 Halla el valor de los siguientes percentiles:

- a) $Z_{0'58}$.
- b) $Z_{0'42}$.
- c) $Z_{0'999}$.
- d) $Z_{0'001}$.

Ejercicio 3.4 Genera 10000 datos aleatorios procedentes de una distribución chi-cuadrado de Pearson con 100 grados de libertad. Calcula la media de esta columna de datos aleatorios. Haz un histograma de los datos aleatorios generados, con la curva Normal superpuesta. ¿Puedes extraer alguna conclusión?

Ejercicio 3.5 Haz la representación gráfica de la función de densidad de una variable aleatoria chi-cuadrado de Pearson con 100 grados de libertad. Los valores del eje horizontal pueden ser todos los comprendidos entre 0 y 200 con un incremento de 0'1.

Ejercicio 3.6 Haz la representación gráfica de la función de distribución de una variable aleatoria chi-cuadrado de Pearson con 100 grados de libertad. Los valores del eje horizontal pueden ser todos los comprendidos entre 0 y 200 con un incremento de 0'1.

Ejercicio 3.7 Calcula el valor de los siguientes percentiles:

- a) $\chi^2_{6, 0'01}$.
- b) $\chi^2_{6, 0'99}$.
- c) $\chi^2_{72, 0'975}$.

Ejercicio 3.8 Sea X una variable aleatoria que sigue una distribución chi-cuadrado de Pearson con 15 grados de libertad. Determina el valor de a que verifica la siguiente igualdad:

- a) $P(X \leq a) = 0'05$.
- b) $P(X > a) = 0'99$.

Ejercicio 3.9 Calcula el valor de los siguientes percentiles:

a) $t_{26, 0'9}$.

b) $t_{26, 0'1}$.

c) $t_{75, 0'8}$.

Ejercicio 3.10 Sea X una variable aleatoria que sigue una distribución t de Student con 20 grados de libertad. Determina el valor de a que verifica la siguiente igualdad:

a) $P(X \leq a) = 0'99$.

b) $P(X \geq a) = 0'25$.

Ejercicio 3.11 Calcula el valor de los siguientes percentiles:

a) $F_{8, 6, 0'975}$.

b) $F_{25, 50, 0'01}$.

c) $F_{45, 35, 0'01}$.

Ejercicio 3.12 Sea X una variable aleatoria que sigue una distribución F de Snedecor con 10 grados de libertad en el numerador y 8 grados de libertad en el denominador. Determina el valor de a que verifica la siguiente igualdad:

a) $P(X < a) = 0'9$.

b) $P(X > a) = 0'05$.

4

Contrastes no paramétricos en una población

Observación importante: Si denotamos el nivel de significación por α , en todos los contrastes de hipótesis que realicemos con *Minitab*, el valor en el que nos tenemos que fijar es el nivel crítico o p-valor, ya que:

Si $p\text{-valor} > \alpha \Rightarrow$ aceptamos la hipótesis nula, H_0 .

Si $p\text{-valor} < \alpha \Rightarrow$ rechazamos la hipótesis nula y, por tanto, aceptamos la hipótesis alternativa, H_1 .

4.1. Contraste de aleatoriedad de la muestra

Con frecuencia las muestras se toman en serie temporal, cabiendo la posibilidad de que una observación dependa de la observación anterior. De ocurrir esto, la muestra no es aleatoria. Como tal propiedad es la base de la Estadística Inferencial, todos los contrastes de hipótesis y todos los intervalos de confianza que se dan en este texto quedarán invalidados si falla la hipótesis de aleatoriedad de la muestra. Resulta, pues, crucial dar procedimientos que permitan contrastar la hipótesis nula H_0 : *la muestra es aleatoria* contra la hipótesis alternativa H_1 : *la muestra no es aleatoria*. Los contrastes para ello son diversos, pero el más utilizado es el que describimos a continuación, que se denomina **contraste de las rachas**.

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

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 las muestras de datos de las columnas **Pulse1**, **Pulse2**, **Height** y **Weight** son aleatorias. Para ello, seleccionamos **Stat** \Rightarrow **Nonparametrics** \Rightarrow **Run Test**. En el cuadro de diálogo resultante, activamos el recuadro **Variables** (haciendo *clic* dentro de él); seleccionamos (haciendo doble *clic* sobre sus nombres) las columnas **Pulse1**, **Pulse2**, **Height** y **Weight**. Como vamos a comprobar la aleatoriedad de más de una muestra, tenemos que dicotomizar mediante las respectivas medias (no podemos dicotomizar mediante las respectivas medianas). Por tanto, activamos la opción **Above and below the mean** y pulsamos en **OK**. Si hubiésemos comprobado la aleatoriedad de una sola muestra, podríamos haber dicotomizado mediante la mediana, para lo cual habríamos calculado previamente el valor de dicha mediana; habríamos activado la opción **Above and below:** y, al lado, habríamos tecleado el resultado de dicha mediana.

En la ventana de sesión nos aparecen los resultados de los cuatro contrastes. Para la variable **Pulse1**, el p-valor es 0'368, mayor que el nivel de significación elegido (0'05), por lo que aceptamos la hipótesis nula; es decir, aceptamos que la muestra de resultados de dicha variable es aleatoria. Para la variable **Pulse2**, el p-valor es 0'002, menor que el nivel de significación elegido (0'05), por lo que rechazamos la hipótesis nula; es decir, rechazamos que la muestra de resultados de dicha variable es aleatoria. Para la variable **Height**, el p-valor es 0, menor que el nivel de significación elegido (0'05), por lo que rechazamos que la muestra de resultados de dicha variable es aleatoria. Para la variable **Weight**, el p-valor es 0'001, menor que el nivel de significación elegido (0'05), por lo que rechazamos que la muestra de resultados de dicha variable es aleatoria.

4.2. Contrastes de normalidad

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

En **Minitab** hay varias técnicas para comprobar el ajuste a una distribución Normal. Una de ellas es la opción **Graph** \Rightarrow **Probability Plot**. Con esta opción es posible comprobar la normalidad de varias variables a la vez.

Vamos a utilizar este método para comprobar qué variables de la hoja de datos **Marks.mtw** se ajustan al modelo Normal (cuando están observadas en toda la población). El archivo **Marks.mtw** es una hoja de datos que **Minitab** tiene de muestra y se encuentra en **C:\Archivos de programa\Minitab 15\English\Sample Data\Student9**. En las aulas de informática de la Universidad de Murcia este archivo de datos se encuentra en **C:\Archivos de programa\UM\Minitab 15\English\Sample Data\Student9**.

En primer lugar, abrimos dicha hoja de datos (**File** \Rightarrow **Open Worksheet**). El archivo muestra las calificaciones (puntuadas de 0 a 100) de 24 estudiantes en tres exámenes de tipo test (**Test1**, **Test2** y **Test3**).

En segundo lugar, vamos a comprobar que las muestras de los datos de las columnas **Test1**, **Test2** y **Test3** son aleatorias.

En tercer lugar, vamos a ver si se puede aceptar que las variables **Test1**, **Test2** y **Test3** son Normales. Para ello, seleccionamos **Graph** \Rightarrow **Probability Plot**. En el cuadro de diálogo resultante seleccionamos **Single** y pulsamos en **OK**. En **Graph variables** seleccionamos, de la lista de variables de la izquierda, las que podrían ajustarse a un modelo Normal; es decir, **Test1**, **Test2** y **Test3**. Pulsamos en **Distribution** y, en el cuadro de diálogo resultante, dejamos lo que está activado por defecto; es decir, **Normal**, y no rellenamos la opción **Historical Parameters** ya que no sabemos los resultados de las estimaciones de la media y de la desviación típica poblacionales.

Nos aparecen tres gráficos, uno para cada una de las variables seleccionadas. Además, vemos que aparecen, en la parte superior derecha de las representaciones gráficas, los resultados de un contraste de normalidad; concretamente, el contraste de Anderson-Darling.

Podemos ver que el gráfico probabilístico de la variable **Test1** se aproxima a una recta. Además, el p-valor del contraste de normalidad es igual a 0,232 y, por tanto, es mayor que los usuales niveles de significación ($\alpha = 0,05$ o $\alpha = 0,01$). En consecuencia, podemos aceptar que la variable **Test1** se ajusta al modelo Normal.

Por otra parte, podemos observar que el gráfico probabilístico de la variable **Test2** también se aproxima a una recta. Además, el p-valor del contraste de normalidad es igual a 0,119 y, por tanto, es mayor que los usuales niveles de significación ($\alpha = 0,05$ o $\alpha = 0,01$). En consecuencia, podemos aceptar que la variable **Test2** se ajusta al modelo Normal.

Por último, el gráfico probabilístico de la variable **Test3** no se aproxima a una recta. Además, el p-valor del contraste de normalidad es menor que 0,007. Tanto si consideramos un nivel de significación de $\alpha = 0,01$ como si consideramos un nivel de significación de $\alpha = 0,05$ resulta que el p-valor es menor que α . En consecuencia, la variable **Test3** no se ajusta al modelo Normal.

Otra opción para realizar un test de ajuste a un modelo Normal es **Stat**⇒**Basic Statistics**⇒**Normality Test**, aunque tiene el inconveniente de que solamente comprueba la normalidad de **una** variable (cada vez que se selecciona esta opción).

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 contraste, **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.

4.3. Ejercicios propuestos

Ejercicio 4.1

- a) Crea un nuevo proyecto de *Minitab*.
- b) Abre la hoja de datos **Prestamos.mtw** (datos del Ejercicio 1.1).
- c) Calcula la 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**

Ejercicio 4.2

- Crea un nuevo proyecto de **Minitab**.
- Abre la hoja de datos **Transacciones.mtw** (datos del Ejercicio 1.2).
- ¿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?
- ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que las variables **TR**, **TRF** y **Porcentaje TRF** son Normales? ¿Por qué?
- Graba el proyecto con el siguiente nombre: **Ejercicio4-2.mpj**

Ejercicio 4.3 En la tabla siguiente aparecen los datos de 10 bibliotecas, en las cuales se ha observado las siguientes variables: número total de títulos catalogados en un año (X), número de horas totales al año que emplea la biblioteca en catalogar sus títulos (Y) y costo, en euros, de una hora de catalogación (Z).

x_i	y_i	z_i
1550	220	15'75
1640	230	14'50
1000	140	16'40
950	135	16'70
750	110	17'10
1700	255	12'50
1650	228	14'80
1860	270	15'25
1900	280	18'50
900	130	17'30

- Crea un nuevo proyecto de **Minitab**.
- Introduce solamente los datos de la variable Z . Guarda la hoja de datos con el nombre **Costo-hora-catalogacion.mtw**
- ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la muestra de datos de la variable Z es aleatoria?
- ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'02$, que la variable aleatoria Z es Normal?
- Graba el proyecto con el siguiente nombre: **Ejercicio4-3.mpj**

Ejercicio 4.4 En la tabla siguiente aparecen los resultados del peso, en gramos, (X) y del precio, en euros, (Y) de una muestra de 12 libros.

x_i	y_i
325	110
890	30
415	75
400	45
515	32
650	69
790	30
890	34
320	42
420	46
620	53
720	97

- Crea un nuevo proyecto de *Minitab*.
- Introduce solamente los datos de la variable Y . Guarda la hoja de datos con el nombre **Precio-libros.mtw**
- ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la muestra de datos de la variable Y es aleatoria?
- ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'02$, que la variable aleatoria Y es Normal?
- Graba el proyecto con el siguiente nombre: **Ejercicio4-4.mpj**

Ejercicio 4.5 En una muestra aleatoria simple de 15 individuos que consultan bases de datos, el tiempo (en minutos) que están utilizando el ordenador para realizar esta tarea es:

22	13	17	14	15	18	19	14	17	20	21	13	15	18	17
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

- Crea un nuevo proyecto de *Minitab*.
- Introduce los datos y grábalos con el nombre **Tiempo-consulta.mtw**
- ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la muestra es aleatoria?
- ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la variable aleatoria “Tiempo empleado en consultar bases de datos por ordenador” es Normal?
- Graba el proyecto con el siguiente nombre: **Ejercicio4-5.mpj**

Ejercicio 4.6 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
----	----	----	----	----	----	----	----	----	----

- Crea un nuevo proyecto de *Minitab*.
- Calcula la mediana de estos datos.

- c) 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?
- d) ¿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?
- e) Graba el proyecto con el siguiente nombre: **Ejercicio4-6.mpj**

Ejercicio 4.7 La tabla siguiente contiene el número mensual de materias buscadas por los usuarios de una biblioteca (X) y el número mensual de materias localizadas por dichos usuarios (Y):

mes	x_i	y_i
1	42	22
2	65	30
3	68	35
4	55	30
5	35	20
6	40	25
7	50	30
8	26	15
9	42	22
10	56	38
11	38	15
12	50	34

- a) Crea un nuevo proyecto de **Minitab**.
- b) Introduce solamente los datos de la variable X .
- c) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la muestra de datos de la variable X es aleatoria?
- d) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la variable aleatoria X es Normal?
- e) Graba el proyecto con el siguiente nombre: **Ejercicio4-7.mpj**

5

Contrastes paramétricos en una población

5.1. Contrastes sobre la media

El contraste de hipótesis sobre una media sirve para tomar decisiones acerca del verdadero valor poblacional de la media de una variable aleatoria.

5.1.1. Contraste sobre la media cuando la desviación típica poblacional es conocida

Esta técnica es válida solamente si la muestra es aleatoria y la población es Normal o el tamaño muestral, n , es grande (en la práctica, $n \geq 30$).

Para hacer este test hay que seleccionar **Stat** \Rightarrow **Basic Statistics** \Rightarrow **1-Sample Z**. Esta opción también nos da el intervalo de confianza para la media poblacional, μ .

Abrimos el archivo de datos **Pulse.mtw**. Vamos a suponer que conocemos el valor de la desviación típica poblacional de la variable **Pulse1** (pulso antes de correr), $\sigma = 10$ pulsaciones por minuto. Comprobemos si se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que el pulso medio poblacional antes de correr es mayor que 70 pulsaciones por minuto. Si μ denota la media poblacional de la variable $X = \text{Pulso antes de correr}$, el contraste es $H_0 : \mu \leq 70$ frente a $H_1 : \mu > 70$.

En la Práctica 4 ya hemos comprobado que la muestra de resultados de la variable **Pulse1** es aleatoria. Además, el tamaño muestral es grande ($n = 92$). Por tanto, podemos utilizar este procedimiento estadístico.

Seleccionamos la opción **Stat** \Rightarrow **Basic Statistics** \Rightarrow **1-Sample Z**. En **Samples in columns** seleccionamos, de la lista de variables de la izquierda, la columna o columnas para las cuales se va a realizar este tipo de contraste; en nuestro caso, '**Pulse1**'. Dejamos desactivada la opción **Summarized data**. En **Standard deviation** tecleamos el valor de la desviación típica poblacional, σ , que suponemos que es **10**. Activamos **Perform hypothesis test** y en **Hypothesized mean** especificamos el valor, μ_0 , con el que se compara la

media poblacional, que es 70. Si pulsamos el botón **Options** nos aparece un nuevo cuadro de diálogo con las siguientes opciones:

Confidence level: Por defecto se muestra un intervalo de confianza al 95 % para la media poblacional μ . Se puede introducir un valor entre 1 y 99 para solicitar otro nivel de confianza. En nuestro caso, podemos dejar lo que aparece por defecto, es decir, 95.

Alternative: Aquí se especifica cuál es la hipótesis alternativa: **less than** significa que la hipótesis alternativa es $H_1 : \mu < \mu_0$, **not equal** significa que la hipótesis alternativa es $H_1 : \mu \neq \mu_0$ y **greater than** significa que la hipótesis alternativa es $H_1 : \mu > \mu_0$. Tengamos en cuenta que con la opción **less than** el intervalo de confianza para la media será del tipo $(-\infty, b)$, con la opción **not equal** el intervalo de confianza para la media será del tipo (a, b) y con la opción **greater than** el intervalo de confianza para la media será del tipo $(a, +\infty)$. En nuestro caso, tenemos que seleccionar **greater than** ya que la hipótesis alternativa es $H_1 : \mu > 70$.

Podemos comprobar, en la ventana de sesión, que el p-valor es 0'003, claramente menor que el nivel de significación, $\alpha = 0'05$. En consecuencia, rechazamos la hipótesis nula y, por tanto, aceptamos la hipótesis alternativa; es decir, aceptamos que la media poblacional de la variable **Pulse 1** es mayor que 70 pulsaciones por minuto. El intervalo de confianza al 95 % para la media poblacional, asociado a este contraste de hipótesis, es $(71'15, +\infty)$.

También se puede realizar este contraste de hipótesis si sabemos el tamaño muestral y el resultado de la media muestral. Veámoslo con un ejemplo:

En el volumen de Julio de 1992 de Economics Abstracts, la media del número de palabras por resumen es 79'56, con una varianza de 615'04. Se extrae una muestra aleatoria simple de 30 resúmenes escritos en alemán y se observa que la media del número de palabras por resumen es 67'47. Se quiere decidir si existe una diferencia significativa entre la media de palabras por resumen de los escritos en alemán y la media de palabras por resumen de todos los de este volumen.

Vamos a suponer que la varianza del número de palabras por resumen de los escritos en alemán coincide con la varianza del número de palabras por resumen de todos los de este volumen. Así pues, los datos que tenemos son los siguientes:

$$\begin{aligned}\mu_0 &= 79'56, \\ \sigma^2 &= 615'04 \Rightarrow \sigma = \sqrt{615'04} = 24'8, \\ \bar{X} &= 67'47, \\ n &= 30.\end{aligned}$$

La variable observada en la población no puede ser Normal pues es discreta, pero como el tamaño muestral es 30, entonces podemos aplicar esta técnica. Así pues, consideramos el siguiente contraste de hipótesis:

$$\begin{aligned}H_0 : \mu &= 79'56, \\ H_1 : \mu &\neq 79'56.\end{aligned}$$

Seleccionamos la opción **Stat** \Rightarrow **Basic Statistics** \Rightarrow **1-Sample Z**. Activamos la opción **Summarized data**, con lo cual se desactiva automáticamente la opción **Samples in columns**. En **Sample size** tenemos que teclear el tamaño muestral, que es 30 y en **Mean** tenemos que teclear el resultado de la media muestral, que es 67,47. En **Standard deviation** tecleamos el valor de la desviación típica poblacional, σ , que suponemos que es 24,8. Activamos **Perform hypothesis test** y en **Hypothesized mean** especificamos el valor,

μ_0 , con el que se compara la media poblacional, que es **79,56**. Pulsamos en **Options** y, en el cuadro de diálogo resultante, en **Alternative** seleccionamos **not equal** puesto que nuestra hipótesis alternativa es $H_1 : \mu \neq 79'56$.

Podemos comprobar, en la ventana de sesión, que el p-valor es 0'008, claramente menor que los niveles de significación usuales ($\alpha = 0'05$ ó $\alpha = 0'01$). En consecuencia, rechazamos la hipótesis nula y, por tanto, aceptamos que existe diferencia significativa entre la media del número de palabras por resumen en alemán y la media del número de palabras por resumen de todos ellos. El intervalo de confianza al 95 % para la media poblacional, asociado a este contraste de hipótesis, es (58'60, 76'34).

5.1.2. Contraste sobre la media cuando la desviación típica poblacional es desconocida

Igual que en el apartado anterior, esta técnica es válida solamente si la muestra es aleatoria y la población es Normal o el tamaño muestral, n , es grande (en la práctica, $n \geq 30$).

Para realizar este contraste paramétrico hay que seleccionar **Stat** \Rightarrow **Basic Statistics** \Rightarrow **1-Sample t**. La manera de utilizar esta opción es la misma que la explicada en el apartado anterior.

Con el archivo de datos **Pulse.mtw**, veamos si se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que el pulso medio poblacional antes de correr es igual a 71 pulsaciones por minuto. Lo que queremos comprobar es si la media poblacional de la variable **Pulse1** es igual a 71 pulsaciones por minuto, suponiendo ahora desconocida la desviación típica poblacional (lo cual es cierto). Si μ denota la media poblacional de la variable **Pulse1**, el contraste es $H_0 : \mu = 71$ frente a $H_1 : \mu \neq 71$.

Podemos comprobar, en la ventana de sesión, que el p-valor es 0'107, claramente mayor que el nivel de significación, $\alpha = 0'05$, por lo que podemos aceptar la hipótesis nula; es decir, aceptamos que la media poblacional del número de pulsaciones por minuto antes de correr es igual a 71. El intervalo de confianza al 95 % para la media poblacional de dicha variable es (70'59, 75'15).

También se puede realizar este contraste de hipótesis si sabemos el tamaño muestral, el resultado de la media muestral y el resultado de la cuasi-desviación típica muestral. Veámoslo con un ejemplo:

El número medio de libros por estante de una biblioteca es 24. Extraída una muestra de 91 estantes de libros de matemáticas se obtiene una media de 25 libros, con una cuasi-desviación típica de 1'5. Queremos decidir si existe diferencia significativa entre el número medio de libros de matemáticas por estante y el número medio de libros por estante.

La variable $X = \text{“Número de libros de matemáticas por estante”}$ no puede ser Normal porque es discreta; pero como $n = 91 \geq 30$ entonces se puede utilizar este procedimiento.

Los datos conocidos son:

$$\begin{aligned}\mu_0 &= 24, \\ S &= 1'5, \\ \bar{X} &= 25, \\ n &= 91.\end{aligned}$$

El contraste de hipótesis que vamos a hacer es el siguiente:

$$\begin{aligned}H_0 &: \mu = 24, \\ H_1 &: \mu \neq 24.\end{aligned}$$

Seleccionamos la opción **Stat** \Rightarrow **Basic Statistics** \Rightarrow **1-Sample t**. Activamos la opción **Summarized data**, con lo cual se desactiva automáticamente la opción **Samples in columns**. En **Sample size** tenemos que teclear el tamaño muestral, que es **91**, en **Mean** tenemos que teclear el resultado de la media muestral, que es **25**, y en **Standard deviation** tenemos que teclear el resultado de la cuasi-desviación típica muestral, que es **1,5**. Activamos **Perform hypothesis test** y en **Hypothesized mean** especificamos el valor, μ_0 , con el que se compara la media poblacional, que es **24**. Pulsamos en **Options** y, en el cuadro de diálogo resultante, en **Alternative** seleccionamos **not equal** puesto que nuestra hipótesis alternativa es $H_1 : \mu \neq 24$.

Podemos comprobar, en la ventana de sesión, que el p-valor es 0, el mínimo posible y, por supuesto, claramente menor que los niveles de significación usuales ($\alpha = 0'05$ ó $\alpha = 0'01$). En consecuencia, rechazamos la hipótesis nula y, por tanto, aceptamos que existe diferencia significativa entre el número medio de libros de matemáticas por estante y el número medio de libros por estante. El intervalo de confianza al 95 % para la media poblacional, asociado a este contraste de hipótesis, es (24'688, 25'312).

5.2. Contrastes sobre la varianza

El contraste de hipótesis sobre una varianza sirve para tomar decisiones acerca del verdadero valor poblacional de la varianza de una variable aleatoria. **Minitab** realiza el contraste solamente en el caso en el que la media poblacional es desconocida.

Esta técnica es válida solamente si la muestra es aleatoria y la población es Normal.

Para hacer el contraste de hipótesis sobre una varianza poblacional hay que seleccionar **Stat** \Rightarrow **Basic Statistics** \Rightarrow **1 Variance**. Esta opción también se utiliza para realizar un test sobre la desviación típica poblacional.

En la Práctica 4 ya hemos comprobado que la muestra de resultados de la variable **Pulse1** (del archivo de datos **Pulse.mtw**) es aleatoria, y que la variable **Pulse1** es Normal. Por tanto, podemos utilizar este procedimiento estadístico para comprobar si se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la varianza poblacional del pulso antes de correr es menor que 130 pulsaciones al cuadrado. Si σ^2 denota la varianza poblacional de la variable $X = \text{Pulso antes de correr}$, el contraste es $H_0 : \sigma \geq 130$ frente a $H_1 : \sigma^2 < 130$.

Seleccionamos, por tanto, la opción **Stat** \Rightarrow **Basic Statistics** \Rightarrow **1 Variance**. En el cuadro de diálogo resultante, arriba a la derecha, seleccionamos **Enter variance** (si quisiéramos realizar un contraste sobre la desviación típica poblacional, seleccionaríamos **Enter standard deviation**); en **Samples in columns** se selecciona, de la lista de variables de la izquierda, la columna o columnas para las cuales se va a realizar este tipo de contraste; en nuestro caso se selecciona '**Pulse1**'. Dejamos desactivada la opción **Summarized data**. Activamos **Perform hypothesis test** y en **Hypothesized variance** se especifica el valor, σ_0^2 , con el que se compara la varianza poblacional, que es **130**. Si pulsamos el botón **Options** nos aparece un nuevo cuadro de diálogo con las siguientes opciones:

Confidence level: Por defecto se muestra un intervalo de confianza al 95 % para la varianza poblacional σ^2 . Se puede introducir un valor entre 1 y 99 para solicitar otro nivel de confianza. En nuestro caso, podemos dejar lo que aparece por defecto, es decir, 95.

Alternative: Aquí se especifica cuál es la hipótesis alternativa: **less than** significa que la hipótesis alternativa es $H_1 : \sigma^2 < \sigma_0^2$, **not equal** significa que la hipótesis alternativa es $H_1 : \sigma^2 \neq \sigma_0^2$ y **greater than** significa que la hipótesis alternativa es $H_1 : \sigma^2 > \sigma_0^2$. Tengamos en cuenta que con

la opción **less than** el intervalo de confianza para la varianza será del tipo $(-\infty, b)$, con la opción **not equal** el intervalo de confianza para la varianza será del tipo (a, b) y con la opción **greater than** el intervalo de confianza para la varianza será del tipo $(a, +\infty)$. En nuestro caso, tenemos que seleccionar **less than** ya que la hipótesis alternativa es $H_1 : \sigma^2 < 130$.

Podemos comprobar, en la ventana de sesión, que el p-valor (para el método Standard) es 0'338, claramente mayor que el nivel de significación, $\alpha = 0'05$. En consecuencia, aceptamos la hipótesis nula y, por tanto, no podemos aceptar la hipótesis alternativa; es decir, no podemos aceptar que la varianza poblacional del pulso antes de correr es menor que 130 pulsaciones al cuadrado. El intervalo de confianza al 95 % para la varianza poblacional, asociado a este contraste de hipótesis (con el método Standard), es $(-\infty, 158)$. El intervalo de confianza al 95 % para la desviación típica poblacional, asociado a este contraste de hipótesis (con el método Standard), es $(-\infty, 12'6)$.

También se puede realizar este contraste de hipótesis si sabemos el tamaño muestral y el resultado de la cuasi-varianza muestral. Veámoslo con un ejemplo:

Se sabe que las calificaciones en la asignatura *A* es una variable Normal de media y varianza desconocidas. Se extrae una muestra aleatoria simple de 81 alumnos de la asignatura *A*, obteniéndose una media de 6'8 puntos, con una cuasi-varianza de 1'69 puntos al cuadrado, en las calificaciones de dichos alumnos. Sabemos que la varianza de las calificaciones en otra asignatura *B* es de 2'6 puntos al cuadrado. Queremos saber si la verdadera varianza de las calificaciones en la asignatura *A* es menor que la varianza en las calificaciones en la asignatura *B*.

Como la cuasi-varianza muestral es $S^2 = 1'69 < 2'6$, esta evidencia debe ser compatible con la hipótesis alternativa. Así pues, vamos a realizar el siguiente contraste:

$$H_0 : \sigma^2 \geq 2'6,$$

$$H_1 : \sigma^2 < 2'6.$$

Seleccionamos la opción **Stat** \Rightarrow **Basic Statistics** \Rightarrow **1 Variance**. En el cuadro de diálogo resultante, arriba a la derecha, seleccionamos **Enter variance**. Activamos la opción **Summarized data**, con lo cual se desactiva automáticamente la opción **Samples in columns**. En **Sample size** tenemos que teclear el tamaño muestral, que es 81, y en **Sample variance** tenemos que teclear el resultado de la cuasi-varianza muestral, que es 1,69. Activamos **Perform hypothesis test** y en **Hypothesized variance** se especifica el valor, σ_0^2 , con el que se compara la varianza poblacional, que es 2,6. Pulsamos en **Options** y, en el cuadro de diálogo resultante, en **Alternative** seleccionamos **less than** puesto que nuestra hipótesis alternativa es $H_1 : \sigma^2 < 2'6$.

Podemos comprobar, en la ventana de sesión, que el p-valor es 0'006, claramente menor que los niveles de significación usuales ($\alpha = 0'05$ ó $\alpha = 0'01$). En consecuencia, rechazamos la hipótesis nula y, por tanto, aceptamos que la varianza de las calificaciones en la asignatura *A* es menor que la varianza de las calificaciones en la asignatura *B*. El intervalo de confianza al 95 % para la varianza poblacional, asociado a este contraste de hipótesis, es $(-\infty, 2'24)$.

5.3. Ejercicios propuestos

Ejercicio 5.1

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

- b) Abre la hoja de datos **Prestamos.mtw** (datos del Ejercicio 1.1).
- c) ¿Se puede aceptar, con un nivel de significación de 0'05, que la media poblacional del porcentaje anual de préstamos por usuario es igual a 70? ¿Por qué?
- d) ¿Se puede aceptar, con un nivel de significación de 0'05, que la varianza poblacional del porcentaje anual de préstamos por usuario es igual a 140? ¿Por qué?
- e) Graba el proyecto con el siguiente nombre: **Ejercicio5-1.mpj**

Ejercicio 5.2

- a) Crea un nuevo proyecto de *Minitab*.
- b) Abre la hoja de datos **Transacciones.mtw** (datos del Ejercicio 1.2).
- c) ¿Se puede aceptar, con un nivel de significación de 0'05, que la media poblacional del porcentaje de transacciones de referencia finalizadas es menor que 86? ¿Por qué?
- d) ¿Se puede aceptar, con un nivel de significación de 0'05, que la desviación típica poblacional del porcentaje de transacciones de referencia finalizadas es mayor que 5? ¿Por qué?
- e) Graba el proyecto con el siguiente nombre: **Ejercicio5-2.mpj**

Ejercicio 5.3

- a) Crea un nuevo proyecto de *Minitab*.
- b) Abre la hoja de datos **Costo-hora-catalogacion.mtw** (datos del Ejercicio 4.3).
- c) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'01$, que la media poblacional del costo de una hora de catalogación es menor que 17 euros?
- d) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'01$, que la desviación típica poblacional del costo de una hora de catalogación es mayor que 2 euros?
- e) Graba el proyecto con el siguiente nombre: **Ejercicio5-3.mpj**

Ejercicio 5.4

- a) Crea un nuevo proyecto de *Minitab*.
- b) Abre la hoja de datos **Precio-libros.mtw** (datos del Ejercicio 4.4).
- c) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'01$, que la media poblacional del precio es igual a 55 euros?
- d) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'01$, que la desviación típica poblacional del precio es igual a 24 euros?
- e) Graba el proyecto con el siguiente nombre: **Ejercicio5-4.mpj**

Ejercicio 5.5

- a) Crea un nuevo proyecto de *Minitab*.
- b) Abre la hoja de datos **Tiempo-consulta.mtw** (datos del Ejercicio 4.5).
- c) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la media poblacional del tiempo empleado en consultar bases de datos por ordenador es mayor que 15 minutos?

d) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la desviación típica poblacional del tiempo empleado en consultar bases de datos por ordenador es menor que 2 minutos?

e) Graba el proyecto con el siguiente nombre: **Ejercicio5-5.mpj**

Ejercicio 5.6 El número medio recomendado de usuarios servidos semanalmente por cada miembro del personal de una biblioteca es de 100. En una muestra aleatoria simple de 81 miembros del personal de las bibliotecas de una determinada región se obtiene una media de 132'88 usuarios servidos semanalmente, con una cuasidesviación típica de 55'19.

a) Crea un nuevo proyecto de *Minitab*.

b) ¿Las bibliotecas de dicha región siguen la recomendación mencionada? ¿Por qué?

c) Graba el proyecto con el siguiente nombre: **Ejercicio5-6.mpj**

Ejercicio 5.7 El precio medio de los libros en rústica es de 63'4 euros, con una desviación típica de 14'8 euros. Una muestra aleatoria simple de 61 libros en rústica con ilustraciones en color tiene un precio medio de 69'5 euros, con una cuasidesviación típica de 16'6 euros.

a) Crea un nuevo proyecto de *Minitab*.

b) ¿Permiten los datos afirmar que los libros en rústica con ilustraciones en color son más caros que el resto de libros en rústica? ¿Por qué?

c) ¿La varianza del precio de los libros en rústica con ilustraciones en color es mayor que la del precio de los libros en rústica? ¿Por qué?

d) Graba el proyecto con el siguiente nombre: **Ejercicio5-7.mpj**

Ejercicio 5.8 Se sabe que el número medio de veces que un artículo científico es citado durante los 5 siguientes años a su publicación es de 6'5. Se eligen aleatoria e independientemente 71 artículos de medicina, obteniéndose una media de 7'8 citas durante los 5 siguientes años a su publicación, con una cuasidesviación típica de 2'3.

a) Crea un nuevo proyecto de *Minitab*.

b) ¿Se puede afirmar que durante los 5 siguientes años a su publicación se citan más los artículos de medicina que el resto de artículos científicos? ¿Por qué?

c) Graba el proyecto con el siguiente nombre: **Ejercicio5-8.mpj**

6

Contrastes paramétricos en dos poblaciones

6.1. Comparación de dos varianzas con muestras independientes

En el apartado siguiente vamos a estudiar el problema de la comparación de dos medias poblacionales en el caso en que observemos dos variables aleatorias Normales (una en cada población), suponiendo que se han extraído dos muestras aleatorias (una de cada población) independientes. Veremos en dicho apartado que necesitamos saber si las varianzas poblacionales (que serán desconocidas) son iguales o distintas. Por este motivo estudiamos ahora el contraste de comparación de varianzas en el caso en que desconozcamos los valores de las medias poblacionales.

Este procedimiento estadístico solamente es válido cuando las dos muestras son aleatorias y las dos poblaciones son Normales.

Para realizar este test paramétrico hay que seleccionar **Stat** \Rightarrow **Basic Statistics** \Rightarrow **2 Variances**.

Ejemplo 1. Con el archivo de datos **Pulse.mtw**, comprobemos si se puede aceptar, con un nivel de significación de $\alpha = 0.05$, que la varianza poblacional del pulso de los hombres antes de correr es igual a la varianza poblacional del pulso de las mujeres antes de correr. Lo que se quiere es comparar la varianza poblacional de la variable **Pulse1** para los grupos en los que la variable **Sex** vale **1** (Hombre) y **2** (Mujer). El contraste que tenemos que hacer es $H_0 : \sigma_1^2 = \sigma_2^2$ frente a $H_1 : \sigma_1^2 \neq \sigma_2^2$, siendo $X_1 = \text{“Pulso de los hombres antes de correr”}$ y $X_2 = \text{“Pulso de las mujeres antes de correr”}$. Como no hay relación alguna entre el grupo de hombres y el grupo de mujeres, podemos afirmar que las muestras son independientes. Por tanto, nos encontramos ante un contraste de comparación de dos varianzas poblacionales, con muestras independientes y medias poblacionales desconocidas. Ya hemos comprobado, en la Práctica 4, que las dos variables, X_1 y X_2 , son Normales.

Para hacer este contraste seleccionamos **Stat** \Rightarrow **Basic Statistics** \Rightarrow **2 Variances**. Activamos la opción **Samples in one column**, con lo cual se desactivan automáticamente las opciones **Samples in different columns** y **Summarized data**. En **Samples** seleccionamos, de la lista de variables de la izquierda, la columna

'Pulse1'; en **Subscripts** seleccionamos, de la lista de la izquierda, la columna '**Sex**'. Si pulsamos el botón **Options** nos aparece un nuevo cuadro de diálogo con las siguientes opciones:

Confidence level: Por defecto se muestra un intervalo de confianza al 95 % para la diferencia de desviaciones típicas poblacionales, $\sigma_1 - \sigma_2$. Se puede introducir un valor entre 1 y 99 para solicitar otro nivel de confianza. En nuestro ejemplo, podemos dejar lo que aparece por defecto, es decir, 95.

Title: Aquí se puede escribir un título para el resultado del contraste. En nuestro ejemplo, podemos dejarlo en blanco.

Como resultado de este contraste obtenemos una nueva ventana que contiene dos gráficos y los resultados de dos tests de hipótesis sobre comparación de dos varianzas (el test F de Snedecor y el test de Levene). Podemos comprobar que el p-valor para el test F de Snedecor es 0'299; claramente mayor que el nivel de significación, $\alpha = 0'05$, por lo que podemos aceptar la hipótesis nula; es decir, podemos aceptar que la varianza poblacional del pulso de los hombres antes de correr es igual a la varianza poblacional del pulso de las mujeres antes de correr. Con el test de Levene también aceptaríamos la hipótesis nula pues el p-valor es igual a 0'148.

Ejemplo 2. Con el archivo de datos **Pulse.mtw**, comprobemos si se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la varianza poblacional del pulso de los hombres después de correr es igual a la varianza poblacional del pulso de las mujeres después de correr. Lo que se quiere es comparar la varianza poblacional de la variable **Pulse2** para los grupos en los que la variable **Sex** vale **1** (Hombre) y **2** (Mujer). El contraste que tenemos que hacer es $H_0 : \sigma_1^2 = \sigma_2^2$ frente a $H_1 : \sigma_1^2 \neq \sigma_2^2$, siendo $X_1 = \text{"Pulso de los hombres después de correr"}$ y $X_2 = \text{"Pulso de las mujeres después de correr"}$.

Para hacer este contraste seleccionamos **Stat** \Rightarrow **Basic Statistics** \Rightarrow **2 Variances**. Activamos la opción **Samples in one column**, con lo cual se desactivan automáticamente las opciones **Samples in different columns** y **Summarized data**. En **Samples** seleccionamos, de la lista de variables de la izquierda, la columna '**Pulse2**'; en **Subscripts** seleccionamos, de la lista de la izquierda, la columna '**Sex**'.

Se puede comprobar que el p-valor para el test F de Snedecor es 0'003, claramente menor que el nivel de significación, $\alpha = 0'05$, por lo que tenemos que rechazar la hipótesis nula y, por tanto, aceptar que la varianza poblacional del pulso de los hombres después de correr es distinta de la varianza poblacional del pulso de las mujeres después de correr. Con el test de Levene llegamos a la misma conclusión pues el p-valor es igual a 0'011.

También se puede realizar este contraste de hipótesis si sabemos los dos tamaños muestrales y los resultados de las dos cuasi-varianzas muestrales. Veámoslo con un nuevo ejemplo:

Ejemplo 3. Supongamos que, de una muestra aleatoria de 21 personas que son socias de una biblioteca, la media del número de horas por semana que pasan en la biblioteca es 10, con una cuasi-varianza de 9. Y para una muestra aleatoria independiente de la primera, de 16 personas que no son socias de la biblioteca, la media es 6, con una cuasi-varianza de 4. ¿Existe diferencia significativa entre las varianzas del número de horas semanales que pasan en la biblioteca los socios y los no socios?

Como la cuasi-varianza muestral en el grupo de los socios es mayor que en el grupo de los no socios, entonces S_1^2 será la cuasi-varianza en el grupo de los socios; es decir, $X_1 = \text{"Tiempo semanal que permanecen en la biblioteca los socios"}$ y $X_2 = \text{"Tiempo semanal que permanecen en la biblioteca los no socios"}$. Hemos de suponer que las variables aleatorias X_1 y X_2 son Normales.

Así pues, se tienen los siguientes datos:

$$\begin{aligned}n_1 &= 21, & S_1^2 &= 9, \\n_2 &= 16, & S_2^2 &= 4.\end{aligned}$$

Vamos a decidir sobre el siguiente contraste de hipótesis:

$$\begin{aligned}H_0 &: \sigma_1^2 = \sigma_2^2, \\H_1 &: \sigma_1^2 \neq \sigma_2^2.\end{aligned}$$

Seleccionamos la opción **Stat** \Rightarrow **Basic Statistics** \Rightarrow **2 Variances**. Activamos la opción **Summarized data**, con lo cual se desactivan automáticamente las opciones **Samples in one column** y **Samples in different columns**. Dentro de **First**, en **Sample size** tenemos que teclear el tamaño muestral de la primera muestra, que es **21**, y en **Variance** tenemos que teclear el resultado de la cuasi-varianza de la primera muestra, que es **9**. Dentro de **Second**, en **Sample size** tenemos que teclear el tamaño muestral de la segunda muestra, que es **16**, y en **Variance** tenemos que teclear el resultado de la cuasi-varianza de la segunda muestra, que es **4**.

Tanto en la ventana de sesión como en el gráfico generado comprobamos que el p-valor para el test F de Snedecor es 0'114, mayor que los niveles de significación usuales ($\alpha = 0'05$ ó $\alpha = 0'01$) y, por tanto, aceptamos la hipótesis nula. En consecuencia, aceptamos que no existe diferencia significativa entre las varianzas del número de horas semanales que pasan en la biblioteca los socios y los no socios.

6.2. Comparación de dos medias con muestras independientes

En general, un contraste para decidir sobre la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a la hipótesis alternativa $H_1 : \mu_1 \neq \mu_2$ es bastante frecuente y constituye uno de los primeros objetivos de cualquier investigador que se inicia en estadística. Los métodos de resolución del problema varían según las muestras sean independientes o apareadas, y según las varianzas poblacionales sean conocidas o desconocidas. Dentro del caso en que las varianzas poblacionales sean desconocidas, el método depende de si son iguales o distintas. El caso de muestras independientes y varianzas poblacionales conocidas no se puede hacer con *Minitab*. Trataremos, a continuación, el resto de los casos.

6.2.1. Comparación de dos medias con muestras independientes y varianzas poblacionales desconocidas pero iguales

Este procedimiento solamente es válido cuando las dos muestras son aleatorias y las dos poblaciones son Normales o los dos tamaños muestrales son grandes (en la práctica $n_1, n_2 \geq 30$).

Para realizar este test paramétrico hay que seleccionar **Stat** \Rightarrow **Basic Statistics** \Rightarrow **2-Sample t**.

Con el archivo de datos **Pulse.mtw**, comprobemos si se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que el pulso medio poblacional de los hombres antes de correr es igual al pulso medio poblacional de las mujeres antes de correr. Lo que se quiere es comparar la media poblacional de la variable **Pulse1** para los grupos en los que la variable **Sex** vale **1** (Hombre) y **2** (Mujer). El contraste

que tenemos que hacer es $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$, siendo X_1 = “Pulso de los hombres antes de correr” y X_2 = “Pulso de las mujeres antes de correr”.

En el **Ejemplo 1** de la sección 6.1 hemos comprobado que se puede aceptar que la varianza poblacional del pulso de los hombres antes de correr es igual a la varianza poblacional del pulso de las mujeres antes de correr. Por tanto, nos encontramos ante un contraste de comparación de dos medias poblacionales, con muestras independientes y varianzas poblacionales desconocidas pero iguales. Aunque las variables aleatorias X_1 y X_2 no fuesen Normales (que sí lo son, pues lo hemos comprobado en la Práctica 4), se puede aplicar este contraste debido a que los tamaños muestrales son suficientemente grandes: $n_1 = 57$ y $n_2 = 35$.

Para hacer este contraste seleccionamos **Stat** \Rightarrow **Basic Statistics** \Rightarrow **2-Sample t**. Activamos la opción **Samples in one column**, con lo cual se desactivan automáticamente las opciones **Samples in different columns** y **Summarized data**. En **Samples** seleccionamos, de la lista de variables de la izquierda, la columna ‘Pulse1’; en **Subscripts** seleccionamos, de la lista de la izquierda, la columna ‘Sex’; y activamos **Assume equal variances** ya que hemos comprobado que las varianzas poblacionales son desconocidas pero iguales. Si pulsamos el botón **Options** nos aparece un nuevo cuadro de diálogo con las siguientes opciones:

Confidence level: Por defecto se muestra un intervalo de confianza al 95 % para la diferencia de medias poblacionales, $\mu_1 - \mu_2$. Se puede introducir un valor entre 1 y 99 para solicitar otro nivel de confianza. En nuestro ejemplo, podemos dejar lo que aparece por defecto, es decir, 95.

Test difference: Aquí se pone el valor con el que se compara la diferencia de medias poblacionales, μ_0 . La hipótesis nula $H_0 : \mu_1 = \mu_2$ es equivalente a $H_0 : \mu_1 - \mu_2 = 0$, por lo que el valor con el que se compara la diferencia de medias poblacionales, en este ejemplo, es cero; es decir, $\mu_0 = 0$. En consecuencia, nosotros dejamos lo que aparece por defecto (cero).

Alternative: Aquí se especifica cuál es la hipótesis alternativa: **less than** significa que la hipótesis alternativa es $H_1 : \mu_1 - \mu_2 < \mu_0$, **not equal** significa que la hipótesis alternativa es $H_1 : \mu_1 - \mu_2 \neq \mu_0$ y **greater than** significa que la hipótesis alternativa es $H_1 : \mu_1 - \mu_2 > \mu_0$. Tengamos en cuenta que con la opción **less than** el intervalo de confianza para $\mu_1 - \mu_2$ será del tipo $(-\infty, b)$, con la opción **not equal** el intervalo de confianza será del tipo (a, b) y con la opción **greater than** el intervalo de confianza será del tipo $(a, +\infty)$. En nuestro ejemplo, tenemos que dejar lo que aparece por defecto, que es **not equal**, ya que la hipótesis alternativa es $H_1 : \mu_1 \neq \mu_2$, que es equivalente a $H_1 : \mu_1 - \mu_2 \neq 0$.

Podemos comprobar, en la ventana de sesión, que el p-valor es 0’006, claramente menor que el nivel de significación, $\alpha = 0’05$, por lo que debemos rechazar la hipótesis nula y, por tanto, aceptar la hipótesis alternativa. Aceptamos que el pulso medio poblacional de los hombres antes de correr es distinto del pulso medio poblacional de las mujeres antes de correr. Como la media muestral del pulso de las mujeres antes de correr (76’9) es mayor que la media muestral del pulso de los hombres antes de correr (70’42) podríamos, incluso, aceptar que la media poblacional del pulso de las mujeres antes de correr es mayor que la media poblacional del pulso de los hombres antes de correr. El intervalo de confianza al 95 % para la diferencia de medias poblacionales, $\mu_1 - \mu_2$, es $(-10’96, -1’91)$.

También se puede realizar este contraste de hipótesis si sabemos los dos tamaños muestrales, los resultados de las dos medias muestrales y los resultados de las dos cuasi-desviaciones típicas muestrales. Veámoslo con un nuevo ejemplo:

Con los datos del **Ejemplo 3** (de la sección 6.1) queremos decidir si existe diferencia significativa entre el número medio de horas semanales que permanecen en la biblioteca los socios y los no socios.

Como en dicho ejemplo hemos decidido aceptar que no existe diferencia significativa entre las varianzas poblacionales, entonces nos encontramos ante un contraste de comparación de dos medias poblacionales, con muestras independientes y varianzas poblacionales desconocidas pero iguales. Realizaremos el siguiente contraste de hipótesis:

$$H_0 : \mu_1 = \mu_2 ,$$

$$H_1 : \mu_1 \neq \mu_2 .$$

Los datos son:

$$n_1 = 21 , \quad \bar{X}_1 = 10 , \quad S_1 = 3 ,$$

$$n_2 = 16 , \quad \bar{X}_2 = 6 , \quad S_2 = 2 .$$

Seleccionamos la opción **Stat** \Rightarrow **Basic Statistics** \Rightarrow **2-Sample t**. Activamos la opción **Summarized data**, con lo cual se desactivan automáticamente las opciones **Samples in one column** y **Samples in different columns**. Dentro de **First**, en **Sample size** tenemos que teclear el tamaño muestral de la primera muestra, que es **21**, en **Mean** tenemos que teclear el resultado de la media de la primera muestra, que es **10**, y en **Standard deviation** tenemos que teclear el resultado de la cuasi-desviación típica de la primera muestra, que es **3**. Dentro de **Second**, en **Sample size** tenemos que teclear el tamaño muestral de la segunda muestra, que es **16**, en **Mean** tenemos que teclear el resultado de la media de la segunda muestra, que es **6**, y en **Standard deviation** tenemos que teclear el resultado de la cuasi-desviación típica de la segunda muestra, que es **2**. Activamos **Assume equal variances** ya que hemos comprobado (en el **Ejemplo 3**, como ya hemos dicho) que las varianzas poblacionales son desconocidas pero iguales. Pulsamos en **Options** y en el cuadro de diálogo resultante dejamos lo que aparece por defecto (**Confidence level: 95**, **Test difference: 0**, **Alternative: not equal**).

Podemos comprobar, en la ventana de sesión, que el p-valor es 0, el mínimo posible y, por supuesto, menor que los niveles de significación usuales ($\alpha = 0'05$ ó $\alpha = 0'01$), por lo que debemos rechazar la hipótesis nula. Aceptamos, en consecuencia, que existe diferencia significativa entre el número medio de horas semanales que permanecen en la biblioteca los socios y los no socios. Como la media muestral del número de horas semanales que permanecen en la biblioteca los socios (10) es mayor que la media muestral del número de horas semanales que permanecen en la biblioteca los no socios (6) podríamos, incluso, aceptar que la media poblacional del número de horas semanales que permanecen en la biblioteca los socios es mayor que la media poblacional del número de horas semanales que permanecen en la biblioteca los no socios. El intervalo de confianza al 95 % para la diferencia de medias poblacionales, $\mu_1 - \mu_2$, es (2'326, 5'674).

6.2.2. Comparación de dos medias con muestras independientes y varianzas poblacionales desconocidas y distintas

Igual que en el apartado anterior, este procedimiento solamente es válido cuando las dos muestras son aleatorias y las dos poblaciones son Normales o los dos tamaños muestrales son grandes (en la práctica $n_1, n_2 \geq 30$).

Para realizar este test paramétrico hay que seleccionar, igual que antes, **Stat** \Rightarrow **Basic Statistics** \Rightarrow **2-Sample t**. Hay que rellenar el cuadro de diálogo de manera similar al apartado anterior, con la salvedad de que, en este caso, hay que desactivar la opción **Assume equal variances**.

Con el archivo de datos **Pulse.mtw**, comprobemos si se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que el pulso medio poblacional de los hombres después de correr es igual al pulso medio poblacional de las mujeres después de correr. Queremos comparar la media poblacional de la variable **Pulse2** para los grupos en los que la variable **Sex** vale **1** (Hombre) y **2** (Mujer). El contraste que tenemos que hacer es $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$, siendo $X_1 = \text{“Pulso de los hombres después de correr”}$ y $X_2 = \text{“Pulso de las mujeres después de correr”}$.

En el **Ejemplo 2** de la sección 6.1 hemos comprobado que se puede aceptar que la varianza poblacional del pulso de los hombres después de correr es distinta de la varianza poblacional del pulso de las mujeres después de correr. Por tanto, nos encontramos ante un contraste de comparación de dos medias poblacionales, con muestras independientes y varianzas poblacionales desconocidas y distintas. Aunque las variables aleatorias X_1 y X_2 no fuesen Normales, se puede aplicar este contraste debido a que los tamaños muestrales son suficientemente grandes: $n_1 = 57$ y $n_2 = 35$.

Para hacer el contraste seleccionamos **Stat** \Rightarrow **Basic Statistics** \Rightarrow **2-Sample t**. Activamos la opción **Samples in one column**, con lo cual se desactivan automáticamente las opciones **Samples in different columns** y **Summarized data**. En **Samples** seleccionamos, de la lista de variables de la izquierda, la columna '**Pulse2**'; y en **Subscripts** seleccionamos, de la lista de la izquierda, la columna '**Sex**'. Si se pulsa el botón **Options** aparece un cuadro de diálogo similar al ejemplo anterior. En este cuadro de diálogo dejamos lo que aparece por defecto (**Confidence level: 95**, **Test difference: 0**, **Alternative: not equal**).

Podemos comprobar, en la ventana de sesión, que el p-valor es 0'007, claramente menor que el nivel de significación, $\alpha = 0'05$, por lo que debemos rechazar la hipótesis nula y, por tanto, aceptar la hipótesis alternativa. Aceptamos que el pulso medio poblacional de los hombres después de correr es distinto del pulso medio poblacional de las mujeres después de correr. Como la media muestral del pulso de las mujeres después de correr (86'7) es mayor que la media muestral del pulso de los hombres después de correr (75'9) podríamos, incluso, aceptar que la media poblacional del pulso de las mujeres después de correr es mayor que la media poblacional del pulso de los hombres después de correr. El intervalo de confianza al 95 % para la diferencia de medias poblacionales, $\mu_1 - \mu_2$, es $(-18'65, -3'02)$.

6.3. Comparación de dos medias con muestras apareadas

Este procedimiento solamente es válido cuando las dos muestras son aleatorias y la variable aleatoria diferencia, $D = X_1 - X_2$, es Normal o el tamaño muestral común, n , es grande (en la práctica, $n \geq 30$).

Para realizar este test paramétrico hay que seleccionar **Stat** \Rightarrow **Basic Statistics** \Rightarrow **Paired t**.

Con el archivo de datos **Pulse.mtw**, comprobemos si se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que el pulso medio poblacional antes de correr es igual al pulso medio poblacional después de correr. Lo que se quiere es comparar la media poblacional de la variable **Pulse1** con la media poblacional de la variable **Pulse2**. El contraste que tenemos que hacer es $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$, siendo $X_1 = \text{“Pulso antes de correr”}$ y $X_2 = \text{“Pulso después de correr”}$. Como las dos variables están observadas en los mismos individuos, podemos afirmar que las muestras están relacionadas; es decir, son apareadas o asociadas. Por tanto, nos encontramos ante un contraste de comparación de dos medias poblacionales con muestras apareadas. Aunque la variable aleatoria di-

ferencia, $D = X_1 - X_2$, no fuese Normal, se puede aplicar este contraste debido a que los tamaños muestrales son suficientemente grandes: $n_1 = n_2 = n = 92$.

Para hacer este contraste seleccionamos **Stat** \Rightarrow **Basic Statistics** \Rightarrow **Paired t**. Activamos la opción **Samples in columns**; en **First sample** seleccionamos, de la lista de variables de la izquierda, la columna 'Pulse1'; en **Second sample** seleccionamos, de la lista de variables de la izquierda, la columna 'Pulse2'. Si pulsamos el botón **Options** nos aparece un cuadro de diálogo similar al de la opción anterior (**2-Sample t** \Rightarrow **Options**). En este cuadro de diálogo dejamos lo que aparece por defecto (**Confidence level: 95**, **Test difference: 0**, **Alternative: not equal**).

Podemos comprobar, en la ventana de sesión, que el p-valor es igual a 0, el mínimo posible y, por supuesto, menor que el nivel de significación, $\alpha = 0'05$, por lo que debemos rechazar la hipótesis nula y, por tanto, aceptar la hipótesis alternativa. Aceptamos, por tanto, que el pulso medio poblacional antes de correr es distinto del pulso medio poblacional después de correr. Como la media muestral del pulso después de correr (80'00) es mayor que la media muestral del pulso antes de correr (72'87) podríamos, incluso, aceptar que la media poblacional del pulso después de correr es mayor que la media poblacional del pulso antes de correr. El intervalo de confianza al 95 % para la diferencia de medias poblacionales, en este caso, es $(-9'92, -4'34)$.

6.4. Ejercicios propuestos

Ejercicio 6.1

- a) Crea un nuevo proyecto de *Minitab*.
- b) Abre la hoja de datos **Transacciones.mtw** (datos del Ejercicio 1.2).
- c) Utilizando el test de Levene, ¿se puede aceptar, con un nivel de significación de 0'05, que la varianza poblacional del número anual de transacciones de referencia de las bibliotecas públicas es igual a la varianza poblacional del número anual de transacciones de referencia de las bibliotecas universitarias? ¿Por qué?
- d) ¿Se puede aceptar, con un nivel de significación de 0'05, que la media poblacional del número anual de transacciones de referencia de las bibliotecas públicas es igual a la media poblacional del número anual de transacciones de referencia de las bibliotecas universitarias? ¿Por qué?
- e) Utilizando el test F de Snedecor, ¿se puede aceptar, con un nivel de significación de 0'05, que la varianza poblacional del porcentaje de transacciones de referencia finalizadas de las bibliotecas públicas es igual a la varianza poblacional del porcentaje de transacciones de referencia finalizadas de las bibliotecas universitarias? ¿Por qué?
- f) ¿Se puede aceptar, con un nivel de significación de 0'05, que la media poblacional del porcentaje de transacciones de referencia finalizadas de las bibliotecas públicas es igual a la media poblacional del porcentaje de transacciones de referencia finalizadas de las bibliotecas universitarias? ¿Por qué?
- g) Graba el proyecto con el siguiente nombre: **Ejercicio6-1.mpj**

Ejercicio 6.2 En la tabla siguiente aparece el precio, en euros, de una muestra aleatoria de 15 libros que se prestan pocas veces (X_1) y el precio, en euros, de una muestra aleatoria de 15 libros que se prestan muchas veces (X_2).

x_{1i}	x_{2i}
75	110
32	30
30	45
34	69
42	46
57	53
51	97
36	43
82	42
45	37
58	48
66	45
40	105
35	61
51	57

- Crea un nuevo proyecto de **Minitab**.
- Guarda los datos en el archivo **PrecioLibros.mtw**
- ¿Se puede aceptar, con un nivel de significación de 0'05, que la varianza poblacional del precio de los libros que se prestan poco es igual a la varianza poblacional del precio de los libros que se prestan mucho? ¿Por qué?
- ¿Se puede aceptar, con un nivel de significación de 0'05, que la media poblacional del precio de los libros que se prestan poco es igual a la media poblacional del precio de los libros que se prestan mucho? ¿Por qué?
- Graba el proyecto con el siguiente nombre: **Ejercicio6-2.mpj**

Ejercicio 6.3 En la siguiente tabla aparece el número de palabras por resumen de una muestra aleatoria de 30 artículos científicos escritos en francés (X_1) y el número de palabras por resumen de una muestra aleatoria de 30 artículos científicos escritos en inglés (X_2).

x_{1i}	70	65	68	74	79	67	75	80	62	69
	61	57	71	74	82	91	70	64	72	67
	74	70	81	85	70	74	75	71	69	54
x_{2i}	80	47	59	67	89	57	72	78	74	72
	104	118	89	87	79	78	101	120	107	95
	85	87	90	98	89	75	90	101	85	94

- Crea un nuevo proyecto de **Minitab**.
- Guarda los datos en el archivo **LongitudResumenes.mtw**

- c) ¿Se puede aceptar, con un nivel de significación de 0'05, que la varianza poblacional de la longitud de los resúmenes de artículos escritos en francés es igual a la varianza poblacional de la longitud de los resúmenes de artículos escritos en inglés? ¿Por qué?
- d) ¿Se puede aceptar, con un nivel de significación de 0'05, que la media poblacional de la longitud de los resúmenes de artículos escritos en francés es igual a la media poblacional de la longitud de los resúmenes de artículos escritos en inglés? ¿Por qué?
- e) Graba el proyecto con el siguiente nombre: **Ejercicio6-3.mpj**

Ejercicio 6.4 Dos expertos califican una muestra aleatoria de 30 libros según su calidad (1=muy mala, 2=mala, 3=regular, 4=buena, 5=muy buena). En la tabla siguiente aparece la opinión del primer experto (X_1) y la opinión del segundo experto (X_2).

x_{1i}	x_{2i}	x_{1i}	x_{2i}
2	1	4	4
5	4	4	3
4	5	5	4
2	3	5	3
3	3	1	2
1	5	2	5
3	3	2	3
1	3	3	2
4	2	4	1
2	5	4	2
3	2	1	3
4	3	2	4
3	3	1	2
1	3	5	5
2	5	5	2

- a) Crea un nuevo proyecto de *Minitab*.
- b) Guarda los datos en el archivo **Opinion.mtw**
- c) ¿Se puede aceptar, con un nivel de significación de 0'05, que la media poblacional de los resultados de la opinión del primer experto es igual a la media poblacional de los resultados de la opinión del segundo experto? ¿Por qué?
- d) Graba el proyecto con el siguiente nombre: **Ejercicio6-4.mpj**

Ejercicio 6.5 Elegimos al azar 30 matrimonios y observamos el número de veces que los hombres han visitado alguna biblioteca en los tres últimos meses (X_1) y el número de veces que las mujeres han visitado alguna biblioteca en los tres últimos meses (X_2). Los resultados se muestran en la siguiente tabla.

x_{1i}	x_{2i}	x_{1i}	x_{2i}	x_{1i}	x_{2i}
12	8	8	10	25	14
30	11	14	15	12	16
10	12	20	12	8	10
20	16	13	19	23	20
15	10	11	6	14	17
14	9	7	7	8	10
11	12	6	7	12	23
9	10	8	6	27	10
7	7	15	20	32	27
5	4	42	35	14	18

- Crea un nuevo proyecto de **Minitab**.
- Guarda los datos en el archivo **VisitasBiblioteca.mtw**
- ¿Podemos afirmar que hay diferencia significativa entre los hombres y las mujeres de los matrimonios en cuanto al número de veces que van a la biblioteca? ¿Por qué?
- Graba el proyecto con el siguiente nombre: **Ejercicio6-5.mpj**

Ejercicio 6.6 En la siguiente tabla aparece el número de usuarios diarios de la biblioteca A (variable X_1) y el número de usuarios diarios de la biblioteca B (variable X_2) en 10 días elegidos al azar.

x_{1i}	x_{2i}
51	45
72	58
35	32
70	56
75	68
98	76
100	88
80	69
72	57
90	75

- Crea un nuevo proyecto de **Minitab**.
- Guarda los datos en el archivo **UsuariosDiarios.mtw**
- Calcula, en una nueva columna, los resultados de la variable diferencia $D = X_1 - X_2$.
- ¿Se puede aceptar, con un nivel de significación de 0'05, que la muestra de las diferencias, $d_i = x_{1i} - x_{2i}$, es aleatoria? ¿Por qué?
- ¿Se puede aceptar, con un nivel de significación de 0'05, que la variable diferencia, $D = X_1 - X_2$, es Normal? ¿Por qué?

- f) ¿Se puede aceptar, con un nivel de significación de 0'05, que la media poblacional del número de usuarios diarios de la biblioteca A es igual a la media poblacional del número de usuarios diarios de la biblioteca B? ¿Por qué?
- g) Graba el proyecto con el siguiente nombre: **Ejercicio6-6.mpj**