Manual de MINITAB 14 para Windows

Autora: Dra. Josefa Marín Fernández
Departamento de Estadística e Investigación Operativa
Facultad de Matemáticas
Universidad de Murcia

Diciembre de 2006

Índice de contenidos

1. Introducción a Minitab 14 para Windows

1.1. Elementos de Minitab 14 para Windows

Al ejecutar ***Minitab*** 14 aparece la pantalla de la Figura 1.

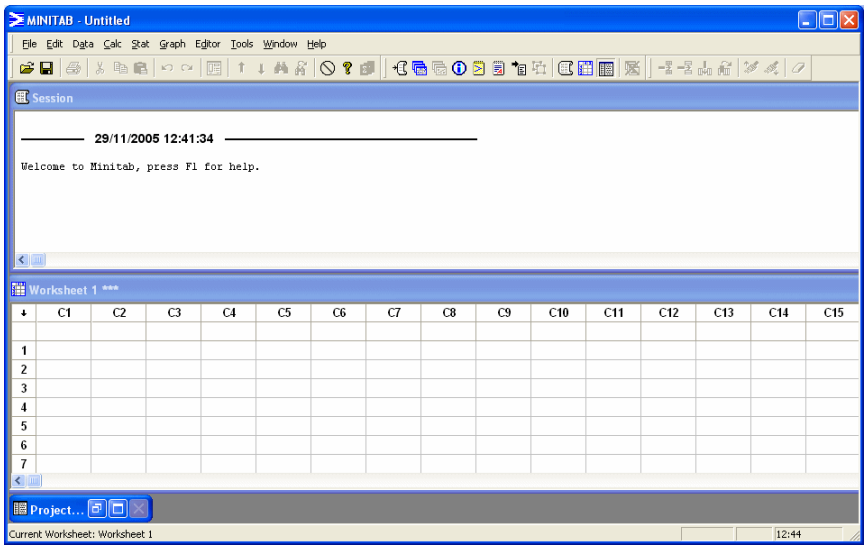

Figura 1: Pantalla inicial de Minitab 14

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

- En la primera línea aparece la **barra de título** con el nombre de la ventana y los botones de minimizar, maximizar y cerrar.
- En la segunda línea está la **barra de menús** con los 10 menús que luego comentaremos.
- La tercera línea es 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 $C1, C2, \ldots$, tal como está escrito, pero también se les puede dar un nombre, escribiéndolo debajo de $C1, C2, \ldots$ Cada columna es una variable y cada fila corresponde a una observación o caso.
- En la parte inferior aparece (minimizada) la **ventana de proyecto (Proyect 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, podemos pulsar Ctrl+m o podemos hacer *clic* sobre su icono en la barra de herramientas (primer icono de la Figura 2). Para activar la hoja de datos (**Worksheet**) podemos hacer *clic* sobre ella, podemos pulsar Ctrl+d o podemos hacer *clic* sobre su icono en la barra de herramientas (segundo icono de la Figura 2). Para activar la ventana de proyecto (**Proyect Manager**) podemos maximizarla, podemos pulsar Ctrl+i 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

1.2. 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 ventana de datos. En la parte superior aparece $C1$, $C2$, $C3$, ... 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, ↓; si se hace *clic* sobre ella, apuntará hacia la derecha, →. 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 la variable es numérica basta con escribir el resultado de dicha variable para cada individuo de la muestra. Si la variable no es numérica, una vez que hemos hecho *clic* sobre CJ (o sobre el nombre que le hayamos puesto a la variable) hay que seleccionar la opción **Editor⇒Format Column** y elegir el tipo de variable.

Por ejemplo, podemos introducir los datos de la Figura 3, correspondientes a las calificaciones de una muestra de 8 alumnos en un determinado examen y el tiempo empleado en realizar dicho examen.

Worksheet 1 ***

↓	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

1.3. Grabación de datos

Una vez introducidos los datos, éstos pueden guardarse en un fichero para poder ser utilizados en cualquier otro momento. En realidad, los datos deberían guardarse muy a menudo, no sólo cuando hayamos terminado de introducirlos todos. ¿Qué pasaría si tenemos que introducir 1000 datos y cuando ya hemos introducido 950 se produce un corte de energía eléctrica?

Para guardar únicamente la ventana de datos hay que seleccionar **File⇒Save Current Worksheet** ó **File⇒Save Current Worksheet As**. Por ejemplo, podemos guardar los datos de la Figura 3 en un archivo que denominaremos **Notas_Tiempo.mtw**.

Si queremos guardar toda la información actual del programa (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** ó **File⇒Save Project As**. Es muy importante diferenciar entre ficheros de datos (**.mtw**) y ficheros 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**.

1.4. 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 Proyect**.

()	Paréntesis
**	Exponenciación
*	Multiplicación
/	División
+	Suma
−	Resta

(a) Operadores aritméticos

<	Menor que
>	Mayor que
<=	Menor o igual que
>=	Mayor o igual que
=	Igual que
<>	No igual que

(b) Operadores relacionales

AND	Operador Y
OR	Operador O
NOT	Operador NO

(c) Operadores lógicos

Cuadro 1: Operaciones aritméticas, relacionales y lógicas

Normalmente los ficheros de datos de ***Minitab*** 14 se encuentran en **C:\Archivos de programa\Minitab 14 \Data** y, como ya sabemos, llevan la extensión **.mtw**.

Por ejemplo, podemos abrir el fichero 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 fichero 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 fichero.

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 tener información más detallada sobre el tipo de ficheros que se pueden leer, se puede seleccionar **File⇒Open Workshhet** y, en el cuadro de diálogo resultante, se hace *clic* sobre **Ayuda**.

1.5. Opciones principales del menú *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 en los siguientes apartados, activamos la ventana de sesión y luego seleccionamos **Editor⇒Enable Commands**.

1.5.1. Operaciones por filas mediante la opción *Calc⇒Calculator*

En este apartado vamos a ver el modo de generar nuevas variables mediante transformaciones efectuadas sobre los valores de las variables ya definidas.

Para practicar esta opción tendremos abierto el fichero de datos **Pulse.mtw**.

En el Cuadro 1 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 el Cuadro 1 (de arriba a abajo).

Para construir una nueva variable mediante transformaciones de otras ya existentes, se tiene que elegir la opción **Calc ⇒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 fichero 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 (activando el cuadro 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 fichero de datos **Pulse.mtw** vamos a calcular la media geométrica de las variables **Pulse1** y **Pulse2** (raíz cuadrada del producto de ambas variables). Para ello, seleccionamos la opción **Calc⇒Calculator**; en **Store result in variable** tenemos que teclear la posición de la columna que contendrá los resultados; por ejemplo, **C4** (si dicha columna está vacía) o el nombre que queremos darle a dicha columna. Si el nombre contiene espacios en blanco, hay que escribirlo entre comillas simples; por ejemplo, vamos a denominar a la nueva variable **'media geométrica Pulse1 Pulse2'**. En **Expression** tenemos que colocar (utilizando, como hemos dicho, la calculadora y la lista de variables) la

operación que se realiza para determinar la media geométrica indicada: `('Pulse1' * 'Pulse2')**(1 / 2)`. Por último, pulsamos en **OK**.

1.5.2. Operaciones por columnas mediante la opción *Calc⇒Column Statistics*

La opción **Calc⇒Column Statistics** calcula, para una columna (o variable), **uno** de los estadísticos siguientes:

Sum	suma	$\sum_{i=1}^{n} x_i$
Mean	media aritmética	$\overline{x} = \left(\sum_{i=1}^{n} x_i\right)/n$
Standard deviation	cuasidesviación típica	$S = \sqrt{\left(\sum_{i=1}^{n} (x_i - \overline{x})^2\right)/(n-1)}$
Minimum	mínimo dato	x_{min}
Maximum	máximo dato	x_{max}
Range	recorrido total	$R = x_{max} - x_{min}$
Median	mediana=valor que deja por debajo de él el 50 % de los datos	
Sum of squares	suma de cuadrados	$\sum_{i=1}^{n} x_i^2$
N total	número total de casos=N nonmissing+N missing	
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	

El resultado del estadístico calculado se puede almacenar (opcionalmente) en una constante, si lo indicamos en **Store result in**.

Por ejemplo, del fichero de datos **Pulse.mtw** vamos a determinar la mediana de los datos de la columna **Height** y vamos a guardar el resultado en una constante que vamos a denominar **Mediana de altura**. Para ello, seleccionamos **Calc⇒Column Statistics**; activamos la opción **Median**; hacemos *clic* en el recuadro que hay a la derecha de **Input variable** y seleccionamos (haciendo doble *clic* sobre su nombre) la columna **Height**; en **Store result in** tecleamos **'Mediana de altura'** y pulsamos en **OK**. ***Minitab*** guarda esta constante también como $K1$. Esta constante se puede consultar, en cualquier momento, en la ventana **Proyect Manager** (concretamente en **Worksheets\Pulse.mtw\constants**) y puede ser utilizada en cálculos posteriores.

1.5.3. Operaciones por filas mediante la opción *Calc⇒Row Statistics*

La opción **Calc⇒Row Statistics** calcula los mismos estadísticos del apartado anterior, pero por filas, en vez de por columnas. En este caso, a diferencia del anterior, es totalmente necesario rellenar el recuadro **Store result in** ya que los resultados forman una nueva variable o columna.

Por ejemplo, del fichero de datos **Pulse.mtw** vamos a hallar la media aritmética (por filas) de la variables **Pulse1** y **Pulse2** y guardar los resultados en una nueva columna (variable) denominada **Media aritmética Pulse1 Pulse2**. Para ello, seleccionamos **Calc⇒Row Statistics**; activamos la opción **Mean**; hacemos *clic* en el recuadro que hay debajo de **Input variables** y seleccionamos (haciendo doble *clic* sobre sus nombres) las columnas **Pulse1** y **Pulse2**; en **Store result in** tecleamos **'Media aritmética Pulse1 Pulse2'** y pulsamos en **OK**.

Las operaciones realizadas con esta opción también pueden realizarse mediante **Calc⇒Calculator**.

1.5.4. Tipificación de datos

Con la opción **Calc⇒Standardize** se calcula, en una nueva columna o variable, los datos tipificados o estandarizados de una de las columnas de nuestra hoja de datos. Hay varias formas de tipificar los datos pero la más usual es la siguiente: Si x_i son los datos de la muestra, $\overline{x}$ es la media y S es la cuasidesviación típica o desviación típica corregida, los datos tipificados o estandarizados son $y_i = (x_i - \overline{x})/S$. Esto se logra dejando activada la opción **subtract mean and divide by std. dev.**.

Por ejemplo, vamos a crear una nueva variable (columna), designada por **Pulse1 Tipificada**, que contenga los datos de **Pulse1** tipificados o estandarizados. Para ello, seleccionamos **Calc⇒Standardize**; en **Input columns** seleccionamos (haciendo doble *clic* sobre su nombre) la columna **Pulse1**; en **Store results in** tecleamos **'Pulse1 Tipificada'**; dejamos activada la opción **Substract mean and divide by std. dev.** y pulsamos en **OK**.

Las operaciones realizadas con esta opción también pueden realizarse mediante **Calc⇒Calculator**.

1.5.5. 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 fichero **Pulse.mtw**, abrimos una nueva hoja de datos con la opción **File⇒New**. En el cuadro de diálogo que aparece seleccionamos **Minitab Woorksheet**. A esta nueva hoja de datos ***Minitab*** le asignará el nombre Worksheet J, siendo J un número natural. Luego podemos 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'**. 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**. Una vez obtenida la nueva columna vamos a denominar **Ejemplo_Practica_1.mtw** a la nueva hoja de datos utilizando la opción **File⇒Save Current Worksheet As**.

1.5.6. Creación de resultados aleatorios de una distribución conocida

En ***Minitab*** podemos generar datos de distribuciones usuales utilizando la opción **Calc⇒Random Data**.

Por ejemplo, en el fichero de datos **Ejemplo_Practica_1.mtw** vamos a generar 100 datos de una distribución Uniforme en el intervalo $(0, 1)$ (100 números aleatorios comprendidos entre 0 y 1). Para ello, seleccionamos la opción **Calc⇒Random Data⇒Uniform**; en el cuadro de diálogo resultante ponemos **Generate** 100 **rows of data**; en **Store in column** escribimos el nombre de la nueva columna: **'100 datos de U(0,1)'**; en **Lower endpoint** tecleamos **0** y en **Upper endpoint** escribimos **1**.

Esta opción será utilizada en posteriores prácticas.

1.6. Opciones principales del menú *Data*

Sólo se explicarán algunas de las opciones más utilizadas del menú **Data**. En el cuadro de diálogo de cada opción existe un botón **Help** que la explica bastante bien.

1.6.1. 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.

Para practicar esta opción vamos a apilar los datos de la columna **Patrón entre 0 y 1** y de la columna **100 datos de U(0,1)** del fichero de datos **Ejemplo_Practica_1.mtw**. Para ello, 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: **'Patrón entre 0 y 1'** **'100 datos de U(0,1)'**; en **Store stacked data in** activamos la opción **Column of current worksheet** y tecleamos la posición de una columna que esté vacía, por ejemplo, **C3**. En **Store subscripts in** tecleamos la posición de la columna en la que queremos guardar la procedencia de cada dato, por ejemplo, **C4**. Es conveniente dejar activada la opción **Use variable names in subscript column**.

1.6.2. Desapilamiento de columnas

La opción **Data⇒Unstack columns** permite separar una columna en varias según los valores de la columna de alguna variable (que contiene los subíndices). Esta opción es la contraria de la explicada en el apartado anterior.

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** (*¿corrió o no?*). Para ello, seleccionamos **Data⇒Unstack Columns**; en **Unstack the data in** seleccionamos (haciendo doble *clic* sobre su nombre) la variable o columna **Pulse2**; en **Using subscripts in** seleccionamos (haciendo doble *clic* sobre su nombre) 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** y dejamos activado **Name the columns containing the unstaked data**.

1.6.3. Ordenación de los 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.mtw** vamos a crear una nueva variable (columna), designada por **Pulse1 ordenado**, que contenga los resultados de la variable **Pulse1** ordenados de menor a mayor. Para ello, seleccionamos **Data⇒Sort**;

en **Sort column** seleccionamos (haciendo doble *clic* sobre su nombre) la variable **Pulse1**; en **By column** volvemos a seleccionar la misma columna. Si dejamos desactivada la opción **Descending** la ordenación se hará de menor a mayor resultado, que es lo que queremos. En **Store sorted data in** activamos **Column of current worksheet** y tecleamos el nombre que queremos ponerle a dicha columna: **'Pulse1 ordenado'**.

Tenemos que 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.

1.6.4. Ordenación por rangos

La opción **Data⇒Rank** crea una nueva columna que indica la posición que ocuparía cada dato si los ordenáramos de menor a mayor. Cuando dos o más valores de la columna son iguales (empates) se asigna a cada uno de ellos el rango medio de los rangos que tendrían si fueran distintos. Por ejemplo, si los dos resultados más pequeños estuviesen empatados, en principio ocuparían los números de orden 1 y 2; pero al estar empatados, los rangos de los dos valores coinciden entre sí y coinciden con $(1+2)/2 = 1'5$.

Con la hoja de datos **Pulse.mtw** podemos practicar esta opción creando una nueva columna, que denominaremos **Rangos de Pulse1**, en la cual aparecerá la posición que ocuparía cada resultado de la variable **Pulse1** si los ordenásemos de menor a mayor (con la corrección mencionada por empates). Para ello, seleccionamos **Data⇒Rank**; en **Rank data in** elegimos (haciendo doble *clic* sobre su nombre) la columna **Pulse1** y en **Store ranks in** escribimos **'Rangos de Pulse1'**.

El primer resultado de **Rangos de Pulse1** es igual a $22'5$ porque el valor 64 (observación primera de la variable **Pulse1**) ha aparecido 4 veces (casos numerados con el 1, 5, 49 y 71 de la variable **Pulse1**) y estos valores ocuparían los números de orden 21, 22, 23 y 24; pero como están empatados se les asigna el mismo rango: la media aritmética de estos cuatro rangos; es decir, $(21+22+23+24)/4 = 22'5$.

1.6.5. 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.mtw** podemos codificar la variable **Pulse1** de la forma siguiente:

intervalo de Pulse1	nueva categoría
[48,65]	Pulso bajo
(65,83]	Pulso medio
(83,100]	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 **Into columns** escribimos el nombre la nueva variable, por ejemplo, **'Codificación de Pulse1'**. En la primera línea de **Original values** escribimos **48:65** (todos los resultados comprendidos entre 48, incluido, y 65, incluido) y en la primera línea de **New** escribimos **Pulso bajo**. En la segunda línea de **Original values** escribimos **65:83** (todos los resultados comprendidos entre 65, sin incluir, y 83, incluido) y en la segunda línea de **New** escribimos **Pulso medio**. En la tercera línea de **Original values** escribimos **83:100** (todos los resultados comprendidos entre 83, sin incluir, y 100, incluido) y en la tercera línea de **New** escribimos **Pulso alto**.

1.7. Algo más sobre la ventana *Session*

Ya hemos visto que una de las utilidades de la ventana de sesión es la de servir para la presentación de los comandos aplicados en cada opción de las que hemos realizado. Además, podemos repasar resultados obtenidos con anterioridad moviéndonos hacia arriba en dicha ventana. También hemos visto que para activar la ventana de sesión (**Session**) podemos hacer *clic* sobre ella, podemos pulsar Ctrl+m o podemos hacer *clic* sobre su icono en la barra de herramientas. Los resultados incluidos en la ventana de sesión pueden grabarse como un fichero de texto (**txt**) activando dicha ventana y seleccionando **File⇒Save Session Window As**. También podemos usar las opciones de marcar, copiar y pegar para pasar los resultados obtenidos a editores de texto. Además, es posible imprimir todos sus contenidos activando dicha ventana y seleccionando **File⇒Print Session Window**.

Una vez seleccionada la ventana de sesión, la activación de la opción **Editor⇒Enable Commands** permite ejecutar los comandos de ***Minitab***. Por ejemplo, si tecleamos en la ventana de sesión (tras **MTB** >) **Mean C1** y pulsamos el botón Intro, el programa calcula media aritmética de los datos de la columna $C1$ de la hoja de datos activa. Si escribimos

Let K2=1/3 y pulsamos el botón Intro, el programa guarda el valor 1/3 en la correspondiente constante. Si tecleamos ahora **Print K2**, el programa nos da el valor de dicha constante.

Lógicamente, es más sencillo el manejo de ***Minitab*** utilizando los menús, pero los comandos pueden incorporarse posteriormente a los programas (macros) que construyamos. Además, una vez habilitado el lenguaje de comandos, cuando ejecutemos una opción del menú, ésta se escribirá en la ventana de sesión, con lo que podremos ver cuál es la sintaxis concreta del comando que queremos utilizar.

Para que el contenido de la ventana de sesión pueda modificarse, debemos activar dicha ventana y seleccionar **Editor⇒Output Editable**, con lo que podemos rectificar fácilmente cualquier error, modificar comandos ejecutados anteriormente o simplemente preparar los resultados para ser imprimidos.

Una vez activada la opción **Editor⇒Output Editable**, la ventana de sesión es el lugar en el que se ejecutan los *macros* o programas, tanto los que construyamos nosotros como los que incluye ***Minitab*** o los realizadas por otros usuarios. Los *macros* llevan la extensión **.mac** y normalmente están incluidos en el directorio **C:\Archivos de programa\Minitab 14\Macros**. En la versión 14 de ***Minitab*** solamente se incluyen cuatro *macros*, pues los resultados del resto de los macros de la versión anterior pueden conseguirse con distintas opciones de los menús.

1.8. Algo más sobre la ventana *Proyect Manager*

Ya sabemos que para activar la ventana de proyecto (**Proyect Manager**) podemos maximizarla, podemos pulsar Ctrl+i o podemos hacer *clic* sobre su icono en la barra de herramientas.

Esta ventana presenta toda la información disponible en forma de directorios. Resulta ser especialmente útil cuando se maneja una gran cantidad de datos. El directorio **Session** nos muestra, de forma resumida y organizada, la información correspondiente a dicha ventana. El directorio **History** presenta (en lenguaje de comandos) todas las operaciones que hemos realizado. A diferencia de lo que ocurría con la ventana de sesión, no sirve para ejecutar comandos ni macros, y en él no se muestran los resultados de la ejecución de los comandos. En este directorio aparece solamente el programa de las operaciones que hemos realizado, y su contenido puede consultarse o copiarse directamente para la realización de macros. Los directorios de datos, **Worksheets**, contienen información sobre las columnas (variables), constantes y matrices manejadas en cada ventana de datos que se esté utilizando. Además, indican el número de datos incluidos en una columna, así como los datos ausentes de la misma (Missing).

1.9. Ejercicios propuestos

1.1. Con la hoja de datos **Pulse.mtw** haz lo siguiente:

- **a)** Crea una nueva variable, designada por **Sexo**, que contenga los datos de la variable **Sex** pero cuyos resultados aparezcan con las palabras **Hombre** (en vez de 1) y **Mujer** (en vez de 2).
- **b)** Desapila los resultados de la variable **Pulse1** según los resultados de la variable **Sexo**. Calcula la media aritmética de estas dos nuevas columnas. Interpreta los resultados.

1.2. Con la hoja de datos **Yield.mtw** haz lo siguiente:

- **a)** Calcula los resultados de la variable media geométrica de las columnas **Time**, **Temp**, **Yield** y **Cost** (raíz cuarta del producto de las cuatro variables). Denomina a la nueva variable **Media geométrica**.
- **b)** Codifica los datos de la variable **Media geométrica** de la forma indicada en la siguiente tabla:

intervalo	categoría
(40,50]	A
(50,60]	B
(60,70]	C

- **c)** Calcula una nueva columna en la que aparezcan los rangos de la variable **Media geométrica**.

1.3. Una determinada universidad ha plantado 6 variedades distintas de alfalfa en 4 campos experimentales diferentes a fin de estudiar si hay diferencias significativas en la producción. Los datos se encuentran en el fichero **Alfalfa.mtw**, donde C1 es la producción, C2 es la variedad y C3 es el campo experimental.

- **a)** Ordena los datos de la producción (**Yield**) en orden creciente. ¿Cuál es el dato que ocupa el décimo lugar?
- **b)** Desapila los datos de la producción (**Yield**) según los distintos campos experimentales (**Field**). Calcula la media aritmética (por filas) de las cuatro columnas resultantes (media de la producción en los cuatro campos experimentales). Denomina a la nueva variable **Media producción 4 campos** y determina su mediana.

2. Estadística descriptiva. Representaciones gráficas

2.1. Distribución de frecuencias

Para hacer la distribución de frecuencias de una o más variables, podemos utilizar la opción **Stat⇒Tables⇒Tally Individual Variables**.

Para practicar esta opción, podemos abrir el fichero de datos (Worksheet) **Pulse.mtw**. 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). Si queremos saber el número de casos (frecuencia absoluta) y el porcentaje de cada una de las categorías de la variable **Activity**, utilizamos la opción **Stat⇒Tables⇒Tally Individual Variables**; en el recuadro **Variables** seleccionamos, de la lista de variables de la izquierda, la columna **Activity** y en **Display** activamos **Counts** y **Percents**. Podemos ver, en la ventana de sesión (**Session**), que hay 21 alumnos con nivel alto de actividad física, y que un 66'3 % de ellos tiene un nivel medio de actividad física.

2.2. Estadística descriptiva con la opción *Stat⇒Basic Statistics⇒Display Descriptive Statistics*

Ya hemos visto que la opción **Calc⇒Column Statistics** calcula, para una columna (o variable), **uno** de los estadísticos siguientes: **Sum** (suma), **Mean** (media arimética), **Standard deviation** (cuasidesviación típica o desviación típica corregida), **Minimum** (mínimo resultado), **Maximum** (máximo resultado), **Range** (recorrido o amplitud total), **Median** (mediana), **Sum of squares** (suma de cuadrados), **N total** (número total de casos o tamaño muestral), **N nonmissing** (número de casos para los cuales sabemos el resultado de la variable) y **N mising** (número de casos para los cuales no sabemos el resultado de la variable). A continuación vamos a trabajar con una opción mucho más amplia, que nos permite, entre otras cosas, calcular más un estadístico y trabajar con más de una variable (columna) a la vez.

La opción **Stat⇒Basic Statistics⇒Display Descriptive Statistics** permite obtener los estadísticos descriptivos más usuales 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 nueva opción, podemos calcular los estadísticos descriptivos más importantes de las variables **Pulse1**, **Height** y **Weight** de la hoja de datos (Worksheet) **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**. En la ventana de sesión nos salen los resultados, para cada una de las tres variables, de los siguientes estadísticos descriptivos:

N	número de casos para los cuales sabemos el resultado de la variable	
N*	número de casos para los cuales no sabemos el resultado de la variable	
Mean	media aritmética	$\overline{x} = \left(\sum_{i=1}^{N} x_i \right) / N$
SE Mean	error estándar de la media	$S/\sqrt{N}$
StDev	cuasidesviación típica	$S = \sqrt{\left(\sum_{i=1}^{N} (x_i - \overline{x})^2 \right) / (N-1)}$
Minimum	mínimo dato	
Q1	primer cuartil=valor que deja por debajo de él el 25 % de los datos	
Median	mediana=segundo cuartil=valor que deja por debajo de él el 50 % de los datos	
Q3	tercer cuartil=valor que deja por debajo de él el 75 % de los datos	
Maximum	máximo dato	

Con la misma hoja de datos, 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.

El botón **Statistics** del cuadro de diálogo que aparece con la opción **Stat⇒Basic Statistics⇒Display Descriptive Statistics** conduce a una nueva ventana en la 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. Algunos de ellos ya han sido explicados anteriormente. Los estadísticos descriptivos que podemos seleccionar (cuando pulsamos el botón **Statistics**) son los siguientes:

Mean	media aritmética	$\overline{x} = \left(\sum_{i=1}^{n} x_i\right)/n$
SE of mean	error estándar de la media	$S/\sqrt{n}$
Standard deviation	cuasidesviación típica	$S = \sqrt{\left(\sum_{i=1}^{n} (x_i - \overline{x})^2\right)/(n-1)}$
Variance	cuasivarianza	S^2
Coefficient of variation	coeficiente de variación	$CV = S/\lvert\overline{x}\rvert$
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 valores más pequeños y el 5 % de los valores más grandes	
Sum	suma	$\sum_{i=1}^{n} x_i$
Minimum	mínimo dato	x_{min}
Maximum	máximo dato	x_{max}
Range	recorrido total	$R = x_{max} - x_{min}$
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 (esto tiene sentido cuando se ha rellenado el recuadro **By variables**)	
Percent	porcentaje de casos (esto tiene sentido cuando se ha rellenado el recuadro **By variables**)	
Cumulative percent	porcentaje acumulado de casos (esto tiene sentido 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	$g_1 = m_3/S^3$, siendo $m_3 = \left(\sum_{i=1}^{n} (x_i - \overline{x})^3\right)/(n-1)$
Kurtosis	coeficiente de apuntamiento	$g_2 = (m_4/S^4) - 3$, siendo $m_4 = \left(\sum_{i=1}^{n} (x_i - \overline{x})^4\right)/(n-1)$
MSSD	media de los cuadrados de las sucesivas diferencias	

Por ejemplo, podemos comprobar que el coeficiente de variación de la variable **Height** de la hoja de datos (Worksheet) **Pulse.mtw** es igual a $5'33$.

2.3. Representaciones gráficas con la opción *Stat⇒Basic Statistics⇒Display Descriptive Statistics*

El botón **Graphs** del cuadro de diálogo que aparece con la opción **Stat⇒Basic Statistics⇒Display Descriptive Statistics** permite elegir alguno de los siguientes gráficos (por defecto no se realiza ninguno) de las variables que hemos seleccionado en el recuadro **Variables**:

Histogram of data o histograma, que agrupa los datos en intervalos, representando sobre ellos rectángulos de área proporcional a la frecuencia absoluta de cada intervalo;

Histogram of data, with normal curve o histograma al que se le superpone la curva de la distribución normal de media igual a media muestral de la variable seleccionada y desviación típica igual a la cuasidesviación típica muestral de dicha variable;

Individual value plot o gráfico de valores individuales, que representa los datos en forma de puntos, y

Boxplot of data o diagrama caja-bigote, que representa los valores mínimo y máximo (extremos de los bigotes), los cuartiles $Q1$ y $Q3$ (extremos de la caja) y la mediana. Dentro de la caja tendremos el 50 % de los datos de la muestra y en cada bigote tendremos el 25 % de los datos más extremos. Este último tipo de gráfico nos permite visualizar tanto el valor central como la dispersión de los datos, y es muy útil a la hora de comparar datos de distintas muestras o grupos.

Por ejemplo, de la hoja de datos (Worksheet) **Pulse.mtw**, podemos realizar el histograma (con la curva normal superpuesta) de la variable **Height**, el gráfico de valores individuales de la variable **Activity** y el diagrama caja-bigote de la variable **Pulse1**.

2.4. Representaciones gráficas con la opción *Graph*

Además de los gráficos que se obtienen con la **Stat⇒Basic Statistics⇒Display Descriptive Statistics**, podemos crear representaciones gráficas con el menú **Graph**.

Una opción 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).

2.4.1. Histograma

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 simple de la variable **Weight** de la hoja de datos **Pulse.mtw**. Para ello, 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**. Para más información sobre las acciones de estos botones, pulsar el botón **Help** del mismo cuadro de diálogo. En principio, podríamos dejar todas las opciones por defecto a la hora de realizar este primer histograma.

El histograma resultante podemos copiarlo en el portapapeles, haciendo *clic* sobre el gráfico con el botón derecho del ratón y seleccionado, 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 como Paint Shop Pro. También podemos almacenarlo en la ventana de proyecto, **Proyect 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 fichero aparte.

Una vez obtenido el histograma es posible cambiar su aspecto. Para ello, hacemos *clic* sobre el gráfico, *clic* sobre la parte del gráfico que queremos cambiar y doble *clic* sobre esa parte. 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 simple de la variable **Weight** de la hoja de datos **Pulse.mtw** de la siguiente manera:

- Que el título sea *Histograma de la variable Peso*.
- 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*.
- Que en el eje vertical se muestren 13 marcas (*ticks*).
- Que el texto del eje vertical sea *Frecuencia absoluta*.

2.4.2. Diagrama de sectores o de *pastel*

Este gráfico resume los datos de una columna contando el número de datos iguales y representándolos mediante sectores proporcionales al número de datos de cada clase. Se utiliza con datos cualitativos o de tipo discreto con pocos resultados distintos. Se obtiene con la opción **Graph⇒Pie Chart**.

Por ejemplo, podríamos hacer el diagrama de *pastel* 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 row data** 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 Chart 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.

Igual que ocurría con el histograma, una vez obtenido el diagrama de *pastel* podemos copiarlo en el portapapeles, o almacenarlo en el directorio **ReportPad** de la ventana **Proyect Manager**, o grabarlo en un fichero aparte. También es posible cambiar su aspecto una vez obtenido, haciendo *clic* sobre el gráfico, *clic* sobre la parte del gráfico que queremos cambiar y doble *clic* sobre esa parte. Para practicar vamos a cambiar el gráfico de sectores anterior de la siguiente manera:

- Que el título sea *Gráfico de sectores de la variable Actividad Física*.
- Que junto a los sectores circulares aparezca la frecuencia absoluta y el porcentaje de cada categoría.

2.4.3. Diagrama de barras

Este tipo de gráfico se utiliza con datos cualitativos o de tipo discreto con pocos resultados distintos. El diagrama de barras se construye colocando en el eje horizontal los resultados (o categorías) de la variable y subiendo, sobre ellos, unas barras (rectángulos o líneas) de altura igual a la frecuencia absoluta (o la frecuencia relativa o el porcentaje) de cada resultado (o categoría). Se obtiene con la opción **Graph⇒Bar Chart**.

Por ejemplo, podríamos hacer el diagrama de barras 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⇒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. Como las categorías son números concretos (0, 1, 2 y 3) es más riguroso que, en vez de barras, aparezcan solamente líneas verticales; por tanto, activamos el botón **Data View** y en el cuadro de diálogo resultante activamos solo la opción **Proyect 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 **Proyect Manager**, o grabarlo en un fichero aparte. También es posible cambiar su aspecto una vez obtenido, haciendo *clic* sobre el gráfico, *clic* sobre la parte del gráfico que queremos cambiar y doble *clic* sobre esa parte. También podemos observar que si hacemos *clic* sobre el gráfico y luego pasamos el ratón por encima de las barras, se nos indica la frecuencia absoluta de cada categoría. Para practicar vamos a cambiar el diagrama de barras anterior de la siguiente manera:

- Que el título sea *Diagrama de barras de la variable Actividad Física*.
- 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*).
- Que el texto del eje vertical sea *Frecuencia absoluta*.
- Que el texto del eje horizontal sea *Actividad Física (0=Ninguna, 1=Baja, 2=Media, 3=Alta)*.

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 fichero de datos **Pulse.mtw** vamos a hacer el diagrama de barras de la variable **Activity** en grupos definidos por la variable **Sex**. Para ello, en el cuadro de diálogo que resulta al seleccionar **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** 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*, escrito con letra Arial y con un tamaño de 10 puntos.
- Que las barras tengan distinto color según los resultados de la variable **Sex** (*clic* sobre el gráfico, doble *clic* sobre cualquiera de las barras, en el cuadro de diálogo resultante seleccionar la carpeta **Groups**, en el recuadro **Categorical variables for attribute assignment** seleccionar la variable **Sex**.)
- Que en el eje vertical se muestren 10 marcas (*ticks*).
- Que el texto del eje vertical sea *Frecuencia absoluta*.

- Que en el eje horizontal todo esté escrito con la fuente Verdana, en negrita y con un tamaño de 8 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*.

2.4.4. Diagramas bivariantes

La opción **Graph⇒Scatterplot** realiza una gráfica con los datos (bivariantes) de dos columnas de la misma longitud.

Por ejemplo, de la hoja de datos **Pulse.mtw** podemos representar la altura en pulgadas, **Height**, frente al peso en libras, **Weight**. Para ello, seleccionamos la opción **Graph⇒Scatterplot**, en el cuadro de diálogo que aparece seleccionamos **Simple**, en el siguiente cuadro de diálogo, en el recuadro **Y Variables** seleccionamos (de la lista de variables de la izquierda) **Height**, en el recuadro **X Variables** seleccionamos **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 la *nube de puntos* resultante se agrupa cerca de una línea recta, lo que significa que hay una clara relación lineal 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 **Proyect Manager**, o grabar en un fichero aparte. También es posible cambiar su aspecto una vez obtenido, haciendo *clic* sobre el gráfico, *clic* sobre la parte del gráfico que queremos cambiar y doble *clic* sobre esa parte. Para practicar vamos a diagrama de dispersión anterior de la siguiente manera:

- Que el título sea *Diagrama de dispersión de la Altura frente al Peso*.
- Que los símbolos sean rombos verdes de tamaño 1.
- Que en el eje horizontal se muestren 14 marcas (*ticks*).
- Que el texto del eje horizontal sea *Peso de los alumnos, en libras*.
- Que en el eje vertical se muestren 10 marcas (*ticks*).
- Que el texto del eje vertical sea *Altura de los alumnos, en pulgadas*.

La opción **Graph⇒Scatterplot** es la que se utiliza para hacer la representación gráfica de una determinada función $f(x)$. Para ello es necesario tener en una columna los valores de x (generalmente creados por patrón) y en otra columna los resultados de $y = f(x)$ (generalmente calculados a partir de la opción **Calc⇒Calculator**). Por ejemplo, vamos a hacer la representación gráfica de la función $f(x) = (1 + x)(1 - x^2)$ en el intervalo $[-3, 3]$. Para ello se procede de la siguiente manera:

1) Se abre una hoja de datos (Worksheet) nueva.

2) Mediante la opción **Calc⇒Make Patterned Data⇒Simple Set of Numbers** se crea una nueva columna que denominaremos **x** y que contendrá todos los números comprendidos entre el -3 y el 3 con un incremento de $0,01$. En la columna **x** habrá un total de 601 números.

3) En otra columna se calculan los resultados de la función función $f(x) = (1 + x)(1 - x^2)$ para cada valor de la columna **x**. Para hacerlo, se selecciona **Calc⇒Calculator**; en **Store result in variable** tecleamos **f(x)**; en **Expression** tenemos que colocar, utilizando la calculadora y la lista de variables que aparecen en este cuadro de diálogo, la siguiente expresión: `(1+'x')*(1-'x'**2)`

4) Para representar gráficamente la función 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)'** y en **X variables** se selecciona la columna **'x'**. 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.

También se puede lograr lo mismo de la siguiente manera: se elige la opción **Graph⇒Scatterplot**; se selecciona **Simple**; en el siguiente cuadro de diálogo, en **Y variables** se selecciona la columna **'f(x)'** y en **X variables** se selecciona la columna **'x'**; se activa el botón **Data View** y en el cuadro de diálogo resultante se deja activada solamente la opción **Connect line**.

2.5. Ejercicios propuestos

2.1. Con la hoja de datos **Wine.mtw** haz lo siguiente:

a) Determina la frecuencia absoluta de cada una de las categorías de las variables **Clarity** y **Region**.

b) Calcula la mediana y el coeficiente de variación de la variable **Flavor** separando sus resultados según los valores de la variable **Region**.

c) Haz el histograma simple de la variable **Flavor**. Una vez obtenido, modifícalo de la manera siguiente:

- Que el título sea *Histograma de la variable Sabor*.
- Que las barras sean de color verde claro con una trama de relleno con puntitos. Que los bordes de las barras sean de color verde oscuro y de tamaño 2.
- Que haya 6 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 *Sabor del vino*.
- Que en el eje vertical se muestren todos los números naturales comprendidos entre el 0 y el 14.
- Que el texto del eje vertical sea *Frecuencia absoluta*.

2.2. Con la hoja de datos **Bears.mtw** haz lo siguiente:

a) Determina la frecuencia absoluta y el porcentaje de cada una de las categorías de las variables **Obs.No** (número de veces que ha sido medido cada oso) y **Sex** (1=macho, 2=hembra).

b) Calcula la media y la cuasivarianza de la variable **Length** (longitud o altura del oso) separando sus resultados según los valores de la variable **Sex** (sexo).

c) Haz el diagrama de dispersión o nube de puntos de la variable **Length** (en el eje vertical) frente a la variable **Weight** (en el eje horizontal). Según este gráfico, ¿cuál es la fuerza de la relación lineal entre las dos variables: muy débil, débil, regular, bastante fuerte o muy fuerte? ¿Por qué?

Una vez obtenido el gráfico, modifícalo de la manera siguiente:

- Que el título sea *Diagrama de dispersión de la Longitud frente al Peso*.
- Que los símbolos sean triángulos azules de tamaño 1.
- Que el texto del eje horizontal sea *Peso de los osos, en libras*.
- Que en el eje horizontal se muestren 15 marcas (*ticks*).
- Que el texto del eje vertical sea *Longitud de los osos, en pulgadas*.
- Que en el eje vertical se muestren 12 marcas (*ticks*).

2.3. Con la hoja de datos **Wine.mtw** haz lo siguiente:

a) Determina la frecuencia absoluta de cada una de las categorías de las variables **Clarity** y **Region**.

b) Calcula la media y el coeficiente de variación de la variable **Aroma** separando sus resultados según los valores de la variable **Region**.

c) Haz el diagrama de barras de la variable **Clarity** agrupado según los resultados de la variable **Region**. Una vez obtenido, modifícalo de la manera siguiente:

- Que las barras tengan distinto color según los resultados de la variable **Region**.
- Que el título sea *Diagrama de barras de la variable Claridad en grupos definidos por la variable Región*, escrito con letra Arial y con un tamaño de 9 puntos.
- Que en el eje vertical se muestren todos los números naturales comprendidos entre el 0 y 12.
- Que el texto del eje vertical sea *Frecuencia absoluta*.
- Que en el eje horizontal todo esté escrito con la fuente Verdana, en negrita y con un tamaño de 8 puntos. Que en dicho eje aparezcan los nombres de las variables en español: *Claridad* en vez de *Clarity*, y *Región* en vez de *Region*.
- Que la fuente de la leyenda sea Verdana.

3. Probabilidad. Variables aleatorias

3.1. Muestras aleatorias de las distribuciones usuales

Como ya se ha visto anteriormente, 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, vamos a crear una nueva hoja de datos que llevará por nombre **Probabilidad.mtw**. A continuación, vamos a crear una columna, en dicha hoja de datos, que lleve por nombre **100 datos de N(5,2)** y que contenga 100 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 **Generate...** tecleamos **100**; en **Store in column** tecleamos el nombre **'100 datos de N(5,2)'**; en **Mean** tecleamos **5** y en **Standard deviation** ponemos un **2**. A continuación vamos a hacer el histograma de la muestra aleatoria obtenida en la columna **'100 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 **'100 datos de N(5,2)'** y pulsamos **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. Genera ahora una muestra de la misma distribución, $\mathcal{N}(5, 2)$, pero de tamaño 10000 y haz el histograma correspondiente a los datos de la nueva muestra. ¿Qué aprecias respecto al ajuste de la gráfica a la curva Normal? ¿Piensas que tiene que ver con el tamaño de la muestra?

La opción **Calc⇒Random Data** también nos puede servir para calcular el valor aproximado de cualquier medida o momento de cualquier distribución. Por ejemplo, para calcular la mediana de una distribución Exponencial de media 15 podemos crear 5000 datos aleatorios de dicha distribución y después determinar la mediana de la columna creada. Para ello, seleccionamos **Calc⇒Random Data⇒Exponential**; en **Generate...** tecleamos **5000**; en **Store in column** tecleamos el nombre **'5000 datos de E(15)'**; en **Scale** tecleamos el valor de la media, que es **15**, y en **Threshold** dejamos lo que aparece por defecto, que es cero. Recordemos que para determinar la mediana de una columna tenemos varias posibilidades, una es la opción **Calc⇒Column Statistics** y otra es la opción **Stat⇒Basic Statistics⇒Display Descriptive Statistics**. Para determinar la mediana de la muestra de la distribución Exponencial de media 15, nosotros vamos a utilizar la opción **Calc⇒Column Statistics**; en **Statistic** activamos **Median**; en **Input variable** seleccionamos (de la lista de variables de la izquierda) la columna **'5000 datos de E(15)'** y no escribimos nada en el recuadro **Store result in**. En la ventana de sesión nos aparece el resultado de la mediana deseada, que podemos comprobar que se aproxima al valor real de la mediana de una distribución Exponencial de media 15, que es $M_e = -15\, ln\, 0'5 = 10'3972077\cdots$. Cuanto más grande sea el tamaño muestral, tanto más se aproximará el valor de la mediana de la muestra al valor teórico de dicha mediana.

La distribución **Discrete** que aparece en el menú de la opción **Calc⇒Random Data** no es un modelo concreto, sino que sirve para cualquier modelo discreto previamente introducido en dos columnas; una para los valores que toma x y otra para los resultados de sus probabilidades $p(x)$. Por ejemplo, podemos generar una muestra aleatoria de tamaño 1000 de la distribución discreta que tiene por función de probabilidad $p(x) = x/55$ para $x = 1, 2, \ldots, 10$ y podemos comprobar gráficamente que aproximadamente se cumplen las probabilidades teóricas. Para ello, se procede de la siguiente manera:

a) Mediante la opción **Calc⇒Make Patterned Data⇒Simple Set of Numbers** se crea una nueva columna, que podemos denominar **x**, con los valores $1, 2, 3, \ldots, 10$. Esta columna contiene los posibles resultados de la variable aleatoria discreta.

b) Mediante la opción **Calc⇒Calculator**, se calculan los resultados de la función de probabilidad para todos y cada uno de los valores de la columna **x**. A la nueva columna la podemos denominar **p(x)**. Recordemos que es mejor emplear la lista de variables y la calculadora de dicho cuadro de diálogo que teclear las operaciones y los nombres de las variables.

c) Se selecciona **Calc⇒Random Data⇒Discrete**; en **Generate...** tecleamos **1000**; en **Store in column** tecleamos **'muestra modelo discreto'**; en **Values in** seleccionamos, de la lista de variables de la izquierda, la columna **x**; en **Probabilities in** seleccionamos, de la lista de variables de la izquierda, la columna **p(x)**. La nueva columna **'muestra modelo discreto'** contiene la muestra deseada. Con una probabilidad muy alta, el dato que más habrá aparecido será el 10 puesto que es el valor más probable, con probabilidad $p(10) = 0'\widehat{18}$ y el dato que menos habrá aparecido será el 1 puesto que es el valor menos probable, con una probabilidad $p(1) = 0'0\,\widehat{18}$.

d) Hacemos un diagrama de barras de la columna **'muestra modelo discreto'** para comprobar que aproximadamente se cumplen las probabilidades teóricas. Para ello, se selecciona **Graph⇒Bar Chart**, dejamos activada la opción **Simple** y hacemos *clic* en **OK**. En el siguiente cuadro de diálogo, en **Categorical variables** seleccionamos, de la lista de variables de la izquierda, la columna **'muestra modelo discreto'**. Si pasamos el cursor sobre cada una de las barras

del gráfico resultante podemos ver la frecuencia absoluta de cada uno de los 10 valores de x. Como tenemos una muestra de tamaño 1000, para averiguar la frecuencia relativa (que es lo que se aproxima a la probabilidad), tenemos que dividir la frecuencia absoluta entre 1000. Comprobemos que la frecuencia absoluta del resultado **10** se aproxima a $0'\widehat{18}\cdot 1000 = 181'\widehat{81}$.

Como ya sabemos, la distribución **Uniforme** genera números aleatorios de tipo continuo comprendidos entre dos números cualesquiera. La distribución **Integer** es su equivalente en el caso discreto; es decir, genera números aleatorios de tipo discreto (números enteros) comprendidos entre dos números enteros cualesquiera. Por ejemplo, vamos a utilizar esta distribución para simular los resultados de 1000 lanzamientos de un dado. Para ello, seleccionamos **Calc⇒Random Data⇒Integer**; en **Generate...** tecleamos **1000**; en **Store in column** tecleamos el nombre **'1000 lanzamientos dado'**; en **Minimum value** tecleamos **1** y en **Maximum value** ponemos un **6**. Ahora podemos comprobar gráficamente que aproximadamente se cumplen las probabilidades teóricas. Para ello, vamos a hacer un diagrama de barras de los datos obtenidos: Se selecciona **Graph⇒Bar Chart⇒Simple** y en **Categorical variables** se elige la columna **'1000 lanzamientos dado'**. Si pasamos el cursor sobre cada una de las barras del gráfico resultante podemos ver la frecuencia absoluta de cada uno de los 6 resultados posibles. Como tenemos una muestra de tamaño 1000, para averiguar la frecuencia relativa (que es lo que se aproxima a la probabilidad), tenemos que dividir la frecuencia absoluta entre 1000. Comprobemos que la frecuencia absoluta de cada resultado se aproxima a $\frac{1}{6}\cdot 1000 = 166'\widehat{6}$.

Si un determinado suceso A tiene por probabilidad p; es decir, $P(A) = p$, podemos aproximarnos al verdadero valor de la probabilidad p generando una columna con una muestra aleatoria de gran tamaño de la distribución de Bernoulli de parámetro p y luego calculando la media de dicha columna (pues la media teórica de la distribución de Bernoulli de parámetro p es igual a p). Vamos a utilizar lo anterior para averiguar, aproximadamente, el valor de la probabilidad de que el valor mínimo de 5 observaciones de una distribución $\mathcal{N}(12,4)$ sea menor que 10. Este suceso lo vamos a denotar por A; es decir, *A=el valor mínimo de 5 observaciones de una distribución* $\mathcal{N}(12,4)$ *es menor que 10*, y a su probabilidad la vamos a denotar por p; es decir, $P(A) = p$. Para averiguar el valor aproximado de la probabilidad p hacemos lo siguiente:

a) Generamos 5 muestras de tamaño grande (por ejemplo, 10000) procedentes de una distribución $\mathcal{N}(12,4)$, cada una de ellas en una columna de ***Minitab***. A estas columnas las podemos denominar $X1$, $X2$, $X3$, $X4$ y $X5$. Cada fila se puede considerar como una muestra de tamaño 5 procedente de una distribución $\mathcal{N}(12,4)$. Por tanto, hemos obtenido 10000 muestras de tamaño 5 de dicha distribución Normal.

b) Utilizamos la opción **Calc⇒Row Statistics** para calcular el mínimo de cada muestra de tamaño 5; es decir, determinamos la función mínimo (por filas) de las columnas $X1$, $X2$, $X3$, $X4$ y $X5$. Denominamos a la nueva columna **'Mínimo X1 a X5'**.

c) Utilizamos la opción **Calc ⇒Calculator** para determinar el resultado de la expresión lógica **'Mínimo X1 a X5'<10**. A la nueva columna la denominamos **Mínimo<10**. Esta nueva columna es una muestra aleatoria de una distribución de Bernoulli de parámetro igual a la probabilidad del suceso A, pues la operación anterior ha asignado el valor 1 si el suceso A ocurre, y ha asignado el valor cero si A no ocurre.

d) Con la opción **Calc⇒Column Statistics** calculamos la media (*Mean*) de la columna **Mínimo<10**. Dicha media es una estimación o aproximación al parámetro p de la distribución de Bernoulli de parámetro igual a la probabilidad del suceso A y por tanto es una estimación o aproximación de la probabilidad del suceso A.

Como ya sabemos, el Teorema Central del Límite nos dice que si tenemos n variables aleatorias independientes idénticamente distribuidas, X_1,..., X_n, con media μ y varianza σ^2, entonces:

I) Cuando n es suficientemente grande, la variable Suma Muestral ($S_n = X_1 + \cdots + X_n$) tiene, aproximadamente, una distribución Normal de media $n\mu$ y varianza $n\sigma^2$.

II) Cuando n es suficientemente grande, la variable Media Muestral ($\overline{X} = S_n/n$) tiene, aproximadamente, una distribución Normal de media μ y varianza σ^2/n.

Para practicar con lo anterior, se propone el siguiente ejercicio: Genera 5 muestras de tamaño grande (por ejemplo, 10000) de cualquier distribución, por ejemplo, de una variable Uniforme en el intervalo $(0,1)$. Mediante la opción **Calc⇒Row Statistics** (utilizada dos veces) crea dos columnas nuevas: la primera mediante la suma (por filas) de las cinco columnas y la segunda mediante la media aritmética (por filas) de las cinco columnas. Haz el histograma (con la curva Normal superpuesta) de estas dos nuevas variables, Suma y Media. ¿Qué se puede observar en estas gráficas? Determina las medias y las varianzas de las dos columnas nuevas, Suma y Media, y compara los resultados con los teóricos. Recordemos que la media de una distribución Uniforme en el intervalo (a,b) es igual a $(a+b)/2$ y la varianza es $(b-a)^2/12$.

3.2. Función de densidad y función de probabilidad

Minitab puede calcular el resultado de la función de densidad (o de la función de probabilidad) 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), Uniform (Uniforme), Binomial, Hypergeometric (Hipergeométrica), Discrete, Integer, Poisson, Beta, Cauchy, Exponential, Gamma, Laplace, 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 $0,01$. En la columna **'x de -4 a 4'** habrá un total de 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.

Para completar el ejemplo anterior, podríamos superponer en un mismo gráfico las curvas de densidad de las distribuciones $\mathcal{N}(0,1)$ (Normal Estándar), t_2 (t de Student con 2 grados de libertad), t_5 (t de Student con 5 grados de libertad) y t_{30} (t de Student con 30 grados de libertad) con el fin de comprobar que la distribución t_n se va aproximando a la distribución $\mathcal{N}(0,1)$ cuando va aumentando el valor del parámetro n. Para ello, se procede de la siguiente manera:

a) Se selecciona **Calc⇒Probability Distributions⇒t**; se activa **Probability density**; en **Degrees of freedom** se teclea **2**; 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 de t_2; por ejemplo, **'f(x) t2'**.

b) Se selecciona **Calc⇒Probability Distributions⇒t**; se activa **Probability density**; en **Degrees of freedom** se teclea **5**; 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 de t_5; por ejemplo, **'f(x) t5'**.

c) Se selecciona **Calc⇒Probability Distributions⇒t**; se activa **Probability density**; en **Degrees of freedom** se teclea **30**; 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 de t_{30}; por ejemplo, **'f(x) t30'**.

d) Se selecciona la opción **Graph⇒Scatterplot⇒With connect line**. En el cuadro de diálogo que aparece, junto al **1** en **Y variables** seleccionamos la columna **'f(x) N(0,1)'** y en **X variables** seleccionamos la columna **'x de -4 a 4'**; junto al **2** en **Y variables** seleccionamos la columna **'f(x) t2'** y en **X variables** seleccionamos otra vez la columna **'x de -4 a 4'**; junto al **3** en **Y variables** seleccionamos **'f(x) t5'** y en **X variables** seleccionamos otra vez **'x de -4 a 4'**; y junto al **4** en **Y variables** seleccionamos la columna **'f(x) t30'** y en **X variables** seleccionamos otra vez la columna **'x de -4 a 4'**. Luego pulsamos **Multiple graphs** y en el cuadro de diálogo resultante activamos la opción **Overlay on the same graph**. Como ya hemos dicho anteriormente, sería conveniente quitar los puntos del gráfico, dejando sólo la línea de conexión.

Ahora vamos a calcular los resultados de la función de probabilidad de la distribución discreta $\mathcal{B}(200, 0'4)$ (Binomial de parámetros $n = 200$ y $p = 0'4$), vamos a hacer su representación grafica y vamos a averiguar el valor de la media de dicha variable aleatoria discreta. Para ello procedemos de la siguiente manera:

a) Mediante la opción **Calc⇒Make Patterned Data⇒Simple Set of Numbers** crearemos una nueva columna que denominaremos **'x de 0 a 200'** y que contendrá todos los resultados posibles de la distribución $\mathcal{B}(200, 0'4)$, que, como sabemos, son: $0, 1, 2, \cdots, 200$.

b) Calculamos los resultados de la función de probabilidad de $\mathcal{B}(200, 0'4)$ para todos y cada uno de los valores de la columna **'x de 0 a 200'**. Para ello, seleccionamos la opción **Calc⇒Probability Distributions⇒Binomial**; activamos **Probability**; en **Numbers of trials** tecleamos **200**; en **Probability of success** tecleamos **0,4**; en **Input column** elegimos, de la lista de variables de la izquierda, la columna **'x de 0 a 200'** y en **Optional storage** tecleamos el nombre de la columna que contendrá los resultados de la función de probabilidad; por ejemplo, **'p(x) B(200,0,4)'**.

c) Ahora vamos a hacer la representación gráfica bidimensional que tiene en el eje horizontal los resultados de la columna **'x de 0 a 200'** y en el eje vertical los resultados de la columna **'p(x) B(200,0,4)'**. Para ello, se selecciona 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 **'p(x) B(200,0,4)'** y en **X variables** se selecciona la columna **'x de 0 a 200'**. Como ya hemos dicho anteriormente, sería conveniente quitar los puntos del gráfico, dejando sólo la línea de conexión. Se puede comprobar que esta representación gráfica se aproxima mucho a la curva de densidad de una distribución Normal, lo cual se debe a lo siguiente: cuando n es grande y p no se acerca a 0 ni a 1, entonces $\mathcal{B}(n, p)$ se aproxima a $\mathcal{N}(np, \sqrt{npq})$, siendo $q = 1 - p$.

d) También vamos a calcular la media teórica de la distribución $\mathcal{B}(200, 0'4)$. Recordemos que la media de una distribución discreta es $E(X) = \sum x_i \cdot p(x_i)$. Por tanto, usamos la opción **Calc ⇒Calculator**. En **Store result in variable** tecleamos el nombre de la columna que contendrá los resultados de los productos $x_i \cdot p(x_i)$; por ejemplo, **'x p(x)'**; en **Expression** ponemos (empleando la lista de variables y la calculadora de dicho cuadro de diálogo) `'x de 0 a 200'*'p(x) B(200,0,4)'`. Ahora tenemos que calcular la suma de todos los resultados de la columna **'x p(x)'**, para lo cual elegimos la opción **Calc⇒Column Statistic**; activamos **Sum**; en **Input variable** seleccionamos, de la lista de variables de la izquierda, la columna **'x p(x)'** y dejamos desactivada la opción **Store result in**. En la ventana de sesión podemos ver el resultado de la media, que es igual a $E(X) = n \cdot p = 200 \cdot 0'4 = 80$. De forma similar podríamos determinar cualquier otro momento de dicha distribución discreta.

3.3. Función de distribución (probabilidad acumulada)

Para calcular el resultado de la función de distribución de una variable aleatoria, $F(t) = P(X \leq t)$, 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**.

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. Si no rellenamos el recuadro **Optional storage**, el resultado aparece 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 $P(X > a)$, $P(a < X < b)$, $P(|X| < |a|)$, $P(|X| > |a|)$, 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: $P(X \leq x) = P(X < x)$, $P(X \geq x) = P(X > x)$, $\cdots$. Pero si X es una variable aleatoria discreta, las probabilidades $P(X \leq x)$ y $P(X < x)$ no son (en general) iguales.

Vamos a hacer algunos ejemplos:

- Si $X \equiv \mathcal{B}(85, 0'55)$, entonces $P(50 \leq X < 60) = P[(X < 60) - (X < 50)] = P(X < 60) - P(X < 50) = P(X \leq 59) - P(X \leq 49) = F(59) - F(49) = 0'997638 - 0'724689 = 0'272949$.
- Si $X \equiv \mathcal{N}(0, 1)$, entonces $P(|X| \geq 1'75) = P[(X \leq -1'75) \cup (X \geq 1'75)] = P(X \leq -1'75) + P(X \geq 1'75) = 2 \cdot P(X \leq -1'75) = 2 \cdot F(-1'75) = 2 \cdot 0'0400592 = 0'080118$.
- Si $X \equiv \mathcal{N}(6'5, 1'85)$, entonces $P(5 \leq X < 7) = P[(X < 7) - (X < 5)] = P(X < 7) - P(X < 5) = P(X \leq 7) - P(X \leq 5) = F(7) - F(5) = 0,606524 - 0,208737 = 0'397787$.

Como ya hemos dicho, cuando n es grande y p no se acerca a 0 ni a 1, entonces $\mathcal{B}(n, p)$ se aproxima a $\mathcal{N}(np, \sqrt{npq})$, siendo $q = 1 - p$. Vamos a poder observarlo con el siguiente ejemplo:

Sea X una variable aleatoria $\mathcal{B}(200, 0'4)$ y sea Y una variable aleatoria Normal de media 80 y desviación típica 6'928203. Vamos a comprobar (mediante una representación gráfica conjunta) que las funciones de distribución de ambas variables son muy parecidas. La solución es la siguiente:

a) Calculamos los resultados de la función de distribución de $\mathcal{B}(200, 0'4)$ para todos y cada uno de los valores de dicha columna **'x de 0 a 200'**. Para ello, seleccionamos la opción **Calc⇒Probability Distributions⇒Binomial**; activamos **Cumulative probability**; en **Numbers of trials** tecleamos **200**; en **Probability of success** tecleamos **0,4**; en **Input column** elegimos, de la lista de variables de la izquierda, la columna **'x de 0 a 200'** y en **Optional storage** tecleamos el nombre de la columna que contendrá los resultados de la función de distribución de la Binomial; por ejemplo, **'F(x) B(200,0,4)'**.

b) Calculamos los resultados de la función de distribución de $\mathcal{N}(80, 6'928203)$ para los mismos valores de x, es decir, para los valores de la columna **'x de 0 a 200'**. Para ello, se elige **Calc⇒Probability Distributions⇒Normal**; se activa **Cumulative probability**; en **Mean** se teclea **80**; en **Standard deviation** se pone **6,928203**; en **Input column** elegimos, de la lista de variables de la izquierda, la columna **'x de 0 a 200'** y en **Optional storage** tecleamos el nombre de la columna que contendrá los resultados de la función de distribución de la Normal; por ejemplo, **'F(x) N(80,6,9)'**.

c) Ahora vamos a superponer, en un mismo gráfico, las dos funciones de distribución. Para ello, se selecciona la opción **Graph⇒Scatterplot⇒With connect line**. En el cuadro de diálogo que aparece, junto al **1** en **Y variables** seleccionamos la columna **'F(x) B(200,0,4)'** y en **X variables** seleccionamos la columna **'x de 0 a 200'**, y junto al **2** en **Y variables** seleccionamos la columna **'F(x) N(80,6,9)'** y en **X variables** seleccionamos otra vez la columna **'x de 0 a 200'**. Luego pulsamos **Multiple graphs** y en el cuadro de diálogo resultante activamos **Overlay on the same graph**. Como ya hemos dicho anteriormente, sería conveniente quitar los puntos del gráfico, dejando sólo la línea de conexión.

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 x que hace que la probabilidad del suceso $(X \leq x)$ sea igual a un valor determinado p; es decir, hallar x para que se cumpla $P(X \leq x) = p$; 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⇒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 x que verifica $P(X \leq x) = 0'98$, cuando $X \equiv \chi^2_{20}$ (Chi-cuadrado de Pearson con 20 grados de libertad). Para ello seleccionamos la opción **Calc⇒Probability Distributions⇒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. Si no rellenamos el recuadro **Optional storage**, el resultado aparece en la ventana de sesión. Se puede comprobar que el valor x que verifica $P(X \leq x) = 0'98$ es $35'0196$; es decir, $P(X \leq 35'0196) = 0'98$, siendo $X \equiv \chi^2_{20}$.

Si queremos calcular los valores a y b tales que las probabilidades de los tipos $P(X > a)$, $P(a < X < b)$, $P(|X| < |a|)$, $P(|X| > |a|)$ sean iguales a un cierto resultado, tenemos que utilizar lápiz y papel, y aplicar las propiedades de la probabilidad para llegar a expresiones en las que sólo aparezcan ecuaciones del tipo $P(X \leq x) = p$ (percentiles), pues éstas son las que calcula ***Minitab***.

Vamos a hacer algunos ejemplos:

- Sea X una variable aleatoria que sigue una distribución t de Student con 30 grados de libertad ($X \equiv t_{30}$). Halla el valor de a que cumple $P(|X| > a) = 0'2$.

  Solución:

$$P(|X| > a) = 0'2 \Rightarrow P[(X < -a) \cup (X > a)] = 0'2 \Rightarrow P(X < -a) + (X > a) = 0'2$$
$$\Rightarrow 2P(X > a) = 0'2 \text{ (por ser simétrica)} \Rightarrow P(X > a) = 0'1$$
$$\Rightarrow P(X \leq a) = 0'9 \Rightarrow F(a) = 0'9 \Rightarrow a = 1'310415$$

- Sea X una variable aleatoria que sigue una distribución F de Snedecor con 10 grados de libertad en el numerador y 20 grados de libertad en el denominador ($X \equiv F_{10,20}$). Halla el valor de a que verifica la siguiente igualdad: $P(|X| \leq a) = 0'9$.

  Solución:

$$P(|X| \leq a) = 0'9 \Rightarrow P[-a \leq X \leq a] = 0'9$$
$$\Rightarrow P[(X \leq a) - (X < -a)] = 0'9 \Rightarrow P(X \leq a) - P(X < -a) = 0'9$$
$$\Rightarrow P(X \leq a) = 0'9 \text{ ya que } P(X < -a) = 0$$
$$\Rightarrow F(a) = 0'9 \Rightarrow a = 1'936738$$

Para distribuciones discretas, en general, fijado un p, no necesariamente existe un valor x que verifique $F(x) = p$, por lo que el programa dará los dos valores de x para los cuales $F(x)$ está más cerca de p. Por ejemplo, para la distribución Binomial $\mathcal{B}(3, 0'5)$ con $p = 0'7$ se obtienen los valores $x = 1$ y $x = 2$. Si almacenamos el resultado en una constante, ***Minitab*** opta por el mayor (en este caso, $x = 2$).

3.5. Ejercicios propuestos

3.1. Utilizando procedimientos similares a los explicados en la sección 3.1 haz los siguientes ejercicios:

a) Determina, de manera aproximada, la probabilidad de superar 310 kilos en un viaje en ascensor el que suben 4 personas cuyos pesos proceden de una distribución Normal de media 75 kilos y desviación típica 7 kilos.

b) Determina, de manera aproximada, la probabilidad de que un sistema, que consta de 3 componentes conectados en serie, siga funcionando después de 800 horas si cada componente tiene tiempo de funcionamiento exponencial de media 1000 horas e independiente de las demás.

c) Aproxima las probabilidades de la suma de dos dados. Representa gráficamente los resultados mediante un diagrama de barras. ¿Cuál es el valor más probable de la suma de dos dados?

d) Calcula el valor aproximado de la probabilidad de que al lanzar 100 monedas al aire se obtengan entre 45 y 55 caras. Basta con que generes una muestra (de tamaño grande, por ejemplo, 10000) de la correspondiente distribución Binomial y después crees una muestra de Bernoulli a partir de la expresión lógica $45 \leq X \leq 55$, donde X es la columna que contiene la muestra de la distribución Binomial.

e) Si seleccionamos al azar dos números comprendidos entre 0 y 1, calcula el valor aproximado de las probabilidades siguientes:

I) La suma de ambos sea menor que 1 (la probabilidad exacta es $0, 5$).
II) El producto de ambos sea menor que $0, 25$ (la probabilidad exacta es $0, 25(1 + \ln 4) \simeq 0, 5965$).

3.2. Utilizando procedimientos similares a los explicados en la sección 3.2 haz los siguientes ejercicios:

a) Representa, en una misma gráfica, distintas funciones de densidad de distribuciones chi-cuadrado de Pearson con n grados de libertad; por ejemplo, para $n = 5$, $n = 10$, $n = 30$ y $n = 50$. Los valores del eje horizontal pueden ser: $1, 2, \cdots, 120$. Comprueba que cuanto más aumenta n, más se aproxima dicha curva de densidad a la del modelo Normal.

b) Sea X una variable aleatoria Binomial de parámetros $n = 100$ y $p = 0'01$ y sea Y una variable aleatoria de Poisson de media $\lambda = 1$. Comprueba (mediante una representación gráfica conjunta) que las funciones de probabilidad de ambas variables son casi iguales.

3.3. Utilizando procedimientos similares a los explicados en la sección 3.3 haz los siguientes ejercicios:

a) Sea X una variable aleatoria que sigue una distribución de Poisson de parámetro 8, $X \equiv \mathcal{P}(\lambda = 8)$. Calcula:

I) $P(X = 8)$.
II) $P(X < 6)$.
III) $P(X > 7)$.
IV) $P(X \leq 5)$.
V) $P(X \geq 9)$.
VI) $P(5 < X < 15)$.
VII) $P(5 \leq X \leq 15)$.

b) Sea X una variable aleatoria que sigue una distribución Chi-cuadrado con n grados de libertad, $X \equiv \chi^2_n$. Calcula:

I) Para $n = 12$, $P(X < 4'8)$.
II) Para $n = 20$, $P(X > 4'8)$.
III) Para $n = 4$, $P(3'3 < X < 9'4)$.
IV) Para $n = 25$, $P(|X| > 1'5)$.
V) Para $n = 14$, $P(|X| < 4'5)$.

c) Sea X una variable aleatoria Chi-cuadrado de Pearson con 200 grados de libertad y sea Y una variable aleatoria Normal de media 200 y desviación típica 20. Comprueba (mediante una representación gráfica conjunta) que las funciones de distribución de ambas variables son muy parecidas.

d) Sea X una variable aleatoria t de Student con 120 grados de libertad y sea Y una variable aleatoria Normal de media 0 y desviación típica $1'008439$. Comprueba (mediante una representación gráfica conjunta) que las funciones de distribución de ambas variables son muy similares.

3.4. Utilizando procedimientos similares a los explicados en la sección 3.4 determina el valor de k que verifica las siguientes igualdades:

a) $P(X < k) = 0'9$.
b) $P(X > k) = 0'05$.
c) $P(|X| < k) = 0'98$.
d) $P(|X| \geq k) = 0'1$.

para cada uno de tres casos siguientes:

I) Si X es una variable aleatoria que sigue una distribución Normal Estándar.
II) Si X es una variable aleatoria que sigue una distribución Chi-cuadrado de Pearson con 50 grados de libertad.
III) Si X es una variable aleatoria que sigue una distribución Exponencial de media igual a 2.

4. Introducción a la inferencia estadística. Muestreo

4.1. Generación de muestras aleatorias

Podemos generar datos de distribuciones usuales utilizando la opción **Calc⇒Random Data**, como ya se ha visto en anteriormente. Esta opción permite generar una muestra aleatoria de cualquier columna de la hoja de datos actualmente abierta o de una de las distribuciones de probabilidad que aparecen listadas. Por ejemplo, vamos a crear una nueva hoja de datos que llevará por nombre **Muestras.mtw** y, a continuación, vamos a crear una columna, en dicha hoja de datos, que lleve por nombre **100 datos de chi50** y que contenga 100 datos aleatorios de una distribución chi-cuadrado de Pearson con 50 grados de libertad (χ^2_{50}).

Para generar una muestra aleatoria de una columna de la hoja de datos actualmente abierta utilizamos la opción **Calc⇒Random Data⇒Sample from Columns**. En esta opción se supone que todos los datos de la columna tienen la misma probabilidad de ocurrir. Podemos elegir entre el muestreo con reemplazamiento o el muestreo sin reemplazamiento. Por ejemplo, vamos a generar una muestra aleatoria de tamaño 30, sin reemplazamiento, de los datos de la columna **100 datos de chi50**. Para ello, seleccionamos la opción **Calc⇒Random Data⇒Sample from Columns**. En **Sample.....rows** tecleamos 30; en el recuadro siguiente (**from columns**) seleccionamos, de la lista de variables que tenemos a la izquierda, la columna **100 datos de chi50**; en **Store samples in** tecleamos el nombre de la columna que contendrá la muestra solicitada, por ejemplo, **submuestra de chi50** y, por último, dejamos desactivada la opción **Sample with replacement**. Hay que tener en cuenta que si el muestreo es sin reemplazamiento, el tamaño muestral no puede superar al número de datos de la columna de la cual procede la muestra.

Para generar muestras aleatorias de modelos discretos no incluidos en la lista de distribuciones utilizamos la opción **Calc⇒Random Data⇒Discrete**, como ya hemos visto anteriormente. Recordemos que previamente a la utilización de esta opción tenemos que introducir en una columna los valores que toma la variable, x_i, y en otra columna los resultados de sus probabilidades, $p(x_i) = P(X = x_i)$.

Para generar muestras aleatorias de modelos continuos no incluidos en la lista de distribuciones tenemos dos alternativas, que se explican en los dos sub-apartados siguientes.

4.1.1. Método de la transformada inversa

Para utilizar este método debemos conocer la expresión explícita de la función de distribución, $F(t)$, de la variable aleatoria continua. El procedimiento es el siguiente:

I) En una columna, que podemos denominar **u**, se genera una muestra aleatoria, del tamaño deseado (n), procedente de una distribución uniforme en el intervalo $(0, 1)$; es decir, se generan n números aleatorios comprendidos entre 0 y 1: $u_1, \ldots, u_n$. Estos serán resultados aleatorios de la función de distribución de la variable aleatoria continua.

II) Se determina la expresión explícita de la inversa de la función distribución, $F^{-1}(u)$.

III) Mediante la opción **Calc⇒Calculator**, se calculan los resultados de la inversa de la función de distribución para todos y cada uno de los valores de la columna **u**; es decir, se calculan $F^{-1}(u_1), \ldots, F^{-1}(u_n)$. A la nueva columna la podemos denominar **F-1(u)** y es la que contiene la muestra del modelo continuo deseado.

Como ejemplo, vamos a generar una muestra aleatoria de tamaño 100 de la variable aleatoria continua cuya función de distribución es $F(x) = x^3$ para $0 < x < 1$, $F(x) = 0$ si $x \leq 0$ y $F(x) = 1$ si $x \geq 1$. Recordemos que la función inversa de $F(x) = x^3$ es $F^{-1}(u) = \sqrt[3]{u} = u^{1/3}$ que, en el recuadro de la expresión numérica de la opción **Calc⇒Calculator** se escribe (empleando la lista de variables y la calculadora de dicho cuadro de diálogo) de la siguiente manera: `'u'**(1/3)`.

4.1.2. Método del rechazo

Para utilizar este método debemos conocer la expresión explícita de la función de densidad, $f(x)$, de la variable aleatoria continua. El procedimiento es el siguiente:

1) Debemos disponer de un intervalo (a, b) tal que $f(x) \neq 0$ para todo $x \in (a, b)$. Y debemos calcular una cota superior de $f(x)$; es decir, un valor k que verifique $f(x) \leq k$ para todo $x \in (a, b)$.

2) En una columna, que podemos denominar **x**, se genera una muestra aleatoria de tamaño grande (al menos del doble del tamaño final deseado), procedente de una distribución uniforme en el intervalo (a, b).

3) En otra columna, que podemos denominar **y**, se genera una muestra aleatoria, del mismo tamaño que en el paso anterior, procedente de una distribución uniforme en el intervalo $(0, k)$.

4) Mediante la opción **Calc⇒Calculator**, se calculan los resultados de la función de densidad para todos los valores de la columna **x**. A la nueva columna la podemos denominar **f(x)**.

5) Mediante la opción **Calc⇒Calculator**, obtenemos una columna que nos indique si $y < f(x)$. La nueva columna (de ceros y unos) la denominaremos **y<f(x)**. Recordemos que un uno significa que sí se cumple la condición y un cero significa que no se cumple la condición.

6) Mediante la opción **Data⇒Unstack Columns** separamos los valores de la columna **x** para los cuales se verifique la condición $y < f(x)$. La columna que contenga estos valores constituirá la muestra deseada.

Este método presenta la desventaja de que no puede elegirse el tamaño muestral resultante.

Como ejemplo, vamos a generar una muestra aleatoria de tamaño grande (no muy lejano de 100) de una variable aleatoria continua cuya función de densidad es $f(x) = \frac{6}{5}(x + x^2)$ si $0 < x < 1$ y cero en el resto. Podemos tener en cuenta las siguientes indicaciones:

- Podemos partir de un tamaño muestral inicial de 250.
- El valor de a es 0 y el valor de b es 1.
- La cota superior de la anterior función de densidad se alcanza en $x = 1$ pues dicha función es creciente. Por tanto:
$$k = \text{cota superior} = f(1) = \frac{6}{5}(1 + 1^2) = \frac{6}{5} \cdot 2 = \frac{12}{5} = 2,4$$
- En el recuadro **Expression** de la opción **Calc⇒Calculator** la función $f(x) = \frac{6}{5}(x + x^2)$ se escribe (empleando la lista de variables y la calculadora de dicho cuadro de diálogo) de la siguiente manera: `(6/5)*('x'+'x'**2)`.
- La opción **Data⇒Unstack Columns** se utiliza, en este eje de la siguiente forma: En **Unstack the data in** seleccionamos, de la lista de variables de la izquierda, la variable **'x'**. En **Using subscripts in** seleccionamos la columna que contiene la procedencia de cada dato, que es **'y<f(x)'**. En **Store unstacked data in** activamos la opción **After last column in use** y dejamos activado **Name the columns containing the unstaked data**.

4.2. Función de distribución empírica

La distribución empírica asociada a una muestra $X_1, \ldots, X_n$ de tamaño n es la distribución de tipo discreto que toma dichos valores con probabilidad igual a $\frac{1}{n}$ para cada uno de ellos. Su correspondiente función de distribución es un estadístico que se aproxima a la verdadera función de distribución de la que proceden los datos de la muestra y se llama función de distribución empírica. Para obtener la función de distribución empírica se procede de la siguiente manera:

I) Se ordenan los valores de la muestra de forma creciente con la opción **Data⇒Sort**, almacenando los nuevos resultados en una columna que podemos denominar **muestra ordenada**.

II) Mediante la opción **Calc⇒Make Patterned Data⇒Simple Set of Numbers** se crea una nueva columna, que podemos denominar **F empírica**, con los valores $\frac{1}{n}$, $\frac{2}{n}$, $\frac{3}{n}$, $\ldots$, 1. Esta columna contiene los resultados de la función de distribución empírica.

Para comprobar que, efectivamente, la función de distribución empírica se aproxima a la función de distribución de la que proceden los datos de la muestra, se puede hacer lo siguiente:

1) Mediante la opción **Calc⇒Probability Distributions** se calculan los resultados de la función de distribución real para todos y cada uno de los valores de la columna **muestra ordenada**. A esta nueva columna la podemos denominar **F real**.

2) Se hace una representación gráfica conjunta de las dos funciones de distribución. Para ello, se selecciona la opción **Graph⇒Scatterplot⇒With connect line**. En el cuadro de diálogo que aparece, junto al **1** en **Y variables** seleccionamos la columna **F empírica** y en **X variables** seleccionamos la columna **muestra ordenada**; junto al **2** en **Y variables** seleccionamos la columna **F real** y en **X variables** seleccionamos otra vez la columna **muestra ordenada**. Luego pulsamos **Multiple graphs** y en el cuadro de diálogo resultante activamos la opción **Overlay on the same graph**. Recordemos que es conveniente quitar los puntos dejando sólo la línea de conexión: para ello, se hace doble *clic* sobre la curva; en **Attributes⇒Symbols** se marca la opción **Custom**, y en **Type** se selecciona **None** (buscando hacia arriba).

Como ejemplo, podemos determinar la función de distribución empírica de la muestra contenida en la columna **100 datos de chi50** (100 datos aleatorios de una distribución chi-cuadrado de Pearson con 50 grados de libertad), comprobando después que, efectivamente, la función de distribución empírica se aproxima a la función de distribución de la que proceden los datos de la muestra. Podemos tener en cuenta las siguientes indicaciones:

- La opción **Data⇒Sort** se utiliza de la siguiente manera: En **Sort column** seleccionamos, de la lista de variables de la izquierda, la variable **'100 datos de chi50'**. En **By column** volvemos a seleccionar la misma columna, **'100 datos de chi50'**. Dejamos desactivada la opción **Descending** para que la ordenación se haga de menor a mayor. En **Store sorted data in** activamos **Column of current worksheet** y lo rellenamos con el nombre que queremos ponerle a dicha columna: **'muestra ordenada'**.
- Para generar una lista con los siguientes 100 números: $1/100 = 0'01$, $2/100 = 0'02$, $3/100 = 0'03$, ..., 1, seguiremos los siguientes pasos: Seleccionamos la opción **Calc⇒Make Patterned Data⇒Simple Set of Numbers**. En **Store patterned data in** tecleamos **'F empírica'**. En **From first value** tecleamos **0,01**, en **To last value** escribimos **1** y en **In steps of** ponemos **0,01**.
- Para calcular los resultados de la función de distribución real para todos y cada uno de los valores de la columna **'muestra ordenada'** hacemos lo siguiente: Se selecciona la opción **Calc⇒Probability Distributions⇒Chi-Square**. En el cuadro de diálogo resultante activamos **Cumulative Probability**. Dejamos lo que sale por defecto en **Noncentrality parameter**. En **Degrees of freedom** ponemos **50**. En **Input column** seleccionamos la columna **'muestra ordenada'** y en **Optional storage** tecleamos el nombre de la nueva columna, **'F real'**.

4.3. Aproximación a la distribución en el muestreo

En general, dado un estadístico T basado en una muestra $X_1, \ldots, X_n$ de algún modelo de probabilidad, no es sencillo encontrar la distribución exacta de T (distribución en el muestreo). Por esta razón, conviene disponer de métodos aproximados para calcularla. Para ello, se puede obtener una muestra de valores de T, $T_1, \ldots, T_m$, generando m muestras de tamaño n del modelo elegido y calculando T para cada una de ellas. La distribución empírica de la muestra $T_1, \ldots, T_m$ es una aproximación de la distribución de T. Este procedimiento se conoce con el nombre de *Método de Montecarlo*. Esta aproximación será mejor cuanto mayor sea m.

Por ejemplo, vamos a aproximar la distribución en el muestreo del estadístico media muestral, $\overline{X} = \frac{X_1+X_2}{2}$ para el modelo normal estándar. Antes de hacer este ejemplo, es conveniente que en la ventana de sesión (**Session**) aparezcan los comandos que va a usar ***Minitab*** en las opciones que vamos a utilizar, para lo cual activamos la ventana de sesión y luego seleccionamos **Editor⇒Enable Commands**. Como ***Minitab*** calcula estadísticos por filas, para aproximar la distribución en el muestreo del estadístico $T = \frac{X_1+X_2}{2}$ en el modelo normal estándar vamos a proceder de la siguiente manera:

1) Mediante la opción **Calc⇒Random Data⇒Normal** generamos, en una columna que denominaremos **X1**, una muestra aleatoria de tamaño 100 procedente de una distribución normal estándar.

2) Generamos, en una columna que denominaremos **X2**, otra muestra aleatoria de tamaño 100 procedente de una distribución normal estándar, otra vez mediante la opción **Calc⇒Random Data⇒Normal**.

3) Mediante la opción **Calc⇒Row Statistics** calculamos la media (**Mean**) por filas de **X1** y **X2**; es decir, determinamos la expresión $\frac{X_1+X_2}{2}$, y guardamos los resultados en una nueva columna que denominaremos **media**. Esta columna contiene una muestra de tamaño 100 del estadístico media muestral, $\overline{X}$.

4) Con la opción **Data⇒Sort** ordenamos los valores de la muestra contenida en la columna **media** de forma creciente, almacenando los nuevos resultados en una columna que denominaremos **media ordenada**.

5) Mediante la opción **Calc⇒Make Patterned Data⇒Simple Set of Numbers** creamos una nueva columna, que denominaremos **F empírica media**, con los valores $1/100 = 0'01$, $2/100 = 0'02$, $3/100 = 0'03$, ..., 1. Esta columna contiene los resultados de la función de distribución empírica del estadístico $\overline{X}$.

6) Mediante la opción **Calc⇒Probability Distributions** calculamos los resultados de la función de distribución real de la media muestral $\overline{X}$ para todos y cada uno de los valores de la columna **media ordenada**. Recordemos que, en este caso, el estadístico media muestral, $\overline{X}$, sigue una distribución normal de media 0 y desviación típica $\frac{1}{\sqrt{2}} = 0,70710678$. A la nueva columna la denominaremos **F media real**.

7) Hacemos una representación gráfica conjunta de las dos funciones de distribución.

Podemos ver, en la ventana de sesión, que los comandos de ***Minitab*** necesarios para hacer este proceso han sido los siguientes:

```
MTB > Name c12 "x1"
MTB > Random 100 'x1';
SUBC>   Normal 0,0 1,0.
MTB > Name c13 "x2"
MTB > Random 100 'x2';
SUBC>   Normal 0,0 1,0.
MTB > Name c14 "media"
MTB > RMean 'x1' 'x2' 'media'.
MTB > Name c15 "media ordenada"
MTB > Sort 'media' 'media ordenada';
SUBC>   By 'media'.
MTB > Name c16 "F empirica media"
MTB > Set 'F empirica media'
DATA>   1( 0,01 : 1 / 0,01 )1
DATA>   End.
MTB > Name c17 "F media real"
MTB > CDF 'media ordenada' 'F media real';
SUBC>   Normal 0,0 0,70710678.
MTB > Plot 'F empirica media'*'media ordenada' 'F media real'*'media ordenada';
SUBC>   Symbol;
SUBC>   Connect;
SUBC>   Overlay.
```

4.3.1. Utilización de macros para la aproximación a la distribución en el muestreo

Minitab contiene un lenguaje de programación sencillo pero potente, que permite elaborar una gran variedad de programas hechos a la medida del usuario. Estos programas se llaman *macros*. Las instrucciones de las macros pueden contener los típicos controladores de flujo que se usan en los lenguajes de programación; por ejemplo:

- `IF/ELSEIF/ELSE/ENDIF` permite ejecutar diferentes bloques de comandos dependiendo de una condición lógica.
- `DO/ENDDO` permite repetir un bloque de comandos una serie de veces.
- `WHILE/ENDWHILE` repite un bloque de comandos mientras la expresión lógica es cierta.
- `NEXT` transfiere el control del flujo a la condición lógica en las sentencias `DO` y `WHILE`.
- `BREAK` sale forzosamente de los bucles `DO` y `WHILE`.
- `GOTO/MLABEL` permite saltar desde la línea `GOTO p` hasta la línea `MLABEL p` saliendo de cualquier bucle, condición, etc. El número p no puede ser una variable, debe ser un dígito.
- `EXIT` termina la macro y devuelve el control a la ventana de sesión de ***Minitab***.

Existen macros globales y macros locales. Las macros locales tienen más posibilidades que las globales. La estructura de una macro local es la siguiente:

`MACRO`	Es obligatorio ponerlo
`[Identificador]`	Nombre + variables de entrada y salida
`# Comentarios`	***Minitab*** no leelas líneas que empiezan por #
`[Declaración de variables]`	Líneas distintas para las constantes, vectores y matrices
`[Cuerpo de la macro]`	
`ENDMACRO`	Es obligatorio ponerlo

Veamos cómo automatizar la aproximación a la distribución en el muestreo del estadístico media muestral para el modelo normal estándar, aprovechando las líneas de comandos de ***Minitab*** que aparecían en la ventana de sesión. Los pasos serán los siguientes:

1) Activamos la ventana de sesión y en el menú **Editor** activamos **Output Editable** y desactivamos **Enable Commands**. Borramos todo el contenido de la ventana de sesión, incluso la fecha.

2) Tecleamos lo siguiente:

```
MACRO
SimulaMedia m n y z
#
# Simula la función de distribución de la media muestral de una variable normal estándar
#
# m: constante que indica el número de muestras
# n: constante que indica el tamaño de las muestras
# y: columna donde se van almacenando las medias y donde luego se ordenan de menor a mayor
# z: columna que almacena la función de distribución empírica de la media muestral
#
MCONSTANT m n i t k     # Declaración de las constantes
MCOLUMN x y z     # Declaración de las variables (vectores)
#
# i: constante que indica el número de iteración
# t: constante auxiliar que va almacenando cada componente del vector y
# k: constante auxiliar que va almacenando cada componente del vector z
# x: columna donde se almacenan las muestras aleatorias
#
DO i=1:m
Random n x;
Normal 0 1.
Mean x t
Let y(i)=t
ENDDO
Sort y y;
By y.
Let k=1/m
Set z
k:1/k
End
Plot z*y;
Connect.
ENDMACRO
```

3) Seleccionamos la opción **File⇒Save Session Windows As**, elegimos el directorio **C:\Archivos de programa\Minitab 14 \Macros** y grabamos el texto de la ventana de sesión con el nombre **SimulaMedia.MAC**.

Para aproximar la distribución en el muestreo del estadístico $\overline{X} = \frac{X_1+X_2}{2}$ del modelo normal estándar podemos hacer lo siguiente. Con la ventana de sesión activada, en el menú **Editor** activamos **Enable Commands** y tecleamos **%SimulaMedia 100 2 c1 c2**. Esto genera 100 muestras aleatorias de tamaño 2 del modelo normal estándar, guarda los resultados de las 100 medias muestrales (ordenadas de menor a mayor) en la columna c1; guarda los resultados de la función de distribución empírica de la media muestral en la columna c2, y representa gráficamente dicha función de distribución empírica, que será la aproximación a la función de distribución en el muestreo del estadístico $\frac{X_1+X_2}{2}$. Aumentando el valor de m se obtiene una mejor aproximación.

4.4. Ejercicios propuestos

4.1. Abrir el fichero de datos (Worksheet) **Acid.mtw** que se encuentra, como ya sabemos, en el directorio **C:\Archivos de programa\Minitab 14 \Data**. Extraer una muestra aleatoria de tamaño 10 (con reemplazamiento) de los datos de la columna **Acid1**. Calcular la media y la cuasi-desviación típica de dicha muestra.

4.2. Mediante el método de la transformada inversa, generar una muestra aleatoria (de tamaño 1000) del modelo cuya función de distribución es $F(x) = x - (x^2/4)$ si $0 < x < 2$, $F(x) = 0$ si $x \leq 0$ y $F(x) = 1$ si $x \geq 2$. La inversa de la función $F(x) = x - (x^2/4)$ para $0 < x < 2$ es $F^{-1}(y) = 2 - 2\sqrt{1-y}$ para $0 < y < 1$.

4.3. Mediante el método del rechazo, generar una muestra aleatoria (de un tamaño no lejano de 100) del modelo cuya función de densidad es $f(x) = x^3/20$ si $1 < x < 3$ (y cero en el resto). ¿Qué tamaño muestral ha salido?

4.4. Obtener una muestra aleatoria de tamaño 1000 del modelo F de Snedecor con 20 grados de libertad en el numerador y 40 grados de libertad en el denominador y comparar (mediante una representación gráfica conjunta) la función de distribución empírica con la función de distribución teórica.

4.5. Aproximar (mediante la creación de una macro) la distribución en el muestreo del estadístico $T = X_1 + X_2 - X_3$ para el modelo normal estándar. Aproximar el valor de la varianza de dicho estadístico. Comparar (mediante una representación gráfica conjunta) la función de distribución empírica y la función de distribución teórica de T. Recordemos que T sigue un modelo normal de media 0 y varianza 3.

5. Inferencia paramétrica y no paramétrica

5.1. Resumen de los contrastes de hipótesis

- *Hipótesis estadística*: afirmación sobre la forma de una o más distribuciones, o sobre el valor de uno o más parámetros de esas distribuciones.
- *Hipótesis nula*: hipótesis estadística que se somete a contraste. Se denota por H_0.
- *Hipótesis alternativa*: es la negación de la hipótesis nula H_0, e incluye todo lo que H_0 excluye. Se denota por H_1.
- *Contraste de hipótesis*: procedimiento que nos capacita para determinar si las muestras observadas difieren significativamente de los resultados esperados, y por tanto nos ayuda a decidir si aceptamos o rechazamos la hipótesis nula.
  - ∗ *Contraste paramétrico*: la hipótesis nula es una afirmación sobre el valor de uno o más parámetros de la variable aleatoria observada en la población.
  - ∗ *Contraste no paramétrico*: la hipótesis nula no es una afirmación sobre el valor de uno o más parámetros de la variable aleatoria observada en la población.
- *Estadístico de contraste*: estadístico que se observa al realizar un contraste de hipótesis, y que nos sirve para aceptar o rechazar la hipótesis nula por poseer una distribución muestral conocida.
- *Región crítica*: zona de la distribución muestral del estadístico de contraste que corresponde a los valores que permiten rechazar la hipótesis nula, y por tanto aceptar la hipótesis alternativa.
- *Región de aceptación*: zona de la distribución muestral del estadístico de contraste que corresponde a los valores que permiten aceptar la hipótesis nula.
- *Contraste unilateral o de una cola*: la región crítica se encuentra en una sola zona de la distribución muestral del estadístico de contraste.
- *Contraste bilateral o de dos colas*: la región crítica se encuentra repartida entre dos zonas de la distribución muestral del estadístico de contraste.
- *Error de tipo I*: error que se comete cuando se decide rechazar una hipótesis nula que en realidad es verdadera.
- ***Nivel de significación***: probabilidad de cometer un error de tipo I. Se denota por $\boldsymbol{\alpha}$.
- *Error de tipo II*: error que se comete cuando se decide aceptar una hipótesis nula que en realidad es falsa. La probabilidad de cometer dicho error se denota por β.
- *Potencia de un contraste*: probabilidad de rechazar la hipótesis nula cuando es falsa. Por tanto, la potencia es igual a $1 - \beta$.
- *p-valor (o nivel crítico)*: es el nivel de significación más pequeño al que una hipótesis nula puede ser rechazada con el estadístico de contraste obtenido. En general, se rechaza H_0 si el p-valor es claramente menor que $0'05$; se acepta H_0 si el p-valor es claramente mayor que $0'05$; y se repite el contraste con una muestra diferente si el p-valor tiene un resultado próximo a $0'05$.

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-valor $> \alpha \Rightarrow$ aceptamos H_0.
Si p-valor $< \alpha \Rightarrow$ rechazamos H_0 y, por tanto, aceptamos H_1.

5.2. Contraste sobre una media. Intervalo de confianza para 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.

5.2.1. Contraste sobre una media cuando la desviación típica poblacional es conocida

condiciones	estadístico	contraste	región crítica
• Muestra aleatoria simple de tamaño n. • σ conocida. • Población Normal ó población cualquiera siempre que $n \geq 30$.	$Z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$Z < -Z_{1-\alpha/2}$ $Z > Z_{1-\alpha/2}$
		$H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$	$Z < -Z_{1-\alpha}$
		$H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$	$Z > Z_{1-\alpha}$

Para hacer este test hay que seleccionar **Stat** ⇒**Basic Statistics** ⇒**1-Sample Z**. Esta opción también nos da el intervalo de confianza para μ.

Para realizar los ejemplos de contrastes paramétricos vamos a utilizar el fichero de datos (Worksheet) **Pulse.mtw**, por lo cual lo abriremos ahora. 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** (1=Baja, 2=Media, 3=Alta).

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 de 70. Si μ denota la media poblacional de la variable X=*Pulso antes de correr*, el contraste que tenemos que hacer es $H_0 : \mu \leq 70$ frente a $H_1 : \mu > 70$.

Como es un test sobre una media poblacional con desviación típica poblacional conocida y como el tamaño muestral es grande ($n = 92$), podemos utilizar la opción **Stat** ⇒**Basic Statistics** ⇒**1-Sample Z**. 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** pues aquí se pondrían los resultados del tamaño muestral y de la media muestral. En **Standard deviation** se teclea el valor de la desviación típica poblacional, σ, que es **10**. En **Test mean** se especifica 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 de 70 pulsaciones por minuto. El intervalo de confianza al 95 % para la media poblacional, asociado a este contraste de hipótesis, es $(71'1547, +\infty)$.

5.2.2. Contraste sobre una media cuando la desviación típica poblacional es desconocida

condiciones	estadístico	contraste	región crítica
• Muestra aleatoria simple de tamaño n. • σ desconocida. • Población Normal ó población cualquiera siempre que $n \geq 30$.	$T = \dfrac{\overline{x} - \mu_0}{S/\sqrt{n}}$	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$T < -t_{n-1\,,\,1-\alpha/2}$ $T > t_{n-1\,,\,1-\alpha/2}$
		$H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$	$T < -t_{n-1\,,\,1-\alpha}$
		$H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$	$T > t_{n-1\,,\,1-\alpha}$

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

Vamos a aplicar este método para comprobar 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 que tenemos que hacer 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 $(70'5897, 75'1494)$.

5.3. Comparación de dos varianzas poblacionales

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.

Una de las técnicas inferenciales para dar solución a este problema es el *test F de Snedecor*, que se puede resumir como sigue:

condiciones	Muestras aleatorias simples independientes de tamaños n_1 y n_2. Poblaciones Normales. μ_1, μ_2 desconocidas.		
estadístico	$F = \dfrac{S_1^2}{S_2^2}$ con $S_1^2 \geq S_2^2$		
contraste	$H_0 : \sigma_1^2 = \sigma_2^2$ $H_1 : \sigma_1^2 \neq \sigma_2^2$	$H_0 : \sigma_1^2 \geq \sigma_2^2$ $H_1 : \sigma_1^2 < \sigma_2^2$	$H_0 : \sigma_1^2 \leq \sigma_2^2$ $H_1 : \sigma_1^2 > \sigma_2^2$
región crítica	$F < \dfrac{1}{F_{n_2-1,n_1-1,1-\alpha/2}}$ $F > F_{n_1-1,n_2-1,1-\alpha/2}$	$F < \dfrac{1}{F_{n_2-1,n_1-1,1-\alpha}}$	$F > F_{n_1-1,n_2-1,1-\alpha}$

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

Ejemplo 1. 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 la variable *Pulso de los hombres antes de correr* y X_2 la variable *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.

Para hacer este contraste se selecciona **Stat** ⇒**Basic Statistics** ⇒**2 Variances**. Se deja activada la opción **Samples in one column**; en **Samples** se selecciona, de la lista de variables de la izquierda, la columna **Pulse1**; en **Subscripts** se selecciona, de la lista de la izquierda, la columna **Sex**; dejamos desactivada la opción **Summarized data** pues aquí se pondrían los resultados de los tamaños muestrales y de las varianzas muestrales. 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.

Ejemplo 2. Comprobemos, ahora, 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 la variable *Pulso de los hombres después de correr* y X_2 la variable *Pulso de las mujeres después de correr*.

Para hacer este contraste se selecciona **Stat** ⇒**Basic Statistics** ⇒**2 Variances**. Se deja activada la opción **Samples in one column**; en **Samples** se selecciona, de la lista de variables de la izquierda, la columna **Pulse2**; en **Subscripts** se selecciona, de la lista de la izquierda, la columna **Sex**; y dejamos desactivada la opción **Summarized data**.

Como resultado de este contraste obtenemos una nueva ventana, en la que 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.

5.4. Comparación de dos medias poblacionales

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.

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

condiciones	Muestras aleatorias simples independientes de tamaños n_1 y n_2. Poblaciones Normales (o cualesquiera si $n_1, n_2 \geq 30$). σ_1, σ_2 desconocidas pero iguales.		
estadístico	$T = \dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}$		
contraste	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$H_0 : \mu_1 \geq \mu_2$ $H_1 : \mu_1 < \mu_2$	$H_0 : \mu_1 \leq \mu_2$ $H_1 : \mu_1 > \mu_2$
región crítica	$T < -t_{n_1+n_2-2\,,\,1-\alpha/2}$ $T > t_{n_1+n_2-2\,,\,1-\alpha/2}$	$T < -t_{n_1+n_2-2\,,\,1-\alpha}$	$T > t_{n_1+n_2-2\,,\,1-\alpha}$

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

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 la variable *Pulso de los hombres antes de correr* y X_2 la variable *Pulso de las mujeres antes de correr*. En el **Ejemplo 1** de la sección 5.3 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 sean normales, 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 se selecciona **Stat** ⇒**Basic Statistics** ⇒**2-Sample t**. Se deja activada la opción **Samples in one column**; en **Samples** se selecciona, de la lista de variables de la izquierda, la columna **Pulse1**; en **Subscripts** se selecciona, de la lista de la izquierda, la columna **Sex**; dejamos desactivada la opción **Summarized data** pues aquí se pondrían los resultados de los tamaños muestrales y de las medias muestrales; 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 ($76'9$) es mayor que la media muestral del pulso de los hombres ($70'42$) podríamos, incluso, aceptar que la media poblacional del pulso de las mujeres es mayor que la media poblacional del pulso de los hombres. El intervalo de confianza al 95 % para la diferencia de medias poblacionales, $\mu_1 - \mu_2$, es $(-10'96232, -1'90986)$.

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

condiciones	Muestras aleatorias simples independientes de tamaños n_1 y n_2. Poblaciones Normales (o cualesquiera si $n_1, n_2 \geq 30$). σ_1, σ_2 desconocidas y distintas.		
estadístico	$T = \dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$		
grados de libertad	g=nº natural más próximo a $\dfrac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$		
contraste	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$H_0 : \mu_1 \geq \mu_2$ $H_1 : \mu_1 < \mu_2$	$H_0 : \mu_1 \leq \mu_2$ $H_1 : \mu_1 > \mu_2$
región crítica	$T < -t_{g\,,\,1-\alpha/2}$ $T > t_{g\,,\,1-\alpha/2}$	$T < -t_{g\,,\,1-\alpha}$	$T > t_{g\,,\,1-\alpha}$

Para realizar este test paramétrico hay que seleccionar, igual que antes, **Stat ⇒Basic Statistics ⇒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**.

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 la variable *Pulso de los hombres después de correr* y X_2 la variable *Pulso de las mujeres después de correr*. En el **Ejemplo 2** de la sección 5.3 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 sean 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 se selecciona **Stat ⇒Basic Statistics ⇒2-Sample t**. Se deja activada la opción **Samples in one column**; en **Samples** se selecciona, de la lista de variables de la izquierda, la columna **Pulse2**; en **Subscripts** se selecciona, de la lista de la izquierda, la columna **Sex**; dejamos desactivadas las opciones **Summarized data** y **Assume equal variances**. Si pulsamos el botón **Options** nos 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 $(-18'6493, -3'0249)$.

5.4.3. Comparación de dos medias con muestras relacionadas (apareadas o asociadas)

condiciones	Muestras aleatorias simples apareadas de tamaño n. La variable aleatoria $D = X_1 - X_2$ es Normal (o cualquiera si $n \geq 30$).		
estadístico	$T = \dfrac{\overline{d}}{\dfrac{S_d}{\sqrt{n}}}$ donde $\overline{d}$ y S_d son la media y la cuasidesviación típica de D		
contraste	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$H_0 : \mu_1 \geq \mu_2$ $H_1 : \mu_1 < \mu_2$	$H_0 : \mu_1 \leq \mu_2$ $H_1 : \mu_1 > \mu_2$
región crítica	$T < -t_{n-1\,,\,1-\alpha/2}$ $T > t_{n-1\,,\,1-\alpha/2}$	$T < -t_{n-1\,,\,1-\alpha}$	$T > t_{n-1\,,\,1-\alpha}$

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

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 la variable *Pulso antes de correr* y X_2 la variable *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 las variables aleatorias X_1 y X_2 no sean normales, se puede aplicar este contraste debido a que los tamaños muestrales son suficientemente grandes: $n_1 = n_2 = n = 92$.

Para hacer este contraste se selecciona **Stat ⇒Basic Statistics ⇒Paired t**. Se deja activada la opción **Samples in columns**; en **First sample** se selecciona, de la lista de variables de la izquierda, la columna **Pulse1**; en **Second sample** se selecciona, de la lista de variables de la izquierda, la columna **Pulse2**; y dejamos desactivada la opción **Summarized data (differences)** pues aquí se pondrían los resultados del tamaño muestral y de la media muestral de las diferencias. Si pulsamos el botón **Options** nos aparece un cuadro de diálogo similar al de la opción anterior (**2-Sample t⇒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 $0'000$, 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, 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 ($80'0000$) es mayor que la media muestral del pulso antes de correr ($72'8696$) 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 $(-9'92027, -4'34060)$.

5.5. Contrastes no paramétricos de bondad de ajuste

Los contrastes de hipótesis presentados en las secciones anteriores coinciden en dos características: permiten contrastar hipótesis referidas a algún parámetro y requieren del cumplimiento de determinadas condiciones sobre las poblaciones originales de las que se extraen los datos (generalmente normalidad). Estas dos características combinadas permiten agrupar a este tipo de contrastes en una gran familia de técnicas denominadas *contrastes paramétricos*. Pero en muchas ocasiones no se cumplen las condiciones necesarias para poder hacer un contraste paramétrico, por lo que se tienen que aplicar otras técnicas que llamaremos *contrastes no paramétricos*.

En los contrastes no paramétricos de bondad de ajuste se trata de determinar, a través de una muestra, si una variable aleatoria se ajusta bien a una cierta distribución dada de antemano (Normal, Exponencial, Weibull, etc.).

5.5.1. Gráficos probabilísticos

Este método de bondad de ajuste se basa en el hecho de que si una muestra, $X_1, \ldots, X_n$, proviene de un modelo con función de distribución F, entonces $F(X_1), \ldots, F(X_n)$ es una muestra del modelo Uniforme en el intervalo $(0, 1)$, por lo que, una vez ordenada, los valores esperados de dicha muestra serán: $1/n$, $2/n$, $\cdots$, 1. De esta forma, si representamos gráficamente los $F(X_i)$ ordenados frente a los i/n, el gráfico debe ser aproximadamente una línea recta. En algunos casos esta linealidad se mantiene aunque se estimen los parámetros desconocidos de F. Es decir, el ajuste será bueno si la gráfica es aproximadamente una recta. Este tipo de técnicas dan sólo una aproximación gráfica, aunque, en algunos casos, van acompañados de algún contraste de bondad de ajuste. Si es así, aceptaremos la hipótesis nula de ajuste a la distribución teórica si el p-valor es mayor que el nivel de significación (que usualmente es $\alpha = 0'05$). Para realizar los gráficos probabilísticos se selecciona **Graph⇒Probability Plot**.

Vamos a utilizar este método para comprobar si las variables aleatorias **Pulse1** (pulso antes de correr) y **Pulse2** (pulso después de correr) pueden considerarse Normales (cuando están observadas en toda la población). Para ello, seleccionamos **Graph⇒Probability Plot⇒Single**. En **Graph variables** seleccionamos, de la lista de variables de la izquierda, las columnas **Pulse1** y **Pulse2**; pulsamos en **Distribution** y, en el cuadro de diálogo resultante, dejamos lo que está por defecto (**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 dos 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 test de Anderson-Darling.

Podemos ver que el gráfico probabilístico de la variable **Pulse1** no se aproxima mucho a una recta. Además, el p-valor del test de normalidad es igual a $0'013$. Si consideramos un nivel de significación de $\alpha = 0'01$ entonces el p-valor es levemente mayor que α, por lo que podríamos aceptar la hipótesis nula de que la variable **Pulse1** es Normal. Pero si consideramos un nivel de significación de $\alpha = 0'05$ (que es lo usual) entonces el p-valor es menor que α, por lo que no podemos aceptar la hipótesis nula de que la variable **Pulse1** es Normal.

Por otra parte, podemos observar que el gráfico probabilístico de la variable **Pulse2** tampoco se aproxima mucho a una recta. Además, el p-valor del test de normalidad es, en este caso, menor que $0'005$. Ahora, 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, por lo que no podemos aceptar la hipótesis nula de que la variable **Pulse2** es Normal. Se puede comprobar que si hacemos el mismo procedimiento para comprobar si **Pulse1** sigue un modelo Lognormal, el gráfico resultante se aproxima a una recta y además, el p-valor es $0'159$, claramente mayor que los habituales niveles de significación ($0'05$ ó $0'01$), por lo que podríamos aceptar que **Pulse1** sigue un modelo Lognormal.

5.5.2. Contraste de normalidad

El problema de comprobar la normalidad de una variable aleatoria, a partir de los datos proporcionados por una muestra, ha sido tratado a menudo debido al uso frecuente de esta hipótesis en la estadística inferencial.

Una de las técnicas para contrastar la hipótesis nula H_0: *la variable aleatoria observada en la población es Normal* frente a la hipótesis alternativa H_1: *la variable aleatoria observada en la población no es Normal* es el método de Kolmogorov-Smirnov (y Lilliefors), que se puede resumir en la tabla siguiente:

contraste	H_0: la variable aleatoria X observada en la población es Normal H_1: la variable aleatoria X observada en la población no es Normal
condiciones	La variable no es cualitativa nominal. Se extrae una muestra aleatoria simple de tamaño n.
estadístico	$L = \text{máx } \lvert F_i^0 - H_i \rvert$, donde F_i^0 = función de distribución de $\mathcal{N}(\overline{x}, S)$ evaluada en el dato muestral i-ésimo (o en el extremo superior del intervalo de clase i-ésimo), H_i = frecuencia relativa acumulada del dato (o del intervalo de clase) i-ésimo.
región crítica	$L \geq l_{n\,,\,\alpha}$ de la tabla de los puntos críticos de este contraste.

Si queremos ajustar a un modelo Normal, en ***Minitab*** podemos usar la opción **Stat⇒Basic Statistics⇒Normality Test**.

Vamos a utilizar esta opción para comprobar si se puede aceptar que la variable **Height** (altura, en pulgadas) puede considerarse Normal. Para ello usamos **Stat⇒Basic Statistics⇒Normality Test**; en **Variable** seleccionamos, de la lista de variables de la izquierda, la columna **Height**; en **Percentile Lines** dejamos lo que está activado por defecto, que es **None**; en **Tests for Normality** podemos activar uno de los siguientes tres tests: Anderson-Darling, Ryan-Joiner o Kolmogorov-Smirnov. Por ejemplo, vamos a activar el último test, **Kolmogorov-Smirnov**. El recuadro **Title** vamos a dejarlo en blanco. El resultado es un gráfico probabilístico en el cual también está indicado (en la parte superior derecha) el p-valor, que es mayor que $0'15$. Este p-valor es claramente mayor que los habituales niveles de significación ($0'05$ ó $0'01$), por lo que podríamos aceptar que la variable **Height** sigue un modelo Normal.

5.6. Contraste chi-cuadrado sobre independencia de dos variables

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

Tenemos n individuos de una muestra, los cuales pueden clasificarse con arreglo a dos variables cualitativas X e Y; la primera de ellas con r clases, y la segunda con k clases, dando así lugar a una *tabla de contingencia* o *de doble entrada* como la siguiente:

$X\backslash Y$	B_1	B_2	$\cdots$	B_j	$\cdots$	B_k	suma
A_1	f_{11}	f_{12}	$\cdots$	f_{1j}	$\cdots$	f_{1k}	f_{1*}
A_2	f_{21}	f_{22}	$\cdots$	f_{2j}	$\cdots$	f_{2k}	f_{2*}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	f_{i1}	f_{i2}	$\cdots$	f_{ij}	$\cdots$	f_{ik}	f_{i*}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_r	f_{r1}	f_{r2}	$\cdots$	f_{rj}	$\cdots$	f_{rk}	f_{r*}
suma	f_{*1}	f_{*2}	$\cdots$	f_{*j}	$\cdots$	f_{*k}	n

En dicha tabla, f_{ij} es igual al número de individuos que pertenecen a la clase A_i de la variable X y a la clase B_j de la variable Y, f_{i*} es igual al número de individuos en la categoría A_i de la variable X independientemente del valor de Y, y f_{*j} es igual al número de individuos en la categoría B_j de la variable Y independientemente del valor de X.

La pregunta a responder es si las variables X e Y están relacionadas o no. La hipótesis nula puede enunciarse de diversas formas: *las variables X e Y son independientes*, o *X e Y no están relacionadas*, o *X e Y no están asociadas*.

Denotaremos por e_{ij} a las frecuencias esperadas si fuese cierta la hipótesis nula. Estas frecuencias se calculan de la siguiente forma:

$$e_{ij} = \frac{f_{i*} \cdot f_{*j}}{n}\,.$$

Consideremos el estadístico:

$$\chi^2_{exp} = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}\,.$$

Si se cumple la hipótesis nula de independencia entre las dos variables aleatorias, se puede demostrar que dicho estadístico sigue una distribución χ^2 de Pearson con $(r-1)(k-1)$ grados de libertad. Por tanto, siguiendo el esquema clásico de los contrastes de hipótesis, rechazaremos la hipótesis nula cuando $\chi^2_{exp} \geq \chi^2_{(r-1)(k-1)\,,\,1-\alpha}$.

La condición de validez de este contraste es que ninguna e_{ij} sea menor que 1, y no más del 20 % de ellas sean inferiores a 5. Cuando la condición no se verifica, el test no puede hacerse, aunque para evitarlo se pueden unir clases o se puede aumentar el tamaño de la muestra.

En ***Minitab*** hay dos formas de aplicar este contraste, según tengamos recogidos los datos:

5.6.1. Datos en una tabla de doble entrada

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

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

	usuarios	no usuarios
hombres	6	24
mujeres	14	16

Este ejemplo se haría, *a mano*, de la siguiente manera: La hipótesis nula es H_0: *no existe relación entre el sexo y el uso de la biblioteca*, y por tanto la hipótesis alternativa es H_1: *existe relación entre el sexo y el uso de la biblioteca*. Las frecuencias esperadas bajo la hipótesis nula son las que aparecen en la tabla siguiente:

	usuarios	no usuarios
hombres	10	20
mujeres	10	20

Así pues, el valor del estadístico de contraste es:

$$\begin{aligned}
\chi^2_{exp} &= \sum\sum \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \\
&= \frac{(6-10)^2}{10} + \frac{(24-20)^2}{20} + \frac{(14-10)^2}{10} + \frac{(16-20)^2}{20} \\
&= 1'6 + 0'8 + 1'6 + 0'8 = 4'8\,.
\end{aligned}$$

En nuestro caso, $r = 2$ y $k = 2$, y por tanto $(r-1)(k-1) = 1$. Si elegimos un nivel de significación $\alpha = 0'1$, obtenemos:

$$\chi^2_{(r-1)(k-1)\,,\,1-\alpha} = \chi^2_{1\,,\,0'9} = 2'70554\,.$$

Como $\chi^2_{exp} \geq \chi^2_{(r-1)(k-1)\,,\,1-\alpha}$ entonces rechazamos la hipótesis nula. Aceptamos que existe asociación (relación o dependencia) entre el sexo y la utilización de la biblioteca.

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

↓	C1	C2
	SI	NO
1	6	24
2	14	16

Ahora seleccionamos **Stat⇒Tables⇒Chi-Square Test (Table in Worksheet)**; en **Columns containing the table** elegimos, de la lista de variables de la izquierda, las columnas **C1** y **C2**; es decir, **SI** y **NO**, y pulsamos en **OK**. En la ventana de sesión podemos ver el resultado del p-valor, que es $0'028$. Si consideramos un nivel de significación de $\alpha = 0'01$ entonces el p-valor es mayor que α, por lo que podríamos aceptar la hipótesis nula de independencia. Pero si consideramos un nivel de significación de $\alpha = 0'05$ (que es lo usual) entonces el p-valor es menor que α, por lo que no podríamos aceptar la hipótesis nula de independencia, aceptando entonces que existe relación entre el sexo y el uso de la biblioteca.

5.6.2. Datos en dos (o tres) columnas

Si en la hoja de datos éstos se encuentran recogidos en dos (o tres) columnas, se utiliza **Stat⇒Tables⇒Cross Tabulation and Chi-Square**.

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

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

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

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

Ejemplo 2. Para utilizar la opción **Stat⇒Tables⇒Cross Tabulation and Chi-Square** no es necesario que tengamos una columna con las frecuencias de cada combinación de resultados de dos variables; también se puede utilizar dicha opción si solamente tenemos **dos columnas** que contienen los resultados de una variable bidimensional, (x_i, y_i), pero es necesario que las dos variables sean de tipo discreto, con pocos resultados distintos; de lo contrario no se puede aplicar este contraste (recordemos que las frecuencias esperadas bajo la hipótesis de independencia han de ser todas mayores que 1 y no más del 20 % de ellas pueden ser menores que 5).

Para hacer un ejemplo de este caso, vamos a activar la hoja de datos **Pulse.mtw**. Vamos a comprobar si existe dependencia entre las variables **Smokes** (la persona es fumadora o no) y **Sex** (sexo). La hipótesis nula es H_0: *no*

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

```
Rows: Smokes   Columns: Sex

              1      2    All

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

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

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

Cell Contents:      Count
                    Expected count


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

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

5.7. Contraste chi-cuadrado sobre homogeneidad de dos poblaciones

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

El contraste chi-cuadrado de homogeneidad es el mismo que el test chi-cuadrado de independencia de variables explicado en el apartado anterior, aunque la hipótesis nula no sea la misma. El test, por tanto, puede resumirse como sigue:

contraste	H_0: las dos poblaciones son homogéneas H_1: las dos poblaciones no son homogéneas
estadístico	$\chi^2_{exp} = \sum_{i=1}^{2}\sum_{j=1}^{k} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$, donde $k =$ número de clases, $f_{ij} =$ frecuencia absoluta observada en la muestra i para la clase j, $f_{i*} =$ tamaño de la muestra i, $f_{*j} =$ número de individuos en la clase j, $e_{ij} = \frac{f_{i*} \cdot f_{*j}}{n} =$ frecuencia esperada bajo H_0.
condiciones	Las dos muestras aleatorias son independientes. $e_{ij} \geq 1$ para todas las clases. $e_{ij} \geq 5$, salvo para un 20 % de las clases como máximo.
región crítica	$\chi^2_{exp} \geq \chi^2_{k-1\,,\,1-\alpha}$

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

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

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

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

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

privadas	públicas
6	30
14	32
17	17
9	3

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

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

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

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

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

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

Total         46        82    128

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

Como solamente una de las frecuencias esperadas es menor que 5, podemos aplicar esta técnica. El resultado del p-valor es $0'001$, claramente menor que los habituales niveles de significación ($0'05$ ó $0'01$) por lo que rechazamos la hipótesis nula y, en consecuencia, aceptamos que la distribución de los resultados de la prueba no es la misma en las universidades públicas que en las privadas.

5.8. Ejercicios propuestos

5.1. 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

Comprobar, mediante el contraste de Kolmogorov-Smirnov, si la variable aleatoria X=*Tiempo empleado en consultar bases de datos por ordenador* es Normal. Si es posible, responder a la siguiente pregunta: ¿se puede aceptar que la media poblacional del tiempo empleado en consultar bases de datos por ordenador es mayor que 15 minutos?

5.2. Los siguientes datos corresponden a las edades de una muestra de 10 personas que visitan un centro de cálculo.

19	24	83	30	17	23	33	19	68	56

Mediante la realización de un gráfico probabilístico, comprobar si la variable aleatoria *X=Edad de las personas que visitan el centro de cálculo* es Normal. Si es posible, responder a la siguiente pregunta: ¿se puede aceptar que la media poblacional de la edad de las personas que visitan el centro de cálculo es menor que 40 años?

5.3. En la siguiente tabla aparece el número de préstamos diarios realizados por dos bibliotecas durante 20 días elegidos al azar.

Biblioteca A	65	74	47	81	71	52	74	81	48	68
Biblioteca B	57	63	38	70	68	46	63	75	39	57

¿Se puede aceptar, con un nivel de significación de 0'05, que la varianza poblacional del número de préstamos diarios realizados por la biblioteca A es igual a la varianza poblacional del número de préstamos diarios realizados por la biblioteca B? ¿Se puede aceptar, con un nivel de significación de 0'05, que el número medio poblacional de préstamos diarios realizados por la biblioteca A es igual al número medio poblacional de préstamos diarios realizados por la biblioteca B?

5.4. Se les preguntó a 30 matrimonios, elegidos al azar, el número de veces que habían ido a alguna biblioteca en los tres últimos meses, siendo los resultados los siguientes:

Hombre	Mujer	Hombre	Mujer	Hombre	Mujer
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

¿Podemos afirmar que hay diferencia significativa entre los hombres y las mujeres de los matrimonios en cuanto al número medio de veces que van a la biblioteca?

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

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

Determinar si existe relación entre el rango académico y la opinión de los profesores respecto del proyecto mencionado.

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

	científicos	de ficción
área A	870	745
área B	304	251

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