Práctica 2. Estadística descriptiva

1. Distribución de frecuencias

Para determinar la distribución de frecuencias de una o más variables, utilizamos la opción **Stat⇒Tables ⇒Tally Individual Variables**.

Para practicar esta opción, podemos abrir el archivo 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. Estadística descriptiva con la opción *Stat ⇒Basic Statistics ⇒Display Descriptive Statistics*

En la **Práctica 1** vimos que la opción **Calc⇒Column Statistics** calcula, para una columna (o variable), **uno** de los estadísticos siguientes: **Sum** (suma), **Mean** (media arimética), **Standard deviation** (cuasi-desviación típica), **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 de 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, vamos a calcular los estadísticos descriptivos más importantes de las variables **Pulse1**, **Height** y **Weight** de la hoja de datos **Pulse.mtw**. Para ello, seleccionamos

Stat⇒Basic Statistics⇒Display Descriptive Statistics y en el recuadro **Variables** del cuadro de diálogo resultante seleccionamos, de la lista de columnas que tenemos a la izquierda, las tres variables **'Pulse1'**, **'Height'** y **'Weight'**. 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	$\overline{x} = \dfrac{\sum_{i=1}^{N} x_i}{N}$
SE Mean	error estándar de la media	$\dfrac{S_x}{\sqrt{N}}$
StDev	cuasi-desviación típica	$S_x = \sqrt{\dfrac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{N-1}}$
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 (**Pulse.mtw**) podemos calcular los estadísticos de la variable **Pulse2** (Pulso después de correr) separando sus resultados según los valores de la variable **Ran** (¿corrió o no corrió?). Para ello, seleccionamos **Stat⇒Basic Statistics⇒Display Descriptive Statistics**; en el recuadro **Variables** del cuadro de diálogo resultante seleccionamos la variable **'Pulse2'**; y en **By variables (Optional)** seleccionamos la variable **'Ran'**. En consecuencia, en la ventana de sesión aparecen los resultados de los mencionados estadísticos de la variable **Pulse2** separados para cada grupo de resultados de la variable **Ran**. Por ejemplo, podemos comprobar que para el grupo de personas que sí corrió (**Ran=1**) la media del pulso es $92'51$ y la mediana es 88, mientras que para el grupo de personas que no corrió (**Ran=2**) la media del pulso es $72'32$ y la mediana es 70.

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	$\overline{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$
SE of mean	error estándar de la media	$\dfrac{S_x}{\sqrt{n}}$
Standard deviation	cuasi-desviación típica	$S_x = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$
Variance	cuasi-varianza	S_x^2
Coefficient of variation	coeficiente de variación insesgado	$CV = 100\,\dfrac{S_x}{\lvert\overline{x}\rvert}$
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 menores y el 5 % de los mayores	
Sum	suma	$\sum_{i=1}^{n} x_i$
Minimum	mínimo dato	x_{min}
Maximum	máximo dato	x_{max}
Range	recorrido 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 (solo cuando se ha rellenado el recuadro **By variables**)	
Percent	porcentaje de casos (solo cuando se ha rellenado el recuadro **By variables**)	
Cumulative percent	porcentaje acumulado de casos (solo cuando se ha rellenado el recuadro **By variables**)	
Sum of squares	suma de cuadrados	$\sum_{i=1}^{n} x_i^2$
Skewness	coeficiente de asimetría	$g_1 = \dfrac{m_3}{S^3}$, con $m_3 = \dfrac{\sum_{i=1}^{n} (x_i - \overline{x})^3}{n-1}$
Kurtosis	coeficiente de apuntamiento	$g_2 = \dfrac{m_4}{S^4} - 3$, con $m_4 = \dfrac{\sum_{i=1}^{n} (x_i - \overline{x})^4}{n-1}$
MSSD	media de los cuadrados de las sucesivas diferencias	

Como ejemplo, vamos a comprobar que la suma de los datos de la variable **Pulse1** de la hoja de datos **Pulse.mtw** es 6704 y la suma de los cuadrados de los datos de la misma variable es 499546.

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 cuasi-desviación 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, con la hoja de datos **Pulse.mtw** vamos a dibujar el histograma (con la curva normal superpuesta) de la variable **Height**.

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).

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 seleccionando, del menú contextual que resulta, la opción **Copy Graph**. De esta manera, podríamos pegarlo en otro programa bajo Windows, por ejemplo, uno de edición de gráficos. También podemos almacenarlo en la ventana de proyecto, **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 archivo aparte.

Una vez obtenido el histograma es posible cambiar su aspecto. Para ello, hacemos doble *clic* sobre la parte del gráfico que queremos cambiar. Aparece, entonces, una nueva ventana que nos permite hacer dicha transformación. 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'*, en letra Arial, cursiva, negrita, de color azul oscuro y con un tamaño de 10 puntos.
- Que las barras sean de color azul claro con una trama de relleno oblicua y con los bordes de color azul oscuro.
- Que haya 7 intervalos de la misma amplitud y que en el eje horizontal aparezcan los límites de los intervalos (no los puntos medios).
- Que el texto del eje horizontal sea *Peso de los alumnos, en libras*, en letra Arial, cursiva, no negrita, de color azul oscuro y con un tamaño de 9 puntos.
- Que en el eje vertical se muestren 13 marcas (*ticks*), en letra Arial, de color azul oscuro y con un tamaño de 8 puntos.
- Que el texto del eje vertical sea *Frecuencia absoluta*, en letra Arial, cursiva, no negrita, de color azul oscuro y con un tamaño de 9 puntos.

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

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

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

Vamos a aprender a hacer un diagrama de sectores cuando tenemos en una columna las categorías de una variable y en otra columna las frecuencias absolutas de dichas categorías. Por ejemplo, vamos a realizar el diagrama de sectores de los datos de la Figura 1, correspondientes a los idiomas en que están escritos los libros de los estantes de una determinada biblioteca.

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

Figura 1: Idioma de los libros de una biblioteca

Como estos datos no tienen nada que ver con los datos del archivo **Pulse.mtw**, abrimos una nueva hoja de datos con la opción **File⇒New**. En el cuadro de diálogo que aparece seleccionamos **Minitab Woorksheet**. A esta nueva hoja de datos ***Minitab*** le asignará el nombre Worksheet J, siendo J un número natural. A continuación introducimos los datos tal como se muestra en la Figura 1. Luego guardamos esta hoja de datos con el nombre **IdiomaLibros.mtw** (**File⇒Save Current Worksheet As**). Para dibujar el diagrama de sectores seleccionamos **Graph⇒Pie Chart**. En el cuadro de diálogo resultante, activamos la opción **Chart values from a table**; seleccionamos la columna **'Idioma'** en el recuadro **Categorical Variable**; seleccionamos la columna **'Nº de estantes'** en el recuadro **Summary variables** y pulsamos en **OK**. Como ya sabemos, podemos modificar este gráfico.

4.3. Diagrama de barras

4.3.1. Diagrama de barras simple

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 segmentos rectilíneos) 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. En el cuadro de diálogo resultante, seleccionamos la columna **'Activity'** en el recuadro **Categorical Variables**. Como las categorías son números concretos (0, 1, 2 y 3) es más riguroso que, en vez de barras, aparezcan solamente segmentos rectilíneos; por tanto, activamos el botón **Data View** y en el cuadro de diálogo resultante activamos solo la opción **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 archivo aparte. También es posible cambiar su aspecto, una vez obtenido, haciendo doble *clic* sobre la parte del gráfico que queremos cambiar. Podemos observar, además, que si hacemos *clic* sobre el gráfico (para activarlo) y luego pasamos el ratón por encima de las barras, se nos indica la frecuencia absoluta de cada categoría. 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'*, en letra Comic Sans MS, cursiva, negrita, de color rojo y con un tamaño de 11 puntos.
- Que las barras (líneas) sean de color rojo y de un tamaño (grosor) de 3 puntos.
- Que en el eje vertical se muestren 13 marcas (*ticks*), en letra Arial, no negrita, de color rojo y con un tamaño de 10 puntos.
- Que el texto del eje vertical sea *Frecuencia absoluta*, en letra Arial, cursiva, no negrita, de color rojo y con un tamaño de 9 puntos.
- Que el texto del eje horizontal sea *Actividad Física (0=Ninguna, 1=Baja, 2=Media, 3=Alta)*, en letra Arial, cursiva, no negrita, de color rojo y con un tamaño de 8 puntos.
- Que en la parte superior de cada barra aparezca la frecuencia absoluta de cada categoría (*clic* sobre una de las barras con el botón derecho del ratón, opción **Add**, **Data Labels**, dejar activado **Use y-values labels**).

Vamos a aprender a hacer un diagrama de barras cuando tenemos en una columna las categorías de una variable y en otra columna las frecuencias absolutas de dichas categorías. Por ejemplo, vamos a realizar el diagrama de barras de los datos de la Figura 1, correspondientes a los idiomas en que están escritos los libros de los estantes de una determinada biblioteca. En primer lugar, abrimos la hoja de datos **IdiomaLibros.mtw**. Para dibujar el diagrama de barras seleccionamos **Graph⇒Bar Chart**, activamos la opción **Values from a table** del apartado **Bars represent**; activamos el modelo **Simple** del apartado **One column of values** y pulsamos en **OK**. En el cuadro de diálogo resultante, seleccionamos la columna **'Nº de estantes'** en el recuadro **Graph variables**; seleccionamos la columna **'Idioma'** en el recuadro **Categorical Variable** y pulsamos en **OK**. Como ya sabemos, podemos modificar este gráfico.

4.3.2. Diagrama de barras agrupado (o apilado)

Con la opción **Graph⇒Bar Chart** existe la posibilidad de seleccionar una nueva variable para determinar las barras dentro de cada grupo; esto se realiza seleccionando **Cluster** (para un diagrama de barras agrupado según los resultados de otra variable) o **Stack** (para un diagrama de barras apilado según los resultados de otra variable). Por ejemplo, con el archivo de datos **Pulse.mtw** vamos a hacer el diagrama de barras de 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'*, en letra Verdana, negrita, de color morado y con un tamaño de 9 puntos.
- Que las barras tengan distinto color según los resultados de la variable **Sex** y que aparezca una leyenda explicativa (doble *clic* sobre una de las barras, en el cuadro de diálogo resultante seleccionar la carpeta **Groups**, en el recuadro **Assign attributes by categorical variables** seleccionar la variable **Sex**.)
- Que en el eje vertical se muestren 10 marcas (*ticks*), en letra Verdana, no negrita, de color morado y con un tamaño de 10 puntos.
- Que el texto del eje vertical sea *Frecuencia absoluta*, en letra Verdana, no negrita, de color morado y con un tamaño de 11 puntos.
- Que en el eje horizontal todo esté escrito con la fuente Verdana, no negrita, de color morado y con un tamaño de 9 puntos. Que en dicho eje aparezcan los nombres de las variables en español: *Actividad Física* en vez de *Activity*, y *Sexo* en vez de *Sex*. Que en el mismo eje los resultados de la variable *Sex* no sean 1 y 2 sino *Hombre* y *Mujer*. Y los resultados de la variable *Activity* no sean 0, 1, 2 y 3 sino *Ninguna*, *Poca*, *Media* y *Alta*.

Vamos a aprender a hacer un diagrama de barras agrupado (o apilado) cuando tenemos los datos en una tabla de doble entrada. Por ejemplo, vamos a realizar el diagrama de barras agrupado de los datos de la Figura 2, correspondientes al número de citas en diferentes campos de investigación y en tres distintos años.

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

Figura 2: Citas anuales en distintos campos de investigación

En primer lugar, abrimos una nueva hoja de datos con la opción **File⇒New**. En el cuadro de diálogo que aparece seleccionamos **Minitab Woorksheet**. A continuación introducimos los datos tal como se muestra en la Figura 2. Luego guardamos esta hoja de datos con el nombre **Citas.mtw**. Para dibujar el diagrama de barras agrupado seleccionamos **Graph⇒Bar Chart**, activamos la opción **Values from a table** del apartado **Bars represent**; activamos el modelo **Cluster** del apartado **Two-way table** y pulsamos en **OK**. En el cuadro de diálogo resultante, seleccionamos las columnas **'1970'**, **'1980'** y **'1990'** en el recuadro **Graph variables**; seleccionamos la columna **'Campo investigación'** en el recuadro **Row labels**; activamos **Rows are outermost categories and columns are innermost** y, por último, pulsamos en **OK**. Como ya sabemos, podemos modificar este gráfico.

4.4. Diagramas bivariantes

4.4.1. Diagrama de dispersión o nube de puntos

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

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

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

- Que el título sea *Diagrama de dispersión de la 'Altura' frente al 'Peso'*, en letra Times New Roman, cursiva, negrita, de color rojo y con un tamaño de 14 puntos.
- Que los símbolos sean rombos rojos de tamaño 1.
- Que en el eje horizontal se muestren 14 marcas (*ticks*), en letra Times New Roman, no negrita, de color rojo y con un tamaño de 12 puntos.
- Que el texto del eje horizontal sea *Peso de los alumnos, en libras*, en letra Times New Roman, cursiva, no negrita, de color rojo y con un tamaño de 12 puntos.
- Que en el eje vertical se muestren 10 marcas (*ticks*), en letra Times New Roman, no negrita, de color rojo y con un tamaño de 12 puntos.
- Que el texto del eje vertical sea *Altura de los alumnos, en pulgadas*, en letra Times New Roman, cursiva, no negrita, de color rojo y con un tamaño de 12 puntos.
- Que la recta de regresión sea de color rojo y de tamaño 2.

4.4.2. Representación gráfica de una función y=f(x)

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 $f(x) = x^2 + 2x - 4$ en el intervalo $[-3, 3]$. Para ello se procede de la siguiente manera:

1) Se abre una hoja de datos nueva (**File**, **New**, **Minitab Worksheet**).

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 $0,01$. Se puede comprobar que en la columna **x** hay un total de 601 números.

3) En otra columna se calculan los resultados de la función función $f(x) = x^2 + 2x - 4$ 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: `'x'**2+2*'x'-4`

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**.

5. Correlación y regresión lineal

En el apartado 4.4 hemos visto cómo obtener (y cómo modificar) el diagrama de dispersión o nube de puntos de una variable estadística bidimensional.

Para obtener el **coeficiente de correlación lineal** de Pearson se selecciona **Stat⇒Basic Statistics ⇒Correlation**. En el cuadro de diálogo que aparece, en el recuadro de la izquierda está la lista de variables, de las cuales podemos seleccionar dos o más.

Por ejemplo, de la hoja de datos **Pulse.mtw** vamos a calcular el coeficiente de correlación lineal de Pearson entre las variables Altura en pulgadas, **Height**, y Peso en libras, **Weight** y lo vamos a guardar para poder aumentar el número de decimales que se obtienen. Para ello, seleccionamos **Stat⇒Basic Statistics⇒Correlation**. En el cuadro de diálogo resultante hacemos *clic* en el recuadro que hay debajo de **Variables** y seleccionamos, de la lista de variables de la izquierda, las columnas **'Height'** y **'Weight'**; desactivamos **Display p-values** y activamos **Store matrix (display nothing)** y pulsamos en **OK**. ***Minitab*** no muestra el resultado en la ventana de sesión pero guarda, con el nombre **CORR1** (en general, **CORRj**, con $j = 1, 2, \ldots$), la matriz de correlaciones siguiente:

$$\begin{array}{cc} 1,00000 & 0,78487 \\ 0,78487 & 1,00000 \end{array}$$

lo cual quiere decir que el coeficiente de correlación lineal entre las variables **Height** y **Weight** es igual a $0'78487$. Por tanto, la fuerza de la relación lineal entre estas dos variables es moderada. El primer 1 significa que el coeficiente de correlación lineal entre **Height** y **Height** es igual a 1 (lo cual es lógico) y, por supuesto, el segundo 1 significa que el coeficiente de correlación lineal entre **Weight** y **Weight** es igual a 1.

Para aumentar el número de decimales del resultado del coeficiente de correlación lineal entre las variables **Height** y **Weight** hacemos lo siguiente: seleccionamos **Data⇒Copy⇒Matrix to Column**; hacemos *clic* en el recuadro que hay debajo de **Copy from matrix** y seleccionamos (haciendo doble *clic* sobre su nombre) la matriz **CORR1**; en **In current worksheet, in columns** tenemos que teclear las posiciones de dos columnas (**CJ** y **CK** que estén vacías) que contendrán las dos columnas de la matriz de correlaciones. Podemos dejar activada la opción **Name the column containing the copied data**. Por último, pulsamos en **OK**. Ahora ya podemos aumentar el número de decimales como hemos visto en la Práctica 1: hacemos *clic* sobre el nombre de la variable (o sobre su número de columna: CJ); pulsamos con el botón derecho del ratón; seleccionamos **Format Column⇒Numeric**; activamos **Fixed decimal** y en **Decimal places** tecleamos, por ejemplo, 8 y pulsamos en **OK**. Podemos observar que el resultado del coeficiente de correlación lineal entre las variables **Height** y **Weight** es igual a $0'78486641$.

La opción **Stat⇒Basic Statistics⇒Covariance** es similar a lo que acabamos de explicar pero en lugar de determinar el coeficiente de correlación lineal entre cada par de variables calcula lo que ***Minitab*** llama covarianza, pero que en realidad es la cuasi-covarianza (similar a la covarianza, pero dividiendo por $(n-1)$ en vez de por n; siendo n el tamaño muestral). La cuasi-covarianza, S_{xy}, está relacionada con la covarianza, s_{xy}, de la siguiente manera:

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_j - \overline{y})}{n-1} = \frac{n}{n-1}\, s_{xy}.$$

De esto se deduce que el coeficiente de correlación lineal de Pearson se puede calcular de cualquiera de las dos formas siguientes:

$$r_{xy} = \frac{s_{xy}}{s_x\, s_y} = \frac{S_{xy}}{S_x\, S_y}.$$

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

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

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

6. Ejercicios propuestos

2.1.

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

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

c) Determina la distribución de frecuencias de la variable **Intervalos PPU**.

d) Para las variables **Usuarios**, **Préstamos** y **PPU** calcula todas las medidas descriptivas que hemos estudiado en las clases teóricas.

e) Dibuja el diagrama de dispersión, con la recta de regresión superpuesta, de la variable **Préstamos** sobre la variable **Usuarios**. Modifícalo de la siguiente forma:

- Que el título sea *Diagrama de dispersión del 'Nº anual de préstamos' frente al 'Nº anual de usuarios'* en letra Verdana, itálica, negrita, de color rojo y con un tamaño de 9 puntos.
- Que los símbolos sean cuadrados rellenos, de color verde oscuro y de tamaño 2.
- Que en el eje horizontal se muestren 20 marcas (*ticks*) y que los números sean de color azul y con un tamaño de 8 puntos.
- Que el texto del eje horizontal sea *Número anual de usuarios*, en letra Verdana, itálica, no negrita, de color rojo y con un tamaño de 11 puntos.
- Que en el eje vertical se muestren 18 marcas (*ticks*) y que los números sean de color azul y de un tamaño de 8 puntos.
- Que el texto del eje vertical sea *Número anual de préstamos*, en letra Verdana, itálica, no negrita, de color rojo y con un tamaño de 11 puntos.
- Que la recta de regresión sea de color rojo y de tamaño 2.

f) Calcula, con una precisión de 6 decimales, el coeficiente de correlación lineal entre las variables **Préstamos** y **Usuarios**.

g) Determina la ecuación de la recta de regresión de la variable **Préstamos** sobre la variable **Usuarios**.

h) Dibuja el histograma simple de la variable **PPU**.

- Que haya 4 intervalos de la misma amplitud y que en el eje horizontal aparezcan los límites de los intervalos (no los puntos medios).
- Que el título sea *Histograma del 'Porcentaje anual de préstamos por usuario'*, en letra Times New Roman, negrita, de color rojo oscuro y con un tamaño de 14 puntos.
- Que las barras sean de color rojo claro con una trama de relleno horizontal y con los bordes de color rojo oscuro, de tamaño 2.
- Que el texto del eje horizontal sea *Porcentaje anual de préstamos por usuario*, en letra Times New Roman, cursiva, no negrita, de color rojo oscuro y con un tamaño de 12 puntos.
- Que en el eje vertical se muestren 7 marcas (*ticks*) y que los números sean de color rojo oscuro y con un tamaño de 12 puntos.
- Que el texto del eje vertical sea *Frecuencia absoluta*, en letra Times New Roman, cursiva, no negrita, de color rojo oscuro y con un tamaño de 12 puntos.

i) Dibuja el gráfico de sectores de la variable **Intervalos PPU**.

- Que el título sea *Gráfico de sectores de la variable 'Intervalos PPU'*, en letra Verdana, cursiva, negrita, de color azul oscuro y con un tamaño de 12 puntos.

- Que junto a los sectores circulares aparezca la frecuencia absoluta y el porcentaje de cada categoría.
- En la leyenda, tanto la fuente de la cabecera como la fuente del cuerpo sea Verdana, de color azul oscuro y con un tamaño de 10 puntos.

j) Graba el proyecto con el siguiente nombre: **Ejercicio2-1.mpj**

2.2.

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

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

c) Determina la distribución de frecuencias de la variable **Intervalos Porcentaje TRF**.

d) Para las variables **TR**, **TRF** y **Porcentaje TRF** calcula las medidas descriptivas siguientes: mínimo, primer cuartil, mediana, tercer cuartil, máximo, recorrido, recorrido intercuartílico, media, cuasi-varianza, cuasi-desviación típica, suma de los datos y suma de los cuadrados de los datos.

e) Calcula la media, la mediana y la cuasi-desviación típica de la variable **Porcentaje TRF** separando sus resultados según los valores de la variable **Tipo Biblioteca**.

f) Dibuja el diagrama de dispersión, con la recta de regresión superpuesta, de la variable **TRF** sobre la variable **TR**. Modifícalo de la siguiente forma:

- Que el título sea *Nube de puntos y recta de regresión* en letra Verdana, negrita, de color azul y con un tamaño de 12 puntos.
- Que los símbolos sean triángulos rellenos, de color magenta y de tamaño 1.
- Que en el eje horizontal se muestren 10 marcas (*ticks*) y que los números sean de color azul y de un tamaño de 9 puntos.
- Que el texto del eje horizontal sea *Número anual de transacciones de referencia*, en letra Verdana, itálica, no negrita, de color azul y con un tamaño de 10 puntos.
- Que en el eje vertical se muestren 10 marcas (*ticks*) y que los números sean de color azul y de un tamaño de 9 puntos.
- Que el texto del eje vertical sea *Número anual de transacciones de referencia finalizadas*, en letra Verdana, itálica, no negrita, de color azul y con un tamaño de 9 puntos.
- Que la recta de regresión sea de color morado y de tamaño 2.

g) Calcula, con una precisión de 6 decimales, el coeficiente de correlación lineal entre las variables **TR** y **TRF**.

h) Determina la ecuación de la recta de regresión de la variable **TRF** sobre la variable **TR**.

i) Dibuja el diagrama de barras de la variable **Intervalos Porcentaje TRF** en grupos definidos por la variable **Tipo Biblioteca**.

- Que las barras tengan distinto color según los resultados de la variable **Tipo Biblioteca** y que aparezca una leyenda explicativa.
- Que el título sea *Diagrama de barras agrupado*, escrito con letra Arial, negrita, de color rojo oscuro y con un tamaño de 16 puntos.
- Que el texto del eje vertical sea *Frecuencia absoluta*, escrito con letra Arial, negrita, de color rojo oscuro y con un tamaño de 12 puntos.

- Que en el eje horizontal todo esté escrito con la fuente Arial, de color rojo oscuro y con un tamaño de 10 puntos. Que en este eje no aparezca *Tipo biblioteca* ni los resultados de esta variable: *bib. pública* y *bib. universitaria* (puesto que ya está en la leyenda).

j) Graba el proyecto con el siguiente nombre: **Ejercicio2-2.mpj**

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

Gasto en personal	6570
Gasto en libros	3450
Otros gastos	2380

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

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

c) Haz un diagrama de barras y modifícalo a tu gusto.

d) Haz un gráfico de sectores y modifícalo a tu gusto.

e) Graba el proyecto con el siguiente nombre: **Ejercicio2-3.mpj**

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

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

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

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

c) Haz un diagrama de barras agrupado y modifícalo a tu gusto.

d) Graba el proyecto con el siguiente nombre: **Ejercicio2-4.mpj**

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

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

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

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

c) Haz un diagrama de barras en el cual las barras sean segmentos rectilíneos. Modifícalo a tu gusto.

d) Graba el proyecto con el siguiente nombre: **Ejercicio2-5.mpj**