PRÁCTICA 1. INTRODUCCIÓN AL USO DEL STATGRAPHICS

Objeto

El objetivo de esta práctica informática consiste en familiarizarse, mediante software estadístico, con la utilización de algunas de las herramientas de Estadística Descriptiva Unidimensional más utilizadas y que ya han sido introducidas en las clases de teoría (**primera parte** de la Unidad Temática 1 (**UT 1**). En concreto se trabajará con **herramientas** de **tablas de frecuencias** y **gráficas**.

Los **datos** que se usarán en la práctica recogen los valores de varias características de una muestra de 104 libros, adquiridos entre 2004 y 2008 y extraídos al azar, del *Catálogo de libros y audiovisuales de la Universidad Politécnica de Valencia*. Los datos están almacenados en el fichero **P1-Catalogos_UPV-1.sf3** disponible en **PoliformaT**.

NOTA. El documento **"Introducción al uso de Statgraphics"** contiene las instrucciones básicas para iniciar el programa, así como para manejar las ventanas de resultados de los análisis.

1. Obtención de tablas de frecuencias para una variable

Comenzaremos por generar e interpretar tablas de frecuencias para una variable. Tal y como se ha visto en clase de teoría, el procedimiento a seguir es distinto, según si la variable es cualitativa o categórica (o numérica con pocos valores diferentes), o bien se trata de una variable cuantitativa (numérica con presencia de muchos valores observados diferentes).

1.1. Variables cualitativas

En los casos en los que las variables son cualitativas o cuantitativas discretas con pocos valores posibles se opera del modo que se describe a continuación.

Vamos a realizar una tabla de frecuencias de la característica materia de un libro.

Los valores de esta característica están contenidos en la variable ***MATERIA*** del fichero que estamos analizando.

Obsérvese que esta característica es **cualitativa nominal** a diferencia de, por ejemplo, la variable ***AÑO*** que es **cualitativa ordinal**.

Para ello, seleccionamos la opción de menú **Describe > Categorical Data > Tabulation...** (**Figura 1**).

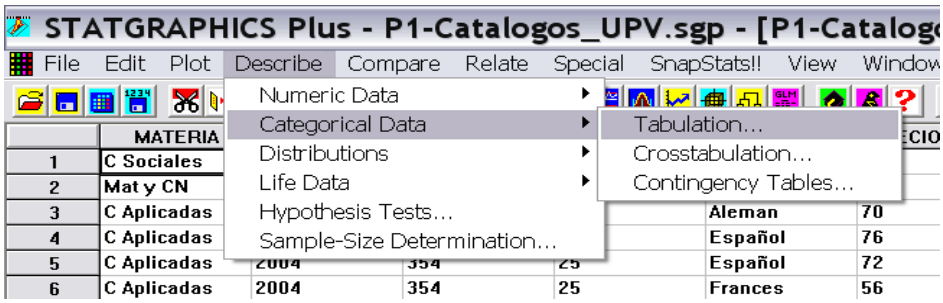


Figura 1. Opción de menú para realizar una tabla de frecuencias de una variable cualitativa.

En el cuadro de diálogo que aparece, indicaremos que queremos realizar la tabla de frecuencias de la variable ***MATERIA***. Para ello, seleccionamos la variable de la lista y la trasladamos al cuadro de texto **"Data"**, simplemente pulsando el botón con el triángulo negro justo al lado de dicho cuadro de texto. También podemos teclear directamente el nombre de la variable. Dejaremos en blanco el cuadro de texto **"Select"** y finalmente, pulsamos en "OK" (**Figura 2**).

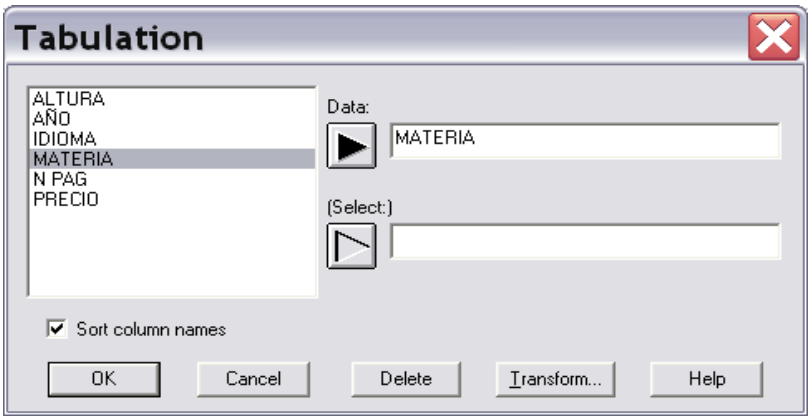


Figura 2. Cuadro de diálogo para la obtención de una tabla de frecuencias sobre una variable cualitativa.

Una vez hecho esto, aparecerá una ventana de análisis, que contendrá un diagrama de barras con la frecuencia de cada valor y la tabla de frecuencias (eventualmente, de acuerdo con las preferencias por defecto del programa, también puede aparecer un diagrama de tarta, u otros resultados). Para visualizar la tabla de frecuencias, si no aparece en alguno de los paneles, debemos hacer clic sobre el botón **Tabular options** de la ventana de análisis (**Figura 3**), y marcar **"Frequency Table"** (**Figura 4**) en la lista de opciones.

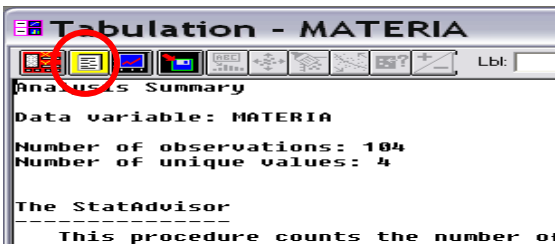


Figura 3. Situación del botón **Tabular options** en la barra de herramientas de la ventana de análisis.

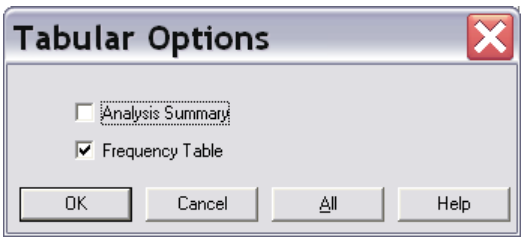


Figura 4. Selección de la opción "Frequency Table" para mostrar la tabla de frecuencias.

A partir de la tabla obtenida, contesta la siguiente cuestión.

Pregunta 1. ¿Cuántas materias diferentes contiene la muestra? ¿Qué porcentaje de los libros de la muestra pertenecen a la materia de Lengua y Literatura? ¿Cuántos libros corresponden a las materias de Ciencias Aplicadas, Medicina, Tecnología, Matemáticas y Ciencias Naturales?

1.2. Variables cuantitativas

Para el caso de variables cuantitativas (numéricas continuas o discretas con un número elevado de valores posibles), utilizaremos otro menú diferente, con el fin de obtener una tabla de frecuencias que agrupe los valores en intervalos.

Realizaremos una tabla de frecuencias de la variable ***PRECIO***, que corresponde al precio de venta, en euros, de los libros. La opción de menú que debemos seleccionar es **Describe > Numeric Data > One Variable Analysis...** (**Figura 5**).

Seleccionamos la variable ***PRECIO*** para el análisis, de manera idéntica a como hemos hecho en el apartado anterior, y pulsamos "OK" (**Figura 6**).

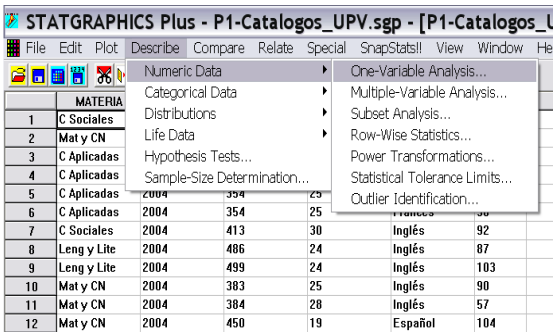


Figura 5. Opción de menú para realizar un análisis sobre una sola variable numérica o cuantitativa.

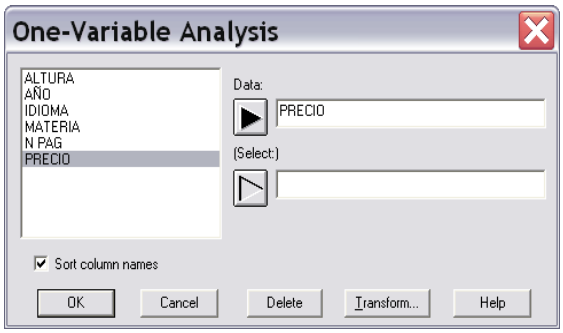


Figura 6. Cuadro de diálogo para seleccionar un análisis sobre una variable numérica o cuantitativa.

De nuevo, obtendremos la ventana de resultados del análisis. Para visualizar la tabla de frecuencias, entramos en **Tabular options** (**Figura 3**), y marcamos la opción **"Frequency Tabulation"** (**Figura 7**).

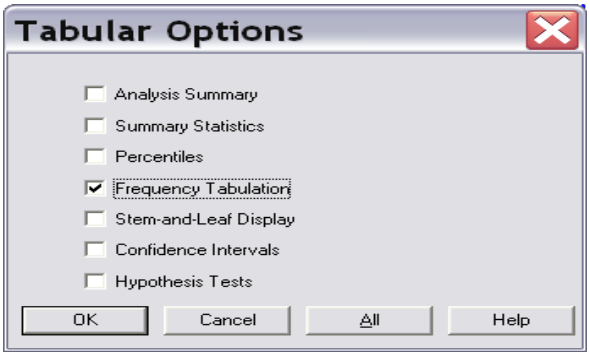


Figura 7. Selección de la opción para mostrar la tabla de frecuencias, dentro del análisis de una variable numérica.

Si pulsamos en el botón **Graphical options** (**Figura 8**) de la ventana de análisis de la variable, podemos acceder a la lista de gráficos disponibles asociados a dicho análisis. Marcando **"Frequency Histogram"** (**Figura 9**), obtendremos el histograma de la variable ***PRECIO***. Este es un procedimiento diferente y más general que el que se explicará en la sección 3 para acceder a todas las representaciones gráficas posibles asociadas al análisis de una variable.

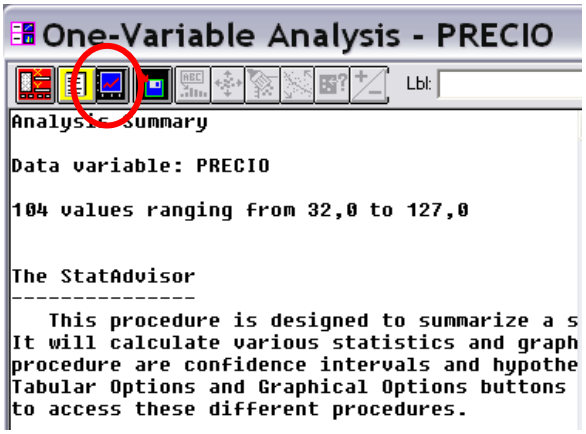


Figura 8. Ubicación del botón **Graphical options** en la barra de herramientas de la ventana de análisis.

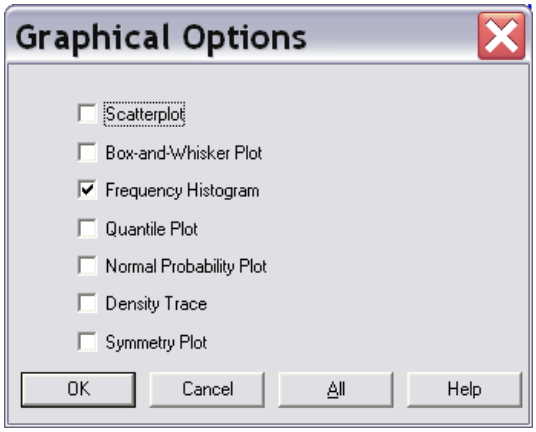


Figura 9. Selección de la opción para mostrar el histograma, dentro de las opciones gráficas para el análisis de una variable cuantitativa.

Si nos fijamos en la ventana del análisis de la variable ***PRECIO*** en este momento, podemos observar al mismo tiempo la tabla de frecuencias y el histograma correspondientes a dicha variable (**Figura 10**). Recordemos que el histograma es una de las posibles representaciones gráficas de una tabla de frecuencias.

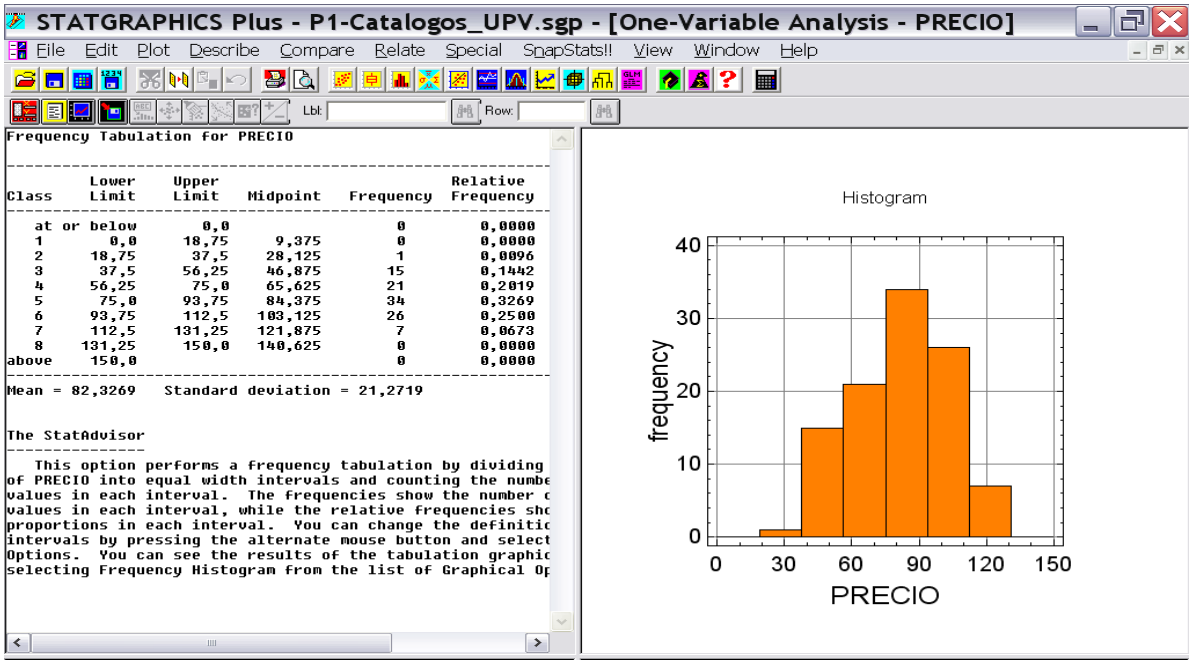


Figura 10. Análisis de una variable cuantitativa con muchos valores, con los paneles correspondientes a la tabla de frecuencias y el histograma de dicha variable.

Como sabemos por lo visto en las clases de teoría, en una tabla de frecuencias de una variable cuantitativa continua o discreta con muchos valores, es importante decidir el **número de intervalos** o clases en que dividiremos el rango de valores observados.

Statgraphics permite cambiar el número de intervalos que el programa ha decidido por defecto. Para ello, hacemos clic con el botón derecho sobre el panel de la tabla de frecuencias, y seleccionamos las opciones del panel (**"Pane Options..."**).

No sólo vamos a modificar el número de intervalos o clases, sino también los límites inferior y superior del rango de valores representado. Concretamente, rellenaremos el cuadro de diálogo de la siguiente manera: **"Number of Classes"** (número de intervalos): **10**; **"Lower Limit"** (límite inferior): 25; y **"Upper Limit"** (límite superior): 135 (**Figura 11**). Terminamos, como siempre, pulsando "OK".

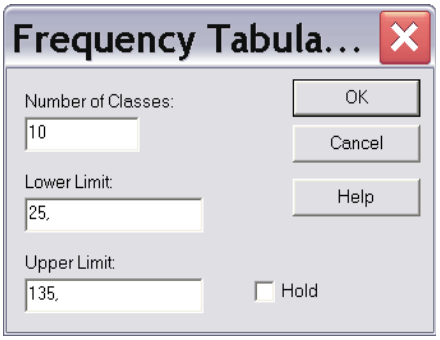


Figura 11. Cuadro de diálogo en el que pueden modificarse las opciones aplicadas por defecto para construir la tabla de frecuencias y el histograma.

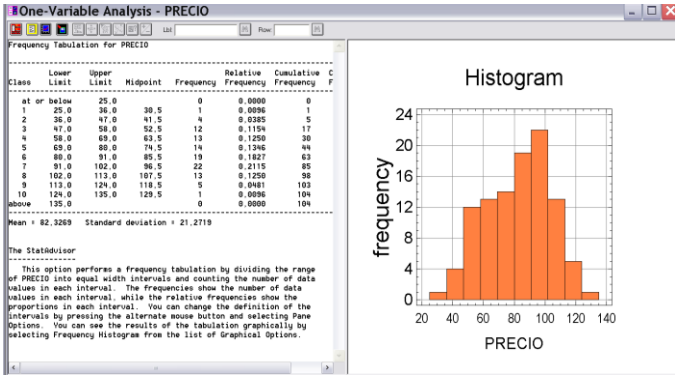


Figura 12. Efecto de realizar simultáneamente las modificaciones sobre la tabla de frecuencias y el histograma.

Puede observarse (**Figura 12**) cómo cualquier cambio realizado sobre la tabla de frecuencias se refleja de manera automática en el histograma de la variable.

A partir de la tabla de frecuencias, y aplicando lo que sobre ella has aprendido en clase de teoría, contesta la siguiente cuestión.

Pregunta 2. ¿Por debajo de qué precio se encuentra aproximadamente el 90% de los libros de la muestra del Catálogo?

2. Obtención de un diagrama de sectores

A continuación, vamos a realizar un estudio descriptivo sobre el año de adquisición de los libros. Con este fin, construiremos un diagrama de sectores o de tarta sobre la variable ***AÑO***.

Para ello, volvemos a seleccionar la opción de menú **Describe > Categorical Data > Tabulation...** (**Figura 1**).

En el cuadro de diálogo (**Figura 2**) seleccionamos ahora la variable ***AÑO*** y pulsamos "OK".

Posteriormente, para mostrar el diagrama de sectores, pulsamos el botón de opciones gráficas, **Graphical options** (**Figura 8**), y marcamos **"Piechart"** (literalmente, gráfico de tarta).

> **¡ATENCIÓN!** Existe también la opción de menú **Plot > Business Charts > Piechart...**, pero ésta requiere que los datos estén ya agrupados, en lugar de representar observaciones individuales, como sucede en nuestro caso.

Es posible modificar el aspecto del diagrama de sectores pulsando sobre él con el botón derecho del ratón y seleccionando **"Pane Options..."**. En concreto, podemos visualizar tanto las frecuencias absolutas (**"Counts"**) como las relativas (**"Percent"**).

Esta vez, puede ser interesante (incluso necesario) entrar también en **"Graphics Options..."** (de nuevo, pulsando con el botón derecho sobre el diagrama de sectores).

> Como ***AÑO*** es una variable cuantitativa ordinal los valores son tratados como números. Sin embargo, si estuviéramos analizando la variable ***MATERIA***, los valores podrían haberse almacenado como caracteres (nombres) o codificados con números.
>
> Si, en nuestro archivo, la variable ***MATERIA*** estuviera codificada de manera numérica, entrando en la pestaña "Legend" de **"Graphics Options..."** podríamos asignar las etiquetas (nombres) correspondientes a cada valor de la variable, como por ejemplo:
>
> **1:** *C_Aplicadas* **2:** *C_Sociales* **3:** *Leng_ y_Lite* **4:** *Mat_ y_CN*

A partir de la información proporcionada por el diagrama de sectores, contesta la siguiente cuestión.

Pregunta 3. Según la muestra, ¿En qué año se adquirieron menos libros? Da el resultado en frecuencias absolutas y en frecuencias relativas.

3. Obtención de un histograma

En este apartado vamos a volver a estudiar el precio de los libros. Lo haremos a través de una representación gráfica de los valores observados para la variable ***PRECIO***; concretamente, y dado que se trata de una variable continua, utilizaremos un histograma.

Como sabemos por las clases de teoría, un histograma se construye agrupando los datos observados en intervalos, y dibujando una barra por cada intervalo, siendo la altura de cada barra proporcional a la cantidad de valores observados en dicho intervalo (es decir, a la frecuencia absoluta).

Para obtener en Statgraphics un histograma de una variable, podemos proceder de diversas formas. Entre ellas, destacamos aquí dos:

- A través de la opción de menú **Plot > Exploratory Plots > Frequency Histogram...** (**Figura 13**).
- Directamente pulsando sobre el botón del histograma en el barra principal de herramientas (**Figura 14**).

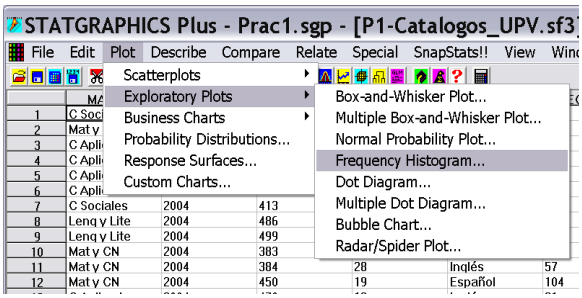


Figura 13. Selección del histograma a través del menú desplegable.

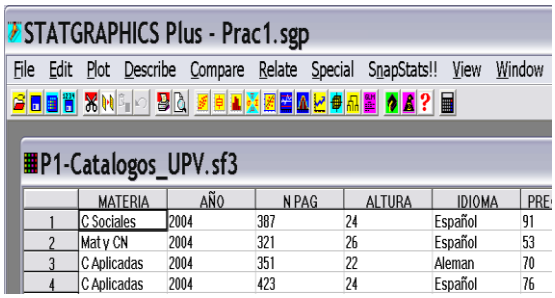


Figura 14. Selección del histograma a través de la barra principal de herramientas.

En el cuadro de diálogo que aparece, seleccionaremos como otras veces la variable que queremos analizar (***PRECIO***, en este caso), y pulsaremos "OK" para generar el histograma.

Pulsando con el botón derecho sobre el histograma, y seleccionando **"Pane Options..."**, llegamos a la ventana donde podemos modificar el aspecto del gráfico. Repetiremos las operaciones relativas al número de intervalos y el rango de valores ya realizadas en la sección 1.2 (**Figura 11**). Además, cabe destacar que en esta ventana podemos elegir entre visualizar las frecuencias absolutas o las relativas (desmarcando o marcando, respectivamente, la opción **"Relative"**). La forma del histograma no cambiará; sólo la escala en que son medidas las barras (y, por tanto, la interpretación de la "altura" de éstas).

NOTA. Con la opción **"Locate"** (disponible al pulsar con el botón derecho sobre el histograma, previamente maximizado con un doble clic), que explicaremos más adelante, es posible determinar exactamente la "altura" de cualquier barra del histograma.

A partir de la información que proporciona el gráfico, contesta la siguiente cuestión.

Pregunta 4. ¿Cuál es el intervalo de precios más frecuente? ¿Qué porcentaje de libros tiene un precio superior a 80 €?

Volviendo a las opciones del panel (**"Pane Options..."**) del histograma, podemos transformar éste en un diagrama de frecuencias acumuladas, si seleccionamos simultáneamente las opciones **"Cumulative"** y **"Polygon"** (**Figura 15**).

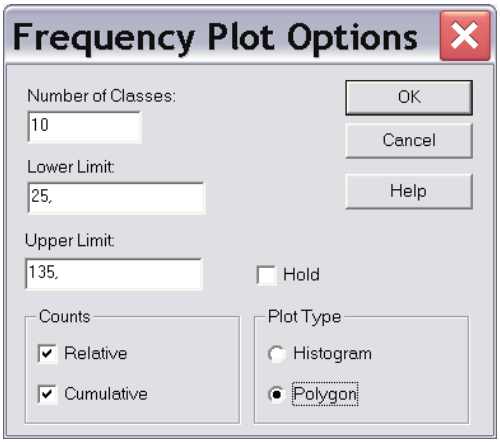


Figura 15. Cuadro de diálogo en el que pueden modificarse las opciones aplicadas por defecto para construir el histograma. Además, seleccionando "Cumulative" y "Polygon" se obtiene un diagrama de frecuencias acumuladas de la variable.

El gráfico resultante (**Figura 16**) permite observar de manera aproximada la evolución de las frecuencias acumuladas. Para generarlo, se dibuja un punto en cada intervalo cuya altura representa la frecuencia (absoluta o relativa) acumulada en dicho intervalo, y posteriormente se unen todos los puntos mediante líneas rectas, siendo el resultado una curva rectilínea y no decreciente, como se muestra a continuación:

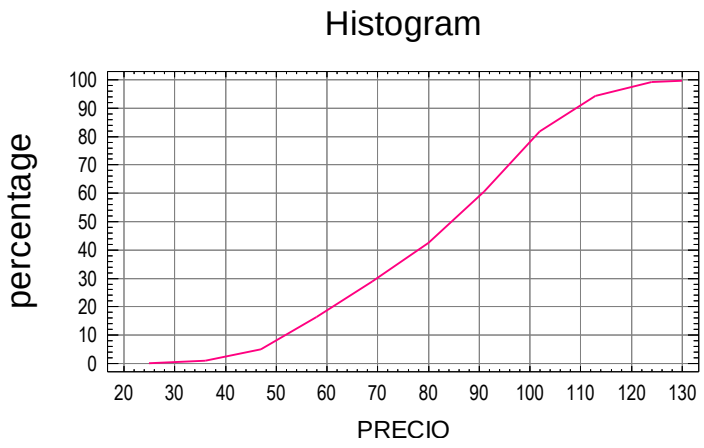


Figura 16. Diagrama de frecuencias acumuladas para la variable *PRECIO*.

NOTA. El diagrama de frecuencias únicamente permite obtener información *aproximada*, ya que la forma de la línea poligonal que en él se muestra varía, según número de intervalos en que se haya dividido el rango de valores.

Para poder realizar una lectura más precisa de la información que proporciona el diagrama de frecuencias acumuladas, el programa facilita las opciones **"Locate"** y **"Zoom"**, que podemos activar pulsando con el botón derecho sobre el histograma.

Estas opciones sólo están habilitadas si el área del gráfico se ha maximizado previamente, haciendo doble clic sobre ella.

Cuando está activada, la opción **"Locate"** permite determinar con exactitud las coordenadas de cualquier punto del gráfico sobre el que hagamos un clic (**Figura 17**).

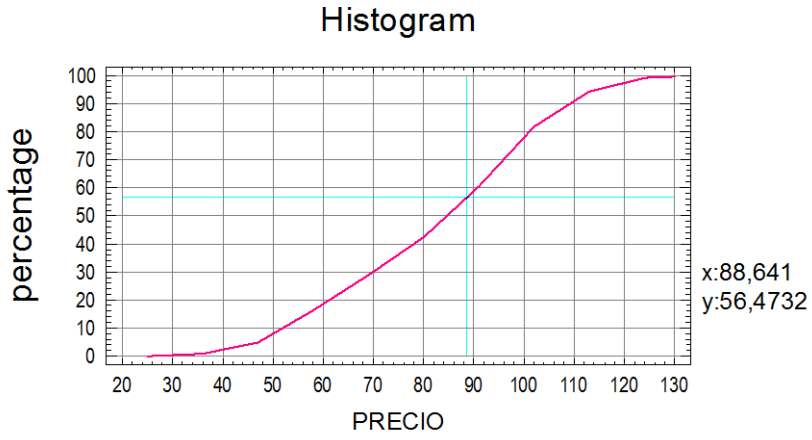


Figura 17. Ejemplo de uso de la función "Locate" sobre un diagrama de frecuencias acumuladas.

RECUERDA: En cualquiera de los análisis que llevemos a cabo, podemos elegir que éste se realice sólo sobre los individuos de la muestra que cumplan una determinada condición, rellenando convenientemente el cuadro de texto **"Select"**.

Por ejemplo, si deseamos realizar un análisis sobre el precio de los libros de la materia Lengua y Literatura únicamente, lo indicaremos añadiendo en el cuadro **"Select"** la condición ***MATERIA* = "Leng_y_Lite"**, o bien ***MATERIA* = 3**, según la manera en que se encuentre codificada dicha variable en nuestro archivo (**Figura 18**).

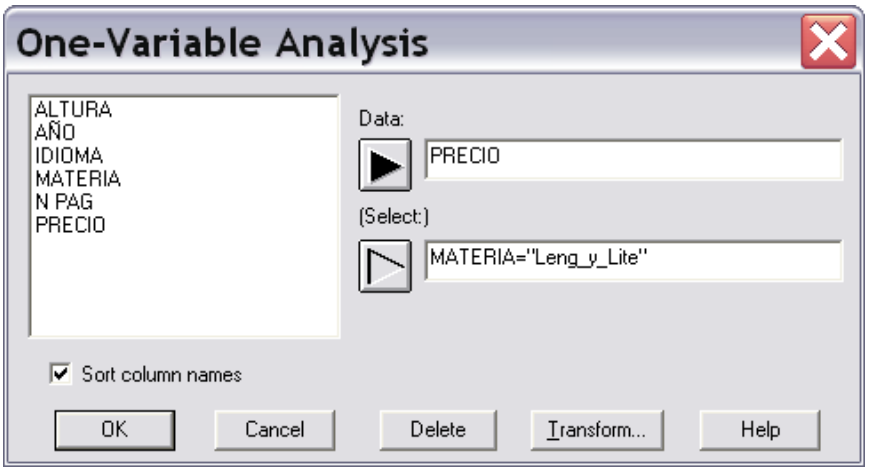


Figura 18. Ejemplo de uso del cuadro de texto "Select" para seleccionar los casos que se desea estudiar.

A partir del diagrama de frecuencias acumuladas de la variable ***PRECIO***, responde las siguientes cuestiones.

Pregunta 5. Aproximadamente, ¿qué porcentaje de los libros tienen un precio inferior a 85 €?

Pregunta 6. ¿Cuál es el precio por encima de cual se encuentra aproximadamente el 60% de los libros?

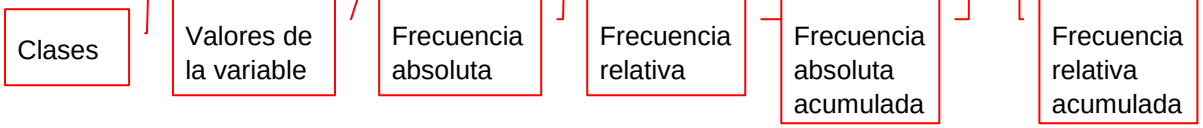
Pregunta 7. Aproximadamente, ¿cuántos libros de la materia Ciencias Aplicadas, Medicina, Tecnología valen menos de 85 €?

Respuestas a las preguntas propuestas

Pregunta 1

Frequency Table for MATERIA

Class	Value	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	C_Aplicadas	42	0,4038	42	0,4038
2	C_Sociales	20	0,1923	62	0,5962
3	Leng_y_Lite	14	0,1346	76	0,7308
4	Mat_y_CN	28	0,2692	104	1,0000



La muestra contiene 4 materias diferentes, que son los 4 valores distintos o **clases** que toma la variable aleatoria ***MATERIA***:

C_Aplicadas → Ciencias Aplicadas, Medicina, Tecnología

C_Sociales → Ciencias Sociales

Leng_y_Lite → Lengua y Literatura

Mat_y_ CN → Matemáticas y Ciencias Naturales

El porcentaje de los libros de la muestra que pertenecen a la materia de Lengua y Literatura nos la da la frecuencia relativa en *tanto por cien*. En este caso, estos libros representan un 13,46%.

$f_{Leng\_y\_Lite}$ = 0,1346 = $n_{Leng\_y\_Lite}$ / N = 14/104 = 0,1346 → 13,46%

El número de libros que corresponden a las materias de Ciencias Aplicadas, Medicina, Tecnología, Matemáticas y Ciencias Naturales se obtiene sumando el número de éstos en los que la variable ***MATERIA*** toma el valor *C Aplicadas* y en los que toma el valor *Mat y CN*, esto es, sumando las frecuencias absolutas de las dos clases.

$n_{C\_Aplicadas}$ + $n_{Mat\_y\_CN}$ = (suma de las frecuencias absolutas) = 42 + 28 = 70

Pregunta 2

Frequency Tabulation for PRECIO

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below		25,0		0	0,0000	0	0,0000
1	25,0	36,0	30,5	1	0,0096	1	0,0096
2	36,0	47,0	41,5	4	0,0385	5	0,0481
3	47,0	58,0	52,5	12	0,1154	17	0,1635
4	58,0	69,0	63,5	13	0,1250	30	0,2885
5	69,0	80,0	74,5	14	0,1346	44	0,4231
6	80,0	91,0	85,5	19	0,1827	63	0,6058
7	**] 91,0**	**102,0]**	96,5	22	0,2115	85	**0,8173**
8	**]102,0**	**113,0]**	**107,5**	13	0,1250	98	**0,9423**
9	113,0	124,0	118,5	5	0,0481	103	0,9904
10	124,0	135,0	129,5	1	0,0096	104	1,0000
above	135,0			0	0,0000	104	1,0000

Límite inferior

Límite superior

Punto medio o marca de clase

Buscamos un valor aproximado de ***PRECIO*** por debajo del cual se encuentra el 90% de los libros de la muestra. Esto es, el 90% de los libros de la muestra tiene un precio inferior o igual a dicho valor.

La frecuencia relativa acumulada de un intervalo dado nos indica qué proporción (o porcentaje si se toma en *tanto por cien*) de los individuos de la muestra presentan un precio contenido en dicho intervalo o en uno de los intervalos anteriores.

Buscaremos, por tanto, aquel intervalo cuya frecuencia relativa acumulada esté alrededor de 90%.

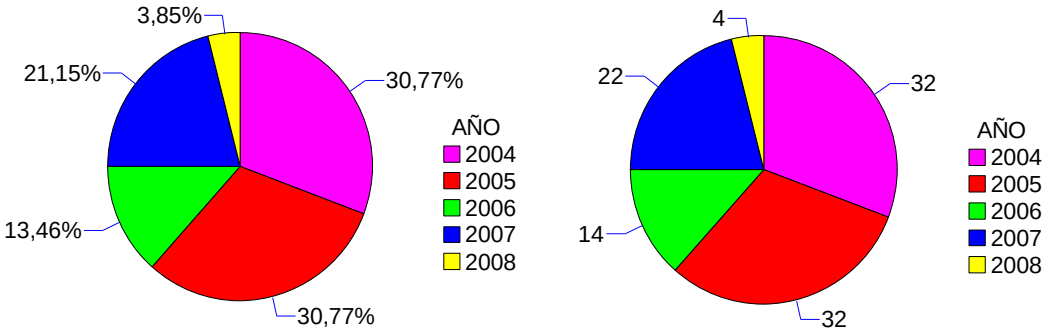
Observamos que las frecuencias relativas acumuladas de los intervalos]91 , 102] y]102 , 113] son 0,8173 y 0,9423, respectivamente.

Esto significa que hay un 81,73% de los libros con una precio inferior o igual a 102 €, mientras que por debajo de 113 € encontramos el 94,23% de los 104 libros de la muestra.

Por tanto, en algún punto entre 102 € y 113 € se encuentra el valor del ***PRECIO*** por debajo de la cual se encuentra el 90% de los libros.

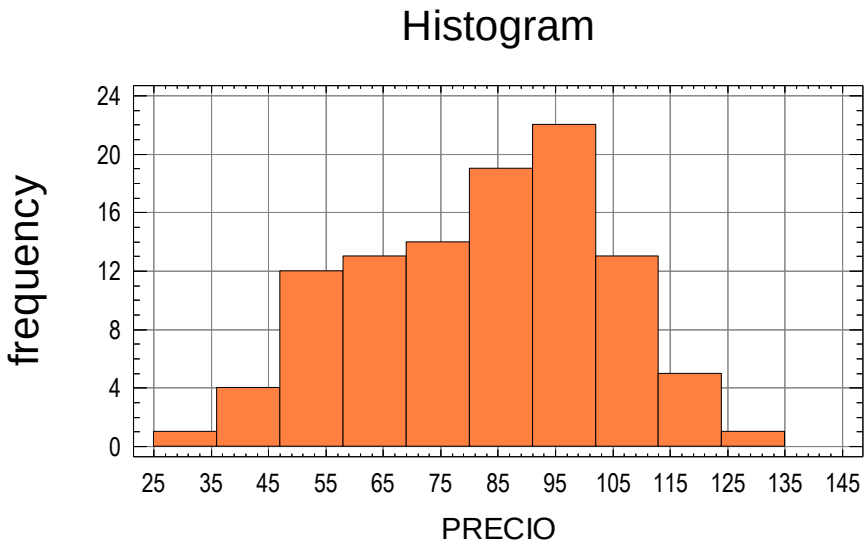
Si se quiere, puede tomarse, como aproximación, el punto medio del intervalo (107,5 €). [Es posible realizar aproximaciones más exactas (aplicando interpolación, etc.).]

Pregunta 3



A la vista de los gráficos, alternando entre las vistas de frecuencias absolutas y relativas, el año en el que se han adquirido menos libros es 2008. En el año 2008 se han adquirido 4 libros, lo cual supone un 3,85% del total de los libros muestreados.

Pregunta 4



A partir del histograma, una vez configurado con 10 intervalos, y variando desde 25 € a 135 €, observamos que la barra más alta corresponde al intervalo]91 , 102], que, por tanto, es el intervalo más frecuente.

Con la ayuda de la función **"Locate"**, o bien acudiendo a la tabla de frecuencias para la variable ***PRECIO***, vemos que el número de observaciones contenidas en dicho intervalo (es decir, su frecuencia absoluta) es 22.

En resumen, el grupo de individuos de la muestra cuyo precio está entre 91 y 102 € es el más numeroso, y está compuesto por 22 libros.

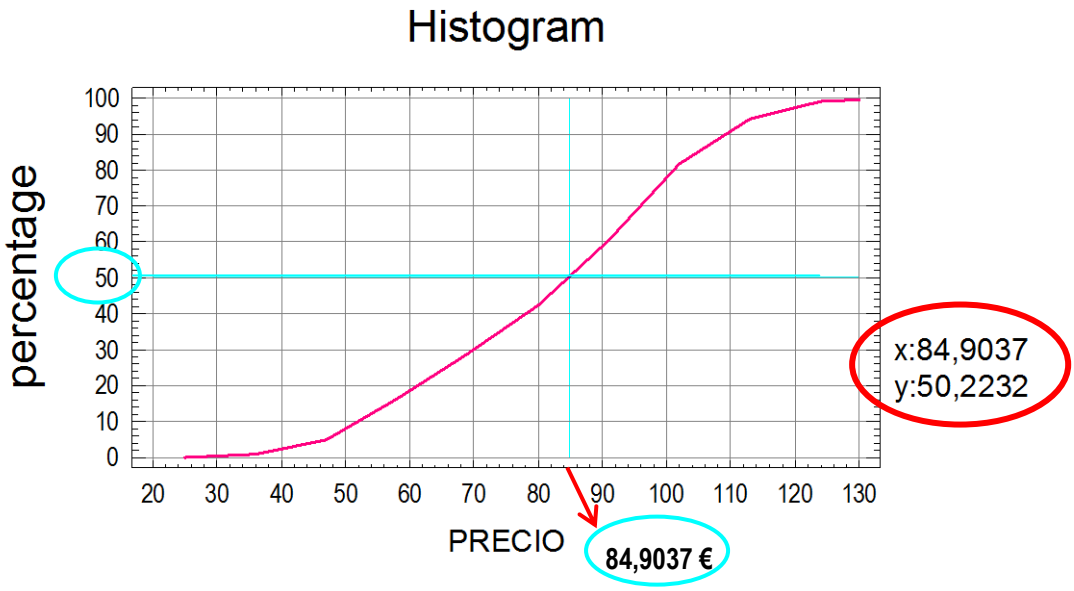
Para calcular, a partir del histograma, qué porcentaje de libros tienen un precio superior a 80 €, lo configuramos (con **"Pane Options..."**) para que nos muestre las frecuencias relativas, y sumamos las correspondientes a los intervalos]80 , 91] ,]91 , 102] ,]102 , 113] ,]113 , 124] y]124 , 135]

El resultado es (con la ayuda de la tabla de frecuencias) 0,1827 + 0,2115 + 0,1250 + 0,0481 + 0,0096 = 0,5967 → 59,67%.

Alternativamente, también podíamos sumar las frecuencias absolutas de los intervalos, y después calcular la frecuencia relativa, dividiendo por el número total de casos (tamaño muestral):

Así, los cálculos serían (19 + 22 + 13 + 5 + 1) / 104 = 60 / 104 = 59,67%

Pregunta 5



NOTA. Las respuestas a las preguntas 5 y 6 presuponen que hemos configurado el diagrama de frecuencias acumuladas con 10 intervalos, comenzando desde 25 € y llegando hasta 135 €.

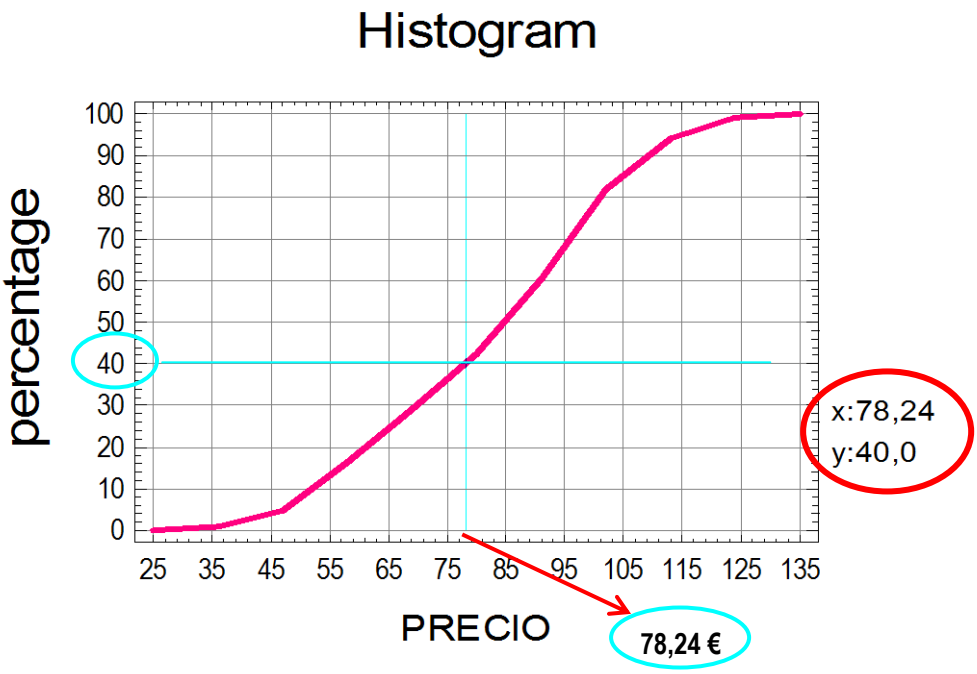
Para saber, a partir del diagrama, qué porcentaje de los libros tienen un precio inferior a 85 €, utilizamos la opción **"Locate"** (combinada con **"Zoom"**, si es necesario), y localizamos las coordenadas del punto sobre la línea del gráfico correspondiente a dicha altura. Como se observa en el gráfico de arriba, estas coordenadas son (84,9037 , 50,2232) → aproximadamente (85 , 50)

Es decir, aproximadamente el 50% de los libros tienen un precio por debajo de 85 €.

Pregunta 6

El precio por encima del cual se encuentra el 60% de los libros es aquel con una frecuencia relativa acumulada de 100 – 60 = 40%, esto es, el precio que deja por debajo de el 40% de la muestra.

Del mismo modo que antes, buscamos en el gráfico el punto sobre la línea que corresponda a un porcentaje de 40%. Las coordenadas de dicho punto son (170,3 , 40,0).



Por tanto, el precio por arriba del cual se sitúa el 60% de los libros es 78,24 €, aproximadamente.

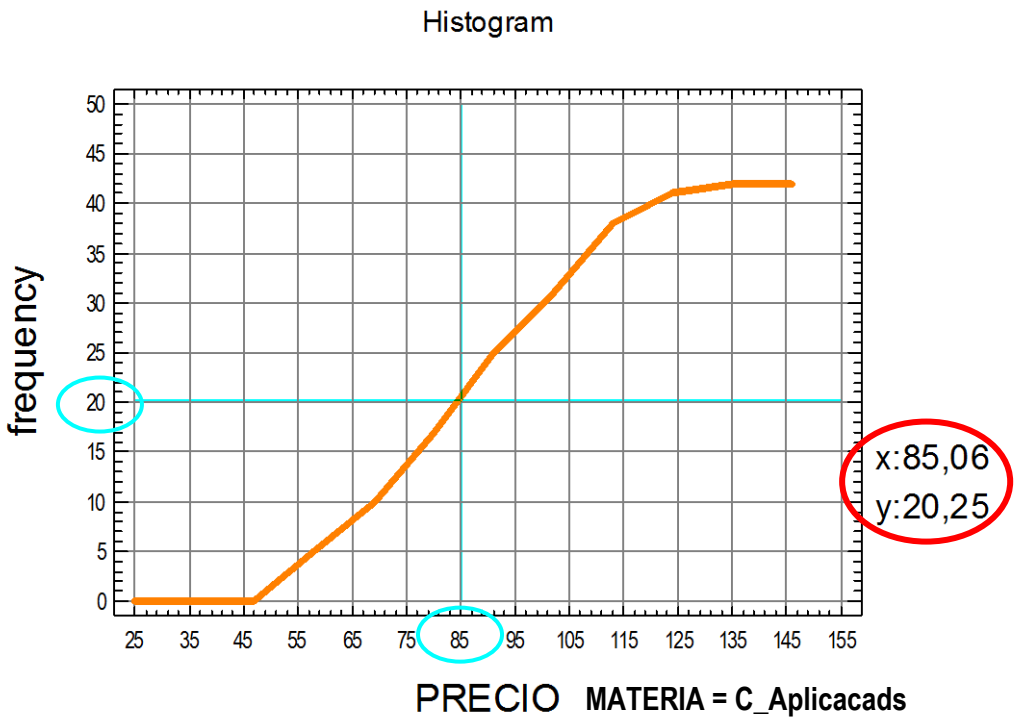
Pregunta 7

Ahora realizamos el estudio únicamente sobre los el precio de los libros que corresponden a la materia de Ciencias Aplicadas, Medicina, Tecnología (*C_Aplicadas*), seleccionándolos tal como se explica en el enunciado de la práctica.

Nos preguntan por la cantidad (frecuencia absoluta) de libros de la materia *C_Aplicadas* en la muestra cuyo precio es menor de 85 €. Para ello, configuraremos el gráfico con **"Pane Options..."** para que muestre las frecuencias absolutas (desmarcamos **"Relative"**).

NOTA. Mantenemos el resto de configuraciones del diagrama por defecto; no modificamos ni el número de intervalos, ni los límites del rango visualizado.

Buscamos con **"Locate"** el punto sobre la línea que corresponde al precio de 85 €; sus coordenadas son (85,06 , 20,25).



Por tanto, podemos decir que, aproximadamente, 20 libros de la materia Ciencias Aplicadas, Medicina, Tecnología en la muestra extraída del Catálogo de libros y audiovisuales de la UPV poseen un precio inferior a 85 €.

En porcentaje, suponen 20/42 = 47,62% sobre el total de libros de Ciencias Aplicadas, Medicina, Tecnología, que es 42.

Fuentes

Material docente previo de E. Vázquez (DEIOAC - UPV) | Catálogo de Libros y Audiovisuales de la UPV | Material docente de V. Giner (DEIOAC - UPV) | Material docente de A. Calduch (DEIOAC - UPV) | http://www.statgraphics.net/ |