Práctica 2 de Estadística

En este tema nos vamos a ocupar de exponer los métodos mas usuales que en *Estadística Descriptiva* se utilizan para estudiar un conjunto de datos relativos a un determinado fenómeno de interés, de manera que pueda ser extraída la máxima información procedente de estos datos. Esta información será utilizada posteriormente no sólo para obtener conclusiones respecto del colectivo constituido por esos datos, sino también para establecer conclusiones respecto del fenómeno del cual se han tomado los datos. De esta labor se ocupa la *Inferencia Estadística.*

Los fenómenos de interés son de muy diversa procedencia. De hecho, el análisis estadístico se puede aplicar a la mayoría de las disciplinas. En ocasiones cuentan con miles de datos, como es el caso de las encuestas sobre la intención de voto en las elecciones generales, cuyos resultados, si el plan de elección de los encuestados es estadísticamente adecuado, dan resultados muy fiables. En otros casos el número de datos estudiados coincide con la población total, como es el caso de la elaboración de los censos de población; sin embargo, cualquier revisión de la calidad de un censo se hace a partir de una **muestra** o conjunto de datos representativa de la población.

En cualquier caso, el estudio estadístico se realiza sobre un conjunto de elementos pertenecientes generalmente a un colectivo mayor que se intenta describir lo más fiablemente posible sin necesidad de estudiar en su totalidad.

2.1 Población y muestra: tamaño muestral y tamaño poblacional.

Se entiende por **población** el conjunto de todos los elementos sobre los que se van a obtener datos para realizar el estudio estadístico. Se entiende por **muestra** un subconjunto finito de elementos tomados de una población. Llamamos **tamaño poblacional** al número de elementos de la población, y **tamaño muestral** al número de elementos de la muestra.

Podemos por su tamaño distinguir dos tipos de poblaciones, las *poblaciones finitas* y las *poblaciones indefinidamente grandes,* cuyo tamaño podemos considerar si no infinito, si suficientemente grande para poder suponerlo. Las muestras se toman porque el número de elementos de la población es demasiado grande para poder ser estudiados en su totalidad. Para que los datos de la muestra obtenida sirvan para caracterizar a la población de donde proviene, es necesario que la muestra sea *representativa,* en el sentido de que las conclusiones obtenidas sirvan para el total de la población. Esta representatividad está relacionada tanto con el tamaño muestral en comparación con el tamaño poblacional, como con la forma de seleccionar los elementos de la muestra, tareas que forman parte del *plan de muestreo* previo al estudio estadístico y que son importantísimas, pues de ello depende la calidad de los resultados posteriores.

2.2 Variables y tipo de variables.

Las características que poseen los elementos de una población y que van a ser objeto de estudio estadístico reciben el nombre de *variables.* Así, por ejemplo, la muestra [1, 2, 1, 0, 3, 3, 1, 1, 0, 2] corresponde a 10 valores de la variable X definida como "número de hijos en una muestra de 10 familias" Esta variable solamente puede tomar los valores 0, 1, 2, 3, 4, ... Los valores de una muestra se llaman *observaciones* (o *casos*) de la variable.
La variable X definida como "diámetro de agujas para jeringuillas" puede tomar cualquier valor de un intervalo, como por ejemplo, [0.45, 0.55]. Los valores de la variable asociada a la encuesta realizada sobre el partido político votado en las últimas elecciones, no son numéricos. Vemos pues que las variables pueden ser de distintos tipos según los valores que toman. Vamos a dar una clasificación de los distintos tipos de variables en dos grandes grupos, las variables *categóricas* y las variables *medibles.*

2.2.1 Variables categóricas.

Son las asociadas a características que no son cuantificables, aunque cada distinto tipo de resultado, que llamaremos *categoría,* puede ser clasificado y asociado a un número o letra. Podemos distinguir las variables categóricas en dos tipos:

- **Variables categóricas nominales.** Son aquéllas en las cuales las asignaciones no suponen ningún orden en los posibles resultados del experimento. Por ejemplo "la profesión del cabeza de familia" en una encuesta realizada a un grupo de familias. Aquellas variables que permiten clasificar a los elementos en dos clases diferenciadas, las llamaremos *variables dicotómicas.* Un ejemplo claro, es la clasificación de un grupo de personas, según el sexo, se le puede asignar a esta variable los valores V y H ó 0 y 1.

- **Variables categóricas ordinales.** Son aquéllas en las que las asignaciones corresponden a un orden de preferencias. Un ejemplo de variable categórica ordinal es la clasificación de los resultados de un examen en: Insuficiente, Suficiente, Notable, Sobresaliente.

2.2.2 Variables medibles.

Las disciplinas científicas poseen instrumentos de medida que permiten cualificar los resultados experimentales. Esto supone al mismo tiempo la adopción de unidades de medida para valorar los distintos resultados. Es el caso del sistema métrico decimal, los sistemas monetarios, las escalas de temperaturas, el conjunto de los números naturales, etc.
Podemos clasificar las variables medibles en dos grandes grupos:

- **Variables medibles discretas.** Son aquéllas cuyos posibles valores constituyen un conjunto de cardinal finito o a lo sumo numerable. La variable que mide el número de hijos en una encuesta realizada en una población, constituyen un conjunto de tipo finito y la variable que mide el número de lanzamientos de una moneda hasta obtener una cara es un conjunto infinito numerable.

- **Variables medibles continuas.** Son aquéllas que pueden tomar cualquier valor de un determinado intervalo, es decir, entre dos valores posibles, pueden tomar también cualquiera que haya entre ellos. Es el caso, por ejemplo, de la duración de una llamada telefónica, el peso de cualquier producto, la estatura de una persona, etc.
  Por extensión del concepto de variable continua, una magnitud que pueda tomar un gran número de valores y muy próximos, aunque sean aislados, será considerada como continua.

2.3 Distribuciones de frecuencias y representaciones gráficas.

2.3.1 Distribución de frecuencias de una variable categórica.

Supongamos que tomamos una muestra de tamaño n de una población, y observamos la variable categórica X. Sean x_1, x_2, ..., x_m los valores que asignamos a cada categoría de la variable.
Definimos como *frecuencia absoluta* de la categoría x_i, $i = 1, 2, \ldots, m$ y la representamos como n_i al número de veces que aparece x_i en la muestra tomada. Evidentemente se cumple:

$$\sum_{i=1}^{m} n_i = n$$

Definimos como *frecuencia relativa* de la categoría x_i, $i = 1, 2, ..., m$ y la representamos como f_i a la proporción sobre el total de la muestra que representa la frecuencia absoluta de x_i es decir:

$$f_i = \frac{n_i}{n}$$

Se cumple:

$$\sum_{i=1}^{m} f_i = 1$$

Las frecuencias relativas pueden expresarse en porcentaje sin más que multiplicar por cien. La relación entre las distintas categorías y sus frecuencias recibe el nombre de *distribución de frecuencias* de la variable categórica X, y los valores obtenidos pueden ser dispuestos en una tabla que se llama *tabla de frecuencias,* y que tiene la forma general:

x_i	n_i	f_i
x_1	n_1	f_1
x_2	n_2	f_2
…	...	...
x_m	n_m	f_m
	n	1

2.4 Representaciones gráficas de variables categóricas.

Los valores expresados en la tabla de frecuencias pueden ser representados en dos ejes para obtener una visión gráfica de la forma de la distribución de los valores. Se indican las distintas categorías en el eje *OX*, y las frecuencias absolutas (o relativas) en el eje *OY*. Esta representación se llama **diagrama de barras**. Otra forma de representar las variables categóricas es el **diagrama de sectores**; este diagrama está formado por sectores circulares cuyo ángulo central es proporcional a las frecuencias absolutas o relativas correspondientes.

2.5. Manejo de SPSS

2.5.1 Análisis de frecuencias

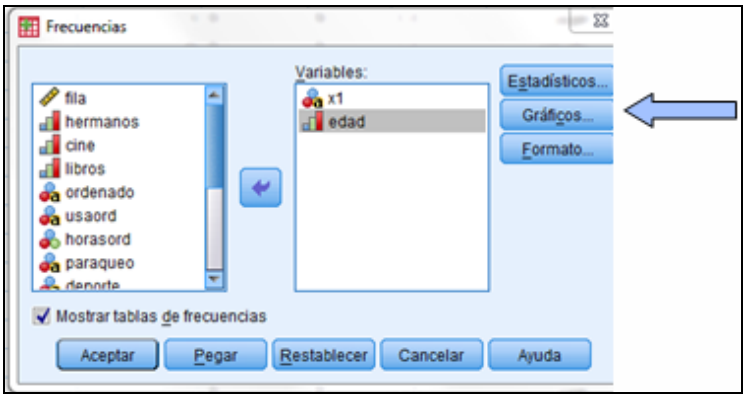
Las operaciones de análisis de variables se agrupan en SPSS en el menú ‘Analizar’. Dentro del mismo encontraremos la opción ‘Estadísticos Descriptivos/Frecuencias’. Para realizar el análisis únicamente hemos de escoger la variable sobre la que realizar el análisis y marcar la opción ‘Mostrar tabla de frecuencias’.

El resultado del análisis se muestra mediante una aplicación independiente, el *Visor de Resultados*.

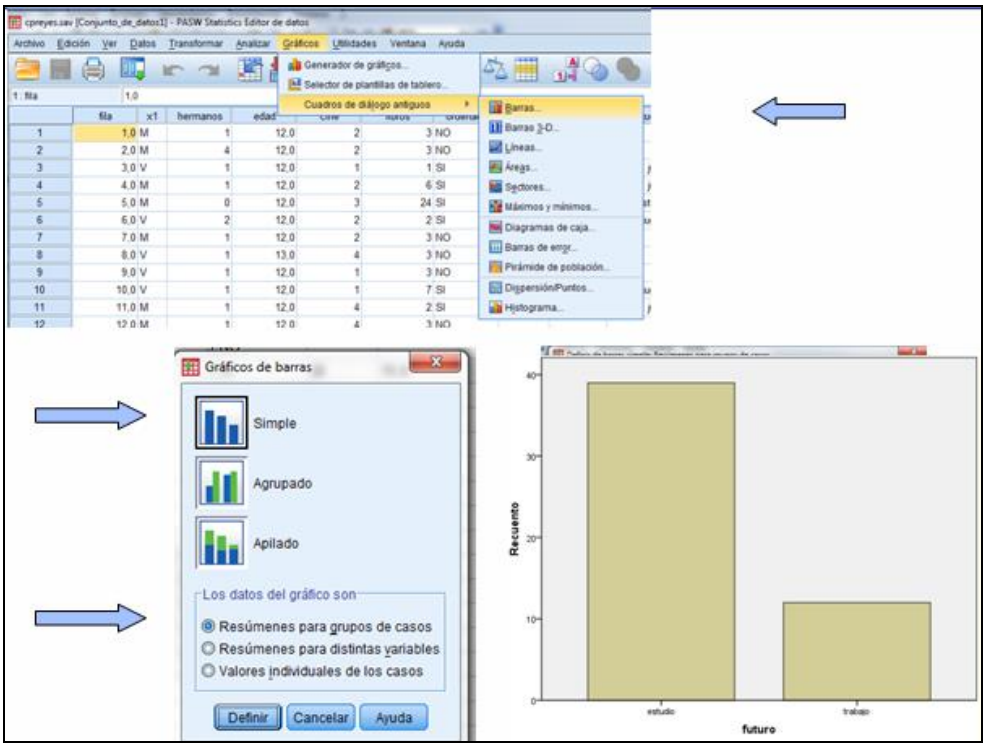
2.5.2. Realización de gráficos

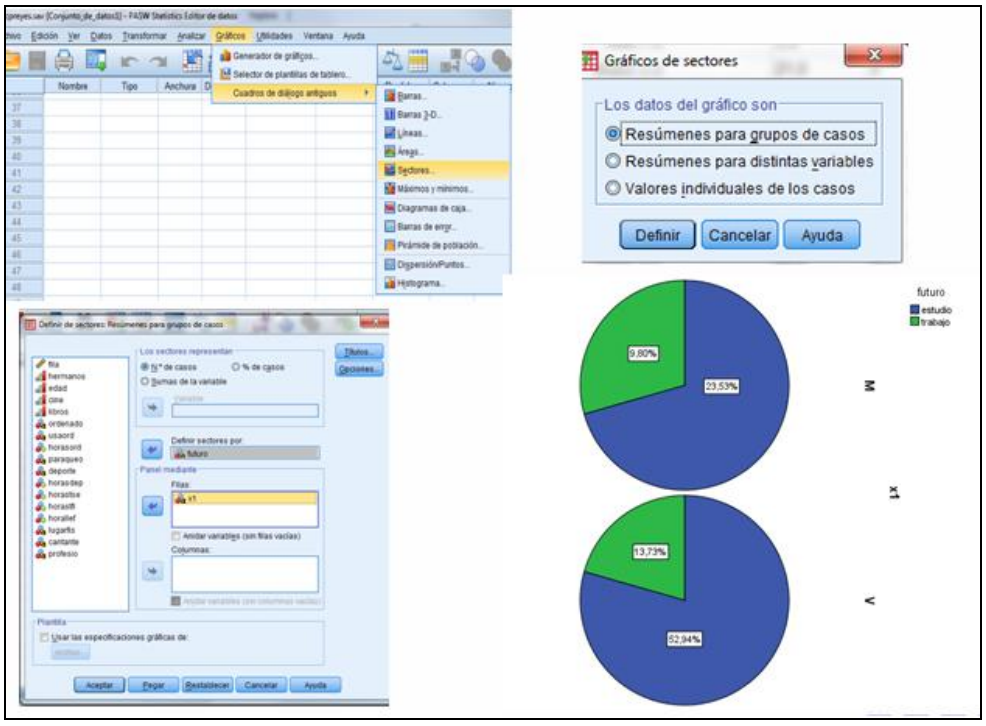
En SPSS existen dos formas de crear un gráfico, todas ellas agrupadas en el menú de 'Gráficos':

- A partir de la opción "Analizar >Estadísticos Descriptivos>Frecuencias.



- De forma rápida, indicando únicamente el tipo de gráfico y la variable a mostrar. Para ello elegimos directamente el tipo de gráfico desde el menú de 'Gráficos'.





2.5.3. Guardar datos y resultados

SPSS almacena en ficheros separados las variables de datos y los resultados de los análisis (tablas y gráficos). Para almacenar las variables lo haremos desde el menú de 'Archivo' del Editor de Datos, mientras que para almacenar los resultados hemos de usar el Visor de Resultados.

2.6 Ejercicios.

1. Clasificar las siguientes variables razonando la respuesta

Nombre	
Edad (en años)	
Peso	
Nota media del expediente (de 0 a 10)	
Número gigabites descargados de Internet en un periodo de tiempo	
Red social a la que perteneces	
Marca personal en salto de longitud	
Color de ojos	
Comprensión lectora (baja, media o alta)	
Estado de conservación de diferentes ordenadores	

2. En la tabla adjunta tenemos la representación de un conjunto de datos obtenidos de una población; se trata de una muestra de 25 ordenadores de una tienda informática, de los cuales observamos varias características; para cada ordenador obtenemos datos correspondientes a las variables:

- X_1: Marca (1)
- X_2: Precio en euros
- X_3: Número de periféricos
- X_4: Sistema Operativo preinstalado (2)
- X_5: Tiempo en segundos de conexión a Internet en un determinado periodo de tiempo

(1)

1. ACER
2. APPLE
3. ASUS
4. DELL
5. HP
6. HUAWEI
7. LENOVO
8. MSI

(2)

1. Windows 10
2. LINUX
3. Mac OS
4. Windows 8
5. UNIX
6. Windows 7

	X_1	X_2	X_3	X_4	X_5
Ordenador	**Marca**	**Precio**	**Periféricos**	**S.O.**	**Tiempo conexión (s)**
1	2	1340	6	3	989
2	6	1025	3	1	1900
3	3	795	5	6	1020
4	5	904	3	5	994
5	1	1020	3	5	1375
6	2	1934	4	3	3000
7	4	1170	3	1	945
8	5	898	4	2	2999
9	7	945	3	1	2184
10	2	1999	4	3	3150
11	4	1085	7	1	1955
12	6	1120	3	1	1025
13	1	955	5	8	895
14	4	1300	7	1	1170
15	3	1095	2	4	1989
16	1	945	3	7	2900
17	7	999	5	1	1999
18	1	1085	3	6	2085
19	6	1120	4	1	3250
20	8	2355	6	1	1855
21	2	1880	5	3	1955
22	8	2170	4	2	925
23	7	1300	2	1	1595
24	2	2500	3	3	1460
25	5	1999	5	5	1895

a) Clasificar las variables. Dar una explicación razonada de la clasificación.
b) Construir las tablas de frecuencias de las variables X_1, X_3 y X_4. ¿Qué conclusiones puedes extraer de las tablas de frecuencias?
c) En general, ¿qué ventajas supone utilizar las frecuencias relativas en vez de las frecuencias absolutas?
d) Representar la variable X_1 de dos formas distintas: mediante un diagrama de barras y mediante un diagrama de sectores en el que se incluyan los porcentajes. ¿Qué conclusiones extraes? ¿Qué ventajas/desventajas ofrece el diagrama de sectores frente al diagrama de barras?
e) Representar mediante un diagrama de barras la variable X_5. Interpreta el resultado.
f) Obtén el gráfico de sectores de la variable X_2 agrupando los precios de los ordenadores en dos categorías: 1- menos de 1000 euros, 2- entre 1000 y 1500 euros, 3- más de 1500 euros. Incluye los porcentajes en el gráfico y explica lo que observas.
g) ¿Qué conclusión global puede extraerse de este estudio?