

## Práctica 3 de Estadística

### 3.1 Distribución de frecuencias de una variable medible.

Las muestras de variables medibles, tanto discretas como continuas, suelen contener en la mayoría de los casos muchos valores distintos. Esto trae consigo tablas de frecuencias en las que los valores de la variable se repiten muy poco, y por tanto, las frecuencias absolutas toman el valor 1 o valores muy bajos. Imaginemos una variable que tome los valores:

X	5	6	7	10	12	13	14	15	16	17	18	20	25	26	27
---	---	---	---	----	----	----	----	----	----	----	----	----	----	----	----

El diagrama de barras es:



La forma que adopta el diagrama no es muy informativa; si la variable toma muchos valores (en el ejemplo sólo son 15), la situación es peor. Por ello conviene agrupar los valores de forma adecuada como se ve en la siguiente sección.

### 3.2 Distribución de frecuencias agrupadas.

Tomemos una muestra de tamaño  $n$  y observemos una variable  $X$ . Supongamos que  $x_1, x_2, \dots, x_n$  son los valores observados, ordenados de menor a mayor. Llamaremos *rango* de la muestra a la diferencia entre el mayor y menor de los valores:

$$R = x_n - x_1$$

Esta longitud la vamos a dividir en una serie de intervalos de igual amplitud, cuyo número dependerá del caso particular en estudio. Generalmente se toman los intervalos de igual amplitud pues producen distribuciones de frecuencias con gráficas más representativas; pero en ocasiones, si hay algún valor anormalmente grande o pequeño,

puede tomarse un intervalo con amplitud mayor para incluir dicho valor, sin necesidad de definir intervalos que no contengan ningún elemento.

Supongamos que decidimos dividir el rango en  $m$  intervalos  $I_1, \dots, I_m$  disjuntos. Entonces la variable  $X$  pasa a ser considerada como categórica con  $m$  categorías, siendo cada categoría uno de los intervalos que llamaremos *Intervalos de clase*.

Se suele representar el intervalo de clase  $I_i$  por su valor central que representaremos por  $c_i$  y que llamaremos *marca de clase*. Este valor representa a todos los valores contenidos en su intervalo.

Se define como *frecuencia absoluta* del intervalo  $I_i$  y la representaremos por  $n_i$ ,  $i = 0, 1, \dots, m$  al número de valores de la muestra comprendidos entre los extremos del intervalo  $I_i$ . Se cumplirá:

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

Se define como *frecuencia relativa* del intervalo  $I_i$ , representada por  $f_i$  a la proporción de valores muestrales que representa el intervalo  $I_i$ :

$$f_i = \frac{n_i}{n} \quad \text{Se cumple que: } \sum_{i=1}^m f_i = 1$$

Un ejemplo de tabla de frecuencias agrupadas es el siguiente:

$I_i$	$n_i$	$f_i$
[70 — 100[	5	0.25
[100 — 130[	6	0.3
[130 — 160[	3	0.15
[160 — 190[	1	0.05
[190 — 220[	1	0.05
[220 — 250[	2	0.1
[250 — 280[	0	0
[280 — 310[	0	0
[310 — 340[	0	0
[340 — 370[	0	0
[370 — 400[	1	0.05
[400 — 430[	1	0.05
	20	1

También podemos definir otro tipo de frecuencias que serán de utilidad en algunos casos: las frecuencias acumuladas.

Llamaremos *frecuencia acumulada absoluta* del valor  $x$  y se representa por  $N_x$  al número de valores de la muestra menores o iguales que  $x$ . Si  $x_1, \dots, x_m$  son los valores distintos ordenados crecientemente, con frecuencias absolutas  $n_1, \dots, n_m$  respectivamente:

$$N_k = N_{x_k} = \sum_{i=1}^k n_i$$

Se cumple:

$$N_1 = n_1$$

$$N_2 = N_1 + n_2$$

...

$$N_i = N_{i-1} + n_i$$

...

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

Llamaremos *frecuencia acumulada relativa* del valor  $x$  y la representaremos por  $F_x$  a la proporción de valores muestrales que representa el valor  $N_x$ , es decir:

$$F_x = \frac{N_x}{n}$$

En particular:

$$F_1 = f_1$$

$$F_2 = F_1 + f_2$$

...

$$F_i = F_{i-1} + f_i$$

...

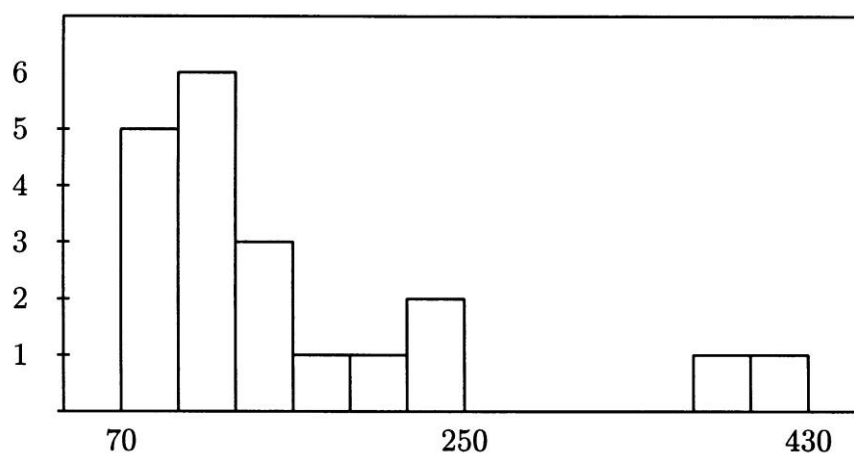
$$F_m = \sum_{i=1}^m f_i = 1$$

Incorporando estos conceptos al ejemplo anterior, la tabla quedaría:

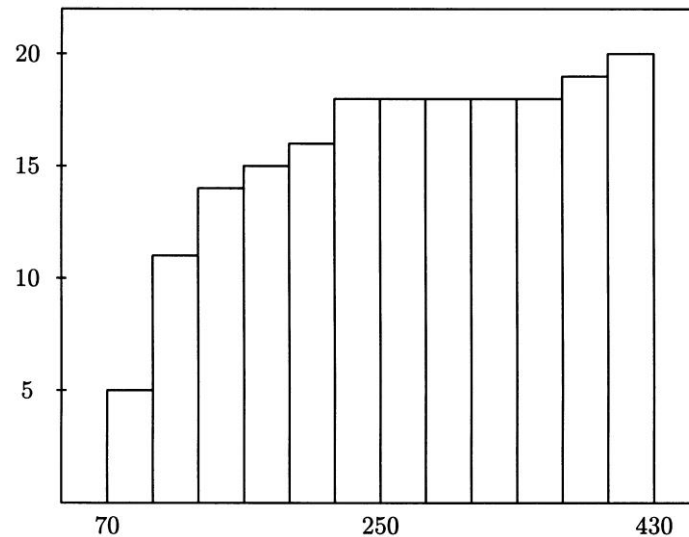
$I_i$	$n_i$	$N_i$	$f_i$	$F_i$
[70 — 100[	5	5	0.25	0.25
[100 — 130[	6	11	0.3	0.55
[130 — 160[	3	14	0.15	0.70
[160 — 190[	1	15	0.05	0.75
[190 — 220[	1	16	0.05	0.80
[220 — 250[	2	18	0.1	0.90
[250 — 280[	0	18	0	0.90
[280 — 310[	0	18	0	0.90
[310 — 340[	0	18	0	0.90
[340 — 370[	0	18	0	0.90
[370 — 400[	1	19	0.05	0.95
[400 — 430[	1	20	0.05	1
	20		1	

### 3.3 Representaciones gráficas

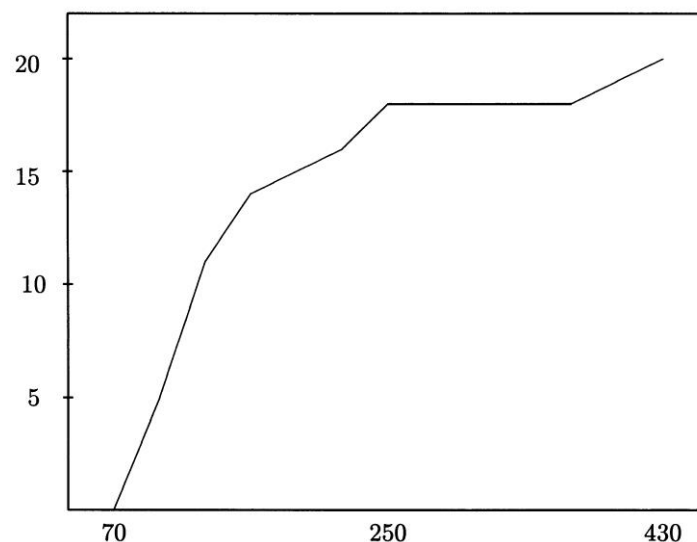
Como vimos, los diagramas de barras, a veces, dan poca información (cuando hay valores de la variable muy próximos); estos agrupamientos de valores pueden ser representados y observados con menor dificultad mediante un adecuado agrupamiento en intervalos. Esta representación recibe el nombre de *histograma*. En el ejemplo anterior tendríamos:



El histograma depende, obviamente, del número de intervalos en que se divide el rango. Si representamos el histograma de frecuencias acumuladas, tendríamos (a otra escala):



Otra representación gráfica interesante para las variables medibles, especialmente cuando son continuas o toman muchos valores distintos, es la gráfica que contiene las distribuciones agrupadas de frecuencias acumuladas, y que recibe el nombre de *Polígono acumulativo*. Se obtiene a partir del histograma de frecuencias acumuladas, sustituyendo los rectángulos por segmentos, formando una poligonal:



Se toma, como frecuencia acumulada del extremo inferior del primer intervalo como 0, y la frecuencia del extremo superior del último intervalo coincidirá con el valor máximo ( $n$  ó 1, según se consideren frecuencias absolutas o relativas). También se pueden unir los puntos medios de los segmentos superiores de los rectángulos.

### 3.4. Manejo de SPSS

#### 3.4.1. Creación de variables definidas en intervalos

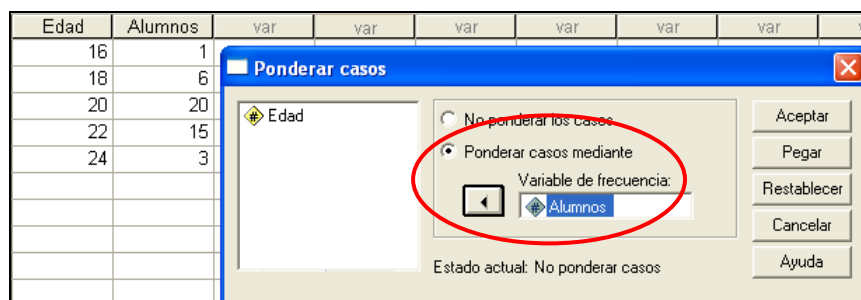
Cuando los valores para realizar un estudio se suministran en forma de tabla de frecuencias agrupada, se han de realizar una serie de operaciones para poder trabajar con la variable. Supongamos que tenemos la siguiente tabla de frecuencias agrupada con las edades de los alumnos en una clase de primero:

Edad del alumno	Núm. Alumnos
<17	1
17-19	6
19-21	20
21-23	15
>23	3

En primer lugar, deberemos calcular el valor central de cada intervalo. Para aquellos casos en que no se pueda (<17, >23), simplemente aplicamos el mismo incremento que al resto de intervalos:

Edad del alumno	Núm. Alumnos
16	1
18	6
20	20
22	15
24	3

Una vez hecho esto, introducimos las dos variables en SPSS, y escogemos la opción 'Datos/Ponderar casos'. De esta manera, indicamos a SPSS que los valores de la variable 'Edad' aparecen tantas veces en la muestra como indica la variable 'Alumnos'.



Una vez hemos ponderado, usaremos únicamente la variable ‘Edad’ para realizar los estudios, es decir, para SPSS es como si tuviéramos 45 casos (alumnos):

Edad	var	var	var	var	var	var	var	var
16								
18								
18								
18								
18								
18								
18								
20								
20								
⋮								

### 3.4.2. Recodificar una variable medible en intervalos

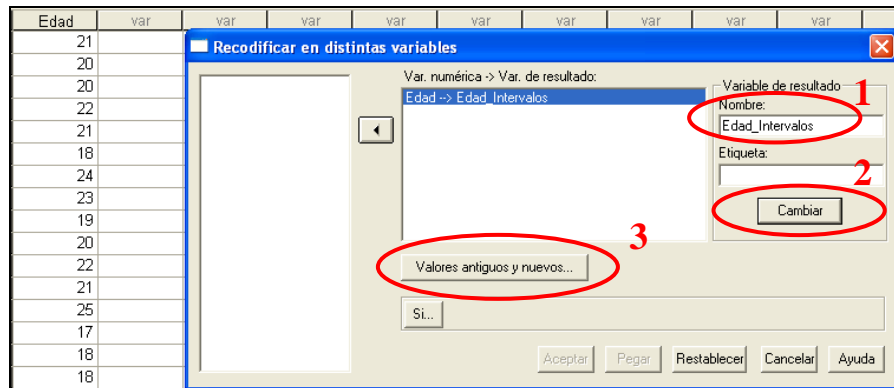
Otra posibilidad que se puede plantear es que nos suministren los valores y debamos agruparlos nosotros en intervalos. Supongamos el siguiente conjunto de datos perteneciente también a las edades de los alumnos de una clase:

21,20,20,22,21,18,24,23,19,20,22,21,25,17,18,18,19,20,19,20

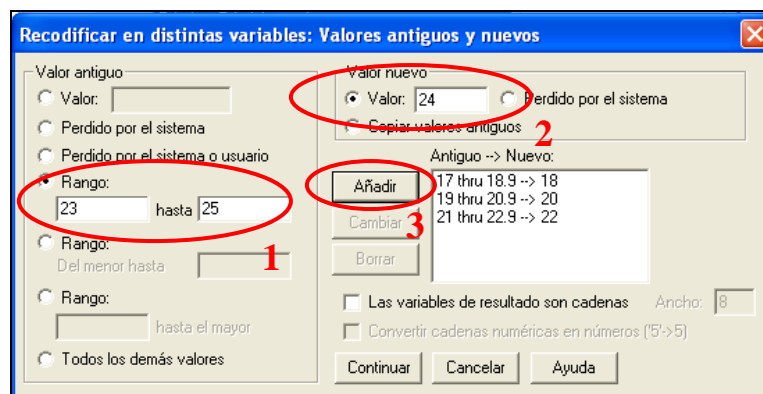
Primero deberemos conocer la amplitud de la muestra. Esto se puede averiguar con la opción ‘Analizar/Estadísticos Descriptivos/Descriptivos’, marcando la casilla ‘amplitud’ en el botón ‘Opciones’. También podemos calcularlo como diferencia entre los valores mínimo y máximo. Una vez hecho esto, escogeremos en cuántos intervalos queremos dividir la muestra o la anchura de los mismos. Para el ejemplo anterior, la amplitud es  $25-17 = 8$ . Si dividimos la muestra en 4 intervalos, la amplitud de cada intervalo será  $8/4 = 2$ . Con estos datos, podemos calcular los intervalos para dividir la muestra, empezando con el valor mínimo y añadiendo la amplitud del intervalo:

[17-19[ , [19-21[ , [21-23[ , [23-25[

Para llevar a cabo este proceso, primero crearemos una variable con los datos originales, y una vez hecho esto, seleccionaremos la opción ‘Transformar/Recodificar/En distintas variables’. Aquí deberemos indicar la variable original y la nueva variable a crear (la que contiene los intervalos). Para completar la operación, deberemos pulsar en el botón ‘Cambiar’.



Una vez hayamos hecho esto, pulsaremos en el botón ‘Valores antiguos y nuevos...’ para especificar los intervalos de la variable original que vamos a agrupar, y el nuevo valor que le asignamos (por lo general, el valor central de dicho intervalo). Hemos de tener siempre en cuenta que los intervalos han de ser disjuntos, esto es, no puede coincidir el final de un intervalo con el inicio del siguiente.

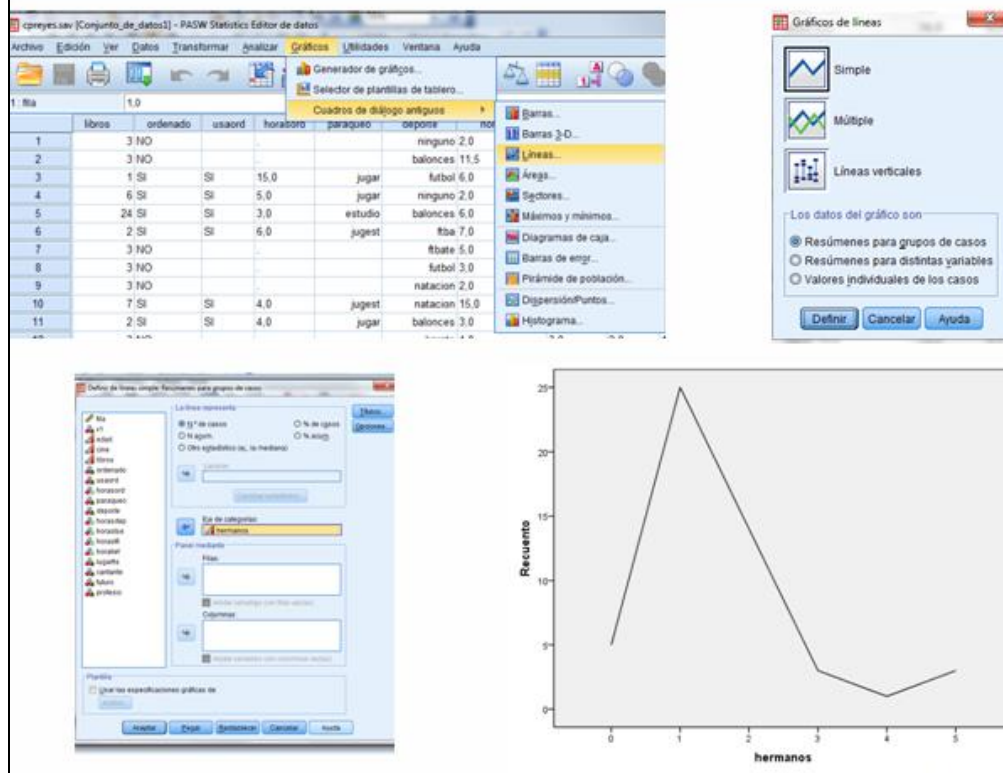


### 3.4.3. Histogramas y Polígonos de Frecuencias

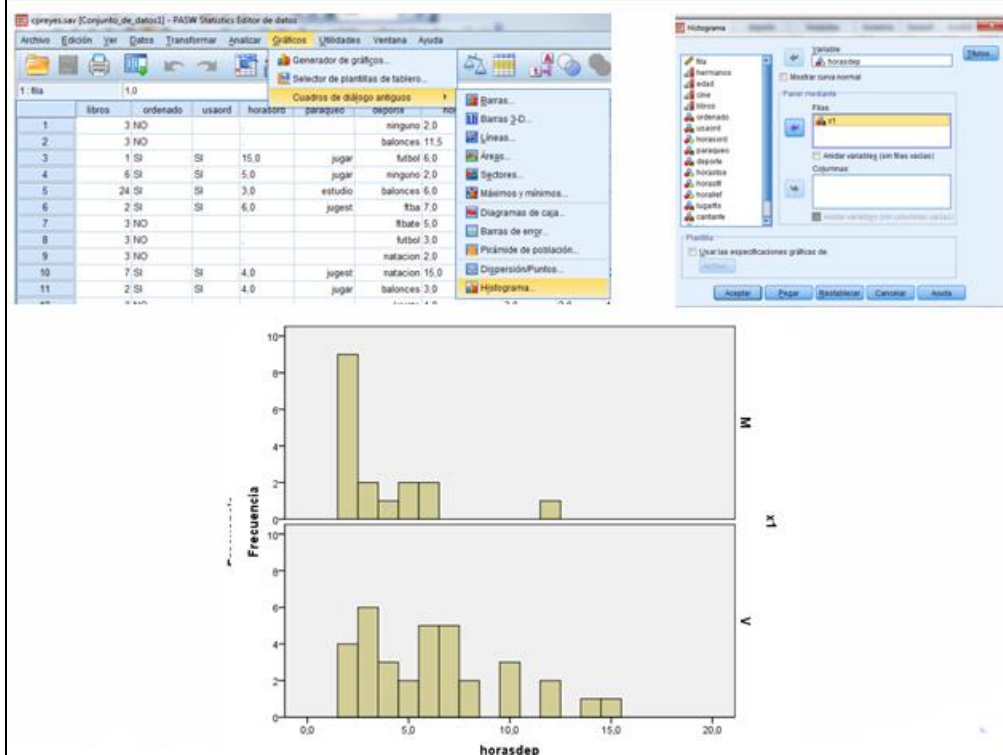
El Histograma y el Polígono de frecuencias son dos herramientas gráficas para el análisis agrupado de variables medibles. Para realizar un Polígono de frecuencias, primero hemos de recodificar la variable con alguno de los métodos vistos en los puntos 3.4.1. y 3.4.2. Una vez hecho esto, elegimos la opción ‘Gráficos>cuadro de diálogo antiguos>Líneas’. En el caso del Histograma escogeremos la opción ‘Gráficos>cuadro de diálogo antiguos>Histograma’ y seleccionaremos la variable.



## Polígono de frecuencias



## Histograma



### 3.5 Ejercicios

1. Los 80 programadores de una empresa disponen de espacio en un servidor de disco para almacenar los datos de sus programas. El uso que hacen del mismo es el siguiente:

Capacidad (Gb)	Número de programadores
Menos de 500	5
De 501 a 1000	12
De 1001 a 1500	21
De 1501 a 2000	25
De 2001 a 2500	7
De 2501 a 3000	1
Más de 3001	9

Para analizar el uso que hacen del espacio en el servidor, se pide:

- a) Construye la tabla de frecuencias completa.
  - b) Representa el histograma.
  - c) Explica e interpreta los resultados obtenidos en los apartados anteriores.
2. Se ha medido cincuenta veces el tiempo de ejecución de un programa en un servidor, obteniéndose los siguientes resultados, en segundos, que varían dependiendo de la carga que haya tenido el servidor durante la ejecución:

66, 70, 68, 81, 77, 73, 66, 81, 78, 74,  
70, 71, 79, 72, 64, 72, 82, 74, 79, 65,  
64, 81, 69, 76, 75, 69, 80, 68, 75, 81,  
67, 78, 82, 74, 79, 63, 80, 77, 76, 68,  
72, 60, 68, 58, 62, 76, 83, 61, 80, 77

Agrupando los tiempos en intervalos de clase de longitud 5, obtener el histograma, el polígono de frecuencias y el polígono de frecuencias acumuladas. Explica e interpreta los resultados obtenidos.

3. Durante un mes se contabiliza el número de visitas a las 1000 páginas de un sitio web. El fichero **PRACTICA3E19-datos-ejer3** con los datos se puede encontrar en UACloud.
- a) Agrupa los datos en intervalos de la misma amplitud y forma la correspondiente tabla de frecuencias. Explica e interpreta los resultados obtenidos.
  - b) Obtén el polígono de frecuencias. Explica e interpreta los resultados obtenidos.
  - c) Dibuja dos histogramas de 15 y 25 intervalos y razona cuál de ellos sería el más adecuado para representar los datos. Explica e interpreta los resultados obtenidos.
  - d) ¿Qué conclusiones generales puedes extraer?