T. 2 – Organización y representación gráfica de los datos

1. **La distribución de frecuencias**
2. **La representación gráfica de una distribución de frecuencias**
3. **Propiedades de las distribuciones de frecuencias**

• La recogida de datos asociada a la realización de un estudio suele representar la obtención de un conjunto más o menos numeroso de datos, ahora bien, la interpretación de los mismos a simple vista suele resultar poco inteligible en la mayoría de los casos. La *estadística descriptiva* nos ofrece herramientas para organizar y resumir los datos que hayamos recogido, de modo que pueda ser extraída e interpretada la información contenida en los mismos que sea de nuestro interés.

1. La distribución de frecuencias

• La distribución de frecuencias constituye una de las formas más intuitiva de organizar los datos de de una variable: se basa en el conteo del número de entidades (casos, sujetos) que tienen cada uno de los valores con que la variable se ha manifestado (modalidades). Es una técnica estadística básica pero, a la vez, muy informativa y relevante en la práctica del análisis de datos.

• El número de veces que aparece una determinada modalidad de una variable (X) es lo que se conoce como la frecuencia absoluta (n_i) de esa modalidad o valor.

- Derivadas de las frecuencias absolutas se pueden obtener las frecuencias relativas o proporciones (p_i):

$$p_i = n_i / n$$

- Las frecuencias relativas también pueden expresarse como porcentajes ($\%_i$) con tan sólo multiplicar su valor por 100:

$$\%_i = p_i \cdot 100$$

Ejemplo para la variable categórica "Estado civil" (X), habiendo sido recogidos datos para una muestra de 50 personas de la ciudad de Castellón (n = 50):

X: {0, 0, 1, 2, 2, 0, 1, 3, 2, 0, 1, 0, 1, 2, 0, 2, 1, 1, 0, 1, 0...}

Codificación: [0: soltero/a; 1: casado/a; 2: separado/a o divorciado/a; 3: viudo/a]

X_i	Frec. absoluta (n_i)	Frec. relativa (p_i)	Porcentaje ($\%_i$)
0	15	0,3	30
1	20	0,4	40
2	11	0,22	22
3	4	0,08	8
	50	1,00	100

• En el caso de las variables cuantitativas y las cuasi-cuantitativas , además de lo anterior, se puede obtener también la siguiente información para cada una de las modalidades:

- las frecuencias absolutas acumuladas (n_a),
- las frecuencias relativas acumuladas (p_a),
- y los porcentajes acumulados ($\%_a$).

Ejemplo para la variable cuantitativa "Nº de hijos/as" (X), con datos para una muestra de 20 familias del barrio de Velluters de la ciudad de Valencia:

X: {2, 1, 0, 3, 2, 2, 3, 1, 1, 0, 1, 2, 1, 2, 0, 2, 4, 2, 3, 1}

X_i	Frec. absoluta (n_i)	Frec. relativa (p_i)	Porcentaje ($\%_i$)	Frec. absoluta acumulada (n_a)	Frec. relativa acumulada (p_a)	Porcentaje acumulado ($\%_a$)
0	3	0,15	15	3	0,15	15
1	6	0,30	30	9	0,45	45
2	7	0,35	35	16	0,80	80
3	3	0,15	15	19	0,95	95
4	1	0,05	5	20	1,00	100
	20	1	100			

• Algunas anotaciones acerca de las distribuciones de frecuencias:

(1) Es costumbre situar los valores correspondientes a la columna de las modalidades de la variable X en sentido creciente de arriba hacia abajo.

(2) Para los valores de la variable que no haya ningún caso es costumbre no dedicar ninguna fila en la tabla de la distribución de frecuencias a fin de que ésa ocupe menos espacio.

(3) Las frecuencias relativas o proporciones se caracterizan por: tomar valores entre 0 y 1; ser la suma de todas ellas igual a la unidad.

Ejercicio 1: Las siguientes datos son de los estudiantes de una clase en la que un observador, durante el tiempo que ha durado una sesión de clase de 2 horas, ha anotado el número de veces que ha participado cada uno de los estudiantes dirigiéndose a todo el grupo en voz alta.

2 2 3 0 3 1 8 0 3 9 1 1 0 4 0 2 9 5 0 1 9 8

Obtener la distribución de frecuencias completa y contestar las siguientes preguntas (utilizar dos decimales en los cálculos y a la hora de presentar resultados):

a) ¿Qué proporción de estudiantes participaron en menos de 2 ocasiones en la sesión de clase?, ¿cuántos estudiantes son?

b) ¿Qué porcentaje de estudiantes participaron 5 veces? ¿Cuántos son?

c) ¿Qué proporción de estudiantes participaron 4 veces o menos? ¿Cuántos son?

d) ¿Qué porcentaje de estudiantes participaron más de 4 veces? ¿Cuántos son?

e) ¿Qué proporción de estudiantes participaron al menos una vez? ¿Cuántos son?

f) ¿Qué porcentaje de estudiantes participaron entre 2 y 5 veces, ambas inclusive? ¿Cuántos son?

g) ¿Qué porcentaje de estudiantes participaron 8 ó 9 veces? ¿Cuántos son?

h) ¿Qué proporción de estudiantes participaron 4 veces o más? ¿Cuántos son?

(4) En el caso de las variables cuantitativas continuas dado que, si la medida de la variable se realiza con cierta precisión, se puede obtener un número amplio de datos diferentes, es práctica habitual que en la columna de las modalidades (X_i) los valores representen a intervalos de valores de igual amplitud.

Ejemplo de la distribución de frecuencias elaborada a partir de los datos de la variable "Peso (*kg*)" (*X*) de los 420 jugadores inscritos en la liga profesional masculina de balonmano en la temporada 2008/09:

X: {82,5; 91,1; 90,6; 83,8; 92,1; 88,3; 93,6; 101,4; 91,7; 80,2; …}

X_i (kg)	n_i
77	1
79	3
80	2
81	6
82	5
83	9
...	...

Así, por ejemplo, el valor 80 de la columna de las modalidades representa, en realidad, al conjunto de valores comprendido entre 79,5 y 80,5 kg; el valor 81 al intervalo de 80,5 a 81,5 kg, y así sucesivamente. Recuérdese que en la enumeración de intervalos que se solapan en un punto es habitual considerar que el primer valor del intervalo forme parte del mismo, mientras que el segundo ya se considere del siguiente intervalo.

(5) Siguiendo con el caso anterior, si el número de modalidades que toma la variable es muy amplio, una alternativa que permite generar una distribución de frecuencias más compacta consiste en organizar la distribución de frecuencias definiendo intervalos de valores.

Ejemplo de la distribución de frecuencias elaborada a partir de los datos de la variable "Altura (*cm*)" para una muestra de 1436 sujetos adultos de la población española:

X_i (cm)	n_i
140-149	15
150-159	131
160-169	345
170-179	623
180-189	267
190-199	42
200-209	13

En este caso, el intervalo 140-149, por poner un ejemplo, representa a todos los valores comprendidos entre 139,5 y 149,5 cm.

Ejercicio 2: A partir de la distribución de frecuencias de la variable "Altura (cm)" presentada más arriba, obtén las correspondientes columnas de frecuencias relativas, porcentajes, frecuencias absolutas acumuladas, frecuencias relativas acumuladas y porcentajes acumulados.

Ejercicio 3: En una encuesta sobre condiciones psicosociales en el lugar de trabajo se preguntó a una muestra de 3420 sujetos sobre "en qué medida su trabajo es desgastador emocionalmente" y se obtuvieron los siguientes resultados:

X_i	p_a
Nunca	0,098
Alguna vez	0,175
A veces	0,332
Muchas veces	0,531
Siempre	1

a) ¿Qué proporción de sujetos considera que su trabajo es desgastador emocionalmente alguna vez?

b) ¿Qué porcentaje de sujetos considera que su trabajo es desgastador emocionalmente siempre?

c) ¿Cuántos sujetos consideran que su trabajo es desgastador emocionalmente a veces? ¿Y cuántos consideran que siempre lo es?

(6) Una distribución de frecuencias condicionada muestra la distribución de frecuencias de una variable para los casos que en una segunda variable tienen un determinado valor.

Sea, por **ejemplo**, la distribución de frecuencias de la variable "Calificaciones examen", siendo el tamaño de la muestra igual a 200 (n = 200)

Calificaciones examen	n_i
Aprobado	130
Notable	45
Sobresaliente	23
Matrícula honor	2
	200

A continuación se muestran las distribuciones de frecuencias de la variable "Calificaciones examen" condicionada a los valores de la variable "Sexo" [Mujer; Hombre]:

Calificaciones examen	(Sexo: Mujer) n_i	(Sexo: Hombre) n_i
Aprobado	80	50
Notable	20	25
Sobresaliente	14	9
Matrícula de honor	1	1
	115	85

Si el tamaño de los subgrupos definidos por la variable condicionante no es igual o bastante similar, es conveniente presentar las distribuciones de frecuencias expresadas en proporciones o porcentajes a fin de que la comparación entre ambas sea más intuitiva. Sea el caso de la variable "Calificaciones examen" condicionada a la variable "Sexo".

Calificaciones examen	(Sexo: Mujer) $\%_i$	(Sexo: Hombre) $\%_i$
Aprobado	69,6	58,8
Notable	17,4	29,4
Sobresaliente	12,2	10,6
Matrícula de honor	0,87	1,18
	100	100

Ejercicio 4: En el contexto de un estudio sobre la percepción de la ciencia y la tecnología en España se preguntó a una muestra de 7054 sujetos (1252 jóvenes y 5802 adultos) cómo valoraban el nivel de la formación científica y técnica recibida. Los resultados fueron:

Nivel de formación	Joven	Adulto
Muy bajo	8%	22,5%
Bajo	28,5%	33,3%
Normal	45,5%	32,8%
Alto	14,3%	9,0%
Muy alto	3,7%	2,4%

a) Expresa las distribuciones anteriores en frecuencias absolutas.

b) Genera la distribución de frecuencias para el conjunto de sujetos de la muestra en frecuencias absolutas y en frecuencias relativas.

c) ¿Qué grupo de sujetos está más satisfecho con el nivel de formación recibido? ¿Crees que hay relación entre las dos variables presentadas en este ejemplo?

• El programa SPSS: Al obtener la distribución de frecuencias de una variable con este programa se muestra n_i, $\%_i$, $\%_i$ *válido* y $\%_a$, pero no la información referida a las frecuencias relativas (p_i y p_a). La diferencia entre $\%_i$ y $\%_i$ *válido* reside que el primero se obtiene dividiendo cada frecuencia relativa entre el número total de casos para los que se plantea la recogida de información, mientras que el segundo se obtiene dividiendo entre el número de casos para los que de hecho se ha recogido algún dato en la variable, no teniéndose en cuenta los sujetos que tienen valores faltantes en la variable.

Véanse los siguientes **ejemplos**: el primero para la variable "Satisfacción con las instalaciones de un centro deportivo" para un grupo de 106 usuarios del mismo, habiendo sido recogido la información a través de una escala de 0 a 20 [0: totalmente insatisfecho; ... ; 20: totalmente satisfecho]; y el segundo para la variable "Ingresos económicos anuales" [Altos; Medios; Bajos] recogida a partir de la pregunta de una encuesta realizada a una muestra de 40 personas entrevistadas a la entrada de un centro comercial.

sat_ins

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	4	1	,9	,9	,9
	7	4	3,8	3,8	4,7
	8	11	10,4	10,4	15,1
	9	7	6,6	6,6	21,7
	10	26	24,5	24,5	46,2
	11	7	6,6	6,6	52,8
	12	24	22,6	22,6	75,5
	13	4	3,8	3,8	79,2
	14	11	10,4	10,4	89,6
	15	4	3,8	3,8	93,4
	16	4	3,8	3,8	97,2
	17	1	,9	,9	98,1
	18	2	1,9	1,9	100,0
	Total	106	100,0	100,0	

Ingresos

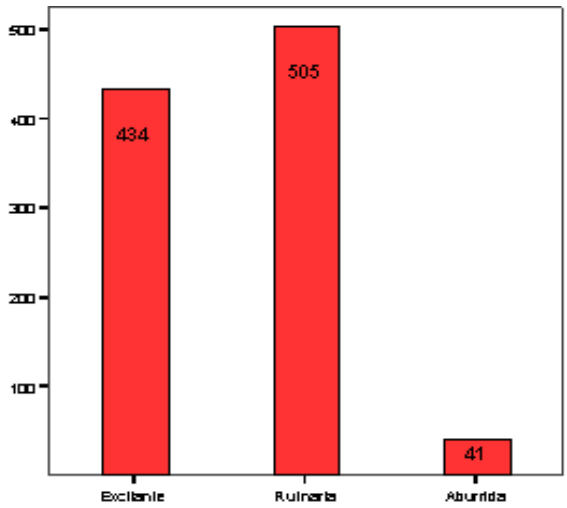
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Altos	4	10,0	21,1	21,1
	Medios	12	30,0	63,2	84,2
	Bajos	3	7,5	15,8	100,0
	Total	19	47,5	100,0	
Perdidos	Sistema	21	52,5		
Total		40	100,0		

2. La representación gráfica de una distribución de frecuencias

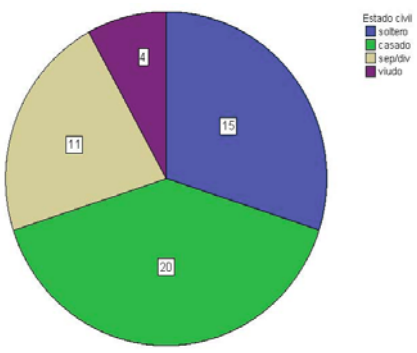
2.1. Para variables categóricas

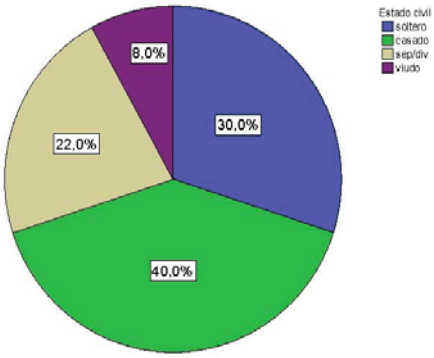
• El diagrama de barras: Las modalidades de la variable se sitúan sobre el eje *X* (abscisas). La altura de las barras es proporcional a la frecuencia absoluta de cada una de las modalidades de la variable. El eje de ordenadas puede aparecer expresado en frecuencias absolutas, en frecuencias relativas o en porcentajes. Los diagramas de barras pueden representarse también de forma horizontal.

Ejemplo para la variable procedente de la siguiente pregunta de un test: "¿Cómo es su vida?"

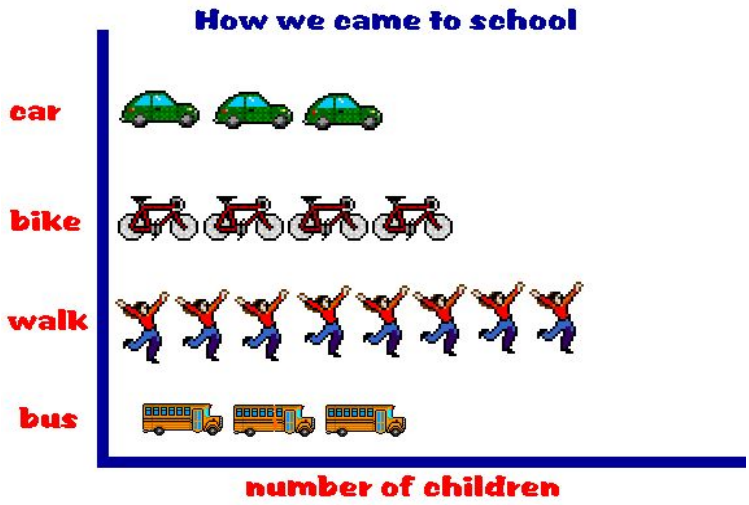


• El diagrama de sectores (pastel, tarta): el área de cada sector es proporcional a la frecuencia o % de la modalidad a la que representa.





•El pictograma: es una variación gráfica de los diagramas de barras.



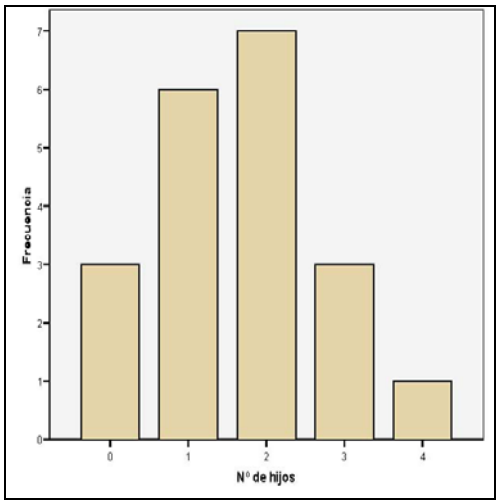
Pueden resultar más atractivos en la presentación de resultados, aunque también más ambiguos en su lectura e interpretación.

Ejercicio 5: A partir de los gráficos de las variables "¿Cómo es su vida?", "Estado civil" y "¿Se encuentra a gusto haciendo lo que hace?", obtener las correspondientes distribuciones de frecuencias con las columnas que correspondan en cada caso.

2.2. Para variables cuasi-cuantitativas y cuantitativas discretas

• El diagrama de barras: se representa de forma análoga a como se hace para las variables categóricas. Señalar que el hueco entre las barras sirve para resaltar que hay valores que no son posibles para la variable representada. A diferencia de las variables categóricas, para este tipo de variables tiene sentido representar no sólo las frecuencias absolutas, las relativas y los porcentajes, sino también las respectivas acumuladas.

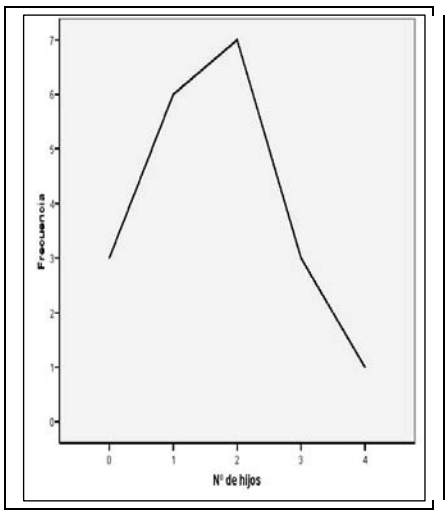
Ejemplo para la variable "Nº de hijos":



Ejercicio 6: Dibujar a su derecha el correspondiente diagrama de barras de frecuencias acumuladas.

• Polígono de frecuencias: polígono que resulta de unir con una línea los valores de las frecuencias o %s (ya sean acumulados o no) correspondientes a las modalidades de la variable.

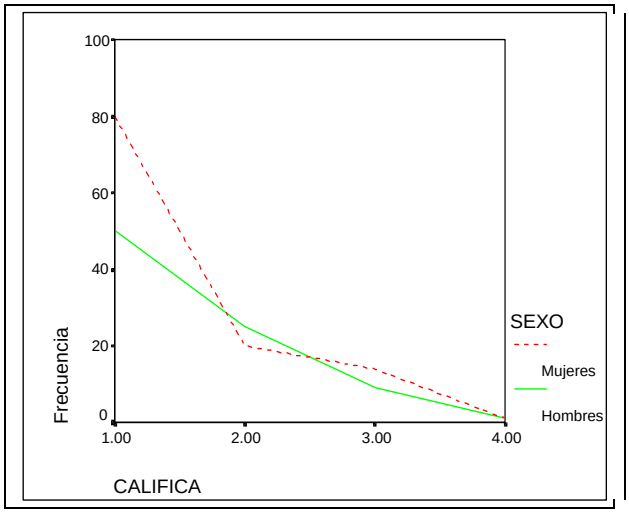
Ejemplo para la variable "Nº de hijos":



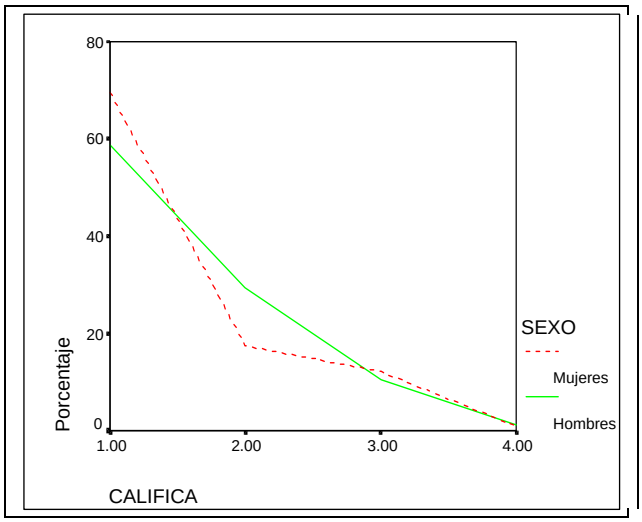
Ejercicio 7: Dibujar a su derecha el correspondiente polígono de frecuencias acumuladas.

- Los polígonos de frecuencias facilitan la superposición gráfica, por ejemplo, para comparar dos variables para un mismo conjunto de casos, o bien, para comparar las distribuciones de frecuencias de una variable condicionada a los valores de una segunda variable.

Ejemplo de polígono de frecuencias superpuesto para la distribución de frecuencias absolutas de la variable "Calificaciones examen" condicionada a la variable "Sexo", cuyos datos se presentaron en un ejemplo anterior al introducir el concepto de distribución de frecuencias condicionada:

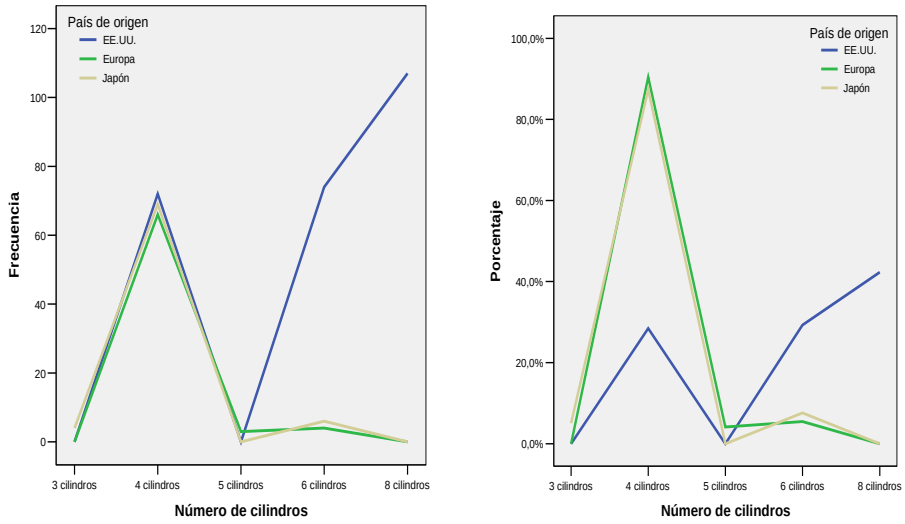


- Si el tamaño de los subgrupos definidos por la variable condicionante no son iguales o bastante similares, es conveniente representar los polígonos de frecuencias superpuestos utilizando frecuencias relativas o porcentajes, a fin de que los polígonos puedan compararse de un modo equitativo, no distorsionado por el diferente tamaño de los subgrupos. Sea el caso de la variable "Calificaciones examen" condicionada a la variable "Sexo".



Un **ejemplo** en el que se aprecia más este hecho es el siguiente con los datos de un estudio que se hizo en los EEUU sobre las características de los diferentes modelos de coches existentes en el mercado en el momento en que se realizó el estudio. En concreto, a continuación se muestra la información correspondiente a la distribución de frecuencias de la variable "Nº de cilindros" condicionada a la variable "País de origen del vehículo", así como los correspondientes gráficos de polígonos de frecuencias superpuestos expresados en frecuencias absolutas y en porcentajes:

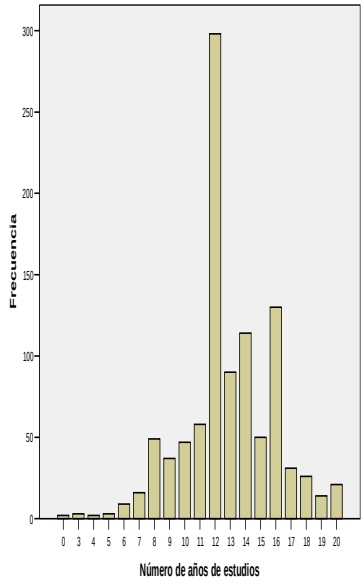
		País de origen			
		EE.UU.	Europa	Japón	Total
Número de cilindros	3 cilindros			4	4
	4 cilindros	72	66	69	207
	5 cilindros		3		3
	6 cilindros	74	4	6	84
	8 cilindros	107			107
Total		253	73	79	405

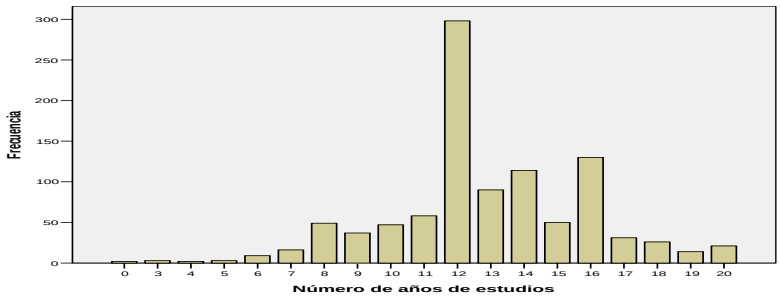


Ejercicio 8: Realizar para la variable "Nº de veces que se participa en clase" (ver ejercicio 1), los diagramas de barras correspondientes a: las frecuencias absolutas; las frecuencias relativas; las frecuencias absolutas acumuladas; las frecuencias relativas acumuladas. Dibujar un polígono de frecuencias a partir de cualquiera de los anteriores.

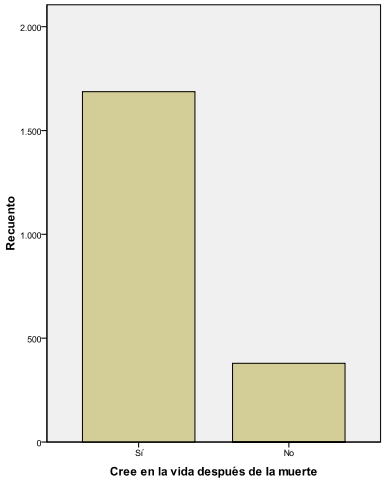
• Un aspecto que puede influir la percepción de una representación gráfica es la relación entre el tamaño del eje X y del eje Y. Por **ejemplo**, los dos siguientes gráficos, aún representando unos

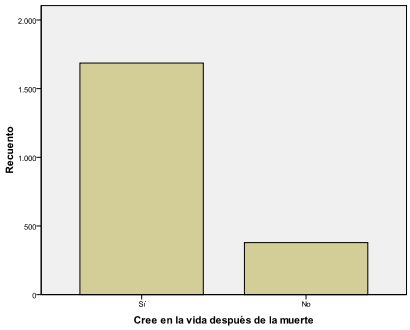
mismos datos (variable "Nº de años de estudios" para una muestra de 1000 personas de la ciudad de Elche), pueden dar lugar a una percepción diferente de la información proporcionada:

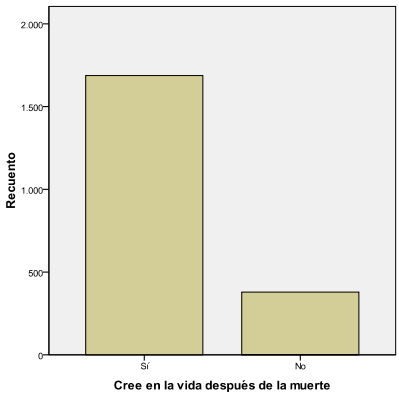




A fin de evitar esta posible fuente de confusión, algunos autores recomiendan que la relación entre la anchura y la altura del gráfico sea de 1,25 a 1. A modo de ejemplo, ¿cuál de las siguientes representaciones gráficas, correspondientes a un mismo conjunto de datos, atiende a esta recomendación (datos obtenidos a partir de una muestra de 2832 encuestados de EEUU)?





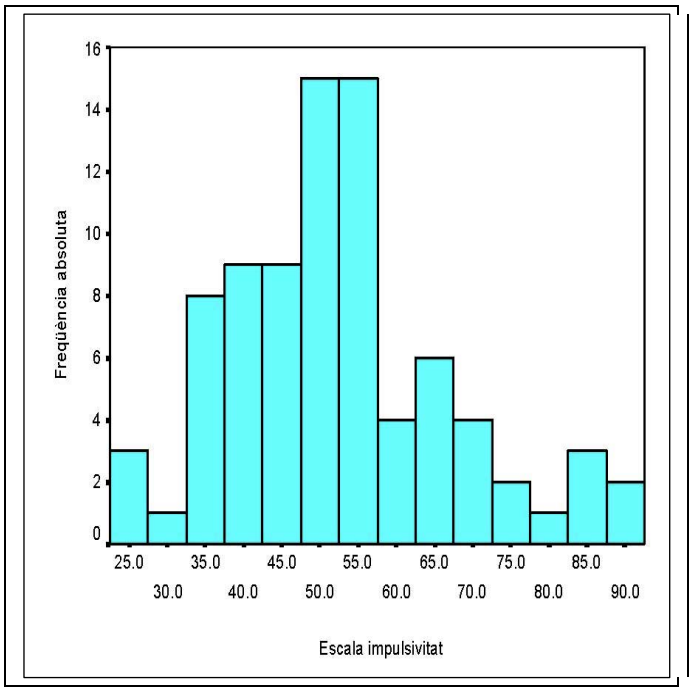


Ejercicio 9: Realizar para la variable "Nº de veces que se participa en clase" (ver ejercicio 1), los diagramas de barras correspondientes a: las frecuencias absolutas; las frecuencias relativas; las frecuencias absolutas acumuladas; las frecuencias relativas acumuladas. Dibujar un polígono de frecuencias a partir de cualquiera de los anteriores.

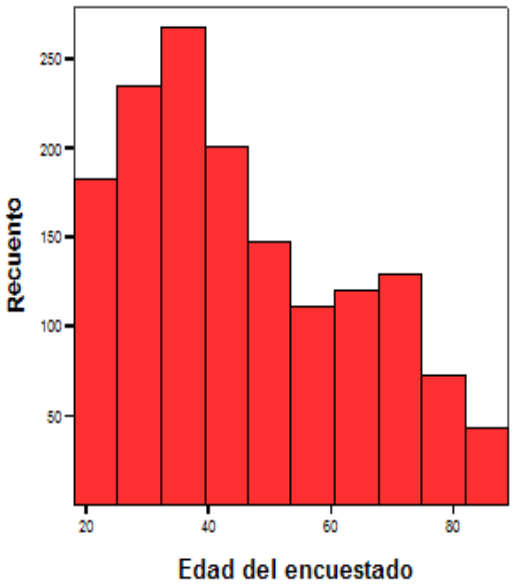
2.3. Para variables cuantitativas continuas

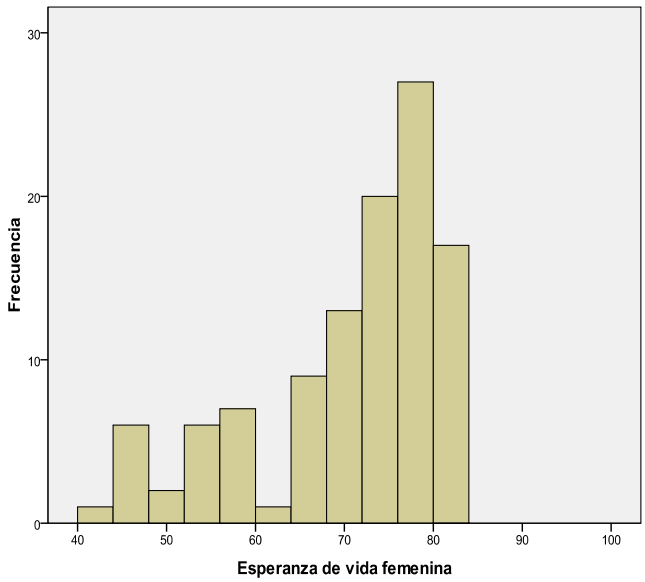
• Histograma: similar al diagrama de barras, si bien, las barras son consecutivas dada la continuidad de la variable. Cada barra representa ahora, no a un valor, sino a un intervalo de valores. A la hora de definir los intervalos de valores se debe tener en cuenta que ninguno de los datos recogidos para la variable se quede fuera de los intervalos. Los intervalos deber ser exhaustivos y excluyentes.

Ejemplo para las puntuaciones obtenidas por un grupo de sujetos en una escala de impulsividad:



Otros **ejemplos** de histogramas obtenidos con el programa SPSS, el de la izquierda para la variable "Edad" de los participantes en una encuesta y, el de la derecha, para la variable "Esperanza de vida femenina" obtenido para un total de 109 países del mundo. En ambos, ¿es conveniente la elección de los intervalos de valores representados?





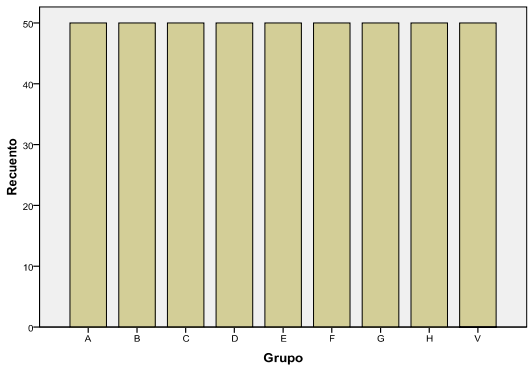
• También es posible dibujar polígonos de frecuencias para las variables cuantitativas continuas uniendo con una línea los valores de las frecuencias o %s (ya sean acumulados o no) correspondientes a los intervalos de valores creados.

3. Propiedades de las distribuciones de frecuencias

• Si bien la representación gráfica de una distribución de frecuencias puede adoptar múltiples formas, existen algunos patrones de distribución que, por lo particular de los mismos y/o por su importancia, han sido denominados de un modo concreto.

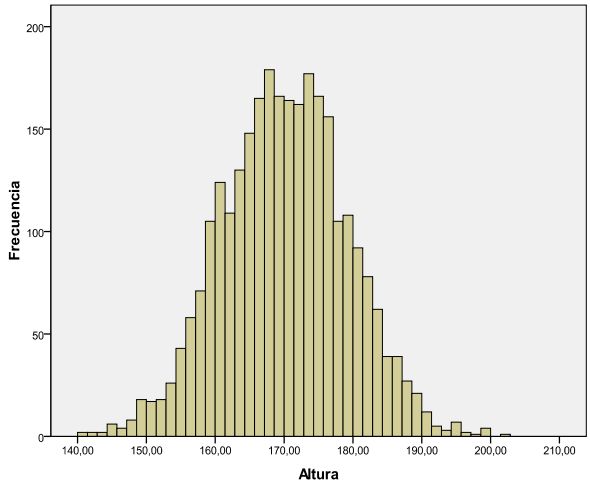
A modo de **ejemplo**, las dos siguientes presentadas en forma gráfica para dos variables concretas:

- La distribución rectangular o uniforme:



Asignatura de Estadística: Nº de estudiantes por grupo.

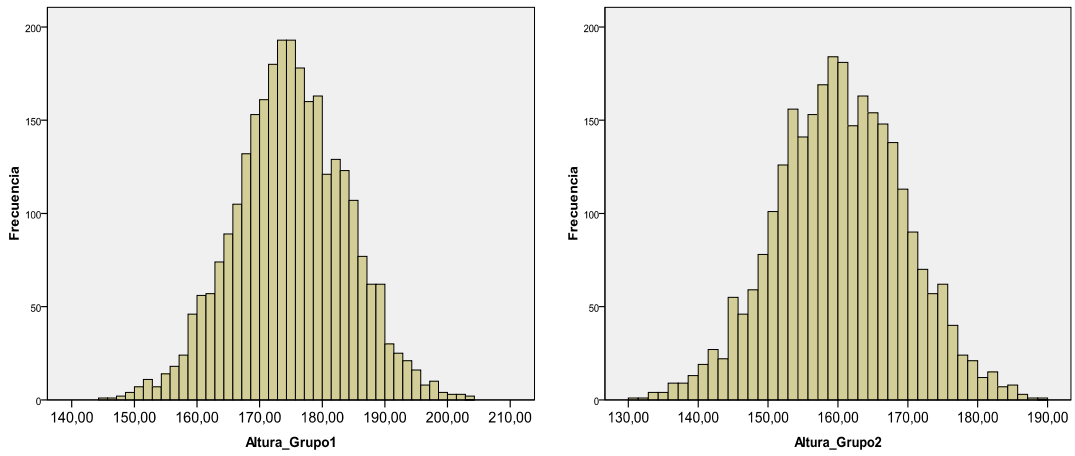
- La distribución normal:



• Sobre estos dos patrones y otros que caracterizan en su conjunto a la distribución de frecuencias de algunas variables se profundizará en un tema posterior. Ahora bien, a la hora de describir una distribución de frecuencias podemos atender, más que a la forma en su conjunto, a diferentes facetas particulares de la misma. Así, los 3 temas que siguen a éste se centran en algunas de estas facetas que permiten sintetizar la información contenida en una distribución de frecuencias. Se trata de facetas como las dos siguientes, las cuales se presentan aquí simplemente a título introductorio y a través de ejemplos gráficos que permitan captar el significado de las mismas:

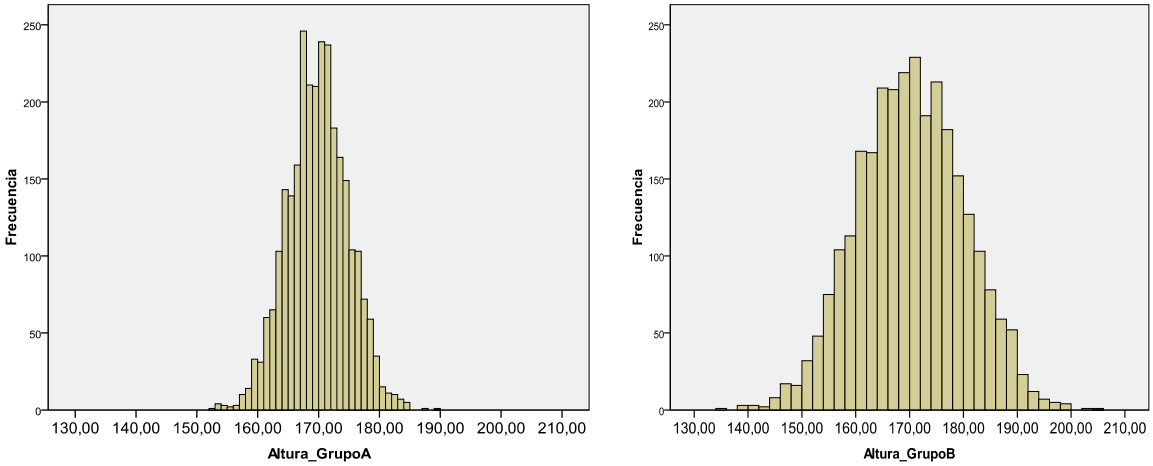
- La posición de la distribución

Ejemplo de la diferente posición de las dos distribuciones de frecuencias de una misma variable, "Altura (cm)", medida en dos grupos de sujetos distintos:



- La dispersión o variabilidad de la distribución

Ejemplo de la diferente dispersión de las dos distribuciones de frecuencias de una misma variable, "Altura (cm)", medida en dos grupos de sujetos distintos -que, en cambio, comparten una posición muy similar:



Referencias

Peña, D. y Romo, J (1997). *Introducción a la estadística para las ciencias sociales*. Madrid: McGraw-Hill.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677-680.