Introducción A la Estadística[1]

INTRODUCCIÓN:

La Estadística *descriptiva* es una parte de la Estadística cuyo objetivo es examinar a todos los individuos de un conjunto para luego describir e interpretar numéricamente la información obtenida.

Sus métodos están basados en la observación y el recuento. Se pretende, una vez realizados, poder simplificar los datos observados para obtener de ellos una información lo más completa posible del total de la población.

En estadística descriptiva el material de trabajo lo constituyen los datos, que son los resultados de las observaciones. Una vez obtenidos los datos hay que ordenarlos y clasificarlos mediante algún criterio racional de modo que sea posible una visión crítica de los mismos.

En general, este tratamiento previo de los datos será de alguno de estos tres tipos:

1) Construcción de **tablas** para ordenar y clasificar los datos.

2) Realización de **gráficos** para representar físicamente los datos.

3) Obtención de **estadísticos** o funciones de los valores de los datos, que pretenden poner de manifiesto ciertas propiedades de los mismos.

1. Conceptos básicos.

Cualquier elemento o ente que sea portador de información sobre alguna propiedad en la cual se está interesado se denomina **individuo.**

El conjunto de todos los individuos en los que se desea estudiar alguna propiedad o característica se llama **población.**

Todo subconjunto finito de la población sobre el que se realice el estudio de la propiedad deseada, es *una* **muestra.** Al número de individuos de este subconjunto se le llama **tamaño** de la muestra.

Ejemplo 1. Para estudiar la evolución del cáncer de mama en la población femenina de un país, se puede considerar que individuo es cada una de las mujeres residentes en el mismo, población es el conjunto de todas ellas y *una* muestra se obtiene al observar el 1% del censo.

Con mucha frecuencia se consideran como población y muestra, *no* los conjuntos de individuos, *sino* las *medidas de la característica asociadas a esos individuos.*

Ejemplo 2. En un banco de sangre se experimenta un nuevo sistema para aumentar el período de conservación de la misma. En este caso cada bolsa de sangre es un individuo; la población es el conjunto de todas las bolsas del banco y una muestra se obtiene tomando un cierto número de bolsas para su análisis.

Obsérvese que el concepto de individuo no va asociado necesariamente con el de persona, sino que puede ser algo de naturaleza más abstracta.

2. Clasificación de los datos.

Conviene también observar que todos los datos no son del mismo tipo. Cuando los datos, es decir los resultados de las observaciones, no son magnitudes *medibles numéricamente*, sino cualidades o atributos, se dice que se trata de datos **cualitativos**, mientras que en caso contrario se habla de datos **cuantitativos.**

Ejemplo 3. Se observan las causas de muerte de 16 individuos de una cierta población, agrupándolas en las cuatro siguientes: enfermedades cardiovasculares (EC), cáncer (C), accidentes (A) y otras causas (O), habiéndose obtenido los siguientes datos:

EC, EC, A, C, O, A, EC, A, O, C,EC, C, O, C y EC.

Como los resultados no son medibles numéricamente, los datos son cualitativos.

Ejemplo 4. Las notas obtenidas en Matemáticas en una clase de COU han sido:

2, 7, 4, 6, 5, 0, 3, 9, 8, 4, 3, 6, 5 y 8.5.

[1] No se pretende hacer un estudio exhaustivo del tema. Los gráficos se han hecho con el procesador de textos Word.

Se trata de datos cuantitativos.

A su vez los datos cuantitativos se denominan **continuos** si los resultados pueden tomar cualquier valor real dentro de un cierto intervalo, o **discretos**, si sólo pueden tomar ciertos valores particulares.

Ejemplo 5. Del estudio de la estatura de un cierto núcleo de población se han obtenido los siguientes datos:

1.62, 1.78, 1.75, 1.58, 1.83, 1.68 y 1.81metros.

Son datos continuos, pues los individuos de una población pueden tener como estatura cualquier número real en un cierto intervalo.

Ejemplo 6. Del alumbramiento de un conjunto de ratas se ha observado el número de crías, obteniéndose los siguientes valores numéricos:

5, 3, 1, 5, 3, 6, 4, 2, 5, 6, 3, 6, 5, 2, 6, 7 y 3.

Por no ser posibles números no naturales, es evidente que se trata de datos cuantitativos discretos.

Es decir los datos se clasifican:

$$Datos \begin{cases} \text{Cuantitativos} \begin{cases} \text{Continuos} \\ \text{Discretos} \end{cases} \\ \text{Cualitativos} \end{cases}$$

Los datos pueden provenir del estudio de un sólo carácter o propiedad (caso **unidimensional**) o de varios simultáneamente (caso **multidimensional**). En este primer tema estudiaremos sólo los datos unidimensionales.

3. *Características de una muestra representativa*

La observación de un determinado carácter en una población puede realizarse de varias formas:

a) Observación *exhaustiva*: si se considera a la totalidad de los individuos.

b) Observación *parcial*: si se utiliza una muestra.

En los casos en que el tamaño de la población es muy grande el estudio estadístico se realiza sobre muestras.

Para seleccionar una muestra han de respetarse dos tipos de criterios:

- De carácter cuantitativo, es decir ¿cuál es le tamaño adecuado de una muestra?
- De carácter cualitativo, o, lo que es lo mismo, ¿cómo debe elegirse la muestra?

Hay múltiples formas de realizar un *muestreo* estadístico, entre otras:

a) *Muestreo aleatorio simple;* se basa en suponer que todos los elementos de la población tienen asignada la misma probabilidad de ser elegidos. Si se numeran los elementos de la población, una tabla de números aleatorios puede facilitar la tarea de selección.

b) *Muestreo por estratos:* Consiste en clasificar previamente a la población en clases o estratos y de ellos obtener muestras aleatorias.

c) *Muestreo por conglomerados:* es en esencia el mismo sistema que el anterior con la diferencia de que ahora la población se divide en clases con determinados caracteres comunes entre ellas (conglomerados).

Nota. De la obtención de muestras de las que se pueden sacar conclusiones válidas para la totalidad de la población se ocupa la *Teoría de muestras.*

4. *Variables estadísticas. Frecuencias.*

Los caracteres estadísticos de una población son las propiedades o cualidades de los individuos que nos interesa estudiar. Un carácter estadístico divide a la población en clases. A cada una de estas clases se la denomina modalidad.

Cuando el carácter es cuantitativo sus diversas modalidades son medibles, es decir se les puede asignar un **número.**

Definición 1. Se llama **variable estadística** a la aplicación que a cada modalidad le hace corresponder ese número, es decir su medida.

Ejemplo 7. En el ejemplo 6 la variable estadística toma los valores: 1, 2, 3, 4, 5, 6 y 7.

La variable estadística será **discreta** cuando sólo pueda tomar un nº finito de valores y **continua** cuando pueda tomar todos los valores de un cierto intervalo.

Ejemplo 8. La variable estadística del ejemplo 5 es continua y discreta la del ejemplo 6.

Definición 2. Se llama **frecuencia absoluta** al número de individuos que toman un determinado valor de una variable estadística (o una modalidad de un atributo).

Para variables estadísticas (es decir, datos cuantitativos) puede definir:

Definición 3. Se llama frecuencia absoluta **acumulada** de un valor a la suma de las frecuencias absolutas de todos los valores menores o iguales que él.

Ejemplo 9. En el ejemplo 6 la frecuencia absoluta del 5 (tener 5 crías) es 4. La frecuencia absoluta acumulada del 2 es 3.

Definición 4. Se llama **frecuencia relativa** a la razón entre la frecuencia absoluta y el número total de datos o tamaño de la población.

Definición 5. Se llama frecuencia relativa **acumulada** de un valor de una variable estadística a la suma de las frecuencias relativas de todos los valores menores o iguales que él.

Ejemplo 10. La frecuencia relativa del 5 es 4/17 y la relativa acumulada del 2 es 3/17.

5. Representación de datos: Tablas.

Las dos formas más comunes de representar los datos son las tablas y los gráficos.

Tablas estadísticas

Las tablas estadísticas aparecen por todas partes y consisten en masas estructuradas de datos.

Están confeccionadas de tal modo que resultan muy fáciles de leer y de interpretar. Hay que utilizar, fundamentalmente, el sentido común.

Para la construcción de tablas de **datos cuantitativos** pueden tratarse éstos individualmente o agrupándolos en clases

❶ Tratamiento individual

Para variable discreta, o que siendo continua tengamos pocos datos.

Si tenemos una muestra de tamaño N, la tabla se estructura así:

Variable estadística : x_i	Frecuencias absolutas		Frecuencias relativas	
	puntuales	acumuladas	puntuales	acumuladas
x_1	n_1	$N_1= n_1$	$f_1= n_1/N$	$F_1= N_1/N$
x_2	n_2	$N_2= n_1 + n_2$	$f_2=n_2/N$	$F_2=N_2/N$
.....	.....	.....	.....	.....
x_k	n_k	$N_k= n_1 + n_2 +..+ n_k$	$f_k= n_k/N$	$F_k= N_k/N$

$$\sum_{i=1}^{k} n_k = \mathbf{N} \qquad \sum_{i=1}^{k} f_i = 1$$

Ejemplo 11. Las notas de los 20 alumnos de una clase son:

4, 3, 3, 5, 6, 7, 9, 0, 5, 4, 9, 10, 2, 7, 2, 2, 5, 6, 5, 0

Vamos a calcular una tabla:

Variable	Frecuencias absolutas		Frecuencias relativas	
estadística : x_i	puntuales n_i	acumuladas N_i	puntuales f_i	acumuladas F_i
0	2	2	1/10	1/10
2	3	5	3/20	5/20=1/4
3	2	7	1/10	7/20
4	2	9	1/10	9/20
5	5	14	1/4	14/20=7/10
7	3	17	3/20	17/20
9	3	20	3/20	20/20=1

Ejercicio 1. En un Instituto hay matriculados 2200 alumnos que se distribuyen por edades en la forma siguiente: 215 de 14 años, 437 de 15, 421 de 16, 396 de 17, 512 de 18, 124 de 19 y 95 de 20. Formar la tabla de distribución y de frecuencias, que incluya frecuencias acumuladas.

❷ Tratamiento por clases

Cuando en la población o muestra que estudiamos existen muchos valores diferentes, es conveniente, aún a costa de perder algo de información, dividir el intervalo de variación en una serie de subintervalos que cubran el total; a cada uno de ellos se le llama una **clase**, a sus extremos, *extremos de clase,* al punto medio de cada clase, **marca de clase** y a la diferencia entre sus extremos, **amplitud** de la clase.

En estos casos la tabla adopta una estructura como la del cuadro siguiente:

Clases (intervalos)	**Marcas de clase** (m_i)	**Frecuencias absolutas**.....		**Frecuencias relativas**...	
		de clase	**acumuladas**	**de clase**	**acumuladas**

Mientras que en el caso del tratamiento individual la tabla quedaba perfectamente determinada por los posibles valores de los datos, en el de clases está claro que no sucede así, pues hay libertad para elegir el número de clase y los extremos de las mismas.

Los intervalos, en general, deben tener la misma **amplitud.**

Para decidir el nº de clases que se deben tomar conviene tener en cuenta que si éste es excesivo con respecto al número de datos, pueden aparecer irregularidades accidentales provenientes de pocas observaciones en algunas clases. Sin embargo, si se toma el número de clases demasiado reducido se producirá una pérdida importante de información.

Un criterio *orientativo* para decidir cuántas clases se deben tomar lo proporciona la siguiente fórmula empírica debida a Sturges: **$k = 1 + 3.3 \log n$**

Ejemplo 12. Se ha pasado un test de 79 preguntas a 600 personas. El número de respuestas correctas se refleja en la siguiente tabla:

intervalos	m_i	f. abs. puntual	f. abs. acumulada	f. rel. puntual	f. rel. acumulado
[0, 10)	5	40	40	1/15	1/15
[10, 20)	15	60	100	1/10	1/6
[20. 30)	25	75	175	1/8	7/24
[30, 40)	35	90	265	3/20	53/120
[40, 50)	45	105	370	7/40	37/60
[50, 60)	55	85	455	17/120	91/120
[60, 70)	65	80	535	2/15	107/120
[70, 80)	75	65	600	13/120	1
		600		1	

Ejemplo 13. En una Caja de Reclutamiento se toma una muestra de tamaño 30 de los pesos de los mozos correspondientes a un cierto reemplazo, obteniéndose los siguientes datos medidos en kg:

71.9, 63.9, 62.3, 72.5, 78.0, 70.7, 71.4, 60.5, 60.9, 68.2, 88.5, 76.1, 82.1, 63.7, 79.8, 67.5, 50.1, 69.5, 66.1, 47.3, 72.1, 59.8, 93.7, 80.7, 61.2, 64.3, 53.7, 74.7, 96.3, 73.2.

Construir una tabla de frecuencias agrupando los datos en clases de la misma amplitud.

Solución

Si bien no es estrictamente necesario, en general, es conveniente ordenar los datos de menor a mayor. A continuación se presenta la misma muestra ordenada:

47.3, 50.1, 53.7, 59.8, 60.5, 60.9, 61.2, 62.3, 63.7, 63.9, 64.3, 66.1, 67.5, 68.2, 69.5, 70.7, 71.4, 71.9, 72.1, 72.5, 73.2, 74.7, 76.1, 78.0, 79.8, 80.7, 82.1, 88.5, 93.7, 96.3.

Como los valores extremos son 47.3 y 96.3 y el número de clases aconsejado para estos datos es **6** (aplicando la fórmula de Sturges), tomaremos 6 intervalos de amplitud 10, la tabla queda estructurada de la siguiente manera:

clases	Marcas de clase	frecuencias absolutas		Frecuencias relativas	
		de clase	acumuladas	de clase	acumuladas
45 -55	50	3	3	0.1	0.1
55 -65	60	8	11	0.266	0.366
65 -75	70	11	22	0.366	0.733
75 -85	80	5	27	0.166	0.900
85 -95	90	2	29	0.066	0.966
95 -105	100	1	30	0.033	1
		30		0.997≈1	

Intervalos no solapados.

Si los datos recogidos están ya agrupados en intervalos no solapados, como por ejemplo:

Intervalo	n_i
120-139	32
140-149	37
150-159	23
160-169	19

Es conveniente tomar unos intervalos que contengan a éstos, pero sin modificar las frecuencias. Esto es:

Intervalo	n_i
[119,5-139,5)	32
[139,5-149,5)	37
[149,5-159,5)	23
[159,5-169,5)	19

Estos nuevos valores se llaman *límites reales de la clase.*

Observación. Las tablas nos dan una visión, de la característica que se está estudiando, mucho más clara que la que da la muestra, tal cómo se presenta inicialmente.

Ejercicio 2. El número de personas que viven en cada uno de los portales de una gran barriada es: 63, 58, 70, 47, 120, 76, 80, 59, 80, 70, 63, 77, 104, 97, 78, 90, 112, 88, 67, 58, 87, 94, 100, 74, 55, 80, 75, 49, 98, 67, 84, 73, 95, 121, 58, 71, 66, 87, 76, 56, 77, 82, 93, 102, 56, 46, 78, 67, 65, 95, 69, 90, 58, 76, 54, 76, 98, 49, 87, 69, 80, 64, 65, 56, 69, 68, 99, 106.

Construye una tabla de frecuencias[2].

Series cronológicas

Se Llaman series cronológicas a unas tablas estadísticas que recogen observaciones hechas a lo largo del tiempo, normalmente a intervalos iguales. Es por tanto una serie estadística en que la variable independiente es el **tiempo.**

Ejemplo 14. El número de médicos colegiados en España en el período de 1984 - 1992:

[2] Aunque la variable es discreta conviene agruparlos en clases ya que hay un número muy grande de datos.

1984	1985	1986	1987	1988	1989	1990	1991	1992
99730	107503	119890	123543	129897	138967	147978	152943	156748

Ejercicio 3. La producción editorial española de libros de sociología y Estadística, en los años que se indica es:

Años	1991	1992	1993	1994	1995	1996	1997
nº	345	487	589	376	479	652	741

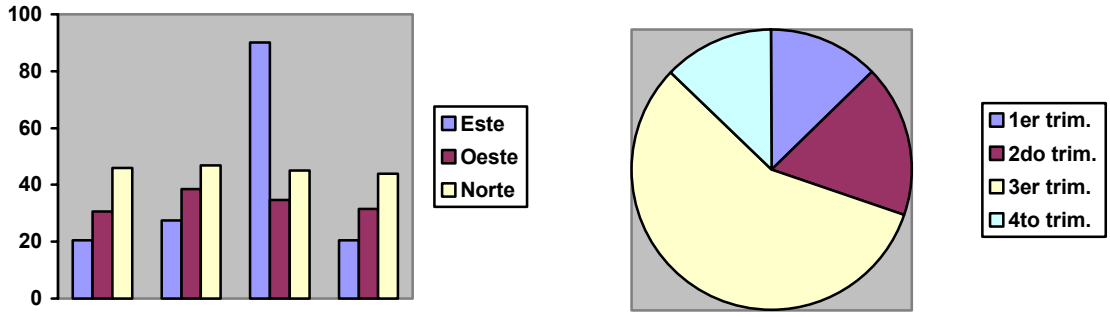
Hacer una tabla de frecuencias absolutas y relativas puntuales. Expresar la relativa en porcentajes.

6. *Representación de datos: Gráficos.*

Los gráficos no son más que traducciones a un dibujo del contenido de las tablas. La finalidad de los gráficos estadísticos es que la información esté al alcance de personas no expertas, que *entre por los ojos*. Los hay de muy diversos tipos pero todos son muy fáciles de interpretar.

❶ Variables cualitativas

Los más usados son los diagramas de rectángulos y los de sectores.



Ejercicio 4. El censo, en miles de cabezas, del ganado en el territorio español, en 1994 fue:

Ganado	Número de cabezas
Bovino	5300
Ovino	18047
Caprino	2601
Porcino	12308
Caballar	264
Mular	153
Asnar	164

Dibujar un diagrama de sectores y otro de rectángulos.

❷ Variables cuantitativas.

Distinguiremos entre variable discreta o continua.

Tratamiento individual

Para el tratamiento individual los medios de representación más utilizados son el gráfico (o diagrama) de barras, el polígono de frecuencias y los gráficos acumulativos.

Diagrama de barras: Se asocia a una tabla de frecuencias ya sea absoluta o relativa.

Sobre un eje horizontal se representan los valores discretos que toman los datos y sobre cada uno de ellos se coloca una barra vertical (o un rectángulo) de longitud (altura) proporcional a la frecuencia.

Ejemplo 15. Vamos a hacer un diagrama de barras de frecuencias absolutas para el ejemplo 6.

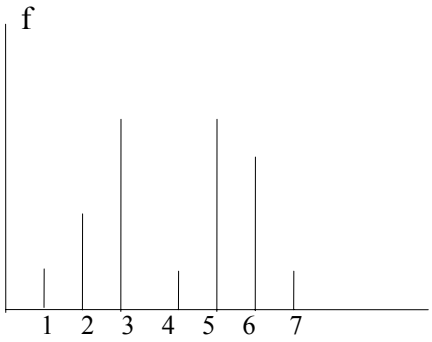
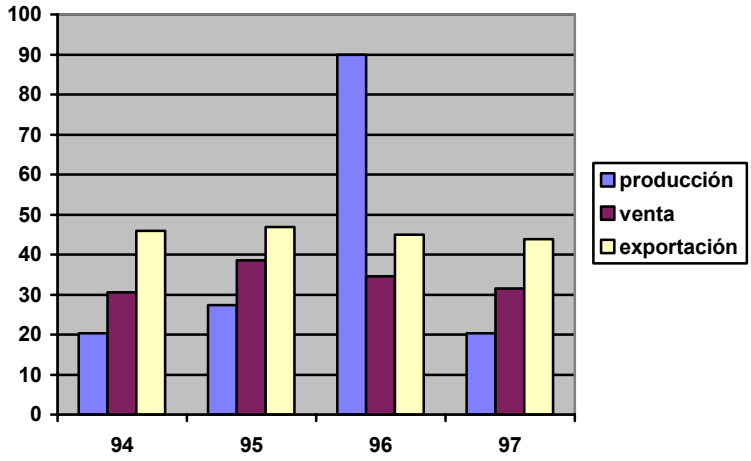


figura 1

En ocasiones se *superponen dos o más diagramas* para comparar datos:
Ejemplo 16: Producción y venta de automóviles en España:



Polígono de frecuencias: Como el anterior se asocia a una tabla de frecuencias.

Se representan en un sistema cartesiano los puntos aislados y luego se unen por medio de segmentos (poligonal). Se usa sobre todo para frecuencias acumuladas (figura 1). También para series cronológicas.

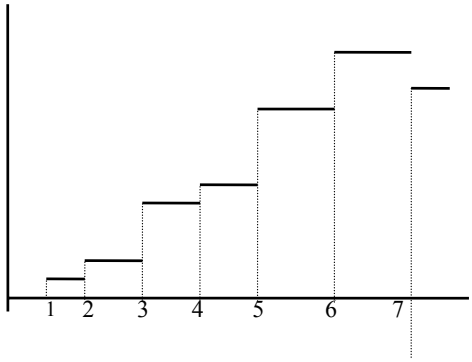
Ejercicio 5. La esperanza de vida al nacimiento ha evolucionado desde 1900, como se refleja en la tabla siguiente:

Años	1900	1910	1920	1930	1940	1950	1960	1970	1980
Varones	33,9	40,9	40,3	48,3	47,1	59,8	67,4	69,6	72,6
Mujeres	35,7	42,6	42,1	51,6	53,2	64,3	72,2	75,1	78,6

Dibujar los polígonos de frecuencias superpuestos para poder compararlos.

Gráficos acumulativos: Se construye a partir del mismo eje horizontal del gráfico de barras, llevando sobre cada valor discreto una vertical de longitud proporcional a la frecuencia acumulada, absoluta o relativa, de dicho valor. Se suele completar el gráfico dándole forma de una escalera de peldaños horizontales.

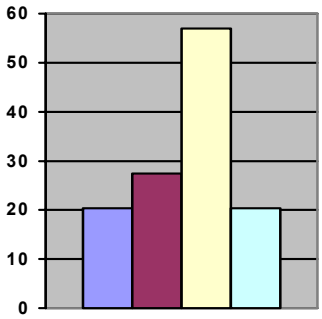
Ejemplo 16. Gráfico de barras acumulativo



Tratamiento por clases

Cuando las variables son continuas, o discretas agrupadas, los gráficos que más se utilizan son: el histograma de frecuencias y los polígonos de frecuencias (absolutas o relativas)

Histogramas de frecuencias. Sobre el eje de abscisas se marcan los extremos de las sucesivas clases y con base en cada clase se dibuja un rectángulo *de altura proporcional a la frecuencia* (absoluta o relativa) observada en dicha clase[3].



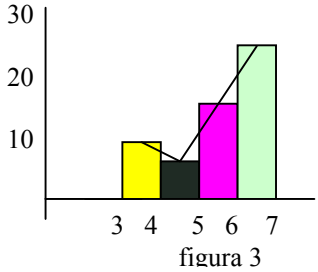


figura 3

Figura 2

Ejercicio 6. En la siguiente tabla se presenta la distribución por edades del número de muertes registradas en España (datos hasta el 30-9-94) a causa del SIDA.

Edad en años	**<3**	**3-9**	**10-12**	**13-14**	**15-19**	**20-24**	**25-29**	**30-34**	**35-39**	**40-49**	**50-59**	**60-69**
Nº de muertes	**411**	**171**	**35**	**31**	**247**	**2888**	**8576**	**7640**	**3292**	**2552**	**909**	**544**

a) Construye la tabla de frecuencias relativas agrupando los datos en las siguientes categorías de edad: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59 y 60-69 años.

b) Representa gráficamente la información obtenida en el apartado a) mediante un histograma,

[3] Cuando se trabaja con clases de amplitudes diferentes es más adecuado el **histograma de frecuencias relativas por unidad de amplitud:** En abscisas se marcan los extremos de las sucesivas clases y con base en cada una de ellas se dibuja un rectángulo de *área proporcional* a la frecuencia relativa.

Polígono de frecuencias. Se asocia a cada clase un punto del plano cartesiano, de abscisa el valor de la marca de clase y de ordenada la frecuencia observada en dicha clase. Uniendo los puntos resulta una línea quebrada que se denomina polígono de frecuencias (figura 3)

Polígono de frecuencias acumuladas.

Partiendo del valor cero en el extremo izquierdo de la primera clase, el polígono acumulado va tomando en los sucesivos extremos derechos de las clases un valor igual a la frecuencia acumulada. Uniendo los puntos así obtenidos resulta el polígono acumulativo de frecuencias (figura 4).

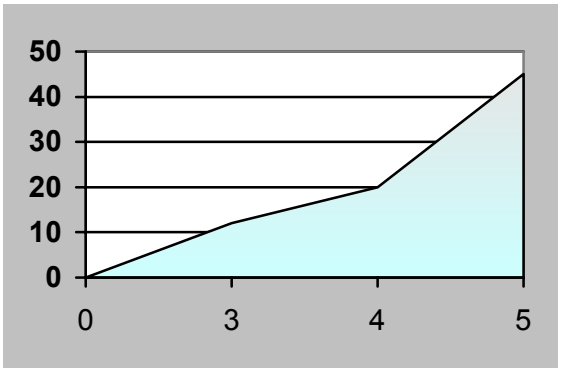


figura 4

Ejercicio 7. Los jugadores de un determinado equipo de baloncesto se clasifican, por altura, según la tabla siguiente:

Altura	1,70-175	1,75-1,80	1,80-185	185-190	1,90-1,95	1,95-2,00
Nº de jugadores	1	3	4	8	5	2

Dibujar el polígono de frecuencias absolutas acumulativo.

7. *Parámetros estadísticos.*

Las tablas estadísticas son una forma organizada de dar toda la información, todos los datos de que disponemos.

Con las gráficas estadísticas se pierde algo de información, pero el mensaje "entra por los ojos", que es lo que se pretende.

En cualquiera de los dos casos, la cantidad de datos que se dan es excesiva para que sea operativo, por ejemplo para la comparación con otras distribuciones.

Por ello se definen los parámetros estadísticos, que nos van a servir para resumir en números aspectos relevantes de la distribución, que puedan dar una idea de la misma o permitir compararlas con otras.

Clases de parámetros estadísticos[4]

♦ Medidas de centralización: media (ya conocida), moda (el valor que se presenta con más frecuencia) y mediana (el valor del individuo que ocuparía el lugar central sí se colocaran ordenados de menor a mayor). Tienen como misión representar con un número a la serie estadística bajo el punto de vista de su posición.

[4] Se amplía el apartado en el Anexo (se darán las fotocopias después de recoger los trabajos)

♦ Medidas de dispersión: rango o recorrido, desviación media, varianza, desviación típica, coeficientes de Pearson... Sirven para medir el grado de alejamiento de los datos respecto de una medida central.

♦ Medidas de posición: cuartiles, deciles, centiles o percentiles. Señalan la situación de algunos valores importantes de la distribución.

En la ordenación que se hizo para la mediana se llaman cuartiles primero, segundo y tercero a los que superan exactamente al 25%, 50% y 75% de los valores. El segundo cuartil es la mediana.

♦ Medidas de asimetría, para señalar si la distribución está sesgada hacia uno u otro lado.

♦ Medidas de apuntamiento o curtosis que indican si la distribución es más o menos puntiaguda.

Para el cálculo práctico de muchos parámetros estadísticos se utilizan tablas que facilitan dichos cálculos (Las **fórmulas** para hallar los parámetros estadísticos más usuales se dan después)

TABLA 1

x_i	n_i	$x_i n_i$	$\lvert x_i - \bar{x}\rvert$	$\lvert x_i - \bar{x}\rvert \, n_i$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 \, n_i$	.....

TABLA 2

x_i	n_i	$x_i n_i$	x_i^2	$x_i^2 \, n_i$	x_i^3	$x_i^3 \, n_i$	.....

Ejemplo 17. Construir la tabla 1 con los datos del ejemplo 11

x_i	n_i	$x_i n_i$	$\lvert x_i - \bar{x}\rvert$	$\lvert x_i - \bar{x}\rvert \, n_i$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 \, n_i$
0	2	0	4,65	9,30	21,62	43,24
2	3	6	2,65	7,95	7,02	21,06
3	2	6	1,65	3,30	2,72	5,44
4	2	8	0,65	1,30	0,42	0,84
5	5	25	0,35	1,75	0,12	0,60
7	3	21	2,35	7,05	5,52	16,56
9	3	27	4,35	13,05	18,92	56,76
	20	93		**la media** es 93/20=4,65		

Ejemplo 18. Construir la tabla 2 con los datos del ejemplo 13.

clases	Marcas de clase x_i	frecuencia n_i	$x_i n_i$	x_i^2	$x_i^2 \, n_i$
45 -55	50	3	150	2500	7500
55 -65	60	8	480	3600	28800
65 -75	70	11	770	4900	53900
75 -85	80	5	400	6400	32000
85 -95	90	2	180	8100	16200
95 -105	100	1	100	10000	10000

Ejemplo 17[5]**.** a) Hallar la media y la varianza de la variable cuyos valores y frecuencias absolutas vienen dadas en la tabla adjunta

Valores de la variable	3	5	4	2	0	8	7
frecuencias	1	3	4	1	3	1	2

[5] Propuesto en selectividad.

b) Representar gráficamente los datos en un diagrama de barras.

Solución a)

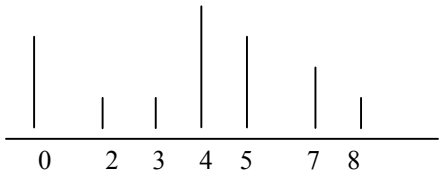
x_i	n_i	$x_i n_i$	x_i^2	$x_i^2 n_i$
0	3	0	0	0
2	1	2	4	4
3	1	3	9	9
4	4	16	16	64
5	3	15	25	75
7	2	14	49	98
8	1	8	64	64
	15	58		314

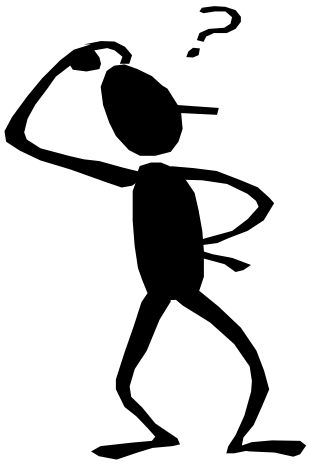
Se tiene : (Ver fórmulas)

$$\overline{x} = \frac{58}{15} = 3{,}87$$

$$\sigma^2 = \frac{314}{15} - (3{,}87)^2 = 5{,}96$$

b)





LOS PARÁMETROS ESTADÍSTICOS) *(ANEXO)*

MEDIA ARITMÉTICA $\bar{x}$

Es el valor

$$\bar{x} = \frac{\sum n_i x_i}{\sum n_i}$$

Si se trabaja con datos agrupados para la fórmula anterior, [1], se toma x_i igual a las marcas de clase.

MODA[6] M_o

Es el valor de la variable de mayor frecuencia.

La distribución puede tener varias modas

Para el caso continuo se habla del intervalo modal (el de mayor frecuencia $\mathbf{n_i}$).

MEDIANA M_E

Es el valor que ocupa el lugar central

CUANTILES

Se llama cuantil de orden α de una distribución al valor de la variable que deja por debajo de él al α % de los elementos de la población.

Los que más se usan son los cuartiles y los centiles o percentiles.

La **mediana coincide con el cuartil segundo Q_2**.

[6] Veremos en los ejercicios resueltos cómo se asigna un valor.

Propiedades

1. Si sumamos una constante a todos los valores la media aumenta en el mismo número.[7] Es decir si $x_i' = x_i + A \Rightarrow \bar{x} = \bar{x} + A$

2. Análogamente, si $x_i' = kx_i$,entonces $\bar{x}' = k\bar{x}$

3. Si $z_i = x_i + y_i \Rightarrow \bar{z} = \bar{x} + \bar{y}$

4. La suma algebraica de las desviaciones respecto de la media es cero, es decir : $\sum (x_i - \bar{x}) = 0$

5. La suma de las desviaciones cuadráticas, $\sum (x_i - a)^2$, es mínima si $a = \bar{x}$.

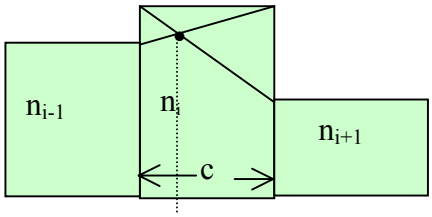
Un *inconveniente* de la media es que los datos con valores extremos pueden influir excesivamente en su evaluación.

Cálculo de la moda[8]

Para calcular la moda, para datos agrupados, se puede usar la fórmula

$$M_o = L_i + \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})} \cdot c$$

[2] (c es la ampitud de la clase modal)



L_i M_o

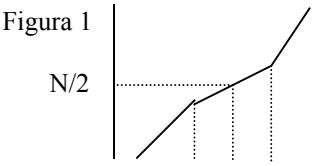
Cálculo de la mediana

Si la distribución tiene un nº impar de datos siempre existe una única mediana y es precisamente el valor central en la relación ordenada de menor a mayor. Si el nº de datos es par se toma como mediana la media de los valores centrales

Para hallar la mediana, cuando los datos estén agrupados, se puede usar el polígono de frecuencias acumuladas (Figura 1)y buscar la abscisa que corresponde a y = N/2 (por interpolación lineal).

$$M_e = L_i + \frac{\frac{N}{2} - N_{i-1}}{n_i} \cdot c$$

Figura 1



L_i M_e

La fórmula anterior [3], nos da dicho valor. En ella:

N_{i-1} es la frecuencia absoluta acumulada hasta llegar a la clase mediana, n_i la frecuencia absoluta de la clase mediana, L_i el límite inferior de la clase mediana y **c** la amplitud de dicha clase.

Los cuartiles y centiles se calculan de forma análoga a la mediana (usando el polígono de frecuencias acumulativo) .

[7] Esta propiedad permite hacer traslaciones de los datos para simplificar los cálculos

RANGO

También llamado recorrido, es la diferencia entre el mayor y el menor de los datos.

DESVIACIÓN MEDIA

Es la media de las desviaciones respecto de la media.

VARIANZA σ^2

Se define como la media de las desviaciones cuadráticas respecto de la media.

DESVIACIÓN TÍPICA

Se define como la ráiz cuadrada de la varianza:

$$\sigma = \sqrt{\sigma^2}.$$

COEFICIENTE DE VARIACIÓN DE PEARSON

Es la razón entre la desviación típica y la media.

No debe usarse para valores muy próximos a cero de la media.

TIPIFICACIÓN

Para comparar dos series datos estadísticos se *normaliza* (o tipifica) la variable

COEFICIENTES DE ASIMETRÍA Y CURTOSIS

Sirven para medir la "simetría" y el "apuntamiento" de las series estadídticas

Si el coeficiente de asimetría es:>0 la curva es sesgada a la derecha, y sí es<0, sesgada a la iizquierda

Cálculo del rango.

Para el caso **continuo**, se toma la diferencia máxima posible entre los límites de intervalos.

Cálculo de la desviación media[9]

Como la suma de las desviaciones respecto de la media da cero lo que se toma son las diferencias en valor absoluto.

La fómula es:

$$D_m = \frac{\sum |x_i - \bar{x}| \cdot n_i}{\sum n_i}$$

Cálculo de la varianza.

De la definición se tiene:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2 n_i}{\sum n_i}$$

Propiedades

1. Si se suma una constante a todos los valores de la variable la desviación típica no varía.

2. Si se multiplican todos los valores de la variable por el mismo número, la desviación típica queda multiplicada por el mismo número

3. Se verifica que

$$\sigma = \sqrt{\frac{\sum x_i^2 \cdot n_i}{\sum n_i} - \bar{x}^2}$$

fórmula que simplifica su cálculo.

Se utiliza para eliminar la influencia de las unidades en el valor de la dispersión y mide la **dispersión relativa** de la muestra..

Por definición se calcula mediante la fórmula: $C_p = \frac{\sigma}{\bar{x}}$

Si**.** X es una variable estadística la variable **normalizada** es:

$$Z = \frac{X - \bar{x}}{\sigma},$$

Se dice que se ha **tipificado** la variable.

Cálculo de los coeficientes de asimetría y apuntamiento.

El coeficiente directo de asimetría se define así:

$$\alpha_3 = \frac{\sum (x_i - \bar{x})^3 \cdot n_i}{\sum n_i} : \sigma^3 :$$

El de apuntamiento : $\alpha_4 = \frac{\sum (x_i - \bar{x})^4 \cdot n_i}{\sum n_i} : \sigma^4$

Observación. Cuando se trabaja con datos agrupados se toma x_i igual a la marca de clase.

EJERCICIOS resueltos

1. a) Completar los datos que faltan en la siguiente tabla estadística, donde f, F y f_r representan, respectivamente, la frecuencia absoluta, acumulada y relativa:

x	f	F	f_r
1	4		0,08
2	4		
3		16	0,16
4	7		0,14
5	5	28	
6		38	
7	7	45	
8			

b) Calcula la media, mediana y moda de esta distribución

Solución

a) La frecuencia relativa de 1 es $0{,}08 = \frac{4}{N}$,

de donde N = 50, lo que nos permite completar la tabla.

x	f	F	f_r
1	4	4	0,08
2	4	8	0,08
3	8	16	0,16
4	7	23	0,14
5	5	28	0,10
6	10	38	0,20
7	7	45	0,14
8	5	50	0,10

b) la media **x** = 4,76; la mediana es 5 y la moda es 6.

2. Observados los alquileres de un conjunto de despachos se ha obtenido:

Alquileres en miles de pesetas	n_i
[0,15)	17
[15,30)	130
[30,45)	180
[45,60)	30
[60,75)	10
[75,90)	5

Calcula la moda y la mediana.

Solución:

Como los datos son agrupados tenemos: para la moda la fórmula:

$$m_0 = L_i + \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i + n_{i+1})}.c =$$

$$= 30 + \frac{50}{200}.15 = 33{,}75$$

Para la mediana usamos el polígono acumulativo de frecuencias

x_i	n_i	N_i
[0,15)	17	17
[15,30)	130	147
[30,45)	180	327
[45,60)	30	357
[60,75)	10	367
[75,90)	5	372

Por interpolación lineal se llega a:

$186\text{-}147 = \frac{327 - 147}{15}(m_e - 30)$, de donde:

m_e=33,25. Comprobar aplicando la fórmula [2]

3. Una empresa petrolera ha tenido unos beneficios anuales de 2000 millones de pesetas. En dicho sector la media es de 1500 millones y la desviación típica de 450 millones. Un comercio tuvo un beneficio de 8 millones. La media del sector es de 6 millones y la desviación típica de 2,5 millones. ¿Cuál tuvo mejor beneficio respecto a su sector?.

Solución

Tipificamos las variables[9]

[9] Al tipificar las variables las medimos en unidades de desviación típica, lo que permite compararlas.

Para la empresa del petróleo: $\frac{2000-1500}{450}=\frac{500}{450}=1{,}1$; para el comercio: $\frac{8-6}{1{,}5}=\frac{2}{1{,}5}=1{,}3$

Luego tuvo mayor beneficio respecto de su sector el comercio, ya que se desvió por encima de la media en 1,3, mientras que la petrolera sólo 1,1.

4. De dos muestras la primera con media 30 y desviación típica 4 y la segunda de media 50 y desviación típica 5, ¿cuál es la que aparece más dispersa?

Solución

Calculamos el coeficiente de variación de Pearson[10], $C_p = \frac{\sigma}{\bar{x}}$ de ambas:

4/30 = 0,13 y 5/50 = 0,1, luego es más dispersa la primera.

© 5. Se quiere hacer una revisión médica a los empleados de una empresa. Se han escogido 3 muestras del mismo número de empleados. De la primera muestra se han revisado 6 personas por hora, de la segunda 5 personas por hora y de la tercera 4 personas por hora. Hallar el promedio de las revisiones.

Solución

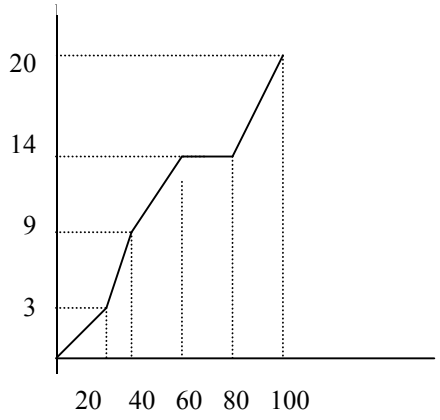
Se trata del cociente entre las magnitudes: número de personas y números de horas. Al calcular los cocientes se ha mantenido fijo el número de personas. Por tanto para hallar el promedio se ha de calcular la media armónica, cuya definición es: $m_h=\frac{N}{\sum_{i=1}\frac{1}{x_i}}$. $m_h = 3/(1/6+1/5+1/4) =4{,}86$

© 6. Un profesor hace 3 exámenes considerando que el 2º es más importante que el 1º y el 3º más importante que el 2º. Para calcular la nota del alumno que ¿ promedio te parece el más indicado?

Solución

La media **ponderada**: que se define así: $m_p=\frac{\sum_{i=1}^{N}x_i p_i}{\sum_{i=1}^{N}p_i}$.

7. Se considera una distribución de datos agrupados en intervalos cuyo polígono de frecuencias acumuladas es el de la figura.



Calcula:

a) Tabla de distribución de frecuencias acumuladas.

b) la media.

Solución

a)

x_i	n_i	N_i
20	3	3
40	6	9
60	5	14
80	0	14
100	6	20

b) $\bar{x}=\frac{20.3+40.6+60.5+80.0+100.6}{20}=60$

[10] Mide la dispersión relativa,

8. En la fabricación de un cierto tipo de clavos, aparecen un cierto nº de ellos defectuosos. Se han estudiado 200 lotes de 500 clavos cada uno obteniendo:

Clavos defectuosos	1	2	3	4	5	6	7	8
nº de lotes	5	15	38	42	49	32	17	2

Calcular la mediana y el percentil 20.

Solución:

Se construye la tabla estadística con las columnas de las frecuencias absolutas acumuladas, siendo ésta:

Nº de piezas	**Nº de lotes (f_a)**	**Frec. absoluta acumulada.**
1	5	5
2	15	20
3	38	58
4	42	100
5	49	149
6	32	181
7	17	198
8	2	200
	200	

El percentil 20

$$N\frac{i}{100} = 200\frac{20}{200} = 40$$

comprendido entre las frecuencias 20 y 58 luego $P_{20} = 3$

Como es par la distribución la mediana es la media de los valores centrales.

Los valores centrales son 4 y 5, por tanto la mediana es 4,5.

9. En el estudio de un cierto fenómeno se obtiene la siguiente tabla:

x_i	7	10	12	16	19	20	21
n_i	6	7	16	17	22	19	17

Calcula los cuartiles Q_1 y Q_3 correspondiente..

Solución

x_i	n_i	N_i
7	6	6
10	7	13
12	16	29
16	17	46
19	22	68
20	19	87
21	17	104

Se tiene: $\sum n_i = 104$, y $\frac{104}{4} = 26$, que corresponde al dato 12; 3.26=78, correspondiente al dato 20. Luego:

Q_1=12, Q_3=20

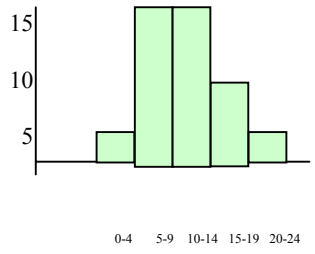
10. La siguiente tabla muestra las frecuencias relativas, f_i, de respuestas correctas contestadas a un test de 24 preguntas por 50 personas.

x	0-4	5-9	10-14	15-19	20-24
f_i	0,1	0,3	0,3	0,2	0,1

Calcular la frecuencia absoluta en cada intervalo y el histograma de frecuencias absolutas.

Solución**.**

x	0-4	5-9	10-14	15-19	20-24
f_i	5	15	15	10	5



11. La distribución de las notas obtenidas por 60 alumnos en un examen, agrupados en intervalos, es:

Notas	[0,1)	[1,2)	[2,3)	[3,4)	[4,5)	[5,6)	[6,7)	[7,8)	[8,9)	[9,10)
nº de alumnos	1	2	5	7	9	15	11	6	3	1

Determine: a) la media; b) la moda; c) el percentil 90.

Solución

Construimos la tabla con las marcas de clase.

x_i	f_i	$x_i f$
0,5	1	0,5
1,5	2	3
2,5	5	12,5
3,5	7	24,5
4,5	9	40,5
5,5	15	82,5
6,5	11	71,5
7,5	6	45
8,5	3	25,5
9,5	1	9,5
	60	315

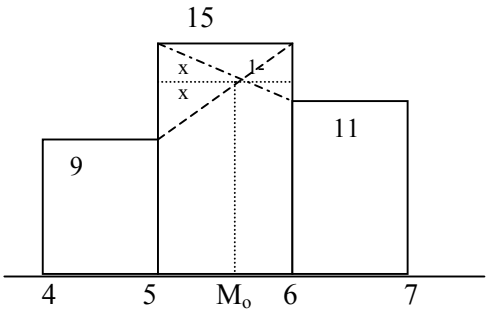
a) La media es $\bar{x} = \frac{315}{60} = 5{,}25$

b) El intervalo modal es [5,6). El valor que se asigna se obtien aplicando la fórmula [3]

$$M_o = 5 + \frac{15-9}{15-9+15-11} \cdot 1 = 5{,}6$$

Veamos de dónde sale esta fórmula.

Como indicamos en la teoría el punto que se asigna es el de la figura



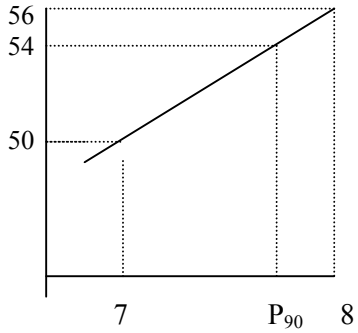
Por semejanza de triángulos :

$\frac{x}{6} = \frac{1-x}{4} \Rightarrow$ x=0,6, luego M_o =5+0,6 =5,6

c) Para el cálculo del percentil 90 necesitamos el plígono de frecuencias acumulada

	n_i	N_i
[0,1)	1	1
[1,2)	2	3
[2,3)	5	8
[3,4)	7	15
[4,5)	9	24
[5,6)	15	39
[6,7)	11	50
[7,8)	6	56
[8,9)	3	59
[9,10)	1	60

$$N\frac{90}{100} = 60\frac{90}{100} = 54$$



Por interpolación lineal.

y –50 = 6 (x-7)

Para y = 54, 4 = 6x – 42, de donde

P_{90}= x = 7,67

12. De un conjunto de 10 datos de media 5 se elimina uno de ellos, de forma que los 9 datos restantes tienen media 4. ¿Qué dato se ha suprimido.

Solución

Por definición de media se tiene: $\frac{x_1 + x_2 + ... + x_{10}}{10} = 5$ y $\frac{x_1 + x_2 + ... + x_9}{9} = 4$, de donde el dato que falta, que hemos llamado x_{10}, vale 10

EJERCICIOS propuestos

1 Los jugadores de un determinado equipo de baloncesto se clasifican según por altura según la tabla siguiente:

altura	1,70-1,75	1,75-1,80	1,80-1,85	1,85-190	1,90-195	1,95-2,00
nº de jugadores	1	3	4	8	5	2

Queremos analizar la variable altura para ello se pide:
a) la media, la moda y la mediana.
b) la desviación típica.
c) los cuartiles 1º y 3º.

2. Los pacientes que acuden a una consulta médica se distribuyen, según la edad, en una tabla:

X(edad)	[0, 10)	[10, 20)	[20,30)	[30, 40)	[40, 50)	[50,60)
N (frecuencia)	7	10	30	18	12	3

Se pide:
a) El histograma de frecuencias.
b) La media, desviación típica, mediana y moda.
c) Porcentaje de pacientes menores de 40 años que acuden a la consulta.

3. a) Calcula la media, moda, mediana y el percentil 70 de la variable del ejercicio 6.

b) Calcular el coeficiente de variación de Pearson ($C_p = \frac{\sigma}{\bar{x}}$)

4. En un Instituto de bachillerato existen dos grupos de COU para la asignatura de Matemáticas II.
Las calificaciones de la 1ª evaluación para una muestra de 10 alumnos de cada grupo fueron las siguientes:

Grupo A	0	1	1	3	5	5	6	8	8	9
Grupo B	2	2	4	4	4	5	5	6	6	8

a) ¿Qué grupo obtuvo mejores resultados?
b) ¿cuál es más homogéneo?
Razone ambas respuestas