¿CURSO BÁSICO DE ESTADÍSTICA DESCRIPTIVA¡

ÍNDICE

CAPÍTULO 1: INTRODUCCIÓN A LA ESTADÍSTICA

Tema 1: Introducción a la estadística

CAPÍTULO 1: INTRODUCCIÓN A LA ESTADÍSTICA

1.1. Introducción a la estadística descriptiva

Tradicionalmente la aplicación del término estadística se ha utilizado en tres ámbitos:

a) Estadística como enumeración de datos.
b) Estadística como descripción, es decir, a través de un análisis de conjuntos coherentes de datos para su posterior comparación y análisis. (ESTADÍSTICA DESCRIPTIVA)
c) Estadística matemática o inferencia, unida a la teoría de de probabilidades. Se encarga de extraer conclusiones a partir de una muestra al total de la población con un pequeño margen de error. (ESTADÍSTICA INDUCTIVA)

Por tanto se podría definir la estadística como ***ñla ciencia que permite estudiar las regularidades o patrones en un conjunto de datos para tomar decisiones racionalesò.***

Todo análisis estadístico requiere seguir una serie de etapas:

1) Definición del problema de estudio y objetivos del mismo.
2) Selección de la información necesaria para realizar el estudio.
3) Recogida de la información que va a depender del presupuesto con el que contemos y de la calidad de los datos exigida.
4) Ordenación y clasificación de la información en tablas y gráficos.
5) Resumen de los datos mediante medidas de posición, dispersión, asimetría y concentración.
6) Análisis estadístico formal obteniendo hipótesis y contrastándolas.
7) Interpretación de resultados y extracción de conclusiones.
8) Extrapolación y predicción.

1.2. Nociones básicas de estadística descriptiva

La estadística descriptiva es la ciencia que analiza series de datos (por ejemplo, edad de una población, peso de los trabajadores de un determinado centro de trabajo, temperatura en los meses de verano, etc) y trata de extraer conclusiones sobre el comportamiento de estos elementos o variables.

Las variables que se observan y analizan pueden ser de dos tipos:

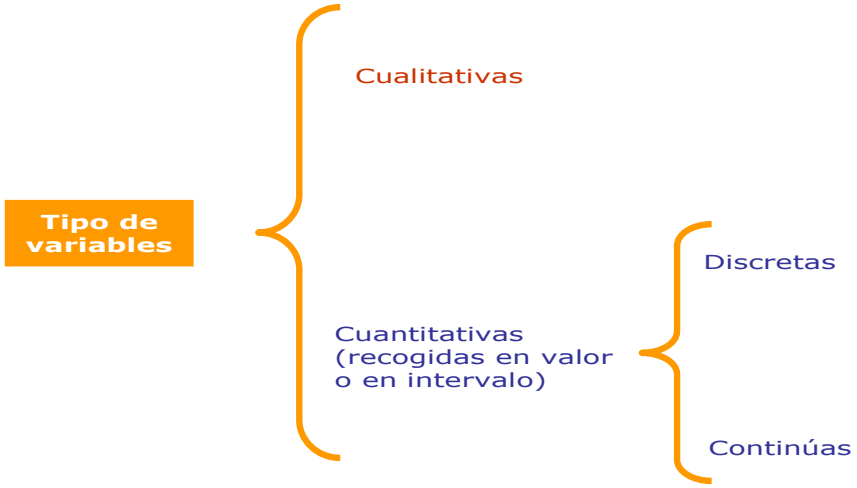
a) Variables cualitativas o atributos: no se pueden medir numéricamente, representan características o atributos de las variables (por ejemplo: nacionalidad, sexo, religión).

b) Variables cuantitativas: tienen valor numérico (edad, altura, precio de un producto, ingresos anuales).

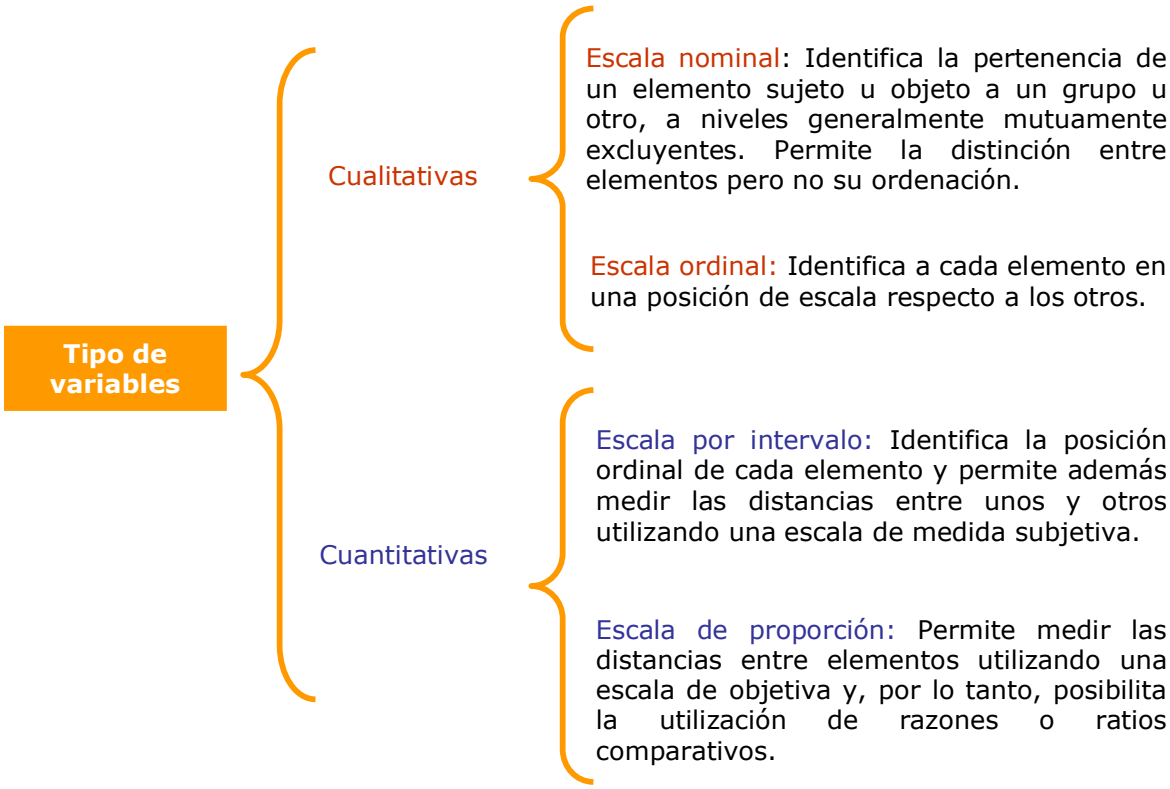
Por su parte, las variables cuantitativas se pueden clasificar atendiendo a los valores que pueden tomar en discretas y continuas:

Discretas**:** sólo pueden tomar valores enteros (1, 2, 8, -4, etc.). Por ejemplo: número de hermanos (puede ser 1, 2, 3....,etc, pero, por ejemplo, nunca podrá ser 3,45).

Continuas: pueden tomar cualquier valor real dentro de un intervalo. Por ejemplo, la velocidad de un vehículo puede ser 80,3 km/h, 94,57 km/h...etc.



Según sea de un tipo u otro la variable podrá medirse de distinta manera, o lo que es lo mismo en la terminología estadística, tendrán distintas escalas de medida.



La información que se recoge de una o varias variables se presenta en tablas que representan la distribución de dichas variables y también se pueden clasificar en:

a) Distribuciones unidimensionales**:** sólo recogen información sobre una característica (por ejemplo: edad de los alumnos/as de una clase).

b) Distribuciones bidimensionales**:** recogen información sobre dos características de cada elemento de la población simultáneamente (por ejemplo: edad y altura de los alumnos/as de una clase) .

c) Distribuciones multidimensionales: recogen información sobre tres o más características de cada elemento (por ejemplo: edad, altura y peso de los alumnos/as de una clase).

1.3. Distribuciones unidimensionales

Después de una primera aproximación a los conceptos estadísticos más importantes y básicos, el analista de información estará preparado para abordar una de las fases más importantes que todo análisis estadístico requiere. Es decir, una vez que hemos definido los objetivos que queremos cubrir con el análisis y obtenido la información relevante, debemos presentarla en tablas y gráficos para conocer mejor el problema que estamos analizando. Las primeras herramientas para **conocer** y por tanto **describir** el problema que estamos analizando nos las proporciona la estadística descriptiva a través de las siguientes maneras de clasificar la información:

1.3.1. Tabulación de la información

Consiste en presentar la información organizada en tablas

❖ Valores de la variable sin agrupar

x_i	n_i	f_i	N_i	F_i
X_1	n_1	N_1/N	N_1	$F_1 = f_1$
X_2	n_2	N_2/N	$N_2 = n_1 + n_2$	$F_2 = f_1 + f_2$
X_n	n_n	n_n/N	$N_n = N$	$F_n = 1$
	$\sum n = N$	$\sum f_i = 1$		

- x_i Valor de la variable
- n_i Frecuencia absoluta: Número de veces que aparece un determinado valor de x
- f_i Frecuencia relativa: Número de veces que aparece un determinado valor de x respecto al total
- N_i Frecuencia absoluta acumulada: Suma de la frecuencia absoluta correspondiente más todas las anteriores

- **F_i** Frecuencia relativa acumulada: Suma de la frecuencia relativa correspondiente más todas las anteriores
- **N** Tamaño de la muestra
- **Distribución**

  Representa los valores de la variable y la frecuencia con que aparecen dichos valores (x_i , n_i)
- **Recorrido**

  Diferencia entre el máximo y el mínimo valor de la variable

Se utiliza este tipo de distribución cuando el número de valores diferentes que toma la variable no es grande, generalmente menos de 15 ó 20 valores (por ejemplo número de hijos).

❖ Datos de la variable agrupados

Cuando el número de valores diferentes que puede tomar la variable es demasiado grande para que resulte fácil presentar la información de manera reducida se utilizan los ó*intervalos*ô(por ejemplo estatura de un grupo de alumnos).

En el caso en que tengamos variables agrupadas en intervalos, introducimos el concepto de marca de clase que es el punto medio del intervalo. En el caso de variables agrupadas en intervalos las frecuencias hacen referencia al intervalo y nunca a valores concretos de dicho intervalo. Puede haber intervalos de la misma o distinta amplitud (c_i). La distribución en este caso viene dada por el extremo inferior (L_{i-1}), el extremo superior (L_i) y la frecuencia (L_{i-1}- L_i, n_i).

Ejemplo 1: Supongamos que queremos hacer un estudio en una clase de universitarios. Entre otras cosas, se les pregunta lo siguiente.

Pregunta 1: Edad del encuestado
Pregunta 2: Ingresos anuales familiares

A la hora de tabular la información la primera pregunta, al referirse a una clase de universitarios donde aproximadamente casi todos los alumnos tienen la misma edad, se hace más interesante recoger la información sin agrupar, es decir, la tabulación quedará de la siguiente manera:

-Cuadro 1-

Edad	**Frecuencia absoluta**	**Frecuencia relativa**	**Frecuencia absoluta acumulada**	**Frecuencia relativa acumulada**
xi	**ni**	**fi**	**Ni**	**Fi**
18	78	0,78	78	0,78
19	15	0,15	93	0,93
20	3	0,03	96	0,96
21	2	0,02	98	0,98
43	1	0,01	99	0,99
45	1	0,01	100	1
N	**100**			

A la hora de tabular la segunda pregunta, y como cada familia puede tener unos ingresos distintos, si representásemos los datos sin agrupar nos podríamos encontrar con una tabla con un dato por individuo, por lo que es más recomendable presentar la información de la variable agrupada en intervalos. De tal manera que la tabla resultante quedará de la siguiente manera:

-Cuadro 2-

Ingresos xi	Frecuencia absoluta ni	Frecuencia relativa fi	Frecuencia absoluta acumulada Ni	Frecuencia relativa acumulada Fi
Menos de 18.000 ú	5	0,05	5	0,05
[18.000ú - 24.000ú	10	0,10	15	0,15
[24.001ú - 30.000ú	10	0,10	25	0,25
[30.001ú - 36.000ú	30	0,30	55	0,55
[36.001ú - 42.000ú	30	0,30	85	0,85
Más de 42.000 ú	15	0,15	100	1
N	**100**			

1.3.2. Representaciones gráficas de la información

Las representaciones gráficas de los datos ofrecen una idea más intuitiva y más fácil de interpretar de un conjunto de datos sometidos a investigación. Por ello las representaciones gráficas se convierten en un medio muy eficaz para el análisis ya que las regularidades se recuerdan con más facilidad cuando se observan gráficamente.

❖ Representaciones gráficas para datos sin agrupar

Diagrama de barras: representa frecuencias sin acumular. Estos gráficos son válidos para datos cuantitativos (de tipo discreto) y cualitativos. En el eje óyô se pueden representar tanto las frecuencias absolutas como relativas

-Gráfico 1- Diagrama de barras

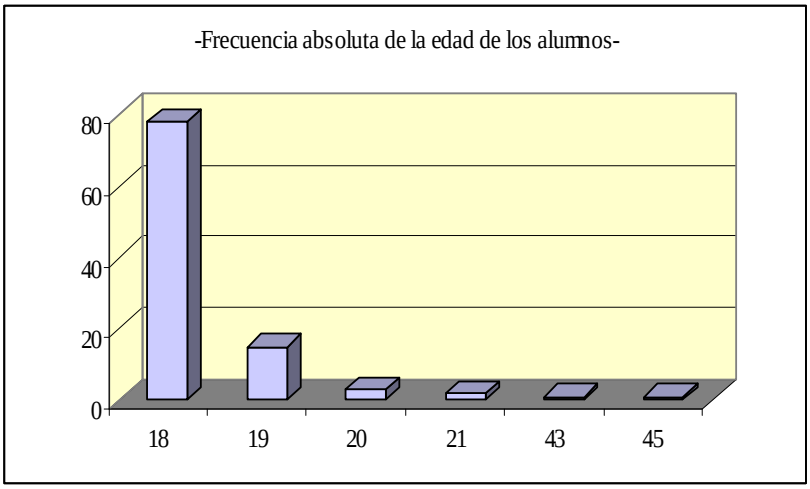
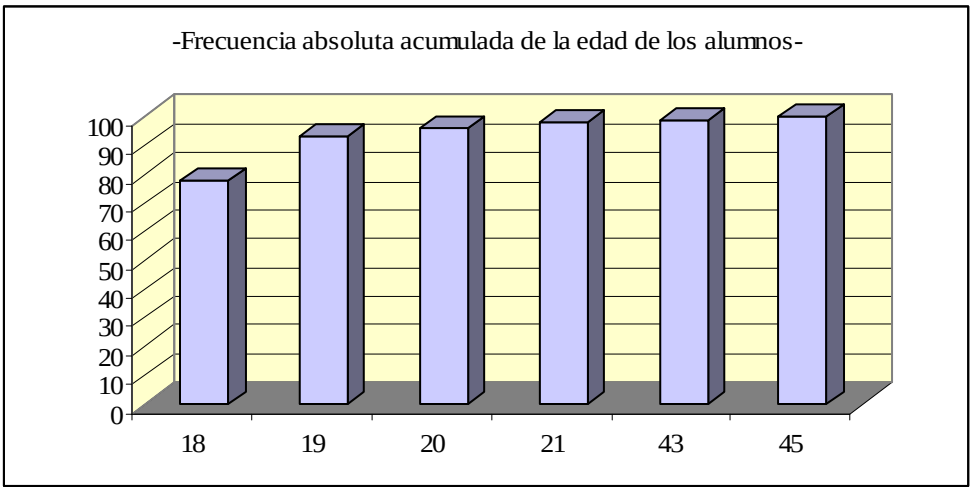


Diagrama de escalera: representa frecuencias acumuladas de un conjunto de datos. Este gráfico puede representar tanto las frecuencias absolutas como relativas.

-Gráfico 2- Diagrama de escalera



❖ Representaciones gráficas para datos agrupados

Histograma: representa frecuencias sin acumular. Este gráfico es válido para datos cuantitativos de tipo continuo o discreto si tiene un gran número de datos. El histograma está formado por rectángulos de área igual o proporcional a la frecuencia observada.

Área = base * altura

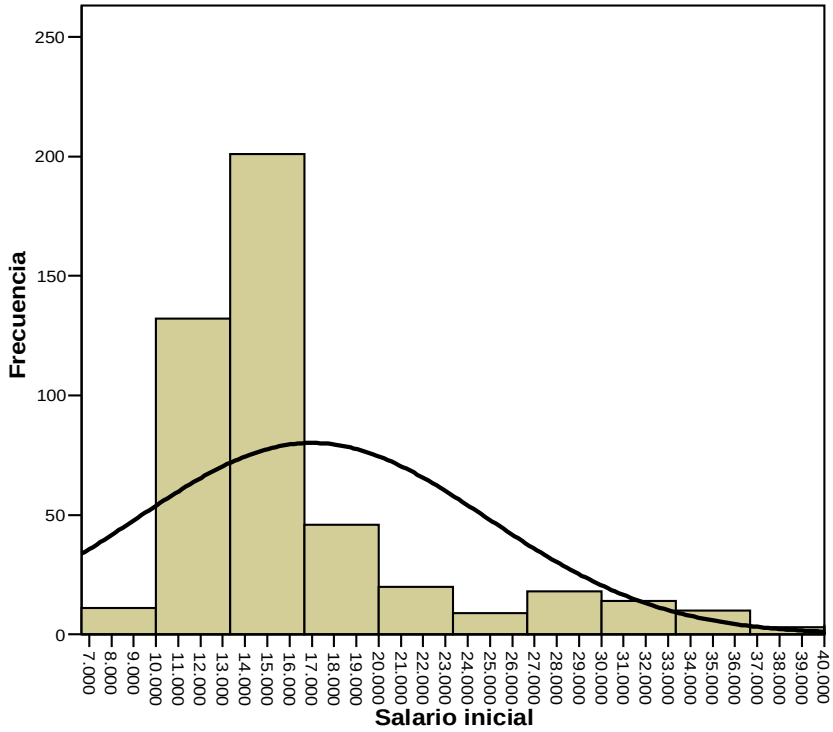
$n_i = c_i$ * altura

altura = densidad de frecuencia = n_i/c_i

Es decir la altura del rectángulo vendrá dada por n_i y será proporcional a dicho valor (también se llama función de densidad). Por tanto en el caso de intervalos iguales, la altura nos está dando una idea de cual es el intervalo más frecuente (aquel cuya barra del histograma sea más alta). En el caso de construir el histograma utilizando f_i la suma total del área del histograma será igual a 1.

A continuación vamos a ver unos ejemplos de histogramas en los dos casos comentados anteriormente, es decir, con intervalos iguales y con intervalos distintos.

Intervalos iguales: **-Gráfico 3- Histograma serie de intervalos iguales**

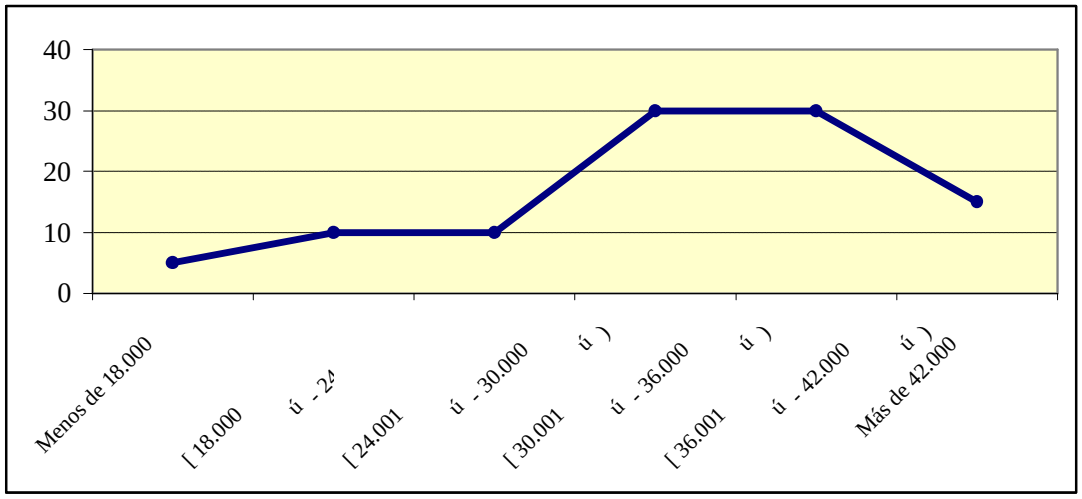


Intervalos distintos **-Gráfico 4- Histograma serie de intervalos distintos**

Polígono de frecuencias acumuladas: representa frecuencias acumuladas. Su construcción se realiza levantando sobre las marcas de clase, localizadas en el eje de abscisas, puntos de altura igual a la frecuencia observada. La unión de estos puntos da lugar a una línea poligonal denominada ópolígono de frecuenciasô

-Gráfico 5- Polígono de frecuencias acumuladas

Ingresos xi	Frecuencia absoluta ni	Marca de clase	Frecuencia relativa fi	Frecuencia absoluta acumulada Ni	Frecuencia relativa acumulada Fi
Menos de 18.000 ú	5	15.000	0,05	5	0,05
[18.000ú - 24.000ú	10	21.000	0,10	15	0,15
[24.001ú - 30.000ú	10	27.000	0,10	25	0,25
[30.001ú - 36.000ú	30	33.000	0,30	55	0,55
[36.001ú - 42.000ú	30	39.000	0,30	85	0,85
Más de 42.000 ú	15	45.000	0,15	100	1
N	**100**				



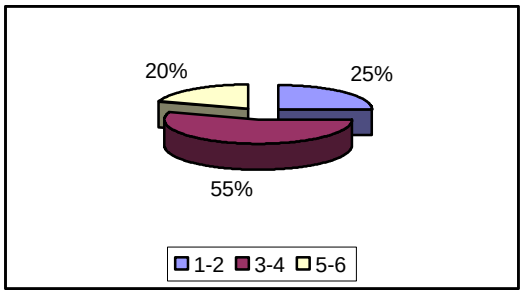
Tanto los histogramas como los polígonos de frecuencia se pueden realizar con frecuencias absolutas o relativas.

Gráficos de sectores

Estos gráficos se basan en un círculo o bien en un semicírculo y consiste en dividir el círculo o semicírculo en sectores cuyas áreas sean proporcionales a cada uno de los términos de la serie. Generalmente se utilizan para representar series de atributos o series cuantitativas presentadas en pocos intervalos.

-Gráfico 6- Gráfico de sectores

Xi	ni
1-2	10
3-4	22
5-6	8
	40

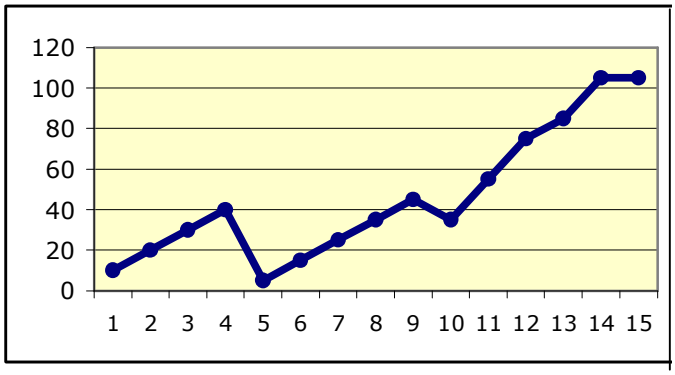


Diagramas Gannt

Estos diagramas nos permiten conocer la evolución de una variable en estudio desde una situación inicial hasta el momento actual. Es un gráfico de mucha utilidad para analizar crecimientos, tendencias, en definitiva, la evolución de la serie en el tiempo.

-Gráfico 7- Diagrama de Gannt

T	Xi
1	10
2	20
3	30
4	40
5	5
6	15
7	25
8	35
9	45
10	35
11	55
12	75
13	85
14	105
15	105



1.3.3. Medidas resumen de las distribuciones de frecuencias

El siguiente paso que debe dar el analista de la información es **resumir** la información que tiene disponible una vez que la ha organizado y representado mediante la tabulación y los gráficos. Para resumir la información dispone de las siguientes medidas que son distintas funciones de la variable:

- Medidas de posición
- Medidas de dispersión
- Medidas de asimetría
- Medidas de apuntamiento o curtosis
- Medidas de concentración

a) Medidas de posición

❖ **Medidas de posición central**

Estas medidas pretenden caracterizar la distribución de la variable/s que estamos analizando por los valores del centro. Es decir, son valores representativos de todos los valores que toma la variable.

➢ Media aritmética:

Representa el centro de gravedad de una distribución y se define como la suma ponderada de los valores de la variable por sus frecuencias relativas y lo denotaremos por $\bar{x}$ y se calcula mediante la expresión:

$$\bar{x} = \sum_{i=1}^{n} x_i * f_i = \sum_{i=1}^{n} \frac{x_i * n_i}{N}$$

dónde $\mathbf{x_i}$ representa el valor de la variable en distribuciones no agrupadas o la marca de clase en distribuciones agrupadas. Es decir, en este último caso, se hace el supuesto que la frecuencia del intervalo está agrupada en la marca de clase.

El inconveniente de la media aritmética es que es muy sensible a los valores extremos de una distribución.

- Media aritmética simple y ponderada

Hay veces donde hay que obtener una media aritmética de variables cuyos valores observados tienen distinta importancia y por tanto se deben ponderar de distinta manera para obtener la media.

En el caso de que la ponderación sea distinta estaremos hablando de una media ponderada y los valores por los cuales se ponderan los distintos valores se llaman pesos o ponderaci ones (w_i)

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$

- Mediana

La mediana es el valor central de la variable, es decir, supuesta la muestra ordenada en orden creciente o decreciente, el valor que divide en dos partes la muestra. Para calcular la mediana debemos tener en cuenta si la variable es discreta o continua.

Cálculo de la mediana en el caso discreto:

Tendremos en cuenta el tamaño de la muestra.

- Si **N es Impar,** hay un término central, el término $X_{\frac{N}{2}+1}$ que será el valor de la mediana.
- Si **N es Par,** hay dos términos centrales, $X_{\frac{N}{2}}, X_{\frac{N}{2}+1}$ la mediana será la media de esos dos valores

Cálculo de la mediana en el caso de datos en intervalo:

Para determinar el valor de la mediana en el caso de tener representada los valores de la variable en intervalos hay que partir de una hipótesis: la variable evoluciona de manera continúa y uniforme dentro del propio intervaloô

En este caso el cálculo de la mediana consta de dos fases, la determinación del intervalo que contiene la mediana y el cálculo de su valor.

1º Para determinar el intervalo en el que se encuentra la mediana se acumulan las frecuencias y el primer intervalo cuya frecuencia acumulada (N_i) sea mayor o igual a $\frac{N}{2}$ es el intervalo que contiene la mediana. Si llamamos L_i y L_{i+1} a los límites del intervalo que contiene la mediana, n_i a la frecuencia ordinaria de dicho intervalo, N_i a la frecuencia acumulada, N_{i-1} la frecuencia acumulada hasta el intervalo anterior y c_i la amplitud del intervalo entonces la fórmula es la siguiente:

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} \times c_i$$

Para calcular la mediana no es preciso que todos los intervalos estén definidos. Del único intervalo que necesitamos conocer la amplitud es del intervalo modal.

- Moda

La moda es el valor de la variable que tenga mayor frecuencia absoluta, la que más se repite, es la única medida de centralización que tiene sentido estudiar en una variable cualitativa, pues no precisa la realización de ningún cálculo.

Por su propia definición, la moda no es única, pues puede haber dos o más valores de la variable que tengan la misma frecuencia siendo esta máxima. En cuyo caso tendremos una distribución bimodal o polimodal según el caso.

Cuando los datos están agrupados en intervalos se puede tomar la marca de clase o realizar una aproximación mediante la siguiente fórmula:

$$Moda = L_1 + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} * c$$

donde :

L_i = límite inferior de la clase modal

c = amplitud del intervalo

La moda se puede utilizar para datos cualitativos pero no tiene porqué situarse en la zona central del gráfico.

❖ Medidas de posición no central

Estas medidas dividen a la población en partes iguales y sirven para clasificar a un individuo dentro de una determinada muestra o población (mismo concepto que la mediana)

- Cuartiles

Medidas de localización que divide a la población en cuatro partes iguales (Q1, Q2 y Q3).

Q1: Valor de la distribución que deja el 75% de los valores por encima
Q2: Valor de la variable que deja el 50% de los valores de la variable por encima (coincide con la mediana)
Q3: Valor de la variable que deja el 25% de los valores de la variable por encima

$$Q_t = L_{i-1} + \frac{\frac{N}{4} - N_{i-1}}{n_i} \times c_i$$

- Deciles

Medidas de localización que divide a la población en diez partes iguales

dk = Decil k-simo es aquel valor de la variable que deja a su izquierda el k·10 % de la distribución.

$$D_t = L_{i-1} + \frac{\frac{N}{10} - N_{i-1}}{n_i} \times c_i$$

- Percentiles

Medidas de localización que divide a la población en cien partes iguales. El primer percentil supera al uno por ciento de los valores y es superado por el noventa y nueve por ciento restante.

Pk = Percentil k-ésimo es aquel valor que deja a su izquierda el K*1% de la distribución

$$P_t = L_{i-1} + \frac{\frac{N}{100} - N_{i-1}}{n_i} \times c_i$$

Reflexiones sobre las medidas de posición central

a) La media, la mediana y la moda coinciden en toda distribución simétrica o normal

b) La media aritmética es la medida de posición que más se utiliza pues normalmente es la que mejor representa los datos, al intervenir todos ellos en su determinación. Por otra parte permite la aplicación del cálculo de probabilidades. Ahora bien, tiene el inconveniente de que en el caso de que exista una gran diferencia entre los valores extremos pierda gran parte de su utilidad al estar afectada por ellos. Por ello en este caso es más conveniente el uso de la mediana.

c) Un promedio puede actuar como medida de tendencia central solamente si existe una cantidad considerable de concentración en la distribución de frecuencias, es decir, que la variación no es demasiado grande.

d) Un promedio sirve como una medida útil de localización para comparar dos o más distribuciones de frecuencias solamente si las que se comparan tienen aproximadamente la misma forma.

b) Medidas de dispersión

Hasta el momento hemos estudiado los valores centrales de la distribución, pero también es importante conocer si los valores en general están cerca o alejados de estos valores centrales, para ver si estos valores son o no son representativos. Es por esto por lo que surge la necesidad de estudiar medidas de dispersión.

Los momentos son valores específicos de la distribución y van íntimamente ligados a las medidas de dispersión y se hallan con la siguiente fórmula:

Momento de orden r $$M_r = \sum_{i=1}^{n} (x_i - o_t)^r \frac{n_i}{N}$$

Momentos respecto al origen (a_1, a_2...) Cuando $O_t = 0$

Momentos respecto a la media (m_1, m_2é) Cuando $O_t = \bar{x}$

El momento de orden r es el promedio de las desviaciones de los valores de una variable, con respecto al origen o a la media, elevadas a la potencia r.

Relación entre momentos:

$m_0 = a_0$

a_1= media

m_1= 0

❖ Medidas de dispersión absolutas

- Rango o recorrido

Es la diferencia entre el mayor valor de una variable y el menor. Depende mucho de los valores extremos y esto puede dar una impresión falsa de la dispersión, por lo que se suele utilizar el rango intercuartílico que es la diferencia entre el tercer y primer cuartel (Q_3 ï Q_1)

- En valor absoluto

Estas medidas tienen las mismas unidades de medidas que la variable a la que hacen referencia (X_i)

$$\sum_{i=1}^{n} ! X_i - promedio \mid n_i / N$$

Con estas medidas de dispersión, sólo se pueden comparar, en principio distribuciones con las mismas unidades de medida.

- Cuadráticas

Las unidades de medida son las de la variable elevada al cuadrado

Varianza ($\mathring{u}^2$, s^2): es la media aritmética de los cuadrados de las desviaciones respecto a la media

$$\sum_{i=1}^{n} (X_i - promedio)^2 n_i / N$$

Al igual que la media, en el caso de que los datos estén agrupados en clases, se tomará la marca de clase como x_i.

El problema de estas medidas es que para comparar variables sí tienen diferentes unidades de medida no se pueden comparar. La solución por tanto es eliminar las unidades de medida y por tanto necesito medidas que no estén afectadas por las unidades.

Para solucionar este inconveniente se hace lo siguiente:

Desviación típica **= $\mathring{u}$= s = +** $\sqrt{\sum_{i=1}^{n} (X_i - promedio)^2 n_i / N}$

Ambas medidas, tanto la varianza como la desviación típica siempre son positivas.

La desviación típica es la mejor medida de dispersión y la más empleada. Cuando las distribuciones de frecuencias se aproximan a una distribución simétrica o normal entonces se verifica una propiedad muy importante que consiste, en que aproximadamente:

⇒ El 68% de los valores de la variable están comprendidos entre $\bar{x} \pm \sigma$

⇒ El 95% de los valores de la variable están comprendidos entre $\bar{x} \pm 2\sigma$

⇒ El 99% de los valores de la variable están comprendidos entre $\bar{x} \pm 3\sigma$

❖ **Medidas de dispersión relativas**

Estas medidas no tienen unidades de medida

Recorrido relativo R_r

Número de veces que el recorrido contiene a la media

$$Rr = \frac{\mathrm{Re}}{\bar{x}}$$

Recorrido semintercuartílico Rκ

$$R_{\delta} = \frac{c_3 - c_1}{c_3 + 1}$$

Coeficiente de apertura Ap

$$Ap = \frac{x_n}{x_1}$$

Coeficiente de variación de Pearson

A veces interesa comparar la variabilidad o dispersión de una población desde dos puntos de vista diferentes e incluso comparar la variabilidad de dos poblaciones o muestras distintas. Cuando no podemos utilizar la desviación típica (porque las distribuciones son muy diferentes o porque las variables presentan distintas unidades de medida) se utiliza el

coeficiente de variación ya que se obtienen medidas homogéneas y por tanto comparables. Aquélla que mayor CV tenga nos indica una mayor dispersión en la distribución

$$CV = \frac{S}{\bar{x}}$$

c) Medidas de asimetría

❖ **Asimetría**

Estas medidas tratan de ver como se distribuye la variable en torno a un eje de simetría. Este eje de simetría se fija en una recta que pase por la media aritmética de la distribución. La asimetría también se utiliza para comparar distribuciones por que se pretende que estas medidas carezcan de unidades.

La medida que da el grado de asimetría de una distribución de datos es el sesgo. Existen varias fórmulas para hallar el sesgo.

- Coeficiente de asimetría**:** cuantía de las desviaciones por encima de la media y la cuantía de las desviaciones por debajo.

Coeficiente de asimetría de Fisher: momento de orden 3 respecto a la media dividido por la desviación típica elevada al cubo. Este coeficiente se calcula para distribuciones acampanadas y en for ma de ójuô

$$g_1 = \frac{\sum_{i=1}^{n} (x - \bar{x})^3 \frac{n_i}{N}}{S^3} = \frac{m_3}{S^3}$$

$g_1 > 0$	Asimétrica positiva (Asimétrica por la izquierda)
$g_1 = 0$	Simétrica
$g_1 < 0$	Asimétrica negativa (Asimétrica por la derecha)

- Coeficiente de asimetría de Pearson: Este coeficiente se calcula para distribuciones en forma de campana.

$$Ap = \frac{\bar{x} - Mo}{S}$$

Ap > 0	Asimétrica por la derecha ($Mo > \bar{x}$)
Ap = 0	Simétrica
Ap < 0	Asimétrica por la izquierda ($Mo < \bar{x}$)

- Coeficiente de asimetría de Bowley

$$Ab = \frac{c_3 + c_1 - 2Me}{c_3 - c_1}$$

Ab > 0	Asimétrica por la derecha
Ab = 0	Simétrica
Ab < 0	Asimétrica por la izquierda

d) Medidas de apuntamiento o curtosis

Con el coeficiente de Curtosis se pretende observar como se distribuyen los valores centrales de nuestra variable. Para ello se compara la distribución que se esté analizando con la distribución normal. Estas medidas nos van a indicar si la distribución tiene una forma de campana más o menos apuntada que la distribución normal.

$$g_2 = \frac{m_4}{s_4} - 3$$

$g_2 > 0$	Leptocúrtica (perfil estirado)
$g_2 = 0$	Mesocúrtica (perfil intermedio)
$g_2 < 0$	Pleticúrtica (perfil achatado)

El apuntamiento tiene como unidad de medida la curtosis. Para medir la curtosis (K) pueden utilizarse los cuartiles y percentiles:

$$k = \frac{Q}{P_{90} - P_{10}}$$

donde:

K= coeficiente de curtosis percentílico

Q= rango semiintercuartílico ($\frac{Q_3 - Q_1}{2}$)

P_{90}= Percentil 90

P_{10}= Percentil 10

e) Medidas de concentración

Estas medidas tienen por finalidad medir la uniformidad del reparto de la frecuencia total de una variable. Por ejemplo, si un grupo de trabajadores, percibieran el mismo salario, la uniformidad de la variable sería absoluta; por el contrario, en un caso hipotético, si la masa total de los salarios fuera percibida por un solo trabajador, entonces la falta de uniformidad sería total- en este caso diremos que la concentración es máxima. Lógicamente, cuando se tiende a la uniformidad absoluta, la media aritmética es perfectamente representativa de la distribución de frecuencias, contrariamente a lo que sucede cuando la concentración es máxima.

Las medias más habituales para la medición de la concentración de una distribución de frecuencia son:

- **Curva de Lorenz: Medida gráfica**

La curva de Lorenz es una representación gráfica que se obtiene de colocar en los ejes de abscisas y coordenadas los porcentajes acumulados del número de observaciones y del total del valor de la variable analizada. Por ser idénticos tanto la escala como el campo de variación de cada uno de los ejes, la curva de Lorenz encaja perfectamente en un cuadrado. Se representa también la diagonal que arranca desde el origen, que se toma como punto de referencia de la curva.

Si la variable analizada fuese totalmente uniforme, la curva de lorenz coincidiría con el dibujo de la diagonal dibujada. En el caso opuesto, la curva de Lorenz estaría formada por los lados inferior y derecho del cuadrado.

❖ **Índice de Gini**

La curva de Lorenz es ilustrativa de la concentración de una distribución. Sin embargo, es conveniente disponer de un indicador que nos permita valor numéricamente dicha concentración y, al mismo tiempo, facilite la comparación entre dos distribuciones. Este es el Índice de Gini o índice de concentración.

El índice de Gini se define como el cociente entre el área rayada entre la curva de Lorenz y la diagonal principal y el área comprendida entre uno de los dos triángulos obtenidos por la diagonal principal.

El Índice de Gini, por tanto, varía entre 0 y 1, aproximándose a 1 cuando la concentración tiende a ser máxima, y a 0 en caso contrario.

Numéricamente, el índice de Gini sólo se puede calcular a través de un sistema de cálculo de áreas.

1.4. Distribuciones bidimensionales

La mayoría de los fenómenos que se estudian en cualquier disciplina están determinados por la observación de distintas variables relativas a dicho fenómeno. Es decir, si queremos estudiar las características de un producto y compararlo con los de la competencia normalmente se recogerá información sobre distintos atributos del producto como por ejemplo tamaño, color, precio, unidades vendidas, etc. Es decir, todas estas características son variables referentes a nuestro producto y por tanto tendremos distribuciones que no serán unidimensionales. En concreto vamos a analizar las distribuciones bidimensionales que consiste en el estudio de dos características a la vez en una muestra.

Los dos caracteres observados no tienen por qué ser de la misma clase, así nos podemos encontr ar con las siguientes situaciones:

Tipos	variables (X, Y)	*Ejemplo*
Variables cualitativas	Categórica / Categórica	Sexo y clase social
Variables cuantitativas	Discreta / Discreta	Número de hermanos y número de hijos.
	Continua / Continua	Peso y altura
	Discreta / Continua	Pulsaciones y temperatura cuerpo
Cualitativa y cuantitativa	Categórica / Discreta	Sexo y número de cigarrillos
	Categórica / Continua	Sexo e ingresos

Otro factor a tener en cuenta es que el número de modalidades distintas que adopta el carácter X no tiene por qué ser el mismo que el que adopta el carácter Y:

$X = \{ x_1, x_2, x_3, ..., x_j \}$; $Y = \{ y_1, y_2, y_3, ..., y_k \}$

a) Tabulación cruzada

En el caso de distribuciones bidimensionales a la hora de organizar los datos y observar la relación entre dos variables se utilizan las tablas de doble entrada. Estas tablas tienen la siguiente estructura:

y / x	Y_1	Y_2	.	Y_j	.	Y_k	$n_{i.}$
X_1	n_{11}	n_{12}		n_{1j}		n_{1k}	$n_{1.}$
X_2		n_{22}		n_{2j}		n_{2k}	$n_{2.}$
.							
X_i				n_{ij}			$n_{i.}$
.							
X_h	n_{h1}	n_{h2}				n_{hk}	$n_{h.}$
n.j	$n._1$	$n._2$		$n._j$		$n._k$	N

- n_{ij} : Frecuencia conjunta
  Número de veces que aparece el valor X_i con Y_j
- $n_{i.:}$ Frecuencia marginal de la variable X
- $n._{j:}$ Frecuencia marginal de la variable y
- N: Suma del total de las observaciones
- $(x_i\, y_j\, n_{ij})$: Distribución conjunta
- $(x_i\, n_{i.})$: Distribución marginal de X
- $(y_j\, n_{j.})$: Distribución marginal de y

En este tipo de representación también podemos representar las frecuencias relativas. Basta con dividir las frecuencias conjuntas entre el número total de observaciones:

$$f_{ij} = \frac{n_{ij}}{N}$$

La suma de las frecuencias absolutas es igual al número de pares observados (N):

$$\sum_{i=1}^{h}\sum_{j=1}^{k} n_{ij} = N$$

La suma de las frecuencias relativas es igual a la unidad:

$$\sum_{i=1}^{h}\sum_{j=1}^{k} f_{ij} = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{n_{ij}}{N} = 1$$

Una tabla de doble entrada también se puede expresar como una *tabla simple o marginal*, de forma que siempre es posible pasar de una a otra según convenga.

Distribuciones Marginales

Si en una tabla de doble entrada utilizamos solamente los valores correspondientes a X, sin que para nada intervengan los valores de la variable y, esta distribución se denomina distribución marginal de la variable X. Análogamente cuando tomamos los valores de la variable y sin tener en cuenta los valores de la variable x estamos ante la distribución marginal de y.

De las frecuencias absolutas marginales se obtienen las frecuencias relativas marginales. Y de igual forma podemos obtener las medias, varianzas y desviaciones típicas marginales.

Frecuencias absolutas marginales

$$\sum_{i} n_{i.} = N \quad ; \quad \sum_{j} n_{.j} = N$$

Frecuencias relativas marginales

$$f_{i.} = \frac{n_{i.}}{N} \quad ; \quad f_{.j} = \frac{n_{.j}}{N}$$

Medias marginales

$$\bar{x} = \frac{\sum_{i=1}^{h} x_i n_{.i}}{N} \quad ; \qquad \bar{y} = \frac{\sum_{j=1}^{k} y_j n_{.i}}{N}$$

Varianzas marginales

$$s_x^2 = \frac{\sum_{i=1}^{h} (x_{i.} - \bar{x})^2 n_{i.}}{N} \quad ; \qquad s_y^2 = \frac{\sum_{j=1}^{k} (y_{.j} - \bar{y})^2 n_{.j}}{N}$$

Desviaciones típicas marginales

$$s_x = \sqrt{\frac{\sum_{i=1}^{h} (x_{i.} - \bar{x})^2 n_{i.}}{N}} \quad ; \qquad s_y = \sqrt{\frac{\sum_{j=1}^{k} (y_{.j} - \bar{y})^2 n_{.j}}{N}}$$

Distribuciones condicionadas

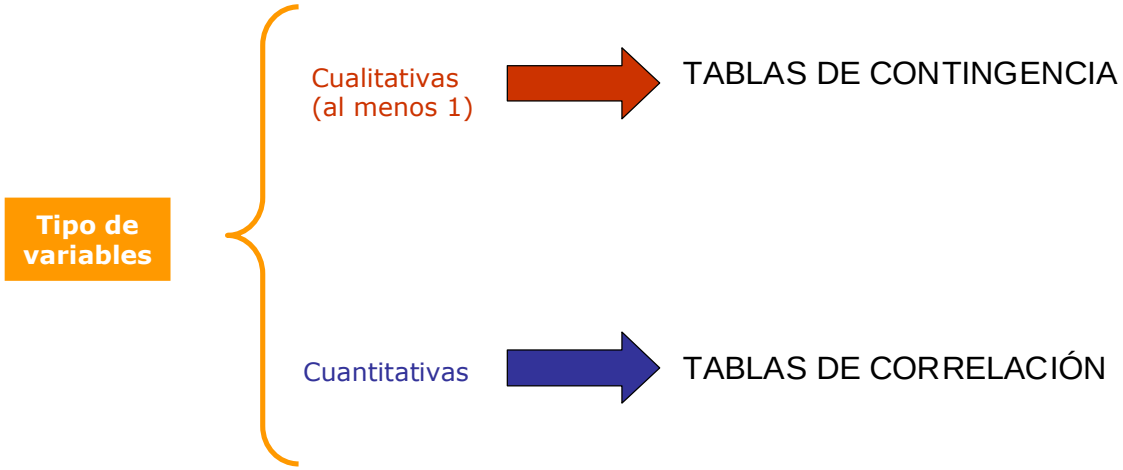
En ocasiones podemos necesitar condicionar los valores de la variable Y a un determinado valor de X o viceversa. Estas distribuciones así obtenidas se denominan: *distribución de la variable Y condicionada a X=x_i o distribución de la variable X condicionada a Y=y_j*

$$\{n(x_i / Y = y_j\} = \{n_{1j}, n_{2j}, ..n_{ij}, n_{hj}\}$$
$$\{n(y_j / X = x_i\} = \{n_{i1}, n_{i2}, ..n_{ij}, n_{ik}\}$$

$$n(x_i / Y = y_j) = \frac{n(x_i / Y = y_j)}{n._j}$$

$$n(y_j / X = x_i) = \frac{n(y_j / X = x_i)}{n_{i.}}$$

Dependiendo del tipo de variables con el que estemos construyendo la tabla hablamos de tablas de contingencia o tablas de correlación:



b) Representación gráfica

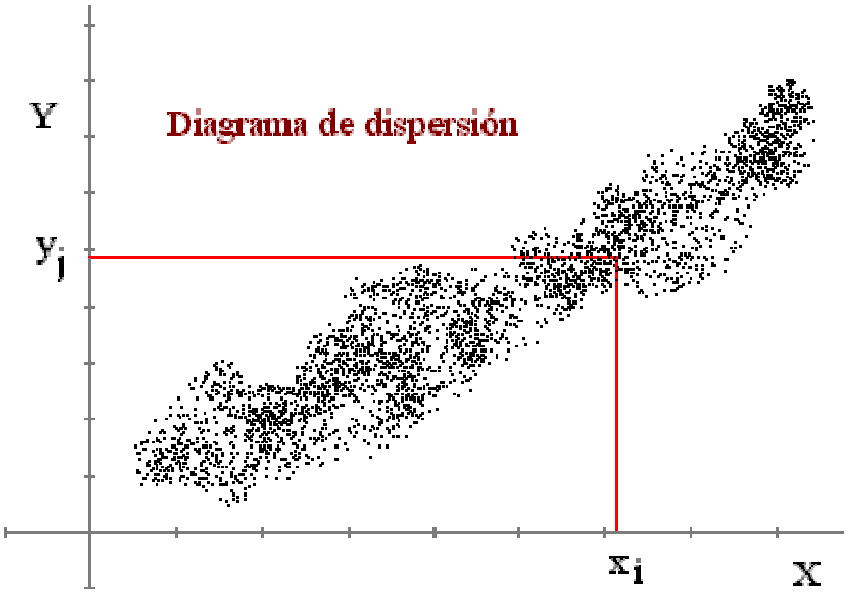
- **DIAGRAMAS DE DISPERSIÓN**

El diagrama de dispersión es la representación sobre unos ejes cartesianos de los distintos valores de la variable (X, Y). En el eje de abscisas representamos los valores de X y en el de ordenadas los valores de Y, de tal forma que cada par viene representado por un punto del plano XY.

En el caso de que las dos variables estén agrupadas en intervalos el diagrama se construye mediante casillas que tienen dentro tantos puntos como el valor de la frecuencia absoluta correspondiente a los intervalos X e Y.

Si las variables que componen el par son una discreta y otra continua se utilizan las marcas de clase, siendo un caso similar al primero

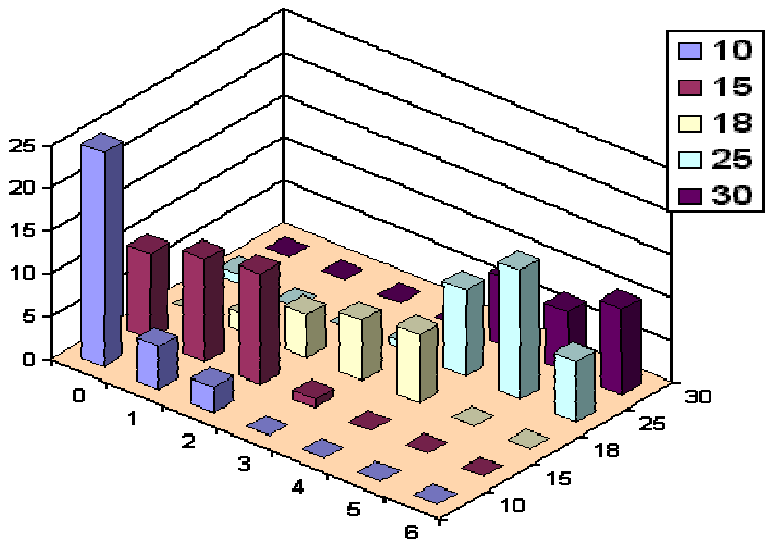
Los diagramas de dispersión también se conocen como nube de puntos.



❖ **DIAGRAMAS DE FRECUENCIAS**

Como en un diagrama de dispersión no puede quedar reflejado las veces que se repite un par o un intervalo, hemos de recurrir a una representación en tres dimensiones de (X, Y). Dos son para la variable bidimensional y una dimensión para expresar las frecuencias.

La figura adjunta representa los datos del ejemplo 1. La variable X toma los valores 10, 15,... y la variable Y los valores 0, 1,2,...; en el eje Z están representadas las frecuencias absolutas del par (X, Y).



c) Medidas de resumen y asociación

A continuación vamos a estudiar las medidas de resumen para el caso de distribuciones bidimensionales con variables cuantitativas.

- Cuando hay pocos datos o están muy agrupados (tablas de 2 o 3 columnas)

Aparece un parámetro nuevo que es la covarianza que es la media aritmética de las desviaciones de cada una de las variables respecto a sus medias respectivas. Es decir, representa la variación conjunta de las dos variables que se estén analizando y pueden tener cualquier signo. Viene representada por la siguiente expresión:

$$S_{xy} = m_{11} = \sum_{i=1}^{n}\sum_{j=1}^{k}(x_i - \bar{x})(y_j - \bar{y})\frac{n_{ij}}{N}$$

Sí S_{xy} es mayor que 0 las dos variables se mueven en el mismo sentido (Δ x Δ y)
Sí S_{xy} es menor que 0 las dos variables se mueven en distinto sentido (Δ x D y)

- Cuando hay muchos datos (tablas de doble entrada)

Puede pasar que se quiera medir la relación que existe entre dos conjuntos de datos, es decir la dependencia o independencia estadística entre dos variables de una distribución bidimensional. Por ejemplo, si se analiza la estatura y el peso de los alumnos de una clase es muy posible que exista relación entre ambas variables: mientras más alto sea el alumno, mayor será su peso. Entonces vamos a obtener la correlación o dependencia entre dos variables. Según sean los diagramas de dispersión podemos establecer los siguientes casos:

- Independencia funcional o correlación nula: cuando no existe ninguna relación entre las variables. ($r = 0$)

- Dependencia funcional o correlación funcional: cuando existe una función tal que todos los valores de la variable la satisfacen (a cada valor de x le corresponde uno solo de y o a la inversa) ($r = \pm 1$)

- Dependencia aleatoria o correlación lineal: cuando los puntos del diagrama se ajustan a una línea recta o a una curva, puede ser positiva o directa, o negativa o inversa ($-1<r<0$ ó $0<r<1$)

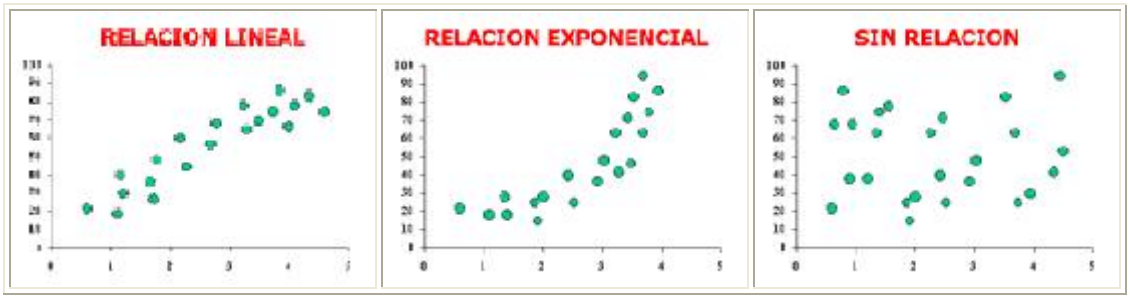
Para establecer estas relaciones tenemos las siguientes medidas

1. Coeficiente de correlación lineal: es una forma de cuantificar más precisa el tipo de correlación que hay entre las dos variables.

2. Regresión: consiste en ajustar lo más posible la nube de puntos de un diagrama de dispersión a una curva. Cuando esta es una recta obtenemos la recta de regresión lineal, cuando es una parábola, regresión parabólica, cuando es una exponencial, regresión exponencial, etc. (lógicamente r debe ser distinto de 0 en todos los casos).

1. **Coeficiente de correlación lineal**

El coeficiente de correlación lineal mide el grado de intensidad de esta posible relación entre las variables. Este coeficiente se aplica cuando la relación que puede existir entre las variables es lineal (es decir, si representáramos en un gráfico los pares de valores de las dos variables la nube de puntos se aproximaría a una recta).



No obstante, puede que exista una relación que no sea lineal, sino exponencial, parabólica, etc. En estos casos, el coeficiente de correlación lineal mediría mal la intensidad de la relación las variables, por lo que convendría utilizar otro tipo de coeficiente más apropiado. Para ver, por tanto, si se puede utilizar el coeficiente de correlación lineal, lo mejor es representar los pares de valores en un gráfico y ver que forma describen.

El **coeficiente de correlación lineal** se calcula aplicando la siguiente fórmula:

$$r = \frac{1/n * \Sigma (x_i - xm) * (y_i - ym)}{\left((1/n * \Sigma (x_i - xm)^2) * (1/n * \Sigma (y_i - ym)^2) \right)^{1/2}}$$

Es decir:

Numerador: se denomina **covarianza**. Se suma el resultado obtenido de todos los pares de valores y este resultado se divide por el tamaño de la muestra.

Denominador: es la raíz cuadrada del producto de las varianzas de "x" y de "y".

Los valores que puede tomar el **coeficiente de correlación "r"** son: $-1 < r < 1$

Si "r" > 0, la correlación lineal es positiva (si sube el valor de una variable sube el de la otra). La correlación es tanto más fuerte cuanto más se apr oxime a 1.

Por ejemplo: altura y peso: los alumnos más altos suelen pesar más.

Si "r" < 0, la correlación lineal es negativa (si sube el valor de una variable disminuye el de la otra). La correlación negativa es tanto más fuerte cuanto más se aproxime a -1.

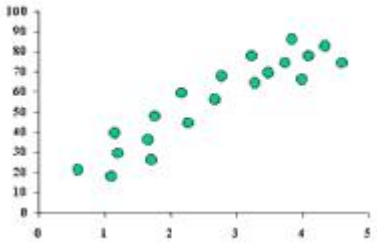
Por ejemplo: peso y velocidad: los alumnos más gordos suelen correr menos.

Si "r" = 0, no existe correlación lineal entre las variables. Aunque podría existir otro tipo de correlación (parabólica, exponencial, etc.)

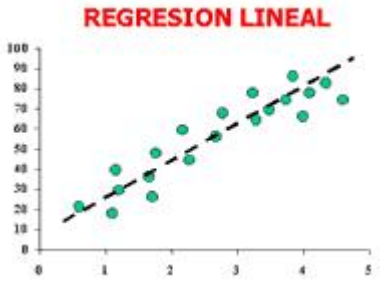
De todos modos, aunque el valor de "r" fuera próximo a 1 o -1, tampoco esto querría decir obligatoriamente que existe una relación de causa-efecto entre las dos variables, ya que este resultado podría haberse debido al puro azar.

2. **Regresión lineal**

Si representamos en un gráfico los pares de valores de una distribución bidimensional: la variable "x" en el eje horizontal o eje de abcisa, y la variable "y" en el eje vertical, o eje de ordenada. Vemos que la nube de puntos sigue una tendencia lineal:



El **coeficiente de correlación lineal** nos permite determinar si, efectivamente, existe relación entre las dos variables. Una vez que se concluye que sí existe relación, la **regresión** nos permite definir la recta que mejor se ajusta a esta nube de puntos.



Una recta viene definida por la siguiente fórmula:

$$y = a + bx$$

Donde "y" sería la variable dependiente, es decir, aquella que viene definida a partir de la otra variable "x" (variable independiente). Para definir la recta hay que determinar los valores de los parámetros "a" y "b":

El **parámetro "a"** es el valor que toma la variable dependiente "y", cuando la variable independiente "x" vale 0, y es el punto donde la recta cruza el eje vertical. El **parámetro "b"** determina la pendiente de la recta, su grado de inclinación. La **regresión lineal** nos permite calcular el valor de estos dos parámetros, definiendo la recta que mejor se ajusta a esta nube de puntos.

El **parámetro "b"** viene determinado por la siguiente fórmula:

$$b = \frac{1/n * \Sigma (x_i - xm) * (y_i - ym)}{1/n * \Sigma (x_i - xm)^2}$$

Es la covarianza de las dos variables, dividida por la varianza de la variable "x".

El **parámetro "a"** viene determinado por:

$$a = y_m - (b * x_m)$$

Es la media de la variable "y", menos la media de la variable "x" multiplicada por el parámetro "b" que hemos calculado.

Ejercicios:

1. El curso MEB de ESCP-EAP obtiene las siguientes puntuaciones en un test de habilidad mental

43 40 41 50 62 35 38 50 32 35 36 45 58 30 33 45 49 46 47 51 64 36 39 51 51 48 49 53 66 38 41 43 71 45 46 55 68 40 53 55 52 49 50 59 62 45 48 60 32 30 40 39 42 30 35 40 38 36 46 45 68 50 69 69

Se pide:

a) Formar una distribución de frecuencias con 14 intervalos
b) Hacer la representación gráfica del polígono de frecuencias
c) Hacer la representación gráfica del histograma
d) Hacer la representación gráfica de las frecuencias acumuladas relativas

2. Las puntuaciones obtenidas por un grupo de alumnos de Primaria en un test de habilidad sicomotora, ha dado las puntuaciones siguientes:

x	xi	ni	ni xi	Ni	fi	Fi
60-63	61,5	2	123	2	1%	1%
56-59	57,5	12	690	14	7%	8%
52-55	53,5	18	963	32	10%	18%
48-51	49,5	36	1782	68	21%	39%
44-47	45,5	38	1729	106	22%	61%
40-43	41,5	20	830	126	11%	72%
36-39	37,5	18	675	144	10%	82%
32-35	33,5	10	335	154	6%	88%
28-31	29,5	8	236	162	5%	93%
24-27	25,5	6	153	168	3%	96%
20-23	21,5	4	86	172	2%	98%
16-19	17,5	2	35	174	1%	99%
12-15	13,5	0	0	174	0%	99%
8-11	9,5	0	0	174	0%	99%
4-7	5,5	1	5,5	175	1%	100%
N		**175**	**7642,5**		**100%**	

Se pide:

a) Hallar la media
b) Hallar la mediana
c) Hallar Q_1 y Q_3
d) Hallar los percentiles 18 y 84
e) Hallar la moda

3. El primer curso de sociología ha obtenido una nota media al final del curso de 5,7 de un total de 110 alumnos. El segundo curso una nota media de 6,6 de un total de 60 alumnos y el curso tercero una nota media de 5,1 de un total de 48 alumnos. ¿Cuál es la nota media de los tres cursos?

4. Dada la tabla siguiente:

15 19 31 30 23 76 13 35 27 32 77 35 24 18 18 15 45 76 81 27 76 23 18 18
75 15 69 14 75 63 29 19 81 15 29 81 45 17 15 41 18 31

Se pide:

a) El recorrido de los datos
b) Agrupar los datos en 8 intervalos
c) Calcular la amplitud de los intervalos
d) La desviación media
e) La desviación típica
f) Los cuatro momentos
g) La asimetría
h) La curtosis

5. Dada la siguiente distribución calcular todos los coeficientes de asimetría y explicar el significado de su valor :

Puntuaciones	ni
80-84	8
75-79	7
70-74	5
65-69	6
60-64	12
55-59	6
50-54	9
45-49	4
40-44	5
	62