ESTADÍSTICA DESCRIPTIVA

INTRODUCCIÓN

1. CONCEPTO DE ESTADÍSTICA

La estadística es la rama de las matemáticas que estudia los fenómenos colectivos recogiendo, ordenando y clasificando y simplificando los datos para analizarlos e interpretarlos.

2. PROCESO HISTÓRICO Y SITUACIÓN ACTUAL.LA ESTADÍSTICA Y LAS OTRAS CIENCIAS

El inicio de la estadística se relaciona con la idea de hacer un censo, un recuento. Se tiene constancia de censos hechos hacia el año 2238 a.C. en China, por el emperador Tao. En Egipto se hacían censos de las riquezas y en Roma del número de habitantes y de su distribución por el territorio.

Hacia 1540 aparece en Alemania la obra de Sebastian Hünster "Cosmografía universal", en donde se recogían prácticamente todos los conocimientos de estadística anteriores.

Los censos tomaron un nuevo valor y se empezaron a utilizar para interpretar fenómenos sociales. En el año 1662 el inglés Graunt publicó un estudio sobre los datos de mortalidad de Londres que se puede considerar como el primer trabajo fundamentado sobre la población. Nacía así la estadística.

La estadística, que en principio se definió como "la ciencia de las cosas que pertenecen al estado", se entiende hoy como una ciencia que estudia los fenómenos sociales, económicos y físicos a partir de datos numéricos.

Con Darwin se produjo una gran revolución en el mundo de la biología. Los estudios de Darwin sobre la evolución de poblaciones animales tenían un atractivo especial desde el punto de vista de la estadística. Los organismos que estuvieran mejor adaptados sobrevivirían mayor tiempo y dejarían un mayor número de descendientes, y por eso, hay una correlación entre las características genéticas transmisibles y el grado de supervivencia.
El primero en acudir a métodos estadísticos para contrastar las teorías de Darwin fue su primo Galton (1822-1911). Galton introdujo exhaustivos estudios de estadística que tuvieron una importancia vital posteriormente e influyeron poderosamente en otros científicos. Fue Weldon quien siguió los trabajos de Galton y quien buscó la colaboración del filósofo y matemático Pearson (1857-1936). La colaboración de estos dos hombres impulsó de una manera decisiva la estadística actual.

En el laboratorio de Pearson se hicieron estudios de estadística que desembocaron en ramas científicas que hoy son independientes. Así por ejemplo, la econometría que nos permite resolver problemas estadísticos dentro de la economía; la investigación operativa o la simulación.

Actualmente, los ordenadores permiten trabajar con masas enormes de datos, de manera que podemos hacer predicciones y resolver problemas estadísticos que hace unos años eran costosos y largos.

Un problema típico de estadística es establecer el grado de relación que hay entre dos variables. Por ejemplo, una empresa e puede preguntar en qué grado un aumento de los gastos en publicidad hacen aumentar las ventas de un producto. Muchos fenómenos en que intervienen dos variables se estudian hoy con los conceptos de correlación y regresión lineal.

3. PARTES DE LA ESTADÍSTICA

Generalmente la estadística se divide en dos partes:

- **Estadística DESCRIPTIVA**, que recoge, ordena y clasifica los datos construyendo además, tablas y gráficos que simplifiquen las observaciones y puedan facilitarnos el estudio. Sólo se realizan deducciones directas de los datos sin hacer predicciones que incluyan el cálculo de probabilidades.
- **Estadística INFERENCIAL**, que extrae conclusiones y realiza predicciones a partir de los resultados descriptivos que se han sacado de una muestra. En esta parte se hace un uso constante del cálculo de probabilidades.

I. DISTRIBUCIONES ESTADÍSTICAS UNIDIMENSIONALES

1. DEFINICIONES BÁSICAS

1.1. POBLACIÓN, INDIVIDUO Y MUESTRA

- **Población:** es el conjunto del cual se realiza el estudio. La población debe estar bien determinada al inicio del estudio y puede estar formada por personas, cosas, períodos temporales….

- **Individuo:** es cada una de las unidades elementales de la población.

La imposibilidad de estudiar todos los elementos de la población, ya sea porque es muy numerosa o porque no disponemos de herramientas o tiempo suficiente para hacerlo, nos obliga a trabajar con muestras:

- **Muestra:** es el subconjunto de individuos de la pobalción del cual se hace el estudio para generalizar las conclusiones a toda la población.

➢ ***El problema de las muestras representativas***

El trabajar con una muestra nos facilita o posibilita la realización del estudio, pero conlleva algunos problemas:

- *¿Es fiable extrapolar los resultados obtenidos a toda la población?*
- *¿Qué criterios hay que seguir a la hora de elegir una muestra que permita hacer esta extrapolación?*

Elegir una muestra adecuada hará que la recogida de datos y la realización del estudio sea sencillo y que las conclusiones extraídas sean fiables.

Una muestra se dice **representativa** cuando reproduce con la máxima exactitud posible las características de toda la población, al menos en los aspectos que nos interesa estudiar.
Para elegir muestras representativas las dos técnicas más usuales son:

- **Muestreo aleatorio simple:** donde todos los elementos de la población tienen las misma probabilidad de ser escogidos para formar parte de la muestra(simple sorteo).

- **Muestreo aleatorio estratificado:** donde basándonos en otros datos estadísticos más exhaustivos, los elementos de la muestra son elegidos por estratos previamente definidos y dentro de cada estrato todos los elementos que lo forman tienen la misma probabilidad de ser escogidos.

1.2. CARACTERES Y MODALIDADES. CLASES

1.2.1. Definiciones

El **carácter:** es el aspecto, el fenómeno o la cualidad que s estudia en cada uno de los individuos de la muestra. Por ejemplo: el color de ojos, la altura, el sexo, la profesión…
Cada una de las diferencias que se pueden establecer dentro de un carácter reciben el nombre de **modalidad** si no se expresan numéricmente y **valor** si se expresan numéricamente. Así del carácter color de ojos algunas modalidades son verdes, marrones, azules…, del sexo, las modalidades son hombre, mujer, de la altura hay infinitos valores.

Cuando el número de modalidades (o valores) es muy grande, por ejemplo los valores del carácter altura, hay que reducir las opciones que tiene este carácter a unas cuantas **clases.**
Así, por ejemplo las modalidades de carreras universitarias se pueden agrupar en las clases ciencias y letras. Los ~~valores~~ del carácter altura se suele dar en forma de intervalos que serán las clases: de 100 cm. a 105cm., de 105 cm. a 110 cm….., más de 210 cm.

A la hora de escoger las clases podemos hacerlo en función de nuestras necesidades e intereses, pero siempre hemos de asegurarnos de que:

- quede claramente definida
- sea exhaustiva, que no quede ninguna modalidad ~~(o valor)~~ sin ser de ninguna clase
- sea exclusiva, ninguna modalidad puede ser simultáneamente de dos clases.

1.2.2. Tipos de caracteres

Hay dos tipos de caracteres:

- **Cualitativos:** si las diferentes modalidades no se pueden determinar numéricamente. Por ejemplo el sexo, el estado civil, la profesión…
  A su vez, los caracteres cualitativos se clasifican en:
  - *Nominales:* cuando no son ordenables. Es el caso del sexo, el color de ojos…
  - *Ordinales:* cuando se pueden ordenar. Es el caso de las calificaciones: I, S, B, N, E.

- **Cuantitativos:** si los valores de los caracteres de los individuos se pueden medir numéricamente. Generalmente este tipo de carácter recibe el nombre de variable estadística, y se representan por una letra X, o Y.

1.3. VARIABLES ESTADÍSTICAS.TIPOS.

Hay que distinguir entre:

- **Variable estadística discreta:** cuando los valores que toma la variable son aislados. El caso más normal es aquel en que los valores que toma la variable son números enteros, por ejemplo:nº de hijos, nº de trabajadores en una empresa…

- **Variable estadística continua:** cuando la variable estadística puede tomar cualquier valor de un determinado intervalo. En general todas las variables relacionadas con el espacio el tiempo o la masa, los son. Más concretamente, la altura, el peso…

2. FRECUENCIAS

2.1. CONCEPTO DE FRECUENCIA

Una vez recogidos los datos lo primero que hay que observar es el número de repeticiones de cada valor.
Sea X una variable estadística y sea x_i uno de los valores que puede tomar, se denomina

- **frecuencia absoluta del valor x_i** el número de veces que se presenta este. Se representa por **f_i**.

- **frecuencia relativa del valor x_i** el cociente entre la frecuencia absoluta f_i y el número total de datos N. Lo representaremos por **h_i.**

$$h_i = \frac{f_i}{N}$$

2.2. FRECUENCIA ACUMULADA

A veces, es útil conocer cuántas observaciones (o qué proporción de observaciones) se encuentran por debajo o por encima de cada valor. Esta información viene dada por lo que denominamos frecuencias acumuladas.
Dada X variable estadística, y xi uno de los valores que puede tomar, se denomina

- **frecuencia absoluta acumulada de x_i** a la suma de las frecuencias absolutas de los valores anteriores o iguales a x_i. La representamos por **F_i**:

$$F_i = \sum_{j=1}^{i} f_j = f_1 + f_2 + \ldots + f_i$$

- **frecuencia relativa acumulada de x_i** a la suma de las frecuencias relativas de los valores anteriores o iguales a x_i. La representaremos por **H_i**:

$$H_i = \sum_{j=1}^{i} h_j = h_1 + h_2 + \ldots + h_i$$

2.3. TABLAS DE FRECUENCIAS

Toda esta información se recoge en una **tabla estadística de frecuencias**.
Cuando el número de valores que puede tomar la variable estadística es muy grande, estos valores se agrupan en intervalos llamados **clases**. (ejemplo 2) A la hora de hacer cálculos tomaremos como representante de cada intervalo su punto medio, valor que se denomina **marca de clase**.

Ejemplo 1.
El número de hermanos de un grupo de 25 alumnos es este:
1, 3, 1, 1, 1, 0, 1, 2, 0, 1, 0, 0, 1, 0, 2, 3, 1, 0, 0, 2, 1, 1, 1, 4, 1

La tabla estadística que corresponde a estos datos es la siguiente:

x_i	f_i	F_i	h_i	H_i
0	7	7	0.28	0.28
1	12	19	0.48	0.76
2	3	22	0.12	0.88
3	2	24	0.08	0.96
4	1	25	0.04	1.00

Ejemplo 2.
Las alturas de un grupo de 25 alumnos son las siguientes:
171, 177, 176, 160, 186, 170, 169, 178, 163, 166, 173, 162, 167, 179, 157, 173, 168, 167, 172, 168, 173, 182, 176, 183, 188

La tabla estadística que corresponde a estos datos es la siguiente (los datos se han agrupado en intervalos de amplitud 5 cm):

Clases	f_i	F_i	h_i	H_i
[155,160)	1	1	0.04	0.04
[160,165)	3	4	0.12	0.16
[166,170)	6	10	0.24	0.40
[170,175)	6	16	0.24	0.64
[175,180)	5	21	0.20	0.84
[180,185)	2	23	0.08	0.92
[185,190)	2	25	0.08	1
	25		1	

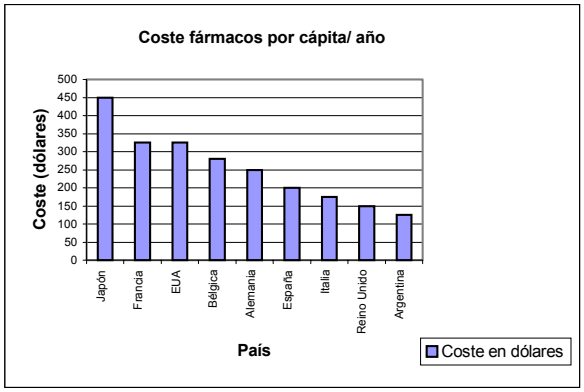
3. GRÁFICOS ESTADÍSTICOS

- **DIAGRAMA DE BARRAS**

Se utiliza para variables estadísticas discretas con datos no agrupados. Constan de dos semiejes perpendiculares que se cortan en un punto. Sobre uno de los semiejes(normalmente el horizontal), situamos las modalidades de la variable y sobre el otro eje las frecuencias. Finalmente, sobre cada modalidad se sitúa una barra de altura proporcional a la frecuencia de la modalidad.

EJEMPLO
Representación en diagrama de barras del coste de los fármacos per cápita/año que se recogen en la tabla siguiente.

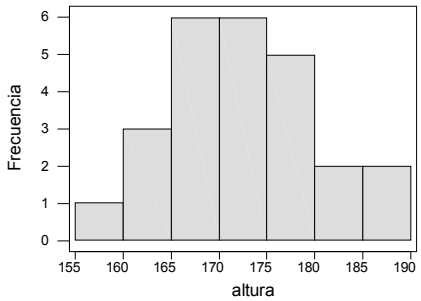
País	**Coste en dólares**
Japón	450
Francia	325
EUA	325
Bélgica	280
Alemania	250
España	200
Italia	175
Reino Unido	150
Argentina	125



- **HISTOGRAMA**

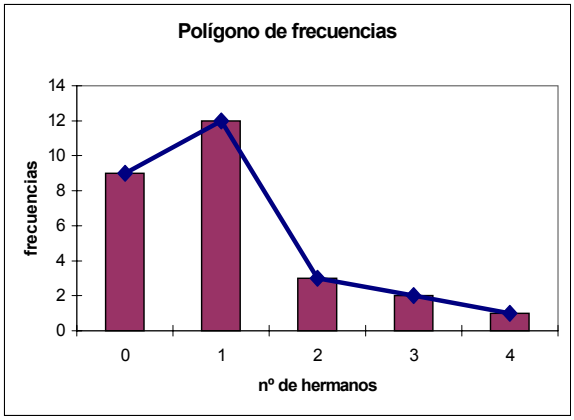
Se utiliza en distribuciones de frecuencias con los datos agrupados en intervalos de clase.
Se construye situando en el eje de abscisas los intervalos y en el de ordenadas las frecuencias.
Después se dibuja para cada clase, un rectángulo con base la amplitud del intervalo y área proporcional a la frecuencia.

EJEMPLO
Histograma correspondiente al *ejemplo 2 de la sección 2.3.*



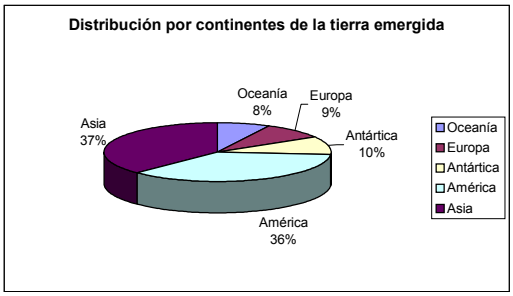
- **POLÍGONO DE FRECUENCIAS**
  En el diagrama de barras o en el histograma, si en vez de dibujar barras, unimos con una poligonal los puntos que tienen como abscisa el punto medio de las barras y como ordenada la frecuencia correspondiente a esa barra, obtenemos el polígono de frecuencias.

  EJEMPLO
  Polígono de frecuencias correspondiente al *ejemplo 1 de la sección2.3.*



- **DIAGRAMA DE SECTORES**
  Es una representación en un círculo que se construye dibujando sobre éste un sector circular de área proporcional a la frecuencia absoluta de cada modalidad de la variable estadística.
  Se utilizan cuando se quiere comparar entre sí las frecuencias correspondientes a las diferentes modalidades de un carácter.

  EJEMPLO



- **CARTOGRAMA**
  Son gráficos que se hacen sobre un mapa, donde, mediante diferentes coloraciones o dibujos, se indica la distribución de un carácter estadístico.

- **PICTOGRAMA**
  Son gráficos que se caracterizan por la utilización de dibujos que hacen referencia al carácter o la población estudiados. La medida, el número de dibujos, o simplemente la sensación visual ha de ser proporcional a la frecuencia de cada modalidad. Se busca más el impacto visual que la exactitud en los datos. Hay que tener, pues, cuidado de que la información no quede desfigurada.

4. PARÁMETROS ESTADÍSTICOS

Son valores que se calculan para resumir la información recogida.

4.1. MEDIDAS (O PARÁMETROS) DE CENTRALIZACIÓN

Se denominan así porque indican alrededor de qué valores se sitúan los datos de la distribución estadística.

4.1.1. La media

La media es la media aritmética de todos los valores de la variable estadística. Se representa por x, y se calcula mediante la expresión:
(Recordar que f_i es la frecuencia absoluta de x_i.)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{N} = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_n f_n}{N}$$

En el caso de datos agrupados en clases, tomaremos como xi la marca de clase de cada intervalo.

4.2. MEDIDAS (O PARÁMETROS) DE DISPERSIÓN

Las medidas de posición no nos dan toda la inforamción que necesitamos sobre la variable estadística. La completaremos con las medidas de dispersión, que nos dan una idea de cómo de agrupados o dispersados tenemos los datos.

4.2.1. La varianza

Es la media aritmética de los cuadrados de las desviaciones de cada valor respecto a la media. Se representa por S_X^2. Su expresión es:

$$S_X^2 = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^2}{N} = \frac{f_1 (x_1 - \bar{x}) + f_2 (x_2 - \bar{x}) + \ldots + f_n (x_n - \bar{x})}{N}$$

Desarrollando la fórmula anterior podemos obtener la siguiente, que a menudo permite simplificar los cálculos:

$$S_X^2 = \frac{\sum_{i=1}^{N} f_i x_i^2}{N} - \bar{x}^2 = \frac{f_1 x_1^2 + f_2 x_2^2 + \ldots f_n x_n^2}{N} - \bar{x}^2$$

¿Cómo pasar de una a otra expresión?

$$S_X^2 = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^2}{N} = \frac{\sum_{i=1}^{n} f_i (x_i^2 - 2x_i \bar{x} + \bar{x}^2)}{N} = \frac{\sum f_i x_i^2}{N} - 2\bar{x}\frac{\sum f_i x_i}{N} + \bar{x}^2 \frac{\sum f_i}{N} =$$

$$= \frac{\sum f_i x_i^2}{N} - 2\bar{x} + \bar{x}^2 = \frac{\sum f_i x_i^2}{N} - \bar{x}^2$$

4.2.2. La desviación típica o desviación estándar

Es la raíz cuadrada positiva de la varianza. Tiene la ventaja sobre la varianza de que se expresa en las mismas unidades que los valores de la variable, mientras que la varianza se expresa con estas unidades al cuadrado, por tanto es más fácil d interpretar la información que nos da la desviación típica.

- ***EJEMPLO (de cálculo de media, varianaza y desviación típica)***
  Calcular la media, la varianza y la desviación típica de la distribución que viene dada por la siguiente tabla:

Número de calzado	*Número de alumnos*
35	*4*
36	*15*
37	*17*
38	*20*
40	*10*
42	*4*

Resolución:
Para hacer los cálculos, iremos creando una tabla con columnas que nos permitirán calcular la media y la varianza. Para calcular la varianza utilizaremos la 2ª expresión que es la que nos permite realizar los cálculos de manera más rápida.

x_i	f_i	$x_i \cdot f_i$	x_i^2	$x_i^2 \cdot f_i$
35	*4*	*140*	*1225*	*4900*
36	*15*	*540*	*1269*	*19440*
37	*17*	*629*	*1369*	*23273*
38	*20*	*760*	*1444*	*28880*
40	*10*	*400*	*1600*	*16000*
42	*4*	*168*	*1764*	*7056*
	70	*2637*		*99549*

La media será:

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i \cdot f_i}{N} = \frac{2637}{70} = 37.67$$

El numerador es la suma de los valores de la 3ª columna de la tabla.

A partir de los valores de la quinta columna podemos calcular la varianza:

$$S_x^2 = \frac{\sum_{i=1}^{6} x_i^2 \cdot f_i}{N} - \bar{x}^2 = \frac{99549}{70} - 37.67^2 = 1422.1286 - 1419.0289 = 3.0997$$

Y extrayendo la raíz cuadrada de la varianza halamos la desviación típica, que vale 1.76.

4.4. PROPIEDADES (DE LOS PARÁMETROS ESTADÍSTICOS)

a) Si se suma una constante a todos los valores de la variable, su media aumenta su valor en dicha constante, mientras que la varianza y la desviación típica no varían.
b) Si se multiplican todos los valores de la variable por una misma constante positiva, la media y la desviación típica quedan multiplicadas por dicha constante, mientras que la varianza queda multiplicada por el cuadrado de la constante.

Estas dos propiedades nos permiten trabajar con valores más sencillos de la variable, en caso de que estos sean muy grandes o muy pequeños.

4.5.INTERPRETACIÓN CONJUNTA DE MEDIA Y DESVIACIÓN TÍPICA

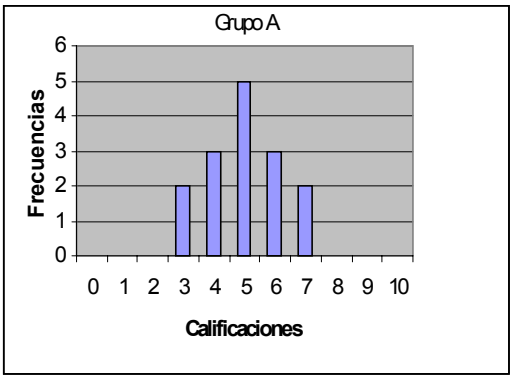
En toda distribución estadística el estudio conjunto de la media y la desviación típica, permiten hacerse una idea clara de cómo se distribuyen los datos. Con la media se sabe cuál es el valor medio de los datos que se tienen de la población, pero este dato es insuficiente.

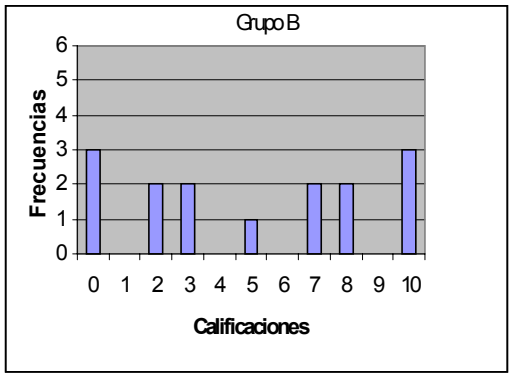
EJEMPLO
Hemos ordenado las calificaciones de un examen de historia de dos grupos de estudiantes:

Grupo A*: 3, 3, 4, 4, 4, 5, 5, 5, 5, 5, 5, 6, 6, 6, 7, 7*
Grupo B*: 0 ,0, 0, 2, 2, 3, 3, 5, 7, 7, 8, 8, 10, 10, 10*

Haciendo los cálculos pertinentes, se ve que la media en ambos grupos es de 5 puntos.

Si se hace una representación de sus distribuciones de frecuencia, se obtienen los gráficos que siguen:





La desviación típica para el grupo A vale S_A=1,21. Eso quiere decir que las calificaciones están próximas a 5.
La desviación típica ara el grupo B vale S_B=3,67. Por tanto, las calificaciones están muy alejadas de 5.

En el primer grupo no hay notas brillantes, pero tampoco notas muy malas. En cambio, las notas del grupo B son muy dispersas.

II. DISTRIBUCIONES ESTADÍSTCAS BIDIMENSIONALES

Hasta ahora hemos considerado experimentos en los que se observaba un único carácter de cada individuo, pero es muy frecuente que se observen dos caracteres de los individuos de la muestra, dando así lugar a los experimentos bidimensionales o bivariantes.

1. CONCEPTO DE VARIABLE ESTADÍSTICA BIDIMENSIONAL O BIVARIANTE

Llamamos variable estadística bidimensional a la que se obtiene al estudiar un fenómeno respecto de dos variables estadísticas unidimensionales.
Se representa por el par (X,Y), donde X, e Y son variables unidimensionales.

2. DATOS DE UNA VARIABLE ESTADÍSTICA BIDIMENSIONAL

Los datos de las variables bidimensionales son los pares (x_1,y_1), (x_2, y_2),…,(x_n,y_n), donde x_1, x_2, …, x_n son los valores de la variable X, e y_1, y_2, …, y_n los valores de la variable Y.

3. REPRESENTACIÓN GRÁFICA: NUBE DE PUNTOS O DIAGRAMA DE DISPERSIÓN

Una nube de puntos es la representación en un sistema de ejes cartesianos de los ares de datos (x_i,y_i) de la variable estadística bidimensional.
La forma del diagrama de dispersión nos permite intuir si existe o no relación entre las dos variables estudiadas, si esta relación es directa o inversa, y la intensidad de esta relación.
EJEMPLO y a ver si puedes comentar un poco ahí lo de la relación.

4. TABLAS DE FRECUENCIAS

4.1. TABLAS SIMPLES

Una tabla de frecuencias simple es la que recoge en filas o columnas las frecuencias de los valores (x_i,y_i) de la variable.

EJEMPLO
Las calificaciones de 40 alumnos en matemáticas y física han sido las siguientes:

X=calificación en matemáticas	3	4	5	6	6	7	7	8	10
Y=calificación en física	2	5	5	6	7	6	7	9	10
Nº de alumnos	4	6	12	4	5	4	2	1	2

La frecuencia absoluta de cada par (X,Y) viene dado por el nº de alumnos que han obtenido las calificaciones correspondientes.
Ejercicio:hacer el diagrama de dispersión correspondiente

4.2. TABLAS DE 'CONTINGENCIA' O 'CRUZADAS' O 'DE DOBLE ENTRADA'

Una tabla de contingencia es la que recoge las frecuencias de los valores (x_i,y_i) de la variable.
Se suelen utilizar cuando hay muchos datos o cuando tenemos los datos agrupados en clases.
En una tabla de doble entrada se incluye siempre una fila y una columna de totales que reciben el nombre de distribuciones marginales.

EJEMPLO
Se han clasificado 50 familias de acuerdo con el número de hijos (X), y de hijas (Y) y se han obtenido los resultados siguientes:

X / Y	0	1	2	3	4	5	6	
0	2	-	4	3	1	-	-	10
1	3	-	9	-	-	3	-	15
2	-	6	-	6	-	-	1	13
3	1	4	-	-	2	1	-	8
4	-	-	2	-	1	-	-	3
5	-	-	-	1	-	-	-	1
	6	10	15	10	4	4	1	50

Se usan cuando se trabaja con muchos datos, o bien cuando los valores están agrupados en clases. Se pueden transformar en tablas simples.

5.PARÁMETROS ESTADÍSTICOS

5.1. PARÁMETROS (O MEDIDAS) DE CENTRALIZACIÓN (MARGINALES) Y PARÁMETROS (O MEDIDAS) DE DIPERSIÓN (MARGINALES)

Se calculan para cada una de las variables X, e Y.
De nuevo tendremos la media, como medida de centralización y la varianza y desviación típica como medidas de dispersión, pero esta vez para cada una de las variables X e Y.

Consideremos una variable estadística bidimensional (X,Y) cuya distribución de frecuencias viene dada por la siguiente tabla:

Variable X x_i	Variable Y y_i	Frec. absoluta f_i
x_1	y_1	f_1
x_2	y_2	f_2
x_3	y_3	f_3
…	…	…
…	…	…
…	…	…
x_n	y_n	f_n
		$\Sigma f_i=N$

Las fórmulas para la media y la varianza quedarán de la siguiente forma:

	Variable X	*Variable Y*
Media	$\overline{X}=\dfrac{\sum x_i \cdot f_i}{N}$	$\overline{Y}=\dfrac{\sum y_i \cdot f_i}{N}$
Varianza	$S_X^2=\dfrac{\sum f_i \cdot (x_i-\overline{x})^2}{N}$	$S_Y^2=\dfrac{\sum f_i \cdot (y_i-\overline{y})^2}{N}$
	$S_X^2=\dfrac{\sum x_i^2 \cdot f_i}{N}-\overline{x}^2$	$S_Y^2=\dfrac{\sum y_i^2 \cdot f_i}{N}-\overline{y}^2$

La raíz cuadrada positiva de las varianzas se denomina desviación típica, y se representa por S_X y S_Y.

5.2. PARÁMETROS (O MEDIDAS) DE CORRELACIÓN

5.2.1. Introducción

Lo que más interesa en el estudio de una variable estadística bidimensional (X,Y) es la relación entre las variables unidimensionales X, e Y, de manera que podamos hacer previsiones posteriores. Estas relaciones no se deben interpretar como si fuesen dependencias funcionales, sino que las hemos de entender como tendencias en la asociación de valores.
Se habla de dependencia o relación estadística cuando el diagrama de dispersión tiende a aproximarse a la representación de una función. Si los valores de una variable no influyen en los de la otra, diremos que las variables X e Y son independientes.
Esto lo estudiaremos en profundidad más adelante.

5.2.2. Definición de covarianza. Organización de datos en una tabla para un ejemplo de cálculo

Se denomina **covarianza** de una variable bidimensional (X, Y) a la media aritmética de los productos de las desviaciones de X e Y respecto de sus medias. Se representa por $\mathbf{S_{XY}}$ y se calcula con las siguientes fórmulas:

Más adelante veremos la interpretación de la covarianza, así como su interpretación según el signo.

$$S_{XY} = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})(y_i - \bar{y})}{N} \qquad \text{o} \qquad S_{XY} = \frac{\sum_{i=1}^{n} f_i x_i y_i}{N} - \bar{x}\bar{y}$$

¿Cómo pasar de una expresión a otra?

$$S_{XY} = \frac{\sum f_i \cdot (x_i - \bar{x}) \cdot (y_i - \bar{y})}{N} = \frac{\sum f_i \cdot (x_i y_i - \bar{y} x_i - \bar{x} y_i + \bar{x}\bar{y})}{N} =$$

$$= \frac{\sum f_i x_i y_i}{N} - \bar{y}\frac{\sum f_i x_i}{N} - \bar{x}\frac{\sum f_i y_i}{N} + \bar{x}\bar{y} = \frac{\sum f_i x_i y_i}{N} - \bar{y}\bar{x} - \bar{x}\bar{y} + \bar{x}\bar{y} = \frac{\sum f_i x_i y_i}{N} - \bar{x}\bar{y}$$

- **Organización de los datos para el cálculo de la covarianza (también de medias y varianzas marginales):**

x_i	y_i	f_i	$x_i \cdot f_i$	$x_i^2 \cdot f_i$	$y_i \cdot f_i$	$y_i^2 \cdot f_i$	$x_i \cdot y_i \cdot f_i$
x_1	y_1	f_1	$x_1 \cdot f_1$	$x_1^2 \cdot f_1$	$y_1 \cdot f_1$	$y_1^2 \cdot f_1$	$x_1 \cdot y_1 \cdot f_1$
x_2	y_2	f_2	$x_2 \cdot f_2$	$x_2^2 \cdot f_2$	$y_2 \cdot f_2$	$y_2^2 \cdot f_2$	$x_2 \cdot y_2 \cdot f_2$
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
x_n	y_n	f_n	$x_n \cdot f_n$	$x_n^2 \cdot f_n$	$y_n \cdot f_n$	$y_n^2 \cdot f_n$	$x_n \cdot y_n \cdot f_n$
		$\Sigma f_i = N$	$\Sigma x_i \cdot f_i$	$\Sigma x_i^2 \cdot f_i$	$\Sigma y_i \cdot f_i$	$\Sigma y_i^2 \cdot f_i$	$\Sigma x_i \cdot y_i \cdot f_i$

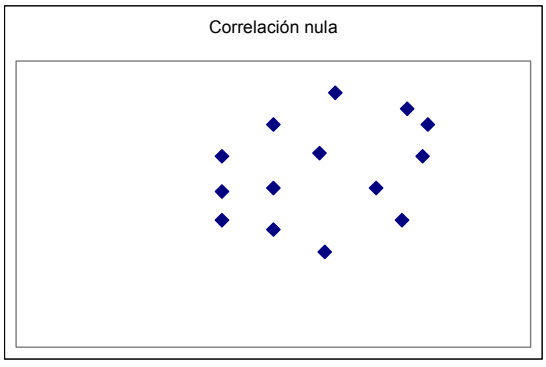
5.2.3. Concepto de correlación

Se denomina **correlación** a la "relación o dependencia" que hay entre las dos variables que intervienen en una distribución bidimensional.

- ***TIPOS DE CORRELACIÓN (VERLO EN LOS DIAGRAMAS DE DISPERSIÓN)***

La correlación puede ser:

a) **Correlación nula:** cuando no hay ninguna relación entre las variables. En este caso los puntos del diagrama están distribuidos al azar , sin formar ninguna línea. Se dice que las variables son incorreladas.

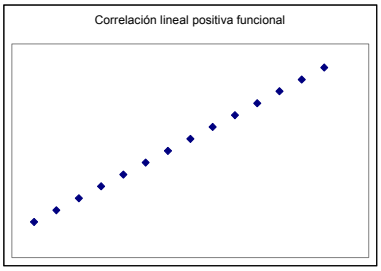


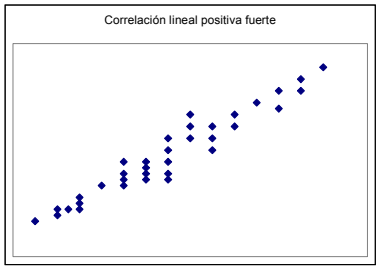
En caso de no ser incorreladas podemos hablar de lo siguiente:

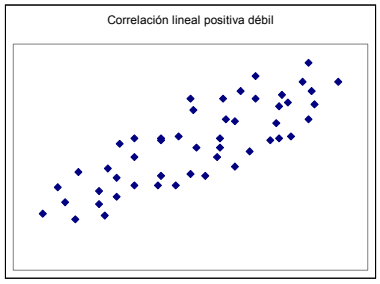
b) **Correlación rectilínea o curvilínea:** según si la nube de puntos se condensa alrededor de una línea recta o una curva, respectivamente.

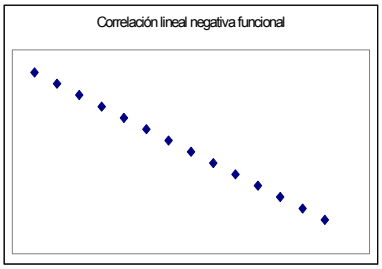
c) **Correlación de tipo funcional:** si hay alguna función que satisfaga todos los valores de la distribución. Si la correlación no es funcional, diremos que ésta es más fuerte o más débil según la mayor o menor tendencia de los valores de la distribución a satisfacer una determinada función.

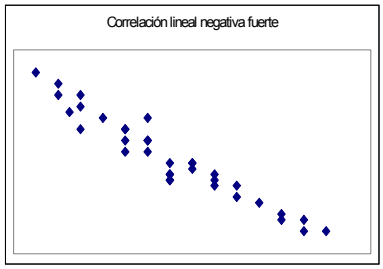
d) **Correlación positiva o directa:** cuando a medida que crece una variable, la otra también crece.
Correlación negativa o inversa: cuando a medida que crece una variable, la otra decrece.

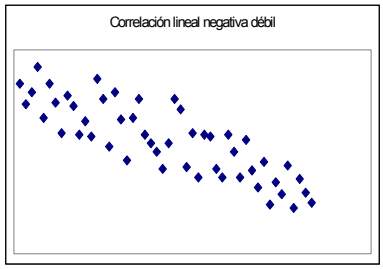


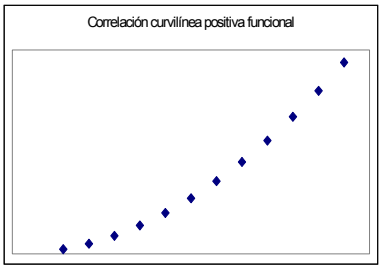


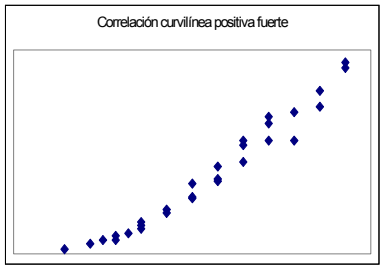


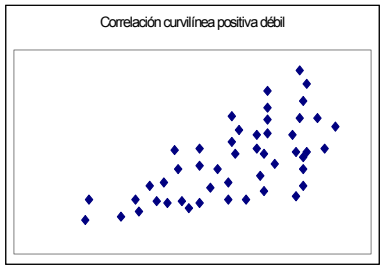


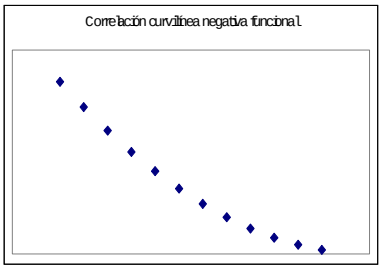


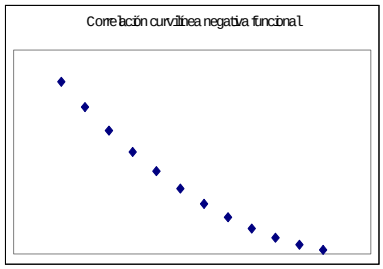


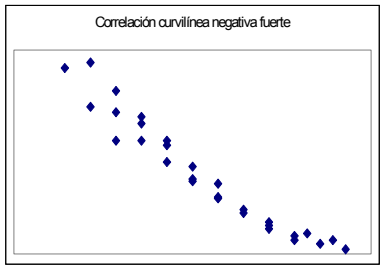












❖ ***COEFICIENTE DE CORRELACIÓN LINEAL O DE PEARSON***

Una vez observada intuitivamente, por medio de la nube de puntos, que hay una relación lineal entre las variables, nos interesa cuantificar de manera más precisa y objetiva esta relación.
Para ello se utiliza el **coeficiente de correlación de Pearson**, que se denota por r se define por medio de la expresión siguiente:

$$r = \frac{S_{XY}}{S_X \cdot S_Y}$$

❖ ***PROPIEDADES Y TIPO DE CORRELACIÓN SEGÚN EL VALOR DEL COEF. DE CORRELACIÓN***

Propiedades de r:

a) El coeficiente de correlación es **adimensional**.

b) **-1<=r<=1**, cumpliendo lo siguiente:

- Si **r=-1**, todos los valores de la variable estadística se sitúan sobre una recta decreciente. En este caso se dice que entre X e Y hay una dependencia funcional (lineal) negativa.
- Si **−1<r<0**, la correlación es (lineal) negativa y será más fuerte (es decir el gráfico se proximará a una recta) a medida que nos acerquemos a −1, y más débil si nos acercamos a 0. En este caso decimos que X e Y están en dependencia aleatoria.
- Si **r=0**, no hay relación entre las X e Y. Decimos que X e Y son aleatoriamente independientes.
- Si **0<r<1**, la correlación es (lineal) positiva y será más fuerte (es decir, el gráfico se aproximará a una recta) a medida que nos acerquemos a 1 y más débil si nos acercamos a 0. En este caso decimos que X e Y están en dependencia aleatoria.
- Si **r=1**, todos los valores de la variable estadística se sitúan sobre una recta creciente. En este caso se dice que entre X e Y hay una dependencia funcional (lineal) positiva.

❖ ***IMPLICACIONES COVARIANZA TIPOS DE CORRELACIÓN)***

Observando la fórmula del coeficiente de correlación podemos afirmar:

- $S_{XY}>0 \Rightarrow r>0$ (correlación directa)
- $S_{XY}=0 \Rightarrow r=0$ (no hay correlación)
- $S_{XY}<0 \Rightarrow r<0$ (correlación inversa)

6.REGRESIÓN LINEAL

6.1. CONCEPTO DE REGRESÓN LINEAL. FINALIDAD.

A veces el estudio de la relación entre dos variables estadísticas tiene como objetivo hacer predicciones sobre los valores de cada una de las variables. El coeficiente de correlación lineal permite estudiar la relación entre las variables, pero no resuelve el problema de la predicción.
El objetivo de la regresión lineal es determinar una relación funcional (que será lineal, es decir una recta) que nos permita hacer predicciones sobre los valores de una de las variables.

Si entre dos variables hay una correlación fuerte, el diagrama de puntos se condensa alrededor de una recta. Si X es la variable independiente e Y la variable dependiente de X, el problema consiste en encontrar la ecuación de la recta que se ajuste mejor a la nube de puntos.

6.2. RECTA DE REGRESIÓN. ECUACIONES.COEFICIENTES DE REGRESIÓN

6.2.1. Definición de recta de regresión

La recta de regresión es la recta que más se aproxima a la nube de puntos de una distribución bidimensional con fuerte correlación lineal.

6.2.2. Ecuaciones de regresión

La ecuación de la *recta de regresión de Y sobre X* es:

$$y - \bar{y} = \frac{S_{XY}}{S_X^2}(x - \bar{x})$$

donde X es la varible predictora o de entrada e Y la variable de respuesta.
Sustituyendo en esta ecuación los valores de X podemos obtener, con cierta aproximación, los valores esperados de Y, que llamaremos predicciones o estimaciones.

Análogamente, la ecuación de la *recta de regresión de X sobre Y* es:
donde Y es la variable predictora o de entrada y X la variable de respuesta.

$$x - \bar{x} = \frac{S_{XY}}{S_Y^2}(y - \bar{y})$$

Servirá para hacer predicciones de X si conocemos los de Y.

6.2.3. Coeficientes de regresión

Los coeficientes de regresión tienen relación con las pendientes de la recta de regresión.

- **Coeficiente de regresión de Y sobre X:**

$$m_{YX} = \frac{S_{XY}}{S_X^2}$$

En este caso el coeficiente de regresión es la pendiente de la recta.

- **Coeficiente de regresión de X sobre Y:**

$$m_{XY} = \frac{S_{XY}}{S_Y^2}$$

En este caso el coeficiente de regresión es el inverso de la pendiente de la recta.

6.4. FIABILIDAD DE LAS PREDICCIONES

Cuanto mayor sea el coeficiente de correlación lineal en valor absoluto, más fiabilidad habrá.

- Si **|r| es muy pequeño**, no tiene sentido hacer estimaciones.
- Si **|r| es cercano a 1**, probablemente los valores reales se acercan a los estimados.
- Si **|r|=1**, las estimaciones coincidirán con los valores reales.

6.5. EJEMPLO

Los pediatras facilitan a los padres la tabla siguiente aon la medias de todos los pesos de los niños según su edad.

Edad (meses)	*0*	*3*	*6*	*9*	*12*	*15*	*18*	*21*	*24*
Peso (kg)	*3,5*	*6,25*	*8*	*9,2*	*10,2*	*11*	*11,6*	*12,05*	*12,6*

Hallar el coeficiente de correlación y la recta de regresión del peso sobre la edad. Representar la recta de regresión en el diagrama de dispersión.
¿Cuál es el incremento mensual del peso?
¿Cuál es el peso esperado de un niño de 14 meses?

El diagrama de dispersión que corresponde a estos datos es el siguiente:
Si hacemos el diagrama de dispersión veremos que hay una dependencia lineal entre las dos variables.

Hacemos los siguientes cálculos para hallar el coeficiente de correlación y la recta de regresión:

	X	**Y**	**X²**	**Y²**	**XY**
	0	3,50	0	12,2500	0,00
	3	6,25	9	39,0625	18,75
	6	8,00	36	64,0000	48,00
	9	9,20	91	84,6400	82,80
	12	10,20	144	104,0400	122,40
	15	11,00	225	121,0000	165,00
	18	11,60	324	134,5600	209,80
	21	12,05	441	145,2025	253,05
	24	12,60	576	158,7600	302,40
Sumas	108	84,40	1836	863,5150	1201,20

$$\overline{X} = \frac{108}{9} = 12 \qquad \overline{Y} = \frac{84,4}{9} \cong 9,378$$

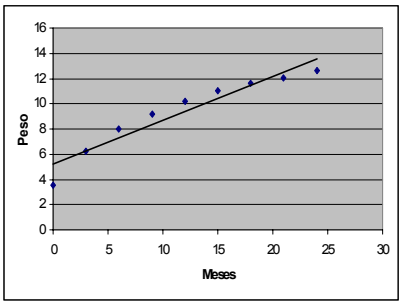
$$S_X = \sqrt{\frac{1836}{9} - 12^2} = \frac{\sqrt{540}}{3} \cong 7,745967 \qquad S_Y = \sqrt{\frac{863,5150}{9} - \left(\frac{84,4}{9}\right)^2} \cong 2,829027$$

$$S_{XY} = \frac{1201,20}{9} - 12\frac{84,4}{9} \cong 20,933333$$

Coeficiente de correlación $r = \dfrac{S_{XY}}{S_X S_Y} \cong 0,955$

Pendiente de la recta de regresión $m = \dfrac{S_{XY}}{S_X^2} \cong 0,349$

Recta de regresión $Y - \dfrac{84,4}{9} = 0,349(X - 12) \Rightarrow Y = 0,349X + 5,190$



Por tanto, el incremento de peso esperado en un mes será de 0,349 kg (pendiente de la recta de regresión).

El peso esperado para un niño de 14 meses es Y=0,349·14+5,190=10,076 kg