MINISTERIO DE EDUCACION Y CIENCIA

DIRECCION GENERAL DE ENSEÑANZAS MEDIAS

SUBDIRECCION GENERAL DE ORDENACION ACADEMICA

BACHILLERATO DE ADMINISTRACION Y GESTION

ASIGNATURA:

ESTADISTICA APLICADA

DOSSIER DIDACTICO DE ESTADISTICA DESCRIPTIVA

AUTOR:

SANTIAGO DE LA FUENTE FERNANDEZ

COLABORANDO EL:

GRUPO MADRID-MARKOV

SILVIA ANA PEREZ MATEO
SONSOLES PERTEGUER MUÑOZ
MARIA LUISA QUIROS PARRA
CONCEPCION ALONSO DELGADO
MARIA DEL CARMEN SOLER LEON

SANTIAGO DE LA FUENTE FERNANDEZ

GRUPO MADRID – MARKOV

ESTADISTICA DESCRIPTIVA

el alumno – el profesor

ISBN: 84 - 404 - 0348 - 6

Depósito legal: M - 29344 - 1987

PROLOGO

Busco el lugar donde estoy en esta historia,
porque yo también soy estadística.

La estadística está desempeñando un importante papel ascendente en casi todas las facetas del progreso humano. Al principio sólo era aplicada a los "asuntos de Estado", de donde procede su nombre; en la actualidad por su enorme interés y múltiples aplicaciones la influencia de la estadística se extiende a la economía, medicina, agricultura, física, política, psicología, sociología, veterinaria y otros muchos campos de las ciencias experimentales.

En esta línea, la estadística tiene un lugar relevante en la nueva curricula de las enseñanzas medias.

El alumno de Bachillerato debe poseer una formación estadística que le permita: por un lado, una fácil comprensión de los gráficos y técnicas tan utilizadas hoy en día por los medios de comunicación; por otro, una interpretación correcta de los conceptos de amplia divulgación como son: salarios, deflación, inflación, IPC, poder adquisitivo, evolución de una población, análisis curricula de distintas disciplinas y tasas de las diferentes variables socioeconómicas.

En este Dossier-didáctico (ESTADISTICA DESCRIPTIVA / ESTADISTICA INFERENCIAL) se pretende seleccionar aquellos conceptos indispensables que sean de utilidad para el alumno.

Dossier-didáctico está enfocado preferentemente hacia actividades resueltas - que sirvan para introducir y esclarecer la teoría de puntos más críticos, en los que el estudiante continuamente se siente más inseguro.

Los objetivos que se pretenden cubrir podemos resumirlos básicamente en los siguientes:

a) La elaboración coherente de series de datos relativos a cualquier fenómeno de la vida real. Ayudándose para ello de las técnicas analíticas y gráficas de esta materia.

b) Saber detectar la asociación o no entre dos o más variables y obtener las consecuencias de este hecho de operar en problemas de decisión social, científica, económica, etc.

c) Interpretar conceptos tan utilizados hoy día por los medios de comunicación como inflación, deflación, IPC, poder adquisitivo, tasas de natalidad, mortalidad, evolución económica, etc. Por último conocer y manejar las principales fuentes estadísticas elaboradas en nuestro país.

La incorporación al aula de los elementos de este Dossier-didáctico (DESCRIPTIVA / INFERENCIAL) y el buen uso de estos, puede plantear que las mismas actividades sean más formativas y el proyecto educativo más fiable.

\+ + + +

Con mi más profundo agradecimiento a los compañeros del Grupo MADRID-MARKOV que han tenido la gentileza de colaborar en profundas discusiones sobre Dossier-didáctico (profesores: Pérez Mateo, Perteguer Muñoz, Alonso Delgado, Soler León, Quirós Parra), y que tanto contribuyeron en su práctica experimental, entrego este libro al lector, deseándole feliz excursión por él ... y un saludo grato.

Madrid, noviembre de 1987.

LAFUENTE

INDICE GENERAL

3. Sucesos y probabilidad: Fenómenos aleatorios. Probabilidad. Probabilidad condicionada. Independencia. Probabilidad total. Análisis combinatorio. Teorema de Bayes. Asignación de probabilidades.

¿ QUE ES LA ESTADISTICA ?

"El pensamiento estadístico será un día tan necesario para el ciudadano eficiente como la capacidad de leer y escribir"

H.G. WELLS

- Estadística en la vida.
- Conceptos Generales (Población estadística. Unidad estadística. Caracteres. Muestra).
- ¿Qué es la Estadística?
- ¿Qué es la Probabilidad?
- Estadística:

  Estadística Descriptiva

  Estadística Inferencial

Estadística en la vida

Habrás observado tú mismo que toda entidad de cierta importancia, tanto oficial como privada, controla sus actividades estableciendo sus estudios estadísticos.

La Estadística tiene un origen etimológico oscuro. Proceda del latín "status" (situación), del griego "statera" (balanza) o del alemán — "staat" (Estado), lo cierto es que encierra estas tres ideas, ya que estudia la situación de personas o cosas midiendo o pesando los hechos sometidos a su estudio en beneficio de una unidad científica, política o social.

Existe, pues, un convencimiento casi general de que el lenguaje estadístico irá asumiendo cada vez mayor relieve en las ciencias aplicadas (Economía, Ciencias de la Información, Historia, Biología, Medicina, — Pedagogía, Psicología, Políticas, Ingeniería, ...)

A juicio de Atkinson, "es un hecho histórico familiar que a medida que la ciencia progresa, sus teorías se van haciendo más y más estadísticas en la forma".

CONCEPTOS GENERALES

- **Población estadística o colectivo**

  Conjunto de elementos (existentes o posibles) sobre el cual van a recaer las observaciones. Por elemento entendemos cualquier persona, animal, cosa, operación, familia, institución, etc.

  Las poblaciones podrán ser finitas o infinitas, dependiendo del número de elementos que las forman.

- **Unidad estadística o individuo**

  Cada uno de los elementos que componen la población estadística. El indi-

viduo es un ente observable que puede ser una persona, animal, cosa, o - incluso algo abstracto.

- Caracteres

La observación del individuo la describimos mediante uno o más caracteres. El carácter es, por tanto, una propiedad inherente en el individuo.

Conviene distinguir entre caracteres cuantitativos y caracteres cualitativos:

a) Carácter cuantitativo: Es el carácter que es medible, esto es, se puede cuantificar, como, por ejemplo, la edad, el peso y la estatura de las personas.

b) Carácter cualitativo: Es el carácter que no es medible, como, por ejemplo, el color del pelo, el sexo, etc.

- Muestra

Cualquier subconjunto de una población. La muestra hace siempre referencia a una población de la cual es parte.

Supongamos que observamos un carácter de la población. Por ejemplo, consideramos la estatura de los universitarios españoles (población). Paralelamente, observamos la estatura de 500 universitarios españoles (muestra).

Es claro que, dada una misma población, podemos tener distintas poblaciones de observaciones y, consiguientemente, distintas muestras, según que estudiemos uno u otro carácter. Así, con los mismos universitarios españoles, podíamos haber considerado su altura, su peso, su capacidad intelectual. etc.

"Se suelen tomar muestras cuando es difícil o costosa la observación de todos los elementos de la población estadística".

El número de elementos de la muestra se llama "tamaño de la muestra".

¿QUE ES LA ESTADISTICA?

Difícil resulta definirla y no se ha llegado a un acuerdo. Quizá se pudiera definir:

> "Es la ciencia o disciplina que recoge, ordena y analiza los datos de una muestra, extraída de cierta población, y que, a partir de esa muestra, valiéndose del Cálculo de Probabilidades, se encarga de hacer inferencias (predicciones y generalizaciones) acerca de la población estadística".

Se hace necesario distinguir entre Estadística, estadísticas y estadístico (o estadísticos).

ESTADISTICA: Es la ciencia acabada de definir.

Estadísticas: Son los resultados numéricos obtenidos mediante la Estadística. Por ejemplo, consumo medio semanal de leche por familia, número de accidentes de tráfico durante un mes, etc.

Estadístico: Valor numérico obtenido a partir de los valores presentados por una muestra. Así, por ejemplo, será estadístico la media aritmética de ocho observaciones de una muestra. Supongamos que éstas son: 3, 12, 8, 7, 7, 10, 6, 8.

¿QUE ES PROBABILIDAD?

El delicado problema de la definición de probabilidad ha preocupado desde la Antigüedad a matemáticos y filósofos, todavía hoy divide a unos y a otros. Para Aristóteles "lo probable es lo que ocurre con frecuencia". A esta "definición" imprecisa (perfeccionada después), se oponen definiciones de tipo subjetivo, así, si en una urna tenemos 8 bolas iguales salvo en el color (blanca, negra, amarilla, roja, azúl) y extraemos una bola al

azar; diremos que todas las bolas tienen "la misma probabilidad" de salir. Si entre las 8 bolas hay 3 blancas diremos que la probabilidad de obtener - bola blanca es $p = \frac{3}{8}$, y en general:

> "La probabilidad p de que ocurra un acontecimiento es igual al número de casos favorables sobre el número de casos posibles:
>
> $$p = \frac{\text{número de casos favorables}}{\text{número de casos posibles}}$$ "

"En Estadística inferencial, la Probabilidad es el puente que nos permite pasar válidamente de la muestra a la población, en otras palabras:

> La Probabilidad legitima el salto desde las características conocidas de la muestra hasta las características desconocidas de la población".

Será un estadístico la media aritmética de las puntuaciones, es decir: - -

$$\frac{3 + 12 + 8 + 7 + 7 + 10 + 6 + 8}{8} = 7'625$$

Por supuesto, estadístico es, también, usado para denominar a la persona que se dedica a la Estadística.

Según la definición dada, la Estadística consta de dos partes fundamentales:

a) <u>Estadística descriptiva</u>

Tiene como cometido describir una muestra, esto es, recoger, ordenar y analizar los datos de una muestra.

b) <u>Estadística inferencial</u>

Cuyo cometido es hacer inferencias sobre la población, a partir de la muestra. Para hacer predicciones y generalizaciones sobre la población es necesario utilizar la Probabilidad.

I

ESTADISTICA DESCRIPTIVA

- Ordenación de datos.
- Distribuciones de un carácter.
- Representaciones gráficas.
- Medidas de tendencia central.
- Medidas de dispersión o concentración.
- Elaboración de una encuesta.

"FIN DE CURSO"

Al terminar el curso, Javi, Pipo y Tony se reunen en su club ...

JAVI:

¡He aprobado todo! ¡Ya he acabado la E.G.B. y he obtenido el Graduado Escolar!.

TONY:

¡Yo tengo un Certificado de Estudios Primarios!.

PIPO:

¡Yo todavía no he acabado la E.G.B.!, pero ... quisiera saber cuánta gente obtiene el Graduado Escolar.

TONY:

¿Por qué no buscamos al profesor de matemáticas del Instituto para preguntárselo?.

TONY:

¡Hola Sr. Profesor!. ¡Queremos saber cuánta gente obtiene el Graduado Escolar!.

PROFESOR:

¡Bueno, bueno!, ¡busquen datos en el colegio del barrio de los últimos años!. ¡Aquí está el nº de Graduados y el nº de Certificados que han sido expedidos en los últimos años en este colegio!.

AÑO	1980	1981	1982	1983	1984	1985	1986	1987
GRADUADO ESCOLAR	110	90	102	95	100	92	84	80
CERTIFICADO DE ESTUDIOS	12	35	20	30	22	32	40	45

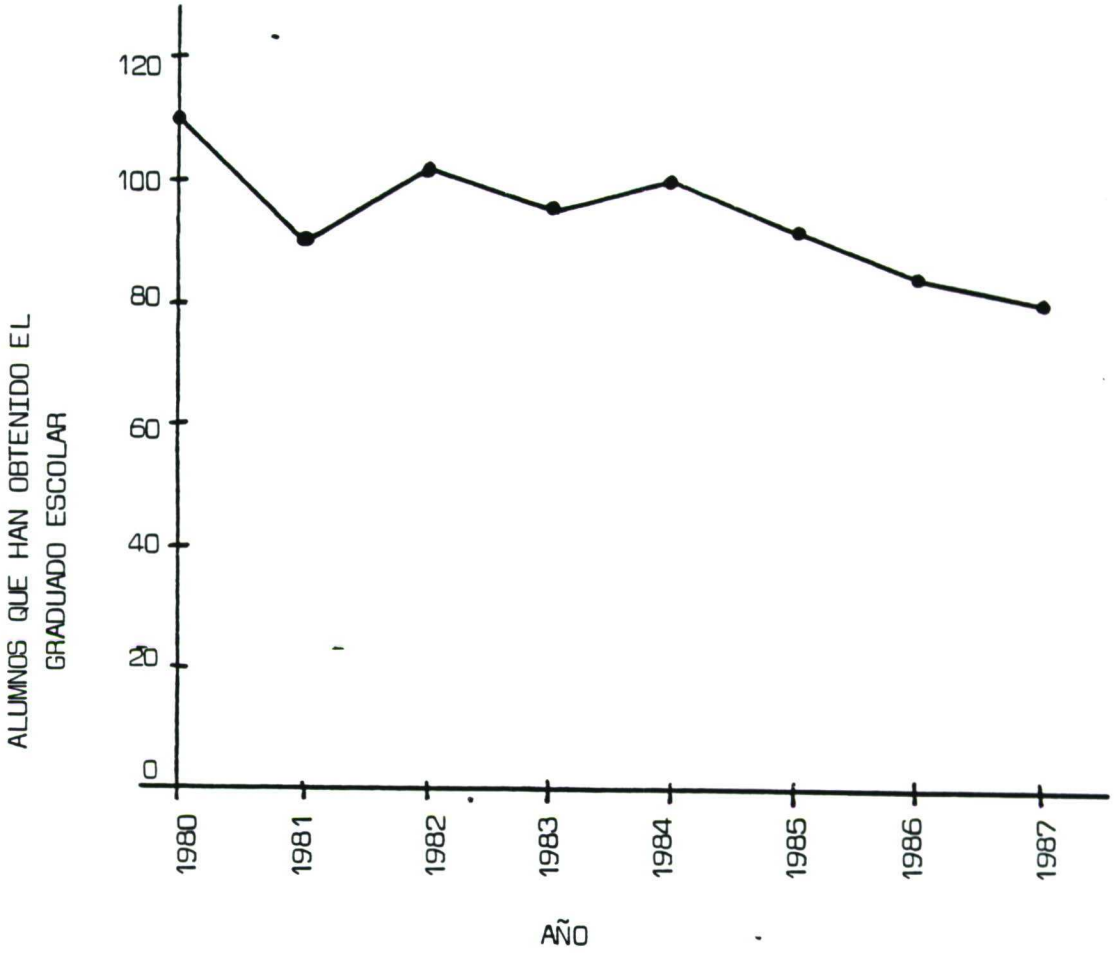
Vamos a representar los datos gráficamente. Hay varios métodos de representación, según el caso estudiado nos interesarán unos u otros.

En nuestro caso, veamos primero el GRAFICO DE LINEA. Estudiaremos cuántas personas han obtenido el Graduado Escolar en los diferentes años.

Sobre el eje OX representaremos la variable independiente ó tiempo y sobre el eje OY la variable dependiente del tiempo; en nuestro caso el nº de personas que han obtenido el Graduado Escolar.

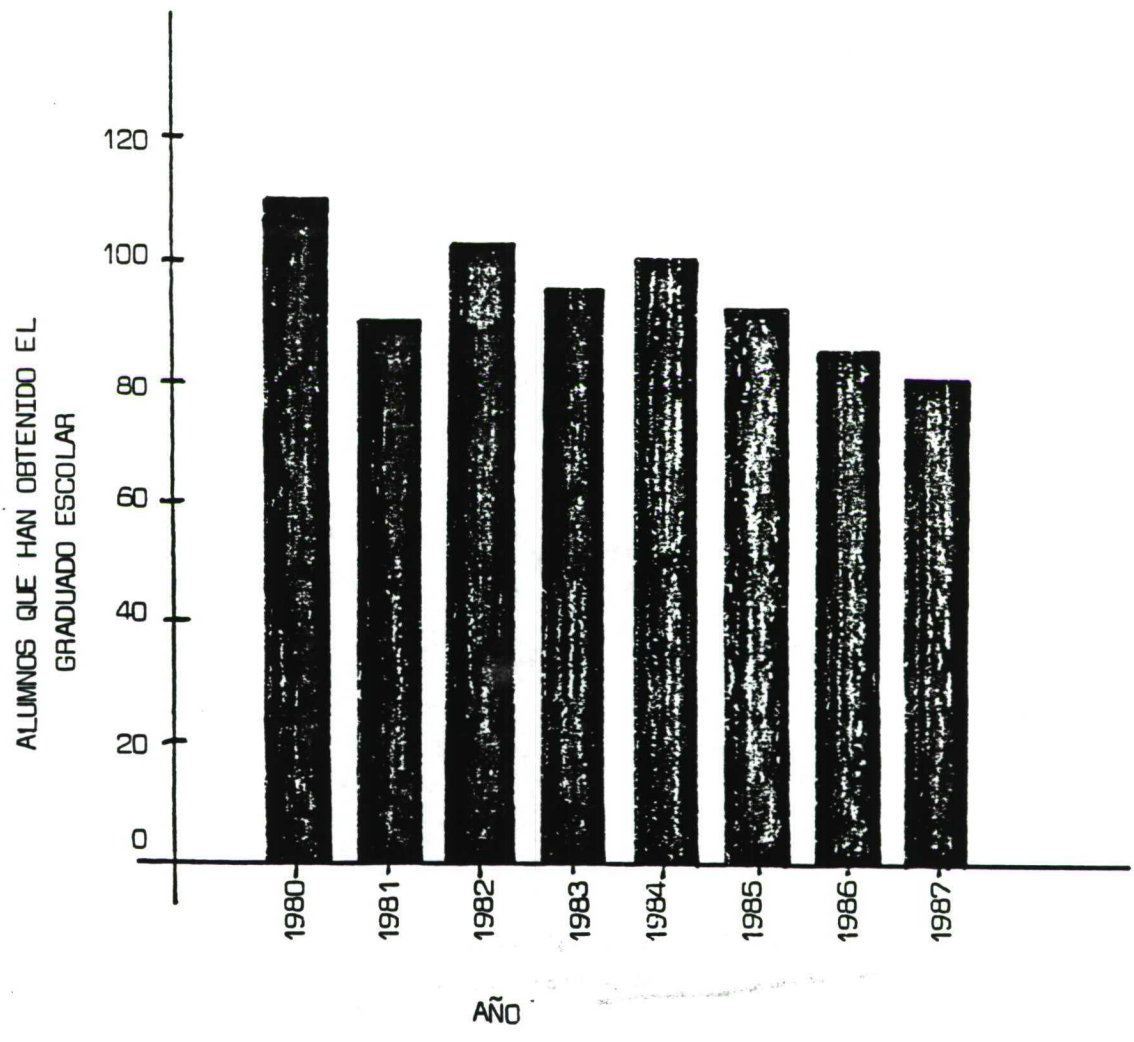
Las unidades sobre los ejes no tienen porqué ser iguales, puesto que las dos variables representan cantidades esencialmente distintas.

El cero se representa sobre el eje vertical, pero no tiene representación sobre el eje horizontal.



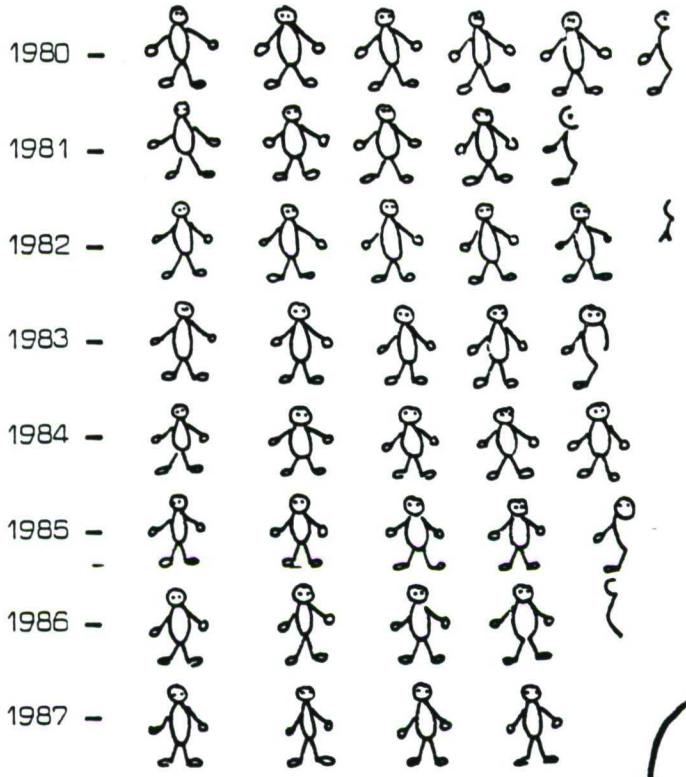
Los puntos se sitúan en el sistema de coordenadas, sacándolos de la tabla, por ejemplo (1983, 95)

Veamos otra forma:



Con estos datos podemos utilizar un tercer método; que es el diagrama de figuras o PICTOGRAMA, y es un método muy llamativo y original de representar datos mediante figuras, por ejemplo:

Si  representa a 20 personas que han obtenido el Graduado Escolar, podríamos formar el siguiente PICTOGRAMA:



Si queremos representar dos variables juntas, por ejemplo, el nº de personas que obtienen el Graduado Escolar y el nº de personas que piden un Certificado de Estudios, podríamos representar los datos de la siguiente manera:

GRAFICO DE LINEA:

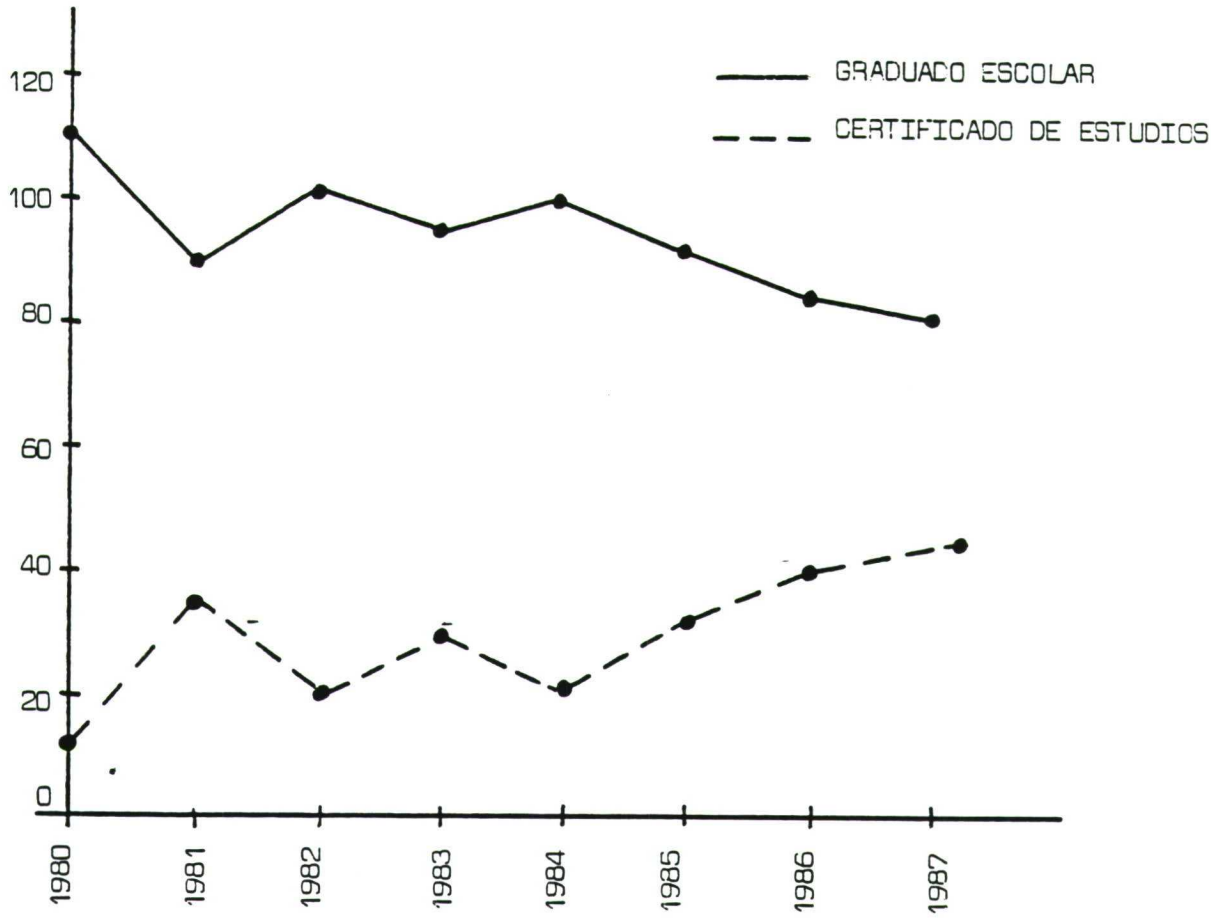
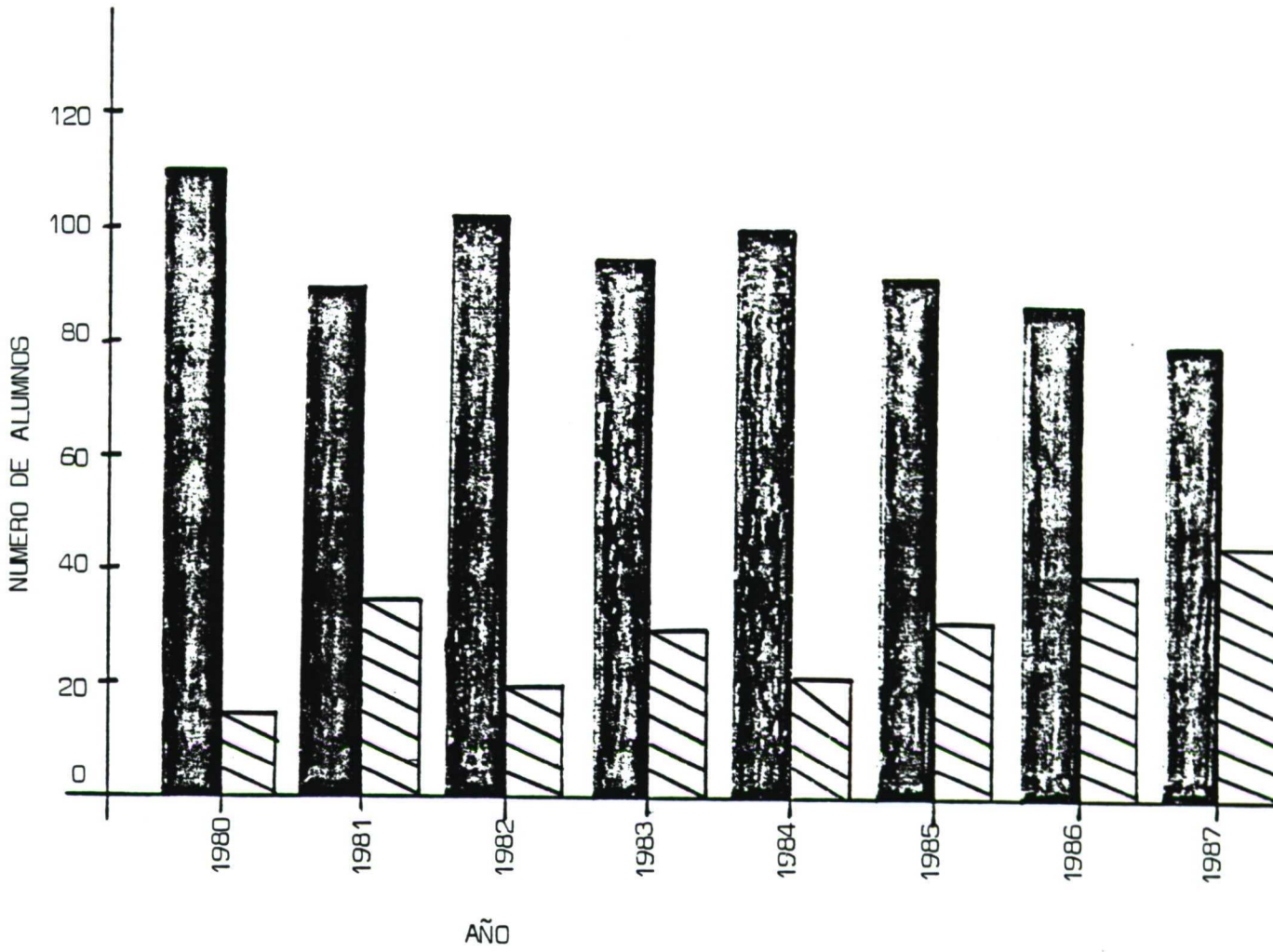


DIAGRAMA DE BARRAS:



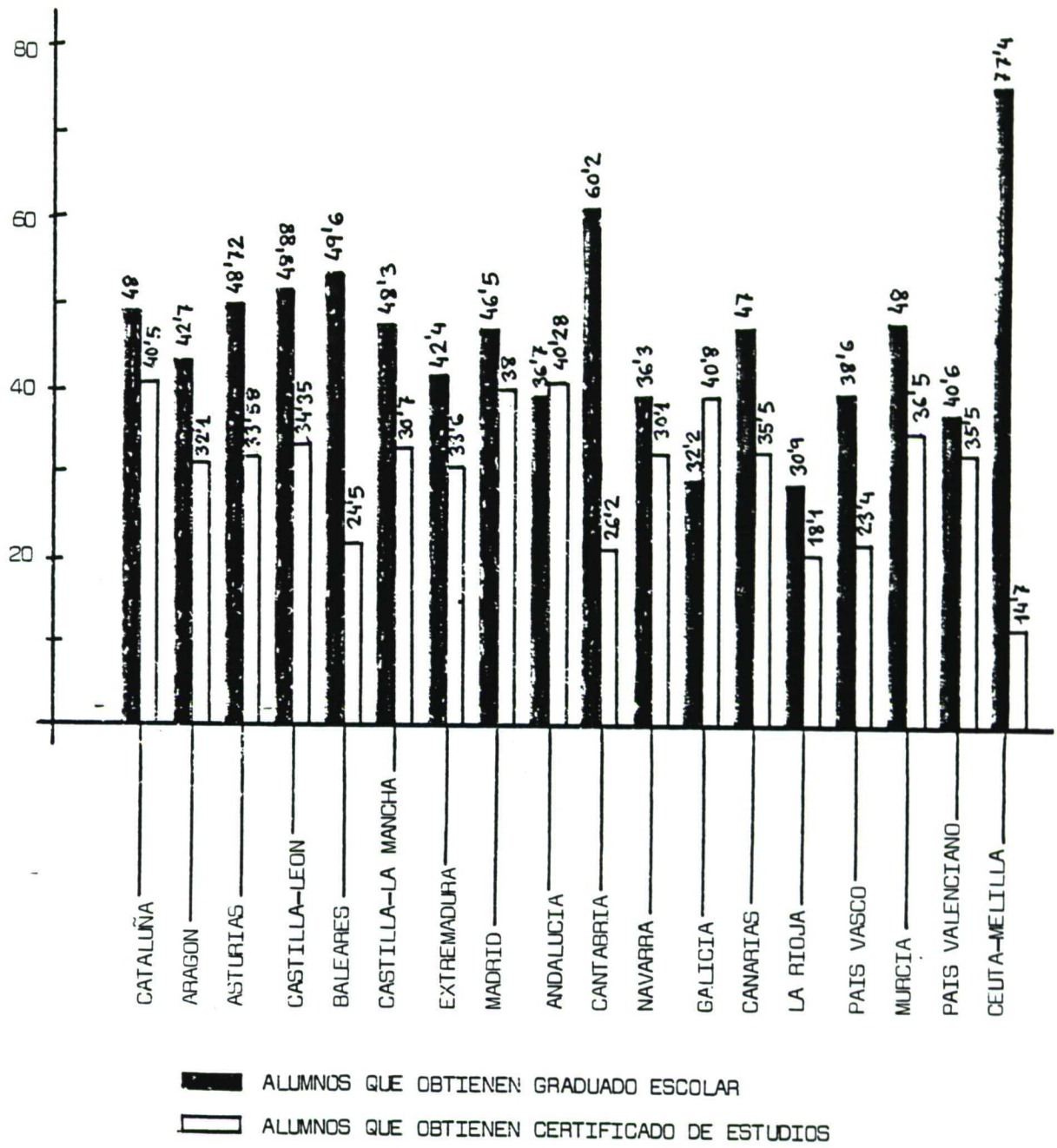
PROFESOR:

Hay un estudio de toda España. Esperar que lo busco ... ¡aquí está! ... (resultados en cientos) ... Veamos:

	GRADUADO ESCOLAR	CERTIFICADO DE ESTUDIOS
CATALUÑA	48	40'5
ARAGON	42'7	32'1
ASTURIAS	48'72	33'58
CASTILLA-LEON	48'88	34'35
BALEARES	49'6	24'5
CASTILLA-LA MANCHA	48'3	30'7
EXTREMADURA	42'4	33'6
MADRID	46'5	38
ANDALUCIA	36'7	40'28
CANTABRIA	60'2	26'2
NAVARRA	36'3	30'1
GALICIA	32'2	40'8
CANARIAS	47	35'5
LA RIOJA	30'9	18'1
VASCONGADAS	38'6	23'4
MURCIA	48	36'5
VALENCIA	40'6	35'5
CEUTA-MELILLA	77'4	14'7
NUMERO TOTAL DE ALUMNOS	823 x 100	568'41 x 100

(resultados imaginarios)

DISTRIBUCION POR AUTONOMIAS DE LOS RESULTADOS AL FINAL DE LOS ESTUDIOS



donde

Alumnos que obtienen Graduado Escolar 823

Alumnos que obtienen Certificado de Estudios ..... 568'41

Los resultados vienen expresados en cientos, en consecuencia, alumnos Graduado Escolar = 82.300, alumnos Certificados de Estudios = 56.841.

Os voy a explicar un método de representación que se utiliza bastante, es el "DIAGRAMA DE SECTOR".

Consiste en representar mediante sectores circulares, las distintas modalidades de una variable. Cada una de las modalidades se representa proporcionalmente a los 360º del círculo.

Supored que el resultado definitivo al terminar la E.G.B. fué el siguiente:

RESULTADOS (en cientos)	
GRADUADO ESCOLAR	823
CERTIFICADO DE ESTUDIOS	568'41
NO ACABAN E.G.B.	108'59
SUMA 1500	

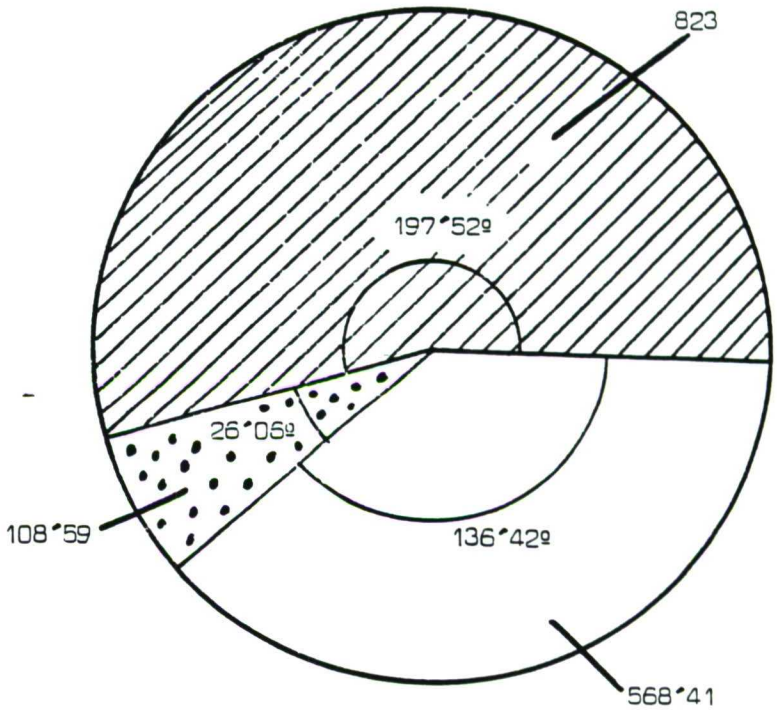
dividimos proporcionalmente los 360º del círculo:

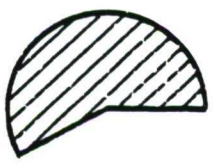
$$\text{Graduado Escolar} = \frac{823}{1500} \times 360º = 197'52º$$

$$\text{Certificado de Estudios} = \frac{568'41}{1500} \times 360º = 136'42º$$

$$\text{No acaban E.G.B.} = \frac{108'59}{1500} \times 360º = 26'06º$$

RESULTADOS OBTENIDOS AL FINAL DE E.G.B. MEDIANTE "DIAGRAMA DE SECTOR".

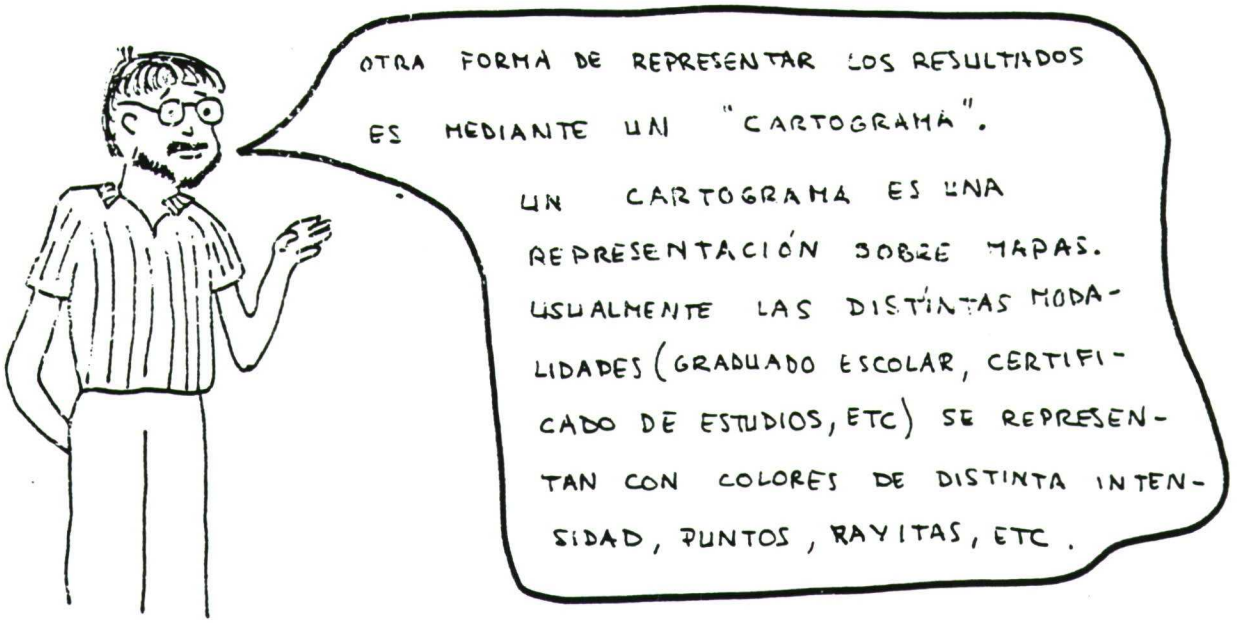




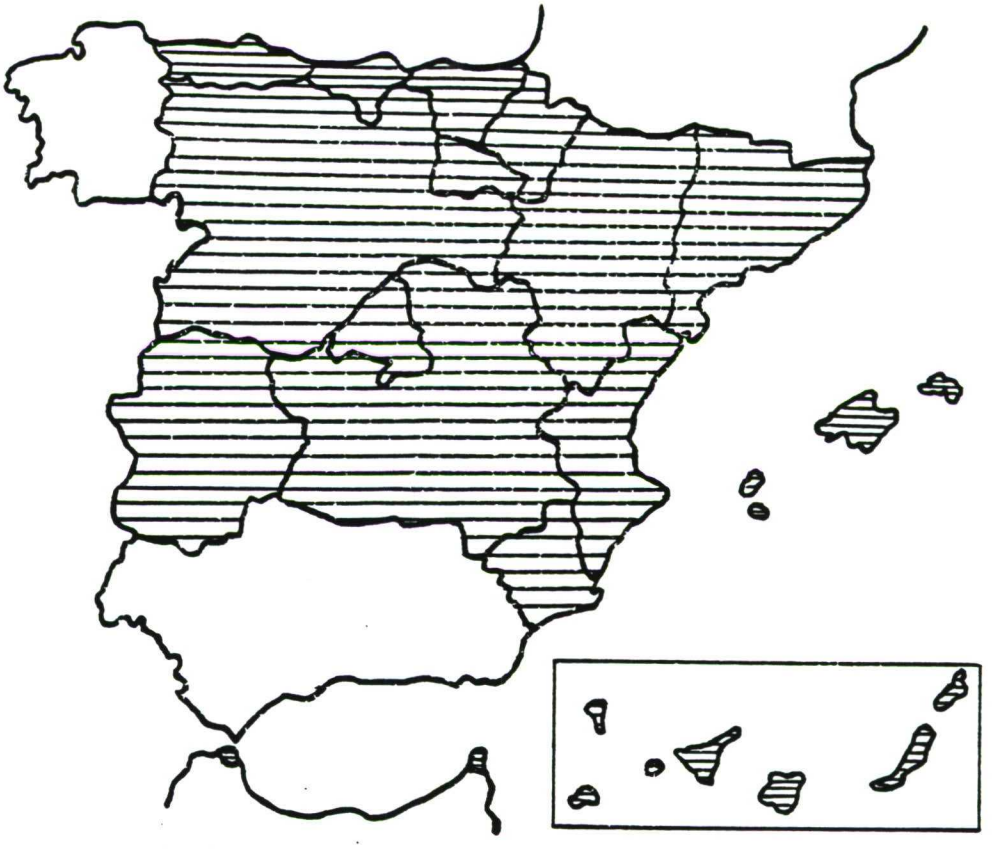
ALUMNOS QUE OBTIENEN EL GRADUADO ESCOLAR

ALUMNOS QUE NO ACABAN E.G.B.

ALUMNOS CON CERTIFICADO DE ESTUDIOS



REPRESENTACION DE LOS RESULTADOS AL FINAL DE LA E.G.B. MEDIANTE UN MAPA.



ACTIVIDADES

R E C U E R D A

Individuo: Es un ente observable que no tiene porqué ser una persona, puede ser un objeto, un animal, una cosa, o algo abstracto.

Población: Es el conjunto de todos los individuos donde van a recaer las observaciones.

Muestra: Es un subconjunto de individuos de la población. Se suelen tomar muestras cuando es difícil o costosa la observación de todos los individuos de la población.

Caracteres: La observación del individuo la describimos mediante uno o más caracteres. Hay caracteres medibles, es decir, se pueden cuantificar: el peso, la altura, etc. Y caracteres no medibles: el sexo, color del pelo, etc.

A los caracteres medibles se les llama "caracteres cuantitativos".

A los caracteres no medibles se les llama "caracteres cualitativos".

Modalidades: Un carácter puede tener distintas modalidades, de esta forma:

- Un carácter cuantitativo: "el peso" puede adoptar las modalidades: ocho años, diez años, quince años, etc.
- Un carácter cualitativo: "el estado civil" puede adoptar las modalidades:

  soltero, casado, separado, viudo, etc.

Variables estadísticas: Las diferentes modalidades de un carácter cuantitativo adoptan valores numéricos. A estos valores numéricos se los llama "variable estadística".

Las variables estadísticas se clasifican:

VARIABLE ESTADISTICA DISCRETA: Son aquellas que toman valores aislados, y no pueden tomar ningún valor entre dos valores consecutivos fijados. Ejemplo:

número de enfermos de un hospital: 3, 4, 5, 6, ...

no puede haber: 3'5, 4'2, 5'7, 6'3, ... enfermos.

VARIABLE ESTADISTICA CONTINUA: Son aquellas que pueden tomar infinitos valores en un intervalo dado, esto es, pueden tomar valores entre dos valores consecutivos fijados. Ejemplo:

altura de las personas: 1'70 metros, 1'80 metros, 1'85 metros, ...

puede haber personas: 1'73 metros, 1'74 metros, 1'76 metros, ...

ORGANIZACION DE DATOS

Nos ocupamos ahora de ordenar los datos obtenidos en la observación de una muestra o población.

Piensa que una variable estadística X puede tomar distintos valores, x_1, x_2, x_3, ..., x_k, y cada uno de éstos valores puede aparecer repetido más de una vez.

• RECORRIDO O AMPLITUD: Es la diferencia entre el mayor y el menor de los valores que forman la variable. Ejemplo:

sea X = edad en años

X : 8, 10, 15, 24, 36, 37, 40, 50, 51, 52

el recorrido sería: 52 - 8 = 44

• FRECUENCIA ABSOLUTA (n_i): La frecuencia absoluta de un valor x_i de la variable estadística X es el número de veces que aparece repetido dicho valor en el conjunto de observaciones realizadas.
La frecuencia absoluta del valor x_i se denota por n_i.

• FRECUENCIA RELATIVA (f_i): La frecuencia relativa de un valor x_i de la variable estadística X es el cociente entre la frecuencia absoluta n_i y el número de observaciones realizadas N.

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

• PORCENTAJE: El porcentaje de un valor x_i de la variable estadística X se obtiene multiplicando la frecuencia relativa f_i por 100.

$$(\%)x_i = P_{x_i} = f_i \times 100 = \frac{n_i}{N} \times 100$$

• FRECUENCIA ABSOLUTA ACUMULADA (N_i): La frecuencia absoluta acumulada de un valor x_i de la variable estadística X es la suma de frecuencias absolutas de los valores inferiores o iguales a x_i.

La frecuencia absoluta acumulada de un valor x_i se denota por N_i:

$$N_i = n_1 + n_2 + n_3 + \dots + n_{i-1} + n_i$$

siendo

$$N_1 = n_1$$

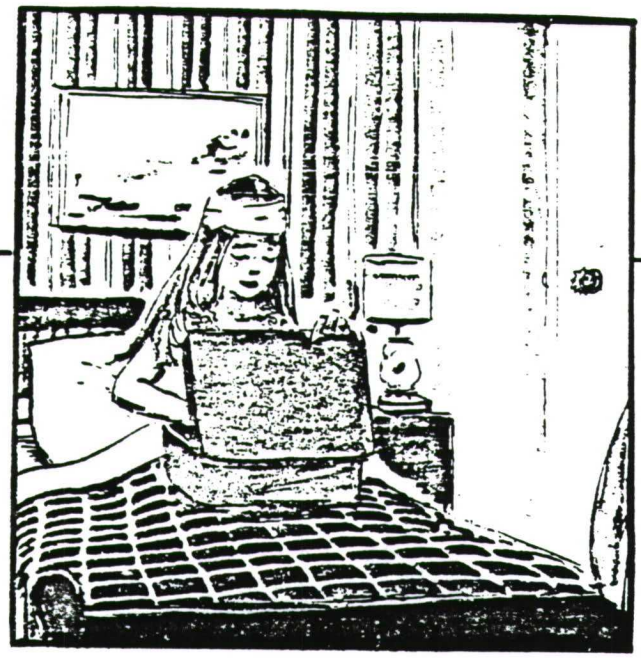
$$N_2 = n_1 + n_2$$

$$N_3 = n_1 + n_2 + n_3$$

$$\dots\dots\dots\dots\dots$$

• FRECUENCIA RELATIVA ACUMULADA (F_i): La frecuencia relativa acumulada de un valor x_i es el cociente entre la frecuencia absoluta acumulada N_i y el número de observaciones realizadas N.

$$F_i = \frac{N_i}{N}$$



ACTIVIDAD - 1

En un determinado hospital se han ido anotando, durante una semana, la distribución de 40 camas, por razón de internamiento. Obteniéndose así la tabla de información:

X= camas necesarias	frecuencia absoluta n_i	frecuencia relativa f_i	Porcentaje $f_i \times 100$
accidentes	12	12/40 = 0'3	30 %
quemaduras	6	6/40 = 0'15	15 %
partos	15	15/40 = 0'375	37'5 %
enfermedades crónicas	7	7/40 = 0'175	17'5 %
SUMA TOTAL.	40	1	100 %

Observa que X es una variable cualitativa (razón de internamiento) que adopta cuatro modalidades:

- accidentes
- quemaduras
- partos
- enfermedades crónicas

Se verifica:

- La suma de las frecuencias absolutas n_i es igual al número total de casos (N = 40).

$$\sum_{i=1}^{4} n_i = 40 = N$$

- La suma de las frecuencias relativas f_i es igual a la unidad.

$$\sum_{i=1}^{4} f_i = 1$$

- El porcentaje se obtiene: $(\%)_i = f_i \times 100$, la suma de todos los porcentajes de cada modalidad es igual a 100.

"Las variables cualitativas (no pueden ser medidas) se representan gráficamente por:

- Diagrama de barras
- Diagrama de sectores
- Pictograma".

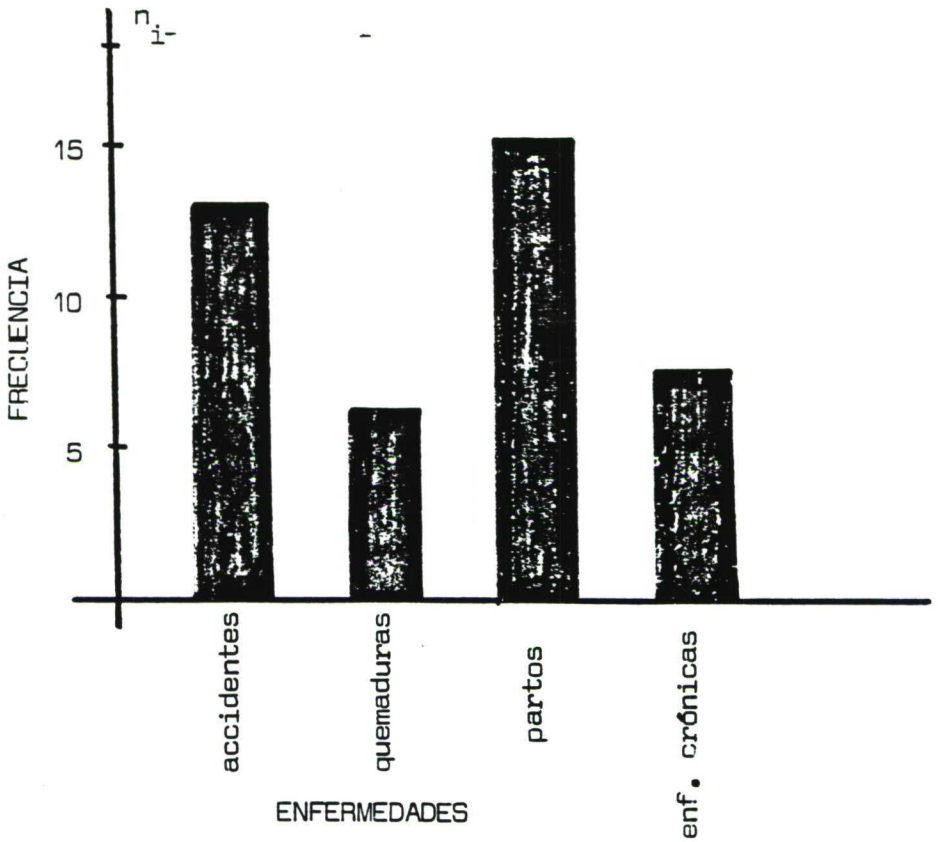
REPRESENTACIONES GRAFICAS DE VARIABLES CUALITATIVAS

Las representaciones gráficas más usuales para caracteres cualitativos son:

• DIAGRAMA DE BARRAS (o de rectángulos)

Consiste en un conjunto de barras o de rectángulos sobre un eje de coordenadas. Se representan en abscisas las distintas modalidades y se levantan sobre ellas rectángulos de bases iguales y cuya altura será la correspondiente a la frecuencia absoluta de cada modalidad.

	frecuencia absoluta n_i
accidentes	12
quemaduras	6
partos	15
crónicas	7



Las barras están separadas por espacios en blanco y pueden estar colocadas en cualquier orden.

• DIAGRAMA DE SECTORES

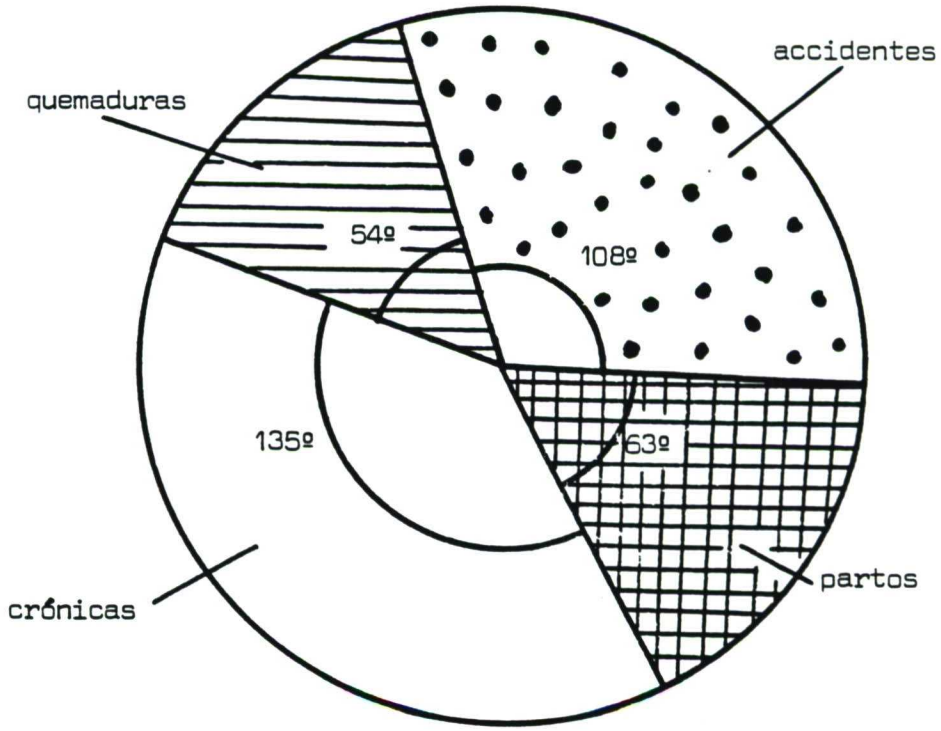
En un círculo se asigna un "sector circular" a cada una de las modalidades, siendo la amplitud del sector proporcional a la frecuencia relativa de cada modalidad.

En otras palabras:

> "Cada una de las modalidades se representa proporcionalmente a los 360º del círculo".

Gráficamente:

	frecuencia absoluta n_i	frecuencia relativa f_i
accidentes	12	12/40 = 0'3
quemaduras	6	6/40 = 0'15
partos	15	15/40 = 0'375
crónicas	7	7/40 = 0'175
SUMA TOTAL	40	1



$$\text{accidentes} = f_1 \times 360º = 0'3 \times 360º = 108º$$

$$\text{quemaduras} = f_2 \times 360º = 0'15 \times 360º = 54º$$

$$\text{partos} = f_3 \times 360º = 0'375 \times 360º = 135º$$

$$\text{crónicas} = f_4 \times 360º = 0'175 \times 360º = 63º$$

$$360º$$

• PICTOGRAMA

Cada modalidad se representa por un dibujo de tamaño proporcional a la frecuencia absoluta de la misma. Es frecuente utilizar un dibujo relacionado con la variable que se estudia.

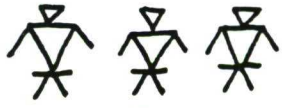
	frecuencia absoluta n_i
accidentes	12
quemaduras	6
partos	15
crónicas	7

E = enfermo

partos accidentes crónicas quemaduras

También es frecuente tomar un dibujo estándar y repetirlo un número de veces proporcional a la frecuencia.

Sea  = 4 enfermos.

accidentes 

quemaduras 

partos 

crónicas 

Estos diagramas son utilizados a menudo para representar datos estadísticos en una forma que llame la atención a todo el público que los vea.

TABLA DE FRECUENCIAS DE VARIABLE DISCRETA

Una variable estadística discreta es aquella que toma valores aislados, y no puede tomar ningún valor entre dos valores consecutivos fijados.

La tabla de frecuencias para una variable discreta se forma ordenando "de menor a mayor" los distintos valores de la variable y anotando las distintas frecuencias: n_i, f_i, N_i, F_i.

Sea X la variable estadística discreta con modalidades $x_i, x_2, x_3, \ldots, x_i, \ldots \ldots, x_k$, siendo $(x_1 < x_2 < x_3 < \ldots < x_i < \ldots < x_k)$, la TABLA DE FRECUENCIAS:

X	frecuencia absoluta n_i	frecuencia relativa f_i	frecuencia absoluta acumulada N_i	frecuencia relativa acumulada F_i
x_1	n_1	f_1	N_1	F_1
x_2	n_2	f_2	N_2	F_2
x_3	n_3	f_3	N_3	F_3
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
x_i	n_i	f_i	N_i	F_i
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
x_k	n_k	f_k	N_k	F_k
SUMA	$\sum_{i=1}^{k} n_i = N$	$\sum_{i=1}^{k} f_i = 1$		

Se verifica que:

- La suma de las frecuencias absolutas n_i es igual al número total de casos (N).

$$\sum_{i=1}^{k} n_i = N$$

- La suma de las frecuencias relativas f_i es igual a la unidad.

$$\sum_{i=1}^{k} f_i = f_1 + f_2 + \dots + f_i + \dots + f_k = 1$$

siendo, $f_i = \frac{n_i}{N}$.

Cuando realizamos la observación de una muestra o población, se pueden presentar los siguientes casos:

1º.- Que se hagan pocas observaciones y, en consecuencia, la variable estadística toma pocos valores.

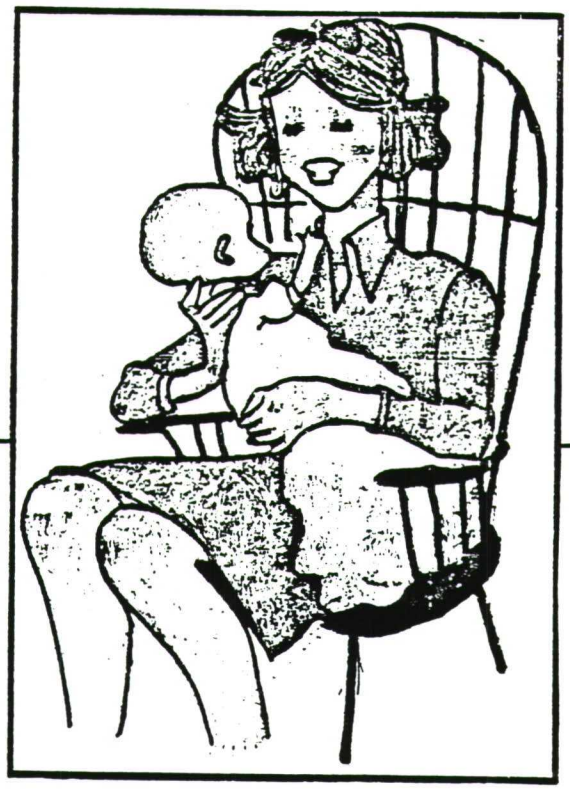
Este es un caso dentro del estudio de una VARIABLE ESTADISTICA DISCRETA.

2º.- Que se hagan muchas observaciones y, no obstante, la variable estadística toma pocos valores ya que los valores se repiten mucho (con mucha frecuencia).

Es un caso dentro del estudio de una VARIABLE ESTADISTICA DISCRETA.

3º.- Que se hagan muchas observaciones y la variable estadística tome muchos valores distintos.

En este caso, agrupamos los valores de la variable estadística en "intervalos" que deben elegirse convenientemente para no perder mucha información.



ACTIVIDAD - 2

En un determinado hospital se han ido anotando, durante una semana, el número de metros que el niño es capaz de andar el primer día que comienza a caminar. Obteniéndose así la tabla de información:

número de metros que el niño anda	0'5	1	2	1'5	2'25	1	1'5	2'25
niños	1	2	3	4	5	6	7	8

Sea la variable X = "niños".

Es una variable estadística discreta, ya que toma ocho valores.

La TABLA DE FRECUENCIA es:

x_i	frecuencia absoluta n_i	frecuencia relativa f_i	frecuencia absoluta acumulada N_i	frecuencia relativa acumulada F_i
1	0'5	0'5/12 = 0'042	0'5	0'5/12
2	1	1/12 = 0'083	1'5	1'5/12
3	2	2/12 = 0'167	3'5	3'5/12
4	1'5	1'5/12 = 0'125	5	5/12
5	2'25	2'25/12 = 0'1875	7'25	7'25/12
6	1	1/12 = 0'083	8'25	8'25/12
7	1'5	1'5/12 = 0'125	9'75	9'75/12
8	2'25	2'25/12 = 0'1875	12	12/12
SUMA	$\sum_{i=1}^{8} n_i = 12$	$\sum_{i=1}^{8} f_i = 1$		

Observa que se verifica:-

- La suma de las frecuencias absolutas n_i es igual al número total de casos (metros) N.

$$\sum_{i=1}^{8} n_i = 12 = N$$

- La suma de las frecuencias relativas f_i es igual a la unidad.

$$\sum_{i=1}^{8} f_i = 1$$

- Narutalmente, la última frecuencia absoluta acumulada N_8 tiene que "coincidir" con el número de observaciones realizadas N, es decir:

$$N_8 = 12 = N$$

- Análogamente, la última frecuencia relativa acumulada F_8 tiene que "coincidir" con la suma de todas las frecuencias relativas.

$$F_8 = 1 = \sum_{i=1}^{8} f_i$$

"Las variables estadísticas discretas se representan gráficamente por:

- Diagrama de barras
- Polígono de frecuencias
- Diagrama de frecuencias acumuladas".

REPRESENTACIONES GRAFICAS DE VARIABLES ESTADISTICAS DISCRETAS

• DIAGRAMA DE BARRAS (o de rectángulos)

Se representan en las abscisas (eje de las x) los distintos valores de la variable y sobre cada uno de los valores se levantan rectángulos de bases iguales, cuya altura es la frecuencia (absoluta o relativa) de cada valor. De esta forma, obtenemos un conjunto de barras (o rectángulos) verticales cuya suma de longitudes debe ser N ó 1, dependiendo de si las frecuencias representadas son absolutas ó relativas.

x_i	n_i	f_i
1	0'5	0'042
2	1	0'083
3	2	0'167
4	1'5	0'125
5	2'25	0'1875
6	1	0'083
7	1'5	0'125
8	2'25	0'1875

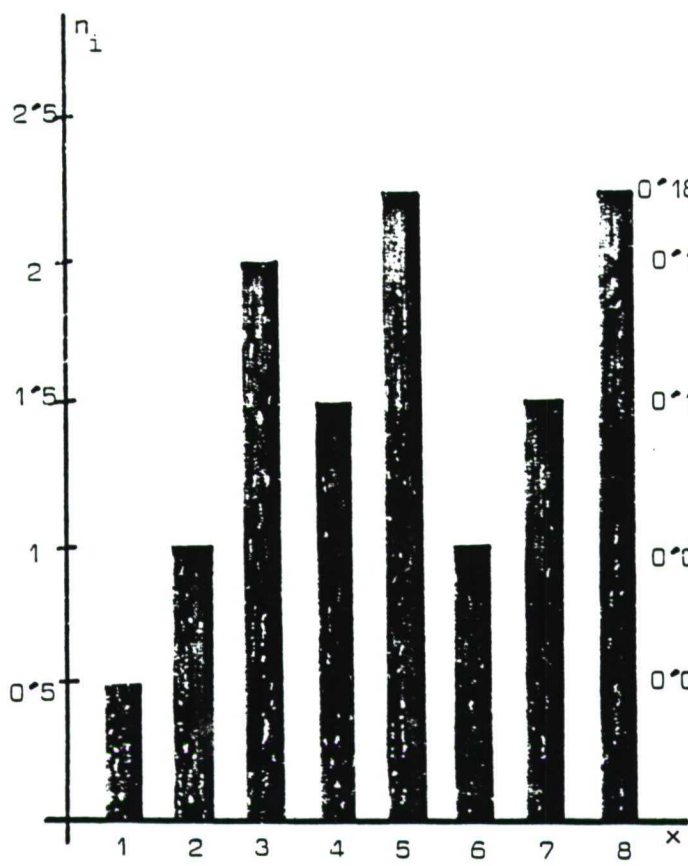


Diagrama de barras
de la frecuencia absoluta

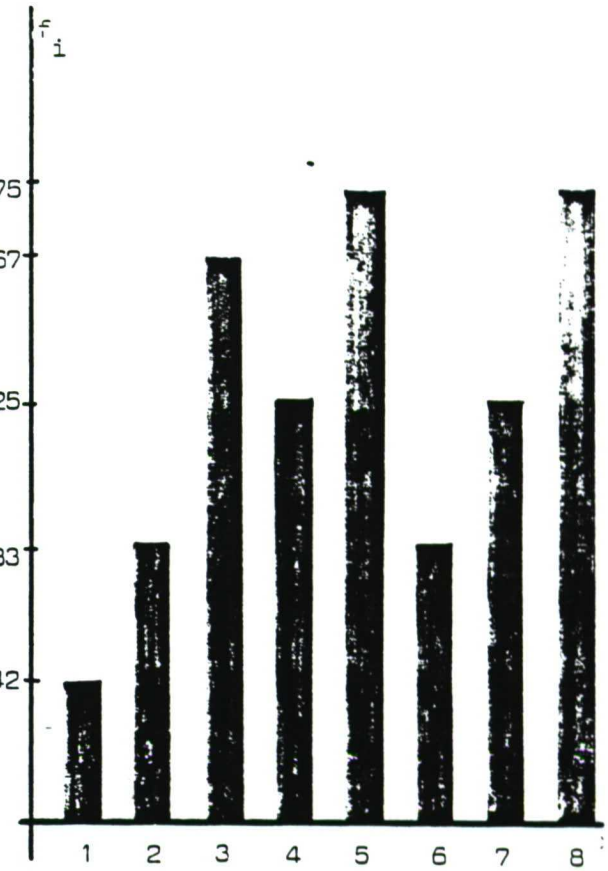
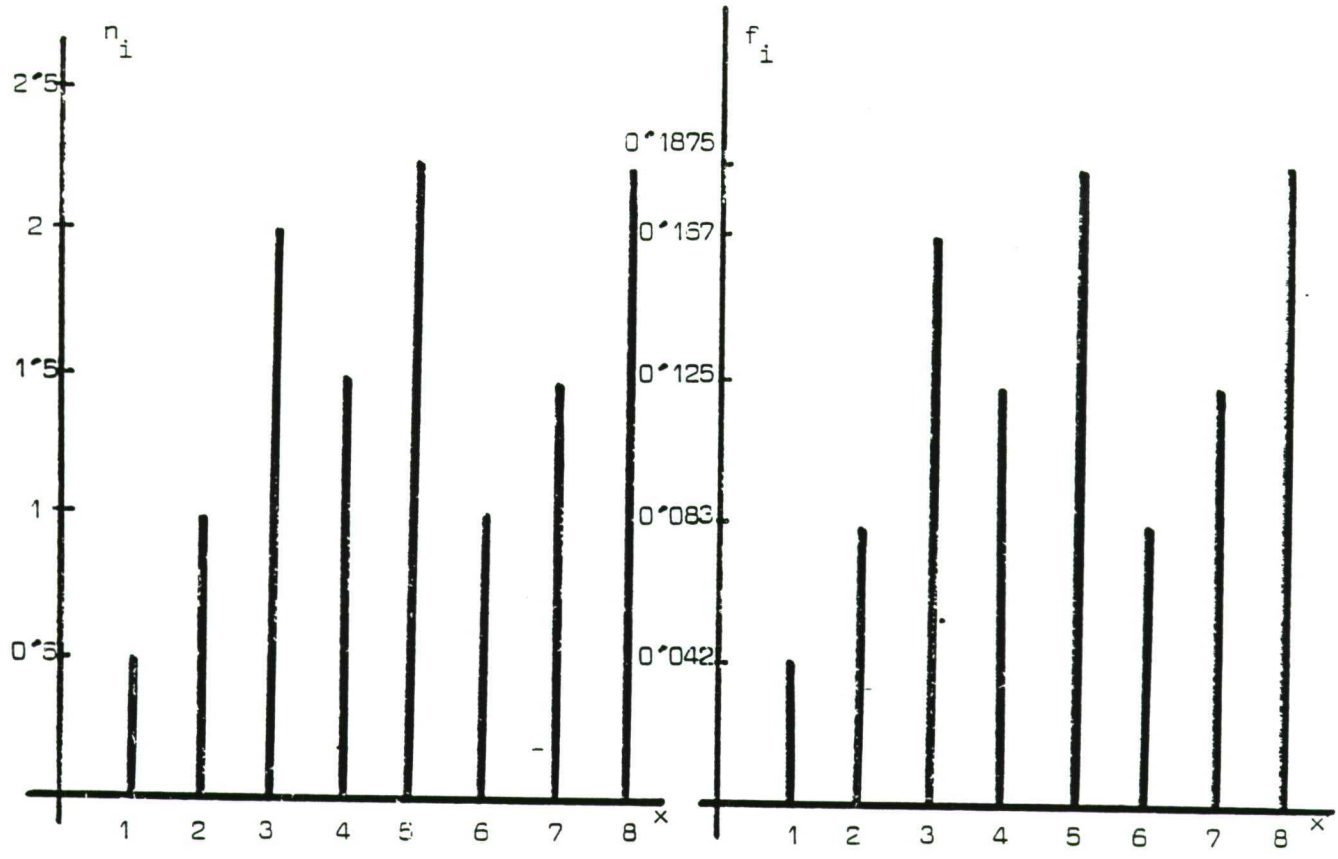


Diagrama de barras
de la frecuencia relativa

Esta forma de representar los diagramas de barras se utilizan a menudo para llamar la atención de todo el público que los vea.

La representación es válida cuando se levanta una línea perpendicular.



- POLIGONO DE FRECUENCIAS

El polígono de frecuencias se obtiene sin más que unir los extremos superiores de las barras en el diagrama de barras.

Veamos:

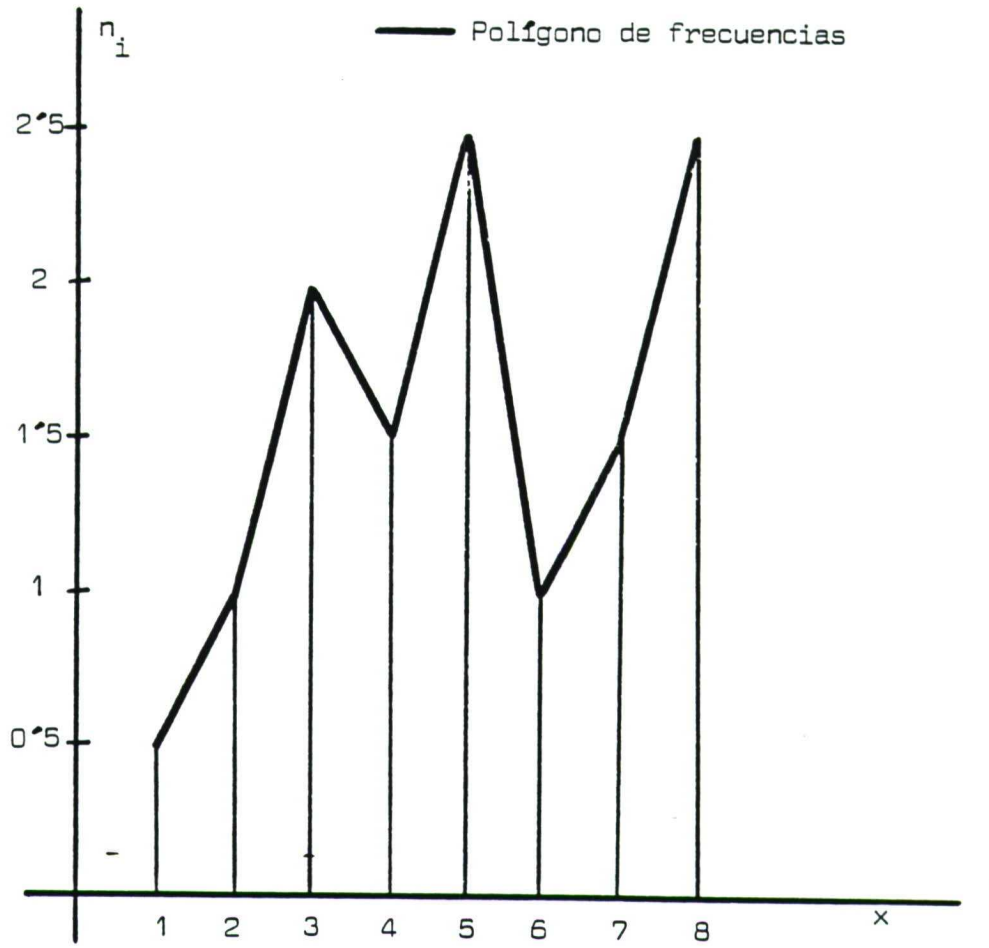


Diagrama de barras de la frecuencia absoluta

Al polígono de frecuencias se le conoce también como "gráfico de línea".

• DIAGRAMA DE FRECUENCIAS ACUMULADAS

Se colocan en las abscisas (eje de las x) los distintos valores de la variable estadística discreta y sobre cada uno de estos valores se levanta una perpendicular de longitud igual a la frecuencia (absoluta o relativa) acumulada correspondiente a ese valor. De esta forma aparecerá un "diagrama de barras creciente". Una vez obtenido este diagrama de barras se trazan segmentos horizontales de cada extremo de barra a cortar la barra inmediata a la derecha.

x_i	n_i	f_i	N_i	F_i
1	0'5	0'5/12	0'5	0'5/12
2	1	1/12	1'5	1'5/12
3	2	2/12	3'5	3'5/12
4	1'5	1'5/12	5	5/12
5	2'25	2'25/12	7'25	7'25/12
6	1	1/12	8'25	8'25/12
7	1'5	1'5/12	9'75	9'75/12
8	2'25	2'25/12	12	12/12
SUMA	12	1		

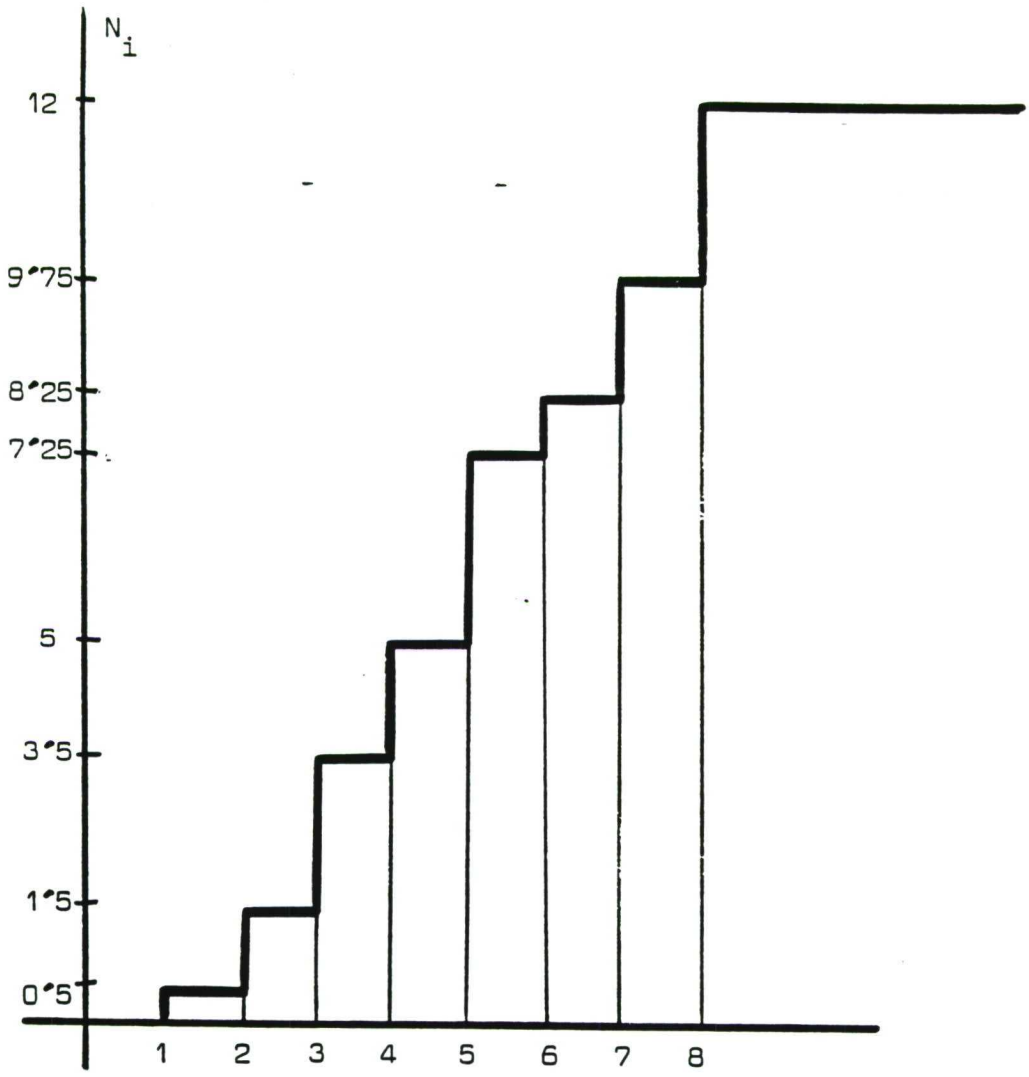


Diagrama de frecuencias absolutas acumuladas

- Dibuja como ejercicio el diagrama de frecuencias absolutas relativas F_i.

Observa que el diagrama de frecuencias acumuladas es continuo a la derecha, esto es, tiene forma de escalera hacia la derecha.

VARIABLES AGRUPADAS EN INTERVALOS DE CLASE

Al realizar la observación de una muestra o población, se puede presentar el siguiente caso:

> "Que se hagan muchas observaciones y la variable estadística tome muchos valores distintos".

Supongamos que el profesor de Matemáticas ha realizado una prueba de inteligencia abstracta a 20 alumnos, obteniendo las siguientes puntuaciones:

91	92	83	81	88
94	91	87	90	94
85	85	93	90	89
86	87	89	85	89

Observa que son muchos los valores distintos que toma la variable estadística "inteligencia abstracta", en consecuencia:

> Estudiar la variable estadística "inteligencia" como si fuese discreta (valores aislados) no sería la forma adecuada.

Es aconsejable agrupar los datos en "intervalos" y hacer un recuento de las

observaciones que caen dentro de cada uno de ellos.

Ahora bien, el estudiar el intervalo (donde pueden caer varios valores) y no cada uno de los valores de la variable tiene un doble significado, a saber:

- nuestro trabajo se simplifica
- perdemos información.

Por tanto, se hace necesario elegir un número de intervalos que equilibre estos dos aspectos.

PROCEDIMIENTO PARA LA CONSTRUCCION DE INTERVALOS

• Amplitud del intervalo: Se calcula restando el valor máximo y el mínimo de cada intervalo. Distinguiremos dos casos:

intervalos de amplitud constante: todos los intervalos tienen la misma amplitud.

intervalos de amplitud variable: los intervalos no tienen la misma amplitud.

Por comodidad elegiremos intervalos de amplitud constante salvo que el enunciado de la Actividad diga lo contrario.

• Elección de los intervalos de clase:

- Se ordenan los valores obtenidos en orden creciente.
- Fijamos una determinada amplitud para cada intervalo.
- Tomamos intervalos semiabiertos de la forma $[a,b)$, de forma que contienen siempre el menor de los valores "a", pero no al superior "b":

$$b - a = \text{amplitud}$$

- Obtenemos la información que contiene cada intervalo, calculando la "MARCA DE CLASE", esto es, el punto medio de cada intervalo:

$\frac{b + a}{2}$ es la "marca de clase" del intervalo $[a,b)$.

- Sean $a_0, a_1, a_2, a_3, a_4, \ldots, a_k$ los k datos obtenidos en la observación y ordenados en orden creciente, es decir, $a_0 < a_1 < a_2 < \ldots < a_k$.

La tabla de frecuencias de una variable agrupada en intervalos es:

TABLA DE FRECUENCIAS

Intervalos	Marcas de clase x_i	n_i	f_i	N_i	F_i
$[a_0,a_1)$	$x_1 = (a_0 + a_1)/2$	n_1	f_1	N_1	F_1
$[a_1,a_2)$	$x_2 = (a_1 + a_2)/2$	n_2	f_2	N_2	F_2
$[a_2,a_3)$	x_3	n_3	f_3	N_3	F_3
$[a_3,a_4)$	x_4	n_4	f_4	N_4	F_4
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
$[a_{l-1},a_l)$	x_l	n_l	f_l	N_l	F_l

Recuerda que en un intervalo semiabierto $[a_{l-1},a_l)$ se encuentra el menor de los valores a_{l-1} pero no se encuentra el superior a_l. Por esta razón, el extremo superior del último intervalo $[a_{l-1},a_l)$ debe ser mayor que el último de los valores a_k de la variable. Es decir:

$$a_k < a_l$$

ACTIVIDAD - 3

Veinte alumnos han obtenido en una prueba de inteligencia abstracta las siguientes puntuaciones:

91	92	83	81	88
94	91	87	90	94
85	85	93	90	89
86	87	89	85	89

Se pide:

Agrupar los datos en cuatro intervalos de amplitud constante. Construir la tabla de frecuencias.

- La ordenación de los valores obtenidos en orden creciente:

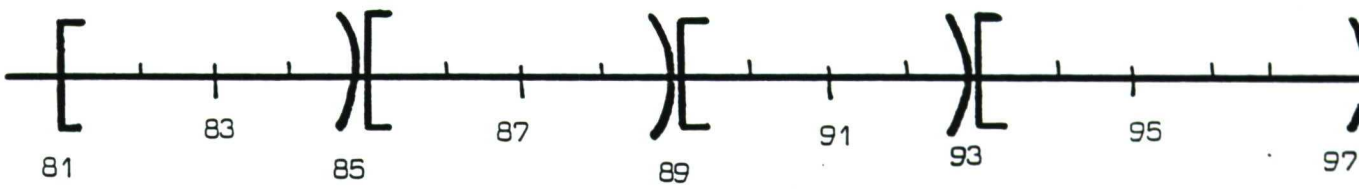
81	83	85	85	85
86	87	87	88	89

89 89 90 90 91

91 92 93 94 94

- Tomamos intervalos semiabiertos $[a,b)$ de manera que el valor inferior "a" pertenezca al intervalo y el valor del extremo superior "b" no. En consecuencia, el último intervalo no podrá tener como extremo superior 94 (último valor de la variable estadística), sino un valor superior a él.

Gráficamente:



se tienen los intervalos semiabiertos:

$[81,85)$ $[85,89)$ $[89,93)$ $[93,97)$

amplitud constante para los intervalos = 4

marcas de clase:

83 87 91 95

- El recuento de los valores que caen dentro de cada intervalo, se obtiene:

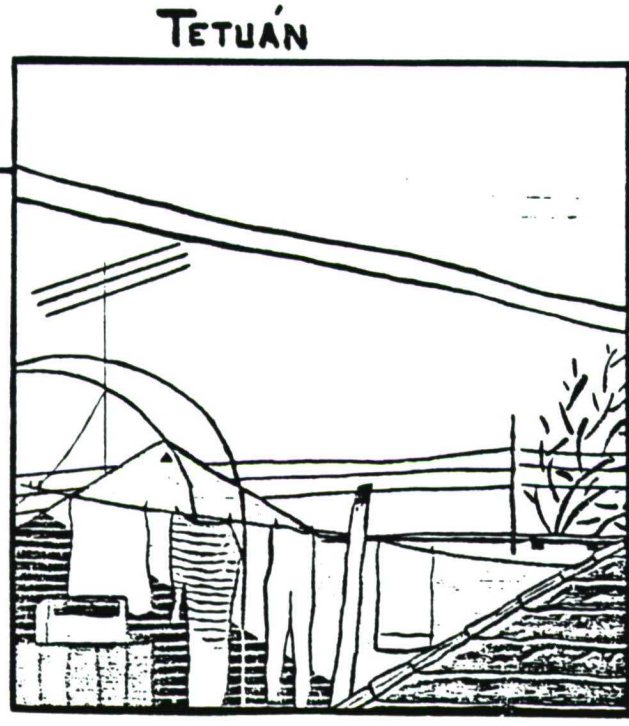
Intervalos	Valores que caen dentro	Recuento	n_i
$[81,85)$	81, 83	II	2
$[85,89)$	85, 85, 85, 86, 87, 87, 88	𝍸 II	7
$[89,93)$	89, 89, 89, 90, 90, 91, 91, 92	𝍸 III	8
$[93,97)$	93, 94, 94	III	3
		SUMA 20	

- TABLA DE FRECUENCIAS

Intervalos	Marcas de clase x_i	n_i	f_i	N_i	F_i
[81,85)	83	2	2/20	2	2/20
[85,89)	87	7	7/20	9	9/20
[89,93)	91	8	8/20	17	17/20
[93,97)	95	3	3/20	20	20/20
SUMA		20	1		

"Las variables agrupadas en intervalos de clase se representan gráficamente por:

- Histograma
- Polígono de frecuencias".



ACTIVIDAD - 4

En el barrio de Tetuán se han tomado las edades a 26 niños, obteniéndose:

3	7	10	10	6
5	4	12	11	10
15	10	6	2	10
9	10	8	15	13
14	12	7	10	6
8				

Se pide:

1) Distribuir las edades en intervalos de amplitud 4.

2) Construir la tabla de frecuencias.

1) Cuando se dispone de gran número de datos, es práctico el distribuirlos en intervalos y determinar el número de individuos pertenecientes a cada intervalo (recuento).

- Ordenamos las observaciones obtenidas en orden creciente:

2	3	4	5	6
6	6	7	7	8
8	9	10	10	10
10	10	10	10	11
12	12	13	14	15
15				

- Observa que el recorrido de la variable estadística X = "edad" es 15 - 2 = 13.

Como el recorrido de la variable X es 13, para saber aproximadamente el número de intervalos que hemos de construir dividimos 13 entre la longitud o amplitud de cada intervalo y obtenemos:

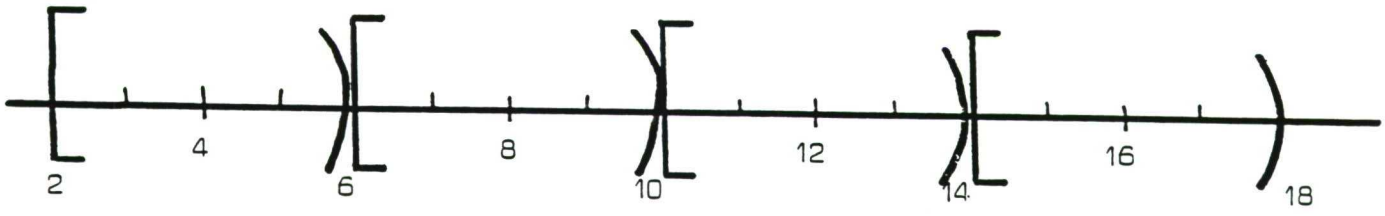
$$\text{número de intervalos} = \frac{\text{recorrido de la variable}}{\text{amplitud de intervalo}} =$$

$$= \frac{13}{4} = 3'25 \text{ (cuatro intervalos)}$$

es decir, hemos de construir cuatro intervalos.

- Tenemos que hacer cuatro intervalos semiabiertos de la forma $[a,b)$, tales que contienen siempre al inferior de los valores "a", pero no al superior "b". Es decir, empezamos en el 2 y haciéndolos semiabiertos $[a,b)$ de amplitud 4 terminamos en el 18.

Gráficamente:



Resultan los intervalos:

$[2,6)$ $[6,10)$ $[10,14)$ $[14,18)$

siendo la amplitud de cada intervalo = 4.

Fíjate que el extremo superior del último intervalo $[14,18)$, b = 18, debe ser mayor que el último de los valores de la variable.

- Las "marcas de clase" serán:

4 8 12 16

(punto medio de cada intervalo)

- Hacemos el "recuento" de los valores que caen dentro de cada intervalo:

Intervalos	Valores que caen dentro	Recuento	n_i
[2, 6)	2, 3, 4, 5	IIII	4
[6,10)	6, 6, 6, 7, 7, 8, 8, 9	𝍸 III	8
[10,14)	10, 10, 10, 10, 10, 10, 10, 11, 12, 12, 13	𝍸 𝍸 I	11
[14,18)	14, 15, 15	III	3
		SUMA ... 26	

2) TABLA DE FRECUENCIAS

Intervalos	Marcas de clase x_i	n_i	f_i	N_i	F_i
[2, 6)	4	4	4/26	4	4/26
[6,10)	8	8	8/26	12	12/26
[10,14)	12	11	11/26	23	23/26
[14,18)	16	3	3/26	26	26/26
SUMA		26	1		

Recuerda:

- El número de intervalos de una variable estadística viene dado por la expresión:

$$\text{nº intervalos} = \frac{\text{recorrido de la variable}}{\text{amplitud del intervalo}} \quad \text{(aproximado)}$$

se supone que los intervalos son de amplitud constante.

- Las variables agrupadas en intervalos de clase se representan gráficamente por:

  - Histograma
  - Polígono de frecuencias.

REPRESENTACIONES GRAFICAS DE VARIABLES AGRUPADAS EN INTERVALOS

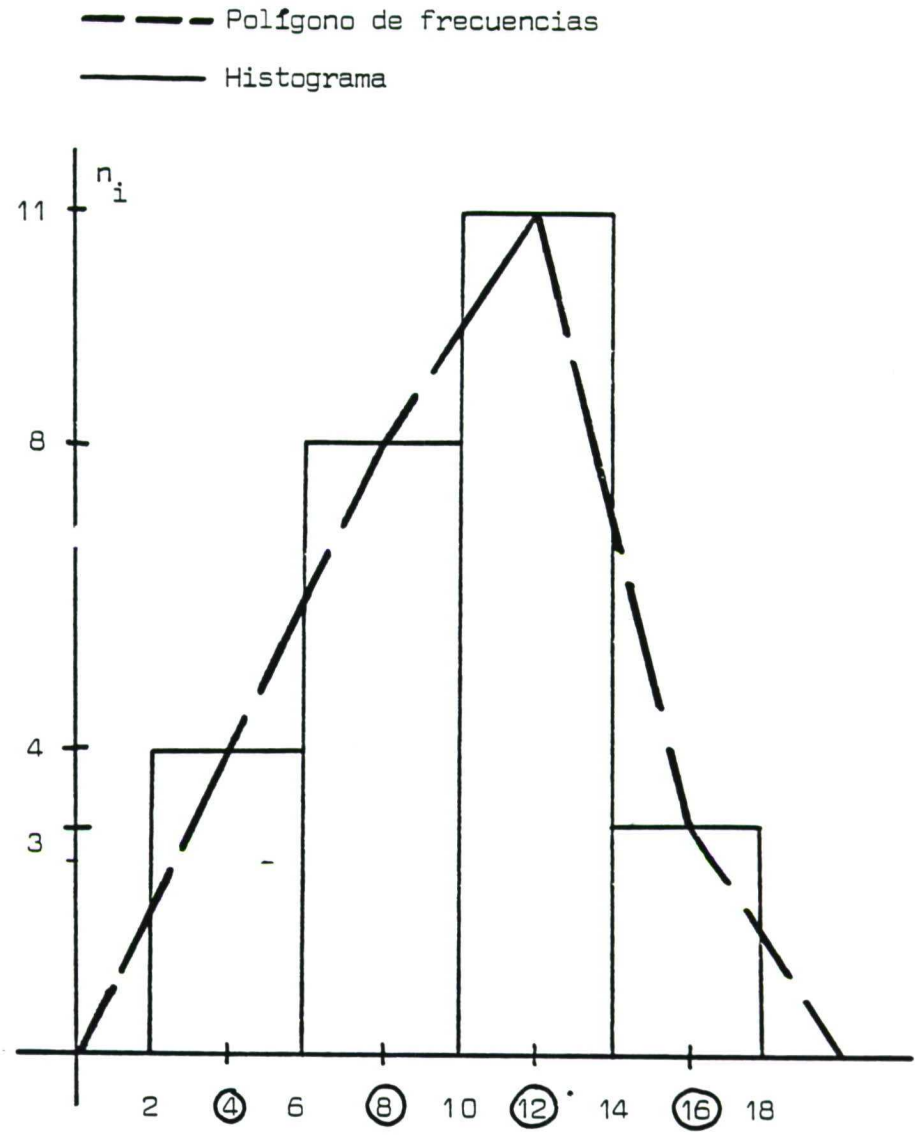
- HISTOGRAMA

Consiste en levantar rectángulos, de tal manera que, cada base es la amplitud de cada intervalo y la altura la frecuencia de dicho intervalo.

- EL POLIGONO DE FRECUENCIAS

Se obtiene uniendo los puntos medios de las bases superiores de cada rectángulo en el histograma.

Intervalos	Marcas de clase x_i	n_i
[2, 6)	4	4
[6, 10)	8	8
[10, 14)	12	11
[14, 18)	16	3



Observa como el polígono de frecuencias (línea quebrada) empieza y termina en el eje horizontal, esto es:

empieza en la marca de clase del intervalo anterior

termina en la marca de clase del intervalo posterior.

CUADRO REPRESENTACIONES

GRAFICAS

VARIABLES	REPRESENTACIONES GRAFICAS
Cualitativas: No pueden ser medidas. Sus valores se presentan como modalidades.	DIAGRAMA DE BARRAS DIAGRAMA DE SECTOR PICTOGRAMA
Discretas: Son aquellas que toman valores aislados, y no pueden tomar ningún valor entre dos valores consecutivos fijados.	DIAGRAMA DE BARRAS POLIGONO DE FRECUENCIAS DIAGRAMA DE FRECUENCIAS ACUMULADAS
Agrupadas en intervalos: Es el tratamiento que se da a las variables cuando toman muchos valores distintos.	HISTOGRAMA POLIGONO DE FRECUENCIAS

El único objetivo de las representaciones gráficas es ayudar a visualizar la información recogida. No olvides que con las representaciones gráficas debemos tener mucho cuidado, puesto que si están mal construídas o no son las adecuadas pueden llevarnos a conclusiones falsas. Es aconsejable, por tanto, emplear las técnicas estadísticas apropiadas para que la visualización no nos lleve a una mentira, o una gran mentira.

Piensa que la estadística no miente si la técnica empleada es la correcta.

ACTIVIDAD – 1: Dibuja el histograma y el polígono de frecuencias de la prueba de inteligencia abstracta que se realizó a veinte alumnos, siendo las puntuaciones:

91	92	83	81	88
94	91	87	90	94

85 85 93 90 89

86 87 89 85 89

ACTIVIDAD - 2: En un examen de Estadística los alumnos han obtenido las siguientes puntuaciones:

16, 15, 18, 17, 16, 21, 22, 17, 16, 13, 15, 23, 17, 24, 25, 26, 27, 18, 18, 15, 16, 28, 30, 15, 15, 16, 17, 21, 21, 22.

Se pide:

1) Agrupar los datos en cuatro intervalos de amplitud constante.

2) Construir la tabla de frecuencias.

3) Dibuja el histograma.

ACTIVIDAD - 3: ¿Crees que es correcta la representación gráfica del PICTOGRAMA en el cómic?.

ACTIVIDAD - 4: En un pueblo de Segovia se ha entrevistado a los varones mayores de veinticinco años. Obteniéndose la siguiente tabla de información:

Estado civil	Varones
solteros	150
casados	180
viudos	20
separados	5

Se pide:

1) Diagrama de barras.

2) Diagrama de sector.

3) Pictograma.

ACTIVIDAD – 5: En una determinada Clínica se han ido anotando, durante un mes, el número de palabras que el niño dice cuando comienza a hablar. Obteniéndose la siguiente tabla de información:

Número de palabras	25	30	10	50	30	25	10	10
niños	1	2	3	4	5	6	7	8

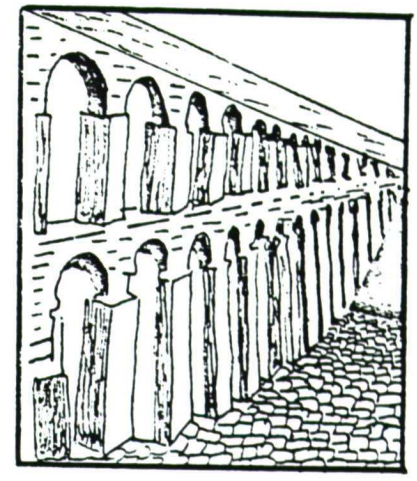
Se pide:

1) Tabla de frecuencias.

2) Diagrama de barras para frecuencias absolutas.

3) Diagrama de barras para frecuencias absolutas acumuladas.

"UN EMPLEADO FURIOSO"

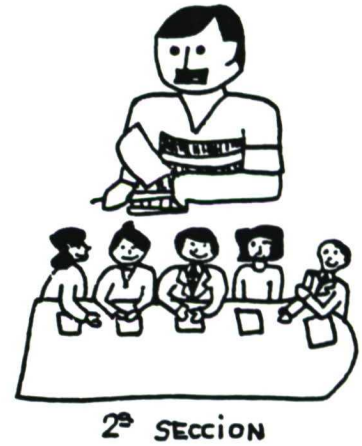


"Muebles Quintana" es una pequeña fabrica en Fuenterrebollo (Segovia).

La empresa familiar está formada por un sobrino, dos tíos y cinco parientes, constituídos en Sociedad Anónima.

La dirección está a cargo del sobrino D. Luís Fernández Quintana.

El bloque laboral está formado por diez operarios y dos encargados de sección.



El año 1.986 es bueno para los negocios, y una de las secciones requiere un empleado más.

El Sr. Quintana entrevista a Raúl que necesita un puesto de trabajo.

SR. QUINTANA : Aquí pagamos bien. El salario medio es de 80.000 pesetas mensuales.
Durante los dos primeros meses de adiestramiento sólo cobrará usted 60.000 pesetas, pero luego le subimos la mensualidad.

Pasan quince días. Raúl furioso, se entrevista con su jefe.

RAUL : ¡ Me ha engañado usted, Sr. Quintana !. He hablado con los otros operarios y ninguno gana más de 50.000 pesetas mensuales. ¿ Cómo puede ser de 80.000 pesetas el salario medio ?.

El Sr. Quintana no comprende la actitud de Raúl y al cabo de un momento, contesta:

SR. QUINTANA : Vamos, vamos, Raúl, no se excite. El salario medio es de 80.000 pesetas. Venga usted a la Dirección, que se lo voy a demostrar.

DIRECCION : Mire usted la nómina mensual.

NOMINA MENSUAL DE " MUEBLES QUINTANA "

Sr. Quintana	330.000,00 pts.
2 tíos a 150.000 pts.	300.000,00 pts.
5 parientes a 70.000 pts.	350.000,00 pts.
2 encargados a 60.000 pts.	120.000,00 pts.
10 operarios a 50.000 pts.	500.000,00 pts.
Total ....1.600.000,00 pts.	

Yo gano 330.000 pts., mis tíos 300.000 pts., mis cinco parientes sacan 70.000 pts. cada uno, los encargados salen a 60.000 pts. y los diez operarios a 50.000 pts. cada uno.
El total mensual es de 1.600.000 pts. para repartir entre 20 personas, ¿estoy equivocado?.

RAUL : Perdón, Sr. Quintana, tiene usted razón. El promedio es de 80.000 pts. Pero, aún así, ¡ usted me ha engañado !.

SR. QUINTANA : No estoy de acuerdo, Raúl. Todavía no ha comprendido nada. Pude haber ido diciéndole los salarios por orden. El salario medio sería entonces de 70.000 pts. Pero, fíjese bien Raúl, eso no es la media, sino la mediana.

RAUL : ¿ Entonces, las 50.000 pts. ?

SR. QUINTANA : Entienda Raúl que el salario ganado por el máximo número de personas es la "moda". La equivocación de usted radica en que no distingue entre media, mediana y moda.

RAUL : ¡ Ya sé la diferencia ...!. ¿ Pero usted, Sr. Quintana, no comprende que mi familia tiene que comer ?. Págueme los quince días de trabajo, ¡ me despido !.

El Sr. Quintana hace ver una serie de confusiones frecuentes entre media, mediana y moda. Esto nos hace pensar que a veces los resultados estadísticos dan información equivocada.

La palabra "media" es la abreviatura de "media aritmética", es una medida central muy valiosa. En ocasiones, cuando los valores extremos son muy — dispares (330.000 - 50.000), como en "Muebles Quintana", el "salario medio" da una impresión falsa.

Es fácil encontrar situaciones donde la media nos conduce a error. En todas ellas, los valores extremos son poco significativos y producen desajustes que desequilibran la información del conjunto de datos.

Como aspecto positivo, hay que señalar que reduce toda la información a un solo valor, tiene en cuenta los datos de todo el conjunto y su cálculo es sencillo.

Las informaciones sobre algunas estadísticas realizadas, producen aún — mayor desconcierto, a causa de que "término medio" se aplica en ocasiones no a la media aritmética, sino a la mediana o a la moda.

La "mediana" es el valor central (cuando el número de datos es impar) o la media aritmética (cuando el número de datos es par) de los dos valores centrales, supuestos todos ordenados en orden de magnitud.

A Raúl la mediana le ofrece una información más real que la media aritmética, pero, aún así, la mediana le da una imagen deformada de los salarios de su empresa.

Realmente a Raúl le conviene saber "la moda", que es el salario que más personas perciben.

En el lenguaje de "andar por la vida", frases como "caso típico" suelen referirse a la moda, pues son aquellos que se presentan con más frecuencia que ningún otro.

RECUERDA

NOTACION SUMATORIA $\sum$

El símbolo $\sum$ es la letra griega mayúscula "sigma", y se utiliza para expresar la suma de los distintos valores de una variable $X(x_1, x_2, x_3, \ldots, x_k)$, es decir, por definición:

$$\sum_{i=1}^{k} x_i = x_1 + x_2 + x_3 + \ldots + x_k$$

La letra i en x_i, la cual puede representar cualquiera de los números 1, 2, 3, ..., k, se llama "subíndice".

Así pues $\sum_{i=1}^{k} x_i$ se lee como "sumatorio de x_i", donde i toma los valores desde 1 hasta k.

$$\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4$$

$$\sum_{i=1}^{4} a = a + a + a + a = 4a$$

cada desarrollo del sumatorio $\sum_{i=1}^{4}$ tendrá tantos sumandos como el número que indica el límite de los valores de i.

PROPIEDADES DE $\sum$:

1. El sumatorio de una constante a es k veces la constante:

$$\sum_{i=1}^{k} a = \underbrace{a + a + a + \ldots + a}_{\text{k veces}} = ka$$

Así pues:

$$\sum_{i=1}^{3} 5 = 5 + 5 + 5 = 3 \cdot 5 = 15$$

2. El sumatorio de una constante a por una variable X, es igual a la constante por el sumatorio de la variable:

$$\sum_{i=1}^{k} a.x_i = ax_1 + ax_2 + ax_3 + \dots + ax_k =$$

$$= a\,(x_1 + x_2 + x_3 + \dots + x_k) = a.\sum_{i=1}^{k} x_i$$

Así pues:

$$\sum_{i=1}^{4} 3x_i = 3.\sum_{i=1}^{4} x_i = 3(x_1 + x_2 + x_3 + x_4)$$

3. El sumatorio de una suma (o diferencia) es igual a la suma (o diferencia) de los sumatorios:

$$\sum_{i=1}^{k} (x_i + y_i) = \sum_{i=1}^{k} x_i + \sum_{i=1}^{k} y_i$$

$$\sum_{i=1}^{k} (x_i - y_i) = \sum_{i=1}^{k} x_i - \sum_{i=1}^{k} y_i$$

Así pues:

$$\sum_{i=1}^{4} (x_i + y_i) = (x_1 + y_1) + (x_2 + y_2) + (x_3 + y_3) + (x_4 + y_4) =$$

$$= (x_1 + x_2 + x_3 + x_4) + (y_1 + y_2 + y_3 + y_4) = \sum_{i=1}^{4} x_i + \sum_{i=1}^{4} y_i$$

$$\sum_{i=1}^{3} (x_i - 2) = (x_1 - 2) + (x_2 - 2) + (x_3 - 2) =$$

$$= (x_1 + x_2 + x_3) - (2 + 2 + 2) = \sum_{i=1}^{3} x_i - \sum_{i=1}^{3} 2$$

MEDIDAS DE TENDENCIA CENTRAL

En ocasiones es conveniente reducir la información obtenida a un solo valor o a un pequeño número de valores para facilitar la comparación entre distintas muestras. De alguna manera estos valores centralizan la información y — reciben el nombre de MEDIDAS DE TENDENCIA CENTRAL.

> Medidas de tendencia central: Son medidas que nos dan el valor de la variable hacia el cual tienden a agruparse los datos. Los más utilizados son: la MEDIA, la MEDIANA y la MODA.

LA MEDIA ARITMETICA $\bar{x}$

La media aritmética o media de un conjunto de k datos $x_1, x_2, x_3, \ldots, x_k$ se representa por $\bar{x}$ y se define como la suma de todos los valores de la variable $X(x_1, x_2, x_3, \ldots, x_k)$ dividida por el número total de datos k.

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i}{k} = \frac{x_1 + x_2 + x_3 + \cdots + x_k}{k}$$

Sea: 2,3,4,5,6 (k= 5 datos)

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{k} = \frac{2 + 3 + 4 + 5 + 6}{5} = 4$$

El cálculo de la media sólo se puede aplicar a variables cuantitativas — (variables que son medibles).

Debemos distinguir:

a) DATOS NO AGRUPADOS

El valor de la variable aparece una sola vez $x_1, x_2, x_3, \ldots, x_k$.

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i}{k}$$

b) DATOS AGRUPADOS

El valor de la variable aparece repetidas veces, es decir, el conjunto k de datos $x_1, x_2, x_3, \ldots, x_k$ se presentan $n_1, n_2, n_3, \ldots, n_k$ veces (frecuencias), respectivamente.

$$\bar{x} = \frac{x_1 \cdot n_1 + x_2 \cdot n_2 + x_3 \cdot n_3 + \ldots + x_k \cdot n_k}{N} = \frac{\sum_{i=1}^{k} x_i \cdot n_i}{N}$$

donde

— $N = n_1 + n_2 + n_3 + \ldots + n_k$ es el número total de casos.

— (n_i) es el número de veces que aparece repetido cada valor x_i de la variable estadística X y que llamaremos FRECUENCIA ABSOLUTA.

Sean los valores 2, 2, 2, 3, 3, 4, 4, 4. Agrupamos los valores en una tabla:

x_i	n_i	$x_i \cdot n_i$
$x_1 = 2$	$n_1 = 3$	$2.3 = 6$
$x_2 = 3$	$n_2 = 2$	$3.2 = 6$
$x_3 = 4$	$n_3 = 3$	$4.3 = 12$
	$\sum_{i=1}^{3} n_i = 8$	$\sum_{i=1}^{3} x_i \cdot n_i = 24$

Observa que $N = \sum_{i=1}^{3} n_i = 8$ (número total de datos).

Luego:

$$\bar{x} = \frac{\sum_{i=1}^{3} x_i \cdot n_i}{N} = \frac{x_1 \cdot n_1 + x_2 \cdot n_2 + x_3 \cdot n_3}{N} = \frac{2.3 + 3.2 + 4.3}{8} = \frac{24}{8} = 3$$

$$\bar{x} = 3$$

CARACTERISTICAS DE LA MEDIA

- Es muy sensible en la variación de cada una de las puntuaciones. En otras palabras, basta con que varíe una sola puntuación para que varíe la media.

Ej. 5 , 8 , 6 , 1 $\quad \bar{x} = 5$

5 , 8 , 6 , 9 $\quad \bar{x} = 7$

- No es recomendable calcular la media cuando las puntuaciones extremas son muy dispares.

Ej. 2 , 3 , 7 , 500 $\quad \bar{x} = 128$

$\bar{x} = 128$ es "poco representativo"

LA MEDIANA (M_e)

La mediana es la medida central que, una vez ordenados los valores de la variable $x_1, x_2, x_3, \ldots, x_n$ en forma creciente o decreciente, deja igual número de observaciones (datos) inferiores que superiores a ella. Por tanto la mediana es el dato u observación que ocupa el valor central de la distribución estadística.

Sean los datos: 2, 4, 3, 6, 5

Colocados en orden creciente quedan: 2 3 **4** 5 6

(4: valor central)

Cuando el número de datos es un número par, la mediana es la semisuma de los valores centrales.

Sean los datos: 2, 4, 3, 6, 5, 7

Colocados en orden creciente quedan: 2 3 **4 5** 6 7

(4 5: valores centrales)

$$M_e = \frac{4 + 5}{2} = 4'5$$

CARACTERISTICAS DE LA MEDIANA

- Es menos sensible que la media. Pueden cambiar algunas puntuaciones y no variar la mediana.

Ej. 2 , 3 , 4 , 6 , 7 $\quad M_e = 4$

2 , 3 , 4 , 6 , 12 $\quad M_e = 4$

- La mediana es más representativa que la media cuando los valores extremos son muy dispares.

Ej. 2 , 3 , 7 , 500 $M_e = 7$

- La mediana suele ser aplicada a variables cuantitativas (que se pueden medir).

Estudiaremos la mediana para DATOS AGRUPADOS en el cómic "extraterrestres — muy particulares".

LA MODA (M_d)

La moda es el valor de la variable que tiene mayor frecuencia. La moda se — puede aplicar tanto a variables cualitativas (no son medibles) como a variables cuantitativas.

La moda no siempre es única; así, si hay dos modas, la distribución de datos se llama bimodal; si tres, trimodal, etc.

Sea: 2, 2, 2, 3, 3, 4
La moda $M_d = 2$ (valor de mayor frecuencia).

Sea: 2, 2, 2, 3, 3, 3, 4
Las modas $M_{d_1} = 2$ y $M_{d_2} = 3$ Es bimodal.

CARACTERISTICAS DE LA MODA

- Es muy sencilla de calcular, generalmente basta con observación de datos.

- Se calcula la moda cuando se trabaja en variables cualitativas (variables que no son medibles) o cuando por alguna razón no se puede calcular la media y la mediana.

- La moda es poco representativa. Sólo se debe calcular cuando los datos se repiten con mucha frecuencia (muchas veces).

Ej. 2 , 2 , 2 , 3 , 5 , 7 , 300 $M_d = 2$

Estudiaremos la moda para DATOS AGRUPADOS en el cómic "extraterrestres muy particulares".

Fijemonos en las características comparativas de la MEDIA, MEDIANA y MODA.

M E D I A	M E D I A N A	M O D A
Es muy sensible: Basta que varíe un solo valor para que varíe la media. Sea: 2,3,4,5,6 $\bar{x} = 4$ 2,3,4,5,11 $\bar{x} = 5$	Es menos sensible que la media: Pueden cambiar algunos datos y no variar la mediana. Sea: 2,3,4,5,6 $M_e = 4$ 2,3,4,5,11 $M_e = 4$	
No se debe calcular la media cuando los valores extremos son muy dispares. Sea: 2,2,3,5,300 $\bar{x} = 156$ "la media es poco representativa".	La mediana es más representativa que la media cuando los valores extremos son muy dispares. Sea: 2,2,3,5,300 $M_e = 3$	La moda es poco representativa. Sólo se debe calcular cuando los datos se repiten con mucha frecuencia. Sea: 2,2,3,5,300 $M_d = 2$
Se calcula la media cuando los datos están distribuidos simétricamente alrededor de un valor central. Sea: 3,4,6,8,9 $\bar{x} = 6$	Se calcula la mediana cuando se desea conocer el valor que deja por encima y por debajo el 50 % de los datos.	Se calcula la moda cuando se trabaja con variables cualitativas (caracteres no mediables) o cuando por alguna razón no se puede calcular la media y la mediana.
La media es el centro de gravedad de la distribución (colección de datos).	La mediana es el valor que divide el número total de datos colocados en orden creciente o decreciente en dos partes de igual área.	La moda es el valor de mayor frecuencia de la distribución.

DIFERENTES MEDIAS

- MEDIA ARITMETICA PONDERADA: A veces se asocia a los números $x_1, x_2, x_3, \ldots, x_k$ ciertos "pesos" $w_1, w_2, w_3, \ldots, w_k$ que dependen de la significación o importancia de cada uno de los números. En este caso

$$\bar{x} = \frac{x_1 \cdot w_1 + x_2 \cdot w_2 + x_3 \cdot w_3 + \ldots + x_k \cdot w_k}{w_1 + w_2 + w_3 + \ldots + w_k} = \frac{\sum_{i=1}^{k} x_i \cdot w_i}{N}$$

siendo $N = w_1 + w_2 + w_3 + \ldots + w_k$

se llama "media aritmética ponderada".

Ej. El profesor de matemáticas valora 3 veces más el examen final de curso que los exámenes parciales. Un estudiante tiene una nota de examen final de 4 y notas de exámenes parciales de 6 y 7. ¿Cual es su nota final?.

$$\bar{x} = \frac{4.3 + 6.1 + 7.1}{3 + 1 + 1} = 5$$

- MEDIA ARMONICA: La media armónica $\bar{x}_A$ de una serie de k números $x_1, x_2, x_3, \ldots, x_k$ que se presentan con $n_1, n_2, n_3, \ldots, n_k$ frecuencias, respectivamente, se define como

$$\bar{x}_A = \frac{N}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \frac{n_3}{x_3} + \ldots + \frac{n_k}{x_k}} = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$$

$$N = n_1 + n_2 + n_3 + \ldots + n_k$$

Si cada $x_1, x_2, x_3, \ldots, x_k$ se presenta una sola vez, entonces:

$$\bar{x}_A = \frac{K}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \ldots + \frac{1}{x_k}} = \frac{K}{\sum_{i=1}^{k} \frac{1}{x_i}}$$

Ej. La media armónica de los números 2, 4, 8 es:

$$\bar{x}_A = \frac{3}{\frac{1}{2} + \frac{1}{4} + \frac{1}{8}} = \frac{3}{\frac{7}{8}} = 3'42$$

Obsérvese que si algún valor x_i es cero, entonces la media armónica no tiene sentido.

- MEDIA GEOMETRICA: La media geométrica de una serie de k números $x_1, x_2, x_3, \ldots, x_k$ que se presentan con $n_1, n_2, n_3, \ldots, n_k$ frecuencias, respectivamente, se representa por $\bar{x}_G$ y se define como

$$\bar{x}_G = \sqrt[N]{x_1^{n_1} \cdot x_2^{n_2} \cdot x_3^{n_3} \cdots x_k^{n_k}}$$

donde $N = n_1 + n_2 + n_3 + \cdots + n_k$ número total de casos.

Como este cálculo es un tanto complicado, se procede tomando logaritmos, o bien manejando la calculadora.

Si cada número $x_1, x_2, x_3, \ldots, x_k$ aparece una sola vez, se tiene:

$$\bar{x}_G = \sqrt[k]{x_1 \cdot x_2 \cdot x_3 \cdots x_k}$$

Ej. La media geométrica de los números 2, 4, 8 es:

$$\bar{x}_G = \sqrt[3]{2 \cdot 4 \cdot 8} = \sqrt[3]{64} = 4$$

Esta medida de centralización tiene la dificultad de que sólo con que un valor sea cero, ya $\bar{x}_G$ sería nula y, por tanto, su representatividad sería dudosa. Sin embargo, para el manejo de algunos valores es muy útil.

- RELACION ENTRE MEDIAS ARITMETICA-ARMONICA-GEOMETRICA:

$$\bar{x}_A \leqslant \bar{x}_G \leqslant \bar{x}$$

La igualdad $\bar{x}_A = \bar{x}_G = \bar{x}$ se presenta sólo cuando los números $x_1, x_2, x_3, \ldots, x_k$ son idénticos.

Ej. Medias aritmética, armónica y geométrica de los números 2, 4, 8.

$\bar{x} = 4'66$

$\bar{x}_A = 3'42$

$\bar{x}_G = 4$

Comprueba la desigualdad $\bar{x}_A < \bar{x}_G < \bar{x}$

ACTIVIDAD – 1: Efectuar los desarrollos de los siguientes sumatorios:

$$\sum_{i=1}^{3} 3x_i \qquad \sum_{i=1}^{6} k \qquad \sum_{i=1}^{5} (x_i - 4)$$

ACTIVIDAD - 2: Las calificaciones de un estudiante en seis pruebas fueron 6, 5, 4, 7, 8 y 9. Hallar:

- la media aritmética de las calificaciones.
- la mediana de las calificaciones.

ACTIVIDAD - 3: Los tiempos de reacción de un individuo a estímulos sensoriales fueron 0'50, 0'47, 0'49, 0'62, 0'53, 0'44, 0'67 y 0'50 segundos, respectivamente. Determinar el tiempo medio de reacción del individuo a los estímulos.

ACTIVIDAD - 4: De un total de 100 números, 20 eran 4, 40 eran 5, 30 eran 6 y el resto eran 7. Hallar la media aritmética de los números.

ACTIVIDAD - 5: Cuatro grupos de estudiantes, formados por 10, 20, 15 y 25 - individuos, registran una media de alturas de 1'62, 1'52, 1'70 y 1'68 metros, respectivamente. Hallar la altura media de todos los estudiantes.

ACTIVIDAD - 6: Los salarios por hora de cinco empleados del INI son 100 pts, 130 pts, 68 pts, 500 pts, 1800 pts. Hallar:

- la mediana del salario horario.
- la media del salario horario.
- ¿Qué medida central da una mejor información?.

ACTIVIDAD - 7: Un hombre viaja de Madrid a Segovia (98 kms.) a una velocidad media de 80 kms. por hora y vuelve de Segovia a Madrid por la misma ruta con una velocidad media de 70 kms. por hora. Hallar la velocidad media del viaje completo. (La velocidad media es la media armónica).

ACTIVIDAD - 8: Hallar la media, mediana y moda para los siguientes conjuntos de datos:

a) 7, 4, 10, 9, 15, 12, 7, 9, 7.

b) 8, 11, 4, 3, 2, 5, 10, 6, 4, 1, 10, 8, 12, 6, 5, 7.

ACTIVIDAD - 9: Hallar la media geométrica y la media aritmética de los números: 2, 4, 8, 16, 32.

ACTIVIDAD - 10: Hallar dos números cuya media aritmética es 5 y cuya media geométrica es 4.

ACTIVIDAD - 11: Hallar la media aritmética, la media geométrica y la media armónica de los números: 0, 2, 4, 6.

ACTIVIDAD – 1: $\sum_{i=1}^{3} 3x_i = 3(x_1 + x_2 + x_3) = 3 \sum_{i=1}^{3} x_i$

$$\sum_{i=1}^{6} k = 6k$$

$$\sum_{i=1}^{5} (x_i - 4) = \left(\sum_{i=1}^{5} x_i \right) - 20$$

ACTIVIDAD – 2: $\bar{x} = 6'5$, $M_e = 6'5$

ACTIVIDAD – 3: $\bar{x} = 0'54$ seg.

ACTIVIDAD – 4: $\bar{x} = 5'30$

ACTIVIDAD – 5: $\bar{x} = 1'63$ metrcs. -

ACTIVIDAD – 6: $M_e = 130$ pesetas , $\bar{x} = 519'6$ pts.

es más representativa la mediana.

ACTIVIDAD – 7: $\bar{x}_A = 74'66$ Km/h velocidad media

ACTIVIDAD – 8: (a) $\bar{x} = 8'\hat{8}$, $M_e = 9$, $M_d = 7$

(b) $\bar{x} = 6'375$, $M_e = 6$. Puesto que los números 4, 5, 6, 8 y 10 aparecen dos veces, se puede considerar que estas son las cinco modas, es más razonable decir que no existe moda.

ACTIVIDAD – 9: $\bar{x}_G = 8$, $\bar{x} = 12'4$

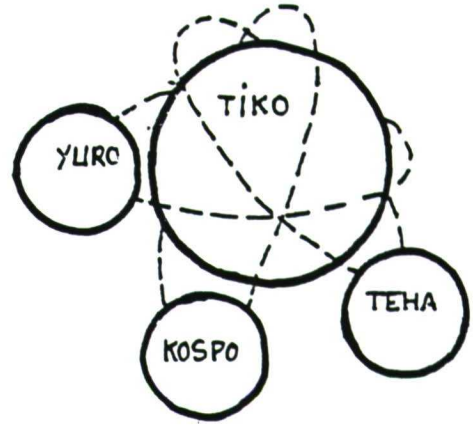
ACTIVIDAD – 10: Los números son 2 y 8.

ACTIVIDAD – 11: $\bar{x} = 3$, $\bar{x}_G = 0$, $\bar{x}_A = 0$

No tiene sentido hablar de media geométrica y media armónica puesto que un valor es cero.

"EXTRATERRESTRES MUY PECULIARES"

Salimos de la Tierra con la idea de explorar el planeta TIKO. Los astrónomos no dudaban de la existencia de tres pequeños satélites girando en torno a él, pero eran escépticos en cuanto a la existencia de vida humana.

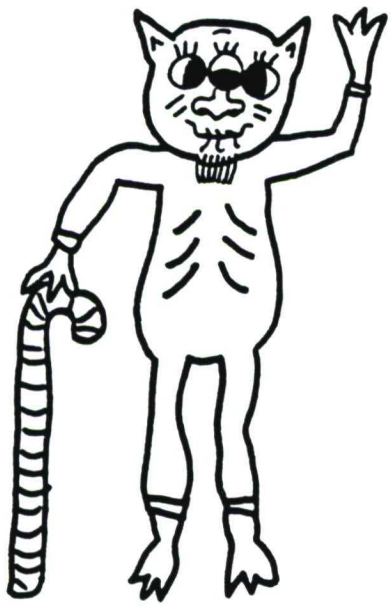


Nosotros intuíamos que sí la había y decidimos aventurarnos.

A continuación os contamos lo que sucedió.

Llegada al primer satélite: YURO

Los "yuristas" son verdes y tienen tres ojos, pero aún siendo tan distintos a nosotros fuimos capaces de notar algo familiar, hasta que uno del grupo dijo: ¡Se mueven todos como mi bisabuelo Rodolfo!.



Esa fue la clave: En general eran muy viejos, se desplazaban despacio, la mayoría con bastón y tenían muchas arrugitas.

Cuando sintonizamos nuestros lenguajes pudimos conocer su modo de vida y también... ¡sus edades!.

Efectivamente, en un pueblo de 250 yuristas, había mucha más gente anciana que la que nunca habíamos conocido. La edad media era de 296 años.

Sea:

- N = "número de yuristas en el grupo" = 250.
- La variable X = "edad de los habitantes de Yuro"

$$X = (x_1 = 100,\ x_2 = 200,\ x_3 = 300,\ x_4 = 400,\ x_5 = 500)$$

- n_i = "número de veces que aparece repetida cada edad x_i de la variable X"

$$(n_1 = 35,\ n_2 = 40,\ n_3 = 80,\ n_4 = 90,\ n_5 = 5)$$

observa $\sum_{i=1}^{5} n_i = n_1 + n_2 + n_3 + n_4 + n_5 = 250$

A n_i se llama "FRECUENCIA ABSOLUTA".

- N_i = "suma de los n_i inferiores o iguales a él"

$N_1 = n_1 = 35$, $N_2 = n_1 + n_2 = 35 + 40 = 75$

$N_3 = n_1 + n_2 + n_3 = 35 + 40 + 80 = 155$

$N_4 = 245$, $N_5 = 250$

Observa $N_5 = 250$

A N_i se llama "FRECUENCIA ABSOLUTA ACUMULADA"

La tabla de frecuencias es:

x_i	n_i	N_i	$x_i \cdot n_i$
100	35	35	3.500
200	40	75	8.000
300	80	155	24.000
400	90	245	36.000
500	5	250	2.500
	$\sum_{i=1}^{5} n_i = 250$		$\sum_{i=1}^{5} x_i \cdot n_i = 74.000$

- La edad media $\bar{x}$ es la media aritmética de edades del grupo de yuristas y se define como:

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i . n_i}{N} = \frac{100.35 + 200.40 + 300.80 + 400.90 + 500.5}{250} =$$

$$= \frac{74.000}{250} = 296 \text{ años.}$$

- La mediana M_e es la edad que ocupa el lugar central, supuestas las edades ordenadas en forma creciente, es decir, la edad que deja igual número de observaciones (edades) inferiores que superiores a ella. Veamos:

35 veces | 40 veces | 80 veces | 90 veces | 5 veces

100,.....,100, 200,.....,200, 300,.....,300, 400,.....,400, 500,.....,500

125

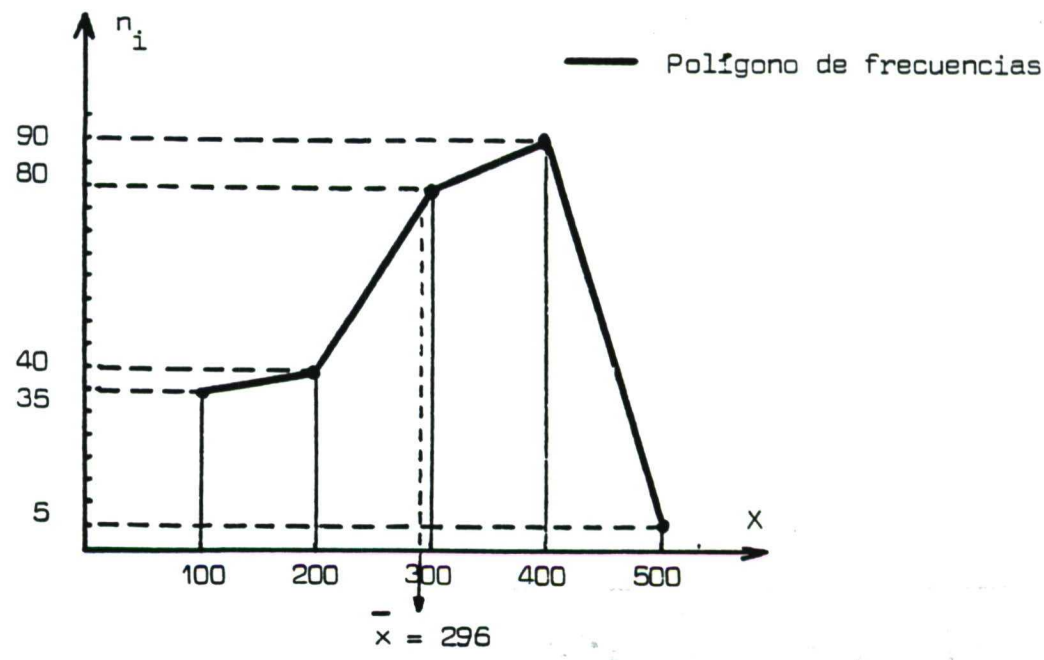
Observa: $\frac{N}{2} = \frac{250}{2} = 125$ observaciones.

- El sujeto que ocupa la posición 125 tiene 300 años. Por tanto, $M_e = 300$.

- La moda M_d es la edad que se repite mayor número de veces (o que tiene máxima frecuencia)

$$M_d = 400$$

EL DIAGRAMA DE BARRAS para la frecuencia absoluta n_i de los habitantes de Yuro es:



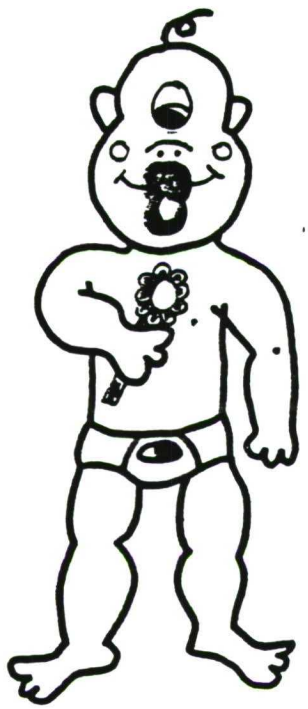
Observa: El polígono de frecuencias se obtiene uniendo los extremos superiores de las barras.

Esta gráfica se considera asimétrica a la izquierda o negativa porque presenta "cola" a la izquierda, es decir, las frecuencias descienden más lentamente por ese lado que por el derecho, en el que la "caída" es más brusca.

Se cumple que $\bar{x} \leq M_e \leq M_d$

Y esto es lo que sucede en el satélite de nuestros amigos "yuristas".

Llegada al segundo satélite: KOSPO



El segundo viaje nos trajo una nueva sorpresa, si en el anterior todos los habitantes eran muy mayores, aquí había el mismo alboroto que en una guardería. La diferencia era que estos —— "bebés" eran azules y con un solo ojo. Nos costó lo nuestro interrogar a un pueblo de "kospitas" porque no tenían aún muy desarrollado su lenguaje, pero la información sobre sus edades es muy curiosa; fijaos:

x_i	n_i	N_i	$x_i \cdot n_i$
100	5	5	500
200	90	95	18.000
300	80	175	24.000
400	40	215	16.000
500	35	250	17.500
$\sum_{1}^{5}$			$\sum_{1}^{5} x_i \cdot n_i = 76.000$

x_i = edad en días

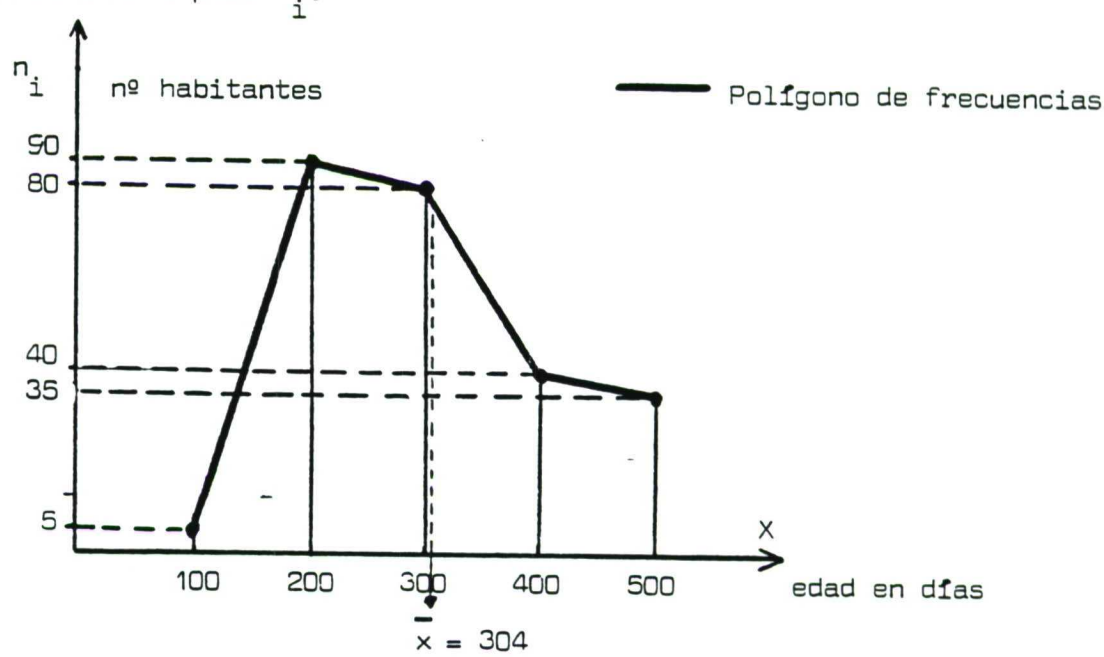
n_i = nº de habitantes

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i \cdot n_i}{N} = \frac{76.000}{250} = 304$$

$M_e = 300$

$M_d = 200$

DIAGRAMA DE BARRAS para n_i:



Al contrario que en la anterior, aquí la asimetría es a la derecha o positiva, ya que la "cola" queda a la derecha, es decir, las frecuencias descienden más lentamente por ese lado que por el izquierdo en el que la "subida" es más brusca.

Se cumple que: $\bar{x} \geqslant M_e \geqslant M_d$

A nuestros amigos los bebés-kospita les quedan los datos igual a la distribución mencionada.

Llegada al tercer satélite: TEHA

El tercer satélite de Tiko apenas nos sorprendió: ¡Pocas cosas ya nos podían asombrar...!

Nos llegamos a sentir como en casa, porque su ambiente y sus habitantes eran muy parecidos a nosotros.

Aparte de ser un poco más bajitos, flacos y de color amarillo:

Se veían familias completas con miembros de todas las edades. Mediante la fácil comunicación que establecimos, llegamos a conocer su vida y sus costumbres, una de ellas era celebrar sus "cumpletrimestres", ya que así es como controlan su edad.

Interrogando a un pueblo de tehistas, elaboramos una tabla de frecuencias:

x_i	n_i	N_i	$x_i \cdot n_i$
100	30	30	3.000
200	60	90	12.000
300	90	180	27.000
400	60	240	24.000
500	30	270	15.000
	$\sum_{i=1}^{5} n_i = 270$		$\sum_{i=1}^{5} x_i \cdot n_i = 81.000$

x_i = edad en trimestres

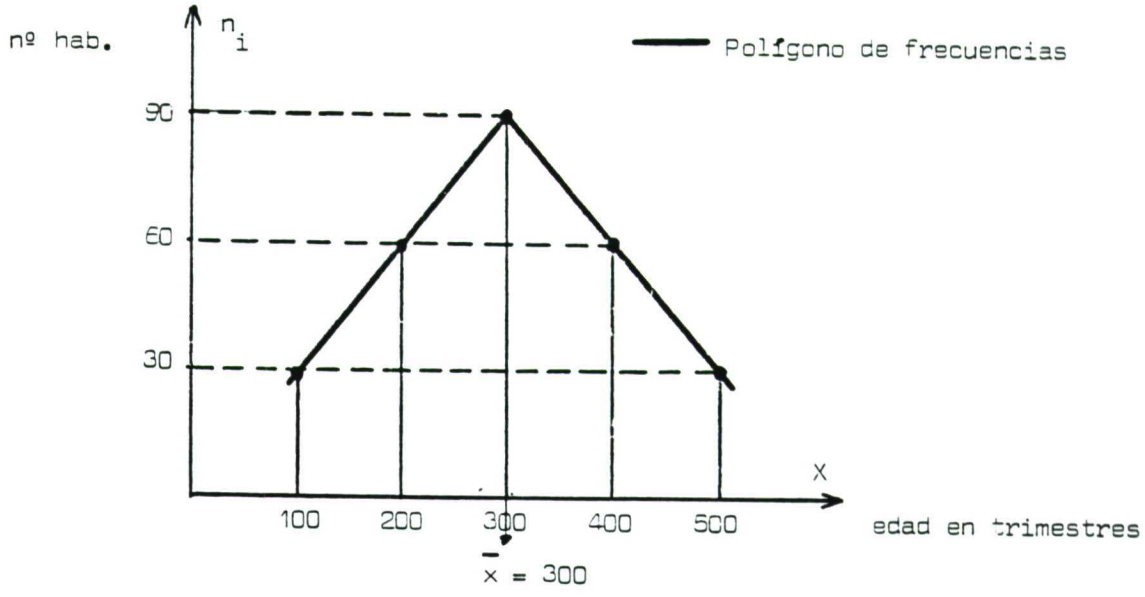
n_i = nº de habitantes

$\bar{x} = 300$

$M_e = 300$

$M_d = 300$

DIAGRAMA DE BARRAS para n_i:



Vemos que esta distribución es simétrica porque los valores de las variables equidistantes de un valor central tienen las mismas frecuencias.

En este caso: $\bar{x} = M_e = M_d$

¡Ah!... Seguro que quereis saber como acabó nuestra aventura - ¿no es así?.

Muy sencillo: nos van a devolver la visita.

¿No sobrará en tu casa alguna cama vacía?.

R E C U E R D A

- FRECUENCIA ABSOLUTA (n_i): Llamaremos frecuencia absoluta de un valor x_i de la variable estadística X al número de veces que aparece repetido dicho valor en el conjunto de observaciones realizadas.

- FRECUENCIA ABSOLUTA ACUMULADA (N_i): La frecuencia absoluta acumulada de un valor x_i es la suma de las frecuencias absolutas (n_i) de los valores o iguales a x_i. Es decir:

$$N_1 = n_1 \ , \quad N_2 = n_1 + n_2 \ , \quad N_3 = n_1 + n_2 + n_3 \ , \quad \ldots$$

Evidentemente, los valores de x_i han de estar ordenados de forma creciente y, por tanto, se tiene:

$$N_k = N$$

siendo N el número total de datos y $X(x_1, x_2, x_3, \ldots, x_k)$

- CALCULO PRACTICO DE LA MEDIANA:

1. Se representa el DIAGRAMA de frecuencias absolutas acumuladas N_i.
2. Dividimos el número de observaciones N entre 2.
3. Comprobamos si el número obtenido (N/2) se encuentra en la tabla de frecuencias absolutas acumuladas N_i.
4. En caso de no encontrarse, estará comprendido entre dos números (N_{k-1}, N_k) de la citada tabla, con lo cual la "mediana" será aquel valor de la variable X que corresponde al mayor N_k. En otras palabras, la "mediana" es el valor de la x_k que corresponde a N/2.

✱ En el caso concreto de los habitantes de Yuro, se tenía:

TABLA DE FRECUENCIAS

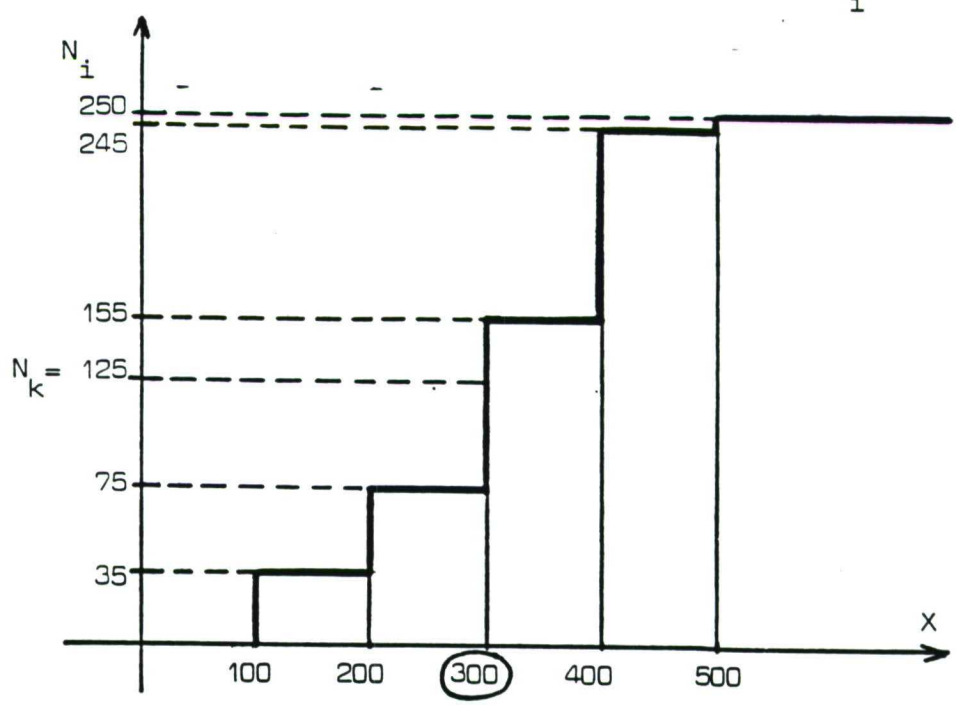
x_i	n_i	N_i	$x_i \cdot n_i$
100	35	35	3.500
200	40	75	8.000
300	80	155 (125)	24.000
400	90	245	36.000
500	5	250	2.500
	$\sum_{i=1}^{5} n_i = 250$		$\sum_{i=1}^{5} x_i \cdot n_i = 74.000$

x_i = edad en años.

n_i = nº habitantes que tienen x_i años.

de donde:

1. Diagrama de frecuencias absolutas acumuladas N_i.



CONSTRUCCION: La abscisa x_k tendrá una ordenada N_k. De esta forma aparece un diagrama de barras creciente. Trazando segmentos horizontales de cada extremo de barra a cortar la barra inmediata a su derecha se obtiene el diagrama de frecuencias acumuladas.

2. $\frac{N}{2} = \frac{250}{2} = 125$

3. Vamos a la columna N_i de la TABLA de frecuencias y observamos que 125 no se encuentra allí. Luego estará comprendido entre 75 y 155.

4. Vamos al Diagrama acumulativo de las frecuencias absolutas N_i y vemos a qué abscisa x_k corresponde la ordenada $N_k = N/2 = 125$.

La mediana $M_e = 300$.

✱ En otro pueblo de 310 yuristas nos encontramos con gente todavía más anciana. La edad media era de casi 334 años. Observa los datos recogidos en la tabla de frecuencias:

TABLA DE FRECUENCIAS

x_i	n_i	N_i	$x_i \cdot n_i$
100	35	35	3.500
200	35	70	7.000
300	85	155 (155)	25.500
400	100	255	40.000
500	55	310	27.500
	$\sum_{i=1}^{5} n_i = 310$		$\sum_{i=1}^{5} x_i \cdot n_i = 103.500$

x_i = edad en años

n_i = nº habitantes de x_i años

- La edad media $\bar{x}$ es la media aritmética de las edades del grupo y se define como:

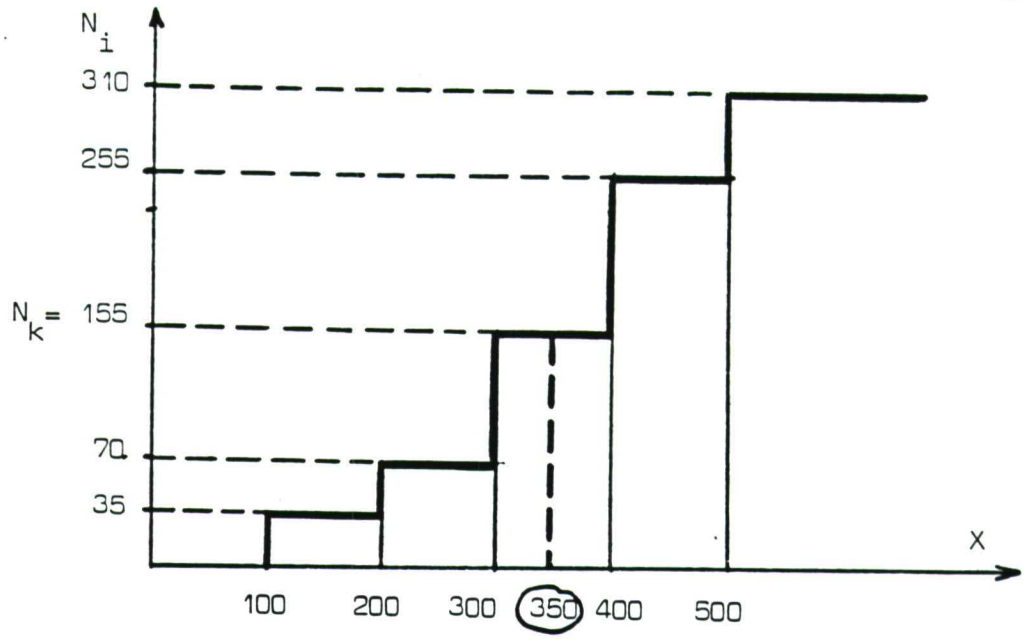
$$\bar{x} = \frac{\sum_{i=1}^{5} x_i \cdot n_i}{N} = \frac{103.500}{310} = 333'87 \text{ años}$$

- La moda M_d es la edad que se repite con mayor frecuencia (mayor número de veces):

$M_d = 400$

• Para el cálculo de la mediana M_e procedemos de la forma siguiente:

1. Representamos el DIAGRAMA de frecuencias absolutas acumuladas N_i.



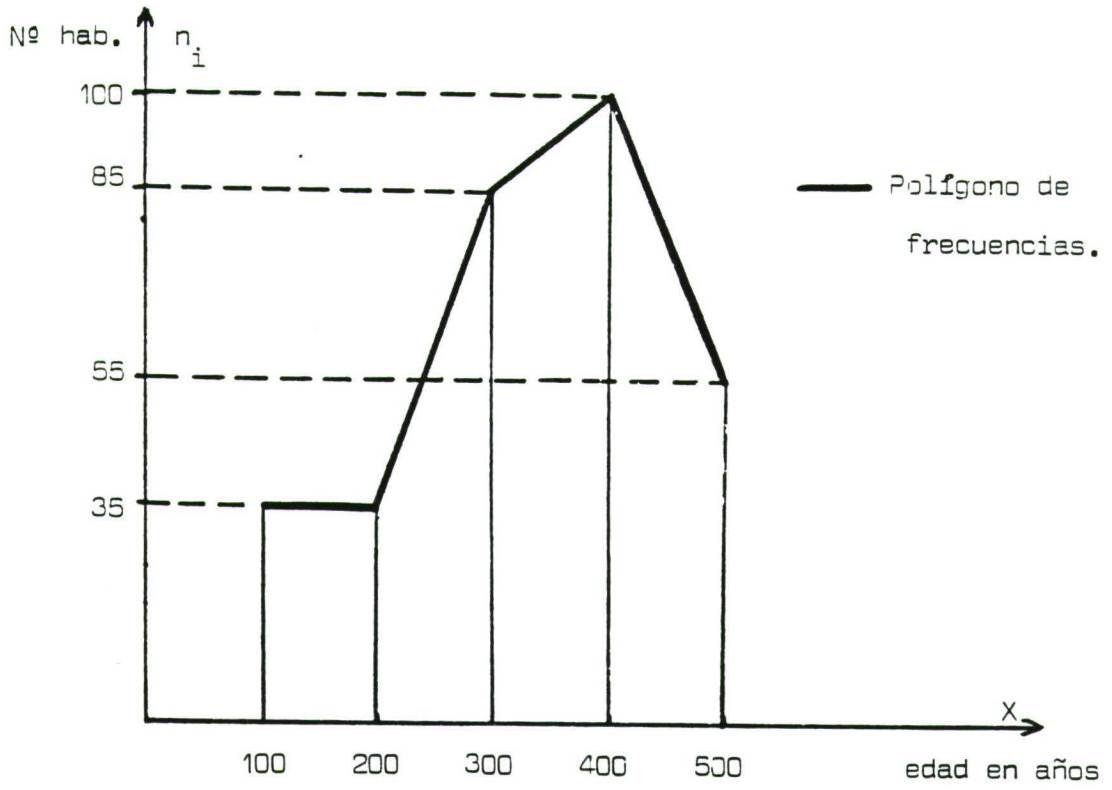
2. $\frac{N}{2} = \frac{310}{2} = 155$

3. Vamos a la columna N_i de la TABLA DE FRECUENCIAS y observamos que el valor $N/2 = 155$ coincide con la frecuencia absoluta acumulada del valor 300 y, por tanto, la ordenada 155 corresponde a todas las edades del intervalo (300,400).

Como la mediana es un valor (edad), la edad media de dicho intervalo será:

$$M_e = \frac{300 + 400}{2} = 350$$

• Para visualizar la información recogida representamos el polígono de frecuencias, uniendo los extremos superiores de las barras en el DIAGRAMA de barras para la frecuencia absoluta n_i.



Resulta una gráfica asimétrica a la izquierda ya que presenta cola a la izquierda, esto es, las frecuencias descienden más lentamente por la izquierda que por la derecha.

Observa: $\bar{x} < M_e < M_d$

ACTIVIDAD – 1: En un pueblo de "Teha" recogimos información sobre la edad de sus habitantes. Obteniéndo así la tabla adjunta:

edad en trimestres	100	200	300	400	500
número de habitantes	40	60	100	70	30

Se pide:

1) Tabla de frecuencias. Diagrama de barras para la frecuencia absoluta n_i. Diagrama de barras para la frecuencia absoluta acumulada N_i.

2) Media, mediana y moda.

3) Razonar el tipo de gráfica que resulta.

ACTIVIDAD - 2: En la clínica La Paz se han ido anotando durante un mes el número de metros que el niño anda el primer día que comienza a caminar, obteniendo:

número de niños	2	3	4	5	6	7	8
número de metros	6	10	5	10	3	4	4

Hallar:

1) Media, mediana y moda.

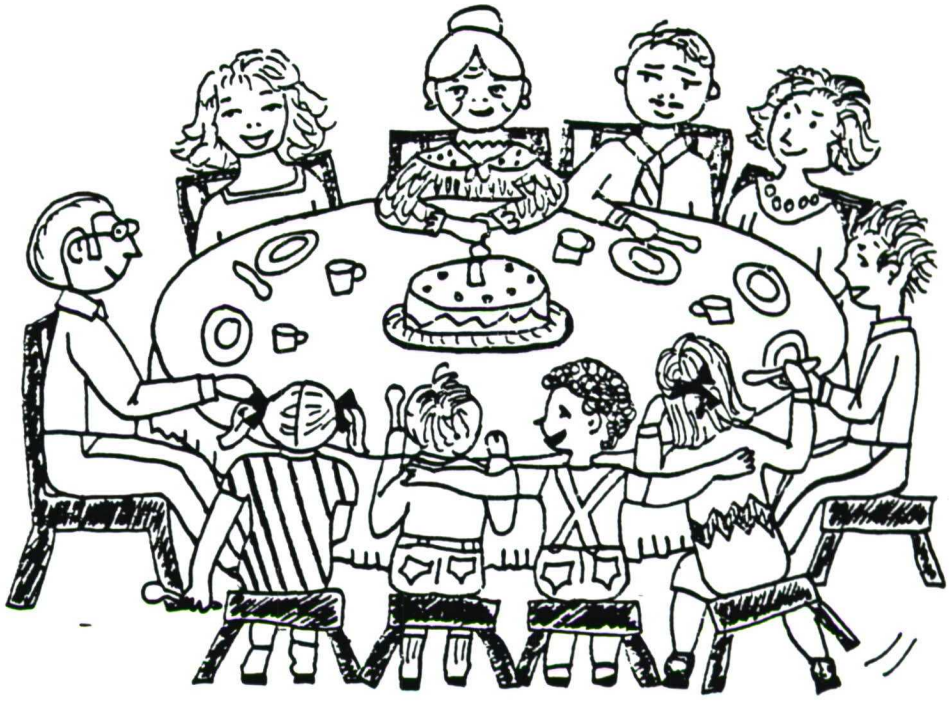
2) Razonar el tipo de gráfica que se obtiene.

ESTADISTICA DESCRIPTIVA

ACTIVIDADES

"LA FIESTA DE CUMPLEAÑOS"

El pasado fin de semana, mis primos: Joaquín, Elena, Susana, Rubén y yo, que soy Javier nos reunimos en casa de la abuela. Celebrábamos su cumpleaños y comimos junta toda la familia.



En la sobremesa, los mayores empezaron a hablar de cosas muy aburridas y decidimos irnos al Parque de Atracciones. Teníamos un problema: ni una peseta en el bolsillo. La abuela nos llamó y nos dijo: "Os daré dinero, pero... según vuestra edad".

Joaquín y Elena:

A nosotros que somos mellizos, nos tienes que dar lo mismo, tenemos seis años cada uno.

Abuela: Repartíos 200 pts. para los dos.

Susana: Abuela, yo tengo diez años.

Abuela: Pues toma 150 pts.

Rubén: ¡Yo tengo trece años!

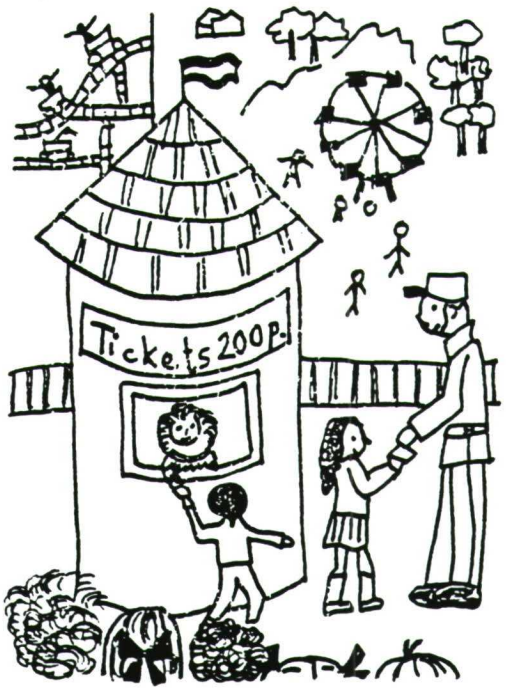
Abuela: Aquí tienes 300 pts.

Javier: A mí me tienes que dar más que a ellos. Tengo quince años y encima tengo que hacer de niñera.

Abuela: Toma 350 pts. y no protestes tanto.

Al llegar al Parque de Atracciones había un cartel que anunciaba:

" tickets para diez atracciones a 200 pts. "



El problema surgió de inmediato: Joaquín, Elena y Susana no tenían suficiente dinero para entrar. Si no venían ellos, nos fastidiaban la tarde, así que propuse: "Si juntamos todo el dinero y lo repartimos entre los cinco podremos comprar tickets para todos".

JOAQUÍN	ELENA	SUSANA	RUBÉN	JAVIER	TODOS
100	100	150	300	350	1000

1.000 pts. : 5 personas = 200 pts. cada una.

Rubén, el Pitagorín, dijo: Mi "profe" a eso lo llama hallar la media aritmética.

Javier: ¡Anda, calla y no seas listillo!. Siempre salgo yo perdiendo, de 350 pts. que llevo, voy a acabar con 200 pts.

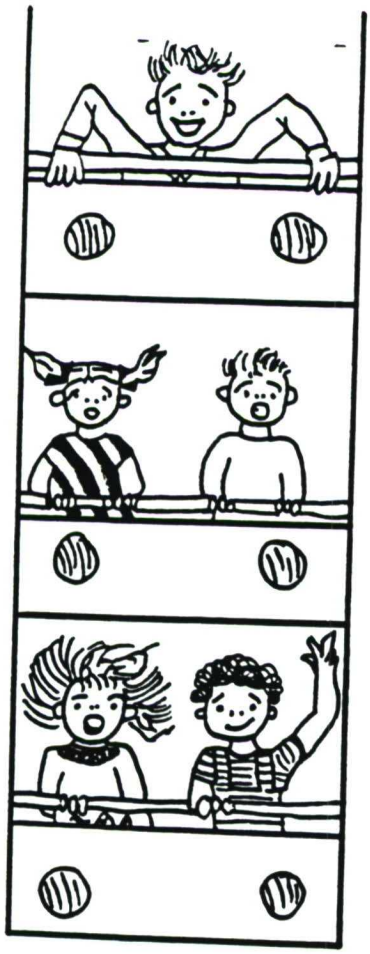
Rubén: No protestes porque es la única solución posible. Si llegamos a calcular la mediana, que es el dinero que tiene el que ocupa el

lugar central sólo dispondríamos de 150 pts. y no podríamos montarnos en nada.

JOAQUÍN	ELENA	SUSANA	RUBÉN	JAVIER
100	100	150	300	350

Javier: Ya, ya lo sé, y si calculáramos la moda, todavía nos iría peor, porque es el valor que más se repite y en este caso es 100 pts.

JOAQUÍN	ELENA	SUSANA	RUBÉN	JAVIER
100	100	150	300	350



Todos: Pasemos a divertirnos que es a lo que hemos venido.

"EL VIAJE FIN DE CURSO"

La clase de Javier prepara la excursión fín de curso.
Se ha propuesto que todos los lunes entregue cada uno el dinero ahorrado en el fín de semana. Así cuando llegué el día del viaje sus padres tendrán que darles menos dinero y ese no será un problema para dejarles ir.

Eduardo, el profesor de Matemáticas, explicó el lunes como iba el "banco" de la clase, y tomando como ejemplo a los grupos uno y dos aprovechó para ampliar algunas nociones de Estadística.

GRUPO 1		GRUPO 2	
TOÑO	470p.	ELISA	200p.
VICTORIA	500p.	JOSE	500p.
JUAN	530p.	JAVIER	800p.

JOSE:

Tenemos los dos grupos una media de 500 pts. Ya hay menos que añadir para el precio del viaje.

TOÑO:

Más o menos tenemos todos lo mismo...

ELISA:

¡Qué bien!. Yo pensaba que sólo me correspondían 200 pts. que es lo que entregué...

JAVIER:

¿Qué dices?. Si yo he reunido 800 pts. ¿Cómo me van a dar sólo 500 pts.?

PROFESOR:

Si os dais cuenta sólo os estáis quejando los del GRUPO 2. Si todos hubiérais entregado aproximadamente lo mismo, no estaría mal hablar de que os corresponde esa media de 500 pts. Pero, en el GRUPO 2 hay mucha diferencia de dinero entre unos y otros.

VICTORIA:

Entonces... ¿Hay situaciones en que la media no da buena idea de cómo están distribuidos los datos?.

¡Claro, eso es! La utilización de las medidas centrales: MEDIA, MEDIANA y MODA, depende de varios factores, pero el que nos afecta ahora es el de la dispersión de los datos. Cuando hay puntuaciones muy extremas la media por sí sola no refleja la realidad del grupo y conviene utilizar otras medidas adicionales.

Toda distribución de frecuencias tiene dos características que la definen:

—Un punto central alrededor del cual tienden a agruparse los datos: media, mediana y moda.

—Una variabilidad o dispersión de los datos respecto a ese valor central.

PROFESOR:

Las variables que describen la dispersión de una distribución de frecuencias son: la VARIANZA y la DESVIACION TIPICA, entre otras.
Sus fórmulas son:

- LA VARIANZA : $\sigma^2 = \frac{\Sigma(x_i-\bar{x})^2}{N}$

- LA DESVIACION TIPICA : $\sigma = \sqrt{\frac{\Sigma(x_i-\bar{x})^2}{N}}$

Cuando los datos están muy alejados de la media, la varianza y la desviación típica son grandes; por el contrario cuando los datos están muy proximos a la media, la varianza y la desviación típica son pequeñas, incluso valdrían cero si todos los datos de la distribución fueran iguales entre sí. Por tanto sólo es recomendable su cálculo en caso en que sea útil la media.

JAVIER:

¿Por qué no calculamos la varianza y la desviación típica en nuestros datos para comprobar lo que nos has explicado?.

GRUPO 1			GRUPO 2		
x_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	x_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$
470	-30	900	200	-300	90.000
500	0	0	500	0	0
530	30	900	800	300	90.000
$\bar{x}=500$		1.800	$\bar{x}=500$		180.000
$S^2=\frac{\Sigma(x_i-\bar{x})^2}{N}=600$			$S^2=\frac{\Sigma(x_i-\bar{x})^2}{N}=60.000$		
$S=\sqrt{S^2}=\sqrt{600}=10\sqrt{6}=24'49$			$S=\sqrt{S^2}=\sqrt{60.000}=100\sqrt{6}=244'94$		

JOSE:

En el primer grupo la desviación típica es de 24'49, mientras que en el segundo es de 244'94: ¡diez veces mayor!

PROFESOR:

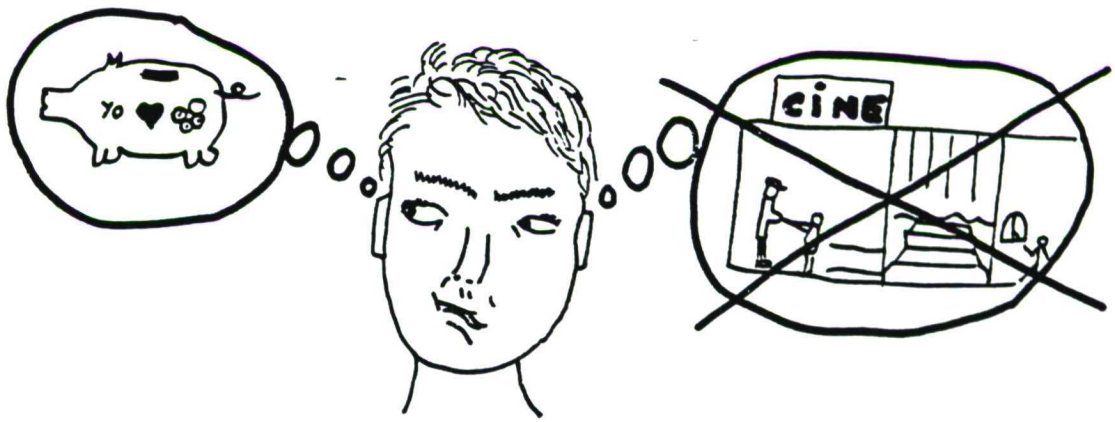
En efecto, la desviación típica del segundo grupo es mucho mayor, lo que hace concluir que la media no refleja totalmente la realidad del grupo.

ELISA:

¡Qué tristeza me quedo con mis pobres ahorros de 200 pts!.

JAVIER:

Vale chicos. Cada uno que arregle con su dinero lo que pueda y si le parece poco, menos cines y más ahorros.



RECUERDA

CONCEPTO DE MEDIDA DE DISPERSION

Ya dijimos que a veces es conveniente reducir toda la información obtenida a un solo valor o un número pequeño de valores. Estos valores que centralizan la información, reciben el nombre de "medidas de tendencia central".

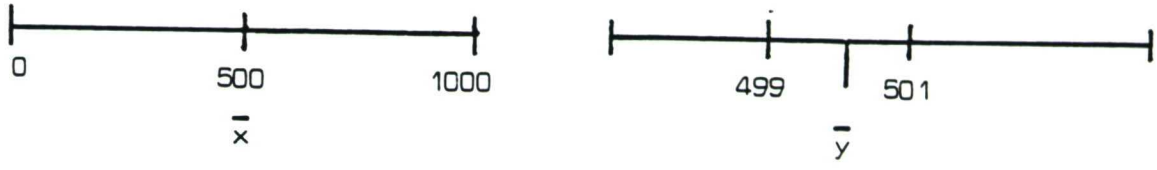
Es evidente que al resumir un conjunto de valores en uno solo se pierde información. Por ejemplo, consideremos las variables X e Y con sus respectivas frecuencias:

X	n_i
0	1
500	1
1000	1

Y	n_i
499	1
501	1

$$\bar{x} = \frac{0.1 + 500.1 + 1000.1}{3} = 500 \qquad \bar{y} = \frac{499.1 + 501.1}{2} = 500$$

En ambos casos la media aritmética es 500; sin embargo la variable X está mucho más dispersa que la variable Y.



Habrás observado tú mismo que al tratar gráficamente los conjuntos de datos, parece lógico pensar que la representatividad de $\bar{y}$ es mayor que la de $\bar{x}$.

Por tanto, se hace necesario cuantificar la representatividad de los valores centrales. Se ve, la necesidad de definir nuevas medidas estadísticas. Estas medidas reciben el nombre de "medidas de dispersión".

MEDIDAS DE DISPERSION

Toda distribución de frecuencias tiene dos "características" que las definen:

- Valor central alrededor del cual tienden a agruparse todos los valores: media, mediana y moda.
- Una dispersión de valores respecto a ese valor central.

En toda investigación sobre una muestra ambas medias descriptivas (centralización, dispersión) deben ir parejas para evitar conclusiones erróneas.

Las medidas de dispersión más utilizadas son: la varianza, la desviación típica, el coeficiente de variación de Pearson, el recorrido y el recorrido semiintercuartílico.

VARIANZA (σ^2)

La varianza σ^2 de una variable estadística X se define de la forma:

$$\sigma^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_i}{N}$$

desarrollando esta fórmula podemos obtener otra expresión que frecuentemente resulta más cómoda:

$$\sigma^2 = \frac{\sum_{i=1}^{k} x_i^2 \cdot n_i}{N} - \bar{x}^2$$

siendo:

x_i = distintos valores de la variable X.

$\bar{x}$ = la media aritmética.

N = número total de datos u observaciones.

Es evidente que si los valores x_1, x_2, ..., x_k aparecen, cada uno de ellos, una sola vez, se tiene:

$$\sigma^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2}{k}$$

k = número total de datos u observaciones.

En el caso que $\sigma^2 = 0$ entendemos que todos los valores x_i coinciden con la media aritmética $\bar{x}$, es decir, todas las observaciones están concentradas en un mismo valor, por lo que la dispersión es mínima (nula).

"CARACTERISTICAS DE LA VARIANZA"

- Al ser la varianza σ^2 una suma de cuadrados es siempre positiva. Por otro lado, estará expresada en unidades al cuadrado mientras que la variable estudiada se expresa en unidades. - Observa la necesidad de definir otra nueva medida de dispersión: la desviación típica.

- No es recomendable su cálculo cuando tampoco lo sea el de la media aritmética como medida de tendencia central.

DESVIACION TIPICA (σ)

La desviación típica σ se define como la raíz cuadrada positiva de la varianza:

$$\sigma = + \sqrt{\sigma^2} = + \sqrt{\frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_i}{N}}$$

o bien

$$\sigma = + \sqrt{\frac{\sum_{i=1}^{k} x_i^2 \cdot n_i}{N} - \bar{x}^2}$$

La desviación típica es la medida de dispersión más utilizada en estadística. Piensa que viene expresada en las mismas unidades que los valores de la variable X, por lo que su interpretación resulta más sencilla.

"CARACTERISTICAS DE LA DESVIACION TIPICA"

- Toma siempre valores positivos.
- No es recomendable su cálculo cuando tampoco lo sea el de la media como medida de tendencia central.

COEFICIENTE DE VARIACION DE PEARSON

El coeficiente de variación de Pearson C.V. se define de la forma:

$$C.V. = \frac{\sigma}{\bar{x}}$$

siendo

σ = desviación típica.

$\bar{x}$ = media aritmética.

A veces este coeficiente se multiplica por 100, para mayor comodidad en el manejo de cifras, ya que así trabajaríamos con porcentajes.

$$C.V. = \frac{\sigma}{\bar{x}} \cdot 100 \quad \text{(expresado en porcentajes)}$$

"CARACTERISTICAS DEL COEFICIENTE DE VARIACION DE PEARSON"

- La utilidad de este coeficiente estriba en la posibilidad de - comparar la dispersión de dos o más grupos no homogéneos.

Ejemplo 1: ¿Qué medidas están más dispersas los pesos o las alturas de un grupo de estudiantes?.

Como son medidas no homogéneas, ya que una está hecha en metros y la otra en kilogramos, no son comparables y hemos de recurrir a alguna medida de dispersión abstracta en la que esa dificultad se salve. Esta medida de dispersión es el "coeficiente de variación de Pearson".

Ejemplo 2: ¿Qué grupo se encuentra más disperso, el peso de un - grupo de niños al nacer, y el peso de un grupo de adultos de 50 años, si tienen la misma desviación típica σ = 10 kg?.

La desviación típica σ = 10 no representa lo mismo para los recién nacidos que para los adultos. En el grupo de recién nacidos sería enormemente desproporcionada, mientras que para el grupo - de adultos resulta normal.

En este caso también hemos de recurrir a una medida de dispersión abstracta, utilizamos el "coeficiente de variación de Pearson".

OTRAS MEDIDAS DE POSICION Y DISPERSION

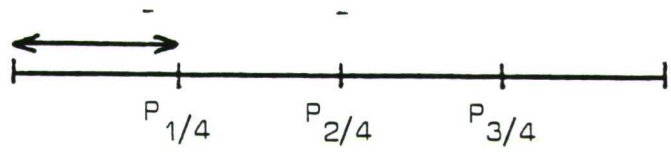
MEDIDAS DE POSICION

Las medidas de posición son aquellas que nos informan del orden o de la posición que ocupa un valor dentro del total de valores observados.

Las medidas de posición más utilizadas son: los cuartiles y percentiles.

- CUARTILES

Se definen los cuartiles como tres valores de la variable que dividen las observaciones realizadas en cuatro partes iguales:



Primer cuartil $P_{1/4}$: es el valor de la variable que deja la cuarta parte de las observaciones menores o iguales a él, en otras palabras, las 3/4 partes de las observaciones superiores a él.

Se calcula igual que la mediana M_e, pero en vez de tomar el número de observaciones N/2, se toma el número de observaciones N/4.

Segundo cuartil $P_{2/4}$: es el valor de la variable que deja inferiores o iguales a él las 2/4 partes (la mitad) de las observaciones.

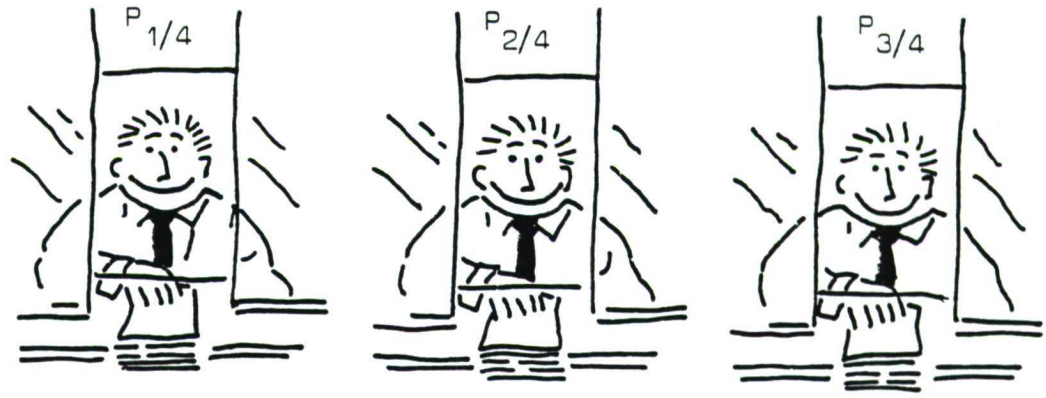
Se calcula tomando 2N/4 = N/2.

Este cuartil es la mediana M_e: $P_{2/4} = M_e$

Tercer cuartil $P_{3/4}$: es el valor de la variable que deja inferiores o iguales a él las 3/4 partes de las observaciones.

Se calcula tomando 3N/4 observaciones.

MEDIDA DE POSICION: LOS CUARTILES



Hemos medido el número de metros que el niño anda seguido antes de caerse en una clínica infantil, obteniendo la siguiente tabla de distribución de frecuencias:

x_i	n_i	N_i
1	2	2
2	6	8
3	10	18
4	5	23
5	10	33
6	3	36
7	2	38
8	2	40

$\sum_{i=1}^{8} n_i = 40$

x_i = edad del niño

n_i = metros que el niño anda

Se quieren calcular los cuartiles de esta distribución

$N = \sum_{i=1}^{8} n_i = 40$ número de observaciones

Primer cuartil $P_{1/4}$ se toma $\frac{N}{4}$

Segundo cuartil $P_{2/4}$ se toma $\frac{2N}{4}$

Tercer cuartil $P_{3/4}$ se toma $\frac{3N}{4}$

Primer cuartil: $1/4 \cdot N = 1/4 \cdot 40 = 10$

En la columna de las frecuencias absolutas acumuladas N_i, el valor 10 se encuentra entre 8 y 18:

$$8 < 10 < 18$$

El cuartil $P_{1/4}$ será el valor de la x_i que corresponde a $N_3 = 18$, es decir:

$$P_{1/4} = 3$$

Segundo cuartil: $2/4 \cdot N = 2/4 \cdot 40 = 20$

$$18 < 20 < 23$$

de donde

$$P_{2/4} = 4 = M_e$$

Tercer cuartil: $3/4 \cdot N = 3/4 \cdot 40 = 30$

$$23 < 30 < 33$$

entonces

$$P_{3/4} = 5$$

- PERCENTILES

Se llaman también "centiles" (de 100). Se define el percentil como el valor de la variable que deja inferiores o iguales a él un porcentaje de observaciones.

Así: el percentil k-ésimo será el valor de la variable que deja inferiores o iguales a él las k/100 partes de las observaciones (el k por 100), donde k puede tomar cualquier valor desde 1 a 99.

El cálculo de los percentiles es idéntico al de la mediana y los cuartiles. Se denotan por P_k, de tal forma:

Percentil 25: es el valor de la variable que deja 25/100 de las observaciones menores o iguales a él. Se denota P_{25} y su cálculo es igual que el de la mediana M_e tomando 25.N/100 observaciones.

Observa que

$$25 \cdot N/100 = N/4 = \text{primer cuartil } P_{1/4}$$

Percentil 50 (P_{50}): es el valor de la variable que deja 50/100 de las observaciones menores o iguales a él. Se calcula igual que la mediana tomando 50.N/100 observaciones.

Observa que

$$50 \cdot N/100 = N/2 = M_e = P_{2/4}$$

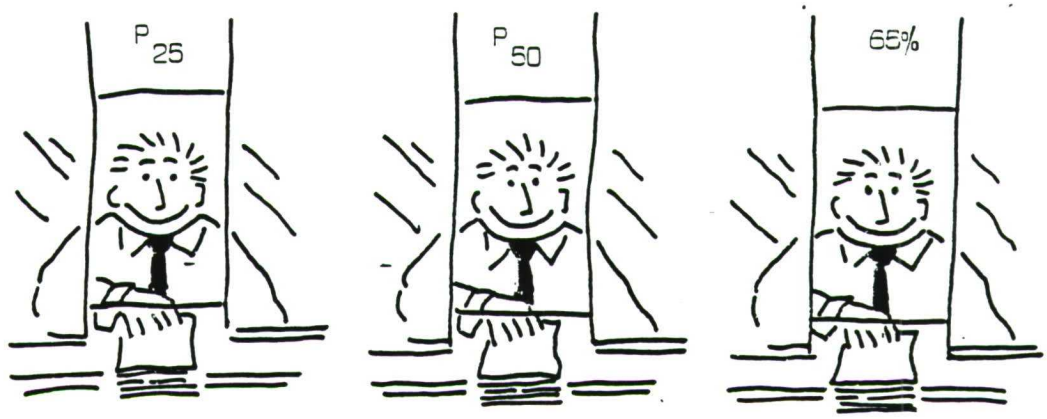
Este percentil también se llama "mediana".

Percentil 75 (P_{75}): Se calcula igual que la mediana tomando 75.N/100 observaciones.

Observa que

$$75 \cdot N/100 = 3N/4 = P_{3/4} \quad \text{(tercer cuartil)}$$

MEDIDA DE POSICION: LOS PERCENTILES



Con los mismos datos de la anterior distribución de frecuencias, calcular:

- Los percentiles 25 y 50.
- El valor de la variable que deja por debajo el 65% de las edades.

Percentil 25(P_{25}): Se toman 25.N/100 observaciones, de donde:

$$25 \cdot N/100 = 25 \cdot 40/100 = 10$$

Primeramente tenemos que localizar el intervalo en el que se encuentra P_{25}, es decir, aquel cuya frecuencia absoluta acumulada N_i sea igual o superior al 25% de N:

$$8 < 10 < 18$$

Y encontramos que a $N_3 = 18$ corresponde un valor $x_3 = 3$, - por tanto:

$$P_{25} = 3$$

Observa que el percentil P_{25} es igual al primer cuartil - $P_{1/4}$.

Percentil 50 (P_{50}): se toman 50.N/100 observaciones:

$$50 \cdot N/100 = 50 \cdot 40/100 = 20 = M_e$$

El percentil P_{50} se encuentra en el intervalo 18 y 23 de la frecuencia absoluta acumulada N_i, es decir:

$$18 < 20 < 23$$

Y encontramos que a $N_4 = 23$ corresponde un valor $x_4 = 4$, - de donde:

$$P_{50} = 4 = M_e \quad \text{(mediana)}$$

El 65% de las edades corresponde al percentil 65, P_{65}. Tenemos que localizar el intervalo en el que se encuentra el P_{65} en la columna de las frecuencias - absolutas acumuladas N_i, por tanto:

$$65 \cdot N/100 = 65 \cdot 40/100 = 26$$

En la columna de las N_i:

$$23 < 26 < 33$$

Encontramos que a $N_5 = 33$ corresponde un valor $x_5 = 5$, entonces:

$$P_{65} = 5$$

esto es, el valor de la variable X = edad que deja por debajo el 65% de las observaciones es de 5 años.

"OTRAS MEDIDAS DE DISPERSION"

- AMPLITUD O RECORRIDO

Se define la amplitud o recorrido de una variable estadística como la diferencia entre su valor máximo y su valor mínimo.

$$\text{recorrido} = R = \text{máximo } (x_i) - \text{mínimo } (x_i)$$

Ejemplo: Las edades de ocho estudiantes son:

15, 16, 10, 20, 21, 9, 7, 24

El valor más alto es 24.

El valor más bajo es 7.

El recorrido es 24 - 7 = 17 años. Esto es, el recorrido de la distribución de edades de los estudiantes es de 17 años.

- RECORRIDO SEMIINTERCUARTILICO

El recorrido semiintercuartílico es la mitad de la diferencia entre el tercer cuartil $P_{3/4}$ y el primer cuartil $P_{1/4}$ o, lo que es lo mismo, la mitad de la diferencia entre el percentil P_{75} y el percentil P_{25}. Sea, por tanto:

$$P = \frac{P_{3/4} - P_{1/4}}{2}$$

o bien

$$P = \frac{P_{75} - P_{25}}{2}$$

recuerda que:

$$P_{3/4} = P_{75}$$

$$P_{1/4} = P_{25}$$

Ejemplo: El recorrido semiintercuartílico de la distribución de frecuencias anterior:

$$P = \frac{P_{3/4} - P_{1/4}}{2} = \frac{5 - 3}{2} = 1$$

¿COMO SE DEBEN UTILIZAR LAS MEDIDAS DE DISPERSION?

Sabemos que una medida de tendencia central nos proporciona poca información. Para describir una información más completa necesitamos cuantificar la representatividad de las medidas centrales, esta información adicional nos la proporciona las medidas de dispersión.

Es necesario, pues, mostrar las "parejas" de medidas estadísticas más empleadas, así como su idoneidad.

MEDIDAS ESTADISTICAS

DE TENDENCIA DENTRAL	DE DISPERSION
Media aritmética = $\bar{x}$	varianza = σ^2 desviación típica = σ
Mediana = M_e	Recorrido semiintercuartílico = P
Moda = M_d	Recorrido = R

MOMENTOS

Definimos el momento de orden r respecto al parámetro c, de la forma:

$$M_r(c) = \frac{\sum_i (x_i - c)^r \cdot n_i}{N}$$

En particular, nos interesan dos casos importantes:

Momentos respecto al origen: cuando el parámetro c = 0, entonces:

$$M_r(0) = \frac{\sum_i (x_i - 0)^r \cdot n_i}{N} = \frac{\sum_i x_i^r \cdot n_i}{N}$$

A estos momentos particulares se los denota por (a_r), de forma que:

$$a_r = \frac{\sum_i x_i^r \cdot n_i}{N}$$

dando valores a r, se obtiene:

$$a_0 = \frac{\sum_i x_i^0 \cdot n_i}{N} = \frac{\sum_i n_i}{N} = \frac{N}{N} = 1$$

$$a_1 = \frac{\sum_i x_i \cdot n_i}{N} = \bar{x} \quad \text{media aritmética}$$

$$a_2 = \frac{\sum_i x_i^2 \cdot n_i}{N} = \text{segundo momento respecto al origen}$$

Momentos respecto a la media: cuando el parámetro $c = \bar{x}$, de donde:

$$M_r(\bar{x}) = \frac{\sum_i (x_i - \bar{x})^r \cdot n_i}{N}$$

A estos momentos particulares se los denota por (m_r), de forma que:

$$m_r = \frac{\sum_i (x_i - \bar{x})^r \cdot n_i}{N}$$

dando valores a r, se obtiene:

$$m_0 = \frac{\sum_i (x_i - \bar{x})^0 \cdot n_i}{N} = \frac{\sum_i n_i}{N} = 1$$

$$m_1 = \frac{\sum_i (x_i - \bar{x}) \cdot n_i}{N} = \frac{0}{N} = 0 \quad \text{primer momento respecto a la media}$$

$$m_2 = \frac{\sum_i (x_i - \bar{x})^2 \cdot n_i}{N} = \sigma^2 \quad \text{varianza}$$

Relación entre momentos: Se pueden encontrar relaciones que liguen los momentos respecto al origen y los momentos respecto a la media. La relación más utilizada viene dada por la expresión:

$$m_2 = a_2 - (a_1)^2$$

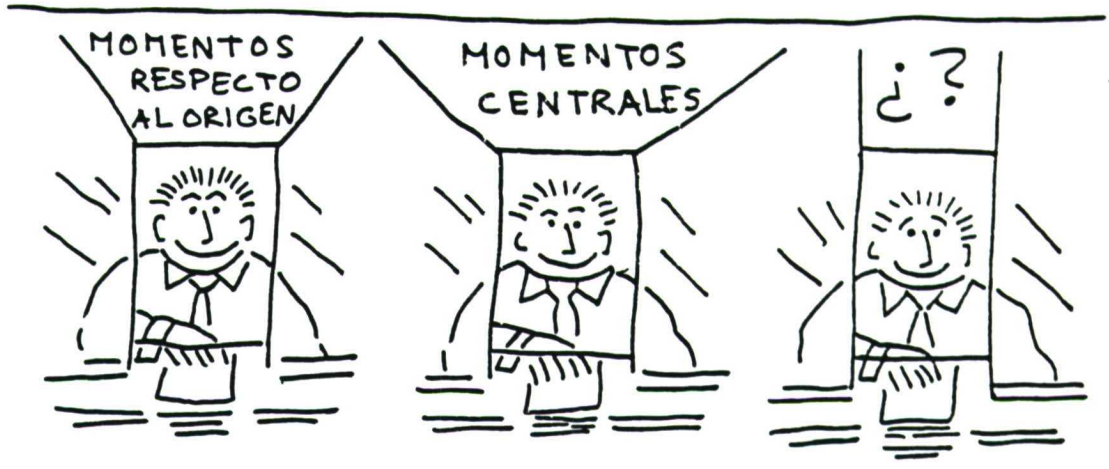
o bien

$$\sigma^2 = a_2 - (\bar{x})^2$$

Piensa que esta relación ya la conocíamos:

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

MEDIDAS ESTADISTICAS: LOS MOMENTOS



Con los mismos datos de la distribución de frecuencias que volvemos a reproducir:

x_i	n_i	N_i
1	2	2
2	6	8
3	10	18
4	5	23
5	10	33
6	3	36
7	2	38
8	2	40

se pide calcular:

1) Momentos respecto al origen de primero, segundo y tercero orden.

2) Momentos centrales de segundo y tercero orden.

1) Los momentos respecto al origen vienen dados por la exoresión:

$$a_r = \frac{\sum_i x_i^r \cdot n_i}{N}$$

de donde:

$$\text{momento respecto origen 1º orden} = a_1 = \frac{\sum_i x_i \cdot n_i}{N}$$

$$\text{momento respecto origen 2º orden} = a_2 = \frac{\sum_i x_i^2 \cdot n_i}{N}$$

$$\text{momento respecto origen 3º orden} = a_3 = \frac{\sum_i x_i^3 \cdot n_i}{N}$$

Para resolver este apartado necesitamos la tabla:

x_i	n_i	$x_i \cdot n_i$	x_i^2	$x_i^2 \cdot n_i$	x_i^3	$x_i^3 \cdot n_i$
1	2	2	1	2	1	2
2	6	12	4	24	8	48
3	10	30	9	90	27	270
4	5	20	16	80	64	320
5	10	50	25	250	125	1250
6	3	18	36	108	216	648
7	2	14	49	98	343	686
8	2	16	64	128	512	1024
	40	162		780		4248

de donde:

$$a_1 = \frac{\sum_{i=1}^{8} x_i \cdot n_i}{N} = \frac{162}{40} = 4'05 = \bar{x} \quad \text{media aritmética}$$

$$a_2 = \frac{\sum_{i=1}^{8} x_i^2 \cdot n_i}{N} = \frac{780}{40} = 19'5$$

$$a_3 = \frac{\sum_{i=1}^{8} x_i^3 \cdot n_i}{N} = \frac{4248}{40} = 106'2$$

2) Los momentos centrales vienen dados por la expresión:

$$m_r = \frac{\sum_i (x_i - \bar{x})^r \cdot n_i}{N} \quad .$$

por tanto:

$$\text{momento central 2º orden} = m_2 = \frac{\sum_i (x_i - \bar{x})^2 \cdot n_i}{N}$$

$$\text{momento central 3º orden} = m_3 = \frac{\sum_i (x_i - \bar{x})^3 \cdot n_i}{N}$$

Para resolver este apartado necesitamos la tabla:

$$\bar{x} = a_1 = 4'05$$

x_i	n_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 \cdot n_i$	$(x_i - \bar{x})^3 \cdot n_i$
1	2	- 3'05	9'30	18'6	- 56'73
2	6	- 2'05	4'20	25'2	- 51'66
3	10	- 1'05	1'10	11	- 11'55
4	5	- 0'05	0'0025	0'0125	- 0'0006
5	10	0'95	0'90	9	8'55
6	3	1'95	3'80	11'4	22'23
7	2	2'95	8'70	17'4	51'33
8	2	3'95	15'60	31'2	123'24
	40			123'81	85'41

en consecuencia:

$$m_2 = \frac{\sum_{i=1}^{8} (x_i - \bar{x})^2 \cdot n_i}{N} = \frac{123'81}{40} = 3'1 = \sigma^2 \quad \text{varianza}$$

efectivamente

$$\sigma^2 = a_2 - (a_1)^2 = \frac{\sum_{i=1}^{8} x_i^2 \cdot n_i}{N} - (\bar{x})^2 = 19'5 - (4'05)^2 = 3'1$$

$$m_3 = \frac{\sum_{i=1}^{8} (x_i - \bar{x})^3 \cdot n_i}{N} = \frac{85'41}{40} = 2'135$$

ACTIVIDAD - 1: En un Instituto se han medido los pesos y las alturas de un grupo de diez estudiantes, obteniéndose los siguientes datos:

Alturas	1'50	1'60	1'60	1'68	1'70	1'65	1'80	1'67	1'75	1'82
Pesos	60	65	62	70	68	68	75	62	74	76

Se desea saber, ¿qué medidas están más dispersas, los pesos o las alturas?.

Sean las variables:

X = alturas de los estudiantes.

Y = pesos de los estudiantes.

A primera vista parece más dispersa la variable peso que la variable altura. Piensa que una variable está hecha en kilogramos y la otra en metros, por tanto, son medidas no comparables y hemos de soslayar este inconveniente re-

curriendo a alguna medida de dispersión que se exprese mediante números carentes de unidades. Esta medida es el coeficiente de variación.

El coeficiente de variación viene dado por:

$$C.V. = \frac{\sigma}{\bar{x}}$$

siendo:

$$\bar{x} = \text{media aritmética} = \frac{\sum_{i=1}^{10} x_i \cdot n_i}{N}$$

$$\sigma = \text{desviación típica} = \sqrt{\frac{\sum_{i=1}^{10} (x_i - \bar{x})^2 \cdot n_i}{N}}$$

Calculemos la media y la desviación típica en cada una de las variables.

a) La tabla de frecuencias de la variable Y = peso.

y_j	n_j	$y_j \cdot n_j$	$(y_j - \bar{y})$	$(y_j - \bar{y})^2$	$(y_j - \bar{y})^2 \cdot n_j$
60	1	60	- 8	64	64
65	1	65	- 3	9	9
62	2	124	- 6	36	72
70	1	70	2	4	4
68	2	136	0	0	0
75	1	75	7	49	49
74	1	74	6	36	36
76	1	76	8	64	64
	10	$\sum_j y_j \cdot n_j = 680$			$\sum_j (y_j - \bar{y})^2 \cdot n_j = 298$

$$\bar{y} = \frac{\sum_j y_j \cdot n_j}{N} = \frac{680}{10} = 68 \text{ kilogramos}$$

$$\sigma_y^2 = \frac{\sum_j (y_j - \bar{y})^2 \cdot n_j}{N} = \frac{298}{10} = 29'8 \text{ kg}^2$$

$$\sigma_y = + \sqrt{\sigma_y^2} = + \sqrt{29'8} = 5'458 \text{ kg.}$$

El coeficiente de variación viene dado:

$$C.V._{pesos} = \frac{\sigma_y}{\bar{y}} = \frac{5'458}{68} = 0'080$$

b) La media aritmética y la desviación típica de la variable X = alturas vienen dadas por las expresiones:

$$\bar{x} = \frac{\sum_i x_i \cdot n_i}{N} = \frac{1}{10} \left[1'50 + (1'60 \times 2) + 1'68 + 1'70 + 1'65 + 1'80 + \right.$$

$$\left. + 1'67 + 1'75 + 1'82 \right] = 1'677 \text{ metros}$$

$$\sigma_x^2 = \frac{\sum_i (x_i - \bar{x})^2 \cdot n_i}{N} = \frac{0'0854}{10} = 0'00854 \text{ m}^2$$

$$\sigma_x = + \sqrt{\sigma_x^2} = + \sqrt{0'00854} = 0'292 \text{ metros}$$

El coeficiente de variación:

$$C.V._{Alturas} = \frac{\sigma_x}{\bar{x}} = \frac{0'292}{1'677} = 0'174$$

Resulta, entonces:

$$C.V._{Pesos} = 0'080 \qquad C.V._{Alturas} = 0'174$$

El coeficiente de variación de Pearson aparece multiplicado por 100, para mayor comodidad en el manejo de las cifras, así trabajamos con porcentajes, teniéndose:

$$C.V._{Pesos} = 8 \qquad C.V._{Alturas} = 17'4$$

ambos números abstractos de fácil comparación expresan que las alturas están más dispersas que los pesos en contra de lo que intuitivamente parecía.

ACTIVIDAD - 2: Una vacuna antituberculosa se administró a un grupo de cuarenta personas, a las veinticuatro horas de su efecto, se tomó la temperatura a las mismas, obteniéndose los siguientes datos:

Número de personas	2	5	20	10	3	0
Temperatura en grados	36'5	37	37'5	38	38'5	40

Se desea saber:

a) La mediana.

b) La moda.

c) Desviación típica.

Sea la variable X = "temperatura en grados que una persona tiene".

La tabla de frecuencias es:

x_i	n_i	N_i
36'5	2	2
37	5	7
37'5	20	27
38	10	37
38'5	3	40
40	0	40
	40	

20

n_i = frecuencia absoluta

N_i = frecuencia absoluta acumulada

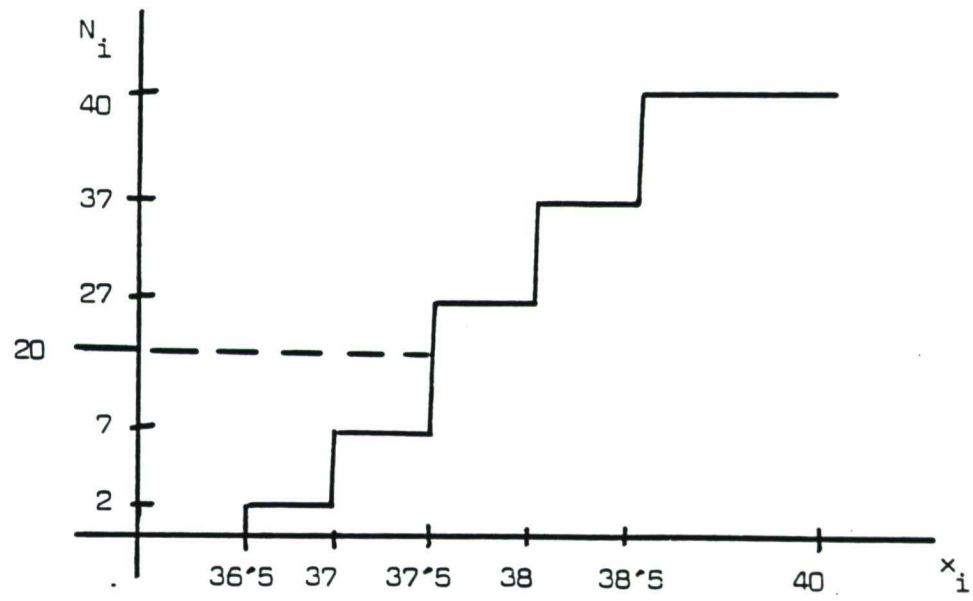
a) Para el cálculo de la mediana M_e procederemos de la siguiente forma:

1. Dividimos N entre 2: $N/2 = 40/2 = 20$.

2. Observamos la columna N_i de la tabla de frecuencias absolutas acumuladas y vemos que este valor no se encuentra allí comprendido entre 7 y 27:

$$7 < 20 < 27$$

3. Dibujamos el diagrama de frecuencias absolutas acumuladas:



y vemos a qué abscisa corresponde la ordenada 20. La mediana es, por tanto:

$$M_e = 37'5$$

b) El valor de la variable que tiene mayor frecuencia es 37'5:

$$M_d = 37'5$$

c) Para hallar de la desviación típica σ, necesitaremos los cálculos:

x_i	n_i	$x_i \cdot n_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 \cdot n_i$
36'5	2	73	- 1'08	1'1664	2'3328
37	5	185	- 0'58	0'3364	1'682
37'5	20	750	- 0'08	0'0064	0'128
38	10	380	0'42	0'1764	1'764
38'5	3	115'5	0'92	0'8464	2'5392
40	0	0	2'42	5'8564	0
40		$\sum_{i=1}^{6} x_i \cdot n_i = 1503'5$			$\sum_{i=1}^{6} (x_i - \bar{x})^2 \cdot n_i = 8'446$

La media aritmética:

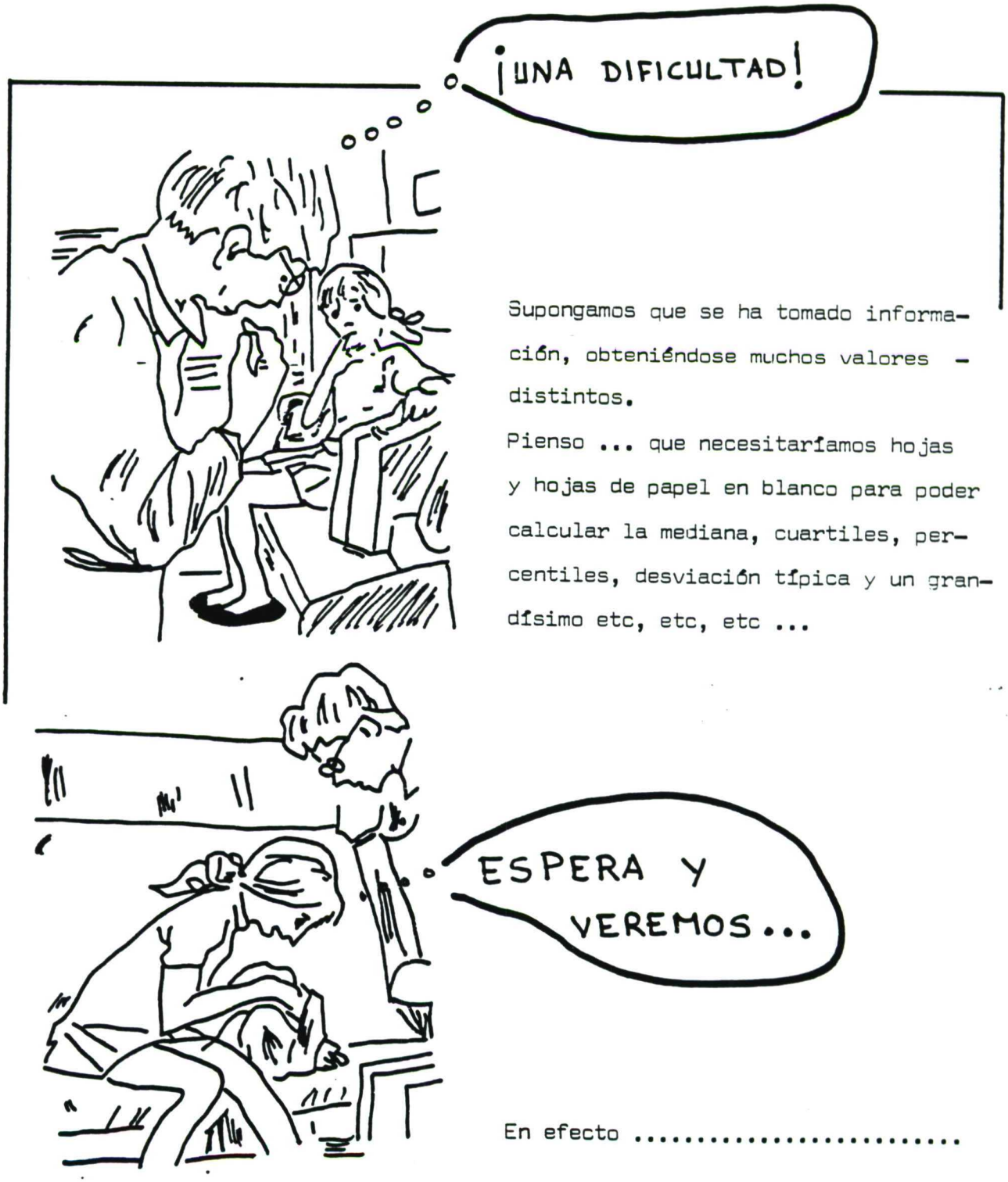
$$\bar{x} = \frac{\sum_{i=1}^{6} x_i \cdot n_i}{N} = \frac{1503'5}{40} = 37'58$$

La varianza:

$$\sigma^2 = \frac{\sum_{i=1}^{6} (x_i - \bar{x})^2 \cdot n_i}{N} = \frac{8'446}{40} = 0'2111$$

La desviación típica:

$$\sigma = + \sqrt{\sigma^2} = + \sqrt{0'2111} = 0'4595$$



Supongamos que cincuenta estudiantes han obtenido en una prueba de inteligencia las siguientes puntuaciones:

8	11	11	8	9	10	16	6	12	19	13	14	9	13	15	9
12	16	8	7	14	11	15	6	14	14	17	11	6	9	10	19
12	11	12	6	15	16	16	12	13	12	12	8	17	13	7	12
14	12														

Es aconsejable, en un caso como este, agrupar los datos en intervalos y hacer un recuento de las observaciones que caen dentro de cada uno de ellos.

Ordenemos estas puntuaciones:

6	6	6	6	7	7	8	8	8	8	9	9	9	9	10	10
11	11	11	11	11	12	12	12	12	12	12	12	12	12	13	13
13	13	14	14	14	14	14	15	15	15	16	16	16	16	17	17
19	19														

No cabe duda que hemos de elegir un número de intervalos de forma que cubra dos objetivos importantísimos:

- Simplifique nuestro trabajo.
- No se pierda mucha información al tomar los datos agrupados.

1. Determinar amplitud de la distribución.
2. Fijar número de intervalos.
3. Calcular la amplitud de los intervalos.
4. Determinar el límite inferior del primer intervalo.

1. Determinar amplitud de la distribución: Se calcula restando la puntuación máxima y la mínima.

$$\text{amplitud} = A = x_{máx} - x_{mín} = 19 - 6 = 13$$

2. Fijar número de intervalos: El número de intervalos que se van a utilizar depende del tamaño de la muestra.

Piensa que si hay pocos intervalos se produce gran pérdida de información y - si hay muchos intervalos la tabla resulta bastante larga.

Algunos autores recomiendan que el número de intervalos no supere el valor de $\sqrt{N}$, siendo N = número total de observaciones.

Por tanto:

$$\text{nº de intervalos} = \sqrt{N} = \sqrt{50} = 7'07 \simeq 7$$

3. Amplitud de los intervalos: La amplitud del intervalo la denotaremos por "c", viene dada por la expresión:

$$\text{amplitud intervalo} = c = \frac{\text{amplitud de la distribución}}{\text{número de intervalos}}$$

de donde:

$$\text{amplitud intervalo} = c = \frac{13}{7} = 1'857$$

esta división no nos da un número entero, no obstante conviene que la amplitud del intervalo sí que lo sea, por lo que redondeamos la división con el - entero superior, es decir:

$$c = 2$$

4. Límite inferior del primer intervalo: Es la puntuación más pequeña a partir de la cual empezamos a contar.

En nuestro caso, decidimos empezar a contar a partir del 5'5, - siendo la amplitud del intervalo c = 2.

En consecuencia

La distribución de frecuencias es la siguiente:

Intervalos	Recuento	n_i	Marcas clase x_i
5'5 - 7'5	𝍸 |	6	6'5
7'5 - 9'5	𝍸 |||	8	8'5
9'5 - 11'5	𝍸 ||	7	10'5
11'5 - 13'5	𝍸 𝍸 |||	13	12'5
13'5 - 15'5	𝍸 |||	8	14'5
15'5 - 17'5	𝍸 |	6	16'5
17'5 - 19'5	||	2	18'5
		50	

Piensa:

- marca de clase: es el punto medio de cada intervalo.
- La elección de intervalos, así como su amplitud es algo personal y no se encuentra sometido a ninguna norma rígida.
- Al operar con las marcas de clase se pierde información.

NUESTRO OBJETIVO ES OBTENER:

a) Los intervalos reales de clase.

b) La media aritmética.

c) La desviación típica.

d) La mediana.

e) La moda.

f) Cuartil tercero.

a) Es conveniente formar la siguiente tabla:

Intervalos	x_i	n_i	$x_i \cdot n_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 \cdot n_i$
5'5 - 7'5	6'5	6	39	- 5'4	29'16	174'96
7'5 - 9'5	8'5	8	68	- 3'4	11'56	92'48
9'5 - 11'5	10'5	7	73'5	- 1'4	1'96	13'72
11'5 - 13'5	12'5	13	162'5	0'6	0'36	4'68
13'5 - 15'5	14'5	8	116	2'6	6'76	54'08
15'5 - 17'5	16'5	6	99	4'6	21'16	126'96
17'5 - 19'5	18'5	2	37	6'6	43'56	87'12
		50	595			554

b) La media aritmética será:

$$\bar{x} = \frac{\sum_{i=1}^{7} x_i \cdot n_i}{N} = \frac{595}{50} = 11'9$$

c) La varianza = $\sigma^2 = \frac{\sum_{i=1}^{7} (x_i - \bar{x})^2 \cdot n_i}{N} = \frac{554}{50} = 11'08$

desviación típica $\sigma = + \sqrt{\sigma^2} = + \sqrt{11'08} = 3'328$

d) Cálculo de la mediana: $\frac{N}{2} = \frac{50}{2} = 25$

Intervalos	x_i	n_i	N_i
5'5 - 7'5	6'5	6	6
7'5 - 9'5	8'5	8	14
9'5 - 11'5	10'5	7	21
11'5 - 13'5	12'5	13	34
13'5 - 15'5	14'5	8	42
15'5 - 17'5	16'5	6	48
17'5 - 19'5	18'5	2	50

(25)

este número 25 no se encuentra en la columna de las frecuencias absolutas acumuladas N_i, vemos que:

$$21 < 25 < 34$$

luego la mediana M_e está en el intervalo 11'5 - 13'5. En otras palabras:

$$11'5 < M_e < 13'5$$

hallaremos la mediana M_e, mediante la proporción correspondiente (interpolación):

$$\frac{34 - 21}{13'5 - 11'5} = \frac{25 - 21}{x}$$

(Observa: $M_e - 11'5 = x$)

$$\frac{13}{2} = \frac{4}{x} \quad , \text{ entonces} \quad x = \frac{8}{13} = 0'615$$

luego

$$M_e = 11'5 + 0'615 = 12'115$$

e) La moda es el intervalo de máxima frecuencia; esto es:

$$11'5 - 13'5$$

Es conveniente que nos vayamos acostumbrando a pensar que cuando agrupamos los datos en intervalos de clase, la marca de clase es a menudo poco representativa, de ahí que no sea acertado tomar - la moda $M_d = 12'5$.

Para hallar la posición exacta de la moda, recurrimos a la expresión:

$$M_d = l + c \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right)$$

donde:

l = límite inferior del intervalo modal.

c = tamaño del intervalo.

Δ_1 = diferencia de la n_i del intervalo modal con la n_{i-1} anterior.

Δ_2 = n_i del intervalo modal - n_{i+1} posterior.

de tal forma que:

$$\left.\begin{array}{l} l = 11'5 \\ c = 13'5 - 11'5 = 2 \\ \Delta_1 = 13 - 7 = 6 \\ \Delta_2 = 13 - 8 = 5 \end{array}\right\} \quad M_d = 11'5 + 2 \cdot \left(\frac{6}{6+5}\right) = 11'5 + \frac{12}{11} = 12'59$$

f) Cálculo de $P_{3/4}$.

$$\frac{3}{4} \cdot N = \frac{3}{4} \cdot 50 = 37'5$$

este número 37'5 no se encuentra en la columna de las N_i, observa que:

$$34 < 37'5 < 42$$

por tanto, el cuartil tercero estará en el intervalo 13'5 - 15'5:

$$13'5 < P_{3/4} < 15'5$$

hallaremos $P_{3/4}$ de forma análoga a la mediana, mediante la proporción correspondiente (interpolación):

$$\frac{42 - 34}{15'5 - 13'5} = \frac{37'5 - 34}{x}$$

$$(P_{3/4} - 13'5 = x)$$

$$\frac{8}{2} = \frac{3'5}{x} \qquad x = \frac{7}{8} = 0'875$$

$$P_{3/4} = 13'5 + 0'875 = 14'375$$

ACTIVIDAD – 1: Calcular la varianza y la desviación típica de los siguientes datos:

a) 5, 2, 1, 5, 3, 8

b) 3, 4, 1, 4

c) 1, 9, 3, 7, 8, 8

ACTIVIDAD – 2: Los valores registrados en dos variables distintas (X = peso; Y = prueba de inteligencia) para un mismo grupo de personas han sido los que aparecen en la tabla adjunta. Deseamos conocer qué variable presenta mayor dispersión.

X	60	65	80	70	60	80	85
Y	8	11	11	15	19	24	12

ACTIVIDAD - 3: Calcular la varianza y la desviación típica a partir de los siguientes datos agrupados en intervalos:

a)

Intervalos	n_i
1 - 3	2
4 - 6	3
7 - 9	4
10 - 12	1

b)

Intervalos	n_i
2 - 4	4
5 - 7	6
8 - 10	3
11 - 13	2

ACTIVIDAD - 4: Sea la siguiente distribución de frecuencias:

Intervalos	$\bar{n}_i$
79 - 84	5
85 - 90	10
91 - 96	16
97 - 102	11
103 - 108	8

Se desea conocer:

1) Los percentiles 25, 36 y 75.

2) Calcular la amplitud semiintercuartílica.

(NOTA: Se recomienda elegir intervalos de la forma 78'5 - 84'5; 84'5 - 90'5; etc, etc ...; c = 6.)

ACTIVIDAD – 5: Los percentiles son valores esencialmente positivos. ¿Si?. ¿No?.

ACTIVIDAD – 6: Dada la siguiente tabla:

X	0	1	2	3	4	5	6	7
n_i	2	3	10	10	5	0	5	0

Se desea conocer:

1) La media aritmética.

2) La moda.

3) Los cuartiles primero y tercero.

4) Recorrido semiintercuartílico.

5) Momento central de segundo orden.

6) Momentos respecto al origen de primero, segundo y tercer orden.

AUTOCOMPROBACION

ACTIVIDAD - 1:

a) $\sigma^2 = 5'33$, $\sigma = 2'308$

b) $\sigma^2 = 1'50$, $\sigma = 1'224$

c) $\sigma^2 = 8'67$, $\sigma = 2'944$

ACTIVIDAD - 2:

$\bar{x} = 71'428$ $\quad \sigma_x^2 = 90'816$ $\quad \sigma_x = 9'529$

$\bar{y} = 14'285$ $\quad \sigma_y^2 = 26'204$ $\quad \sigma_y = 5'118$

$$C.V._x = \frac{9'529}{71'428} = 0'133 \qquad C.V._y = \frac{5'118}{14'285} = 0'358$$

La variable Y presenta mayor dispersión que la variable X.

ACTIVIDAD - 3:

a) $\sigma^2 = 7'56$; $\sigma = 2'749$

b) $\sigma^2 = 8'64$; $\sigma = 2'939$

ACTIVIDAD - 4:

1) $P_{25} = 89$; $P_{36} = 91'625$; $P_{75} = 100'045$

2) $P = (P_{75} - P_{25})/2 = 5'5225$

ACTIVIDAD - 5:

No. Un percentil puede ser una puntuación positiva, negativa o - nula.

ACTIVIDAD - 6:

1. $\bar{x} = 103/35 = 2'942$

2. Hay dos modas: $M_{d_1} = 2$, $M_{d_2} = 3$

3. $P_{1/4} = 2$ ya que $5 < 8'75 < 15$

   $P_{3/4} = 4$ pues $25 < 26'25 < 30$

4. $P = 1$

5. $$m_2 = \frac{\sum_i (x_i - \bar{x})^2 \cdot n_i}{N} = \sigma^2 = 2'5681$$

6. $a_1 = \bar{x} = 2'942$

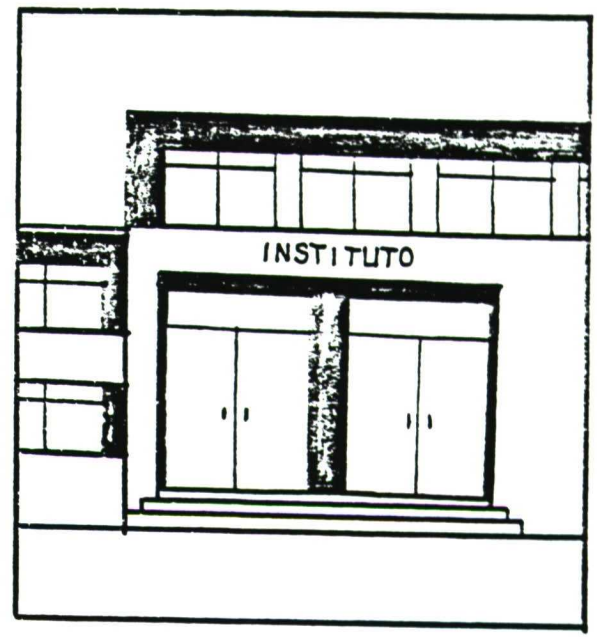
   $a_2 = 11'228$

   $$a_3 = \frac{\sum_i x_i^3 \cdot n_i}{N} = \frac{1753}{35} = 50'085$$

"ENCUESTA EN EL INSTITUTO"

Raúl después de marcharse de "Muebles Quintana", encuentra un trabajo eventual en la Secretaría del Instituto - Cantalejo.

Una tarde hablando con el director, - Don Pablo, tuvo lugar la siguiente conversación:



Don Pablo: Raúl estoy pensando que sería interesante conocer los estudios universitarios que pretenden hacer nuestros alumnos.

Raúl: A tí, se te ocurre hasta pensar con tal de hacernos trabajar.

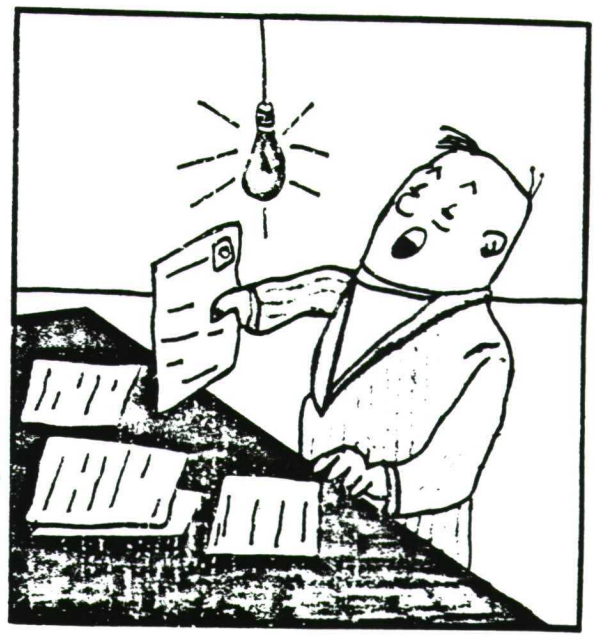
Don Pablo: Trabajo, trabajo..., no es tanto, sólo se trata de saber hacer bien las cosas, elaborar una "estadística"; así es como se solucionan estos problemas ahora.

Ocúpate de ello, para empezar estaría bien que mañana me dijeses cuantos de nuestros alumnos quieren estudiar Empresariales.

Raúl: De acuerdo, está bien.

Realmente, Raúl no estaba de acuerdo, él qué sabía de eso de la estadística (sin ir más lejos, debido a no saber estadística, tuvo un disgusto en su — último trabajo, ¡parecía que le perseguía!), pero debería hacerlo.

Raúl (pensando): ¡Vaya! ¿Cómo podría hacerlo? Hum... ¡Claro!, ahora recuerdo que los alumnos me indicaron su inclinación con los papeles de matrícula. ¡Está resuelto! Pero creo — que hoy dormiré poco, revisar esos papeles me — llevará tiempo.



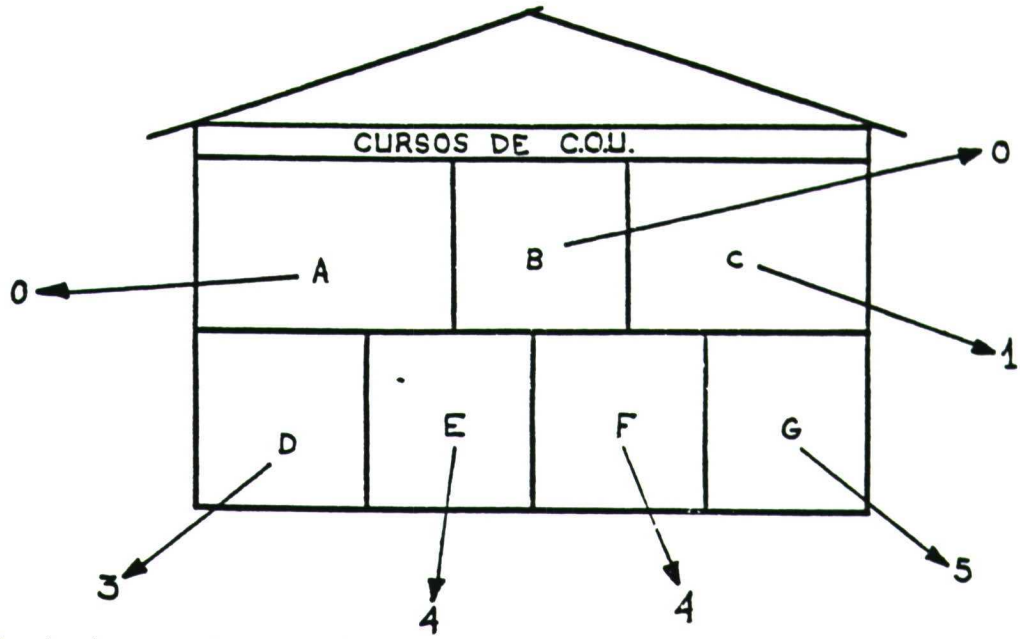
Así Raúl consiguió elaborar las siguientes listas:

ALUMNOS INSTITUTO CANTALEJO

OPCION EN COU \ CURSOS DE COU	COU A	COU B	COU C	COU D	COU E	COU F	COU G	
CIENCIAS	0	0	0	30	40	42	38	
LETRAS	40	41	0	0	0	0	0	
MIXTAS	0	0	39	0	0	0	0	
TOTAL	40	41	39	30	40	42	38	270

DECIDEN EMPRESARIALES

COU A	0
COU B	0
COU C	1
COU D	3
COU E	4
COU F	4
COU G	5
	17



Por lo tanto podría decirle a Don Pablo que 17 alumnos estudiarían Ciencias Empresariales.

Mientras tanto Don Pablo pensaba:

Don Pablo: Voy a darle una sorpresa a Raúl. Si tenemos en C.O.U. 270 alumnos bastará tomar una representación de alrededor de 40 alumnos y preguntarles:

¿Piensas estudiar Empresariales?, a lo a lo cual ellos responderán:

Sí, No, No saben.

Mañana tengo clase con los de C.O.U. "C", les preguntaré a ellos.

De esta manera Pablo obtuvo los siguientes resultados:

NO SABEN	SI	NO
2	1	36

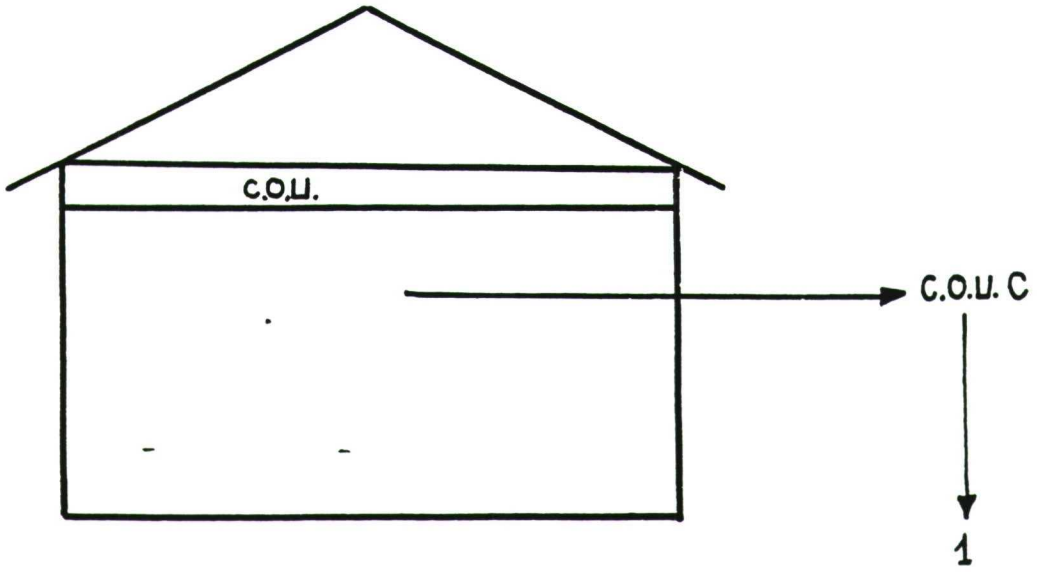
con lo cual

Total de alumnos 270

Número de representantes elegidos 39

contestaron sí 1

$\frac{39}{1} = \frac{270}{x}$ $\quad x = \frac{270.1}{39} = 7$ alumnos de C.O.U. estudiarán Empresariales.



A la mañana siguiente Raúl tenía cara de sueño.

Raúl: Ya tengo resuelto su encargo. Piensan estudiar Empresariales 17 alumnos.

Don Pablo: ¡Es imposible tienen que ser menos!

Raúl: Revise sus cálculos. De mi resultado estoy seguro.

Don Pablo fue rápidamente a buscar a la profesora de Matemáticas, Srta. Ana y le explicó el problema.

Srta. Ana: De acuerdo. Basta con que elijas una representación de alumnos de cada especialidad de la siguiente manera:

Total de alumnos 270

Alumnos de ciencias 150

Alumnos de letras 81

Alumnos de mixtas 39

Tamaño de la muestra 40

Sean : c representación de alumnos de ciencias.

l representación de alumnos de letras.

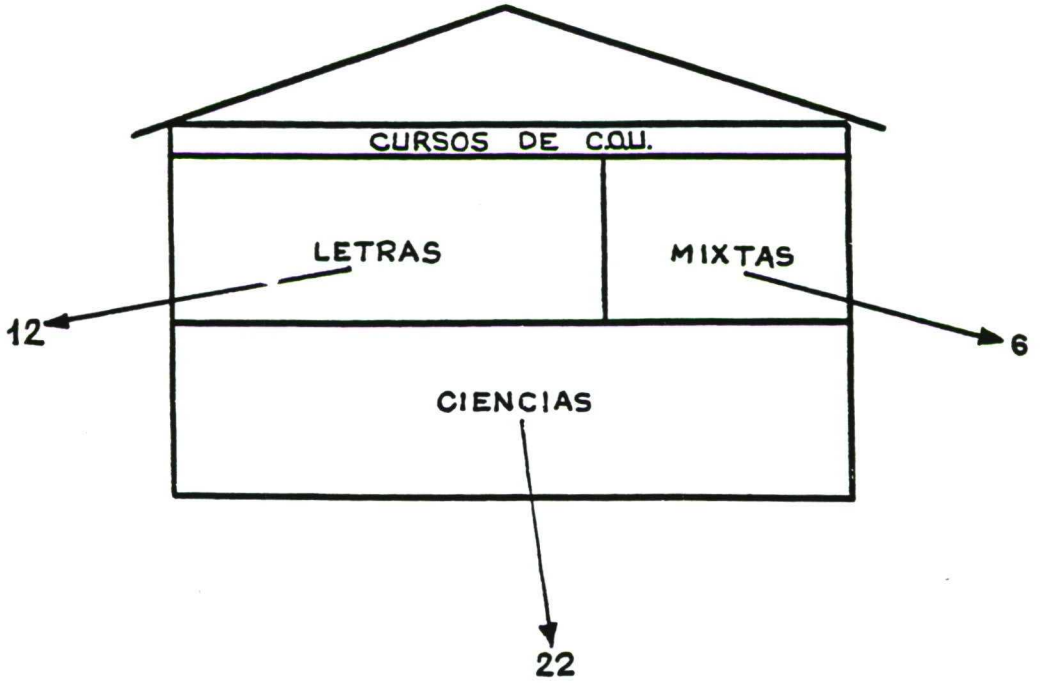
m representación de alumnos de mixtas.

entonces:

$$\frac{270}{150} = \frac{40}{c} \qquad c = \frac{150.40}{270} = 22 \text{ alumnos de ciencias.}$$

$$\frac{270}{81} = \frac{40}{l} \qquad l = \frac{81.40}{270} = 12 \text{ alumnos de letras.}$$

$$\frac{270}{39} = \frac{40}{m} \qquad m = \frac{39.40}{270} = 6 \text{ alumnos de mixtas.}$$



Trás esto el director quedó bastante contrariado. Ana tenía razón, su método era válido. Por otra parte él se había pasado de listo, Raúl aunque ignorante, había — actuado mejor que él. De esta forma había quedado en ridículo ante Ana y ante Raúl. Con Ana no podía hacer nada, pero el contrato con Raúl era temporal. Tendría que tomar medidas.

Al día siguiente Raúl recibió una notificación del director. ¡Había sido despedido!.

DEBES CONOCER

Cuando el estadístico toma información de todos y cada uno de los elementos de una población se dice que realiza un "censo".

A menudo, no es posible realizar un censo, bien sea:

- La población tiene infinitos elementos.
- La toma de información resulta demasiado costosa.
- La población está formada por entes potenciales (personas con una determinada enfermedad).
- La toma de información lleva consigo la destrucción del ente en cuestión (toma de controles de tiempo hasta que se funde una lámpara del televisor).

Este problema lleva al estadístico a tomar la información de unos cuantos elementos de la población. El conjunto de elementos de los que toma información se llama "muestra", y el número de elementos que la componen "tamaño de la muestra".

El proceso de recoger una muestra recibe el nombre de "muestreo". Tiene gran interés en muchos aspectos de la estadística. Por ejemplo, permite elaborar una encuesta.

Para que las conclusiones de la teoría del muestreo sean válidas, la muestra (muestras) debe elegirse de forma que sea "representativa" de la población.

Una "muestra es representativa" cuando cada elemento de la población tiene la misma posibilidad de ser incluido en la muestra. En otras palabras, una muestra está bien elegida:

- Cuando el "tamaño muestral" es representativo.
- Cuando el tanto por ciento de elementos con un determinado carácter que componen la muestra es el mismo tanto por ciento de elementos con idéntico carácter de la población total.

P I E N S A

El Instituto "María de Molina" tiene 600 estudiantes de Ciencias, 400 estudiantes de Mixto y 200 estudiantes de Letras. La Dirección está interesada en hacer una encuesta a 60 estudiantes para saber cuántos estudiantes del Instituto desearían realizar Económicas. ¿De qué modo elegimos los estudiantes para que la muestra sea representativa?

Número de estudiantes: 600 + 400 + 200 = 1200

Muestra: 60 estudiantes

Sean:

x estudiantes de Ciencias

y estudiantes de Mixto

z estudiantes de Letras

$\frac{1200}{600} = \frac{60}{x}$ $\qquad x = \frac{600 \cdot 60}{1200} = 30$ estudiantes de Ciencias

$\frac{1200}{400} = \frac{60}{y}$ $\qquad y = \frac{400 \cdot 60}{1200} = 20$ estudiantes de Mixto

$\frac{1200}{200} = \frac{60}{z}$ $\qquad z = \frac{200 \cdot 60}{1200} = 10$ estudiantes de Letras

De donde:

Los 1200 estudiantes del Instituto "María de Molina" quedan distribuidos:

600 Ciencias $\qquad$ 400 Mixto $\qquad$ 200 Letras

La muestra de 60 estudiantes es representativa si está formada por:

30 Ciencias $\qquad$ 20 Mixto $\qquad$ 10 Letras

RECUERDA

- El estadístico utiliza la "muestra" para la toma de información.
- Si una muestra es "representativa" de una población, se pueden deducir importantes conclusiones acerca de ésta, bastará analiza la muestra.
- Una población puede ser finita o infinita. Por ejemplo, la población formada por todos los coches producidos por SEAT en un día es finita, mientras que la población formada por todos los posibles sucesos (caras, cruces) en tiradas sucesivas de una moneda es infinita.

  En casos prácticos, el muestreo de una población finita que es muy grande, puede considerarse como muestreo de una población infinita.

HACER UNA ENCUESTA

Una "encuesta" es el trabajo estadístico que trata de hacer predicciones y generalizaciones sobre toda la población, mediante la información recogida en la muestra.

La información obtenida en la muestra (a base de preguntas) puede referirse a hechos determinados:

- Número de coches matriculados en Madrid durante 1986.
- Número de estudiantes que cursan C.O.U.
- Número de personas que votan al P.S.O.E.
- Alturas y pesos de los estudiantes de la Universidad Autónoma de Madrid.

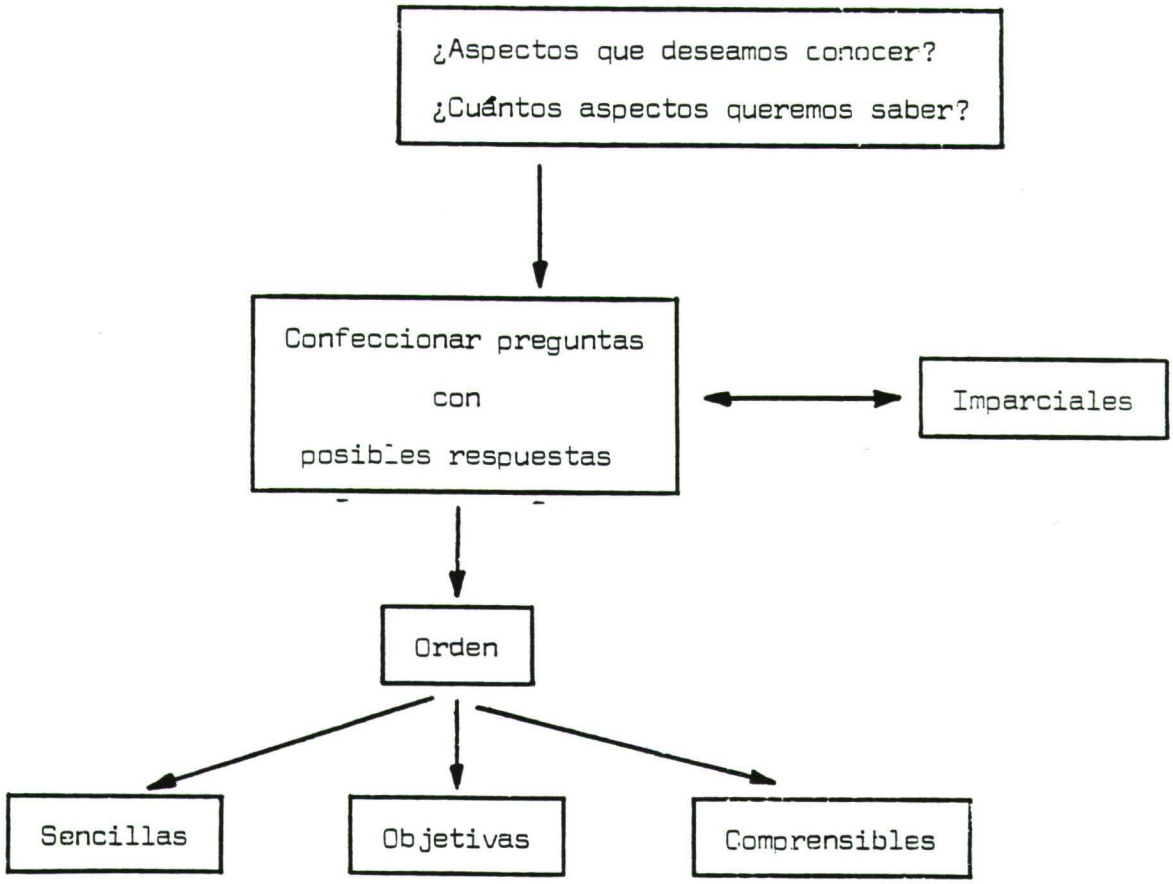
Piensa que hay encuestas más fáciles de confeccionar que otras:

- Estudiantes que no enseñan las notas trimestrales.
- Estudiantes que no dicen el número de hermanos que tienen.

Parece que los resultados de la primera encuesta sean menos precisos que los de la segunda encuesta.

Para recoger información veraz de un conjunto significativo (tamaño) de personas, habrán de redactarse preguntas claras y concretas con un número - limitado de respuestas, para que cada persona pueda elegir la que considere más conveniente.

Análisis de la redacción de preguntas:



P I E N S A

ORDEN:	Una respuesta no tiene que condicionar la respuesta de una pregunta posterior.
OBJETIVAS:	Las alternativas ofrecidas en las preguntas no tienen que destacar.
SENCILLAS:	Las preguntas tienen que estar redactadas de forma breve, escuetas y claras.
COMPRENSIBLE:	Las preguntas han de ser asequibles para todos.

II

VARIABLE ESTADISTICA BIDIMENSIONAL. REGRESION Y CORRELACION

- Variables estadísticas bidimensionales.
- Regresión.
- Correlación.

UN ENCUENTRO INESPERADO

Un día, después de muchos años, Carmen se encuentra con una compañera de estudios:

CARMEN:

¡María! ¡Cuántos años sin vernos! ¿Qué has hecho durante estos años?.

MARIA:

¡Nada en especial!. Estoy casada y tengo un niño.

CARMEN:

Caramba, ¿sólo uno?; ¿te has fijado que hoy día la gente tiene menos hijos que antes?. ¡Mira por ahí viene el que era nuestro profesor de matemáticas!

DON GREGORIO:

¡Hola! ¿Qué es de vuestra vida?

MARIA:

Estabamos hablando del número de hijos que tiene la gente hoy día.

DON GREGORIO:

¿Qué os parece si hacemos una encuesta y vemos la relación que hay entre la edad de las personas y el número de hijos que tienen o piensan tener?.

CARMEN:

¡De acuerdo! Haremos una encuesta y reuniremos los datos.

MARIA:

¿Quedamos en mi casa la semana que viene?.

DON GREGORIO:

¡Hasta la semana que viene entonces!

Al cabo de una semana habían reunido los siguientes datos:

(X = edad , Y = nº de hijos)

X \ Y	0	1	2	3	4	5	6
25	25	10	2	0	0	0	0
30	20	15	3	1	0	0	0
35	5	20	7	5	0	0	0
40	2	8	9	16	6	2	1
45	1	6	10	17	7	3	1

Para responder a estas preguntas hemos de tener en cuenta solamente la variable X= edad y el número de veces que aparece repetida cada edad (x_1= 25 años, x_2= 30 años, x_3= 35 años, x_4= 40 años y x_5= 45 años) sin que intervengan para nada los valores de la variable Y= número de hijos.

Toda la información recogida la representamos en una TABLA DE EDAD.

TABLA DE LA EDAD

x_i=edad	Suma de frecuencias	n_{x_i} (Frecuencias marginales)
25	25 + 10 + 2 + 0 + 0 + 0 + 0	37
30	20 + 15 + 3 + 1 + 0 + 0 + 0	39
35	5 + 20 + 7 + 5 + 0 + 0 + 0	37
40	2 + 8 + 9 + 16 + 6 + 2 + 1	44
45	1 + 6 + 10 + 17 + 7 + 3 + 1	45

- nº total de personas encuestadas 202

siendo n_{x_i} el número de veces que aparece repetida la edad x_i. En otras palabras:

x_1= 25 años tienen 25 + 10 + 2 + 0 + 0 + 0 + 0 = 37 personas = n_{x_1}

x_2= 30 años tienen 20 + 15 + 3 + 1 + 0 + 0 + 0 = 39 personas = n_{x_2}

x_3= 35 años tienen 5 + 20 + 7 + 5 + 0 + 0 + 0 = 37 personas = n_{x_3}

x_4= 40 años tienen 2 + 8 + 9 + 16 + 6 + 2 + 1 = 44 personas = n_{x_4}

x_5= 45 años tienen 1 + 6 + 10 + 17 + 7 + 3 + 1 = 45 personas = n_{x_5}

Como habrás tenido ocasión de observar:

$$\sum_{i=1}^{5} n_{x_i} = n_{x_1} + n_{x_2} + n_{x_3} + n_{x_4} + n_{x_5} = 202 \text{ personas encuestadas.}$$

Pues bien, n_{x_i} recibe el nombre de "FRECUENCIA MARGINAL de la edad x_i", y la tabla de la edad representada con la variable X y sus frecuencias (número de veces que aparece repetida cada edad x_i) forman la "DISTRIBUCION MARGINAL de la X".

Para responder a la pregunta de cuántas personas quieren o tienen un niño hemos de proceder de manera análoga para la variable Y= nº de hijos. Es decir, hemos de tener en cuenta solamente la variable Y= número de hijos y el número de hijos que aparecen, sin que para nada intervengan los valores de la X.

Sea:

TABLA DEL Nº DE HIJOS

y_j = nº de hijos	0	1	2	3	4	5	6	
	25 + 20 + 5 + 2 + 1	10 + 15 + 20 + 8 + 6	2 + 3 + 7 + 9 + 10	0 + 1 + 5 + 16 + 17	0 + 0 + 0 + 6 + 7	0 + 0 + 0 + 2 + 3	0 + 0 + 0 + 1 + 1	
(Frecuencias marginales) n_{y_j}	53	59	31	39	13	5	2	202 (nº total)

siendo n_{y_j} el número de personas entrevistadas que tienen y_j hijos. En otras palabras:

y_1= 0 hijos tienen 25 + 20 + 5 + 2 + 1 = 53 personas = n_{y_1}

y_2= 1 hijo tienen 10 + 15 + 20 + 8 + 6 = 59 personas = n_{y_2}

y_3= 2 hijos tienen 2 + 3 + 7 + 9 + 10 = 31 personas = n_{y_3}

y_4= 3 hijos tienen 0 + 1 + 5 + 16 + 17 = 39 personas = n_{y_4}

y_5= 4 hijos tienen 0 + 0 + 0 + 6 + 7 = 13 personas = n_{y_5}

y_6= 5 hijos tienen 0 + 0 + 0 + 2 + 3 = 5 personas = n_{y_6}

y_7= 6 hijos tienen 0 + 0 + 0 + 1 + 1 = 2 personas = n_{y_7}

Habrás podido observar:

$$\sum_{j=1}^{7} n_{y_j} = n_{y_1} + n_{y_2} + n_{y_3} + n_{y_4} + n_{y_5} + n_{y_6} + n_{y_7} = 202 \text{ personas encuestadas}$$

donde n_{y_j} recibe el nombre de "FRECUENCIA MARGINAL del valor y_j", y la tabla representada con la variable Y= número de hijos y sus frecuencias marginales

n_{y_j}, recibe el nombre de "DISTRIBUCION MARGINAL de la Y".

En efecto, D. Gregorio indica a Carmen y María que se pueden obtener las distribuciones marginales de la X=edad e Y=número de hijos, utilizando la tabla de doble entrada. Veamos:

TABLA DE DOBLE ENTRADA

X \ Y	0	1	2	3	4	5	6	n_{x_i}
25	25	10	2	0	0	0	0	37
30	20	15	3	1	0	0	0	39
35	5	20	7	5	0	0	0	37
40	2	8	9	16	6	2	1	44
45	1	6	10	17	7	3	1	45
n_{y_j}	53	59	31	39	13	5	2	202 = N

X = edad Y = nº de hijos N = nº total de personas entrevistadas

DISTRIBUCION MARGINAL DE LA X

x_i	n_{x_i}	f_{x_i}
25	37	37/202
30	39	39/202
35	37	37/202
40	44	44/202
45	45	45/202
SUMA	202	1

(frecuencias relativas marginales)

DISTRIBUCION MARGINAL DE LA Y

y_j	n_{y_j}	f_{y_j}
0	53	53/202
1	59	59/202
2	31	31/202
3	39	39/202
4	13	13/202
5	5	5/202
6	2	2/202
SUMA	202	1

(frecuencias relativas marginales)

A veces aparecen también las "frecuencias relativas marginales". La frecuencia relativa marginal de un valor observado es el cociente entre su frecuencia absoluta marginal y el total de observaciones realizadas:

$$f_{x_i} = \frac{n_{x_i}}{N} \qquad f_{y_j} = \frac{n_{y_j}}{N}$$

"El porcentaje" de veces que aparece un determinado valor marginal observado

se obtiene multiplicando su frecuencia relativa marginal por 100. De esta manera:

- ¿Qué porcentaje de personas encuestadas tienen 25 años?:

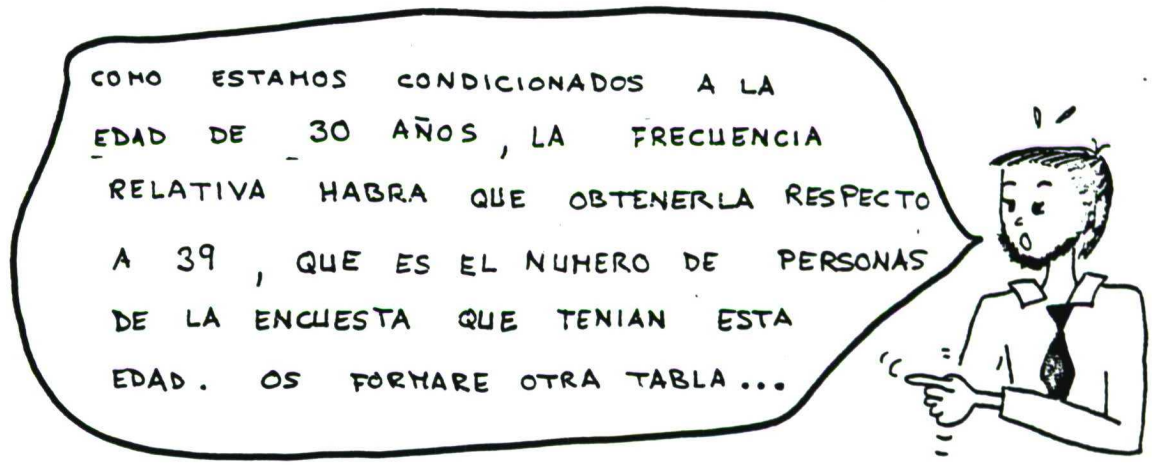
$$f_{x_i} \cdot 100 = \frac{37}{202} \cdot 100 = 18'31\ \%$$

- ¿Qué porcentaje de personas tienen 2 hijos?:

$$f_{y_3} \cdot 100 = \frac{31}{202} \cdot 100 = 15'34\ \%$$

TABLA DE DOBLE ENTRADA

Y= nº hijos / X= edad	0	1	2	3	4	5	6	n_{x_i}
x_1= 25								
x_2= 30	20	15	3	1	0	0		39
35								
40								



DISTRIBUCION DE LA VARIABLE Y= nº de hijos CONDICIONADA A X= 30 AÑOS.

Y	0	1	2	3	4	5	6	
n(Y/X= 30)	20	15	3	1	0	0	0	20+15+3+1 = 39
f(Y/X= 30)	20/39	15/39	3/39	1/39	0	0	0	$\frac{20+15+3+1}{39} = 1$

DISTRIBUCION DE LA VARIABLE X= edad CONDICIONADA A TENER Y= 2 HIJOS.

X = edad	n(X/Y= 2)	f(X/Y= 2)
25	2	2/31
30	3	3/31
35	7	7/31
40	9	9/31
45	10	10/31
	31	1

A partir de las distribuciones marginales de la X= edad y de la Y= número de hijos podemos definir sus medias, varianzas y desviaciones típicas:

- Para la distribución marginal de la variable X= edad:

$$\bar{x} = \frac{\sum x_i \cdot n_{x_i}}{N} \quad , \quad \sigma_x^2 = \frac{\sum (x_i - \bar{x})^2 n_{x_i}}{N} \quad , \quad \sigma_x = \sqrt{\sigma_x^2}$$

siendo, N= número total de personas encuestadas.

- Para la distribución marginal de la variable Y= número de hijos:

$$\bar{y} = \frac{\sum y_j \cdot n_{y_j}}{N} \quad , \quad \sigma_y^2 = \frac{\sum (y_j - \bar{y})^2 n_{y_j}}{N} \quad , \quad \sigma_y = \sqrt{\sigma_y^2}$$

N = número total de personas encuestadas.

* Vamos a la distribución marginal de la X.

x_i	n_{x_i}	$x_i \cdot n_{x_i}$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 n_{x_i}$
25	37	925	- 10'51	110'46	4087'02
30	39	1170	- 5'51	30'36	1184'04
35	37	1295	- 0'51	0'26	9'62
40	44	1760	4'48	20'07	883'08
45	45	2025	9'48	89'87	4044'15
	$\sum_{i=1}^{5} n_{x_i} = 202$	$\sum_{i=1}^{5} x_i . n_{x_i} = 7175$			$\sum_{i=1}^{5} (x_i - \bar{x})^2 n_{x_i} = 10.207'91$

Para hallar la media $\bar{x}$:

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i \cdot n_{x_i}}{N} = \frac{7175}{202} = 35'51$$

La varianza σ_x^2 será:

$$\sigma_x^2 = \frac{\sum_{i=1}^{5} (x_i - \bar{x})^2 n_{x_i}}{N} = \frac{10.207'91}{202} = 50'53$$

La desviación típica σ_x es:

$$\sigma_x = \sqrt{50'53} = 7'10$$

en otras palabras, entre las personas encuestadas las edades están dispersas aproximadamente 7 años.

* Análogamente, si quisiésemos obtener la media, varianza y desviación típica de la variable Y = número de hijos, se tiene:

DISTRIBUCION MARGINAL DE LA Y

y_j	n_{y_j}	$y_j \cdot n_{y_j}$	$(y_j - \bar{y})$	$(y_j - \bar{y})^2$	$(y_j - \bar{y})^2 n_{y_j}$
0	53	0	- 1'61	2'59	137'27
1	59	59	- 0'61	0'37	21'83
2	31	62	0'39	0'15	4'65
3	39	117	1'39	1'93	75'27
4	13	52	2'39	5'71	74'23
5	5	25	3'39	11'49	57'45
6	2	12	4'39	19'27	38'54
	$\sum_{j=1}^{7} n_{y_j} = 202$	$\sum_{j=1}^{7} y_j \cdot n_{y_j} = 327$			$\sum_{j=1}^{7} (y_j - \bar{y})^2 n_{y_j} = 409'24$

Para hallar la media $\bar{y}$:

$$\bar{y} = \frac{\sum_{j=1}^{7} y_j \cdot n_{y_j}}{N} = \frac{327}{202} = 1'61$$

La varianza σ_y^2 es:

$$\sigma_y^2 = \frac{\sum_{j=1}^{7} (y_j - \bar{y})^2 n_{y_j}}{N} = \frac{409'24}{202} = 2'025$$

La desviación típica σ_y :

$$\sigma_y = \sqrt{2'025} = 1'42$$

esto nos indica que existe una dispersión entre 1 y 2 hijos en las personas que hemos realizado la encuesta.


¡PARA FASTIDIAR!
ENTONCES...
LA VERDAD ES QUE NO HACIA FALTA PERDER TANTO TIEMPO. LA CALCULADORA FACILITA RAPIDAMENTE: $\bar{x}, \bar{y}, \sigma_x^2, \sigma_y^2, \sigma_x, \sigma_y$

RECUERDA

• VARIABLE ESTADISTICA BIDIMENSIONAL

Hasta aquí hemos considerado una sola variable. Ahora vamos a estudiar conjuntamente dos variables. Por ejemplo:

- peso y altura de un grupo de estudiantes.
- aptitud para una asignatura y aprovechamiento en la misma.
- consumo de tabaco y el cáncer de pulmón.
- provincia de origen y carrera estudiada.

En estas situaciones el estadístico realiza la observación simultánea de dos caracteres en el individuo, obteniéndose, por tanto, pares de resultados.

Piensa que un carácter puede tomar distintas modalidades. Así:

- el carácter X puede adoptar las modalidades $(x_1, x_2, \ldots, x_k)$
- el carácter Y puede adoptar las modalidades $(y_1, y_2, \ldots, y_l)$

Los distintos valores de las modalidades que pueden adoptar estos caracteres forman un conjunto de pares, que representamos por (X,Y), y llamaremos "variable estadística bidimensional".

Ahora bien, recuerda que el carácter es una cualidad o propiedad inherente en el individuo. Hay caracteres que son medibles (se pueden cuantificar), como, por ejemplo, la edad, el peso y la estatura de las personas. A estos caracteres se les llama "cuantitativos". Pero hay otros caracteres que no son medibles, como, por ejemplo, el color de los ojos, el sexo, etc. A estos caracteres se les llama "cualitativos".

Cuando el estadístico realiza la observación de dos caracteres, estos no tienen por qué ser de la misma clase. Así, se pueden presentar posibles situaciones:

- Dos caracteres cuantitativos: peso y estatura de una persona.
- Dos caracteres cualitativos: sexo y color del pelo de una persona.
- Uno cuantitativo y otro cualitativo: peso y color del pelo de una persona.

Desde luego, podríamos ir estudiando por separado cada una de las tres situaciones. Sin embargo, no seguiremos este camino por una doble razón: En primer lugar, sería enormemente extenso. En segundo lugar, ello nos llevaría a repetirnos ya que lo dicho para uno de los casos, vale prácticamente para los restantes, salvo diferencias accidentales.

Consiguientemente, sean las variables X e Y estrictamente cuantitativas, por ser el caso más común.

• ORDENACION DE DATOS

Observa que el número de modalidades distintas que adopta el carácter $X = (x_1, x_2, x_3, \ldots, x_k)$ no tiene por qué ser el mismo número que el que adopta el carácter Y.

Nos encontramos con el problema de ordenar los datos de forma que tengan cabida los k valores distintos de la variable X y los l valores distintos de la variable Y.

En una "tabla de doble entrada" podemos expresar el número de modalidades distintas que adoptan los caracteres (X,Y), de tal forma que, allí podremos reflejar el número de veces que se repite cada par de valores posibles.

Sea, la TABLA DE DOBLE ENTRADA:

X \ Y	y_1	y_2	y_3	y_j	y_l
x_1	n_{11}	n_{12}	n_{13}	n_{1j}	n_{1l}
x_2	n_{21}	n_{22}	n_{23}	n_{2j}	n_{2l}
⋮				⋮	
x_i	n_{i1}	n_{i2}	n_{i3}	n_{ij}	n_{il}
⋮					
x_k	n_{k1}	n_{k2}	n_{k3}	n_{kj}	n_{kl}

siendo

- n_{ij} el número de veces que aparece repetido el par (x_i, y_i), y que llamaremos "frecuencia absoluta del par (x_i, y_i)".

- f_{ij} es la frecuencia relativa del par observado (x_i, y_i), que vendrá dada por el cociente entre su frecuencia absoluta n_{ij} y el total de pares observados N, es decir:

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

El "porcentaje" de veces que aparece el par observado (x_i, y_i) se obtiene multiplicando $(f_{ij} \times 100)$.

X	1	1	3	5	5	3	1	3	3	1	3	5	5
Y	1	4	1	2	6	6	1	1	1	6	2	6	6

Observa que el número de modalidades distintas que adopta el carácter X no es el mismo número de modalidades que el que adopta el carácter Y:

$$X = \{1, 3, 5\}$$

$$Y = \{1, 2, 4, 6\}$$

Parece lógico ordenar los datos de la mejor forma posible en una tabla de doble entrada donde tengan cabida los 3 valores distintos de la variable X ($i = 1, 2, 3$) y los 4 valores distintos de la variable Y ($j = 1, 2, 3, 4$).

Allí podremos expresar el número de veces que se repite cada par (x_i, y_j):

$$(1,1)\ ,\ (1,4)\ ,\ (1,6)$$

$$(3,1)\ ,\ (3,6)\ ,\ (3,2)$$

$$(5,2)\ ,\ (5,6)$$

Sea la TABLA DE DOBLE ENTRADA:

X \ Y	1	2	4	6	
1	2	0	1	1 →	4
3	3	1	0	1	5
5	0	1	0	3	4
	5	2	1	5 →	N = 13 = $\sum_{i=1}^{3}\sum_{j=1}^{4} n_{ij}$

donde $n_{22} = 1$ es el número de veces que aparece repetido el par (x_2, y_2) y que llamaremos "frecuencia absoluta" del par (x_2, y_2).

Notaremos por f_{22} la "frecuencia relativa" correspondiente al par (x_2, y_2), que viene dada por la expresión:

$$f_{22} = \frac{n_{22}}{N} = \frac{1}{13}$$

siendo N el número total de pares observados.

OBSERVA LAS PROPIEDADES SIGUIENTES:

- La suma de frecuencias absolutas es igual al número de pares observados:

$$\sum_{i=1}^{3} \sum_{j=1}^{4} n_{ij} = 13 = N$$

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

$$\sum_{i=1}^{3} \sum_{j=1}^{4} f_{ij} = \sum_{i=1}^{3} \sum_{j=1}^{4} \frac{n_{ij}}{N} = \frac{1}{N} \sum_{i=1}^{3} \sum_{j=1}^{4} n_{ij} =$$

$$= \frac{1}{N} \sum_{i=1}^{3} (n_{i1} + n_{i2} + n_{i3} + n_{i4}) =$$

$$= \frac{1}{N} \left[(n_{11} + n_{12} + n_{13} + n_{14}) + (n_{21} + n_{22} + n_{23} + n_{24}) + (n_{31} + n_{32} + n_{33} + n_{34}) \right] =$$

$$= \frac{1}{13} \left[(2 + 0 + 1 + 1) + (3 + 1 + 0 + 1) + (0 + 1 + 0 + 3) \right] =$$

$$= \frac{1}{13} (4 + 5 + 4) = 1$$

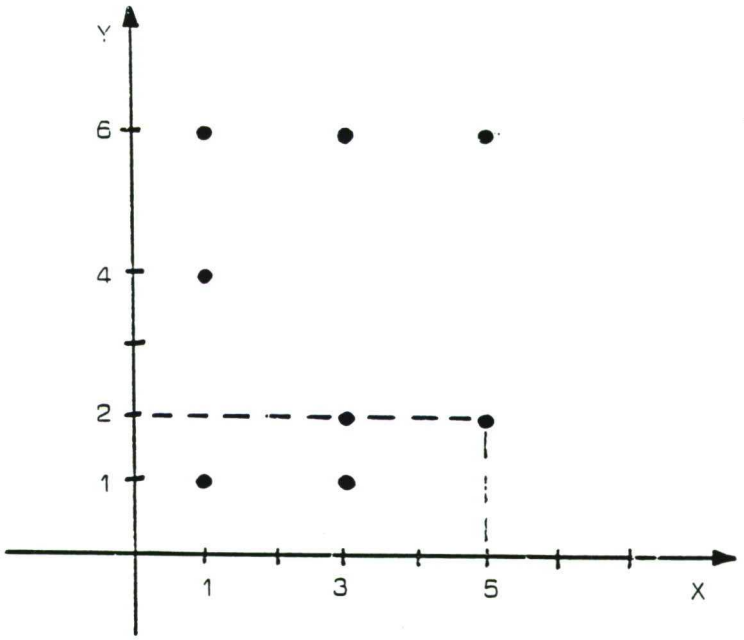
EN GENERAL

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

$$\sum_{i=1}^{k} \sum_{j=1}^{l} f_{ij} = 1$$

¿ LA REPRESENTACIÓN GRÁFICA?

Los valores de la variable X se indican sobre el eje horizontal y los valores de la variable Y en el eje vertical, obteniéndose, por tanto, el DIAGRAMA DE DISPERSION:



• DISTRIBUCIONES MARGINALES

En una tabla de doble entrada recogemos la información de pares de resultados (X,Y):

TABLA DE DOBLE ENTRADA

X \ Y	y_1	y_2	...	y_j	...	y_l	
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1l}	n_{x_1}
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2l}	n_{x_2}
⋮				⋮			
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{il}	n_{x_i}
⋮				⋮			
x_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kl}	n_{x_k}
	n_{y_1}	n_{y_2}	...	n_{y_j}	...	n_{y_l}	N

siendo:

- n_{x_i} es el número de observaciones realizadas cuando el valor x_i. Es decir:

$$n_{x_i} = n_{i1} + n_{i2} + n_{i3} + \cdots + n_{ij} + \cdots + n_{il}$$

o también:

$$n_{x_i} = \sum_{j=1}^{l} n_{ij} = n_{i1} + n_{i2} + \cdots + n_{ij} + \cdots + n_{il}$$

Pues bien, n_{x_i} recibe el nombre de "frecuencia marginal del valor x_i".

- n_{y_j} es el número de observaciones realizadas cuando el valor y_j. Esto es:

$$n_{y_j} = n_{1j} + n_{2j} + \dots + n_{ij} + \dots + n_{kj}$$

o también:

$$n_{y_j} = \sum_{i=1}^{k} n_{ij} = n_{1j} + n_{2j} + \dots + n_{ij} + \dots + n_{kj}$$

donde n_{y_j} recibe el nombre de "frecuencia marginal del valor y_j".

Piensa que ahora estamos interesados en responder a dos tipos de preguntas:

- Preguntas que hagan referencia solamente a la variable X y el recuento de sus frecuencias, sin que intervengan para nada los valores de la Y.
- Preguntas que hagan referencia solamente a la variable Y y el recuento de sus frecuencias, sin que para nada intervengan los valores de la X.

Por tanto, se tiene:

a) DISTRIBUCION MARGINAL DE X

Llamamos distribución marginal de X a la distribución en X de todas las observaciones, independientemente de sus puntuaciones en Y. Viene dada, en la tabla de doble entrada, por las columnas situadas en los extremos. Es decir:

$$X = (x_1, x_2, \dots, x_i, \dots, x_k)$$

$(n_{x_1}, n_{x_2}, \dots, n_{x_i}, \dots, n_{x_k})$ respectivas frecuencias marginales

De una manera formal, la "distribución marginal de X" es la siguiente:

X	n_{x_i}	f_{x_i}
x_1	n_{x_1}	$f_{x_1} = n_{x_1}/N$
x_2	n_{x_2}	$f_{x_2} = n_{x_2}/N$
$\vdots$	$\vdots$	$\vdots$
x_i	n_{x_i}	$f_{x_i} = n_{x_i}/N$
$\vdots$	$\vdots$	$\vdots$
x_k	n_{x_k}	$f_{x_k} = n_{x_k}/N$
	$\sum_{i=1}^{k} n_{x_i} = N$	$\sum_{i=1}^{k} f_{x_i} = 1$

A partir de las frecuencias absolutas marginales n_{x_i} se obtienen las frecuencias relativas marginales f_{x_i}, de forma que:

$$f_{x_i} = \frac{n_{x_i}}{N}$$ frecuencia relativa marginal del valor x_i

"Observa las propiedades de las frecuencias marginales":

$$- \sum_{i=1}^{k} n_{x_i} = n_{x_1} + n_{x_2} + n_{x_3} + \dots + n_{x_i} + \dots + n_{x_k} = N$$

$$- \sum_{i=1}^{k} f_{x_i} = f_{x_1} + f_{x_2} + f_{x_3} + \dots + f_{x_i} + \dots + f_{x_k} =$$

$$= \frac{1}{N} (n_{x_1} + n_{x_2} + n_{x_3} + \dots + n_{x_i} + \dots + n_{x_k}) = \frac{N}{N} = 1$$

siendo N el número total de observaciones.

Con la distribución marginal de X tendremos una media $\bar{x}$, una varianza σ_x^2, y una desviación típica σ_x, que llamaremos media, varianza, y desviación típica marginales. Vienen dadas por las siguientes expresiones:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot n_{x_i}}{N} \quad \text{media marginal de la X}$$

$$\sigma_x^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_{x_i}}{N} \quad \text{varianza marginal de la X}$$

$$\sigma_x = +\sqrt{\sigma_x^2} \quad \text{desviación típica marginal de la X}$$

b) DISTRIBUCION MARGINAL DE Y

Llamamos distribución marginal de Y a la distribución en Y de todas las observaciones, independientemente de sus puntuaciones en X. Viene dada, en la tabla de doble entrada, por las filas situadas en los márgenes superior e inferior. Es decir:

$$Y = (y_1, y_2, \ldots, y_j, \ldots, y_l)$$

$(n_{y_1}, n_{y_2}, \ldots, n_{y_j}, \ldots, n_{y_l})$ respectivas frecuencias marginales

De un modo formal, la "distribución marginal de Y" es la siguiente:

Y	n_{y_j}	f_{y_j}
y_1	n_{y_1}	$f_{y_1} = n_{y_1}/N$
y_2	n_{y_2}	$f_{y_2} = n_{y_2}/N$
⋮	⋮	⋮
y_j	n_{y_j}	$f_{y_j} = n_{y_j}/N$
⋮	⋮	⋮
y_l	n_{y_l}	$f_{y_l} = n_{y_l}/N$
	$\sum_{j=1}^{l} n_{y_j} = N$	$\sum_{j=1}^{l} f_{y_j} = 1$

A partir de las frecuencias absolutas marginales n_{y_j} se obtienen las frecuencias relativas marginales f_{y_j}, de la siguiente manera:

$$f_{y_j} = \frac{n_{y_j}}{N}$$ frecuencia relativa marginal del valor y_j

"Observa las propiedades de las frecuencias marginales:

$$- \sum_{j=1}^{l} n_{y_j} = n_{y_1} + n_{y_2} + \dots + n_{y_j} + \dots + n_{y_l} = N$$

$$- \sum_{j=1}^{l} f_{y_j} = f_{y_1} + f_{y_2} + \dots + f_{y_j} + \dots + f_{y_l} =$$

$$= \frac{1}{N} (n_{y_1} + n_{y_2} + \dots + n_{y_j} + \dots + n_{y_l}) = \frac{N}{N} = 1$$

siendo N el número total de observaciones.

Con la distribución marginal de Y tendremos una media marginal $\bar{y}$, una varianza marginal σ_y^2, y una desviación típica marginal σ_y, tales que:

$$\bar{y} = \frac{\sum_{j=1}^{l} y_j \cdot n_{y_j}}{N}$$ media marginal de la Y

$$\sigma_y^2 = \frac{\sum_{j=1}^{l} (y_j - \bar{y})^2 \cdot n_{y_j}}{N}$$ varianza marginal de la Y

$$\sigma_y = + \sqrt{\sigma_y^2}$$ desviación típica marginal de la Y

- DISTRIBUCIONES CONDICIONADAS

Sea como referencia la tabla de doble entrada:

Y / X	y_1	y_2	...	y_j	...	y_l	MARGINALES DE LA X
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1l}	n_{x_1}
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2l}	n_{x_2}
⋮				⋮			⋮
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{il}	n_{x_i}
⋮				⋮			⋮
x_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kl}	n_{x_k}
MARGINALES DE LA Y	n_{y_1}	n_{y_2}	...	n_{y_j}	...	n_{y_l}	N

Ahora estamos interesados en dos clases de preguntas:

- Cuando "condicionamos" la variable X y el recuento de sus frecuencias, solamente a un valor determinado de la Y, sea $Y = y_j$.

- Cuando "condicionamos" la variable Y y el recuento de sus frecuencias, solamente a un valor determinado de la X, sea este - $X = x_i$.

Estos dos tipos de preguntas dan lugar a:

a) DISTRIBUCION CONDICIONADA DE X A y_j

Llamamos distribución condicionada de X, para $Y = y_j$, a la distribución en X de todas, y solas, las observaciones con valores y_j.

Es decir, estamos condicionando los valores de la variable X al valor y_j.

Nos fijamos en la tabla de doble entrada:

X \ Y	... y_j ...
x_1	n_{1j}
x_2	n_{2j}
⋮	⋮
x_i	n_{ij}
⋮	⋮
x_k	n_{kj}
	n_{y_j}

Se tendrá, por tan...

DISTRIBUCION CONDICIONADA DE X A y_j

X	$n(X/Y = y_j)$	$f(X/Y = y_j)$
x_1	n_{1j}	$f_{1j} = n_{1j}/n_{y_j}$
x_2	n_{2j}	$f_{2j} = n_{2j}/n_{y_j}$
⋮	⋮	⋮
x_i	n_{ij}	$f_{ij} = n_{ij}/n_{y_j}$
⋮	⋮	⋮
x_k	n_{kj}	$f_{kj} = n_{kj}/n_{y_j}$
	n_{y_j}	1

De una manera formal, queda:

$$n(X/Y = y_j) = \left\{n_{1j},\ n_{2j},\ \dots,\ n_{ij},\ \dots,\ n_{kj}\right\}$$

$$f(X/Y = y_j) = \left\{f_{1j},\ f_{2j},\ \dots,\ f_{ij},\ \dots,\ f_{kj}\right\}$$

siendo las frecuencias relativas condicionadas:

$$f(x_i/Y = y_j) = \frac{n(x_i/Y = y_j)}{n_{y_j}}$$

"Observa la propiedad de la frecuencia relativa condicionada":

$$\sum_{i=1}^{k} f(X/y_j) = f_{1j} + f_{2j} + \dots + f_{ij} + \dots + f_{kj} =$$

$$= \frac{1}{n_{y_j}} (n_{1j} + n_{2j} + \dots + n_{ij} + \dots + n_{kj}) = \frac{n_{y_j}}{n_{y_j}} = 1$$

b) DISTRIBUCION CONDICIONADA DE Y A x_i

Análogamente, ahora estamos condicionando los valores de la variable Y al valor x_i.

Nos fijamos en la tabla de doble entrada:

X \ Y	y_1	y_2	...	y_j	...	y_l
⋮						
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{x_i}
⋮						

se sigue que

DISTRIBUCION CONDICIONADA DE Y A x_i

Y	y_1	y_2	...	y_j	...	y_l	
$n(Y/X = x_i)$	n_{i1}	n_{i2}	...	n_{ij}	...	n_{il}	.. n_{x_i}
$f(Y/X = x_i)$	f_{i1}	f_{i2}	...	f_{ij}	...	f_{il}	.. 1

de esta forma, tendremos:

- frecuencia absoluta condicionada:

$$n(Y/X = x_i) = \{n_{i1}, n_{i2}, \ldots, n_{ij}, \ldots, n_{il}\}$$

- frecuencia relativa condicionada:

$$f(Y/X = x_i) = \{f_{i1}, f_{i2}, \ldots, f_{ij}, \ldots, f_{il}\}$$

es decir:

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

"Observa la propiedad de la frecuencia relativa condicionada":

$$\sum_{j=1}^{l} f(Y/x_i) = f_{i1} + f_{i2} + \dots + f_{ij} + f_{il} =$$

$$= \frac{1}{n_{x_i}} (n_{i1} + n_{i2} + \dots + n_{ij} + \dots + n_{il}) =$$

$$= \frac{n_{x_i}}{n_{x_i}} = 1$$

- MOMENTOS

Definimos el momento de órdenes r y s respecto al par de parámetros (c,v), de la forma:

$$M_{r,s}(c,v) = \frac{\sum_{i=1}^{k} \sum_{j=1}^{l} (x_i - c)^r . (y_j - v)^s . n_{ij}}{N}$$

En particular, observa dos casos importantes:

a) Decimos que un "momento es respecto al origen" cuando los parámetros c = 0, v = 0, entonces:

$$M_{r,s}(0,0) = \frac{\sum_{i=1}^{k} \sum_{j=1}^{l} (x_i - 0)^r . (y_j - 0)^s . n_{ij}}{N}$$

A los momentos respecto al origen se les denota por $a_{r,s}$, de forma que:

$$a_{r,s} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{l} x_i^r . y_j^s . n_{ij}}{N}$$

dando valores a r y s, son de interés posterior los momentos:

$$a_{00} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} x_i^0 \cdot y_j^0 \cdot n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}}{N} = \frac{N}{N} = 1$$

$$a_{10} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} x_i^1 \cdot y_j^0 \cdot n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} x_i \cdot n_{ij}}{N} = \frac{\sum_{i=1}^{k} x_i \cdot n_{x_i}}{N} = \bar{x}$$

$$a_{01} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} x_i^0 \cdot y_j^1 \cdot n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} y_j \cdot n_{ij}}{N} = \frac{\sum_{j=1}^{l} y_j \cdot n_{ij}}{N} = \bar{y}$$

$$a_{11} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} x_i \cdot y_j \cdot n_{ij}}{N}$$

$$a_{20} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} x_i^2 \cdot y_j^0 \cdot n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} x_i^2 \cdot n_{ij}}{N} =$$

$$= \frac{\sum_{i=1}^{k} x_i^2 \cdot n_{x_i}}{N}$$ (segundo momento respecto de X)

$$a_{02} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} x_i^0 \cdot y_j^2 \cdot n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} y_j^2 \cdot n_{ij}}{N} =$$

$$= \frac{\sum_{j=1}^{l} y_j^2 \cdot n_{y_j}}{N}$$ (segundo momento respecto de Y)

Recuerda:

$$\sum_{i=1}^{k} n_{ij} = n_{1j} + n_{2j} + \cdots + n_{ij} + \cdots + n_{kj} = n_{y_j}$$

$$\sum_{j=1}^{l} n_{ij} = n_{i1} + n_{i2} + \dots + n_{ij} + \dots + n_{il} = n_{x_i}$$

CONCLUSIONES:

$$a_{00} = 1$$

$$a_{10} = \bar{x}$$

$$a_{01} = \bar{y}$$

$$a_{11} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{l} x_i \cdot y_j \cdot n_{ij}}{N}$$

$$a_{20} = \frac{\sum_{i=1}^{k} x_i^2 \cdot n_{x_i}}{N}$$

$$a_{02} = \frac{\sum_{j=1}^{l} y_j^2 \cdot n_{y_j}}{N}$$

b) Decimos que un "momento es respecto a la media" cuando los parámetros - $c = \bar{x}$, $v = \bar{y}$, entonces:

$$M_{r,s}(\bar{x},\bar{y}) = \frac{\sum_{i=1}^{k} \sum_{j=1}^{l} (x_i - \bar{x})^r \cdot (y_j - \bar{y})^s \cdot n_{ij}}{N}$$

A los momentos respecto a la media se les denota por $m_{r,s}$, de forma que:

$$m_{r,s} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{l} (x_i - \bar{x})^r \cdot (y_j - \bar{y})^s \cdot n_{ij}}{N}$$

dando valores a r y s, son de interés posterior los momentos:

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

$$m_{20} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}(x_i - \bar{x})^2 \cdot (y_j - \bar{y})^0 \cdot n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}(x_i - \bar{x})^2 \cdot n_{ij}}{N} =$$

$$= \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot n_{x_i}}{N} = \sigma_x^2$$

$$m_{02} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}(x_i - \bar{x})^0 \cdot (y_j - \bar{y})^2 \cdot n_{ij}}{N} = \frac{\sum_{j=1}^{l}(y_j - \bar{y})^2 \cdot n_{y_j}}{N} = \sigma_y^2$$

CONCLUSIONES:

m_{11} = covarianza

$m_{20} = \sigma_x^2$ varianza marginal de X

$m_{02} = \sigma_y^2$ varianza marginal de Y

- En la práctica, para hallar la covarianza m_{11} es bastante más rápido, utilizar la expresión:

$$m_{11} = a_{11} - a_{10} \cdot a_{01}$$

esto es

$$m_{11} = a_{11} - \bar{x} \cdot \bar{y}$$

- Análogamente, las varianzas marginales de la X e Y tienen un cálculo más rápido:

$$\sigma_x^2 = m_{20} = a_{20} - (a_{10})^2$$

esto es

$$\sigma_x^2 = a_{20} - (\bar{x})^2$$

y por tanto:

$$\sigma_y^2 = a_{02} - (a_{01})^2$$

o bien

$$\sigma_y^2 = a_{02} - (\bar{y})^2$$

ACTIVIDAD – 1: Dada la variable bidimensional (X,Y) con la tabla de frecuencias

X \ Y	1	2	3	4
2	2	1	0	1
3	0	1	3	0
4	0	0	2	1

Se pide:

1) $\sum\sum n_{ij}$

2) a_{10}, a_{01}, a_{11}

3) m_{11}

4) f_{21}, f_{23}

ACTIVIDAD – 2: Dada la siguiente tabla:

X	1	2	2	3	3	4	4	5	5	6
Y	1	3	2	4	2	1	4	4	5	4

Se pide:

1) Medias marginales de la X y la Y

2) Varianzas marginales de la X y la Y

3) Covarianza

ACTIVIDAD – 3: Sabiendo que $\bar{x} = 6$, $\bar{y} = 8$, $m_{11} = 13$, en la siguiente tabla:

X	Y	X.Y
2	■	■
■	4	■
8	10	80
10	14	140

Se pide:

1) Poner los valores que faltan en la tabla.

2) Varianzas marginales de la X y la Y.

AUTOCOMPROBACION

ACTIVIDAD - 1:

1) $\sum\sum n_{ij} = 11 = N$

2) $\begin{cases} a_{10} = \bar{x} = 2'90 \\ a_{01} = \bar{y} = 2'63 \\ a_{11} = 8'09 \end{cases}$

3) m_{11} = covarianza = 0'463

4) $\begin{cases} f_{21} = \dfrac{n_{21}}{N} = \dfrac{0}{11} = 0 \\ f_{23} = \dfrac{n_{23}}{N} = \dfrac{3}{11} \end{cases}$

ACTIVIDAD – 2:

N = 10 observaciones

1) $\begin{cases} \bar{x} = a_{10} = 3'5 & \text{media marginal de la X} \\ \bar{y} = a_{01} = 3 & \text{media marginal de la Y} \end{cases}$

2) $\begin{cases} a_{20} = \dfrac{145}{10} = 14'5 \\ \sigma_x^2 = a_{20} - (a_{10})^2 = 14'5 - (3'5)^2 = 2'25 \text{ varianza marginal de la X.} \end{cases}$

$\begin{cases} a_{02} = \dfrac{108}{10} = 10'8 \\ \sigma_y^2 = a_{02} - (a_{01})^2 = 10'8 - (3)^2 = 1'8 \text{ varianza marginal de la Y.} \end{cases}$

3) $\begin{cases} a_{11} = \dfrac{118}{10} = 11'8 \\ m_{11} = a_{11} - a_{10} \cdot a_{01} = 11'8 - (3'5).(3) = 1'3 \text{ covarianza} \end{cases}$

ACTIVIDAD – 3:

N = 4 observaciones

1) $\begin{cases} \bar{x} = \dfrac{\sum_{i=1}^{4} x_i}{N} = \dfrac{2 + a + 8 + 10}{4} = 6 \\ 20 + a = 24 \qquad a = 4 \end{cases}$

$$\begin{cases} \bar{y} = \dfrac{\sum_{i=1}^{4} y_i}{N} = \dfrac{b + 4 + 10 + 14}{4} = 8 \\ \\ 28 + b = 32 \qquad b = 4 \end{cases}$$

de donde

X	Y	X.Y
2	4	8
4	4	16
8	10	80
10	14	140
SUMA 24	32	244

Observa que la covarianza $m_{11} = 13$ es un dato innecesario.

2)
$$\begin{cases} a_{10} = \bar{x} = 6 \\ \\ a_{20} = \dfrac{184}{4} = 46 \\ \\ \sigma_x^2 = a_{20} - (a_{10})^2 = 46 - (6)^2 = 10 \text{ varianza marginal de la X} \end{cases}$$

$$\begin{cases} a_{01} = \bar{y} = 8 \\ \\ a_{02} = \dfrac{328}{4} = 82 \\ \\ \sigma_y^2 = a_{02} - (a_{01})^2 = 82 - (8)^2 = 18 \text{ varianza marginal de la Y} \end{cases}$$

$$\begin{cases} a_{11} = \dfrac{244}{4} = 61 \\ \\ m_{11} = a_{11} - a_{10} \cdot a_{01} = 61 - (6).(8) = 13 \\ \\ \text{En efecto, la covarianza } m_{11} = 13 \end{cases}$$

ESTADISTICA DESCRIPTIVA

ACTIVIDADES

INSPECCION VETERINARIA

El negocio de Raúl y Chicha estaba en marcha. En un mes habían conseguido un buen número de conejos y en poco tiempo estarían dispuestos para empezar a vender.

Mientras tanto, Raúl se aburría. Su mujer se ocupaba de casi todo y él apenas tenía nada que hacer, así que iba a menudo al bar del pueblo. Allí se encontró un día con Don Justo, el veterinario.

D. JUSTO:

¡Hola, Raúl!. ¿Qué haces tú por aquí?.

RAUL:

Ya ve, pasando el rato. Como no tengo trabajo, me aburro de lo lindo.

D. JUSTO:

¡Vaya, hombre!. Tal vez puedas echarme una mano y ganar un dinerito.

RAUL:

Dígame en qué le podría ayudar.

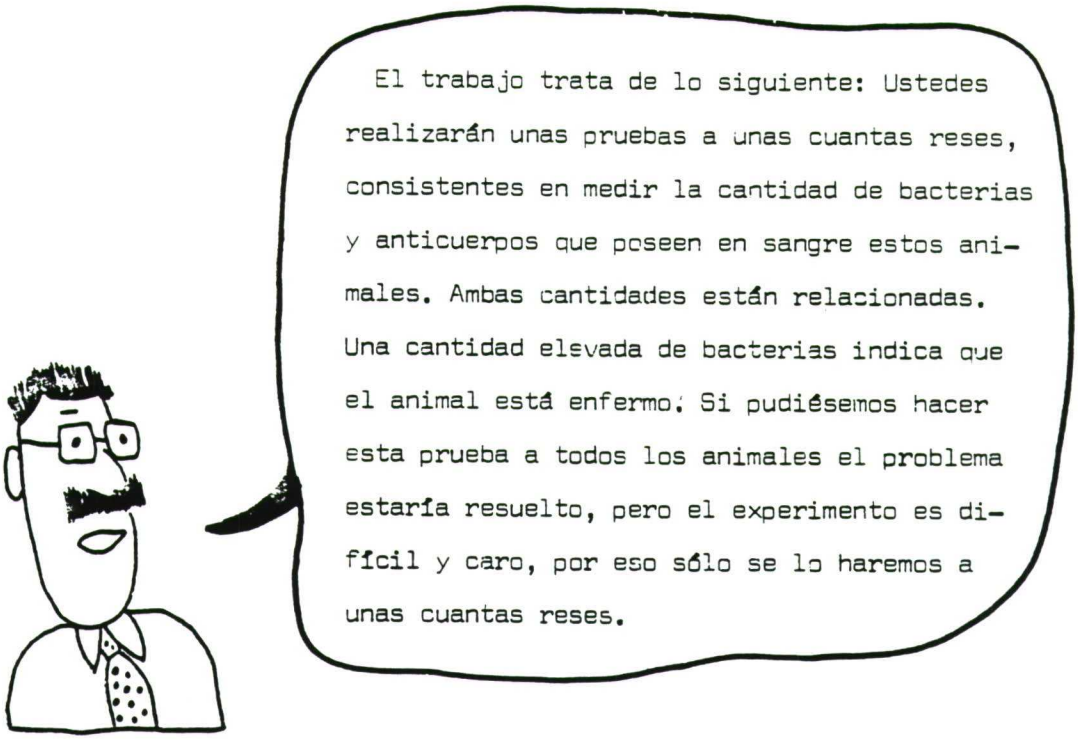
D. JUSTO:

Es muy sencillo. Han enviado una orden del Ministerio por la que tenemos que averiguar si nuestro ganado está enfermo y para ello tenemos que realizar unas pruebas. Nos envían un "experto", pero, aún así, necesitaremos ayuda.

RAUL:

Si el trabajo no es difícil, por mi parte no hay inconveniente en ayudarle.

A los dos días se presentó el experto, Sr. Montes, que dió rápidamente las oportunas indicaciones a Raúl y a D. Justo.



Sr. Montes

RAUL:

Perdón, Sr. Montes ..., no lo entiendo. Si sólo se lo hace a unas cuantas reses, no podrá saber qué pasa con las demás.

SR. MONTES:

Amigo Raúl, ya le he dicho que ambas cantidades están relacionadas. Realizar la primera prueba es fácil y barato, ésta es la que haremos a todas las reses. Conociendo la relación entre ambas podremos preveer aproximadamente qué cantidad de ganado está enfermo en este pueblo.

A la mañana siguiente nuestros amigos se pusieron a trabajar. La tarea de Raúl era sencilla, sólo tenía que ayudar a D. Justo a hacer los análisis a las reses. De la labor técnica se ocuparían después D. Justo y el Sr. Montes, aunque Raúl intentaría enterarse de algo.

D. JUSTO:

Sr. Montes, ya hemos realizado los análisis a seis reses, tal como usted dijo. Los resultados han sido los siguientes:

1ª res	...	10 bacterias	...	13 anticuerpos
2ª "	...	30 "	...	45 "
3ª "	...	20 "	...	18 "
4ª "	...	24 "	...	23 "
5ª "	...	16 "	...	20 "
6ª "	...	25 "	...	30 "

SR. MONTES:

Bien, ahora comienza mi trabajo. Empezaré por ordenar los datos.

RAUL:

¿Ordenarlos?. ¿Cómo?. ¿De mayor a menor?. No entiendo cómo va a hacerlo, porque tiene usted dos tipos de datos distintos, el número de bacterias y el de anticuerpos.

SR. MONTES:

Y ambos corresponden a la misma vaca. Es decir, estudio dos características diferentes de un mismo individuo (con perdón). Cada una de ellas recibe el nombre de variable y en este caso, por ser dos, se trata de una variable bidimensional. Para ordenar los datos, construyo una "tabla de doble entrada" de la manera siguiente:

nº anticuerpos \ nº bacterias	10	16	20	24	25	30	TOTAL
13	1	0	0	0	0	0	1
18	0	0	1	0	0	0	1
20	0	1	0	0	0	0	1
23	0	0	0	1	0	0	1
30	0	0	0	0	1	0	1
45	0	0	0	0	0	1	1
TOTAL	1	1	1	1	1	1	6

Esta tabla me indica que, por ejemplo, he obtenido una vez el resultado 18 anticuerpos y 20 bacterias, mientras que no he obtenido nunca 13 anticuerpos y 25 bacterias. Así mismo, me indica el número de observaciones que he realizado.

D. JUSTO:

Muy ocurrente, de un sólo vistazo podemos saber cuáles han sido nuestro resultados. Sólo veo un inconveniente en esto, es muy pesado a la hora de escribir el arrastrar continuamente las palabras: número de anticuerpos, número de bacterias.

SR. MONTES:

Eso es cierto, para abreviar y reflejar de la forma más clara posible las variables que tenemos, necesitamos algo que recoja la información que éstas llevan. Por lo tanto, vamos a llamar: X = nº de bacterias, Y = nº de anti - cuerpos.

RAUL:

¿Y cómo se las ingenia usted para reflejar los distintos valores que pueden tomar estas variables?.

SR. MONTES:

Es sencillo, utilizo subíndices.

RAUL:

¿Sub- qué?.

D. JUSTO:

Subíndices, Raúl. Calla y escucha.

SR. MONTES:

De esta manera, la variable x_i indica el número de bacterias y toma los valores:

$x_1 = 10$, $x_2 = 16$, $x_3 = 20$,

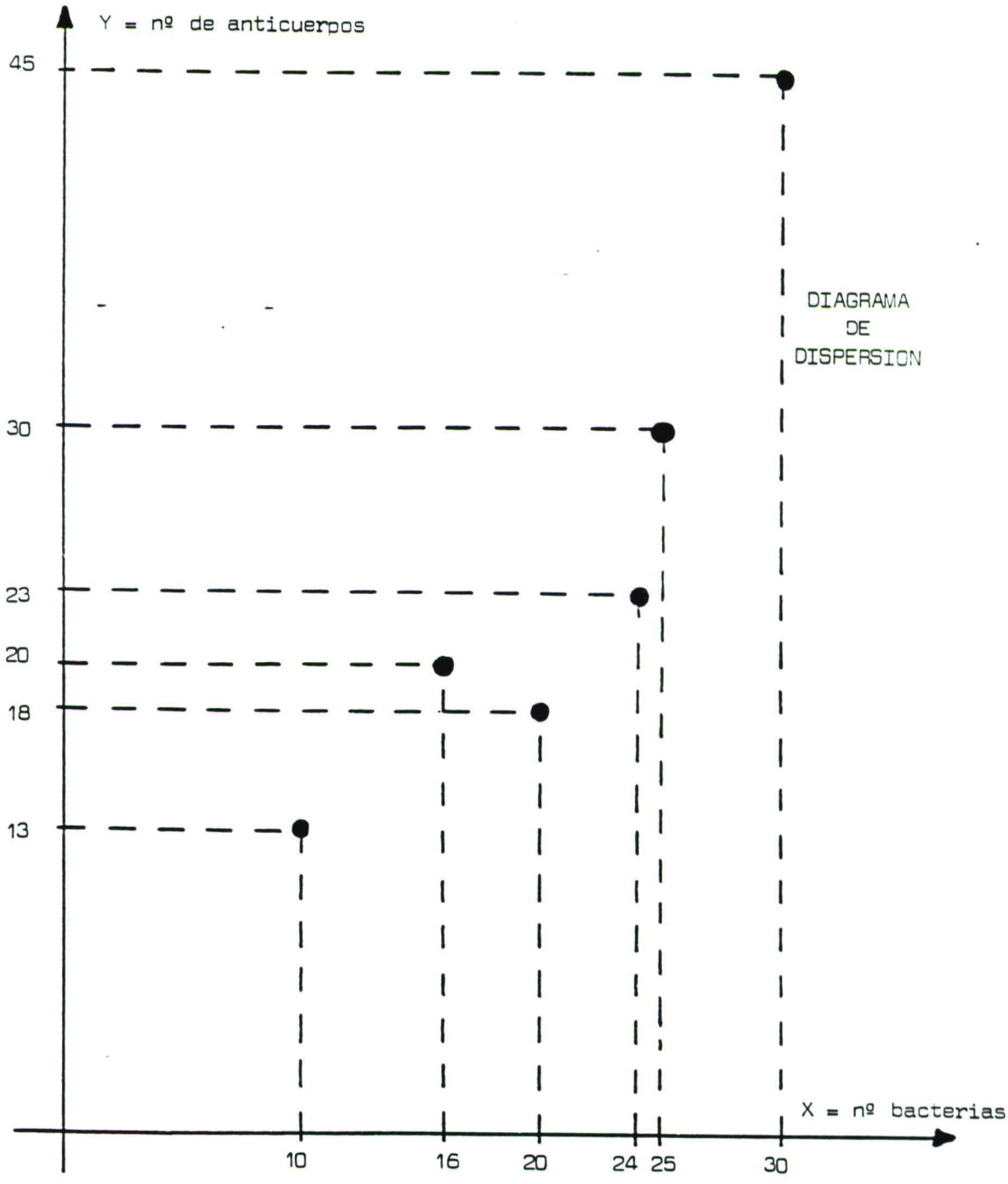
$x_4 = 24$, $x_5 = 25$, $x_6 = 30$.

Y la variable y_j indica el número de anticuerpos, tomando los valores:

$y_1 = 13$, $y_2 = 18$, $y_3 = 20$,

$y_4 = 23$, $y_5 = 30$, $y_6 = 45$.

Tengo además otro método para representar conjuntamente estos datos. Utilizo para ello un sistema cartesiano de ejes:



Los puntos reflejan los pares de valores que están relacionados. Se llama "diagrama de dispersión".

D. JUSTO:

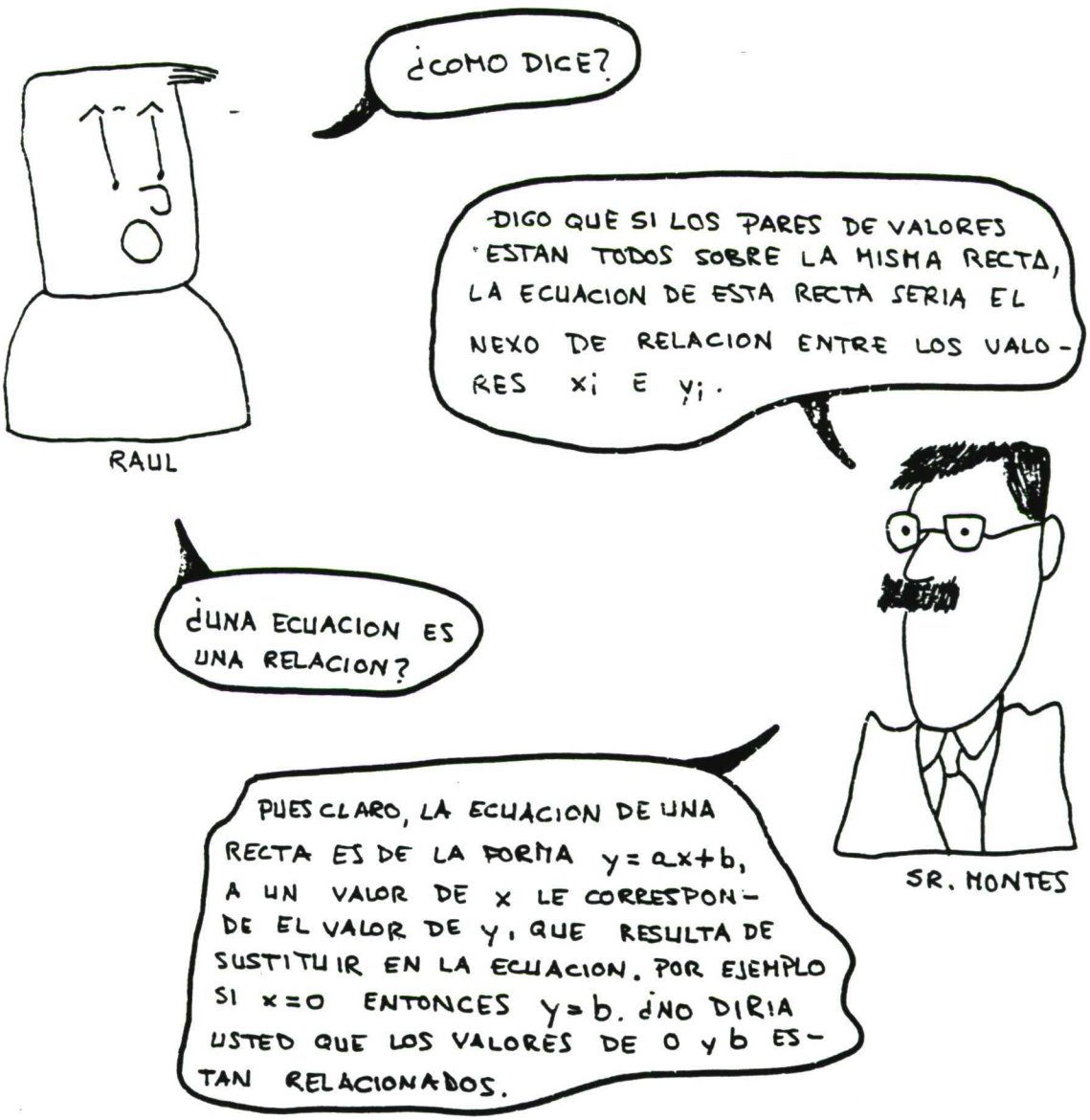
Esto está muy bien, pero lo que no puedo imaginarme es cómo va a conseguir esa famosa relación de la que nos hablaba al principio.

RAUL:

Yo tampoco.

SR. MONTES:

Vamos a ello. Como podeis observar en la nube de puntos, éstos no están alineados. Para que lo entendais, vamos a dividir nuestro problema en dos pasos. Supongamos que los puntos estuviesen alineados. Si así fuera, la relación entre ambas variables estaría clara.



RAUL:

Están ustedes haciéndome un lío. Rectas, ecuaciones, ... Empezamos haciendo unos análisis a unas vacas y ahora hablan de subíndices y rectas. ¿Dónde está esa recta?.

SR. MONTES:

Calma, amigo Raúl. No se impaciente. Ya le he dicho que nuestro problema tiene dos partes. ¿Estamos de acuerdo en que si los pares de valores estuviesen sobre una recta, habríamos encontrado la relación?.

RAUL:

Sí, bueno ... Eso lo entiendo. Pero nuestros puntos no están alineados.

SR. MONTES:

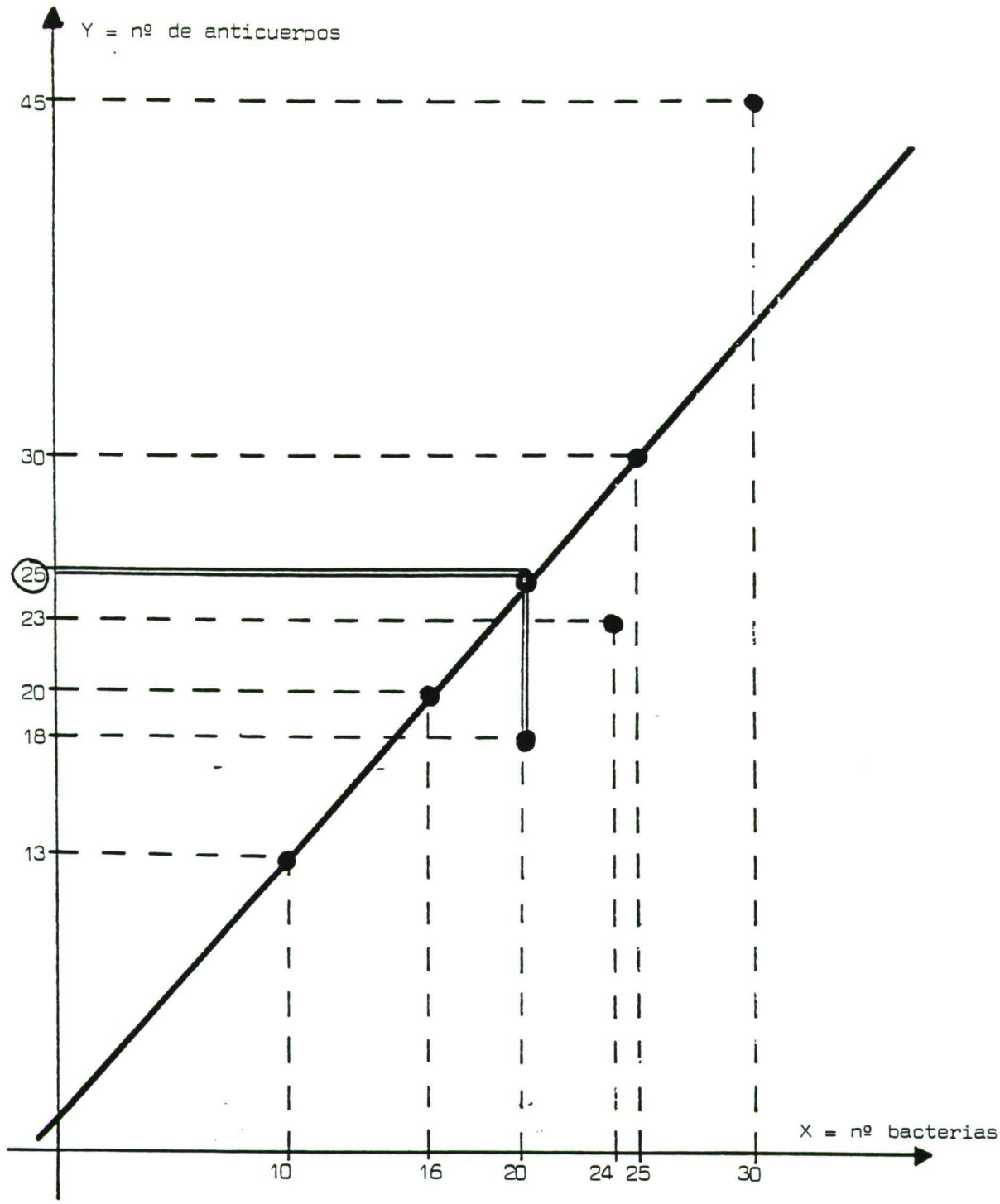
Esa es la segunda parte. Vamos a intentar encontrar una recta que se aproxime a todos estos puntos.

RAUL:

¿Una recta que se aproxime a seis puntos que no están alineados?.

SR. MONTES:

Se se lo dibujo, quizás lo vea usted mejor.



D. JUSTO:

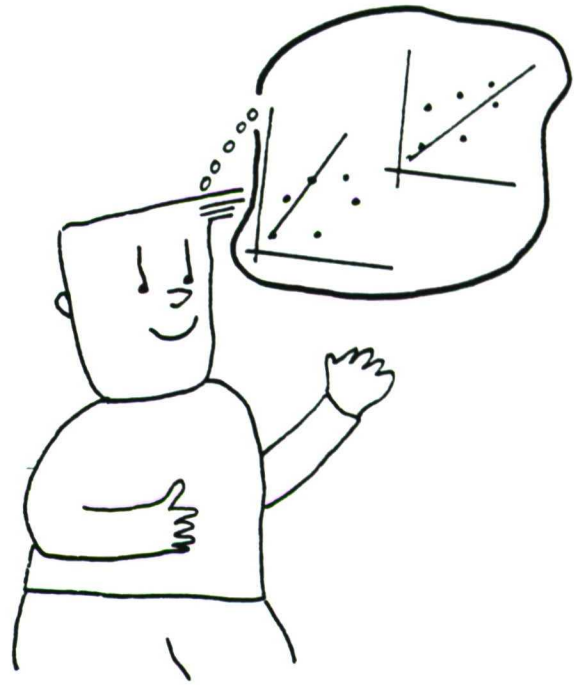
Pero, si esa recta es nuestra relación, ésta es falsa; ya que según eso el valor x_3= 20 le corresponde y_3= 25, cuando el valor real es 18.

SR. MONTES:

Claro, ese es el error que cometemos. Fíjense que desde el principio les he dicho que lo que intentamos es preveer el número de animales enfermos. Preveer y no asegurar, por lo tanto, los resultados no serán seguros.

RAUL:

De acuerdo en eso. Pero ¿ha seguido usted algún método para trazar esa recta?. ¿Podría yo trazar ahora otra recta cualquiera?.



SR. MONTES:

Claro que podría, pero habría que ver quién de los dos comete más error, usted con su recta o yo con la mía.

RAUL:

Pues vaya follón. Si a cada uno se nos ocurre hacer una cosa, no acabaremos nunca.

SR. MONTES:

Para arreglar esto, existe un método generalizado para calcularla. Además, la recta obtenida de esta manera resulta ser aquella con la que se comete menor error, es decir, la recta que más se ajuste a la nube de puntos.

RAUL:

¿Es muy difícil calcularla?.

SR. MONTES:

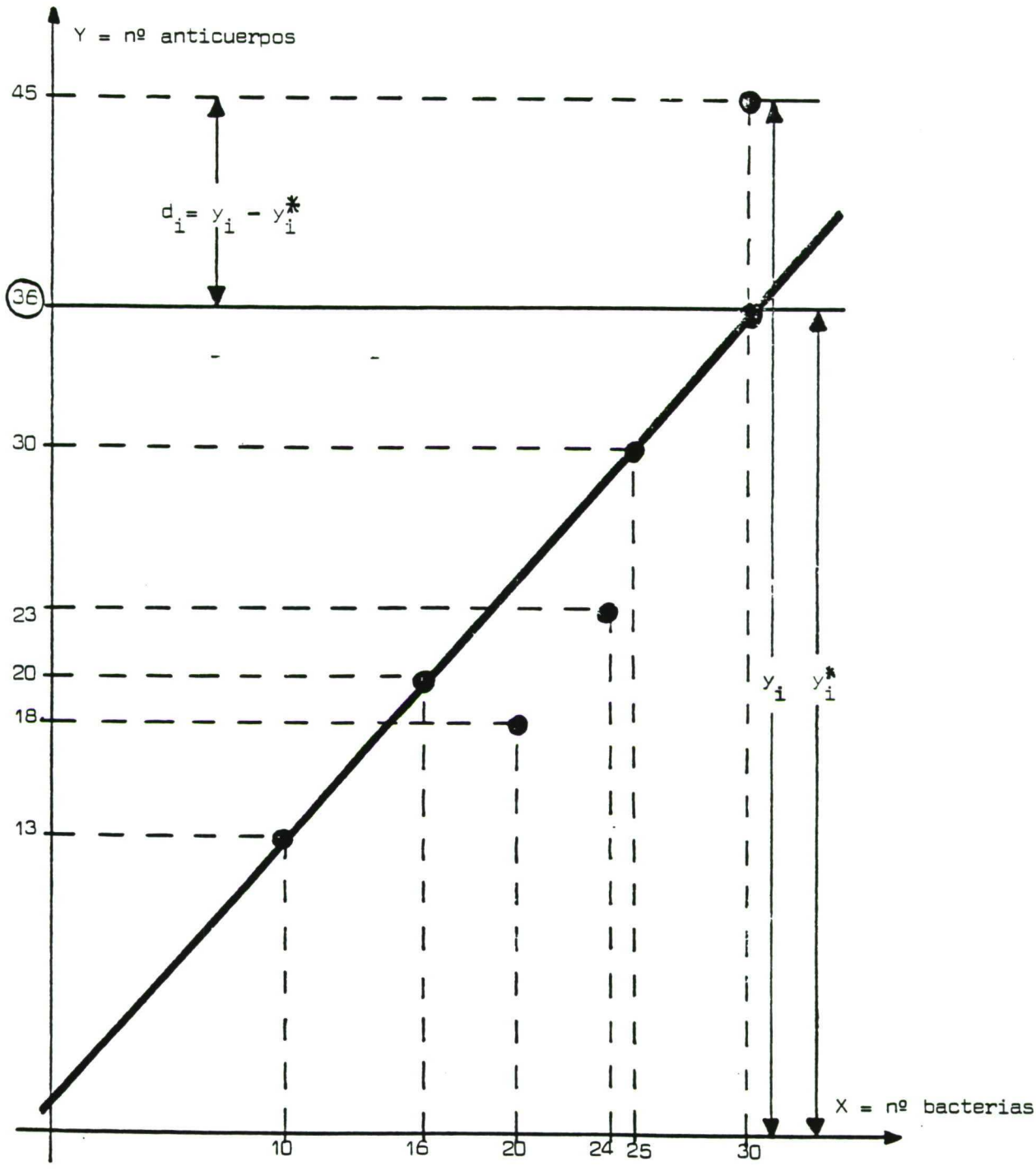
Como parece que les interesa, se lo explicaré. Pero antes pongamos nombres a todos esos conceptos que estamos manejando. Esa recta que construímos se llama "recta de regresión" y el proceso por el que llegamos a ella "regresión", que equivale a predicción, pronóstico o estimación. El método del que les hablaba se llama "ajuste por mínimos cuadrados".

D. JUSTO:

Eso suena muy complicado.

SR. MONTES:

Creo que no lo es. Verán, como bien nos indicó antes D. Justo, tenemos un valor de x_i al que corresponde un valor real y_i y, sin embargo, según la recta le corresponde otro valor que nosotros asignamos teóricamente por y_i^* de - modo que $y_i^* = a.x_i + b$.



El error que cometemos en el ajuste de la recta es:

$$d_i = y_i - y_i^* \qquad \text{positivo}$$

que como pueden apreciar en el dibujo representa la distancia entre el valor observado y el valor teórico.

En otro punto, el error de ajuste de la recta es:

$$d_i = y_i - y_i^* \qquad \text{negativo}$$

RAUL:

Cuanto más pequeño sea este número mejor será la recta que hayamos tomado.

D. JUSTO:

Déjeme adivinar Sr. Montes, si yo consigo que la suma de todas estas distancias, es decir:

$$d_1 + d_2 + d_3 + d_4 + d_5 + d_6$$

sea lo más pequeña posible, la recta que me la proporcionó será la buena.

SR. MONTES:

Casi D. Justo, casi. Pero fíjese que tenemos puntos que quedan encima de la recta y otros bajo ella, esto quiere decir que algunos d_i son positivos y otros negativos.

RAUL:

¿Y eso qué quiere decir?.

SR. MONTES:

Pues que así corremos el peligro de que se anulen los valores positivos con los negativos y entonces la suma no representa la suma real de las distancias.

RAUL:

Vaya y ¿cómo lo arreglamos?.

SR. MONTES:

En vez de utilizar d_i, utilizamos d_i^2, así todos los sumandos son positivos:

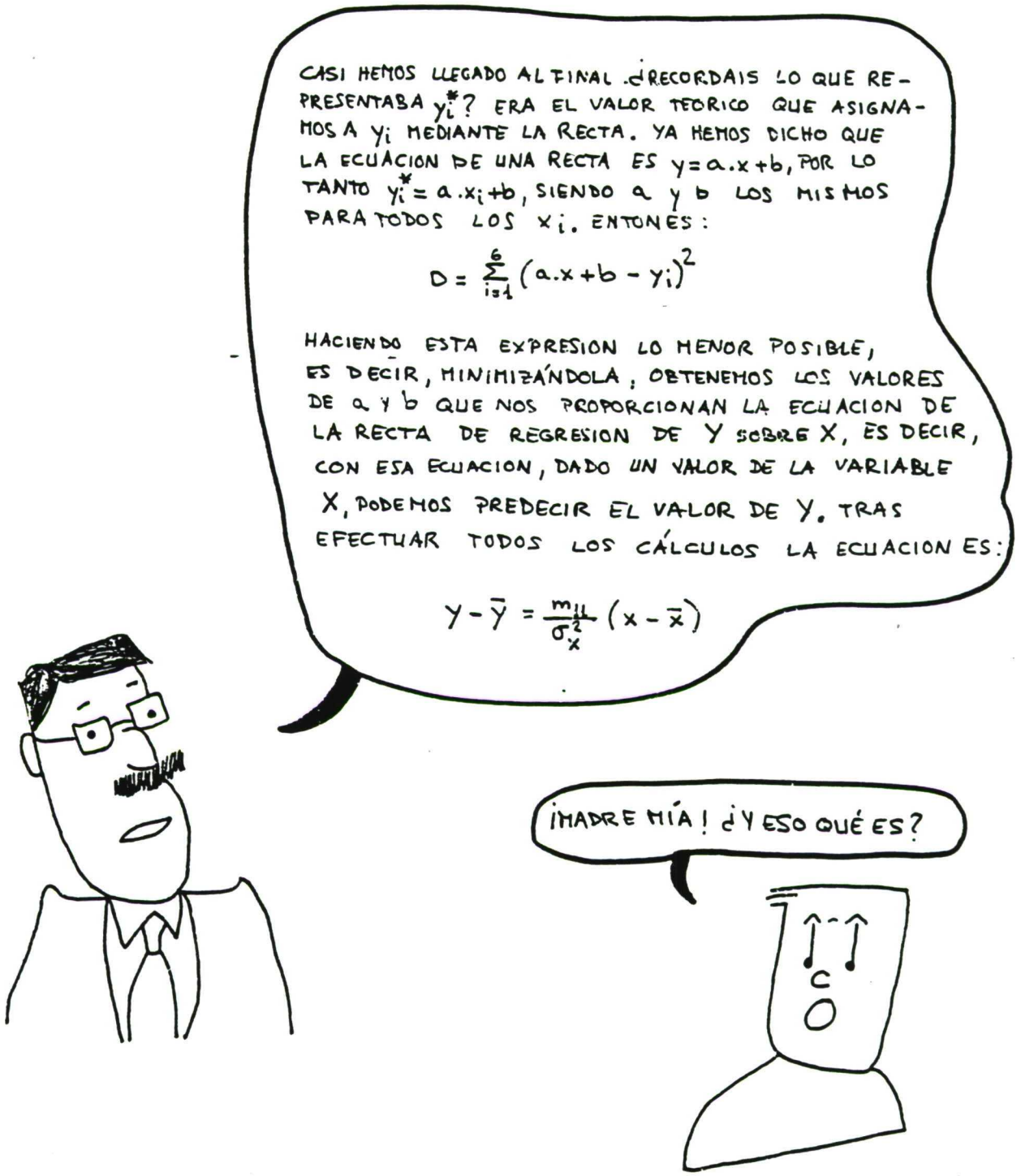
$$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2$$

que podemos escribir utilizando la notación de sumatorios:

$$D = \sum_{i=1}^{6} d_i^2$$

RAUL:

Y ahora ¿qué?.



SR. MONTES:

Eso es. Además para calcular el resto de los coeficientes que aparecen en la ecuación necesitaré realizar una serie de cálculos que ordenaré en una tabla.

x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2
10	13	130	100	169
16	16	288	256	324
20	20	400	400	400
24	23	552	576	529
25	30	750	625	900
30	45	1350	900	2025
SUMA 125	149	3470	2857	4347

Y ahora escuchad, utilizando esta tabla calcularemos sencillamente lo que necesitamos:

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i}{N} = \frac{10 + 16 + 20 + 24 + 25 + 30}{6} = 20'83$$

$$\bar{y} = \frac{\sum_{i=1}^{6} y_i}{N} = \frac{13 + 18 + 20 + 23 + 30 + 45}{6} = 24'83$$

$$m_{11} = a_{11} - a_{10}\, a_{01}$$

LLAMADO COVARIANZA

a_{11}, a_{10}, a_{01} MOMENTOS RESPECTO AL ORIGEN

$$\left.\begin{aligned}
\bar{x} = a_{10} &= \frac{\sum_{i=1}^{6} x_i}{N} = 20'83 \\
\bar{y} = a_{01} &= \frac{\sum_{i=1}^{6} y_i}{N} = 24'83 \\
a_{11} &= \frac{\sum_{i=1}^{6} x_i\, y_i}{N} = \frac{3470}{6} = 578'33 \\
a_{20} &= \frac{\sum_{i=1}^{6} x_i^2}{N} = \frac{2857}{6} = 476'16
\end{aligned}\right\} \quad m_{11} = 578'33 - 517'20 = 61'12$$

$$\sigma_x^2 = a_{20} - a_{10}^2 = 476'16 - 433'88 = 42'27 \quad \text{(varianza de X)}$$

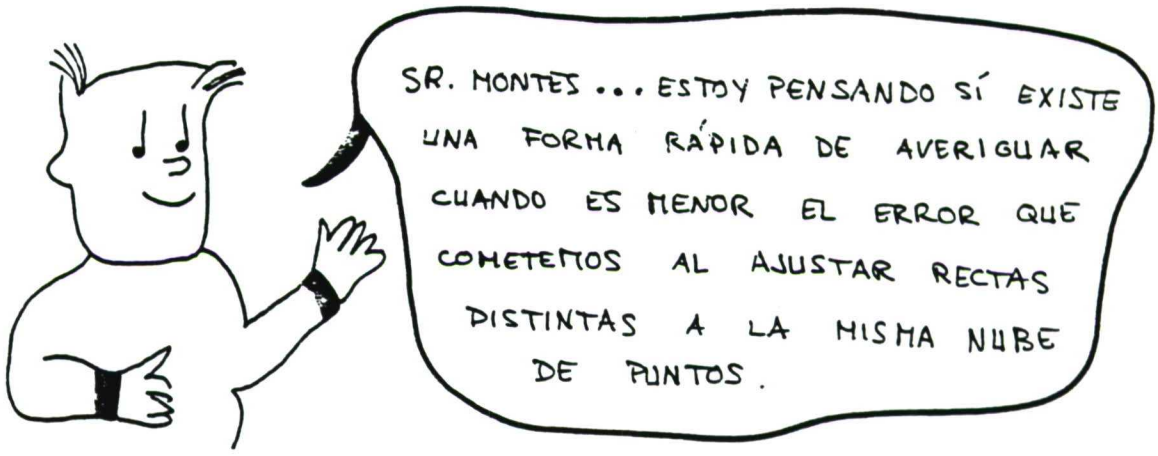
SR. MONTES:

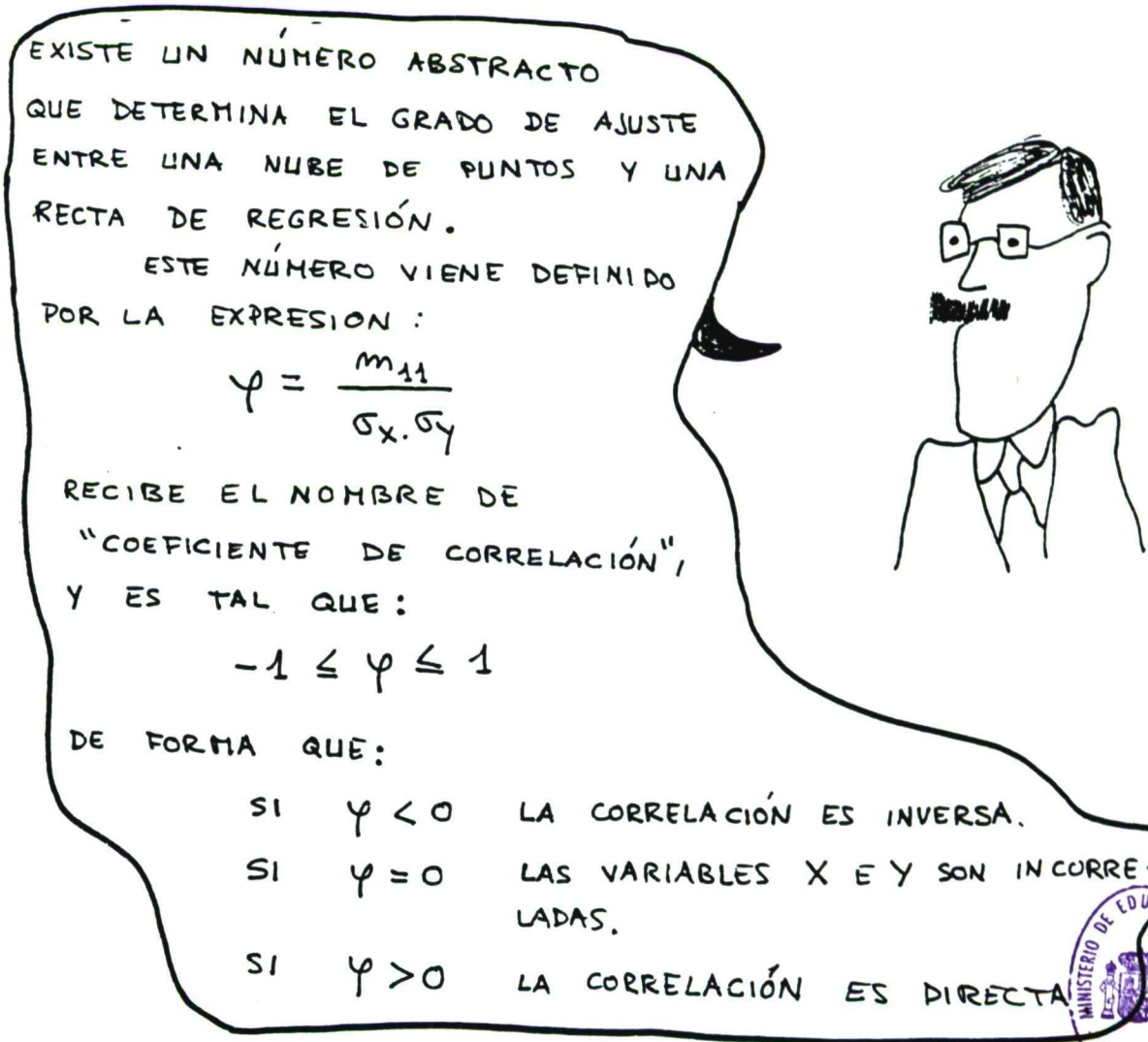
Así, una vez sustituído lo que acabamos de calcular la ecuación de la recta de regresión es:

$$y - 24'83 = \frac{61'12}{42'27}\,(x - 20'83)$$

Ahora ya pueden ustedes empezar a trabajar de nuevo. Realizarán análisis al resto de las reses del pueblo, para obtener los valores x_i, a los que

corresponderán unos valores y_i según nuestra ecuación. Dependiendo de estos valores podremos predecir si el ganado del pueblo está enfermo y nuestro trabajo habrá acabado.





RAUL:

Otra vez está usted haciéndome un lío. Correlación inversa, Correlación directa, Incorreladas, ... ¿Puede explicármelo de otra forma haber si lo entiendo?.

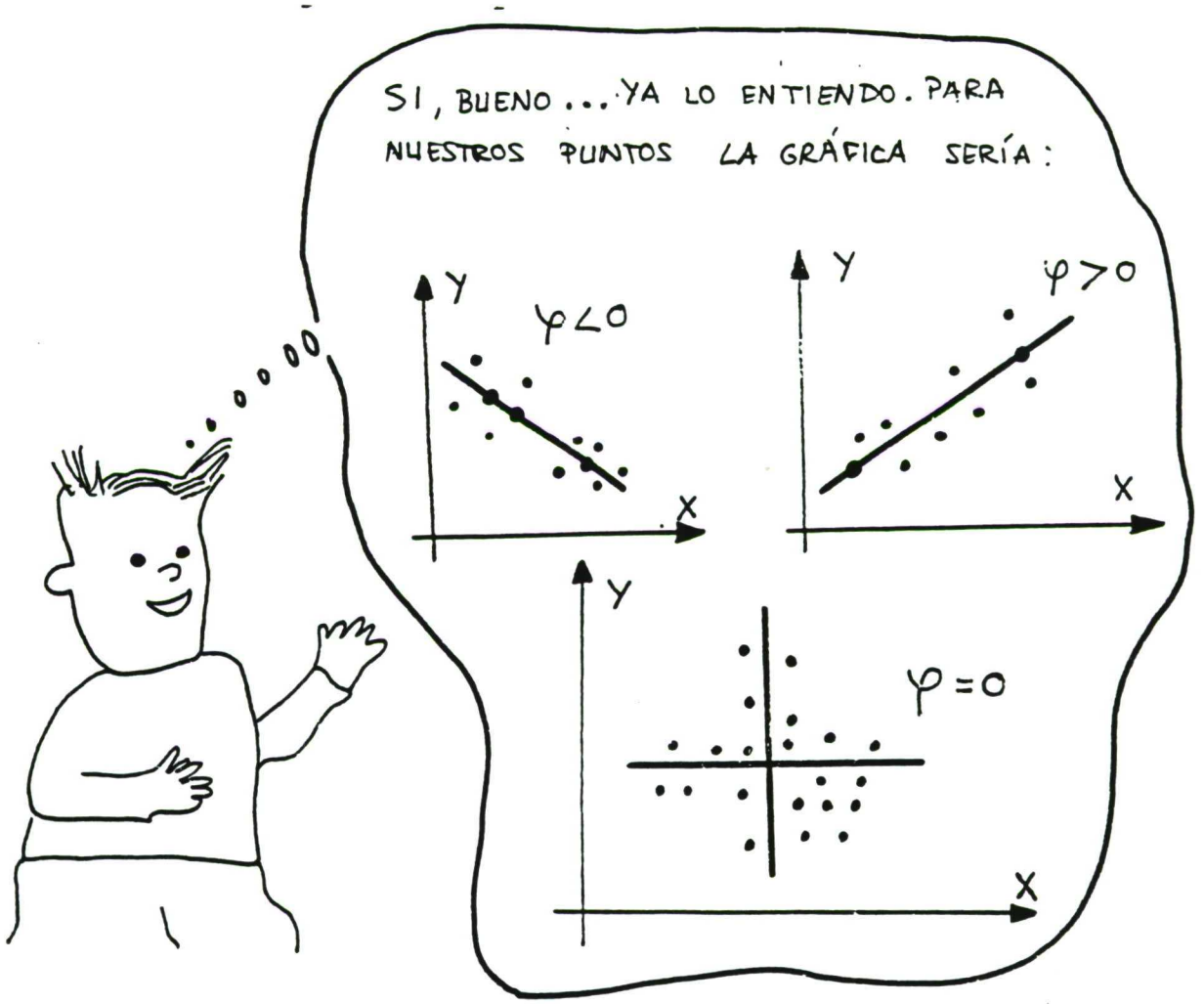
SR. MONTES:

Calma, amigo Raúl. No se impaciente, trataré de explicarme:

"CORRELACION INVERSA": Es cuando a medida que aumentan los valores de la X, disminuyen los valores de la Y. Es decir, cuando el coeficiente de correlación φ es negativo.

"CORRELACION DIRECTA": Cuando a medida que aumentan los valores de la X, aumentan los valores de la Y. Esto es, cuando el coeficiente de correlación φ es positivo.

"VARIABLES INCORRELADAS": Cuando las variables X e Y son independientes, esto es, cuando no existe ningún tipo de correlación. Entonces $\varphi = 0$.



SR. MONTES:

Las gráficas son correctas. Si observa usted Raúl:

- Cuando $\varphi < 0$, se tiene una recta de regresión decreciente.
- Cuando $\varphi > 0$, la recta de regresión es creciente.
- Si $\varphi = 0$, las dos rectas son perpendiculares.

RAUL:

Bueno, bueno, ..., esto no es muy complicado que digamos ... Entonces, el coeficiente de correlación φ de la recta de regresión que habíamos calculado:

$$y - 24'83 = \frac{61'12}{42'27}(x - 20'83)$$

donde:

la covarianza: $m_{11} = a_{11} - a_{10} \cdot a_{01} = a_{11} - \bar{x} \cdot \bar{y} = 61'12$

la varianza de la X: $\sigma_x^2 = 42'27$, por tanto, $\sigma_x = \sqrt{42'27} = 6'50$

la varianza de la Y: $\sigma_y^2 = a_{02} - (a_{01})^2$, es decir:

$$\sigma_y^2 = \frac{4347}{6} - (24'83)^2 = 107'97$$

$$\sigma_y = \sqrt{107'97} = 10'39$$

Vendrá dado por la expresión:

$$\varphi = \frac{m_{11}}{\sigma_x \cdot \sigma_y} = \frac{61'12}{(6'50).(10'39)} = 0'90$$

Esto nos indica que entre las variables X = nº de bacterias e Y = nº de anticuerpos, existe una "correlación directa" con un grado de validez de 0'90, ó del 90%. En otras palabras, a medida que aumenta el número de bacterias - aumenta el número de anticuerpos en un 90%.

Raúl quedó muy impresionado con aquello que les había contado el Sr. Montes y decidió practicar.

RAUL:

D. Justo, ¿usted cree que esto que hemos hecho con las vacas del pueblo es aplicable a los conejos?.

D. JUSTO:

Claro, la bacteria que estábamos estudiando también puede atacar a los conejos. Sólo habría que realizarles los análisis a los correspondientes animalitos.

RAUL:

Pues verá ..., yo tengo una granja de conejos y me gustaría saber si están enfermos.

Raúl y D. Justo realizaron las pruebas obteniendo los siguientes resultados:

Nº BACTERIAS	Nº ANTICUERPOS
10	9
15	10
14	15
9	8

¿Podrías ayudar a Raúl a realizar los cálculos?. Necesita saber urgentemente si los conejos están enfermos, su mujer está a punto de llevarlos al mercado.

RECUERDA

• VARIABLE ESTADISTICA BIDIMENSIONAL

Se considerarán aquellas situaciones donde se realiza la observación simultánea de dos caracteres en el individuo, sea:

- peso y altura de una persona
- consumo de tabaco y el cáncer de pulmón
- aptitud para una asignatura y aprovechamiento en la misma
- provincia de origen y carrera estudiada.

Cada uno de estos caracteres en observación puede adoptar distintas modalidades, así:

carácter X adopta las modalidades $x_1, x_2, x_3, \ldots, x_i, \ldots, x_k$

carácter Y adopta las modalidades $y_1, y_2, \ldots, y_j, \ldots, y_l$

Los distintos valores de las modalidades que pueden adoptar estos caracteres forman un conjunto de pares, que representamos por (X,Y), y llamaremos "variable estadística bidimensional".

• DIAGRAMA DE DISPERSION

Dado que estamos estudiando conjuntamente dos caracteres en el individuo, la representación gráfica de los distintos valores de la variable estadística bidimensional (X,Y) se construye del siguiente modo:

En el eje de abscisas se representan los valores de la X.

En el eje de ordenadas se representan los valores de la Y.

De esta forma, obtenemos una nube de puntos que llamaremos "diagrama de dispersión":

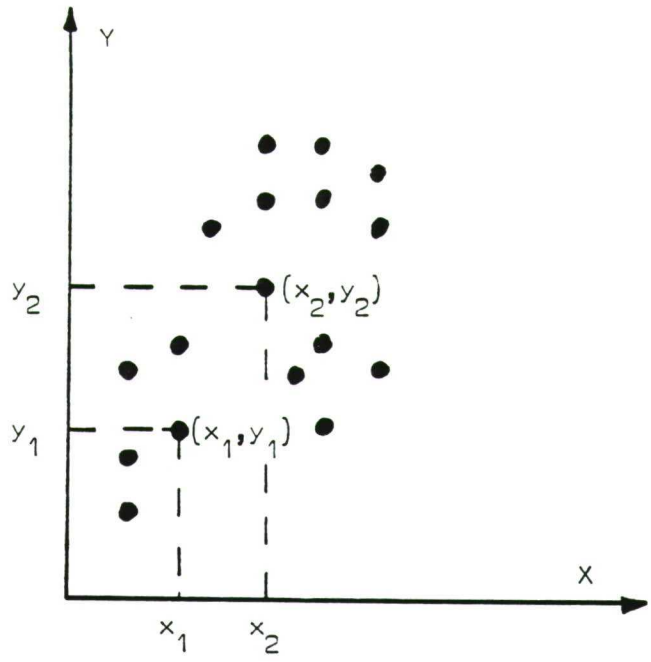


DIAGRAMA DE DISPERSION

• DEPENDENCIA ENTRE LAS VARIABLES X E Y

Como estamos observando dos caracteres en cada individuo, se nos presenta el problema de determinar si existe o no existe dependencia entre ellos. En esta clase de análisis podemos distinguir dos tipos de dependencia:

- DEPENDENCIA FUNCIONAL: Cuando existe una ecuación matemática que relaciona a las variables X e Y, sea:

  X = radios

  Y = longitudes de circunferencias

  observamos:

  $$Y = 2\pi X$$

  es decir, conociendo el valor de la X, se puede conocer con exactitud el valor de la Y.

- DEPENDENCIA ALEATORIA: Cuando no existe una ecuación matemática que relacione a las variables X e Y, sea:

  X = peso

  Y = altura

observa que:

> no podemos encontrar ninguna ecuación matemática que nos dé la altura exacta Y que tiene una persona de X kilos de peso.

En otras palabras, podemos decir que existen variables entre las que no existe ningún tipo de dependencia.

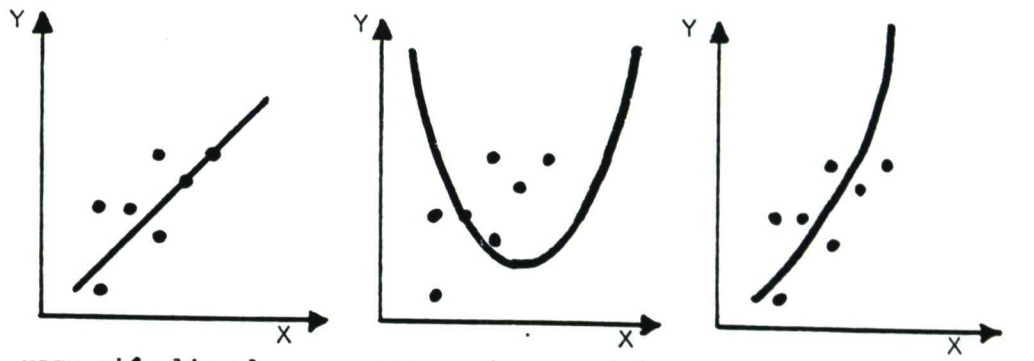
REGRESION

La observación de una variable estadística bidimensional (X,Y) lleva consigo la representación de los puntos obtenidos en una nube o diagrama de dispersión.

> "Estudiaremos bajo el título de "regresión" los problemas referentes a la predicción o pronóstico de los resultados de una de las dos variables, conocidos los resultados en la otra".

De tal forma, el problema general que se plantea en "regresión" consiste en ajustar una función de ecuación conocida a la nube de puntos con el objetivo de poder obtener una "predicción" aproximada de una de las variables a partir de la otra.

La función que pretendemos obtener será aquella cuya ecuación se adopte mejor a la nube de puntos. Por ejemplo:



regresión lineal $y = a + bx$

regresión parabólica $y = a + bx + cx^2$

regresión exponencial $y = k\, a^{bx}$

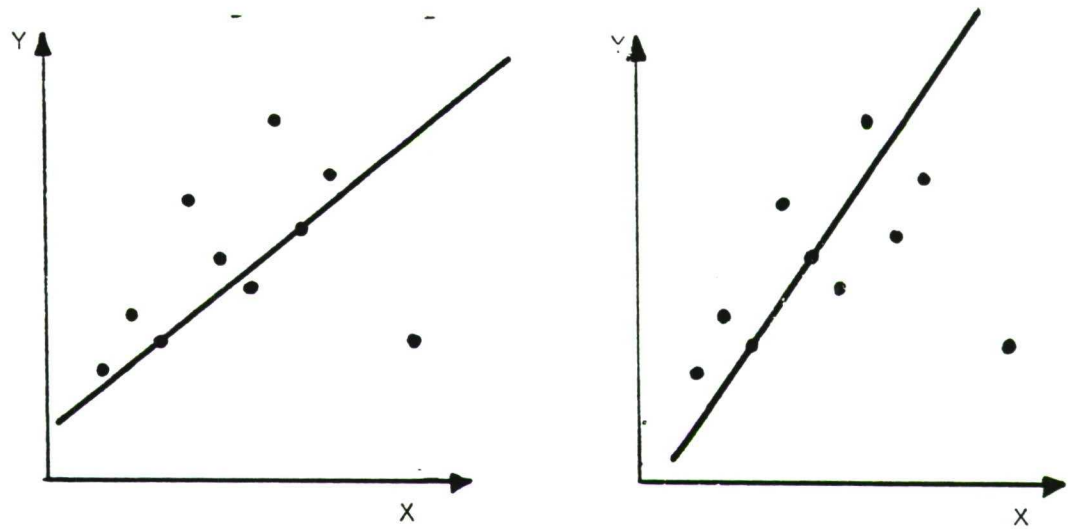
Así, cuando es una recta, tenemos la regresión lineal; cuando es una parábola, tenemos la regresión parabólica; cuando es una función exponencial, tenemos la regresión exponencial, etc.

En este capítulo, estudiaremos únicamente la regresión lineal, es decir, la función que pretendemos obtener será una recta de ecuación $y = a + bx$.

• REGRESION LINEAL

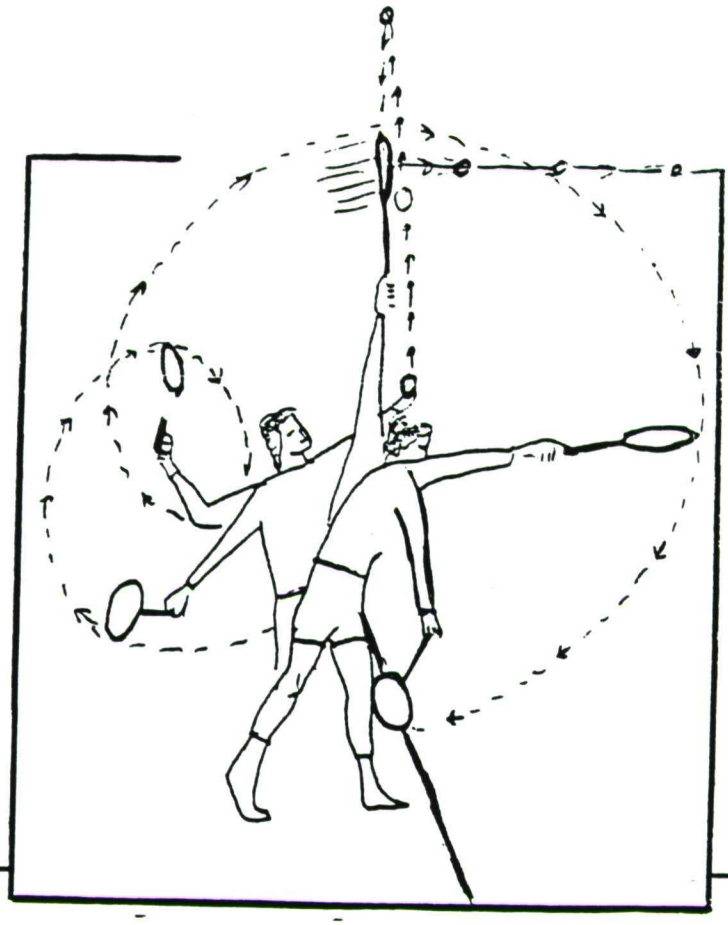
La ecuación de regresión será la ecuación de la recta $y = a + bx$ que mejor se adapte a todos los puntos del diagrama de dispersión, y que nos permitirá pronosticar el valor de una variable en función de la otra con la que está relacionada.

Gráficamente:



regresión lineal

Naturalmente que entre todas las rectas que se pueden elegir para ajustarlas a la nube de puntos hemos de seleccionar la óptima, esto es, la que mejor se encaje sobre los puntos que tenemos. Esta selección la realizamos por el "método de los mínimos cuadrados".



Ecuación de la recta

La ecuación de la recta viene dada por la fórmula:

$$y = a + bx$$

donde

b = pendiente de la recta

a = punto donde la recta corta al eje Y, se llama "ordenada en el origen".

Sea la recta de ecuación $y = 2 + 3x$

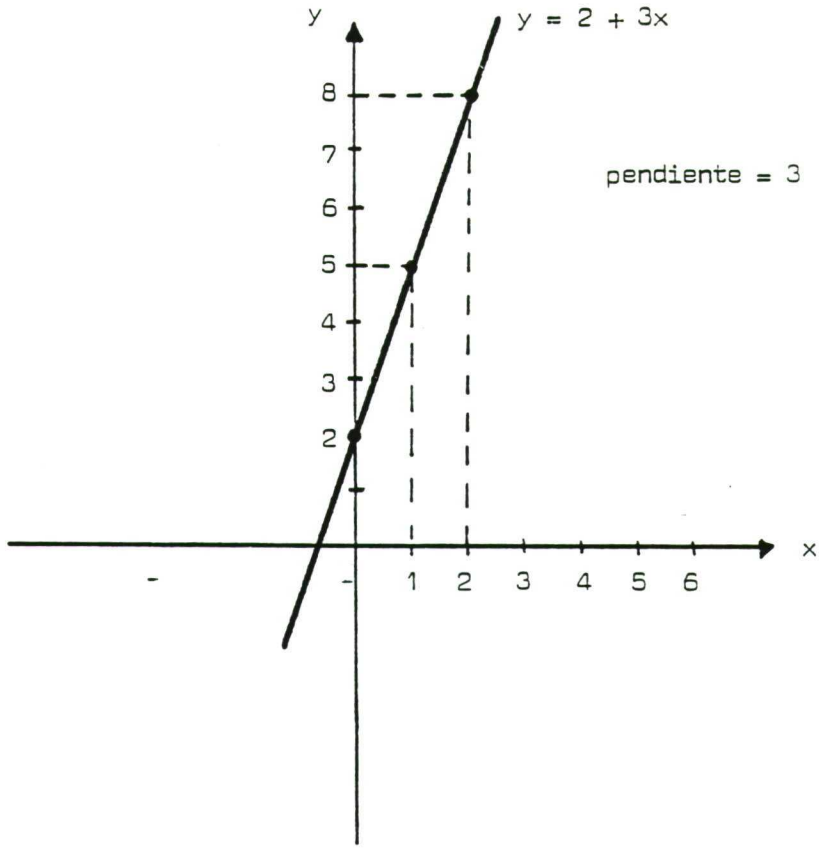
En esta ecuación los valores de la variable Y dependen de los valores que tome la variable X. veamos:

para	$x = 0$	$y = 2 + 3.0 = 2$	$y = 2$
"	$x = 1$	$y = 2 + 3.1 = 5$	$y = 5$
"	$x = 2$	$y = 2 + 3.2 = 8$	$y = 8$

Piensa que una recta queda definida por dos puntos, por tanto, representando

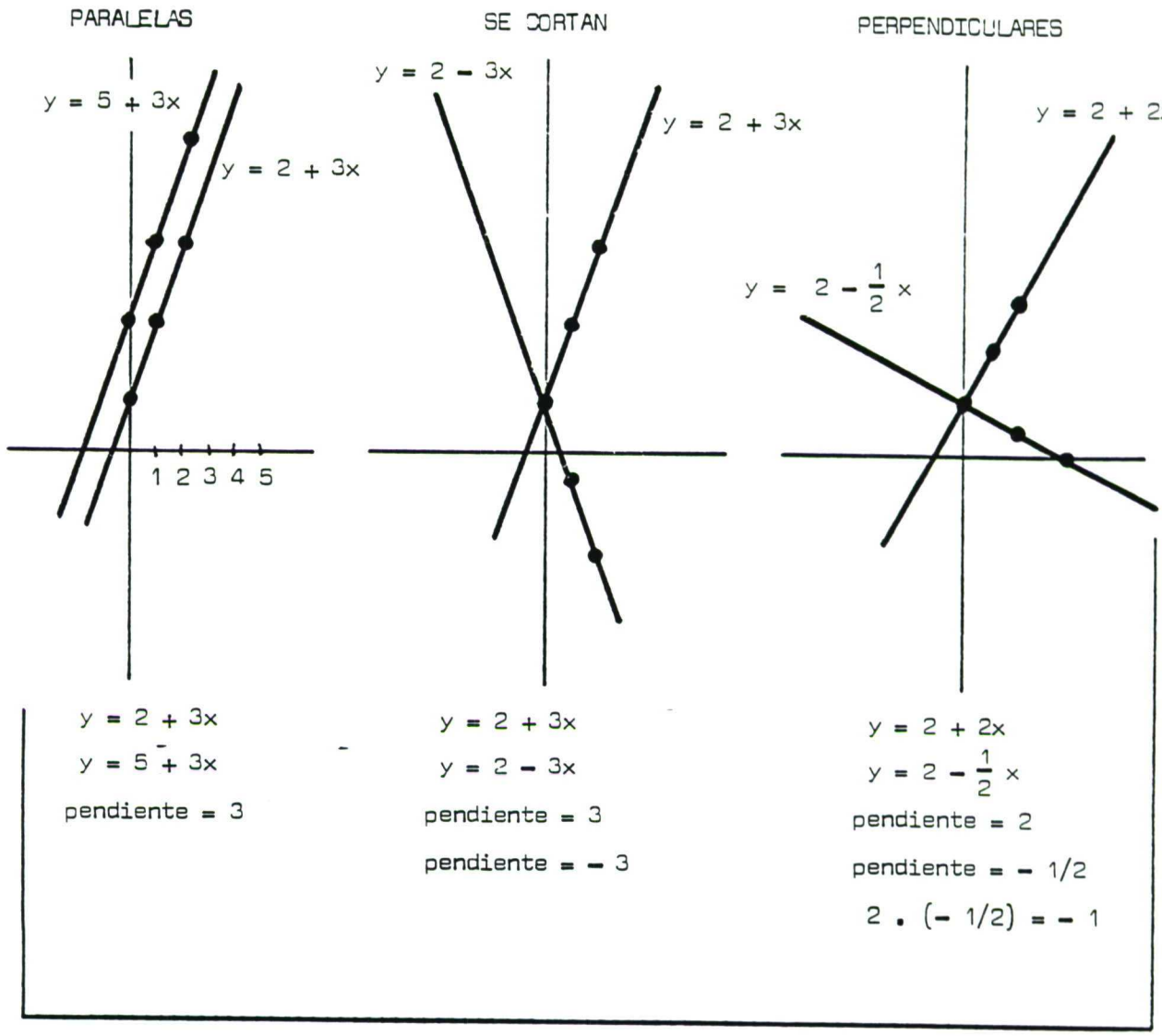
dos puntos cualesquiera en un eje de coordenadas obtenemos la recta de ecuación:

$$y = 2 + 3x$$



Observa que cuando dos rectas:

- tienen la misma pendiente son paralelas
- tienen la misma pendiente pero de distinto signo se cortan
- tienen el producto de sus pendientes (- 1) son perpendiculares.



$y = 2 + 3x$
$y = 5 + 3x$
pendiente = 3

$y = 2 + 3x$
$y = 2 - 3x$
pendiente = 3
pendiente = − 3

$y = 2 + 2x$
$y = 2 - \frac{1}{2} x$
pendiente = 2
pendiente = − 1/2
$2 \cdot (-1/2) = -1$

- METODO DE LOS MINIMOS CUADRADOS: REGRESION LINEAL

Ante un problema de regresión lineal distinguiremos entre:

a) Recta de regresión de Y sobre X.

b) Recta de regresión de X sobre Y.

En el primer caso obtendremos los valores aproximados de la variable Y conocidos los valores de la variable X. Mientras que, en el segundo caso, obtendre-

mos los valores aproximados de la variable X conocidos los valores de la variable Y.

Estudiaremos con más detalle el primer caso y, en el segundo, por un proceso exactamente igual estableceremos la ecuación de la recta de regresión de X sobre Y.

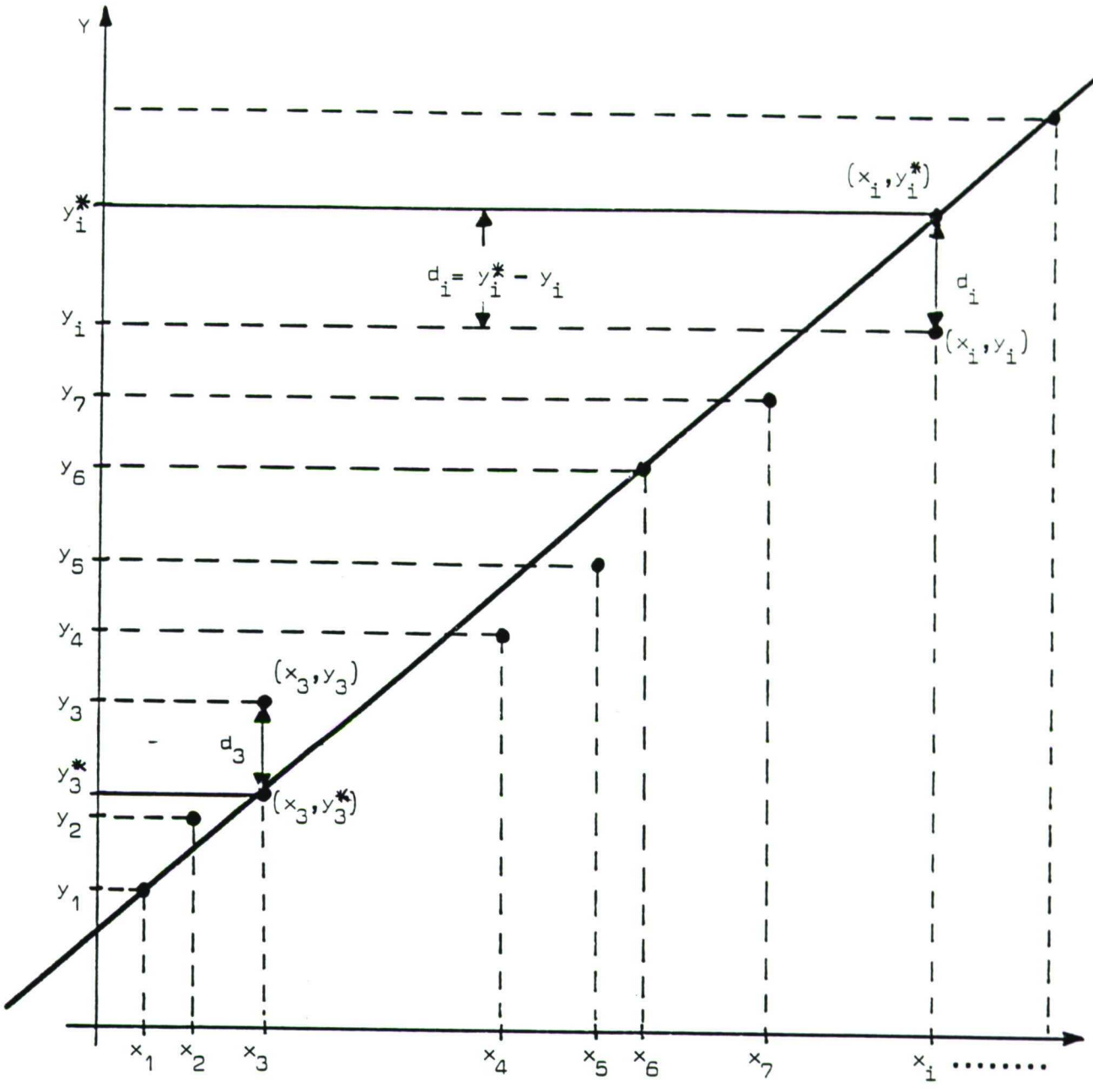
a) RECTA DE REGRESION DE Y SOBRE X

La ecuación de la recta es $y = a + bx$.

El problema que se plantea es intentar ajustar la recta $y = a + bx$ a la nube de puntos que nos preocupa. Para seleccionar la recta óptima, lo que hemos de hacer es estimar los parámetros a y b con los datos observados y empleando el "método de los mínimos cuadrados", cuyo contenido explicamos a continuación, llegar a la ecuación de la recta de regresión de Y sobre X.

Para la exposición, nos basaremos en la tabla siguiente:

X	x_1	x_2	x_3	...	x_i	...	x_k
Y	y_1	y_2	y_3	...	y_i	...	y_l



Observa que los puntos representados tienden a estar en línea recta. Sin embargo, es imposible encontrar una línea recta que pase por todos los puntos simultáneamente. La solución a este problema está en encontrar una línea recta que más se aproxime a estos puntos.

Podemos considerar en cada par (X,Y) que al valor observado x_i le corresponde un valor observado y_i y otro valor teórico y_i^* que sería el que le correspondería en la recta.

En otras palabras, el par (x_i, y_i^*) está en la recta cuando verifica:

$$y_i^* = a + bx_i$$

La distancia entre estos dos valores, teórico y observado, la llamaremos - "error de predicción" y se denota por d_i, siendo:

$$d_i = y_i^* - y_i$$

Observa que los errores de predicción $(d_1, d_2, d_3, \ldots, d_i, \ldots)$ pueden ser de signo positivo y negativo y, por tanto, hacer mínima su suma, esto es:

$$d_1 + d_2 + d_3 + \ldots + d_i + \ldots + \ldots = \text{mínima}$$

Pues bien, el "método de los mínimos cuadrados", para la obtención de los - parámetros a y b, consiste en tomar estas distancias al cuadrado para que - no se puedan contrarrestar los signos positivos y negativos, y hacer mínima su suma.

Tenemos, por tanto, que hacer mínima la expresión:

$$D = \sum_i d_i^2 = \sum_i (y_i^* - y_i)^2$$

esto es

$$D = \sum_i (a + bx_i - y_i)^2$$

De la ecuación de la recta $y = a + bx$. Por las propiedades de la derivada - (para minimizar D) y del sumatorio, nos queda:

$$a = \bar{y} - b.\bar{x}$$

$$b = \frac{m_{11}}{\sigma_x^2}$$

donde:

$$\bar{y} = \text{media de } Y = \frac{\sum_{j=1}^{l} y_j \cdot n_{y_j}}{N} = a_{01}$$

$$\bar{x} = \text{media de } X = \frac{\sum_{i=1}^{k} x_i \cdot n_{x_i}}{N} = a_{10}$$

$$\sigma_x^2 = \text{varianza de } X = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_{x_i}}{N} = a_{20} - (a_{10})^2$$

$$m_{11} = \text{covarianza} = a_{11} - a_{10} \cdot a_{01}$$

$$a_{11} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{l} (x_i - \bar{x}).(y_j - \bar{y}). n_{i_j}}{N}$$

$$a_{20} = \frac{\sum_{i=1}^{k} x_i^2 \cdot n_{x_i}}{N}$$ segundo momento respecto al origen

N = número de pares de datos o bien el número de puntos del diagrama de dispersión.

Como la ecuación de la recta era $y = a + bx$, sustituyendo los valores obtenidos, se tiene:

Ecuación de la recta de regresión de Y sobre X:

$$y - \bar{y} = \frac{m_{11}}{\sigma_x^2} (x - \bar{x})$$

Hay que puntualizar que el cociente (m_{11}/σ_x^2) es la pendiente de la recta de Y sobre X, y se denota:

$b_{yx} = \dfrac{m_{11}}{\sigma_x^2}$ pendiente de la recta, que recibe el nombre de

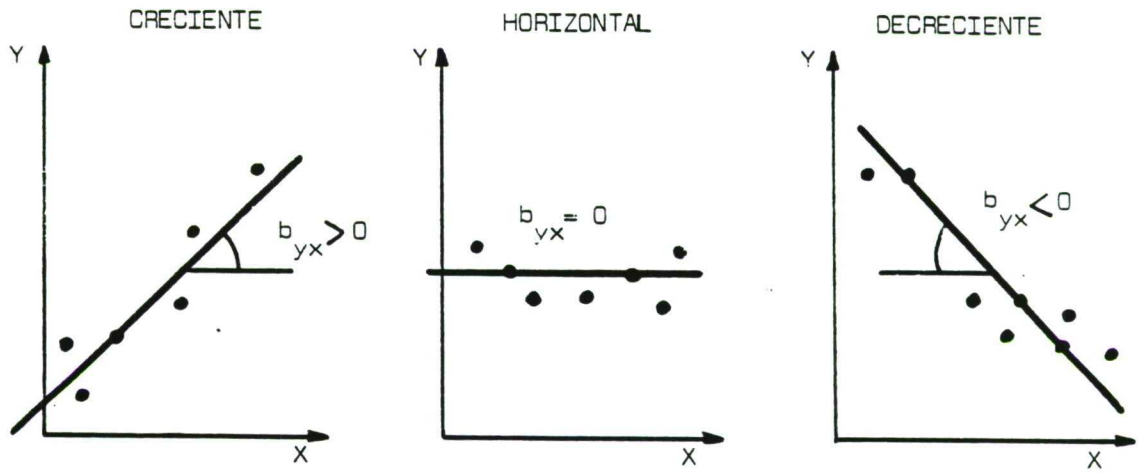
COEFICIENTE DE REGRESION .

Observa que cuando el coeficiente de regresión:

b_{yx} $\begin{cases} > 0 & \text{Recta de regresión de Y sobre X creciente} \\ = 0 & \text{" " " " " " " horizontal} \\ < 0 & \text{" " " " " " " decreciente} \end{cases}$

Gráficamente:

RECTA DE REGRESION DE Y SOBRE X



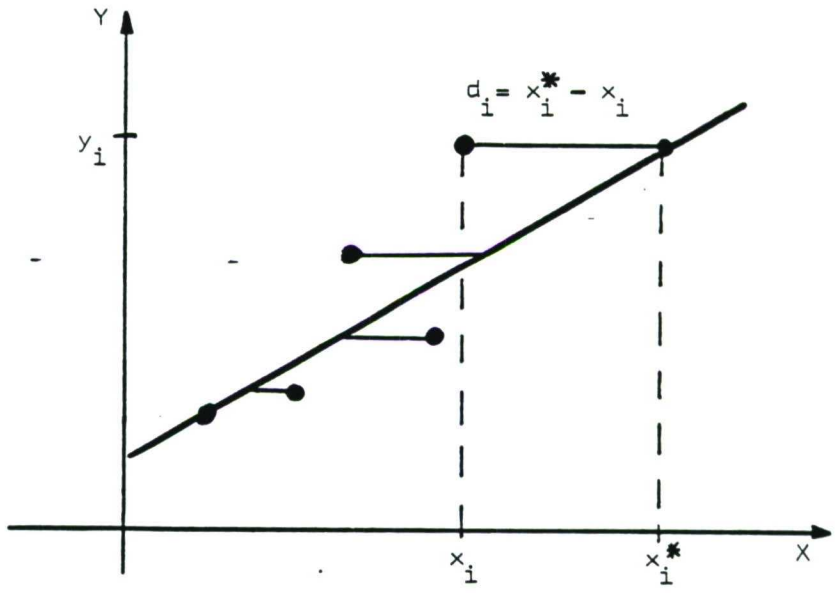
coeficiente de regresión $b_{yx} = \dfrac{m_{11}}{\sigma_x^2}$

b) RECTA DE REGRESION DE X SOBRE Y

Recta que nos permite hacer predicciones de la variable X conocidos los valores de la variable Y.

La ecuación de la recta es $x = c + dy$

Si en lugar de tomar las distancias d_i (error de predicción) sobre el eje de la Y, y se toman sobre el eje de la X (horizontales) y utilizamos el método de los mínimos cuadrados, se tiene:



En otras palabras, el par (x_i^*, y_i) se encuentra en la recta cuando verifica:

$$x_i^* = c + dy_i$$

La distancia entre el valor observado x_i y el valor teórico x_i^* se denomina "error de predicción", siendo:

$$d_i = x_i^* - x_i$$

Por el método de los mínimos cuadrados, hemos de hacer mínima la expresión:

$$D = \sum_i d_i^2 = \sum_i (c + dy_i - x_i)^2$$

por un proceso exactamente igual, se llega a la ecuación de la recta de regresión de X sobre Y:

$$\boxed{x - \bar{x} = \frac{m_{11}}{\sigma_y^2}(y - \bar{y})}$$

El cociente (m_{11}/σ_y^2) es la pendiente de la recta, que se denomina "COEFICIENTE DE REGRESION" y denotamos por b_{xy}:

$$b_{xy} = \frac{m_{11}}{\sigma_y^2}$$

Observa:

$$b_{xy} \begin{cases} > 0 & \text{Recta de regresión de X sobre Y creciente} \\ = 0 & \text{" " " " " " " horizontal} \\ < 0 & \text{" " " " " " " decreciente} \end{cases}$$

CORRELACION LINEAL

La correlación estudia el grado de dependencia que existe entre dos variables estadísticas, intentando cuantificarla mediante el cálculo de los coeficientes de correlación.

Nos limitaremos a estudiar el coeficiente de correlación lineal.

• COEFICIENTE DE CORRELACION LINEAL

El coeficiente de correlación lineal φ es un número abstracto que determina el grado de ajuste entre una nube de puntos y una recta de regresión.

Intuitivamente:

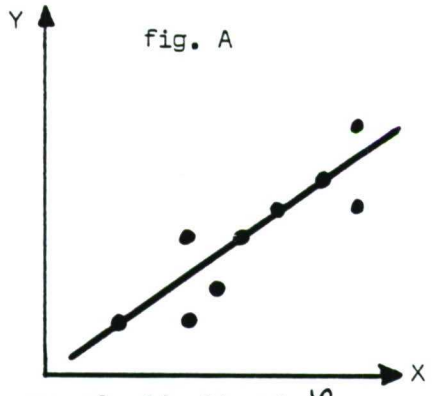


fig. A

correlación lineal φ es más alta

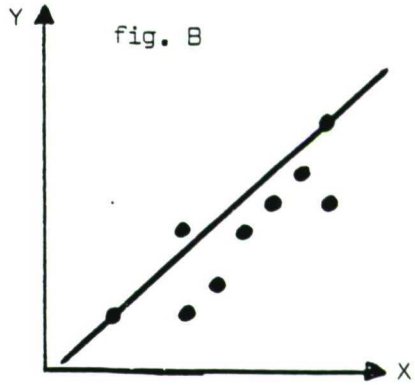


fig. B

correlación lineal φ es más baja

Es decir, en la figura A el grado de ajuste es más alto que en la figura B.

El coeficiente de correlación lineal φ viene definido por la media geométrica de los coeficientes de regresión lineal, esto es:

$$\varphi = \sqrt{b_{yx} \cdot b_{xy}} = \sqrt{\frac{m_{11}}{\sigma_x^2} \cdot \frac{m_{11}}{\sigma_y^2}} = \frac{m_{11}}{\sigma_x \cdot \sigma_y}$$

donde:

σ_x = desviación típica de la variable X

σ_y = desviación típica de la variable Y

m_{11} = covarianza = $a_{11} - a_{10} \cdot a_{01}$

De la expresión anterior:

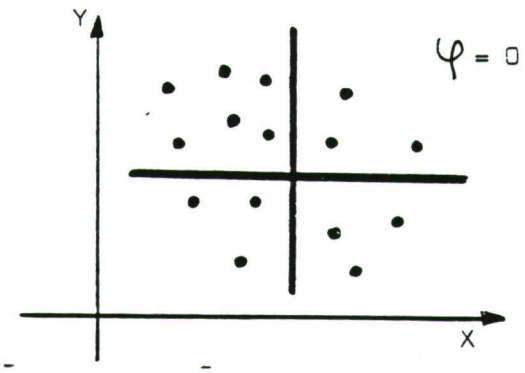
$$\varphi = \frac{m_{11}}{\sigma_x \cdot \sigma_y}$$

se deduce que:

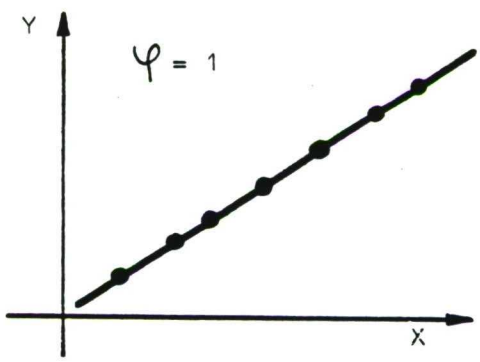
$$-1 \leq \varphi \leq +1$$

es decir, el coeficiente de correlación lineal φ se encuentra acotado entre los valores - 1 y + 1.

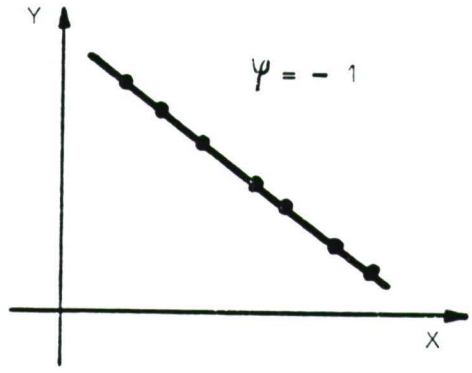
- Un valor de $\varphi = 0$ indica "ausencia de correlación", es decir, que las dos variables son independientes o "incorreladas". Gráficamente:



- Un valor de $\varphi = +1$ lo que nos dice es que todos los puntos de la nube están situados sobre la recta de regresión, existiendo, por tanto, entre las dos variables una "dependencia funcional". Adviértase que cuando $\varphi = 1$, la recta de regresión es creciente. Gráficamente:

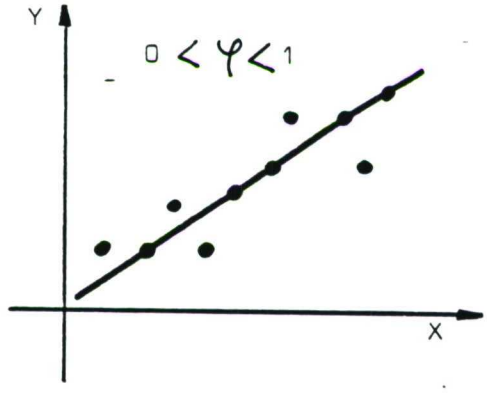


- Un valor de $\varphi = -1$ indica que todos los puntos de la nube se encuentran sobre la recta de regresión y existe entre las dos variables "dependencia funcional" (recta de regresión decreciente). Gráficamente:



- Cuando se tiene un valor grnade de φ (bien sea positivo o negativo) indica que existe una fuerte dependencia entre las dos variables:

- Si $\varphi > 0$, las variables están tanto más "correladas" en cuanto el coeficiente φ se aproxime más a 1. Existe una "dependencia aleatoria" entre las dos variables.

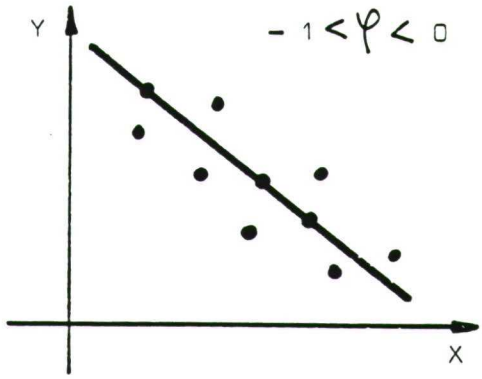


En efecto, si $\varphi > 0$ la "correlacion lineal es directa" y se interpreta:

Los valores grandes de la variable X se asocian con los valores grandes de la variable Y.

Los valores bajos de la variable X se asocian con los valores bajos de la variable Y.

- Si $\varphi < 0$, las variables están tanto más "correladas" en cuanto el coeficiente φ se aproxime más a - 1. Existe una "dependencia aleatoria" entre las dos variables.



En efecto, si $\varphi < 0$ la "correlación lineal es inversa" y se interpreta:

Los valores grandes de la variable X se asocian con valores bajos de la variable Y.

Los valores bajos de la variable X se asocian con valores grandes de la variable Y.

- INTERPRETACION DEL COEFICIENTE DE CORRELACION LINEAL

El coeficiente de correlación lineal φ indica únicamente que dos variables estadísticas (X,Y) varían conjuntamente, y que esta variación conjunta no tiene que tener el significado de "casualidad". De tal modo, sabemos que si $\varphi = \pm 1$ indica que las dos variables dependen de la ecuación de una recta (recta de regresión).

Esto supuesto:

¿qué significa que encontrásemos una correlación $\varphi = 0'65$?. Siendo las variables estadísticas:

X = capacidad intelectual

Y = éxito escolar

Nos preguntamos:

¿Es una correlación alta, media o baja?.

Esta pregunta no tiene sentido considerada considerada absolutamente. No podemos afirmar que la capacidad intelectual sea la causa del éxito o del fracaso en el éxito escolar. Es claro que entran otros factores: disponibilidad de tiempo de estudio, el grado de interés, etc.

Por tanto:

> "No existen normas para indicar si un coeficiente de correlación es alto, medio o bajo. En general, el único criterio que podemos seguir es compararlo con los coeficientes de correlación encontrados por estudios análogos entre las mismas variables y en circunstancias semejantes".

De esta manera:

> Si estudios anteriores entre las mismas variables (X,Y) y en circunstancias semejantes reflejan una correlación $\varphi = 0'80$. Encontrar una correlación $\varphi = 0'65$ resulta una correlación baja, ya que $0'65 < 0'80$.

- RELACION ENTRE LOS COEFICIENTES DE CORRELACION Y REGRESION

a) Relación entre el coeficiente de regresión de Y sobre X y el coeficiente de correlación:

$$\left.\begin{array}{l} b_{yx} = \dfrac{m_{11}}{\sigma_x^2} \\ \varphi = \dfrac{m_{11}}{\sigma_x \cdot \sigma_y} \end{array}\right\} \text{de donde} \quad \begin{array}{l} m_{11} = \sigma_x^2 \cdot b_{yx} \\ m_{11} = \varphi \cdot \sigma_x \cdot \sigma_y \end{array}$$

Igualando se tiene:

$$\sigma_x^2 \cdot b_{yx} = \varphi \cdot \sigma_x \cdot \sigma_y$$

dividiendo por σ_x en la expresión, queda:

$$\sigma_x \cdot b_{yx} = \varphi \cdot \sigma_y$$

entonces

$$b_{yx} = \varphi \cdot \frac{\sigma_y}{\sigma_x} \quad (\text{siendo } \sigma_y \geq 0,\ \sigma_x \geq 0)$$

b) Relación entre el coeficiente de regresión de X sobre Y y el coeficiente de correlación:

$$\left.\begin{array}{l} b_{xy} = \dfrac{m_{11}}{\sigma_y^2} \\ \varphi = \dfrac{m_{11}}{\sigma_x \cdot \sigma_y} \end{array}\right\} \Longrightarrow \begin{array}{l} m_{11} = \sigma_y^2 \cdot b_{xy} \\ m_{11} = \varphi \cdot \sigma_x \cdot \sigma_y \end{array}$$

de donde:

$$\sigma_y^2 \cdot b_{xy} = \varphi \cdot \sigma_x \cdot \sigma_y$$

simplificando, queda:

$$\sigma_y \cdot b_{xy} = \varphi \cdot \sigma_x$$

por tanto:

$$b_{xy} = \varphi \cdot \frac{\sigma_x}{\sigma_y} \quad (\text{siendo } \sigma_x \geq 0,\ \sigma_y \geq 0)$$

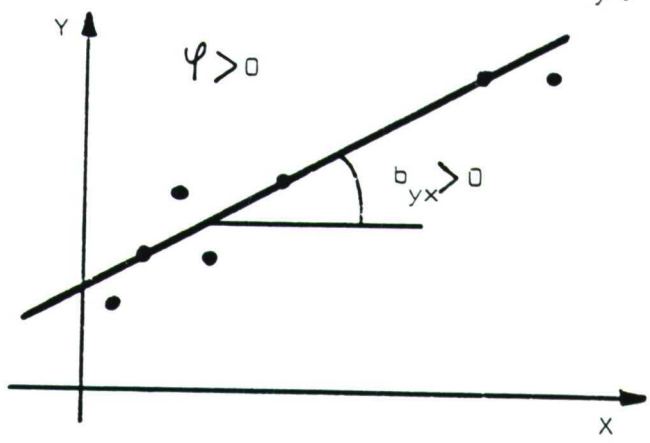
Teniendo en cuenta estas dos relaciones que ligan los coeficientes de regresión y el de correlación, pasamos a estudiar la correlación lineal directa e inversa.

- CORRELACION LINEAL DIRECTA E INVERSA

A) Relación entre b_{yx} y φ:

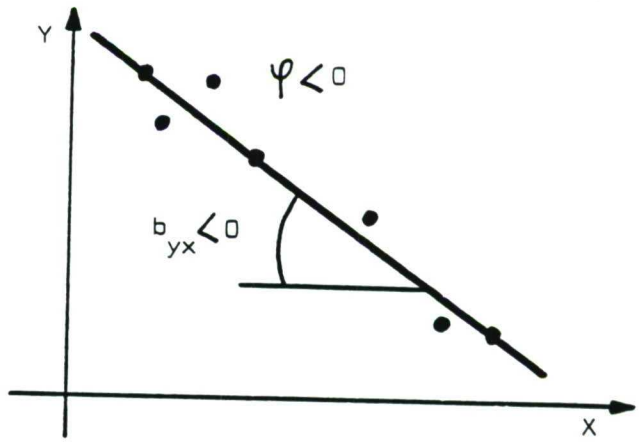
$$b_{yx} = \varphi \cdot \frac{\sigma_y}{\sigma_x} \quad (\text{siendo } \sigma_y \geq 0,\ \sigma_x \geq 0)$$

i) Si $b_{yx} > 0 \iff \varphi > 0$ (puesto que $\sigma_y \geq 0,\ \sigma_x \geq 0$)



En este caso decimos que hay "correlación lineal directa" entre las variables X e Y. De otra forma, la recta de regresión de Y sobre X es "creciente".

ii) Si $b_{yx} < 0 \iff \varphi < 0$ (puesto que $\sigma_y \geq 0,\ \sigma_x \geq 0$)



En este caso decimos que hay "correlación lineal inversa" entre las variables X e Y. Dicho de otra forma, la recta de regresión de Y sobre X es decreciente".

B) Relación entre b_{xy} y φ:

$$b_{xy} = \varphi \cdot \frac{\sigma_x}{\sigma_y} \quad (\text{siendo } \sigma_x \geq 0,\ \sigma_y \geq 0)$$

El razonamiento es análogo.

$$i)\ b_{xy} > 0 \iff \varphi > 0 \quad (\text{ya que } \sigma_y \geq 0,\ \sigma_x \geq 0)$$

La correlación lineal es directa.

La recta de regresión de X sobre Y es "creciente".

$$ii)\ b_{xy} < 0 \iff \varphi < 0 \quad (\text{ya que } \sigma_y \geq 0,\ \sigma_x \geq 0)$$

La correlación lineal es inversa.

La recta de regresión de X sobre Y es "decreciente".

ACTIVIDAD - 1

El equipo médico de una determinada clínica afirma que a medida que el niño desarrolla la capacidad intelectual, mayor es el éxito escolar. En base a los datos obtenidos en una prueba experimental:

Capacidad intelectual	5	5	7	8	2	3	6	6	9
Exito escolar	6	4	8	9	4	6	5	6	8

1) Raúl tiene 6'5 de capacidad intelectual, participa en la prueba experimental, ¿cuál es el éxito escolar que obtiene?.

2) ¿Es válida la afirmación del equipo médico?

Sean las variables estadísticas:

X = capacidad intelectual

Y = éxito escolar

1) Para poder "predecir" el éxito escolar Y que obtiene Raúl conocidos los valores de la capacidad intelectual X es necesario obtener la ecuación de la recta de regresión de Y sobre X.

La recta de regresión de Y sobre X es:

$$y - \bar{y} = \frac{m_{11}}{\sigma_x^2}(x - \bar{x})$$

donde:

$\bar{y}$ = media de la variable Y = a_{01}

$\bar{x}$ = media de la variable X = a_{10}

m_{11} = covarianza = $a_{11} - a_{10} \cdot a_{01}$

σ_x^2 = varianza de la X = $a_{20} - (a_{10})^2$

Para ello, formamos la correspondiente tabla de cálculos:

N = número de pares de datos = 9

Observa que tenemos un conjunto de 9 pares de datos de la forma (x_i, y_i) que representa:

"A la capacidad intelectual x_i le corresponde un éxito escolar - y_i".

Por tanto, la frecuencia de cada par (x_i, y_i) es la unidad.

	x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2
	5	6	30	25	36
	5	4	20	25	16
	7	8	56	49	64
	8	9	72	64	81
	2	4	8	4	16
	3	6	18	9	36
	6	5	30	36	25
	6	6	36	36	36
	9	8	72	81	64
SUMA	51	56	342	329	374

De esta forma tenemos la información suficiente para hallar:

$$a_{10},\ a_{01},\ a_{20},\ a_{02},\ a_{11},\ m_{11},\ \sigma_x^2,\ \sigma_y^2$$

En efecto:

$$\bar{x} = a_{10} = \frac{\sum_{i=1}^{9} x_i}{N} = \frac{51}{9} = 5'67$$

$$\bar{y} = a_{01} = \frac{\sum_{i=1}^{9} y_i}{N} = \frac{56}{9} = 6'22$$

$$a_{20} = \frac{\sum_{i=1}^{9} x_i^2}{N} = \frac{329}{9} = 36'55$$

$$a_{02} = \frac{\sum_{i=1}^{9} y_i^2}{N} = \frac{374}{9} = 41'55$$

$$a_{11} = \frac{\sum_{i=1}^{9} x_i \cdot y_i}{N} = \frac{342}{9} = 38$$

$$\sigma_x^2 = a_{20} - (a_{10})^2 = 36'55 - (5'67)^2 = 4'40$$

$$\sigma_x = + \sqrt{\sigma_x^2} = \sqrt{4'40} = 2'10$$

$$\sigma_y^2 = a_{02} - (a_{01})^2 = 41'55 - (6'22)^2 = 2'86$$

$$\sigma_y = + \sqrt{\sigma_y^2} = \sqrt{2'86} = 1'70$$

La covarianza m_{11} será:

$$m_{11} = a_{11} - a_{10} \cdot a_{01} = 38 - (5'67).(6'22) = 2'73$$

de donde:

La recta de regresión de Y sobre X es:

$$y - \bar{y} = \frac{m_{11}}{\sigma_x^2} (x - \bar{x})$$

esto es

$$y - 6'22 = \frac{2'73}{4'40} (x - 5'67)$$

o bien

$$y - 6'22 = 0'62\ (x - 5'67)$$

simplificando

$$y - 6'22 = 0'62.x - 3'52$$

$$\boxed{y = 2'70 + 0'62.x}$$

Por tanto, para la capacidad intelectual de Raúl ($x = 6'5$), el éxito escolar que se puede "predecir" viene dado por la ecuación:

$$y = 2'70 + (0'62).(6'5) = 6'73 \simeq 7 \quad \text{(aproximadamente)}$$

2) La validez de la afirmación del equipo de médicos se obtiene mediante el coeficiente de correlación lineal φ :

$$\varphi = \frac{m_{11}}{\sigma_x \cdot \sigma_y}$$

bastará sustituir los valores calculados:

m_{11} = covarianza = 2'73

σ_x = desviación típica de la X = 2'10

σ_y = desviación típica de la Y = 1'70

de donde

$$\varphi = \frac{2'73}{(2'10).(1'70)} = 0'76$$

Observa que $-1 < 0'76 < 1$

Se concluye, la validez de la afirmación es de 0'76 en correlación directa,

puesto que $\varphi = 0'76 > 0$. Es decir, a medida que aumenta la capacidad intelectual, aumenta el éxito escolar.

En porcentajes, la validez pedida es de un 76 por 100 en correlación directa.

$$\text{porcentaje} = \varphi \cdot 100 = (\%)_{\varphi}$$

ACTIVIDAD – 1: Estudiar la correlación entre las variables X e Y, a partir de los siguientes datos:

a)

X	Y
2	3
4	5
3	2
5	1
6	3

b)

X	Y
1	5
3	2
4	4
6	9
3	3

ACTIVIDAD – 2: Se han tomado datos entre el número de cigarrillos consumidos diariamente y la mortalidad. Obteniéndose la siguiente tabla:

nº de cigarrillos X	3	5	6	15	20	40	45
Mortalidad por 1000 habitantes Y	0'2	0'3	0'3	0'5	0'7	1'4	1'5

Se pide:

1) Ajustar una recta a los datos obtenidos: recta de regresión de Y sobre X.

2) Estudiar la correlación entre ambas variables.

3) ¿Qué mortalidad se podría predecir para 60 cigarrillos?.

ACTIVIDAD – 3: En un pueblo de Segovia se observó el precio de la resina y la cantidad de kilos producidos, obteniéndose la siguiente tabla:

Precio en pesetas X	20	25	30	32	37	48	50	51
Miles de kilos Y	100	110	120	140	160	200	205	210

Se pide:

1) La recta de regresión de Y sobre X.

2) Coeficiente de correlación.

ACTIVIDAD - 4: Dada la tabla

X	2	4	6	7	8	9	10
Y	3	5	5	6	9	10	12

Hallar:

1) Diagrama de dispersión.

2) Recta de regresión de Y sobre X.

3) Coeficiente de correlación.

4) ¿Qué valor se espera para la variable Y cuando $x = 15$?.

AUTOCOMPROBACION

ACTIVIDAD - 1:

a) $\varphi = -0'10$ correlación inversa

b) $\varphi = 0'60$ correlación directa

ACTIVIDAD - 2:

1) $y = 0'098 + 0'03.x$ recta regresión

2) $\varphi = 0'99$ correlación directa

3) 1'898 por 1000 habitantes

ACTIVIDAD - 3:

1) $y = 18'56 + 3'74.x$ recta regresión

2) $\varphi = 0'99$ correlación directa

ACTIVIDAD - 4:

2) $y = 0'059 + 1'07.x$ recta regresión

3) $\varphi = 0'93$ correlación directa

4) Se puede predecir y = 16'109

ESTADISTICA DESCRIPTIVA

ACTIVIDADES

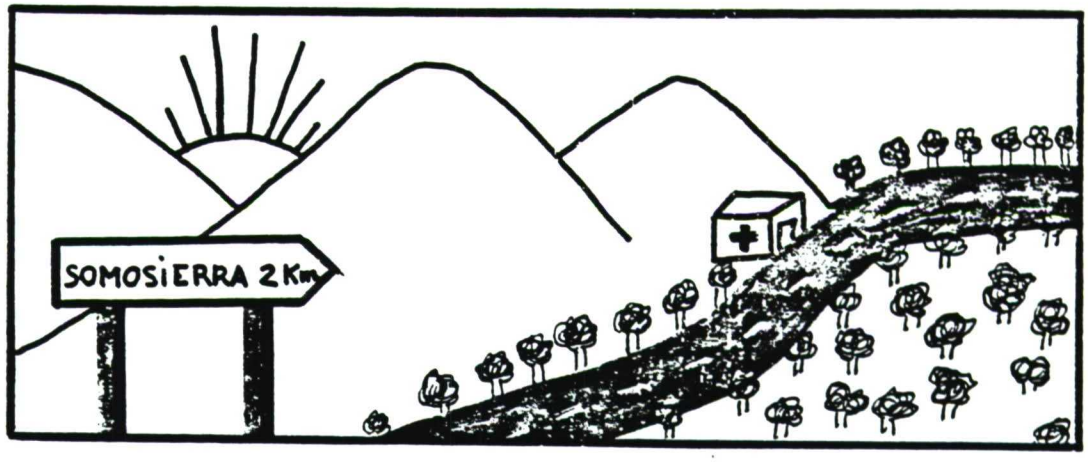
III

SUCESOS Y PROBABILIDAD

- Fenómenos aleatorios.
- Probabilidad.
- Probabilidad condicionada.
- Independencia.
- Probabilidad total.
- Teorema de Bayes.
- Asignación de probabilidades.
- Análisis combinatorio.

"SOMOSIERRA EN FIESTAS"

Todos los veranos en las fiestas del pueblo de Somosierra se elige un "comité de fiestas" formado por dos jóvenes que se encargan de gobernar el pueblo.



En el verano de 1.986 se presentaron cuatro jóvenes para ser elegidos dos de ellos.


MIGUEL

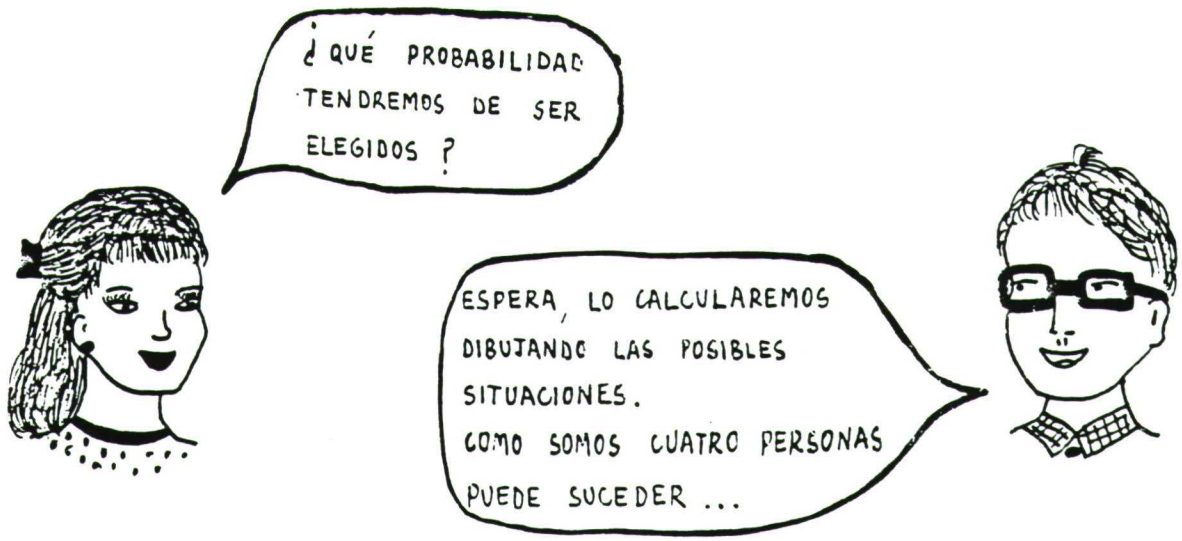

ANA

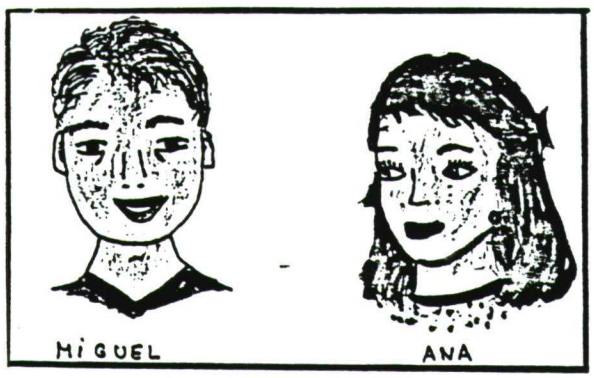

SANTIAGO


BLANCA

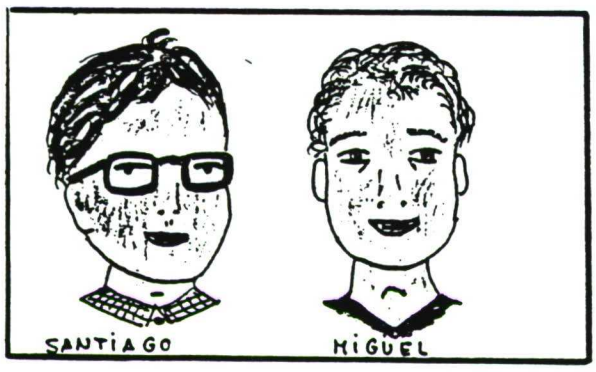
Ana y Santiago son muy amigos y quieren ser elegidos como autoridad de las fiestas.

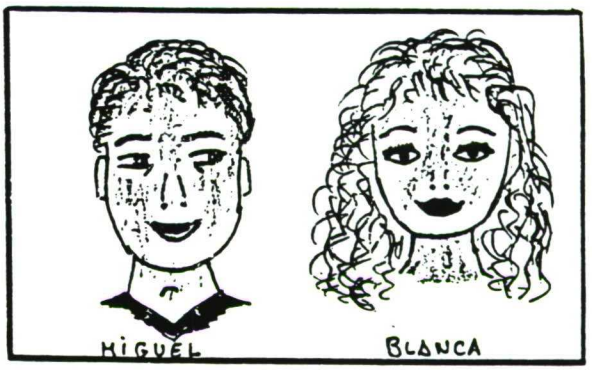
A medida que las fiestas se acercaban se oían los siguientes comentarios:

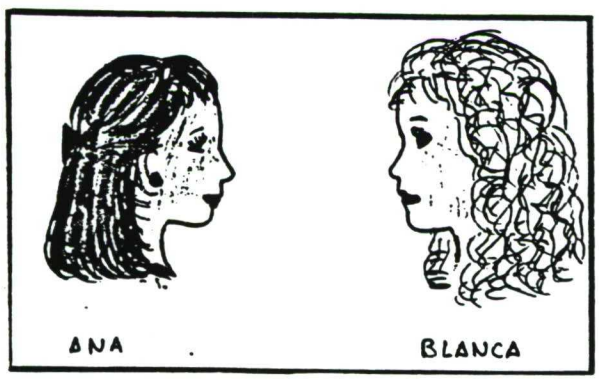

¿QUÉ PROBABILIDAD TENDREMOS DE SER ELEGIDOS ?
ESPERA, LO CALCULAREMOS DIBUJANDO LAS POSIBLES SITUACIONES.
COMO SOMOS CUATRO PERSONAS PUEDE SUCEDER ...


MIGUEL
ANA


SANTIAGO
BLANCA


SANTIAGO
MIGUEL


MIGUEL
BLANCA


ANA
BLANCA


SANTIAGO
ANA

ENTONCES ...
SE PUEDEN FORMAR
SEIS PAREJAS DIFERENTES

ES EVIDENTE.
YA QUE PUEDEN SER PAREJAS
MIXTAS Ó FORMADAS SÓLO
POR DOS CHICOS Ó DOS CHICAS

¡AH! YA ENTIENDO ...
POR ESO LA PROFESORA DE MATEMÁTICAS
EXPLICABA LO DE LAS "COMBINACIONES".

REALMENTE LO QUE HACEMOS
SON TODAS LAS POSIBLES COMBINACIONES
QUE PUEDEN FORMAR CUATRO PERSONAS
TOMADAS DE DOS EN DOS.

MIRA SANTIAGO, QUE RÁPIDO SALE APLICANDO LA FÓRMULA:

$$C_{4,2} = \binom{4}{2} = \frac{4!}{2!\,(4-2)!} =$$

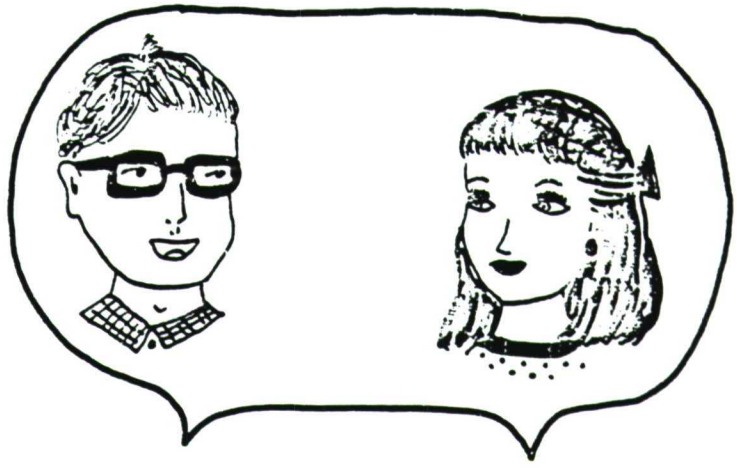
$$= \frac{4!}{2!\,2!} = \frac{4\cdot 3\cdot 2\cdot 1}{(2\cdot 1)\cdot(2\cdot 1)} = 6 \text{ FORMAS DISTINTAS}$$

TE LO EXPLIQUÉ CON DIBUJOS PORQUE CREÍ QUE CON LA FÓRMULA NO LO VERÍAS TAN CLARO. PERO LO QUE REALMENTE QUEREMOS SABER NO SON LAS PAREJAS QUE SE PUEDEN FORMAR, SINO LA PROBABILIDAD DE QUE ESA PAREJA SEAMOS NOSOTROS DOS

¡ESO SE VE FÁCIL CON TUS DIBUJOS! COMO SON SEIS DIBUJOS Y NOSOTROS DOS SÓLO ESTAMOS JUNTOS EN UNO... LA PROBABILIDAD DE QUE SEAMOS LOS DOS LOS ELEGIDOS SERÁ: $\frac{1}{6}$

SÍ, Y COMO SIEMPRE HAY UNA FÓRMULA QUE EXPLICA ESTO.

$$\text{PROBABILIDAD} = \frac{\text{Nº CASOS FAVORABLES}}{\text{Nº CASOS POSIBLES}} = \frac{1}{6}$$



¡BAH! QUÉ PROBABILIDAD TAN PEQUEÑA
¡CONFIEMOS EN LA SUERTE!

FENOMENOS ALEATORIOS:

En el mundo real hay fenómenos regidos por leyes determinadas, es decir, bajo condiciones dadas; el resultado es previsible; a estos fenómenos se les llama "fenómenos deterministas", piensa como un ejemplo de ellos en la caída de un objeto desde una determinada altura.

También existen otros fenómenos que no siguen unas leyes determinadas. - Así, en el lanzamiento de una moneda no se puede predecir con certeza si sale cara o cruz.

Un fenómeno o experimento se denomina "aleatorio" cuando puede dar lugar a varios resultados, sin que sea posible anunciar con certeza cuál de éstos va a ocurrir.

En muchas ciencias se estudian fenómenos aleatorios: biología, economía, física, medicina, ingeniería, sociales, política, etc.

Al describir un experimento aleatorio, es esencial saber qué aspecto del resultado nos interesa observar, dicho de otro modo, cuál es nuestro criterio

para considerar dos resultados como diferentes. Este objetivo se logra mediante el "ESPACIO MUESTRAL".

ESPACIO MUESTRAL:

Dado un experimento aleatorio, el conjunto de los posibles resultados diferentes del mismo, recibe el nombre de espacio muestral asociado al experimento aleatorio.

"Al espacio muestral lo denotaremos por Ω "

Así, en el lanzamiento de un dado pueden presentarse seis casos, que son los números 1, 2, 3, 4, 5 ó 6.

Se puede tomar, por tanto, como espacio muestral:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

SUCESO:

Un suceso asociado a un experimento aleatorio corresponde a la cuestión de que tenga o no tenga respuesta después de realizado el experimento.

Piensa en la pregunta: "¿Es posible que aparezca un 3 ó un 4 en la tirada de un dado?".

Esta pregunta tiene respuesta y es, por tanto, un "suceso".

El subconjunto de Ω que responde "sí" a la pregunta es:

$$A = \{3, 4\}$$

De donde:

"Un suceso de un experimento aleatorio es un subconjunto del espacio muestral Ω".

OPERACIONES CON SUCESOS:

Dados el experimento aleatorio y su espacio muestral asociado Ω, hemos

definido los sucesos. Vamos a definir ciertas operaciones con los sucesos:

• UNION DE SUCESOS: Dados dos sucesos A y B, de un cierto experimento aleatorio, se define la "unión" de A y B, que denotamos por $A \cup B$, a otro suceso que ocurre siempre que ocurra A o siempre que ocurra B.

$A \cup B$: Se verifica cuando se verifica A ó B

• INTERSECCION DE SUCESOS: Dados dos sucesos A y B, de un cierto experimento aleatorio, se define la "intersección" de A y B, que denotamos por $A \cap B$, a otro suceso que ocurre siempre que ocurra A y que ocurra B.

$A \cap B$: Se verifica cuando se verifica A y B

• SUCESO COMPLEMENTARIO: Dado un suceso A de un experimento aleatorio, se define el complementario de A, que representamos por $\bar{A}$, a otro suceso que ocurre siempre que no ocurre A.

Piensa en la pregunta: "¿Cuál será el suceso complementario en la tirada de un dado al suceso en que aparezca un 3 ó un 4?".

En el lanzamiento de un dado, el espacio muestral es:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

El suceso en que aparecen un 3 ó un 4 es:

$$A = \{3, 4\}$$

El suceso complementario $\bar{A}$ será aquel suceso en que no aparecen ni un 3 ni un 4, es decir:

$$\bar{A} = \{1, 2, 5, 6\}$$

Piensa: $A \cup \bar{A} = \{1, 2, 3, 4, 5, 6\} = \Omega$

• SUCESO IMPOSIBLE: Es el suceso que no ocurre nunca, en otras palabras, dados el suceso A y su complementario $\bar{A}$, junto con la operación de intersección, se define un suceso que no ocurre nunca.

"Al suceso imposible se le denota por $\emptyset$".

$$\boxed{A \cap \bar{A} = \emptyset}$$

Piensa en la pregunta: "¿Cuál será el suceso complementario del suceso imposible $\emptyset$?".

El suceso complementario de $\emptyset$ será el suceso que ocurre siempre, por tanto:

$$\emptyset = \Omega$$

• SUCESO CONTENIDO EN OTRO: Dados dos sucesos A y B de un experimento aleatorio, se dice que el suceso A está contenido en el suceso B si siempre que —

ocurra A ocurre B, se denota por:

$$A \subset B : \text{Si ocurre A ocurre B}$$

Piensa en la pregunta: "¿Es cierto que cualquier suceso está contenido en el espacio muestral Ω?".

$$\text{Siempre ocurre} : \emptyset \subset A \subset \Omega$$

En este caso, A está contenido en B, por tanto:

$$\emptyset \subset A \subset B \subset \Omega$$

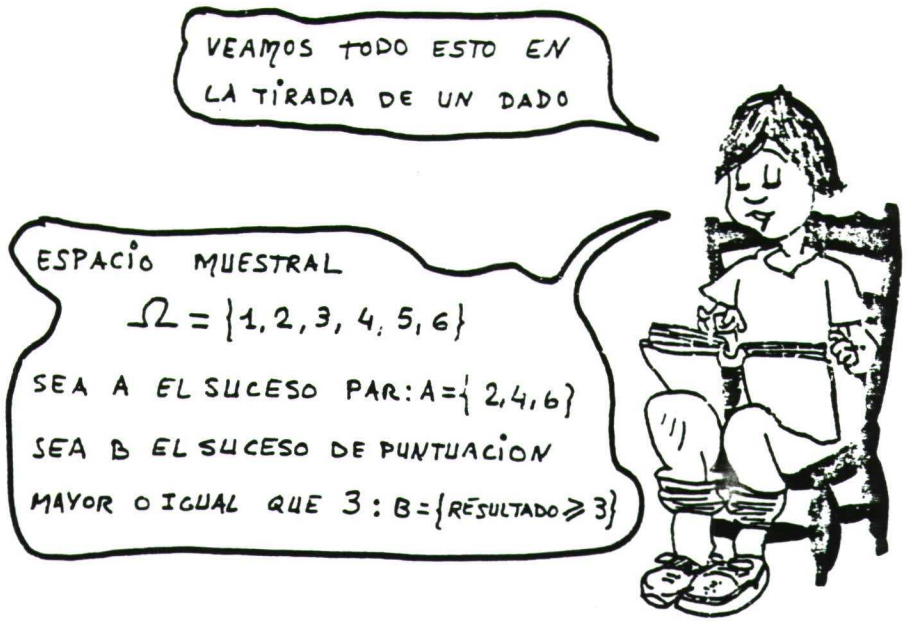
• SUCESOS INCOMPATIBLES: Dados dos sucesos A y B de un experimento aleatorio, se dice que son dos sucesos incompatibles cuando su intersección es el suceso imposible $\emptyset$.

De donde:

$$A \cap B = \emptyset : \text{A y B son dos sucesos incompatibles}$$

• DIFERENCIA DE SUCESOS: Sean A y B dos sucesos de un experimento aleatorio. La diferencia de los sucesos A y B, se denota por A − B, es el suceso que ocurra A y no ocurra B.

$$A - B : \text{ocurra A y no ocurra B}$$



Sea el experimento aleatorio del lanzamiento de un dado.

El espacio muestral: $\Omega = \{1, 2, 3, 4, 5, 6\}$

Sea A el suceso el resultado es par: $A = \{2, 4, 6\}$

Sea B el resultado es mayor o igual que 3: $B = \{3, 4, 5, 6\}$

"Ambos sucesos tienen respuesta después de realizado el experimento".

Veamos:

$A \cup B = \{$ el resultado es par o es mayor o igual a $3\}$

$A \cup B = \{2, 3, 4, 5, 6\}$

$A \cap B = \{$ el resultado es par y es mayor o igual a $3\}$

$A \cap B = \{4, 6\}$

Piensa que los sucesos A y B "no son incompatibles" puesto que la intersección no da el suceso imposible $\emptyset$.

$\bar{A} = \{$ suceso que ocurre siempre que no ocurre A$\}$

$\bar{A} = \{$ el resultado no es par$\}$

$\bar{A} = \{1, 3, 5\}$

$\bar{B} = \{$ suceso que ocurre siempre que no ocurre B $\}$

$\bar{B} = \{$ el resultado no es mayor o igual a 3 $\}$

$\bar{B} = \{1, 2\}$

$A - B = \{$ el resultado es par pero no es mayor o igual a 3 $\}$

$A - B = \{2\}$

DEFINICION CLASICA DE PROBABILIDAD:

La probabilidad p de un suceso o acontecimiento E es igual al número f de casos favorables (casos en que se verifica E) dividido por el número total n de casos posibles:

$$p = P\{E\} = \frac{f}{n}$$

La probabilidad de no aparición del suceso E viene dada por:

$$q = P\{\text{no } E\} = \frac{n - f}{n} = \frac{n}{n} - \frac{f}{n} = 1 - \frac{f}{n} = 1 - P\{E\}$$

Así, pues, $p + q = 1$, en otras palabras:

$$P\{E\} + P\{\text{no } E\} = 1$$

El suceso "no E" se denota por $\bar{E}$.

En la tirada de un dado pueden presentarse seis casos, que son los números 1, 2, 3, 4, 5 ó 6, por tanto, el espacio muestral es:

$$\Omega = \{ 1, 2, 3, 4, 5, 6 \}$$

Llamemos E al suceso de que aparezcan los números 3 ó 4 en una sola tirada de un dado, es decir:

$$E = \{ 3 \text{ ó } 4 \}$$

Puesto que E puede presentarse con dos de estos casos, se tiene:

$$p = P\{E\} = P\{3 \text{ ó } 4\} = \frac{2}{6} = \frac{1}{3} = 0'333 \ldots$$

La probabilidad de no obtener un 3 ó 4 (es decir, obtener 1, 2, 5 ó 6) es:

$$p = P\{\bar{E}\} = P\{1, 2, 5, 6\} = \frac{4}{6} = \frac{2}{3} = 0'666 \ldots$$

P I E N S A

La probabilidad de un suceso es un número comprendido entre 0 y 1.

Si el suceso es imposible (no puede ocurrir), su probabilidad es 0.

Si el suceso es cierto (ocurre siempre), su probabilidad es 1.

En otras palabras:

$$0 \leq P\{E\} \leq 1$$

$$P(\Omega) = = 1$$

$$P(\emptyset) = 0$$

ACTIVIDAD – 1: Un número es seleccionado al azar entre los números 1 al 10. Sea A el suceso "el número elegido es par". Sea B el suceso "el número elegido es primo". Se pide expresar los siguientes sucesos: $A \cap B$, $A \cup B$, $\bar{A} \cap \bar{B}$ y $\bar{A} \cup \bar{B}$.

El espacio muestral es:

$$\Omega = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

El suceso $A = \{2, 4, 6, 8, 10\}$, de donde, $\bar{A} = \{1, 3, 5, 7, 9\}$

El suceso $B = \{1, 2, 3, 5, 7\}$, de donde, $\bar{B} = \{4, 6, 8, 9, 10\}$

con lo cual tenemos:

$A \cap B = \{2\}$ el único primo y par es el número $\{2\}$

$A \cup B = \{1, 2, 3, 4, 5, 6, 7, 8, 10\}$ son los números pares o primos de 1 a 10.

$\bar{A} \cap \bar{B} = \{9\}$ números impares y no primos de 1 a 10

$\bar{A} \cup \bar{B} = \{1, 3, 4, 5, 6, 7, 8, 9, 10\}$ números impares o no primos de 1 a 10

P I E N S A

$(\overline{A \cup B}) = \{9\} = \bar{A} \cap \bar{B}$

$(\overline{A \cap B}) = \{1, 3, 4, 5, 6, 7, 8, 9, 10\} = \bar{A} \cup \bar{B}$

recuerda:

Primera ley de Morgan: $(\overline{A \cup B}) = \bar{A} \cap \bar{B}$

Segunda ley de Morgan: $(\overline{A \cap B}) = \bar{A} \cup \bar{B}$

ACTIVIDAD - 2: Un número es seleccionado al azar entre los números 1 al 10. Sea A el suceso "el número elegido es par". Sea C el suceso "el número elegido es múltiplo de 3". Se pide expresar los siguientes sucesos: $A \cap C$, $A - C$, $A \cap \bar{C}$.

El espacio muestral es:

$$\Omega = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

El suceso $A = \{2, 4, 6, 8, 10\}$, luego, $\bar{A} = \{1, 3, 5, 7, 9\}$

El suceso $C = \{3, 6, 9\}$, entonces, $\bar{C} = \{1, 2, 4, 5, 7, 8, 10\}$

por tanto tenemos:

$A \cap C = \{6\}$ único número par y múltiplo de 3 de 1 a 10

$A - C = \{2, 4, 8, 10\}$ números pares que no son múltiplos de 3 de 1 a 10

$A \cap \bar{C} = \{2, 4, 8, 10\}$ números pares que no son múltiplos de 3 de 1 a 10

P I E N S A

$A - C = \{2, 4, 8, 10\} = A \cap \bar{C}$

ACTIVIDAD - 3: Un número es seleccionado al azar entre los números 1 al 7. Sea A el suceso "el número elegido es par". Sea B el suceso "el número elegido es primo". Ambos sucesos son tales que $P(A) = 0'45$ y $P(B) = 0'80$.

Se pide:

1) Calcular la probabilidad $P(A \cap B)$.
2) Calcular la probabilidad $P(A \cup \bar{B})$.
3) Calcular la probabilidad $P(\bar{A} \cup B)$.
4) Calcular la probabilidad $P(\bar{A} \cup \bar{B})$.

El espacio muestral es $\Omega = \{1, 2, 3, 4, 5, 6, 7\}$

El suceso $A = \{2, 4, 6\}$, de donde, $\bar{A} = \{1, 3, 5, 7\}$

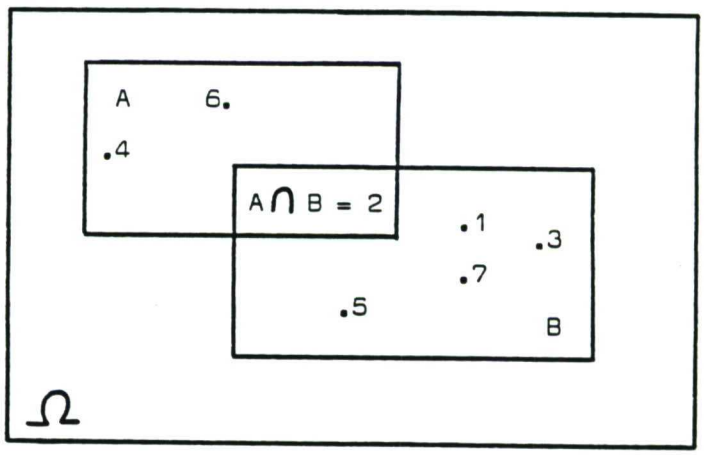
El suceso $B = \{1, 2, 3, 5, 7\}$, de donde, $\bar{B} = \{4, 6\}$

- con lo cual tenemos:

$A \cup B = \{1, 2, 3, 4, 5, 6, 7\} = \Omega$

$A \cap B = \{2\}$

Gráficamente:



En la figura observamos que los dos sucesos no son disjuntos (incompatibles), pues $A \cap B = \{2\}$.

entonces

$$P(A \cup B) = P(A) - P(A \cap B) + P(B)$$

1) Como por hipótesis $P(A) = 0'45$ y $P(B) = 0'80$, tenemos:

$$P(A \cup B) = P(\Omega) = 1$$

de donde

$$1 = 0'45 - P(A \cap B) + 0'80$$

siendo

$$P(A \cap B) = 0'25$$

2) $A \cup \bar{B} = \{2, 4, 6\} = A$

con lo cual

$$P(A \cup \bar{B}) = P(A) = 0'45$$

3) $\bar{A} \cup B = \{1, 2, 3, 5, -7\} = B$

Por otra parte

$$P(\bar{A} \cup B) = P(B) = 0'80$$

4) $\bar{A} \cup \bar{B}$ = aplicando la segunda ley de Morgan = $\overline{(A \cap B)} = 1 - (A \cap B)$

de donde

$$P(\bar{A} \cup \bar{B}) = P\overline{(A \cap B)} = 1 - P(A \cap B) = 1 - 0'25 = 0'75$$

P I E N S A

- Sean A y B dos sucesos cualesquiera del espacio muestral Ω, entonces

$$P(A \cup B) = P(A) - P(A \cap B) + P(B)$$

dos sucesos A y B son incompatibles cuando $A \cap B = \emptyset$.

- Cuando A y B son dos sucesos incompatibles, se tiene:

$$P(A \cup B) = P(A) - P(A \cap B) + P(B)$$

como por hipótesis $P(A \cap B) = P(\emptyset) = 0$

con lo cual

$$P(A \cup B) = P(A) + P(B)$$

ACTIVIDAD – 4: Sean A, B y C tres sucesos incompatibles, con $P(A) = 0'6$, $P(B) = 0'25$, $P(C) = 0'10$.

Calcular las siguientes probabilidades:

1) $P(\bar{A} \cap \bar{B})$.

2) $P(\bar{A} \cap \bar{B} \cap \bar{C})$.

3) $P(\bar{A} \cap \bar{B} \cap C)$.

1) Sea el suceso $\bar{A} \cap \bar{B}$, por la primera ley de Morgan:

$$\overline{(A \cup B)} = \bar{A} \cap \bar{B}$$

como los sucesos A y B son incompatibles:

$$P(A \cup B) = P(A) + P(B)$$

con lo cual

$$P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - \left[P(A) + P(B)\right] =$$

$$= 1 - P(A) - P(B) = 1 - 0'6 - 0'25 = 0'15$$

2) Sea el suceso $\bar{A} \cap \bar{B} \cap \bar{C}$, considerando $\bar{A} \cap \bar{B}$ como un solo suceso, queda:

$\left((\bar{A} \cap \bar{B}) \cap \bar{C}\right)$ = aplicando ley de Morgan =

$= \left(\overline{(A \cup B)} \cap \bar{C}\right)$ = aplicando ley de Morgan =

$= (\overline{A \cup B \cup C})$

como los sucesos A, B y C son incompatibles:

$$P(\bar{A} \cap \bar{B} \cap \bar{C}) = P(\overline{A \cup B \cup C}) = 1 - P(A \cup B \cup C) =$$

$$= 1 - \left[P(A) + P(B) + P(C)\right] = 1 - (0'60 + 0'25 + 0'10) =$$

$$= 1 - 0'95 = 0'05$$

3) Sea el suceso $\bar{A} \cap \bar{B} \cap \bar{C}$

- A y C sucesos incompatibles si $A \cap C = \emptyset$, esto quiere decir que $C \subset \bar{A}$.
- B y C sucesos incompatibles si $B \cap C = \emptyset$, esto quiere decir que $C \subset \bar{B}$.

de donde

$\bar{A} \cap C = C$

$\bar{B} \cap C = C$

con lo cual, $\bar{A} \cap \bar{B} \cap C = C$

se deduce que

$$P(\bar{A} \cap \bar{B} \cap C) = P(C) = 0'10$$

ACTIVIDADES

ACTIVIDAD - 5: Un número es seleccionado al azar entre los números 1 al 7. Sea A el suceso "el número elegido es par". Sea B el suceso "el número elegido es primo". Sea C el suceso "el número elegido es múltiplo de 3".

Expresa los siguientes sucesos:

$A \cap C$, $B \cap C$, $\bar{A} \cap C$, $(A \cup B) \cap \bar{C}$

ACTIVIDAD - 6: Sean A, B y C tres sucesos incompatibles con $P(A) = 0'2$, $P(B) = 0'4$, $P(C) = 0'5$.

Calcular las siguientes probabilidades:

1) $P(\bar{A} \cap \bar{B})$

2) $P(\bar{A} \cup \bar{B})$

3) $P(\bar{A} \cap \bar{B} \cap \bar{C})$

4) $P(\bar{A} \cap \bar{B} \cap C)$

5) $P(B - A)$ • Nota: $B - A = B \cap \bar{A}$

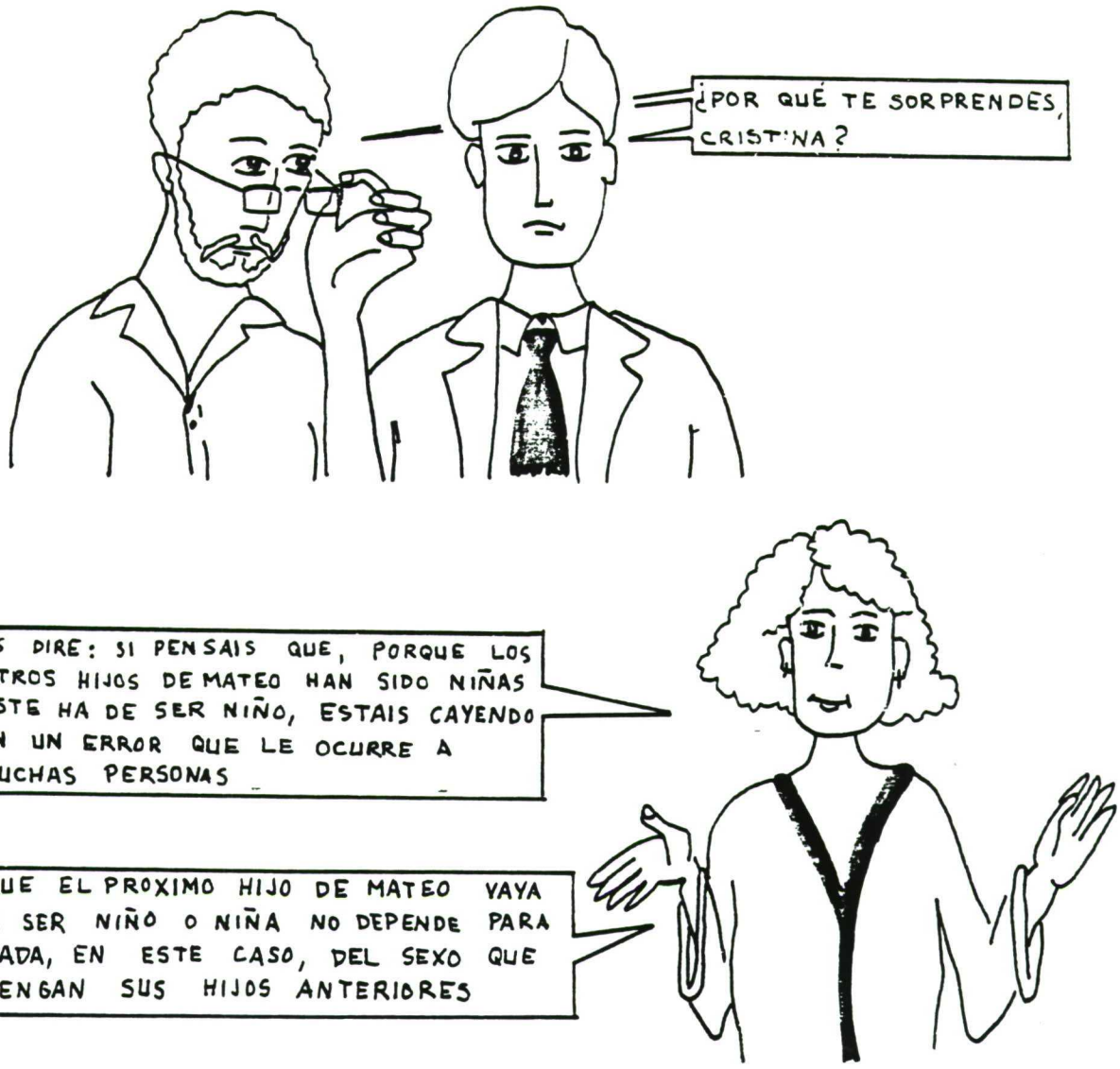
$B \cap A = \emptyset$, $B \subset \bar{A}$

Mateo va a casa de su amigo Raúl para darle la buena noticia de que va a volver a ser padre.

Después de felicitarle, su amigo Raúl asegura que, como Mateo ya ha tenido cuatro niñas, su quinto hijo tiene que ser un niño.

Cristina, la mujer de Raúl, pone una cara extraña y no parece estar de acuerdo con lo que acaba de decir su marido.

Mateo y Raúl se sorprenden de la actitud de Cristina.



Cristina tiene razón. Mateo y Raúl han caído en la trampa conocida como "la falacia del jugador". En este caso, el resultado del siguiente acontecimiento (sexo del futuro hijo), no depende de los anteriores. La probabilidad de que Mateo y su mujer tengan otra niña es la misma de que su primer hijo ya lo fuera.

Supongamos que Mateo va lanzando una moneda equilibrada y saca cuatro caras seguidas. La probabilidad de que en un nuevo lanzamiento salga cara es idéntica a la de antes: un cincuenta por ciento. Así mismo ocurre con cada nuevo embarazo. La probabilidad de tener niño o niña siempre es del 50 % (independientemente del sexo de los hijos anteriores).


BUENO, EN CUALQUIER CASO, NIÑO O NIÑA, LO IMPORTANTE ES QUE HE PEDIDO UNAS VACACIONES ADELANTADAS PARA NOVIEMBRE, QUE ES CUANDO ESPERAMOS QUE NAZCA NUESTRO HIJO.
VAMOS A REALIZAR UN VIAJE EN BARCO A BRASIL DONDE MI MUJER TIENE A SUS PADRES, PARA QUE CONOZCAN AL BEBE
SI, EL BARCO ES UN MEDIO DE TRANSPORTE BASTANTE SEGURO
ES VERDAD, PERO TAMBIEN DEPENDE DEL DIA QUE HAGA: PUEDE QUE HAYA NIEBLA Y PUEDE QUE NO.
SI HAY NIEBLA, LAS PROBABILIDADES DE ACCIDENTE SON ALGO MAYORES, AUNQUE, DESGRACIADAMENTE, UN ACCIDENTE PUEDE OCURRIR CON NIEBLA O SIN ELLA.

YA ENTIENDO LO QUE QUIERES DECIR; EN UN DIA CUALQUIERA PUEDE OCURRIR:

- Que haya niebla ese día : N
- Que no haya niebla ese día : $\bar{N}$

Y ADEMAS, PUEDE SUCEDER:

- Que en un día con niebla ocurra un accidente. $A|N$
- Que en un día con niebla no ocurra ningún accidente $\bar{A}|N$
- Que en un día sin niebla ocurra un accidente $A|\bar{N}$
- Que en un día sin niebla no ocurra ningún accidente $\bar{A}|\bar{N}$

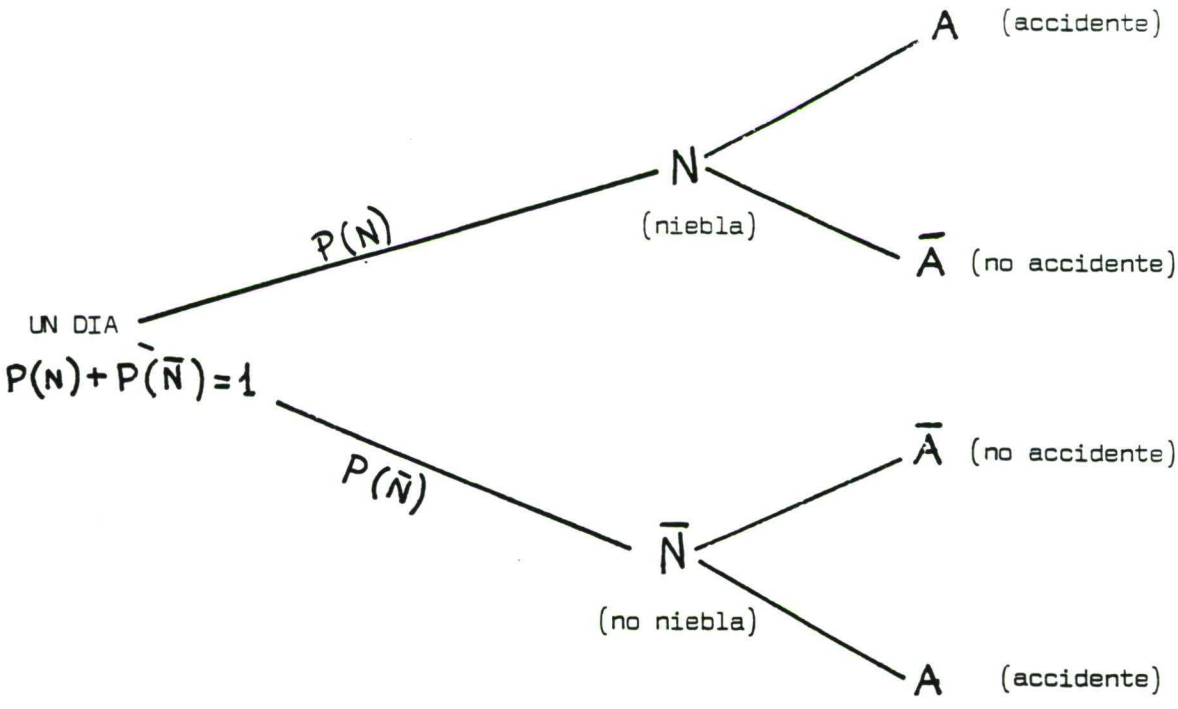
N = Niebla

$\bar{N}$ = No niebla

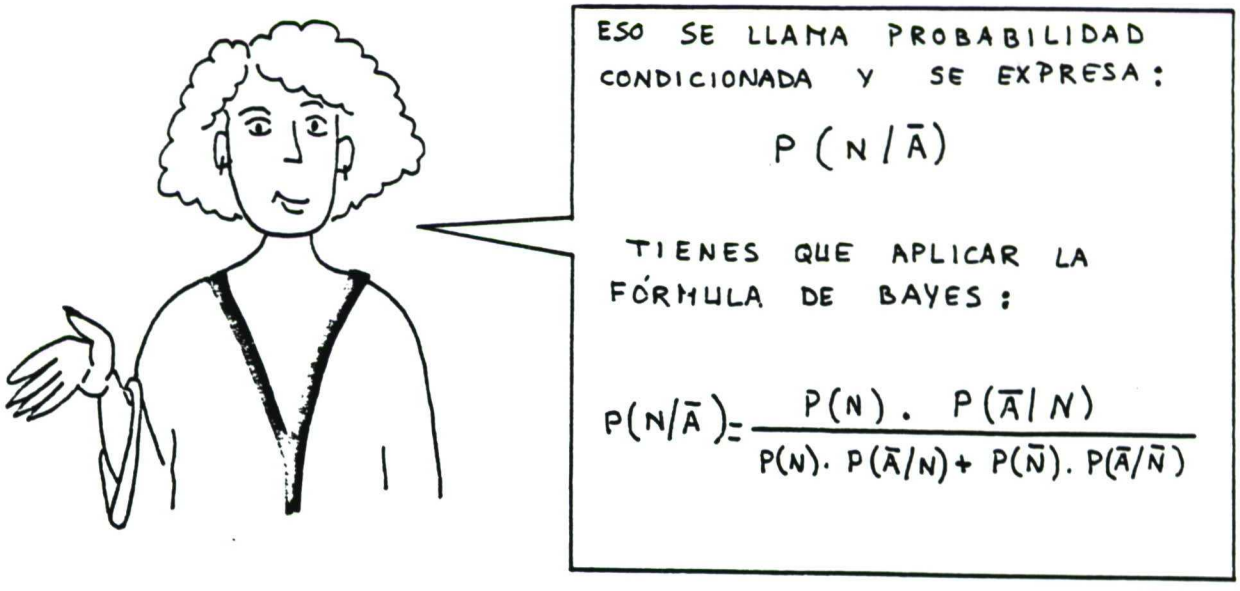
A = Accidente

$\bar{A}$ = No accidente

El "Diagrama de árbol" queda:




ACCIDENTES CASÍ NO OCURREN
¡SEAMOS OPTIMISTAS!
¿QUÉ PROBABILIDAD TENEMOS PARA QUE SEA UN DÍA CON NIEBLA NO OCURRIENDO ACCIDENTE?


ESO SE LLAMA PROBABILIDAD CONDICIONADA Y SE EXPRESA:
$P(N/\bar{A})$
TIENES QUE APLICAR LA FÓRMULA DE BAYES:
$P(N/\bar{A}) = \frac{P(N) \cdot P(\bar{A}/N)}{P(N) \cdot P(\bar{A}/N) + P(\bar{N}) \cdot P(\bar{A}/\bar{N})}$

PIENSA:

$$\text{Probabilidad} = \frac{\text{número de casos favorables}}{\text{número de casos posibles}}$$

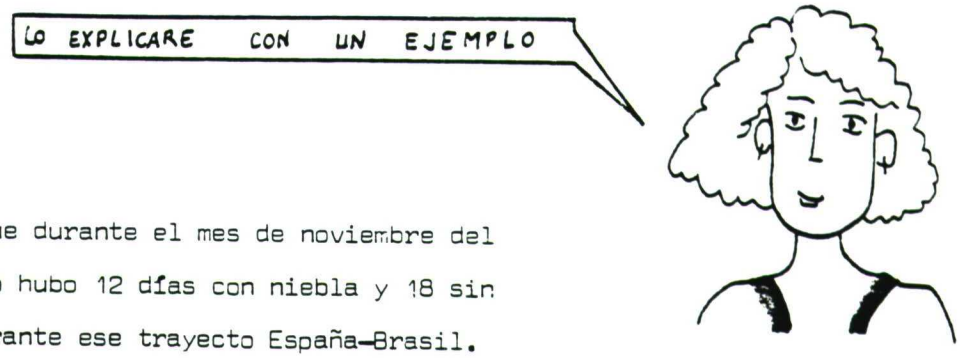
Efectivamente, según BAYES: la probabilidad de que no habiendo accidente sea un día con niebla, $P(N/\bar{A})$ es igual:

- CASOS FAVORABLES: La probabilidad de que el día sea con niebla $P(N)$ por la probabilidad de que siendo el día con niebla no haya accidente $P(\bar{A}/N)$.

- CASOS POSIBLES: Son todos los casos. Es decir, a los casos favorables mencionados, se les suma los restantes, que son: la probabilidad de que el día sea sin niebla $P(\bar{N})$ por la probabilidad de que siendo el día sin niebla no ocurra accidente $P(\bar{A}/\bar{N})$.

Por tanto:

$$P(N/\bar{A}) = \frac{P(N) \cdot P(\bar{A}/N)}{P(N) \cdot P(\bar{A}/N) + P(\bar{N}) \cdot P(\bar{A}/\bar{N})}$$



Supcned que durante el mes de noviembre del año pasado hubo 12 días con niebla y 18 sin niebla durante ese trayecto España-Brasil.

MES DE NOVIEMBRE						
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

DIA SIN NIEBLA ($\bar{N}$) DIA CON NIEBLA (N)

- La probabilidad de que un día cualquiera tuviese niebla P(N) es:

$$P(N) = \frac{12}{30} = 0'4$$

- La probabilidad de que un día del mes no tuviese niebla $P(\bar{N})$ es:

$$P(\bar{N}) = \frac{18}{30} = 0'6$$

PIENSA: $P(N) + P(\bar{N}) = 0'4 + 0'6 = 1$

Suponed que los actuarios de la Empresa Marítima tienen previsto que en el mes de Noviembre:

- Las probabilidades de accidente en un día sin niebla $P(A/\bar{N})$ son:

$$P(A/\bar{N}) = 0'0001$$

de donde

Las probabilidades de que no ocurra un accidente un día sin niebla, son:

$$P(\bar{A}/\bar{N}) = 1 - 0'0001 = 0'9999$$

- Las probabilidades de accidente cuando el día tiene niebla $P(A/N)$ son:

$$P(A/N) = 0'0003$$

de donde

Las probabilidades de que no ocurra un accidente un día con niebla, son:

$$P(\bar{A}/N) = 1 - 0'0003 = 0'9997$$

Habrás observado tú mismo que conocemos todos los datos para poder aplicar la fórmula de Bayes:

$$P(N/\bar{A}) = \frac{P(N) \cdot P(\bar{A}/N)}{P(N) \cdot P(\bar{A}/N) + P(\bar{N}) \cdot P(\bar{A}/\bar{N})} =$$

$$= \frac{(0'4) \cdot (0'9997)}{(0'4) \cdot (0'9997) + (0'6) \cdot (0'9999)} =$$

$$= \frac{0'39988}{0'39988 + 0'59994} = 0'3999 \simeq 0'40 \text{ (aproximadamente)}$$

La probabilidad de que no habiendo accidente el día sea con niebla es $P(N/\bar{A}) = 0'40$.

Parece lógico pensar que, si hace un tiempo muy parecido al del año anterior y no ocurre accidente, se tenga aproximadamente la misma probabilidad de que un día cualquiera tuviese niebla $P(N) = 0'40$.

- PROBABILIDAD CONDICIONADA
- INDEPENDENCIA
- BAYES

PROBABILIDAD CONDICIONADA:

Sean A y B dos sucesos de un experimento aleatorio, la probabilidad de que ocurra B, dado que ha ocurrido A, se denota por P(B/A) y se llama "probabilidad condicionada" de B dado que A se ha presentado.

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

Sea el ejemplo siguiente:

El lanzamiento de un dado tiene un espacio muestral asociado:

$$\Omega = \{1,2,3,4,5,6\}$$

siendo

$$P\{1\} = P\{2\} = P\{3\} = P\{4\} = P\{5\} = P\{6\} = \frac{1}{6}$$

es decir, la probabilidad de que salga cualquier resultado es igualmente posible, se tiene que $p = \frac{1}{6}$, "los resultados son equiprobables".

- Sea A el suceso que el resultado sea par:

$$A = \{\text{resultado par}\} = \{2,4,6\}$$

- Sea B el suceso que el resultado sea menor que 4:

$$B = \{1,2,3\}$$

De donde:

$$P(A) = \frac{\text{casos favorables}}{\text{casos posibles}} = \frac{3}{6} = \frac{1}{2}$$

$$P(B) = \frac{3}{6} = \frac{1}{2}$$

- $A \cap B = \{\text{el resultado sea par menor que } 4\}$

$A \cap B = \{2\}$

$P\{A \cap B\} = \frac{1}{6}$

- Piensa en la pregunta: "¿Dado que ha salido un número par en el lanzamiento de un dado qué probabilidad existe de que sea menor que 4?".

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{1/6}{1/2} = \frac{2}{6} = \frac{1}{3}$$

Observa que en este ejemplo:

$$P(B) = \frac{1}{2}, \qquad P(B/A) = \frac{1}{3}$$

de donde, $P(B) \neq P(B/A)$, los sucesos A y B son DEPENDIENTES.

Decimos que el suceso B depende del suceso A cuando se verifica:
$P(B/A) \neq P(B)$

SUCESOS INDEPENDIENTES:

Dados dos sucesos A y B de un experimento aleatorio decimos que son independientes si:

$$P(B/A) = P(B)$$

en otras palabras:

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = P(B)$$

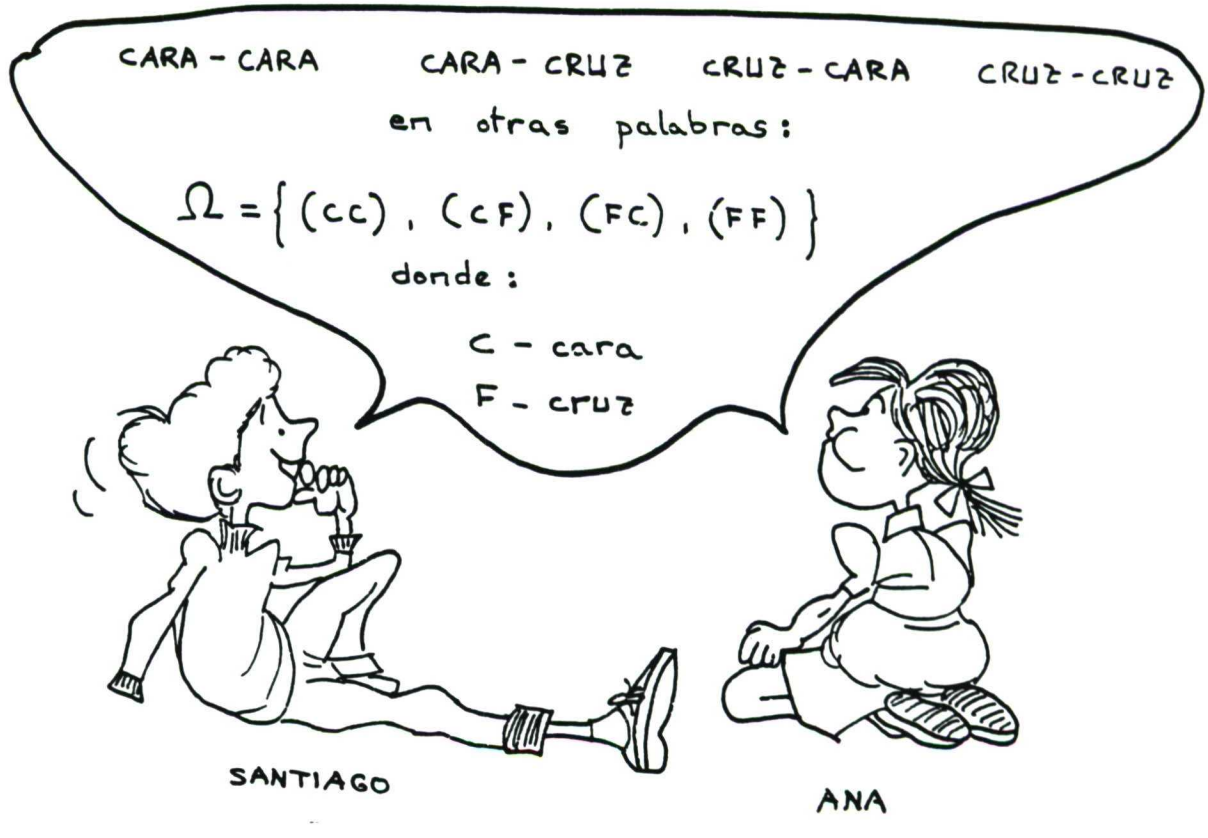
de donde

$$P(A \cap B) = P(A) \cdot P(B)$$

Es decir:

> Se dice que un suceso B es independiente de un suceso A cuando se verifica:
>
> $$P(A \cap B) = P(A) \cdot P(B)$$

PIENSA: El experimento aleatorio consiste en que Ana y Santiago lanzan dos monedas, el espacio muestral es:



Sea el espacio muestral:

$$\Omega = \{ (CC) , (CF) , (FC) , (FF) \}$$

con probabilidades iguales para cada uno de los resultados del espacio muestral:

$$P(CC) = P(CF) = P(FC) = P(FF) = \frac{1}{4}$$

Sean los sucesos:

$$A = \{ (CC), (FC) \} \qquad y \qquad B = \{ (FC), (FF) \}$$

- La probabilidad del suceso A es:

$$P(A) = P\{ (CC), (FC) \} = P\{ (CC) \} + P\{ (FC) \} = \frac{1}{4} + \frac{1}{4} = \frac{2}{4} = \frac{1}{2}$$

- La probabilidad del suceso B es:

$$P(B) = P\{ (FC), (FF) \} = P\{ (FC) \} + P\{ (FF) \} = \frac{1}{4} + \frac{1}{4} = \frac{2}{4} = \frac{1}{2}$$

Entonces:

$$\left.\begin{array}{l} P(A) = \frac{1}{2} \\ P(B) = \frac{1}{2} \end{array}\right\} \quad P(A) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

- El suceso $A \cap B$ ocurre siempre que ocurra A y ocurra B simultáneamente.

siendo

$$A = \{ (CC), (FC) \} \qquad y \qquad B = \{ (FC), (FF) \}$$

se tiene:

$$A \cap B = \{ (FC) \}$$

de donde

$$P(A \cap B) = P\{ (FC) \} = \frac{1}{4} = P(A) \cdot P(B) = \frac{1}{4}$$

Luego:

> Los sucesos A y B son independientes, pues verifican la expresión:
>
> $$P(A \cap B) = P(A) \cdot P(B)$$

TEOREMA DE BAYES:

Sean A_1 y A_2 dos sucesos incompatibles de un espacio muestral Ω, es decir:

$A_1 \cup A_2 = \Omega$ "suceso seguro"

$A_1 \cap A_2 = \emptyset$ "suceso imposible"

" Estamos interesados en conocer la probabilidad de que ocurrido el suceso B la causa que lo haya producido sea el suceso A_1 "

Esto es, queremos calcular:

$$P(A_1/B)$$

El teorema de Bayes para dos sucesos establece:

$$P(A_1/B) = \frac{P(A_1) \cdot P(B/A_1)}{P(A_1) \cdot P(B/A_1) + P(A_2) \cdot P(B/A_2)}$$

ACTIVIDAD - 1: La tabla de mortalidad muestra el número de supervivientes a varias edades, desde el nacimiento, de un grupo de mil varones de un pueblo de Segovia.

TABLA DE MORTALIDAD DE FUENTERREBOLLO

Edad	Supervivientes
0	1000
10	996
20	990
30	910
40	880
50	790
60	405
70	60
80	15
90	5
100	1

Se desea saber:

1) La probabilidad de que un recién nacido alcance los cincuenta años de edad.

2) La probabilidad de que una persona de treinta años viva hasta los sesenta.

3) La probabilidad de que una persona de cincuenta años viva hasta los ochenta.

COMPRUEBA

1. P(50 años)= 0'79
2. P(60 años/30 años)= 0'445
3. P(80 años/50 años)= 0'0189

ACTIVIDAD - 2: Un padre ha estimado, por experiencia de años anteriores, que la probabilidad de que su hijo falte al Instituto es de 0'2. También ha estimado que el 30 por 100 de los días que no asiste son dedicados al deporte y el 70 por 100 de los días de clase se dedican al deporte.

Se desea saber:

1) Probabilidad de que un día que se dedica al deporte no fuera al Instituto.

2) Probabilidad de que un día no dedicado al deporte asiste al Instituto.

COMPRUEBA

1. $P(A/B)= 0'096$
2. $P(\overline{A}/\overline{B})= 0'631$

A = suceso falta al Instituto

B = suceso lo dedica al deporte

ACTIVIDAD – 3: Sean dos sucesos A y B tales que:

$$P(A/B)= \frac{1}{5}\ , \qquad P(\bar{B})= \frac{1}{3}\ , \qquad P(B/A)= \frac{1}{4}$$

Se pide encontrar P(A) y P(B).

ESTADISTICA DESCRIPTIVA

ACTIVIDADES

Después de ser despedido del Instituto Cantalejo, Raúl estaba muy preocupado; ¡Tenía que encontrar trabajo, y eso en el pueblo estaba muy difícil!.

A los pocos días de saber que su amigo Mateo iba a ser padre se le ocurrió una graciosa idea que enseguida propuso a su mujer:

RAUL:

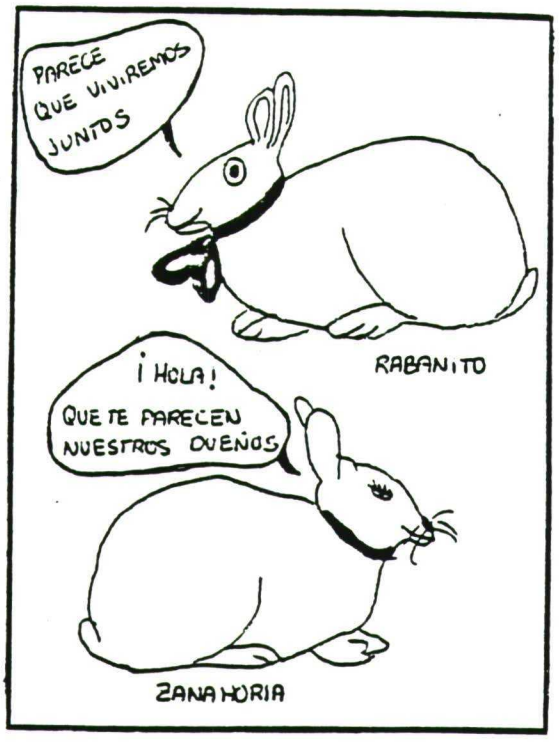
Chica..., he pensado, ... que como las condiciones del pueblo son favorables, y puesto que yo no encuentro trabajo en ningún sitio ... debería poner mi propio negocio, así que ...

CHICA:

No entiendo nada, qué negocio puedes poner, si no sabes hacer nada.

RAUL:

Para criar conejos no hay que saber demasiadas cosas. Eso es lo que quiero hacer, poner una granja de conejos.



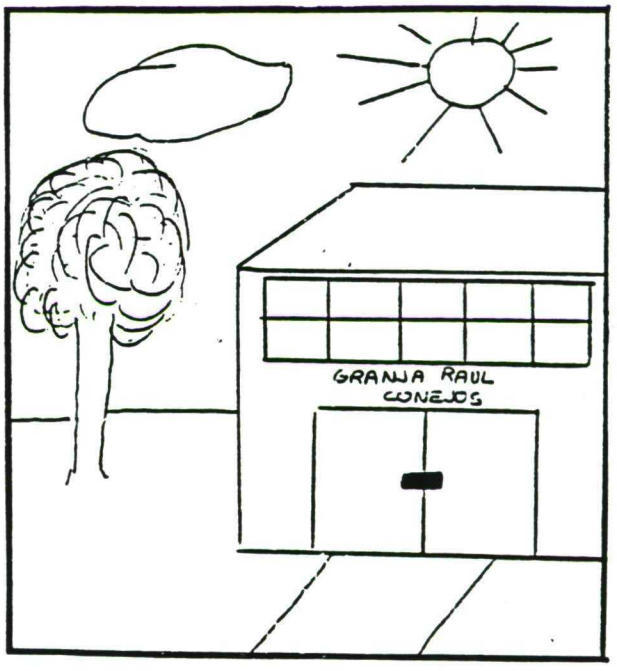
La mujer de Raúl era muy decidida, - además pensó que teniendo a Raúl en casa podría controlarlo mejor. Dicho y hecho, se marchó a comprar conejos.

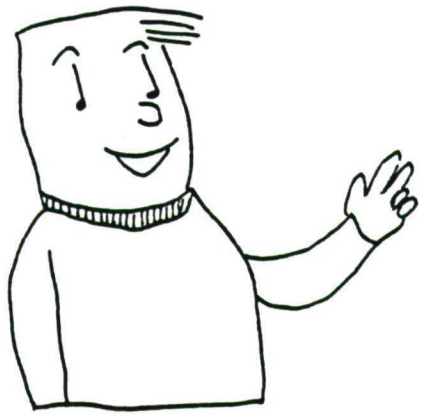
CHICHA:

El primer conejo macho y la primera hembra serán los fundadores de - nuestro negocio, los llamaremos Rabanito y Zanahoria.

A los pocos meses el negocio estaba en marcha, Raúl y Chicha tenían un montón de conejos.

Al poco tiempo, Raúl se entero de que sus fundadores habían vuelto a tener crías. Entusiasmado fue - a contarselo a su mujer.

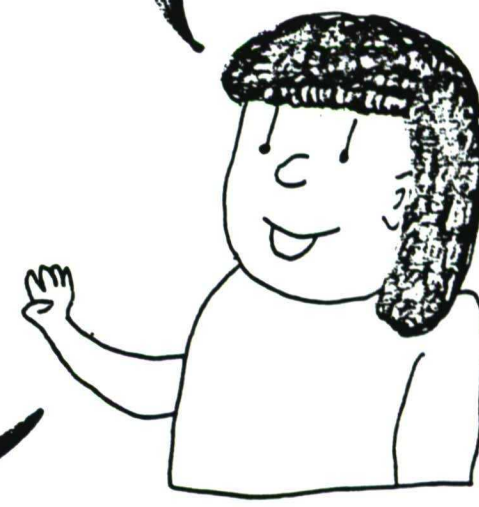




ZANA Y RABANITO HA TENIDO 4 CRÍAS ¿CUÁNTAS SERÁN MACHOS Y CUÁNTAS HEMBRAS?

POSIBLEMENTE SEAN 2 MACHOS Y 2 HEMBRAS

NO LO SE. ¿TU QUE CREES? ¿SERÁN TODOS DEL MISMO SEXO?



POR LO QUE HE VISTO HASTA AHORA, CREO QUE ESO NO ES ASI

Bueno, la cuestión no era muy importante, pero a Raúl le había picado la curiosidad.

RAUL:

No es tan difícil hacer una previsión lógica. Que una cría sea macho es tan posible como que sea hembra, por lo tanto, la mitad de los conejos serán machos y la otra mitad hembras. Es decir, Zana y Rabanito han tenido dos conejitos y dos conejitas.

Por lo que Chicha había observado la suposición de Raúl no la convenció demasiado.

Así Chicha se puso a pensar en las posibilidades que había en cuanto al sexo de las cuatro crías.

El problema no estaba muy claro, por lo tanto pensó en buscar ayuda. Se fue a la biblioteca del Instituto y allí buscó un libro que hablase de como se combinaban diferentes elementos. El libro decía:

ANALISIS COMBINATORIO

Las combinaciones nos resuelven problemas en los que tenemos que seleccionar agrupaciones de elementos escogidos de entre un número de ellos, con la condición de que dos agrupaciones serán distintas si, y sólo si, están formadas por elementos distintos, en otras palabras, no se tiene en cuenta el orden de los elementos en la agrupación.

Así:

$$C_{4,0} = \binom{4}{0} = \frac{4!}{0!\ (4-0)!} = 1$$

representa el número de grupos de 4 elementos que podemos formar en los que, en nuestro caso, no aparece ningún macho.
Se llama "número combinatorio 4 sobre 0".

Análogamente:

$$C_{4,1} = \binom{4}{1} = \frac{4!}{1!\ 3!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{1\ (3 \cdot 2 \cdot 1)} = 4$$

por tanto, son 4 casos en los que hay un solo macho y, en consecuencia, tres hembras.

de esta forma:

$$C_{4,2} = \binom{4}{2} = \frac{4!}{2!\ 2!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{(2.1)\ (2.1)} = 6$$

hay 6 casos en los que en un grupo de 4 aparecen 2 machos.

$$C_{4,3} = \binom{4}{3} = \frac{4!}{3!\ 1!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{(3 \cdot 2 \cdot 1)\ 1} = 4$$

en 4 ocasiones aparecen 3 machos y 1 sola hembra.

Y por último:

$$C_{4,4} = \binom{4}{4} = \frac{4!}{4!\ 0!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{(4.3.2.1).1} = 1$$

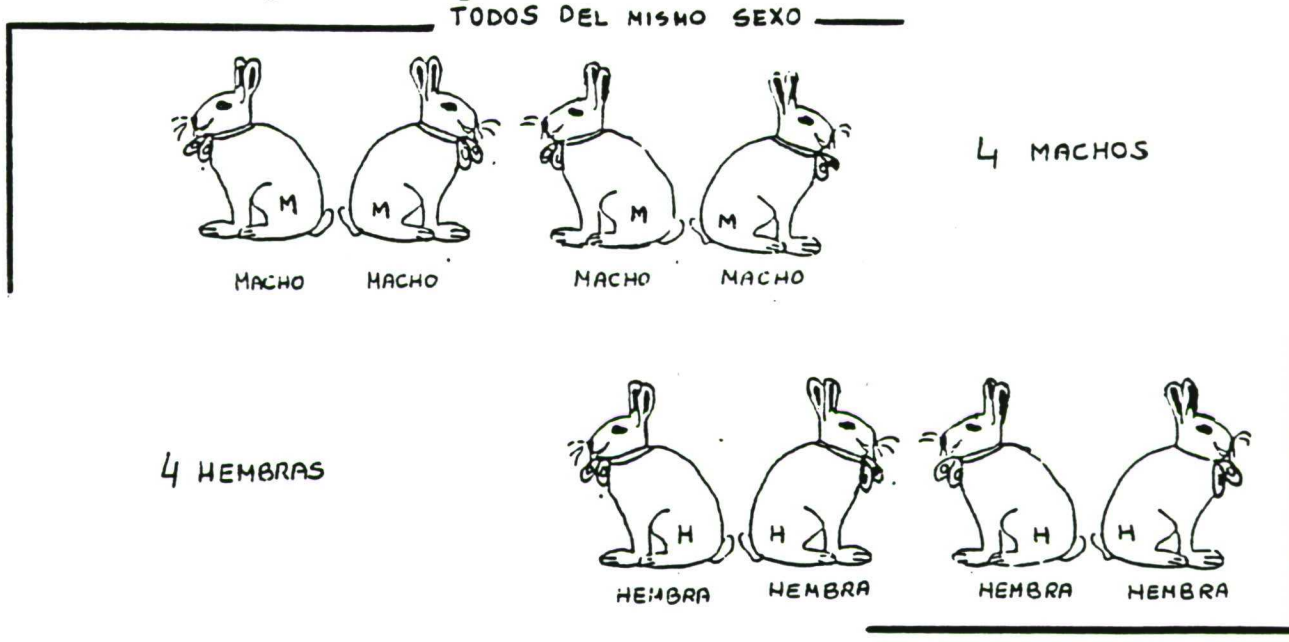
en un solo caso, aparecen 4 machos en un grupo de 4.

El número total de posibles combinaciones era:

$$C_{4,0} + C_{4,1} + C_{4,2} + C_{4,3} + C_{4,4} = 1 + 4 + 6 + 4 + 1 = 16$$

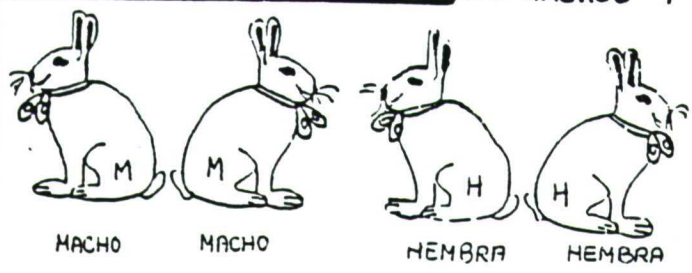
Chica pensó que para comprobarlo podía construir todas las combinaciones.

Había dos casos en los que las cuatro crías eran del mismo sexo:



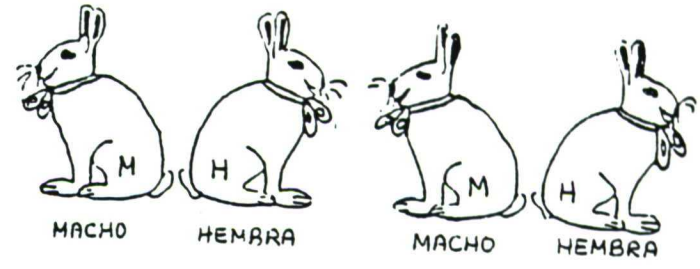
Se presentaba el caso de dos machos y dos hembras en un total de 6 casos de los 16 posibles.

DOS MACHOS Y DOS HEMBRAS

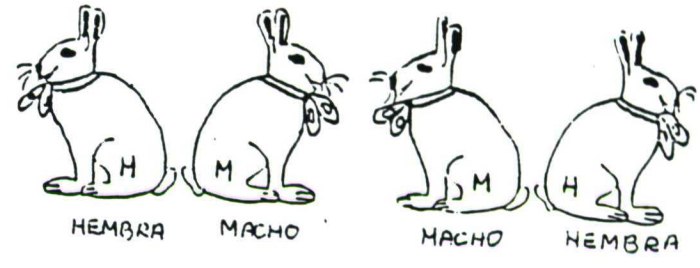


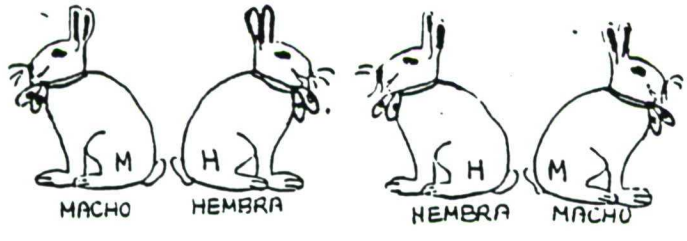
2 MACHOS Y 2 HEMBRAS

2 MACHOS Y 2 HEMBRAS

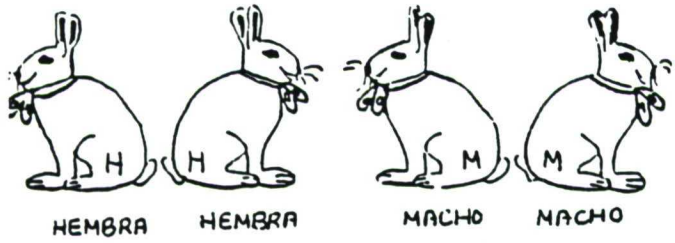


2 MACHOS Y 2 HEMBRAS

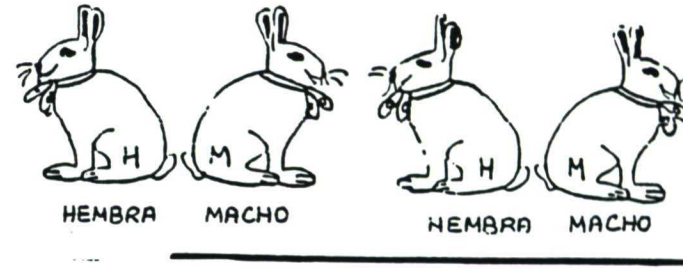




2 MACHOS Y 2 HEMBRAS

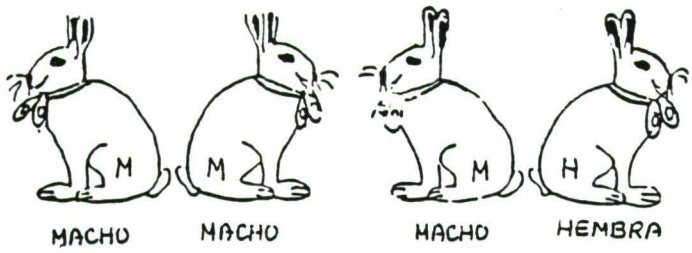


2 MACHOS Y 2 HEMBRAS



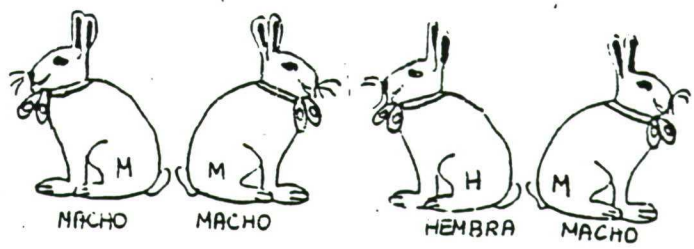
Por último, 8 veces de las 16 se daba el caso de ser 3 de un sexo y 1 del otro, es decir 3 machos y 1 hembra ó bien 1 macho y 3 hembras.

TRES MACHOS UNA HEMBRA

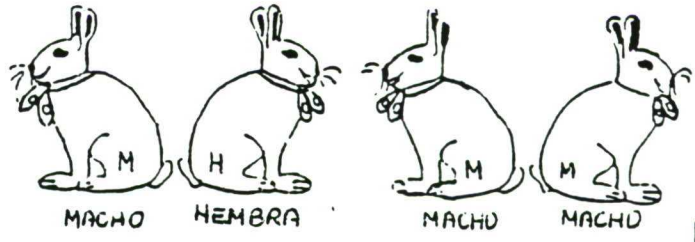


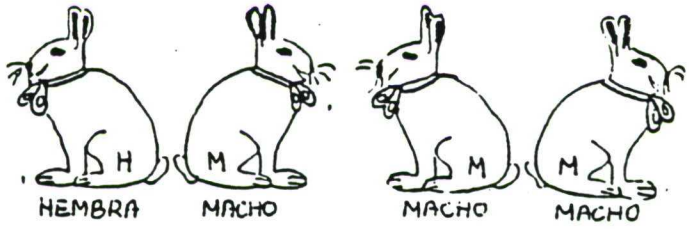
3 MACHOS Y 1 HEMBRA

3 MACHOS Y 1 HEMBRA



3 MACHOS Y 1 HEMBRA

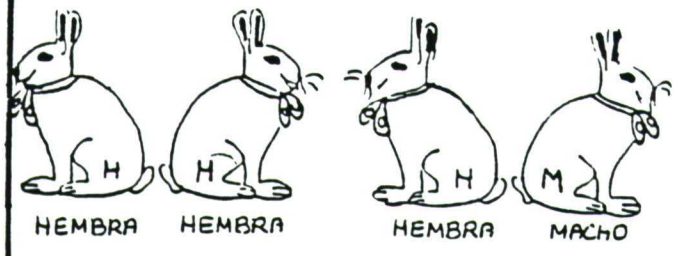




3 MACHOS Y 1 HEMBRA

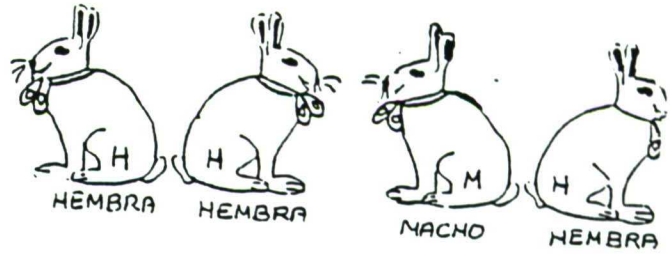
Análogamente:

TRES HEMBRAS Y UN MACHO

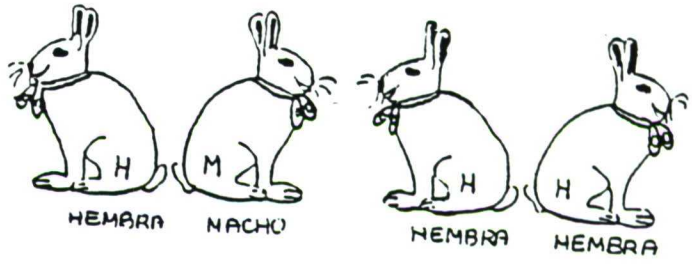


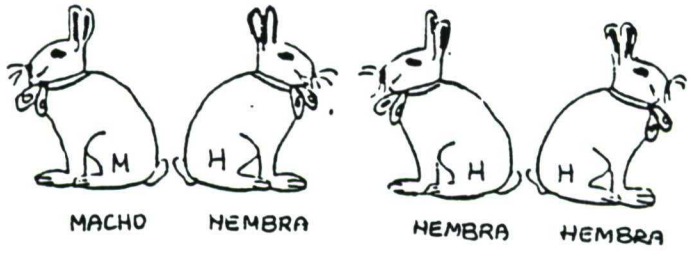
3 HEMBRAS Y 1 MACHO

3 HEMBRAS Y 1 MACHO



3 HEMBRAS Y 1 MACHO





3 HEMBRAS Y 1 MACHO

Analizando todos los casos resultaban 16.

Chicha pensó que habían estado en lo cierto al pensar que era muy difícil que las cuatro crías tuvieran el mismo sexo. Sin embargo Raúl se había equivocado al pensar que lo más lógico era que Zana y Rabanito hubiesen tenido 2 conejitos y 2 conejitas. ¡Iría a contárselo rápidamente!.

CHICHA:

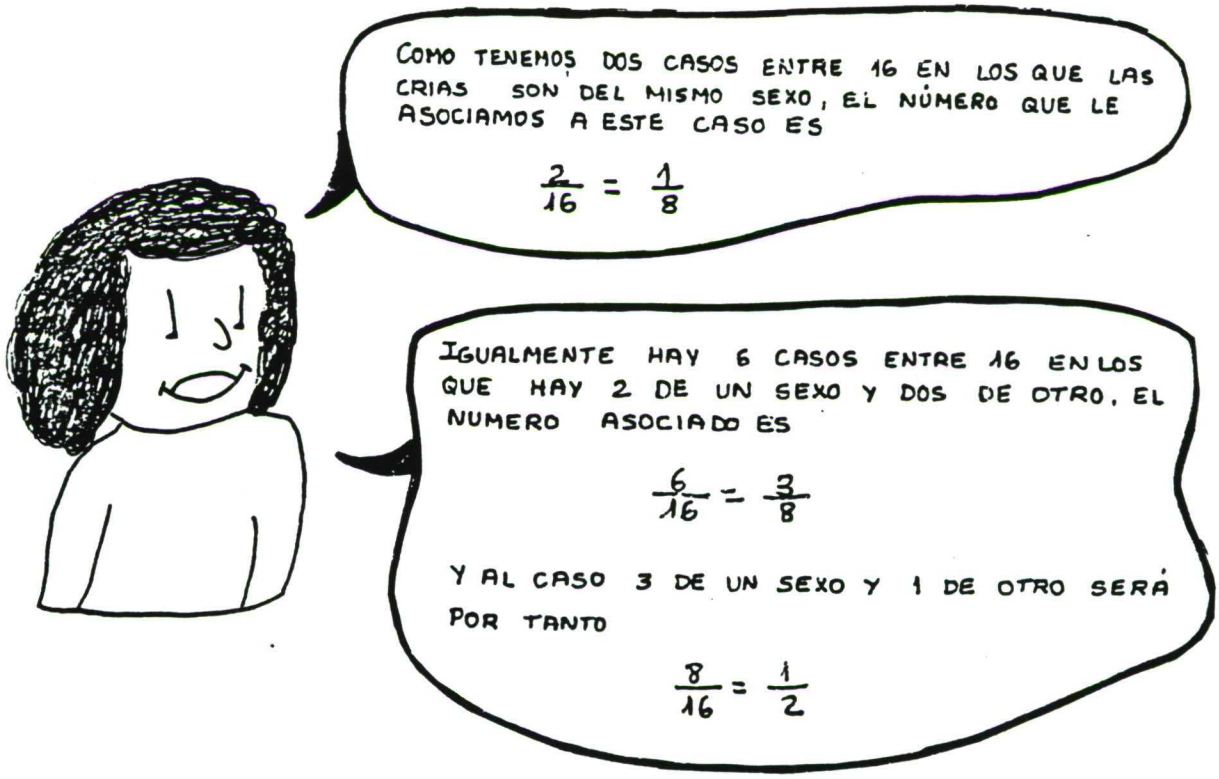
Raúl estabas equivocado. El caso más verosímil es que los conejos sean 3 de un sexo y 1 de otro.

RAUL:

¡No lo entiendo, Chicha, explícamelo!.

CHICHA:

Es sencillo una vez que estudias todas las posibilidades, que son 16, de las cuales 6 veces se presenta la forma dos de un sexo y dos de otro, mientras que tres de un sexo y uno de otro sexo se da en 8 casos. Si todavía esto no te aclara la cuestión, vamos a intentar comparar los resultados. La forma de hacerlo es asociar a cada uno de los casos un número y luego comparar estos números entre sí.

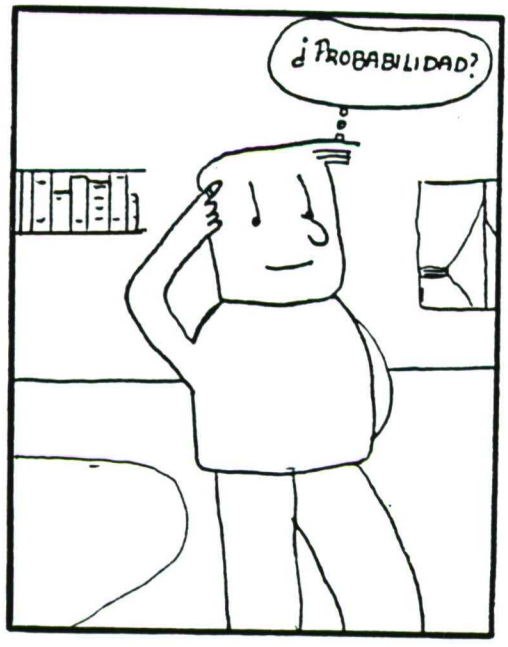


RAUL:

Y ese número, ¿se llama de alguna forma?.

CHICHA:

Sí, me he informado en la biblicteca del Instituto. Se llama "probabilidad".



RAUL:

Luego, ¿puedo decir que la probabilidad de que Zana y Rabanito tengan las cuatro crías del mismo sexo es $\frac{1}{8}$?

CHICHA:

Eso es, fíjate además que la probabilidad de que sean 2 machos y 2 hembras es $\frac{3}{8}$, mientras que el que sean 3 de un sexo y 1 de otro tiene probabilidad $\frac{1}{2}$. Como $\frac{3}{8}$ es menor que $\frac{1}{2}$ ($\frac{1}{2} = \frac{4}{8}$), es más probable que los conejitos sean 3 de un sexo y 1 de otro.

RAUL:

¿Eso quiere decir que podemos asegurar que Zana y Rabanito tendrán siempre 3 conejos de un sexo y 1 de otro?.

CHICHA:

No hombre, no. Quiere decir que es más posible el caso "3 y 1" que el caso "2 y 2", y que el caso "todos del mismo sexo" es el más raro. Pero puede ocurrir cualquiera de ellos.

RAUL:

Total que no podemos asegurar nada.

CHICHA:

Exactamente. Lo único seguro es que han nacido 4 conejitos y que estos van a tener un sexo, además la combinación de sexos que se dé tiene que

ser necesariamente una de las 16 que te cité antes.

Al poco tiempo Raúl se enteró que otra de sus parejas de conejos había tenido 5 crías y comenzó a pensar en los diferentes casos que existían. ¿Podrías hacer tú lo mismo?.

ANALISIS COMBINATORIO:

Es una herramienta indispensable en el estudio de los experimentos aleatorios con espacio muestral asociado Ω.

Muchos experimentos aleatorios corresponden a la idea de que los resultados son EQUIPROBABLES, es decir, cada resultado tiene la misma probabilidad de ocurrir. "Pensemos en el lanzamiento de un dado":

Hay seis casos que pueden presentarse, que son los números 1, 2, 3, 4, 5 ó 6.

Por tanto:

Ω = 1, 2, 3, 4, 5, 6 es el espacio muestral.

Podemos suponer que estos seis casos son igualmente posibles, en otras palabras:

$$P\{1\} = P\{2\} = P\{3\} = P\{4\} = P\{5\} = P\{6\} = \frac{1}{6}$$

esto es, la probabilidad de que salga cualquiera de los seis números $\{1, 2, 3, 4, 5, 6\}$ es:

$$p = \frac{\text{casos favorables}}{\text{casos posibles}}$$

$$p = P\{1\} = \frac{1}{6}\ ,\ \ldots\ ,\ p = P\{5\} = \frac{1}{6}\ ;\ \ldots$$

Imagínate un experimento aleatorio en que los resultados son EQUIPROBABLES y en donde el espacio muestral Ω está formado por "n" sucesos, tendríamos:

$$\Omega = \{a_1, a_2, a_3, \ldots, a_n\}$$

donde la probabilidad de cada suceso a_i es precisamente $\frac{1}{n}$, siendo n el número total de sucesos.

Piensa que para representar todos los sucesos $\{a_1, a_2, a_3, \ldots, a_n\}$, escribimos a_i donde i puede ser cualquiera de los números 1, 2, 3, ..., n; de tal forma:

$a_i = a_1$ cuando $i = 1$

$a_i = a_2$ cuando $i = 2$

$a_i = a_3$ cuando $i = 3$

.....

.....

$a_i = a_n$ cuando $i = n$

De donde:

El símbolo a_i denota cualquiera de los n valores $\{a_1, a_2, a_3, \ldots, a_n\}$. La letra i en a_i se llama "subíndice". Análogamente puede utilizarse como subíndice cualquiera otra letra distinta de i, como puede ser j, k, m, p, ...

Para estos experimentos aleatorios es esencial analizar las probabilidades de los sucesos.

Para facilitar esta tarea es preciso echar mano del "análisis combinatorio".

PRINCIPIO FUNDAMENTAL:

Si un suceso puede presentarse con cualquiera de n formas distintas y si cuando esto ha sucedido otro suceso puede presentarse con cualquiera de m formas distintas, entonces el número de formas en que ambos sucesos pueden presentarse en el orden dado es n.m.

Piensa en la pregunta: "¿De cuántas formas pueden ocuparse los cargos de concejal y alcalde si hay 4 candidatos para concejal y 2 para alcalde?".

Los dos cargos pueden ocuparse de:

$$4 \cdot 2 = 8 \text{ formas}$$

FACTORIAL DE n:

El factorial de n se escribe por n! y viene definido por:

$$n! = n(n-1)\ (n-2)\ (n-3) \ldots 4 \cdot 3 \cdot 2 \cdot 1$$

Piensa:

$$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$$

$$3! = 3 \cdot 2 \cdot 1 = 6$$

$$2! = 2 \cdot 1 = 2$$

$$1! = 1 = 1$$

Piensa ahora: 4! 3! / 6! 0!

$$4! \cdot 3! = (4 \cdot 3 \cdot 2 \cdot 1)\ (3 \cdot 2 \cdot 1) = 24 \cdot 6 = 144$$

$$6! \cdot 0! = (6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1)\ (0) = 720 \cdot 1 = 720$$

$$\boxed{0! = 1}$$

COMBINACIONES DE m ELEMENTOS TOMADOS DE n EN n:

Una combinación de m elementos diferentes tomados de n en n es una "selección de n de los m elementos sin atender a la ordenación de los mismos".

El número de combinaciones de m elementos tomados de n en n se representa por $C_{m,n}$ ó por la notación $\binom{m}{n}$ debida al gran matemático suizo Euler. Sea:

$$C_{m,n} = \binom{m}{n} = \frac{m!}{n!\,(m-n)!}$$

Piensa en los ejemplos: $C_{3,2}$ y $C_{5,3}$

$$C_{3,2} = \binom{3}{2} = \frac{3!}{2!\,(3-2)!} = \frac{3!}{2!\,1!} = \frac{3 \cdot 2 \cdot 1}{(2 \cdot 1) \cdot 1} = \frac{6}{2} = 3$$

$$C_{5,3} = \binom{5}{3} = \frac{5!}{3!\,(5-3)!} = \frac{5!}{3!\,2!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(3 \cdot 2 \cdot 1)\,(2 \cdot 1)} = \frac{120}{12} = 10$$

Piensa en la pregunta: "¿Cuál es el número de combinaciones de las letras a, b, c tomadas de dos en dos?".

Como a, b, c son tres letras, el número de combinaciones que pueden formarse con ellas tomadas de dos en dos vendrá dado por la fórmula:

$$C_{3,2} = \binom{3}{2} = \frac{3!}{2!\,(3-2)!} = \frac{3 \cdot 2 \cdot 1}{(2 \cdot 1) \cdot 1} = \frac{6}{2} = 3$$

Estas son:

ab ac bc

Observa que ab es la misma combinación que ba

ac es la misma combinación que ca

cb es la misma combinación que bc

¡¡ ya sabemos que en las combinaciones no cuenta el orden de colocación de los elementos !!

VARIACIONES DE m ELEMENTOS TOMADOS DE n EN n:

Una variación de m elementos diferentes tomados de n en n es una "selección" de n elementos de entre los m tomados, con la condición de que se tiene en cuenta el orden de los elementos.

Se representa por:

$$V_{m,n} = \frac{m(m-1)\ (m-2)\ \ldots\ (m-n+1)}{n \text{ elementos}}$$

Piensa en los ejemplos: $V_{5,3}$ y $V_{6,2}$

$$V_{5,3} = 5 \cdot 4 \cdot 3 = 60$$

$$V_{6,2} = \frac{6 \cdot 5}{2} = 30$$

Piensa en la pregunta: "¿Cuántos números de dos cifras distintas se pueden formar con los números 1, 2, 3?".

Es el número de selecciones que se pueden formar tomando dos elementos de un conjunto de tres elementos. Como el orden de los números da lugar a distintas cifras, se trata de variaciones de tres elementos tomados de dos en dos. Sea:

$V_{3,2} = 3 \cdot 2 = 6$ números distintos.

Los números son:

12, 13, 23, 21, 31, 32

PERMUTACIONES DE m ELEMENTOS:

Una permutación de m elementos es un caso particular de una variación de m elementos tomados de m en m.

Se representa por:

$$P_m = V_{m,m} = m(m-1)(m-2)\ \ldots\ (m-n+1) = m(m-1)(m-2)\ \ldots\ 1 = m!$$

$$\boxed{P_m = m!}$$

Piensa que como una permutación es un caso particular se tiene en cuenta el orden de los elementos.

Veamos un ejemplo: P_7 , P_3 ,

$P_7 = 7! = 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot \ \cdot 1 = 5040$

$P_3 = 3! = 3 \cdot 2 \cdot 1 = 6$

$P_0 = 0! = 1$

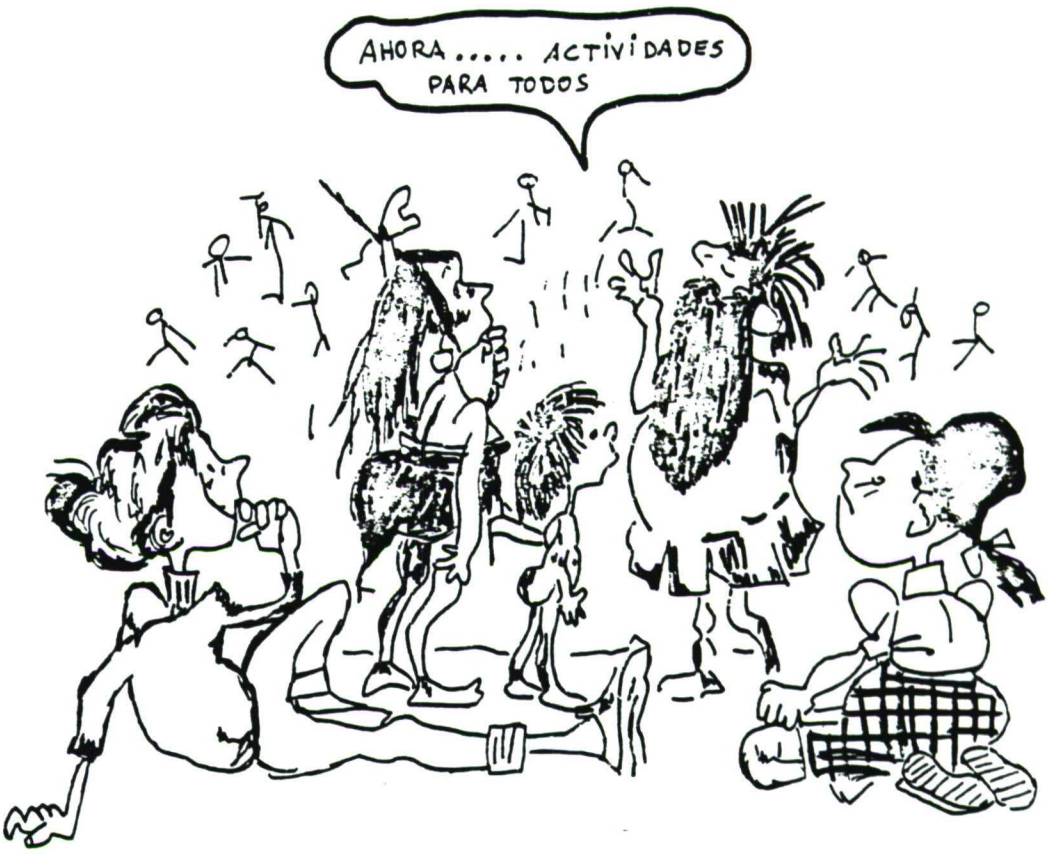
Piensa en la pregunta: "¿De cuánta formas pueden sentarse Pedro, Juan y Ana en una fila?".

Es el número de colocaciones que p den formarse tomando tres elementos de un conjunto de tres, pero en las q el orden de las colocaciones son - distintas cuando el orden de las pers as sea distinto. Se trata de permutaciones de tres elementos:

$P_3 = 3! = 3 \cdot 2 \cdot 1 = 6$ formas istintas

Las formas de colocarse en fila so

Pedro - Juan - Ana	Ped	- Ana - Juan	Juan - Ana - Pedro
Juan - Pedro - Ana	Ana	Pedro - Juan	Ana - Juan - Pedro



ACTIVIDAD – 1: ¿De cuántas formas puede elegirse un comité de 5 personas de entre 10 personas?

Es el número de disposiciones (selecciones) que se pueden formar tomando 5 personas de un conjunto de 10 personas. Como el orden dentro de la disposición es indiferente, resulta que se trata de combinaciones de 10 elementos tomados de 5 en 5. Sea:

$$C_{10,5} = \binom{10}{5} = \frac{10!}{5!\,(10-5)!} = \frac{10!}{5!\,5!} =$$

$$= \frac{10.9.8.7.6.5!}{(5.4.3.2.1).5!} = \frac{10.9.8.7.6}{5.4.3.2.1} =$$

$$= \frac{30240}{120} = 252 \text{ formas distintas}$$

ACTIVIDAD - 2: De un total de 5 chicos y 7 chicas, se forma un comité de 2 chicos y 3 chicas. ¿De cuántas formas puede formarse si puede pertenecer a él cualquier chico o chica?

El número de selecciones que se pueden formar es:

- 2 chicos de un total de 5. Como el orden de la selección no cuenta, resulta que se trata de combinaciones de 5 chicos tomados de 2 en 2. De donde:

$$C_{5,2} = \binom{5}{2} = \frac{5!}{2!\ (5-2)!} = \frac{5!}{2!\ 3!} =$$

$$= \frac{5.4.3!}{2!\ \ 3!} = \frac{5.4}{2!} = \frac{5.4}{2.1} = 10 \text{ formas}$$

- 3 chicas de un total de 7, donde el orden de la selección no cuenta. Por tanto:

$$C_{7,3} = \binom{7}{3} = \frac{7!}{3!\ (7-3)!} = \frac{7!}{3!\ 4!} =$$

$$= \frac{7.6.5.4!}{(3.2.1)!} = \frac{7.6.5}{3.2.1} = 35 \text{ formas}$$

- El número pedido es:

$$C_{5,2} \cdot C_{7,3} = 10 \cdot 35 = 350$$ formas diferentes se puede formar un comité de 2 chicos y 3 chicas.

ACTIVIDAD - 3: De un grupo de once personas (6 hombres y 5 mujeres) se quiere formar un tribunal de selectividad que conste de 3 hombres y 2 mujeres. ¿De cuántas formas puede formarse el tribunal si dos hombres determinados no pueden estar en el tribunal?.

Se trata de averiguar el número de selecciones que pueden hacerse teniendo presente que:

- no cuenta el orden dentro de la selección,
- dos hombres determinados no pueden estar en el tribunal.

Como el grupo está formado por 6 hombres y 2 hombres no pueden estar en el tribunal, se tiene:

3 hombres de un total de 4 que pueden elegirse:

$$C_{4,3} = \binom{4}{3} = \frac{4!}{3!\ (4-3)!} = \frac{4!}{3!\ 1!} =$$

$$= \frac{4.3.2.1}{(3.2.1).1} = 4 \text{ formas de elegir}$$

2 mujeres de un total de 5:

$$C_{5,2} = \binom{5}{2} = \frac{5!}{2!\ (5-2)!} = \frac{5!}{2!\ 3!} = 10 \text{ formas de elegir mujeres}$$

El tribunal compuesto de 3 hombres y 2 mujeres puede elegirse:

$$C_{4,3} \cdot C_{5,2} = 4 \cdot 10 = 40 \text{ formas diferentes.}$$

ACTIVIDAD - 4: ¿De cuántas maneras se pueden extraer cuatro cartas de una baraja española de 40 cartas?.

Es el número de disposiciones que se pueden formar tomando cuatro elementos de un conjunto de cuarenta. Como el orden dentro de la disposición es indiferente, se trata de combinaciones de cuarenta tomadas de 4 en 4, sea:

$$C_{40,4} = \binom{40}{4} = \frac{40!}{4!\ 36!} = \frac{40.39.38.37.36!}{(4.3.2.1)\ .\ 36!} =$$

$$= \frac{40.39.38.37}{4.3.2.1} = 91390 \text{ maneras.}$$

ACTIVIDAD - 5: ¿De cuántas maneras pueden ser colocadas en una fila 5 bolas de diferentes colores?.

Es el conjunto de todas las elecciones distintas que se pueden formar tomando 5 elementos de entre 5 elementos, con la condición de que se tiene en cuenta el orden de los elementos.

De donde:

$$V_{5,5} = \frac{5.4.3.2.1}{5} = 120 \text{ maneras}$$

Recuerda: $P_5 = 5! = 5.4.3.2.1 = V_{5,5}$

Como se trata de elegir 5 bolas entre 5, se tiene $V_{5,5}$. Este caso particular de variaciones recibe el nombre de permutaciones de 5 elementos.

ESTADISTICA DESCRIPTIVA

ACTIVIDADES

"ASIGNAR PROBABILIDADES"

El profesor de Matemáticas todos los días tiene que desplazarse desde Plaza Castilla hasta el CEP LATINA; después de probar varios itinerarios, encuentra que dos son los mejores: uno es atravesando Atocha, y el otro es ir por la M-30.

A primeras horas de la mañana, por Atocha, sólo tarda 25 minutos, pero cuando viaja al mediodía, la duración es de 50 minutos. En caso de que decida ir por la M-30 el tráfico es siempre uniforme y la duración es de 35 minutos.

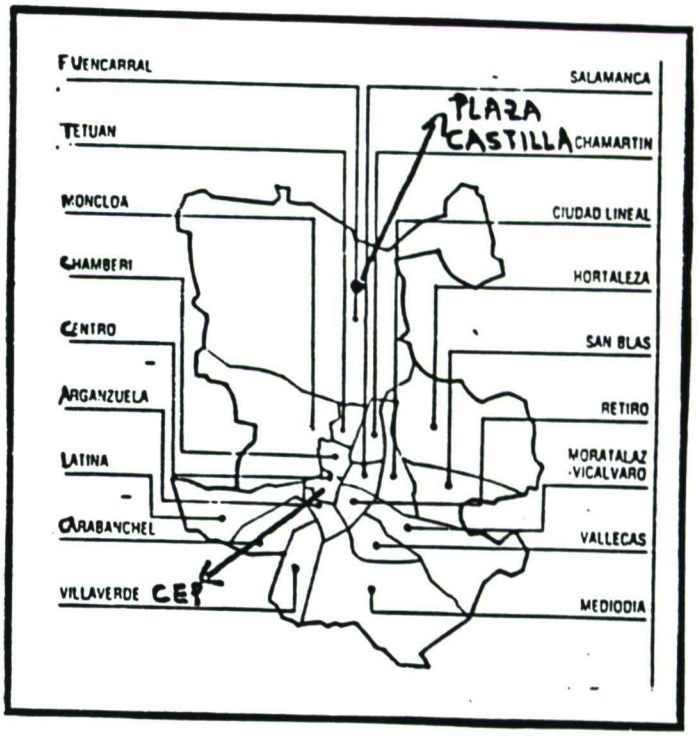
¿QUE CAMINO DEBE ELEGIR?

El profesor que tiene que desplazarse es el "decisor".

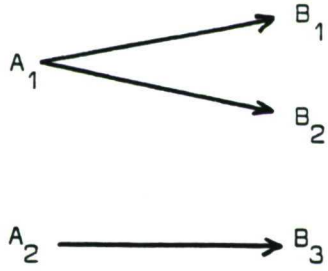
Se le presentan dos alternativas o acciones:

A_1= ir por Atocha.

A_2= ir por la M-30



ALTERNATIVAS RESULTADOS



• Si se decide por la alternativa A_1= ir por Atocha, se presentan dos consecuencias o resultados:

A_1 = ir por Atocha
- B_1/A_1 = tráfico fluido mañana (25 minutos)
- B_2/A_1 = tráfico denso mediodía (50 minutos)

• Si decide la alternativa A_2= ir por la M-30, la consecuencia que se presenta es:

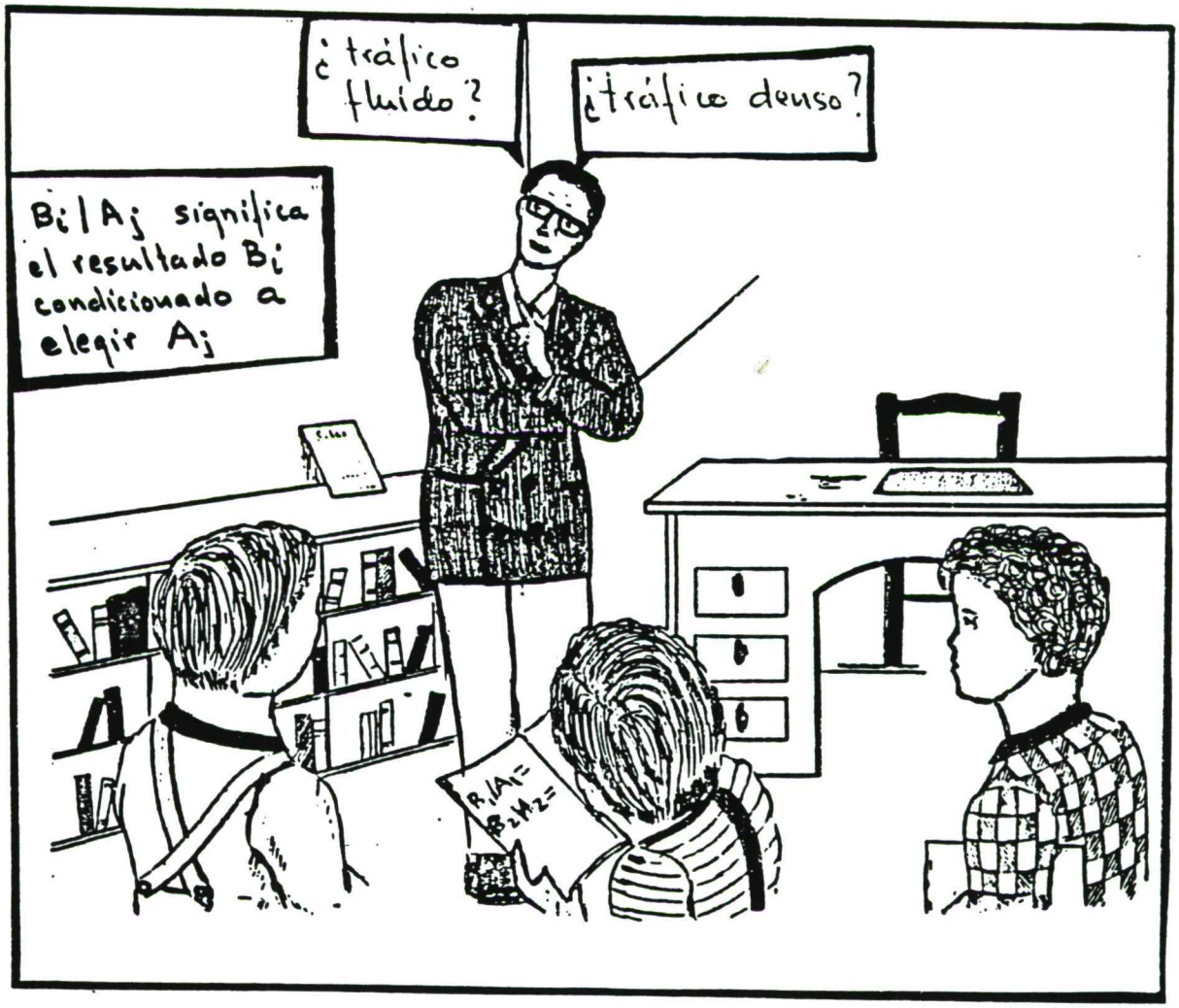
A_2 = ir por la M-30 $\longrightarrow$ B_3/A_2 = tráfico uniforme (35 minutos)

De donde, la valoración de las consecuencias es:

$V(B_1/A_1)$ = 25 minutos

$V(B_2/A_1)$ = 50 minutos

$V(B_3/A_2)$ = 35 minutos



Hoy es un día anormal, el profesor se ha visto obligado a desplazarse al CEP LATINA a las 11 de la mañana.

Si él supiese cómo va a estar el tráfico, entonces el ambiente o contexto sería de CERTIDUMBRE y razonaría:

- Si sé que el tráfico es fluido elijo ir por Atocha.

- Si sé que el tráfico es denso elijo ir por la M-30.

Seguramente, desconoce cómo se encuentra el tráfico, pero debido a la rutina de todos los días, conoce que el tráfico está fluido con una probabilidad relativa $p = 0'80$ y que está denso con una probabilidad $1 - p$, es decir, $0'20$.

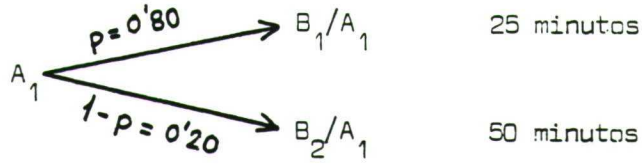
En este caso, la experiencia que tiene le conduce a poder afirmar que el tráfico es fluido en un 80 % y es denso en un 20 %. El profesor se encuentra en un ambiente de RIESGO y razona así:

- Si $p > \frac{1}{2}$ elijo ir por Atocha.

- Si $p \leq \frac{1}{2}$ elijo ir por la M-30.

Como $p = 0'80 > \frac{1}{2}$ voy por Atocha.

Otro criterio que puede seguir en el ambiente de RIESGO es calcular la esperanza matemática de las dos alternativas y compararlas para elegir la óptima. Veamos:

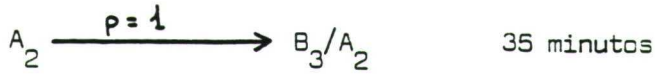


A_1 = (Tarda 25 minutos con probabilidad p = 0'80 ó bien tarda 50 minutos con probabilidad 1 − p = 0'20)

La esperanza matemática de la alternativa A_1 es:

$$25p + 50(1 - p) = 25 \cdot 0'8 + 50(1 - 0'80) = 25 \cdot 0'8 + 50 \cdot 0'20 =$$

$$= 30 \text{ minutos}$$



A_2 = (Tarda 35 minutos con probabilidad 1 "certeza").

La esperanza matemática de la alternativa A_2 es:

$$35 \cdot 1 = 35$$

Realmente, $25p + 50(1 - p) < 35$, entonces elige ir por Atocha.

Hoy, además de ser un día "anormal" es un mal día, el profesor de Matemáticas creía que la experiencia en el viajecito le permitía poder asignar una probabilidad determinada a cada alternativa, A_1 y A_2, pero la realidad le ha demostrado que estaba equivocado. ¡Los viajes durante un año no ofrecían información suficiente!.

En este caso el profesor no sabe asignar una probabilidad a la alternativa o acción A_1 (ir por Atocha), se mueve en un ambiente de INCERTIDUMBRE, procede a razonar así:

- Si las cosas van bien, es más divertido ir por la ciudad, luego me quedo con la alternativa A_1.

- Por otro lado, si las cosas van mal, mejor voy por la M-30 porque puede ocurrir que el tráfico sea denso y llegaría muy tarde.

BIBLIOGRAFIA

WAYNE W., DANIEL : Estadística con aolicaciones a las ciencias sociales. MacGraw Hill. México, 1981.

SEGANE, J.: Psicología matemática I. UNED.

TEJEDOR, F.J.: Estadística inferencial. Universidad Autónoma de Madrid. 1981.

CALVO, F.: Estadística aplicada. Ediciones Deusto. 1978.

GUENTHER, W.: Introducción a la Inferencia estadística. Ed. Del Castillo. Madrid, 1968.

OSTLE, B.: Estadística aplicada. Ed. Limusa Wiley.

RIOS, S.: Iniciación-estadística. Ed. ICE.

ROHATGI : And introduction to probability theory & mathematical statistics. Ed. Wiley.

TUCKER, H.: Introducción a la teoría matemática de las probabilidades y a la estadística. Ed. Vicens Vives.

LARSON, H.J.: Introducción a la teoría de probabilidades e inferencia estadística. Ed. Limusa.

KAZMIER, L.J.: Estadística aplicada a la Administración y la Economía. Ed. Schanm.

KAZMIER, L.J.: Basic statistics for business and economics. Ed. MacGraw Hill.

GREGORY HARTLEY LEWIS : Estadística Básica. Ed. Del Castillo. Madrid, 1973.

SAMUELSON : Curso de Economía Moderna. Ed. Aguilar.

G. BARBANCHO, A.: Estadística elemental moderna. Ed. Ariel. Barcelona, 1981.

MOOD/GRAYBILL : Introducción a la teoría estadística. Ed. Aguilar.

MERRIL, W.: Introducción a la estadística económica. Ed. Amorrortu.

LINDLEY, D.: Introduction to probability and statistics, part 2, "Inference". Ed. Cambridge University Press.

ALCAIDE INCHAUSTI, A.: Estadística aplicada a las ciencias sociales. Ed. Pirámide.

A. HBER Y R. RUNYON : Estadística General. Fondo Educativo Interamericano. México, 1973.

GRAIS, B.: Statistique descriptive. Ed. Dunod.

FELLER, W.: Introducción a la teoría de probabilidades y aplicaciones, I e II. Ed. Limusa.

CRAMER, H.: Métodos matemáticos a la Estadística. Ed. Aguilar.

V. QUESADA, ISIDORO, LOPEZ : Curso y Ejercicios de Estadística. Ed. Alhambra.

VIZMANOS, J.R.: Bioestadística. Centro de Promoción Reprográfica. Madrid.

DOMENECH I MASSONS, J.M.: Métodos estadísticos para investigadores. Ed. Herder.

FREEMAN, H.: Introducción a la inferencia estadística. Ed. Trillas.

AZORIN : Curso de muestreo y aplicaciones. Ed. Aguilar.

MURRAY R. SPIEGEL : Estadística. Ed. MacGraw Hill.

DE LA FUENTE, S.: Matemáticas en forma de problemas. CIE 1986.

MINISTERIO DE EDUCACION Y CIENCIA

DIRECCION GENERAL DE ENSEÑANZAS MEDIAS

SUBDIRECCION GENERAL DE ORDENACION ACADEMICA

DOSSIER DIDACTICO
DE
ESTADISTICA DESCRIPTIVA

AUTOR: SANTIAGO DE LA FUENTE FERNANDEZ

GRUPO MADRID - MARKOV

ISBN: 84 - 404 - 0348 - 6
Depósito legal: M - 29344 - 1987

Tiene como cometido describir una muestra, es decir, recoger, ordenar y analizar los datos de una muestra. Esto lo hacemos mediante dos medidas que representan dos aspectos fundamentales

- su tendencia central
- su dispersión.

MEDIDAS DE TENDENCIA CENTRAL

A veces es conveniente reducir la información recogida a un solo valor o a un número pequeño de valores para facilitar la comparación entre distintas muestras o poblaciones. Estos valores, que centralizan la información, se denominan "medidas de tendencia central". Los más estudiados son: la media, mediana y moda.

- Media aritmética $(\bar{x})$

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \dots + x_k n_k}{N} = \frac{\sum_{i=1}^{k} x_i n_i}{N}$$

cuando los datos $x_1, x_2, \dots, x_k$ aparecen una sola vez, se tiene:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i}{k}$$

Piensa que hay k valores.

Como la media es función de todas y cada una de las puntuaciones, se verá muy afectada cuando las puntuaciones extremas son muy dispares. Entonces, no es recomendable calcular la media - por ser poco representativa.

- Media geométrica ($\bar{x}_G$)

$$\bar{x}_G = \sqrt[N]{x_1^{n_1} \cdot x_2^{n_2} \cdots x_k^{n_k}}$$

- Media armónica ($\bar{x}_A$)

$$\bar{x}_A = \frac{N}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \cdots + \frac{n_k}{x_k}} = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$$

observa que si algún valor x_i es cero, entonces la media armónica no tiene sentido.

- La relación entre media aritmética, geométrica y armónica - viene dada por la expresión:

$$\bar{x}_A \leq \bar{x}_G \leq \bar{x}$$

Cuando $\bar{x}_A = \bar{x}_G = \bar{x}$ todos los números $x_1, x_2, x_3, \ldots, x_k$ son - idénticos.

- Mediana (M_e)

La mediana de una serie de N datos ordenados en orden creciente o decreciente es la puntuación que ocupa el lugar central de la distribución estadística. En otras palabras, la mediana M_e es - la medida central que deja igual número de observaciones inferiores que superiores a ella.

DATOS NO AGRUPADOS: Cuando la frecuencia de cada valor sea la unidad (aparece una sola vez cada valor), su cálculo es muy sencillo:

- Si hubiese un número impar de valores, la mediana M_e es el valor central.

- Si hubiese un número par de valores, entonces tomaríamos como mediana M_e la media aritmética de -

los dos valores centrales.

DATOS AGRUPADOS: Cuando los valores $(x_1, x_2, \dots, x_k)$ se presentan, respectivamente, con una frecuencia $n_1, n_2, \dots \dots, n_k$, el valor de la mediana M_e se obtiene mediante las siguientes etapas:

1) Dividimos el número N de observaciones entre 2.

2) Comprobamos si el número N/2 se encuentra en la tabla de las frecuencias absolutas acumuladas N_i.

3) Si el número N/2 no está en la tabla de las N_i, la mediana M_e será aquel valor x_k de la variable X que corresponde al mayor número entre los que se encuentra N/2 $(N_{k-1} < N/2 < N_k)$.

$$M_e = x_k$$

4) Si el número N/2 se encuentra en la tabla de las N_i es que coincide con la frecuencia absoluta acumulada de algún valor x_k de la variable X, en este caso, la mediana M_e viene dada por la expresión:

$$M_e = \frac{x_k + x_{k+1}}{2}$$

Piensa: La mediana M_e es la medida central que divide las observaciones (la distribución) en dos partes iguales. Esto hace que sea más representativa que la media cuando las puntuaciones extremas son muy dispares, puesto que la mediana depende sólo de las puntuaciones centrales de la distribución y no es afectada por los valores extremos.

• La moda M_d

La moda M_d corresponde al valor de la variable que se presenta con mayor frecuencia, es decir, es el valor más común.

La moda puede no existir, incluso si existe puede no ser única; así, si hay dos modas, la distribución se llama bimodal, si tres trimodal, etc.

Características de la moda:

- es sencilla de calcular
- es poco representativa. Sólo se calcula cuando lo que se desea es una medida de tendencia central aproximada y muy común.

MEDIDAS DE DISPERSION

Las medidas de tendencia central reducen toda la información recogida de una muestra a un solo valor, en algunos casos, éste único valor estará más próximo a la realidad de las observaciones que en otros. Por tanto, se hace necesario cuantificar la representatividad de los valores centrales.

Se vé, la necesidad de definir unas nuevas medidas estadísticas. Estas se denominan "medidas de dispersión".

¿Qué es una medida de dispersión?: Es un índice estadístico que nos permite conocer el grado de variabilidad o dispersión de los datos de una distribución. Los más utilizados son: la varianza y la desviación típica.

- VARIANZA (σ^2)

Se define la varianza σ^2 de una variable estadística X de la siguiente forma:

$$\sigma^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_i}{N}$$

siendo:

x_i = distintos valores de la variable X.

$\bar{x}$ = la media aritmética.

N = número total de datos u observaciones.

Es evidente que si los valores $x_1, x_2, \dots, x_k$ aparecen, cada uno de ellos, una sola vez, se tiene:

$$\sigma^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2}{k}$$

En el caso que $\sigma^2 = 0$ entendemos que todos los valores x_i coinciden con la media $\bar{x}$, es decir, todas las observaciones están concentradas en un mismo valor, por lo que la dispersión es mínima (nula).

PIENSA EN LAS CARACTERISTICAS DE LA VARIANZA:

- Puesto que la varianza σ^2 se obtiene como suma de cuadrados es siempre positiva. Por otro lado, estará expresada en unidades al cuadrado cuando la variable estudiada se expresa en unidades. Observa la necesidad de definir otra nueva medida de dispersión: la desviación típica.

- No es recomendable su cálculo cuando tampoco lo sea el de la media como medida de tendencia central.

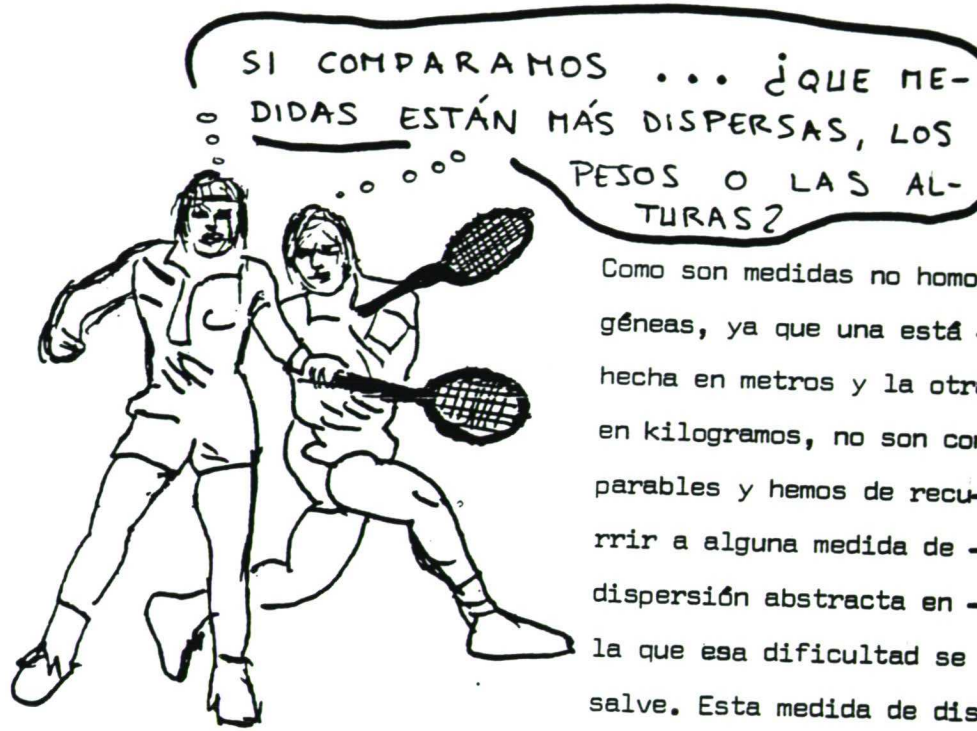
• DESVIACION TIPICA (σ)

Se define la desviación típica σ como la raíz cuadrada positiva de la varianza:

$$\sigma = + \sqrt{\sigma^2} = + \sqrt{\frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_i}{N}}$$

es la medida de dispersión más utilizada en estadística. Piensa que viene expresada en las mismas unidades que los valores de la variable X, por lo que su interpretación resulta más sencilla.

- Toma siempre valores positivos.
- No es recomendable su cálculo cuando tampoco lo sea el de la media como medida de tendencia central.



Como son medidas no homogéneas, ya que una está hecha en metros y la otra en kilogramos, no son comparables y hemos de recurrir a alguna medida de dispersión abstracta en la que esa dificultad se salve. Esta medida de dispersión es el "coeficiente de variación de Pearson".

La medida más dispersa será la que tiene mayor coeficiente de variación de Pearson.

• COEFICIENTE DE VARIACION DE PEARSON

Se define el coeficiente de variación de Pearson C.V. de la forma:

$$C.V. = \frac{\sigma}{\bar{x}}$$

A veces este coeficiente se multiplica por 100 para así trabajar en porcentajes.

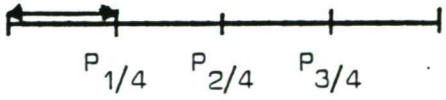
Este coeficiente no se puede hallar cuando $\bar{x} = 0$.

OTRAS MEDIDAS DE TENDENCIA CENTRAL Y DISPERSION

DE TENDENCIA CENTRAL

- CUARTILES

Se definen los cuartiles como tres valores de la variable que dividen las observaciones realizadas en cuatro partes iguales:



Primer cuartil ($P_{1/4}$): es el valor de la variable que deja la cuarta parte de las observaciones menores o iguales a él y, por tanto, las 3/4 partes superiores a él.

Se calcula igual que la mediana, pero en vez de tomar el número de observaciones N/2, se toma el número de observaciones N/4.

Segundo cuartil ($P_{2/4}$): es el valor de la variable que deja inferiores o iguales a él las 2/4 partes (la mitad) de las observaciones. Este cuartil es la "mediana".

Evidentemente, se toma $\frac{2N}{4} = N/2$ número de observaciones.

Tercer cuartil ($P_{3/4}$): es el valor de la variable que deja inferiores o iguales a él las 3/4 partes de las observaciones.

Se calcula de forma análoga a la mediana tomando $\frac{3}{4}$ N observaciones.

- PERCENTILES

Se llaman también "centiles" (de 100).

Se define el percentil como el valor de la variable que deja inferiores o iguales a él un porcentaje determinado de observaciones.

Así: el percentil k-ésimo será el valor de la variable que deja inferiores o iguales a él las k/100 partes de las observaciones (el k por 100), donde k puede tomar cualquier valor desde 1 a 99.

Su cálculo es idéntico al de la mediana y los cuartiles.

Se denotan por P_k.

de tal forma:

Percentil 25: es el valor de la variable que deja 25/100 de las observaciones menores o iguales a él. Se denota P_{25} y su cálculo es igual que el de la mediana tomando $\frac{25}{100}$ N observaciones.

observa que

$$\frac{25}{100} N = \frac{1}{4} N = P_{1/4}$$

Percentil 50 (P_{50}): es el valor de la variable que deja $\frac{50}{100}$ de las observaciones menores o iguales a él. Se calcula igual que la mediana tomando $\frac{50}{100}$ N observaciones.

observa

$$\frac{50}{100} N = N/2$$

Este percentil también se llama "mediana".

Percentil 75 (P_{75}):

$$P_{75} = P_{3/4} \quad \text{(tercer cuartil)}$$

DE DISPERSION

- Amplitud o Recorrido

Se define la amplitud o recorrido de una variable estadística como la diferencia entre su valor máximo y su valor mínimo.

- Recorrido semiintercuartílico

El recorrido semiintercuartílico es la mitad de la diferencia entre el tercer cuartil ($P_{3/4}$) y el primero ($P_{1/4}$) o, lo que es lo mismo, entre el percentil 75 (P_{75}) y el percentil 25 (P_{25}).

Sea, por tanto:

$$P = \frac{P_{3/4} - P_{1/4}}{2}$$

o bien

$$P = \frac{P_{75} - P_{25}}{2}$$

¿COMO SE DEBEN UTILIZAR LAS MEDIDAS DE DISPERSION?

Sabemos que una medida de tendencia central nos proporciona poca información. Para describir una información más completa necesitamos cuantificar la representatividad de las medidas centrales, esta información adicional nos la proporciona las medidas de dispersión.

Es necesario, pues, mostrar las "parejas" de medidas estadísticas más empleadas, así como su idoneidad:

media ($\bar{x}$): cuando las observaciones están distribuidas simétricamente alrededor de un valor central.

varianza (σ^2): cuando es recomendable el cálculo de la media como media de tendencia central. Presenta el inconveniente de expresar las unidades al cuadrado.

desviación típica (σ): cuando es recomendable el cálculo de la media como media de tendencia central. Tiene la ventaja de expresarse en las mismas unidades que los valores de la variable.

mediana (M_e): cuando existen valores extremos dispares que afectan a la media. O bien cuando se desea conocer el valor medio exacto de las observaciones.

recorrido semiintercuartílico: cuando es recomendable el cálculo de la mediana como medida de tendencia central.

Moda (M_d): Es aconsejable su cálculo cuando alguno de los valores se presenta con mucha frecuencia.

Amplitud o Recorrido: cuando es recomendable el cálculo de la moda como medida de tendencia central.

MOMENTOS

Definimos el momento de orden r respecto al parámetro c, de la forma:

$$M_r(c) = \frac{\sum_i (x_i - c)^r \cdot n_i}{N}$$

Observa, en particular, dos casos importantes:

- Decimos que un momento es respecto al origen cuando c = 0, entonces, se tiene:

$$M_r(0) = \frac{\sum_i (x_i - 0)^r \cdot n_i}{N} = \frac{\sum_i x_i^r \cdot n_i}{N}$$

A este momento particular se le denota por a_r, de forma que

$$a_r = \frac{\sum_i x_i^r \cdot n_i}{N}$$

dando valores a r, se obtiene:

$$a_0 = \frac{\sum_i x_i^0 \cdot n_i}{N} = \frac{\sum_i n_i}{N} = \frac{N}{N} = 1$$

$$a_1 = \frac{\sum_i x_i \cdot n_i}{N} = \bar{x} \text{ (primer momento respecto al origen)}$$

$$a_2 = \frac{\sum_i x_i^2 \cdot n_i}{N} \text{ (segundo momento respecto al origen)}$$

CONCLUSIONES:

$$a_0 = 1$$

$$a_1 = \bar{x} \text{ (media aritmética)}$$

- Decimos que un momento es respecto a la media - cuando $c = \bar{x}$, de donde:

$$M_r(\bar{x}) = \frac{\sum_i (x_i - \bar{x})^r \cdot n_i}{N}$$

A este momento particular se le denota por m_r, de forma que:

$$m_r = \frac{\sum_i (x_i - \bar{x})^r \cdot n_i}{N}$$

dando valores a r se obtiene:

$$m_0 = \frac{\sum_i (x_i - \bar{x})^0 \cdot n_i}{N} = \frac{\sum_i n_i}{N} = \frac{N}{N} = 1$$

$$m_1 = \frac{\sum_i (x_i - \bar{x}) \cdot n_i}{N} = \frac{0}{N} = 0$$ (primer momento respecto a la media)

$$m_2 = \frac{\sum_i (x_i - \bar{x})^2 \cdot n_i}{N} = \sigma^2$$ (segundo momento respecto a la media)

CONCLUSIONES:

$m_0 = 1$

$m_1 = 0$

$m_2 = \sigma^2$ (varianza)

- Para calcular la varianza σ^2 podemos emplear - también la relación:

$$m_2 = a_2 - (a_1)^2$$

o bien

$$\sigma^2 = a_2 - (\bar{x})^2$$

siendo

a_2 = momento respecto al origen de orden 2 = $\frac{\sum_i x_i^2 \cdot n_i}{N}$

a_1 = momento respecto al origen de orden 1 = $\frac{\sum_i x_i \cdot n_i}{N}$

= media aritmética.

En la práctica, a veces, éste método de calcular la varia es mucho más rápido y sencillo.

• ESPACIO MUESTRAL: En un experimento aleatorio, el conjunto de posibles resultados diferentes del mismo, recibe el nombre de "espacio muestral" asociado al experimento aleatorio. Se denota por Ω.

• SUCESO: Un "suceso" de un experimento aleatorio corresponde a la pregunta de que tenga o no tenga respuesta después de realizado el experimento.

"OPERACIONES CON SUCESOS"

- Unión de sucesos A y B: $A \cup B$
- Intersección de sucesos A y B: $A \cap B$
- Suceso complementario de A: $\bar{A}$
- Suceso imposible: $A \cap \bar{A} = \emptyset$
- Sucesos incompatibles: $A \cap B = \emptyset$, A y B incompatibles.
- Diferencia de sucesos: $A - B = A \cap \bar{B}$
- Suceso A contenido en el suceso B: $A \subset B$
- Asociativas:

$$A \cup (B \cup C) = (A \cup B) \cup C$$
$$A \cap (B \cap C) = (A \cap B) \cap C$$

- Conmutativas:

$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$

- Distributivas de la unión y de la intersección:

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

- Elemento neutro:

$A \cup \emptyset = A$ $\quad \emptyset$ para la unión

$A \cap \Omega = A$ $\quad \Omega$ para la intersección

• PROBABILIDAD: Dado el espacio muestral Ω, se define la probabilidad como una aplicación del espacio muestral Ω en el intervalo $[0,1]$, esto es:

$$P : \Omega \longrightarrow [0,1]$$

de tal forma:

$P(\Omega) = 1$ $\quad$ suceso seguro

$P(\emptyset) = 0$ $\quad$ suceso imposible

$P(A \cup B) = P(A) + P(B)$ cuando los sucesos A y B son incompatibles

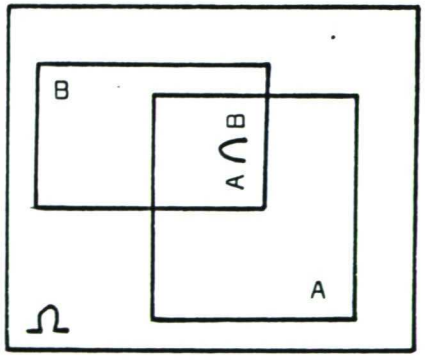
CONSECUENCIAS:

$$\left.\begin{array}{l} P(\Omega) = 1 \\ A \cup \bar{A} = \Omega \\ A \text{ y } \bar{A} \text{ incompatibles: } A \cap \bar{A} = \emptyset \end{array}\right\} \Rightarrow$$

$$\Rightarrow \begin{array}{l} P(\bar{A}) + P(A) = 1 \\ P(\bar{A}) = 1 - P(A) \end{array}$$

$$A \subset B \Rightarrow P(A) \leq P(B)$$

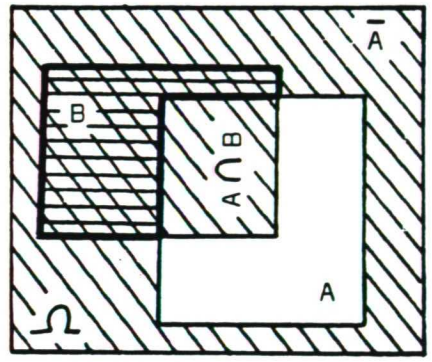
Sean A y B dos sucesos cualesquiera del espacio muestral Ω, entonces:



$A \cup B = A + B - A \cap B$

de donde

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$



$\bar{A} \cap B = B - A \cap B$

de donde

$P(\bar{A} \cap B) = P(B) - P(A \cap B)$

• PROBABILIDAD CONDICIONADA: La probabilidad de que ocurra el suceso B cuando haya ocurrido el suceso A se llama probabilidad condicionada y se denota:

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

probabilidad del suceso B condicionado al suceso A.

• INDEPENDENCIA: Dados dos sucesos A y B de un espacio muestral Ω, diremos que son independientes si:

$$P(B/A) = P(B)$$

en otras palabras:

$$P(A \cap B) = P(A) \cdot P(B)$$

• PROBABILIDAD TOTAL: Consideremos un conjunto de sucesos A_i, siendo i = 1, 2, ... , n del espacio muestral Ω tales que:

$$\bigcup_{i=1}^{n} A_i = \Omega$$

Si el suceso B puede ocurrir por alguna de las causas A_i, la probabilidad de que ocurra el suceso B viene dada por:

$$P(B) = \sum_{i=1}^{n} P(A_i) \cdot P(B/A_i)$$

esto es:

> "La probabilidad de que ocurra el suceso B es la suma de las probabilidades de las causas $P(A_i)$ - por la probabilidad del suceso B condicionado a la causa A_i, $P(B/A_i)$".

• TEOREMA DE BAYES: La probabilidad de que ocurrido el suceso B la causa que lo haya producido sea la A_i, establece que:

$$P(A_i/B) = \frac{P(A_i) \cdot P(B/A_i)}{\sum_{i=1}^{n} P(A_i) \cdot P(B/A_i)}$$

En el caso particular, cuando las causas son A_1 y A_2, el teorema de Bayes establece:

$$P(A_1/B) = \frac{P(A_1) \cdot P(B/A_1)}{P(A_1) \cdot P(B/A_1) + P(A_2) \cdot P(B/A_2)}$$

VARIABLE ESTADISTICA BIDIMENSIONAL

Se considerarán situaciones en las que el estadístico realiza la observación simultánea de dos caracteres en el individuo, obteniéndose, por tanto, pares de resultados.

Un carácter puede tomar distintas MODALIDADES:

carácter X puede tomar las modalidades $(x_1, x_2, \dots, x_k)$

carácter Y puede tomar las modalidades $(y_1, y_2, \dots, y_l)$

> Los distintos valores de las modalidades que pueden adoptar estos caracteres forman un conjunto de pares, que representaremos por (X,Y), y llamaremos "variable estadística bidiomensional".

ORDENACION DE DATOS

TABLA DE DOBLE ENTRADA

X \ Y	y_1	y_2	...	y_j	....	y_j	
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1l}	n_{x_1}
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2l}	n_{x_2}
⋮							
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{il}	n_{x_i}
⋮							
x_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kl}	n_{x_k}
	n_{y_1}	n_{y_2}	...	n_{y_j}	...	n_{y_l}	N

siendo

n_{ij} = número de veces que aparece repetido el par (x_i, y_j).
Llamaremos "frecuencia absoluta" del par (x_i, y_j).

f_{ij} = es la "frecuencia relativa" del par observado (x_i, y_j), viene dada por el cociente:

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

PROPIEDADES:

$$\sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} = N \text{ (número total de observaciones)}$$

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

DISTRIBUCIONES MARGINALES

Se pueden obtener las distribuciones marginales simultáneamente de la tabla de doble entrada.

Así:

- DISTRIBUCION MARGINAL DE LA X:

X	n_{x_i}	f_{x_i}
x_1	n_{x_1}	$f_{x_1} = n_{x_1}/N$
x_2	n_{x_2}	$f_{x_2} = n_{x_2}/N$
⋮	⋮	⋮
x_i	n_{x_i}	$f_{x_i} = n_{x_i}/N$
⋮	⋮	⋮
x_k	n_{x_k}	$f_{x_k} = n_{x_k}/N$
	$\sum_{i=1}^{k} n_{x_i} = N$	$\sum_{i=1}^{k} f_{x_i} = 1$

Media y varianza marginal de la X:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot n_{x_i}}{N} \text{ (media)}$$

$$\sigma_x^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_{x_i}}{N} \text{ (varianza)}$$

$$\sigma_x = + \sqrt{\sigma_x^2} \quad \text{(desviación típica)}$$

Observa:

$$\sum_{i=1}^{k} n_{x_i} = n_{x_1} + n_{x_2} + \dots + n_{x_i} + \dots + n_{x_k} = N$$

$$\sum_{i=1}^{k} f_{x_i} = f_{x_1} + f_{x_2} + \dots + f_{x_i} + \dots + f_{x_k} = 1$$

donde

n_{x_i} = frecuencia absoluta marginal del valor x_i

f_{x_i} = frecuencia relativa marginal del valor x_i

- DISTRIBUCION MARGINAL DE LA Y:

Y	n_{y_j}	f_{y_j}
y_1	n_{y_1}	$f_{y_1} = n_{y_1}/N$
y_2	n_{y_2}	$f_{y_2} = n_{y_2}/N$
⋮	⋮	⋮
y_j	n_{y_j}	$f_{y_j} = n_{y_j}/N$
⋮	⋮	⋮
y_l	n_{y_l}	$f_{y_l} = n_{y_l}/N$
	$\sum_{j=1}^{l} n_{y_j} = N$	$\sum_{j=1}^{l} f_{y_j} = 1$

Media y varianza marginal de la Y:

$$\bar{y} = \frac{\sum_{j=1}^{l} y_j \cdot n_{y_j}}{N} \quad \text{(media)}$$

$$\sigma_y^2 = \frac{\sum_{j=1}^{l} (y_j - \bar{y})^2 \cdot n_{y_j}}{N} \quad \text{(varianza)}$$

$$\sigma_y = + \sqrt{\sigma_y^2} \quad \text{(desviación típica)}$$

Observa:

$$\sum_{j=1}^{l} n_{y_j} = n_{y_1} + n_{y_2} + \dots + n_{y_j} + \dots + n_{y_l} = N$$

$$\sum_{j=1}^{l} f_{y_j} = f_{y_1} + f_{y_2} + \dots + f_{y_j} + \dots + f_{y_l} = 1$$

donde

n_{y_j} = frecuencia absoluta marginal del valor y_j

f_{y_j} = frecuencia relativa marginal del valor y_j

DISTRIBUCIONES CONDICIONADAS

Cuando se desea conocer la distribución de una variable solamente para un único valor de la otra variable, estamos condicionando los valores de la variable X al valor y_j, o bien condicionando los valores de la variable Y al valor x_i.

De donde:

- DISTRIBUCION CONDICIONADA DE X A Y = y_j:

Nos fijamos en la tabla de doble entrada, y formamos la tabla donde aparecen todos los valores de la variable X y el valor y_j de la variable Y, de esta forma:

X \ Y	. . . . . y_j . . . . .
x_1	n_{1j}
x_2	n_{2j}
⋮	⋮
x_i	. . . . . n_{ij}
⋮	⋮
x_k	n_{kj}
	n_{y_j}

Se tendrá, por tanto:

X	$n(X/Y = y_j)$	$f(X/Y = y_j)$
x_1	n_{1j}	$f_{1j} = n_{1j}/n_{y_j}$
x_2	n_{2j}	$f_{2j} = n_{2j}/n_{y_j}$
⋮	⋮	⋮
x_i	$\cdots \; n_{ij}$	$f_{ij} = n_{ij}/n_{y_j}$
⋮	⋮	⋮
x_k	n_{kj}	$f_{kj} = n_{kj}/n_{y_j}$
	n_{y_j}	1

De una manera formal, queda:

- $n(X/Y = y_j) = \left\{n_{1j}, n_{2j}, \cdots, n_{ij}, \cdots, n_{kj}\right\}$

- $f(X/Y = y_j) = \dfrac{n(X/Y = y_j)}{n_{y_j}}$

- DISTRIBUCION CONDICIONADA DE Y A $X = x_i$:

Análogamente, si deseamos obtener la distribución de la variable Y solamente cuando $X = x_i$, se tendrá:

X \ Y	y_1	y_2	⋯	y_j	⋯	y_l	
⋮				⋮			⋮
x_i	n_{i1}	n_{i2}	⋯	n_{ij}	⋯	n_{il}	$\cdots \; n_{x_i}$
⋮				⋮			⋮

se sigue que

Y	y_1	y_2	⋯	y_j	⋯	y_l	
$n(Y/X = x_i)$	n_{i1}	n_{i2}	⋯	n_{ij}	⋯	n_{il}	n_{x_i}
$f(Y/X = x_i)$	f_{i1}	f_{i2}	⋯	f_{ij}	⋯	f_{il}	1

de esta forma, tendremos:

- frecuencia absoluta condicionada

$$n(Y/X = x_i) = \left\{n_{i1}, n_{i2}, \cdots, n_{ij}, \cdots, n_{il}\right\}$$

• frecuencia relativa condicionada

$$f(Y/X = x_i) = \frac{n(Y/X = x_i)}{n_{x_i}}$$

MOMENTOS

Definimos el momento de órdenes r y s respecto al par de parámetros (c,v), de la forma:

$$M_{r,s}(c,v) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}(x_i - c)^r \cdot (y_j - v)^s \, n_{ij}}{N}$$

Observa, en particular, dos casos importantes:

- Decimos que un momento es respecto al origen cuando c = 0, v = 0, entonces:

$$M_{r,s}(0,0) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}(x_i - 0)^r \cdot (y_j - 0)^s \, n_{ij}}{N}$$

A los momentos respecto al origen se les denota por $a_{r,s}$, de forma que:

$$a_{r,s} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} x_i^r \cdot y_j^s \cdot n_{ij}}{N}$$

dando valores a r y s, son de interés posterior los momentos:

$$a_{00} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} x_i^0 \cdot y_j^0 \cdot n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}}{N} = \frac{N}{N} = 1$$

$$a_{10} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} x_i^1 \cdot y_j^0 \cdot n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} x_i \cdot n_{ij}}{N} = \frac{\sum_{i=1}^{k} x_i \cdot n_{x_i}}{N} = \bar{x}$$

$$a_{01} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} x_i^0 \cdot y_j^1 \cdot n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} y_j \cdot n_{ij}}{N} = \frac{\sum_{j=1}^{l} y_j \cdot n_{y_j}}{N} = \bar{y}$$

$$a_{11} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} x_i \cdot y_j \cdot n_{ij}}{N}$$

$$a_{20} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} x_i^2 \cdot y_j^0 \cdot n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} x_i^2 \cdot n_{ij}}{N} = \frac{\sum_{i=1}^{k} x_i^2 \cdot n_{x_i}}{N}$$

$$a_{02} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} x_i^0 \cdot y_j^2 \cdot n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} y_j^2 \cdot n_{ij}}{N} = \frac{\sum_{j=1}^{l} y_j^2 \cdot n_{y_j}}{N}$$

CONCLUSIONES IMPORTANTES:

$$a_{00} = 1$$

$$a_{10} = \bar{x}$$

$$a_{01} = \bar{y}$$

$$a_{11} = \frac{\sum_{i}\sum_{j} x_i \cdot y_j \cdot n_{ij}}{N}$$

$$a_{20} = \frac{\sum_{i} x_i^2 \cdot n_{x_i}}{N}$$

$$a_{02} = \frac{\sum_{j} y_j^2 \cdot n_{y_j}}{N}$$

- Decimos que un momento es respecto a la media cuando $c = \bar{x}$ y $v = \bar{y}$, de donde:

$$M_{r,s}(\bar{x},\bar{y}) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} (x_i - \bar{x})^r \cdot (y_j - \bar{y})^s \cdot n_{ij}}{N}$$

A los momentos respecto a la media se les denota por $m_{r,s}$, de forma que:

$$m_{r,s} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} (x_i - \bar{x})^r \cdot (y_j - \bar{y})^s \cdot n_{ij}}{N}$$

dando valores a r y s, son de interés posterior los momentos:

$$m_{11} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}(x_i - \bar{x}).(y_j - \bar{y})\, n_{ij}}{N} \quad \text{covarianza}$$

$$m_{20} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}(x_i - \bar{x})^2 . (y_j - \bar{y})^0 . n_{ij}}{N} =$$

$$= \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}(x_i - \bar{x})^2 . n_{ij}}{N} = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 . n_{x_i}}{N} = \sigma_x^2$$

$$m_{02} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}(x_i - \bar{x})^0 . (y_j - \bar{y})^2 . n_{ij}}{N} =$$

$$= \frac{\sum_{j=1}^{l}(y_j - \bar{y})^2 . n_{y_j}}{N} = \sigma_y^2$$

CONCLUSIONES IMPORTANTES:

m_{11} = covarianza

$m_{20} = \sigma_x^2$ varianza de la X

$m_{02} = \sigma_y^2$ varianza de la Y

- La covarianza m_{11} es la media aritmética de los productos entre la diferencia $(x_i - \bar{x})$ y la diferencia $(y_j - \bar{y})$ correspondientes a cada uno n_{ij} de los elementos que componen un grupo.

Recuerda:

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

En la práctica, para hallar la covarianza es bastante más fácil y rápido, la expresión:

$$m_{11} = a_{11} - a_{10} \cdot a_{01}$$

donde

$$a_{11} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} x_i \cdot y_j \cdot n_{ij}}{N}$$

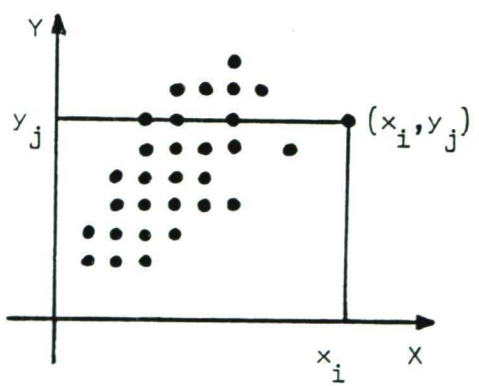
$a_{10} = \bar{x}$ media aritmética de X.

$a_{01} = \bar{y}$ media aritmética de Y.

REGRESION

Como estamos observando dos caracteres en cada individuo, se presenta ahora el problema de determinar la existencia de algún tipo de dependencia entre ellos. Un ejemplo de este tipo de análisis consiste en estudiar la relación entre el consumo de tabaco y el cáncer de pulmón.

DIAGRAMA DE DISPERSION: Es la representación sobre uncs ejes cartesianos de los distintos valores de la variable (X,Y).



REGRESION: Consiste en obtener una ecuación que se pueda usar para predecir los valores de una variable a partir de la otra.

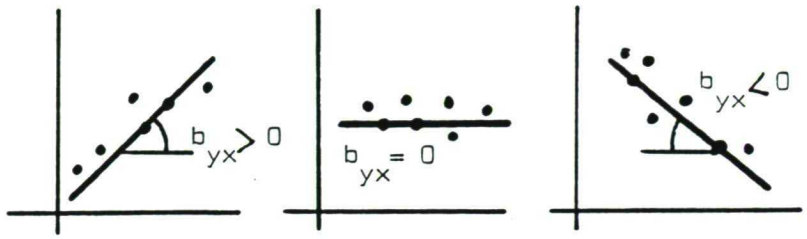
Así, cuando la ecuación es una recta, tenemos la "regresión lineal". De esta forma:

- RECTA DE REGRESION DE Y SOBRE X: Estudia la posible "predicción" de los valores de la variable Y a partir de valores conocidos de la variable X. La ecuación viene dada por la expresión:

$$y - \bar{y} = \frac{m_{11}}{\sigma_x^2}(x - \bar{x})$$

siendo, $b_{yx} = \frac{m_{11}}{\sigma_x^2}$ la pendiente de la recta de regresión de Y sobre X.

$$b_{yx} = \frac{m_{11}}{\sigma_x^2} = \begin{cases} > 0 & \text{recta regresión creciente} \\ = 0 & \text{" " horizontal} \\ < 0 & \text{" " decreciente} \end{cases}$$



recta regresión creciente | recta regresión horizontal | recta regresión decreciente

A la pendiente b_{yx} también se le llama "coeficiente de regresión" de Y sobre X.

- RECTA DE REGRESION DE X SOBRE Y: Estudia la posible "predicción" de los valores de la variable X a partir de los valores conocidos de la variable Y.

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

$$x - \bar{x} = \frac{m_{11}}{\sigma_y^2}(y - \bar{y})$$

La pendiente de la recta de regresión, o también el "coeficiente de regresión de X sobre Y" es tal que:

$$b_{xy} = \frac{m_{11}}{\sigma_y^2} = \begin{cases} > 0 & \text{recta regresión creciente} \\ = 0 & \text{" " horizontal} \\ < 0 & \text{" " decreciente} \end{cases}$$

CORRELACION

Estudia el tipo de dependencia que existe entre las variables X e Y, intentando cuantificarla mediante el cálculo del coeficiente de correlación.

En otras palabras, se entiende por correlación el grado de asociación entre dos variables. Nos limitaremos a estudiar el coeficiente de correlación lineal.

- COEFICIENTE DE CORRELACION LINEAL: Es un número abstracto que determina el grado de ajuste entre una nube de puntos (diagrama de dispersión) y una recta de regresión.

Viene definido por la expresión:

$$r = \sqrt{b_{yx} \cdot b_{xy}} = \frac{m_{11}}{\sigma_x^2} \cdot \frac{m_{11}}{\sigma_y^2} = \frac{m_{11}}{\sigma_x \cdot \sigma_y}$$

El coeficiente de correlación lineal φ se encuentra acotado entre los valores - 1 y 1, por tanto:

$$-1 \leq \varphi \leq 1$$

Si el coeficiente de correlación φ aparece multiplicado por 100, se trabajaría con "porcentajes".

- RELACION ENTRE LOS COEFICIENTES DE REGRESION Y DE CORRELACION:

$$\left.\begin{array}{l} b_{yx} = \dfrac{m_{11}}{\sigma_x^2} \\ \varphi = \dfrac{m_{11}}{\sigma_x \cdot \sigma_y} \end{array}\right\} \Rightarrow b_{yx} = \varphi \cdot \frac{\sigma_y}{\sigma_x}$$

$$\left.\begin{array}{l} b_{xy} = \dfrac{m_{11}}{\sigma_y^2} \\ \varphi = \dfrac{m_{11}}{\sigma_x \cdot \sigma_y} \end{array}\right\} \Rightarrow b_{xy} = \varphi \cdot \frac{\sigma_x}{\sigma_y}$$

- CORRELACION LINEAL DIRECTA E INVERSA:

a) Sea la recta de regresión de Y sobre X, observando la relación entre los coeficientes de regresión y correlación:

$$b_{yx} = \varphi \cdot \frac{\sigma_y}{\sigma_x}$$

se tiene:

- $b_{yx} > 0 \iff \varphi > 0$ (ya que $\sigma_x \geq 0$, $\sigma_y \geq 0$)

en este caso, decimos que hay "correlación directa" entre las variables.

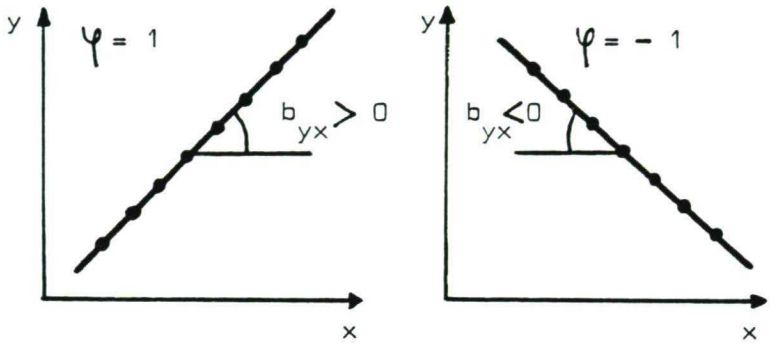
- $b_{yx} < 0 \iff \varphi < 0$ (ya que $\sigma_x \geq 0$, $\sigma_y \geq 0$)

en este caso, decimos que hay "correlación inversa" entre las variables.

b) En la recta de regresión de X sobre Y, se sigue un razonamiento análogo.

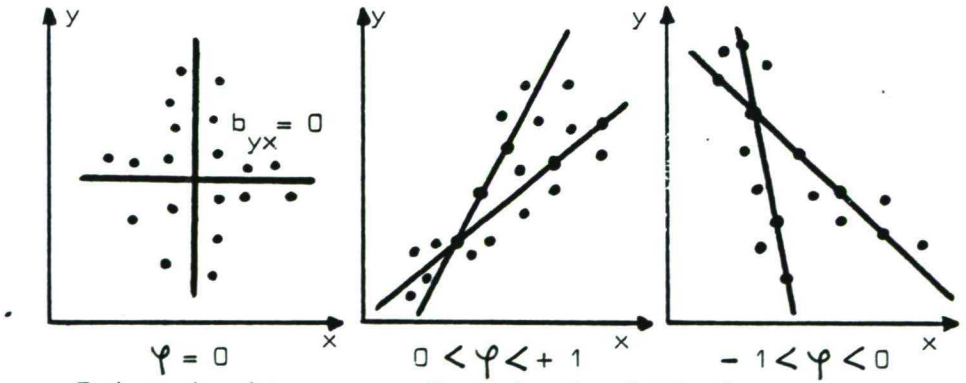
- INTERPRETACION

a) DEPENDENCIA FUNCIONAL: Los puntos de la nube se encuentran sobre la recta de regresión:



recta regresión cociente recta regresión decreciente

b) DEPENDENCIA ALEATORIA: No existe una ecuación que relacione las variables X e Y.



Independencia aleatoria Dependencia aleatoria