

Estadística

Tema 1: Descripción de una variable

Estadística descriptiva e inductiva

► **Estadística descriptiva**

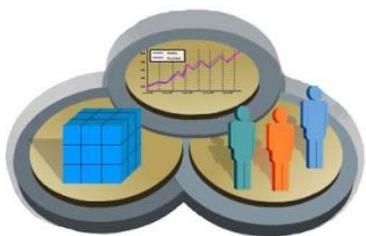
- ▶ recolección, presentación, descripción, análisis e interpretación de una colección de datos.
- ▶ es el método de obtener de un conjunto de datos **conclusiones** sobre sí mismos,
- ▶ Puede utilizarse para resumir o describir cualquier conjunto, se trate de una población o de una muestra
- ▶ Ejemplos:
 - Los datos del Censo de población de un año determinado.
 - La cantidad de robos ocurridos el último mes en una ciudad concreta.



Estadística descriptiva e inductiva

► Estadística inductiva

- ▶ proceso de lograr generalizaciones acerca de las propiedades del todo, población, partiendo de lo específico, muestra
- ▶ La muestra debe ser representativa de la población



- ▶ La estadística inferencial es el conjunto de técnicas que se utiliza para obtener conclusiones que sobrepasan los límites del conocimiento aportado por los datos

► 3

Estadística descriptiva e inductiva

► Ejercicios: ¿Qué campo de la estadística será necesario utilizar?.

- 1) Un material que es fabricado en un proceso continuo, antes de ser cortado y enrollado en grandes rollos, debe supervisarse su espesor (mediante un calibrador). Se registraron diez mediciones de papel, en mm, y el promedio resultó de 30,1 mm.
- 2) Un lote de 1.000 CDs debe pasar por un control de calidad. Se eligen al azar 30 CDs para decidir si el lote pasa o no dicho control de calidad y pueda distribuirse.

► 4

Estadística descriptiva

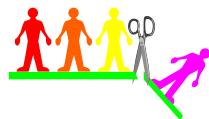
► Población:

- ▶ Es el conjunto de individuos sobre los que se quiere obtener información
- ▶ Normalmente es demasiado grande para poder abarcarla



► Muestra

- ▶ Un subconjunto representativo de la población
- ▶ Es un subconjunto al que tenemos acceso y sobre el que realmente hacemos las observaciones (mediciones)
- ▶ Está formado por miembros “seleccionados” de la población (individuos, unidades experimentales) de acuerdo con un plan o regla, con el fin de obtener información de la población



► 5

Estadística descriptiva

► Individuo o elemento

- ▶ Un miembro de la población

► Variable

- ▶ Cada una de las características que se miden en los individuos de la población
- ▶ La información de que disponemos de cada individuo se resume en variables
- ▶ Las variables estadísticas presentan valores o modalidades
 - ▶ Los valores o modalidades deben ser incompatibles (todo individuo puede presentar una y solamente una modalidad)

► 6

Estadística descriptiva

► Ejemplo



- **Muestra:** 60 trabajadores de empresas de comunicación
- **Unidad de análisis o Individuo:** Trabajador de empresa de comunicación
- **Variables:** sexo, edad, salario, Nº de horas de trabajo, etc.

► 7

Variables estadísticas

► Variables cuantitativas

- Miden alguna propiedad cuantificable del individuo cuyos valores se expresan en escala de intervalos o de razón
 - **Discretas:** aquéllas que son numerables y toman sólo ciertos valores en el intervalo considerado y no admiten valores intermedios entre dos valores consecutivos
 - Por ejemplo el número de hijos de una familia, número de artículos defectuosos, etc.
 - **Continuas:** aquéllas que son medibles y que pueden tomar cualquier valor en el intervalo considerado.
 - Por ejemplo el peso de una persona, la temperatura de ignición de un gas, el tiempo de acceso a un servidor, altura, etc..

► 8

Variables estadísticas

► Variables cualitativas:

- ▶ Son cualidades o características de los individuos, con cuyos valores no se pueden realizar operaciones aritméticas y, por tanto, no se pueden asociar **naturalmente** a un número (no se pueden hacer operaciones algebraicas con ellos)
- ▶ Sólo pueden expresarse en escala nominal u ordinal
- ▶ También se denominan categóricas, ya que agrupan a los individuos en categorías
 - ▶ **Nominales:** Si sus valores no se pueden ordenar
 - Asignatura del primer curso del grado, sexo, estado civil, tipo de envase, grupo sanguíneo, religión, nacionalidad, fumador (Sí/No)
 - ▶ **Ordinales:** Si sus valores se pueden ordenar
 - Mejoría a un tratamiento, Grado de satisfacción, Intensidad del dolor

► 9

Variable estadística

► Variables ordinales

- ▶ Son aquéllas cuyo valor puede ser ordenado de mayor a menor.
- ▶ Por ejemplo una pregunta dirigida a cada uno de los individuos de una población:

¿Está usted satisfecho con la política laboral del gobierno?
Marque la opción elegida

Muy satisfecho
Satisfecho
Poco satisfecho
Nada satisfecho

► 10

Estadística descriptiva

► Ejemplo (i)

- Entre los individuos de la *población española* se podrían escoger como *variables*:

- **El grupo sanguíneo**
 ▸ {A, B, AB, O} ← Var. Cualitativa
- **Su nivel de felicidad “declarado”**
 ▸ {Deprimido, Ni fu ni fa, Muy Feliz} ← Var. Ordinal
- **El número de hijos**
 ▸ {0,1,2,3,...} ← Var. Numérica discreta
- **La altura**
 ▸ {1'62 ; 1'74; ...} ← Var. Numérica continua



► 11

VARIABLES ESTADÍSTICAS

► Ejemplo (ii)

- **Codificación** de las variables con números para poder procesarlas con facilidad (SPSS)
- Es conveniente asignar “etiquetas” a los valores de las variables para recordar qué significan los códigos numéricos.
 - **Sexo** (Cualit: Códigos arbitrarios)
 ▸ 1 = Hombre; 2 = Mujer
 - **Raza** (Cualit: Códigos arbitrarios)
 ▸ 1 = Blanca; 2 = Negra
 - **Felicidad** (Ordinal: Respetar un orden al codificar)
 ▸ 1 = Muy feliz
 ▸ 2 = Bastante feliz;
 ▸ 3 = No demasiado feliz

The figure consists of two side-by-side screenshots of the SPSS Data Editor. Both show a dataset with 10 observations and 10 variables: sexo, raza, región, feliz, vida, herma, hijos, educ, edad, and ed. In the top screenshot, all variables are labeled with their respective names. In the bottom screenshot, the variables are labeled with numbers (1 through 9) except for 'ed' which is labeled 'edad'. The data values remain the same in both cases.

	sexo	raza	región	feliz	vida	herma	hijos	educ	edad	ed
1	Mujer	Blanca	Nor-E	Muy feliz	Excitante	1	2	12	61	No r
2	Mujer	Blanca	Nor-E	Bastante	Excitante	2	1	20	32	
3	Hombre	Blanca	Nor-E	Muy feliz	No proced	2	1	20	35	
4	Mujer	Blanca	Nor-E	No conte	Rutinaria	2	0	20	26	
5	Mujer	Negra	Nor-E	Bastante	Excitante	4	0	12	25	No
6	Hombre	Negra	Nor-E	Bastante	No proced	7	5	10	59	
7	Hombre	Negra	Nor-E	Muy feliz	Excitante	7	3	10	46	
8	Mujer	Negra	Nor-E	Bastante	No proced	7	4	16	Nn	
9										

► 12

Variables estadísticas

► Ejemplo (iii)

- ▶ Aunque se codifiquen con números, debemos recordar siempre el verdadero tipo de las variables y su significado cuando vayamos a usar programas de cálculo estadístico.
- ▶ No todo está permitido con cualquier tipo de variable.

	Nombre	Tipo	Anch	Deci	Etiqueta	Valo▲
1	sexo	Numérico	1	0	Sexo del encuestado	{1, Hombre}...
2	raza	Numérico	1	0	Raza del encuestado	{1, Blanca}...
3	región	Numérico	8	0	Región de los Estados Unidos	{1, Nor-Este}...
4	feliz	Numérico	1	0	Nivel de felicidad	{0, No proce...
5	vida	Numérico	1	0	¿Su vida es excitante o aburrida?	{0, No proce...
6	hermanos	Numérico	2	0	Número de hermanos y hermanas	{98, No sabe}...
7	hijos	Numérico	1	0	Número de hijos	{8, Ocho o m...
8	educ	Numérico	2	0	Número de años de escolarización	{97, No proce...
9	edad	Numérico	2	0	Edad del encuestado	{98, No sabe}...

► 13

Variables estadísticas

- ▶ Los posibles valores de una variable suelen denominarse **modalidades**.
- ▶ Las modalidades pueden agruparse en **clases o intervalos**
 - ▶ Edades: Menos de 20 años, de 20 a 50 años, más de 50 años
 - ▶ Hijos: Menos de 3 hijos, De 3 a 5, 6 o más hijos
- ▶ Las **modalidades/clases** o (valores/intervalos) deben formar un sistema exhaustivo y excluyente
 - ▶ **Exhaustivo:** No podemos olvidar ningún posible valor de la variable
 - **Mal:** ¿Cuál es su color del pelo? (**Rubio, Moreno**)?
 - **Bien:** ¿Cuál es su grupo sanguíneo?
 - ▶ **Excluyente:** No pueden presentarse 2 valores simultáneos de la variable
 - ▶ Estudio sobre el ocio
 - **Mal:** De los siguientes, qué le gusta: (deporte, cine)
 - **Bien:** Le gusta el deporte: (**Sí, No**)
 - **Bien:** Le gusta el cine: (**Sí, No**)
 - **Mal:** Cuántos hijos tiene: (Ninguno, Menos de 5, Más de 2)

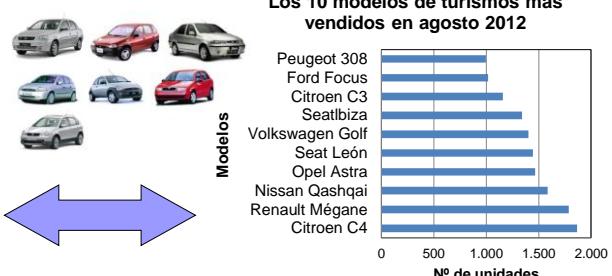
► 14

Distribuciones de frecuencia

► Tratamiento de los datos

- Deben ser ordenados para facilitar su análisis

ESPAÑA- AGOSTO 2012	
MODELO	NUMERO
Citroen C4	1.865
Renault Mégane	1.787
Nissan Qashqai	1.585
Opel Astra	1.466
Seat León	1.446
Volkswagen Golf	1.402
Seatbiza	1.341
Citroen C3	1.159
Ford Focus	1.017
Peugeot 308	998



- Las tablas de frecuencias y las representaciones gráficas son dos maneras equivalentes de presentar la información.
- Las dos exponen ordenadamente la información recogida en una muestra.
- Ejemplo:** Asoc. Nal. de Import. de Automov.: <http://www.aniacam.com/datos/default.php>

► 15

Distribuciones de frecuencia

► Clases o intervalos de clases

- Grupo de valores que describen una característica
- Se utilizan cuando existe gran cantidad de valores discretos o valores continuos

► Rango de las clases

- Es la variación entre el valor mínimo y el máximo
- El rango se divide en intervalos (suelen ser de la misma longitud)

► Ejemplo: Altura de una persona

- Rango de valores entre 1,50 m y 2,00 m
- Intervalo de clases 5 cm.
- Existen 10 clases

► 17

Distribuciones de frecuencia

► Gráficos estadísticos

- ▶ Se utilizan para:
 - Explorar los datos
 - Visualizar la forma de la distribución de los datos
 - Observar patrones o tendencias
 - Agrupar información por factores
 - Observar relaciones
 - Comparar distribuciones
 - Comparar medidas estadísticas

► 18

Distribuciones de frecuencia

► Gráficos estadísticos

- ▶ Diagrama de barras
- ▶ Diagrama de pastel



► 19

Distribuciones de frecuencia

▶ Frecuencias absolutas (f_i)

Es el número de veces que se presenta una modalidad

Ejemplo: Datos de hijos por familia

Nº DE HIJOS	Valor	Frecuencia absoluta (F)
0	26	
1	42	
2	32	
3	21	
4	14	
5	11	
6	4	
Total	150	

► 20

Distribuciones de frecuencia

▶ Frecuencias relativas (n_i)

► Cociente entre la frecuencia absoluta (f_i) y el tamaño de la muestra (N)

$$n_i = \frac{\text{frecuencia absoluta}}{\text{tamaño de la muestra}} = \frac{f_i (\text{nº de observaciones del valor } i)}{N (\text{nº de observaciones})}$$

► Es la proporción que representa la frecuencia de cada intervalo de clase en relación al total

► Es útil para comparar distribuciones de frecuencias de poblaciones distintas

▶ Porcentaje (p_i)

► Es la frecuencia relativa expresada en %:

$$\cdot P_i = n_i * 100$$

Valor	España		Portugal	
	f. Absoluta (f_i)	f. Relativa (n_i)	f. Absoluta (f_i)	f. Relativa (n_i)
0	26	0,173	6	0,060
1	42	0,28	21	0,210
2	32	0,213	26	0,260
3	21	0,14	23	0,230
4	14	0,093	13	0,130
5	11	0,073	8	0,080
6	4	0,027	3	0,030
Total	150	1.000	100	1.000

► 21

Distribuciones de frecuencia

► Frecuencia acumulada (F_i)

- ▶ Indica cuántos casos hay por debajo o arriba de un determinado valor o límite de clase.

► Frecuencia relativa acumulada (N_i)

- ▶ Es el cociente entre la Frecuencia acumulada (F_i) y el tamaño de la muestra (N)

$$N_i = \frac{F_i \text{ (frecuencia acumulada)}}{N(n^{\circ} \text{ de observaciones})}$$

► Porcentaje acumulado (P_i)

- ▶ Es la frecuencia relativa acumulada (N_i) expresada en %

NOTA IMPORTANTE: Las notaciones según textos pueden cambiar.

► 22

Distribuciones de frecuencia

► Ejemplo I (I)

Datos de hijos por familia

Nº DE HIJOS	Valor	Frecuencia absoluta (F)
	0	26
	1	42
	2	32
	3	21
	4	14
	5	11
	6	4
	Total	150

► 23

Distribuciones de frecuencia

► Ejemplo I (2)

- Para poder estimar con precisión en n° de hijos se realiza encuesta a 1.517 familias. Los resultados aparecen en la siguiente tabla:

	Nº de hijos	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0	419	27,6	27,8	27,8
	1	255	16,8	16,9	44,7
	2	375	24,7	24,9	69,5
	3	215	14,2	14,2	83,8
	4	127	8,4	8,4	92,2
	5	54	3,6	3,6	95,8
	6	24	1,6	1,6	97,3
	7	23	1,5	1,5	98,9
	Ocho o más	17	1,1	1,1	100,0
Total		1.509	99,5	100,0	
Perdidos	No contesta	8	0,5		
Total		1517	100,0		

► Determine

- 1) ¿Qué porcentaje de individuos tiene menos de 3 hijos?
 - Sol: 69,5%
(27,8+16,9+24,9)
 - 2) ¿Y entre 4 y 6 hijos?
 - Soluc 1^a forma:
8,4%+3,6%+1,6% = 13,6%.
 - Soluc 2^a forma:
97,3% - 83,8% = 13,5%
- NOTA: Porcentajes válidos

► 24

Distribuciones de frecuencia

► Ejemplo I (3)

- ¿Cuántos individuos tienen menos de 2 hijos?
 - frec. indiv. sin hijos + frec. indiv. con 1 hijo
= 419 + 255 = 674 individuos
- ¿Qué porcentaje de individuos tiene 6 hijos o menos?
 - 97,3%
- ¿Qué cantidad de hijos es tal que al menos el 50% de la población tiene una cantidad inferior o igual?
 - 2 hijos

Número de hijos

	Frec.	Porcent. (válido)	Porcent. acum.
0	419	27,8	27,8
1	255	16,9	44,7
2	375	24,9	69,5
3	215	14,2	83,8
4	127	8,4	92,2
5	54	3,6	95,8
6	24	1,6	97,3
7	23	1,5	98,9
Ocho+	17	1,1	100,0
Total	1509	100,0	

► 25

Distribuciones de frecuencia

► Tabla de frecuencia para variables continuas

- Las variables se agrupan en clases (intervalos que tienen la misma amplitud)
- La amplitud de la clase es la diferencia entre el límite superior e inferior de la clase.

- Se toma un valor, marca de clase (MC ó C) , que representa a todo el intervalo o clase para el cálculo de algunos parámetros y es el punto medio de cada intervalo
- A cada clase se le asigna su frecuencia correspondiente.

Peso	M. Clase	f	F
[40 – 50)	45	5	5
[50 – 60)	55	10	15
[60 – 70)	65	21	36
[70 – 80)	75	11	47
[80 – 90)	85	5	52
[90 – 100)	95	3	55
[100 – 130]	115	3	58

► 26

Distribuciones de frecuencia

CONSTRUCCIÓN DE UNA TABLA CON VALORES AGRUPADOS

1) Determinar el número de clases o intervalos (k) (C_1, C_2, \dots, C_k):

- Total de unidades de análisis (n)
- Regla de Sturges: $k = 1 + 3,3 \log n$ (\log en base 10)

2) Determinar amplitud del intervalo

- Valor mínimo que toma la variable en el grupo, $\min(x_i)$ $i=1, 2, \dots, n$.
- Valor máximo que toma la variable en el grupo, $\max(x_i)$ $i=1, 2, \dots, n$.
- Rango = $\max(x_i) - \min(x_i) = R$; Amplitud = $(R+1)/k = a$

3) Construir los intervalos: Límite inferior y Límite superior de cada intervalo

- L_{ij} = Límite inferior de la clase j, $j=1, 2, \dots, k$
- L_{sj} = Límite superior de la clase j, $j=1, 2, \dots, k$
- $L_{i1} = \min(x_i) - (1/2)a$
- $L_{i1} = L_{i1} + a$
- $L_{i2} = L_{i1}$
- $L_{i2} = L_{i2} + a$
-

► 27

Distribuciones de frecuencia

CONSTRUCCIÓN DE UNA TABLA CON VALORES AGRUPADOS

Elementos de una tabla de frecuencia cuando la variable es continua (x)

Intervalo	Centro de clase	Amplitud	f	F	n	N
$[L_{i1} : L_{s1})$	c_1	a_1				
$[L_{i2} : L_{s2})$	c_2	a_2				
.	.					
$[L_{ik} : L_{sk})$	c_k	a_k			N^o	I
Total				N^o	I	

$c_j = (L_{ij} + L_{sj})/2$ $a_j = (L_{sj} - L_{ij})$

► 28

Distribuciones de frecuencia

EJEMPLO: Construcción de una tabla de datos agrupados

3,9	4,1	4,2	3,2	1,6
2,5	1,1	9,1	5,1	2,7
1,9	7,3	2,4	4,9	1,6
5	2,5	6,5	1,9	5,2
6,3	1,2	3,3	1,8	4,4

Pasos para la agrupación en intervalos de clase de igual amplitud

1) Ordenar la tabla 1,1 1,2 1,6 1,6 1,8 1,9 1,9 2,4 2,5 2,5 2,7 3,2 3,3 3,9 4,1 4,2 4,4 4,9 5 5,1 5,2 6,3 6,5 7,3 9,1

2) Se calcula el rango, R, (ver Medidas de Dispersión) de las muestras:

$$Rango = R = \max(x_i) - \min(x_i) = 9,1 - 1,1, I = 8$$

3) El número de intervalos de clase que podemos tomar para agrupar los datos es:

$$k = 1 + 3/3 \log 25 = 5,64 \approx 6 \text{ (se aproxima por el número natural siguiente)}$$

4) Por tanto, la amplitud de cada intervalo es: Amplitud = $a = (R+I)/k = 1,5$

5) Construir los intervalos: Límite inferior y Límite superior de cada intervalo:

$$\begin{aligned} Li_1 &= \min(x_i) - (1/2) = 1,1 - 0,5 = 0,6; & Ls_1 &= Li_1 + a = 2,1; \\ Li_2 &= Ls_1 = 2,1; & Ls_2 &= Li_2 + a = 3,5; \\ \dots & & & \\ Li_6 &= Ls_1 = 8,1; & Ls_6 &= 9,6; \end{aligned}$$

Intervalo	Li	Ls	f	n	p	F	N
1	0,6	2,1	7	0,28	28%	7	0,28
2	2,1	3,6	6	0,24	24%	13	0,52
3	3,6	5,1	6	0,24	24%	19	0,76
4	5,1	6,6	4	0,16	16%	23	0,92
5	6,6	8,1	1	0,04	4%	24	0,96
6	8,1	9,6	1	0,04	4%	25	1
			25	1	100%		

► 29

Parámetros y estadísticos

► **Parámetro:** Es una cantidad numérica calculada sobre una población

- La altura media de los individuos de un país
- La idea es resumir toda la información que hay en la población en unos pocos números (parámetros).
- Los **parámetros estadísticos** sirven para sintetizar la información dada por una tabla o por una gráfica.

► **Estadístico:** Es una cantidad numérica calculada sobre una muestra

- La altura media de los que estamos en esta aula.
 - ¿Somos una muestra *representativa* de la población?
- Si un estadístico se usa para aproximar un parámetro también se le suele llamar **estimador**.
- NOTA:** Normalmente interesa conocer un parámetro, pero por la dificultad que conlleva estudiar a *toda* la población, se calcula un estimador sobre una muestra y se confía en que sean próximos.

► 30

TIPOS DE PARÁMETROS

► Habitualmente se agrupan los **parámetros** en las siguientes categorías:

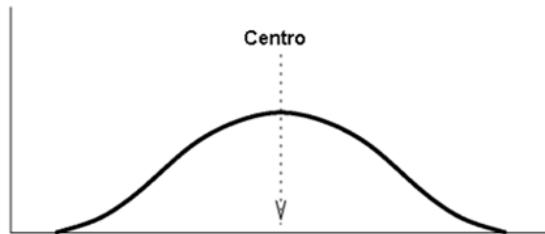
- **Medidas de centralización.**
- **Medidas de dispersión.**
- **Medidas de posición.**

► 31

Medidas

► Medidas de centralización

- ▶ Con frecuencia se hace necesario resumir los numerosos datos observados con unas pocas características numéricas (Ej: IPC)
- ▶ Se suelen situar en el centro de la distribución de datos
- ▶ Los más destacados son la media aritmética, la media geométrica , la media armónica, la mediana y la moda

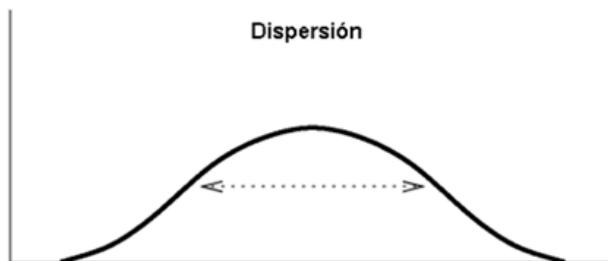


► 32

Medidas

► Medidas de dispersión

- ▶ Son indicadores de cuánto se dispersan los datos
- ▶ Indican cuánto se alejan los datos de las medidas centrales
- ▶ Rango o recorrido, varianza, desviación típica



► 33

Medidas

► **Medidas de posición**

- ▶ Las medidas de posición dividen un conjunto de datos en grupos con el mismo número de individuos.
- ▶ Para calcular las medidas de posición es necesario que los datos estén ordenados de menor a mayor.
- ▶ La medidas de posición son: Cuartiles, Deciles, Percentiles.

► 34

Medidas de centralización: Media

► **Media aritmética**

Es el valor medio de una serie de valores, es iguala la suma de los valores dividido por el número de los mismos

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_i x_i}{n}$$

► **Ejemplo**

Se tienen edades de cinco estudiantes, 27, 23, 26, 30 y 24 años

$$\bar{x} = \frac{27+23+26+30+24}{5} = \frac{130}{5} = 26$$

► 35

Medidas de centralización (Resumen)

-MEDIA ARITMÉTICA (PROMEDIO)

-MEDIANA

-MODA

DATOS CUANTITATIVOS

x
x_1
x_2
\vdots
x_n

Media Aritmética o Promedio

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

DATOS CUANTITATIVOS ORDENADOS DE MENOR A MAYOR

Mediana

$$M_E = x_{(k)} \quad \text{Si } n \text{ es impar}$$

$$M_E = \frac{x_{(k)} + x_{(k+1)}}{2} \quad \text{Si } n \text{ es par}$$

$x_{(k)}$ = dato del centro

DATOS CUALITATIVOS Y CUANTITATIVOS

Moda

M_o = "el dato que más se repite"

► 36

Medidas de centralización: Media

► Media aritmética calculada a partir de datos agrupados

- Si los datos se presentan en datos agrupados en tablas de frecuencia no es posible obtener los valores individuales
- Se suponen datos distribuidos uniformemente dentro de cada clase

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{f_1 + f_2 + \cdots + f_n} = \frac{\sum x_i f_i}{N} = \sum x_i n_i$$

► 37

Medidas de centralización: Media

► Ejemplo

Valor (x_i)	f. Absoluta (f_i)	f. Relativa (n_i)	$x_i f_i$	$x_i n_i$
0	26	0,173	0	0
1	42	0,28	42	0,28
2	32	0,213	64	0,426
3	21	0,14	63	0,42
4	14	0,093	56	0,372
5	11	0,073	55	0,365
6	4	0,027	24	0,162
Total (suma)	150	1.000	304	2,03
		Suma/N		2,03

► 38

Medidas de centralización: Media

► Ejemplo con variable en intervalos

Peso	Marca de Clase	f_i	F_i
40 – 50	45	5	5
50 – 60	55	10	15
60 – 70	65	21	36
70 - 80	75	11	47
80 - 90	85	5	52
90 - 100	95	3	55
100 – 130	115	3	58

► Para calcular la media es necesario elegir un punto representante del intervalo:

La marca de clase (MC)

► Determinando la media:

$$\bar{x} = \frac{\sum_i MC_i \cdot f_i}{N} = \frac{45 \cdot 5 + 55 \cdot 10 + \dots + 115 \cdot 3}{58} = 69,3$$

► 39

Medidas de centralización: Media

► **Media aritmética ponderada**

- ▶ Se utiliza cuando los datos a promediar tienen diferentes pesos
- ▶ Se tienen una serie de observaciones x_1, x_2, \dots, x_n ; y un conjunto de factores de ponderación por observación p_1, p_2, \dots, p_n

$$\bar{x} = \frac{x_1 p_1 + x_2 p_2 + \dots + x_n p_n}{p_1 + p_2 + \dots + p_n} = \frac{\sum x_i p_i}{\sum p_i}$$

► 40

Medidas de centralización: Media

► **Media armónica**, es la inversa de la media aritmética de las inversas de los valores de la variable

$$\bar{x}_a = \frac{n}{1/x_1 + 1/x_2 + \dots + 1/x_n} = \frac{n}{\sum 1/x_i}$$

- ▶ Se utiliza para promediar velocidades, tiempos, rendimiento, etc. (cuando influyen los valores pequeños)
- ▶ Cuando algún valor es cero, o próximo a cero no se puede calcular
- ▶ **Ejemplo:** Promediar la velocidad de un automóvil que viaja a 80Km/h en el trayecto de ida y a 120 Km/h en el de vuelta

► 41

Medidas de centralización: Media

- **Media geométrica**, es la raíz n (índice de frecuencia) cuyo radicando es el producto de las potencias de cada valor elevado a sus respectivas frecuencias absolutas (en el siguiente ej. no se elevan, sino que se repiten, pero es lo mismo)

$$\overline{x_g} = \sqrt[n]{x_1 \cdot x_1 \cdot \dots \cdot x_n}$$

- Se utiliza cuando los valores a promediar siguen una progresión geométrica. También para promediar porcentajes, tasas, nº índices, etc.

► 42

Medidas de centralización: Media

► Propiedades de la media aritmética

- Puede ser calculada de distribuciones con escala relativa y de intervalos
- Todos los valores son incluidos en el cómputo de la media
- Una serie de datos sólo tiene una media
- Es una medida muy útil para comparar dos o más poblaciones
- Conveniente cuando los datos se concentran simétricamente con respecto a ese valor
- Centro de gravedad de los datos

► Desventajas

- Si alguno de los valores es extremadamente grande o pequeño, la media no es el promedio adecuado para representar la serie de datos. Es muy sensible a valores extremos.

► 43

Medidas de centralización: Mediana

- ▶ Cuando una serie de datos contiene algún dato muy pequeño o muy grande, la media aritmética no es representativa.
- ▶ En esos casos se utiliza la **mediana**
- ▶ **Mediana:** es el valor comprendido entre el menor valor de la variable para el que su frecuencia absoluta acumulada supera o es igual a la mitad del número de elementos de la población y el menor valor de la variable para el que su frecuencia absoluta acumulada supera, estrictamente a la mitad del número de elementos de la población

► 44

Medidas de centralización: Mediana

► **Mediana (Valores discretos)**

x
$x_{(1)}$
$x_{(2)}$
\vdots
$x_{(n)}$

- ▶ Datos Cuantitativos ordenados de menor a mayor
- ▶ Mediana es un valor que divide a las observaciones en dos grupos con el mismo número de individuos
- ▶ Si n es impar: $M_E = x_{(k)}$ donde $x_{(k)}$ = dato del centro
- ▶ Si n es par: $M_E = \frac{x_{(k)} + x_{(k+1)}}{2}$ media de los dos datos centrales
- ▶ **Ejemplo:**
 - ▶ Mediana de 1, 2, 4, 5, 6, 6, 8 es 5
 - ▶ Mediana de 1, 2, 4, 5, 6, 6, 8, 9 es $(5+6)/2=5,5$
 - ▶ Mediana de 1, 2, 4, 5, 6, 6, 800 es 5. ¡La media es 117,7!!

► 45

Medidas de centralización: Mediana

► Ejemplo 1

Valor	f. Absoluta (f_i)	f. Relativa (n_i)	f. Relativa acumulada (F_i)
0	26	0,173	0,173
1	42	0,28	0,453
2	32	0,213	0,667
3	21	0,14	0,807
4	14	0,093	0,9
5	11	0,073	0,973
6	4	0,027	1
Total	150	1	

► 46

Medidas de centralización: Mediana

► Mediana (Valores continuos y datos agrupados)

- 1) Se determina la clase de la mediana o la clase por debajo de la cual se encuentra menos del 50% de los datos
- 2) Dentro de clase se determina el valor exacto de la mediana mediante la expresión:

$$M_E = x_{med} = L_i + (L_i - L_{i-1}) \cdot \frac{\frac{N}{2} - F_{i-1}}{f_i}$$

donde:

L_i : Límite inferior de la clase mediana

F_{i-1} : Frecuencia acumulada hasta la clase mediana

N : Conjunto de datos

$L_i - L_{i-1}$: Anchura de la clase (C)

f_i : Frecuencia absoluta de la clase mediana

► 47

Medidas de centralización: Mediana

► Cálculo de la Mediana (Ejemplo 2):

	f_i	F_i
[60, 63)	5	5
[63, 66)	18	23
[66, 69)	42	65
[69, 72)	27	92
[72, 75]	8	100
	100	

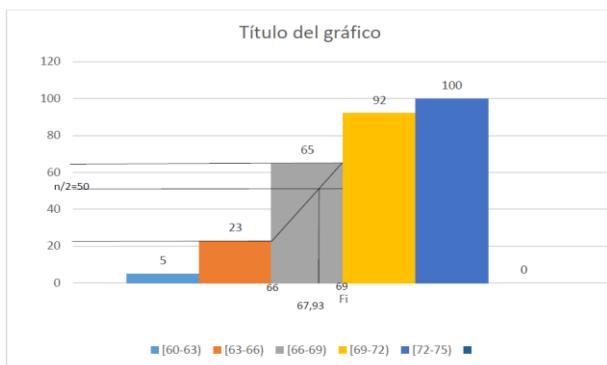
- Clase mediana [66,69)
- $L_i=66$
- $F_{i-1}=23$
- $N=100$
- $C=3$
- $f_i=42$

$$M_E = x_{med} = 66 + \frac{50 - 23}{42} 3 = 67,93$$

► 48

Medidas de centralización: Mediana

► Cálculo de la Mediana (Ejemplo):



$$x_{med} = 66 + \frac{50 - 23}{42} 3 = 67.93$$

► 49

Medidas de centralización: Media, Mediana y Robustez

► Comparación de la Media y la Mediana: Robustez

- ✓ Los datos atípicos son datos extremos o lejanos de la mayoría de las observaciones.
- ✓ La media y la mediana tienen un comportamiento diferente frente a los datos atípicos
- ✓ La media en su calculo considera todos los datos, incluyendo los datos atípicos.
- ✓ La mediana es una medida que se ve poco afectada por los datos atípicos, no los considera en su calculo dado que separa los datos
- **la mediana es una medida robusta en comparación con la media.**

► 50

Medidas de centralización: Moda

► Moda

- Moda es el valor más frecuente de la variable
- La moda se puede determinar en todo tipo de variables
- No se ve afectada por valores extremos
- En muchas series de datos no hay modas porque ningún valor aparece más de una vez
- En algunas series de datos existe más de una moda

► 51

Medidas de centralización: Moda

► Cálculo de la moda en datos agrupados

$$Mo = L_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} C$$

- L_i = Límite inferior de la clase modal (clase de mayor frecuencia)
- Δ_1 = Diferencia de frecuencias absolutas de la clase modal y la premodal
- Δ_2 = Diferencia de frecuencias absolutas de la clase modal y la postmodal
- C = Amplitud de la clase modal

► 52

Medidas de centralización: Moda

► Cálculo de la moda (Ejemplo 2)

► Datos no agrupados

x_i	61	64	67	70	73
f_i	5	18	42	27	8

- $Mo = 67$

► 53

Medidas de centralización: Moda

► Cálculo de la moda (ejemplo 2)

► Datos agrupados

	f_i	F_i
[60, 63)	5	5
[63, 66)	18	23
Clase modal → [66, 69)	42	65
[69, 72)	27	92
[72, 75]	8	100
	100	

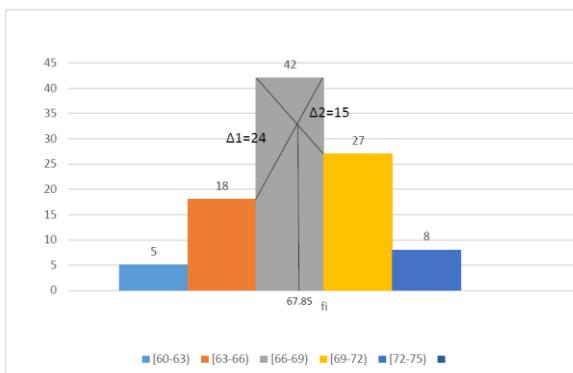
- Clase modal [66,69)
- $L_i = 66$
- $\Delta_1 = 24 = 42 - 18$
- $\Delta_2 = 15 = 42 - 27$
- $C = 3$

$$Mo = 66 + \frac{24}{24+15} \cdot 3 = 67,85$$

► 54

Medidas de centralización: Moda

► Cálculo de la moda, ejemplo



$$Mo = 66 + \frac{24}{24+15} \cdot 3 = 67.85$$

► 55

Medidas de centralización: Moda

► Cálculo de la moda (ejemplo 3)

► Datos agrupados

salario k€)	nº empleados
[10.000; 15.000)	100
[15.000; 20.000)	200
[20.000; 25.000)	200
[25.000; 30.000)	300
[30.000; 35.000)	400
[35.000; 40.000)	1.000
[40.000; 45.000)	800
[45.000; 50.000]	200

• Clase modal [35.000; 40.000)

• $L_i = 35.000$

• $\Delta_1 = 1.000 - 400 = 600$

• $\Delta_2 = 1.000 - 800 = 200$

• $C = 5.000$

$$Mo = 35.000 + \frac{600}{600 + 200} 5.000 = 38.750 \text{ €}$$

► 56

Medidas de centralización

► Ejemplo

► Media aritmética de datos agrupados

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \dots + x_n f_n}{f_1 + f_2 + \dots + f_n} = \frac{\sum x_i f_i}{N} = \sum x_i n_i$$

i	x_i	frecuencia (f_i)	frec. Relativa (n_i)
1	0	40	0,44
2	1	26	0,29
3	2	14	0,16
4	3	6	0,07
5	4	3	0,03
6	5	0	0,00
7	6	1	0,01
TOTAL		90	1,00

Media aritmética de datos agrupados

i	$x_i \cdot n_i$	$x_i \cdot f_i$
1	0,00	0
2	0,29	26
3	0,31	28
4	0,20	18
5	0,13	12
6	0,00	0
7	0,07	6
media=	1	1

MEDIANA

$$\begin{aligned} fr(x \leq 1) &= 0,73 > 0,5 \\ fr(x=0) &= 0,44 < 0,5 \\ fr(x \geq 1) &= 0,44 < 0,5 \end{aligned}$$

$$\text{mediana=} \quad 1$$

MODA

$$x_i=0 \quad 0,44 \quad 0,44 > \text{resto}$$

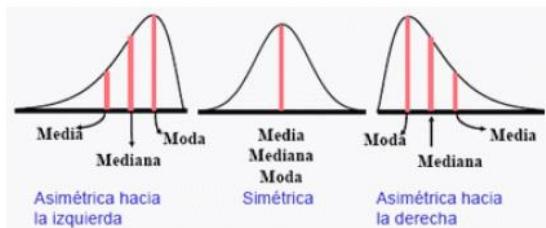
$$\text{moda=} \quad 0$$

► 57

Medidas de centralización: Media, mediana, moda

- ▶ Si las distribuciones de datos son totalmente simétricas, la media, la mediana y la moda coinciden.
- ▶ En distribuciones moderadamente asimétricas, se mantiene aproximadamente la siguiente relación:

$$\text{Media} - \text{Moda} = 3(\text{Media} - \text{Mediana})$$



► 58

Medidas de dispersión

- RANGO
- VARIANZA
- DESVIACIÓN ESTÁNDAR

x
x_1
x_2
\vdots
x_n

Rango

$$R = \max(x_i) - \min(x_i)$$

Varianza

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

Desviación Estándar

$$s = \sqrt{s^2}$$

Comparación entre Variables

Se refiere al comportamiento de las variables cuantitativas en un grupo

Coeficiente de Variación

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

► 59

Medidas de dispersión: Varianza

- ▶ Es la media de los cuadrados de las desviaciones de los datos respecto a la media
 - ▶ Cálculo de la varianza para el conjunto de los datos

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

- ▶ Para datos agrupados

$$S^2 = \frac{\sum f_i(x_i - \bar{x})^2}{N} = \frac{\sum f_i x_i^2}{N} - \bar{x}^2$$

- ▶ O empleando las frecuencias relativas:

$$S^2 = \sum n_i(x_i - \bar{x})^2 = \sum n_i \cdot x_i^2 - \bar{x}^2$$

► 60

Medidas de dispersión: Desviación típica

- ▶ Es la raíz cuadrada positiva de la varianza
 - ▶ Cálculo de la desviación típica para el conjunto de los datos

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

- ▶ Para datos agrupados

$$S = \sqrt{\frac{\sum f_i(x_i - \bar{x})^2}{N}} = \sqrt{\frac{\sum f_i x_i^2}{N} - \bar{x}^2}$$

- ▶ O empleando las frecuencias relativas:

$$S = \sqrt{\sum n_i \cdot (x_i - \bar{x})^2} = \sqrt{\sum n_i \cdot x_i^2 - \bar{x}^2}$$



► 61

Medidas de dispersión: Varianza

▶ Propiedades de la varianza

- ▶ Es sensible a valores extremos alejados de la media
- ▶ Sus unidades son el cuadrado de las de la variable lo que hace que las grandes diferencias se destaque
- ▶ Elevar cada diferencia al cuadrado hace que todos los números sean positivos (para evitar que los números negativos reduzcan la varianza)

▶ Ejemplo:

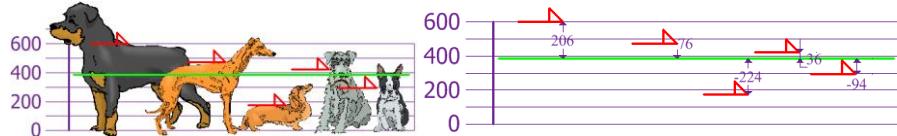
- La varianza contiene la información geométrica relevante en muchas situaciones donde la energía interna de un sistema depende de la posición de sus partículas.
 - Energía de rotación (coeficiente de inercia): ej. patinadores con brazos extendidos (dispersos) o recogidos (poco dispersos)
 - Energía elástica: Muelles 'estirados' con respecto a su posición de equilibrio

▶ 62

Medidas de dispersión: Varianza

▶ Ejemplo: Cálculo de la media, la varianza y la desviación estándar

- ▶ Se ha medido la altura en mm de cinco perros hasta sus hombros obteniendo los siguientes resultados: 600 mm, 470 mm, 170 mm, 430 mm y 300 mm.
- Media = $(600 + 470 + 170 + 430 + 300)/5 = 1970/5 = 394 \text{ mm}$



- Para calcular **la varianza**, se calcula la diferencia de cada altura con la media se eleva al cuadrado, y se hace la media:
- Varianza= $S^2 = [206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2] = 108.520/5 = 21.704 \text{ mm}^2$

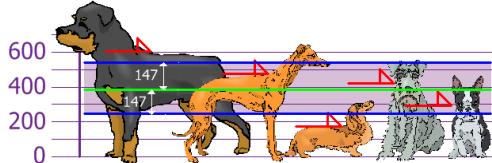
▶ 63

Medidas de dispersión: Varianza

- Ejemplo....: Calculo de la media, la varianza y la desviación estándar

- La **desviación estándar** es la raíz de la varianza:

- Desviación estándar: $S = \sqrt{21.704} = 147 \text{ mm}$
- La utilidad de la desviación estándar es comparar si las alturas están, por ejemplo, a una distancia ($\pm S$) menor que la desviación estándar (147mm) de la media.



- Los Rottweilers son perros grandes, y los Dachshunds son un poco menudos.

► 64

Medidas de dispersión: Coeficiente de variación

- Coeficiente de variación: se define como el cociente:

$$CV = \frac{s}{|\bar{x}|}$$

donde se supone $\bar{x} \neq 0$, $|\bar{x}|$ es el valor absoluto de \bar{x} , de manera que $CV > 0$

- El coeficiente de variación es una medida relativa de variabilidad.
- En ingeniería se utiliza mucho el *coeficiente inverso*, $|\bar{x}|/s$ (*coeficiente señal-ruido*)
- Para datos que representen distintas mediciones de una misma magnitud, CV indica la magnitud del error promedio de medición, s , como % de la cantidad medida

► 65

Medidas de dispersión

► Ejemplo

Distribución de frecuencias de la variable:
Nº de llamadas recibidas en una centralita en 1 minuto

<i>i</i>	x_i	frecuencia (f_i)	frec. Relativa (n_i)
1	0	40	0,44
2	1	26	0,29
3	2	14	0,16
4	3	6	0,07
5	4	3	0,03
6	5	0	0,00
7	6	1	0,01
TOTAL		90	1,00

VARIANZA	
x_i	$(x_i - \text{media})^2 \cdot n_i$
0	0,44
1	0,00
2	0,16
3	0,28
4	0,27
5	0,00
6	0,25
VARIANZA	
	1,4

► 66

Medidas de dispersión: MEDA

- MEDA es la Mediana de las Desviaciones Absolutas y es una medida de dispersión asociada a la mediana, Med, definida por:

$$\text{MEDA} = \text{mediana} |x_i - M_E|$$

- MEDA tiene la ventaja de no verse afectada por datos extremos (medida robusta o resistente)
- Si conocemos la mediana y la MEDA de datos no agrupados, sabemos que, al menos, el 50% de los datos están en el intervalo ($\text{Med} \pm \text{MEDA}$).

► 67

Medidas de dispersión: Rango

- ▶ Rango o recorrido de una variable es la diferencia entre su valor máximo y mínimo.

$$R = \max(x_i) - \min(x_i)$$

- ▶ Ejemplo

- ▶ Los siguientes datos representan el tiempo (en segundos) que 30 trabajadores estuvieron al control de la unidad central de procesos (CPU) de una computadora mainframe grande.

Tiempo de acceso (segundos) a un cluster de CPUs en un periodo de 1 minuto por parte de procesos ejecutados multitarea por diferentes terminales						
	Terminal 1	Terminal 2	Terminal 3	Terminal 4	Terminal 5	Terminal 6
Sala 1	0,02	0,75	1,16	1,38	1,94	3,07
Sala 2	0,15	0,82	1,17	1,4	2,01	3,53
sala 3	0,19	0,84	1,19	1,42	2,16	3,76
Sala 4	0,47	0,92	1,22	1,59	2,41	4,5
Sala 5	0,71	0,96	1,23	1,61	2,59	4,75

▶ $R = 4.75 - 0.02 = 4.73\text{s}$

► 68

Otras medidas de dispersión

- ASIMETRÍA
- KURTOSIS O APUNTAMIENTO

Además de la **posición** y la **dispersión** de los datos, otra medida de interés en una distribución de frecuencias es la simetría/asimetría y el apuntamiento o Kurtosis.

Coeficiente de Asimetría

$$CA = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{N \cdot s^3}$$

{ Si $CA=0$ si la distribución es simétrica alrededor de la media.
Si $CA<0$ si la distribución es asimétrica a la izquierda
Si $CA>0$ si la distribución es asimétrica a la derecha

Coeficiente de Apuntamiento

$$Cap = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{N \cdot s^4}$$

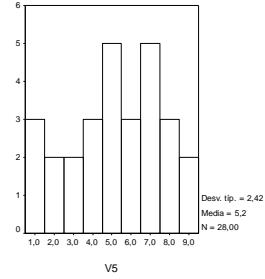
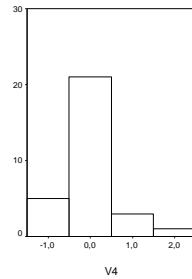
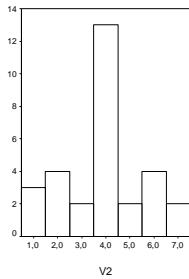
{ - Si $Cap=3$ la distribución se dice normal (similar a la distribución normal de Gauss) y recibe el nombre de **mesocúrtica**.
- Si $Cap>3$, la distribución es más puntiaguda que la anterior y se llama **leptocúrtica**, (mayor concentración de los datos en torno a la media).
- Si $Cap<3$ la distribución es más plana y se llama **platicúrtica**.

► 69

Otras medidas de dispersión

Otras medidas o Coeficientes de asimetría y apuntamiento

✓ Ejemplos



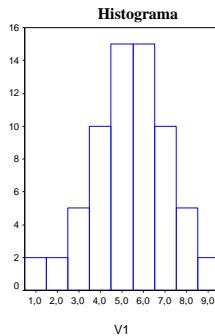
► 70

Otras medidas de dispersión

Otras medidas o Coeficientes de asimetría y apuntamiento

✓ Ejemplos

Datos
1 4 4
1 4 4
1 4 5
2 4 5
2 4 6
2 4 6
2 4 6
3 4 6
3 4 7
4 4 7



Medidas descriptivas	
Media	3,9
Error típico	0,30
Mediana	4
Moda	4
Desviación estándar	1,67
Varianza de la muestra	2,78
kurtosis	2,28
Coeficiente de asimetría	-0,02
Rango	6
Mínimo	1
Máximo	7
Cuenta	30

► 71

Otras medidas de dispersión

Otras medidas o Coeficientes de asimetría y apuntamiento

Tabla de frecuencia

Intervalo	MC	Amplitud	f	n
I ₁	c ₁	a ₁	f ₁	n ₁
I ₂	c ₂	a ₂	f ₂	n ₂
.	⋮	⋮	⋮	⋮
I _k	c _k	a _k	f _k	n _k
Total			n	1

Sea MC_j la marca de clase y f_j la frecuencia relativa de la clase j, donde j=1, 2,..., k.

El **Coeficiente de Asimetría** para datos agrupados esta dado por:

$$CA_c = \frac{\sum_{j=1}^k (c_j - \bar{x}_c)^3 f_j}{N \cdot s_c^3}$$

El **Coeficiente de apuntamiento** para datos agrupados esta dada por:

$$CAp_c = \frac{\sum_{j=1}^k (c_j - \bar{x}_c)^4 f_j}{N \cdot s_c^4}$$

► 72

Medidas de posición

- Las **medidas de posición** dividen un conjunto de datos en grupos con el mismo número de individuos.
- Para calcular las medidas de posición es necesario que los datos estén ordenados de menor a mayor.
- La **medidas de posición** son:
 - Cuartiles
 - Deciles
 - Percentiles
 - Cuantiles

► 73

Medidas de posición: Percentiles, Deciles o Cuartiles

Percentil, Decil o Cuartil: corresponde al valor que toma la variable (cuantitativa), cuando los n datos están ordenados de Menor a Mayor

El Percentil va de 1 a 100

El percentil 25 (25/100): es el valor de la variable que reúne al menos el 25% de los datos

Ejemplo: Si $N=80$, el 25% de 80 es 20; por lo tanto, se busca el dato que este en la posición 20.

Si $N=85$, el 25% de 85 es 21,25; por lo tanto se busca el dato que este en la posición 22.

El Decil va de 1 a 10

El Decil 4 (4/10): es el valor de la variable que reúne al menos el 40% de los datos

Ejemplo: Si $N=80$, el 40% de 80 es 32; por lo tanto, se busca el dato que este en la posición 32.

Si $N=85$, el 40% de 85 es 34; por lo tanto se busca el dato que este en la posición 34.

El Cuartil va de 1 a 4

El Cuartil 3 (3/4): es el valor de la variable que reúne al menos el 75% de los datos

Ejemplo: Si $N=80$, el 75% de 80 es 60; por lo tanto, se busca el dato que este en la posición 60.

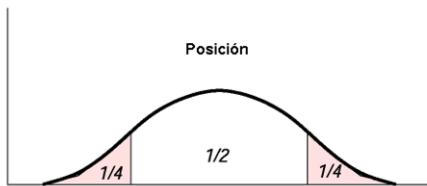
Si $N=85$, el 75% de 85 es 63,75; por lo tanto se busca el dato que este en la posición 64.

► 74

Medidas de posición: Cuartiles

► **Cuartiles.** Cuando se divide a la serie en cuatro partes

- Los **cuartiles** son los **tres valores** de la variable que **dividen a un conjunto de datos ordenados en cuatro partes**.
- Primer cuartil. Valor tal que la cuarta parte de los elementos de la población presentan valores de la variable menores que él
- Lo mismo para el segundo, tercero y cuarto cuartil
- **Q_1 , Q_2 y Q_3** determinan los valores correspondientes al **25%, al 50% y al 75% de los datos**
- **Q_2 coincide con la mediana.**



► 75

Medidas de posición: Cuartiles

► Cálculo de los cuartiles

1. Ordenamos los datos de menor a mayor.
2. Buscamos el lugar que ocupa cada **cuartil** mediante la expresión: $K^*N/4$ donde $K = 1, 2, 3$

Número impar de datos	Número par de datos
2, 5, 3, 6, 7, 4, 9	2, 5, 3, 4, 6, 7, 1, 9
2, 3, 4, 5, 6, 7, 9 ↓ ↓ ↓ Q_1 Q_2 Q_3	1, <u>2, 3</u> , <u>4, 5</u> , <u>6, 7</u> , 9 2.5 4.5 6.5 ↓ ↓ ↓ Q_1 Q_2 Q_3

► 76

Medidas de posición: Cuartiles y Rango Intercuatílico

► El Cuartil va de 1 a 4

El Cuartil 3 (3/4): es el valor de la variable que reúne al menos el 75% de los datos

Ejemplo:

- Si $N=80$, el 75% de 80 es 60; por lo tanto, se busca el dato que este en la posición 60.
- Si $N=85$, el 75% de 85 es 63,75; por lo tanto se busca el dato que este en la posición 64

► Rango Intercuatílico (RI):

es la diferencia entre el tercer y el primer cuartil $RI = Q_3 - Q_1$

► 77

Medidas de posición: Cuartiles

► Cuartiles datos agrupados:

• Cálculo de los cuartiles para datos agrupados

En primer lugar buscamos la **clase** donde se encuentra $k^*N/4$ donde $k=1,2,3$, en la **tabla de las frecuencias acumuladas**.

$$Q_k = L_{i-1} + \frac{\frac{k.N}{4} - F_{i-1}}{f_i} \cdot a_i \quad k = 1, 2, 3$$

L_{i-1} es el límite inferior de la clase donde se encuentra $k^*N/4$.

N es la suma de las frecuencias absolutas.

F_{i-1} es la **frecuencia acumulada** anterior a la clase donde se encuentra $k^*N/4$.

a_i es la amplitud de la clase.

► 78

Medidas de posición: Deciles

► Los **deciles** son los **nueve valores** que dividen la **serie de datos** en **diez partes iguales**.

► Los deciles dan los valores correspondientes al 10%, al 20%... y al 90% de los datos. D_5 coincide con la mediana.

► Deciles de datos agrupados:

► 1º se busca la clase donde se encuentra $k^*N/10$ donde $k = 1,2,\dots,9$, en la tabla de las frecuencias acumuladas.

$$D_k = L_{i-1} + \frac{\frac{k.N}{10} - F_{i-1}}{f_i} \cdot a_i \quad k = 1, 2, 3, \dots, 9$$

• L_{i-1} es el límite inferior de la clase donde se encuentra $K^*N/10$.

• N es la suma de las frecuencias absolutas.

• F_{i-1} es la frecuencia acumulada anterior donde se encuentra $K^*N/10$.

• a_i es la amplitud de la clase.

► 79

Medidas de posición: Deciles

► El Decil va de 1 a 10

El Decil 4 (4/10): es el valor de la variable que reúne al menos el 40% de los datos

Ejemplo:

- Si N=80, el 40% de 80 es 32; por lo tanto, se busca el dato que este en la posición 32.
- Si N=85, el 40% de 85 es 34; por lo tanto se busca el dato que este en la posición 34

► 80

Medidas de posición: Percentiles

- El percentil p% es un valor tal que el p% de los elementos de la población presentan valores de la variable menores que él
- Los percentiles son los 99 valores que dividen la serie de datos en 100 partes iguales. Los percentiles dan los valores correspondientes al 1%, al 2%... y al 99% de los datos. P_{50} coincide con la mediana.

► Percentiles de datos agrupados:

- En primer lugar buscamos la clase donde se encuentra $k \cdot N / 100$ donde $k = 1, 2, \dots, 99$, en la tabla de las frecuencias acumuladas.

$$P_k = L_{i-1} + \frac{\frac{k \cdot N}{100} - F_{i-1}}{f_i} \cdot a_i \quad = 1, 2, 3, \dots, 99$$

- L_{i-1} es el límite inferior de la clase donde se encuentra la mediana.
- N es la suma de las frecuencias absolutas.
- F_{i-1} es la frecuencia acumulada anterior a la clase mediana.
- a_i es la amplitud de la clase.

► 81

Medidas de posición

► El Percentil va de 1 a 100

El percentil 25 (25/100): es el valor de la variable que reúne al menos el 25% de los datos

Ejemplo:

- Si N=80, el 25% de 80 es 20; por lo tanto, se busca el dato que este en la posición 20.
- Si N=85, el 25% de 85 es 21,25; por lo tanto se busca el dato que este en la posición 22.

► 82

Medidas de posición

► Cuantil de orden α

Variable	f	F
$L_0 - L_1$	x_1	f_1
$L_1 - L_2$	x_2	f_2
...		
$L_{k-1} - L_k$	x_k	f_k
N		

- i es el menor intervalo que tiene frecuencia acumulada superior a $\alpha \cdot N$
- $\alpha=0,5$ es mediana

$$C_\alpha = L_{i-1} + (L_i - L_{i-1}) \cdot \frac{\alpha \cdot N - F_{i-1}}{f_i}$$

► 83

Medidas de posición: Ejemplo

► Cálculo el percentil 75 (75/100) o cuartil 3 (3/4)

► Datos agrupados

M. Clase	f	Fa
40 - 50	45	5
50 - 60	55	10
60 - 70	65	21
70 - 80	75	36
80 - 90	85	52
90 - 100	95	55
100 - 130	115	58

- $\alpha N = 0,75 \cdot 58 = 43,5$
- $F_i > 43,5? \rightarrow F_i = 47; f_i = 11 \rightarrow MC = 75$
- $F_{i-1} = 36$
- $L_{i-1} = 70; L_i = 80$

$$P_{75} = C_{0,75} = L_{i-1} + \frac{0,75 \cdot 58 - F_{i-1}}{f_i} (L_i - L_{i-1}) = 70 + \frac{43,5 - 36}{11} (80 - 70) = 76,8$$

► 84

Medidas de posición: Ejemplo

► Cálculo el percentil 50 (50/100) o cuartil 2 (2/4)

► Datos agrupados

M. Clase	f	Fa
40 - 50	45	5
50 - 60	55	10
60 - 70	65	21
70 - 80	75	36
80 - 90	85	52
90 - 100	95	55
100 - 130	115	58

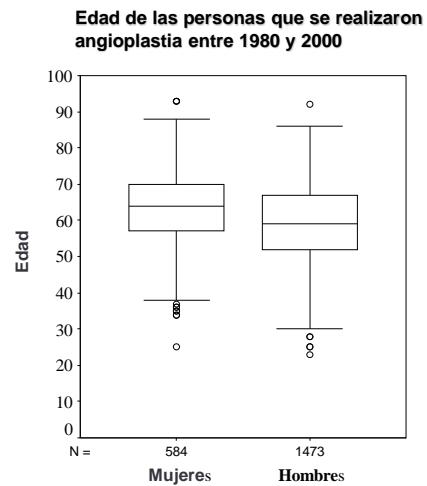
- $\alpha N = 0,5 \cdot 58 = 29$
- $F_i > 29? \rightarrow F_i = 36; f_i = 21 \rightarrow MC = 65$
- $F_{i-1} = 15$
- $L_{i-1} = 60; L_i = 70$

$$P_{50} = C_{0,5} = L_{i-1} + \frac{0,5 \cdot 58 - f_{i-1}}{f_i} (L_i - L_{i-1}) = 60 + \frac{36 - 29}{21} (70 - 60) = 66,6$$

► 85

Diagrama de caja

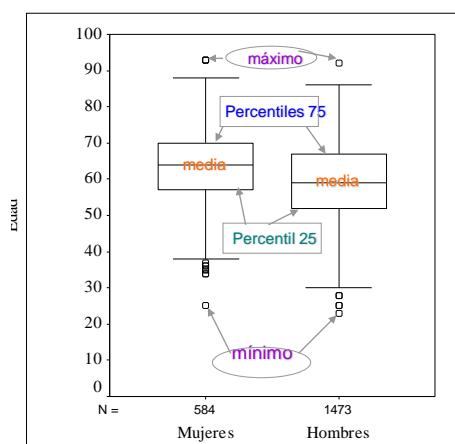
- ▶ Permite identificar gráficamente la media, los percentiles 25 y 75, mínimo y máximo de una variable.
- ▶ Sólo es útil para variables cuantitativas.
- ▶ El eje x permite identificar la población en estudio.
- ▶ El eje y representa los valores de la variable en estudio.



► 86

Diagrama de caja

Edad de las personas que se realizaron angioplastia entre 1980 y 2000

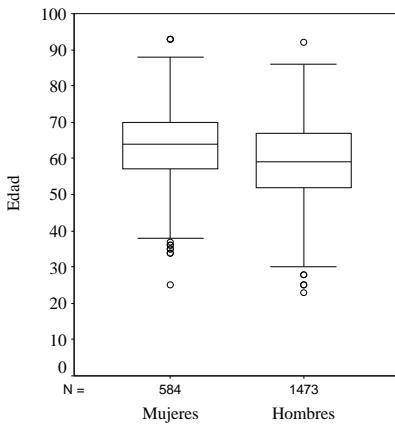


- Permite identificar gráficamente la **media**, los **percentiles 25 y 75**, **mínimo** y **máximo** de una variable.
- Sólo es útil para variables **cuantitativas**.
- El **eje x** permite identificar la población en estudio.
- El **eje y** representa los valores de la variable en estudio.

► 87

Diagrama de caja

- Edad de las personas que se realizaron angioplastia entre 1980 y 2000



Medidas Descriptivas	Mujeres	Hombres
N	584	1473
Media (o promedio)	63,3	59,2
Varianza	109,6	111,9
Desv. Típica (o Desv. Estándar)	10,5	10,6
Coeficiente Variación	0,2	0,2
Mínimo	25	23
Percentil 25	57	52
mediana	64	59
Percentil 75	70	67
Máximo	93	92
Moda	66	56

► 88

Transformaciones lineales

- ▶ El objetivo de la estadística descriptiva es obtener una visión lo más clara y simple posible y las unidades de medida de la variable deben escogerse con este criterio.
- ▶ Ejemplo:
 - Si x es la estatura en metros y se han observado los valores [1,75; 1,68; 1,80; ...]; con 1,65 como menor valor
 - Si realizamos la transformación $y = 100(x - 1,65)$ conduce al conjunto de datos: [10; 3; 15; ...] de tratamiento más simple.
- ▶ Conviene en la descripción de los datos representarlos con 2 ó 3 dígitos, escogiendo apropiadamente las unidades.
- ▶ Esto equivale a efectuar una transformación lineal:

$$y = a + bx$$

► 89

Transformaciones lineales

- ▶ Las medidas características de la variable transformada, y , se obtienen fácilmente a partir de las calculadas para la variable original, x .

$$\bar{y} = \frac{\sum y}{N} = \frac{\sum(a + bx)}{N} = a + b\bar{x}$$

$$S_y^2 = b^2 S_x^2$$

$$S_y = |b| S_x$$

▶ 90

Transformaciones lineales

Ejemplo

- ▶ Dos compañías aseguradoras tienen formas diferentes de pagar a sus empleados. La compañía A lo hace mediante un sueldo fijo mensual y la compañía B a través de un porcentaje sobre los seguros realizados. La distribución de los salarios por categorías es:

Compañía A		Compañía B	
Sueldo (€)	Nº empleados	Sueldo (€)	Nº empleados
500-800	35	500-800	21
800-1000	21	800-1000	25
1000-1500	14	1000-1400	34
		1400-2000	15

1. Por término medio, ¿gana más un empleado de la compañía A ó de la B?
2. Calcular y comentar la representatividad de los sueldos medios.
3. ¿Cuál es el sueldo más frecuente en la compañía A?
4. Si en la compañía B el salario fuese el anterior más un extra fijo de 1000 €, ¿Cuál sería el salario medio y la desviación típica?

▶ 91

Transformaciones lineales

► Ejemplo (solución)

1. Por término medio, ¿gana más un empleado de la compañía A ó de la B?

Compañía A			Compañía B		
Sueldo (€)	Marca (€)	Nº empleados	Sueldo (€)	Marca (€)	Nº empleados
500-800	650	35	500-800	650	21
800-1.000	900	21	800-1.000	900	25
1000-1.500	1.250	14	1.000-1.400	1.200	34
			1.400-2.000	1.700	15

$$\bar{x}_A = 845 \text{ €}$$

$$\bar{x}_B = 1.078,42 \text{ €}$$

$$\bar{x}_B > \bar{x}_A$$

¡Gana más, de media, un empleado de la compañía B!

► 92

Transformaciones lineales

► Ejemplo (solución)

2. Calcular y comentar la representatividad de los sueldos medios.

Compañía A			
Sueldo (€)	Marca (€)	Nº empleados (f _i)	x _i ² *f _i (€ ²)
500-800	650	35	14.787.500,00
800-1.000	900	21	17.010.000,00
1000-1.500	1.250	14	21.875.000,00
	Tot	70	53.672.500,00

Media =	845,0 €
---------	---------

Var _A = S _A ²	52.725,0 € ²
S _A = desviación	229,6 €

► 93

Transformaciones lineales

► Ejemplo (solución)

2. Calcular y comentar la representatividad de los sueldos medios.

Compañía B			
Sueldo (€)	Marca (€)	Nº empleados (f_i)	$x_i^2 * f_i$ (€ ²)
500-800	650	21	8.872.500,00
800-1.000	900	25	20.250.000,00
1.000-1.400	1.200	34	48.960.000,00
1.400-2.000	1.700	15	43.350.000,00
	Tot	95	121.432.500,00

$$\text{Media} = 1.078,4 \text{ €}$$

$$\begin{aligned} \text{Var}_B &= S_B^2 \text{ varianza} & 115.244,9 \text{ €}^2 \\ S_B &= \text{desviación} & 339,5 \text{ €} \end{aligned}$$

$S_A < S_B \rightarrow$ menor dispersión de los sueldos en la empresa A

► 94

Transformaciones lineales

► Ejemplo (solución)

3. ¿Cuál es el sueldo más frecuente en la compañía A?

Compañía A			
Sueldo (€) (x_i)	Marca (€)	Nº empleados (f_i)	ACUMULADO (F_i)
500-800	650	35	35
800-1.000	900	21	56
1000-1.500	1.250	14	70
	Tot	70	

Clase modal [500-800], en el centro del intervalo tenemos la marca : Moda = 650 €

► 95

Transformaciones lineales

► Ejemplo (solución)

3. Si en la compañía B el salario fuese el anterior más extra fijo de 100 €, ¿Cuál sería el salario medio y la desviación típica?

$$Y_B = X_B + 100$$

$$\bar{Y}_B = \bar{X}_B + 100 = 1.078,4 + 100 = 1.178,4 \text{€}$$

$$S_{Y_B}^2 = S_{X_B}^2 \rightarrow S_{Y_B} = S_{X_B} = 339,5 \text{€}$$

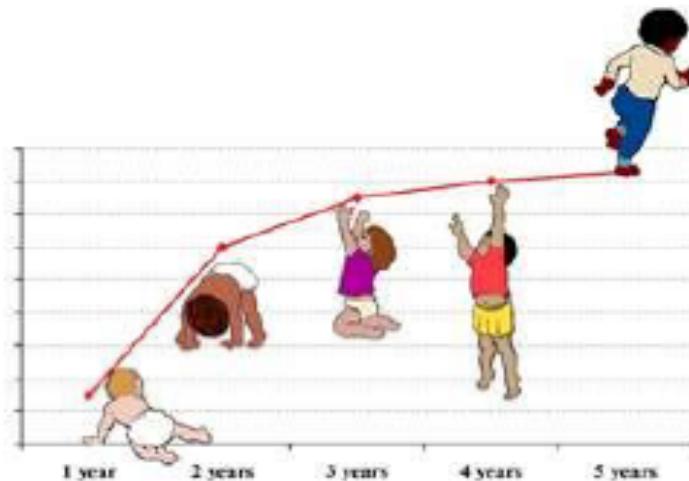
Estadística

Tema 2

T2. Descripción conjunta de varias variables

Introducción

- ▶ Es habitual encontrar en la práctica una relación entre dos o más variables deseando expresar esta relación de forma matemática.
 - Ejemplo: relación entre pesos y alturas
- ▶ En este tema se va a describir de diferentes formas la relación entre dos variables cuando éstas son numéricas.
 - peso de una persona conociendo su altura y contorno de cintura
- ▶ El estudio conjunto de dos variables cualitativas lo aplazamos hasta que veamos contrastes de hipótesis.
 - ▶ ¿Hay relación entre fumar y padecer enfermedad de pulmón?



T2. Descripción conjunta de varias variables

- ▶ El primer paso es **coleccionar los datos si x representan las alturas e y los pesos: $(x_1, y_1), (x_2, y_2) \dots \dots (x_n, y_n)$.**
- ▶ El segundo, es **representarlos en un sistema de coordenadas y el conjunto resultante se llama *diagrama de dispersión o nube de puntos (scatterplot)***

EJEMPLO

- ▶ A la derecha tenemos una posible manera de recoger los datos obtenidos observando dos variables en varios individuos de una muestra.
 - En cada *fila* tenemos los datos de un individuo
 - Cada *columna* representa los valores que toma una variable sobre los mismos.
 - Las individuos no se muestran en *ningún orden particular*.

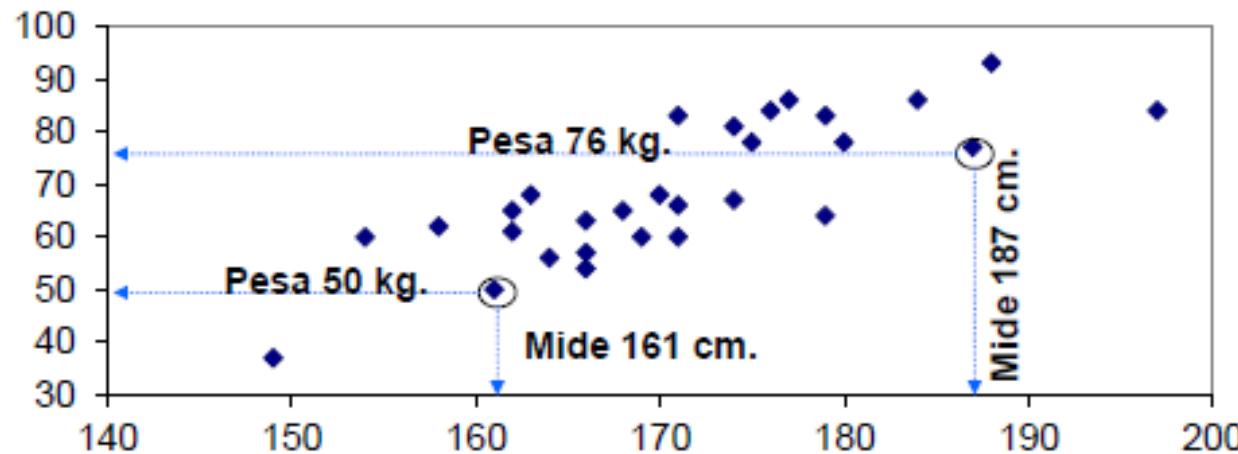
Altura en cm.	Peso en Kg.
162	61
154	60
180	78
158	62
171	66
169	60
166	54
176	84
163	68
...	...

T2. Descripción conjunta de varias variables

Diagramas de dispersión

EJEMPLO

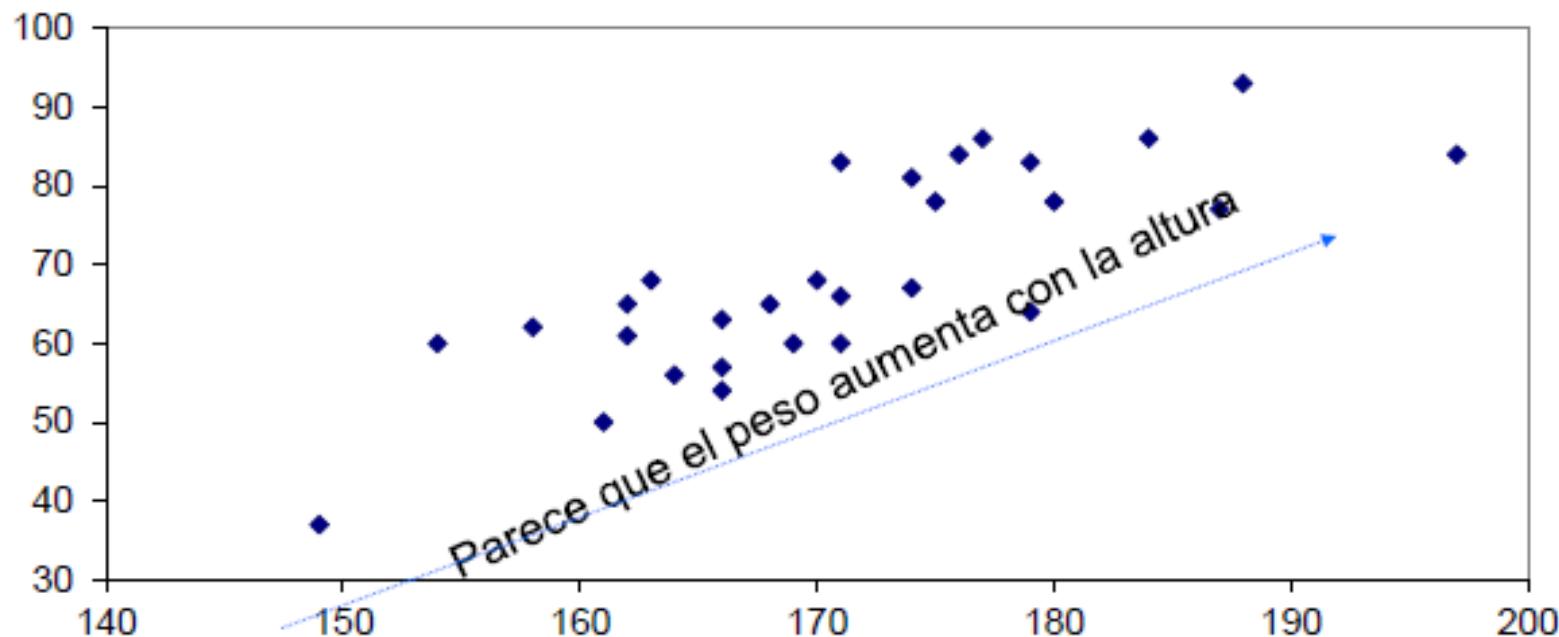
- Se representa el **diagrama de dispersión**: cada individuo es un punto cuyas coordenadas son los valores de las variables.
- El objetivo de este diagrama es intentar reconocer si hay **relación entre las variables**, de qué tipo, y si es posible predecir el valor de una de ellas en función de la otra
- Ej: Tenemos las alturas y los pesos de 30 individuos representados en un diagrama de dispersión.



T2. Descripción conjunta de varias variables

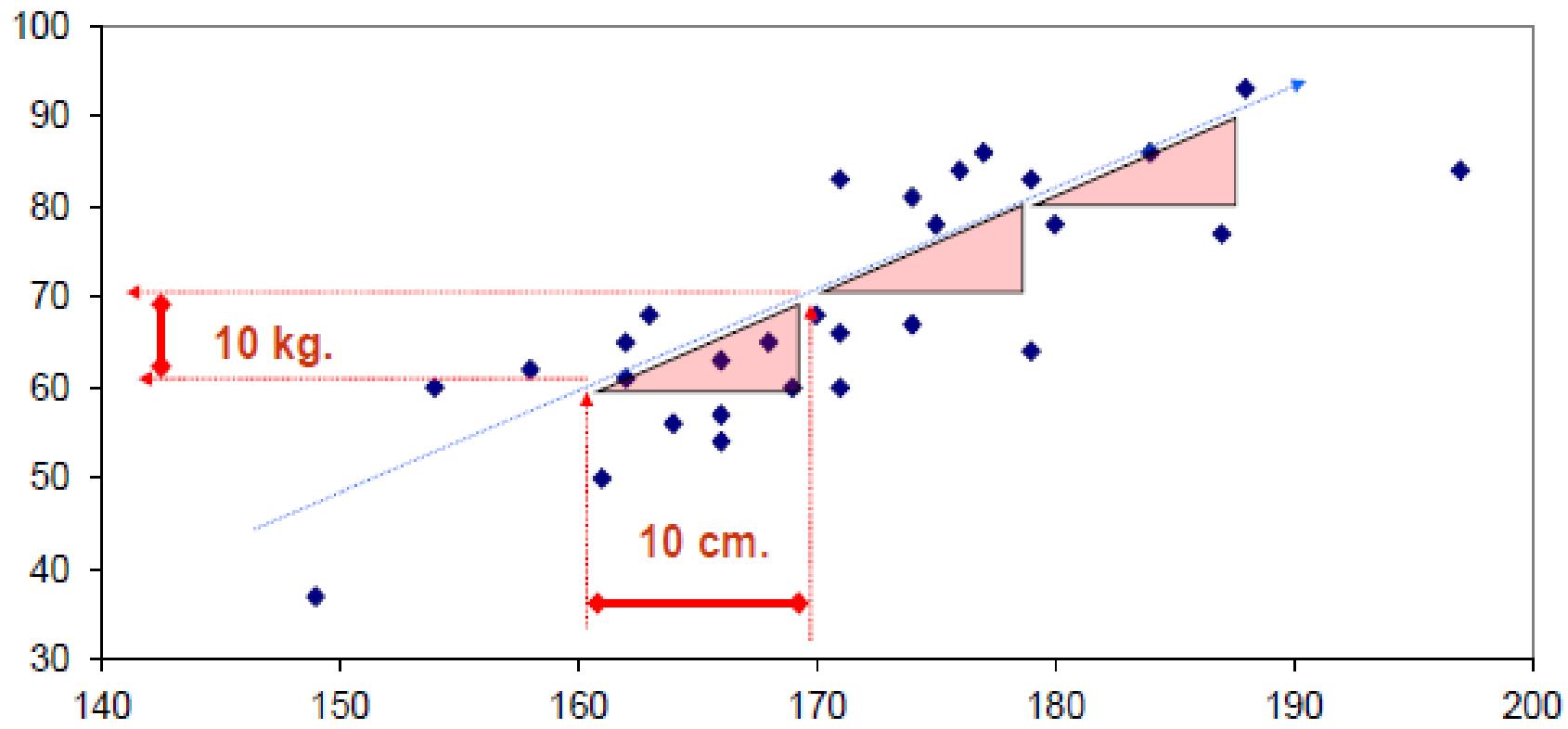
Relación entre variables

- ▶ Tenemos las alturas y los pesos de 30 individuos representados en un diagrama de dispersión.



T2. Descripción conjunta de varias variables

- ▶ **EJEMPLO:** aparentemente el peso aumenta 10 Kg por cada 10 cm de altura: el peso aumenta en una unidad por cada unidad de altura.



T2. Descripción conjunta de varias variables

Llamaremos **distribución conjunta de frecuencias** de dos variables (x, y) a una **tabla** que representa los **valores** observados de **ambas variables** y se puede expresar también mediante las **frecuencias relativas** de aparición de **cada par de valores**.

Variable Bidimensional (X,Y) Sobre una población se observan simultáneamente dos variables X e Y.

La distribución de frecuencias bidimensional de (X,Y) es el conjunto de valores

$\{(x_i, y_j); n_{ij}\} \quad i=1, \dots, p; j=1, \dots, q$ tal que

$$\sum_{i=1}^p \sum_{j=1}^q n_{ij} = N \quad \text{O equivalente:} \quad \sum_{i=1}^p \sum_{j=1}^q f_{ij} = 1$$

donde n_{ij} es la frecuencia absoluta conjunta o total de elementos en la población que presenta el valor bidimensional (x_i, y_j) .

La frecuencia relativa conjunta f_{ij} es la proporción de elementos en la población que presenta el valor (x_i, y_j) .

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

T2. Descripción conjunta de varias variables

Ejemplo distribución bidimensional (en frecuencias absolutas y en relativas):

Un grupo de 91 niños se clasifica según su edad (X) y puntuación en un test (Y)

Frecuencias absolutas

Edad	TEST		
	120	125	130
5	10	8	2
6	7	8	6
7	2	10	13
8	1	4	20

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

Frecuencias relativas

Edad	TEST		
	120	125	130
5	0,110	0,088	0,022
6	0,077	0,088	0,066
7	0,022	0,110	0,143
8	0,011	0,044	0,220

$$0,110 = \frac{10}{91}$$

$$0,220 = \frac{20}{91}$$

¿Cómo se expresa la distribución bidimensional en frecuencias relativas a partir de las de frecuencias absolutas?

¡Es muy fácil! Se divide cada casilla (frecuencia absoluta) entre N (91)

Observa que la fila y columna marginales (sombreadas) representan las frecuencias marginales (las absolutas en tabla de la derecha y las relativas en la de la izquierda).

T2. Descripción conjunta de varias variables

La distribución de frecuencias bidimensional de (X,Y) se puede expresar en una tabla bidimensional (frecuencias absolutas):

	y_1	y_2	...	y_j	...	y_q	
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1q}	n_{1*}
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2q}	n_{2*}
...
x_i	n_{i1}	n_{i2}	...	n_{ij}	n_{iq}	n_{i*}	
...
x_p	n_{p1}	n_{p2}	...	n_{pj}	...	n_{pq}	n_{p*}
	n_{*1}	n_{*2}	...	n_{*j}	...	n_{*q}	N

COLUMNA DE FRECUENCIAS MARGINALES

Total fila 1

$$n_{1*} = \sum_{j=1}^q n_{ij}$$

Total de elementos que presentan el valor x_i

$$n_{i*} = \sum_{j=1}^q n_{ij}$$

Frecuencia absoluta

fila columna

Total de elementos que presentan x_i e y_j

$$n_{ij} = \sum_{j=1}^q n_{ij}$$

Total fila p

$$n_{p*} = \sum_{j=1}^q n_{pj}$$

Total de elementos en la población

Total columna j

$$n_{*j} = \sum_{i=1}^p n_{ij}$$

Total de elementos que presentan el valor y_j

T2. Descripción conjunta de varias variables

La distribución de frecuencias bidimensional de (X,Y) se puede expresar en una tabla bidimensional (frecuencias relativas):

	y_1	y_2	...	y_j	...	y_q	
x_1	f_{11}	f_{12}	...	f_{1j}	...	f_{1q}	f_{1*}
x_2	f_{21}	f_{22}	...	f_{2j}	...	f_{2q}	f_{2*}
...
x_i	f_{i1}	f_{i2}	...	f_{ij}	...	f_{iq}	f_{i*}
...
x_p	f_{p1}	f_{p2}	...	f_{pj}	...	f_{pq}	f_{p*}
	f_{*1}	f_{*2}	...	f_{*j}	...	f_{*q}	1

Proporción de elementos que presenta el valor y_j

Total columna j
Total columna q

$$f_{*j} = \sum_{i=1}^p f_{ij}$$

Proporción de elementos que presenta el valor x_i

Total fila 1

Total fila 2

$$f_{i*} = \sum_{j=1}^q f_{ij}$$

Proporción de elementos que presenta x_i e y_j

$$1 = \sum_{j=1}^q \sum_{i=1}^p f_{ij}$$

T2. Descripción conjunta de varias variables

- Uno de los objetivos del análisis de distribuciones bidimensionales es estudiar si existe **asociación o relación** entre las variables X e Y.
- A partir de una distribución bidimensional se obtendrán distribuciones **unidimensionales** de dos tipos: **marginales** y **condicionadas**.
- Dos distribuciones marginales:
 - Marginal de X
 - Marginal de Y
- Condicionadas:
 - q distribuciones condicionadas de los valores de X a los q valores de Y
 - p distribuciones condicionadas de los valores de Y a los p valores de X

T2. Descripción conjunta de varias variables

A partir de una distribución bidimensional se pueden obtener 2 distribuciones unidimensionales MARGINALES: Marginal de X y Marginal de Y.

MARGINAL DE X

X	n _{i*}	f _{i*}
x ₁	n _{1*}	f _{1*}
x ₂	n _{2*}	f _{2*}
...
x _i	n _{i*}	f _{i*}
...
x _p	n _{p*}	f _{p*}
	N	1

$$f_{i*} = \frac{n_{i*}}{N}$$

Marginal de X: expresa cómo se distribuye X en la población total, al margen de la otra variable

Marginal de Y: expresa cómo se distribuye Y en la población total, al margen de la otra variable

MARGINAL DE Y

Y	y ₁	y ₂	...	y _j	...	y _q	
n _{*j}	n _{*1}	n _{*2}	...	n _{*j}	...	n _{*q}	N
f _{*j}	f _{*1}	f _{*2}	...	f _{*j}	...	f _{*q}	1

$$f_{*j} = \frac{n_{*j}}{N}$$

T2. Descripción conjunta de varias variables

Ejemplo (continuación)

Distribuciones marginales de la Edad y Test

Distribución marginal
de la Edad

Edad	Número alumnos	Proporción de alumnos
5	20	0,220
6	21	0,231
7	25	0,275
8	25	0,275
	91	1

Distribución marginal
Del Test

TEST	número de alumnos	proporción de alumnos
120	20	0,220
125	30	0,330
130	41	0,451
	91	1

Observa que el total de individuos observados en cada marginal es 91. Todos.

¿qué porcentaje de niños tiene edad igual 5?

¿qué proporción de alumnos obtiene en el test más de 125 puntos?

T2. Descripción conjunta de varias variables

Ejemplo distribución bidimensional (en frecuencias absolutas y en relativas):
Un grupo de 91 niños se clasifica según su edad (X) y puntuación en un test (Y)

En frecuencias absolutas

Edad	TEST			Marginal ↓
	120	125	130	
5	10	8	2	20
6	7	8	6	21
7	2	10	13	25
8	1	4	20	25

Marginal → 20 30 41 91

En frecuencias relativas

Edad	TEST			Marginal ↓
	120	125	130	
5	0,110	0,088	0,022	0,220
6	0,077	0,088	0,066	0,231
7	0,022	0,110	0,143	0,275
8	0,011	0,044	0,220	0,275

Marginal → 0,220 0,330 0,451 1,000

¿Cómo se interpretan los valores 10 y 20?

Hay 10 niños que tienen 7 años y puntuación 125 en el test. Hay 20 niños con puntuación igual a 120.

¿Cómo se interpretan los valores 0,110 y 0,220?

Hay una proporción de 0,11 niños que tiene 7 años y puntuación 125 en el test. El 22% de los niños tiene puntuación igual a 120.

T2. Descripción conjunta de varias variables

A partir de una distribución bidimensional se pueden obtener distribuciones unidimensionales CONDICIONADAS: de X y de Y.

CONDICIONAL DE X / Y=y_j

X	n _{ij}	f _{i/j}
x ₁	n _{1j}	n _{1j} /n _{*j} =f _{1/j}
x ₂	n _{2j}	n _{2j} /n _{*j} =f _{2/j}
...
x _i	n _{ij}	n _{ij} /n _{*j} =f _{i/j}
...
x _p	n _{pj}	n _{pj} /n _{*j} =f _{p/j}
	n _{*j}	1

Total de elementos en la subpoblación

Condicional de X dado Y=y_j: expresa cómo se distribuye X en la subpoblación que cumple la condición de presentar el valor Y=y_j

Condicional de Y dado X=x_i: expresa cómo se distribuye Y en la subpoblación que cumple la condición de presentar el valor X=x_i

CONDICIONAL DE Y / X=x_i

Y	y ₁	y ₂	...	y _j	...	y _q	
n _{ij}	n _{i1}	n _{i2}	...	n _{ij}	...	n _{iq}	n _{i*}
f _{j/i}	n _{i1} /n _{i*} =f _{1/i}	n _{i2} /n _{i*} =f _{2/i}	...	n _{ij} /n _{i*} =f _{j/i}	...	n _{iq} /n _{i*} =f _{q/i}	1

Total de elementos en la subpoblación

T2. Descripción conjunta de varias variables

Ejemplo (continuación)

Distribuciones condicionadas de la Edad a los valores del test

Distribución bidimensional

Edad	TEST		
	120	125	130
5	0,110	0,088	0,022
6	0,077	0,088	0,066
7	0,022	0,110	0,143
8	0,011	0,044	0,220
	0,220	0,330	0,451
			1,000

Distribuciones condicionadas de la Edad

Edad	TEST		
	120	125	130
5	0,500	0,267	0,049
6	0,350	0,267	0,146
7	0,100	0,333	0,317
8	0,050	0,133	0,488
	1,000	1,000	1,000
			1,000

¿Cómo se hace si la distribución bidimensional está en frecuencias relativas?

Igual que antes. Se divide cada casilla de la bidimensional (tabla izquierda) entre el total de columna.

Las flechas de la tabla indican la dirección en que se han de hacer los cálculos

Por ejemplo, para obtener la distribución condicionada de la Edad / test =120 se divide cada casilla de la columna encabezada por 120 por el total de columna (0,022). Observa que la población que cumple esa condición es de una proporción igual a 0,022 niños.

Observa que la última fila está formada por unos. Hay 3 distribuciones condicionadas de la Edad. Una marginal de la Edad.

T2. Descripción conjunta de varias variables

Ejemplo (continuación)

Distribuciones condicionadas del Test a los valores de la edad

Distribución bidimensional

		TEST		
		120	125	130
Edad	120	0,110	0,088	0,022
	125	0,077	0,088	0,066
5	130	0,022	0,110	0,143
		0,011	0,044	0,220
		0,220	0,330	0,451
		1,000		

Distribuciones condicionadas del Test

Edad	TEST		
	120	125	130
5	0,500	0,400	0,100
6	0,333	0,381	0,286
7	0,080	0,400	0,520
8	0,040	0,160	0,800
	0,220	0,330	0,451

¿Cómo se hace?

Las flechas de la tabla indican la dirección en que se han de hacer los cálculos

Por ejemplo, para obtener la distribución condicionada del test /Edad=6 años se divide cada casilla de la fila encabezada por 6 entre el total de fila (0,231). Observa que la población que cumple esa condición es de una proporción igual a 0,231 niños.

Observa que la última columna está formada por unos. Hay 4 distribuciones condicionadas del test. Y la marginal del test.

T2. Descripción conjunta de varias variables

- Uno de los objetivos del análisis de distribuciones bidimensionales es estudiar si son **independientes** o por el contrario, existe **asociación o relación** entre las variables X e Y.
- Las variables X e Y se dicen que son **independientes** si los valores de una de ellas no afecta a la distribución de la otra. Esto equivale a decir que **todas las distribuciones condicionadas sean iguales**.
- De modo equivalente se dice que las variables X e Y son independientes si se cumple que la frecuencia relativa conjunta es igual al producto de las frecuencias relativas marginales.
- Si las variables no son independientes se dice que están relacionadas o asociadas. Las distribuciones condicionadas NO son iguales.

T2. Descripción conjunta de varias variables

Ejemplo:

Comprueba si son o no independientes las variables X e Y de la distribución bidimensional (X, Y) siguiente:

	y1	y2	
x1	23	69	92
x2	12	36	48
x3	15	45	60
x4	7	21	28
	57	171	228

Cálculo

Basta ver que las distribuciones condicionadas son iguales. Por ejemplo, las condicionadas de X/Y

Condicionadas de X a los valores de Y: X/Y

	y1	y2	
x1	0,404	0,404	0,404
x2	0,211	0,211	0,211
x3	0,263	0,263	0,263
x4	0,123	0,123	0,123
	1	1	1

¿Cómo se hacen los cálculos?

Verticalmente: Dividiendo cada casilla (frecuencia) entre el total de columna

Observa que la variable X se distribuye igual en el conjunto de individuos que presenta la condición $Y=y_1$, que en el grupo que cumple $Y=y_2$.

La lectura de la tabla de condicionadas se hace en sentido contrario al que se hayan realizado los cálculos; es decir, en el ejemplo la lectura es horizontal: Fila 1: $0,404 = 0,404$; Fila 2: $0,211=0,211$; Fila 3: $0,263=0,263$; Fila 4: $0,123=0,123$.

Todas las condicionadas son iguales. Por tanto las variables X e Y son INDEPENDIENTES

T2. Descripción conjunta de varias variables

Ejemplo (Continuación):

Comprueba si son o no independientes las variables X e Y de la distribución bidimensional (X, Y) siguiente:

	y1	y2	
x1	23	69	92
x2	12	36	48
x3	15	45	60
x4	7	21	28
	57	171	228

Cálculo

Otro modo de ver que son independientes es comprobando que las distribuciones condicionadas de Y/X son todas iguales.

Condicionadas de Y a los valores de X: Y/X

	y1	y2	
x1	0,250	0,750	1,000
x2	0,250	0,750	1,000
x3	0,250	0,750	1,000
x4	0,250	0,750	1,000
	0,25	0,75	1

¿Cómo se hacen los cálculos?

Horizontalmente: Dividiendo cada casilla (frecuencia) entre el total de fila

Observa que la variable Y se distribuye igual en el conjunto de individuos que presenta la condición $X=x_1$, que en el grupo que cumple $X=x_2, \dots$, y que en el grupo $X=x_4$.

La lectura de la tabla de condicionadas se hace en sentido contrario al que se hayan realizado los cálculos; es decir, en el ejemplo la lectura es vertical: Columna 1: $0,250 = 0,250 = 0,250 = 0,250$; Columna 2: $0,750 = 0,750 = 0,750 = 0,750$.

Todas las condicionadas son iguales. Por tanto las variables X e Y son INDEPENDIENTES

T2. Descripción conjunta de varias variables

Ejemplo (Continuación):

Comprueba si son o no independientes las variables X e Y de la distribución bidimensional (X, Y) siguiente: (Puedes hacerlo con frecuencias absolutas o con relativas)

	y1	y2	
x1	23	69	92
x2	12	36	48
x3	15	45	60
x4	7	21	28
	57	171	228

Otro modo de ver que son independientes es comprobando que las frecuencias relativas conjuntas verifican la ecuación:

$$f_{ij} = f_{i*} \cdot f_{*j}$$

O la equivalente

$$n_{ij} = \frac{n_{i*} \cdot n_{*j}}{N}$$

¿Cómo?

Comprueba que cada frecuencia absoluta verifica la ecuación. Por ejemplo,

$$15 = \frac{60 \cdot 57}{228}$$

	y1	y2	
x1	0,101	0,303	0,404
x2	0,053	0,158	0,211
x3	0,066	0,197	0,263
x4	0,031	0,092	0,123
	0,250	0,750	1,000

¿Cómo?

si prefieres usar la primera ecuación:

Se obtiene la distribución bidimensional en frecuencias relativas. Para ello divide cada casilla correspondiente a una frecuencia absoluta entre 228

Por ejemplo, $0,101=23/228$.

Comprueba luego que se verifica $0,101=0,0404$ por $0,250$; $0,303=0,404$ por $0,750$;, $0,092=0,123$ por $0,750$.

T2. Descripción conjunta de varias variables

Medidas de dependencia lineal: Covarianza

- ▶ La **covarianza entre dos variables**, S_{xy} , nos indica si la posible relación entre dos variables es directa o inversa.
 - Directa: $S_{xy} > 0$
 - Inversa: $S_{xy} < 0$
 - Incorreladas: $S_{xy} = 0$
- ▶ El signo de la covarianza nos dice si el aspecto de la nube de puntos es creciente o no, pero no nos dice nada sobre el **grado de relación** entre las variables.

$$S_{xy} = \frac{1}{N} \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \frac{\sum_i x_i \cdot y_i}{N} - \bar{x} \cdot \bar{y}$$

- ▶ Para datos agrupados:

$$S_{xy} = \frac{1}{N} \sum_i f_i \cdot (x_i - \bar{x})(y_i - \bar{y}) = \frac{\sum_i f_i \cdot x_i \cdot y_i}{N} - \bar{x} \cdot \bar{y}$$

T2. Descripción conjunta de varias variables

Ejemplo

Teniendo en cuenta que los datos de radio... y densidad son los siguientes de los planetas son:

(radio ecuatorial, densidad): Mercurio, 2.4, 5.4; Venus, 6.1, 5.2; Tierra, 6.4, 5.5; Marte, 3.4, 3.9. Calcular su covarianza.

Solución:

$$\begin{aligned}s_{xy} &= \frac{\sum_{i=1}^n \sum_{j=1}^m f_{ij} x_i y_j}{\sum_{i=1}^n f_i} - \left(\frac{\sum_{i=1}^n f_i x_i}{\sum_{i=1}^n f_i} \right) \cdot \left(\frac{\sum_{j=1}^m f_j y_j}{\sum_{j=1}^m f_j} \right) \\&= \frac{2,4 * 5,4 + 6,1 * 5,2 + 6,4 * 5,5 + 3,4 * 3,9}{4} \\&\quad - \left(\frac{2,4 + 6,1 + 6,4 + 3,4}{4} \right) \left(\frac{5,4 + 5,2 + 5,5 + 3,9}{4} \right) \\&= 0,41\end{aligned}$$

T2. Descripción conjunta de varias variables

Medidas de dependencia lineal: Correlación

Correlación: Medida de la dependencia existente entre variantes aleatorias.

La correlación lineal se calcula a través de la ecuación:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

r toma valores entre -1 y 1.

- Si la correlación lineal es perfecta, es decir si los valores de *x* e *y* están sobre una recta el valor de *r* será -1 si tiene pendiente negativa y 1 si la tiene positiva.
- Si *r* es igual a 0 no hay dependencia lineal entre las variables, lo cual implica o bien que las variables son independientes, o bien que hay una dependencia no lineal entre las mismas.

T2. Descripción conjunta de varias variables

Teniendo en cuenta que los datos de radio... y densidad son los siguientes:
(radio ecuatorial, densidad): Mercurio, 2.4, 5.4; Venus, 6.1, 5.2; Tierra, 6.4, 5.5; Marte, 3.4, 3.9.) Calcular la correlación existente entre el radio ecuatorial y la densidad para dichos planetas.

$$s_x = \sqrt{\frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{\sum_{i=1}^n f_i}} = \sqrt{\frac{(2,4 - 4,575)^2 + \dots + (3,4 - 4,575)^2}{4}} = 1,715$$

$$s_y = \sqrt{\frac{\sum_{j=1}^m f_j (y_j - \bar{y})^2}{\sum_{j=1}^m f_j}} = \sqrt{\frac{(5,4 - 5)^2 + \dots + (3,9 - 5)^2}{4}} = 0,6442$$

$$s_{xy} = \frac{\sum_{i=1}^n \sum_{j=1}^m f_{ij} x_i y_j}{\sum_{i=1}^n f_i} - \left(\frac{\sum_{i=1}^n f_i x_i}{\sum_{i=1}^n f_i} \right) \cdot \left(\frac{\sum_{j=1}^m f_j y_j}{\sum_{j=1}^m f_j} \right)$$

$$= \frac{2,4 \cdot 5,4 + 6,1 \cdot 5,2 + 6,4 \cdot 5,5 + 3,4 \cdot 3,9}{4} - \left(\frac{2,4 + 6,1 + 6,4 + 3,4}{4} \right) \left(\frac{5,4 + 5,2 + 5,5 + 3,9}{4} \right) = 0,41$$

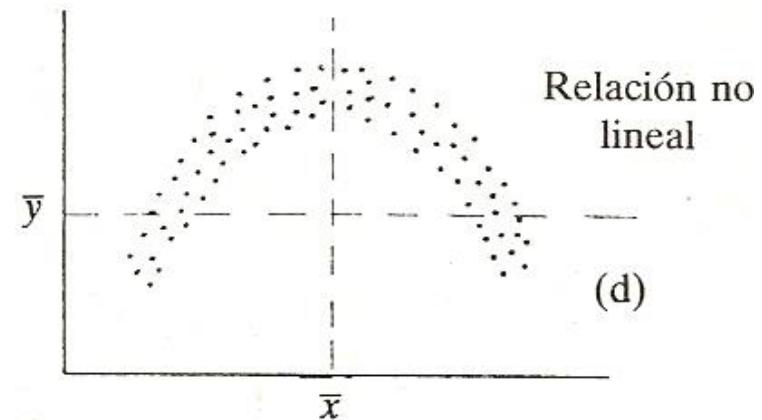
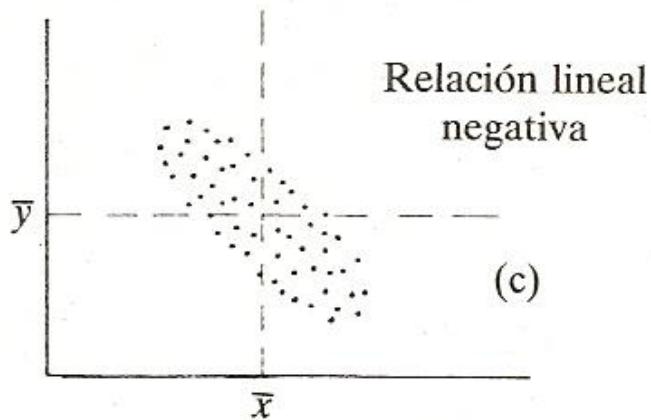
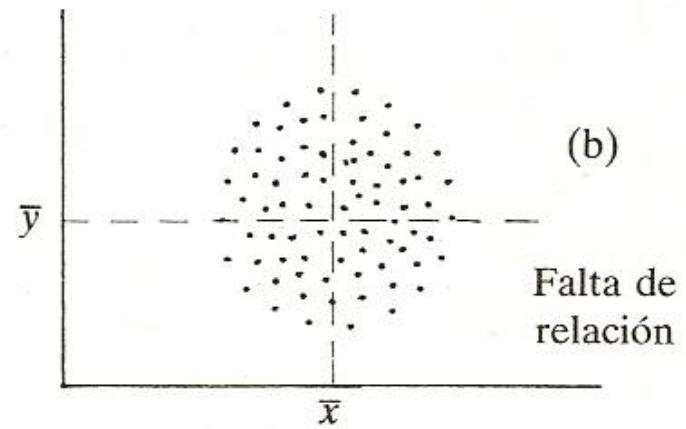
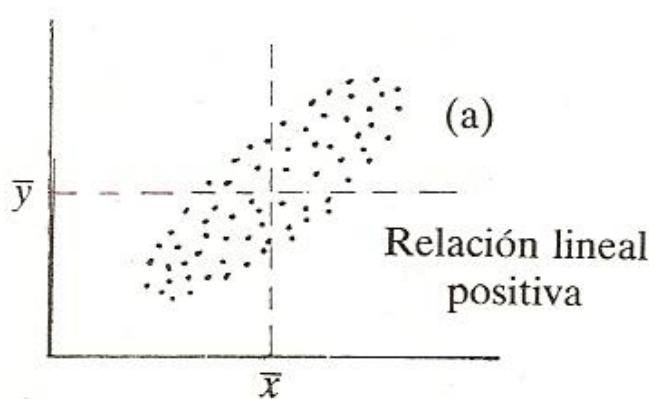
$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{0,41}{1,715 \cdot 0,6442} = 0,371$$

T2. Descripción conjunta de varias variables

- ▶ Propiedades de la correlación:

- Tiene el mismo signo de la covarianza e indica si la relación entre las variables es creciente o decreciente
- El coeficiente de correlación es adimensional
- Sólo toma valores en $[-1, 1]$
- Las variables son incorreladas $\Leftrightarrow r=0$
- Relación lineal perfecta entre dos variables $\Leftrightarrow r=+1$ ó $r=-1$
- Cuanto más cerca esté r de $+1$ ó -1 mejor será el grado de relación lineal.

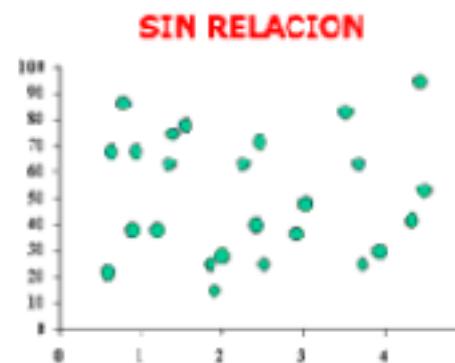
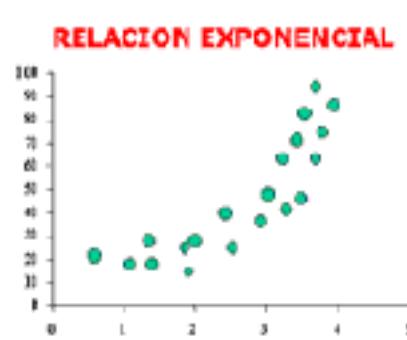
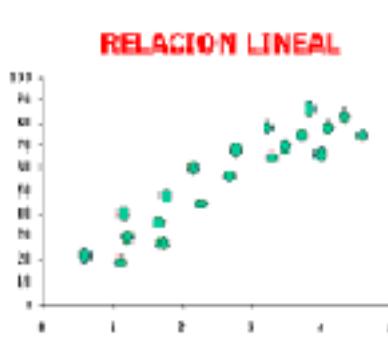
T2. Descripción conjunta de varias variables



T2. Descripción conjunta de varias variables

Curva de Ajuste

- ▶ Del diagrama de dispersión se puede visualizar frecuentemente una curva de ajuste.



- ▶ El problema de hallar ecuaciones de curvas que se ajusten a conjuntos de datos se llama **curva de ajuste**
- ▶ Para los ejemplo anteriores esas curvas podrían ser de la forma

$$y = a + bx$$

$$y = a + bx + cx^2$$

T2. Descripción conjunta de varias variables

Una vez establecida la **función de regresión** podremos obtener el valor de una de las variables si conocemos el valor de la otra.

La relación entre dos variables más básica es una **relación lineal**.

La función de relación lineal entre dos variables es una recta.

$$y = a + bx$$

Se deben determinar **a** y **b**

El método de ajuste más utilizado es el de *mínimos cuadrados*, que está basado en encontrar una función para calcular la y a partir de x , tal que minimice el valor de la suma del cuadrado de las restas de todos los valores de y observados menos los calculados.

$$\sum_{i=1}^n (y_i - y_{ci})^2 \text{ Mínimo}$$

Los valores de **a** y **b** a partir del método de mínimos cuadrados son:

$$b = \frac{s_{xy}}{s_x^2} = r_{xy} \frac{s_y}{s_x} \quad a = \bar{y} - b\bar{x}$$

T2. Descripción conjunta de varias variables

Teniendo en cuenta que los datos de radio... y densidad son los siguientes:

(radio ecuatorial, densidad): Mercurio, 2.4, 5.4; Venus, 6.1, 5.2; Tierra, 6.4, 5.5; Marte, 3.4, 3.9.) Hallar la recta de regresión lineal entre ambas magnitudes.

Solución:

$$s_{xy} = 0,41$$

$$s_x = 1,715 \quad s_y = 0,6442$$

$$r_{xy} = 0,371$$

$$\bar{x} = \left(\frac{2,4+6,1+6,4+3,4}{4} \right) = 4,575 \quad \bar{y} = \left(\frac{5,4+5,2+5,5+3,9}{4} \right) = 5$$

$$b = \frac{s_{xy}}{s_x^2} = \frac{0,41}{1,715^2} = 0,14 \quad \text{o bien} \quad b = r_{xy} \frac{s_y}{s_x} = 0,371 \frac{1,715}{0,6442} = 0,14$$

$$a = \bar{y} - b\bar{x} = 5 - 0,14 * 4,575 = 4,36$$

$$y = 4,36 + 0,14x$$

T2. Descripción conjunta de varias variables

Recta de Regresión

En resumen los parámetros de la recta b y a se obtienen mediante las formulas siguientes .

$$b = \frac{S_{xy}}{S_x^2}$$

$$a = \bar{y} - \frac{S_{xy}}{S_x^2}(\bar{x})$$

- Sustituyendo “a” y “b” en la ecuación de la recta de regresión se obtiene que la ecuación de la recta que más se aproxima a la nube de puntos es:

$$y = \bar{y} + \frac{S_{xy}}{S_x^2}(x - \bar{x})$$

- De forma similar se obtiene la ecuación de la recta de regresión de “x” sobre “y”

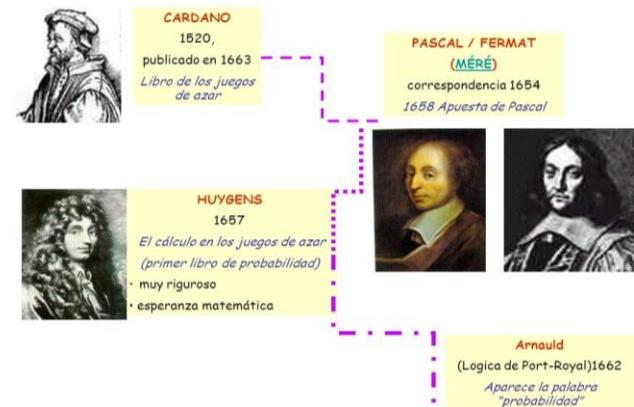
$$x = \bar{x} + \frac{S_{xy}}{S_y^2}(y - \bar{y})$$

Estadística

Tema 3: Probabilidad

Introducción

- ▶ El concepto de probabilidad nace con el deseo del hombre de conocer con certeza los eventos futuros. Es por ello que el estudio de probabilidades surge como una herramienta utilizada por los nobles para ganar en los juegos y pasatiempos de la época.
- ▶ El desarrollo de estas herramientas fue asignado a los matemáticos de la corte.



- ▶ Con el tiempo estas técnicas matemáticas se perfeccionaron y encontraron otros usos muy diferentes para la que fueron creadas.
- ▶ Actualmente se ha continuado con el estudio de nuevas metodologías que permitan maximizar el uso de la computación en el estudio de las probabilidades disminuyendo, de este modo, los márgenes de error en los cálculos.

Introducción

- ▶ En la vida cotidiana aparecen muchas situaciones en las que los resultados observados son diferentes aunque las condiciones iniciales en las que se produce la experiencia sean las mismas.

Por ejemplo, al lanzar una moneda unas veces resultará cara y otras cruz, o lanzar dados resultando distintos valores...

Estos fenómenos, **denominados aleatorios**, se ven afectados por la **incertidumbre**.

En el lenguaje habitual, frases como

"probablemente...", "es poco probable que...", "hay muchas posibilidades de que..." hacen referencia a esta incertidumbre.

- ▶ **La teoría de la probabilidad** pretende ser una herramienta para modelizar y tratar con situaciones de este tipo.
- ▶ Por otra parte, cuando aplicamos las técnicas estadísticas a la recogida, análisis e interpretación de los datos, la teoría de la probabilidad proporciona una base para evaluar la fiabilidad de las conclusiones alcanzadas y las inferencias realizadas.
- ▶ Debido al importante papel desempeñado por la probabilidad dentro de la estadística, es necesario familiarizarse con sus elementos básicos



Introducción

¿Qué es la probabilidad?

Es poco probable que mi amigo Pepe Pérez me haga trampas jugando al "tute"

Hay poca probabilidad de obtener un producto defectuoso en este proceso de fabricación

Es baja la probabilidad de obtener 5, 4, 1 al lanzar 3 dados

Es poco probable que en la centralita de mi empresa se reciban más de 100 llamadas entre las 17:00 y las 17:05 horas

Hay poca probabilidad de que el acusado sea culpable

Es poco probable que yo llegue a ser Premio Nobel de Economía

Es baja la probabilidad de acertar la lotería primitiva

Existe poca probabilidad de que el paciente sufra el síndrome de Algeman

Es baja la probabilidad de que una botella de leche fresca dure más de cuatro meses en buenas condiciones



Introducción

La probabilidad es una medida de la incertidumbre

En cada una de las afirmaciones anteriores, no se está seguro de cada suceso

Pero se conoce algo, en diferentes grados.

En algunos casos, las propiedades físicas permiten obtener una estimación bastante exacta de la probabilidad.

En otros casos, incluso un intervalo amplio no sería suficiente para estimarla.

La diferencia está en el grado de precisión

El verdadero problema está en si podemos utilizar la probabilidad para medir cualquier tipo de incertidumbre.

Los tres colores de las afirmaciones anteriores nos las clasifican en tres categorías de precisión.

Piensa sobre las características comunes en cada uno de los colores

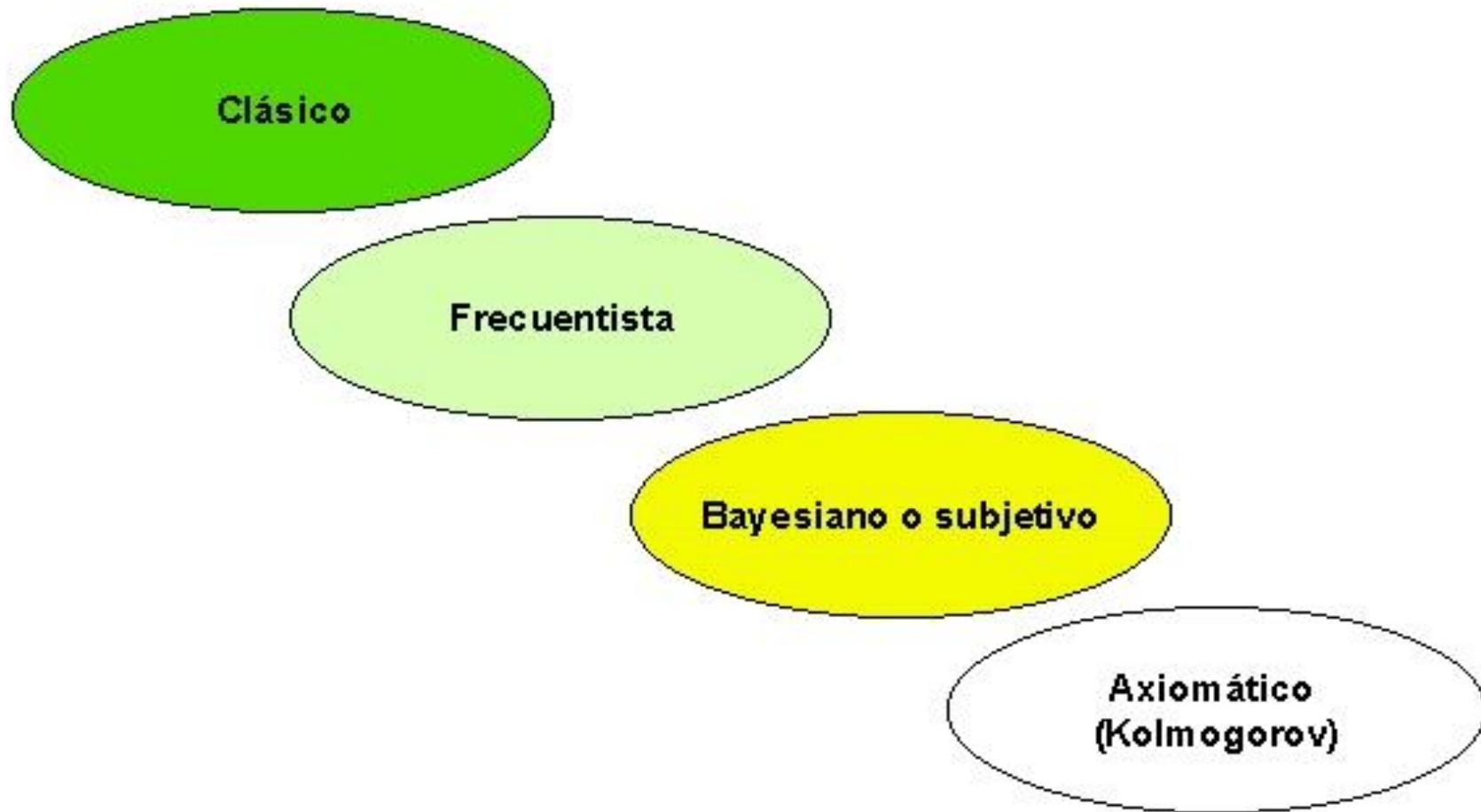
Tratamos con individuos, objetos, conjuntos, propiedades...

Introducción

- ▶ En la actualidad no existe una definición universal del **concepto probabilidad**.
- ▶ De hecho, a lo largo de la historia se han dado diferentes interpretaciones y definiciones de este concepto y aun hoy en día existe una gran controversia entre los probabilistas sobre como debe interpretarse la probabilidad y dar una definición formal de acuerdo a la interpretación, así como el tipo de situaciones a las que debe aplicarse.
- ▶ Antes de establecer la **definición axiomática de probabilidad**, que nos proporcionará las bases para el desarrollo matemático formal de la Teoría de la Probabilidad vamos a exponer las interpretaciones mas significativas de la probabilidad.
- ▶ Las dos primeras son las apropiadas para aplicar la Teoría de la Probabilidad a distintas situaciones.

Introducción

Enfoques de la Probabilidad



Introducción

Definición clásica

Es baja la probabilidad de obtener 5, 4, 1 al lanzar 3 dados

Probabilidad: es el cociente entre el número de resultados favorables y los posibles si todos son igualmente verosímiles

Ej: Probabilidad de obtener 4 al lanzar un dado= $1/6$

Probabilidad de obtener 5,4,1 al lanzar tres dados= $1/216$

¡ La probabilidad siempre está comprendida entre 0 y 1 !

Igualmente verosímil = Igualmente probable

Este enfoque se invalida ante resultados no igualmente verosímiles

Es baja la probabilidad de acertar la lotería primitiva

Ej: el éxito o el fracaso de un proyecto

Introducción

Definición frecuentista

Hay poca probabilidad de obtener un producto defectuoso en este proceso de fabricación

Estos procesos se pueden repetir un número grande de veces en condiciones similares

Probabilidad: es el cociente entre la frecuencia observada del suceso y el total de observaciones cuando el experimento se realiza un número grande de veces.

Es poco probable que en la centralita de mi empresa se reciban más de 100 llamadas entre las 17:00 y las 17:05 horas

Es baja la probabilidad de que una botella de leche fresca dure más de cuatro meses en buenas condiciones

¡Este enfoque excluye sucesos que no se pueden repetir!

Este enfoque es muy utilizado

Introducción

Definición bayesiana o subjetiva

Probabilidad: es el grado de creencia o juicio personal

Es poco probable que mi amigo Pepe Pérez me haga trampas jugando al “tute”

Se necesita coherencia ante la incertidumbre.

No cualquier grado de creencia puede ser considerado como probabilidad.

Deben incorporarse informaciones y opiniones sobre la verosimilitud del resultado

Hay poca probabilidad de que el acusado sea culpable

Existe poca probabilidad de que el paciente sufra el síndrome de Algeman

Es poco probable que yo llegue a ser Premio Nobel de Economía

¡Pero la coherencia no es fácil de definir!



Espacio Muestral y Sucesos

Un experimento aleatorio es aquel que antes de realizarlo no se puede predecir el resultado que se va a obtener. En caso contrario se dice determinista.

Aunque en un experimento aleatorio no sepamos lo que ocurrirá al realizar una "prueba" si que conocemos de antemano todos sus posibles resultados.

El experimento puede repetirse tantas veces como sea necesario en idénticas condiciones

- ▶ El **espacio muestral** es el conjunto de todos los resultados posibles de un experimento aleatorio. Se suele designar con la letra E.
- ▶ Cada uno de estos posibles resultados se llama **suceso elemental**.
- ▶ En general llamaremos **suceso o evento**, elemental o compuesto, a cualquier subconjunto del espacio muestral. Por lo general, se denotan con mayúsculas.
- ▶ El espacio muestral E es un suceso, denominado **suceso seguro** y el conjunto vacío, \emptyset , se denomina **suceso imposible** ya que no se verifica nunca.

Espacio Muestral y Sucesos

► Ejemplos de Espacios Muestrales (EM)

- 1) **EM₁**: Si $x = \%$ de gases contaminantes, los posibles resultados en E₁ son

$$\{x \mid 0 \leq x \leq 100\}$$

- 2) **EM₂**: Si $x=Nº$ en la cara superior de un dado, los posibles resultados en E₂ son:

$$\{x \mid x = 1, 2, 3, 4, 5, 6\}$$

- 3) **EM₃**: Si $x=\text{peso ganado}$, $y=\text{estatura ganada}$; los posibles resultados en E₃ son

$$\{(x; y) \mid -\infty \leq x \leq \infty; y \geq 0\}$$

Espacio Muestral y Sucesos

► Del ejemplo anterior

- 1) En **EM₁**, $\{x \mid 0 \leq x \leq 100\}$ algunos eventos son:

$$A = \{x \mid 0 \leq x \leq 5\} \quad B = \{x \mid 10 \leq x \leq 20\}$$

- 2) **EM₂** Lanzar un dado.

Espacio muestral : $E = \{1, 2, 3, 4, 5, 6\}$.

Suceso A: obtener un 5 $\rightarrow A = \{5\}$. A es un suceso elemental.

Suceso B: obtener un número par $\rightarrow B = \{2, 4, 6\}$. B es un suceso compuesto.

Suceso C: obtener un número mayor que 6 $\rightarrow C = \emptyset$. C es un suceso imposible.

Suceso D: obtener par o impar $\rightarrow D = E$. D es un suceso seguro.

- 3) En **EM₃**, $\{(x; y) \mid -\infty \leq x \leq \infty; y \geq 0\}$ algunos eventos son

$$A = \{(x; y) \mid x \geq 30; y = 0\}$$

Espacio Muestral y Sucesos

Se llama **espacio de sucesos** al **conjunto S** formado por todos los sucesos (elementales y compuestos), incluidos el suceso imposible y el suceso seguro.

Este conjunto puede ser finito, infinito numerable o infinito no numerable.

Ejemplo : Experimento aleatorio: lanzar un dado.

Espacio muestral: $E = \{1, 2, 3, 4, 5, 6\}$.

Espacio de sucesos:

$$S = \{\emptyset, E, \{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{1,2\}, \{1,3\}, \dots, \{5,6\}, \{1,2,3\}, \{1,2,4\}, \dots, \{4,5,6\}, \dots, \{1,2,3,4,5\}, \dots, \{2,3,4,5,6\}\}$$

En este caso, el número de elementos de S es 2^6 .

Teoría de conjuntos

Sean **A y B dos sucesos** cualesquiera de E asociados a un experimento aleatorio, entonces:

- Llamamos **suceso unión** de A y B y lo designamos por $A \cup B$, al suceso que resulta cuando ocurre A o B o ambos a la vez.
- Llamamos **suceso intersección** de A y B y lo designamos por $A \cap B$, al suceso que resulta cuando ocurren a la vez A y B .

Decimos que A y B son **mutuamente excluyentes, disjuntos o incompatibles** si

$$A \cap B = \emptyset.$$

- Llamamos **suceso contrario o complementario** de A y lo designamos por \bar{A} , al que se verifica cuando no lo hace A .
- Llamamos **suceso diferencia** de A y B y lo designamos por $A - B$ al que resulta cuando ocurre A y no ocurre B . Observemos que $A - B = A \cap \bar{B}$.
- Decimos que A **está contenido en** B (A implica B) y lo designamos por $A \subset B$, si siempre que ocurre A ocurre B .

Teoría de conjuntos

► Ejemplos:

1) E: Lanzar un dado y observar el número en la cara superior.

$$\mathbf{EM} = \{1, 2, 3, 4, 5, 6\}$$

$$\mathbf{A} = \{4\}, \quad \mathbf{B} = \text{Nº impar} = \{1, 3, 5\}$$

$$(\mathbf{A} \cap \mathbf{B}) = \emptyset \quad (\mathbf{A} \text{ y } \mathbf{B} \text{ son excluyentes})$$

2) E: Determinar si una persona porta o no un arma blanca.

$$\mathbf{EM} = \{\text{si}, \text{no}\}$$

$$\mathbf{A} = \{\text{si}\}, \quad \mathbf{B} = \{\text{no}\}$$

$$(\mathbf{A} \cap \mathbf{B}) = \emptyset \quad (\mathbf{A} \text{ y } \mathbf{B} \text{ son excluyentes})$$

Teoría de conjuntos

PROPIEDADES DE LAS OPERACIONES CON SUCESOS

<ul style="list-style-type: none">• Commutativa : $A \cup B = B \cup A$ $A \cap B = B \cap A$	<ul style="list-style-type: none">• $A \cup \emptyset = A; A \cap \emptyset = \emptyset$• $A \cup E = E; A \cap E = A$
<ul style="list-style-type: none">• Asociativa: $(A \cup B) \cup C = A \cup (B \cup C)$ $(A \cap B) \cap C = A \cap (B \cap C)$	<ul style="list-style-type: none">• $A \cup A = A; A \cap A = A$
<ul style="list-style-type: none">• Distributiva: $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$	$A = (A \cap B) \cup (A \cap \bar{B})$ <ul style="list-style-type: none">• para cualquier B
<ul style="list-style-type: none">• Leyes de De Morgan: $\overline{A \cup B} = \bar{A} \cap \bar{B}$ $\overline{A \cap B} = \bar{A} \cup \bar{B}$	<ul style="list-style-type: none">• $A \cup \bar{A} = E; A \cap \bar{A} = \emptyset$



Probabilidad- Definición Axiomática

Sea Ω : espacio muestral, $P(\Omega)$ conjunto de las partes de Ω , o conjunto de sucesos, o álgebra de sucesos. Se define **probabilidad**, o **función de probabilidad**, a cualquier función $p: P(\Omega) \rightarrow \mathbb{R}$ (es decir, una regla bien definida por la que se asigna a cada suceso un, y un solo un, número real) que cumpla los axiomas siguientes:

- i. $p(A) \geq 0 \quad \forall A \in P(\Omega)$
- ii. $p(A_1 \cup A_2 \cup A_3 \cup \dots) = p(A_1) + p(A_2) + p(A_3) + \dots$
si $A_i \cap A_j = \emptyset \quad \forall i \neq j$ (sucesos mutuamente excluyentes)
- iii. $p(\Omega) = 1$

A la estructura $(\Omega, P(\Omega), p)$ se le denomina **espacio de probabilidad**. Establecer claramente el espacio de probabilidad será el primer paso imprescindible para estudiar una experiencia aleatoria.

Probabilidad- Propiedades

Las **propiedades de una Probabilidad** son análogas a las de la frecuencia relativa y son:

1. $0 \leq P(A) \leq 1$
2. $P(\bar{A}) = 1 - P(A)$
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
4. Si $A \subset B = P(A) \leq P(B)$
5. $P(E) = 1$
6. $P(\emptyset) = 0$

Recordamos además, la definición clásica de **Probabilidad o Regla de Laplace**: si todos los resultados del experimento son igualmente probables, la probabilidad de que aparezca un determinado suceso A, es el cociente entre el número de casos favorables a ese suceso y el número total de casos

$$P(A) = \frac{n_A}{N}$$

Para aplicar la regla de Laplace hay que contar el número de elementos de un conjunto. Necesitaremos utilizar el Análisis Combinatorio

Probabilidad

Se ha formado un equipo de investigación y análisis estadístico formado por 10 estudiantes procedentes de las titulaciones de los grados en Ingeniería Informática y en Ingeniería de Computadores. De Informática se han elegido 4 chicas y un chico; y de Ingeniería de Computadores 2 chicas y tres chicos.

- Determinar el espacio muestral del experimento escoger al azar de entre los miembros del equipo de investigación (a) un chico o una chica, (b) escoger un estudiante de una titulación y (c) escoger un chico o una chica teniendo en cuenta la titulación.

(a) {chico, chica}, (b) {II, IC}, (c) {(chico, II), (chico, IC), (chica, II), (chica, IC)}

Calcular la probabilidad de escoger al azar de entre los miembros del equipo de investigación (a) un estudiante que sea de informática, (b) un estudiante que sea de Ingeniería de Computadores, (c) un chico, (d) una chica

(a) $P(\text{II})=n_{\text{II}}/N = 5/10=0.5$

(b) $P(\text{IC})= n_{\text{IC}}/N =5/10 =0.5$

(c) $P(\text{chica})=n_{\text{chica}}/N= 6/10=0.6$ (d) $P(\text{chico})=n_{\text{chico}}/N=4/10=0.4$

Probabilidad

Sabiendo como es la constitución del equipo de investigación en estadística formado por estudiantes:

- a) Verificar que la probabilidad de elegir al azar de entre los miembros del equipo un estudiante que sea de II, de IC, chico, chica, son menores que 1.

Según la propiedad 1: $0 \leq P(A) \leq 1 \rightarrow P(IC) = 0.5 \quad P(II) = 0.5$

$$P(chica) = 0.6 \quad P(chico) = 0.4$$

- b) Verificar que las probabilidades de elegir al azar de entre los miembros del equipo un estudiante que sea de II y IC; chico o chica, son complementarias

¿Lo son que sea de la IC y chico?

Según la propiedad 2: $P(\bar{A}) = 1 - P(A)$
 $P(IC) = 1 - P(II) = 0.5;$
 $P(chica) = 1 - P(chico) = 0.6;$
No

Probabilidad

c) Calcular la probabilidad de elegir al azar de entre los miembros del equipo (a) un estudiante que sea chica de II, y (b) chica o de II.

(a) $P(\text{Chica} \cap \text{II}) = \frac{n_{\text{chica II}}}{N} = \frac{4}{10} = 0.4$

(b) según la propiedad 3 (b): $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$$P(\text{chica} \cup \text{II}) = P(\text{chica}) + P(\text{II}) - P(\text{chica} \cap \text{II}) = 0.6 + 0.5 - 0.4 = 0.7$$

d) Verificar que la probabilidad de que un estudiante sea chica siendo de la II es menor que sea de II

Según la propiedad 4 Si $A \subset B = P(A) \leq P(B) \rightarrow P(\text{Chica} \cap \text{II}) = 0.4 \leq P(\text{II}) = 0.5$

e) Calcular (a) la probabilidad de que uno de los 10 estudiantes sea chico o chica y (b) que no sea de ninguna titulación

Según la propiedad 5 $P(E) = 1$ y según 6 $P(\emptyset) = 0$

Probabilidad Condicionada

Como la probabilidad está ligada a nuestra ignorancia sobre los resultados de la experiencia, el hecho de que ocurra un suceso, puede cambiar la probabilidad de los demás.

La probabilidad de que ocurra el suceso A si ha ocurrido el suceso B se denomina **probabilidad condicionada** y se define

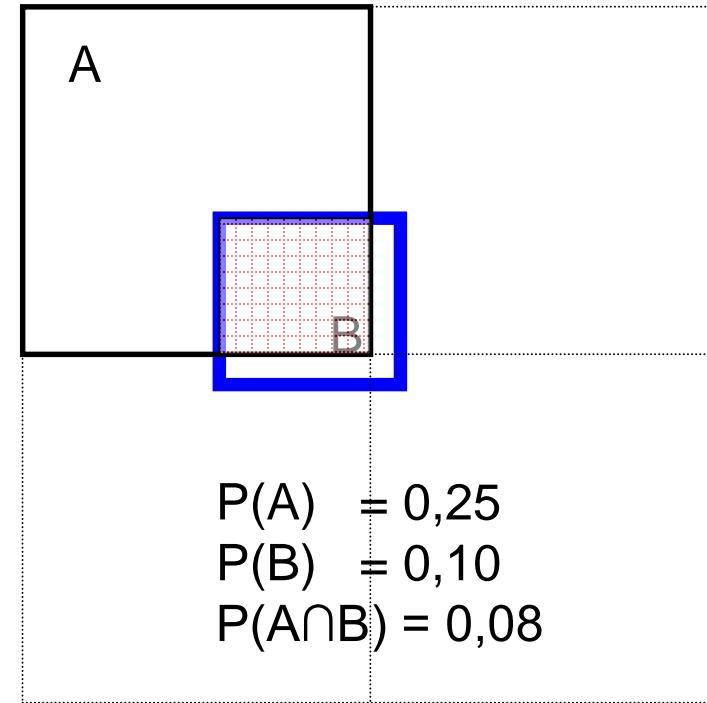
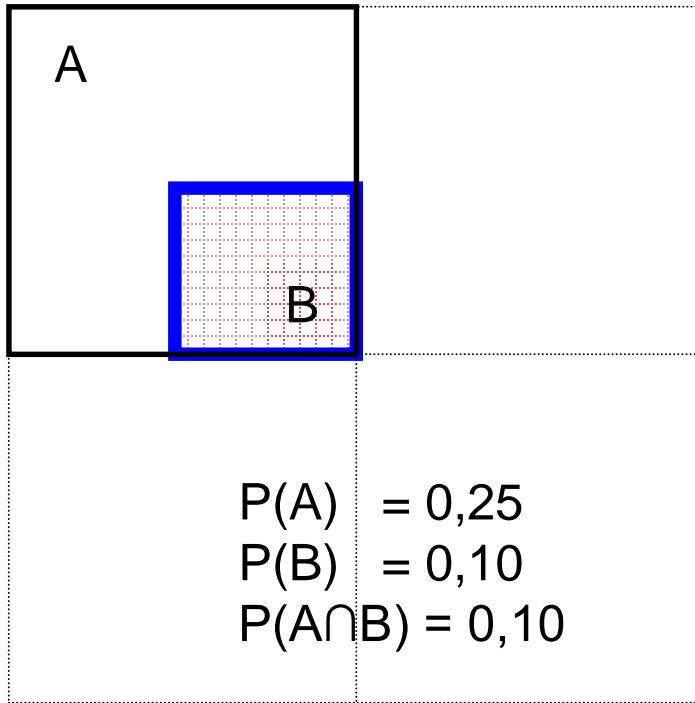
$$P(A/B) = \frac{P(A \cap B)}{P(B)} \quad \text{si } P(B) \neq 0$$

Equivalentemente:

$$P(B/A) = \frac{P(A \cap B)}{P(A)} \quad \text{si } P(A) \neq 0$$

Esta definición es consistente, es decir cumple los axiomas de probabilidad.

Probabilidad condicional: Ejemplo

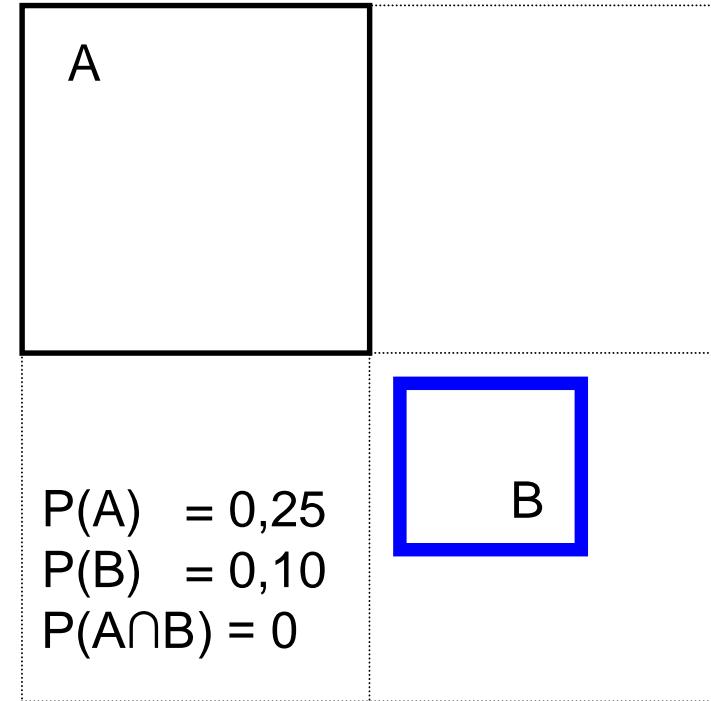
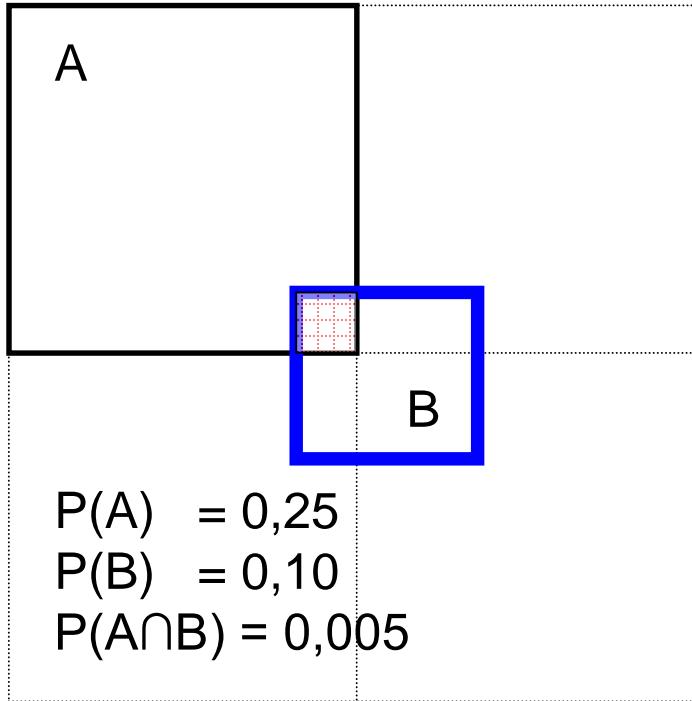


¿Probabilidad de A sabiendo que ha pasado B?

$$P(A|B)=1$$

$$P(A|B)=0,8$$

Probabilidad condicional: Ejemplo



¿Probabilidad de A sabiendo que ha pasado B?

$$\mathbf{P(A|B)=0,05}$$

$$\mathbf{P(A|B)=0}$$

Probabilidad Condicional

- ▶ Dados dos eventos A y B

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- ▶ También se puede expresar como

$$P(A \cap B) = P(B) \cdot P(A|B) = P(A) \cdot P(B|A)$$

- ▶ Esta definición puede extenderse a cualquier número de eventos. Por ejemplo, se puede demostrar que para tres eventos A, B y C

$$P(A \cap B \cap C) = P((A \cap B) \cap C) = P(A \cap B)$$

$$P(C|A \cap B) = P(A) P(B|A) P(C|A \cap B)$$

Probabilidad condicional

- ▶ Se sabe que el 50% de la población fuma y que el 10% fuma y es hipertensa. ¿Cuál es la probabilidad de que un fumador sea hipertenso?

$$A = \{\text{ser hipertenso}\} \quad B = \{\text{ser fumador}\} \quad A \cap B = \{\text{ser hipertenso y fumador}\}$$

$$P(A|B) = 0,10 / 0,50 = 0,20$$

- ▶ Una urna contiene 10 bolas, de las cuales 3 son rojas, 5 verdes y 2 azules. Se extraen al azar 3 bolas. Calcular la probabilidad de que la primera sea azul, y las otras dos verdes.

Definimos $A_1 = \{\text{la 1ª bola es azul}\}; A_2 = \{\text{la 2ª bola es verde}\}; A_3 = \{\text{la 3ª bola es verde}\}$

$$P(A_1 \cap A_2 \cap A_3)$$

$P(A_1) = 2/10$ aplicando la definición clásica de probabilidad, puesto que hay 10 bolas y 2 son verdes.

$P(A_2|A_1) = 5/9$; si la primera bola extraída es azul, en la urna quedan 9 bolas, 5 de ellas verdes.

$P(A_3|A_1 \cap A_2) = 4/8$; si la primera bola extraída es azul y la segunda verde en la urna quedan 8 bolas, 4 de ellas verdes.

$$P(A_1 \cap A_2 \cap A_3) = 2/10 \times 5/9 \times 4/8 = 1/18$$

Probabilidad - independencia

- ▶ Dos sucesos A y $B \in \Omega$ son independientes si y sólo si

$$P(A|B) = P(A) \Leftrightarrow P(B|A) = P(B)$$

En caso contrario, se dice que los sucesos son **dependientes**

- ▶ Si dos sucesos son independientes se verifica

$$A \text{ y } B \text{ son independientes} \Leftrightarrow P(A \cap B) = P(A) P(B)$$

- ▶ Propiedad:

$$A \text{ y } B \text{ son independientes} \Leftrightarrow \bar{A} \text{ y } \bar{B} \text{ son independientes}$$

$$\Leftrightarrow \bar{A} \text{ y } B \text{ son independientes} \Leftrightarrow A \text{ y } \bar{B} \text{ son independientes.}$$

Eventos estadísticamente independientes

Definición: Sean A_1, \dots, A_n sucesos cualesquiera, se dice que son **sucesos independientes** si, para todo subconjunto $\{A_{i_1}, \dots, A_{i_k}\}$ de $\{A_1, \dots, A_n\}$ se verifica que

$$P(A_{i_1} \cap \dots \cap A_{i_k}) = P(A_{i_1}) \cdot \dots \cdot P(A_{i_k}).$$

- Por ejemplo para tres sucesos, **los eventos A,B y C son estadísticamente independientes sí y sólo si**

$$P(A \cap B) = P(A) \cdot P(B)$$

$$P(A \cap C) = P(A) \cdot P(C)$$

$$P(B \cap C) = P(B) \cdot P(C)$$

$$P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$$

Eventos estadísticamente independientes

Se lanzan tres dados.

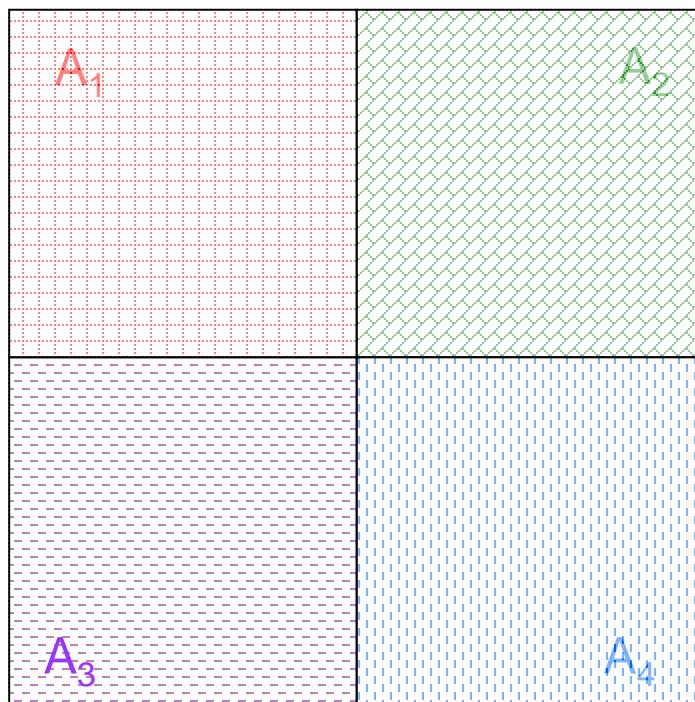
Encontrar la probabilidad de que: Salga 6 en todos.

Son sucesos independientes, el resultado de un dado no afecta a los otros

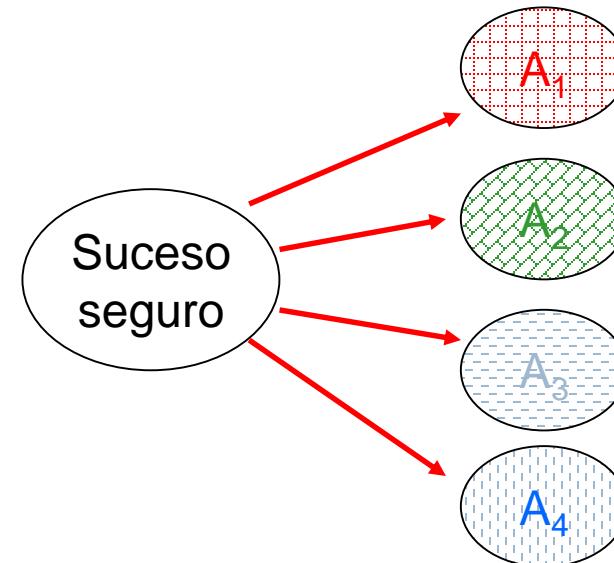
$$P(6_1 \cap 6_2 \cap 6_3) = \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{216}$$

Sistema exhaustivo y excluyente de sucesos

Sea una colección de sucesos **A₁, A₂, A₃, A₄...**, tales que la unión de todos ellos forman el espacio muestral, y sus intersecciones son **disjuntas** ($P(A_i \cap A_j) = 0; i \neq j$)

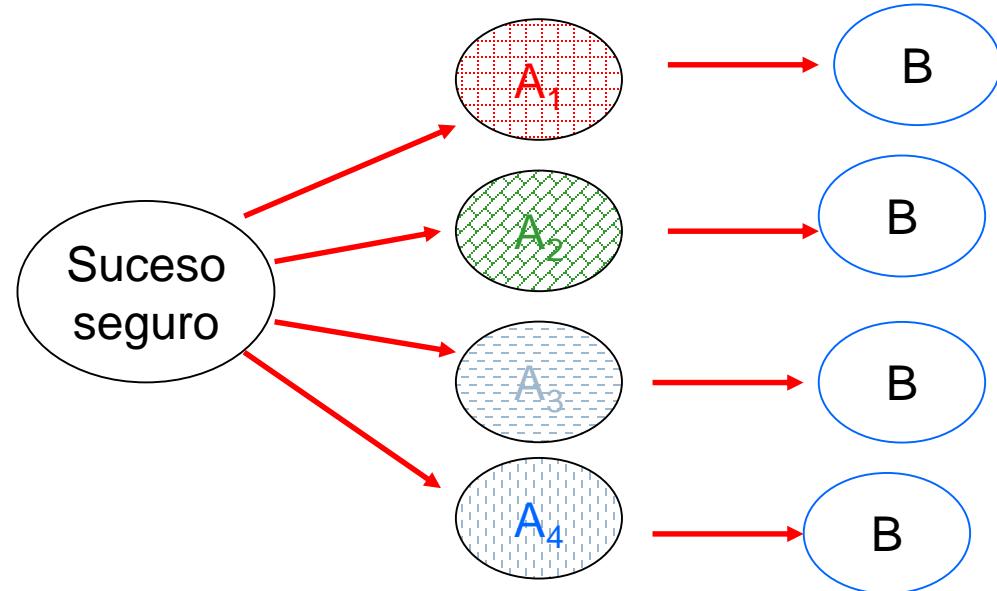
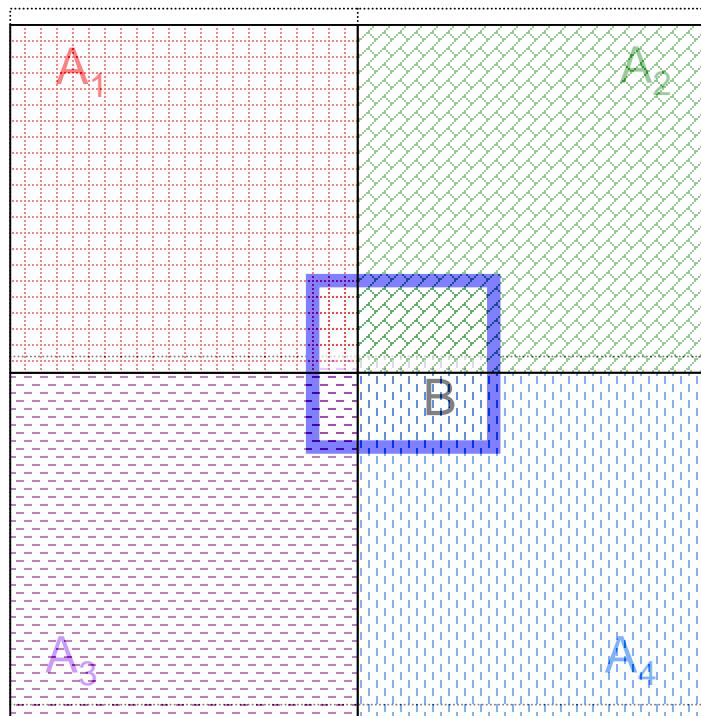


$$\begin{aligned} P(A) &= P(\text{suceso seguro}) = \\ &= P(A_1) + P(A_2) + P(A_3) + P(A_4) = 1 \end{aligned}$$



Sistema exhaustivo y excluyente de sucesos

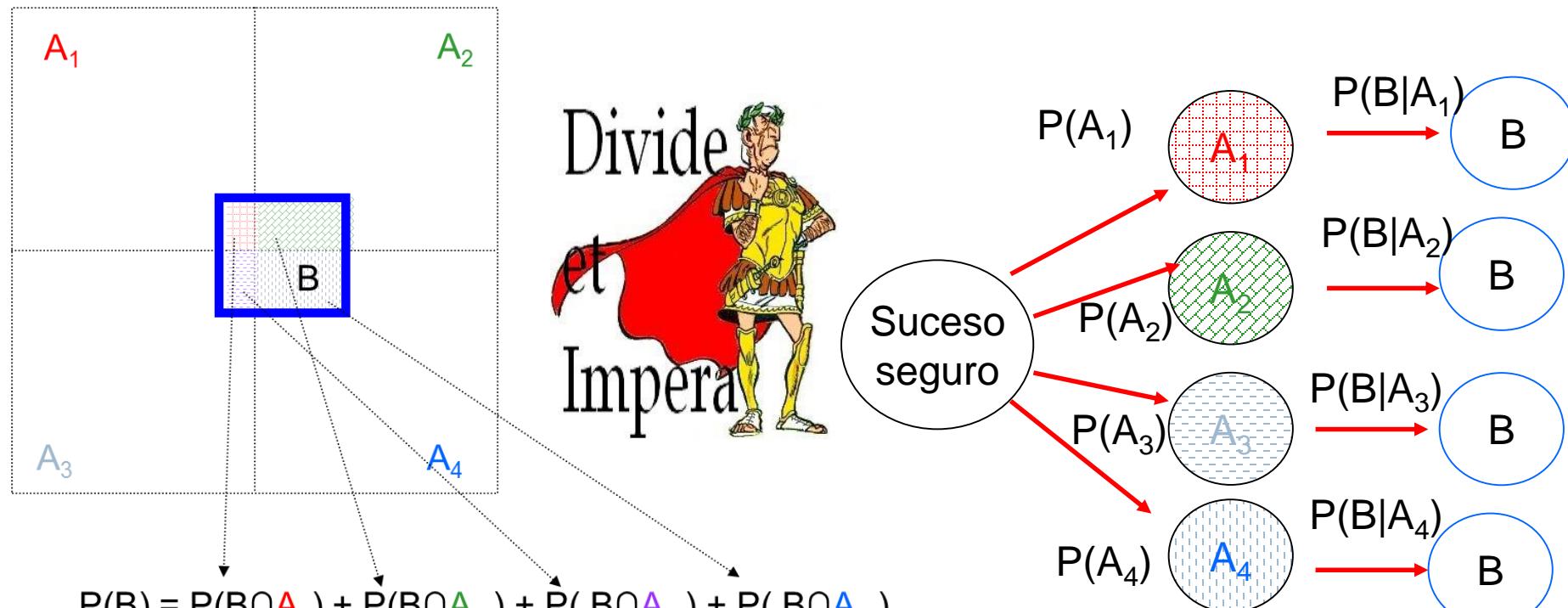
Todo suceso **B**, puede ser **descompuesto** en componentes de dicho sistema: $B = (B \cap A_1) \cup (B \cap A_2) \cup (B \cap A_3) \cup (B \cap A_4)$



Podemos por tanto **descomponer el problema B en subproblemas** más simples.

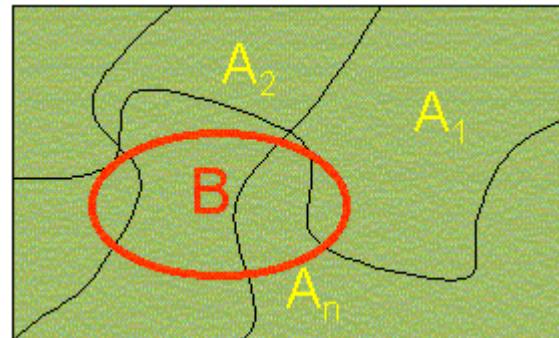
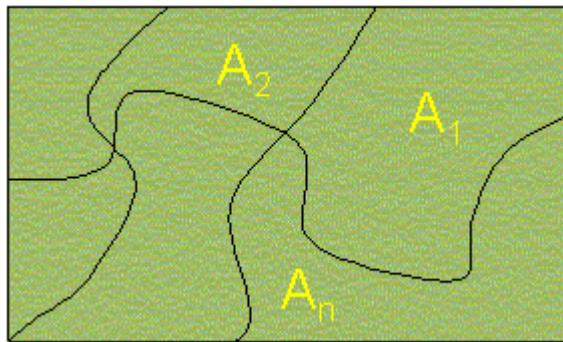
Teorema de la probabilidad total

Si **conocemos** la probabilidad de **B** en cada uno de los componentes de un sistema **exhaustivo y excluyente de sucesos**, entonces **podemos calcular la probabilidad de B**.



Teorema de la probabilidad total

Se llama **partición** a conjunto de sucesos A_i tales que $A_1 \cup A_2 \cup \dots \cup A_n = \Omega$ y $A_i \cap A_j = \emptyset \forall i \neq j$ es decir un conjunto de sucesos mutuamente excluyentes y que cubren todo el espacio muestral



Regla de la probabilidad total:

Si un conjunto de sucesos A_i forman una partición del espacio muestral y $p(A_i) \neq 0 \forall A_i$, para cualquier otro suceso B se cumple

$$p(B) = p(B|A_1)p(A_1) + p(B|A_2)p(A_2) + \dots + p(B|A_n)p(A_n) = \sum_{i=1}^n p(B|A_i)p(A_i)$$

Teorema de la probabilidad total

▶ Ejemplo

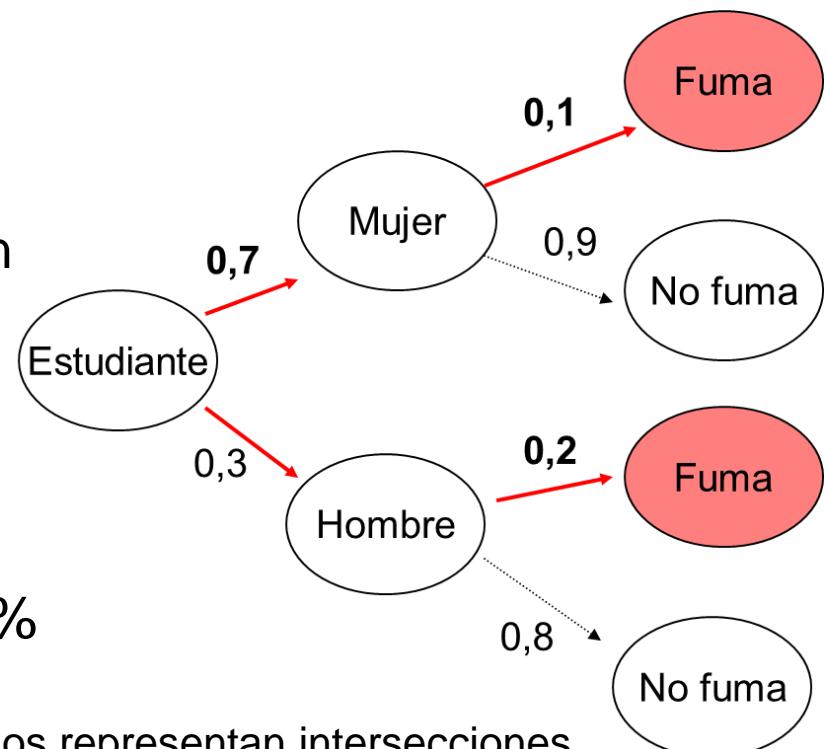
En un aula el 70% de las alumnas son mujeres. De ellas el 10% son fumadoras. De los hombres, son fumadores el 20%.

- ¿Qué porcentaje de fumadores hay?

■ Solución:

Aplicamos el teorema de Probabilidad Total ya que Hombres y mujeres forman un sistema exhaustivo y excluyente de sucesos

$$\begin{aligned} P(F) &= P(M \cap F) + P(H \cap F) = \\ P(M)P(F|M) + P(H)P(F|H) &= \\ = 0,7 \times 0,1 + 0,3 \times 0,2 &= 0,13 = 13\% \end{aligned}$$



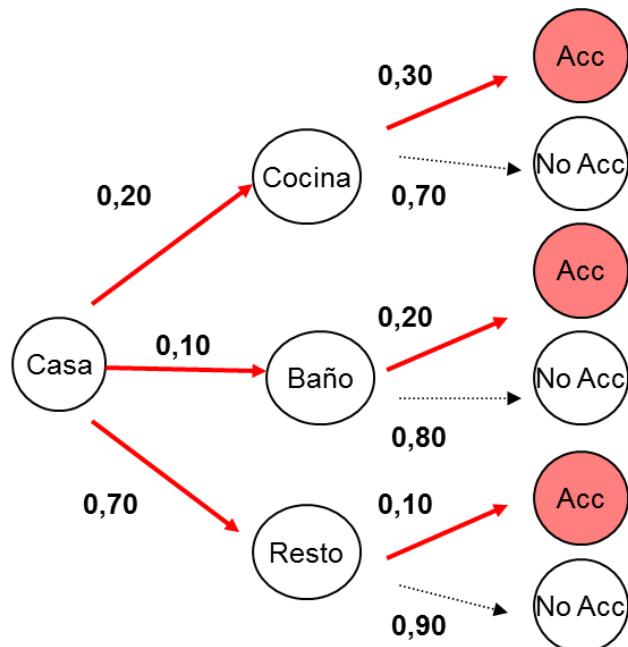
- Los caminos a través de nodos representan intersecciones
- Las bifurcaciones representan uniones disjuntas

Teorema de la probabilidad total

► Ejemplo 2

El 20% del tiempo que se está en una casa transcurre en la cocina, el 10% en el baño y el resto entre el salón y el dormitorio. Por otro lado la probabilidad de tener un accidente doméstico estando en la cocina es de 0,30 de tenerlo estando en el baño es de 0,20 y de tenerlo fuera de ambos de 0,10.

- ¿Cuál es la probabilidad de tener un accidente doméstico?



SOLUCIÓN:

Aplicamos el teorema de Probabilidad Total ya que C (Cocina), B (Baño) y R (salón-Dormitorio) forman un sistema exhaustivo y excluyente de sucesos

$$\begin{aligned} P(A) &= P(A \cap C) + P(A \cap B) + P(A \cap R) = \\ P(C)P(A|C) + P(B)P(A|B) + P(R)P(A|R) &= 0,2 \times \\ 0,3 + 0,1 \times 0,2 + 0,7 \times 0,1 &= 0,15 = 15\% \end{aligned}$$

Teorema de la probabilidad total

► Ejemplo 3

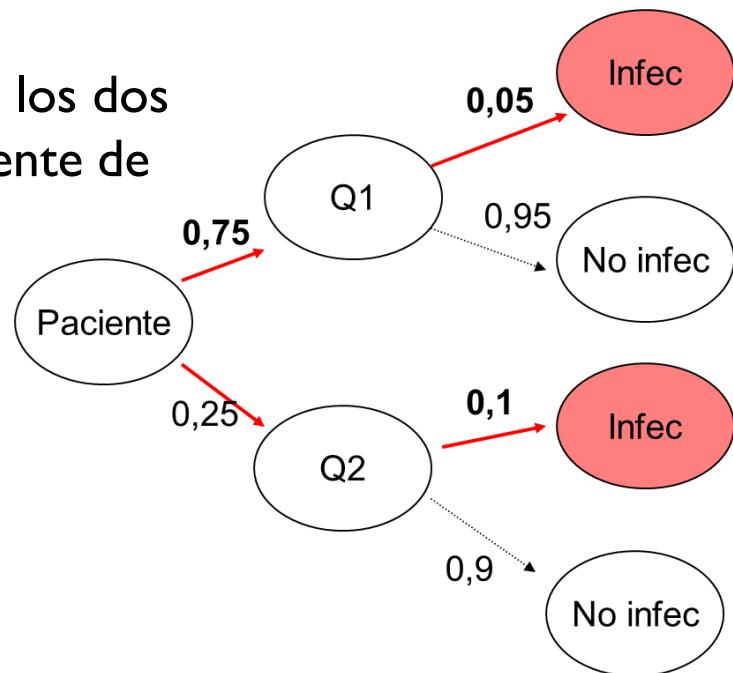
En un centro hospitalario hay dos quirófanos. El 1º se usa el 75% de veces para operar. En el 1º la frecuencia de infección es del 5% y en el 2º del 10%.

- ¿Qué probabilidad de infección hay?

SOLUCIÓN:

Aplicamos el teorema de Probabilidad Total ya que los dos quirófanos forman un sistema exhaustivo y excluyente de sucesos

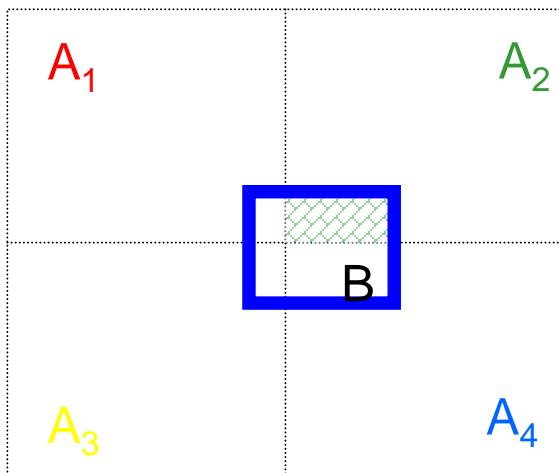
$$\begin{aligned} P(I) &= P(Q1 \cap I) + P(Q2 \cap I) \\ &= P(Q1)P(I|Q1) + P(Q2)P(I|Q2) \\ &= 0,75 \times 0,05 + 0,25 \times 0,1 = 0,0625 = 6,25\% \end{aligned}$$



Teorema de Bayes

- ▶ Sean A_1, A_2, \dots, A_n eventos mutuamente excluyentes, de los cuales uno debe ocurrir, entonces:

$$P(A_j | B) = \frac{P(B \cap A_j)}{P(B)}$$



donde $P(B)$ se puede calcular usando el teorema de la probabilidad total:

$$\begin{aligned} P(B) &= P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_3) + P(B \cap A_4) \\ &= P(B|A_1) P(A_1) + P(B|A_2) P(A_2) + \dots \end{aligned}$$

$$P(A_j | B) = \frac{P(A_j) \cdot P(B | A_j)}{\sum_{j=1}^n P(A_j) \cdot P(B | A_j)}, \quad j = 1, 2, \dots, n$$

Dada la probabilidad de que ocurra un suceso B en cada uno de los componentes de un sistema exhaustivo y excluyente de sucesos, entonces puede determinarse la probabilidad (a posteriori) de ocurrencia de cada A_i .

Teorema de Bayes

▶ Ejemplo I

- Se hace un estudio en un centro médico, se comprueba que el 90% de los fumadores de los que se sospechaba que tenían cáncer lo tenía, mientras que el 5% de los no fumadores lo padecía. Si la proporción de fumadores es de 0,45 ¿Cuál es la probabilidad de que un paciente con cáncer pulmonar elegido al azar sea fumador?

B representa el evento paciente con cáncer y A_1 y A_2 los eventos paciente fumador y no fumador

Aplicando el teorema de Bayes

$$P(A_1 | B) = \frac{P(A_1) \cdot P(B | A_1)}{P(A_1) \cdot P(B | A_1) + P(A_2) \cdot P(B | A_2)}$$

$$P(A_1 | B) = \frac{0,45 \times 0,9}{0,45 \times 0,9 + 0,55 \times 0,05} = 0,9364$$

Teorema de Bayes

▶ Ejemplo 2

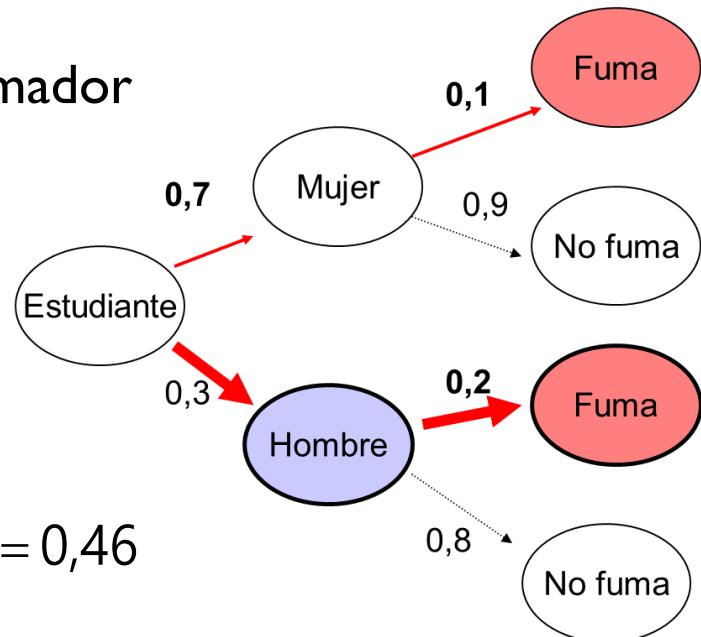
En este aula el 70% de los alumnos son mujeres. De ellas el 10% son fumadoras. De los varones, son fumadores el 20%.

1. ¿Qué porcentaje de fumadores hay?
2. Se elige a un individuo al azar y es... fumador
¿Probabilidad de que sea un hombre?

SOLUCIÓN:

1) $P(F) = 0,7 \times 0,1 + 0,3 \times 0,2 = 0,13$
(Resuelto antes)

2) $P(H|F) = \frac{P(H \cap F)}{P(F)} = \frac{P(H) \cdot P(F|H)}{P(F)} = \frac{0,3 \cdot 0,2}{0,13} = 0,46$



Teorema de Bayes

▶ Ejemplo 3

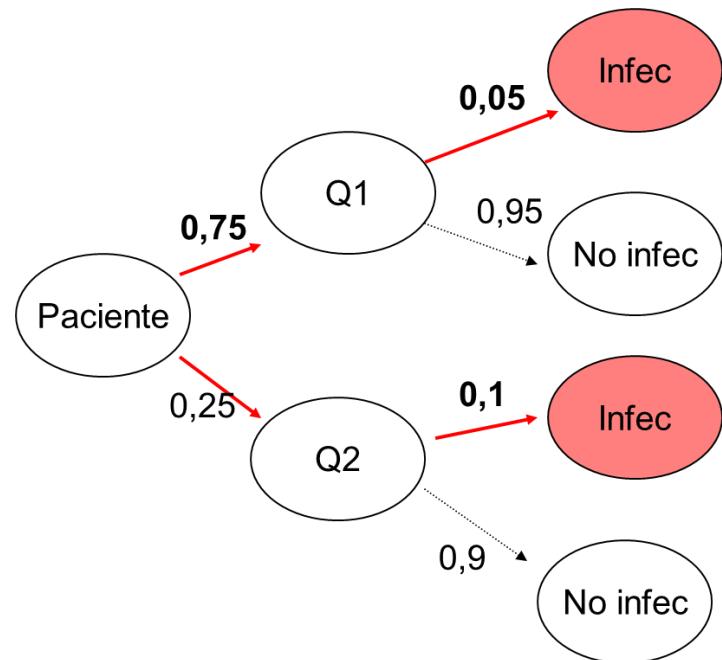
En un centro hay dos quirófanos. El 1º se usa el 75% de veces para operar. En el 1º la frecuencia de infección es del 5% y en el 2º del 10%.

- 1) ¿Qué probabilidad existe de que se produzca una infección? Si se ha producido una infección
- 2) ¿Qué probabilidad hay de que sea en el Q1?

SOLUCIÓN

1) $P(I) = 0,0625 = 6,25\%$ (ejemplo anterior)

$$2) P(Q1|I) = \frac{P(Q1 \cap I)}{P(I)} = \frac{P(Q1) \cdot P(I|Q1)}{P(I)}$$
$$= \frac{0,75 \cdot 0,05}{0,0625} = 0,6$$



Teorema de Bayes

► Ejemplo 4

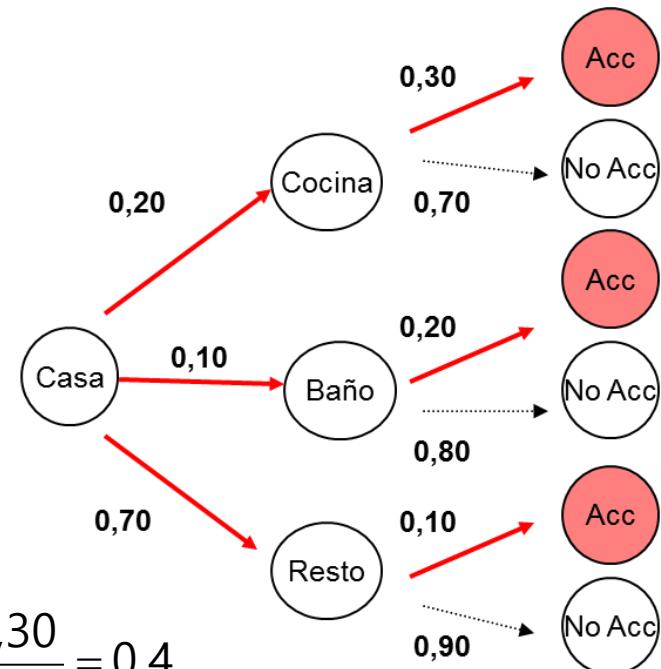
El 20% del tiempo que se está en una casa transcurre en la cocina, el 10% en el baño y el resto entre el salón y el dormitorio. Por otro lado la probabilidad de tener un accidente doméstico estando en la cocina es de 0,30 de tenerlo estando en el baño es de 0,20 y de tenerlo fuera de ambos de 0,10.

1. ¿Cuál es la probabilidad de que se produzca un accidente doméstico?
2. Se ha producido un accidente, ¿cuál es la probabilidad de que haya sido en la cocina?

Solución

1. $P(A) = 0,15$ (ya calculado)

2.
$$P(C|A) = \frac{P(C \cap A)}{P(A)} = \frac{P(C) \cdot P(A|C)}{P(A)} = \frac{0,20 \cdot 0,30}{0,15} = 0,4$$



EJERCICIO

El portero titular de un equipo de fútbol para 8 de cada 10 penaltis, mientras que el suplente solo para 5. El portero suplente juega, por término medio, 15 minutos en cada partido (90 minutos).

- a) Si en un partido se lanzan tres penaltis contra este equipo, ¿cuál es la probabilidad de que se paren los tres?

- b) Si se lanza un penalti y no se para ¿cuál es la probabilidad de que estuviera jugando el portero titular?

EJERCICIO

SOLUCIÓN

Se consideran los sucesos:

P = el portero para un penalti (\bar{P} = el portero no para el penalti)

T = juega el portero titular

S = juega el portero suplente ($S = \bar{T}$)

Con probabilidades:

$$P(S) = \frac{15\text{min}}{90\text{min}} = \frac{1}{6}$$

$$P(T) = 1 - P(S) = 1 - \frac{1}{6} = \frac{5}{6}$$

$$P(P/T) = \frac{8}{10} \rightarrow P(\bar{P}/T) = 1 - P(P/T) = 1 - \frac{8}{10} = \frac{2}{10}$$

$$P(P/S) = \frac{5}{10} \rightarrow P(\bar{P}/S) = 1 - P(P/S) = 1 - \frac{5}{10} = \frac{5}{10}$$

EJERCICIO

- a) Si en un partido se lanzan tres penaltis contra este equipo, ¿cuál es la probabilidad de que se paren los tres?

La probabilidad de que se pare un penalti cualquiera es, utilizando el teorema de la probabilidad total, con los sucesos T y S como sistema exhaustivo y excluyente de sucesos:

$$P(P) = P(P/T) * P(T) + P(P/S) * P(S) = 0'75 = 75\% \text{ (ramas en verde)}$$

Si se lanzan 3 penaltis, se consideran los sucesos:

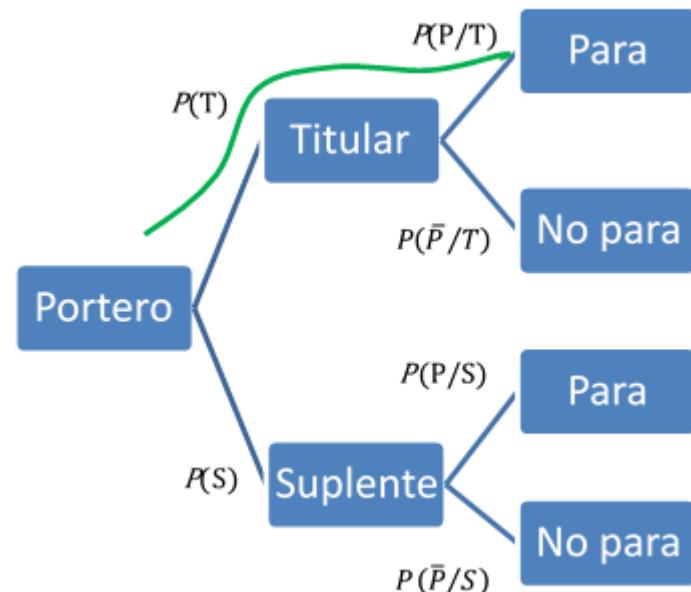
- P_1 = para el primer penalti
- P_2 = para el segundo penalti
- P_3 = para el tercer penalti

Estos tres sucesos son independientes entre si y con probabilidades:

$$P(P_1) = P(P_2) = P(P_3) = 0'75$$

La probabilidad de que se paren los tres penaltis es:

$$P(P_1 \cap P_2 \cap P_3) = P(P_1) * P(P_2) * P(P_3) = (0'75)^3 \approx 0,4219 = 42'19\%$$



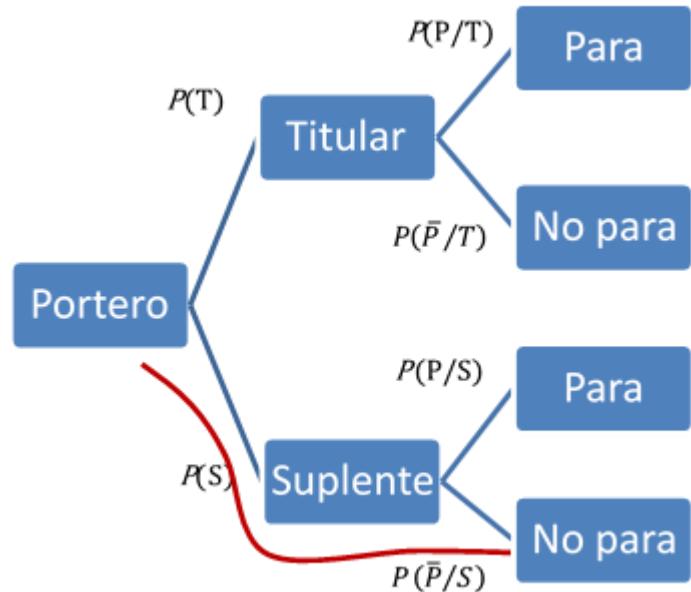
EJERCICIO

- b) Si se lanza un penalti y no se para ¿cuál es la probabilidad de que estuviera jugando el portero titular?

Para calcular $P(\bar{P}/T)$ se aplica el teorema de Bayes con los sucesos T y S como un sistema completo y exhaustivo:

$$P(T/\bar{P}) = \frac{P(\bar{P}/T)*P(T)}{P(\bar{P}/T)*P(T)+P(\bar{P}/S)*P(S)} \quad (\text{rama (1)/ ramas (1) y (2) en rojo})$$

$$\begin{aligned} P(T/\bar{P}) &= \frac{P(\bar{P} / T) * P(T)}{P(\bar{P}/T) * P(T) + P(\bar{P}/S) * P(S)} = \\ &= \frac{\frac{1}{5} \cdot \frac{5}{6}}{\frac{1}{5} \cdot \frac{5}{6} + \frac{1}{2} \cdot \frac{1}{6}} = \frac{2}{3} = 0'6667 = 66'67\% \end{aligned}$$



Combinatoria – Variaciones

Se llama **variaciones ordinarias de m elementos tomados de n en n** ($m \geq n$) a los distintos grupos formados por n elementos de forma que:

No entran todos los elementos

Sí importa el orden

No se repiten los elementos

$$V_m^n = m(m-1)(m-2)(m-3) \cdots (m-n+1)$$

También podemos calcular las **variaciones mediante factoriales**:

$$V_m^n = \frac{m!}{(m-n)!}$$

Se llaman **variaciones con repetición de m elementos tomados de n en n** a los distintos grupos formados por n elementos de manera que:

No entran todos los elementos si $m > n$. Sí pueden entrar todos los elementos si $m \leq n$

Sí importa el orden.

Sí se repiten los elementos

$$VR_m^n = m^n$$

$$VR_{m,n}$$

También se puede escribir como:

Combinatoria – Variaciones

Se va a celebrar la final de salto de longitud en un torneo de atletismo. Participan 8 atletas. ¿De cuántas formas pueden repartirse las tres medallas: oro, plata y bronce?

Elementos disponibles: 8 atletas, $m = 8$.

Elementos por grupo: 3 medallas, $n = 3$

$$V_m^n = m \cdot (m - 1) \cdot (m - 2) \cdots (m - n + 1)$$

$$V_8^3 = 8 \cdot 7 \cdot 6 = 336 \text{ formas}$$

¿Influye el orden de colocación de los elementos?

Si, no es lo mismo recibir oro, plata o bronce.

Pueden ser variaciones o permutaciones.

¿Cogemos todos los elementos disponibles? $m = 8, n = 3$

No sólo 3 de ellos son por tanto variaciones.

¿Se pueden repetir los elementos?

No un mismo atleta no puede llevarse más de una medalla.

Variaciones sin repetición de 8 elementos (m) tomados de 3 en 3 (n).

Utilizando la fórmula $\Rightarrow V_m^n = \frac{m!}{(m - n)!}$

$$V_8^3 = \frac{8!}{(8 - 3)!} = \frac{8 \cdot 7 \cdot 6 \cdot 5!}{5!} = 336 \text{ formas}$$

Combinatoria – Variaciones

El sistema de matrículas de vehículos consiste en un número de 4 dígitos seguido de un bloque de 3 letras consonantes.
(Ejemplo: 1614 - MRM).

¿Cuantas placas hay con un determinado bloque de letras?

¿Cuantas placas hay con un determinado bloque de letras?

Disponemos de 22 consonantes, $m = 22$.

Formamos grupos de 3 letras, $n = 3$.

¿Influye el orden de colocación de los elementos?

Sí, si cambiamos el orden tenemos matrículas distintas.

Pueden ser **variaciones** o **permutaciones**.

¿Cogemos todos los elementos disponibles? $m = 22$, $n = 3$

No sólo 3 de ellos son por tanto **variaciones**.

¿Se pueden repetir los elementos?

Sí.

Variaciones con repetición de 22 elementos (m) tomados de 3 en 3 (n).

$$VR_{\frac{n}{m}}^n = m^n \Rightarrow VR_{22}^3 = 22^3 = 10648 \text{ placas.}$$

Combinatoria – Permutaciones

Las **permutaciones** o, también llamadas, ordenaciones son aquellas formas de agrupar los elementos de un conjunto teniendo en cuenta que:

Influye el orden en que se colocan.

Tomamos todos los elementos de que se disponen.

Permutaciones SIN repetición cuando todos los elementos de que disponemos son distintos.

$$P_n = n!$$

Permutaciones con repetición de n elementos donde el primer elemento se repite a veces, el segundo b veces, el tercero c veces, ... $n = a + b + c + \dots$

Son los distintos grupos que pueden formarse con esos n elementos de forma que:

Sí entran todos los elementos, Sí importa el orden, Sí se repiten los elementos

$$PR_n^{a,b,c,\dots} = \frac{P_n}{a! \cdot b! \cdot c! \cdot \dots}$$

Combinatoria – Permutaciones

¿Cuántos números de 5 cifras diferentes se puede formar con los dígitos: 1, 2, 3, 4, 5?

$$m = 5 \quad n = 5$$

Sí entran todos los elementos

Sí importa el orden

No se repiten los elementos. El enunciado nos pide que las cifras sean diferentes

$$P_5 = 5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$$

¿De cuántas formas distintas pueden sentarse ocho personas en una fila de butacas?

Sí entran todos los elementos. Tienen que sentarse las 8 personas

Sí importa el orden

No se repiten los elementos. Una persona no se puede repetir

$$P_8 = 8! = 40320$$

Combinatoria – Permutaciones

Con las cifras 2, 2, 2, 3, 3, 3, 3, 4, 4; ¿cuántos números de nueve cifras se pueden formar?

$$m = 9 \quad a = 3 \quad b = 4 \quad c = 2 \quad a + b + c = 9$$

Sí entran todos los elementos

Sí importa el orden

Sí se repiten los elementos

$$PR_9^{3,4,2} = \frac{9!}{3! \cdot 4! \cdot 2!} = 1260$$

En el palo de señales de un barco se pueden izar tres banderas rojas, dos azules y cuatro verdes. ¿Cuántas señales distintas pueden indicarse con la colocación de las nueve banderas?

Sí entran todos los elementos

Sí importa el orden

Sí se repiten los elementos

$$PR_9^{3,2,4} = \frac{9!}{3! \cdot 2! \cdot 4!} = 1260$$

Combinatoria – Combinaciones

Se llama **combinaciones de m elementos tomados de n en n ($m \geq n$)** a todas las agrupaciones posibles que pueden hacerse con los m elementos de forma que:

No entran todos los elementos

No importa el orden

No se repiten los elementos

$$C_m^n = \frac{V_m^n}{P_n}$$

También podemos calcular las **combinaciones** mediante **factoriales**:

$$C_m^n = \frac{m!}{n!(m-n)!}$$

Las **combinaciones** se denotan por

$$C_m^n \text{ o } C_{m,n}$$

Combinatoria – Combinaciones

Calcular el número de combinaciones de 10 elementos tomados de 4 en 4.

$$C_{10}^4 = \frac{10 \cdot 9 \cdot 8 \cdot 7}{4 \cdot 3 \cdot 2 \cdot 1} = 210$$

$$C_{10}^4 = \frac{10!}{4! \cdot 6!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6!}{4 \cdot 3 \cdot 2 \cdot 1 \cdot 6!} = 10 \cdot 3 \cdot 7 = 210$$

En una clase de 35 alumnos se quiere elegir un comité formado por tres alumnos.
¿Cuántos comités diferentes se pueden formar?

No entran todos los elementos

No importa el orden: Juan, Ana

No se repiten los elementos

$$C_{35}^3 = \frac{35 \cdot 34 \cdot 33}{3 \cdot 2 \cdot 1} = 6545$$

Combinatoria – Combinaciones

Las **combinaciones con repetición de m elementos tomados de n en n ($m \geq n$)**, son los distintos grupos formados por n elementos de manera que:

No entran todos los elementos

No importa el orden

Sí se repiten los elementos

$$CR_m^n = \binom{m+n-1}{n} = \frac{(m+n-1)!}{n!(m-1)!}$$

Ejemplo

En una bodega hay en unos cinco tipos diferentes de botellas. ¿De cuántas formas se pueden elegir cuatro botellas?

No entran todos los elementos. Sólo elije 4

No importa el orden. Da igual que elija 2 botellas de anís y 2 de ron, que 2 de ron y 2 de anís

Sí se repiten los elementos. Puede elegir más de una botella del mismo tipo

$$CR_5^4 = \frac{(5+4-1)}{4!(5-1)!} = \frac{8!}{4!.4!} = 70$$

Combinatoria – Combinaciones

ANALISIS COMBINATORIO

Si tengo n elementos y quiero contar cuántos grupos de k elementos puedo hacer, debo de responder a tres preguntas:

- P1: ¿Importa el orden de los k elementos dentro de un grupo?
- P2: ¿Se pueden repetir los elementos dentro de un grupo?
- P3: ¿ $k < n$ o $k = n$? (Si hay repetición puede pasar que $k > n$)

	Variaciones	Variaciones con repetición	Combinaciones	Permutaciones	Permutaciones con repetición
P1	SI	SI	NO	SI	SI
P2	NO	SI	NO	NO	SI
P3	$k < n$	$k < n, k = n$ ó $k > n$	$k < n$	$k = n$	$k = n$
	$V_{n,k} = n \cdot (n-1) \cdot \dots \cdot (n-k+1)$	$VR_{n,k} = n^k$	$C_{n,k} = \binom{n}{k} = \frac{n!}{k!(n-k)!}$	$P_n = n!$	$PR_{k_1, k_2, \dots, k_r} = \frac{n!}{k_1! k_2! \dots k_r!}$ con $k_1 + k_2 + \dots + k_r = n$

Estadística

Tema 4: parte I - Variables Aleatorias

Variables Aleatorias

Estudio de la Población a partir de la Distribución de Probabilidad de las Variables.

No se conocen los valores exactos de las variables, como en estadística descriptiva, pero, a través del estudio de una muestra, se han determinado sus probabilidades de aparición.

Una **variable aleatoria** es una variable que toma **valores numéricos** determinados por el resultado de un experimento aleatorio, es decir es una función que asocia un valor numérico a cada posible resultado de un experimento aleatorio.

- ▶ N° de reclamaciones al mes en un servicio
- ▶ N° de servicios prestados
- ▶ N° de piezas defectuosas producidas la máquina
- ▶ Retraso promedio en los trenes
- ▶ N° de caras al lanzar 6 veces una moneda (valores: 0, 1, 2...)
- ▶ N° de llamadas que recibe un teléfono en una hora
- ▶ Tiempo que esperan los clientes para pagar en un supermercado...

Variables Aleatorias

Las variables aleatorias pueden ser discretas o continuas:

Discretas: el conjunto de posibles valores es numerable. Suelen estar asociadas a experimentos en que se mide el número de veces que sucede algo.

Continuas: el conjunto de posibles valores es no numerable. Puede tomar todos los valores de un intervalo. Son el resultado de medir.

Ejemplos:

- ▶ N° de páginas de un libro → discreta
- ▶ Tiempo de respuesta de un servicio → continua
- ▶ N° de preguntas en una clase de una hora → discreta
- ▶ Cantidad de agua consumida en un mes → continua

En la práctica se consideran discretas aquellas variables para las que merece la pena asignar probabilidades a todos los posibles sucesos elementales.

Variables Aleatorias Discretas

Función de probabilidad

Sea X una variable aleatoria discreta con posibles valores $\{x_1, x_2, \dots\}$. Se llama **función de probabilidad** o **función de masa**, al conjunto de probabilidades con las que X toma cada uno de sus valores, es decir, $p_i = P[X = x_i]$, para $i = 1, 2, \dots$.

$$p(x_i) = P(x = x_i) \quad \text{Si } E \text{ es el espacio muestral} \rightarrow \sum_{i \in E} p(x_i) = 1$$

Ejemplo

X = resultado de lanzar un dado. La función de probabilidad es

x	1	2	3	4	5	6
$P[X = x]$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Variables Aleatorias Discretas

Función de probabilidad. Propiedades

- $0 \leq P[X = x_i] \leq 1.$
- $\sum_i P[X = x_i] = 1.$
- $P[X \leq x] = \sum_{i, x_i \leq x} P[X = x_i].$
- $P[X > x] = 1 - P[X \leq x].$

Variables Aleatorias Discretas

Función de Distribución: Función matemática discreta que especifica para cada valor del espacio muestral de una variable discreta la probabilidad de aparición de un valor menor o igual.

$$F(x_i) = P(x \leq x_i)$$

- $F(x_k) = P(x \leq x_k) = \sum_{i=1}^k p(x_i)$ con n valores $\sum_{i=1}^n p(x_i) = 1$

Ejemplo

X = resultado de lanzar un dado. La función de distribución es

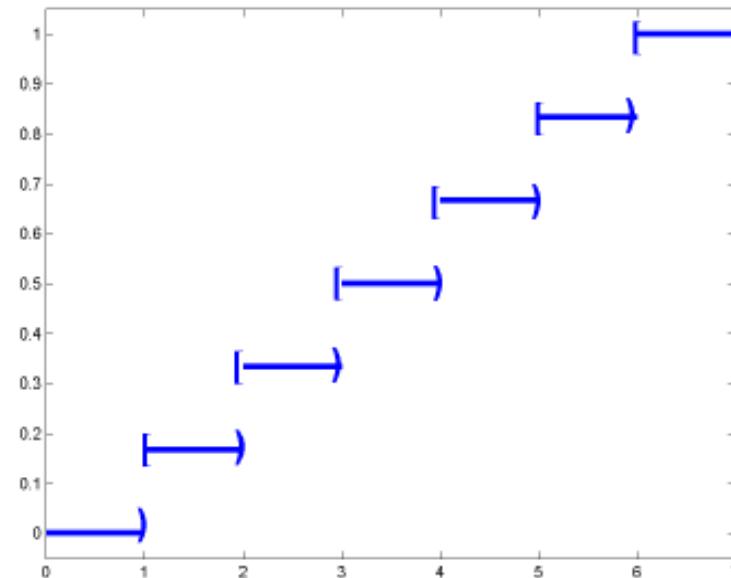
x	1	2	3	4	5	6
$P[X = x]$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
$F(x)$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	$\frac{6}{6}$

Variables Aleatorias Discretas

Función de Distribución

Propiedades

- $F(-\infty) = 0.$
- $F(\infty) = 1.$
- Si $x \leq y$, entonces $F(x) \leq F(y).$



Para X discreta, la función de distribución es de tipo escalón. Cada escalón corresponde a un valor posible de X y el salto corresponde a la probabilidad.

Variables Aleatorias Continuas

La función de probabilidad no tiene sentido en variables aleatorias continuas, porque

$$P(X = x) = 0$$

Para sustituir la función de probabilidad, en variables aleatorias continuas usaremos la **función de densidad**.

Función de Densidad: Función matemática continua que especifica la probabilidad de aparición de cada valor del espacio muestral de una variable continua.

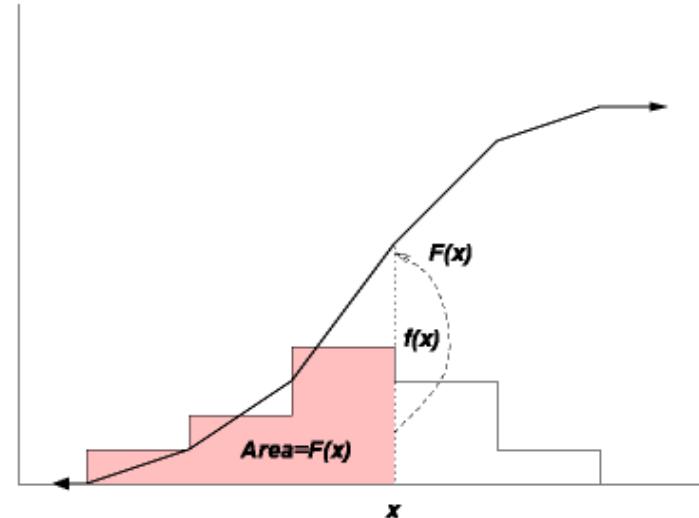
$$p(x_i) = f(X = x_i) = f(x) \quad \text{Como } f \text{ es continua} \rightarrow \int_{-\infty}^{\infty} f(x) dx = 1$$

Variables Aleatorias Continuas

- ▶ La **función distribución (acumulativa)** para una variable aleatoria **X continua** se define, como en el caso discreto, por la probabilidad de que X tome un valor menor o igual a algún x específico

$$P(X \leq x) = F(x) = \int_{-\infty}^x f(t)dt$$

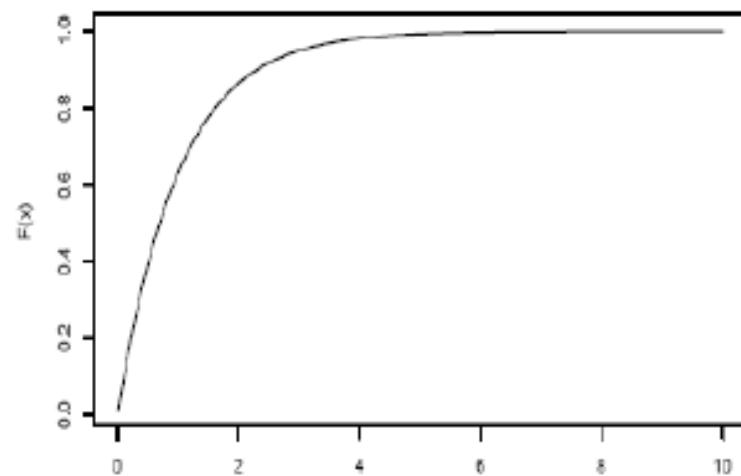
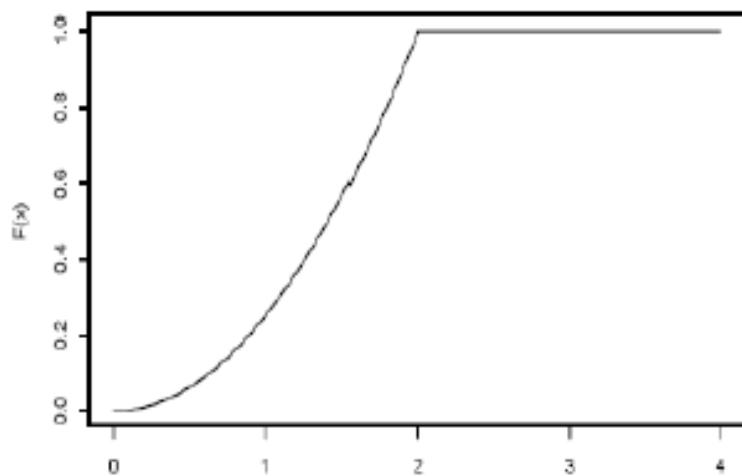
- ▶ Es la **función que asocia a cada valor de una variable, la probabilidad acumulada de los valores inferiores o iguales**
- ▶ Puede entenderse también como la generalización de las frecuencias acumuladas.



Variables Aleatorias Continuas

Propiedades

- $F(-\infty) = 0$.
- $F(\infty) = 1$.
- Si $x \leq y$, entonces $F(x) \leq F(y)$.
- $F(x)$ es continua.



No son funciones de tipo escalón, sino suaves.

Variables Aleatorias Continuas

Función de densidad

Para una variable aleatoria continua X con función de distribución $F(x)$, la **función de densidad** de X es:

$$f(x) = \frac{dF(x)}{dx} = F'(x)$$

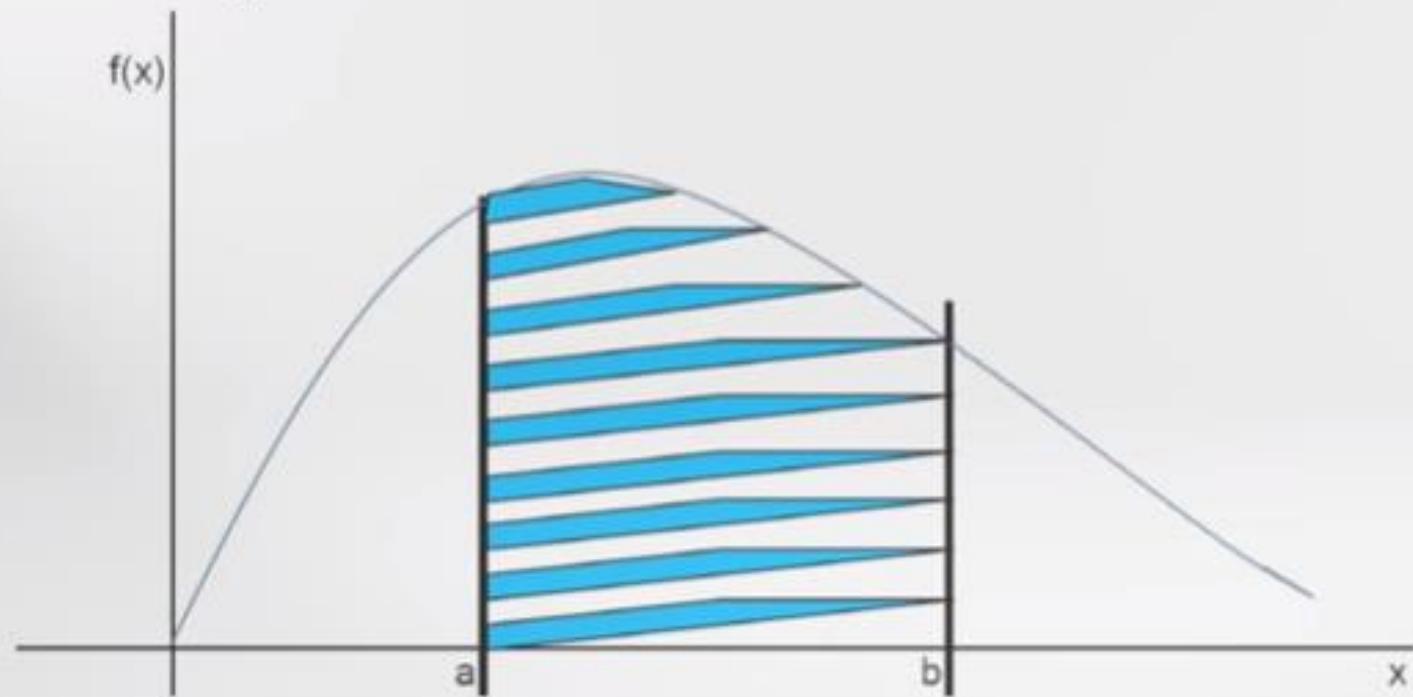
Propiedades

- $f(x) \geq 0 \quad \forall x \in \mathbb{R}$
- $P(a \leq X \leq b) = \int_a^b f(x)dx \quad \forall a, b \in \mathbb{R}$
- $F(x) = P(X \leq x) = \int_{-\infty}^x f(u)du$
- $\int_{-\infty}^{\infty} f(x)dx = 1$

Variables Aleatorias Continuas

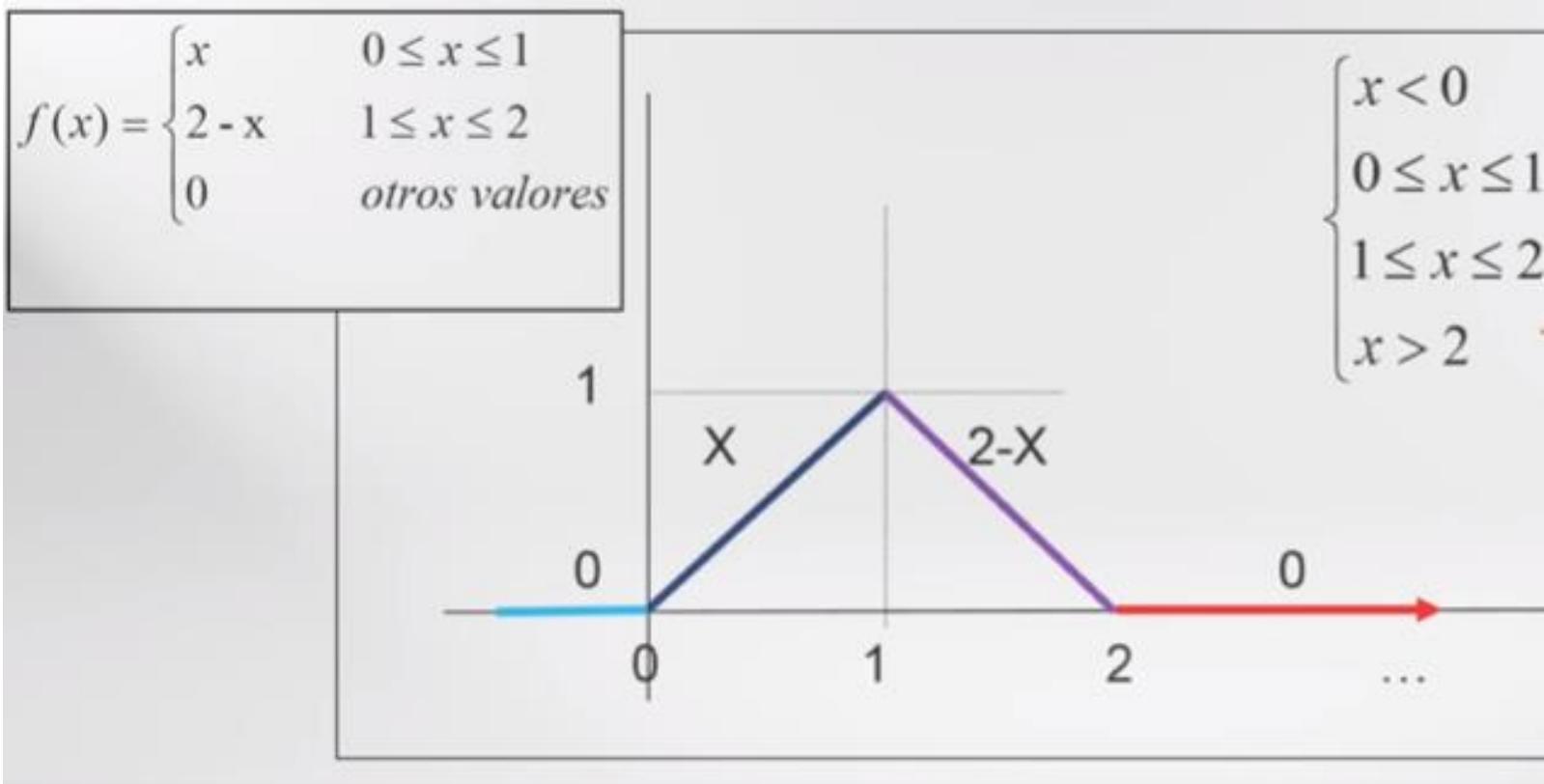
FUNCIÓN DE DENSIDAD DE PROBABILIDAD DE UNA VARIABLE ALEATORIA

$$P(a \leq X \leq b) = \int_a^b f_x(x) dx$$



Variables Aleatorias

Obtén su función de distribución.

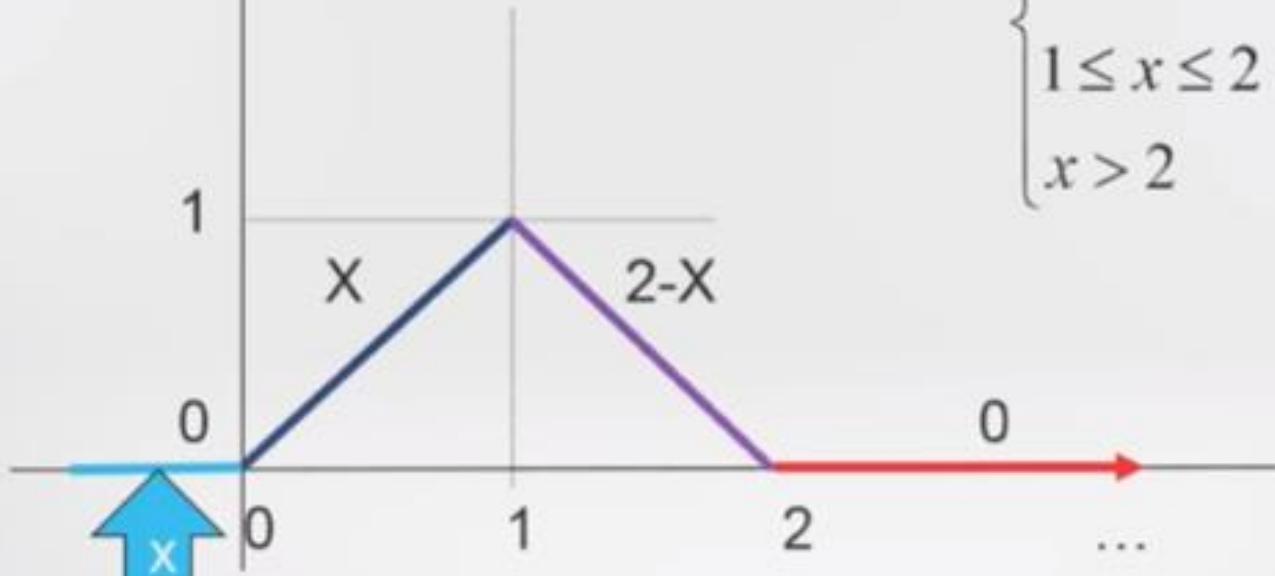


Variables Aleatorias

Primer Tramo F(x):

$$\int_{-\infty}^x 0 \, dx = 0$$

$$\begin{cases} x < 0 \\ 0 \leq x \leq 1 \\ 1 \leq x \leq 2 \\ x > 2 \end{cases}$$



Variables Aleatorias

Segundo Tramo F(x):

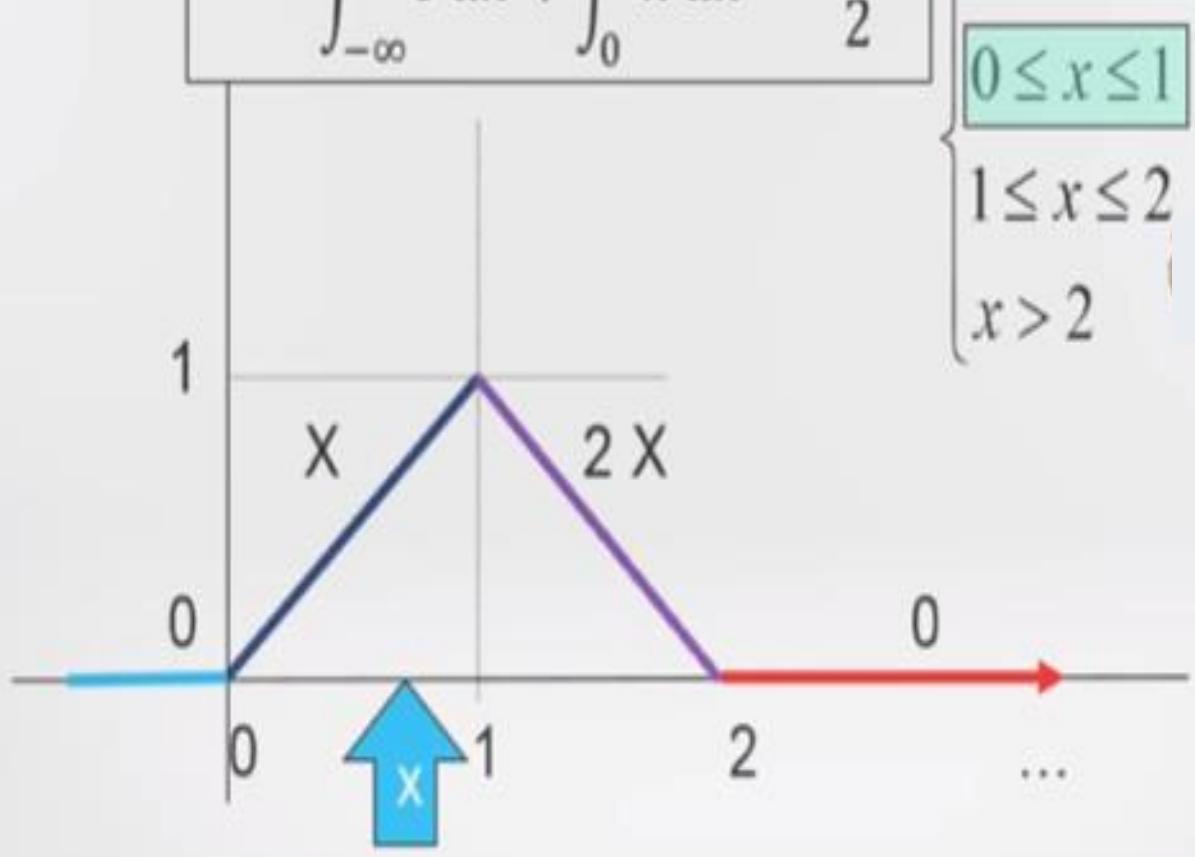
$$\int_{-\infty}^0 0 \, dx + \int_0^x x \, dx = \frac{x^2}{2}$$

$$x < 0$$

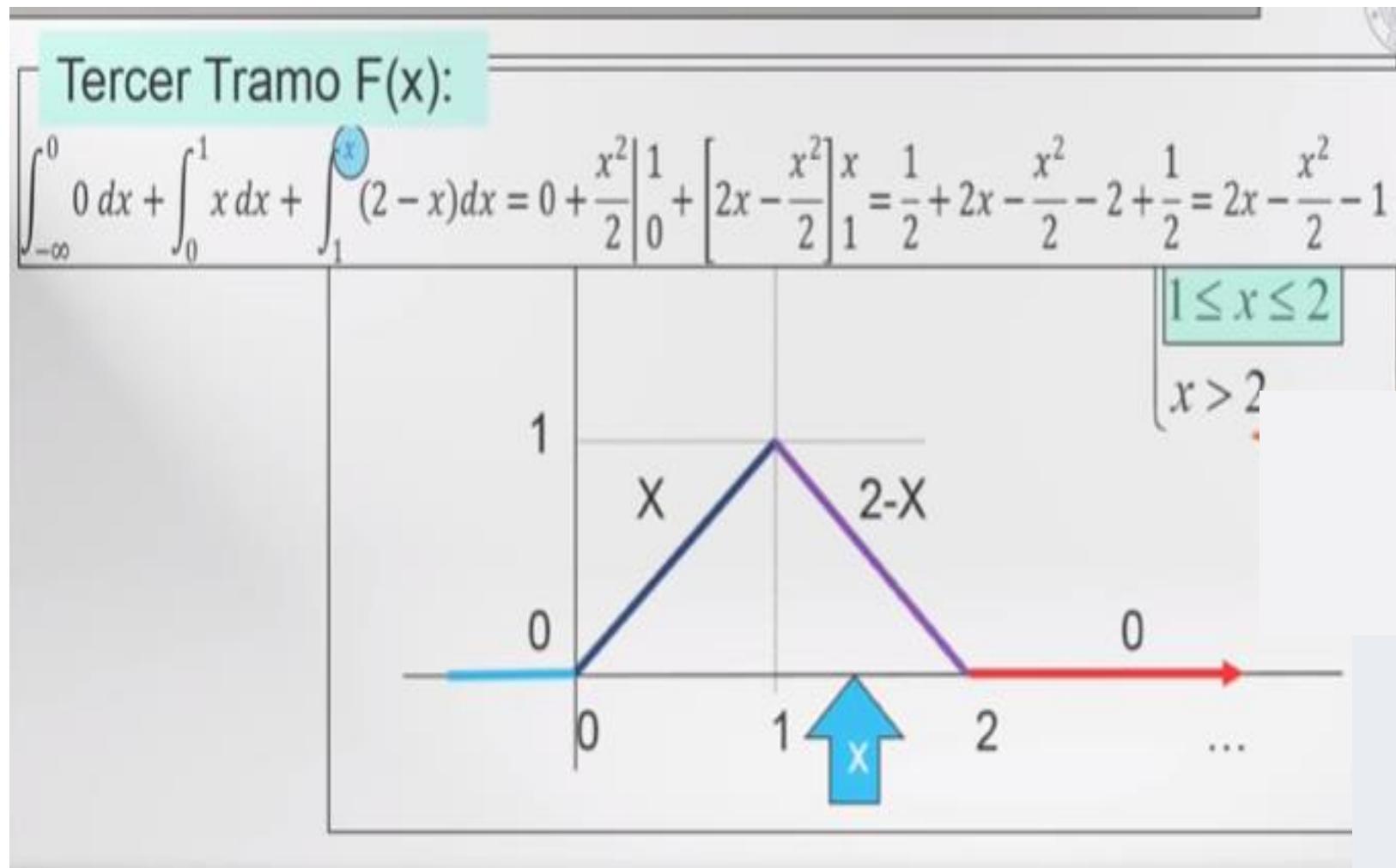
$$0 \leq x \leq 1$$

$$1 \leq x \leq 2$$

$$x > 2$$



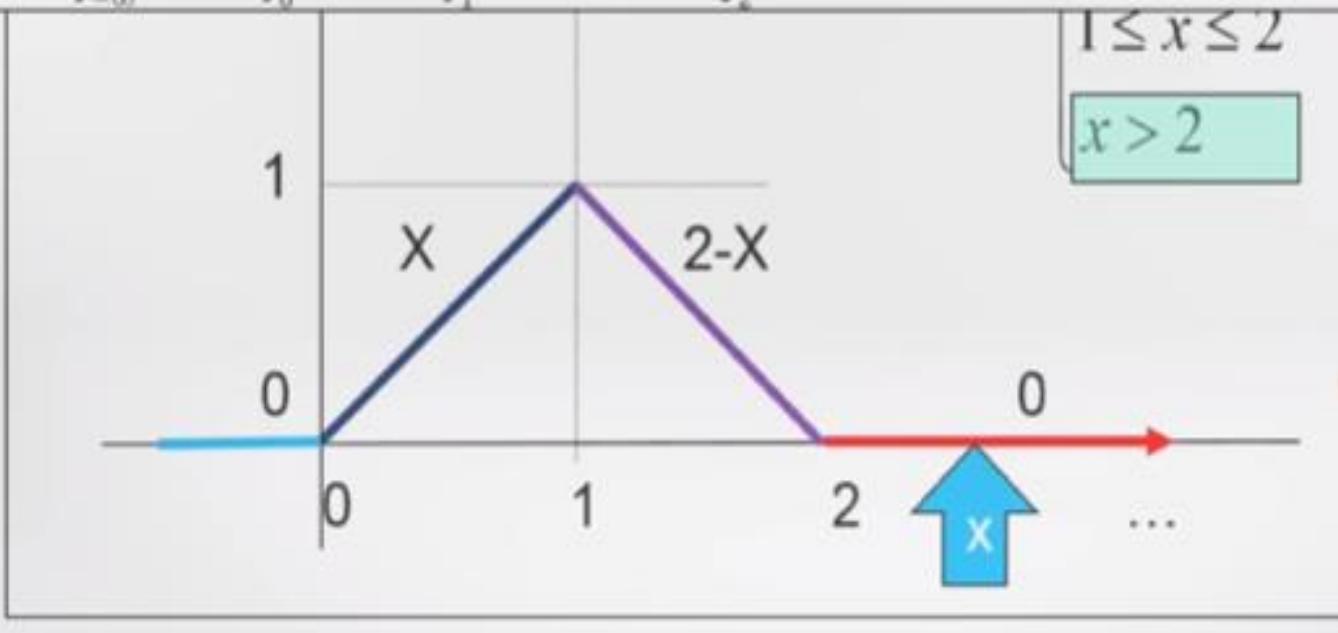
Variables Aleatorias



Variables Aleatorias

Cuarto Tramo F(x):

$$\int_{-\infty}^x 0 dx + \int_0^1 x dx + \int_1^2 (2-x)dx + \int_2^{\infty} 0 dx = 1$$



Variables Aleatorias

Dada la función de densidad de la variable aleatoria X:

$$f(x) = \begin{cases} x & 0 \leq x \leq 1 \\ 2 - x & 1 \leq x \leq 2 \\ 0 & \text{others values} \end{cases}$$

Obtén su función de distribución.

$$F(x) = \begin{cases} 0, & x < 0 \\ \frac{x^2}{2}, & 0 \leq x \leq 1 \\ 2x - \frac{x^2}{2} - 1, & 1 \leq x \leq 2 \\ 1, & x > 2 \end{cases}$$

Variables Aleatorias Continuas

Ejemplo

Una variable aleatoria X tiene función de densidad

$$f(x) = \begin{cases} 12x^2(1-x) & \text{si } 0 < x < 1 \\ 0 & \text{si no} \end{cases}$$

Calcular la $P(X \leq 0,5)$

Calcular la $P(0,2 \leq X \leq 0,5)$

Calcular la $P(X \leq x)$

Variables Aleatorias Continuas

Ejemplo

Una variable aleatoria X tiene función de densidad

$$f(x) = \begin{cases} 12x^2(1-x) & \text{si } 0 < x < 1 \\ 0 & \text{si no} \end{cases}$$

$$P(X \leq 0'5) = \int_{-\infty}^{0'5} f(u)du = \int_0^{0'5} 12u^2(1-u)du = 0'3125$$

$$P(0'2 \leq X \leq 0'5) = \int_{0'2}^{0'5} f(u)du = \int_{0'2}^{0'5} 12u^2(1-u)du = 0'2853$$

$$P(X \leq x) = \int_{-\infty}^x f(u)du = \begin{cases} 0 & \text{si } x \leq 0 \\ 12 \left(\frac{x^3}{3} - \frac{x^4}{4} \right) & \text{si } 0 < x \leq 1 \\ 1 & \text{si } x > 1 \end{cases}$$

Medidas características de una variable aleatoria

Cuando se tiene una variable Discreta o Continua, cuyos valores se pueden inferir con un grado de probabilidad mediante una Función de Probabilidad o una Función de Densidad, respectivamente, se pueden inferir también los mismos parámetros de caracterización de la población que se obtenían para caracterizar poblaciones en Estadística Descriptiva.

Éstos son la media (aquí denominada **Esperanza**) y la **varianza**.

Esperanza (o media):

- Variable Discreta: $\mu_x = E(x) = \sum_{i=1}^n x_i \cdot p(x_i)$
- Variable Continua: $\mu_x = E(x) = \int_{-\infty}^{\infty} x \cdot f(x) dx$

Varianza:

- Variable Discreta: $\sigma_x^2 = \sum_{i=1}^{\infty} (x_i - \mu)^2 \cdot p(x_i)$
- Variable Continua: $\sigma_x^2 = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx = \int_{-\infty}^{\infty} x^2 \cdot f(x) dx - \mu_x^2$

Medidas características de una variable aleatoria

Propiedades de la Esperanza

$$E[aX] = a E[X], \quad a \in \mathbb{R}$$

$$E[X+Y] = E[X] + E[Y]$$

$$E[aX+bY] = aE[X] + bE[Y]; \quad a, b \in \mathbb{R}$$

Propiedades de la Varianza

$$\text{Var}[X] = 0 \Leftrightarrow X \text{ es constante}$$

$$a \text{ constante} \Rightarrow \text{Var}[aX] = a^2 \text{Var}[X]$$

$$a, b \text{ constantes} \Rightarrow \text{Var}[aX+b] = a^2 \text{Var}[X]$$

Medidas características de una variable aleatoria

▶ Ejemplo

- Calcular el beneficio esperado (beneficio medio) con una apuesta de 100 € a la ruleta:
 - a) un número cualquiera
 - b) rojo frente a negro
- a) Los resultados posibles de una jugada en la ruleta son los números (0, 1, ..., 36) con probabilidades 1/37. Si apostamos 100 € a un número, la variable aleatoria X, beneficio obtenido, tomará los valores siguientes:
 - $x = -100$, si ocurre cualquier número distinto al apostado, $P(x = -100) = 36/37$.
 - $x = 3.500$, si ocurre el número elegido, $P(x = 3.500) = 1/37$

Por tanto:

$$E(x) = -100 \cdot (36/37) + 3.500 \cdot (1/37) = -2,7 \text{ €}$$

que supone una pérdida del 2,7% de la cantidad invertida: ¡**La banca gana!!**

Medidas características de una variable aleatoria

b) En el segundo caso hay dos resultados posibles:

- +100 (si sale rojo),
- -100 (si sale negro o el cero).

Entonces:

$$E(x) = -100 (19/37) + 100 (18/37) = -2,7 \text{ €}$$

que es el mismo resultado anterior.

Todas las apuestas de la ruleta tienen la misma esperanza de pérdida.

¡¡La banca vuelve a ganar!!

Medidas características de una variable aleatoria

Ejercicio 1:

Una variable aleatoria X puede tomar los valores 30,40,50 y 60 con probabilidades 0.4,0.2,0.1 y 0.3.

Representar en una tabla la función de probabilidad, y la función de distribución de probabilidad.

Calcular la Esperanza matemática, la varianza y la desviación típica

Ejercicio 2 :

La altura de un cierto árbol sigue una v.a. con la siguiente función de densidad:

$$f(x) = x/12 \quad \text{si } 1 < x < 5 \text{ y } 0 \text{ en el resto.}$$

Calcular la esperanza y la varianza de X.

Solución Ejercicio 1

La tabla de la función de probabilidad, y la función de distribución de probabilidad se han realizado en la sesión de teoría.

$$E(X) = \sum_{i=1}^k x_i P(X = x_i) = 12 + 8 + 5 + 18 = 43$$

Cálculo de la varianza y desviación típica

X	P(X=x)	xP(X = x)	x ² P(X = x)
30	0.4	12	360
40	0.2	8	320
50	0.1	5	250
60	0.3	18	1080
	1	45	2010

$$V(X) = \sum_{i=1}^k x_i^2 P(X = x_i) - E(X)^2 = 2010 - 43^2 = 161$$
$$\sigma = \sqrt{161} = 12,69$$

Solución Ejercicio 2

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx = \int_1^5 \frac{x^2}{12} dx = \frac{1}{36} \left[x^3 \right]_1^5 = \frac{31}{9}$$

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x) dx = \frac{1}{12} \int_1^5 x^3 dx = \frac{1}{48} \left[x^4 \right]_1^5 = 13$$

La esperanza de X ya fue calculada y es: $E[X] = 31/9$.

Por lo tanto:

$$Var[X] = E[X^2] - E[X]^2 = 13 - \left(\frac{31}{9} \right)^2 = 1.1358$$