Estadística Aplicada al Marketing

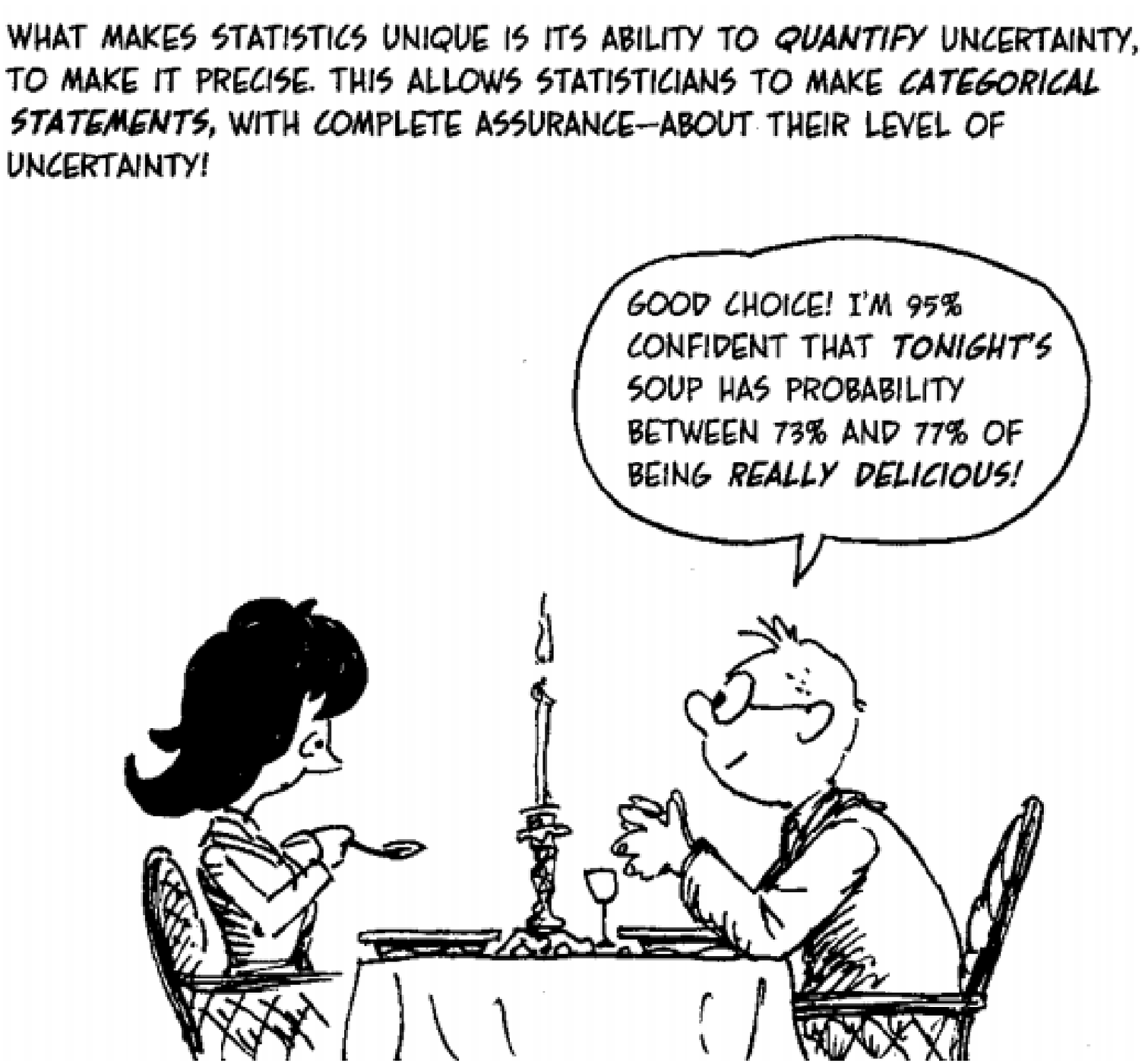
T.1. Introducción a la Estadística

Juan Miguel Navarro Ruiz

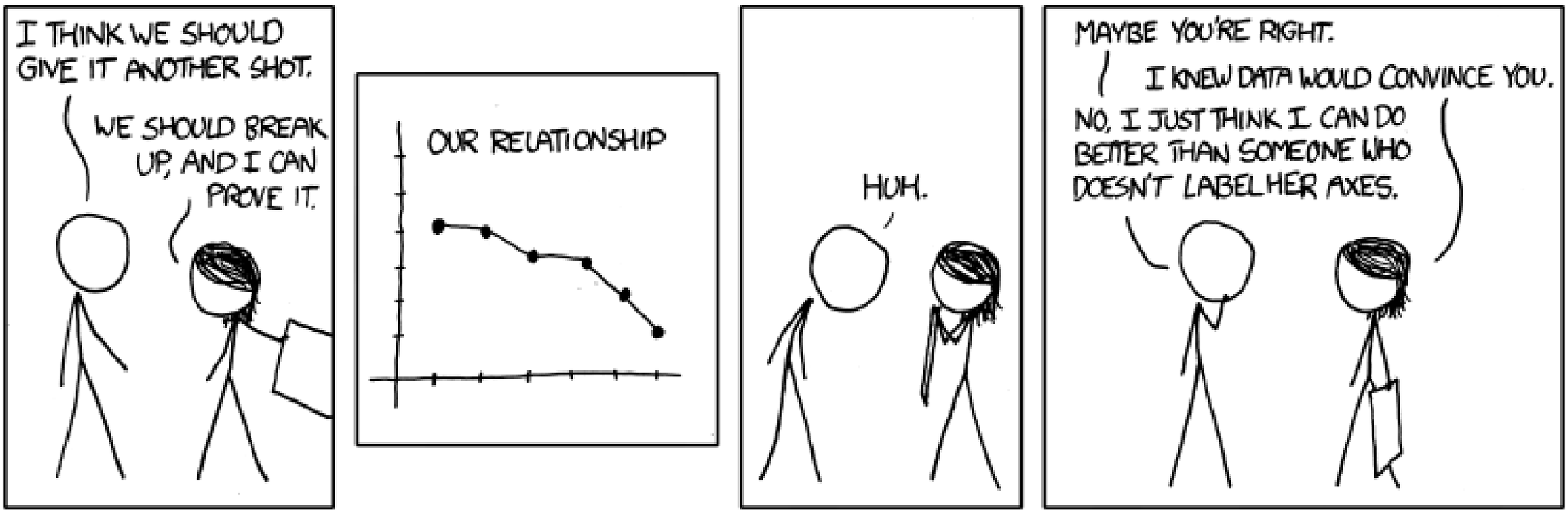
Grado en Marketing y Dirección Comercial

Juan M Navarro- Tlf: (+34) 968 27 88 25
Universidad Católica San Antonio de Murcia - Tlf: (+34) 968 27 88 00 info@ucam.edu - www.ucam.edu

Motivación


WHAT MAKES STATISTICS UNIQUE IS ITS ABILITY TO QUANTIFY UNCERTAINTY, TO MAKE IT PRECISE. THIS ALLOWS STATISTICIANS TO MAKE CATEGORICAL STATEMENTS, WITH COMPLETE ASSURANCE—ABOUT THEIR LEVEL OF UNCERTAINTY!
GOOD CHOICE! I'M 95% CONFIDENT THAT TONIGHT'S SOUP HAS PROBABILITY BETWEEN 73% AND 77% OF BEING REALLY DELICIOUS!

Motivación


I THINK WE SHOULD GIVE IT ANOTHER SHOT.
WE SHOULD BREAK UP, AND I CAN PROVE IT.
OUR RELATIONSHIP
HUH.
MAYBE YOU'RE RIGHT.
I KNEW DATA WOULD CONVINCE YOU.
NO, I JUST THINK I CAN DO BETTER THAN SOMEONE WHO DOESN'T LABEL HER AXES.

1. Estadística Descriptiva e Inferencial

El análisis estadístico es un conjunto de procedimientos diseñados para **resumir**, **organizar** los datos con el objetivo de extraer **información** y elaborar **conclusiones**.

Todas las ciencias empíricas hacen uso de la estadística para generar nuevo conocimiento, y en nuestra vida cotidiana continuamente hacemos uso de sus razonamiento

1. Estadística Descriptiva e Inferencial

La **estadística** moderna surge de la unión de dos disciplinas independientes:

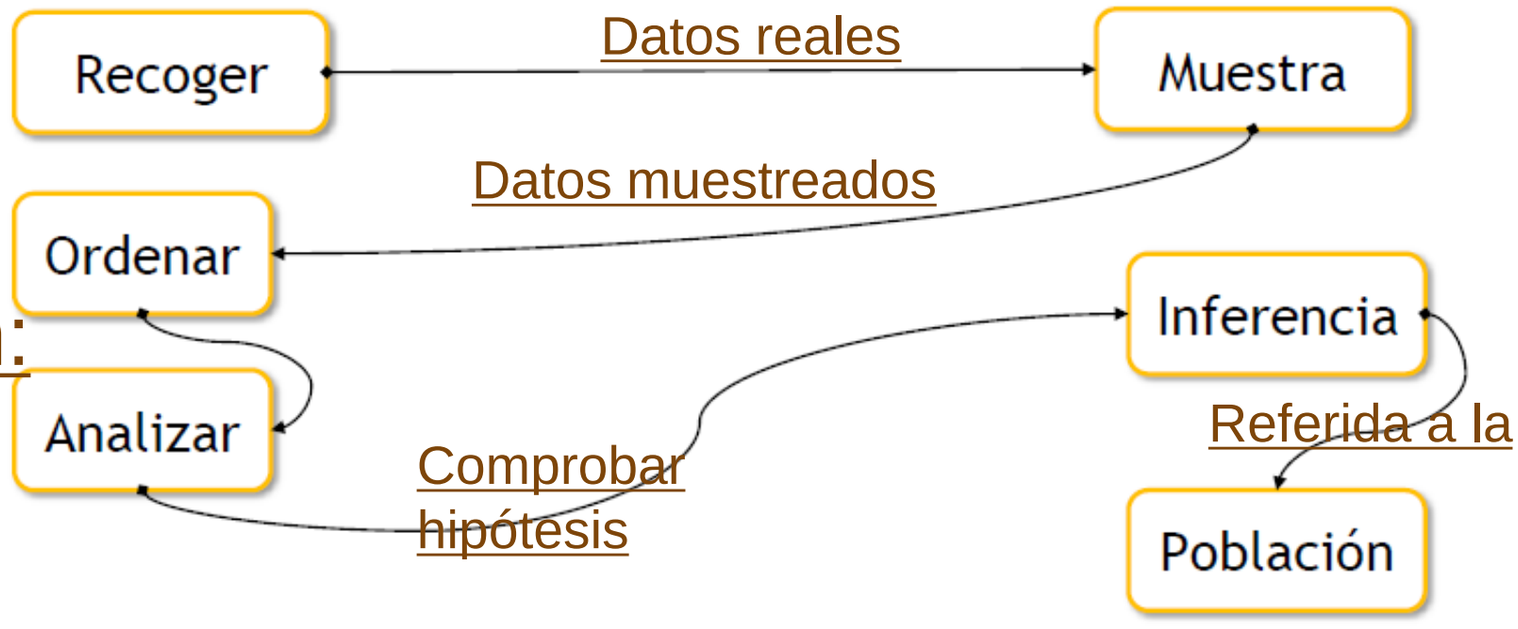
1.- El **cálculo de probabilidades** que apareció como una aproximación matemáticas a los juegos de azar

2.- La estadística o **ciencia de estado** dedicada a llevar los registros ordenados (contar, tabular, clasificar, censar, etc) de los datos del Estado.

1. Estadística Descriptiva e Inferencial

A finales del siglo XIX surge la nueva ciencia que busca extraer conclusiones de la investigación empírica mediante el uso de modelos matemáticos.

Flujo de la información:



1.1. Clasificación

Estadística descriptiva: procedimientos diseñados para describir la información contendida en un conjunto de datos. **Resume y organiza datos.**

Estadística inferencial: procedimientos que permiten generalizar (inferir, deducir) la información contenida en ese conjunto particular de datos (muestra) al conjunto de datos (población) a los que representan. **Extrae información y elabora conclusiones.**

1.1 Clasificación

Cálculo de probabilidades: es una parte de la estadística pero debe entenderse más como una herramienta tanto para hacer estadística descriptiva como inferencial.

Se habla también de **estadística teórica** (teoriza, propone y desarrolla métodos de análisis) y **estadística aplicada** (aplica a datos reales los métodos propuestos por la estadística teórica) = análisis de datos.

1.2 Utilidad del análisis de datos

Las ciencias pueden clasificarse en dos tipos:

Formales o deterministas: se parte de premisas verdaderas y por un proceso de **deducción lógica** se llega a conclusiones verdaderas.

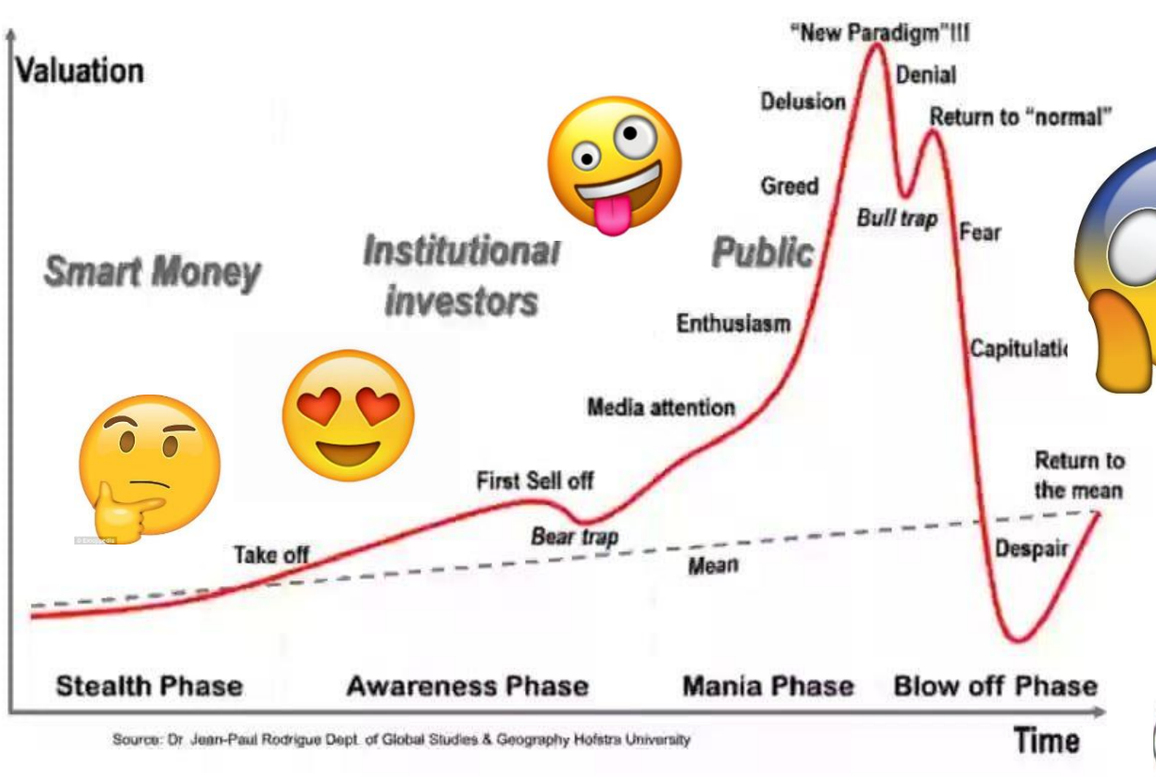
Matemáticas, lógica y ciencias de la computación.

Empíricas: se intenta ir desde lo que se observa en un *conjunto reducido* de observaciones hasta la afirmación de que eso es válido para el *total.* Por un proceso **inductivo**.

Comprenden las ciencias sociales y las ciencias naturales

Fenómenos de la realidad

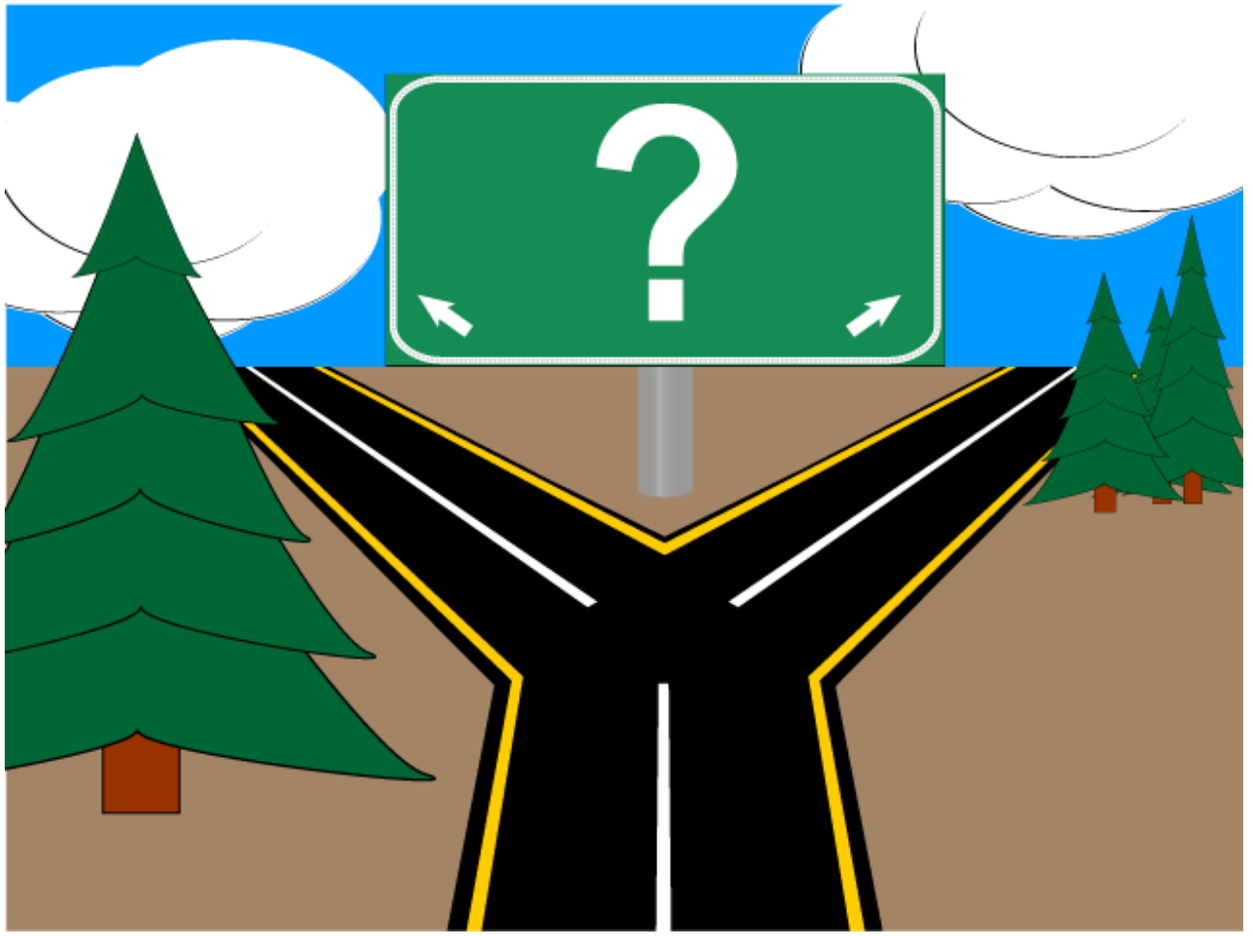
RIESGO. Posibilidad de que suceda o no suceda un contratiempo





Fenómenos de la realidad

INCERTIDUMBRE: Falta de conocimiento seguro y claro de algo



Fenómenos de la realidad

VARIABILIDAD: Ciertas características de seres y procesos presentan variaciones no deterministas





Fenómenos de la realidad

AZAR: Situaciones, ocurrencias y casualidades que no se pueden prever de manera determinista



Fenómenos de la realidad

SESGO: En general, alude a una distorsión, bien en la selección de sujetos para un estudio (sesgo muestral) bien en el procesamiento mental de datos (sesgo cognitivo)



1.3 Niveles de estudio

I.- Nivel descriptivo: se intenta obtener conocimiento sobre algo desconocido, identificar problemas de investigación y generar ideas futuras. ¿Cómo son las cosas?

II. Nivel relacional: permite descubrir qué hechos se encuentran relacionados y en qué grado y además poder predecir unos hechos a partir de otros. ¿Cómo se relacionan las cosas?

III. Nivel explicativo: permite establecer relaciones de tipo causal entre los eventos de manera que lo que ocurre con uno o varios de ellos puede ser explicado recurriendo a otro o varios diferentes. ¿Por qué las cosas son como son?

1.4 Fases de un estudio estadístico

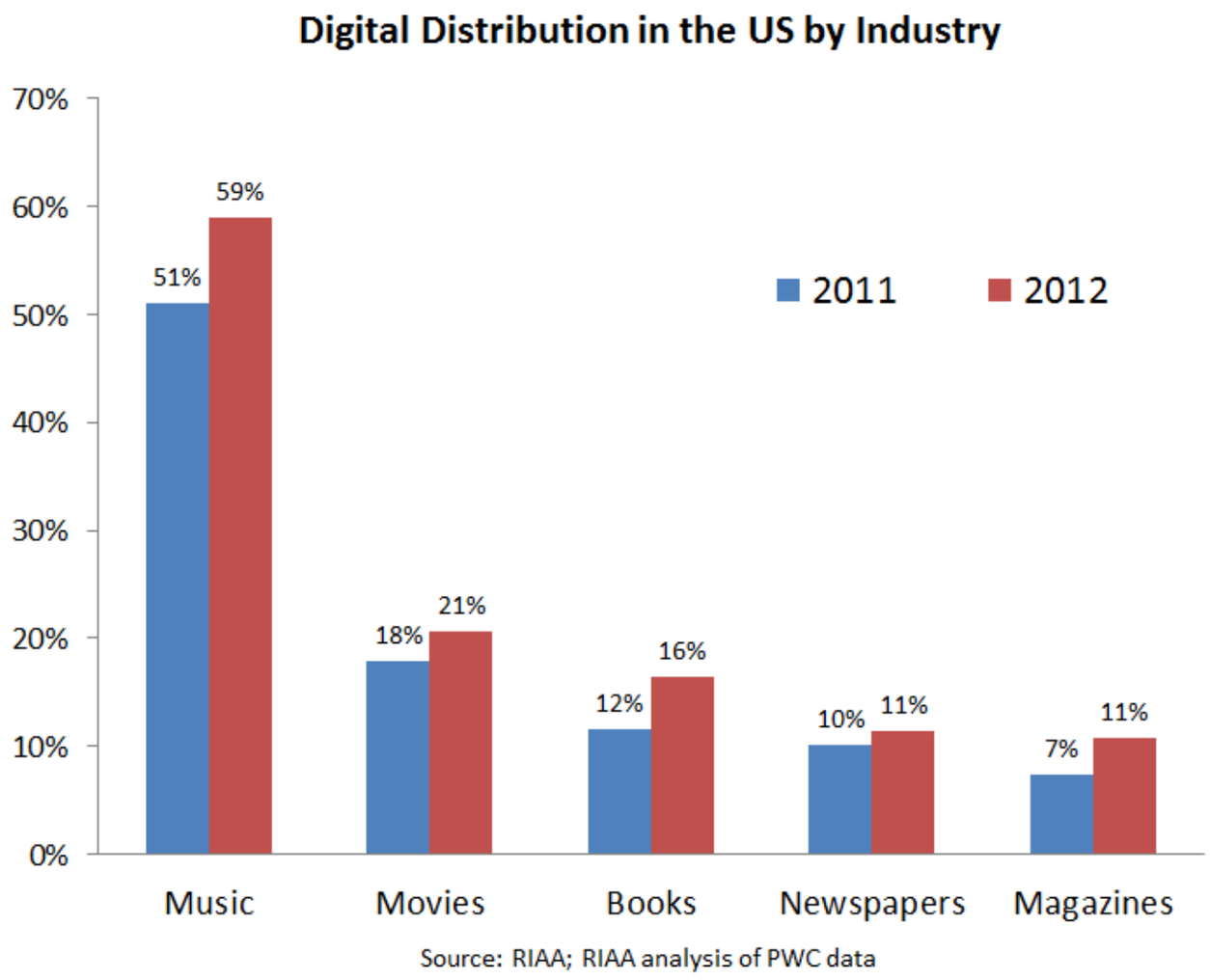
1. **Planteamiento de hipótesis**
   Enunciado del resultado que se quiere probar
2. **Diseño del estudio**
   Diseño del resto del proceso (variables a considerar, procedimiento de muestreo, proceso de análisis e inferencia, etc.)
3. **Muestreo**
   Extracción y registro de datos reales (muestra)
4. **Descripción y resumen de los datos**
   Análisis descriptivo de los datos
5. **Inferencia sobre población. Confianza**
   Análisis inferencial para validación de la hipótesis de partida

2. Conceptos Básicos

Datos. Elementos de información que se recopilan durante un estudio. Están directamente relacionados con el tipo de variable.

Conjunto de datos. Totalidad el conjunto de datos de un estudio.

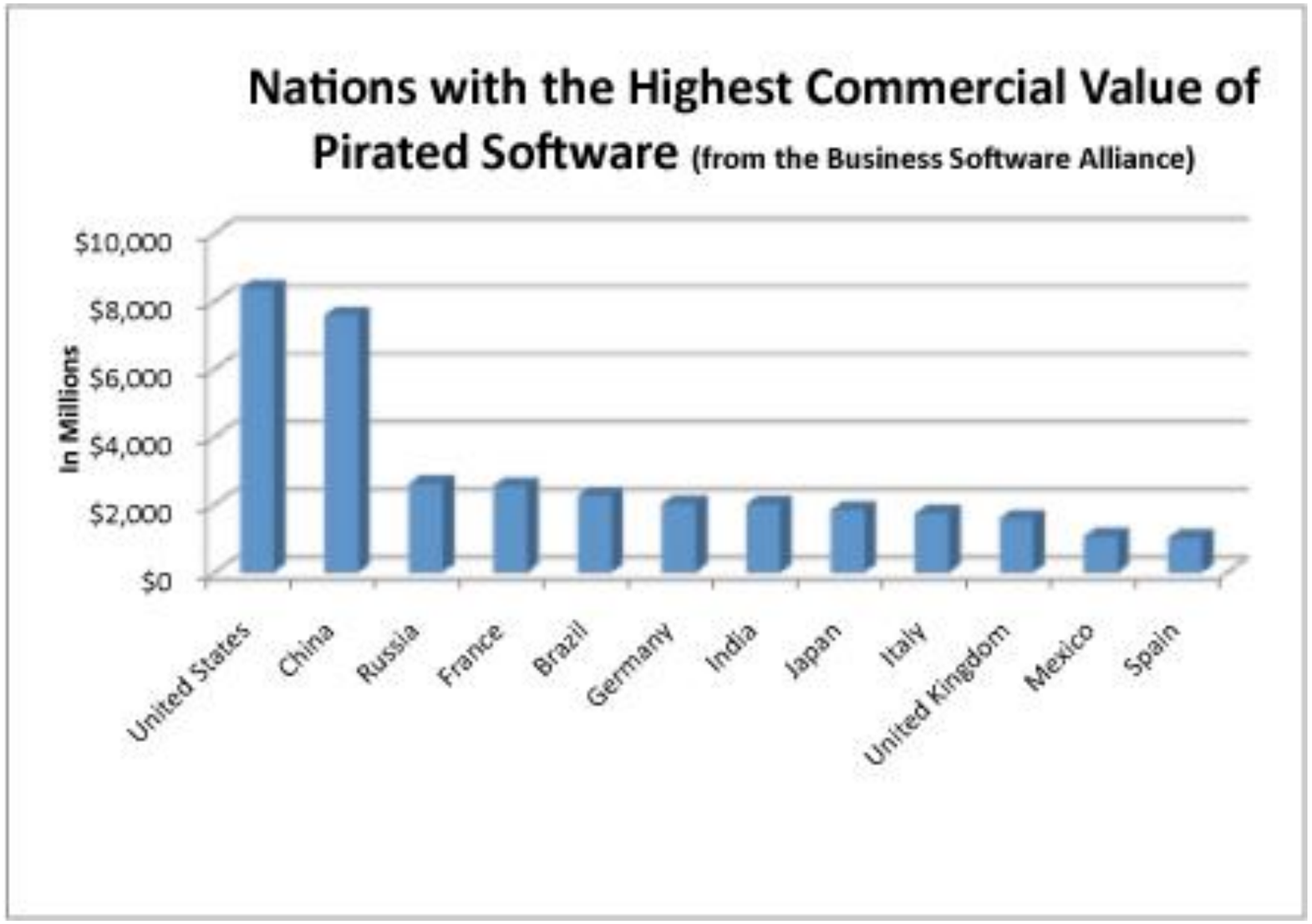
Variable. Es una característica o un valor numérico que puede ser diferente de un individuo a otro (o de una unidad de análisis a otra).


Digital Distribution in the US by Industry
70%
60%
50%
40%
30%
20%
10%
0%
2011
2012
51%
59%
18%
21%
12%
16%
10%
11%
7%
11%
Music
Movies
Books
Newspapers
Magazines
Source: RIAA; RIAA analysis of PWC data

Compamy Name / Title 25 items ☑ Auto-Fit Colums with Data Form

Form | Data Check | Sort | Filters | Fixed Width Columns | Hide Rows & Columns | Display All Fields | Help | Purchase / Register

Name	Member ID	Joined	Membership	Status	Age	Gender	Marital Status	Street	City	State
Arnold Basner	C10010	2001	Gold	Active	28	Male	Married	1070 Massive Ave	Van Nuys	California
Bill Millerton	C10020	2007	Silver	Paused	28	Male	Single	1212 Congo Road	Valencia	California
Bob Franklin	C10030	2004	Bronze	Active	50	Male	Divorced	12112 Duper Road	Valencia	California
Carrie Dater	C10220	2004	Silver	Active	41	Female	Divorced	4447 Willow Drive	Orange County	California
Cathy Deleronsco	C10250	2006	Bronze	Paused	39	Female	Single	2547 Hill Street	Northridge	California
Diane Jenkins	C10180	2003	Gold	Active	38	Female	Married	3247 Zipper Street	Orange County	California
Donna Friendly	C10090	2002	Silver	Active	36	Female	Divorced	1456 Carton Street	Northridge	California
Eric Williams	C10140	2001	Gold	Paused	37	Male	Single	21607 Blender Ave	Tarzana	California
Jill Macy	C10240	2001	Silver	Active	46	Female	Married	666 Bendover Drive	Los Angeles	California
Jim Holloween	C10060	2005	Silver	Active	36	Male	Divorced	13447 Rader Lane	Westlake	California
Karen Hotchick	C10070	2005	Silver	Active	44	Female	Married	13447 Small Lane	Westlake	California
Karen Jacobsen	C10080	2008	Bronze	Active	31	Female	Divorced	1447 Willow Lane Drive	Orange County	California
Kathy Cruster	C10050	2006	Gold	Active	32	Female	Single	13334 Mills Station Road	Los Angeles	California
Larry Write	C10190	2008	Silver	Active	49	Male	Married	3264 Pappy Street	Northridge	California
Lisa Thompsom	C10200	2001	Bronze	Active	37	Female	Divorced	333 Happy Street	Northridge	California
Lisa Wantssome	C10040	2001	Silver	Active	36	Female	Married	1321 Stone Creekside Road	Los Angeles	California
Mary Hadalamb	C10100	2004	Gold	Active	28	Female	Single	1607 Blender Ave	Tarzana	California
Mike Stoner	C10230	2002	Gold	Active	43	Male	Single	6321 Bend Drive	Los Angeles	California
Monica Candoit	C10110	2006	Silver	Active	25	Female	Married	166 Over Drive	Los Angeles	California
Paul Jones	C10210	2007	Bronze	Paused	43	Male	Married	33447 Toaste Lane	Westlake	California
Paula Mitchel	C10120	2001	Bronze	Paused	26	Female	Divorced	1687 Stone Road	Los Angeles	California
Peter Notool	C10130	2001	Silver	Active	48	Male	Divorced	21000 Bloated Ave	Van Nuys	California
Scott Miller	C10170	2002	Silver	Paused	42	Male	Divorced	3212 Smile Road	Valencia	California
Todd Willinters	C10150	2004	Silver	Active	25	Male	Married	2727 Pico Contante Drive	Los Angeles	California
Tom Hasserian	C10160	2004	Bronze	Active	43	Male	Single	27527 Juicer Drive	Los Angeles	California


Nations with the Highest Commercial Value of Pirated Software (from the Business Software Alliance)
In Millions
$10,000
$8,000
$6,000
$4,000
$2,000
$0
United States
China
Russia
France
Brazil
Germany
India
Japan
Italy
United Kingdom
Mexico
Spain

2. Conceptos Básicos

Población. Conjunto concreto de unidades de análisis que interesa estudiar. Es muy importante que quede bien definida en el diseño de estudio.

Muestra. Cada uno de los posibles subconjuntos de la población. Nos interesa que sea **representativa y aleatoria**.

2. Conceptos Básicos

Estadístico. Es un número que resume los datos recopilados en una **muestra.** Un porcentaje, un promedio, la mediana, la desviación típica, etc son estadísticos si se calculan sobre los datos de la **muestra.** Se suelen representar con *caracteres latinos.*

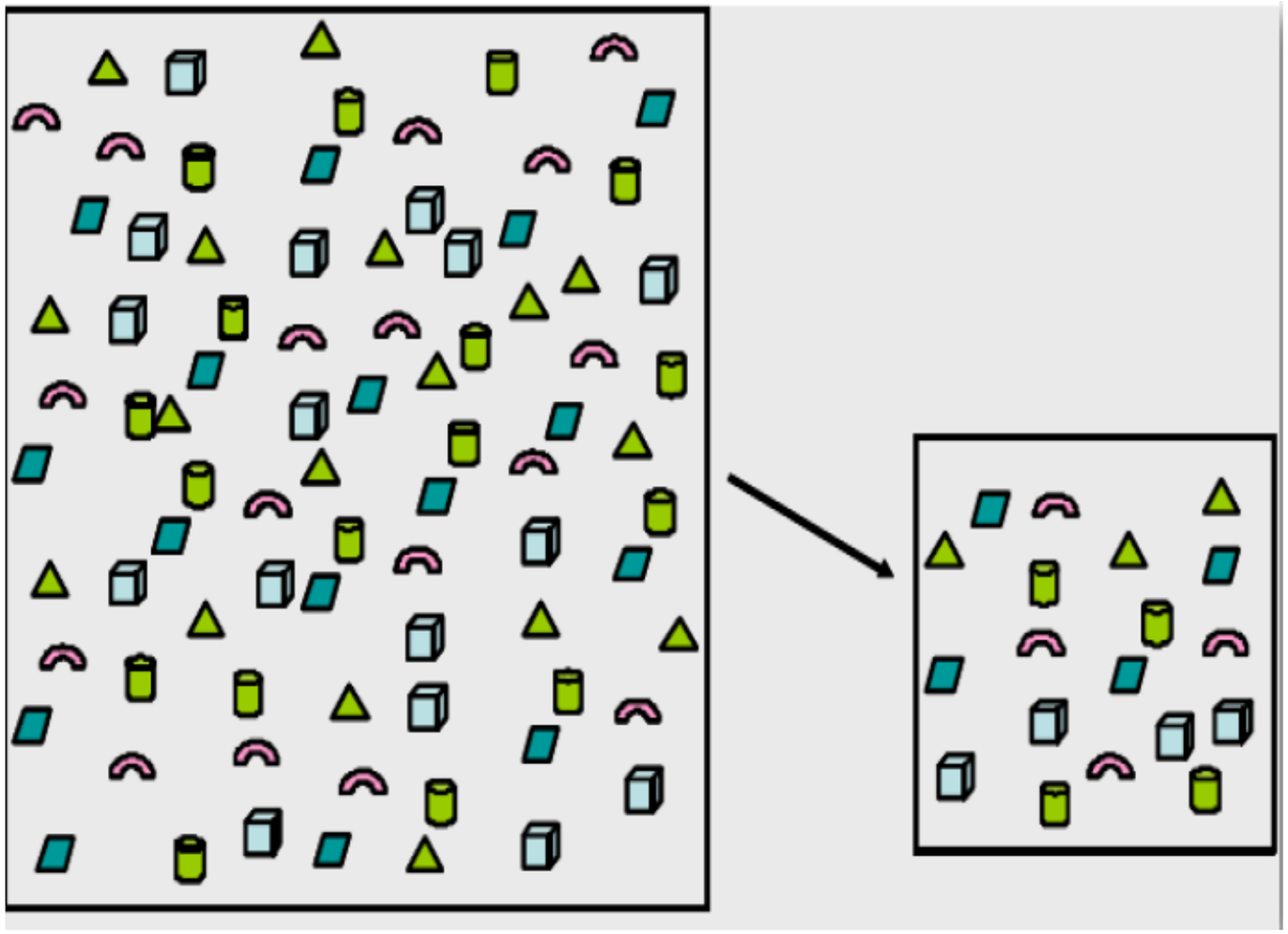
Parámetro. Es el equivalente a cada estadístico pero en la **población.** En general son desconocidos y se *estiman* a partir del correspondiente estadístico muestral. Se representan con *letras griegas.*

2.1 Tipos de muestreo

Algunos ejemplos:

- Simple.
- Sistemático.
- Estratificado.
- ……..

Tipos de muestreo - simple



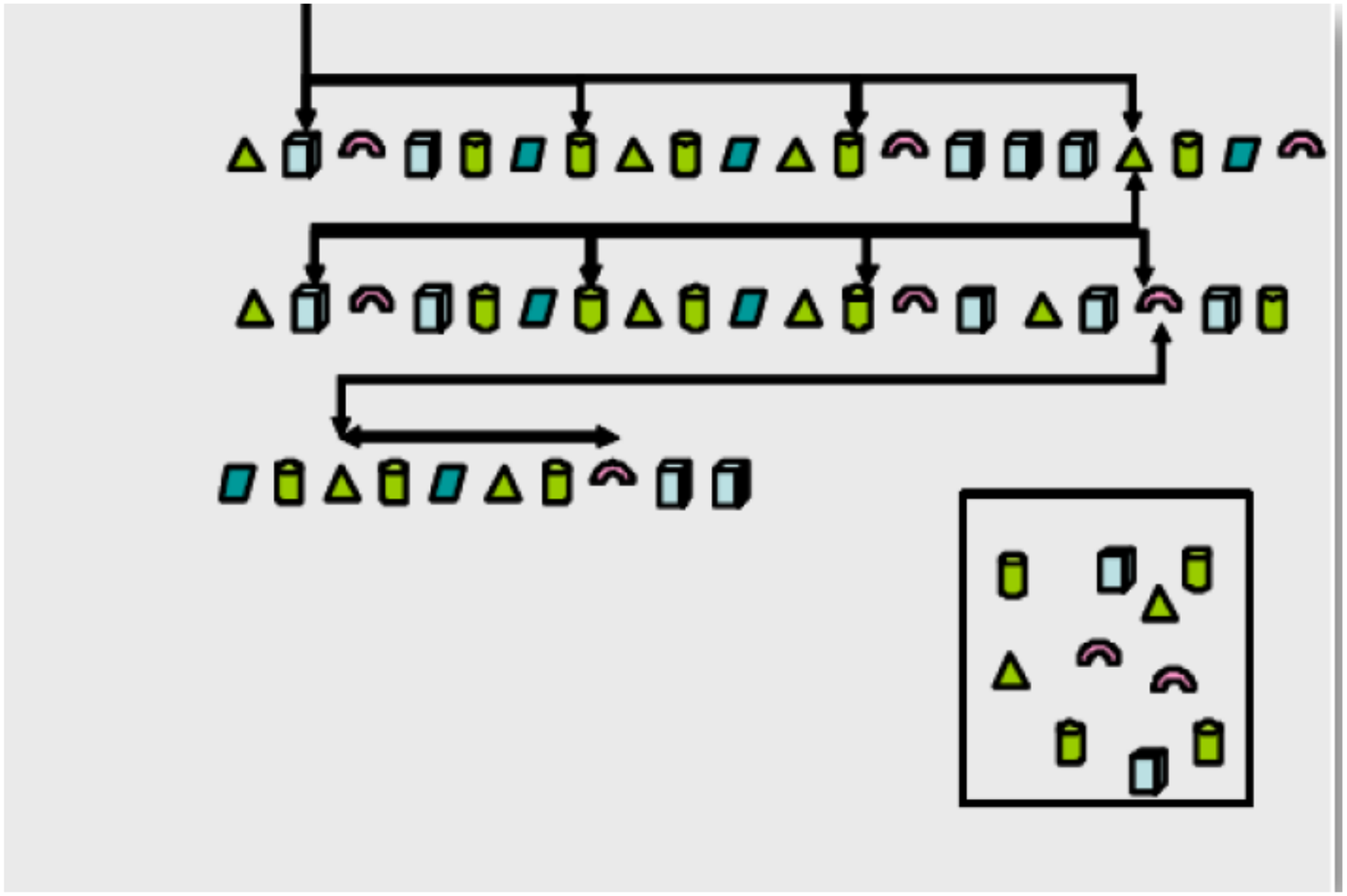
Tipos de muestreo - simple

Selecciona muestras mediante métodos que aseguran que:

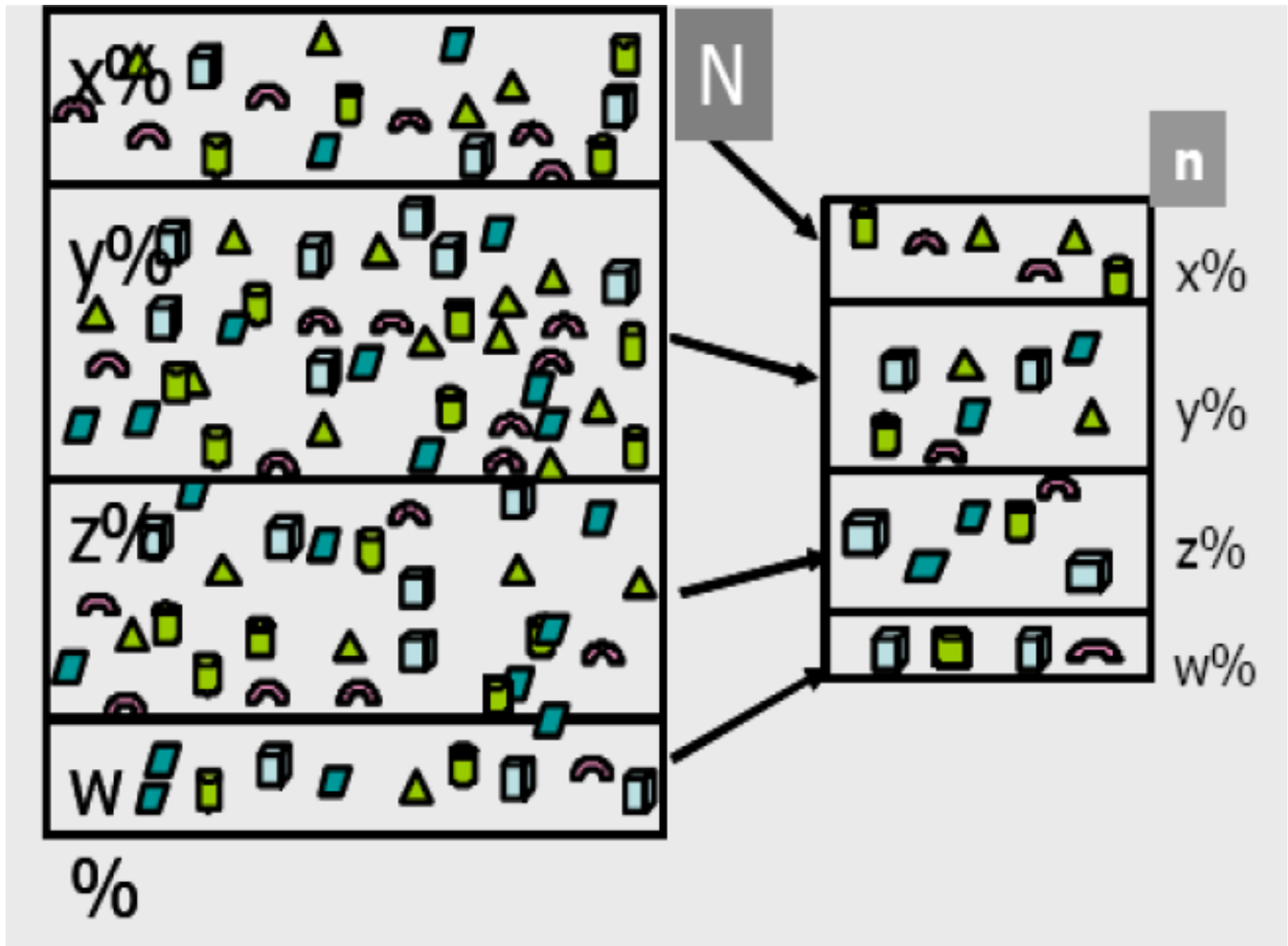
- Cada posible muestra tiene una igual probabilidad de ser seleccionada y
- Cada elemento de la población total tiene una oportunidad igual de ser incluido en la muestra.

Es el más usado en la práctica

Tipos de muestreo - sistemático



Tipos de muestreo - estratificado



3. Variable

Una variable es una **característica** o propiedad de cada unidad de análisis de un estudio.

Es importante conocer en la **fase de diseño** de una investigación cómo se va a recoger la información y definir todas y cada una de las variables del estudio.

3.1 Tipos de variable

En los estudios en los que se evalúa **relación causal:**

Variable independiente

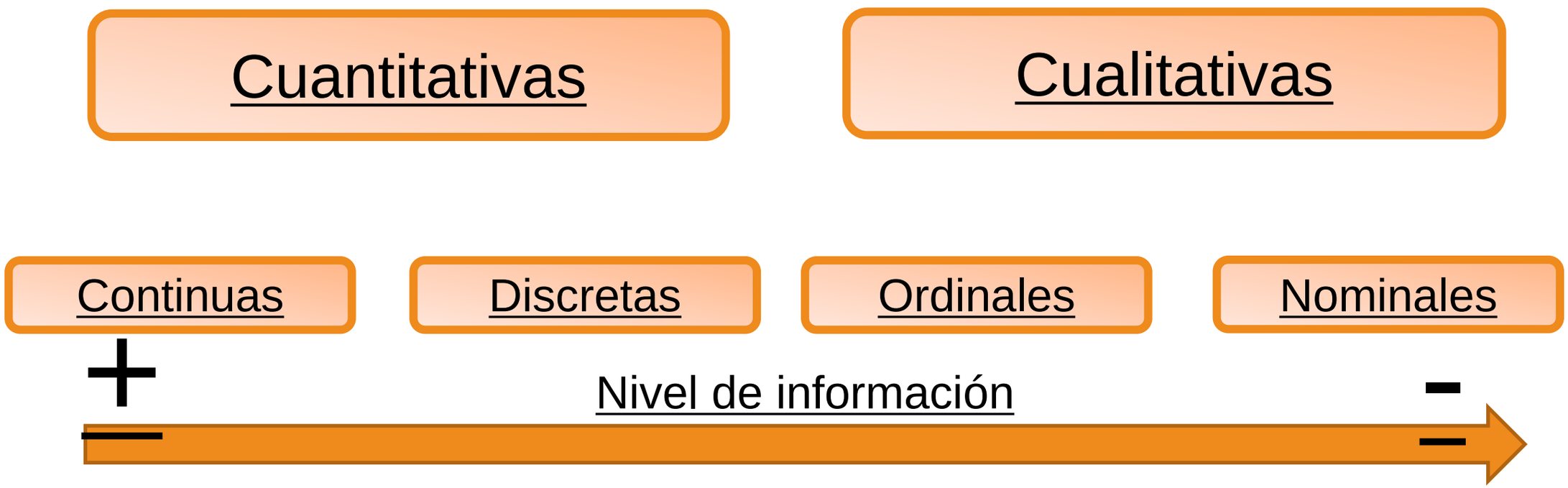
Aquella característica que suponemos **causa** u **origen** del fenómeno estudiado

Variable dependiente

Aquella característica que suponemos **consecuencia** o **efecto** del fenómeno estudiado

3.1 Tipos de variable

Si nos fijamos en el tipo de datos:



3.1 Tipos de variable

Una variable **cuantitativa continua** es una variable de tipo numérico que puede, en teoría, tomar cualquier valor comprendido entre dos números.

El hecho de que registremos la variable con más o menos decimales (o sin decimales) no depende de la naturaleza de la variable, depende de la **precisión** con la que queramos registrar la información.

P. ej.: Peso

3.1 Tipos de variable

Una variable **cuantitativa discreta** es una variable de tipo numérico sólo puede tomar números enteros porque no tiene sentido que tomara decimales.

P.ej.: El número de productos vendidos. Sólo puede tomar valores enteros. Nadie va al fisioterapeuta 2,38 veces.

3.1 Tipos de variable

Una variable **cualitativa ordinal** es una variable no mensurable, no cuantificable con varias categorías en las que cada sujeto puede clasificarse.

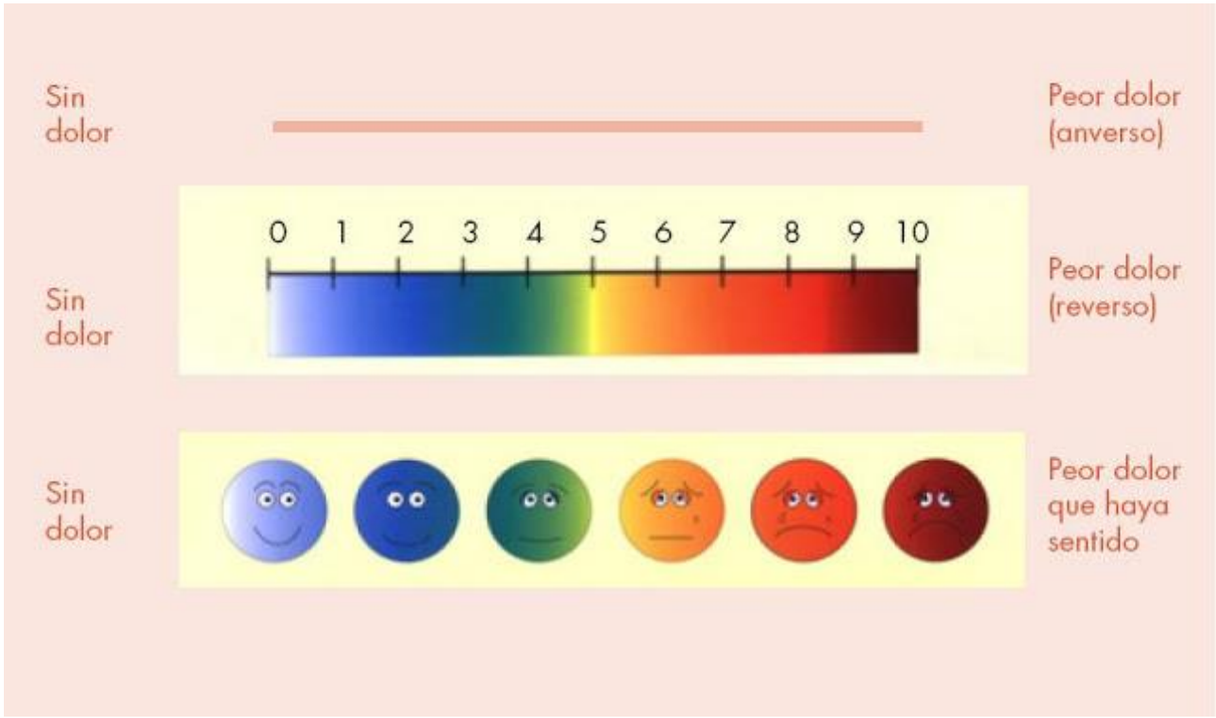
Las categorías pueden **ordenarse** según algún criterio lógico.

P.ej.: grande, mediano, pequeño.

3.1 Tipos de variable

En ocasiones se las denomina ***cuasi-cuantitativas,*** en especial a aquellas que se han obtenido a través de puntuaciones en diferentes escalas: EVA, cuestionarios de calidad de vida, de disfuncionalidad, de depresión, etc.

Desde el punto de vista operativo se suelen tratar como las cuantitativas. **Cuidado** con las interpretaciones posteriores.



3.1 Tipos de variable

Una variable **cualitativa nominal** es una variable no mensurable, no cuantificable con varias categorías en las que cada sujeto puede clasificarse.

Las categorías **no pueden ordenarse** según algún criterio lógico.

Tipos de variable

Ejemplo 1.3.

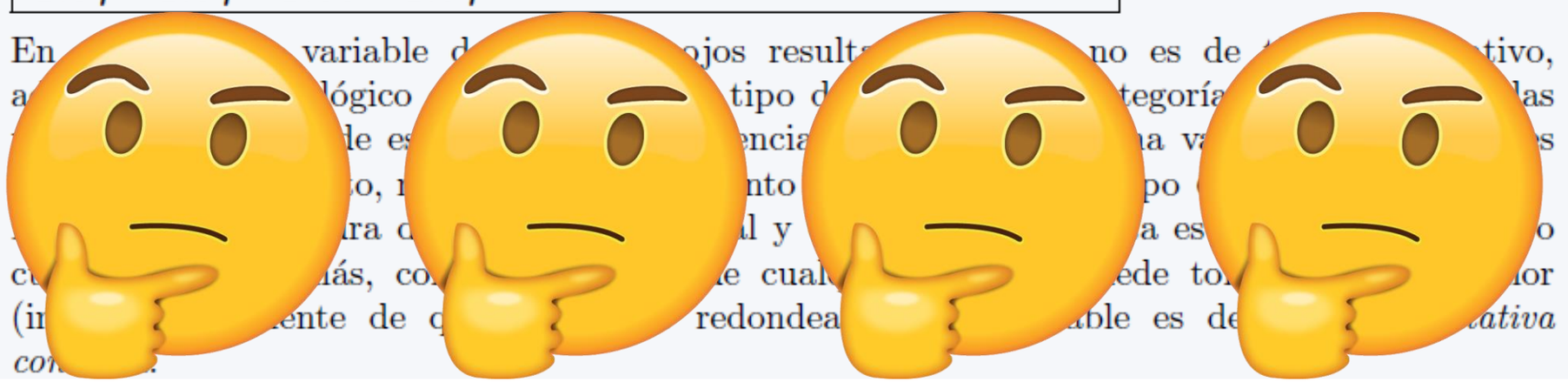
Clasifica las siguientes variables según los criterios anteriores:
Color de los ojos de una muestra de alumnos

Marrón	*Verde*	*Verde*	*Azul*	*Marrón*	*Azul*
Marrón	*Marrón*	*Azul*	*Marrón*	*Azul*	*Marrón*
Verde	*Marrón*	*Verde*	*Marrón*	*Azul*	*Verde*
Marrón	*Marrón*	*Azul*	*Marrón*	*Verde*	*Marrón*
Verde	*Verde*	*Marrón*	*Marrón*	*Marrón*	*Marrón*
Azul					

Estatura de alumnos de la muestra anterior

1.48	*1.56*	*1.56*	*1.59*	*1.60*	*1.61*	*1.63*	*1.64*	*1.64*
1.67	*1.68*	*1.68*	*1.68*	*1.68*	*1.69*	*1.70*	*1.71*	*1.72*
1.72	*1.75*	*1.76*	*1.76*	*1.77*	*1.77*	*1.79*	*1.81*	*1.81*
1.84	*1.84*	*1.88*	*1.94*					

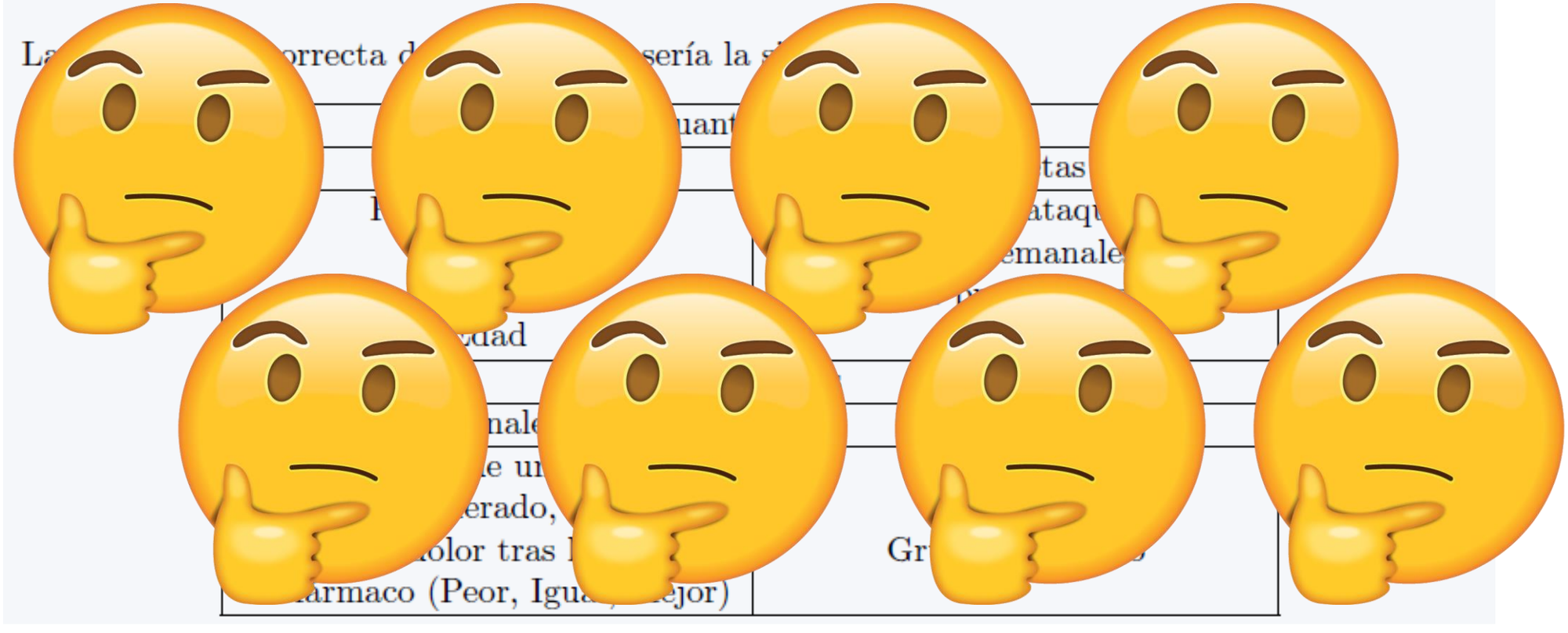
En variable d ojos result no es de tivo, a lógico tipo d tegoría las de es encia a va s , r nto po ra d l y a es o c ás, co e cual ede to ior (i ente de q redondea able es de *cativa co* .



Tipos de variable

Ejemplo 1.4.

Clasifica las siguientes variables según el criterio que acabamos de introducir: Gravedad de un infarto (leve, moderado, fuerte), Número de ataques de asma semanales, Sexo, Presión arterial, Estatura, Peso, Estado de dolor tras la toma de un fármaco (Peor, Igual, Mejor), Provincia, Edad, Número de preguntas acertadas en un test, Grupo sanguíneo.



3.2 Intervalos de clase

En ocasiones una variable que se ha registrado como cuantitativa, interesa **transformarla** en una variable cualitativa (en general ordinal).

En el proceso se pierde información pero se gana en claridad.

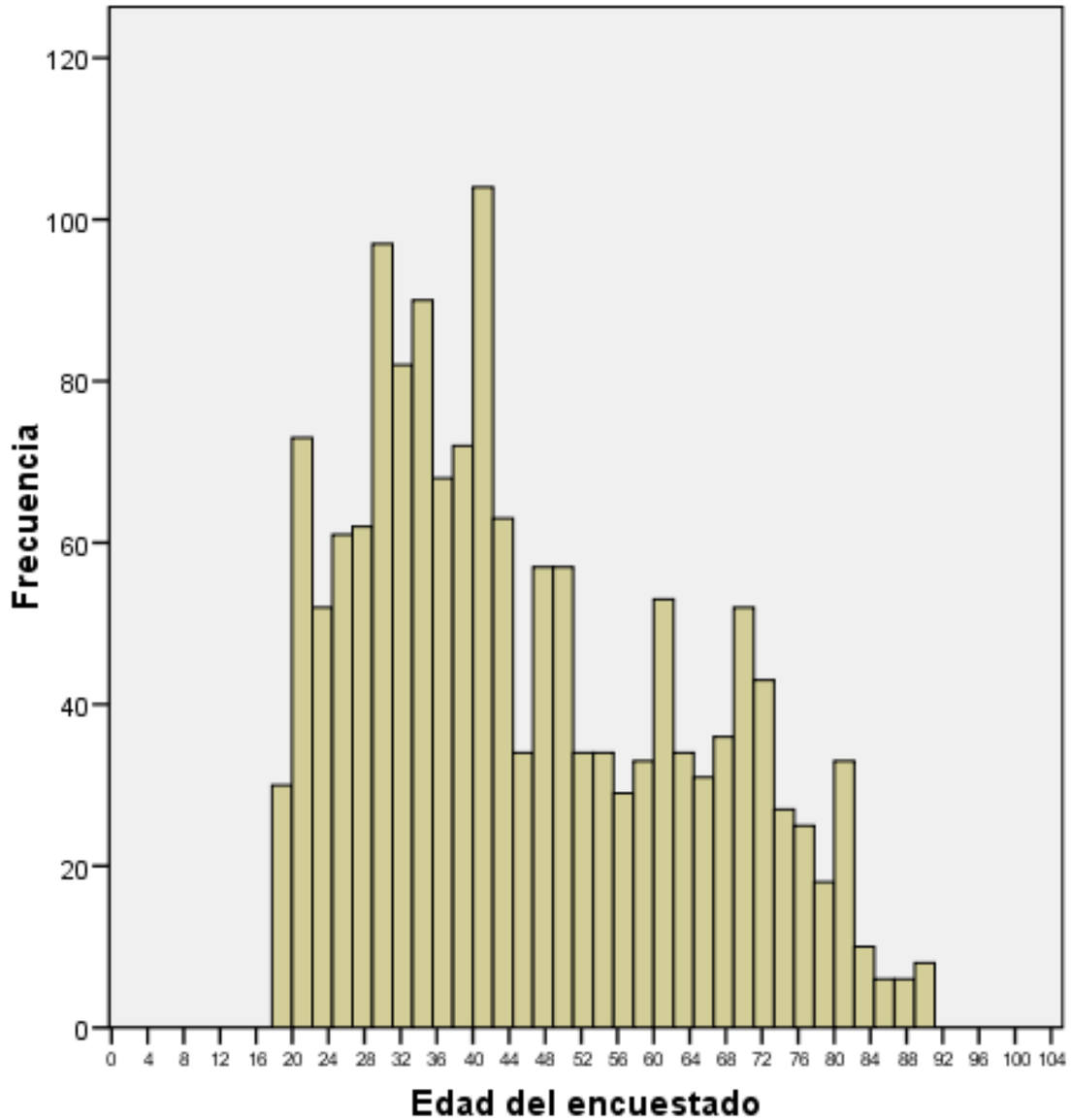
3.2 Intervalos de clase

Se establecen puntos de corte que permiten clasificar a **todos** los sujetos en **alguna** de las nuevas categorías.

Esas categorías reciben el nombre de **intervalos de clase.**

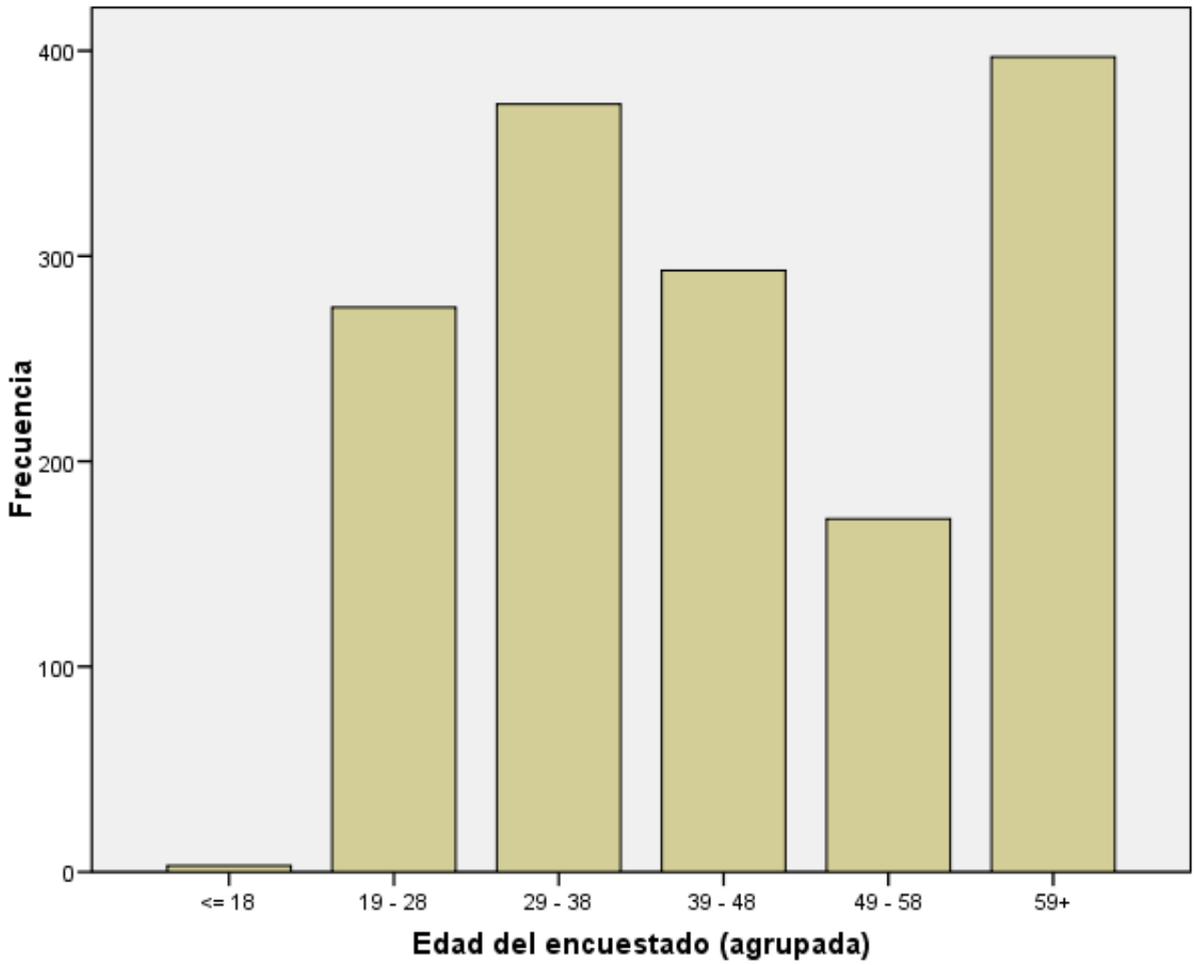
3.2 Intervalos de clase

Tenemos la variable **“edad”** que es cuantitativa continua y la queremos transformar en otra variable que llamaremos **“edad categorizada”**



3.2 Intervalos de clase

En esta nueva variable hemos perdido la información particular, pero ahora tenemos clasificados a los sujetos por ***categorías ordenadas de edad***



3.3 Etiquetado y codificación

Cuando se está preparando la base de datos es necesario definir los nombres de las variables (etiquetado) y los posibles códigos de las diferentes categorías (codificación) para poder trabajar adecuadamente con los datos

No es complicado y solamente deben tener en cuenta unas reglas muy sencillas

3.3 Etiquetado y codificación

1.- El nombre de la variable no debe contener más de 8-10 caracteres alfa-numéricos. Cada nombre debe ser inequívoco y único.

2.- Deben evitarse los espacios. En caso de que sea necesario introducir alguna separación visual, puede hacerse con “punto” o con “guión bajo”

3.- Evitar utilizar caracteres que pueden tener otro significado para los programas de análisis: @, %, &, /, (), #, etc.

3.3 Etiquetado y codificación

4.- Debe definirse también el número de decimales que se quieren visualizar. Para trabajar con matrices de datos debe intentarse que sea visualmente cómoda.

5.- Cuidado con los redondeos. **NUNCA** se debe redondear para hacer los cálculos, en todo caso se redondea al final. Los programas de análisis nunca redondean.

6.- Debe definirse cómo se van a registrar los **valores perdidos**: perdidos por el sistema o perdidos por el usuario.

3.3 Etiquetado y codificación

EJEMPLOS ETIQUETADO CORRECTO

ALTURA
GENERO
NUM_HIJOS
INGRESOS
NACIONALIDAD
PESO
NIVEL_ESTUD
EST_CIVIL
EDAD

3.3 Etiquetado y codificación

EJEMPLOS ETIQUETADO INCORRECTO

ESTATURA_DE_LA_PERSONA

MUJER O HOMBRE

NUM_HIJ@S

NIVEL ESTUD

FECHA#NACIMIENTO

3.3 Etiquetado y codificación

EJEMPLO de codificación

Veremos a continuación un ejemplo de codificación

La variable SEXO puede tomar las categorías Hombre / Mujer que se codificarán como 1 / 2

La variable RAZA puede tomar las categorías Blanca / Negra / Otra que se codificarán como 1 / 2 / 3

La variable FELIZ puede tomar las categorías Muy feliz / Bastante feliz / No demasiado feliz / No contesta que se codificarán como 1 / 2 / 3 / 9 (9 es el código para valores perdidos)

La variable VIDA puede tomar las categorías Excelente / Rutinaria / No procede que se codificarán como 1 / 2 / 0 (0 código para valores perdidos)

3.3 Etiquetado y codificación

	sexo	raza	región	feliz	vida	hermanos	hijos	edad
1	Mujer	Blanca	Nor-Este	Muy feliz	Excitante	1	2	61
2	Mujer	Blanca	Nor-Este	Bastante feliz	Excitante	2	1	32
3	Hombre	Blanca	Nor-Este	Muy feliz	No procede	2	1	35
4	Mujer	Blanca	Nor-Este	No contesta	Rutinaria	2	0	26
5	Mujer	Negra	Nor-Este	Bastante feliz	Excitante	4	0	25
6	Hombre	Negra	Nor-Este	Bastante feliz	No procede	7	5	59
7	Hombre	Negra	Nor-Este	Muy feliz	Excitante	7	3	46
8	Mujer	Negra	Nor-Este	Bastante feliz	No procede	7	4	No contest
9	Mujer	Negra	Nor-Este	Bastante feliz	Rutinaria	7	3	57
10	Mujer	Blanca	Nor-Este	Bastante feliz	Excitante	1	2	64
11	Hombre	Blanca	Nor-Este	Bastante feliz	Excitante	6	0	72
12	Mujer	Blanca	Nor-Este	Muy feliz	No procede	2	5	67
13	Hombre	Blanca	Nor-Este	Bastante feliz	No procede	1	0	33
14	Hombre	Otra	Nor-Este	Bastante feliz	Rutinaria	2	1	23
15	Mujer	Blanca	Nor-Este	Bastante feliz	Rutinaria	7	1	33
16	Mujer	Blanca	Nor-Este	Muy feliz	Rutinaria	6	2	59
17	Hombre	Blanca	Nor-Este	Bastante feliz	No procede	4	1	60
18	Hombre	Blanca	Nor-Este	Muy feliz	Rutinaria	6	2	77
19	Mujer	Negra	Nor-Este	Bastante feliz	No procede	12	2	52
20	Hombre	Negra	Nor-Este	Muy feliz	Aburrida	5	1	55
21	Mujer	Negra	Nor-Este	Muy feliz	Rutinaria	2	1	37
22	Mujer	Otra	Nor-Este	Bastante feliz	No procede	7	0	45
23	Mujer	Negra	Nor-Este	No demasiado fel	Rutinaria	4	0	34
24	Mujer	Negra	Nor-Este	No demasiado fel	Excitante	7	3	35
25	Mujer	Negra	Nor-Este	Bastante feliz	Rutinaria	6	1	35
26	Hombre	Negra	Nor-Este	No demasiado fel	Excitante	7	0	47

	sexo	raza	región	feliz	vida	hermanos	hijos	edad
1	2	1	1	1	1	1	2	61
2	2	1	1	2	1	2	1	32
3	1	1	1	1	0	2	1	35
4	2	1	1	9	2	2	0	26
5	2	2	1	2	1	4	0	25
6	1	2	1	2	0	7	5	59
7	1	2	1	1	1	7	3	46
8	2	2	1	2	0	7	4	99
9	2	2	1	2	2	7	3	57
10	2	1	1	2	1	1	2	64
11	1	1	1	2	1	6	0	72
12	2	1	1	1	0	2	5	67
13	1	1	1	2	0	1	0	33
14	1	3	1	2	2	2	1	23
15	2	1	1	2	2	7	1	33
16	2	1	1	1	2	6	2	59
17	1	1	1	2	0	4	1	60
18	1	1	1	1	2	6	2	77
19	2	2	1	2	0	12	2	52
20	1	2	1	1	3	5	1	55
21	2	2	1	1	2	2	1	37
22	2	3	1	2	0	7	0	45
23	2	2	1	3	2	4	0	34
24	2	2	1	3	1	7	3	35
25	2	2	1	2	2	6	1	35
26	1	2	1	3	1	7	0	47

Para saber más

Pardo Merino A, Ruiz MÁ, San Martín R. **Análisis de datos en ciencias sociales y de la salud I.** Madrid: Síntesis; 2009.

Feinstein AR. **Principles of medical statistics**. FL: Chapman & Hall/CRC; 2002.

Hurley WL, Denegar CR, Hertel J, Vazquez JC. **Métodos de investigación fundamentos de una práctica clínica basada en la evidencia**. Barcelona:Wolter Kluwer; 2012.

Para saber más

https://www.genbeta.com/web/once-cursos-universitarios-gratuitos-sobre-estadistica

https://www.tareasplus.com/Curso-Probabilidad-y-Estadistica/Marcela-Gomez

https://www.tareasplus.com/Estadistica-Empresarial/Neumann-BS

https://miriadax.net/web/curso-practico-bioestadistica-r-3edicion

Estadística Aplicada al Marketing

T.2. Análisis Descriptivo Unidimensional

Juan Miguel Navarro Ruiz

Grado en Marketing y Dirección Comercial

Motivación

THIS CHAPTER CONCENTRATES ON DATA *DESCRIPTION*. HOW CAN WE REPRESENT DATA IN USEFUL WAYS? HOW CAN WE SEE UNDERLYING PATTERNS IN A HEAP OF NAKED NUMBERS? HOW CAN WE SUMMARIZE THE DATA'S BASIC SHAPE?



Recordatorio: Variables

Es algo que nos interesa conocer de un grupo de individuos. La indicaremos por X

- Edad
- Sexo
- Municipio de nacimiento
- Enseñanza de procedencia
- Municipio de residencia
- Profesión de los padres
- Orientación profesional

Pueden ser cuantificables o no

- No cuantificable → (Atributo) Variable cualitativa
- Cuantificable → Variable cuantitativa.

Recordatorio: Variables

Variable cualitativa (atributo). Las distintas opciones que puede tomar se denominan **modalidades**

Variable cuantitativa. Las distintas opciones que puede tomar se denominan **valores**

X

$x_1, \ldots, x_r$

r = nº de modalidades o valores distintos que puede tomar la variable

Recordatorio: Variables

X = sexo

x_1 = Varón, x_2 = Hembra r=2

X = nº de asignaturas matriculado

x_1=1, …. , x_{12}=12 r=12

X= edad

¿CON QUÉ MEDIDA?

Recordatorio: Variables

X= edad	¿CON QUÉ MEDIDA?	
Años cumplidos	r = finito y pequeño	Variable discreta
Meses cumplidos	r = finito y mediano	
Días cumplidos	r = finito y grande	Variable continua
Horas cumplidas	r = finito y muygrande	

1. Análisis de datos

- ¿Para qué **organizamos** los datos y los representamos **gráficamente?**

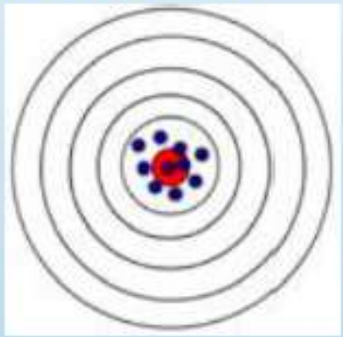
Tener visión de conjunto

Facilitar cálculos posteriores y decisiones a tomar

Comunicar y presentar resultados

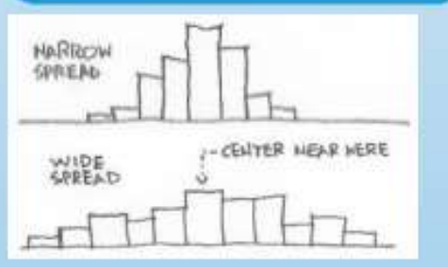
1. Análisis de datos

- En general, tanto si una variable es cualitativa como si es cuantitativa se deberá prestar atención tres características



El centro

Valor hacia donde tienden a agruparse las medidas



Grado de variabilidad de las medidas respecto a un valor concreto

La dispersión

La forma de distribución de los datos

Indicación de la simetría o asimetría y de la tendencia de la distribución

2. Tablas de frecuencias

Dada una v.e. cada uno de sus valores puede aparecer un determinado número de veces.

Al número de veces que aparece el valor x_i se le llama frecuencia absoluta y se indicará por n_i.

Distribución de frecuencias $\{(x_i, n_i)\}$, $i=1,\ldots,r$,
El conjunto formado por todos los valores que puede tomar la variable estadística junto con el número de veces (frecuencia), que aparece cada uno de ellos.

2. Tablas de frecuencias

- Una **tabla** o **distribución de frecuencias** es una forma particular de ordenar los datos basada en los valores concretos que adopta una **variable categórica** y el número de veces que se repite cada valor
- El objetivo principal es organizar la información y **resumirla**.

¿Qué te ha parecido esta clase?					
Categorías	**n**	**P**	**%**	**FAA**	**FRA**
"Me ha parecido muy interesante"	17	0,23	22,97	17	0,23
"Me ha parecido bastante interesante"	22	0,30	29,73	39	0,53
"Me ha parecido algo interesante"	29	0,39	39,19	68	0,92
"No me ha parecido interesante"	6	0,08	8,11	74	1,00
	74	1	100		

2. Tablas de frecuencias

Concretando:

x_i = valor (o cualidad) de la característica observada. Se ordenan de menor a mayor, i=1,…,r

n_i = número de veces que aparece x_i

N = número total de observaciones realizadas (individuos) = $\sum n_i$, i=1,…,r

f_i = frecuencia relativa de aparición de x_i; no es más que la proporción de veces que aparece x_i y por tanto es n_i/N.

p_i = % de veces que aparece x_i; no es más que la f_ix100

2. Tablas de frecuencias

Varon	Hembra	Hembra
Hembra	Varon	Hembra
Hembra	Hembra	Varon
Hembra	Varon	Hembra
Hembra	Varon	Hembra
Varon	Hembra	Varon
Hembra	Hembra	Hembra
Hembra	Varon	Hembra

Característica: SEXO

2. Tablas de frecuencias

Tipo de variable: cualitativa (atributo)

Valores (modalidades)

- X_1 = Varón ; n_1= 8
- X_2 = Hembra; n_2= 16

Distribución de frecuencias: {(Varón,8); (Hembra,16)}

2. Tablas de frecuencias

Tabla de frecuencias

x_i	n_i	f_i	p_i
Varón	8	0,33	33,3
Hembra	16	0,67	66,7
Total	24	1	100

2. Tablas de frecuencias

1	4	5
2	3	4
5	3	2
2	6	2
3	2	4
2	2	5
1	4	3
5	2	4

Característica: Nº de miembros de la unidad familiar

2. Tablas de frecuencias

Tipo de variable: cuantitativa discreta

Valores

1,2,3,4,5,6,

Distribución de frecuencias: {(1,2);…, (6,1)}

x_i	n_i
1	2
2	8
3	4
4	5
5	4
6	1

2. Tablas de frecuencias

x_i	n_i	f_i	p_i
1	2	0,083	8,3
2	8	0,333	33,3
3	4	0,167	16,7
4	5	0,208	20,8
5	4	0,167	16,7
6	1	0,042	4,2
	24	1	100

Al ser una v.e. cuantitativa cabe hablar de cuántos valores y también de…

2. Tablas de frecuencias

- N_i = nº de observaciones ≤ x_i
- F_i = proporción de observaciones ≤ x_i
- P_i = % de observaciones ≤ x_i

Esto nos lleva a los conceptos de frecuencias acumuladas

Que se denominan, respectivamente, frecuencia absoluta acumulada, frecuencia relativa acumulada y porcentaje acumulado

2. Tablas de frecuencias

x_i	n_i	f_i	p_i	N_i	F_i	P
1	2	0,083	8,3	2	0,083	8,3
2	8	0,333	33,3	10	0,417	41,7
3	4	0,167	16,7	14	0,583	58,3
4	5	0,208	20,8	19	0,792	79,2
5	4	0,167	16,7	23	0,958	95,8
6	1	0,042	4,2	24	1	100
	24	1	100			

2. Tablas de frecuencias

Su cálculo es inmediato y corresponde a:

$$N_i = \sum_{1}^{i} n_j = n_1 + n_2 + \ldots + n_i$$

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

$$P_i = \sum_{1}^{r} p_j = p_1 + p_2 + \ldots + p_i$$

2. Tablas de frecuencias

NIVEL DE ESTUDIOS						
	Fr. absolutas (n_i)	Fr. relativas (P_i)	Fr. porcentuales (%)	Fr. Abs. acumuladas (na_i)	Fr. Rel. acumuladas (Pa_i)	Fr. Porc acumuladas (%)
Sin estudios	5	0,01	1,0	5	0,01	1,0
Primarios	20	0,04	4,0	25	0,05	5,0
Secundarios	270	0,54	54,0	295	0,59	59,0
Medios	115	0,23	23,0	410	0,82	82,0
Superiores	90	0,18	18,0	500	1,00	100,0
Total	500	1,00	100,0			

2. Tablas de frecuencias

- La tabla de frecuencias ofrece información sobre 3 aspectos importantes

¿Qué valores toma la variable?

¿Qué valores son los más frecuentes?

En las ordinales: ¿Cuántos sujetos alcanzan un determinado valor?

¿Qué te ha parecido esta clase?					
Categorías	**F**	**P**	**%**	**FAA**	**FRA**
"Me ha parecido muy interesante"	17	0,23	22,97	17	0,23
"Me ha parecido bastante interesante"	22	0,30	29,73	39	0,53
"Me ha parecido algo interesante"	29	0,39	39,19	68	0,92
"No me ha parecido interesante"	6	0,08	8,11	74	1,00
	74	1	100		

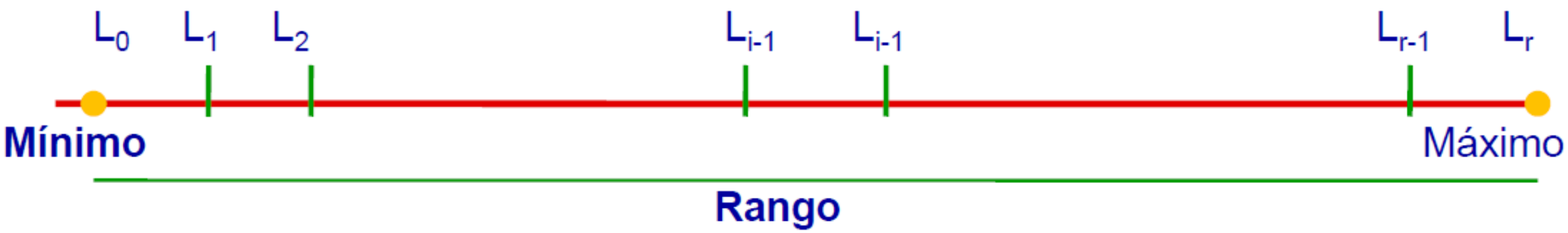
2. Tablas de frecuencias

Tipo de variable: cuantitativa y continua

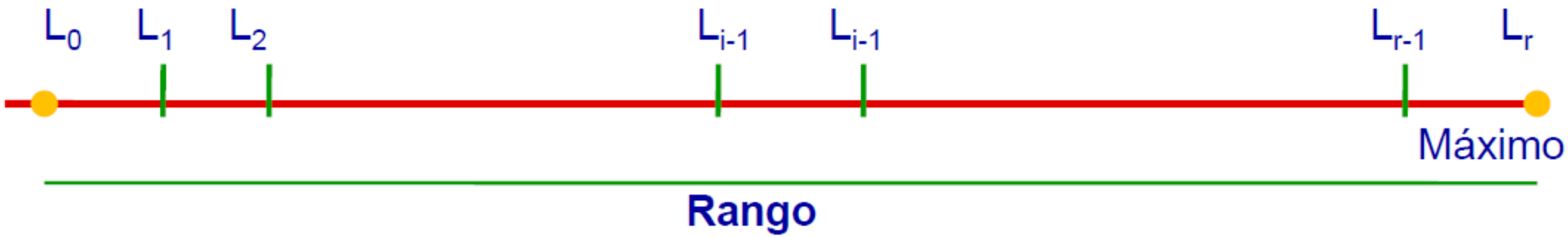
Valores: muchos y diferentes

Se clasifican en intervalos para poder manejarlos.
Pero los intervalos han de formar un sistema exclusivo y excluyente:

- Intervalos disjuntos (excluyentes)
- Que cubran todas las opciones (exclusivo)



2. Tablas de frecuencias



El siguiente paso es “contar” cuántos valores hay en cada intervalo (frecuencia absoluta) y con ello queda definida la distribución de frecuencias, pero ahora como: $\{(L_{i-1}, L_i; n_i)\}$

2. Tablas de frecuencias

Calificaciones de 120 exámenes

4,34	7,19	2,85	6,55	3,49	8,93	7,62	3,63
7,53	2,22	1,59	2,74	4,74	3,28	4,16	7,05
9,39	5,59	7,23	6,08	5,63	3,43	4,29	2,33
5,49	3,85	6,76	5,95	6,15	9,73	8,39	3,18
4,96	9,42	4,95	8,01	7,16	3,24	1,98	9,70
9,87	1,48	9,44	7,66	3,88	1,34	6,33	2,52
3,24	1,20	1,49	0,30	9,41	7,05	7,56	9,74
5,59	9,07	9,51	8,53	2,57	4,62	8,07	4,66
0,52	9,68	1,66	5,21	9,64	7,57	6,85	4,06
4,88	3,51	6,64	3,19	1,96	1,23	7,35	2,85
0,98	4,71	4,96	6,29	9,42	1,40	6,65	3,72
0,88	9,15	8,21	5,84	3,33	8,58	3,63	0,13
3,65	8,50	0,78	1,87	9,30	6,30	2,37	9,30
5,37	3,58	1,47	7,05	0,08	7,90	6,11	8,00
7,61	6,52	7,73	7,98	9,43	5,55	5,59	6,87

2. Tablas de frecuencias

L_{i-1} - L_i	n_i	f_i	p_i
0 - 1	7	0,0583	5,83
1 - 2	12	0,10	10,00
2 - 3	8	0,0667	6,67
3 - 4	16	0,133	13,33
4 - 5	12	0,10	10,00
5 - 6	10	0,0833	8,33
6 -7	13	0,1083	10,83
7 - 8	16	0,1333	13,33
8 - 9	9	0,0750	7,50
9 - 10	17	0,141	14,17
Sumas	120	1	100

Consideremos los intervalos cerrados por la derecha y abiertos por la izquierda, excepto el primero que es cerrado por ambos lados

Tabla de frecuencias completa

L_{i-1}-L_i	n_i	f_i	p_i	N_i	F_i	P
0 - 1	7	0,058	5,83	7	0,058	5,8
1 - 2	12	0,10	10,00	19	0,158	15,8
2 - 3	8	0,066	6,67	27	0,225	22,5
3 - 4	16	0,13	13,33	43	0,358	35,8
4 - 5	12	0,10	10,00	55	0,458	45,8
5 - 6	10	0,083	8,33	65	0,542	54,2
6 -7	13	0,108	10,83	78	0,650	65,0
7 - 8	16	0,13	13,33	94	0,783	78,3
8 - 9	9	0,075	7,50	103	0,858	85,8
9 - 10	17	0,14	14,17	120	1	100
	120	1	100			

2. Tablas de frecuencias

La introducción de los intervalos no permite realizar cálculos con los valores de la variable. Se hace preciso "discretizar" la variable continua para poder trabajar con ella. Para eso se introduce el concepto de marca de clase, que no es más que sustituir cada intervalo por su centro.

De este modo sustituiremos, a efectos de posteriores cálculos y cuando sea preciso, cada intervalo L_{i-1}-L_i por su "representante" X_i

$$x_i = \frac{L_{i-1} + L_i}{2}$$

En cuanto a la diferencia L_i-L_{i-1} = a_i la denominaremos amplitud del intervalo i-ésimo

Con lo cual la tabla de frecuencias anterior, quedaría como sigue y en su tratamiento posterior la variable se va a tratar, a efectos dé cálculo, como discreta.

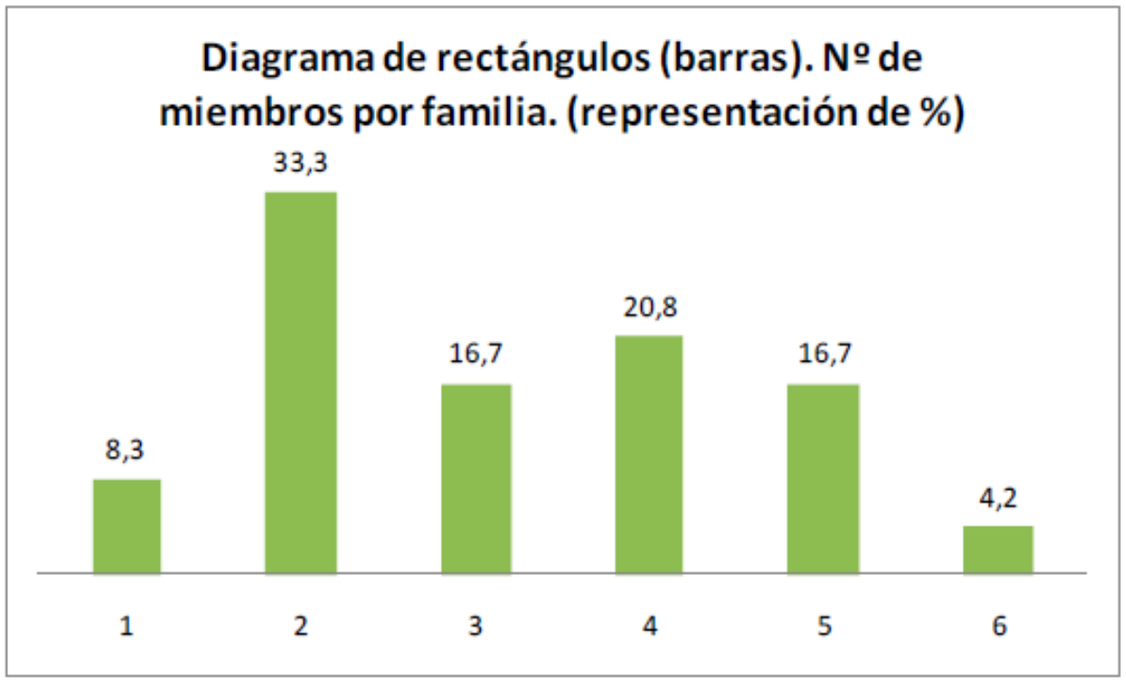
L_{i-1}-L_i	x_i	n_i	f_i	p_i	N_i	F_i	P
0 - 1	0,5	7	0,05	5,8	7	0,05	5,8
1 - 2	1,5	12	0,10	10,0	19	0,15	15,8
2 - 3	2,5	8	0,06	6,7	27	0,22	22,5
3 - 4	3,5	16	0,13	13,3	43	0,35	35,8
4 - 5	4,5	12	0,10	10,0	55	0,45	45,8
5 - 6	5,5	10	0,08	8,3	65	0,54	54,2
6 -7	6,5	13	0,10	10,8	78	0,65	65,0
7 - 8	7,5	16	0,13	13,3	94	0,78	78,3
8 - 9	8,5	9	0,07	7,5	103	0,85	85,8
9 - 10	9,5	17	0,14	14,2	120	1	100
		120	1	100			

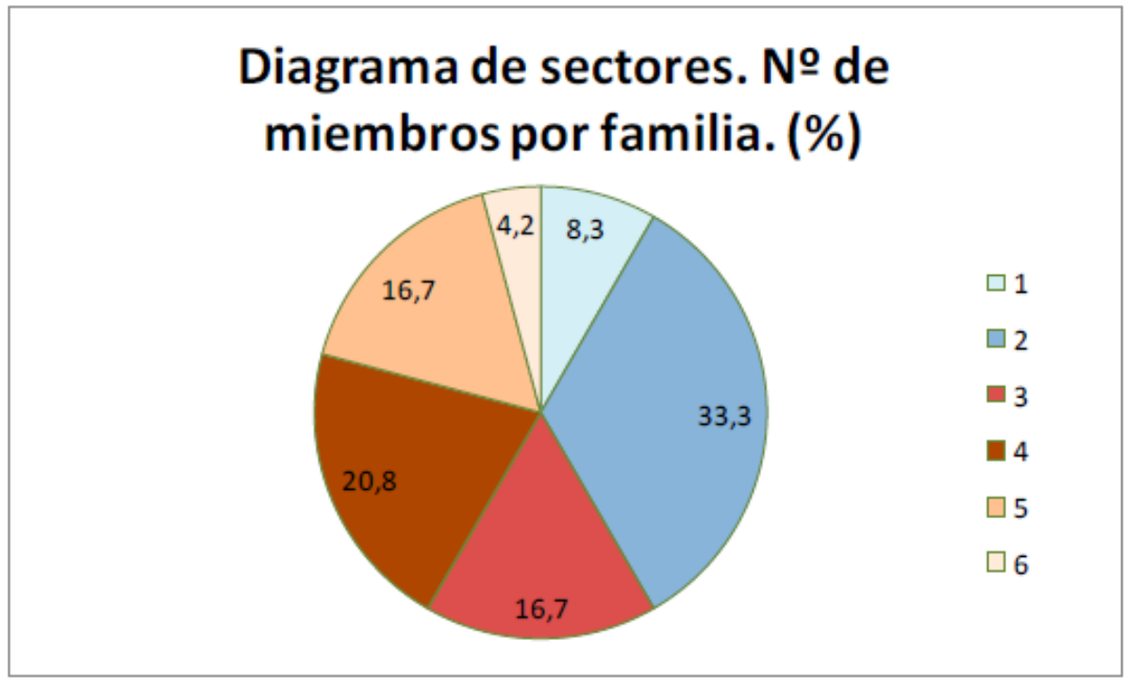
3. Representaciones Gráficas

- En el análisis descriptivo **conviene** hacer representaciones **gráficas** de las tablas de frecuencias para hacerse una idea rápida de lo que está ocurriendo.
- Los gráficos adecuados para representar variables cualitativas son los **gráficos de barras** y los **diagramas de sectores**.
- Para las frecuencias acumuladas se pueden usar **gráficos de barras** y los **gráficos de líneas**.

3. Representaciones Gráficas

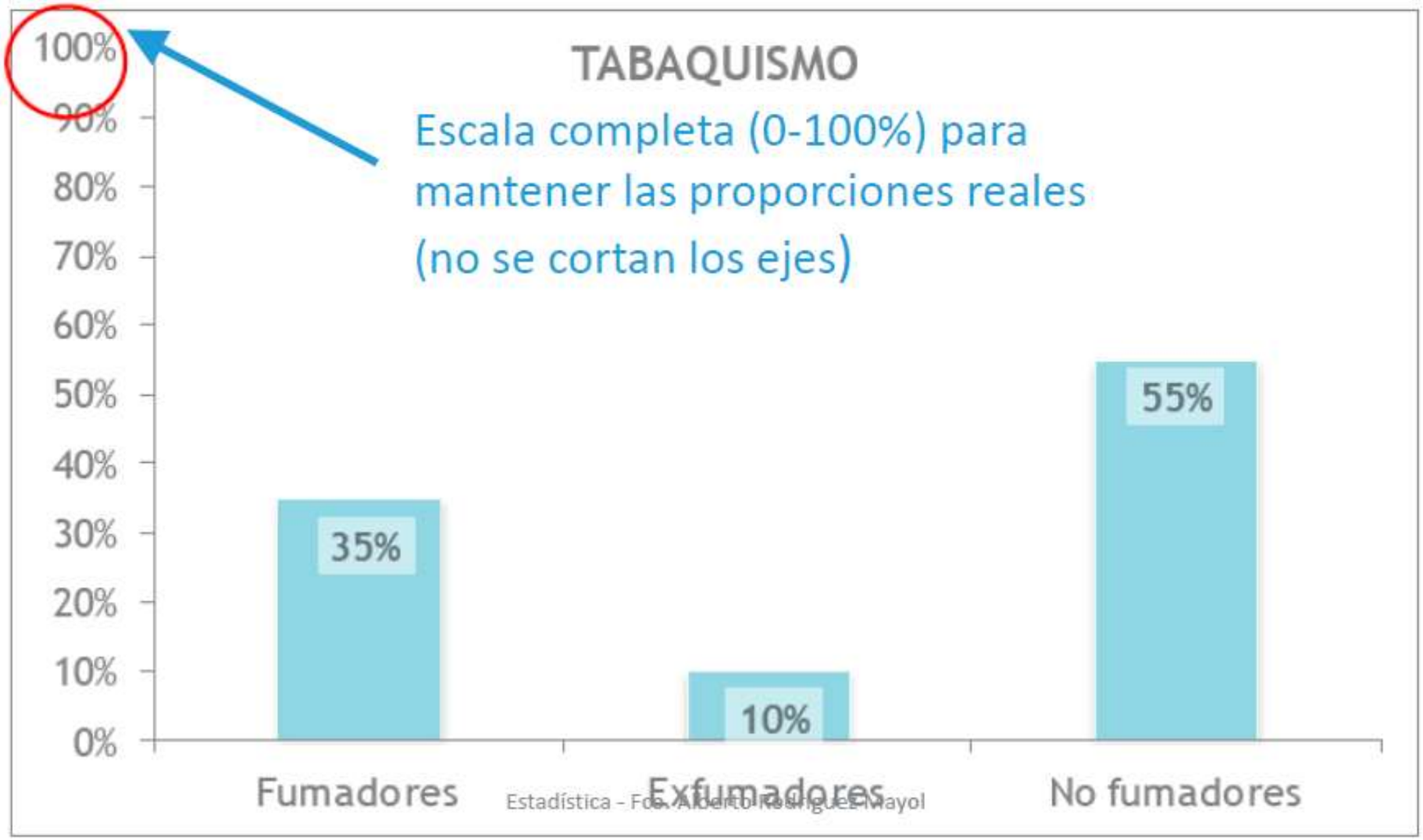
DIAGRAMAS DE RECTANGULOS Y/O SECTORES



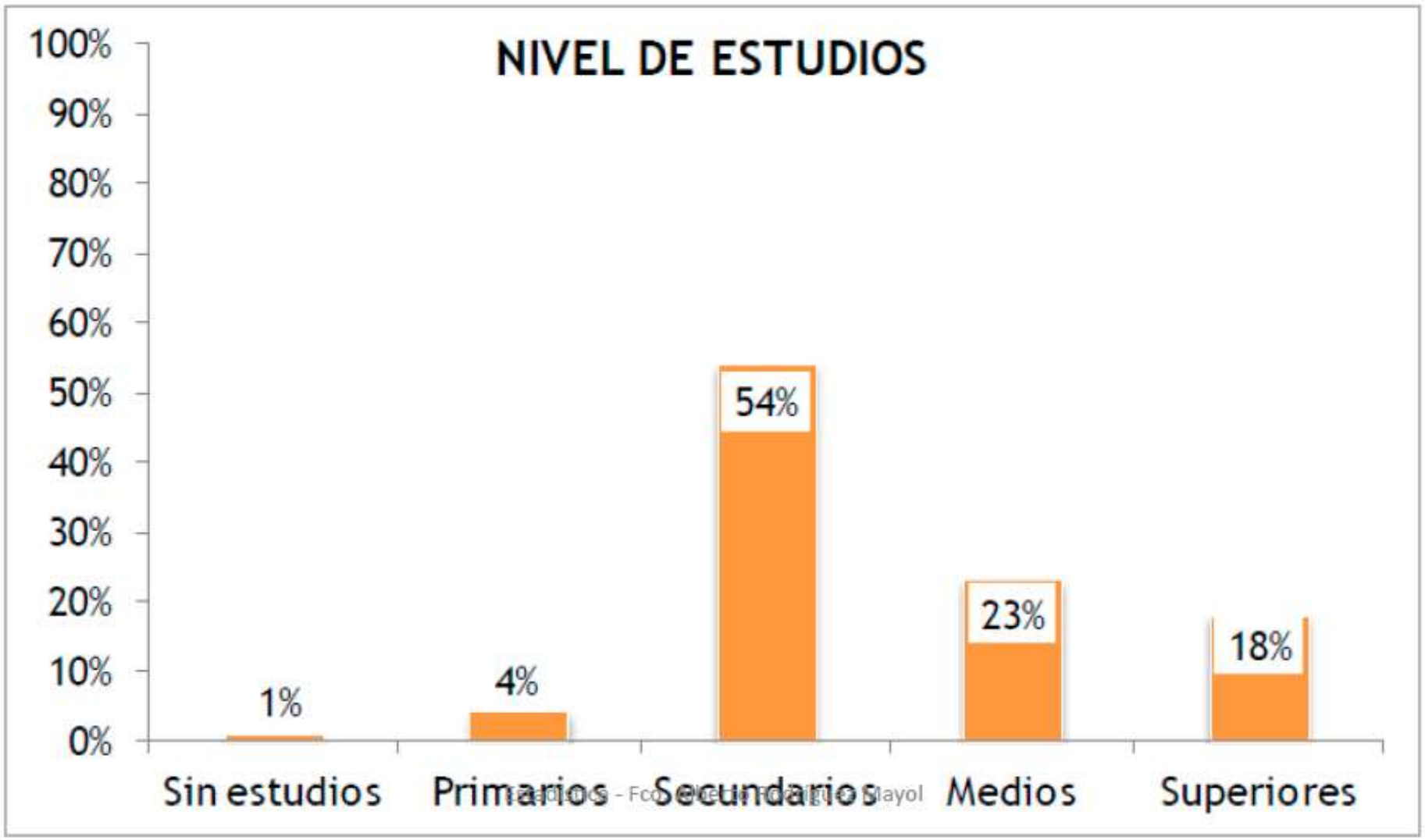


Válidos para variables cualitativas y cuantitativas discretas

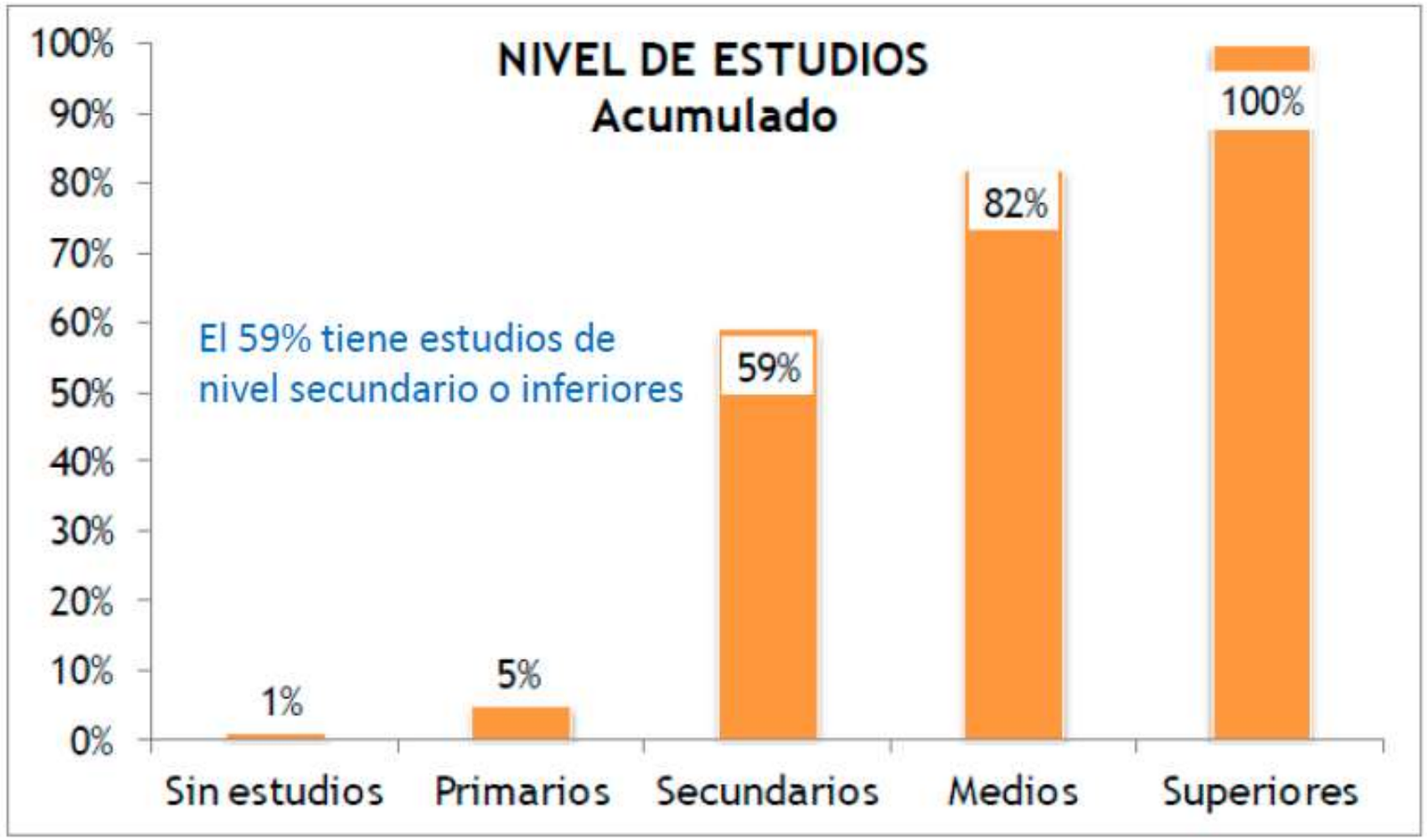
3. Representaciones Gráficas



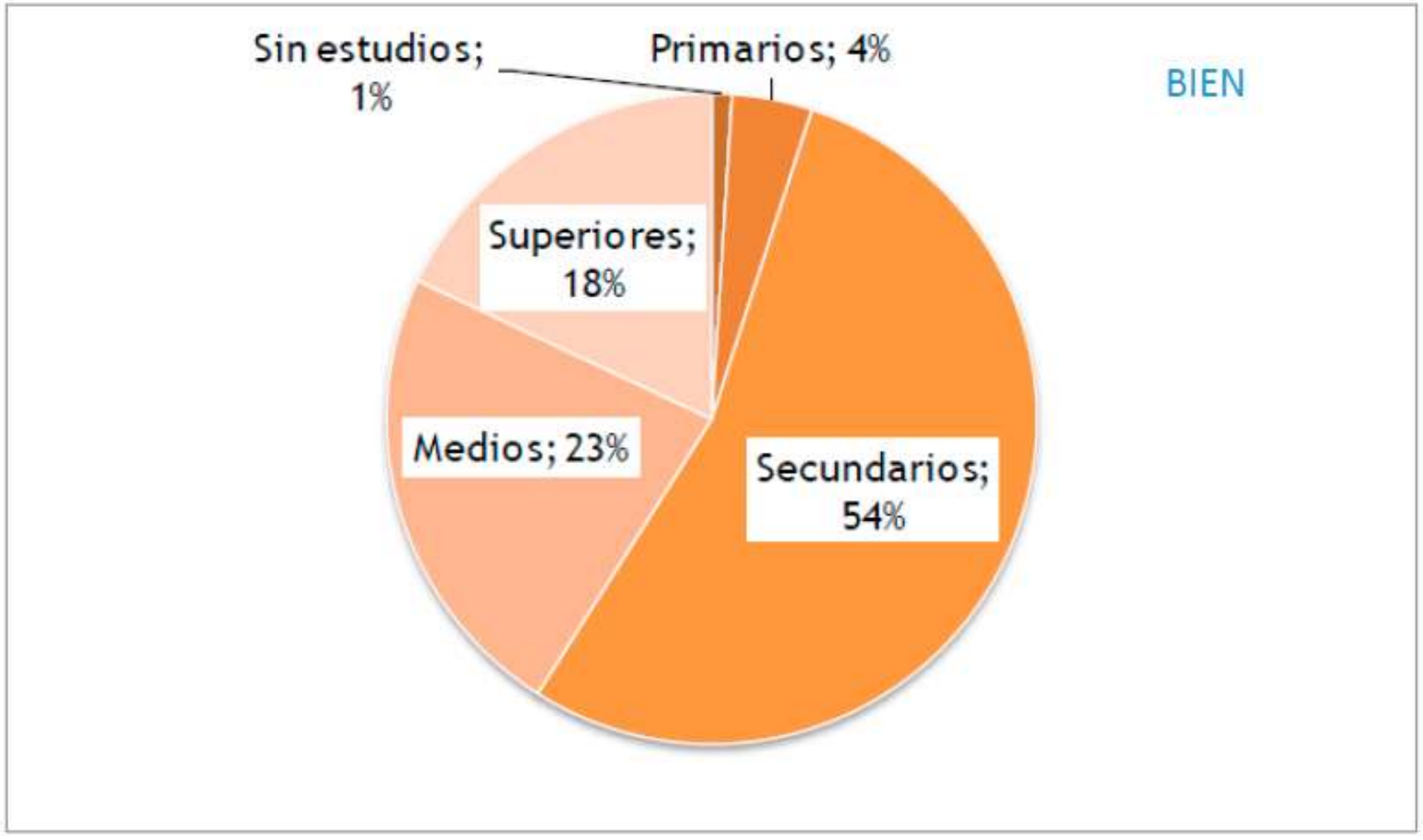
3. Representaciones Gráficas



3. Representaciones Gráficas

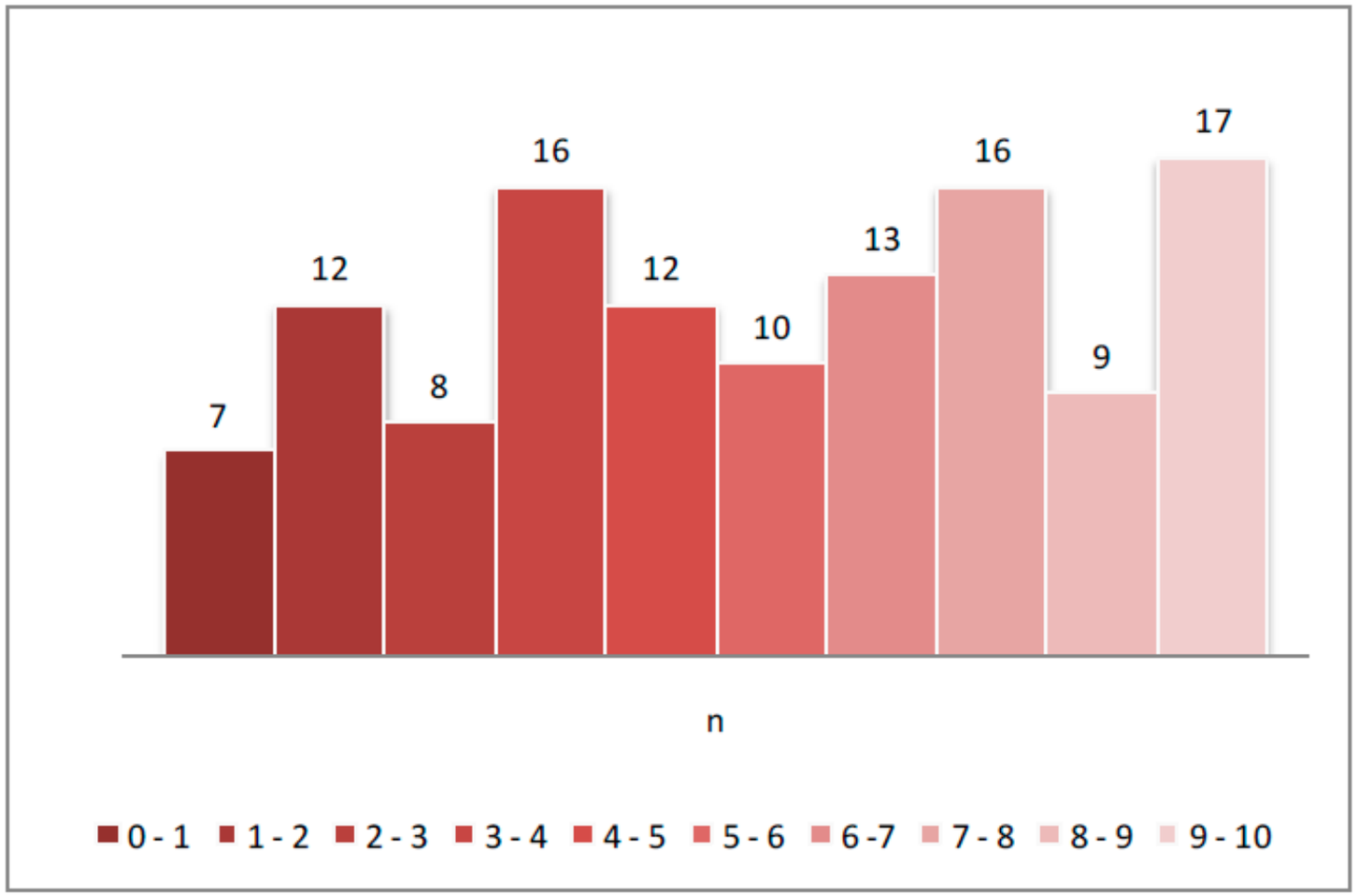


3. Representaciones Gráficas



3. Representaciones Gráficas

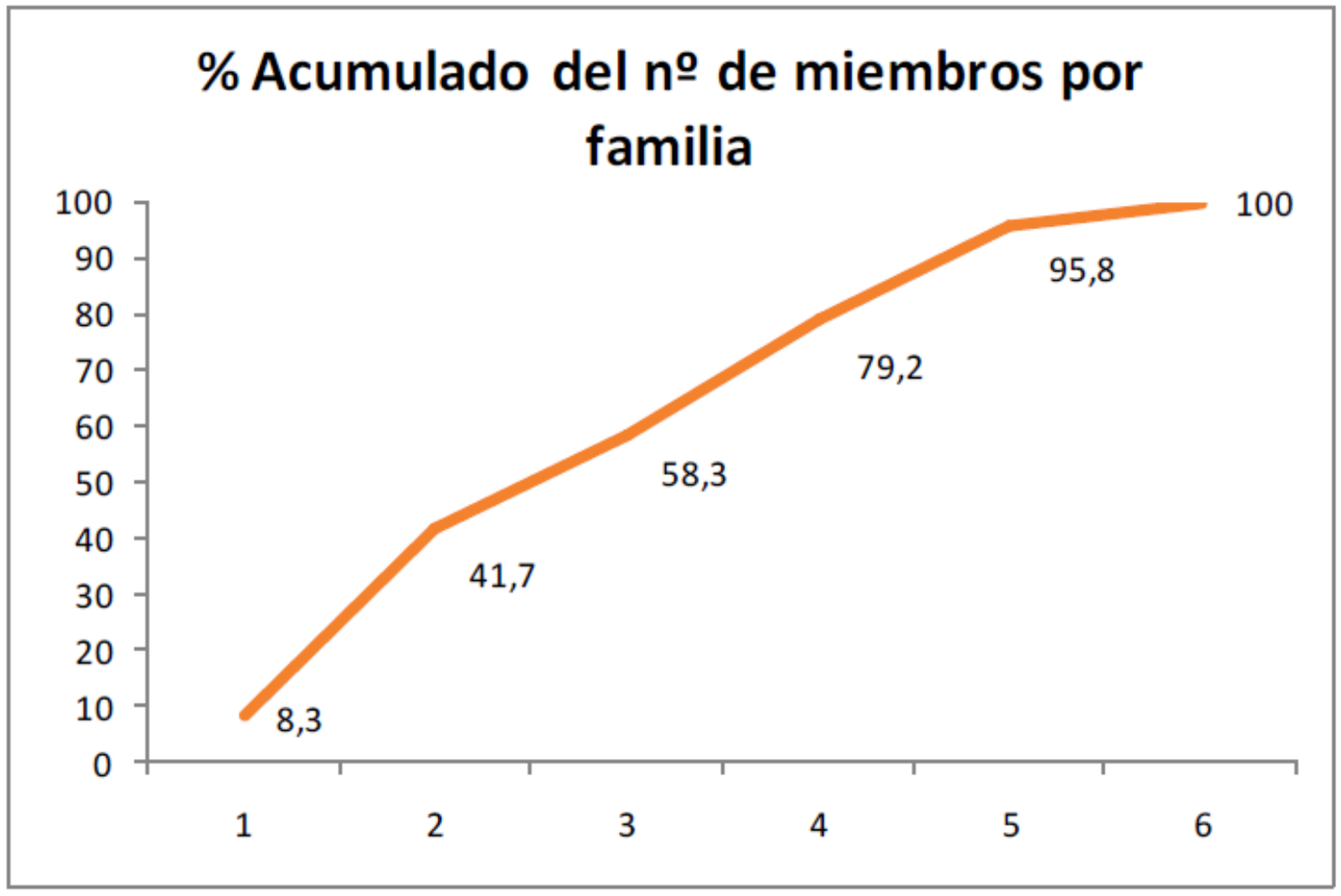
Histograma: calificaciones de examen



Caso de variables cuantitativas continuas

3. Representaciones Gráficas

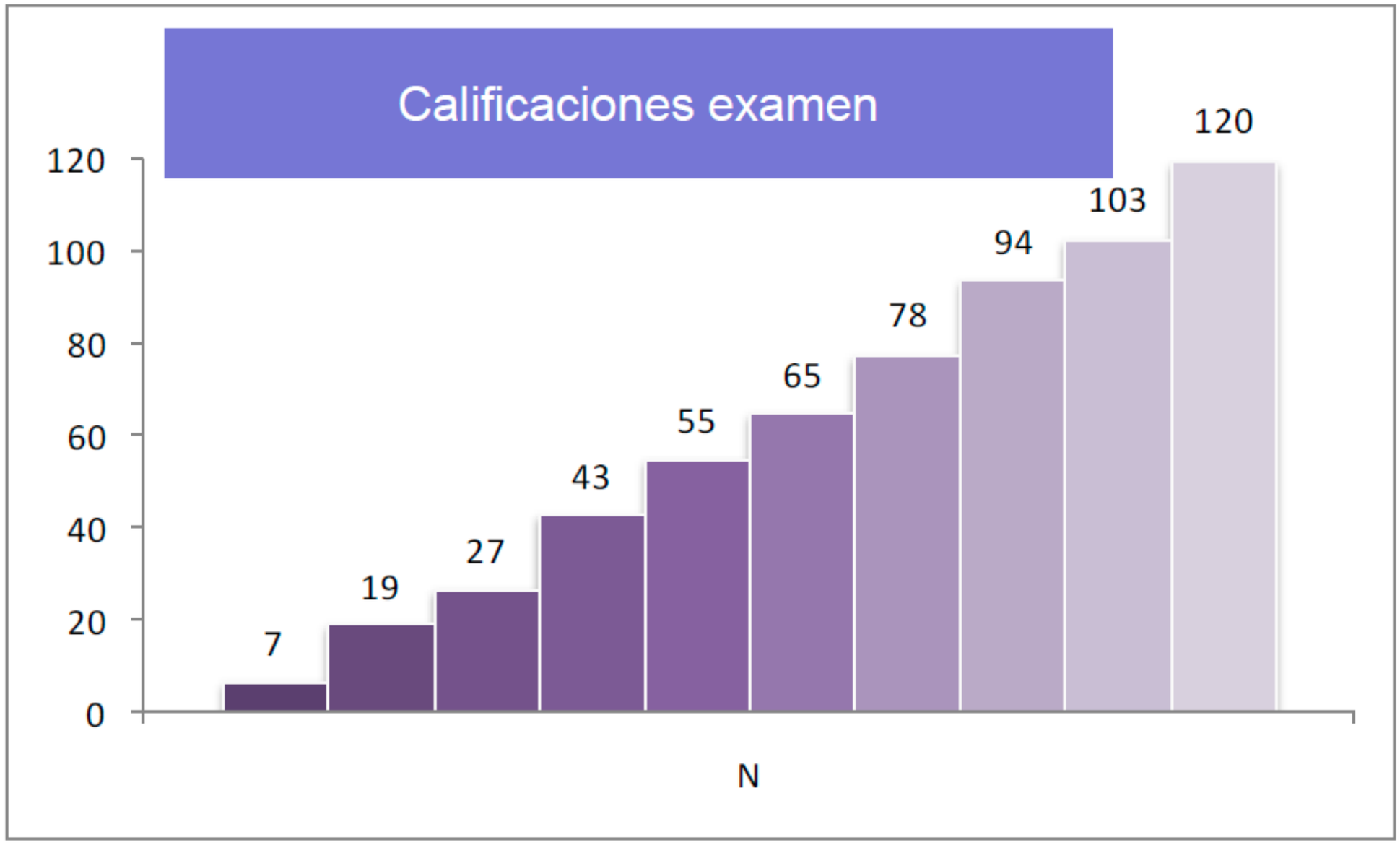
CURVAS ACUMULATIVAS. Para frecuencias acumuladas



Solo tiene sentido para v.e. cuantitativas discretas

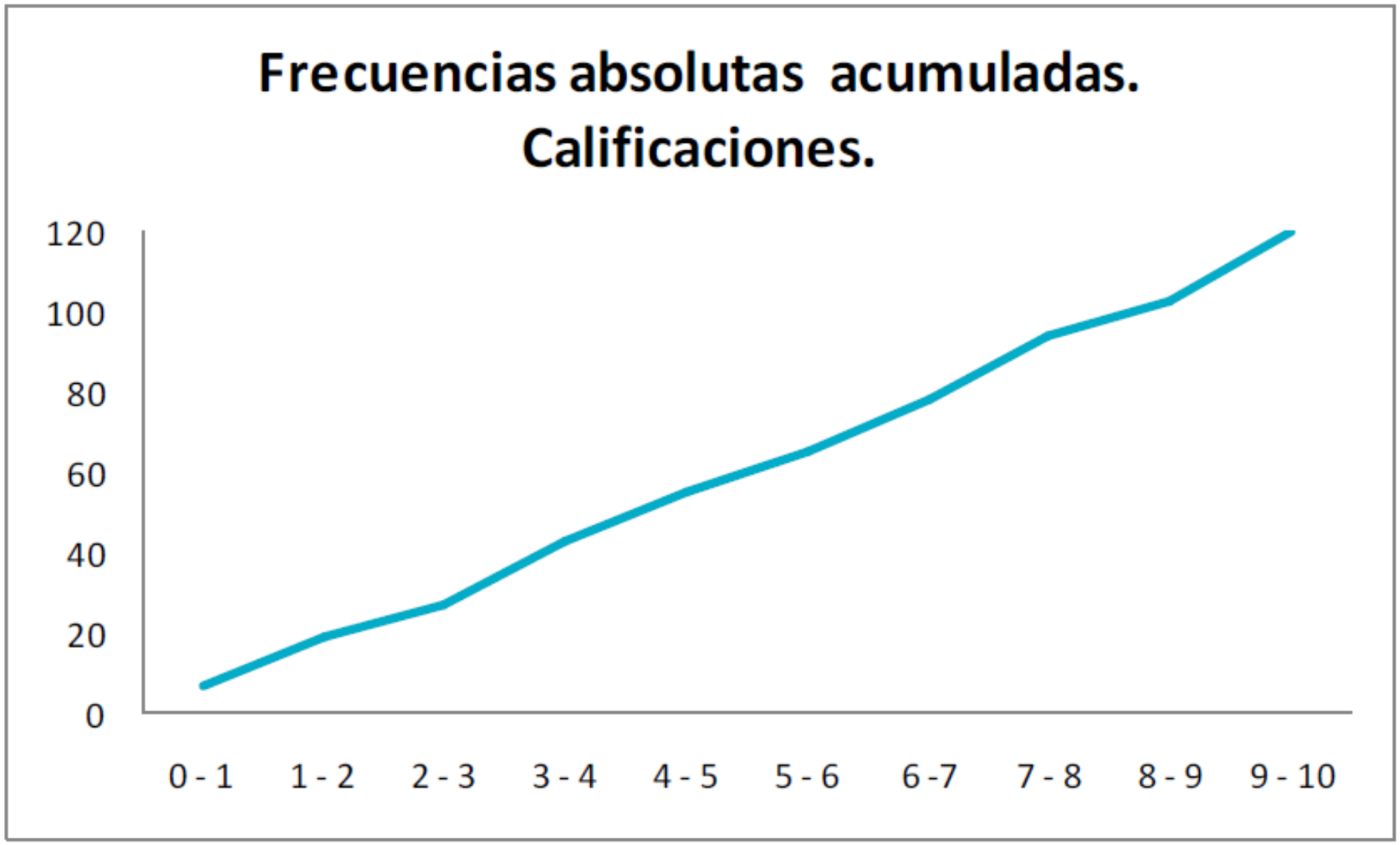
3. Representaciones Gráficas

HISTOGRAMA ACUMULATIVO. Para frecuencias acumuladas



3. Representaciones Gráficas

CURVA ACUMULATIVA. Para frecuencias acumuladas



3. Representaciones Gráficas



4. Medidas de posición

OBJETIVO: Buscar indicadores que representen y sinteticen el conjunto de datos.

Sintetizar: refundir toda la información en un solo indicador
Representar: conocer hasta qué punto ese indicador representa al conjunto de datos

Hacen falta dos clases de indicadores:

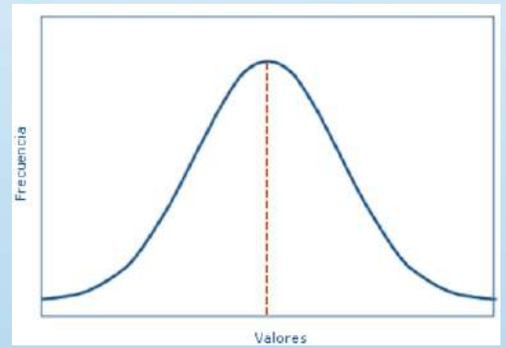
- De posición, es decir indicadores que informen y representen al conjunto de datos

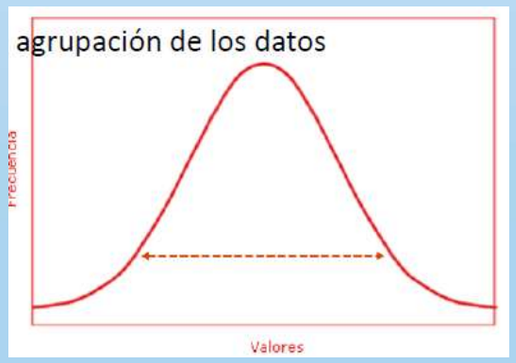
- De dispersión: que informen o bien de la representatividad de los anteriores, o bien nos informe del grado de acumulación o dispersión de los datos y por tanto de su grado de variabilidad.

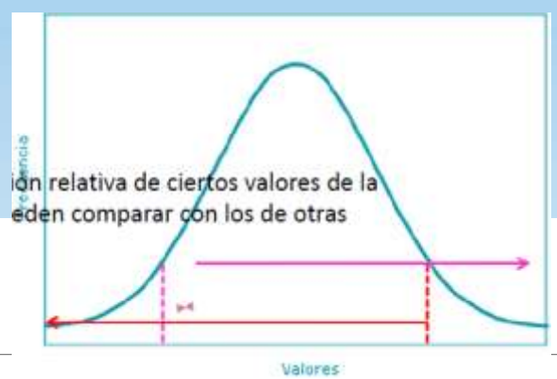
4. Medidas de posición

- Existen más herramientas para describir el conjunto de valores que toma una variable cuantitativa que cualitativa (aunque siempre se puede categorizar)

Medidas de centralización	Media Mediana Moda
Medidas de dispersión	Rango Varianza Desviación típica Coeficiente de variación
Medidas de posición	Cuantiles Percentiles Deciles Cuartiles
Medidas de forma	Asimetría (skewness) Apuntamiento (curtosis)







4. Medidas de posición

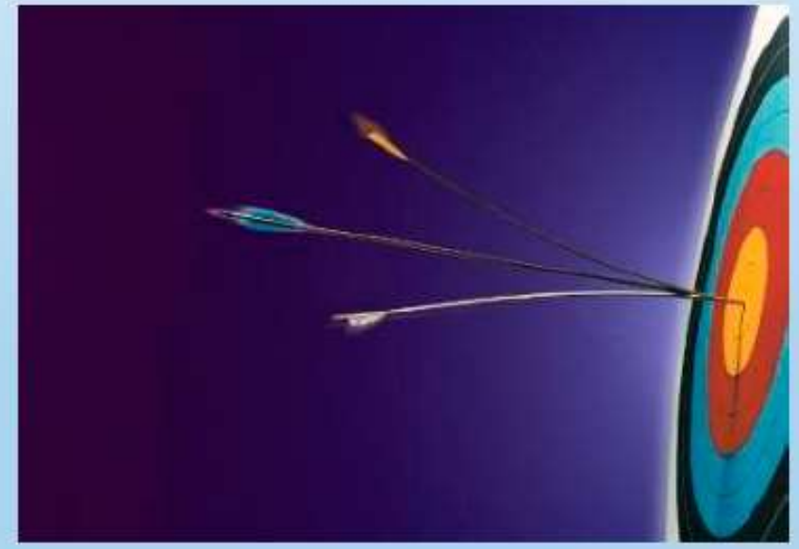
Medidas de centralización

- **Indicadores de hacia qué valores tienden las medidas a agruparse**

Media

Mediana

Moda



4. Medidas de posición

MEDIA ARITMÉTICA: Es el promedio de todos los datos obtenidos. Al producirse repeticiones de valores (frecuencias), su cálculo se realiza como:

$$\bar{X} = \frac{\sum_{1}^{r} x_i n_i}{N} = \sum_{1}^{r} x_i f_i$$

La frecuencia puede interpretarse como la “importancia” que tiene cada valor, dentro del conjunto de datos

4. Medidas de posición

La ordenación de cálculos para la media es muy simple, basta construir lo siguiente sobre la tabla de frecuencias:

x_i	n_i	$x_i n_i$
1	2	2
2	8	16
3	4	12
4	5	20
5	4	20
6	1	6
Sumas	24	76
	MEDIA	**3,17**

$$\overline{X} = \frac{\sum_{1}^{r} x_i n_i}{N} = \frac{76}{24} = 3{,}17$$

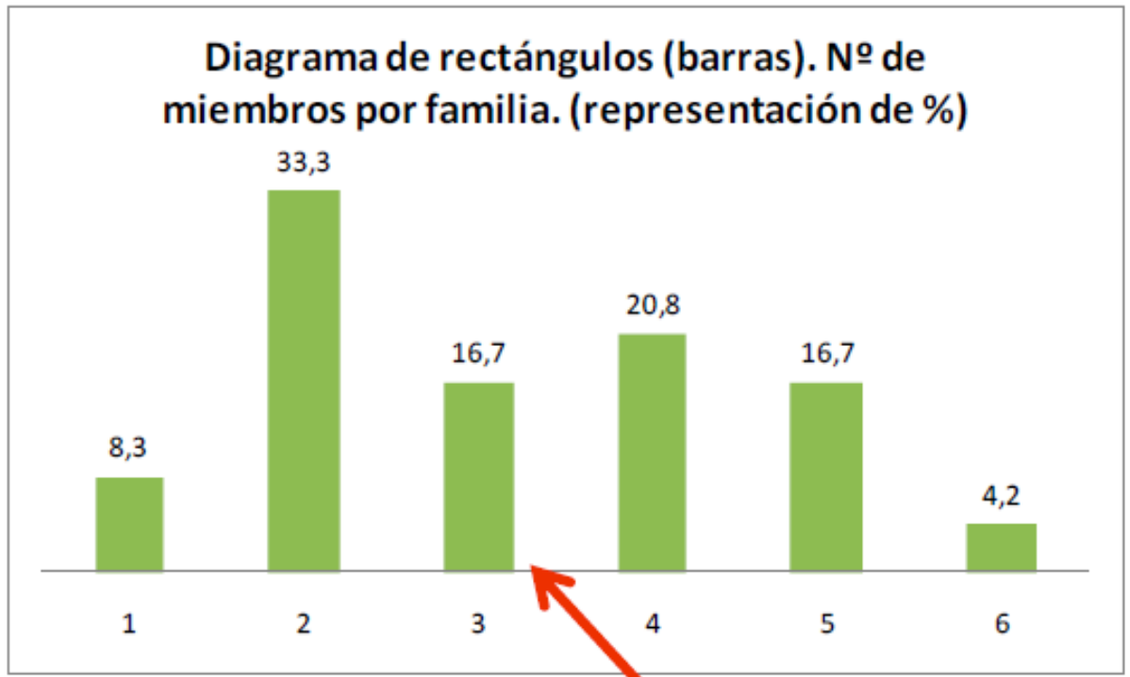
4. Medidas de posición

- Cálculo de la **media**:
  - 1º Suma de todos los valores
  - 2º Dividir entre el número de valores (tamaño muestral).
  - **(46 + 54 + …+ 78)/20 = 1308/20 = 65,4**

	Peso
1	46
2	54
3	58
4	58
5	60
6	61
7	62
8	62
9	63
10	66
11	67
12	67
13	67
14	68
15	72
16	73
17	73
18	76
19	77
20	78

4. Medidas de posición

MEDIA ARITMÉTICA: Al ser un promedio que combina valores con frecuencias, la resultante del cálculo puede interpretarse como el centro de la distribución de frecuencias. Y de hecho es el equivalente a su centro de gravedad, hablando en términos físicos.



4. Medidas de posición

Una cuestión a tener siempre en cuenta es hasta qué punto la media representa adecuadamente el conjunto de datos, además de ser su centro. Dicho de otro modo, para ser un buen indicador debe estar suficientemente cerca de todos los valores de la variable. Esto supone introducir un concepto adicional: la representatividad de la media aritmética

Si 5 personas tienen un salario de 800 euros y una sexta persona tiene un salario de 5.000 euros mensuales. El salario medio del conjunto es de 1.500 euros. ¿Es un buen indicador de la distribución de los salarios de esa pyme?

Si por el contrario las primeras 5 personas mantuvieran su salario de 800 euros y la sexta lo tuviera de 1.200 la media sería de 866,6 euros. ¿Es mejor representante que en el caso anterior?

4. Medidas de posición

Una propiedad útil de la media. ¿Qué ocurre si hacemos un cambio de variable?.
Por ejemplo si tenemos una distribución de salarios en euros y queremos convertirla a dólares. ¿Qué ocurre con la media? ¿Hay que calcularla de nuevo?

$$Y = a + bX \quad \Rightarrow \quad \overline{Y} = a + b\overline{X}$$

$$\overline{Y} = \frac{\sum_1^r y_i n_i}{N} = \frac{\sum_1^r (a + bx_i) n_i}{N} = \frac{\sum_1^r a n_i}{N} + \frac{\sum_1^r b x_i n_i}{N} = \frac{a\sum_1^r n_i}{N} + \frac{b\sum_1^r x_i n_i}{N} =$$

$$= \frac{aN}{N} + b\overline{X} = a + b\overline{X}$$

4. Medidas de posición

MEDIANA: Es aquel valor que es superado por al menos el 50% de las observaciones y que, a la vez, supera al menos al 50% de las observaciones. A efectos prácticos es el valor que se corresponde con una frecuencia acumulada del 50% cuando ese valor existe.

En el caso discreto se puede obtener ordenando todos los valores de menor a mayor, tantas veces como frecuencia tengan, y designando como mediana el que ocupe el lugar central (o la media de los dos valores centrales, en su caso)

En el caso continuo hay que calcular en primer lugar el intervalo mediano, que es primero en que el valor de la frecuencia acumulada es al menos del 50%. Dentro de él estará la mediana y su expresión se obtiene mediante un cálculo geométrico.

4. Medidas de posición

MEDIANA: Me=3

x_i	n_i	f_i	p_i	N_i	F_i	P
1	2	0,1	8,3	2	0,1	8,3
2	8	0,3	33,3	10	0,4	41,7
3	4	0,2	16,7	14	0,6	58,3
4	5	0,2	20,8	19	0,8	79,2
5	4	0,2	16,7	23	1,0	95,8
6	1	0,0	4,2	24	1	100
	24	1	100			

4. Medidas de posición

Medidas de centralización

- Cálculo de la **mediana**:
  - 1º Ordenación de los datos
  - 2º Localización del valor que deja a un lado y otro el mismo número de casos.
- Si hay dos valores centrales se calcula la media de los dos.

- **(66 + 67)/2 = 66,5**

- Si hubiera sólo 19 valores => Sería el valor 10 => El **66**
- Si hubiera 21 valores => Sería el valor 11 => El **67**

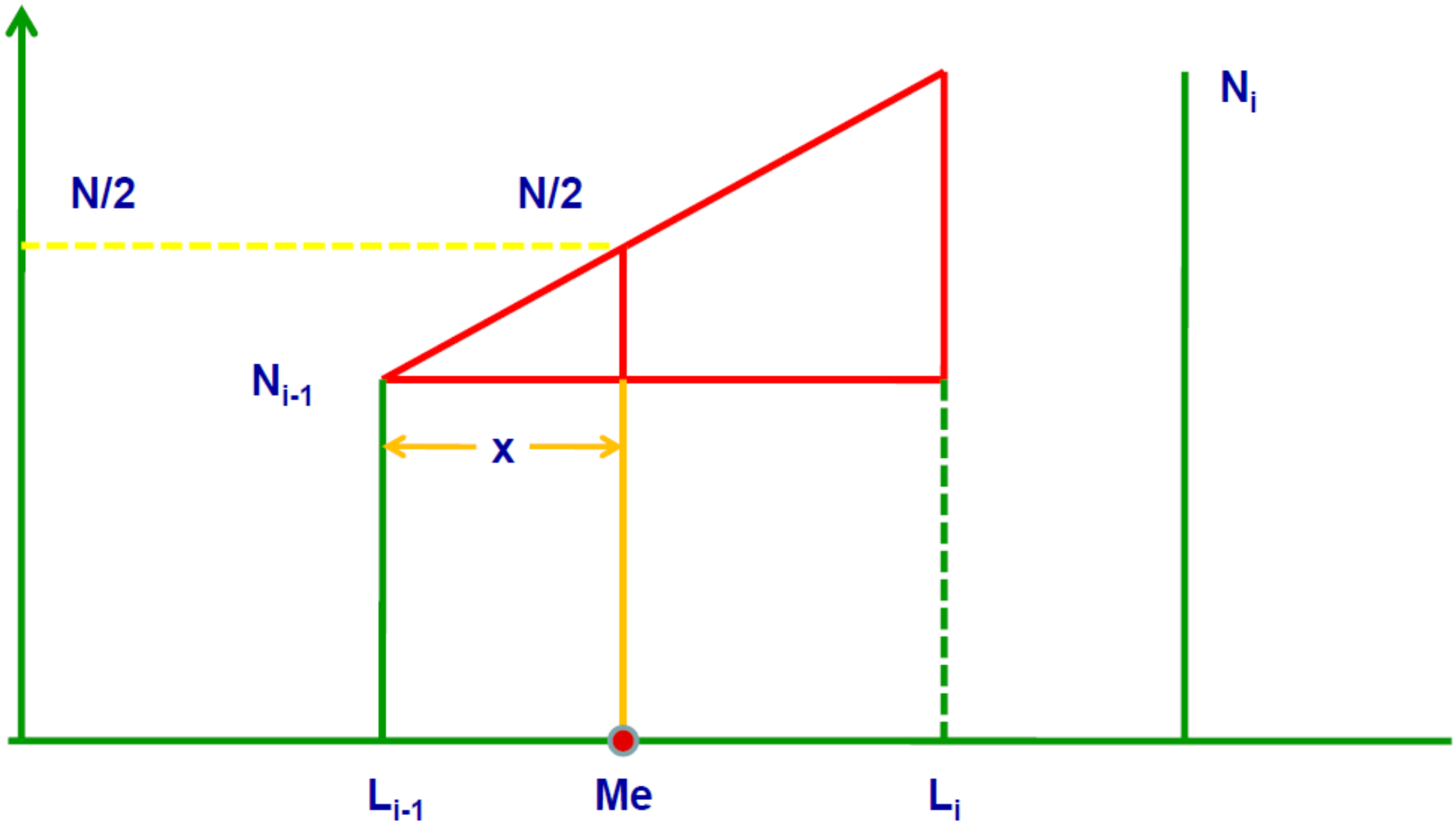
	Peso
1	46
2	54
3	58
4	58
5	60
6	61
7	62
8	62
9	63
10	66
11	67
12	67
13	67
14	68
15	72
16	73
17	73
18	76
19	77
20	78

4. Medidas de posición

L_{i-1}-L_i	N_i	F_i	P_i
0 - 1	7	0,058	5,8
1 - 2	19	0,158	15,8
2 - 3	27	0,225	22,5
3 - 4	43	0,358	35,8
4 - 5	55	0,458	45,8
5 - 6	65	0,542	54,2
6 -7	78	0,65	65,0
7 - 8	94	0,783	78,3
8 - 9	103	0,858	85,8
9 - 10	120	1	100

Intervalo mediano: 5-6

4. Medidas de posición



Intervalo mediano

Mediana $Me=L_{i-1}+x$

4. Medidas de posición

$$x = \frac{\frac{N}{2} - N_{i-1}}{N_i - N_{i-1}}(L_i - L_{i-1}) = \frac{\frac{N}{2} - N_{i-1}}{n_i} a_i$$

$$x = \frac{60 - 55}{10} 1 = \frac{5}{10} 1 = 0{,}5$$

Mediana Me= 5 + 0,5 = 5,5

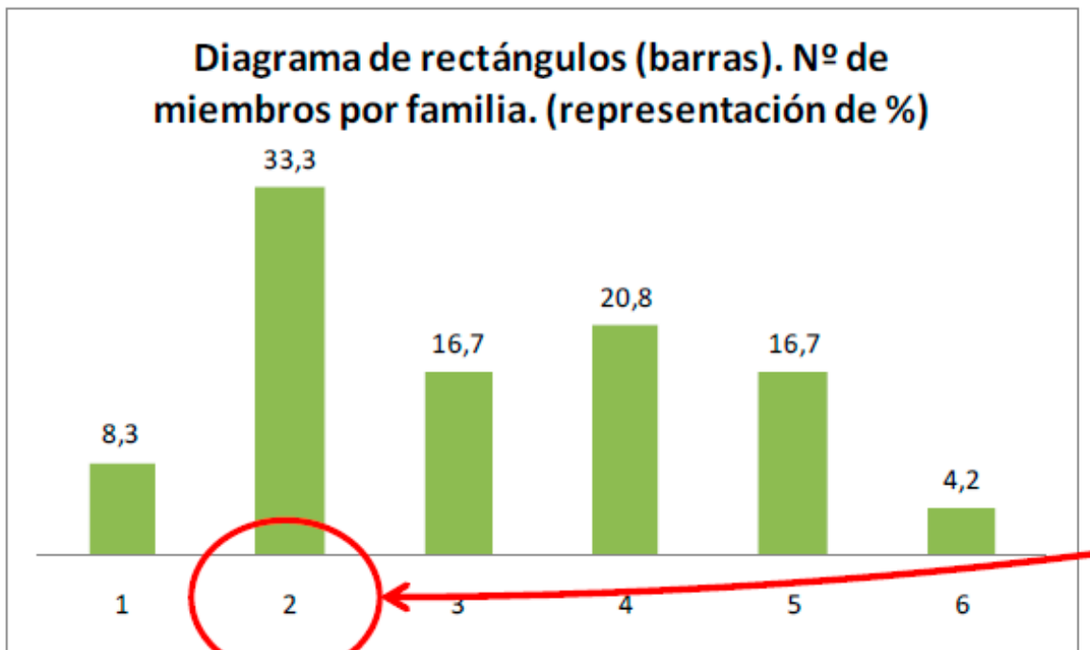
4. Medidas de posición

MODA: Es el valor que más se repite o que más veces aparece. Puede utilizarse tanto para variables cualitativas como cuantitativas.

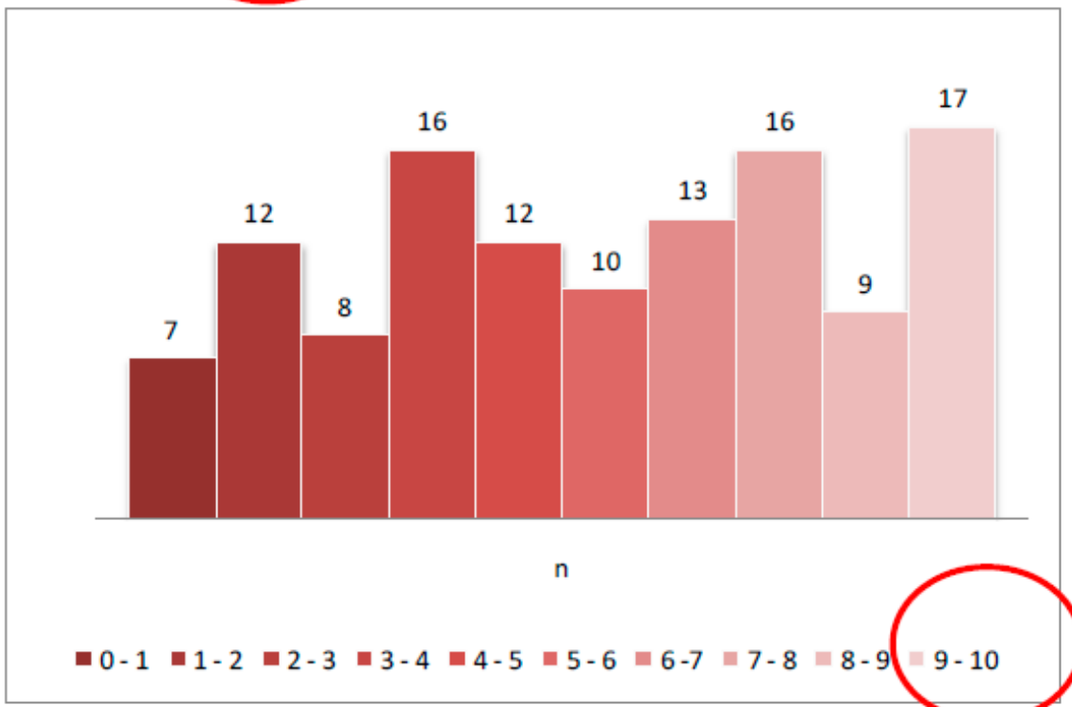
En el caso de v.e. cualitativas o de cuantitativas discretas es simplemente el valor que presente mayor frecuencia. Esto supone que una distribución de frecuencias no tiene por qué tener solo una moda o valor modal (distribución unimodal), sino que puede ocurrir que tenga varias modas (distribución multimodal)

Cuando se trata de v.e. cuantitativas continuas se hablará de intervalo modal, que será aquel que presente mayor frecuencia.

4. Medidas de posición



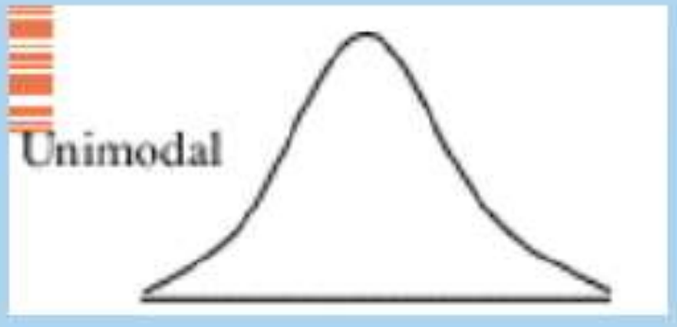
MODA: Mo=2

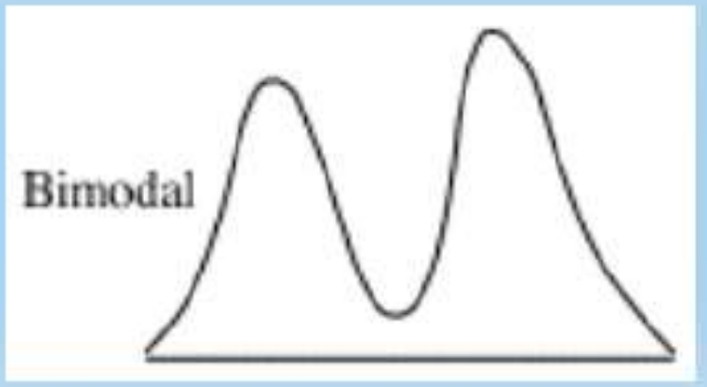


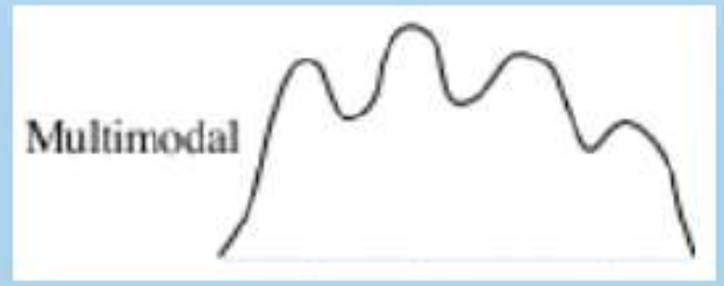
INTERVALO MODAL

4. Medidas de posición

- El valor que más veces aparece repetido en la distribución (con mayor frecuencia).
  - Pueden existir distribuciones con más de una moda (bimodales, multimodales, etc.)







4. Medidas de posición

Medidas de centralización

- Cálculo de la **moda**:
  - 1º Localizar la observación que se presenta con más frecuencia.
  - **67 (frecuencia absoluta 3, frecuencia relativa 15%)**

	Peso
1	46
2	54
3	58
4	58
5	60
6	61
7	62
8	62
9	63
10	66
11	67
12	67
13	67
14	68
15	72
16	73
17	73
18	76
19	77
20	78

4. Medidas de posición

- Cuantiles
- Percentiles
- Deciles
- Cuartiles

- Ofrecen información sobre la posición relativa de ciertos valores de la variable a partir de los cuales se pueden comparar con los de otras
- Ayudarán a distinguir si una observación es relativamente pequeña o grande, respecto de estos valores que vamos a definir

4. Medidas de posición

Cuantiles (de 0 a 1)

Percentiles (de 0 a 100)

Deciles (de 0 a 10)

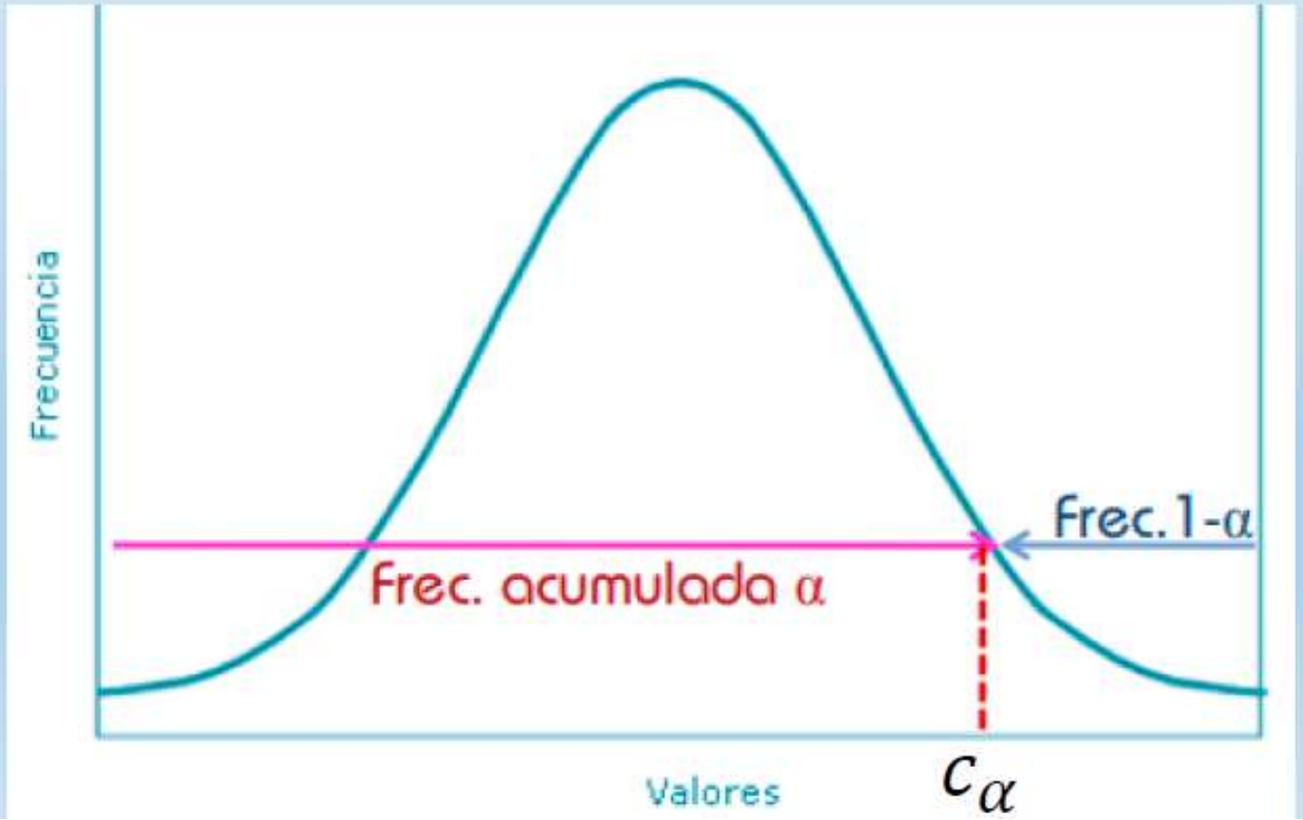
Cuartiles (de 1 a 3)

- **Cuantil** de orden α es un valor de la variable por debajo del cual se encuentra una frecuencia acumulada de α
- **Percentil** de orden k es el cuantil de orden α=k/100. Tiene el mismo significado que el cuantil, pero se expresa en %
- **Decil** de orden k es el cuantil de orden α=k/10. Tiene el mismo significado que el cuantil, pero en escala de 0 a 10
- Los **cuartiles** dividen a la muestra en cuatro grupos iguales.

4. Medidas de posición

Cuantiles (de 0 a 1)

- **Cuantil** de orden α, c_α es un valor de la variable por debajo del cual se encuentra una frecuencia acumulada de α



4. Medidas de posición

Percentiles (de 0 a 100)

$\alpha = k/100$

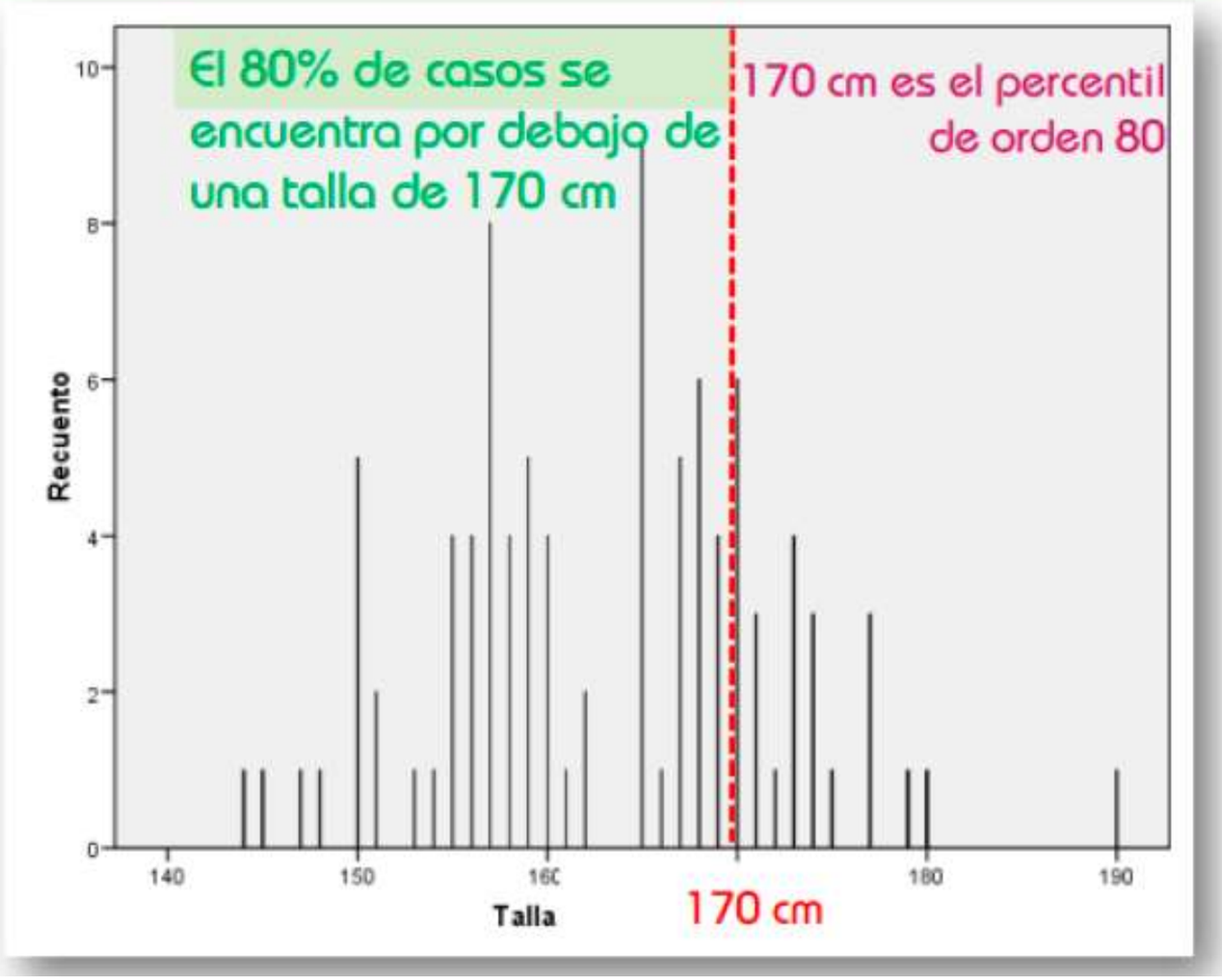
- **Percentil** de orden k es el cuantil de orden $\alpha=k/100$. Tiene el mismo significado que el cuantil, pero se expresa en %
- **Percentil** de orden k es el valor por debajo del cual se encuentra una frecuencia acumulada del k %
- El **percentil 50** es el valor de la variable que deja a un lado y a otro el mismo número de casos = **MEDIANA.**
- *Para calcularlos es necesario ordenar los datos (de menor a mayor o al revés)*

4. Medidas de posición

Talla del paciente (cm)

		Porcentaje acumulado
Válidos	144	1,1
	145	2,1
	147	3,2
	148	4,3
	150	9,6
	151	11,7
	153	12,8
	154	13,8
	155	18,1
	156	22,3
	157	30,9
	158	35,1
	159	40,4
	160	44,7
	161	45,7
	162	47,9
	165	57,4
	166	58,5
	167	63,8
	168	70,2
	169	74,5
	170	80,9
	171	84,0
	172	85,1
	173	89,4
	174	92,6
	175	93,6
	177	96,8
	179	97,9
	180	98,9
	190	100,0

Queremos conocer qué valor de la variable nos deja por debajo de él aprox. Al 80% de los casos



4. Medidas de posición

Deciles (de 0 a 10)

$\alpha = k/10$

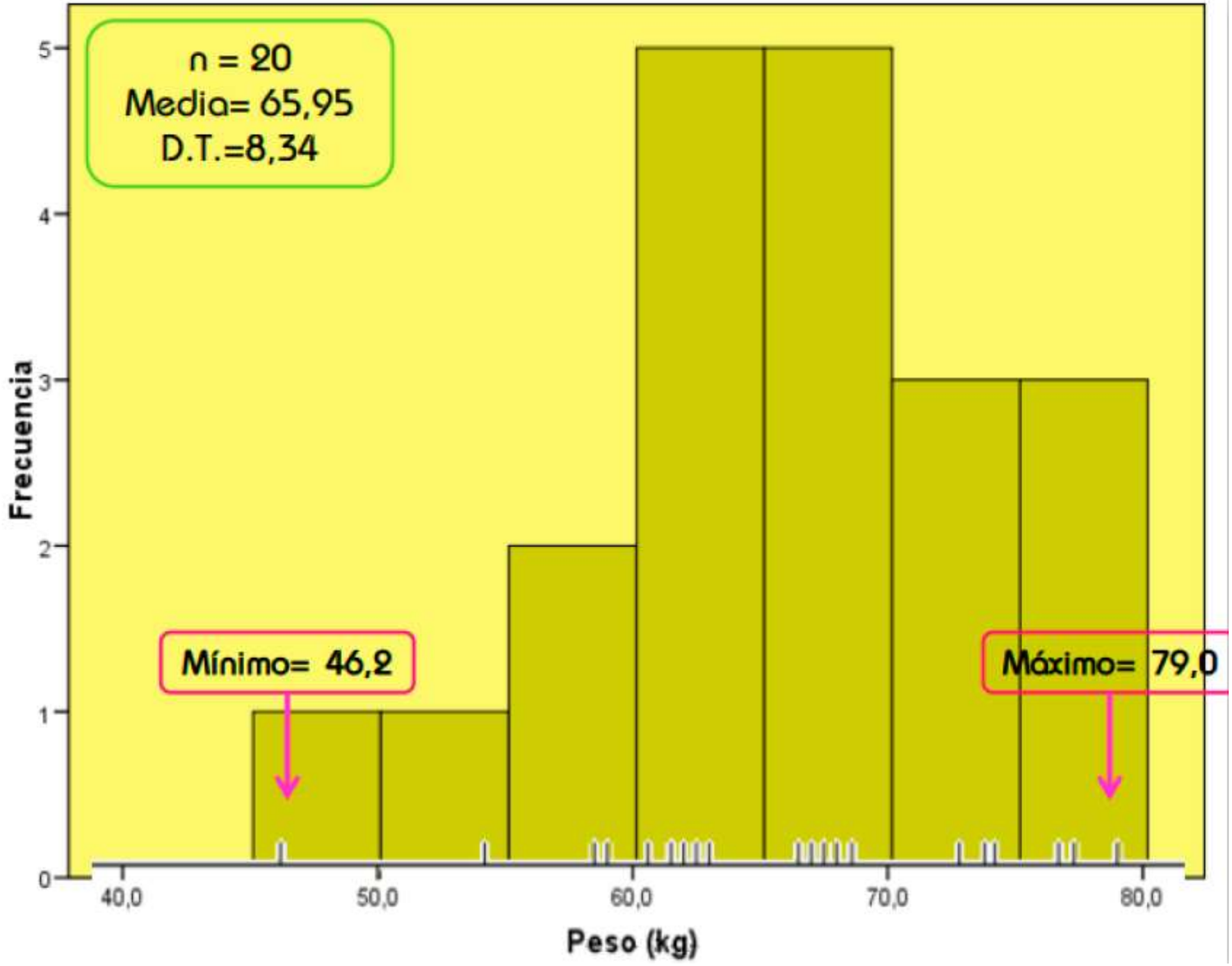
- **Decil** de orden k es el cuantil de orden $\alpha=k/10$. Tiene el mismo significado que el cuantil, pero se expresa en escala de 0 a 10
- **Decil** de orden k es el valor por debajo del cual se encuentra una frecuencia acumulada de (10·k) %
- El **decil 5** es el valor de la variable que deja a un lado y a otro el mismo número de casos = **MEDIANA.**
- *Para calcularlos es necesario ordenar los datos (de menor a mayor o al revés)*

4. Medidas de posición

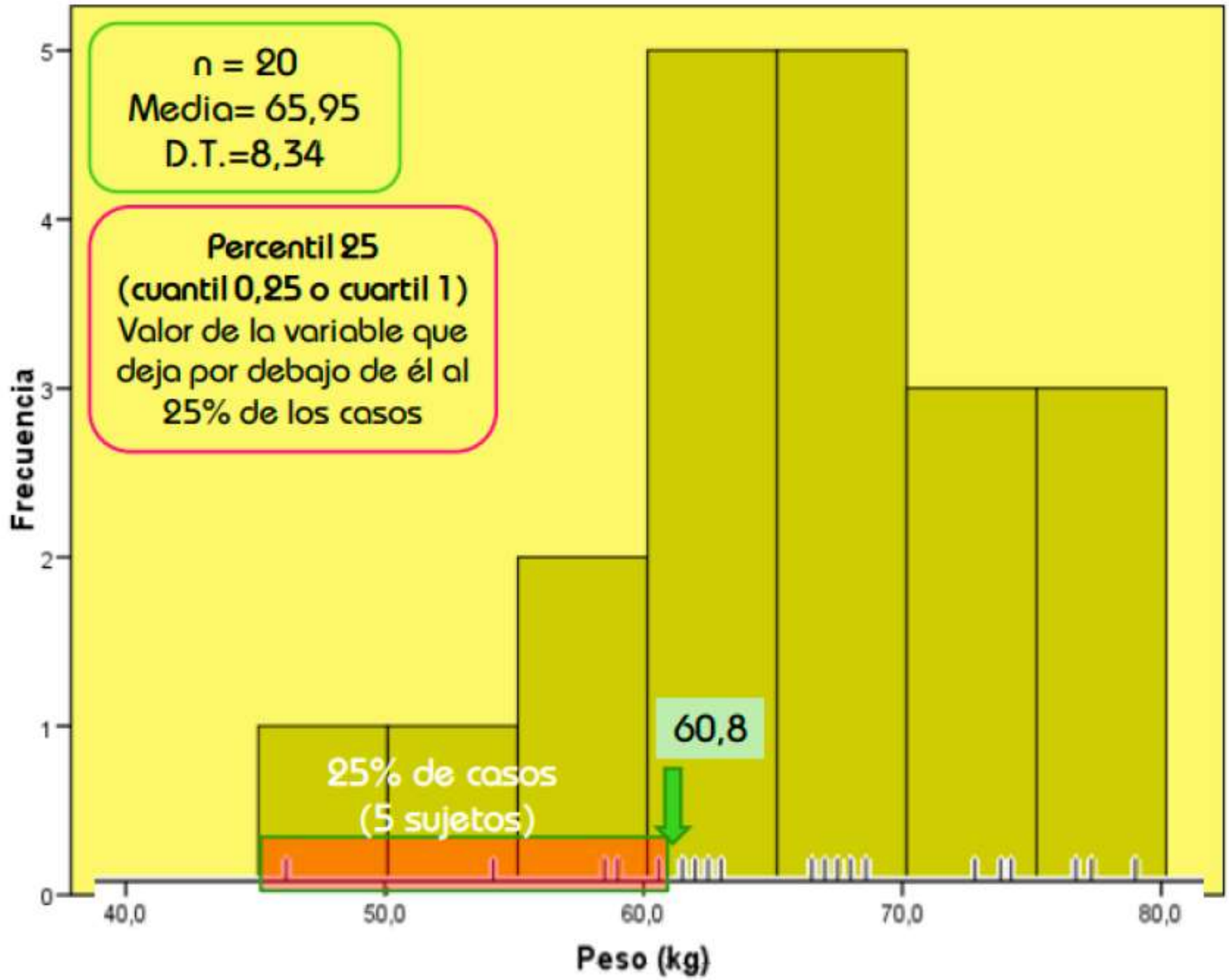
Cuartiles (de 1 a 3)

- Los **cuartiles** dividen a la muestra en cuatro grupos iguales.
- Existen tres cuartiles:
  - Q1 = percentil 25 = cuantil 0,25 (= α)
  - Q2 = percentil 50 = cuantil 0,50 (= α) = decil 5 = mediana
  - Q3 = percentil 75 = cuantil 0,75 (= α)

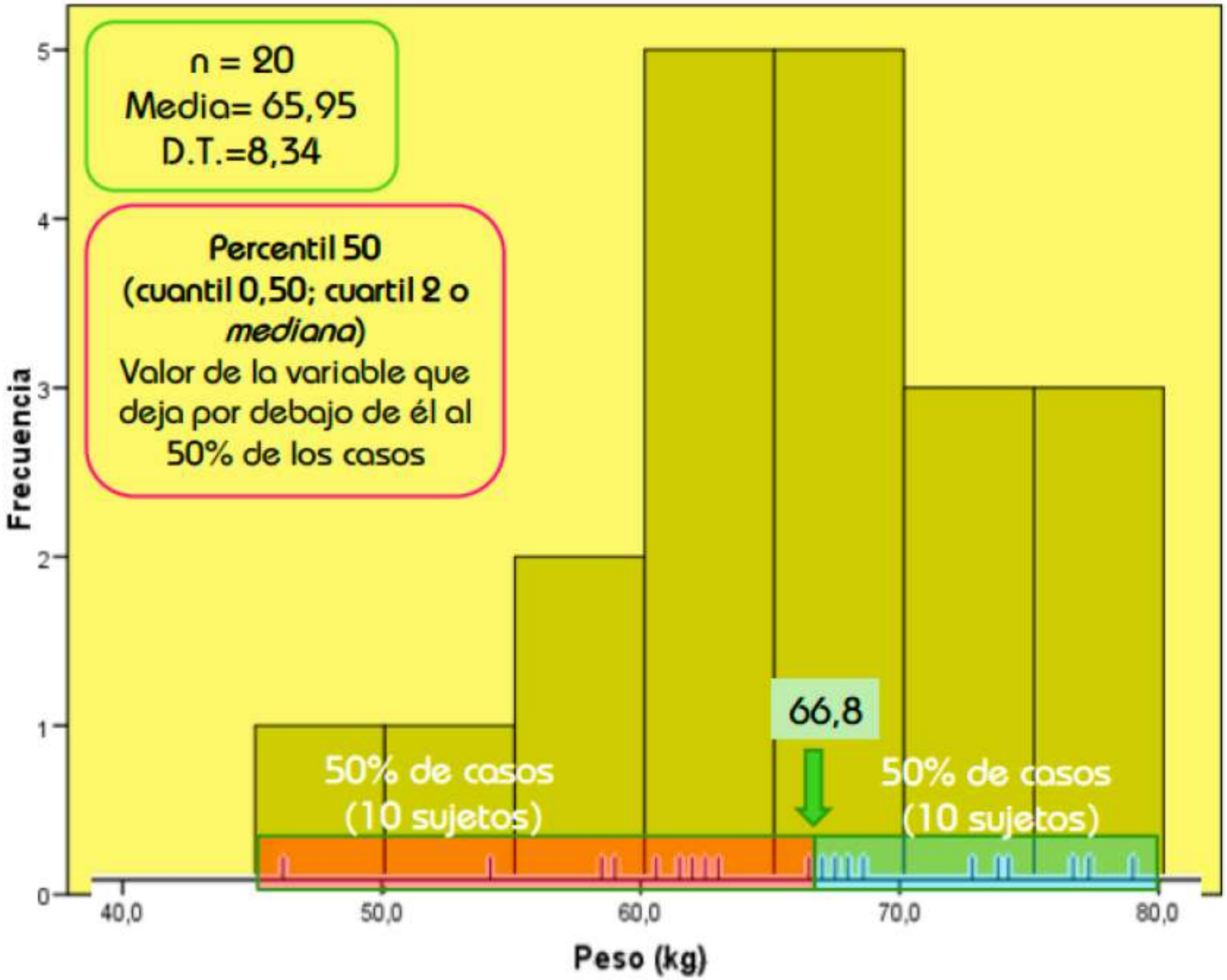
4. Medidas de posición



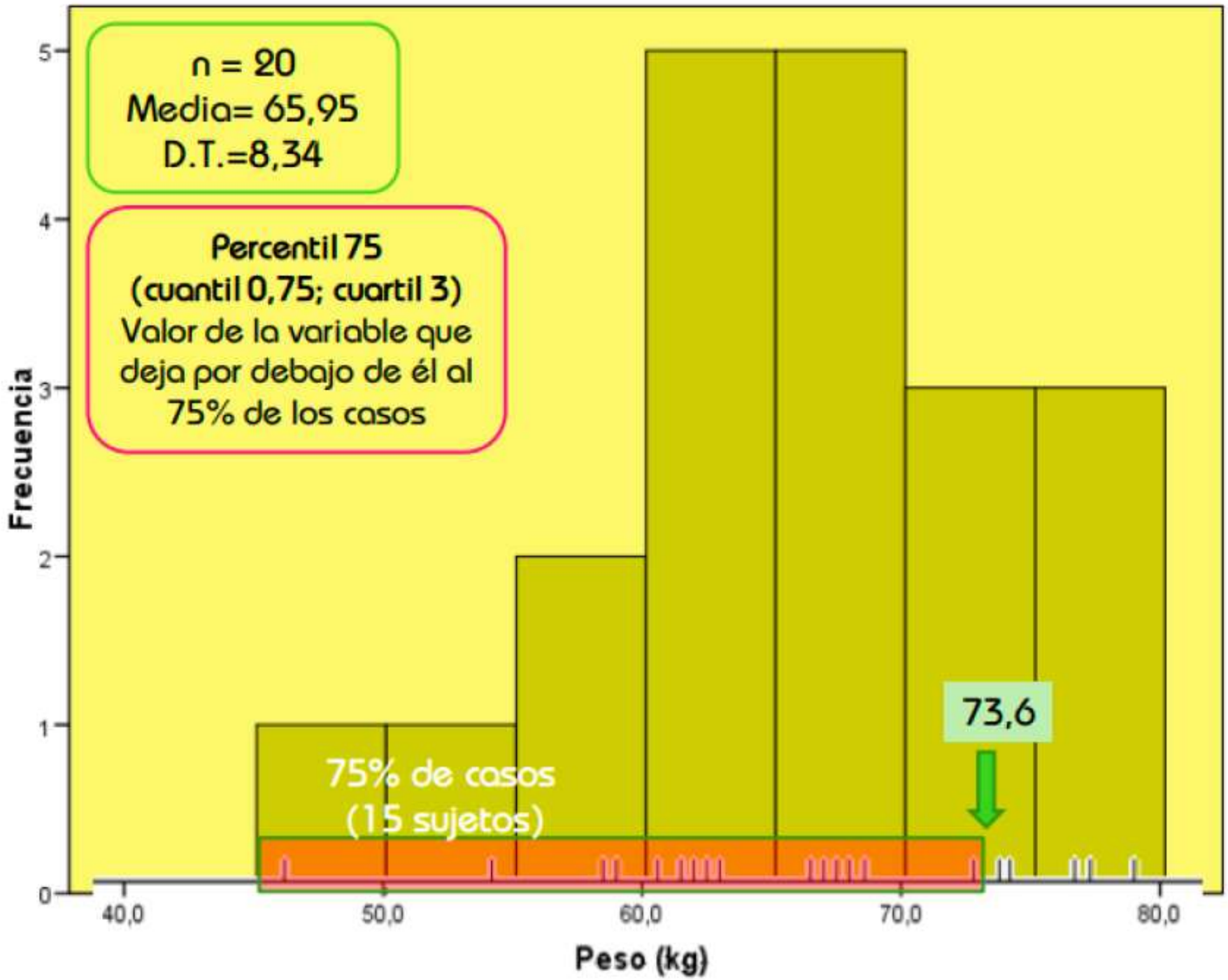
4. Medidas de posición



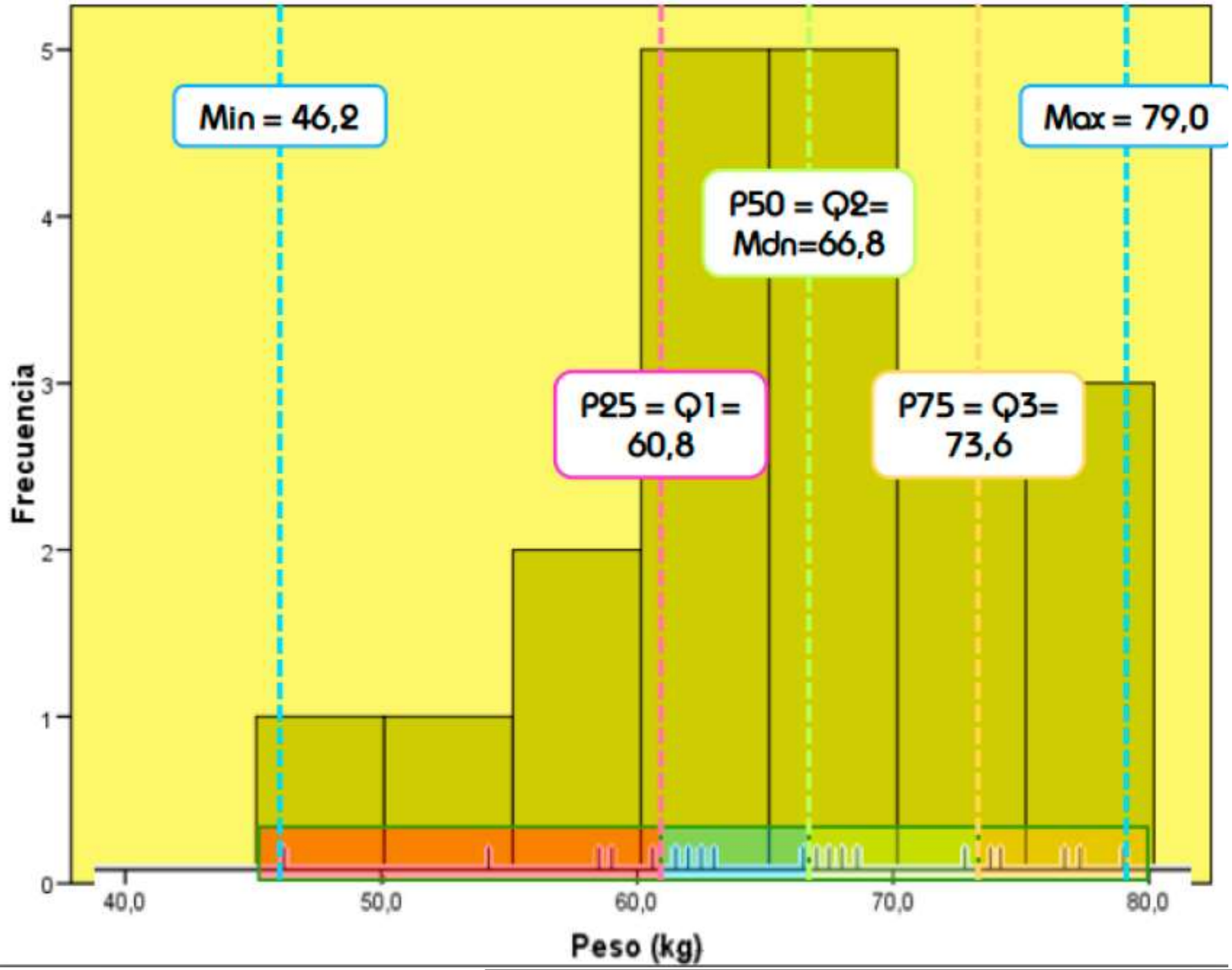
4. Medidas de posición



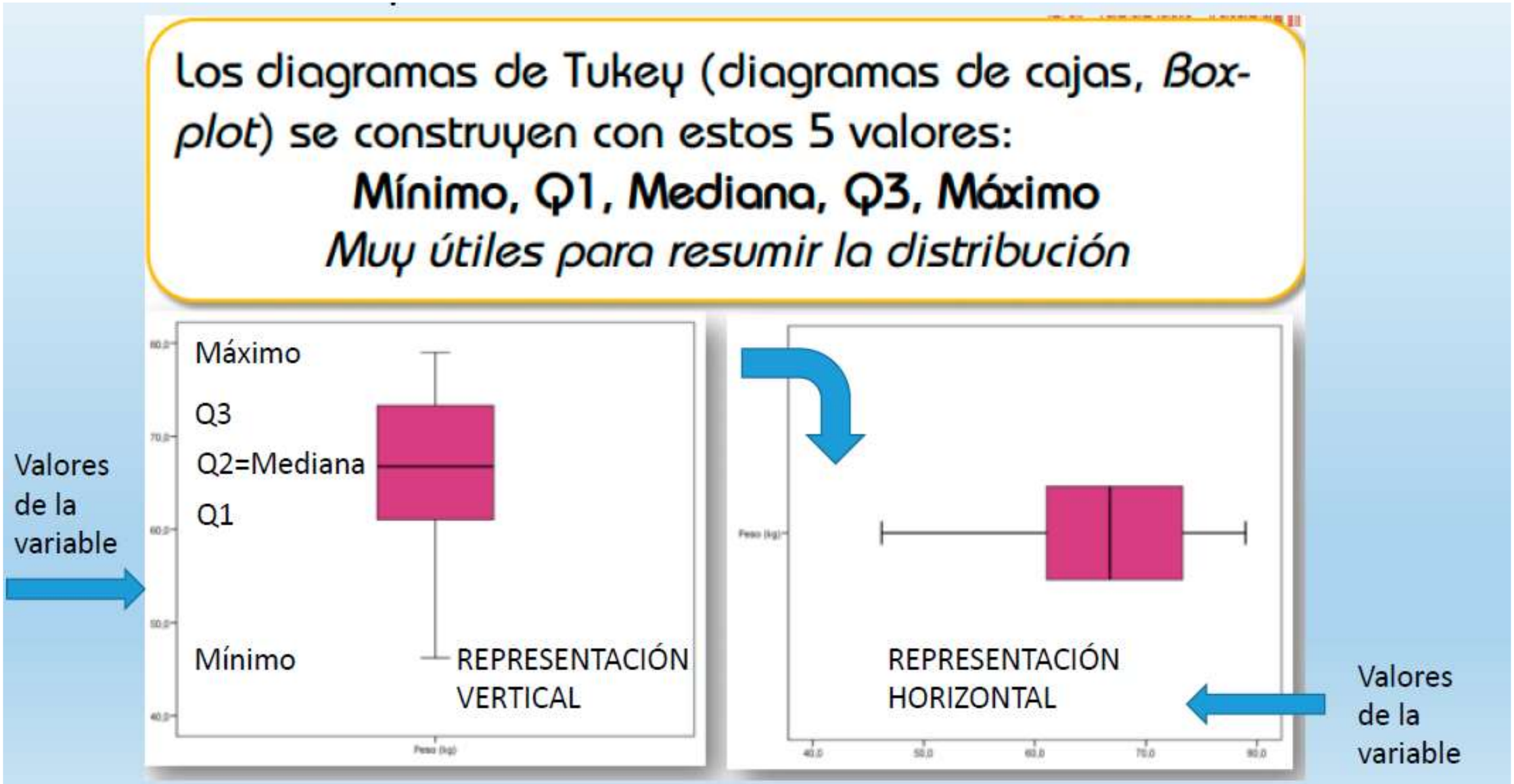
4. Medidas de posición



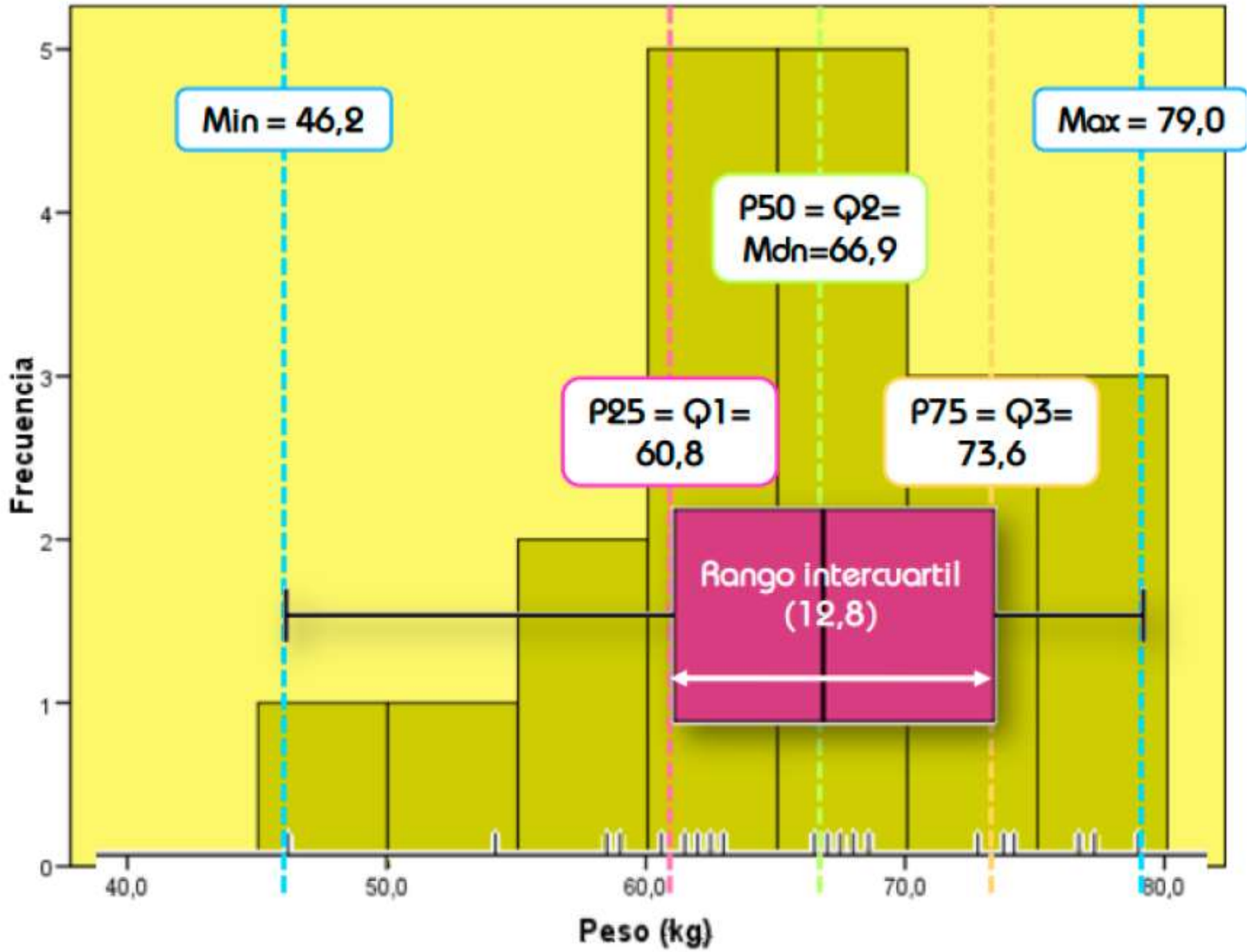
4. Medidas de posición



4. Medidas de posición



4. Medidas de posición



4. Medidas de posición

¿Cómo sacar el percentil 90?
1. Hallar el porcentaje acumulado inmediatamente superior a 90%
(en este caso es el 90,9%)
2. Mirar a qué valor de la variable corresponde
(en este caso 8)

¿Cómo sacar el percentil 25?
1. Hallar el porcentaje acumulado inmediatamente superior a 25%
(en este caso 38,9%)
2. Mirar a qué valor de la variable corresponde
(en este caso 2)

¿Cómo sacar la mediana?
1. Porcentaje acumulado inmediatamente superior a 50% => 54,6%
2. Valor correspondiente => 3

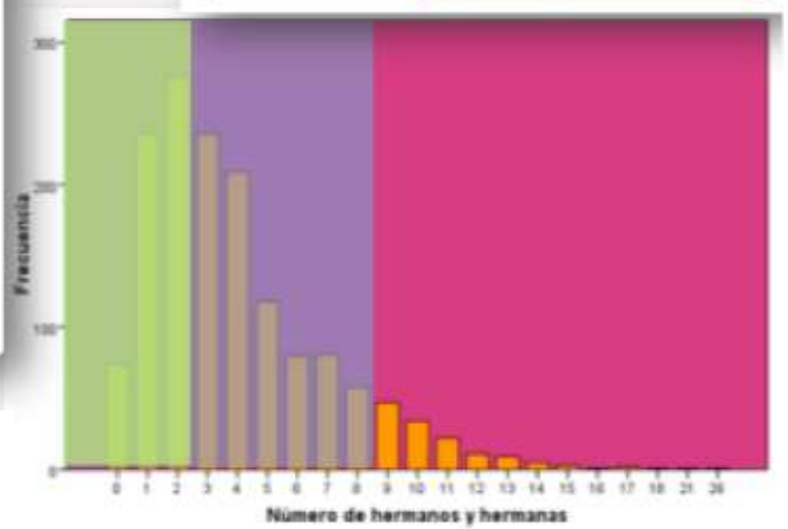
El percentil 25 tiene un valor de 2 (hijos)

El percentil 90 tiene un valor de 8 (hijos)

		Frecuencia	Porcentaje	Porcentaje acumulado	
Válidos	0	74	4,9	4,9	
	1	236	15,7	20,6	
	2	276	18,3	38,9	≥25
	3	236	15,7	54,6	
	4	209	13,9	68,5	
	5	118	7,8	76,3	
	6	80	5,3	81,7	
	7	81	5,4	87,0	
	8	58	3,9	90,9	≥90
	9	47	3,1	94,0	
	10	34	2,3	96,3	
	11	22	1,5	97,7	
	12	11	,7	98,5	
	13	9	,6	99,1	
	14	5	,3	99,4	
	15	3	,2	99,6	
	16	1	,1	99,7	
	17	2	,1	99,8	
	18	1	,1	99,9	
	21	1	,1	99,9	
	26	1	,1	100,0	
	Total	1505	100,0		

Número de hermanos y hermanas

N	Válidos	1505
Media		3,93
Mediana		3,00
Mínimo		0
Máximo		26
Percentiles	10	1
	20	1
	25	2
	30	2
	40	3
	50	3
	60	4
	70	5
	75	5
	80	6
	90	8



5. Medidas de dispersión

- Información de la variabilidad (distancia) que presentan los casos respecto a un valor concreto.

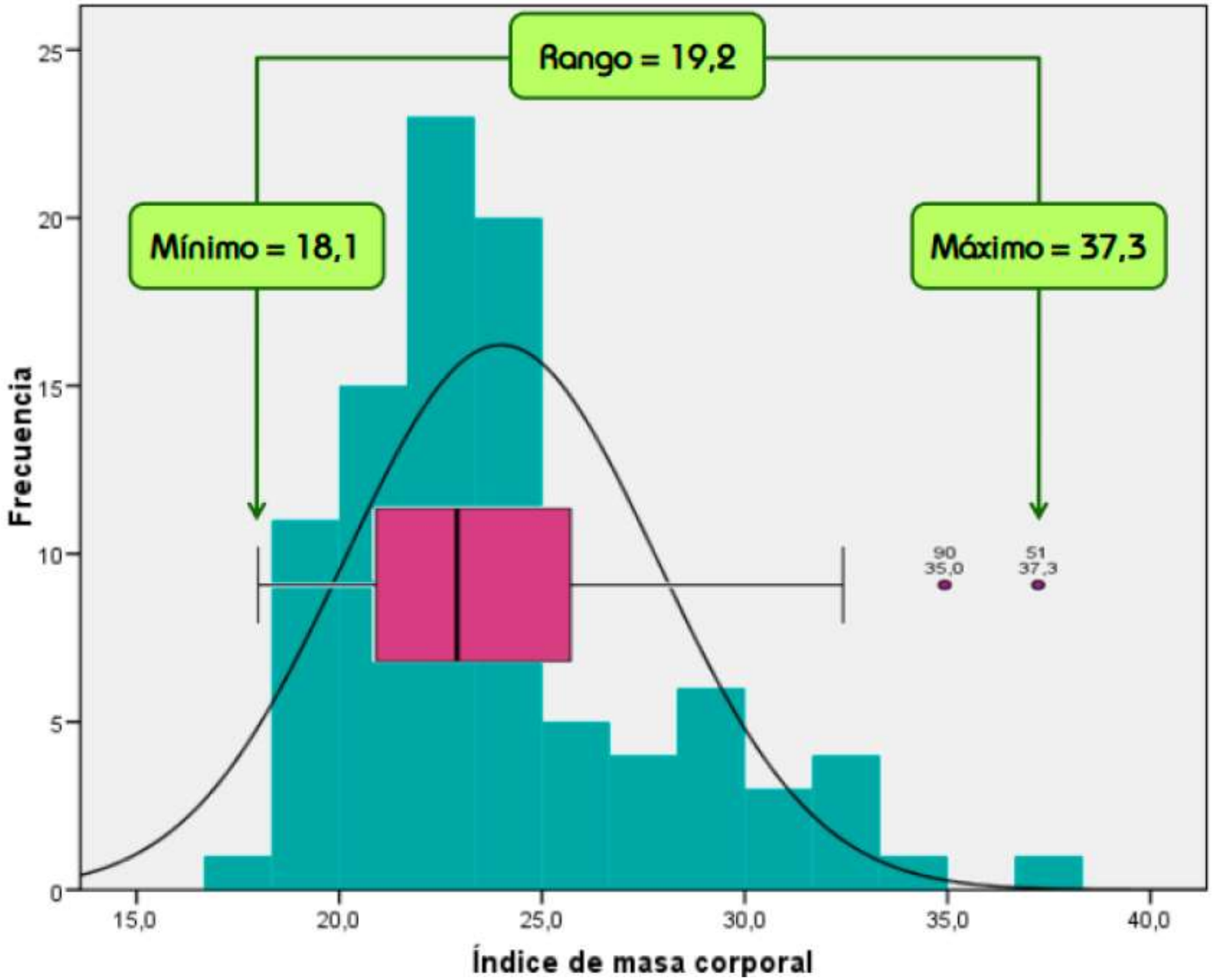
Rango	• La distancia (diferencia) entre las dos observaciones más extremas (máximo y mínimo)
Rango intercuartílico	• La distancia (diferencia) entre el tercer y el primer cuartil
Varianza (desv. típica)	• El promedio de las desviaciones, al cuadrado, de cada una de las observaciones respecto de la media
Coeficiente de variación	• Índice que relaciona la dispersión con la media, en %, para comparar el grado de variabilidad de distintas variables

5. Medidas de dispersión

- Rango
- Rango intercuartílico
- Varianza / D.E.
- Coeficiente de variación

- La distancia (diferencia) entre las dos observaciones más extremas (máximo y mínimo)
- VENTAJAS: Es de cálculo sencillo. No está de más calcularla para detectar casos extremos
- DESVENTAJAS: Sólo implica a dos valores de la variable, el mayor y el menor. Es muy sensible a los casos extremos

5. Medidas de dispersión

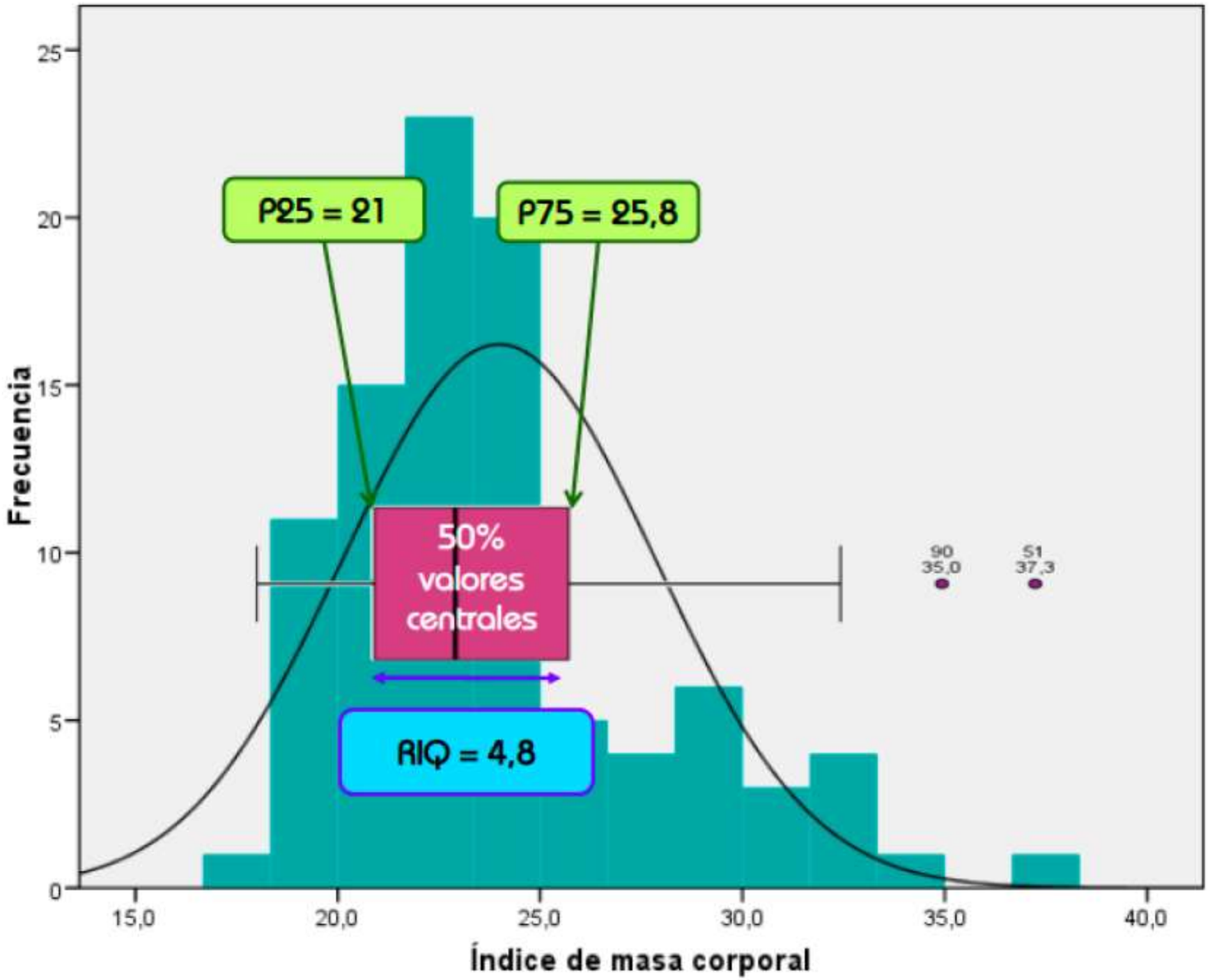


5. Medidas de dispersión

- Rango
- Rango intercuartílico
- Varianza / D.E.
- Coeficiente de variación

- Distancia (diferencia) entre el tercer y primer cuartil.
- VENTAJAS: Es de cálculo sencillo. Excluye 25% de los casos más grandes y el 25% de los casos más pequeños y por tanto utiliza en su cálculo el 50% de las observaciones centrales. No es afectada por casos extremos.
- DESVENTAJAS: Sólo participan en su cálculo el valor del Cuartil 1 (P25) y el del Cuartil 3 (P75%).

5. Medidas de dispersión



5. Medidas de dispersión

- Rango
- Rango intercuartílico
- Varianza / D.E.
- Coeficiente de variación

- **Varianza** (S^2, σ^2) es el promedio de las desviaciones, al cuadrado, de cada una de las observaciones respecto de la media
- VENTAJAS: Participa en los test estadísticos que se aplican si la distribución se parece a una normal (pruebas paramétricas).
- DESVENTAJAS:
  - Sensible a los valores extremos alejados de la media.
  - Está en unidades "al cuadrado" por lo que su interpretación se complica (por ello no se suele usar para interpretación)

5. Medidas de dispersión

- Rango
- Rango intercuartílico
- Varianza / D.E.
- Coeficiente de variación

- **Desviación típica** o **desviación estándar** (S, σ) es la raíz cuadrada de la varianza: $S = \sqrt{S^2}, \sigma = \sqrt{\sigma^2}$
- VENTAJAS: Se encuentra en las mismas unidades que la variable y se interpreta más fácilmente.
- DESVENTAJAS: No es una medida de dispersión adecuada cuando existan valores extremos (los extremos pueden afectarle demasiado)
- Se llama también **D.E**. = Desviación Estándar o **D.T.** = Desviación Típica

5. Medidas de dispersión

Edad (años)	21	32	15	59	60	61	64	60	71	80

Media

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{21+32+15+\ldots+80}{10} = 52{,}3 \text{ años}$$

Varianza

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

$$S^2 = \frac{(21-52{,}3)^2 + (32-52{,}3)^2 + \ldots + (80-52{,}3)^2}{10} = 427{,}61 \text{ (años)}^2$$

Desviación típica

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

$$S = \sqrt{\frac{(21-52{,}3)^2 + (32-52{,}3)^2 + \ldots + (80-52{,}3)^2}{10}} = \sqrt{427{,}61} = 21{,}79 \text{ años}$$

5. Medidas de dispersión

VARIANZA S^2 :expresión general y expresión reducida

$$S^2 = \sum_1^r (x_i - \bar{x})^2 f_i = \sum_1^r (x_i^2 + \bar{x}^2 - 2x_i\bar{x}) f_i = \sum_1^r x_i^2 f_i + \sum_1^r \bar{x}^2 f_i - 2\sum_1^r x_i\bar{x} f_i =$$

$$= \sum_1^r x_i^2 f_i + \bar{x}^2 \sum_1^r f_i - 2\bar{x} \sum_1^r x_i f = \sum_1^r x_i^2 f_i + \bar{x}^2 - 2\bar{x}^2 = \sum_1^r x_i^2 f_i - \bar{x}^2$$

Mide la dispersión en torno a la media. Qué tan cerca o lejos se encuentran, en promedio, todos los valores de la variable de su media.
Cuanto menor sea su valor más representa la media al conjunto de datos

5. Medidas de dispersión

Cambio de variable

$$Y = a + bX \Rightarrow S^2(y) = b^2 S^2(X)$$

$$S^2(y) = \sum_1^r (y_i - \bar{y})^2 f_i = \sum_1^r [(a + bx_i) - (a + b\bar{x})]^2 f_i =$$

$$= \sum_1^r [(bx_i - b\bar{x})]^2 f_i = b^2 \sum_1^r (x_i - \bar{x})^2 f_i = b^2 S^2(X)$$

5. Medidas de dispersión

- Rango
- Rango intercuartílico
- Varianza (desv. típica)
- Coeficiente de variación

- CV = Es un índice adimensional (sin unidades, en %) que relaciona la dispersión y la media. Sirve para comparar el grado de variabilidad de dos variables procedentes de una muestra diferente o que están medidas en unidades diferentes

$$CV = \frac{S}{\bar{x}} \cdot 100$$

5. Medidas de dispersión

$$CV = \frac{S}{\bar{x}} \cdot 100$$

PESO: Media =69,6 Kg; D.T.= 10,44 Kg

CV_{peso}= (10,44/69,6)x 100 ≈ **15%**

TAS: Media = 166 mmHg; D.T.= 21,3 mmHg

CV_{TAS}= (21,3/166)x100 ≈ **13%**

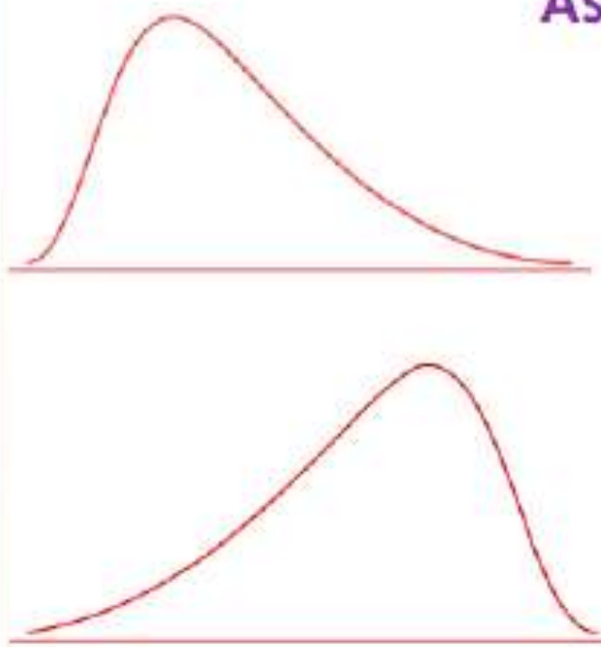
Cuidado porque no debe utilizarse cuando la variable pueda tomar valores negativos

6. Medidas de forma

Medidas de forma

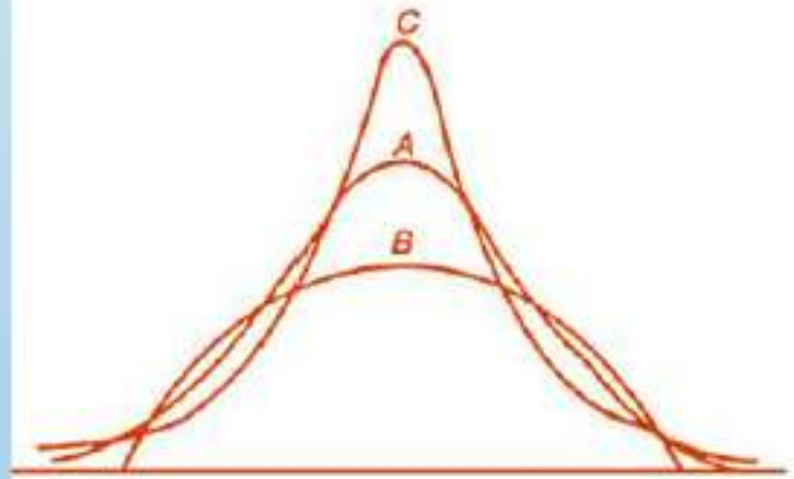
Asimetría (skewness)
Apuntamiento (curtosis)

Asimetría



Simetría respecto a la normal

Apuntamiento



"Aplastamiento" frente a la normal

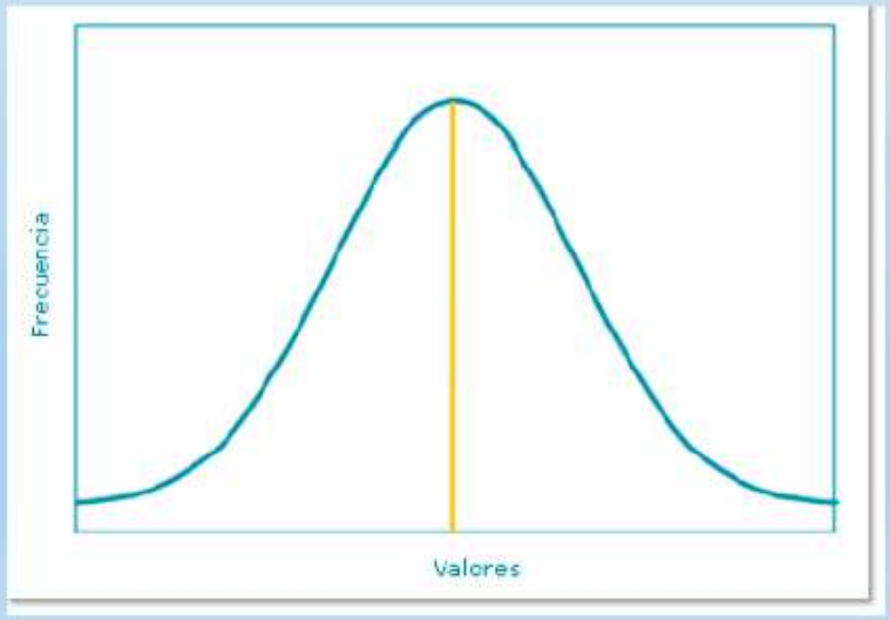
Informan sobre el parecido de la distribución de los datos respecto a la distribución normal

6. Medidas de forma

Asimetría (skewness)

Apuntamiento (curtosis)

Asimetría: Si la distribución es simétrica, la distribución es dividida por la media en dos mitades iguales (asimetría = 0)

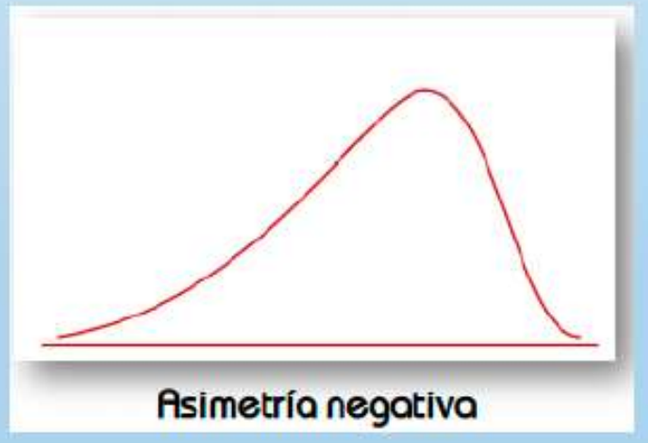


6. Medidas de forma

Asimetría (skewness)

Apuntamiento (curtosis)

Cuando la distribución tiene una cola hacia la derecha presenta una **asimetría positiva**; cuando la cola aparece a la izquierda la **asimetría** es **negativa**.





6. Medidas de forma

Asimetría (skewness)

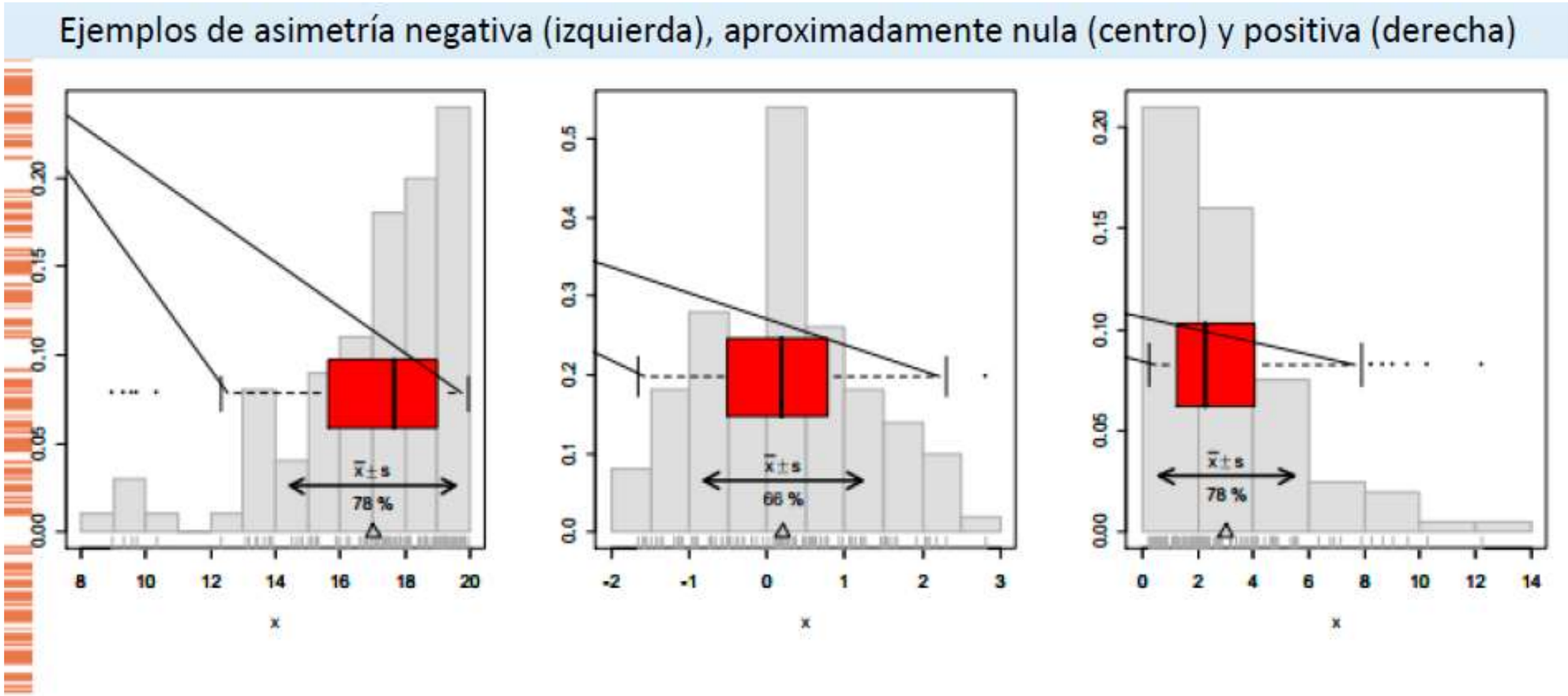
Existen diferentes formas para "cuantificar" la asimetría.
1) Basados en estadísticos de tendencia central (es el coeficiente de asimetría que habitualmente proporcionan los programas de estadística)
2) Basados en las diferencias de los cuartiles respecto de la mediana.

Apuntamiento (curtosis)

$$A = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n}$$

$$A = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$$

6. Medidas de forma



$$A = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{n}$$

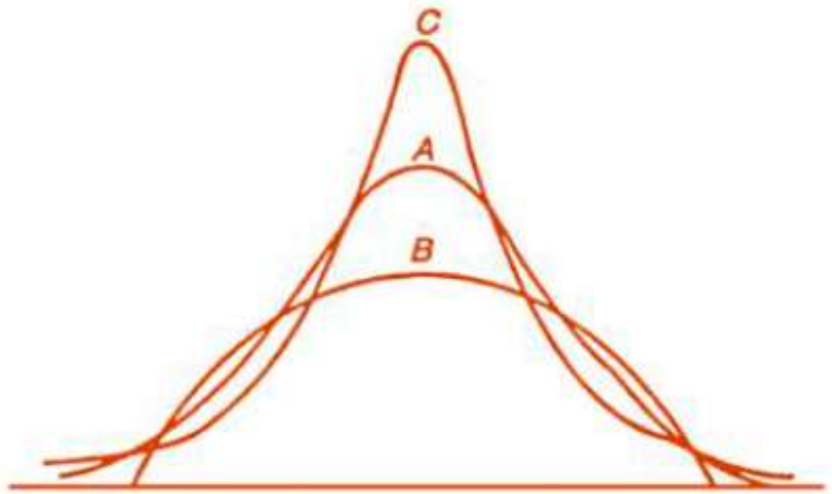
$$A = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$$

6. Medidas de forma

Asimetría (skewness)

Apuntamiento (curtosis)

- **Curtosis**: se refiere a la altura de la distribución respecto a la distribución normal



A= Mesocúrtica (como la normal)

B= Platicúrtica (aplanada)

C=Leptocúrtica (apuntada)

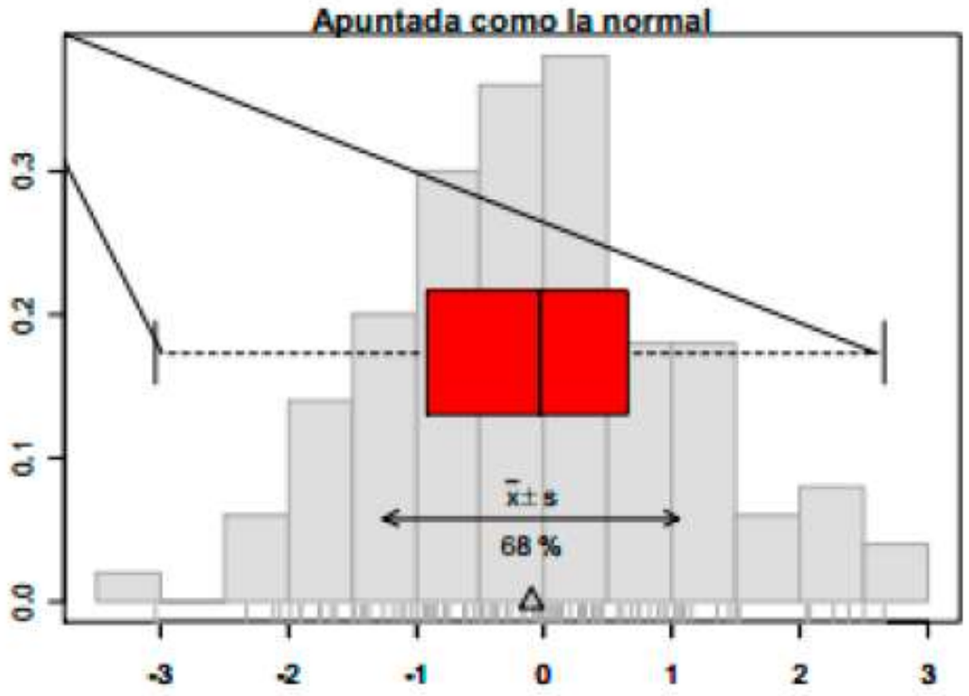
6. Medidas de forma

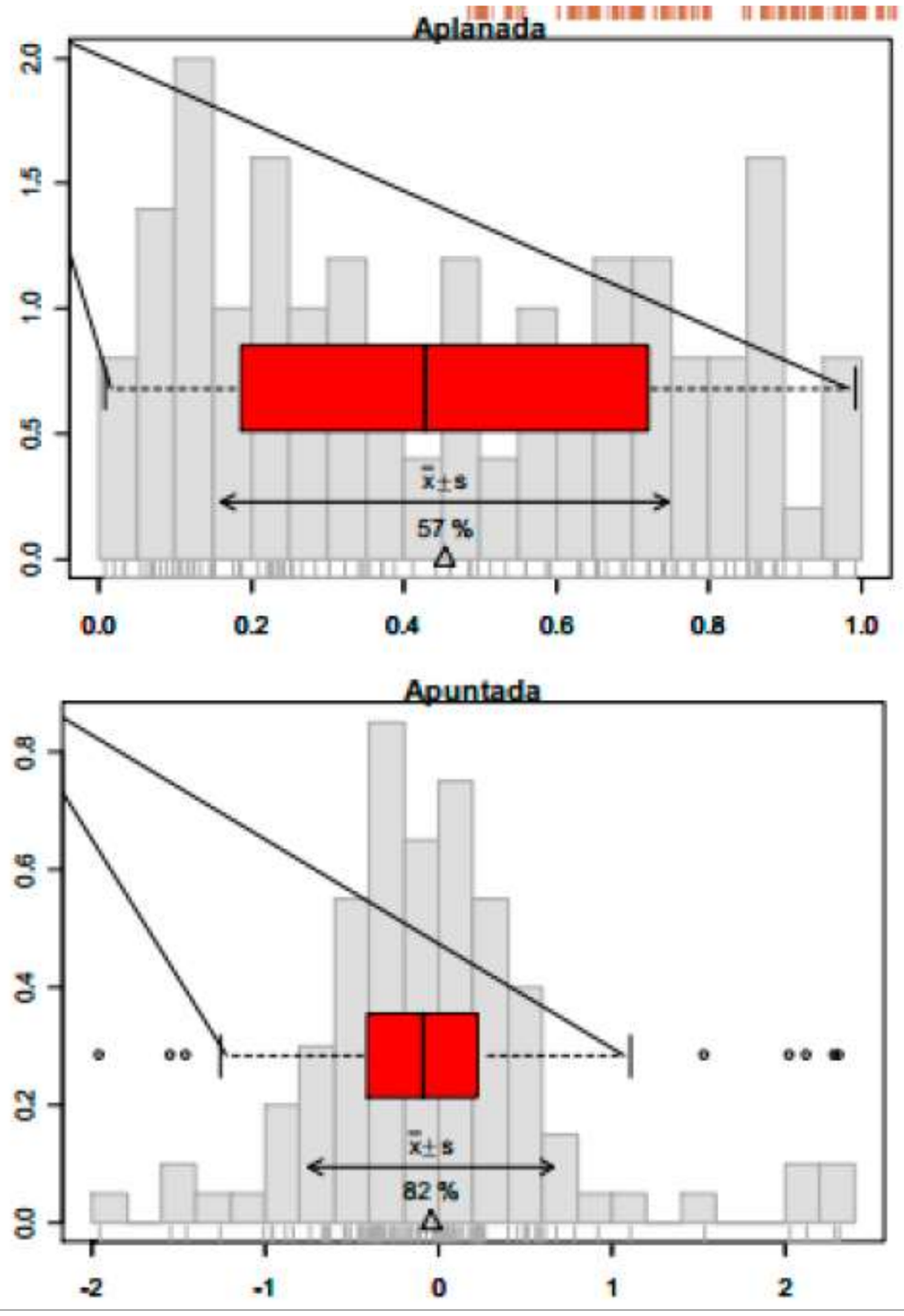
Asimetría (skewness)

Apuntamiento (curtosis)

- **Curtosis**: se refiere a la altura de la distribución
- El **coeficiente de aplastamiento de Fisher** (curtosis) es un índice que nos indica si la distribución es más o menos apuntada que la distribución normal.
  - Coeficiente = 0 distribución mesocúrtica (como la normal)
  - Coeficiente > 0 distribución leptocúrtica (apuntada)
  - Coeficiente es < 0, distribución platicúrtica (aplanada)

6. Medidas de forma



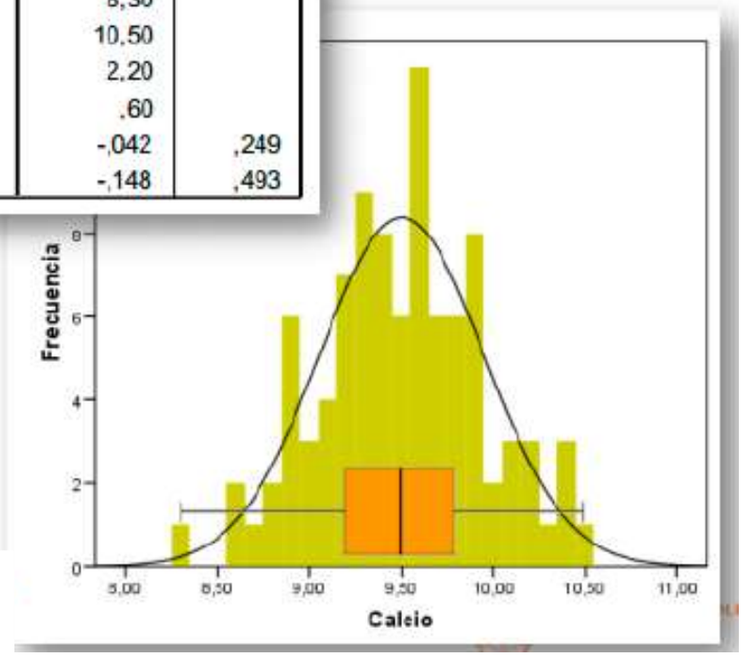


6. Medidas de forma

- Vamos a ver la simetría y la curtosis con un ejemplo de programa de estadística
- Esta variable tiene:
  - Tiene gran simetría porque el valor de asimetría es casi 0
  - Curtosis prácticamente nula, ligeramente negativa (forma mesocúrtica)

Descriptivos n=94

			Estadístico	Error típ.
Calcio	Media		9,50	,04613
	Intervalo de confianza para la media al 95%	Límite inferior	9,41	
		Límite superior	9,59	
	Media recortada al 5%		9,50	
	Mediana		9,50	
	Varianza		,200	
	Desv. típ.		,447	
	Mínimo		8,30	
	Máximo		10,50	
	Rango		2,20	
	Amplitud intercuartil		,60	
	Asimetría		-,042	,249
	Curtosis		-,148	,493



Estadística Aplicada al Marketing

T.3. Análisis Descriptivo Bidimensional

Juan Miguel Navarro Ruiz

Grado en Marketing y Dirección Comercial

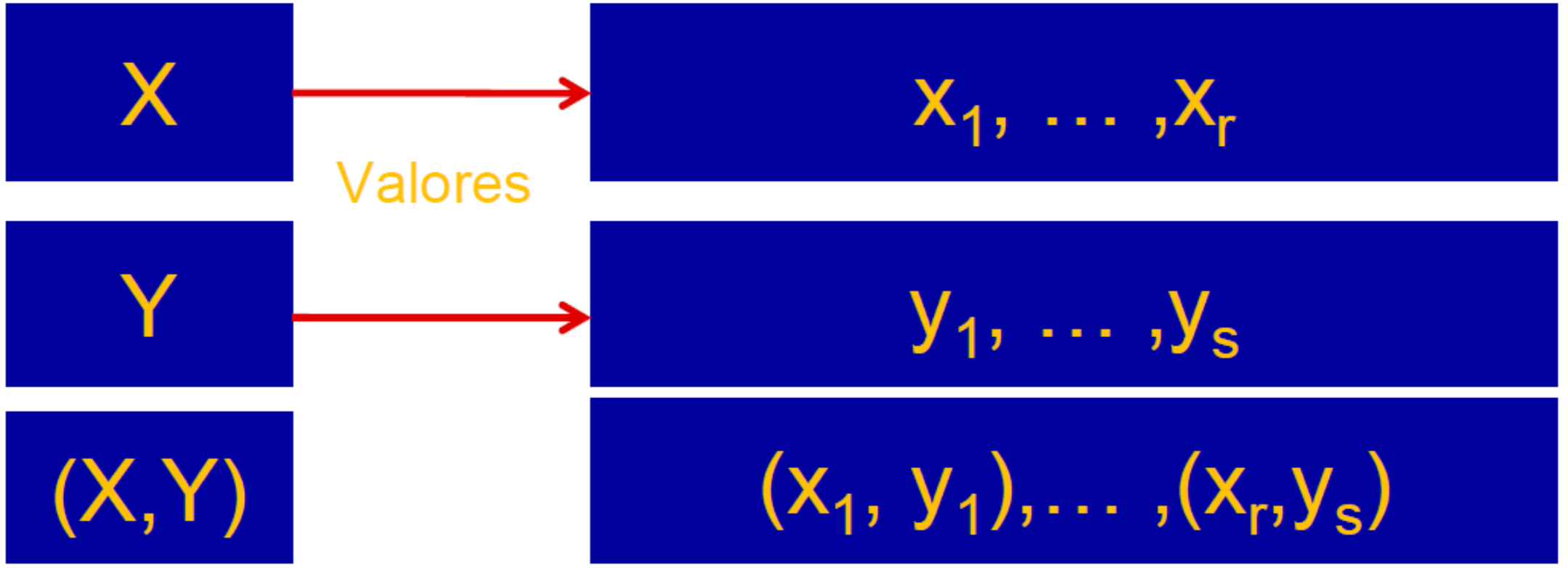
1. Variables Bidimensionales

En este contexto el objetivo se centra en analizar la in formación correspondiente a dos características de cada individuo. Por lo tanto los datos con los que trabajaremos vendrán siempre apareados denominado X a la primera característica observada e Y a la segunda. En definitiva ahora hablaremos de la variable estadistica bidimensional (X,Y)

Edad y sexo
Altura y peso
Municipio de nacimiento y de residencia
Profesión de los padres y orientación profesional

1. Variables Bidimensionales

Como en el caso unidimensional, ahora las variables X e Y pueden ser cualitativas, cuantitativas, discretas o continuas, pudiendo presentarse cualquier combinación de ambas opciones. Se utilizará la notación siguiente



2. Distribuciones Bidimensionales

La ordenación *de datos se realiza ahora "contando" cuántas veces aparece cada pareja* (x_i, y_j) y a ese número, que indicaremos por n_{ij} lo llamaremos frecuencia conjunta (absoluta) correspondiente a dicha pareja.

La distribución conjunta de frecuencias queda ahora definida conociendo:

$$\{(x_i, y_j), n_{ij}\}, \; i=1,\ldots,r; \; j=1,\ldots,s$$

2. Distribuciones Bidimensionales

La ordenación y presentación de datos se realizará mediante una tabla bidimensional de frecuencias (absolutas), que es la representación de la distribución de frecuencias.

	Y_1	...	Y_j	...	Y_s
X_1	n_{11}		n_{1j}		n_{1s}
X_2	n_{21}		n_{2j}		n_{2s}
...	...	...	...	...	...
X_i	n_{i1}		n_{ij}		n_{is}
...	...	...	...	...	...
...	...	...	...	...	...
X_r	n_{r1}	...	n_{rj}	...	n_{rs}

2. Distribuciones Bidimensionales

- Una **tabla de contingencia** es una **tabla de frecuencias conjuntas** que combina las categorías de ambas variables.

Sexo	**ni**
Hombres	94
Mujeres	106
Total	200

Tabaquismo	**ni**
Fumadores	60
Exfumadores	13
No fumadores	127
Total	200

Esto no es todavía una tabla de contingencia porque no expresa las frecuencias de manera conjunta

2. Distribuciones Bidimensionales

Tabla de contingencia Sexo * Tabaquismo					
		Tabaquismo			
		Fumadores	*Exfumadores*	*No fumadores*	**Total**
Sexo	*Hombres*	18	7	69	94
	Mujeres	42	6	58	106
	Total	60	13	127	200

Frecuencias conjuntas: las frecuencias absolutas tomadas las dos variables conjuntamente.

Número de sujetos mujeres **Y** no fumadores

2. Distribuciones Bidimensionales

Igualmente se puede hablar de la tabla de frecuencias relativas, definidas ahora como:

$$f_{ij} = \frac{n_{ij}}{N} \quad \Rightarrow \quad N = \sum_{i=1}^{i=r} \sum_{j=1}^{j=s} n_{ij}$$

	Y_1	...	Y_j	...	Y_s
X_1	f_{11}		f_{1j}		f_{1s}
X_2	f_{21}		f_{2j}		f_{2s}
...	...	...	...	...	...
X_i	f_{i1}		f_{ij}		f_{is}
...	...	...	...	...	...
...	...	...	...	...	...
X_r	f_{r1}	...	f_{rj}	...	f_{rs}

2. Distribuciones Bidimensionales

Porcentaje de hombres fumadores entre los que todos los sujetos del estudio

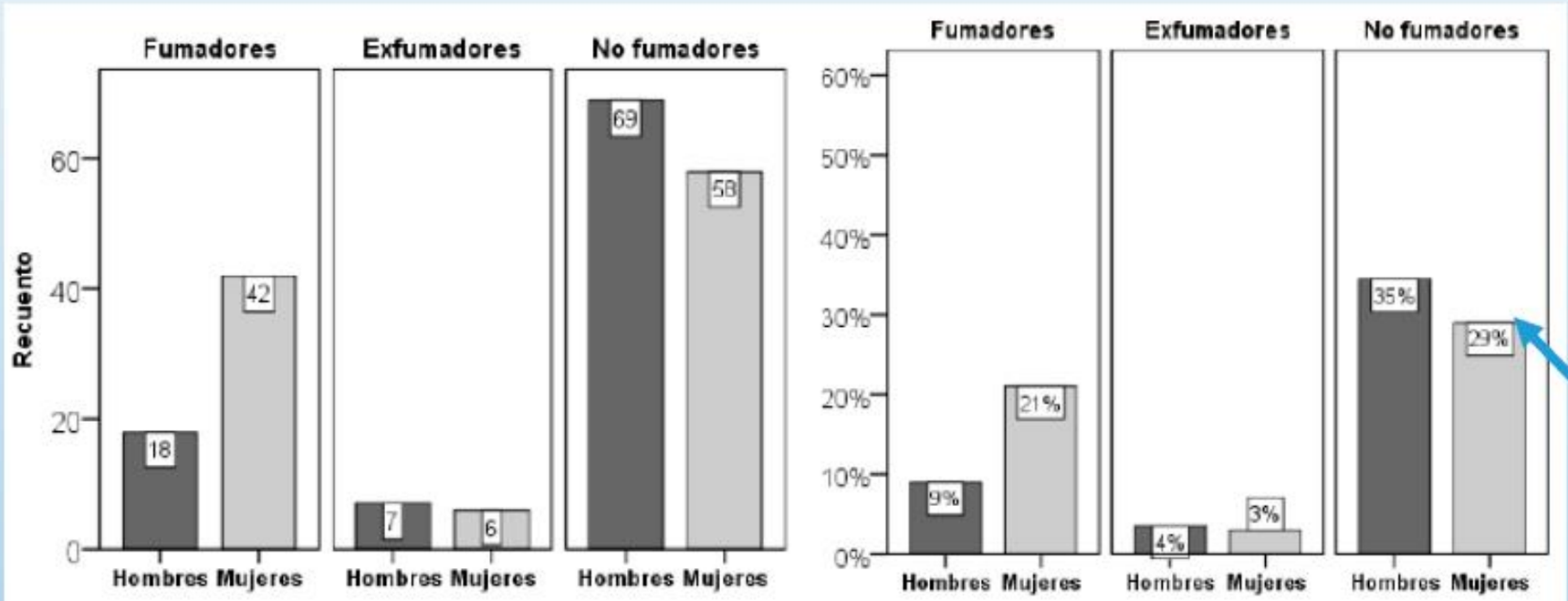
Sexo		Tabaquismo			
		Fumadores	*Exfumadores*	*No fumadores*	**Total**
Hombres	Recuento	18	7	69	94
	% del total	9,0%	3,5%	34,5%	47,0%
Mujeres	Recuento	42	6	58	106
	% del total	21,0%	3,0%	29,0%	53,0%
Total	Recuento	60	13	127	200
	% del total	30,0%	6,5%	63,5%	100,0%

Los porcentajes del total indican el porcentaje de cada frecuencia conjunta respecto al número total de casos de la tabla.

Cada porcentaje es la **PROBABILIDAD CONJUNTA** de que sea hombre y fumador

2. Distribuciones Bidimensionales

Frecuencias absolutas



En este caso no se representan gráficos agrupados, sino segmentados y los porcentajes se refieren al total de sujetos.

Porcentajes referidos al total

2. Distribuciones Bidimensionales

Se pueden utilizar los conceptos de frecuencias acumuladas (absolutas, relativas, porcentajes)

N_{hk}= Número de observaciones para las que $X \leq x_h$, $Y \leq y_k$

F_{hk}= Proporción de observaciones para las que $X \leq x_h$, $Y \leq y_k$

P_{hk}= %de observaciones para las que $X \leq x_h$, $Y \leq y_k$

$$N_{hk} = \sum_{i=1}^{i=h} \sum_{j=1}^{j=k} n_{ij} \qquad F_{hk} = \sum_{i=1}^{i=h} \sum_{j=1}^{j=k} f_{ij} \qquad P_{hk} = \sum_{i=1}^{i=h} \sum_{j=1}^{j=k} p_{ij}$$

3. Distribuciones Marginales y Condicionadas

A partir del comportamiento conjunto de (X,Y) se puede conocer el **comportamiento individual** de cada variable. Es decir, sus distribuciones de frecuencias que, ahora, se denominan distribuciones marginales

La **distribución marginal** de X estará formada por los valores que tome la variable (ya conocidos) junto con la frecuencia (ahora llamada marginal) con la que aparece cada valor.

$$n_{i.} = n_{i1} + n_{i2} + \ldots + n_{ij} + \ldots + n_{is} = \sum_{j=1}^{j=s} n_{ij}$$

Se pueden construir de forma análoga las frecuencias relativas y los porcentajes marginales

3. Distribuciones Marginales y Condicionadas

Dispondremos entonces de la **distribución marginal de X**, definida como $\{(x_i,n_{i.})\}$, i=1,…,r. Y, construida por analogía, la **distribución marginal de Y** como $\{(y_j,n_{.j})\}$, j=1,…,s.

	Y_1	…	Y_j	…	Y_s	$n_{i.}$
X_1	n_{11}		n_{1j}		n_{1s}	$n_{1.}$
X_2	n_{21}		n_{2j}		n_{2s}	$n_{2.}$
…	…	…	…	…	…	
X_i	n_{i1}		n_{ij}		n_{is}	$n_{i.}$
…	…	…	…	…	…	
…	…	…	…	…	…	
X_r	n_{r1}	…	n_{rj}	…	n_{rs}	$n_{r.}$
$n_{.j}$	$n_{.1}$		$n_{.j}$		$n_{.r}$	N

3. Distribuciones Marginales y Condicionadas

Marginales

Tabla de contingencia Sexo * Tabaquismo

		Tabaquismo			
		Fumadores	*Exfumadores*	*No fumadores*	**Total**
Sexo	*Hombres*	18	7	69	94
	Mujeres	42	6	58	106
	Total	60	13	127	200

Frecuencias marginales: son los totales de cada fila o de cada columna. Equivalen a las tablas de frecuencias individuales

Marginales para la variable Tabaquismo

Marginales para la variable Sexo

3. Distribuciones Marginales y Condicionadas

Cada distribución marginal tiene su propio comportamiento y en cada una de ellas se pueden obtener todos los indicadores puestos de manifiesto en el análisis de distribuciones unidimensionales

$$\bar{x} = \sum_{i=1}^{i=r}\sum_{j=1}^{j=s} x_i f_{ij} = \sum_{i=1}^{i=r} x_i \sum_{j=1}^{j=s} f_{ij} = \sum_{i=1}^{i=r} x_i f_{i.}$$

$$S_x^2 = \sum_{i=1}^{i=r}\sum_{j=1}^{j=s} (x_i - \bar{x})^2 f_{ij} = \sum_{i=1}^{i=r} (x_i - \bar{x})^2 \sum_{j=1}^{j=s} f_{ij} = \sum_{i=1}^{i=r} (x_i - \bar{x})^2 f_{i.}$$

3. Distribuciones Marginales y Condicionadas

Un paso más en cuanto a obtener distribuciones unidimensionales es el de intentar conocer cómo se comporta la variable X bajo la condición de Y tome un determinado valor. Es el concepto de **distribución condicionada** que se representa por $(X/Y=y_j)$

Para ello hay que situarse en el hecho de que nos hemos de restringir a los individuos para los cuales $Y=y_j$ y, sobre ellos centrar el análisis. Esto supone que contamos con un total de $n_{.j}$ observaciones, referencia a tener en cuenta.

3. Distribuciones Marginales y Condicionadas

En este caso la distribución de frecuencias vendrá dada por los valores que puede tomar la variable X junto con la frecuencia con las que los toma n_{ij} con j fijo y, acompañada del dato relevante de que el total de observaciones es $n_{.j}$

Con el fin de tener esta situación en cuenta se introduce el concepto de frecuencia (relativa) condicionada $f_{i/j}$ definida como:

$$f_{i/j} = \frac{n_{ij}}{n_{.j}}$$

Con ello la distribución de frecuencias vendrá dada, ahora, por el conjunto $\{(x_i,y_j), f_{i/j}\}$, a la que se le pueden aplicar todos los conceptos ya conocidos de las distribuciones unidimensionales

3. Distribuciones Marginales y Condicionadas

El concepto de distribución condicionada puede generalizarse planteando cuál es el comportamiento de una variable, por ejemplo la X, cuando la otra Y cumple una cierta condición.

En este caso tendríamos la distribución (X/Y∈A), con lo cual los valores de la variable seguirán siendo los x_i en tanto que las frecuencias hay que construirlas como:

$$f_{i/Y\in A} = \frac{\text{Nº de observaciones con } X=x_i,\ Y\in A}{\text{Nº de observaciones con, } Y\in A}$$

3. Distribuciones Marginales y Condicionadas

Condicionadas

Porcentaje de fumadores respecto al total de hombres

Sexo		Tabaquismo			
		Fumadores	*Exfumadores*	*No fumadores*	**Total**
Hombres	Recuento	18	7	69	94
	% dentro de Sexo	19,1%	7,4%	73,4%	100,0%
Mujeres	Recuento	42	6	58	106
	% dentro de Sexo	39,6%	5,7%	54,7%	100,0%
Total	Recuento	60	13	127	200
	% dentro de Sexo	30,0%	6,5%	63,5%	100,0%

Los porcentajes por fila indican el porcentaje de cada frecuencia conjunta respecto a la frecuencia marginal de su fila. Es una **distribución condicionada.**

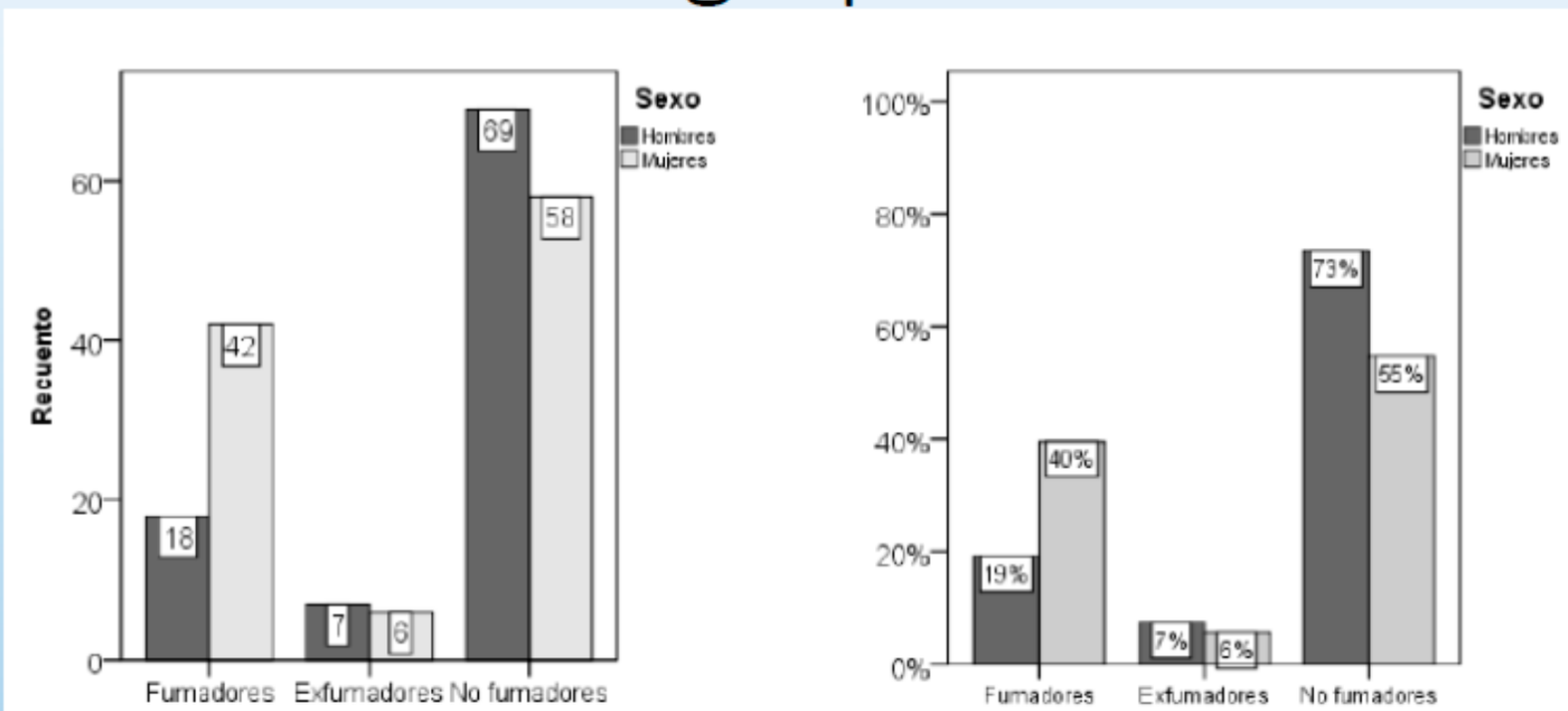
Porcentaje de exfumadores respecto al total de mujeres

Condicionada quiere decir: "La probabilidad de que sea fumador, **CUANDO SÉ que es** un hombre, es de 19.1%"
La variable tabaquismo **ESTÁ CONDICIONADA** a la variable sexo

3. Distribuciones Marginales y Condicionadas

Condicionadas

Gráficos de barras agrupadas



El gráfico de frecuencias absolutas y el de frecuencias relativas tienen un aspecto similar sólo porque el número de mujeres (106) es similar al de hombres (94)

Gráfico de barras para la variable sexo agrupadas por tabaquismo

3. Distribuciones Marginales y Condicionadas

Condicionadas

Porcentaje de hombres entre los que todos los fumadores

Sexo		Tabaquismo			
		Fumadores	*Exfumadores*	*No fumadores*	**Total**
Hombres	Recuento	18	7	69	94
	% dentro de Tab	30,0%	53,8%	54,3%	47,0%
Mujeres	Recuento	42	6	58	106
	% dentro de Tab	70,0%	46,2%	45,7%	53,0%
Total	Recuento	60	13	127	200
	% dentro de Tab	100,0%	100,0%	100,0%	100,0%

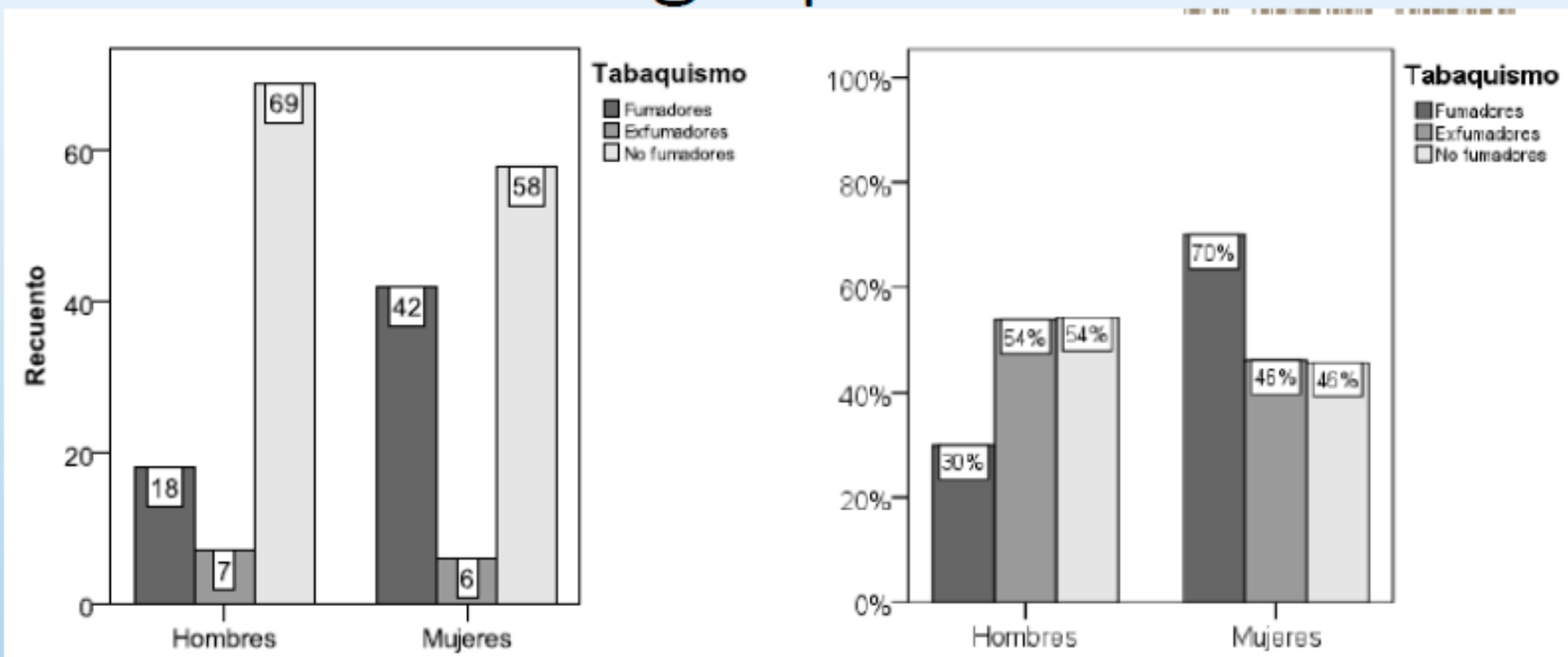
Los **porcentajes por columna** indican el porcentaje de cada frecuencia conjunta respecto a la frecuencia marginal de su columna. Es una **distribución condicionada.**

Ahora es la probabilidad de que sea hombre, **CUANDO SÉ que es** fumador
La variable sexo **ESTÁ CONDICIONADA** a la variable tabaquismo

3. Distribuciones Marginales y Condicionadas

Condicionadas

Gráficos de barras agrupadas



El gráfico de frecuencias absolutas y el de frecuencias relativas tienen un aspecto muy distinto porque el número de fumadores (60), exfumadores (13) y no fumadores (127) es diferente.

Gráfico de barras para la variable tabaquismo agrupadas por sexo

3. Distribuciones Marginales y Condicionadas

Tabla completa

Sexo		Tabaquismo			
		Fumadores	*Exfumadores*	*No fumadores*	**Total**
Hombres	Recuento	18	7	69	94
	% dentro de Sexo	19,1%	7,4%	73,4%	100,0%
	% dentro de Tab	30,0%	53,8%	54,3%	47,0%
	% del total	9,0%	3,5%	34,5%	47,0%
Mujeres	Recuento	42	6	58	106
	% dentro de Sexo	39,6%	5,7%	54,7%	100,0%
	% dentro de Tab	70,0%	46,2%	45,7%	53,0%
	% del total	21,0%	3,0%	29,0%	53,0%
Total	Recuento	60	13	127	200
	% dentro de Sexo	30,0%	6,5%	63,5%	100,0%
	% dentro de Tab	100,0%	100,0%	100,0%	100,0%
	% del total	30,0%	6,5%	63,5%	100,0%

4. Independencia Estadística

- La finalidad de las tablas de correlación (variables cuantitativas) o las tablas de contingencia (cualitativas) es averiguar si dos (o más) variables tienen o no que ver algo entre sí, si están o no relacionadas.

- Con dos variables categóricas sólo caben dos posibilidades: **independencia** o **relación**.
- Dos variables son **independientes** cuando el comportamiento de una de ellas no se ve alterado por la otra.

4. Independencia Estadística

Dado que estamos observando dos características de un mismo individuo cabe pensar que entre ambas puede existir alguna relación o, por el contrario que no exista. En este segundo caso se dirá que las variables que describen dichas características son **independientes**.

La cuestión de **cómo se detecta la independencia** puede plantearse de varias formas, siendo una de ellas el considerar que ese supuesto se tiene que reflejar en las distribuciones condicionadas.

4. Independencia Estadística

En concreto si la variable Y no influye en la X (son independientes), debemos esperar que el comportamiento de la distribución de frecuencias de $X/Y=y_j$ debe ser el mismo si la condición es otra cualquiera, referida a la variable Y. Esto nos llevaría a poder decir que en el caso de la independencia se cumpliría la igualdad:

$$f_{i/1} = f_{i/2} = \ldots = f_{i/j} = \ldots = f_{i/s}$$

4. Independencia Estadística. Cualitativas

Sexo		Tabaquismo			
		Fumadores	*Exfumadores*	*No fumadores*	**Total**
Hombres	Recuento	18	7	69	94
	% dentro de Sexo	19,1%	7,4%	73,4%	100,0%
Mujeres	Recuento	42	6	58	106
	% dentro de Sexo	39,6%	5,7%	54,7%	100,0%
Total	Recuento	60	13	127	200
	% dentro de Sexo	30,0%	6,5%	63,5%	100,0%

Si las variables tabaquismo y sexo fuesen independientes, los porcentajes de tabaco dentro de sexo deberían ser iguales a los totales.

Es lo mismo que decir que dos variables son independientes cuando las distribuciones condicionales son iguales sus respectivas distribuciones marginales.

4. Independencia Estadística. Cualitativas

Sexo		Tabaquismo			
		Fumadores	*Exfumadores*	*No fumadores*	**Total**
Hombres	Recuento	18	7	69	94
	% dentro de Sexo	19,1%	7,4%	73,4%	100,0%
Mujeres	Recuento	42	6	58	106
	% dentro de Sexo	39,6%	5,7%	54,7%	100,0%
Total	Recuento	60	13	127	200
	% dentro de Sexo	30,0%	6,5%	63,5%	100,0%

"A ojo", podemos decir que dos variables son independientes si las frecuencias observadas son parecidas al producto de las frecuencias marginales => veamos cómo se hace

$$\frac{n_{i1}}{n_{.j}} = \frac{n_{i2}}{n_{.j}} = \ldots = \frac{n_{is}}{n_{.j}} = \frac{\sum_{j=1}^{j=s} n_{ij}}{\sum_{j=1}^{j=s} n_{.j}} = \frac{n_{i.}}{N}$$

$$\frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{N}$$

$$\forall ij$$

$$n_{ij} = \frac{n_{i.}n_{.j}}{N} \Rightarrow \frac{n_{ij}}{N} = \frac{n_{i.}n_{.j}}{NN} \Rightarrow f_{ij} = f_{i.}f_{.j}$$

CONDICIÓN DE INDEPENDENCIA

4. Independencia Estadística. Cualitativas

Sexo		Tabaquismo			
		Fumadores	*Exfumadores*	*No fumadores*	**Total**
Hombres	Frec. observada	18	7	69	94
	Frec. esperada	28,2	6,1	59,7	94
Mujeres	Frec. observada	42	6	58	106
	Frec. esperada	31,8	6,9	67,3	106
	Total	60	13	127	200

La frecuencia **observada** para mujeres **Y** fumadoras es de 42 (la muestra real)

La frecuencia **esperada** para mujeres **Y** fumadoras sería de 31,8 mujeres en el caso de las variables fuesen independientes => (106 * 60)/200

Este número corresponde al número de mujeres (106) por el número de fumadores (60) partido por el número total de sujetos en el estudio (200)

4. Independencia Estadística. Prueba Chi^2 de Pearson

- Por el momento nos vale con saber que si las frecuencias esperadas no coinciden con las observadas, existe relación entre las variables.

- Si queremos ser más exactos y saber si esas diferencias son “grandes” o “pequeñas” debemos recurrir a una **prueba estadística** denominada **χ2 de Pearson** (se lee *ji-cuadrado*).

4. Independencia Estadística. Prueba Chi^2 de Pearson

Sexo		Tabaquismo			
		Fumadores	*Exfumadores*	*No fumadores*	**Total**
Hombres	Frec. observada	18	7	69	94
	Frec. esperada	28,2	6,1	59,7	94
Mujeres	Frec. observada	42	6	58	106
	Frec. esperada	31,8	6,9	67,3	106
	Total	60	13	127	200

Se hace un promedio de las diferencias al cuadrado entre las frecuencias observadas y las esperadas. El resultado se comprueba en la tablas para la distribución **χ^2** que nos permite comprobar si esas diferencias son lo suficientemente grandes para considerar la dependencia.

4. Independencia Estadística. Prueba Chi^2 de Pearson

Sexo		Tabaquismo			
		Fumadores	*Exfumadores*	*No fumadores*	**Total**
Hombres	Frec. observada	18	7	69	94
	Frec. esperada	28,2	6,1	59,7	94
Mujeres	Frec. observada	42	6	58	106
	Frec. esperada	31,8	6,9	67,3	106
	Total	60	13	127	200

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$$

Observada

Esperada

4. Independencia Estadística. Prueba Chi^2 de Pearson

Sexo		Tabaquismo			
		Fumadores	*Exfumadores*	*No fumadores*	**Total**
Hombres	Frec. observada	18	7	69	94
	Frec. esperada	28,2	6,1	59,7	94
Mujeres	Frec. observada	42	6	58	106
	Frec. esperada	31,8	6,9	67,3	106
	Total	60	13	127	200

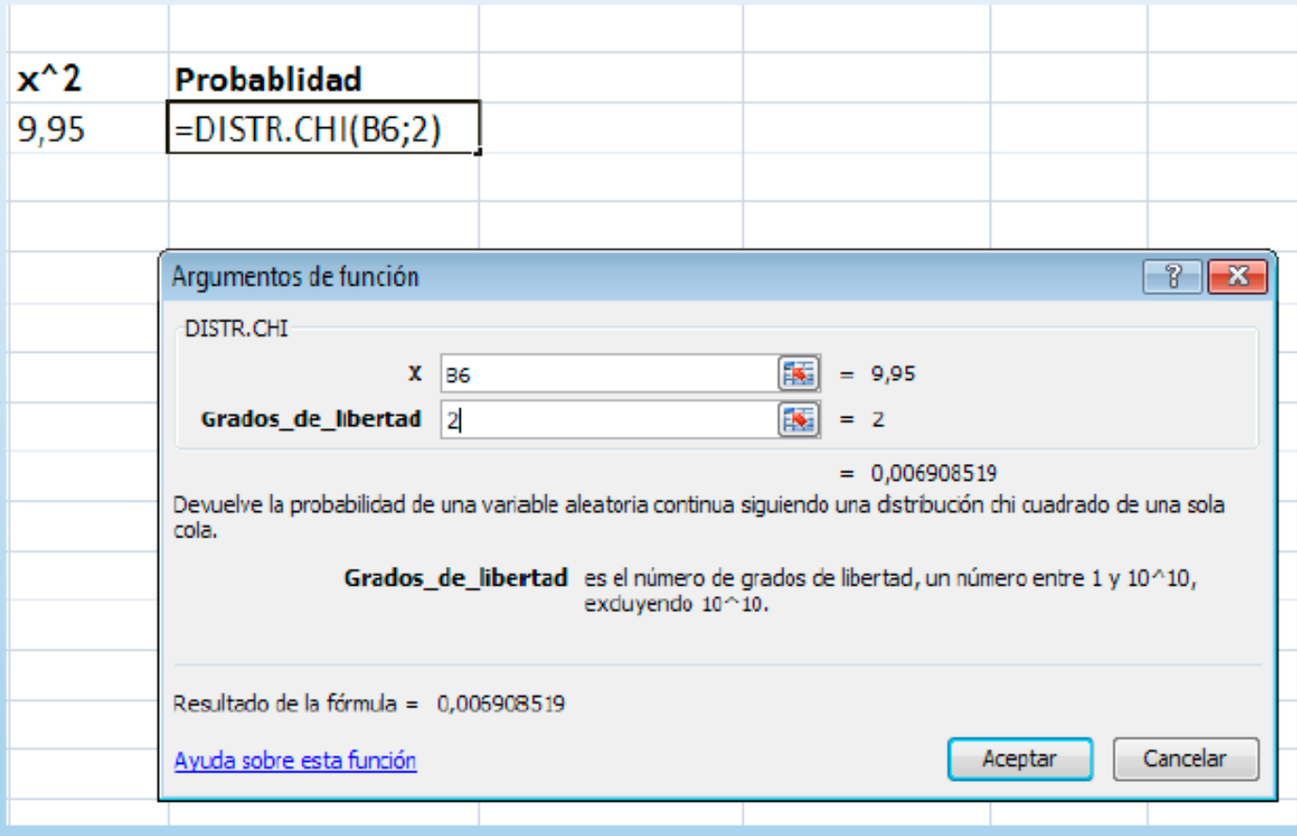
$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$$

$$\chi^2 = \frac{(18-28{,}2)^2}{28{,}2} + \frac{(7-6{,}1)^2}{6{,}1} + \ldots + \frac{(58-67{,}3)^2}{67{,}3} = 9{,}95$$

Todas las categorías conjuntas

Ejemplo: hombres y exfumadores

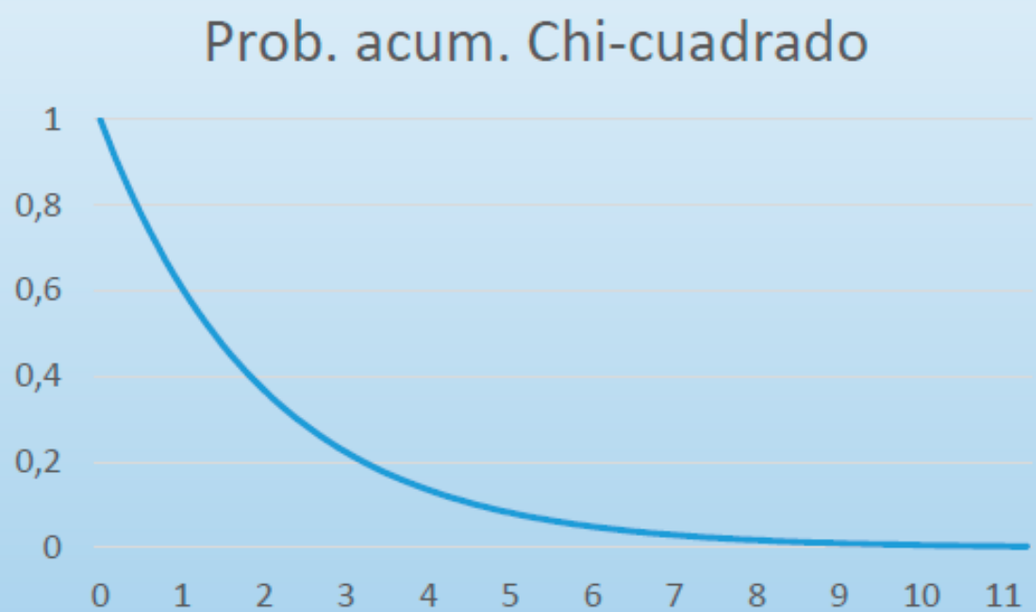
4. Independencia Estadística. Prueba Chi^2 de Pearson



- **Suponiendo que fuesen independientes,** la probabilidad encontrar un valor de 9,95 es 0,007 (0,7%), por tanto, pensamos que no son independientes => **están RELACIONADAS.**
- Esto quiere decir que, según el estudio, el hecho de ser hombre o mujer **SÍ INFLUYE** en la probabilidad de que sea o no fumador

4. Independencia Estadística. Prueba Chi^2 de Pearson

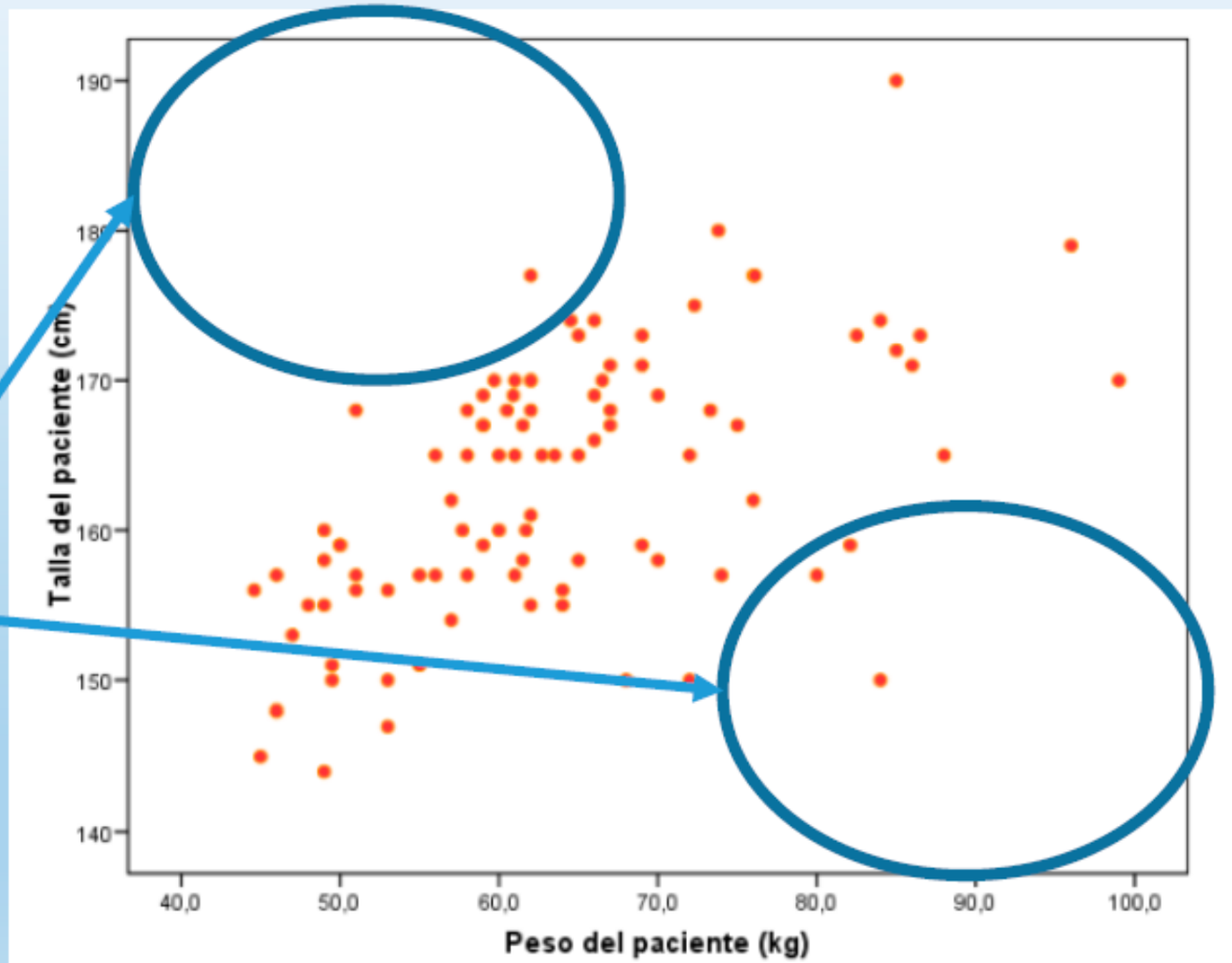
X^2	Prob
0	1
0,1	0,95122942
0,2	0,90483742
0,3	0,86070798
0,4	0,81873075
0,5	0,77880078
0,6	0,74081822
0,7	0,70468809
0,8	0,67032005
0,9	0,63762815
1	0,60653066
1,1	0,57694981
1,2	0,54881164
1,3	0,52204578
1,4	0,4965853



- Cuanto mayor es el valor del resultado de X^2, MENOS probabilidad de que estén relacionadas.

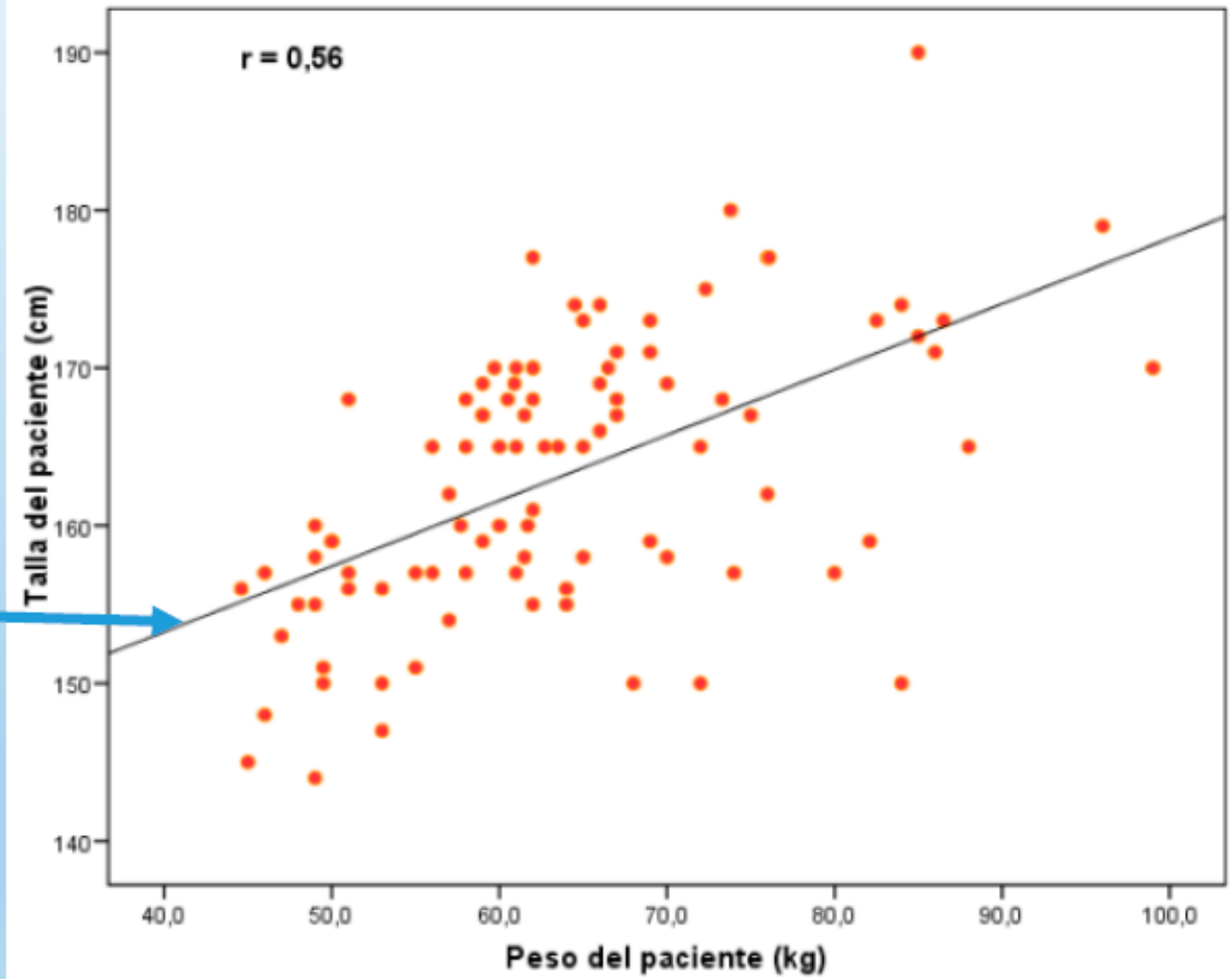
5. Covarianza y coeficiente de correlación

- Entre las variables talla y peso del paciente (cuantitativas) se puede apreciar cierta relación.
- Cada punto es un paciente con su peso (eje horizontal) y talla (eje vertical)
- No hay muchos pacientes con poca talla y mucho peso o mucho peso y poca talla



5. Covarianza y coeficiente de correlación

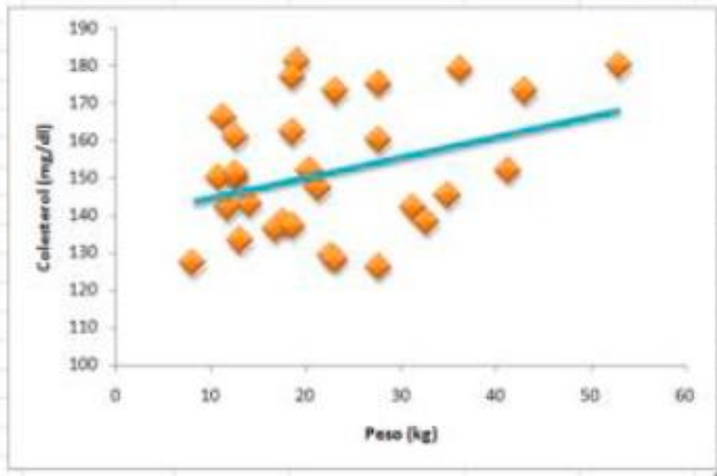
- Entre las variables talla y peso del paciente (cuantitativas) se puede apreciar cierta relación.
- Una técnica permite predecir la recta que "pasa más cerca" de todos los puntos
- Se trata de un "diagrama de dispersión"



5. Covarianza y coeficiente de correlación

	A	B	C	D
1	Edad (años)	Peso (kg)	Colesterol (mg/dl)	Altura (cm)
2	2	13	161	81
3	4	12	142	102
4	4	19	181	97
5	10	33	138	130
6	4	28	175	94
7	2	19	162	84
8	5	23	129	109
9	7	14	143	112
10	3	13	150	84
11	5	23	128	102
12	4	18	138	102
13	6	28	126	112
14	4	13	133	94
15	4	11	150	99
16	3	23	173	84
17	10	35	145	140
18	3	8	127	86
19	6	17	136	109
20	4	13	151	94
	10	53	180	132

CORRELACIÓN= 0,34

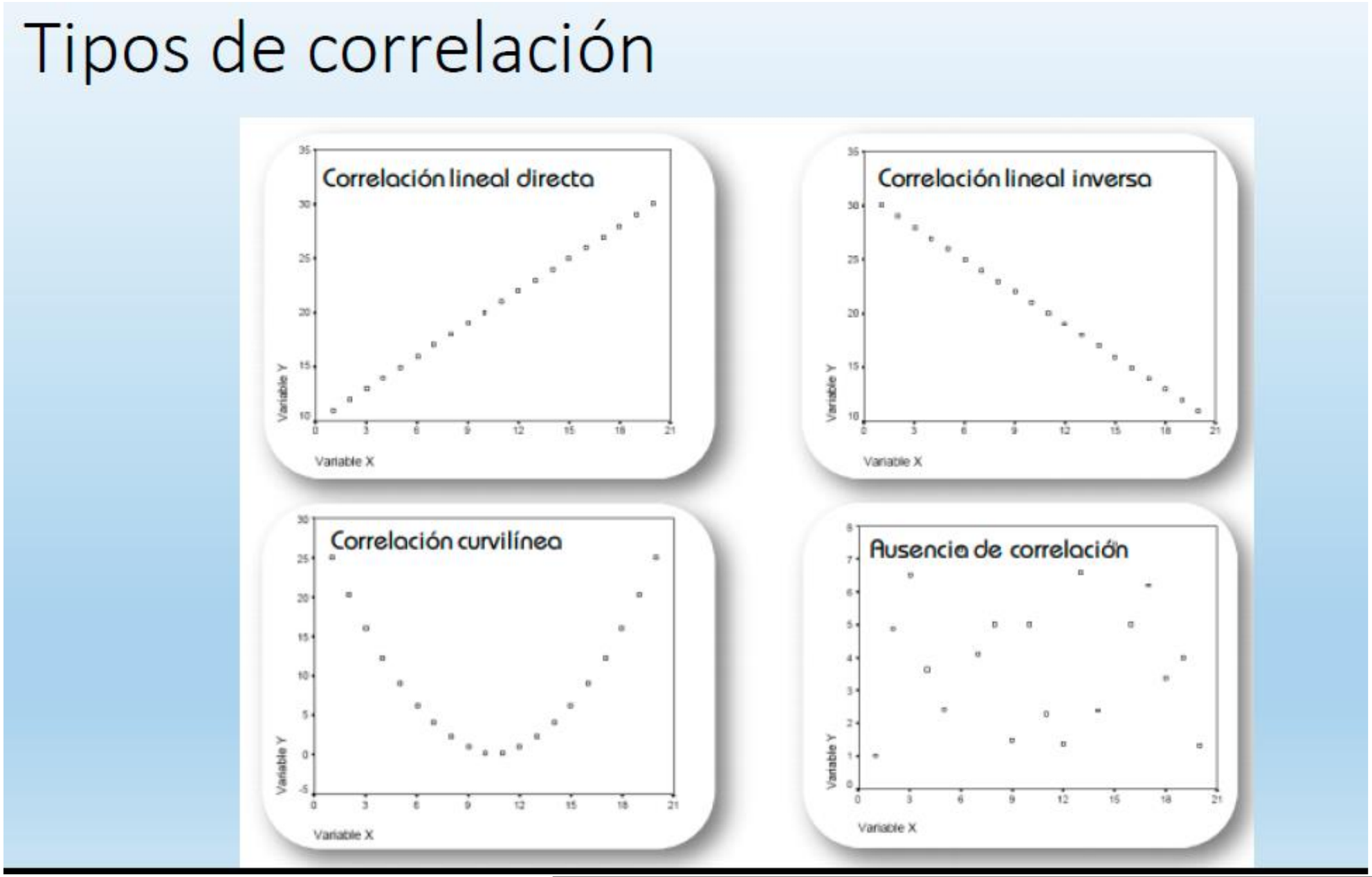


Coeficiente de correlación nos indica tanto el sentido de la relación lineal como la fuerza de la relación

r = 0 Ausencia de relación
r= +1 máxima relación directa
r = -1 máxima relación inversa

- Correlación = 0,34
- Están relacionados "positivamente" (a mayor peso mayor colesterol) pero no totalmente (hay puntos por encima y por debajo de la línea) por eso no vale exactamente 1

5. Covarianza y coeficiente de correlación



5. Covarianza y coeficiente de correlación

Covarianza

- Es una medida de **dispersión conjunta** de dos variables
- Si recordamos la varianza (S2):

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

- Es el sumatorio de las diferencias cada valor de la variable y su media elevada al cuadrado y divido entre el tamaño muestral.
- Es una medida de dispersión de la variable.

5. Covarianza y coeficiente de correlación

- La **covarianza es similar** pero ahora se tiene en cuenta la dispersión de las dos variables (x, y)

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

- Es el sumatorio de las diferencias cada valor de la variable *X* respecto de la media de *X* multiplicado por las diferencias de cada valor de la variable *Y* respecto de la media de *Y*, todo ello divido entre el tamaño muestral.
- Es una medida de **dispersión conjunta**

5. Covarianza y coeficiente de correlación

- Interpretación de la covarianza

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Si $S_{xy}>0$ Relación directa
Si $S_{xy}<0$ Relación inversa
Si $S_{xy}=0$ No hay relación

CUIDADO PORQUE **NO INFORMA** DE LA FUERZA DE LA RELACIÓN ¿por qué?

- Habría que **descontar** la influencia que puedan tener las varianzas de X e Y

5. Covarianza y coeficiente de correlación

- El coeficiente de correlación lineal de Pearson (r) mide el grado de relación entre dos variables

$$r = \frac{S_{xy}}{S_x \cdot S_y}$$

- Tiene el mismo signo que la covarianza y nos informa sobre el grado de relación lineal entre las dos variables.
- Al ser un coeficiente, es adimensional y podrá tomar valores comprendidos entre +1 (correlación lineal directa máxima) y -1 (correlación lineal inversa máxima). r = 0 (o próximo) cuando no exista correlación

5. Covarianza y coeficiente de correlación

- El coeficiente de correlación lineal de Pearson (r) mide el grado de relación entre dos variables
- **Son sólo valores informativos** (depende de varios factores)
  - Sólo se representan los positivos. En el caso de que el coeficiente sea negativo el criterio es el mismo perola relación es inversa

Coeficiente de Correlación	Descripción
$0 < r_{xy} \leq 0,2$	Muy Baja
$0,2 < r_{xy} \leq 0,4$	Baja
$0,4 < r_{xy} \leq 0,7$	Moderada o media
$0,7 < r_{xy} \leq 0,9$	Alta
$0,9 < r_{xy} < 1$	Muy alta

6. Recta de Regresión

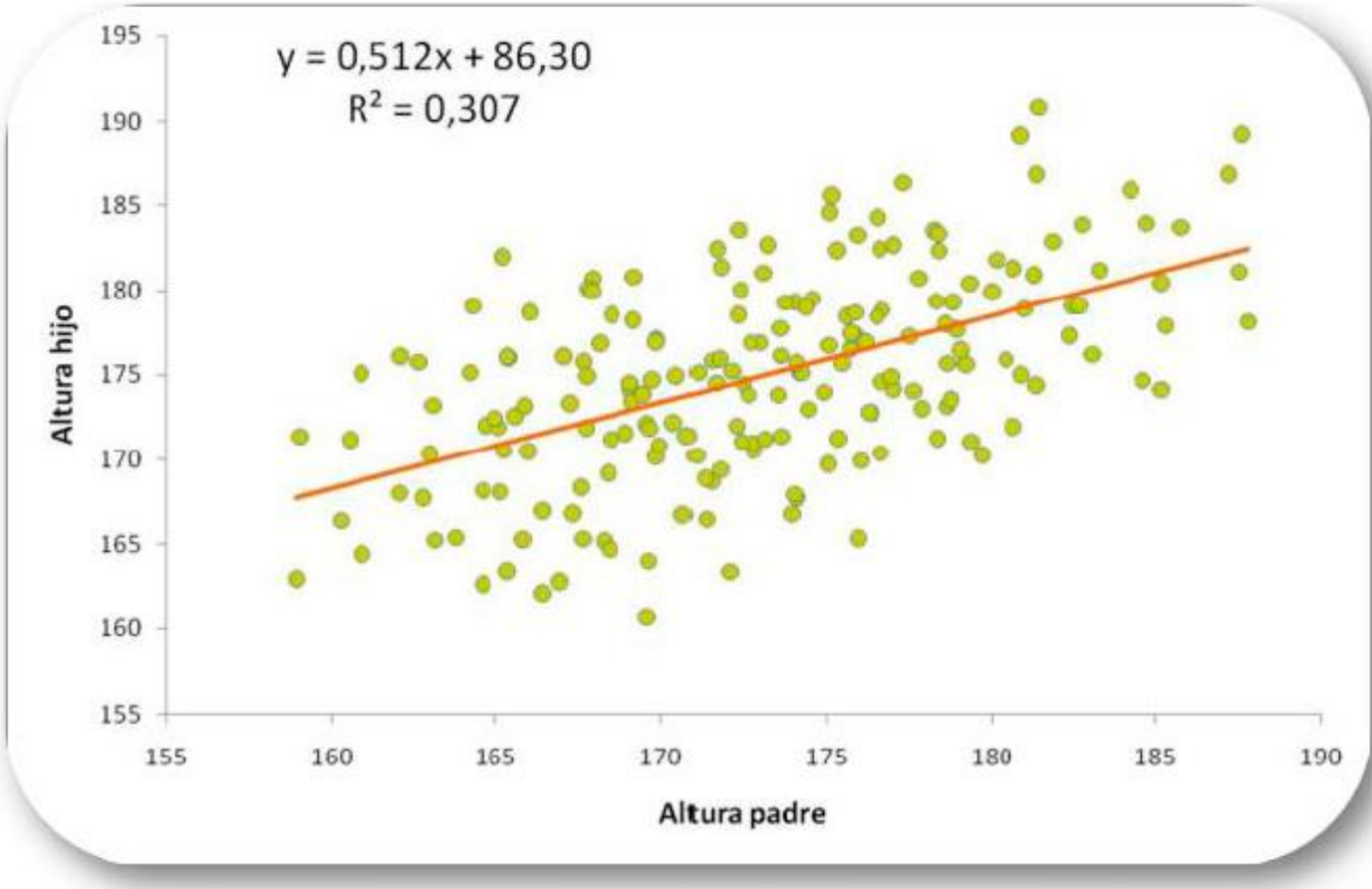
Regresión lineal

- Es una técnica estadística que se utiliza para estudiar la **relación entre variables** cuantitativas y normalmente lleva implícito el objetivo de predecir el comportamiento de una variable (dependiente) a partir de otra u otras variables conocidas (independientes o predictoras).
- Tras el análisis se obtendrá una ecuación denominada **ecuación de regresión**

6. Recta de Regresión

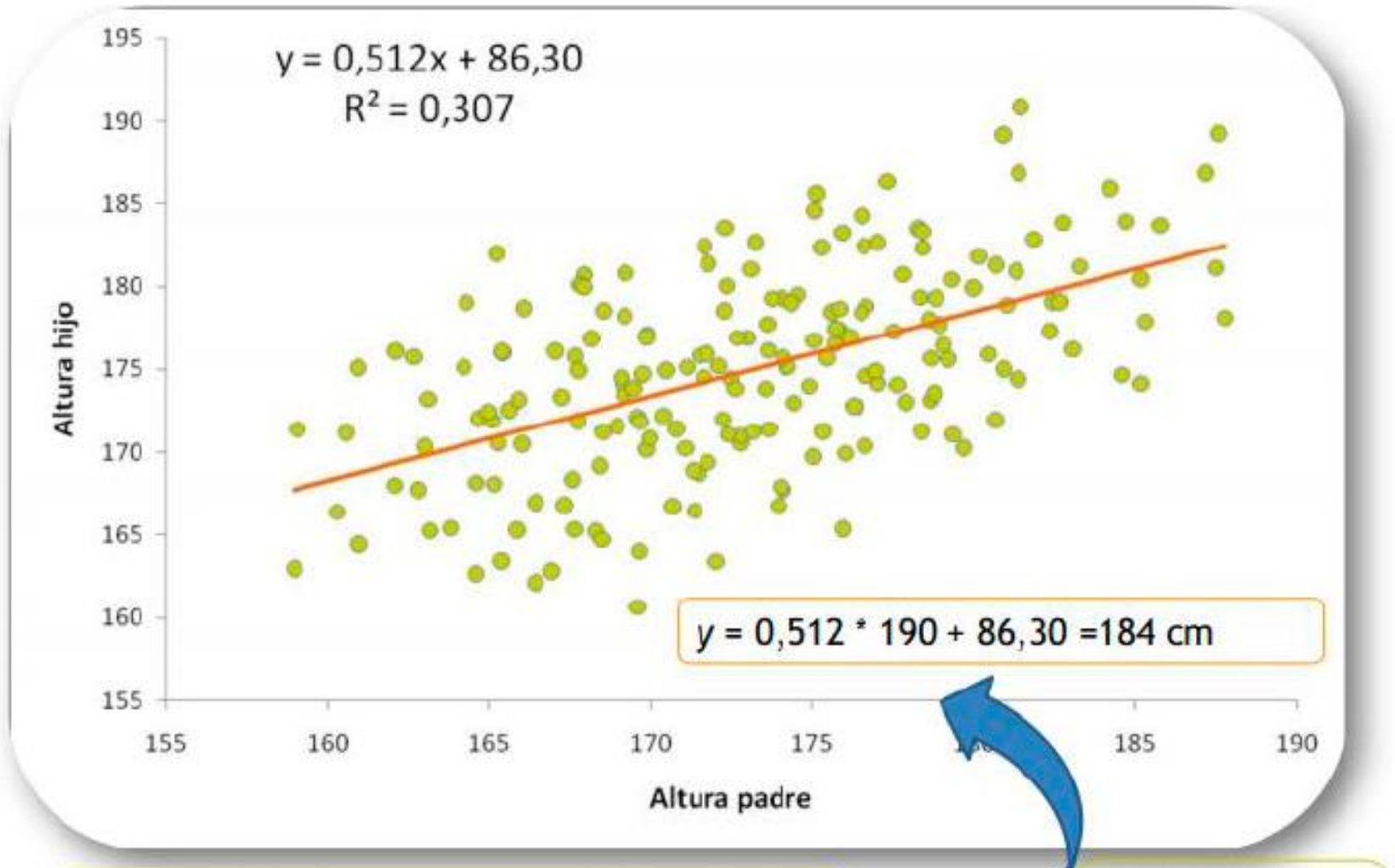
Regresión a la media

- Fue el origen de estas técnicas (la predicción de la altura del hijo a partir de la de los padres)
- Lo llevó a cabo Pearson. A partir de él desarrolló estas técnicas de regresión
- El modelo nos permite predecir la altura del hijo a partir de la edad del padre
- Este modelo explicó un fenómeno observado: si la altura del padre se salía de la media, la altura del hijo tendía a volver a la media, y al revés



Y = variable dependiente, altura del hijo
X =variable independiente, altura del padre

6. Recta de Regresión



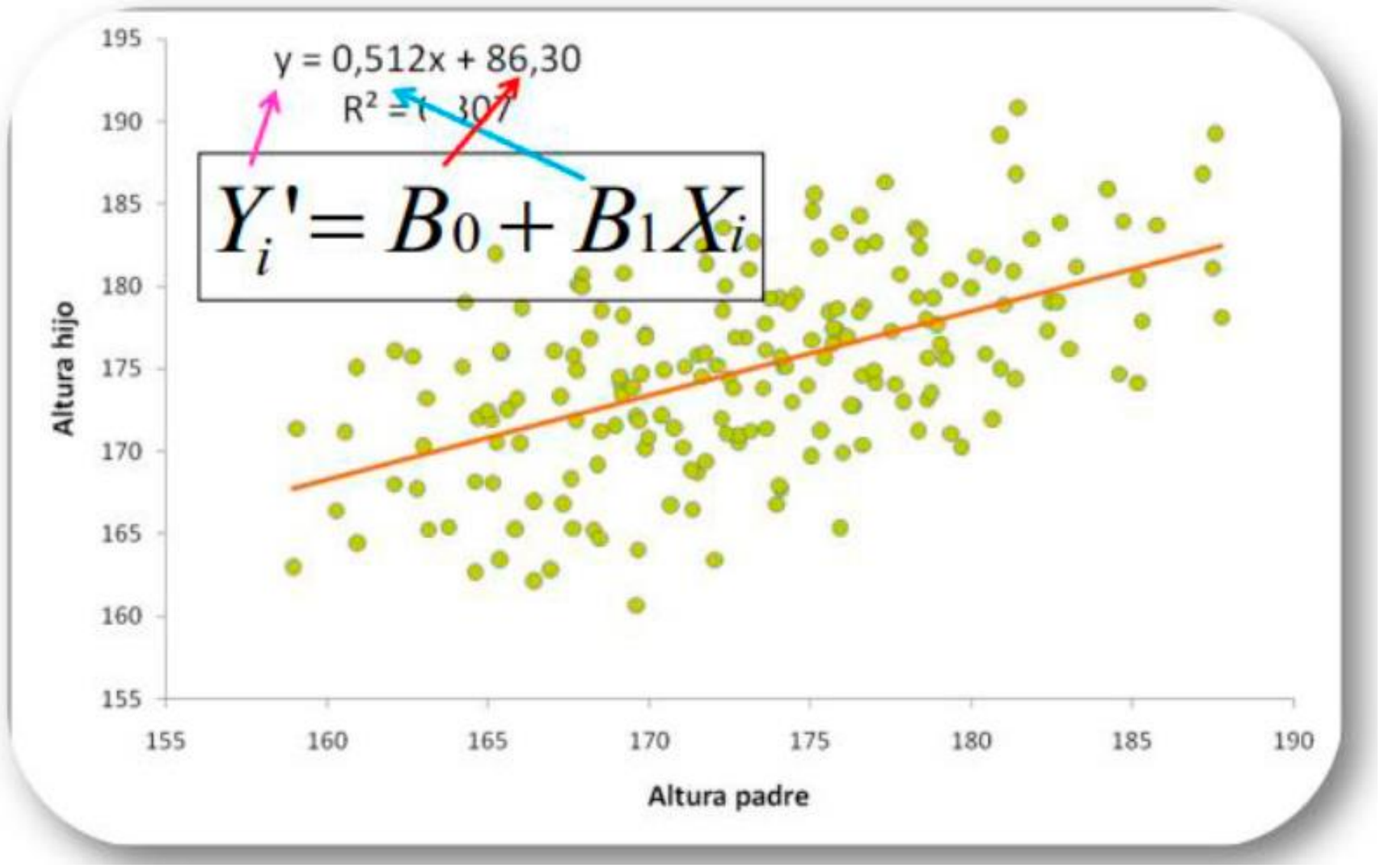
6. Recta de Regresión. Regresión lineal Simple

El modelo de regresión lineal simple es sencillo porque se corresponde con la función de una recta

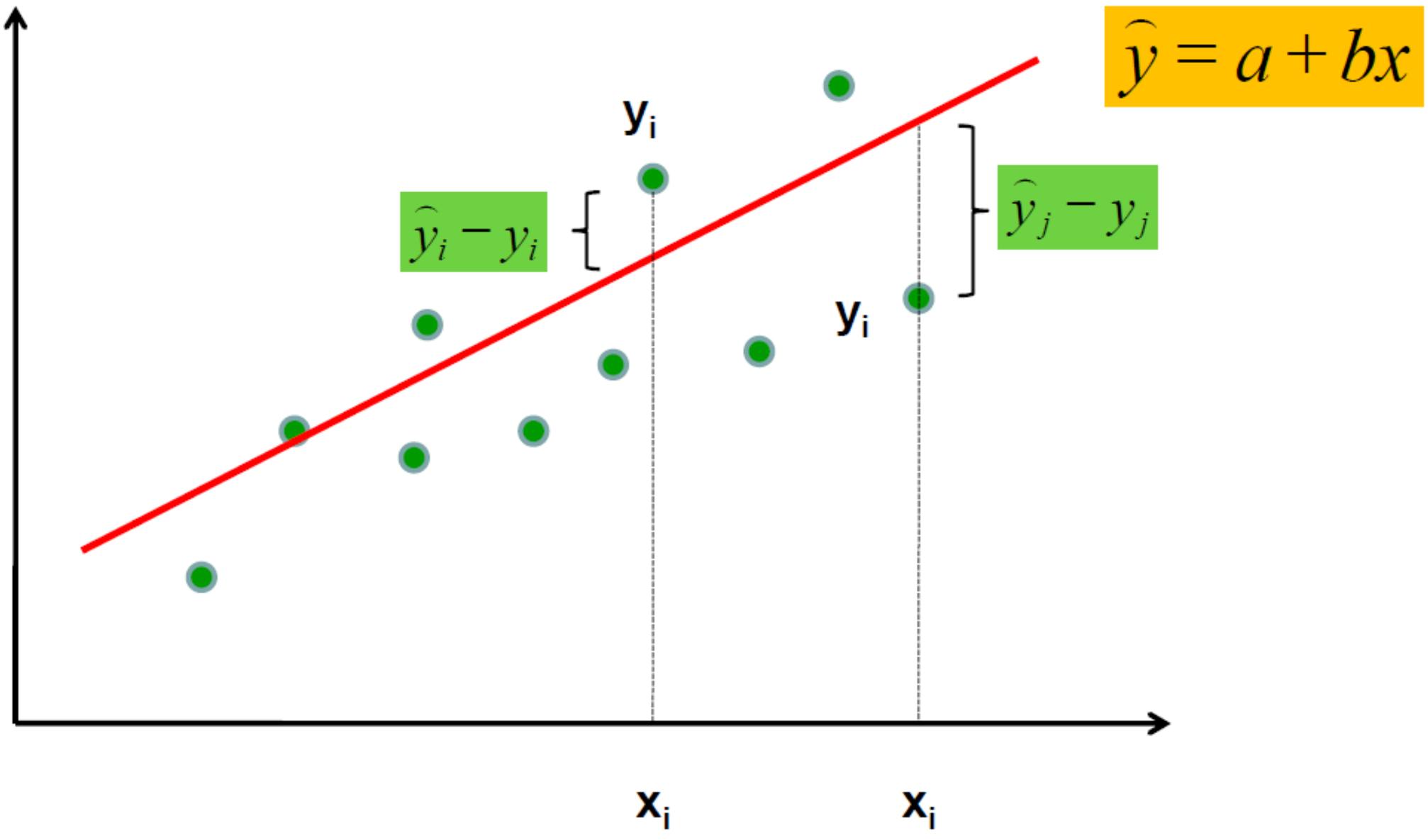
$$Y_i' = B_0 + B_1 X_i$$

Donde Y_i'= variable dependiente ***estimada*** por el modelo.
B_0= ***punto de corte*** de la recta sobre el eje de ordenadas (eje *Y*) y es una constante en el modelo.
B_1= pendiente de la recta de regresión cuyo signo determina la dirección de la relación (directa o inversa) y cuyo valor es un indicador de ***cuántas unidades varía*** la variable *Y* (dependiente) en relación a la variable *X* (independiente).

6. Recta de Regresión



6. Recta de Regresión



6. Recta de Regresión

Queremos encontrar la recta que mejor se ajuste a los datos y por tanto ésta debe ser la que haga mínimos cada uno de los "errores" anteriormente señalados.

Para ello se utiliza el método de los mínimos cuadrados, en virtud del cual se trata de encontrar la recta que haga mínima la suma de los cuadrados de todos los errores.

Es decir encontrar los valores de los parámetros a y b que hagan mínima la siguiente expresión

$$\sum_{j=1}^{j=r} (\hat{y}_j - y_j)^2 f_{.j} = \sum_{i=1}^{i=r} (a + bx_j - y_j)^2 f_{.j}$$

6. Recta de Regresión

El proceso de optimización que se realiza puede plantearse para cualquier tipo de ecuación (a decidir usualmente por el investigador). En el caso de que se haya decidido por una regresión lineal, los valores estimados para los parámetros a y b son los siguientes:

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{Sxy}{S_x^2}$$

6. Recta de Regresión

Se puede definir por tanto la recta de regresión de Y sobre X (variable explicativa la X) como:

$$y - \bar{y} = b(x - \bar{x}) = \frac{Sxy}{S_x^2}(x - \bar{x})$$

Del mismo modo se puede encontrar la recta de regresión de X sobre Y (variable explicativa Y) como:

$$x - \bar{x} = b^*(y - \bar{y}) = \frac{Sxy}{S_y^2}(y - \bar{y})$$

6. Recta de Regresión

Recta de regresión de Y sobre X

$$\hat{y} = a + bx$$

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{Sxy}{S_x^2}$$

Recta de regresión de X sobre Y

$$\hat{x} = a^* + b^* y$$

$$a^* = \bar{x} - b^* \bar{y}$$

$$\frac{1}{b^*} = \frac{S_y^2}{S_{xy}}$$

Pendiente de ambas rectas

6. Recta de Regresión

Dadas las expresiones de las pendientes b y b* se observa que quien marca su signo es precisamente la covarianza S_{xy}.
De este modo si la covarianza es >0 entonces ambas variables varían en el mismo sentido.
Si por el contrario la covarianza es <0 entonces varían en sentidos contrarios.
Cuando la covarianza es cero las rectas se reduce particularmente y ló único que se puede afirmar es que no existe relación lineal. A todos los valores de x les corresponde un único valor de y, y viceversa en la otra recta

6. Recta de Regresión

En el caso de que se pueda afirmar que existe algún tipo de relación lineal, se hace preciso introducir un indicador del grado de relación. Dicho de otro modo, de hasta qué punto la recta de regresión representa o se ajusta adecuadamente al conjunto de puntos.

Esto se realiza a partir de la introducción del coeficiente de correlación lineal r

$$r = \frac{S_{xy}}{S_x S_y}$$

6. Recta de Regresión

Propiedades

$$-1 \leq r = \frac{S_{xy}}{S_x S_y} \leq 1$$

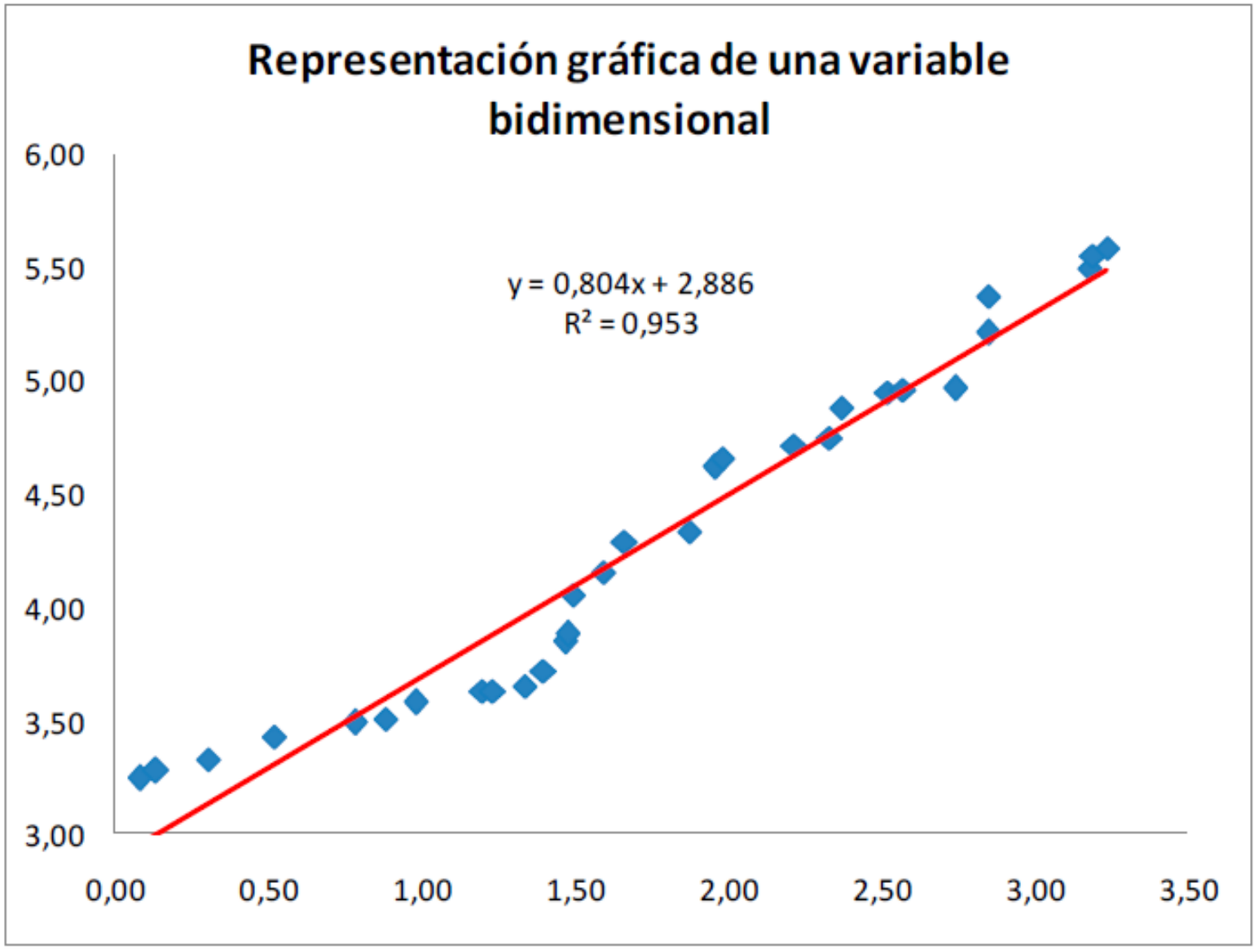
Si r es 0 no existe relación lineal.
Cuando más próximo se encuentra a 0 menos se puede hablar de relación lineal.
Cuando tome el valor 1 o -1 se puede hablar de relación lineal directa o inversa, respectivamente.
No existe un valor concreto de r a partir del cual se pueda considerar admisible la relación lineal. En todo caso no se debe optar por la linealidad para valores inferiores a 0,75 (-0,75)

6. Recta de Regresión

$$\frac{S_{xy}^{\ 2}}{S_x^{\ 2} S_y^{\ 2}} = r^2$$

A esta expresión, en el caso de regresión lineal, se la denomina Coeficiente de Determinación R_2 y puede considerarse igualmente como una medida del grado de ajuste de la recta al conjunto de puntos, variando ahora entre 0 y 1

6. Recta de Regresión



6. Recta de Regresión

La recta de regresión (y cualquier otra curva que pudiéramos considerar) tiene como objeto ajustar una recta al conjunto de puntos. Por tanto se calcula solo con la información disponible y no se puede afirmar que se siga realizando el ajuste en la misma medida cuando se consideran puntos fuera del rango con el que se ha trabajado.

Esto conlleva a que la realización de predicciones con la recta de regresión debe realizarse de forma muy cuidados y prudente puesto que a pesar de que tengamos un muy buen ajuste, cuando actuamos fuera del rango (predicciones) la situación puede cambiar.

Si la utilizamos para realizar predicciones lo hacemos bajo el supuesto de que en los nuevos puntos el comportamiento es el mismo que en el de los conocidos, pero eso no tenemos forma de asegurarlo.

6. Recta de Regresión

Los datos de la producción de trigo en toneladas (X) y el precio del kilo de harina en pesetas (Y) en la década de los 80 en España fueron:

Producción de trigo	30	28	32	25	25	25	22	24	35	40
Precio de la harina	25	30	27	40	42	40	50	45	30	25

Ajusta la recta de regresión por el método de mínimos cuadrados

Resultados

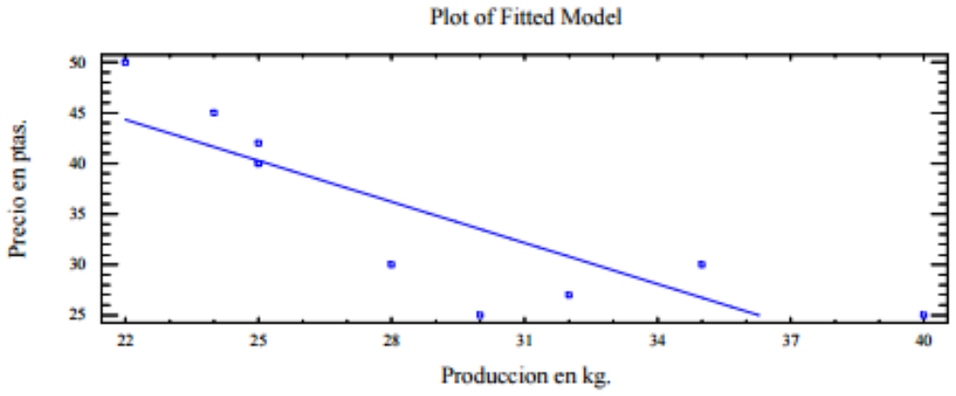
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{10} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{10} x_i^2 - n\bar{x}^2} = \frac{9734 - 10 \times 28{,}6 \times 35{,}4}{8468 - 10 \times 28{,}6^2} = -1{,}3537$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 35{,}4 + 1{,}3537 \times 28{,}6 = 74{,}116$$

La recta de regresión es:

$$\hat{y} = 74{,}116 - 1{,}3537x$$

6. Recta de Regresión



Regression Analysis - Linear model: Y = a + b*X

Dependent variable: Precio en ptas.
Independent variable: Produccion en kg.

Parameter	Estimate	Standard Error	T Statistic	P-Value
Intercept	74,1151	8,73577	8,4841	0,0000
Slope	-1,35368	0,3002	-4,50924	0,0020

$\hat{\beta}_0$ ← Intercept, $\hat{\beta}_1$ ← Slope

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	528,475	1	528,475	20,33	0,0020
Residual	207,925	8	25,9906		
Total (Corr.)	736,4	9			

Correlation Coefficient = -0,84714
R-squared = 71,7647 percent
Standard Error of Est. = 5,0981

6. Recta de Regresión

Las notas de 12 alumnos de una clase en Matemáticas y Física son las siguientes:

Matemáticas	Física
2	1
3	3
4	2
4	4
5	4
6	4
6	6
7	4
7	6
8	7
10	9
10	10

Hallar las **rectas de regresión** y representarlas.

x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2
2	1	2	4	1
3	3	9	9	9
4	2	8	16	4
4	4	16	16	16
5	4	20	25	16
6	4	24	36	16
6	6	36	36	36
7	4	28	49	16
7	6	42	49	36
8	7	56	64	49
10	9	90	100	81
10	10	100	100	100
72	60	431	504	380

1º Hallamos las **medias aritméticas**.

$$\bar{x} = \frac{72}{12} = 6 \qquad \bar{y} = \frac{60}{12} = 5$$

2º Calculamos la **covarianza**.

$$\sigma_{xy} = \frac{431}{12} - 6 \cdot 5 = 5.92$$

3º Calculamos las **varianzas**.

$$\sigma_x^2 = \frac{504}{12} - 6^2 = 6 \qquad \sigma_y^2 = \frac{380}{12} - 25 = 6.66$$

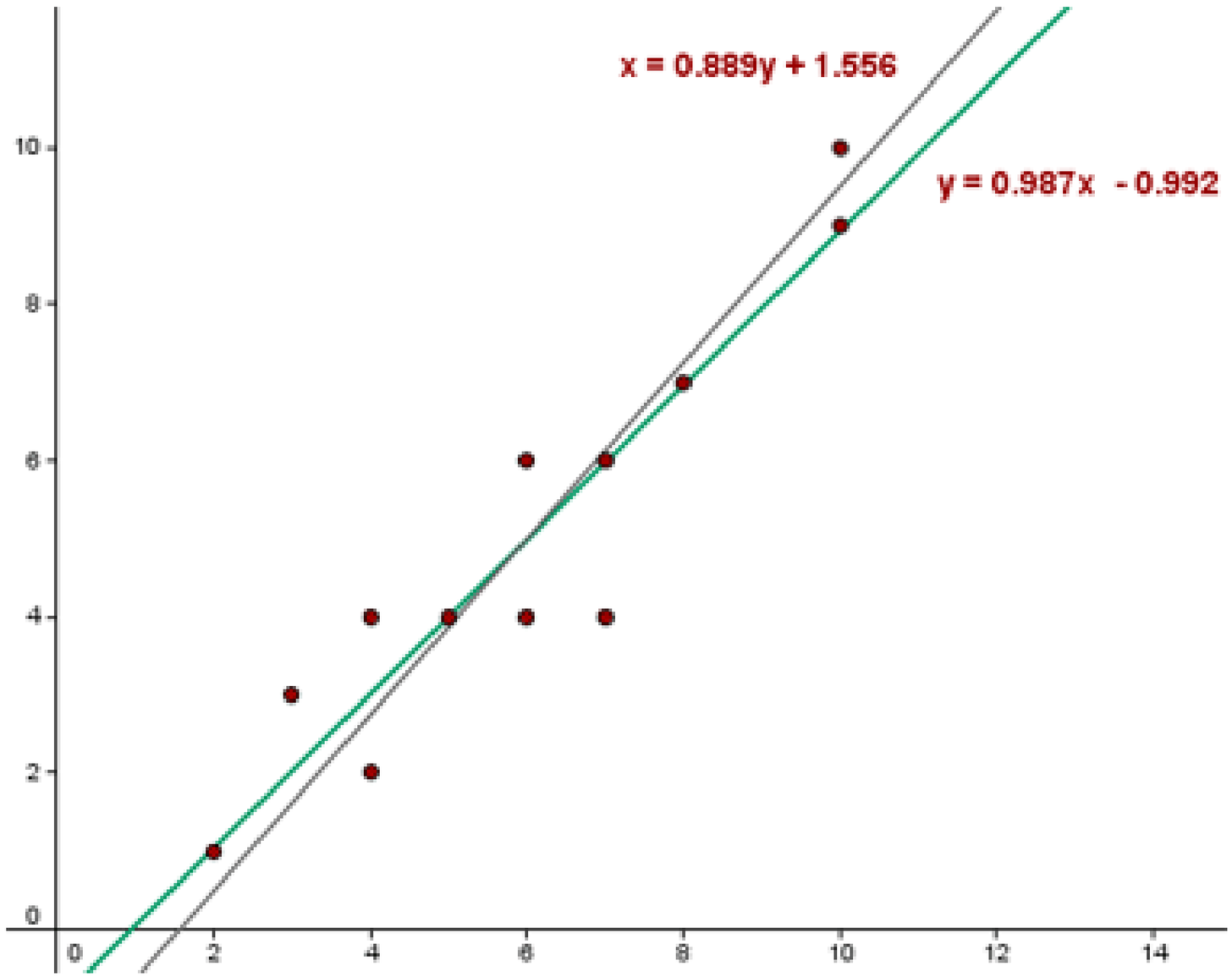
4º Recta de regresión de Y sobre X.

$$y - 5 = \frac{5.92}{6}(x - 6) \qquad y = 0.987x - 0.922$$

4º Recta de regresión de X sobre Y.

$$x - 6 = \frac{5.92}{6.66}(y - 5) \qquad x = 0.889y + 1.556$$

6. Recta de Regresión



1. Variables bidimensionales
2. Tabulación de datos: distribución de frecuencias bidimensionales
3. Distribuciones marginales y condicionadas
4. Independencia estadística
5. Covarianza y coeficiente de correlación
6. Recta de regresión. Estimación de coeficiente. Bondad de ajuste. Predicción

Estadística Aplicada al Marketing

T.4. Fundamentos de Probabilidad

Juan Miguel Navarro Ruiz

Grado en Marketing y Dirección Comercial

1. Conceptos básicos: experimento aleatorio, espacio muestral y sucesos.
2. Definición e interpretación de la probabilidad.
3. Axiomas del cálculo de la probabilidad y sus consecuencias.
4. Probabilidad condicionada.
5. Teorema de la probabilidad total y teorema de Bayes.
6. Independencia de sucesos.

1. Conceptos Básicos: Inferencia

La **inferencia estadística**, o **estadística inductiva**, se usa cuando no se observa la totalidad de la población sino sólo a un subconjunto o muestra, pero con la intención de aplicar los resultados y conclusiones de manera generalizada a toda la población.

La información obtenida no permite hacer afirmaciones absolutas sobre las características de la población pero sí hacer pronósticos en términos de **probabilidad**.

1. Conceptos Básicos: Experimento Aleatorio

Un **experimento** es cualquier operación en la que se puede presentar uno o varios resultados de un conjunto definido de resultados posibles.

Un **experimento determinístico** es aquel que al repetirlo bajo las mismas condiciones ofrece siempre el mismo resultado.

Un **experimento aleatorio** es el que al repetirlo bajo las mismas condiciones no siempre ofrece el mismo resultado.

Ejemplo

Al lanzar una moneda puede salir cara o cruz. Si en un primer lanzamiento sale cara, al repetir el experimento puede salir cara o puede salir cruz. No se obtiene siempre el mismo resultado aunque el experimento se repita en idénticas condiciones. Es un experimento aleatorio.

1. Conceptos Básicos

Características de un experimento aleatorio:

- Se puede repetir indefinidamente bajo las mismas condiciones.
- Se conocen los posibles resultados del experimento pero no se puede saber el resultado concreto que saldrá al realizarlo.

NOTAR QUE:

- Si varían los condiciones iniciales el experimento sería otro distinto y los resultados no serían comparables.
- Se espera que subyazca algún modelo de regularidad o patrón en los resultados de forma que se pueda estudiar el experimento y predecir el funcionamiento del sistema de aparición de resultados.

1. Conceptos Básicos: Espacio Muestral

Al conjunto de todos los posibles resultados de un experimento aleatorio se le llama **espacio muestral** , espacio de resultados o espacio de comportamiento, y cada uno de esos posibles resultados individuales es un **punto muestral, elemento** o comportamiento individual.

El espacio muestral puede ser:

- Espacio muestral **finito** es el que tiene un número finito de puntos o resultados posibles.
- Espacio muestral **infinito numerable** o espacio muestral discreto es el que tiene un número infinito de elementos pero numerables, es decir, asociable a los números naturales.
- Espacio muestral **continuo** es el que tiene un número infinito no numerable de posibles resultados.

1. Conceptos Básicos: Suceso

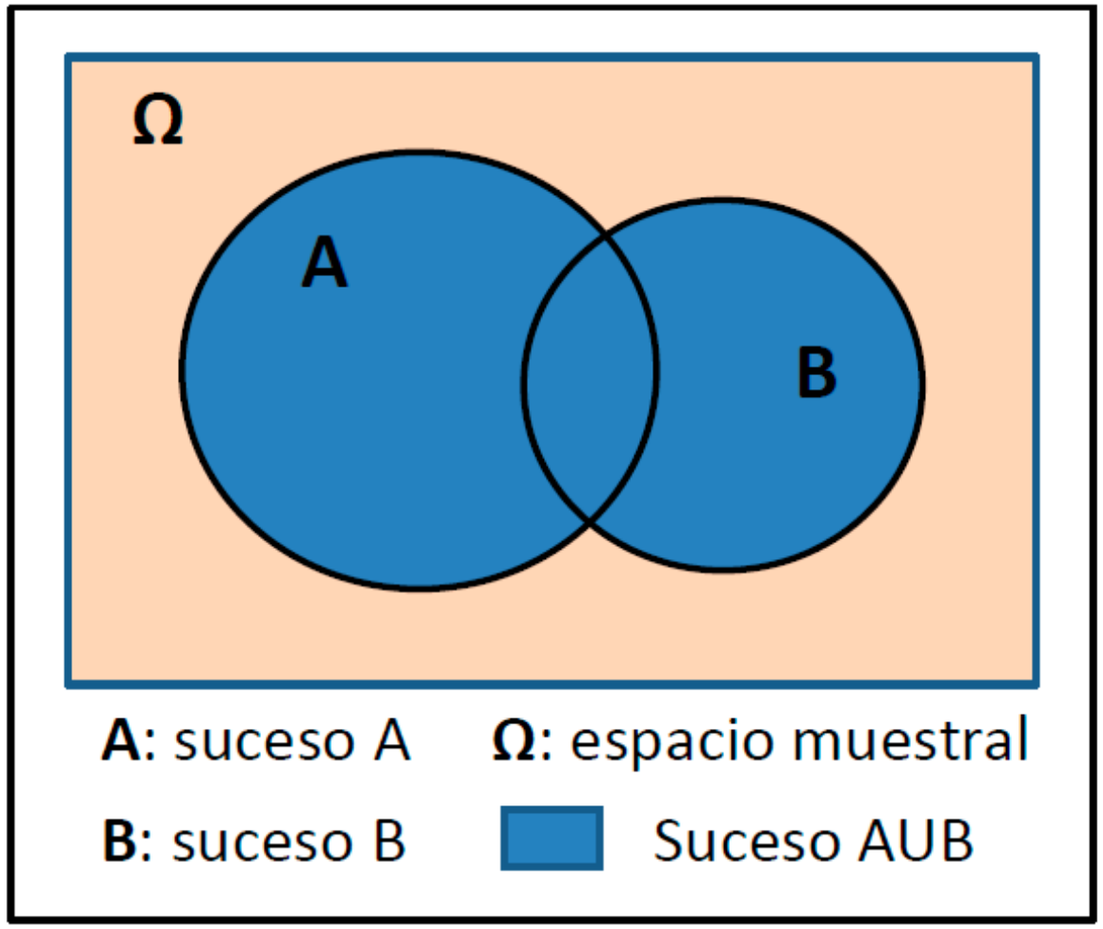
Un **suceso** es un subconjunto del espacio muestral, es decir, de resultados posibles del experimento. Se dice que ocurre o se presenta ese suceso si al realizar el experimento aparece uno de los resultados pertenecientes al subconjunto definido.

- **Suceso elemental** o simple es cada uno de los resultados posibles del experimento. Consta de un solo punto muestral.
- **Suceso compuesto** es el que está formado por varios sucesos elementales.
- **Suceso seguro**, cierto o universal es aquel que contiene a todos los sucesos elementales del espacio muestral. Ocurre siempre.
- **Suceso imposible** es el que no contiene ningún suceso elemental. No ocurre nunca.

1. Conceptos Básicos: Operaciones con Sucesos

Unión de sucesos

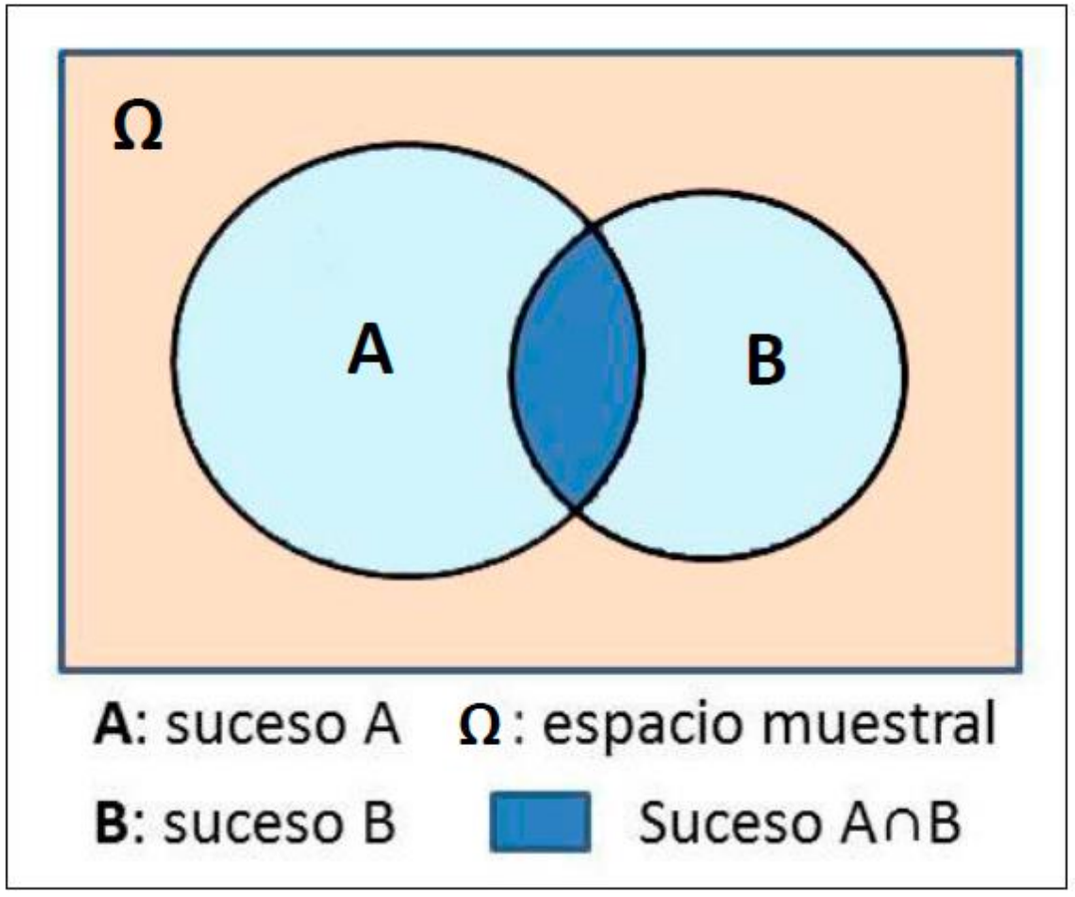
Dados dos sucesos, A y B, la unión de ambos es otro suceso, denotado por $A \cup B$, que está formado por los elementos pertenecientes a A, o a B, o a ambos. El suceso unión se verifica si ocurre cualquiera de los dos.



1. Conceptos Básicos: Operaciones con Sucesos

Intersección de sucesos

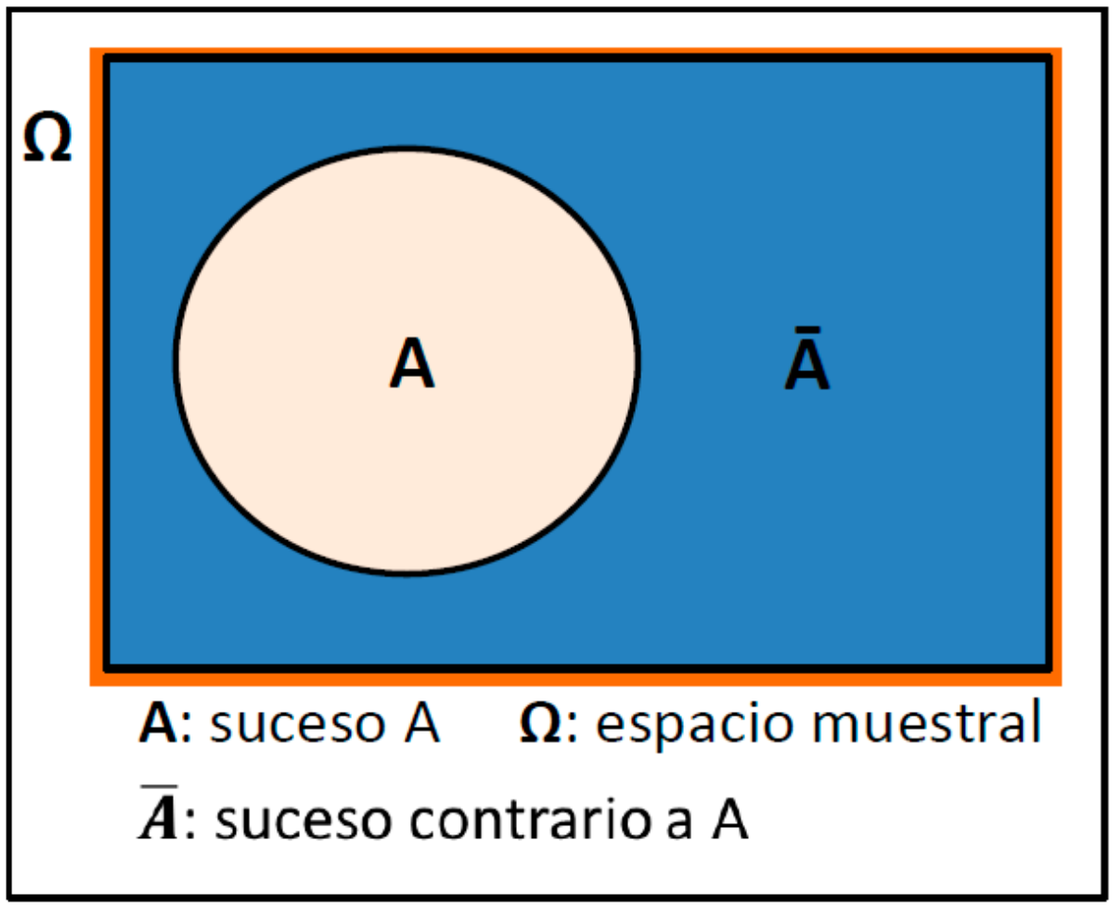
Dados dos sucesos, A y B, se define la intersección de ambos como otro suceso, denotado por $A \cap B$, compuesto por los elementos pertenecientes a A y a B simultáneamente. Este suceso sólo se verifica si ocurre A y ocurre B.



1. Conceptos Básicos: Operaciones con Sucesos

Suceso complementario o contrario

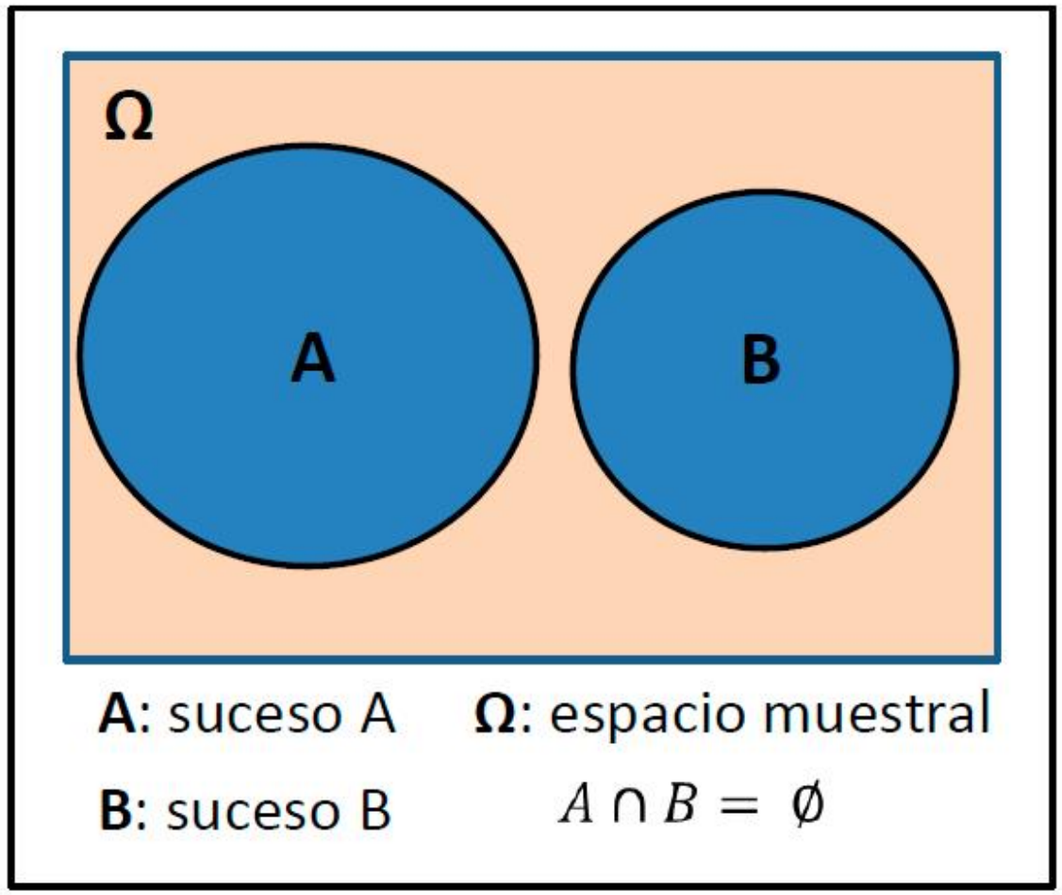
Suceso complementario o contrario de un suceso A es otro suceso, $\bar{A}$, compuesto por los elementos del espacio muestral que no están incluidos en A. Si no ocurre A, ocurre su suceso contrario $\bar{A}$.



1. Conceptos Básicos: Operaciones con Sucesos

Sucesos incompatibles

Dados dos sucesos, A y B, se dice que son incompatibles, disjuntos o excluyentes si la intersección de ambos es un conjunto vacío, $A \cap B = \emptyset$. Esto ocurre porque A y B no tienen ningún elemento en común, no pueden ocurrir simultáneamente.

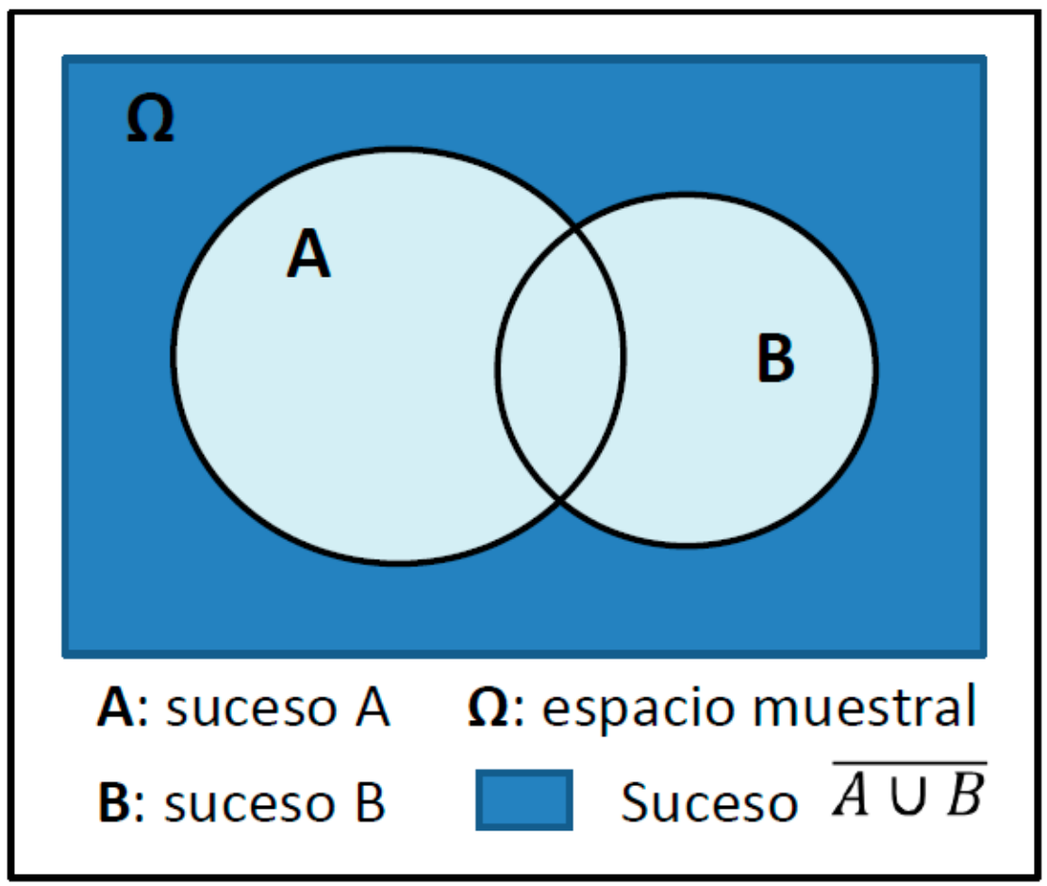


1. Conceptos Básicos: Operaciones con Sucesos

Contrario de la unión de sucesos

Dados dos sucesos, A y B, se define el contrario de la unión de ambos como otro suceso, denotado por $\overline{A \cup B}$, que está formado por los elementos no pertenecientes a A ni a B, es decir, no ocurre A ni ocurre B.

$$\overline{A \cup B} = \bar{A} \cap \bar{B}$$ **(Ley de Morgan)**



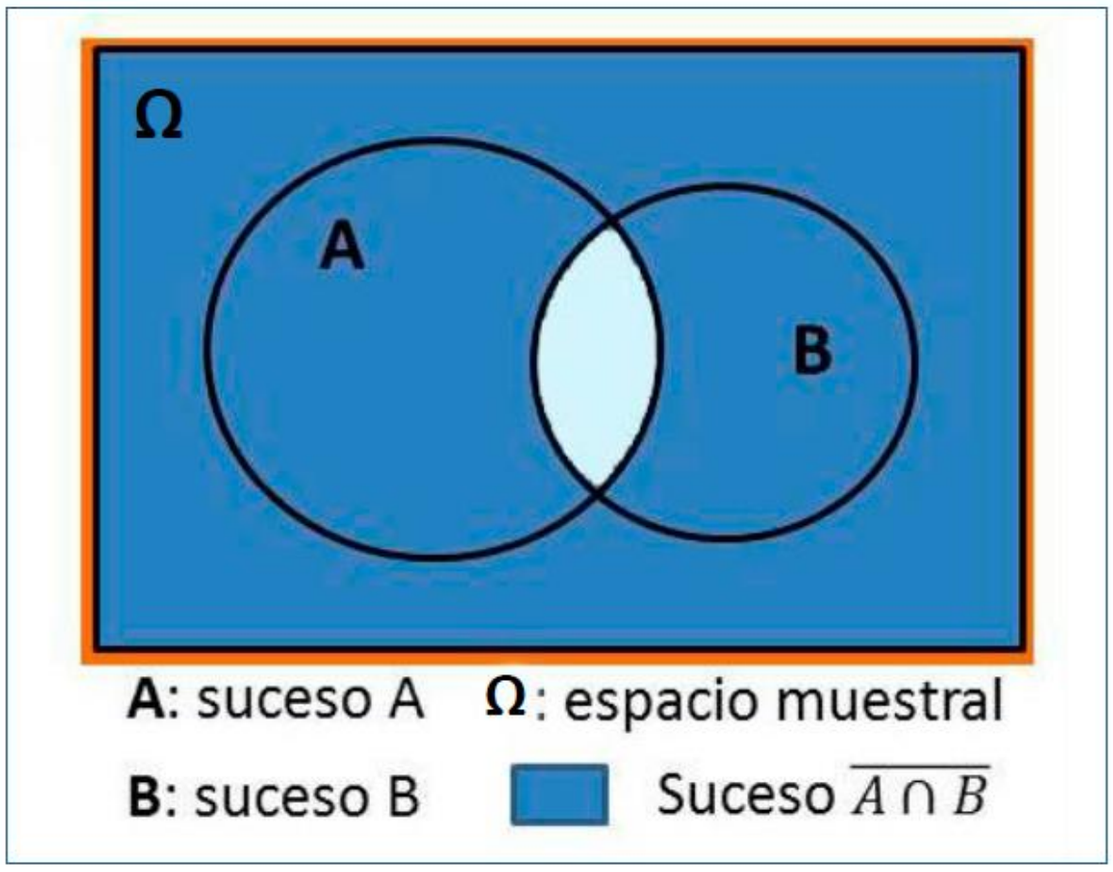
1. Conceptos Básicos: Operaciones con Sucesos

Contrario de la intersección de sucesos

Dados dos sucesos, A y B, se define el contrario de la intersección de ambos como otro suceso, denotado por $\overline{A \cap B}$, que está formado por todos los elementos muestrales excepto aquellos pertenecientes a A y a B simultáneamente. Puede ocurrir A o B o ninguno de los dos, pero no ambos simultáneamente.

$$\overline{A \cap B} = \bar{A} \cup \bar{B}$$

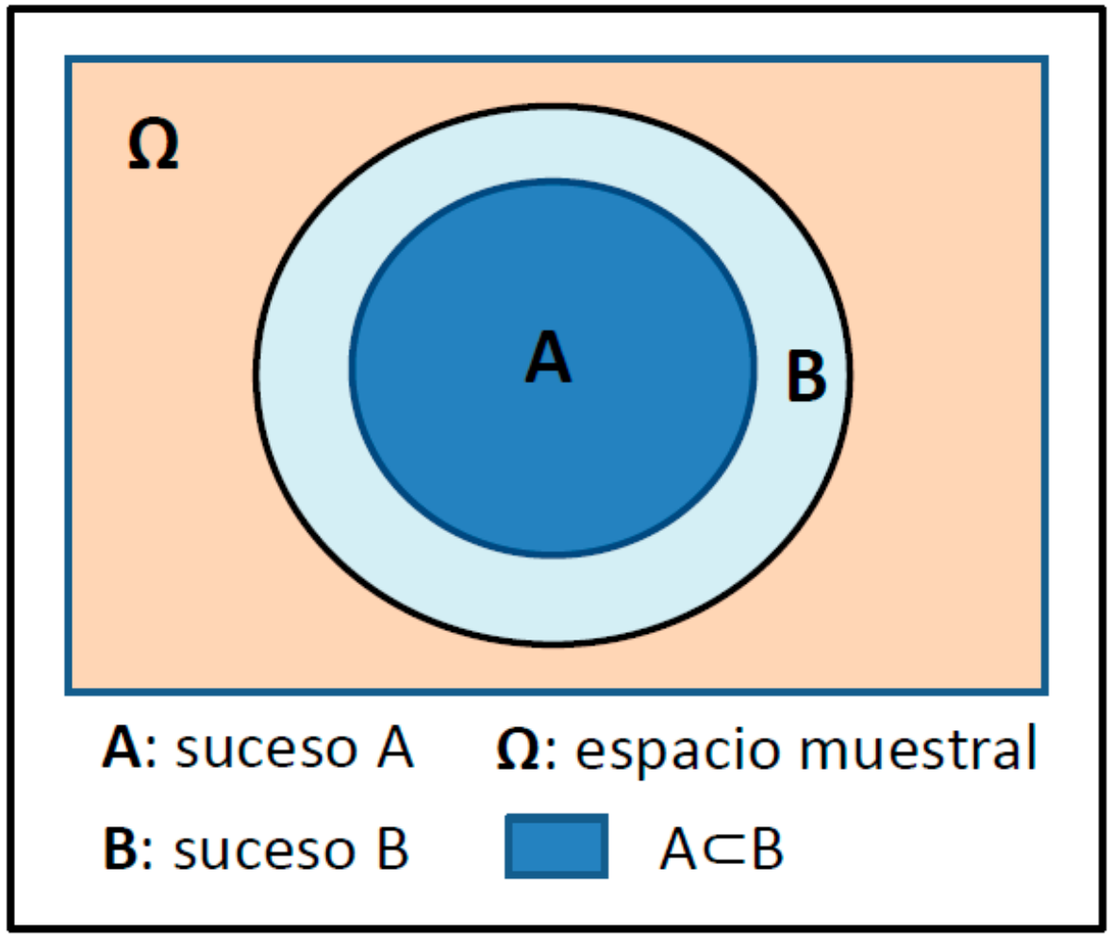
(Ley de Morgan)



1. Conceptos Básicos: Operaciones con Sucesos

Inclusión de sucesos

Dados dos sucesos, A y B, se dice que A está incluido en B, y se denota $A \subset B$, si todos los elementos de A pertenecen también a B, es decir, cada vez que ocurre A está ocurriendo también B.



1. Conceptos Básicos

Propiedades de la unión e intersección

Sean A, B y C conjuntos arbitrarios cualquiera, entonces

1. *Asociativa.* $(A \cup B) \cup C = A \cup (B \cup C)$
   $(A \cap B) \cap C = A \cap (B \cap C)$.
2. *Conmutativa.* $A \cup B = B \cup A$
   $A \cap B = B \cap A$.
3. *Idempotente.* $A \cup A = A$
   $A \cap A = A$.
4. *Elemento ínfimo.* $A \cup \emptyset = A$
   $A \cap \emptyset = \emptyset$.
5. *Distributiva.* $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
   $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

1. Conceptos Básicos: Ejemplo

Sean A, B, C, tres sucesos cualesquiera. Determine expresiones para los siguientes sucesos:

a) Ocurre sólo A.

1. Conceptos Básicos: Ejemplo

b) Ocurren A y B, pero no C.

1. Conceptos Básicos: Ejemplo

c) Ocurren los tres sucesos.

1. Conceptos Básicos: Ejemplo

d) Ocurre al menos uno.

1. Conceptos Básicos: Ejemplo

e) Ocurren al menos dos.

1. Conceptos Básicos: Ejemplo

f) Ocurre uno y sólo uno.

1. Conceptos Básicos: Ejemplo

g) Ocurren dos y sólo dos.

1. Conceptos Básicos: Ejemplo

h) No ocurre ninguno.

2. Definición e interpretación de la probabilidad

Probabilidad de un suceso es una medida numérica de la posibilidad de que ocurra ese suceso.

La *probabilidad de un suceso* A, P(A), toma valores entre 0 y 1. El valor 0 corresponde al suceso imposible, mientras que el valor 1 es para el suceso seguro. Para un suceso, cuanto más probable sea que ocurra, mayor será su *probabilidad*.

$$0 \leq P(A) \leq 1$$

Los sucesos que tienen la misma probabilidad de ocurrir se llaman **sucesos equiprobables**.

2. Definición e interpretación de la probabilidad

Definición clásica de probabilidad. Regla de Laplace

En un espacio muestral finito, Ω, donde todos los elementos tienen la misma probabilidad de aparición, la probabilidad de que ocurra un suceso A es

$$P(A) = \frac{Número\ de\ casos\ favorables\ de\ A}{Número\ de\ casos\ posibles\ de\ \Omega}$$

Ejemplo

Dado un experimento aleatorio consistente en lanzar un dado, el espacio muestral está formado por seis elementos equiprobables.

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

Si el suceso A es sacar un número par, está compuesto por tres puntos muestrales

$$A = \{2, 4, 6\}$$

y su probabilidad es

$$P(A) = \frac{3}{6} = 0{,}5$$

2. Definición e interpretación de la probabilidad

Definición frecuentista de la probabilidad

Cuando el espacio muestral es infinito, o cuando se desconoce si los sucesos son equiprobables, no podríamos aplicar la definición clásica de probabilidad.

Si el experimento se repite n veces y el suceso A_i ocurre n_i veces, la frecuencia relativa observada del suceso es $f(A_i) = n_i / n$

Cabría esperar que, si n se hace muy grande, la frecuencia relativa se estabilizase en torno a la probabilidad de ocurrencia de A.
Así, basado en el principio de regularidad estadística, la probabilidad de un suceso A se obtiene como el límite de su frecuencia relativa de aparición cuando el experimento se repite un número de veces muy grande .

$$P(A) = \lim_{n} f(A_i) = \lim_{n} \frac{n_i}{n}$$

2. Definición e interpretación de la probabilidad

Ejemplo

Si de 150.000 menús servidos en una hamburguesería en 2013, 45.000 son hamburguesa normal (HN), 75.000 son hamburguesa con queso (HQ) y 30.000 son hamburguesa de pollo (HP),

$$P(HN) = \frac{45.000}{150.000} = 0{,}3$$

$$P(HQ) = \frac{75.000}{150.000} = 0{,}5$$

$$P(HP) = \frac{30.000}{150.000} = 0{,}2$$

3. Axiomas del cálculo de probabilidad y sus consecuencias

Dado un experimento con espacio muestral Ω se define una estructura de resultados más compleja que se denomina sigma-algebra de sucesos (Λ) generada con todas las operaciones de unión, intersección y complementariedad que se puedan hacer con los resultados del experimento.

Para que la función de probabilidad esté bien definida, deberá estar definida sobre la sigma-algebra Λ , lo que significa que podremos asignar la probabilidad de ocurrencia a cualquier suceso que definamos.

3. Axiomas de Kolmogorov

Sea Ω un espacio muestral y sea Λ una sigma-álgebra de sucesos sobre Ω. La función P() definida sobre Λ (P: Λ → R) es una función de probabilidad, si verifica los siguientes axiomas (axiomas de Kolmogorov):

$1^o) \; \forall A \in \Lambda \quad P(A) \geq 0$

$2^o) \; P(\Omega) = 1$

$3^o)$ Para cualquier $A_1, A_2 .... \in \Lambda$ tal que $A_i \cap A_j = \phi \; \forall \; i \neq j$

se verifica que $P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$

La terna (Ω, Λ, P) se denomina espacio de probabilidad

3. Axiomas de Kolmogorov

Textualmente, los tres axiomas pueden expresarse como:

I. La probabilidad de un suceso es mayor o igual que cero.

II. La probabilidad del suceso seguro vale 1.

III. La probabilidad de la unión de varios sucesos incompatibles es igual a la suma de sus probabilidades.

La probabilidad definida como frecuencia relativa verifica las condiciones impuestas por Kolmogorov:

a) La probabilidad de un suceso es mayor o igual que cero.

Dado que $0 \leq n_i \leq n$ entonces $0 \leq \frac{n_i}{n} \leq 1$

En ocasiones se enuncia este axioma como que la probabilidad es un número no negativo menor o igual que uno.

b) La probabilidad del suceso seguro vale 1.

Si consideramos como suceso seguro la ocurrencia de cualquiera de los sucesos definidos, su probabilidad será

dado que $n = n_1 + n_2 + \ldots + n_k$

su probabilidad será $\frac{n_1 + n_2 + \ldots + n_k}{n} = 1$

c) La probabilidad de la unión de varios sucesos incompatibles es igual a la suma de sus probabilidades.

Dados dos sucesos incompatibles A_i y A_j, con frecuencias absolutas n_i y n_j, la unión de los dos sucesos tiene como frecuencia absoluta $n_i + n_j$, y su probabilidad será

$$P(A_i \cup A_j) = \frac{n_i + n_j}{n} = \frac{n_i}{n} + \frac{n_j}{n} = P(A_i) + P(A_j)$$

Ejemplo

Si de 150.000 menús servidos en una hamburguesería en 2011, 45.000 son hamburguesa normal (HN), 75.000 son hamburguesa con queso (HQ) y 30.000 son hamburguesa de pollo (HP),

$$P(HN) = 0{,}3 \qquad P(HQ) = 0{,}5 \qquad P(HN) + P(HQ) = 0{,}3 + 0{,}5 = 0{,}8$$

La probabilidad de que el menú pedido sea de HN o HQ es

$$P(HN \cup HQ) = \frac{45.000 + 75.000}{150.000} = 0{,}8$$

Consecuencias de los axiomas

De los tres axiomas de Kolmogorov puede deducirse un conjunto de propiedades o consecuencias que completan la operatividad del cálculo de probabilidades.

1) Dado un suceso A, la probabilidad del suceso complementario $\overline{A}$ es: $P(\overline{A}) = 1 - P(A)$

Como A y $\overline{A}$ son sucesos complementarios $\rightarrow A \cup \overline{A} = \Omega$

$$\text{Entonces} \begin{cases} P(A \cup \overline{A}) = P(\Omega(= 1 & \leftarrow \text{Axioma II} \\ P(A \cup \overline{A}) = P(A) + P(\overline{A}) & \leftarrow \text{Axioma III} \end{cases}$$

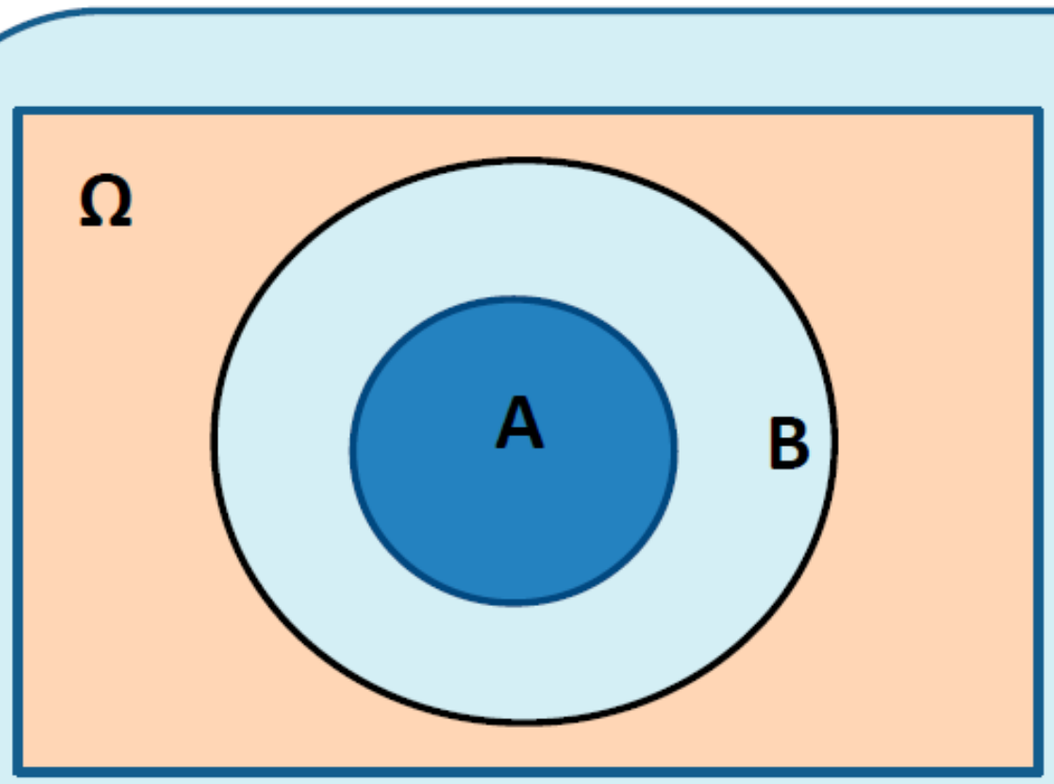
Por lo tanto $\Rightarrow P(A) + P(\overline{A}) = 1 \rightarrow P(\overline{A}) = 1 - P(A)$

2) La probabilidad del suceso imposible es nula.

$$\phi = \overline{\Omega} \rightarrow P(\phi) = P(\overline{\Omega}) = 1 - P(\Omega) = 1 - 1 = 0$$

3. Axiomas del cálculo de probabilidad y sus consecuencias

3) Si $A \subseteq B$ entonces $P(A) \leq P(B)$.



$B = A \cup (B-A) \rightarrow P(B) = P(A) + P(B-A)$ (Axioma III)

$P(B-A) \geq 0 \rightarrow P(B) \geq P(A)$ (Axioma I)

3. Axiomas del cálculo de probabilidad y sus consecuencias

4) La probabilidad de un suceso A es menor o igual que uno.

$$P(A) \leq 1$$

Cualquier suceso A está formado por un subconjunto de puntos del espacio muestral, es decir, está contenido en el espacio muestral

$$A \subset \Omega$$

entonces,

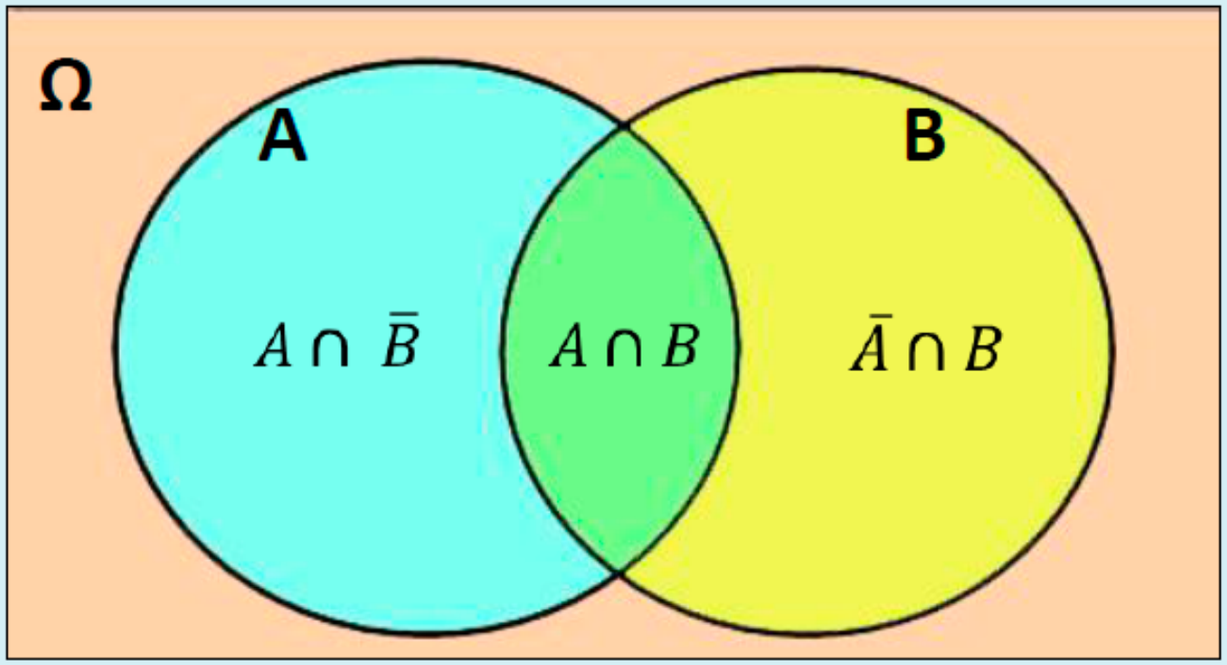
$$P(A) \leq P(\Omega) = 1 \qquad \text{(Axioma II)}$$

Por lo tanto,

$$P(A) \leq 1$$

5) Para dos sucesos cualesquiera, A y B, la probabilidad de la unión de esos sucesos es

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

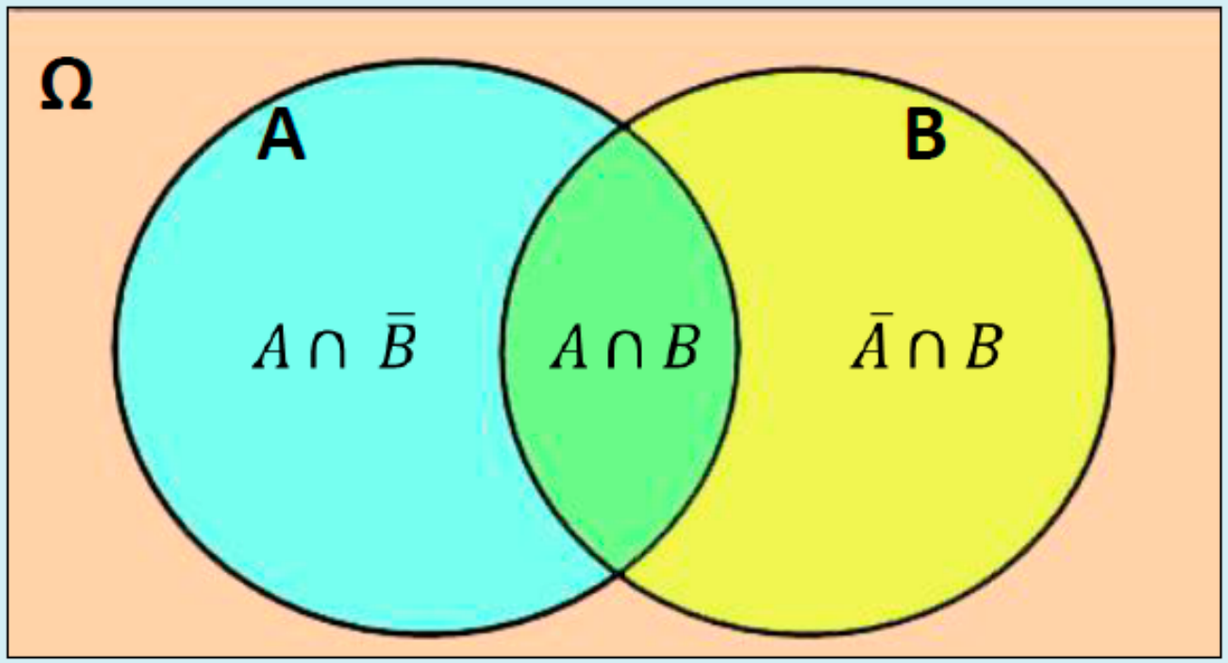


Ω: espacio muestral

A: suceso A

B: suceso B

Si sumamos las probabilidades de A y de B, la probabilidad del suceso A ∩ B se está sumando dos veces, por tanto, si se desea obtener la probabilidad de A ∪ B como la suma de las probabilidades de A y B, debe restarse la probabilidad de la intersección.



$(A \cap \bar{B}), (A \cap B)$ y $(\bar{A} \cap B)$

Son sucesos incompatibles

$$A = (A \cap \bar{B}) \cup (A \cap B) \rightarrow P(A) = P(A \cap \bar{B}) + P(A \cap B)$$

$$B = (\bar{A} \cap B) \cup (A \cap B) \rightarrow P(B) = P(\bar{A} \cap B) + P(A \cap B)$$

$$A \cup B = (A \cap \bar{B}) \cup (\bar{A} \cap B) \cup (A \cap B)$$

$$P(A \cup B) = P(A \cap \bar{B}) + P(\bar{A} \cap B) + P(A \cap B)$$

entonces,

$$P(A) + P(B) = P(A \cap \bar{B}) + P(A \cap B) + P(\bar{A} \cap B) + P(A \cap B) =$$

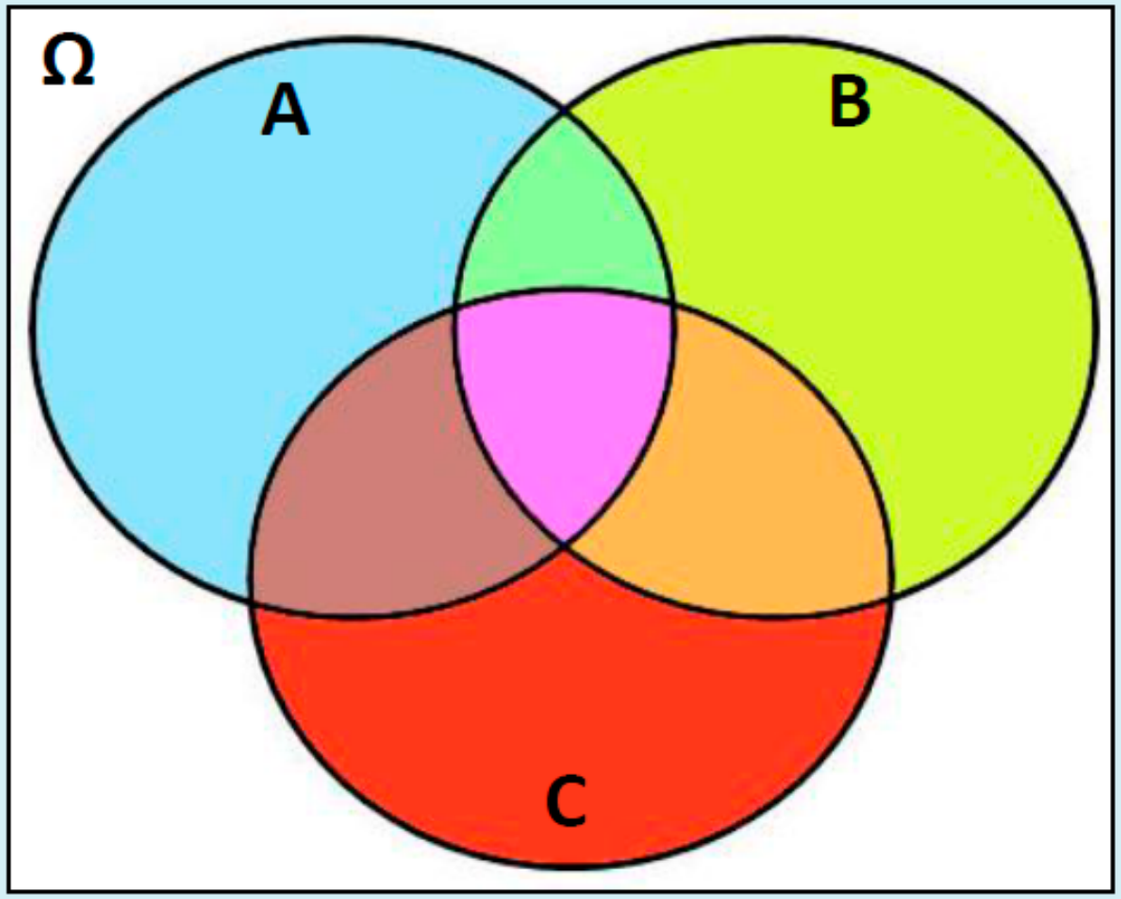
$$= P(A \cup B) + P(A \cap B)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Este resultado es extensible para la unión de más sucesos. Para el caso de tres sucesos cualesquiera, A, B y C, la probabilidad de la unión de esos sucesos es

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B)$$

$$-P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$



Si sumamos las probabilidades de A, de B y de C, las probabilidades de los sucesos de las distintas intersecciones se están sumando repetidas veces, por tanto, si se desea obtener la probabilidad de A U B U C como la suma de las probabilidades de A, B y C, debe depurarse la repetición de las intersecciones.

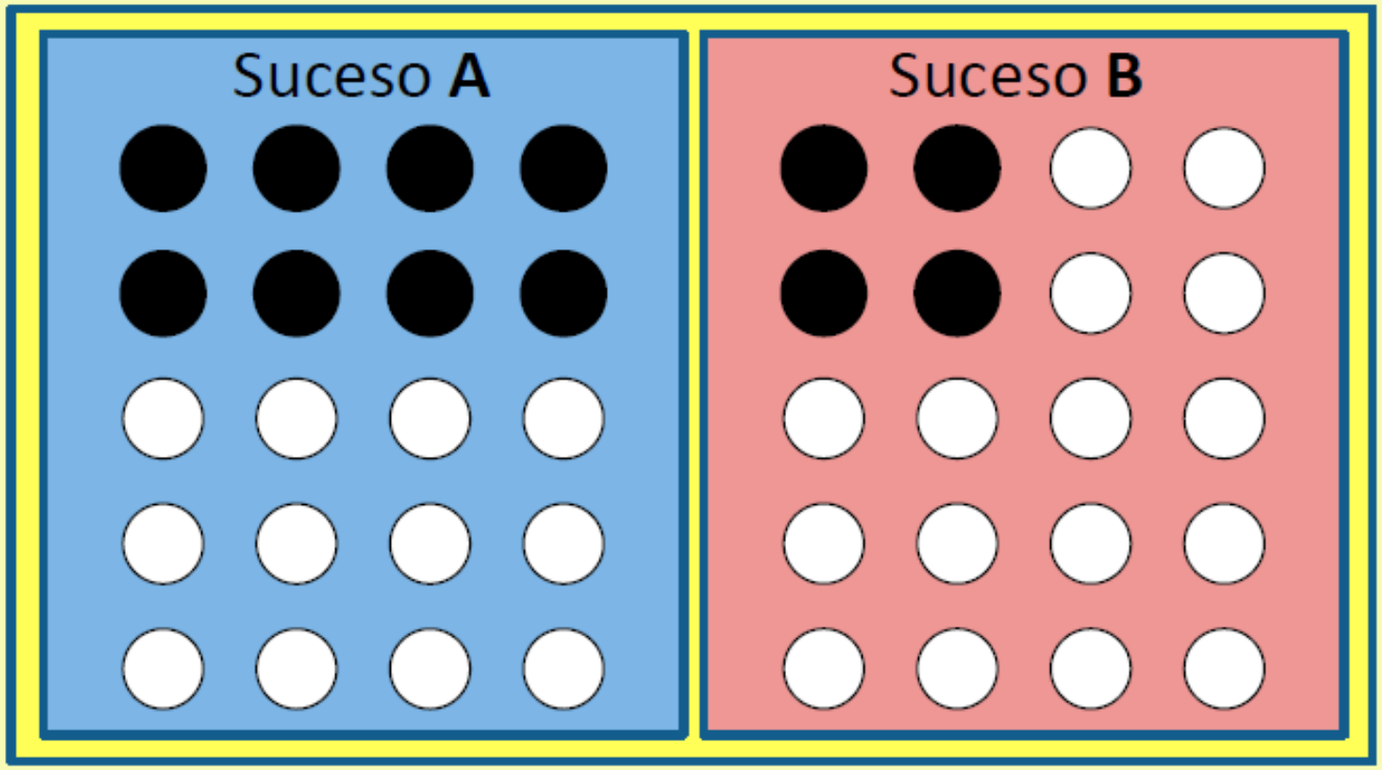
4. Probabilidad Condicionada

Para dos sucesos cualesquiera, A y B, contenidos en un espacio muestral, podemos considerar la probabilidad de que ocurra B, P(B), la probabilidad de que ocurran simultáneamente A y B, $P(A \cap B)$ o bien podemos considerar la probabilidad de que ocurra B una vez que ya ha ocurrido A. En este último caso el espacio muestral queda reducido a A y se busca la probabilidad de B en ese reducto. Esta es la **probabilidad condicionada de B dado A**, P(B/A) y se define, para P(A) > 0, como

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

Ejemplo

Sean los sucesos
A: “coger la caja azul”
B: “coger la caja roja”
C: “sacar una bola negra”
D: “sacar una bola blanca”



La probabilidad de sacar una bola negra, mezclando las dos cajas, es

$$P(C) = \frac{12}{40} = 0{,}3$$

Pero si ya ha ocurrido que se ha elegido la caja azul, la probabilidad de sacar una bola negra es

$$P(C/A) = \frac{8}{20} = 0{,}4 \quad \frac{P(A \cap C)}{P(A)} = \frac{8/40}{20/40} = 0{,}4$$

¿Cuál es la probabilidad de sacar una bola blanca?, ¿y si se ha elegido la caja roja?

Ejemplo

Una hamburguesería sirve 150.000 menús a hombres y mujeres en 2011, repartidos de la siguiente manera:

Hamburguesa	Hombre (H)	Mujer (M)	Total
Normal (HN)	30.000	15.000	45.000
Con queso (HQ)	50.000	25.000	75.000
De pollo (HP)	10.000	20.000	30.000
Total	90.000	60.000	150.000

$$P(H \cap HN) = \frac{30.000}{150.000} = 0{,}2 \qquad P(H) = \frac{90.000}{150.000} = 0{,}6$$

La probabilidad de que un cliente que es hombre pida una hamburguesa normal es:

$$P(HN/H) = \frac{30.000}{90.000} = 0{,}33$$

4. Probabilidad Condicionada

Probabilidad condicionada y los axiomas de Kolmogorov

La probabilidad condicionada cumple los tres axiomas de Kolmogorov.

a) Dado un suceso B, la probabilidad condicionada de B dado A es mayor o igual que cero.

$$P(B/A) = \frac{P(A \cap B)}{P(A)} \geq 0$$

b) Dado el espacio muestral Ω, la probabilidad condicionada de Ω dado A es igual a uno.

$$P(\Omega/A) = \frac{P(\Omega \cap A)}{P(A)} = \frac{P(A)}{P(A)} = 1$$

c) La probabilidad condicionada de la unión de varios sucesos incompatibles es igual a la suma de sus respectivas probabilidades condicionadas.

$$P(B \cup C / A) = \frac{P((B \cup C) \cap A)}{P(A)} = \frac{P((B \cap A) \cup (C \cap A))}{P(A)}$$

como B y C son disjuntos, la probabilidad del numerador se obtiene mediante suma de los dos términos de la unión

$$P(B \cup C / A) = \frac{P(B \cap A) + P(C \cap A)}{P(A)} = \frac{P(B \cap A)}{P(A)} + \frac{P(C \cap A)}{P(A)} =$$

$$= P(B / A) + P(C / A)$$

Si se cumple para dos sucesos incompatibles, por extensión se cumple para la unión de cualquier número de sucesos incompatibles.

Regla de la multiplicación de probabilidades

Partiendo de la definición de probabilidad condicionada, P(B/A), se puede expresar en forma de producto y obtener la regla de la multiplicación o probabilidad compuesta:

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

Y análogamente,

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

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

De lo anterior se concluye que

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

4. Probabilidad Condicionada

La regla de la multiplicación de probabilidades se puede extender a n sucesos. Por ejemplo, para tres sucesos,

$$P\left(C/(A\cap B)\right)=\frac{P\left(C\cap(A\cap B)\right)}{P\left((A\cap B)\right)}=\frac{P\left(A\cap B\cap C\right)}{P\left(A\cap B\right)}$$

de donde, $P\left(A\cap B\cap C\right)=P\left(A\cap B\right)\cdot P\left(C/(A\cap B)\right)$

a su vez, $P\left(A\cap B\right)=P(A)\cdot P(B/A)$

y por tanto:

$$P\left(A\cap B\cap C\right)=P(A)\cdot P\left(B/A\right)\cdot P\left(C/A\cap B\right)$$

Ejemplo

Una hamburguesería sirve 150.000 menús a hombres y mujeres en 2011, repartidos de la siguiente manera:

Hamburguesa	Hombre (H)	Mujer (M)	Total
Normal (HN)	30.000	15.000	45.000
Con queso (HQ)	50.000	25.000	75.000
De pollo (HP)	10.000	20.000	30.000
Total	90.000	60.000	150.000

Las probabilidades de que un cliente sea hombre o mujer son:

$$P(H) = \frac{90.000}{150.000} = 0{,}6$$

$$P(M) = \frac{60.000}{150.000} = 0{,}4$$

Las probabilidades de que un cliente pida uno u otro menú son:

$$P(HN) = \frac{45.000}{150.000} = 0{,}3$$

$$P(HQ) = \frac{75.000}{150.000} = 0{,}5$$

$$P(HP) = \frac{30.000}{150.000} = 0{,}2$$

Ejemplo

Las probabilidades de cada menú condicionadas al sexo del cliente son:

Si el cliente es hombre

$$P(HN/H) = \frac{30.000}{90.000} = 0{,}3333$$

$$P(HQ/H) = \frac{50.000}{90.000} = 0{,}5556$$

$$P(HP/H) = \frac{10.000}{90.000} = 0{,}1111$$

Si el cliente es mujer

$$P(HN/M) = \frac{15.000}{60.000} = 0{,}25$$

$$P(HQ/M) = \frac{25.000}{60.000} = 0{,}4167$$

$$P(HP/M) = \frac{20.000}{60.000} = 0{,}3333$$

4. Probabilidad Condicionada

Ejemplo

Se pueden obtener las probabilidades de las intersecciones de sucesos a partir de la regla de la multiplicación:

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

$$P(H \cap HN) = 0{,}6 \cdot 0{,}3333 = 0{,}2$$

$$P(H \cap HQ) = 0{,}6 \cdot 0{,}5556 = 0{,}3333$$

$$P(H \cap HP) = 0{,}6 \cdot 0{,}1111 = 0{,}0667$$

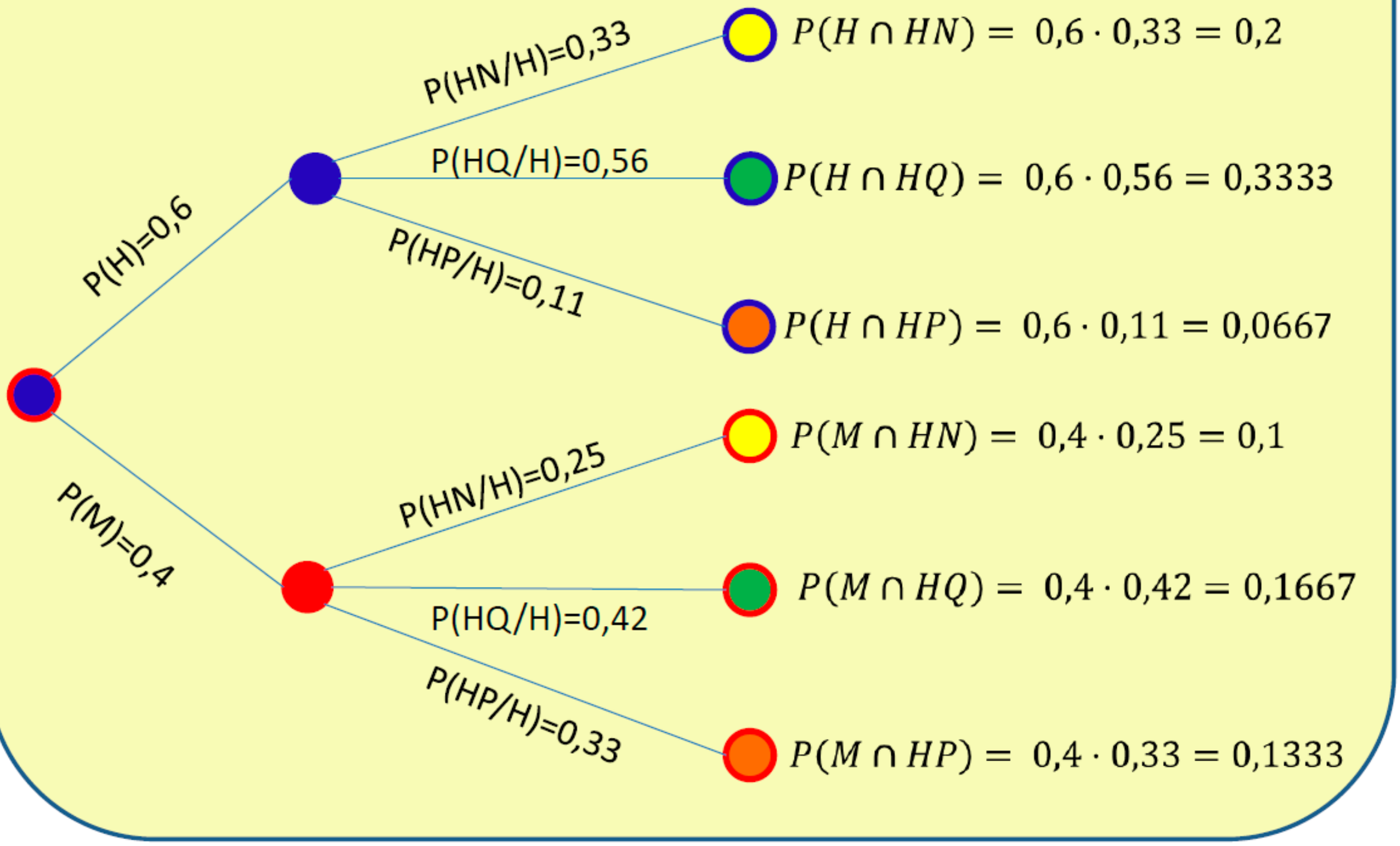
$$P(M \cap HN) = 0{,}4 \cdot 0{,}25 = 0{,}1$$

$$P(M \cap HQ) = 0{,}4 \cdot 0{,}4167 = 0{,}1667$$

$$P(M \cap HP) = 0{,}4 \cdot 0{,}3333 = 0{,}1333$$

Ejemplo

También se puede construir un árbol de probabilidades:



5. Teorema de la probabilidad total y teorema de Bayes

Teorema de la probabilidad total

Un conjunto de sucesos $\{A_i\}$ es una PARTICIÓN de Ω si:

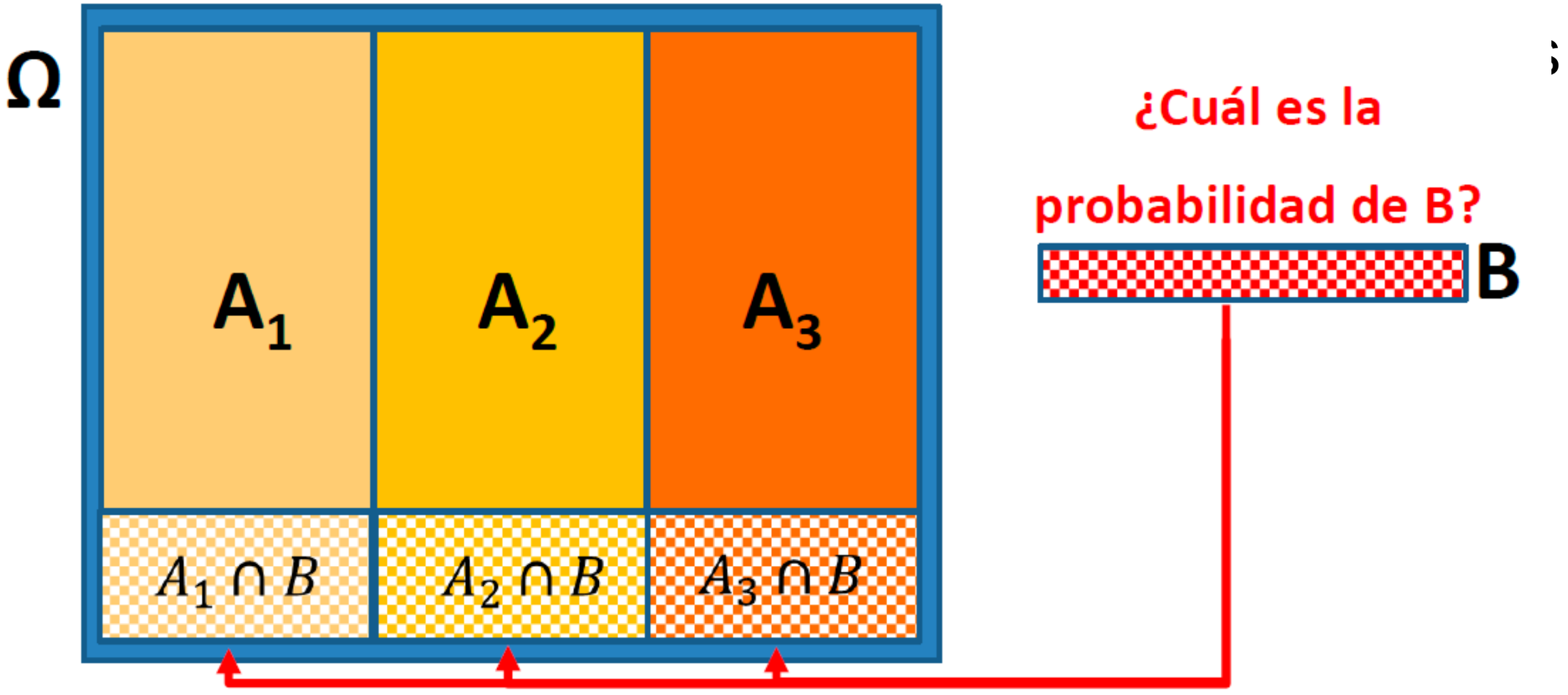
Su probabilidad no es nula $P(A_i) > 0, \quad i = 1,2,\dots,n$

Son incompatibles $A_i \cap A_j = \emptyset, \quad i \neq j$

Forman un sistema completo $\bigcup_{i=1}^{n} A_i = \Omega$

Entonces, para cualquier suceso B, su probabilidad puede expresarse en función de las probabilidades condicionadas de B dados los diferentes sucesos A_i.

$$P(B) = \sum_i P(A_i) \cdot P(B / A_i)$$



Dado que los sucesos A_i son incompatibles, la probabilidad de B es la suma de las probabilidades de las tres intersecciones.

$$P(B) = \sum_{i=1}^{3} P(A_i \cap B) = \sum_{i=1}^{3} P(A_i) \cdot P(B/A_i)$$

5. Teorema de la probabilidad total y teorema de Bayes: Ejemplo

Una entidad bancaria trabaja con empresas de distintos sectores de actividad. Del historial de préstamos de los últimos años, se han sacado los siguientes datos sobre la composición de la clientela según el sector de actividad al que pertenece y del porcentaje de préstamos no devueltos en cada uno de los tres sectores:

Sectores	% Empresas	% Morosos
Primario	20 %	4 %
Secundario	30 %	6 %
Terciario	50 %	5 %

5. Teorema de la probabilidad total y teorema de Bayes

Podemos definir los sucesos y probabilidades

S_1: la empresa es del sector primario

S_2: la empresa es del sector secundario

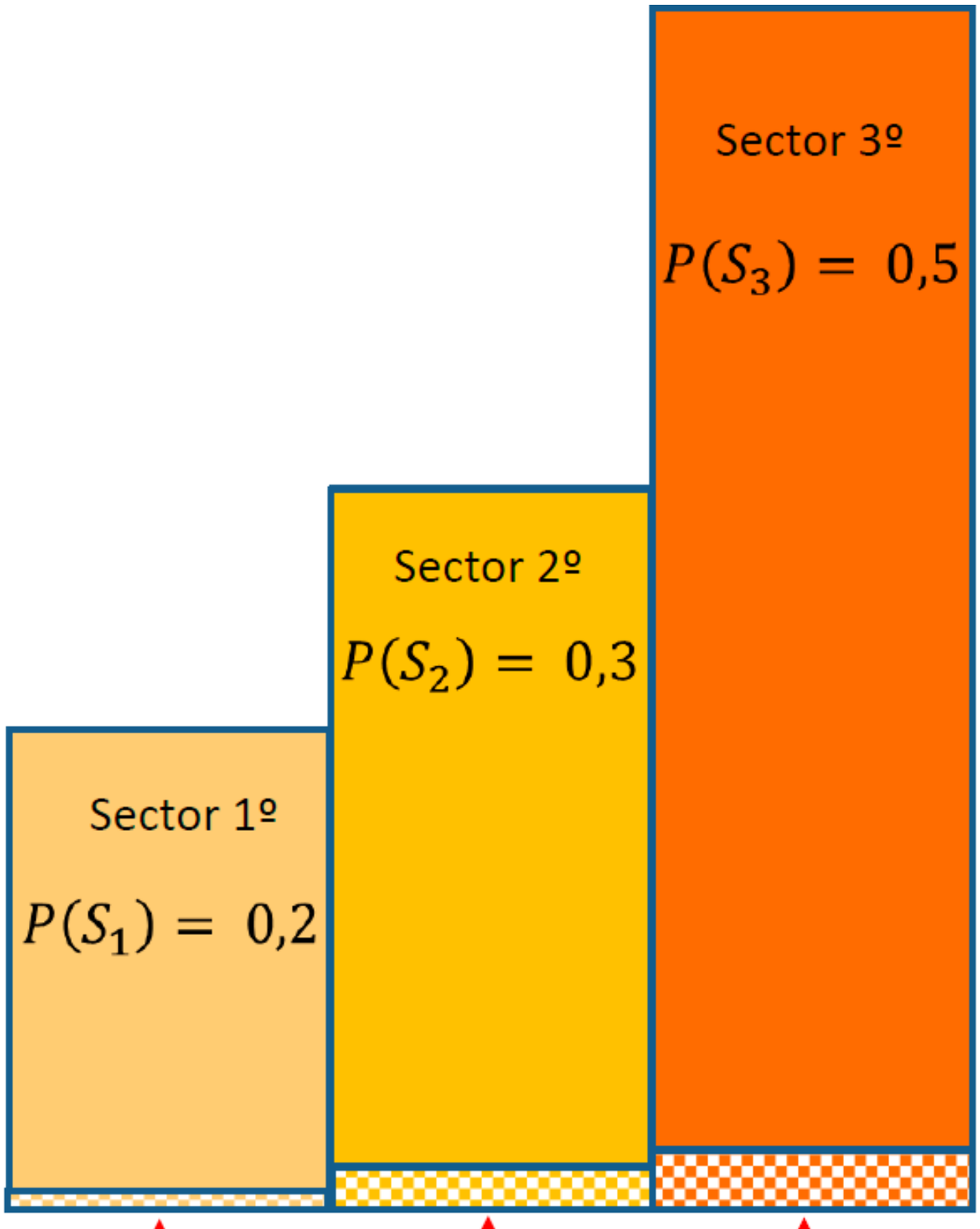
S_3 : la empresa es del sector terciario

M: la empresa no devuelve los préstamos

$$P(S_1) = 0{,}2 \qquad P(S_2) = 0{,}3 \qquad P(S_3) = 0{,}5$$

$$P(M/S_1) = 0{,}04 \qquad P(M/S_2) = 0{,}06 \qquad P(M/S_3) = 0{,}05$$

5. Teorema de la probabilidad total y teorema de Bayes



$$P(M/S_1) = 0{,}04$$

$$P(M/S_2) = 0{,}06$$

$$P(M/S_3) = 0{,}05$$

¿Cuál es la probabilidad de que una empresa sea morosa?

5. Teorema de la probabilidad total y teorema de Bayes

¿Cuál es la probabilidad de que una empresa sea morosa?

Las empresas morosas son:

El 4 % de las empresas del sector primario, que son el 20 % del total, es decir, el 4 % del 20 % del total

$0{,}2 \cdot 0{,}04$ es decir, $P(S_1) \cdot P(M/S_1)$

El 6 % de las empresas del sector secundario, que son el 30 % del total, es decir, el 6 % del 30 % del total

$0{,}3 \cdot 0{,}06$ es decir, $P(S_2) \cdot P(M/S_2)$

El 5 % de las empresas del sector secundario, que son el 50 % del total, es decir, el 5 % del 50 % del total

$0{,}5 \cdot 0{,}05$ es decir, $P(S_3) \cdot P(M/S_3)$

5. Teorema de la probabilidad total y teorema de Bayes

El total de morosidad entre los clientes del banco será:

$$0{,}2 \cdot \ 0{,}04 + 0{,}3 \cdot 0{,}06 + 0{,}5 \cdot 0{,}05 = 0{,}051 \quad (5{,}1\ \%)$$

es decir,

$$P(S_1) \cdot P(M/S_1) + P(S_2) \cdot P(M/S_2) + P(S_3) \cdot P(M/S_3)$$

que es la expresión del teorema de la probabilidad total:

$$P(M) = \sum_{i=1}^{3} P(S_i) \cdot P(M/S_i)$$

5. Teorema de la probabilidad total y teorema de Bayes

Teorema de Bayes (1)

Sean dos sucesos A y B tales que P(A)>0 y P(B)>0

Entonces :

$$P(A \cap B) = P(A)\,P(B/A) = P(B)\,P(A/B)$$

Y despejando tendremos:

$$P(A/B) = \frac{P(B/A)\cdot P(A)}{P(B)}$$

5. Teorema de la probabilidad total y teorema de Bayes

5. Teorema de la probabilidad total y teorema de Bayes

Teorema de Bayes (2)

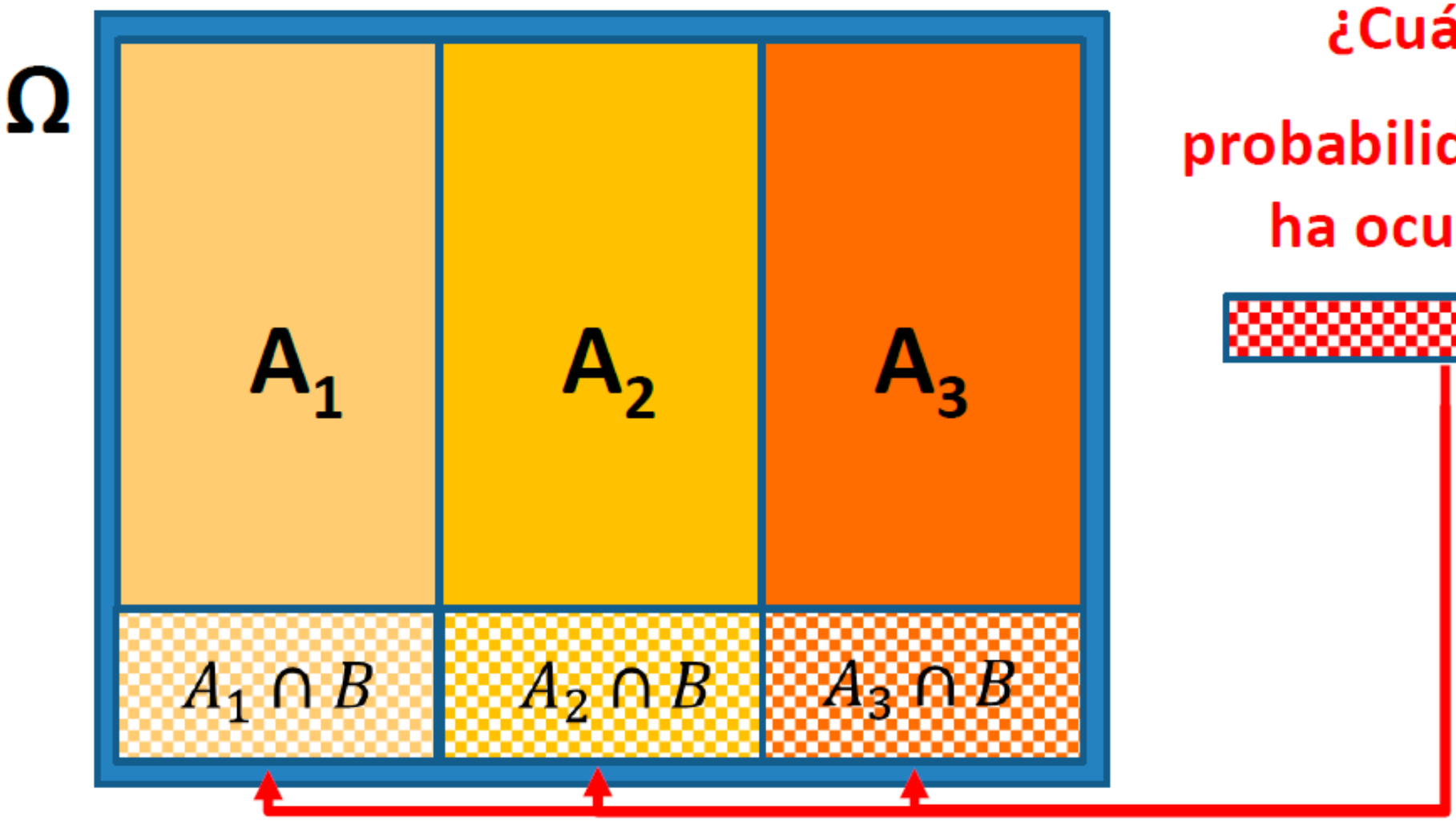
El resultado anterior puede ser generalizado usando una partición de Ω y haciendo uso del teorema de la probabilidad total.

Sea $\{A_i\}$ es una partición de Ω y sea B tal que P(B)>0

Entonces:

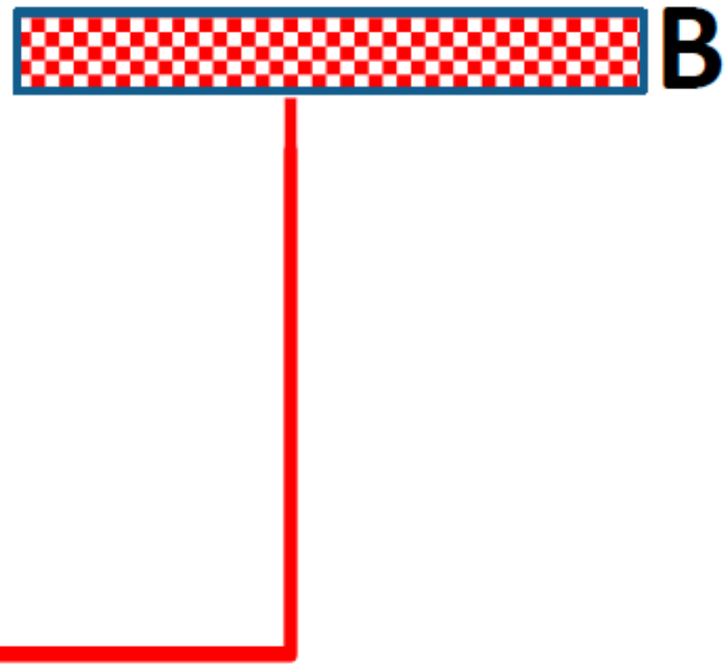
$$P(A_j / B) = \frac{P(B / A_j) \cdot P(A_j)}{\sum_i P(B / A_i) \cdot P(A_i)}$$

5



S

¿Cuál es la probabilidad de A_1 si ha ocurrido B?



Dado que los sucesos A_i son incompatibles, la probabilidad de B es la suma de las probabilidades de las tres intersecciones. Entonces,

$$P(A_1/B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{P(A_1) \cdot P(B/A_1)}{\sum_{i=1}^{3} P(A_i) \cdot P(B/A_i)}$$

5. Teorema de la probabilidad total y teorema de Bayes: Ejemplo

¿Cuál es la probabilidad de que una empresa sea del sector primario entre las empresas morosas?

El "porcentaje" de empresas morosas lo hemos calculado como

$$0{,}2 \cdot 0{,}04 + 0{,}3 \cdot 0{,}06 + 0{,}5 \cdot 0{,}05 = 0{,}051$$

$0{,}2 \cdot 0{,}04$ es la parte de ese porcentaje debida al sector primario

y el cociente, $\dfrac{0{,}2 \cdot 0{,}04}{0{,}051}$

es "la parte" de morosos que pertenecen al sector primario.

5. Teorema de la probabilidad total y teorema de Bayes

El cociente $\dfrac{0{,}2 \cdot 0{,}04}{0{,}051} = 0{,}1569 \quad (15{,}69\ \%)$

es la proporción de empresas del sector primario que hay en el conjunto de las empresas morosas, es decir, la probabilidad de que una empresa morosa sea del sector primario.

En términos de probabilidades tenemos que,

$$P(S_1/M) = \frac{P(S_1) \cdot P(M/S_1)}{P(S_1) \cdot P(M/S_1) + P(S_2) \cdot P(M/S_2) + P(S_3) \cdot P(M/S_3)}$$

que es la expresión del teorema de Bayes.

6. Independencia de Sucesos

Al presentar la **probabilidad condicionada** se afirmaba que no es lo mismo la probabilidad inicial o *a priori* de un suceso B que la probabilidad "condicionada" a que sepamos que haya ocurrido otro suceso A.

Puede ser que el que otro suceso A sea cierto:

- **favorezca** las posibilidades de ocurrencia de B

$$P(B/A) > P(B)$$

- o que las **desfavorezca** , es decir,

$$P(B/A) < P(B)$$

6. Independencia de Sucesos

Pero también puede suceder que el hecho de que haya ocurrido A no tenga ningún efecto sobre la probabilidad de ocurrencia de B, es decir,

$$P(B/A) = P(B)$$

Se dice entonces que A y B son **sucesos independientes**.

Se puede definir la independencia de sucesos por el cumplimiento de una de las tres siguientes condiciones, que son equivalentes:

1. $P(B/A) = P(B) \qquad para\ P(B) > 0$

2. $P(A/B) = P(A) \qquad para\ P(A) > 0$

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

6. Independencia de Sucesos

Según lo anterior, si el suceso A es independiente del suceso B, es lo mismo que decir que B es independiente de A, pero, además, también se puede afirmar lo mismo sobre sus sucesos complementarios.

Si A y B son dos sucesos independientes, entonces también lo son:

$$A \text{ y } \overline{B}, \qquad \overline{A} \text{ y } B \qquad \text{y} \qquad \overline{A} \text{ y } \overline{B}$$

6. Independencia de Sucesos

La **independencia de n sucesos** es una extensión de la definición anterior, para lo que se exige que sean independientes dos a dos (condición que se conoce como **mutuamente independientes**) y además la probabilidad de la intersección de todos ellos debe ser el producto de las n probabilidades.

Por ejemplo, para el caso de tres sucesos, A, B y C, son independientes si se cumplen **todas** las siguientes condiciones:

1. $P(A \cap B) = P(A) \cdot P(B)$
2. $P(A \cap C) = P(A) \cdot P(C)$
3. $P(B \cap C) = P(B) \cdot P(C)$
4. $P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$

6. Independencia de Sucesos: Ejemplo

Un campus universitario ofrece a sus alumnos unas instalaciones deportivas, un servicio de comedores y el acceso por medio de transporte público. Los alumnos usan estos servicios del campus según los siguientes datos:

- El 20 % usa las instalaciones deportivas.
- El 30 % es usuario de los servicios de comedor.
- El 50 % acceden en transporte público.
- El 6 % usan las instalaciones deportivas y el comedor.
- El 20 % utilizan el comedor y el transporte público.
- El 10 % son usuarios de las instalaciones deportivas y el transporte público.

6. Independencia de Sucesos

¿Son independientes el uso de uno u otro servicio?

Consideremos los sucesos

A: Usar las instalaciones deportivas.

B: Utilizar el servicio de comedores.

C: Acceder en transporte público.

En términos de probabilidades, la información anterior se puede expresar como

$$P(A) = 0{,}2 \qquad P(B) = 0{,}3 \qquad P(C) = 0{,}5$$

$$P(A \cap B) = 0{,}06 \qquad P(B \cap C) = 0{,}20 \qquad P(A \cap C) = 0{,}10$$

6. Independencia de Sucesos

Se puede comprobar la independencia mediante el cumplimiento de la primera o la segunda condición:

$$¿P(B/A) = P(B)?$$

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{0{,}06}{0{,}2} = 0{,}3 = P(B) \rightarrow$$ A y B son independientes

$$P(B/C) = \frac{P(B \cap C)}{P(C)} = \frac{0{,}20}{0{,}5} = 0{,}4 \neq P(B) \rightarrow$$ B y C no son independientes

$$P(A/C) = \frac{P(A \cap C)}{P(C)} = \frac{0{,}10}{0{,}5} = 0{,}2 = P(A) \rightarrow$$ A y C son independientes

La condición "usar el transporte público" favorece el uso de comedores pues *P(B/C) > P(B).*

6. Independencia de Sucesos

También se puede comprobar la independencia mediante el cumplimiento de la tercera condición:

$$¿P(A \cap B) = P(A) \cdot P(B)?$$

$P(A \cap B) = 0{,}06$

$P(A) \cdot P(B) = 0{,}2 \cdot 0{,}3 = 0{,}06$

→ A y B son independientes

$P(B \cap C) = 0{,}20$

$P(B) \cdot P(C) = 0{,}3 \cdot 0{,}5 = 0{,}15$

→ B y C no son independientes

$P(A \cap C) = 0{,}10$

$P(A) \cdot P(C) = 0{,}2 \cdot 0{,}5 = 0{,}10$

→ A y C son independientes

Ejemplo

Una entidad bancaria tiene tres sucursales en la ciudad. Cada una está situada en un barrio con distintas características socioeconómicas.

Cuando un préstamo supera los 300.000 euros, la información para su análisis y autorización se pasa a la central de riesgos de esa entidad.

Del total de préstamos recibidos en la central el primer trimestre del pasado año, 120 provenían de la sucursal A, 140 de B y 100 de C. Ahora, transcurrido un año, se conocen los porcentajes de morosidad en cada sucursal para ese tipo de operaciones: 3 % en A, 4 % en B y 2,5 % en C.

Las probabilidades resultantes de la experiencia pasada pretenden aplicarse a las operaciones de este año.

Ejemplo

¿Qué probabilidades se obtienen directamente de los datos facilitados?

Los datos referentes al número de préstamos concedidos permiten hallar la probabilidad de que uno de los préstamos superiores a 300.000 euros pertenezca a una u otra sucursal.

Definimos los sucesos

A: El préstamo proviene de la sucursal A.

B: El préstamo proviene de la sucursal B.

C: El préstamo proviene de la sucursal C.

Y denotamos P(A), P(B) y P(C) respectivamente a la probabilidad de que un préstamo provenga de la sucursal A, B o C.

Ejemplo

Las probabilidades de ocurrencia de los diferentes sucesos se obtienen a través de las frecuencias relativas.

$$P(A) = \frac{120}{360} = 0{,}3333 \quad (33{,}33\%)$$

$$P(B) = \frac{140}{360} = 0{,}3889 \quad (38{,}89\%)$$

$$P(C) = \frac{100}{360} = 0{,}2778 \quad (27{,}78\%)$$

Ejemplo

Por otra parte, los datos sobre la morosidad se refieren al porcentaje de morosos dentro de cada sucursal, es decir, el suceso "morosidad" condicionado a que el préstamo provenga de una u otra sucursal. Por lo tanto los porcentajes facilitados son las probabilidades de la morosidad condicionada a la sucursal de origen.

$$P(M/A) = 0{,}03$$
$$P(M/B) = 0{,}04$$
$$P(M/C) = 0{,}025$$

Ejemplo

¿Qué probabilidades se pueden obtener a partir del teorema de la probabilidad total?

El teorema de la probabilidad total permite aunar las informaciones sobre probabilidades condicionadas para ofrecer la probabilidad del suceso libre de condiciones.

Se precisa disponer de las probabilidades condicionadas a sucesos que sean excluyentes (incompatibles) y completos (que su unión forme el espacio muestral al completo).

En este caso, el espacio muestral son los préstamos superiores a 300.000 euros y provienen de una de las tres sucursales A, B o C, por lo que los sucesos A, B y C cumplen esas premisas.

Ejemplo

Dado que disponemos de las probabilidades del suceso "morosidad", M, condicionado a la procedencia de cada una de las tres sucursales, se puede calcular, a partir de la aplicación del teorema de la probabilidad total, la probabilidad "total" del suceso morosidad, es decir, la morosidad independientemente de dónde proceda ese préstamo.

Anteriormente se enunció el teorema de la probabilidad total como

$$P(B) = \sum_{i=1}^{n} P(A_i) \cdot P(B/A_i)$$

En este ejemplo, B se corresponde al suceso "morosidad", M, y los distintos sucesos A_i son los sucesos de pertenencia a las sucursales, A, B y C.

Ejemplo

Por lo tanto, la aplicación del teorema de la probabilidad total a este ejemplo será:

$$P(M) = P(A) \cdot P(M/A) + P(B) \cdot P(M/B) + P(C) \cdot P(M/C)$$

Y sustituyendo por los valores conocidos de las probabilidades,

$$P(M) = 0{,}3333 \cdot 0{,}03 + 0{,}3889 \cdot 0{,}04 + 0{,}2778 \cdot 0{,}025 = 0{,}0325$$

$$P(M) = 0{,}0325 \quad (3{,}25\ \%)$$

Luego la probabilidad de que un préstamo analizado por la central de riesgos sea impagado es del 3,25 %. Es la probabilidad, para esta entidad bancaria, de que una operación de préstamo superior a 300.000 euros sea impagada.

Ejemplo

¿Qué probabilidades se pueden obtener a partir del teorema de Bayes?

Visto lo anterior, el teorema de Bayes permite "invertir" la condición.

Si estamos ante un préstamo impagado, ¿qué probabilidad tiene de ser un préstamo procedente de la sucursal A?, por ejemplo.

Ahora se trata de la probabilidad de que un préstamo sea de una u otra sucursal condicionado a que ya ha resultado impagado.

Hallar

$P(A/M)$ o $P(B/M)$ o $P(C/M)$

Conocidas

$P(A)$, $P(B)$, $P(C)$, $P(M/A)$, $P(M/B)$, $P(M/C)$

Ejemplo

El teorema de Bayes ha sido anteriormente enunciado como

$$P(A_i/B) = \frac{P(A_i) \cdot P(B/A_i)}{\sum_{i=1}^{n} P(A_i) \cdot P(B/A_i)}$$

De nuevo, cabe decir que en este ejemplo, B se corresponde al suceso “morosidad”, M, y los distintos sucesos A_i son los sucesos de pertenencia a las sucursales, A, B y C.

Obsérvese que el denominador se corresponde con la expresión del teorema de la probabilidad total y así, por ejemplo, la probabilidad de ser de A una empresa morosa es

$$P(A/M) = \frac{P(A) \cdot P(M/A)}{P(M)}$$

Ejemplo

Si el analista de riesgos encuentra un préstamo moroso:

- La probabilidad de ser un préstamo procedente de la sucursal A es

$$P(A/M) = \frac{P(A) \cdot P(M/A)}{P(M)} = \frac{0{,}3333 \cdot 0{,}03}{0{,}0325} = 0{,}3077 \quad (30{,}77\,\%)$$

- La probabilidad de que el préstamo provenga de la sucursal B es

$$P(B/M) = \frac{P(B) \cdot P(M/B)}{P(M)} = \frac{0{,}3889 \cdot 0{,}04}{0{,}0325} = 0{,}4786 \quad (47{,}86\,\%)$$

- La probabilidad de que el préstamo sea de la sucursal C es

$$P(C/M) = \frac{P(C) \cdot P(M/C)}{P(M)} = \frac{0{,}2778 \cdot 0{,}025}{0{,}0325} = 0{,}2137 \quad (21{,}37\,\%)$$

En resumen,

Conocidas	Teorema	Se puede hallar	Expresión
$P(A_1)$ $P(A_2)$ . . . $P(A_n)$	Probabilidad total	P(B)	$\sum_{i=1}^{n} P(A_i) \cdot P(B/A_i)$
$P(B/A_1)$ $P(B/A_2)$ . . . $P(B/A_n)$	Bayes	$P(A_i/B)$	$\frac{P(A_i) \cdot P(B/A_i)}{\sum_{i=1}^{n} P(A_i) \cdot P(B/A_i)}$

Ejemplo

Ejemplo

Estadística Aplicada al Marketing

T.5. Variables aleatorias unidimensionales

Juan Miguel Navarro Ruiz

Grado en Marketing y Dirección Comercial

Juan M Navarro- Tlf: (+34) 968 27 88 25
Universidad Católica San Antonio de Murcia - Tlf: (+34) 968 27 88 00 info@ucam.edu - www.ucam.edu

1. Definición de variable aleatoria.
2. Función de distribución.
3. Variables aleatorias discretas: función de probabilidad y función de distribución.
4. Variables aleatorias continuas: función de densidad y función de distribución.
5. Esperanza matemática. Propiedades.
6. Otras características de una variable aleatoria.

1. Definición Variable Aleatoria

En un experimento aleatorio, los sucesos pueden estar definidos en términos cuantitativos o cualitativos, no obstante, se suele asignar un valor numérico a cada uno de los posibles elementos que componen el espacio muestral. Esta relación entre los puntos muestrales y el valor numérico se establece mediante la variable aleatoria. La variable aleatoria es una variable cuyos valores están asociados a los sucesos elementales de un experimento aleatorio.

Una **variable aleatoria** es una función que asigna un valor numérico a cada punto del espacio muestral.

1. Definición Variable Aleatoria

Los valores posibles de la variable aleatoria pueden ser un conjunto numerable o no numerable, por lo que se pueden distinguir dos tipos de variables aleatorias: discretas y continuas.

Variable aleatoria discreta es aquella que puede tomar un número finito, o infinito numerable, de valores posibles.

Variable aleatoria continua es aquella que puede tomar un número infinito no numerable de posibles valores, incluso aunque se limiten a un grupo de intervalos de la recta real.

2. Función de distribución

Función de distribución

Dada una variable aleatoria X, llamaremos **función de distribución**, F(x), a una función acumulativa de las probabilidades de la variable X, de modo que F(x) es la probabilidad de que la variable X tome valores iguales o inferiores a x.

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

por tanto : $0 \leq F(x) \leq 1$

2. Función de distribución

Propiedades de la función de distribución

1. $F(-\infty) = 0$; $F(+\infty) = 1$

Esta propiedad se debe entender como que el límite de F(x) es cero cuando x tiende a menos infinito, y que vale 1 cuando tiende a infinito.

$$F(-\infty) = \lim_{x \to -\infty} F(x) = 0$$

$$F(+\infty) = \lim_{x \to +\infty} F(x) = 1$$

Si bien, puede entenderse intuitivamente de la siguiente manera:

$$F(-\infty) = P(X \leq -\infty) = P(\emptyset) = 0$$

$$F(+\infty) = P(X \leq +\infty) = P(\Omega) = 1$$

2. Función de distribución

2. F(x) es una función no decreciente.

La función de distribución es monótona no decreciente, es decir, que si $x < y$, entonces $F(x) \leq F(y)$.

Si $x < y$, entonces todos los valores menores que x son a su vez menores que y, y por tanto el conjunto de sucesos asociados a los valores de la variable menores o iguales a x está contenido en el conjunto de sucesos asociados a los valores de la variable inferiores o iguales a y. De este modo, la $P(X \leq x)$ es menor o igual a la $P(X \leq y)$, o lo que es lo mismo, $F(x) \leq F(y)$.

$$\text{Si } x < y \Rightarrow \{X \leq x\} \subset \{Y \leq y\} \Rightarrow P(X \leq x) \leq P(y \leq y) \Rightarrow$$
$$\Rightarrow F(x) \leq F(y)$$

2. Función de distribución

3. La función de distribución es continua por la derecha.

$$\lim_{\substack{y \to x \\ y > x}} F(y) = F(x)$$

Nota: más adelante se verá que la función de distribución puede no ser continua por la izquierda, en particular para las variables discretas.

Toda función de distribución debe cumplir esas tres propiedades, y cualquier función que las cumpla es una función de distribución de una variable aleatoria.

Observar además que:

1) Sea a < b $\Rightarrow P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$

2) $P(X > x) = 1 - P(X \leq x) = 1 - F(x)$

3. Variables aleatorias discretas: función de probabilidad y función de distribución

Función de probabilidad (v.a. discretas)

Dada una variable aleatoria discreta X, llamaremos **función de probabilidad** o distribución de probabilidad, P(x), a una función que asigna probabilidades de ocurrencia a los posibles valores de esa variable cumpliendo las dos siguientes condiciones:.

$$1)\ P(x_i) \geq 0 \quad \forall i$$

$$2)\ \sum_i P(x_i) = 1$$

3. Variables aleatorias discretas: función de probabilidad y función de distribución

Ejemplo 1: Función de probabilidad

Sea el experimento lanzar un dado perfecto.

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

X=resultado del lanzamiento

X_i	1	2	3	4	5	6
$P(X_i)$	1/6	1/6	1/6	1/6	1/6	1/6

3. Variables aleatorias discretas: función de probabilidad y función de distribución

Función de distribución (v.a. discreta)

Para una variable aleatoria discreta X, la **función de distribución** se obtiene a partir de la **función de probabilidad**. Así, F(x) resultará ser la suma de las probabilidades de los x_i inferiores o iguales al valor x:

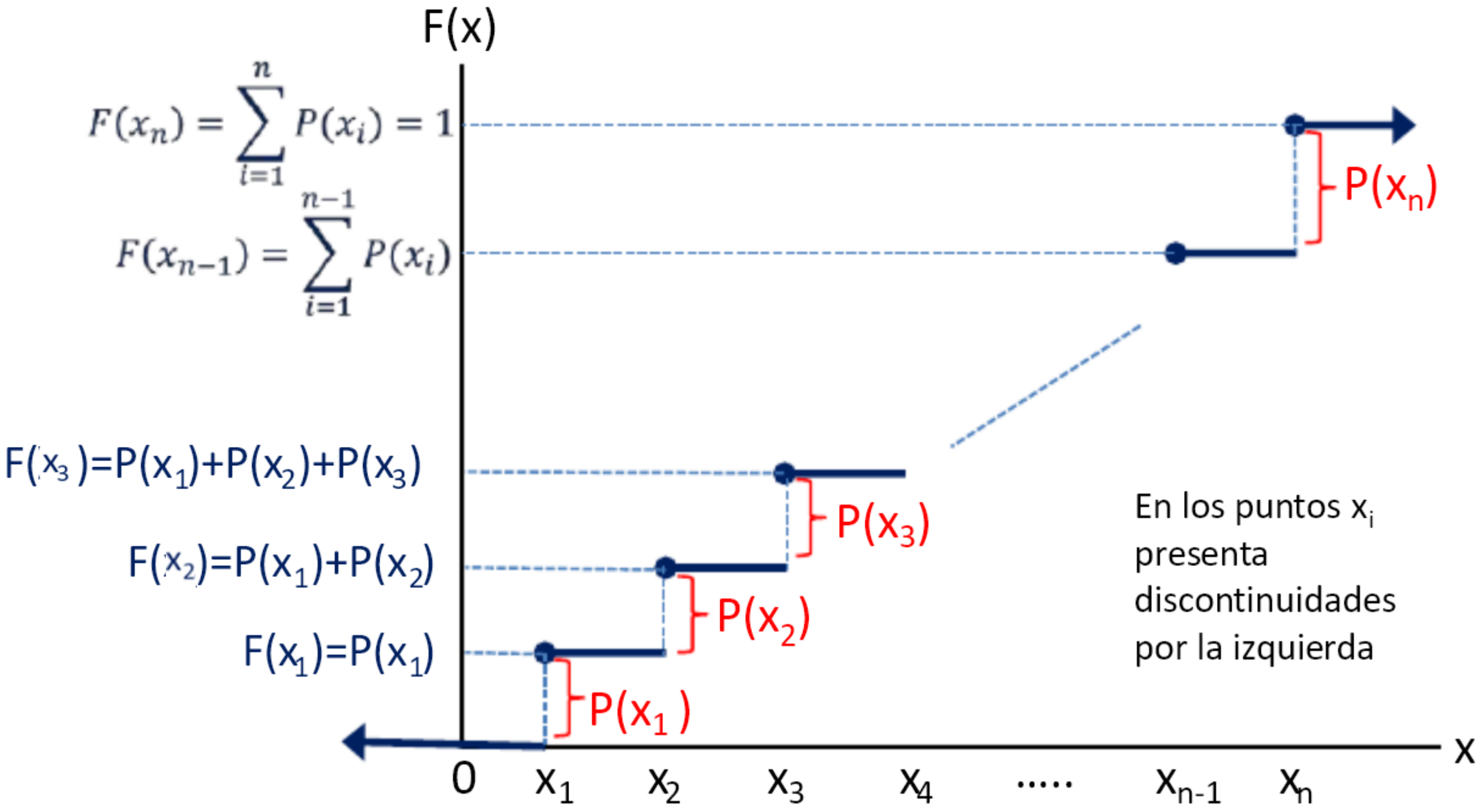
$$F(x) = P(X \leq x) = \sum_{x_i \leq x} p(x_i)$$

A su vez, si para una variable aleatoria discreta X conocemos la **función de distribución**, F(x), podremos obtener su **función de probabilidad**:

$$P(x_i) = P(X = x_i) = P(x_{i-1} < X \leq x_i) = F(x_i) - F(x_{i-1})$$

3. Variables aleatorias discretas: función de probabilidad y función de distribución

De manera gráfica, la función de distribución de una variable discreta se comporta del siguiente modo:



3. Variables aleatorias discretas: función de probabilidad y función de distribución

Ejemplo 2: Funciones de probabilidad y de distribución

Sea el experimento lanzar dos monedas equilibradas.

Ω={(cara , cara); (cruz , cara); (cara , cruz); (cruz , cruz)}

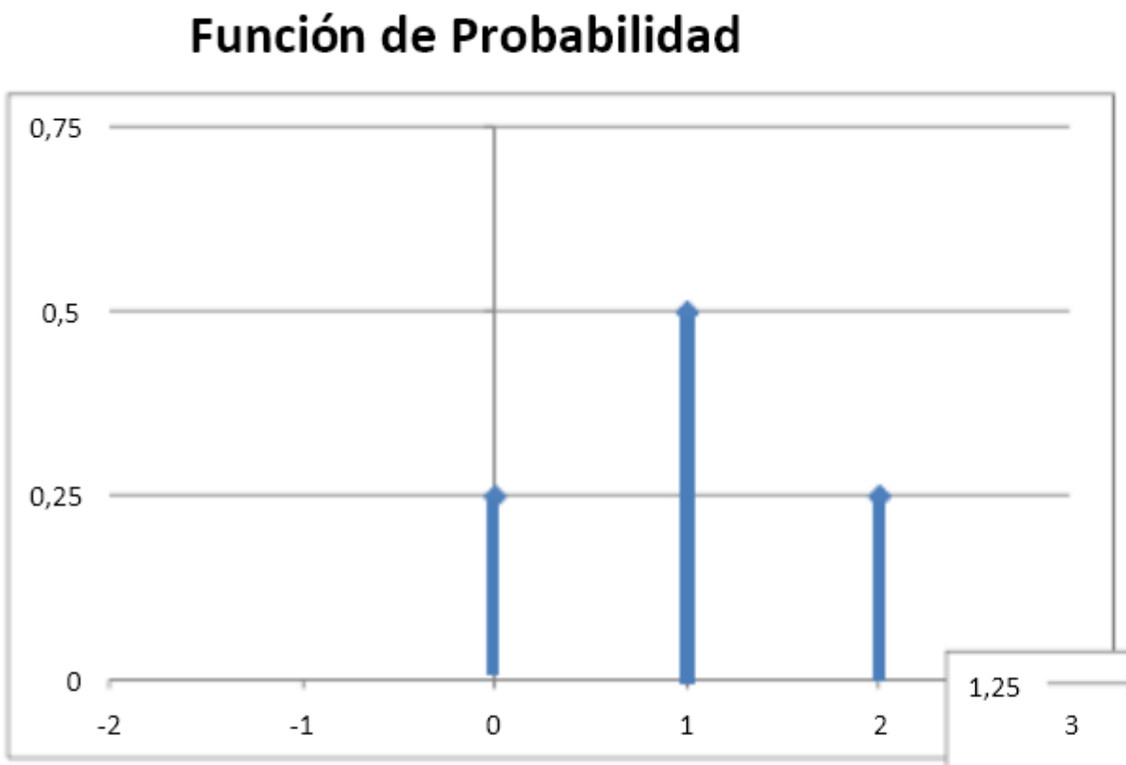
X=número de caras

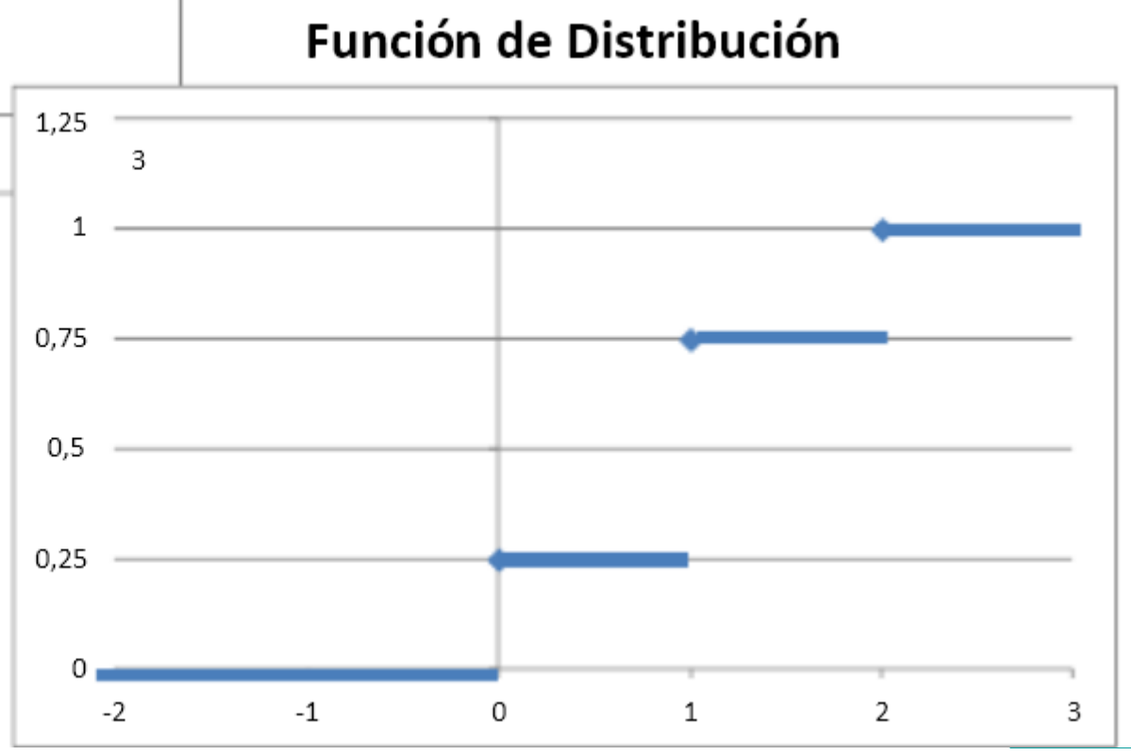
X_i	0	1	2
$P(X_i)$	¼	½	¼

$$F(x)=\begin{cases} 0 & si \quad x<0 \\ 1/4 & si \quad 0\le x<1 \\ 3/4 & si \quad 1\le x<2 \\ 1 & si \quad 2\le x \end{cases}$$

3. Variables aleatorias discretas: función de probabilidad y función de distribución

Ejemplo 2: Funciones de probabilidad y de distribución





3. Variables aleatorias discretas: función de probabilidad y función de distribución

Ejemplo 3: Funciones de probabilidad y de distribución

Sea X una v.a. con función de probabilidad:

$$p_k = P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \text{ para } k = 0,1,2,\ldots\ldots \text{ con } \lambda > 0$$

Podemos comprobar que es una función de probabilidad:

$$p_k \geq 0$$

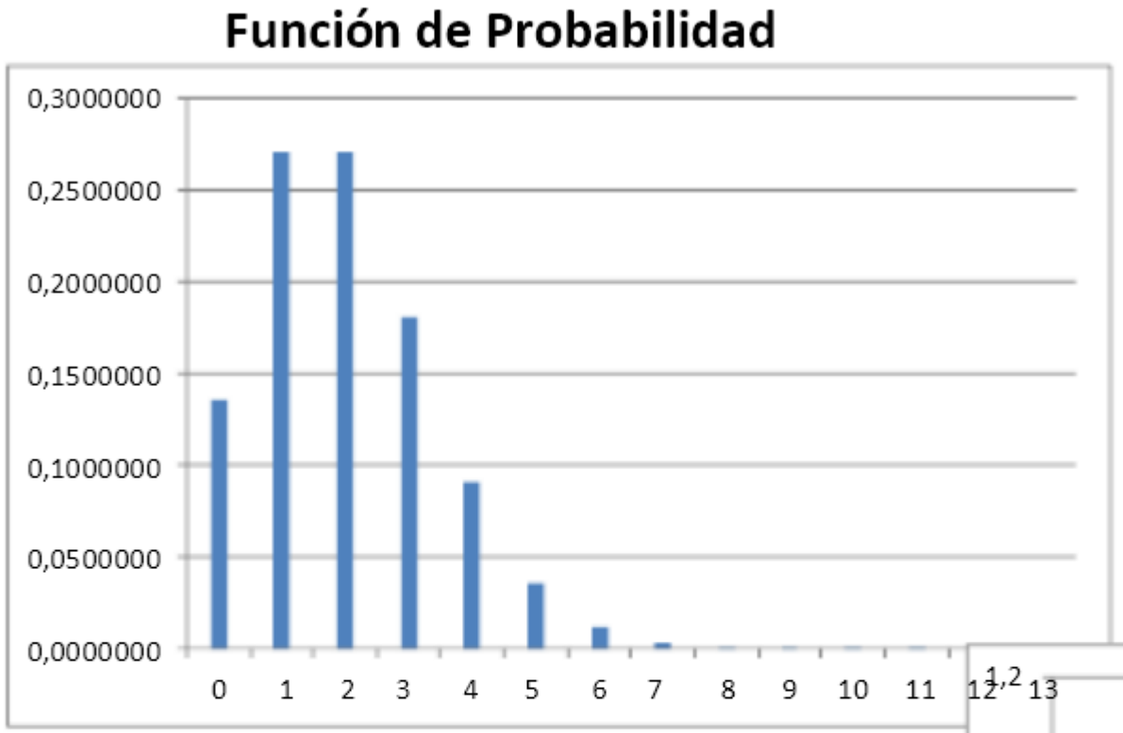
$$\sum_{k=0}^{\infty} p_k = \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = e^0 = 1$$

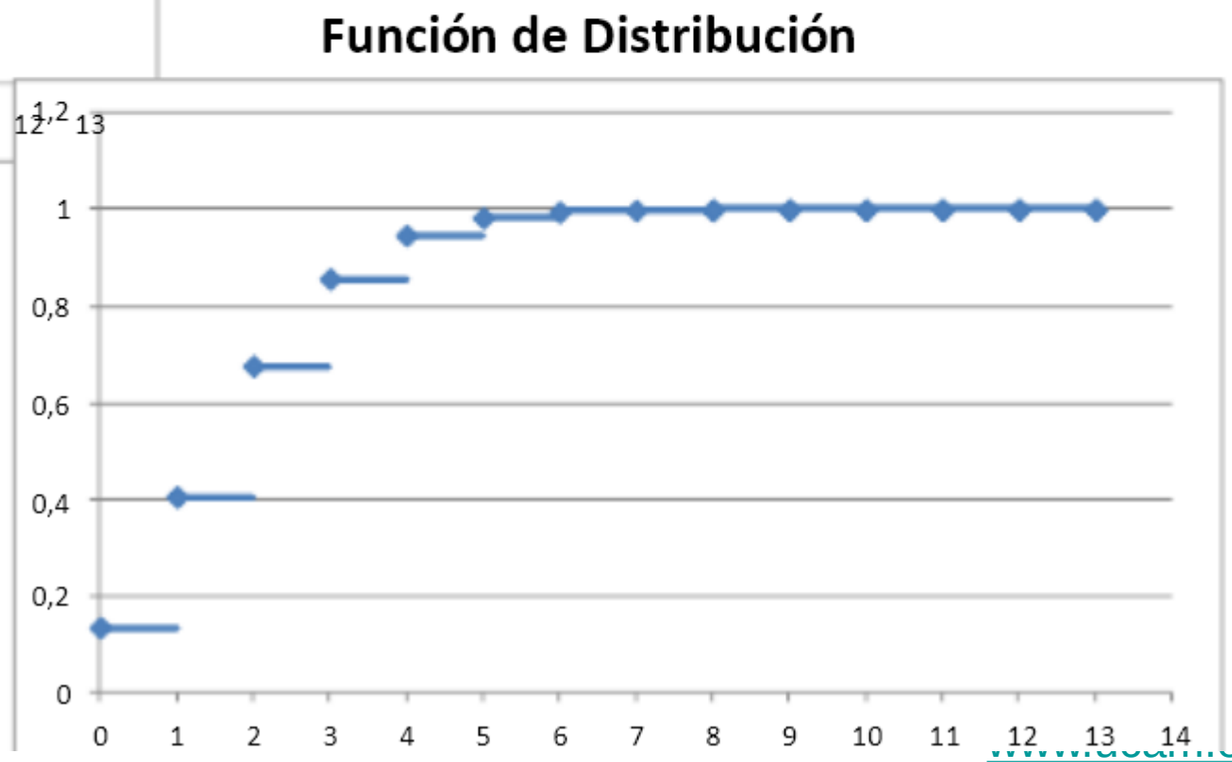
Para $\lambda=2$

K	0	1	2	3	4	5	6	7
P_k	0,1353	0,2707	0,2707	0,1804	0,0902	0,0361	0,0120	0,0034
F(k)	0,1353	0,4060	0,6767	0,8571	0,9473	0,9834	0,9955	0,9989

3. Variables aleatorias discretas: función de probabilidad y función de distribución

Ejemplo 3: Funciones de probabilidad y de distribución





4. Variables aleatorias continuas: función de densidad y función de distribución

Introducción

Cuando la variable aleatoria puede tomar cualquier valor sobre un intervalo de la recta real se trata de una variable continua.

De manera intuitiva, dado que los valores que puede tomar la variable son infinitos no numerables no les podremos asignar individualmente una probabilidad y que el total sume la unidad, como ocurre en el caso discreto.

De hecho, como veremos más adelante, en el caso continuo la probabilidad de un valor puntual de la variable es cero.

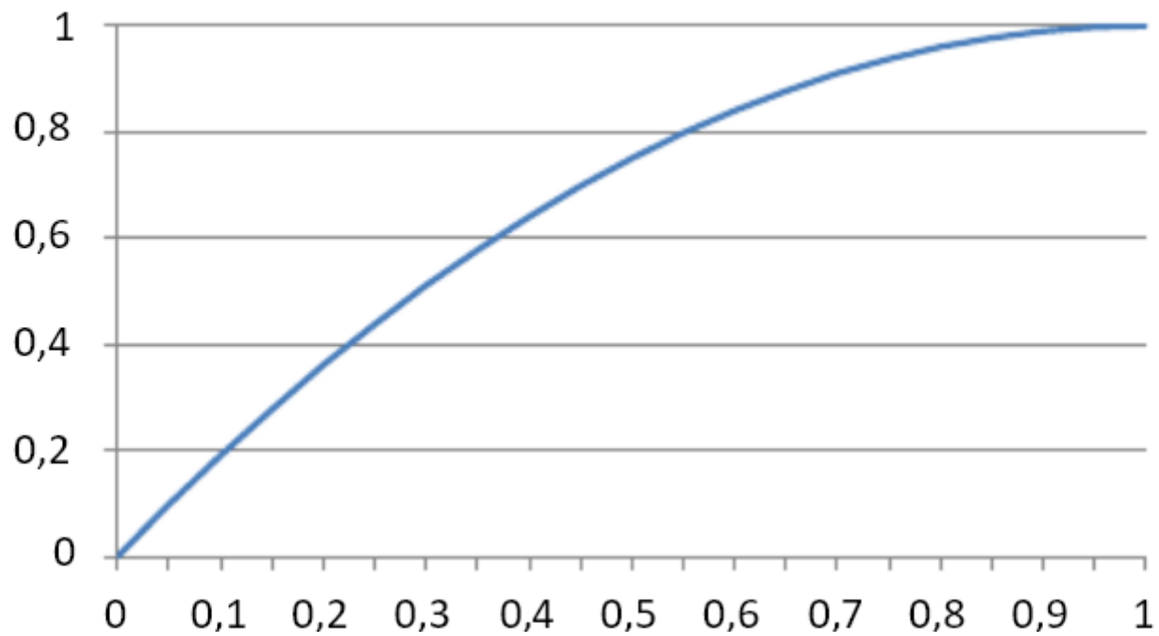
4. Variables aleatorias continuas: función de densidad y función de distribución

Variable aleatoria continua

Una variable aleatoria es continua si su función de distribución, F(x), es absolutamente continua (es continua y su primera derivada existe y es continua) en el intervalo en el que esté definida dicha variable.

La función de $F(x)= 2x-x^2$ en el intervalo [0 ; 1] es la función de distribución de una variable aleatoria continua definida en el intervalo [0; 1], (asumimos que F(x)=0 si x<0 y F(x)=1 si x>1), puesto que:

a) F(0)=0 y F(1)=1
b) F(x) es continua, derivable y con derivada F'(x)=2-2x continua
c) Es no decreciente, puesto que F'(x)=2-2x ≥0 en el intervalo [0 ; 1]



4. Variables aleatorias continuas: función de densidad y función de distribución

Como F(x) es continua, la probabilidad de X tome un valor concreto x es nula:

$$P(x) = \lim_{\Delta x \to 0} P(x < X \leq x + \Delta x) = \lim_{\Delta x \to 0} \left(F(x + \Delta x) - F(x)\right) = F(x) - F(x) = 0$$

Por ello, introducimos un nuevo concepto:

Función de densidad

Dada una v. a. continua con función de distribución F(x), definimos su función de densidad como la derivada de dicha función en el campo de variación de la variable.

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

Entonces, la función de distribución puede expresarse como

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

4. Variables aleatorias continuas: función de densidad y función de distribución

Ejemplo 4: Función de distribución y de densidad

Si X es una variable aleatoria con función de distribución

$F(x)= 2x-x^2$;en el intervalo [0 ; 1]

Entonces, la función de densidad es

$f(x)=2-2x$; en el intervalo [0 ; 1]

Ejemplo 5: Concepto intuitivo de función de densidad

Un experimento mide la afluencia de clientes a un establecimiento comercial. En concreto, la variable aleatoria mide el tiempo que transcurre entre la llegada de un cliente y el siguiente.

La tabla siguiente muestra el tiempo transcurrido (en intervalos de igual amplitud), el número de observaciones en cada intervalo y las frecuencias relativas correspondiente a un conjunto de 100 mediciones.

4. Variables aleatorias continuas: función de densidad y función de distribución

Ejemplo 5: Función de densidad

Tiempo	Nº de clientes	Frecuencia relativa
0 – 1	4	0,04
1 – 2	5	0,05
2 – 3	8	0,08
3 – 4	16	0,16
4 – 5	22	0,22
5 – 6	16	0,16
6 – 7	15	0,15
7 – 8	9	0,09
8 – 9	3	0,03
9 – 10	2	0,02

4. Variables aleatorias continuas: función de densidad y función de distribución

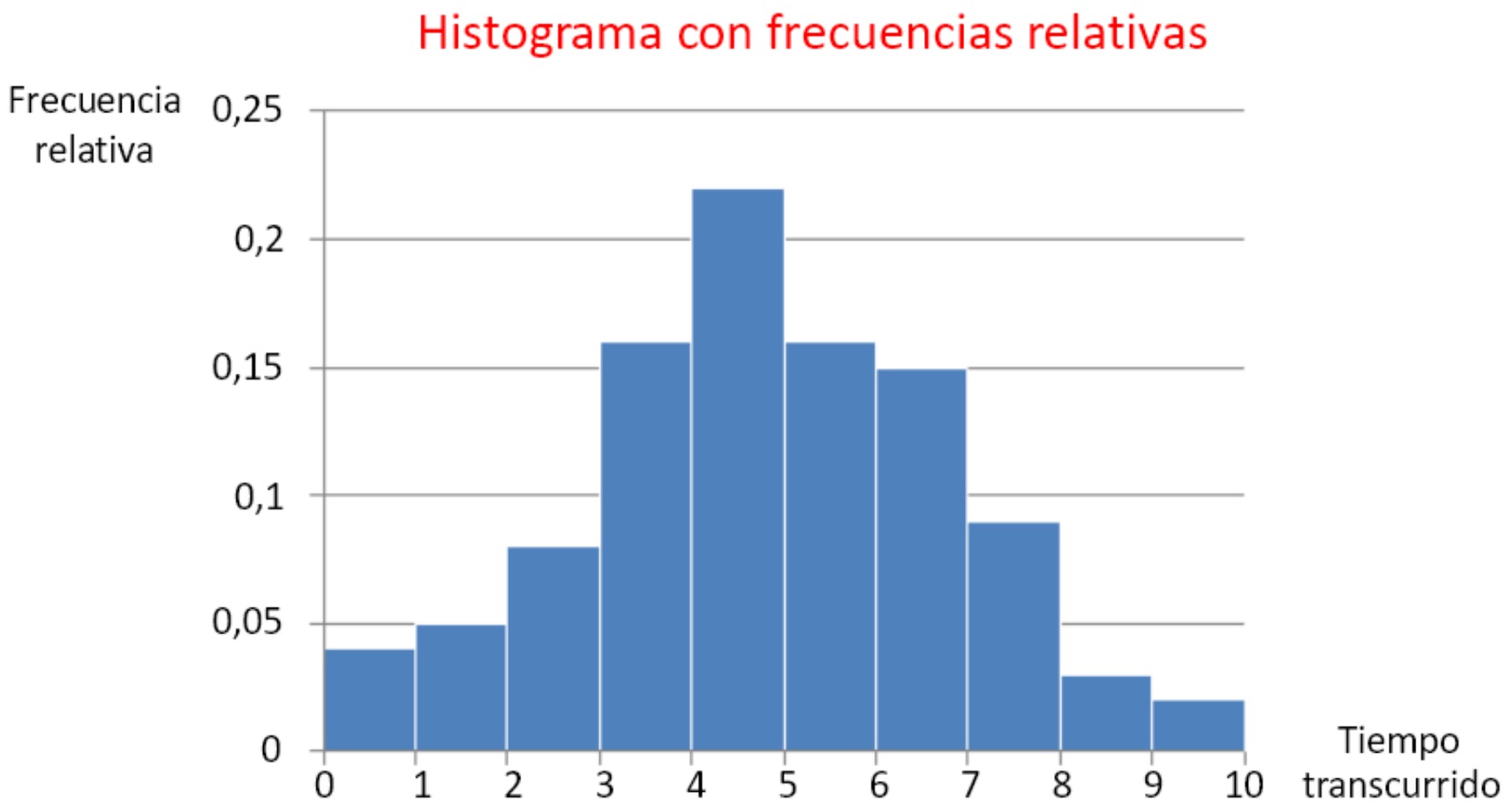
Ejemplo 5: Función de densidad

Vamos a presentar los datos en un gráfico, que, como estamos ante una variable continua, será un histograma pero dibujado a partir de las frecuencias relativas.

De este modo se cumple que el área del rectángulo sea proporcional a la frecuencia relativa (en este caso es igual a la frecuencia relativa ya que la amplitud de los intervalos es 1) y que la suma de todas las áreas será igual a la unidad.

Dado que los intervalos son de amplitud uno, el histograma puede construirse con la frecuencia relativa. La base de los rectángulos es la amplitud del intervalo (1) y la altura la **frecuencia relativa**.

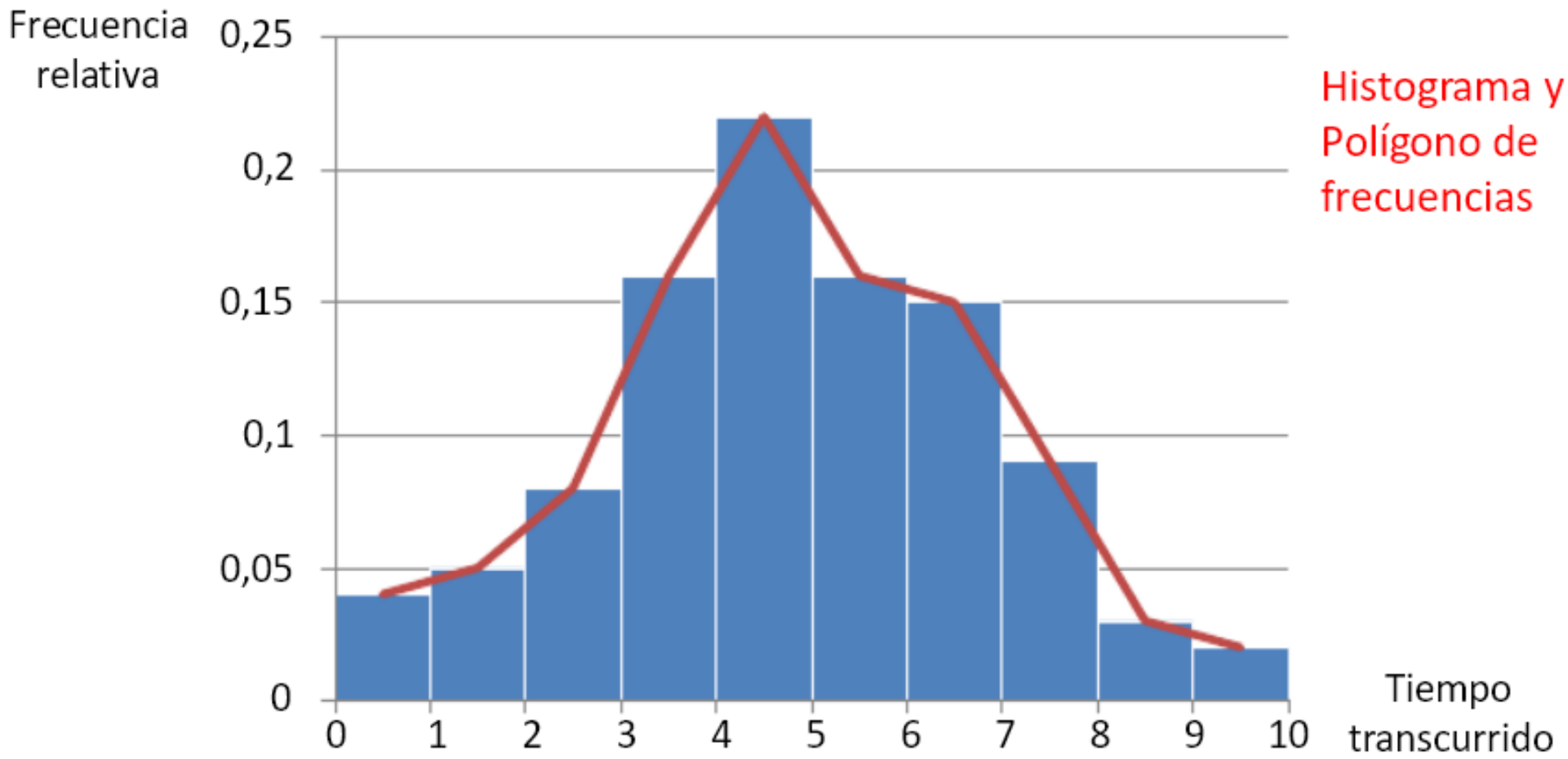
4. Variables aleatorias continuas: función de densidad y función de distribución



Como cada rectángulo tiene por área su frecuencia relativa, el área encerrada por el conjunto de todos los rectángulos es 1.

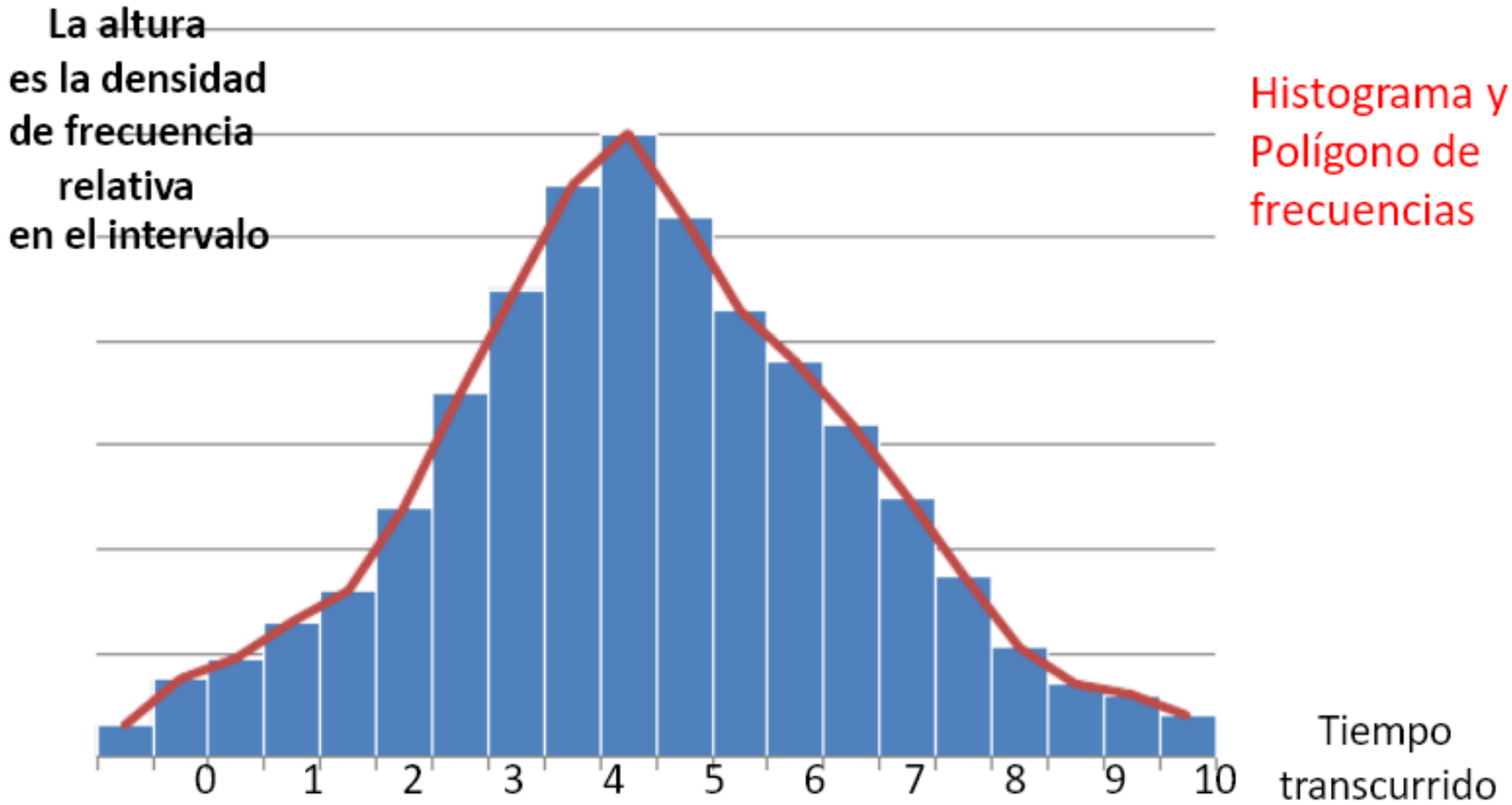
4. Variables aleatorias continuas: función de densidad y función de distribución

Si unimos los extremos de cada barra con una línea poligonal, se obtiene el **polígono de frecuencias**.



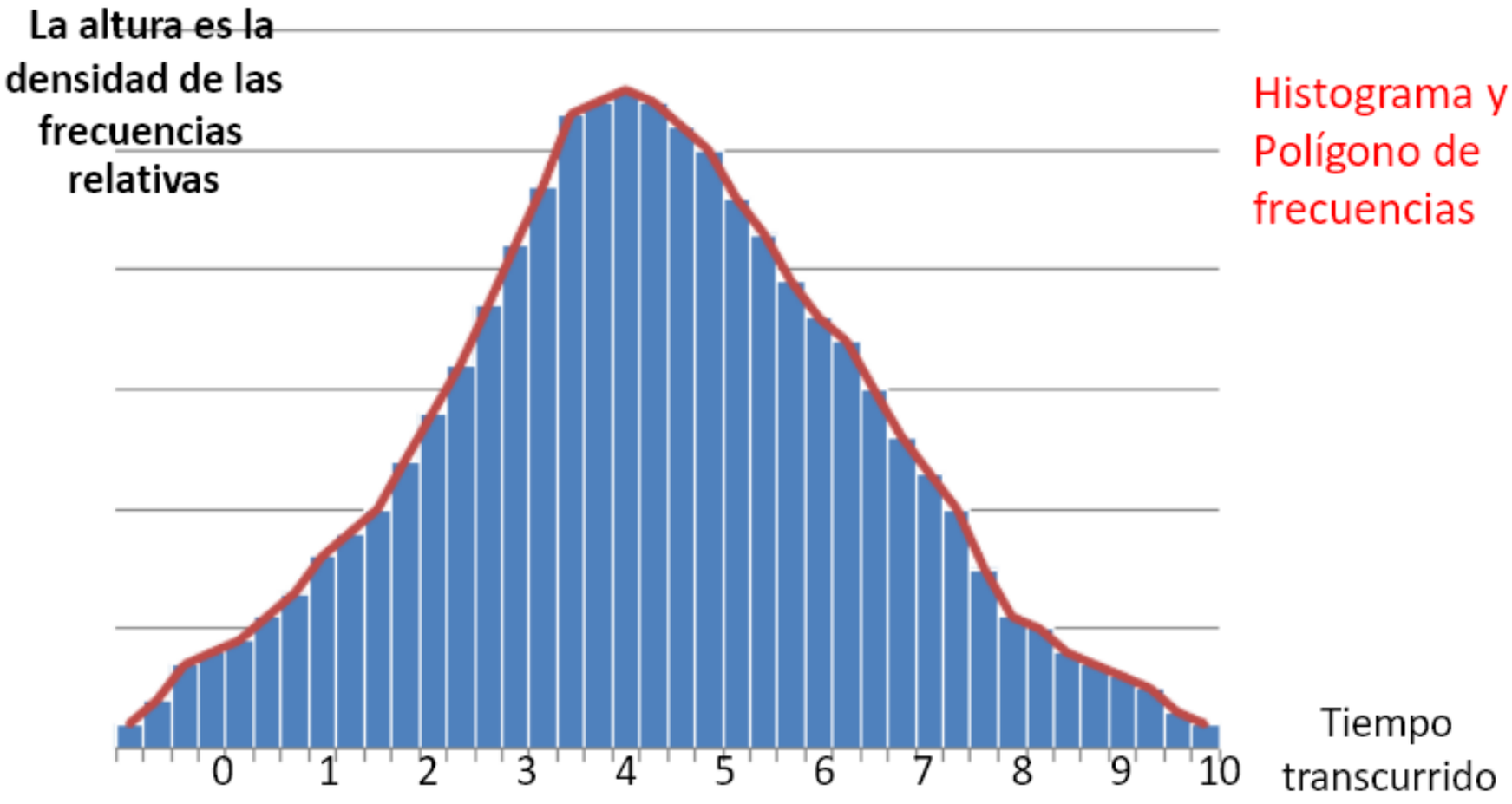
4. Variables aleatorias continuas: función de densidad y función de distribución

Si el experimento se realiza para muchos más clientes podría agruparse la variable en intervalos de menor amplitud.



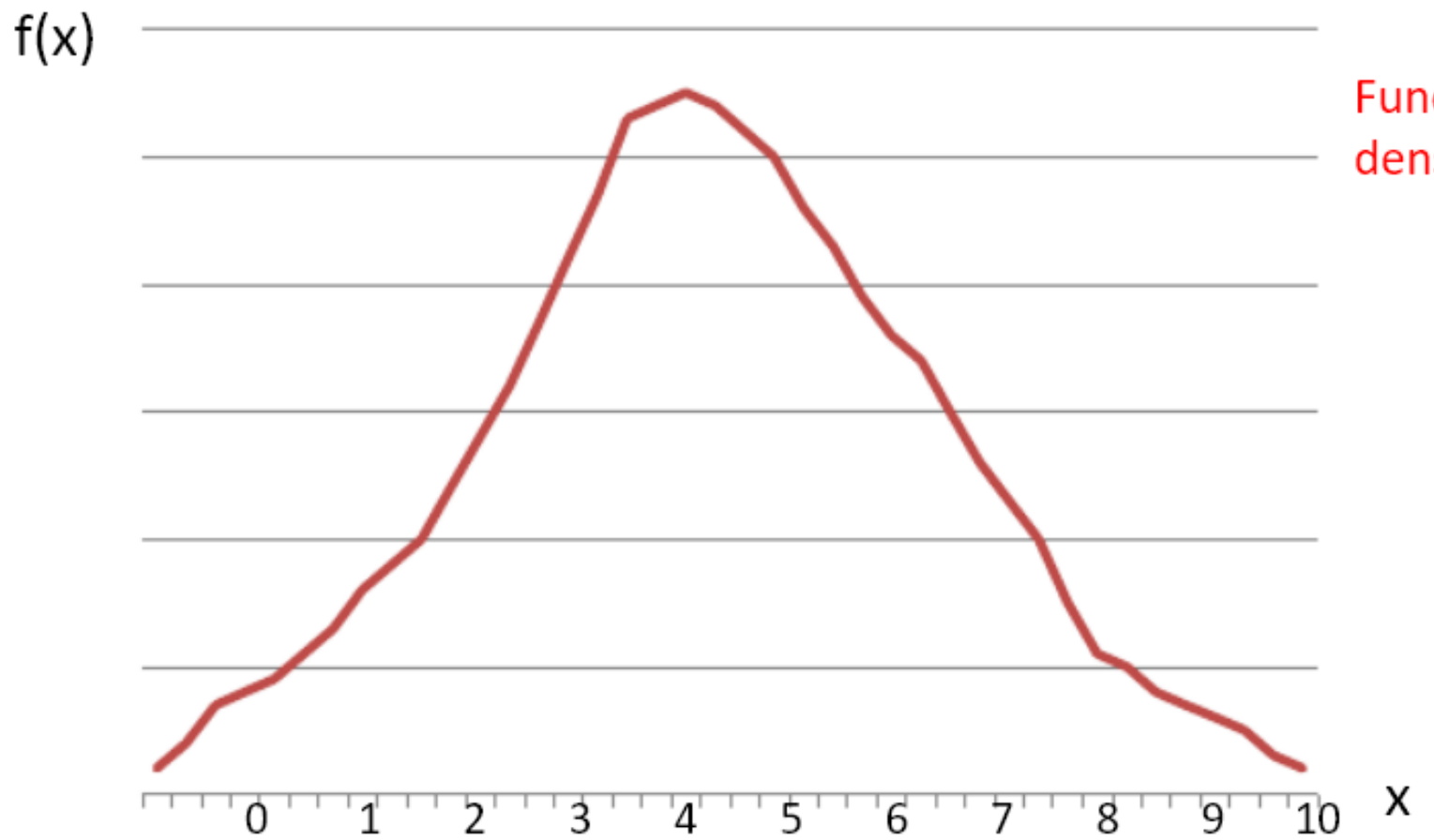
4. Variables aleatorias continuas: función de densidad y función de distribución

Si las amplitudes se hacen muy pequeñas, la línea poligonal del polígono de frecuencias toma el aspecto de una curva.



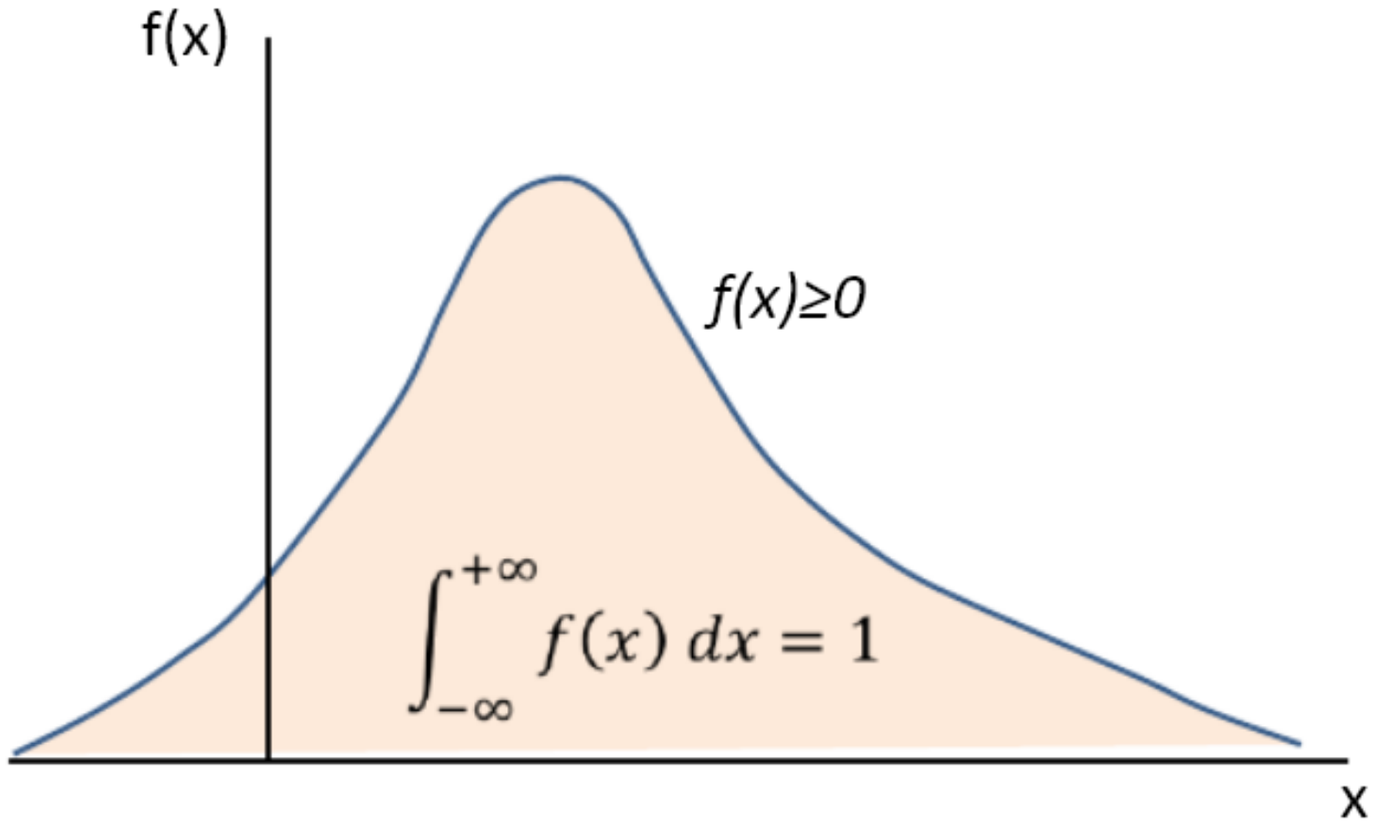
4. Variables aleatorias continuas: función de densidad y función de distribución

Esta curva, f(x), llevada al límite en cuanto al número de intervalos y número de observaciones, sería la **función de densidad** de la variable aleatoria continua X.



4. Variables aleatorias continuas: función de densidad y función de distribución

El área encerrada por la curva de la función de densidad es la probabilidad del conjunto de sucesos, es decir, la unidad, y se calcula como la integral definida de la función f(x).



4. Variables aleatorias continuas: función de densidad y función de distribución

Se puede definir genéricamente la función de densidad como:

Función de densidad

Llamaremos **función de densidad**, f(x), a una función **continua y no negativa** en un intervalo I y con valor igual a cero en el resto del dominio R

tal que: $$\int_{-\infty}^{+\infty} f(x)\partial x = \int_{I} f(x)\partial x = 1$$

4. Variables aleatorias continuas: función de densidad y función de distribución

Toda "función de densidad" lleva asociada una variable aleatoria X de tipo continuo cuya función de distribución quedará definida de la siguiente forma:

Función de distribución

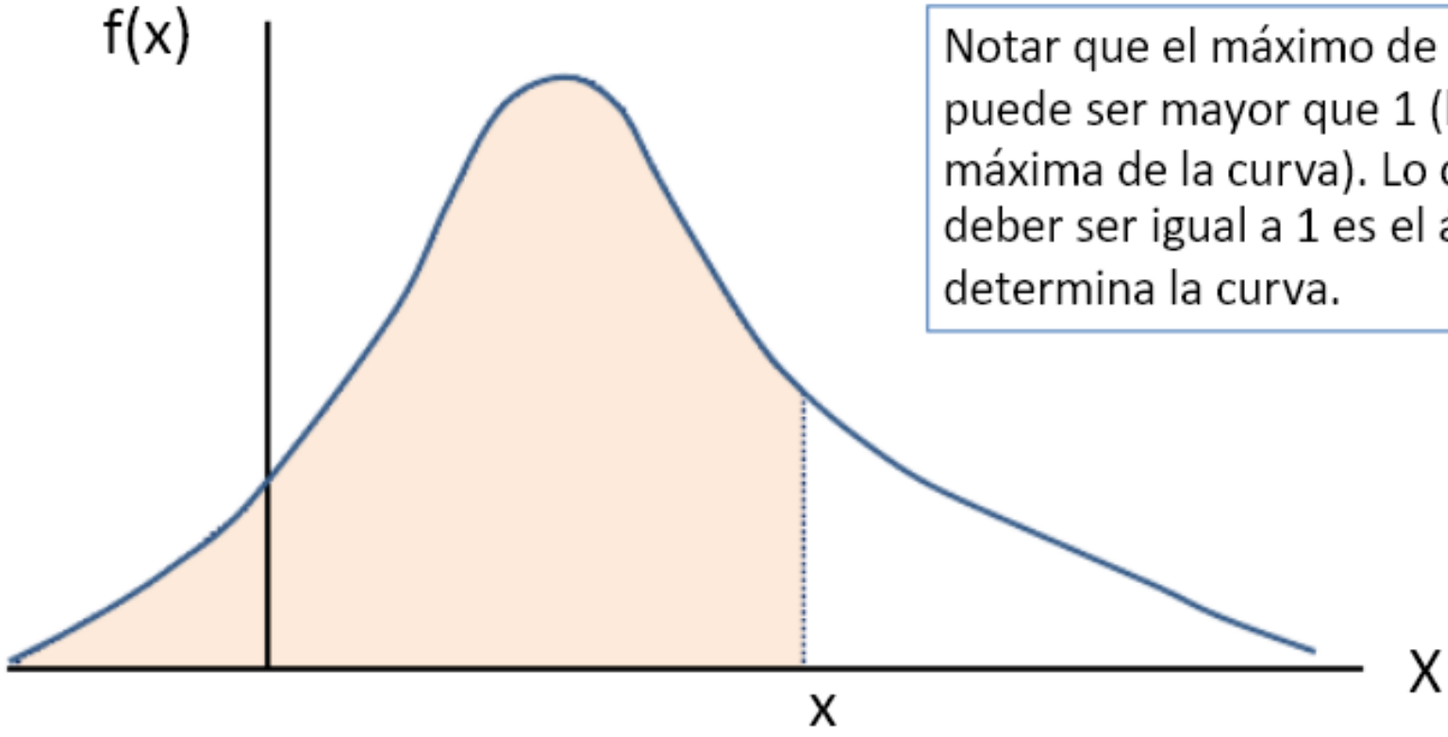
Dada una función de densidad es f(x), queda completamente determinada una variable aleatoria X continua con función de distribución F(x) definida como:

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

4. Variables aleatorias continuas: función de densidad y función de distribución

F(x) es el área encerrada por f(x) a la izquierda del valor x.

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



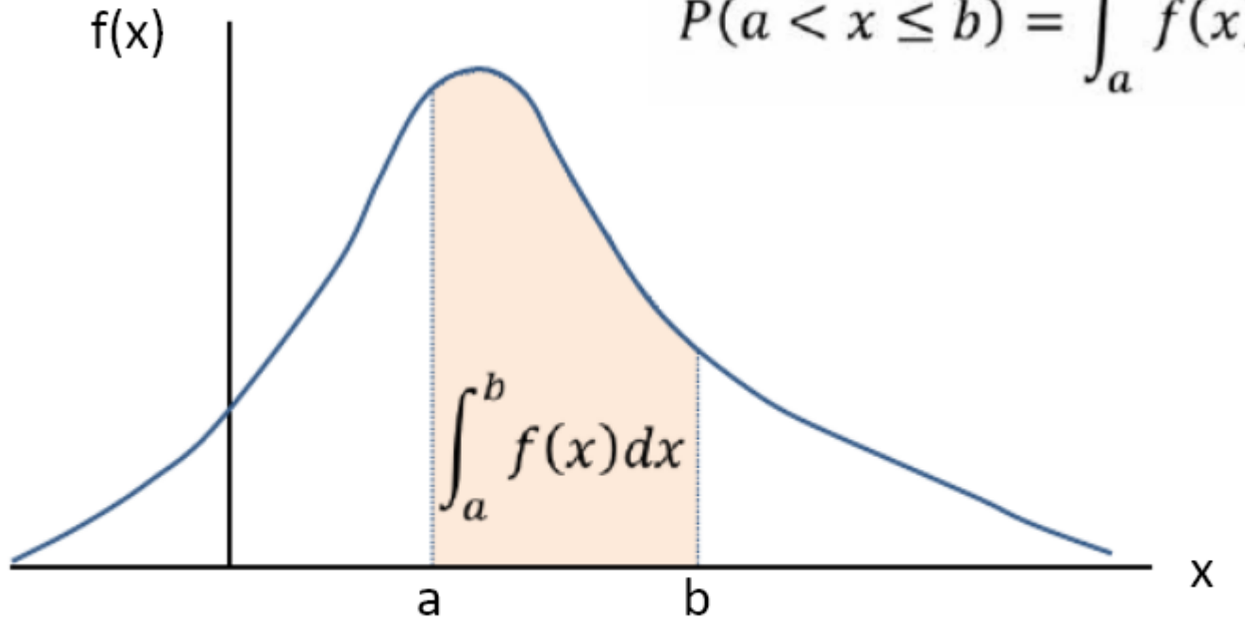
Notar que el máximo de f(x) puede ser mayor que 1 (la altura máxima de la curva). Lo que deber ser igual a 1 es el área que determina la curva.

4. Variables aleatorias continuas: función de densidad y función de distribución

Recordar que para una variable continua la probabilidad de un valor puntual es nula.

Para un intervalo (a, b] la probabilidad es el área encerrada bajo la curva para ese intervalo, es decir, la integral definida de la función de densidad entre a y b.

$$P(a < x \leq b) = \int_a^b f(x)dx$$



Además P(a<X≤b) = P(a ≤ X≤b) = P(a ≤ X<b) = P(a<X<b).

4. Variables aleatorias continuas: función de densidad y función de distribución

Ejemplo 6: Variables aleatorias continuas

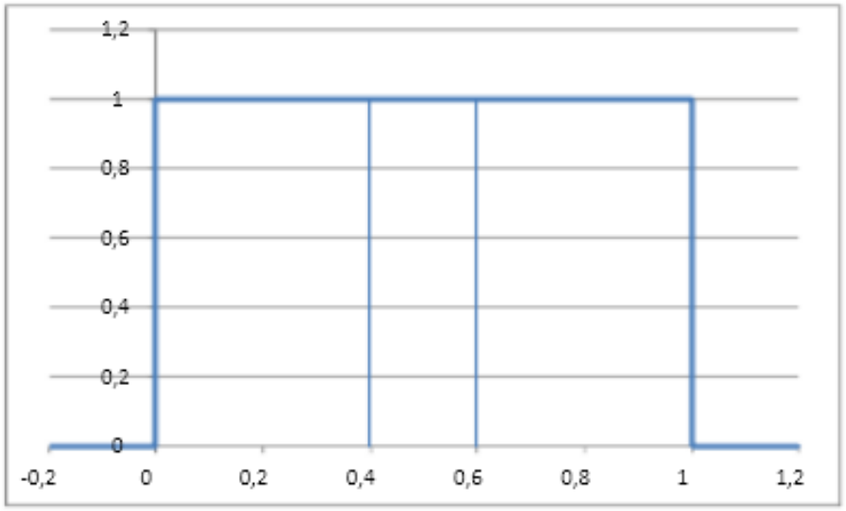
Si X es una variable aleatoria con función de densidad

f(x)= 1 ;en el intervalo [0 ; 1]

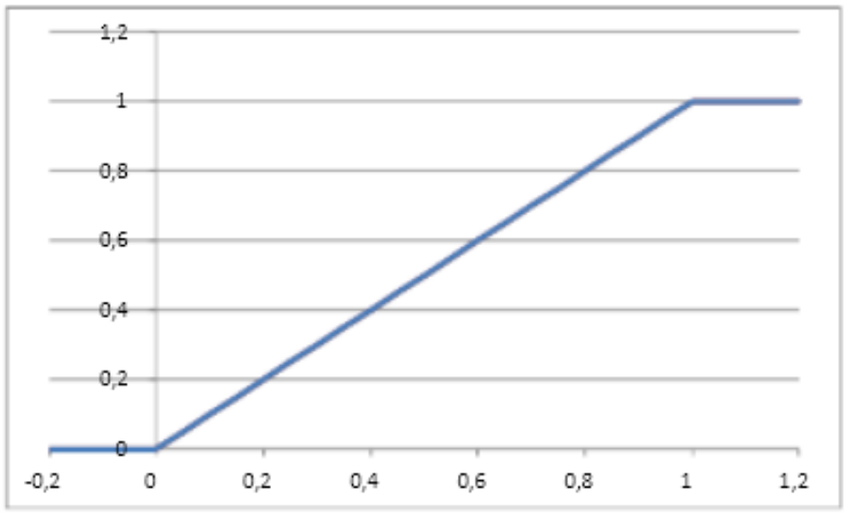
Entonces, su función de distribución es

$$F(x) = \int_0^x f(t)dt = \int_0^x dt = x \quad \text{para } x \in [0;1]$$

f(x)



F(x)



$$P(0{,}4 \le X \le 0{,}6) = 0{,}2$$

$$F(x) = \begin{cases} 0 & x \le 0 \\ x & 0 \le x \le 1 \\ 1 & 1 \le x \end{cases}$$

4. Variables aleatorias continuas: función de densidad y función de distribución

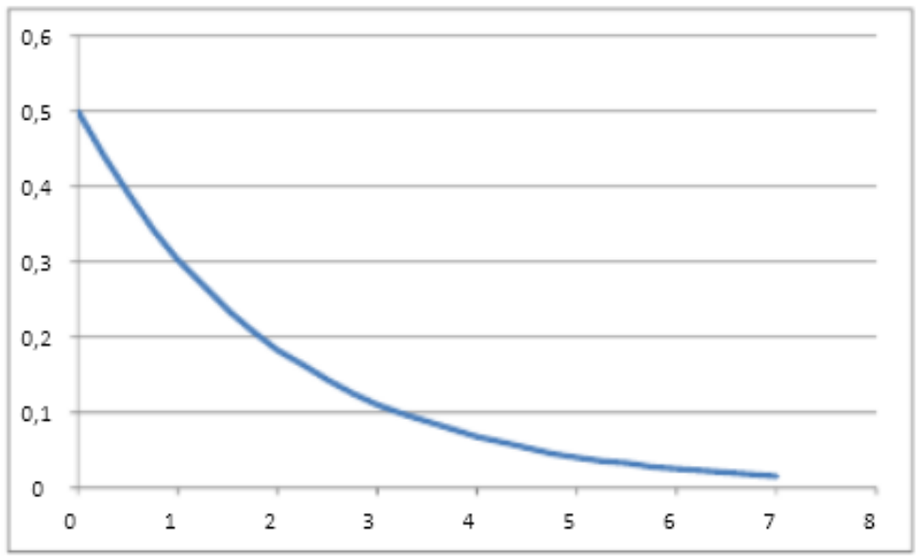
Ejemplo 7: Variables aleatorias continuas

Sea una función f(x) tal que: $f(x) = \lambda e^{-\lambda x} \quad si\ x > 0\ (siendo\ \lambda > 0)$

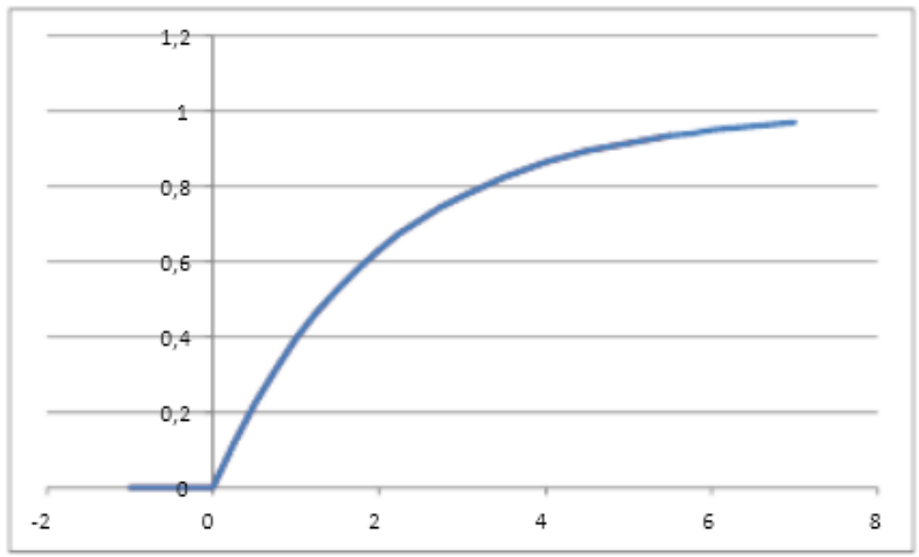
1. $f(x) \geq 0$
2. $\int_0^{+\infty} \lambda e^{-\lambda t}\, dt = \left[-e^{-\lambda t} \right]_0^{+\infty} = 1$

$$F(x) = \int_0^x \lambda e^{-\lambda t} dt = \left[-e^{-\lambda t} \right]_0^x = 1 - e^{-\lambda x} \quad \text{si } x > 0$$

f(x) para λ=0,5



F(x) para λ=0,5



4. Variables aleatorias continuas: función de densidad y función de distribución

RESUMEN: v.a. DISCRETA

Si X es una v.a. discreta con función de distribución F(x) y de probabilidad $p_i = p(x_i)$, entonces:

1. $F(-\infty) = 0 \quad ; \quad F(+\infty) = 1$
2. $F(x)$ *es continua por la derecha y no decreciente*
3. $p(x_i) \geq 0 \quad ; \quad \sum_i p(x_i) = 1$
4. $F(x) = \sum_{x_i \leq x} p(x_i)$
5. $p(x_i) = P(X = x_i) = F(x_i) - F(x_{i-1})$
6. $P(a < X \leq b) = F(b) - F(a) = \sum_{x_{i-1} < x \leq x_i} p(x_i)$

4. Variables aleatorias continuas: función de densidad y función de distribución

RESUMEN: v.a. CONTINUA

Si X es una v.a. continua con función de distribución F(x) y de densidad f(x), entonces :

1. $F(-\infty)=0 \quad ; \quad F(+\infty)=1$
2. $F(x)$ *es absolutamente continua y no decreciente*
3. $f(x)\geq 0 \quad ; \quad \int_{-\infty}^{+\infty} f(t)dt = 1$
4. $\frac{dF(x)}{dx} = f(x) \quad ; \quad F(x) = \int_{-\infty}^{x} f(t)\,dt$
5. $P(X=a)=0$
6. $P(a\leq X\leq b)= F(b)- F(a)= \int_{a}^{b} f(t)dt$

5. Esperanza matemática

5. Esperanza matemática. Propiedades.

Al realizar un experimento aleatorio, por ejemplo un juego de azar con premios, ¿qué ganancia o pérdida se puede esperar que ocurra por término medio?

Si el juego se repite numerosas veces, se puede calcular la ganancia media (la media de los resultados) y concluir que:

- Si el resultado medio (la ganancia) fuese cero significaría que el juego es equitativo, unas veces se gana y otras se pierde, dando una ganancia media de cero.
- Si la ganancia media es positiva significa que el juego proporciona ganancia a los jugadores (de media).
- Si la ganancia media es negativa el juego aporta ganancia al organizador del mismo a costa de pérdidas de los jugadores.

5. Esperanza matemática

Ejemplo 8: Esperanza matemática

Consideremos un juego de azar consistente en lanzar un dado y obtener una ganancia según la puntuación obtenida, tal como se indica en la tabla:

Puntuación del dado	Ganancia (euros)
1	20
2 ó 3	-12
4 ó 5	-18
6	40

Cada una de las puntuaciones es un suceso elemental de probabilidad 1/6 y como son excluyentes, la probabilidad de la unión de algunos de ellos es la suma de sus probabilidades.

5. Esperanza matemática

Si consideramos la variable aleatoria X como la ganancia que se obtiene del juego, sus valores y probabilidades son:

X: Ganancia	Probabilidad P(x)
20	1/6
-12	2/6
-18	2/6
40	1/6

La media ponderada de las ganancias por las veces que se producirán nos da la ganancia media, y comprobamos que es un juego equitativo:

$$20 \cdot \frac{1}{6} + (-12) \cdot \frac{2}{6} + (-18) \cdot \frac{2}{6} + 40 \cdot \frac{1}{6} = 0$$

5. Esperanza matemática

La suma ponderada de la variable por las probabilidades es lo que se denomina esperanza matemática o valor esperado, E[x], de la variable aleatoria.

$$E[x] = 20 \cdot \frac{1}{6} + (-12) \cdot \frac{2}{6} + (-18) \cdot \frac{2}{6} + 40 \cdot \frac{1}{6} = 0$$

Como este juego es equitativo, si se repite una infinidad de veces, el jugador no habrá ganado nada, ni el organizador del juego tampoco. El resultado global del juego es nulo.

5. Esperanza matemática

Esperanza matemática de una variable discreta

Dada una variable aleatoria discreta X, que toma los valores $x_1, x_2, \ldots, x_n$, con probabilidades $P(x_1), P(x_2), \ldots, P(x_n)$, se define el valor esperado o esperanza matemática de la variable aleatoria como

$$\mu = E[X] = \sum_{i=1}^{n} x_i \cdot P(x_i)$$

Esperanza matemática de una variable continua

Dada una variable aleatoria continua X, con función de densidad f(x), se define el valor esperado o esperanza matemática de la variable aleatoria como

$$\mu = E[X] = \int_{-\infty}^{+\infty} x\, f(x)\, dx$$

5. Esperanza matemática

El concepto de esperanza matemática se puede ampliar para el caso general de **esperanza matemática de una función g(x)** definida sobre la variable aleatoria X.

Caso discreto

Dada una variable aleatoria discreta X, con probabilidad P(x), se define el valor esperado o esperanza matemática de una función g(x) como

$$E[g(X)] = \sum_{i=1}^{n} g(x_i) \cdot P(x_i)$$

Caso continuo

Dada una variable aleatoria continua X, con función de densidad f(x), se define el valor esperado o esperanza matemática de una función g(x) como

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x)\, f(x)\, dx$$

5. Esperanza matemática

Propiedades de la esperanza matemática

1. La esperanza de una constante es la propia constante.

$$E[k] = k$$

Suponiendo que una función de la variable X la transforma en una constante k,

Caso discreto

$$E[k] = \sum_{i=1}^{n} k \cdot P(x_i) = k \cdot \sum_{i=1}^{n} P(x_i) = k \cdot 1 = k$$

Caso continuo

$$E[k] = \int_{-\infty}^{+\infty} k\, f(x)\, dx = k \cdot \int_{-\infty}^{+\infty} f(x)\, dx = k \cdot 1 = k$$

5. Esperanza matemática

1.a. La esperanza de una constante por una variable (o una función de la variable) es la constante por la esperanza de la variable (o función de la variable).

$$E[k \cdot g(x)] = k \cdot E[g(x)]$$

1.b. Dada una variable aleatoria, la esperanza de la suma de dos funciones de la variable es la suma de las esperanzas de las funciones de la variable.

$$E[g(x) + h(x)] = E[g(x)] + E[h(x)]$$

Estas dos propiedades se resumen en una:

2. Dada una variable aleatoria, X, y sean dos funciones g(x) y h(x), y sean a y b dos constantes cualesquiera,

$$E[a \cdot g(x) + b \cdot h(x)] = a \cdot E[g(x)] + b \cdot E[h(x)]$$

5. Esperanza matemática

Caso discreto

$$E[a \cdot g(x) + b \cdot h(x)] = \sum_{i=1}^{n} (a \cdot g(x_i) + b \cdot h(x_i)) \cdot P(x_i) =$$

$$= a \cdot \sum_{i=1}^{n} g(x_i) \cdot P(x_i) + b \cdot \sum_{i=1}^{n} h(x_i) \cdot P(x_i) = a \cdot E[g(x)] + b \cdot E[h(x)]$$

Caso continuo

$$E[a \cdot g(x) + b \cdot h(x)] = \int_{-\infty}^{+\infty} \big(a \cdot g(x) + b \cdot h(x)\big) f(x) dx =$$

$$= a \int_{-\infty}^{+\infty} g(x) f(x) dx + b \cdot \int_{-\infty}^{+\infty} h(x) f(x) dx = a \cdot E[g(x)] + b \cdot E[h(x)]$$

5. Esperanza matemática

Puesto que se puede asociar la idea de esperanza matemática de una variable aleatoria a la de media de una variable, la anterior propiedad es equivalente a lo visto para la media aritmética de una variable estadística (estadística descriptiva).

Del mismo modo cabe preguntarse cómo le afecta un **cambio de origen y de escala a la esperanza matemática**.

Dada una variable aleatoria X, y dos constantes cualesquiera *a* y *b*, puede plantearse un cambio de origen y de escala de la variable, $a + b \cdot X$.

Como resultado de las propiedades anteriores, la esperanza matemática reproduce dichos cambios:

$$E[a+bX]= a + bE[X]$$

5. Esperanza matemática

3. Para cualquier aleatoria, X, se verifica que

$$E\left[X - E(X)\right] = E\left[X - \mu\right] = 0$$

Caso discreto

$$E[x - E[x]] = \sum_{i=1}^{n} (x_i - E[x]) \cdot P(x_i) =$$

$$= (\sum_{i=1}^{n} x_i \cdot P(x_i)) - (E[x] \cdot \sum_{i=1}^{n} P(x_i)) \ = E[x] - E[x] \cdot 1 = 0$$

Caso continuo

$$E = [x - E[x]] = \int_{-\infty}^{+\infty} (x - E[x])\, f(x)\, dx =$$

$$= \int_{-\infty}^{+\infty} x\, f(x)\, dx + E[x] \int_{-\infty}^{+\infty} f(x)\, dx = E[x] - E[x] \cdot 1 = 0$$

6. Otras Características de la variable aleatoria

6. Otras características de una variable aleatoria.

Momentos y momentos centrales

Los momentos son unas medidas que sirven para caracterizar la distribución de una variable y así poder estudiar sus características y establecer comparaciones entre distribuciones.

Momento de orden r

Dada una variable aleatoria X, y un número entero positivo r, se define el momento de orden r de la variable X, α_r , como

$$\propto_r = E[x^r] \qquad para\ r = 1, 2, \ldots$$

Momento central de orden r

Dada una variable aleatoria X, y un número entero positivo r, se define el momento central de orden r de la variable X, μ_r , como

$$\mu_r = E\left[(X - E(X))^r\right] = E\left[(X - \mu)^r\right] \quad para\ r = 2,3,\ \mathrm{K}$$

6. Otras Características de la variable aleatoria

El cálculo de los momentos difiere según la variable sea discreta o continua.

Caso discreto

- Momento de orden r

$$\propto_r = E[x^r] = \sum_{i=1}^{n} x^r{}_i \cdot P(x_i) \qquad para\, r = 1, 2, \ldots$$

- Momento central de orden r

$$\mu_r = E[(x - E[x])^r] = \sum_{i=1}^{n} (x_i - E[x])^r \cdot P(x_i) \quad para\, r = 2, 3, \ldots$$

6. Otras Características de la variable aleatoria

Caso continuo

- Momento de orden r

$$\propto_r = E[x^r] = \int_{-\infty}^{+\infty} x^r \, f(x) \, dx \quad para\, r = 1, 2, \ldots$$

- Momento central de orden r

$$\mu_r = E[(x - E[x])^r] = \int_{-\infty}^{+\infty} (x - E[x])^r \, f(x) \, dx \quad para\, r = 2, 3, \ldots$$

Obsérvese que en los momentos centrales, tanto para el caso discreto como para el continuo, se ha excluido el caso r=1. Esto es porque para r=1 ya se ha demostrado, en la tercera propiedad de la esperanza matemática, que el resultado es cero, $E[x - E[x]] = 0$.

6. Otras Características de la variable aleatoria

De toda la familia de momentos hay dos que merecen una atención especial. Son los equivalentes a la media y la varianza.

Llamaremos **media de la distribución** de una variable aleatoria al momento de orden 1.

$$\mu = \alpha_1 = E(X)$$

Se define la **varianza de la distribución** de una variable aleatoria al momento central de orden 2.

$$\sigma^2 = \mu_2 = E\left[(X - \mu)^2\right]$$

6. Otras Características de la variable aleatoria

Varianza de una distribución

La varianza de una distribución es una medida de la dispersión de los valores de la variable respecto a su media. La denotaremos como

$$Var(X) = \sigma_X^2 = \mu_2 = E\left[(X - E(X))^2\right] = E\left[(X - \mu)^2\right]$$

La varianza se expresa en las mismas unidades que la variable, pero al cuadrado. Para reducir la medida a unidades de la variable se calcula la **desviación típica**, σ_x, que se define como la raíz cuadrada positiva de la varianza.

$$\sigma_X = +\sqrt{\sigma_X^2}$$

6. Otras Características de la variable aleatoria

Propiedades de la varianza

1. La varianza es un número mayor o igual que cero.

$$Var\,(X) \geq 0$$

2. La varianza de una constante *k* es cero.

$$Var\,(k) = 0$$

3. La varianza se puede expresar como

$$Var(X) = E[(X-\mu)^2] = E(X^2) - \mu^2$$

Expresado en términos de momentos sería:

$$Var(X) = \alpha_2 - \alpha^2$$

6. Otras Características de la variable aleatoria

4. A la varianza, y a la desviación típica, le afectan los cambios de escala pero no los cambios de origen.

Varianza $$Var(a+bX) = b^2 Var(X)$$

Desviación típica $$\sigma_{a+bx} = |b|\sigma_x$$

De forma análoga a lo estudiado en el tema 1, podríamos definir el coeficiente de variación como el cociente de la desviación típica y la media (σ / μ).

Los cambios de escala no afectaría al coeficiente de variación, pero un cambio de origen sí lo haría al afectar solo al denominador del cociente.

6. Otras Características de la variable aleatoria

Cuantiles de una distribución

Dada una variable aleatoria X, se define el cuantil de orden q , para $0 \leq q \leq 1$, como el valor C_q que verifica:

En el caso discreto, el **primer** valor x_i tal que
$F(x_i) = P(X \leq x_i) \geq q$
En el supuesto que $F(x_i) = q$; $C_q = (x_i + x_{i+1})/2$

En el caso continuo, el valor C_q tal que $F(C_q) = q$

Los cuantiles pueden expresarse en términos porcentuales, con $0 \leq q \leq 100$ y las probabilidades expresadas en tanto por ciento.

6. Otras Características de la variable aleatoria

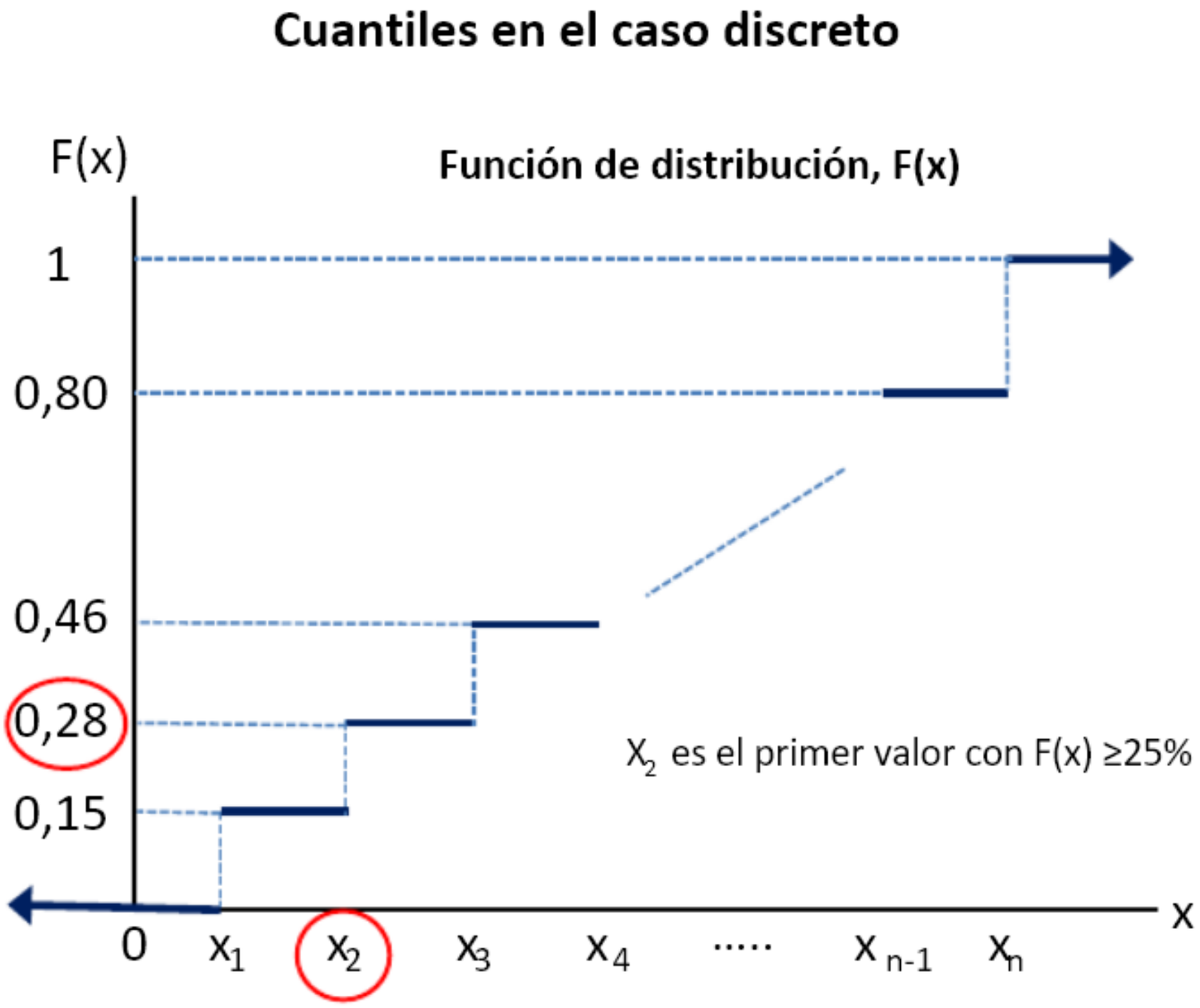
Caso discreto

Podemos buscar los cuantiles, de manera análoga a como se hacía en el tema 2.

Por ejemplo, para hallar el cuantil 25, buscaremos el valor de la variable que deja tras de sí una probabilidad acumulada de 0,25 (el 25% del total), esto es, que la Función de Distribución tome el valor 0,25.

En el caso discreto no tiene por que existir ese valor y se toma el primero tal que F(x) que iguala o supera ese 25 %.

6. Otras Características de la variable aleatoria

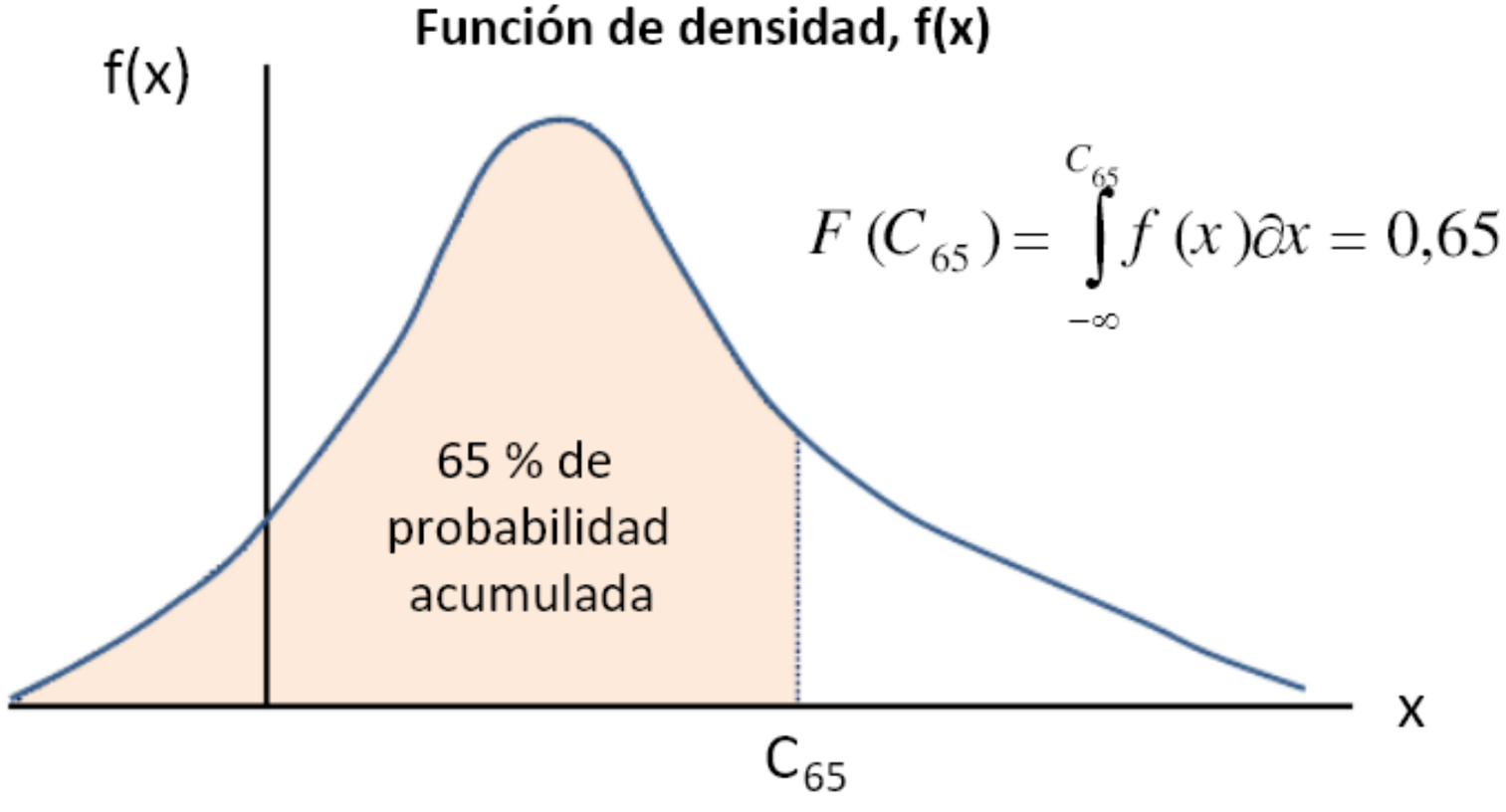


6. Otras Características de la variable aleatoria

Caso continuo

En el caso continuo existe un valor de X para el que su función de distribución vale exactamente el valor del cuantil buscado.

Por ejemplo, para el cuantil 65, que es el que acumula el 65 % de probabilidad por debajo de él,



Estadística Aplicada al Marketing

T.5. Variables aleatorias unidimensionales

Juan Miguel Navarro Ruiz

Grado en Marketing y Dirección Comercial

Curso 2019-2020
Juan M Navarro- Tlf: (+34) 968 27 88 25
Universidad Católica San Antonio de Murcia - Tlf: (+34) 968 27 88 00 info@ucam.edu - www.ucam.edu

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Modelos aleatorios mas frecuentes

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Binomial

La distribución Binomial es un caso particular de **probabilidad de variable aleatoria discreta**, y por sus aplicaciones, es posiblemente la más importante.

Esta distribución corresponde a la realización de un experimento aleatorio que cumple con las siguientes condiciones:

* Al realizar el experimento sólo son posible dos resultados: el suceso A, llamado éxito, o su contrario A', llamado fracaso.

* Al repetir el experimento, el resultado obtenido es independiente de los resultados obtenidos anteriormente.

* La probabilidad del suceso A es constante, es decir, no varía de una prueba del experimento a otra. Si llamamos p a la probabilidad de A, p(A) = P, entonces p(A')= 1 – p = q

* En cada experimento se realizan n pruebas idénticas.

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Binomial

Un experimento aleatorio con las siguientes características sigue el modelo de una **distribución binomial**

a) En cada prueba del experimento sólo son posibles dos resultados.

b) El resultado de cada prueba es independiente de los anteriores.

c) La probabilidad de cada uno de los dos sucesos posibles permanece constante

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Binomial

Todo experimento que tenga estas características se dice que sigue el modelo de la **distribución Binomial o distribución de Bernoulli**.

En general, si se tienen *n* ensayos Bernoulli con probabilidad de éxito *p* y de fracaso *q*, entonces la distribución de probabilidad que la modela es la **distribución de probabilidad binomial** y su regla de correspondencia es:

$$P(X=x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

Como el cálculo de estas probabilidades puede resultar algo tedioso se han construido tablas para algunos valores de *n* y *p* que nos facilitan el trabajo (Ver las tablas de la función de probabilidad Binomial).

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Binomial

Variable aleatoria binomial

Definición: Representamos por *p* la probabilidad de un suceso *A* (éxito) y por *q* la probabilidad del contrario de *A* (fracaso). A la variable *X* que representa el número de éxitos la llamamos **variable aleatoria binomial**.

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Binomial

Esta distribución la representamos como:

$$B(n, p)$$

siendo *n* y *p* los parámetros de la misma

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Binomial

Ejemplo: Una compañía de tabaco ha determinado que el porcentaje de fumadores en una ciudad es del 35%. Se escoge al azar una muestra formada por 10 personas. La variable que representa el número de fumadores dentro de la muestra sigue una distribución binomial .

$$B(10\ ,\ 0.35)$$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Binomial

Función de probabilidad

Se puede demostrar que en una distribución $B(n, p)$, la probabilidad de obtener *r* éxitos viene dada por:

$$p(x = r) = \binom{n}{r} \cdot p^r \cdot q^{n-r}$$

que es la función de probabilidad de la distribución binomial.

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Binomial

***Ejemplo**:* Una prueba de inteligencia está compuesta de 10 preguntas, cada una de las cuales tiene 4 respuestas, siendo sólo una de ellas correcta. Si se contesta aleatoriamente se pide:

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Binomial

a) Probabilidad de acertar 4 preguntas

Se trata de una distribución binomial con ,

$p = \frac{1}{4}$, n=10 y r=4

$$p(x=4) = \binom{10}{4} \cdot \left(\frac{1}{4}\right)^4 \cdot \left(\frac{3}{4}\right)^6 = 0{,}1460$$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Binomial

b) Probabilidad de no acertar ninguna

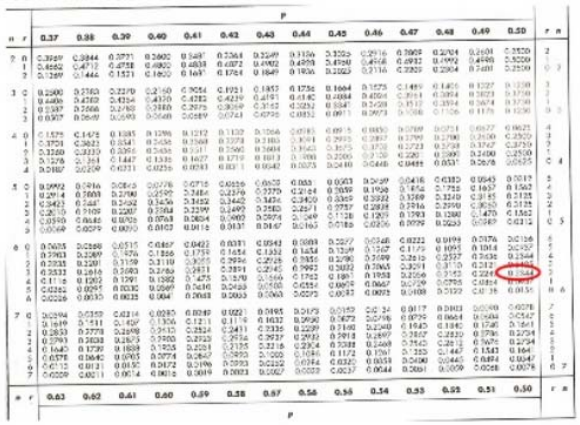
En este caso *r=0:*

$$p(x=0) = \binom{10}{0} \cdot \cdot \left(\frac{3}{4}\right)^{10} = 0{,}0563$$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Binomial

* Tablas binomiales.



- Para una combinación de n y p, la entrada indica una probabilidad de obtener un valor específico de r.
- Para localizar la entrada, cuando p≤0.50, localice p a lo largo del encabezado de la tabla, y en la columna correspondiente localice n y r en el margen izquierdo; cuando p≥0.50, localice el valor de p en la parte inferior de la tabla, y n y r arriba, en el margen derecho.

Para el ejemplo que estamos viendo: ¿Cuál es la probabilidad de obtener exactamente 2 caras al lanzar una misma moneda 6 veces ?
Tenemos p = 0.50, n = 6 y r = 2 obteniendo resultado directo de tablas

P(2 caras) = 0.2344

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Binomial

FUNCIÓN DE DISTRIBUCIÓN DE LA VARIABLE $B(n,p)$

$$F(k) = \sum_{x=0}^{k} \binom{n}{x} p^x (1-p)^{n-x} \quad k=0,1,2,\ldots,n$$

Para calcular las probabilidades puntuales utilizar:

$$P(X=0) = F(0)$$
$$P(X=k) = F(k) - F(k-1) \quad ; \quad k=1,2,\ldots,n$$

		p											
n	k	0,01	0,05	0,10	0,15	0,20	0,25	0,30	1/3	0,35	0,40	0,45	0,50
2	0	0,9801	0,9025	0,8100	0,7225	0,6400	0,5625	0,4900	0,4444	0,4225	0,3600	0,3025	0,2500
	1	0,9999	0,9975	0,9900	0,9775	0,9600	0,9375	0,9100	0,8889	0,8775	0,8400	0,7975	0,7500
	2	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
3	0	0,9703	0,8574	0,7290	0,6141	0,5120	0,4219	0,3430	0,2963	0,2746	0,2160	0,1664	0,1250
	1	0,9997	0,9928	0,9720	0,9392	0,8960	0,8438	0,7840	0,7407	0,7182	0,6480	0,5748	0,5000
	2	1,0000	0,9999	0,9990	0,9966	0,9920	0,9844	0,9730	0,9630	0,9571	0,9360	0,9089	0,8750
			1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
4	0	0,9606	0,8145	0,6561	0,5220	0,4096	0,3164	0,2401	0,1975	0,1785	0,1296	0,0915	0,0625
	1	0,9994	0,9860	0,9477	0,8905	0,8192	0,7383	0,6517	0,5926	0,5630	0,4752	0,3910	0,3125
	2	1,0000	0,9995	0,9963	0,9880	0,9728	0,9492	0,9163	0,8889	0,8735	0,8208	0,7585	0,6875
	3		1,0000	0,9999	0,9995	0,9984	0,9961	0,9919	0,9876	0,9850	0,9744	0,9590	0,9375
	4			1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Binomial

Media y Varianza

Se puede demostrar que para una distribución binomial

$$B(n, p)$$

la media y la varianza vienen dadas por:

$$\mu = n \cdot p \qquad \sigma^2 = n \cdot p \cdot q$$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Binomial

***Ejemplo**:* Una prueba de inteligencia está compuesta de 10 preguntas, cada una de las cuales tiene 4 respuestas, siendo sólo una de ellas correcta. Si se contesta aleatoriamente se pide:

$$\mu = 10 \cdot \frac{1}{4} = 2{,}5 \qquad \sigma^2 = 10 \cdot \frac{1}{4} \cdot \frac{3}{4} = 1{,}875$$

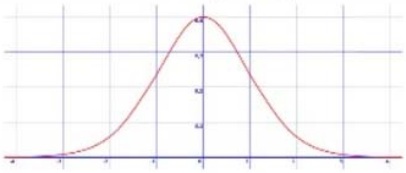
UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal

La distribución normal es también un caso particular de **probabilidad de variable aleatoria contínua**, fue reconocida por primera vez por el francés Abraham de Moivre (1667-1754). Posteriormente, Carl Friedrich Gauss (1777-1855) elaboró desarrollos más profundos y formuló la ecuación de la curva; de ahí que también se le conozca, más comúnmente, como la **"campana de Gauss"**. La distribución de una variable normal está completamente determinada por dos parámetros, su media (μ) y su desviación estándar (σ). Con esta notación, la densidad de la normal viene dada por la ecuación:

$$f(x) = \int_{-\alpha}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \, dx$$

que determina la curva en forma de campana que tan bien conocemos



UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal

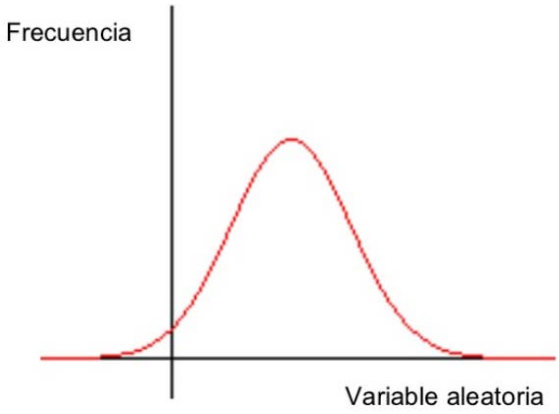
Existen dos razones básicas por las cuales la distribución normal ocupa un lugar tan prominente en la estadística :

- Tiene algunas propiedades que la hacen aplicable a un gran número de situaciones en la que es necesario hacer inferencias mediante la toma de muestras.

- La distribución normal casi se ajusta a las distribuciones de frecuencias reales observadas en muchos fenómenos, incluyendo características humanas, resultados de procesos físicos y muchas otras medidas de interés para los administradores, tanto en el sector público como en el privado.

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal



UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal

Son curvas llamadas "campanas de Gauss".que tienen las siguientes características:

a) En una distribución normal media, moda y mediana coinciden en el máximo de la curva

b) El área encerrada bajo la campana y el eje *x* es igual a la unidad.

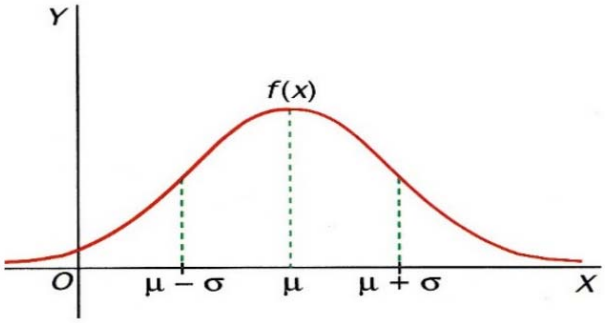
c) Los puntos de inflexión son los puntos

$\mu + \sigma$ y $\mu - \sigma$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal

Propiedad:
No importa cuáles sean los valores de μ y σ para un distribución de probabilidad normal, el área total bajo la curva siempre es 1, de manera que podemos pensar en áreas bajo la curva como si fueran probabilidades. Matemáticamente es verdad que:

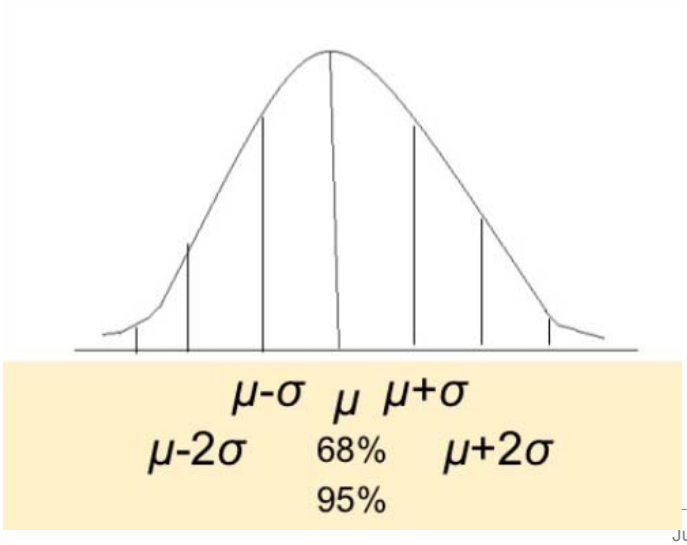
1.Aproximadamente el 68% de todos los valores de una población normalmente distribuida se encuentra dentro de ± 1 desviación estándar de la media.

2.Aproximadamente el 95.5% de todos los valores de una población normalmente distribuida se encuentra dentro de ± 2 desviaciones estándar de la media.

3.Aproximadamente el 99.7% de todos los valores de una población normalmente distribuida se encuentra dentro de ± 3 desviaciones estándar de la media.

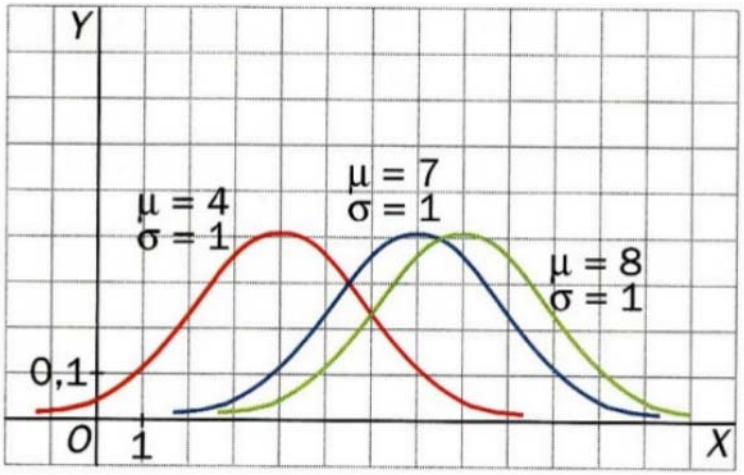
UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal



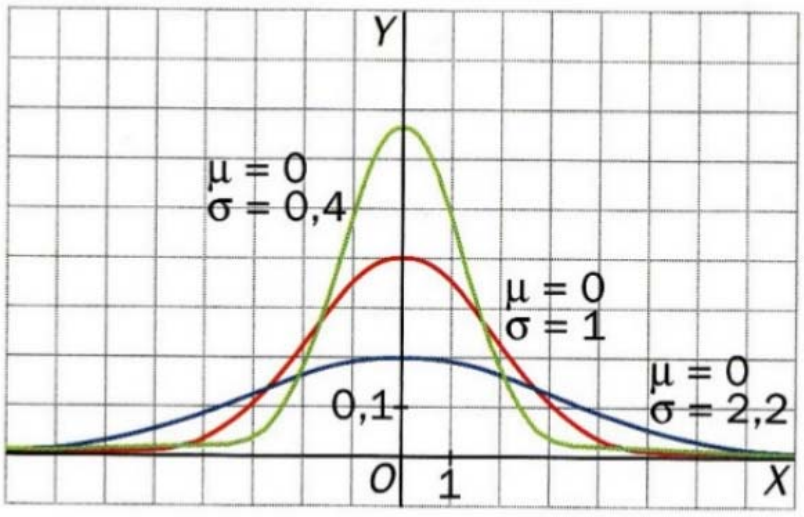
UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal



UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

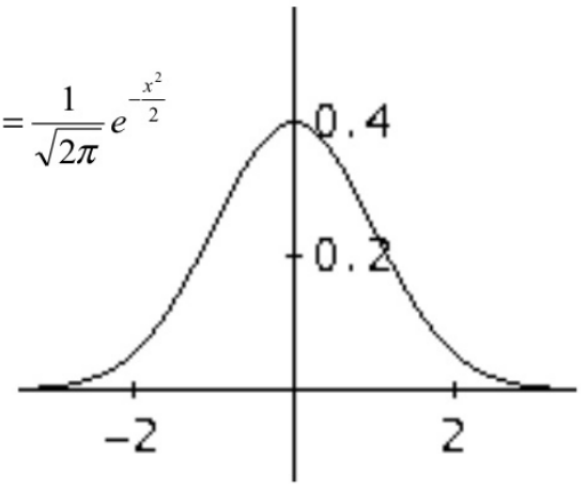
Distribución Normal



UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal

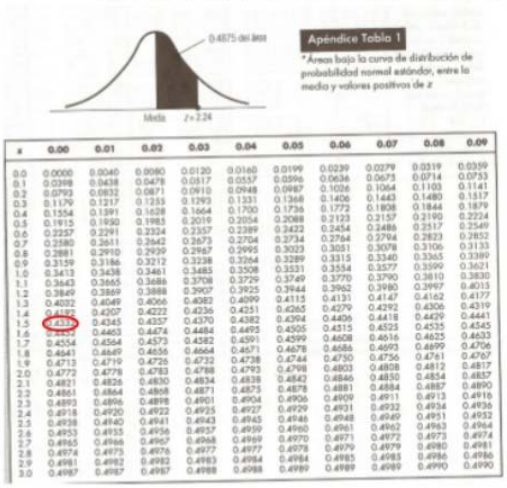
$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$



UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal

Uso de la tabla de distribución de probabilidad normal estándar



Apéndice Tabla 1

*Áreas bajo la curva de distribución de probabilidad normal estándar, entre la media y valores positivos de z

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Para cualquier distribución normal de probabilidad, todos los intervalos que contienen el mismo número de desviaciones estándar a partir de la media contendrán la misma fracción del área total bajo la curva para cualquier distribución de probabilidad normal. Esto hace que sea posible usar solamente una tabla (Apéndice Tabla 1) de la distribución de probabilidad normal estándar.

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal:

$$\mathbb{P}(Z \leq t) = \phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx$$

t	$\mathbb{P}(Z \leq t)$	t	$\mathbb{P}(Z \leq t)$	t	$\mathbb{P}(Z \leq t)$	t	$\mathbb{P}(Z \leq t)$
0.00	0.5000	0.80	0.7881	1.60	0.9452	2.40	0.9918
0.02	0.5080	0.82	0.7939	1.62	0.9474	2.42	0.9922
0.04	0.5160	0.84	0.7995	1.64	0.9495	2.44	0.9927
0.06	0.5239	0.86	0.8051	1.66	0.9515	2.46	0.9931
0.08	0.5319	0.88	0.8106	1.68	0.9535	2.48	0.9934
0.10	0.5398	0.90	0.8159	1.70	0.9554	2.50	0.9938
0.12	0.5478	0.92	0.8212	1.72	0.9573	2.52	0.9941
0.14	0.5557	0.94	0.8264	1.74	0.9591	2.54	0.9945
0.16	0.5636	0.96	0.8315	1.76	0.9608	2.56	0.9948
0.18	0.5714	0.98	0.8365	1.78	0.9625	2.58	0.9951
0.20	0.5793	1.00	0.8413	1.80	0.9641	2.60	0.9953
0.22	0.5871	1.02	0.8461	1.82	0.9656	2.62	0.9956
0.24	0.5948	1.04	0.8508	1.84	0.9671	2.64	0.9959
0.26	0.6026	1.06	0.8554	1.86	0.9686	2.66	0.9961
0.28	0.6103	1.08	0.8599	1.88	0.9699	2.68	0.9963
0.30	0.6179	1.10	0.8643	1.90	0.9713	2.70	0.9965

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal

El valor de z está derivado de la fórmula:

$$Z = \frac{x - \mu}{\sigma}$$

En la que:

- x = valor de la variable aleatoria que nos preocupa.
- μ = media de la distribución de la variable aleatoria.
- σ = desviación estándar de la distribución.
- z = número de desviaciones estándar que hay desde x a la media de la distribución. (el uso de z es solamente un cambio de escala de medición del eje horizontal)

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

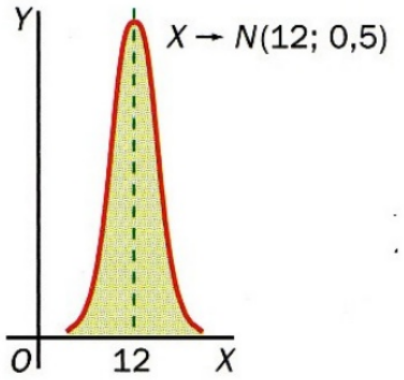
Distribución Normal

k	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal

Tipificación de la variable



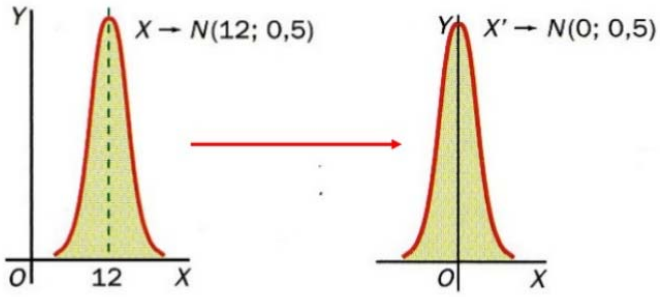
UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal

Centrar

Hacemos $\mu = 0$ con cambio de variable

$$X' = X - 12$$



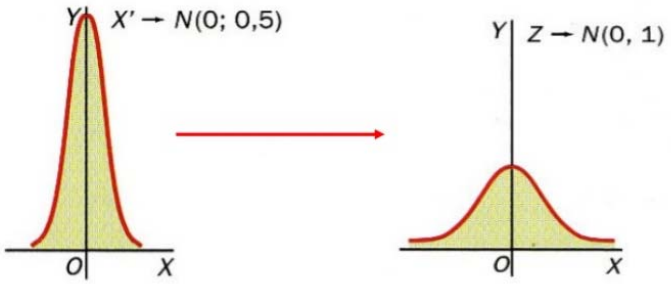
UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal

Contraer o dilatar

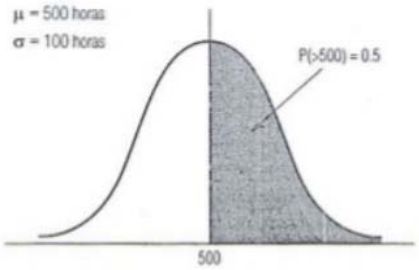
Hacemos $\sigma = 1$ con cambio de variable

$$Z = \frac{X - 12}{0{,}5}$$



UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Ejemplo 1:
Hay un programa de entrenamiento diseñado para mejorar la calidad de las habilidades de supervisión de los supervisores de la línea de producción. Debido a que el programa es autoadministrado, los supervisores requieren un número diferente de horas para terminarlo. Un estudio de los participantes anteriores indica que el tiempo medio que se lleva completar el programa es de 500 horas, y que esta variable aleatoria normalmente distribuida tiene una desviación estándar de 100 horas. ¿Cuál es la probabilidad de que un candidato elegido al azar requiera más de 500 horas para completar el programa de entrenamiento?



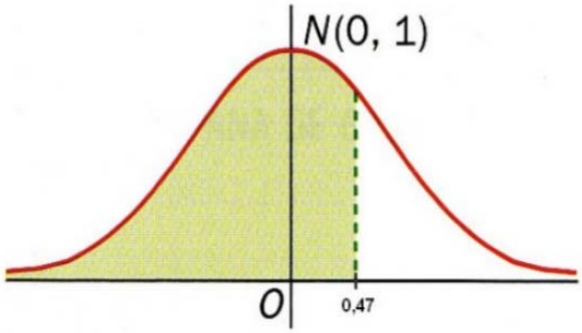
De la gráfica podemos observar que la mitad del área bajo la curva está localizada a ambos lados de la media de 500 horas. **Por tanto, podemos deducir que la probabilidad de que la variable aleatoria tome un valor mayor a 500 es el área sombreada, es decir, 0.5**

30

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal

$$P(Z < 0{,}47)$$



UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal

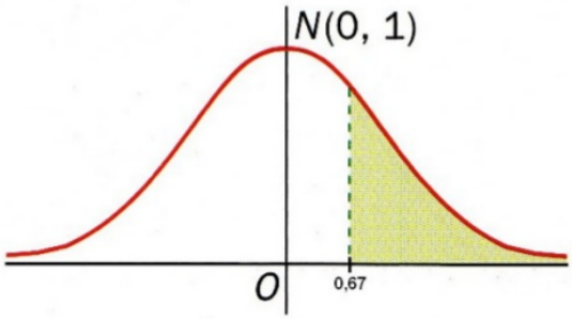
k	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549

$$P(Z < 0,,47) = 0{,}6808$$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal

$$P(Z > 0{,}67)$$



$$P(Z > 0{,}67) = 1 - P(Z \leq 0{,}67)$$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal

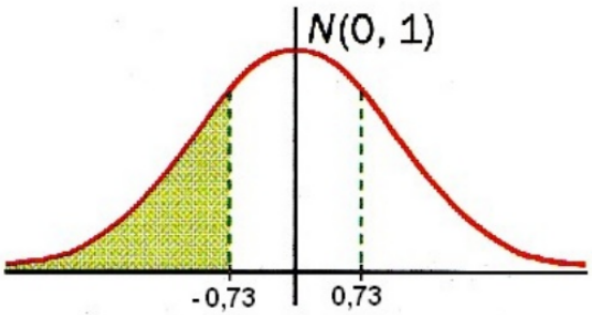
k	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794

$$P(Z > 0{,}67) = 1 - 0{,}7486 = 0{,}2514$$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal

$$P(Z < -0{,}73)$$



$$P(Z < -0{,}73) = P(Z > 0{,}73) = 1 - P(Z \leq 0{,}73)$$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal

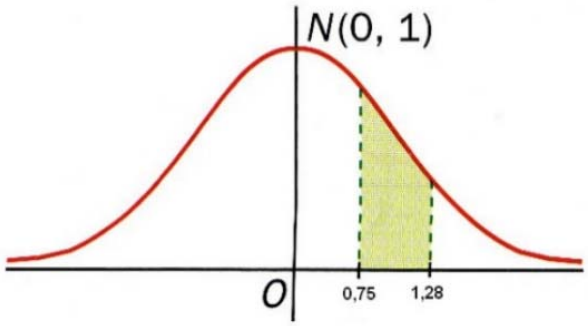
k	0,00	0,01	0,02	0,03	0,04	0,05
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749

$$P(Z < -0{,}73) = 1 - 0{,}7673 = 0{,}2327$$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal

$$P(0{,}75 < Z < 1{,}28)$$



$$P(0{,}75 < Z < 1{,}28) = P(Z < 1{,}28) - P(Z < 0{,}75)$$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal

k	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306

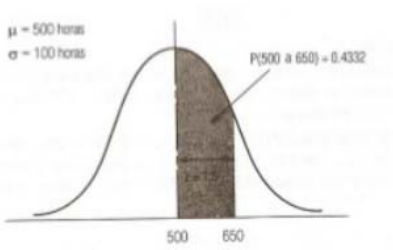
$$P(0{,}75 < Z < 1{,}28) = 0{,}8997 - 0{,}7734 = 0{,}1263$$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Ejemplo 2:
Partiendo de la misma premisa, μ = 500 y σ = 100. ¿Cuál es la probabilidad de que un candidato elegido al azar se tome entre 500 y 650 horas en completar el programa de entrenamiento?

$$Z = \frac{x - \mu}{\sigma} \qquad Z = \frac{650 - 500}{100}$$

Z=1.5 desviaciones estándar

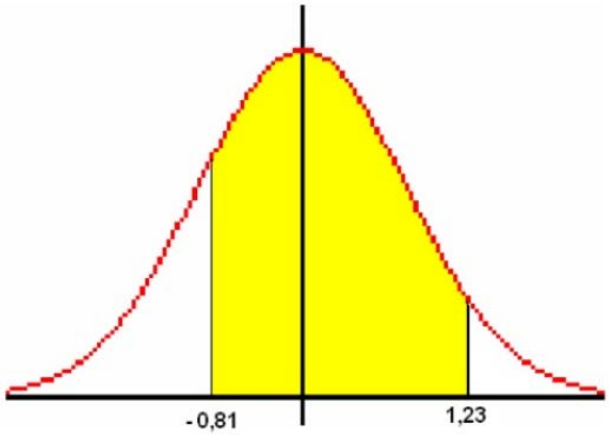


Si buscamos Z=1.5 (refiérase a la tabla de la pág. 28), encontramos una probabilidad de 0.4332.

Por lo tanto, la probabilidad de que un candidato escogido al azar requiera entre 500 y 650 horas para terminar el programa de entrenamiento es de 0.4332

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Distribución Normal . Ejercicio para hacer.



UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

- Se sabe que algunas distribuciones convergen a la normal a medida que sus parámetros se aproximan a ciertos límites.
  - Cuando n es considerablemente grande

- Distribución binomial
  - dos resultados posibles: éxito,fracaso
  - n: tamaño de muestra
  - p: probabilidad de éxito
  - q=1-p: probabilidad de fracaso

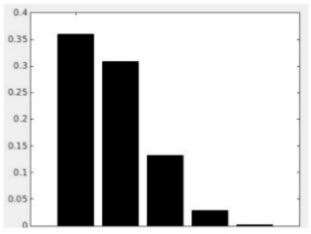
- Distribución normal
  - μ: media de la población **μ=np**
  - σ: desviacion estandar de la población **σ=√(n.p.q)**

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

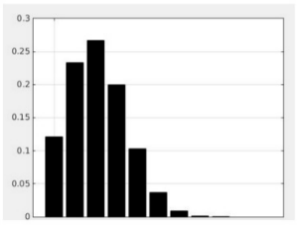
Aproximación de Binomial a Normal

Ejemplo de distribución binomial: p=0.3 q=0.7

- Si n=5 p=0.3 q=0.7



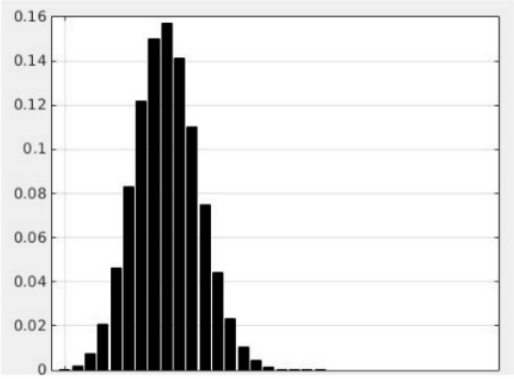
- Si n=10 p=0.3 q=0.7



UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

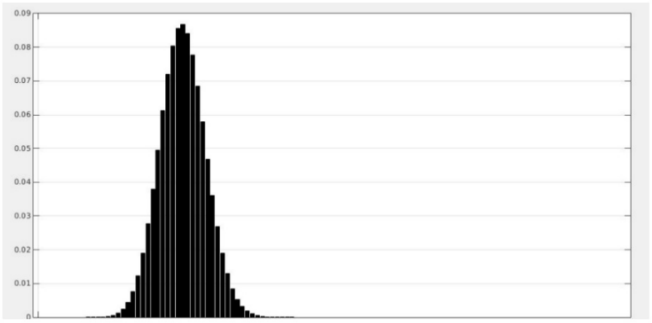
- Si n=30 p=0.3 q=0.7



UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

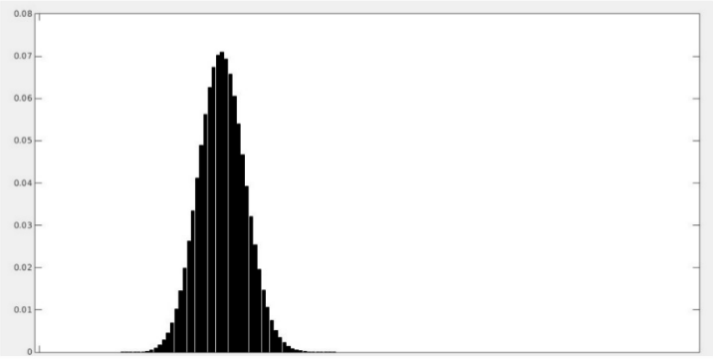
- Si n=100 p=0.3 q=0.7



UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

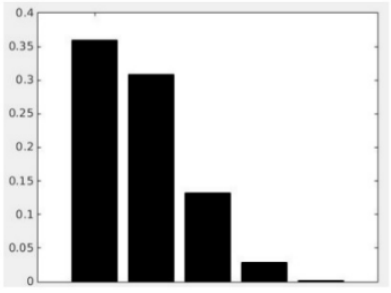
- Si n=150 p=0.3 q=0.7



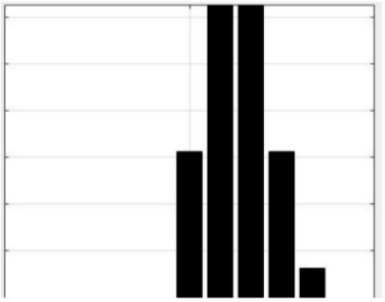
UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

- Ejemplo 1: n=5 p=0.3 q=0.7



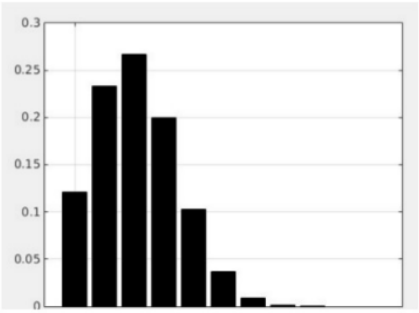
- Ejemplo 2: n=5 p=0.5 q=0.5



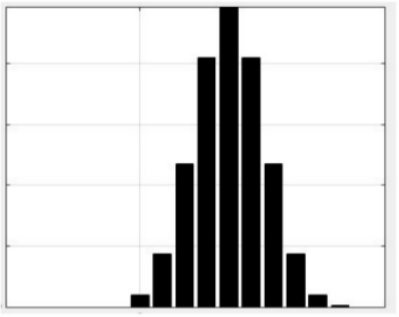
UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

- Ejemplo 1: n=10 p=0.3 q=0.7



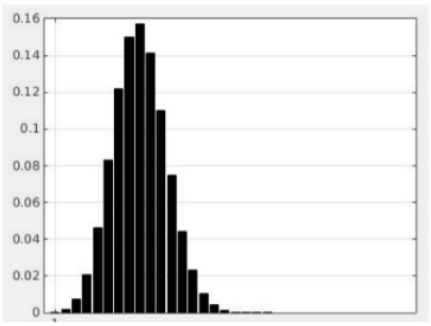
- Ejemplo 2: n=10 p=0.5 q=0.5



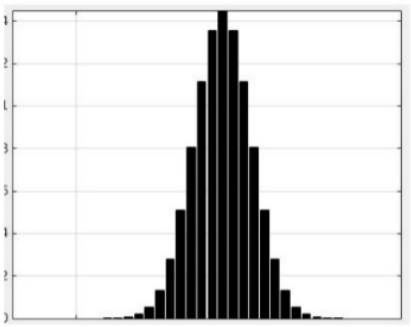
UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

- Ejemplo 1: n=30 p=0.3 q=0.7



- Ejemplo 2: n=30 p=0.5 q=0.5



UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

1.- p = 0.50 ó 50%.

Si la proporción de éxitos **"p"** es igual al **50%**, independientemente del tamaño de la muestra, la aproximación es viable por derecho propio.

2.- p ≠ 0.50 ó 50%.

Si la proporción de éxitos **"p"** es diferente al **50%**, otra guía posible para lograr la aproximación es a través de la regla empírica que prueba el número esperado de éxitos y fracasos. Es decir,

$$\left.\begin{aligned}\mu_{np} &= np > 5\\ \mu_{nq} &= nq > 5\end{aligned}\right\}\ \text{Simultanea}$$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

- La distribución normal con **μ=np** y **σ²= npq** ofrece una muy buena aproximación cuando:
  - **n** es grande y **p** no esta extremadamente cerca de 0 o 1
  - **n** es pequeña pero **p** esta razonablemente cerca de 1/2

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

Si X es una variable aleatoria binomial con media **μ = np** y desviación estandar **σ= √(npq)**, entonces la forma limitante de la distribución de

$$Z = \frac{X-\mu}{\sigma}$$

conforme n--> ∞, es la distribución normal estandar

Permitirá utilizar áreas bajo la curva normal para aproximar propiedades binomiales cuando n es suficientemente grande.

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

FACTOR DE CORRECCIÓN POR CONTINUIDAD

Al aproximar estas distribuciones, la probabilidad que está por estimarse se verá afectada debido a que existe un número infinito de valores posibles en una distribución continua y la probabilidad de que la variable aleatoria sea exactamente igual a algún valor específico, es cero.

Por lo anterior, cuando se utiliza una distribución continua para estimar una variable aleatoria discreta, es necesario hacer un ajuste. Este método se denomina factor de corrección por continuidad y consiste en sumar o restar, según sea el caso, 0.5 del valor de la variable debido a su origen o naturaleza.

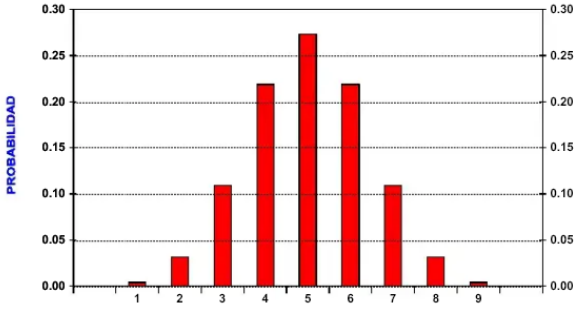
$$X_{corregido} = X_c = X \pm 0.5$$

De esta forma, se crean los límites o intervalos para dar continuidad a la variable.

Aproximación de Binomial a Normal

DISTRIBUCIÓN PARA VALORES DISCRETOS

DISTRIBUCIÓN DE PROBABILIDADES
PARA EL NÚMERO DE CLIENTES QUE PRESENTAN QUEJAS
DURANTE EL PERÍODO DE GARANTÍA



NÚMERO DE CLIENTES QUE PRESENTAN QUEJAS

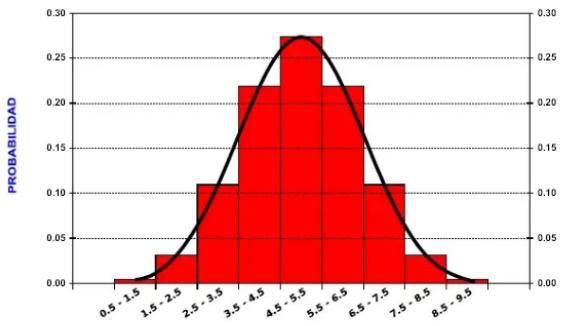
UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

Se crean los límites de los intervalos para dar continuidad a la variable formando así un histograma.

DISTRIBUCIÓN PARA VALORES CONTINUOS

DISTRIBUCIÓN DE PROBABILIDADES
PARA EL NÚMERO DE CLIENTES QUE PRESENTAN QUEJAS
DURANTE EL PERÍODO DE GARANTÍA



NÚMERO DE CLIENTES QUE PRESENTAN QUEJAS

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

Finalmente, la función de la variable tipificada queda como sigue:

$$Z = \frac{X_c - \mu_{np}}{\sigma_{np}} \quad \left\} \quad \begin{array}{l} \mu_{np} = np \\ \sigma_{np} = \sqrt{npq} \end{array} \right.$$

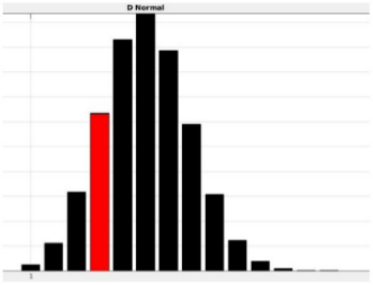
UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

- La probabilidad exacta de que la variable aleatoria binomial X tome un valor determinado x es igual al área de la barra cuya base se centra en x.
  - Por ejemplo, la probabilidad exacta de que X tome el valor 4 es igual al área del rectángulo con base centrada en x = 4.

P (X = 4) = b(4; 15, 0.4) = 0.1268

μ=0.4*15= 6
σ=√(15*0.4*0.6)=1.8974



UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

- Que es aproximadamente igual al área de la región sombreada bajo la curva normal entre las dos ordenadas x1= 3.5 y x2= 4.5

$$Z1 = \frac{3.5-6}{1.897} = -1.32 \qquad Z2 = \frac{4.5-6}{1.897} = -0.79$$

P (X = 4) = b(4; 15, 0.4) ≈ P (−1.32 < Z < −0.79)

= P (Z <−0.79) − P (Z < −1.32) = 0.2148 − 0.0934

= 0.1214

0.1268 ≈ 0.1214

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

- La probabilidad de que X tome un valor de 7 a 9 es dada por:

$$P(7 \le X \le 9) = \sum_{x=0}^{9} b(x; 15, 0.4) - \sum_{x=0}^{6} b(x; 15, 0.4)$$

= 0.9662 − 0.6098 = 0.3564

  - Que es igual a la sumatoria de las áreas de los rectángulos cuyas bases están centradas en x = 7, 8 y 9

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

- Para la aproximación normal calculamos el área de la región sombreada bajo la curva entre las ordenadas x1=6.5 y x2=9.5

$$Z1 = \frac{6.5-6}{1.897} = 0.26 \qquad Z2 = \frac{9.5-6}{1.897} = 1.85$$

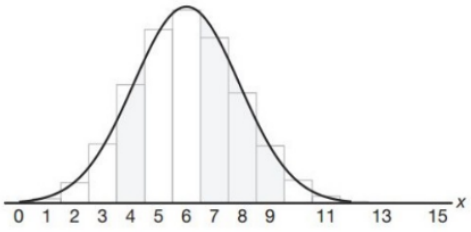
- $P(7 \leq X \leq 9) \approx P(0.26 < Z < 1.85) = P(Z < 1.85) - P(Z < 0.26)$
  $= 0.9678 - 0.6026 = 0.3652$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

- En los ejemplos anteriores se vio que si buscamos el área bajo la curva normal hacia la izquierda de, digamos x, es más preciso utilizar x + 0.5(**corrección de continuidad**)

$P(X = 4) = b(4; 15, 0.4) = 0.1268$ $\quad P(-1.32 < Z < -0.79) = 0.1214$



UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

- Sea X una variable aleatoria binomial con parámetros n y p. Para una n grande, X tiene aproximadamente una distribución normal con μ= np y σ= √(npq)

$$P(X \le x) = \sum_{k=0}^{x} b(k;n,p) \approx \text{área bajo la curva normal a la izq de } x+0.5$$

$$= P\left(Z \le \frac{x+0.5-np}{\sqrt{npq}}\right)$$

y la aproximación será buena si np y nq son mayores que 5

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

- Un paciente que padece una rara enfermedad de la sangre tiene 0.4 de probabilidad de recuperarse. Si se sabe que 100 personas contrajeron esta enfermedad, ¿cuál es la probabilidad de que sobrevivan menos de 30?
  - n es grande
  - p esta relativamente cerca de 1/2

μ=100*0.4= 40
σ=√(100*0.4*0.6)=4.899

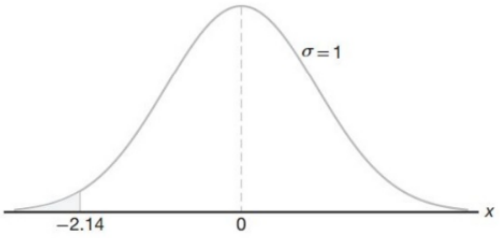
UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

- Para obtener la probabilidad que se desea, tenemos que calcular el área a la izquierda de x = 29.5

$$Z = \frac{29.5-40}{4.899} = -2.14$$

- P (X < 30) ≈ P (Z <−2.14) = 0.0162



UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

- Un examen de opción múltiple tiene 200 preguntas, cada una con 4 respuestas posibles, de las que sólo una es la correcta. ¿Cuál es la probabilidad de que solamente adivinando se obtengan de 25 a 30 respuestas correctas para 80 de los 200 problemas sobre los que el estudiante no tiene conocimientos?

$$P(25 \le X \le 30) = \sum_{x=25}^{30} b(x; 80, 1/4)$$

μ=80*1/4= 20

σ=√(80*1/4*3/4)=3.873

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

- Se necesita el área entre x1 = 24.5 y x2 = 30.5. Los valores z correspondientes son

$$z_1 = \frac{24.5 - 20}{3.873} = 1.16 \qquad z_2 = \frac{30.5 - 20}{3.873} = 2.71$$

$$P(25 \le X \le 30) = \sum_{x=25}^{30} b(x; 80, 1/4) \approx P(1.16 < Z < 2.71)$$

$$= P(Z < 2.71) - P(Z < 1.16) = 0.9966 - 0.8770 = 0.1196$$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

- La probabilidad de que cierta clase de componente sobreviva a una prueba de choque es de 3/4. Calcule la probabilidad de que sobrevivan exactamente 2 de los siguientes 4 componentes que se prueben.

$\mu = 4*3/4 = 3$

$\sigma = \sqrt{(4*3/4*1/4)} = 0.75$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Aproximación de Binomial a Normal

- Se necesita el área entre x1 = 1.5 y x2 = 2.5. Los valores z correspondientes son

$$Z1 = \frac{1.5-3}{0.75} = -2 \qquad Z2 = \frac{2.5-3}{0.75} = -0.66$$

$$b\left(2; 4, \frac{3}{4}\right) = \binom{4}{2}\left(\frac{3}{4}\right)^2\left(\frac{1}{4}\right)^2 = \left(\frac{4!}{2!\,2!}\right)\left(\frac{3^2}{4^4}\right) = \frac{27}{128}. \quad =0.2109$$

- ≈ P (-2 < Z < -0.66) = P (Z < -0.66) − P (Z < -2)
  = 0.2877- 0.0228 = 0.2649



UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

Índice

1. Intro a la Inferencia Estadística
2. Estadísticos mas usuales
3. Estimación Puntual
4. Intervalos de Confianza
5. Contraste de Hipotesis

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

1. Intro a la Inferencia Estadística

Elementos poblacionales =

= características de la variable en estudio sobre la población

habitualmente desconocidos ya que no son directamente observados

Media, varianza, una probabilidad, cierto parámetro de la distribución, …

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

1. Intro a la Inferencia Estadística

¿Cómo obtener más información sobre ellos?

El censo de toda la población suele ser impracticable

Mediante el examen de la variable sobre un pequeño grupo o muestra de la población

1. Intro a la Inferencia Estadística

Resumiendo:

se pretende obtener información (realizar inferencia) acerca del elemento poblacional de interés,

a partir del examen del valor de la variable en un pequeño grupo o muestra (inducción)

Inferencia inductiva estadística ó

Estadística inductiva ó inferencial

será la inducción que se realice empleando herramientas de tipo estadístico

1. Intro a la Inferencia Estadística

Formalización del concepto de muestra de una v.a. X

se escogen aleatoriamente n (tamaño muestral) individuos de la población

para medir X en cada uno de ellos: $X_1 \ldots . X_n$

Muestra de X es ese conjunto de n variables $X_1 \ldots . X_n$ que cumplen $X_i \sim X \ \forall i \not\Rightarrow X_i = X$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

1. Intro a la Inferencia Estadística

Una muestra puede ser:

Muestra aleatoria simple (m.a.s.):
individuos escogidos independientes entre sí
X_i independiente de X_j si $i \neq j$

Otros tipos: existe dependencia.

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

1. Intro a la Inferencia Estadística

Realización muestral:
conjunto de valores numéricos que toman las n variables una vez seleccionada la muestra y medidas las n variables

$x_1,\ldots,x_n$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

1. Intro a la Inferencia Estadística

La inferencia estadística se realiza mediante **estadísticos**, con ellos se extrae la información de la muestra:

función de la muestra, $T(X_1,\ldots,X_n)$, resultante de aplicarle una función determinista $T(x_1,\ldots,x_n)$: $R^n \rightarrow R^m$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

1. Intro a la Inferencia Estadística

Un estadístico, $T(X_1,\ldots,X_n)$, es también una v.a.

Tendrá su propia distribución de probabilidad (distribución muestral)

Tendrá su propia realización muestral, $T(x_1,\ldots,x_n)=t$, una vez disponible la realización de la muestra, $(x_1,\ldots,x_n)$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

1. Intro a la Inferencia Estadística

Ejemplo

Población analizada:
familias de cierta localidad

Variable en estudio:
renta familiar disponible en miles de € (X)

Parámetro poblacional de interés:
renta media disponible (E(X))

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

1. Intro a la Inferencia Estadística

Censo no factible.

Se decide examinar una muestra de n=5 familias elegidas al azar …

… para obtener información (hacer inferencia) sobre la renta media a partir de las rentas de tales familias (inducción)

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

1. Intro a la Inferencia Estadística

Las rentas de las 5 familias

$$X_1, X_2, X_3, X_4, X_5$$

constituyen una m.a.s. de X de tamaño 5:

$X_i \sim X \ \forall i$ e independientes entre si.

Una vez seleccionadas las familias, los valores de sus rentas constituyen una realización muestral

$x_1=47{,}2$, $x_2=20{,}1$, $x_3=7{,}8$, $x_4=28{,}3$, $x_5=15{,}9$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

1. Intro a la Inferencia Estadística

Para hacer inducción sobre E(X), el estadístico habitual que se maneja es

$$T(X_1,\ldots,X_5)=\overline{X}=\sum_{i=1}^{5} X_i/5$$

que es resultado de aplicar a la muestra la función determinista

$$T(x_1,\ldots,x_n)=\sum_{i=1}^{n} x_i/n : R^n \to R$$

en el caso n=5.

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

1. Intro a la Inferencia Estadística

$\overline{X}$ es una v.a con su propia distribución (distribución muestral) derivada de la de las X_i (~X)

(Por ejemplo, será una $N(\mu,\sigma/5^{½})$ si X~ $N(\mu,\sigma)$)

$\overline{X}$ tomará su propia realización muestral una vez disponible la de la muestra

$$t = T(x_1=47,2\ ,\ \dots\ ,\ x_5=15,9) =$$
$$= (47,2+\dots+15,9)/5 = 23,86\ €$$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

1. Intro a la Inferencia Estadística

Importante:

Los resultados inferenciales

están siempre sujetos a error y por ello

solo tendrán cierto grado de fiabilidad

que deberá siempre cuantificarse

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

2. Estadísticos mas usuales

MEDIA MUESTRAL

(inferencia sobre E(X))

$$\overline{X} = \sum_{i=1}^{n} X_i / n$$

Propiedades:

$E(\overline{X}) = E(X)$

$V(\overline{X}) = V(X)/n$ si m.a.s.

$\overline{X} \approx N(E(X), dt(X)/n^{1/2})$ bajo condiciones TCL

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

2. Estadísticos mas usuales

VARIANZA Y CUASIVARIANZA MUESTRALES

(inferencia sobre V(X))

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n} = \frac{\sum_{i=1}^{n} X_i^2}{n} - \overline{X}^2$$

$$S_c^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{n}{n-1} S^2$$

2. Estadísticos mas usuales

Asociadas a ellas están:

Desviación típica muestral: $S=(S^2)^{1/2}$

Cuasidesviación típica muestral: $S_c=(S_c^2)^{1/2}$

implicadas en la inferencia sobre dt(X), desviación típica poblacional

2. Estadísticos mas usuales

FRECUENCIA EMPÍRICA O MUESTRAL

(inferencia sobre $p=P(X\in S)$ con S subconjunto de R)

Un determinado comportamiento de interés de X

$p_{emp.} = P_{emp.}(X\in S) = $ nº de $X_i\in S/n =$

= tanto por uno de $X_i\in S$ en la muestra

Propiedades:

$E(p_{emp.}) = p$

$V(p_{emp.}) = p(1-p)/n$ si m.a.s.

$p_{emp.} \approx N(p,(p(1-p)/n)^{1/2})$ si m.a.s.

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

2. Estadísticos mas usuales

Ejemplo (continuación)

Con la realización muestral disponible de la m.a.s de tamaño 5 de la renta familiar

47,2 20,1 7,8 28,3 15,9

es posible calcular la realización muestral de $\overline{X}$, S^2, S_c^2, S, S_c y de la frecuencia empírica de que la renta familiar sea baja (por debajo de 12.000 €)

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

2. Estadísticos mas usuales

Basta con realizar los correspondientes cálculos:

$$\overline{X} = \sum_{i=1}^{n} X_i / n = 23{,}86$$

$$S^2 = \left(\sum_{i=1}^{n} (X_i - \overline{X})^2\right) / n = 179{,}9784$$

$$S_C^2 = \left(\sum_{i=1}^{n} (X_i - \overline{X})^2\right) / (n-1) = 224{,}973$$

2. Estadísticos mas usuales

Basta con realizar los correspondientes cálculos:

$S = (S^2)^{½} = 13,4156$

$S_c = (S_c^2)^{½} = 14,9991$

$p_{emp.}(\text{renta baja}) = p_{emp.}(X \leq 12) =$
$= \text{nº de } X_i \leq 12 \,/n = 1/5 = 0,20 \;\; (20\%)$

2. Estadísticos mas usuales

Técnicas de inferencia estadística

- Estimación puntual.
- Estimación por intervalos
- Contraste de hipótesis

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

3. Estimación puntual

Resultados muestrales en poblaciones normales

X~N(μ,σ) (Población normal)

distribución de uso amplio de forma exacta o aproximada (recordar TCL)

Se supondrá: $X_1,...,X_n$ m.a.s. de X

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

3. Estimación puntual

Los problemas de inferencia relacionados con la distribución normal pueden referirse a una sola población

en las condiciones generales que se han descrito hasta ahora para cualquier variable

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

3. Estimación puntual

pero al caracterizarse por dos parámetros (μ y σ), la inferencia en una población normal puede ser de cuatro tipos distintos:

SOBRE μ con σ conocida

SOBRE μ con σ desconocida

SOBRE σ con μ conocida

SOBRE σ con μ desconocida

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

3. Estimación puntual

1 muestra - INFERENCIA SOBRE μ con σ conocida

$$\overline{X} \sim N(\mu, \sigma/\sqrt{n}) \text{ ó } Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

3. Estimación puntual

1 muestra - INFERENCIA SOBRE μ con σ desconocida

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$ resultado inútil por

contener σ. En su lugar se emplea:

$$T = \frac{\overline{X} - \mu}{\frac{S_c}{\sqrt{n}}} = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n-1}}} \sim t_{n-1}$$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

3. Estimación puntual

Pero cuando se trabaja con normales, también es habitual que la inferencia a realizar tenga como objetivo comparar dos poblaciones independientes:

Se estudian dos poblaciones independientes con respectivas X e Y

Se desea realizar inferencia para comparar elementos de sus distribuciones

Se dispone de

$X_1,\ldots,X_n \sim X$ e $Y_1,\ldots,Y_m \sim Y$

dos muestras independientes entre sí

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

3. Estimación puntual

Ejemplo:

Al confrontar la renta (X) de las familias murcianas con la de las alicantinas (Y)

interesa comparar sus medias, E(X) y E(Y).

Se dispone de dos muestras:

200 familias murcianas: $X_1,...,X_{200} \sim X$

100 familias alicantinas: $Y_1,...,Y_{100} \sim Y$

independientes entre sí

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

3. Estimación puntual

Si, adicionalmente, se supone normalidad

$$X \sim N(\mu_X,\sigma_X) \text{ e } Y \sim N(\mu_Y,\sigma_Y)$$

lo que habitualmente se compara es

medias entre sí (mediante $\mu_X-\mu_Y$)

ó

varianzas entre sí (mediante σ_X/σ_Y ó σ^2_X/σ^2_Y)

Empleándose siempre mm.aa.ss.

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

3. Estimación puntual

La inferencia en dos poblaciones normales independientes puede ser de varios tipos distintos:

SOBRE μ_X-μ_Y

- CASO 1: se sabe que σ_X=σ_Y
- CASO 2: varianzas completamente desconocidas

SOBRE σ_X/σ_Y

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

3. Estimación puntual

2 muestras indep. - INFERENCIA SOBRE μ_X-μ_Y

- CASO 1: se sabe que σ_X=σ_Y

El siguiente resultado es fácil de demostrar pero no es útil porque σ_X y σ_Y aparecen en él ('parámetros molestos' desconocidos)

$$\frac{\overline{X}-\overline{Y}-(\mu_x-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \sim N(0,1)$$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

3. Estimación puntual

2 muestras indep. - INFERENCIA SOBRE μ_X-μ_Y

- CASO 1: se sabe que $\sigma_X=\sigma_Y$

Hay que estimarlas mediante $S_{c,X}$ y $S_{c,Y}$ y con una demostración similar a la anterior, llegar a :

$$\sqrt{nm\frac{n+m-2}{n+m}}\left(\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{(n-1)S_{c,X}^2+(m-1)S_{c,Y}^2}}\right)\sim t_{n+m-2}$$

2 gg.ll. perdidos

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

3. Estimación puntual

2 muestras indep. - INFERENCIA SOBRE μ_X-μ_Y

- CASO 2: σ_X y σ_Y complet. desconocidas

En este caso, tras estimar mediante $S_{c,X}$ y $S_{c,Y}$ solo se llega a una distribución asintótica para el estadístico resultante:

$$\frac{\bar{X}-\bar{Y}-(\mu_x-\mu_Y)}{\sqrt{\frac{S_{c,X}^2}{n}+\frac{S_{c,Y}^2}{m}}}\approx t_\nu$$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

3. Estimación puntual

2 muestras indep. - INFERENCIA SOBRE μ_X-μ_Y

- CASO 2: σ_X y σ_Y complet. desconocidas

donde los grados de libertad de la t valen:

ν = entero más cercano a

$$\frac{\left(S_{c,X}^2/n + S_{c,Y}^2/m\right)^2}{\frac{\left(S_{c,X}^2/n\right)^2}{n-1} + \frac{\left(S_{c,Y}^2/m\right)^2}{m-1}}$$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

3. Estimación puntual

Resultados muestrales en poblaciones de Bernoulli

X~b(p) (Población de Bernoulli)

Se supondrá: $X_1,\ldots,X_n$ m.a.s. de X

Resultados asintóticos para una sola población y para la comparación de dos poblaciones independientes también disponibles en las tablas; fáciles de entender y demostrar

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

3. Estimación puntual

UNA SOLA POBLACIÓN

INFERENCIA SOBRE p=prob. de éxito=P(X=1)

$p_{emp.} = P_{emp.}(X=1) = (nº \text{ de } X_i=1)/n =$

= tanto por uno de éxitos en la muestra =

$= \overline{X} = \sum_{i=1}^{n} X_i / n \approx N(p,(p(1-p)/n)^{1/2})$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza

- La estimación puntual no es un resultado inductivo.
- c_0 = *VALOR de la magnitud*
  e_0 = *error MÁXIMO cometido*

$$\begin{cases} c_0 \pm e_0 \\ (c_0 - e_0 \, , \, c_0 + e_0) \end{cases}$$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza

- La Estadística adopta el formato de intervalo

y considera el carácter inductivo de sus resultados

$(c_0 - e_0 , c_0 + e_0)$

+

Medida de la fiabilidad de e_0



Intervalo de confianza

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalos de confianza

- El proceso de **INFERENCIA** será:
  - Definir una variable y su población
  - Extraer una muestra
  - Estimar a partir de la muestra el intervalo de valores dónde se hallará la media de la población con cierta probabilidad
- EJEMPLO
  - De la población de España, encuestamos a 1000 personas sobre si fuman o no (muestra), y con los datos obtenidos estimaremos en qué intervalo de valores se encuentra la proporción de fumadores en toda la población de España (por ejemplo, entre 0.25 y 0.30)
  - Como este proceso no es exacto, queremos saber qué probabilidad hay de que efectivamente la proporción de fumadores en España esté entre 0.25 y 0.30 (por ejemplo, con probabilidad del 95%)

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

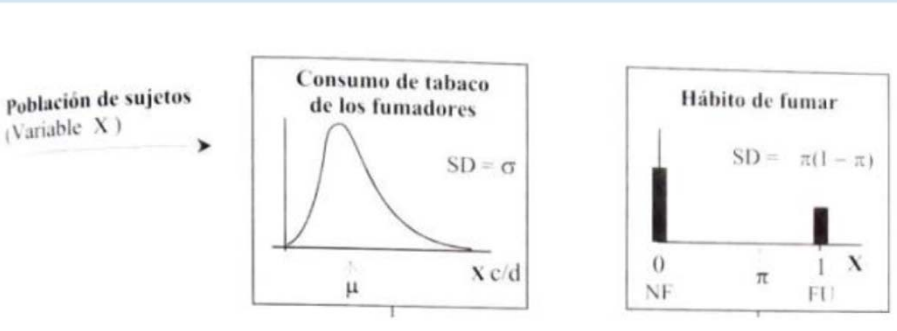
4. Intervalo de confianza

- Veamos primero qué relación hay entre **la distribución de una muestra** y **la distribución de la población** con un supuesto:
- SUPUESTO: Población de fumadores
  - Supongamos que tenemos una población de fumadores (ejemplo, todos los fumadores de un país), y que conocemos la **cantidad de tabaco** que consume cada uno (variable a estudiar).
  - Supongamos que conocemos su distribución, es decir, su media (μ) y desviación estándar(σ o SD en inglés o D.E. en español)
  - También podemos conocer el valor exacto de la **proporción** de fumadores (π)
  - Queremos saber la relación entre la distribución de todas las posibles muestras y la distribución de la población

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza

DISTRIBUCIÓN DE LA POBLACIÓN



UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza

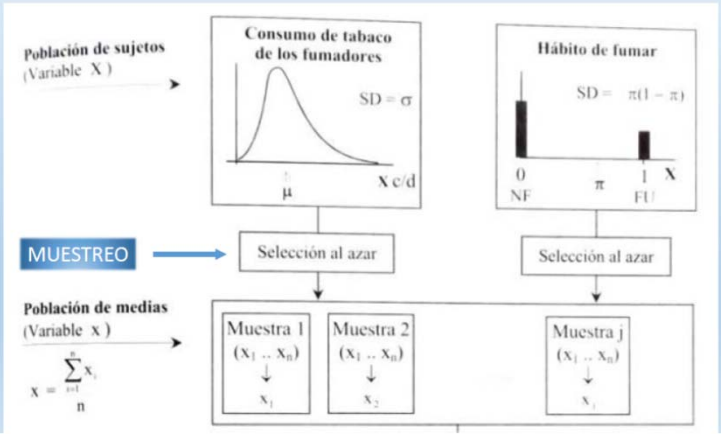
- Vamos a construir una **población de medias** (ejercicio "teórico"):
  - Se construye el **conjunto de todas las posibles muestras**, de un mismo tamaño **n** seleccionados al azar, procedentes de la población (de la población conocemos sus parámetros μ, σ, π)
  - Para cada muestra extraída se calcula por un lado su media $\bar{x}_i$ y por otro la proporción π_{mi} y se repite el proceso todas las veces que haga falta. Se obtiene así una **población de medias** $\bar{x}, \pi_m$



UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza

- Vamos a construir una **población de medias**:



UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza

- Es de esperar que las medias de las diferentes muestras sean muy parecidas entre sí pero fluctuarán con un error de tipo aleatorio (debido al azar) ➔ **VARIABILIDAD**
  - Ejemplo, la población tiene una proporción de π=28% de fumadores, pero en 5 muestras de n=1000 individuos, las proporciones muestrales pueden oscilar, $\pi_1 = 27\%, \pi_2 = 27{,}5\%, \pi_3 = 30\%, \pi_4 = 29\%, \pi_5 = 28{,}5\%,$
- La media de la distribución de medias coincide con la media de la población ($\mu \approx \bar{x}$ o $\pi \approx \pi_m$) aunque con cierta variabilidad.
  - En el ejemplo anterior sería $\pi_m = \frac{(27+27{,}5+30+29+28{,}5)}{5} = 28{,}4$
- Al aumentar el tamaño n de cada muestra, se obtendrán medias cada vez más parecidas a la de la población

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza
Teorema central del Límite

- **IMPORTANTE:** La distribución de una variable media, cuando es el promedio de un número n suficientemente grande de observaciones independientes entre sí, se aproxima a la distribución normal.
- Se considera muestra grande cuando n>30 (tamaño de la muestra)
- ¡Cuidado! Si la asimetría de la variable es muy marcada se requieren tamaños mayores

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza

- ¿Qué pasa cuándo el tamaño es pequeño y los datos no se distribuyen de forma normal?
  - Si la variable es cualitativa (de tipo binario) la distribución muestral se explica por la distribución binomial*.
    - Binomial de tipo binario: Sólo dos categorías
    - Ejemplos: Hombre / mujer. Cara / cruz. Diagnóstico positivo / negativo. Fumador / no fumador.
    - *EXPLICACIÓN: Distribución binomial es distribución de una v.a. binaria en la que la probabilidad es fija. Ejemplos:
      - Probabilidad de sacar cara al tirar una moneda: 0.5.
      - Probabilidad de ser fumador: 0.27

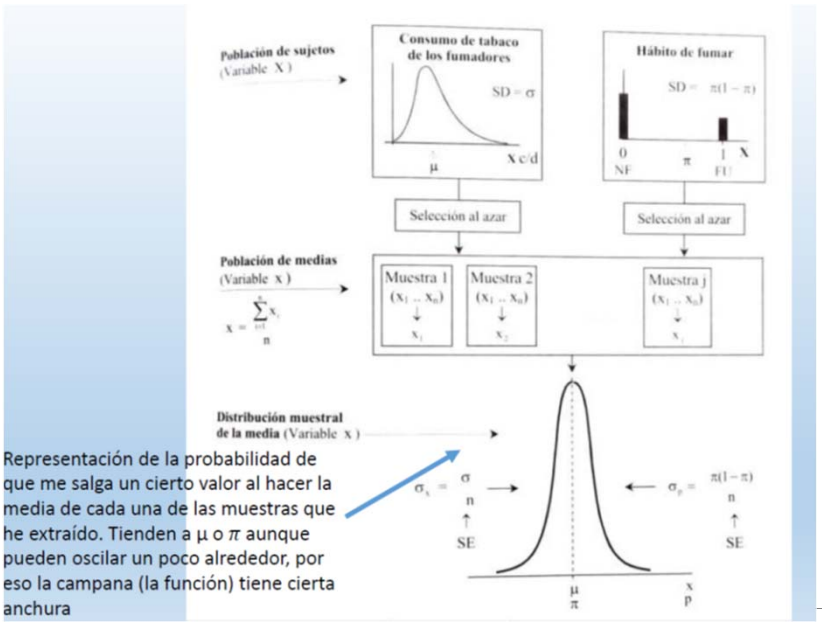
UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza

- ¿Qué pasa cuándo el tamaño es grande aunque los datos no se distribuyen de forma normal?
  - Si la variable es cuantitativa, y las muestras son muy grandes, supondremos que la distribución de la media muestral será **normal**.

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

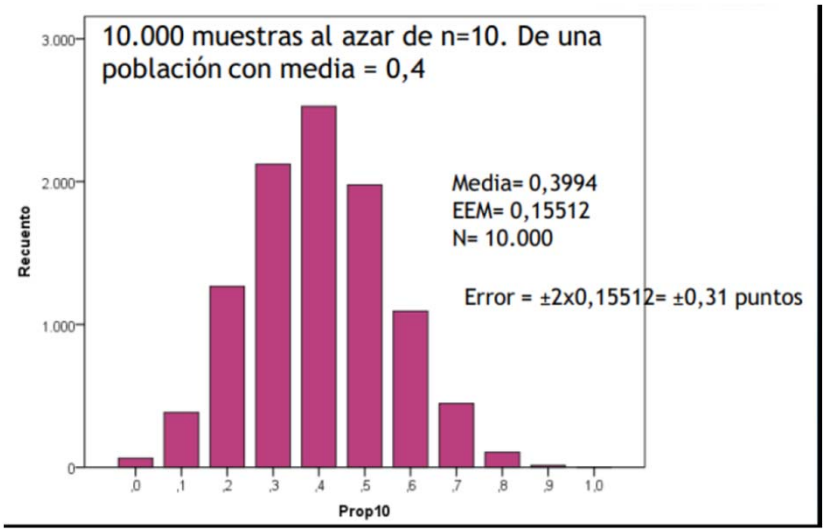
4. Intervalo de confianza



Representación de la probabilidad de que me salga un cierto valor al hacer la media de cada una de las muestras que he extraído. Tienden a μ o π aunque pueden oscilar un poco alrededor, por eso la campana (la función) tiene cierta anchura

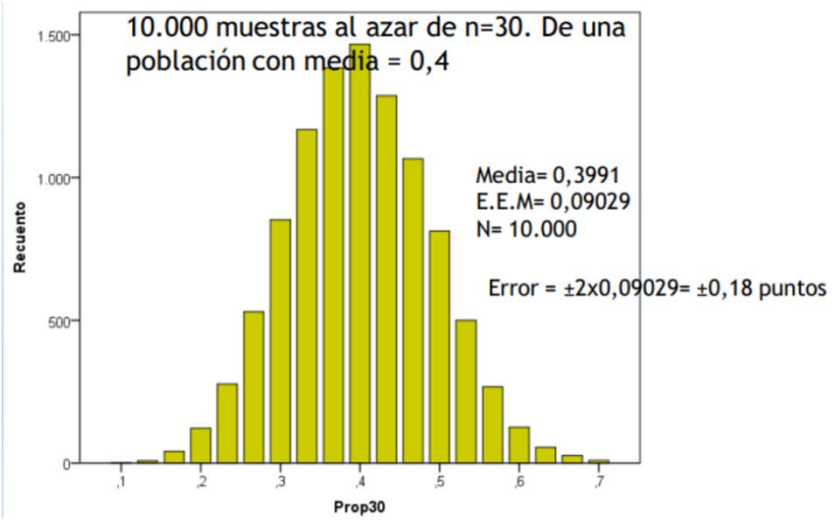
UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza



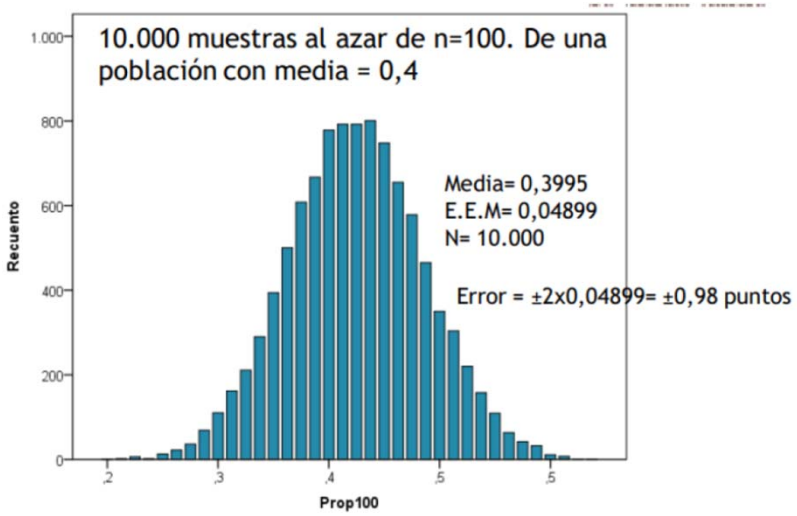
UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza



UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza



UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza

- Para predecir la media poblacional nos gustaría hacerlo sin error, pero el **azar** y el propio **error de muestreo** lo hacen imposible.
- Si queremos predecir el intervalo para la altura media de una población de humanos, podríamos decir que oscilará entre 1,00 m y 2,5 m.
- Al investigador le interesa cerrar al máximo el intervalo (hacerlo pequeño) para localizar en él a la media poblacional **con una cierta confianza**.

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza

Ejemplo 1

Reducir el tiempo de confección empleando nueva maquinaria.

18 obreros elegidos aleatoriamente prueban esa maquinaria:

Minutos consumidos por cada trabajador								
81.2	79.5	71.9	72.1	73.5	69.9	63.6	65.4	74.6
74.4	75.7	81.6	74.1	74.4	77.4	66.8	70.4	69.3

La desviación típica se mantiene en 5.5 minutos.

Estimar por intervalo el tiempo medio que tarda cada trabajador.

La variable se distribuye normalmente.

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza

Población analizada:

$$X \sim N(\mu, \sigma=5.5)$$

Muestra disponible:

$X_1,\ldots,X_n$ con n=18 y $X_i \sim N(\mu, 5.5)$ independientes.

Estimación puntual de μ: $\overline{X} = \frac{\sum_{i=1}^{18} X_i}{18} = 73.1$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza

$$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}} = \frac{5.5}{\sqrt{18}} = 1.296\right) \Rightarrow Z \equiv \frac{\overline{X}-\mu}{1.296} \sim N(0,1)$$

$$\Rightarrow P(-1.96 < Z = \frac{\overline{X}-\mu}{1.296} < 1.96) = 0.95$$

$$-1.96 < \frac{\overline{X}-\mu}{1.296} < 1.96 \Leftrightarrow$$

$$\Leftrightarrow -1.96 \times 1.296 - \overline{X} < -\mu < 1.96 \times 1.296 - \overline{X} \Leftrightarrow$$

$$\Leftrightarrow \overline{X} - 2.54 < \mu < \overline{X} + 2.54$$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza

$$-1.96 < \frac{\overline{X}-\mu}{1.296} < 1.96 \Leftrightarrow \overline{X}-2.54 < \mu < \overline{X}+2.54$$

$$P(-1.96 < Z = \frac{\overline{X}-\mu}{1.296} < 1.96) = 0.95 \Rightarrow$$

$$P_{\mu}(\overline{X}-2.54 < \mu < \overline{X}+2.54) = 95\%$$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza

Como $P_{\mu}(\overline{X}-2.54 < \mu < \overline{X}+2.54) = 95\%$

entonces $(\overline{X}-2.54\,,\,\overline{X}+2.54)$

- ➔ es un intervalo aleatorio
- ➔ contiene a μ con probabilidad 95%

Su realización muestral

(73.1–2.54 , 73.1+2.54) = (70.56 , 75.64)

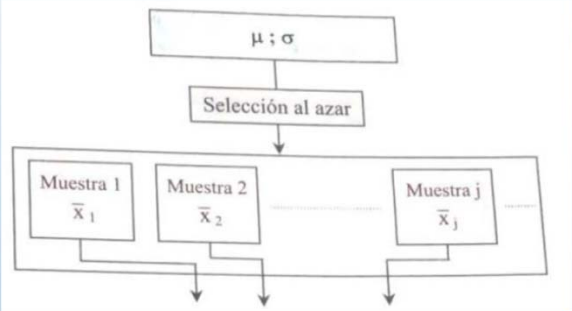
4. Intervalo de confianza

$(73.1-2.54\ ,\ 73.1+2.54) = (70.56\ ,\ 75.64)$

- Acerca de ese intervalo:
  - ➔ No sabemos si contiene a μ.
  - ➔ Erróneo afirmar que tiene una probabilidad del 95% de contener a μ.
  - ➔ Hay una confianza del 95% en que sea una realización de $(\overline{X}-2.54\ ,\ \overline{X}+2.54)$ que contenga a μ.
- Es un intervalo para μ con un nivel de confianza del 95%.
- Es la **estimación por intervalo** buscada para μ.

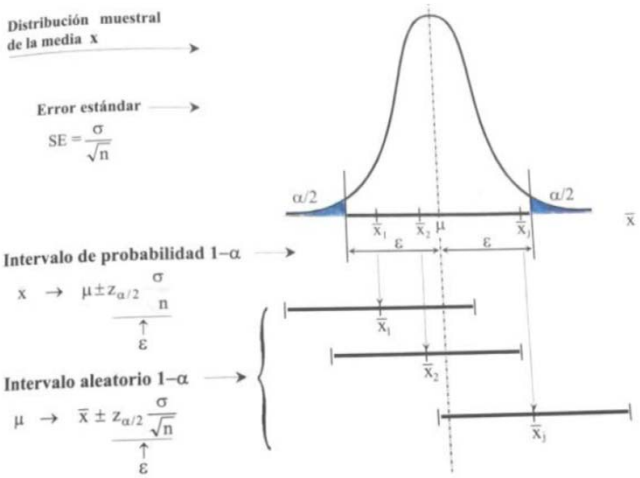
4. Intervalo de confianza

- La población sigue su propia distribución
- Me interesa saber la distribución de las muestras que puedo realizar
- A partir de esa distribución, puedo estimar el intervalo donde CAPTURO el valor de la media de la población con cierta probabilidad



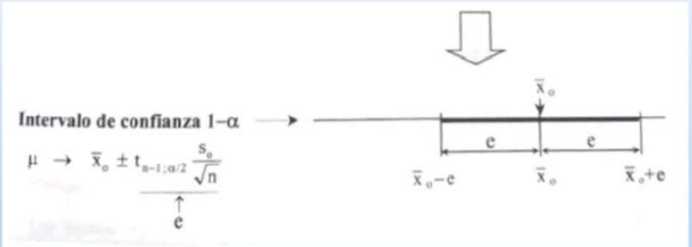
4. Intervalo de confianza

- Tomo una muestra y calculo el intervalo $\bar{x} \pm z_{\alpha/2} {}^{\sigma}/_{\sqrt{n}}$
- Es un intervalo con centro en la media muestral $\bar{x}$ y mide de ancho 2 veces $z_{\alpha/2} {}^{\sigma}/_{\sqrt{n}}$
- Este intervalo captura dentro de ellos el valor de la media de la población μ con probabilidad $1 - \alpha$



4. Intervalo de confianza

- Si conozco la μ de la población, puedo hallar el intervalo alrededor de ella donde se encuentran los valores de $\bar{x}$ con probabilidad $1 - \alpha$, que será $\mu \pm z_{\alpha/2} {}^{\sigma}/_{\sqrt{n}}$
- α es el **riesgo**
- 1- α es la **confianza**



- Los intervalos del tipo $\bar{x} \pm z_{\alpha/2} {}^{\sigma}/_{\sqrt{n}}$ capturan dentro de ellos el valor de μ con probabilidad $1 - \alpha$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza

Valor de riesgo α (alfa)

El valor es arbitrario: no interesa escoger un riesgo excesivo y tampoco muy pequeño.

Lo decidimos nosotros:

Probabilidad del 95%	1-α = 0.95	α = 0.05
Probabilidad del 90%	1-α = 0.90	α = 0.10
Probabilidad del 99%	1-α = 0.99	α = 0.01

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

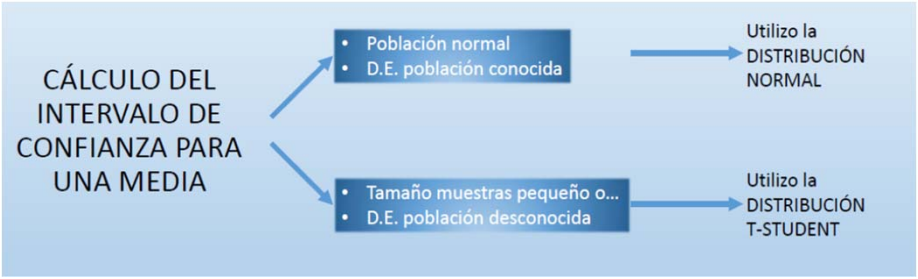
4. Intervalo de confianza

Intervalo de confianza

El valor 1-α es la probabilidad que tiene el intervalo de contener a la media poblacional (μ) antes de extraer la muestra.

Si aceptamos un riesgo α del 5%, podremos afirmar que de cada 100 muestras extraídas, en promedio, 95 intervalos contendrán el valor de la media (μ)

4. Intervalo de confianza



4. Intervalo de confianza

Cálculo del intervalo de confianza para una media

1) Si conocemos la desviación estándar de la población o nuestra muestra es grande podremos calcular el intervalo apoyándonos en las tablas de la **distribución normal estandarizada (tipificada)**

$\alpha/2$ $(1-\alpha)$	0,25	0,125	0,10	0,05	0,025	0,01	0,005	0,001	0,0005
$Z_{\alpha/2}$	0,674	1,150	1,282	1,645	**1,96**	2,326	2,576	3,090	3,291

4. Intervalo de confianza

Cálculo del intervalo de confianza para una media

El enunciado me dice que los datos de media y D.E. se refieren a una población que sigue una ley normal, luego puedo usar la **tabla normal**

Suponemos que la distribución de pesos al nacer de los niños de una población sigue una ley Normal de **media 3,0 kg y D.E.= 0,15 kg.** ¿Cuál es el intervalo con probabilidad de 0,95 de contener la media de los pesos de 100 niños extraídos al azar de dicha población?

4. Intervalo de confianza

- Población normal
- D.E. población conocida

Cálculo del intervalo de confianza para una media

Tabla normal

$1-\alpha = 0,95 \rightarrow \alpha = 0,05 \rightarrow \alpha/2 = 0,025 \rightarrow Z_{\alpha/2} = 1,96$

Error estándar: $SE = \frac{\sigma}{\sqrt{n}} = \frac{0,15}{\sqrt{100}} = \frac{0,15}{10} = 0,015$

Desvío ε: $\varepsilon = Z_{\alpha/2} \times SE = 1,96 \times 0,015 = 0,0294$

IP 95% de μ: $\mu \pm \varepsilon \rightarrow$ 3,0 kg $\pm$ 0,0294 $\rightarrow$ 2,97 a 3,03 kg

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza

- Población normal
- D.E. población conocida

Cálculo del intervalo de confianza para una media

IP 95% de μ: $\mu \pm \varepsilon \rightarrow 3{,}0 \text{ kg} \pm 0{,}0294 \rightarrow 2{,}97 \text{ a } 3{,}03 \text{ kg}$

Conclusión: El intervalo de probabilidad 2,97 a 3,03 kg tiene una probabilidad del 95% de contener las medias observadas de muestras de tamaño n=100 extraídas al azar de una población con media = 3,0 kg y una desviación típica de 0,15 kg.

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza

- Tamaño muestras pequeño o...
- D.E. población desconocida

Cálculo del intervalo de confianza para una media

Se debe recurrir a las tablas de una nueva distribución que también se conoce muy bien y también está tabulada: **distribución de Student-Fisher** que depende del tamaño de la muestra y de la desviación estándar muestral (S)

Con tabla normal — Con t-Student

Desviación estándar de la población → $\varepsilon = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \rightarrow e = t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}$ ← Desviación estándar de la muestra

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza

- Tamaño muestras pequeño o...
- D.E. población desconocida

Cálculo del intervalo de confianza para una media

El enunciado me dice que los datos (media y desviación estándar S) se refieren a una muestra. Por tanto, no tengo los datos de la población y tengo que usar la t-Student

La distribución de pesos de una muestra aleatoria de 10 neonatos presenta una media de 3,2 kg y una D.E. de 0,834 kg. Estimar, con una confianza del 95%, el peso medio de esta población.

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza

- Tamaño muestras pequeño o...
- D.E. población desconocida

Cálculo del intervalo de confianza para una media

$1-\alpha = 0,95 \rightarrow \alpha = 0,05 \rightarrow \alpha/2 = 0,025 \rightarrow t_{n-1;\alpha/2} = 2,262$

Error estándar: $SE = \frac{S}{\sqrt{n}} = \frac{0,834}{\sqrt{10}} = 0,2637$

Desvío ε: $\varepsilon = t_{n-1;\alpha/2} \times SE = 2,262 \times 0,2637 = 0,5966$

IP 95% de μ: $\mu \pm \varepsilon \rightarrow 3,2 \text{ kg} \pm 0,5966 \rightarrow 2,6 \text{ a } 3,8 \text{ kg}$

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

4. Intervalo de confianza

- Tamaño muestras pequeño o...
- D.E. población desconocida

Cálculo del intervalo de confianza para una media

IP 95% de μ: $\mu \pm \varepsilon \rightarrow 3{,}2 \text{ kg} \pm 0{,}5966 \rightarrow 2{,}6 \text{ a } 3{,}8 \text{ kg}$

Conclusión: El peso medio de los niños al nacer se encuentra situado entre 2,6 y 3,8 kg con una confianza del 95%

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

5. Contraste de Hipótesis

Los contrastes de hipótesis proveen de una regla de decisión para rechazar o no una hipótesis sobre X

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

5. Contraste de Hipótesis

No son una demostración irrefutable ...

... solo establecen si la hipótesis es plausible vistos los datos de la muestra.

Pueden llevar a la decisión equivocada y esa posibilidad deberá cuantificarse.

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

5. Contraste de Hipótesis

En este tema se revisarán contrastes PARAMÉTRICOS ...

... y también NO PARAMÉTRICOS.

5. Contraste de Hipótesis

En ambos casos, un test da reglas de decisión para rechazar o no esa hipótesis

y en el caso paramétrico lo hace siempre bajo el supuesto paramétrico

5. Contraste de Hipótesis

Una vez planteada una hipótesis

Hipótesis H_0 ($\theta \in \Theta_0 \subseteq \Theta$ en el caso paramétrico)

automáticamente surge su contraria

Hipótesis H_1 ($\theta \in \Theta_1 = \Theta - \Theta_0$ en el caso paramétrico)

Se las llama, respectivamente, hipótesis nula e hipótesis alternativa
y sus papeles no son simétricos ni intercambiables

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

5. Contraste de Hipótesis

Un test aporta una regla para decidir si rechazar H_0 (aceptar H_1) o no, basándose en la observación de una muestra de X.

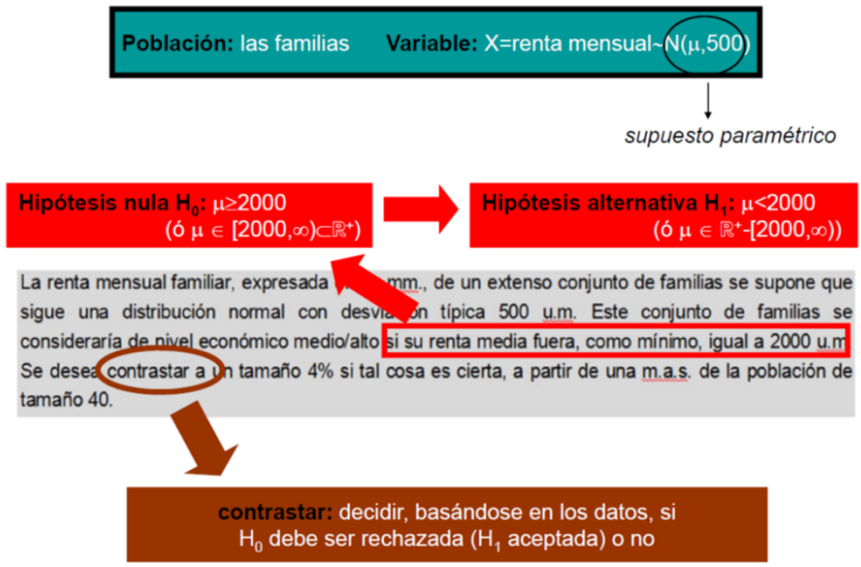
UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

5. Contraste de Hipótesis

EJEMPLO:

La renta mensual familiar, expresada en uu.mm., de un extenso conjunto de familias se supone que sigue una distribución normal con desviación típica 500 u.m. Este conjunto de familias se consideraría de nivel económico medio/alto si su renta media fuera, como mínimo, igual a 2000 u.m. Se desea contrastar a un tamaño 4% si tal cosa es cierta, a partir de una m.a.s. de la población de tamaño 40.

5. Contraste de Hipótesis



5. Contraste de Hipótesis

La regla para rechazar con un test se basa en un estadístico muestral $T(X_1,...,X_n)$ *1:

regla (o criterio) de rechazo: rechazar H_0 si la realización muestral de $X_1,...,X_n$ es tal que el valor muestral de $T(X_1,...,X_n)$ cumple cierta condición*2

*1 T independiente de parámetros desconocidos

*2 Condición que suele adoptar la forma: T>c, T<c, c<T<d ...
Las constantes c,d... se denominan valores críticos.

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

5. Contraste de Hipótesis

Cuestiones a resolver:

¿cómo encontrar T (estadístico del contraste)?

¿cómo establecer la condición en la que se basa la regla de rechazo?

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA

5. Contraste de Hipótesis

Algunos contrastes para una población normal

Contrastes de tamaño α obtenidos a partir de una m.a.s. de tamaño n de una población $N(\mu,\sigma)$

Estadístico	H_0	H_1	Criterio de rechazo de H_0
Contrastes para μ si σ es conocida			
$Z=\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$	$\mu\leq\mu_0$	$\mu>\mu_0$	$Z>z_\alpha$
	$\mu\geq\mu_0$	$\mu<\mu_0$	$Z<-z_\alpha$
$\sim N(0,1)$ si $\mu=\mu_0$	$\mu=\mu_0$	$\mu\neq\mu_0$	$Z>z_{\alpha/2}$ ó $Z<-z_{\alpha/2}$
Contrastes para μ si σ es desconocida			
$T=\frac{\overline{X}-\mu_0}{S/\sqrt{n-1}}$	$\mu\leq\mu_0$	$\mu>\mu_0$	$T>t_{n-1,\alpha}$
	$\mu\geq\mu_0$	$\mu<\mu_0$	$T<-t_{n-1,\alpha}$
$\sim t_{n-1}$ si $\mu=\mu_0$	$\mu=\mu_0$	$\mu\neq\mu_0$	$T>t_{n-1,\alpha/2}$ ó $T<-t_{n-1,\alpha/2}$
Contrastes para σ si μ es conocida			
$\chi^2=\frac{\sum(X_i-\mu)^2}{\sigma_0^2}$	$\sigma\leq\sigma_0$	$\sigma>\sigma_0$	$\chi^2>\chi^2_{n,\alpha}$
	$\sigma\geq\sigma_0$	$\sigma<\sigma_0$	$\chi^2<\chi^2_{n,1-\alpha}$
$\sim\chi^2_n$ si $\sigma=\sigma_0$	$\sigma=\sigma_0$	$\sigma\neq\sigma_0$	$\chi^2>\chi^2_{n,\alpha/2}$ ó $\chi^2<\chi^2_{n,1-\alpha/2}$
Contrastes para σ si μ es desconocida			
$\chi^2=\frac{nS^2}{\sigma_0^2}$	$\sigma\leq\sigma_0$	$\sigma>\sigma_0$	$\chi^2>\chi^2_{n-1,\alpha}$
	$\sigma\geq\sigma_0$	$\sigma<\sigma_0$	$\chi^2<\chi^2_{n-1,1-\alpha}$
$\sim\chi^2_{n-1}$ si $\sigma=\sigma_0$	$\sigma=\sigma_0$	$\sigma\neq\sigma_0$	$\chi^2>\chi^2_{n-1,\alpha/2}$ ó $\chi^2<\chi^2_{n-1,1-\alpha/2}$

La renta mensual familiar, expresada en uu.mm., de un extenso conjunto de familias se supone que sigue una distribución normal con desviación típica 500 u.m. Este conjunto de familias se consideraría de nivel económico medio/alto si su renta media fuera, como mínimo, igual a 2000 u.m. Se desea contrastar [Algunos contrastes para una población normal] de la población de tamaño 40.

Contrastes de tamaño α obtenidos a partir de una m.a.s. de tamaño n de una población $N(\mu,\sigma)$

Estadístico	H_0	H_1	Criterio de rechazo de H_0
Contrastes para μ si σ es conocida			
$Z=\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$	$\mu\leq\mu_0$	$\mu>\mu_0$	$Z>z_\alpha$
	$\mu\geq\mu_0$	$\mu<\mu_0$	$Z<-z_\alpha$
$\sim N(0,1)$ si $\mu=\mu_0$	$\mu=\mu_0$	$\mu\neq\mu_0$	$Z>z_{\alpha/2}$ ó $Z<-z_{\alpha/2}$
Contrastes para μ si σ es desconocida			
$T=\frac{\overline{X}-\mu_0}{S/\sqrt{n-1}}$	$\mu\leq\mu_0$	$\mu>\mu_0$	$T>t_{n-1,\alpha}$
	$\mu\geq\mu_0$	$\mu<\mu_0$	$T<$
$\sim t_{n-1}$ si $\mu=\mu_0$	$\mu=\mu_0$	$\mu\neq\mu_0$	$T>t_{n-1,\alpha/}$
Contrastes para σ si μ es conoc			
$\chi^2=\frac{\sum(X_i-\mu)^2}{\sigma_0^2}$	$\sigma\leq\sigma_0$	$\sigma>\sigma_0$	$\chi^2>$
	$\sigma\geq\sigma_0$	$\sigma<\sigma_0$	$\chi^2<$
$\sim\chi^2_n$ si $\sigma=\sigma_0$	$\sigma=\sigma_0$	$\sigma\neq\sigma_0$	$\chi^2>\chi^2_{n,\alpha/}$
Contrastes para σ si μ es desconocida			
$\chi^2=\frac{nS^2}{\sigma_0^2}$	$\sigma\leq\sigma_0$	$\sigma>\sigma_0$	$\chi^2>\chi^2_{n-1,\alpha}$
	$\sigma\geq\sigma_0$	$\sigma<\sigma_0$	$\chi^2<\chi^2_{n-1,1-\alpha}$
$\sim\chi^2_{n-1}$ si $\sigma=\sigma_0$	$\sigma=\sigma_0$	$\sigma\neq\sigma_0$	$\chi^2>\chi^2_{n-1,\alpha/2}$ ó $\chi^2<\chi^2_{n-1,1-\alpha/2}$

El estadístico T de este contraste es este Z

Esta es la regla de rechazo: se tomará una muestra y si la realización muestral de Z cumple esa desigualdad, entonces se rechazará H_0

La renta mensual familiar, expresada en uu.mm., de un extenso conjunto de familias se supone que sigue una distribución normal con desviación típica 500 u.m. Este conjunto de familias se consideraría de nivel económico medio/alto si su renta media fuera, como mínimo, igual a 2000 u.m. Se desea contrastar a un tamaño 4% si tal cosa es cierta, a partir de una m.a.s. de la población de tamaño 40.

$Z=\frac{X-\mu_0}{\sigma/\sqrt{n}}$	$\mu\leq\mu_0$	$\mu>\mu_0$	$Z>z_\alpha$
	$\mu\geq\mu_0$	$\mu<\mu_0$	$Z<-z_\alpha$
	$\mu=\mu_0$	$\mu\neq\mu_0$	

Estadístico del contraste

Condición de rechazo

Explicación de los algunos de esos símbolos:

- μ_0=2000 (mínimo valor admisible para μ en H_0)
- σ=500 (valor conocido de la desviación típica)
- n=40 (tamaño de muestra)
- α%=4% (su significado está por explicar)
- $-z_\alpha = -z_{4\%} = -1{,}75$ (valor crítico = valor mínimo para Z a partir del cuál ya no se rechazaría H_0)

La renta mensual familiar, expresada en uu.mm., de un extenso conjunto de familias se supone que sigue una distribución normal con desviación típica 500 u.m. Este conjunto de familias se consideraría de nivel económico medio/alto si su renta media fuera, como mínimo, igual a 2000 u.m. Se desea contrastar a un tamaño 4% si tal cosa es cierta, a partir de una m.a.s. de la población de tamaño 40.

$Z=\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$			
	$\mu\leq\mu_0$	$\mu>\mu_0$	$Z>z_\alpha$
	$\mu\geq\mu_0$	$\mu<\mu_0$	$Z<-z_\alpha$
$\sim N(0,1)$ si $\mu=\mu_0$	$\mu=\mu_0$	$\mu\neq\mu_0$	$Z>z_{\alpha/2}$ ó $Z<-z_{\alpha/2}$

Información adicional de interés: indica que Z~N(0,1) caso de que $\mu=\mu_0=2000$ que es un posible valor para μ dentro de H_0.

¡Ojo!, esa ya no sería la distribución de Z si la nula fuera cierta pero por tomar μ otro valor distinto de μ_0 (p.e. 2050, 3100,….).

La renta mensual familiar, expresada en uu.mm., de un extenso conjunto de familias se supone que sigue una distribución normal con desviación típica 500 u.m. Este conjunto de familias se consideraría de nivel económico medio/alto si su renta media fuera, como mínimo, igual a 2000 u.m. Se desea contrastar a un tamaño 4% si tal cosa es cierta, a partir de una m.a.s. de la población de tamaño 40.

Finalmente tenemos el siguiente

criterio de rechazo (c.r.) de H_0:

$$Z=\frac{(\sum X_i/40)-2000}{500/\sqrt{40}}<-1{,}75$$

que se usa de la siguiente forma:

Cuando se midan las cuarenta rentas $(x_1,\ldots,x_{40})$ para formar la correspondiente realización muestral, si cumplen $\frac{(\sum x_i/40)-2000}{500/\sqrt{40}}<-1{,}75$ entonces rechazaremos H_0.

Abreviadamente:

RH_0 si $X_1,\cdots,X_{40}$ cumple c.r.

suceso aleatorio cuyo cumplimiento equivale a rechazar H_0

Una regla así conlleva posibilidad de tomar decisiones erróneas. Veámoslo en el caso paramétrico:

Si el verdadero $\theta \in \Theta_0$ (H_0 cierta) y si $P_\theta(X_1,...,X_n$ cumpla c.r.)$\neq 0$, a veces ocurrirá que rechazaremos H_0 y nos equivocaremos.

Si el verdadero $\theta \in \Theta_1$ (H_1 cierta) y si $P_\theta(X_1,...,X_n$ no cumpla c.r.)$\neq 0$, a veces ocurrirá que no rechazaremos H_0 y nos equivocaremos.

Esas son las dos decisiones de tipo erróneo posibles (incluso en el caso no paramétrico).

La renta mensual familiar, expresada en uu.mm., de un extenso conjunto de familias se supone que sigue una distribución normal con desviación típica 500 u.m. Este conjunto de familias se consideraría de nivel económico medio/alto si su renta media fuera, como mínimo, igual a 2000 u.m. Se desea contrastar a un tamaño 4% si tal cosa es cierta, a partir de una m.a.s. de la población de tamaño 40.

Rechazo H_0 si $Z = \dfrac{(\sum X_i / 40) - 2000}{500/\sqrt{40}} < -1{,}75$

Algunos ejemplos de decisiones erróneas que es posible cometer con este test:

Si H_0 fuera cierta porque, por ejemplo, μ=2025,

$P_{\mu=2025}$(rechazar H_0) = $P_{\mu=2025}(X_1,..,X_{40}$ cumpla c.r.) = $P_{\mu=2025}$(Z<-1,75) =

= $P_{\mu=2025}$(N((2025-2000)/(500/40 ½);1)<-1,75) = 1,94% ≠ 0%

y por ello, en este caso, con ese 1,94% de muestras que cumple

la condición Z<-1,75, rechazaremos H_0 y nos equivocaremos

La renta mensual familiar, expresada en uu.mm., de un extenso conjunto de familias se supone que sigue una distribución normal con desviación típica 500 u.m. Este conjunto de familias se consideraría de nivel económico medio/alto si su renta media fuera, como mínimo, igual a 2000 u.m. Se desea contrastar a un tamaño 4% si tal cosa es cierta, a partir de una m.a.s. de la población de tamaño 40.

Rechazo H_0 si $Z = \dfrac{(\sum X_i / 40) - 2000}{500/\sqrt{40}} < -1{,}75$

Algunos ejemplos de decisiones erróneas que es posible cometer con este test:

Si H_1 fuera cierta porque, por ejemplo, μ=1800,

$P_{\mu=1800}$(no rechazar H_0) = $P_{\mu=1800}$(X_1,..,X_{40} no cumpla c.r.) = $P_{\mu=1800}$(Z>-1,75) =

= $P_{\mu=1800}$(N((1800-2000)/(500/40 ½);1)>-1,75) = 21,77% ≠ 0%

y por ello, en este caso, con ese 21,77% de muestras que cumple

la condición Z>-1,75, no rechazaremos H_0 y nos equivocaremos

5. Contraste de Hipótesis

De forma resumida:

Estado de la naturaleza ⇒ / ⇓ Decisión tomada	H_0 cierta	H_1 cierta
No rechazar H_0	Decisión correcta	Error (de tipo II)
Rechazar H_0	Error (de tipo I)	Decisión correcta