



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Departamento de Estadística e
Investigación Operativa Aplicadas
y Calidad

Grado en Ingeniería Informática

 etsinf

UD 2 – Parte 1

Estadística Descriptiva
Unidimensional



Nota importante



- Estas transparencias son un extracto de la Unidad Didáctica 2 publicada en PoliformaT, y a los capítulos 1, 2 y 7 (apartado 7.7) y 11 (apartado 11.1) del libro Métodos Estadísticos para Ingenieros y sirven sólo como material complementario.
- Esta presentación se utilizará para exponer los contenidos básicos de la UD2 durante la primera semana de clase:
 - Estadística descriptiva unidimensional
 - Estadística descriptiva bidimensional: sólo tablas de frecuencias cruzadas. El resto se verá más adelante.
- Además, los contenidos de esta UD se trabajarán en la primera y segunda Práctica Informática que se llevará a cabo en el laboratorio B del DEIOAC.



Contenido

- 1. Introducción. Conceptos básicos (población, Muestra, v.a. ...)**
- 2. Estadística Descriptiva Unidimensional**
- 3. Estadística Descriptiva Bidimensional**

Resumen

Mapa mental

Glosario

Ejercicios

Práctica *Statgraphics*



Contenido

- **Introducción**
 - Definición, objetivos, aplicaciones,...
- **Conceptos básicos**
 - Población y muestra
 - Tipos
 - Estadística descriptiva e Inferencia
 - Característica aleatoria y variable aleatoria
 - Tipos
 - Dimensiones
 - Datos
- **Estadística descriptiva unidimensional**
 - Tablas de Frecuencias
 - Gráficos: Diagramas de Barras, de Sectores, Histogramas, Box & Whisker....
 - Parámetros:
 - Posición
 - Dispersión
 - Forma:
 - Asimetría
 - Curtosis
- **Estadística Descriptiva Bidimensional**
 - Tablas de frecuencias cruzadas.
 - Distribuciones marginales y condicionales
 - Diagramas de dispersión
 - Covarianza y Coeficiente de Correlación



ystadegau
estatistikak
estadística
statistiques
статистикών
إحصائيات
statistică
statistikë
आँकडे
statistik
statistiek
statistieke

amar
статыстыка
statystyka
統計 statistica
পরিসংখ্যন
statistikk
statistika
takwimu
estatística
statistiko
統計
statistikat
statistic
statistika

istatistika
istatisticí
istatistika
istatistická
istatistiké
staitistiké
statistiek
statistieke

statistics

Conceptos básicos



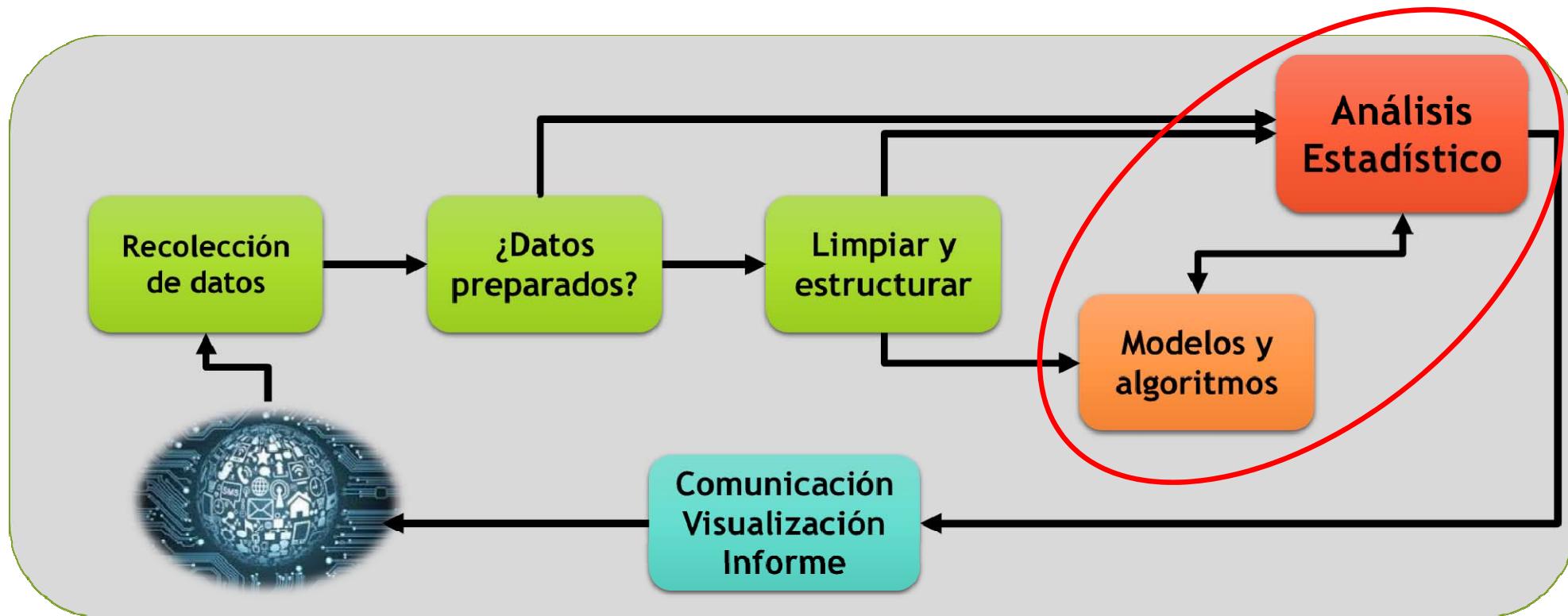
Estadística: una definición

"LA ESTADÍSTICA ES LA CIENCIA CUYO OBJETO ES LA OBTENCIÓN Y EL ANÁLISIS DE DATOS MEDIANTE EL RECURSO A MODELOS MATEMÁTICOS Y A HERRAMIENTAS INFORMÁTICAS"

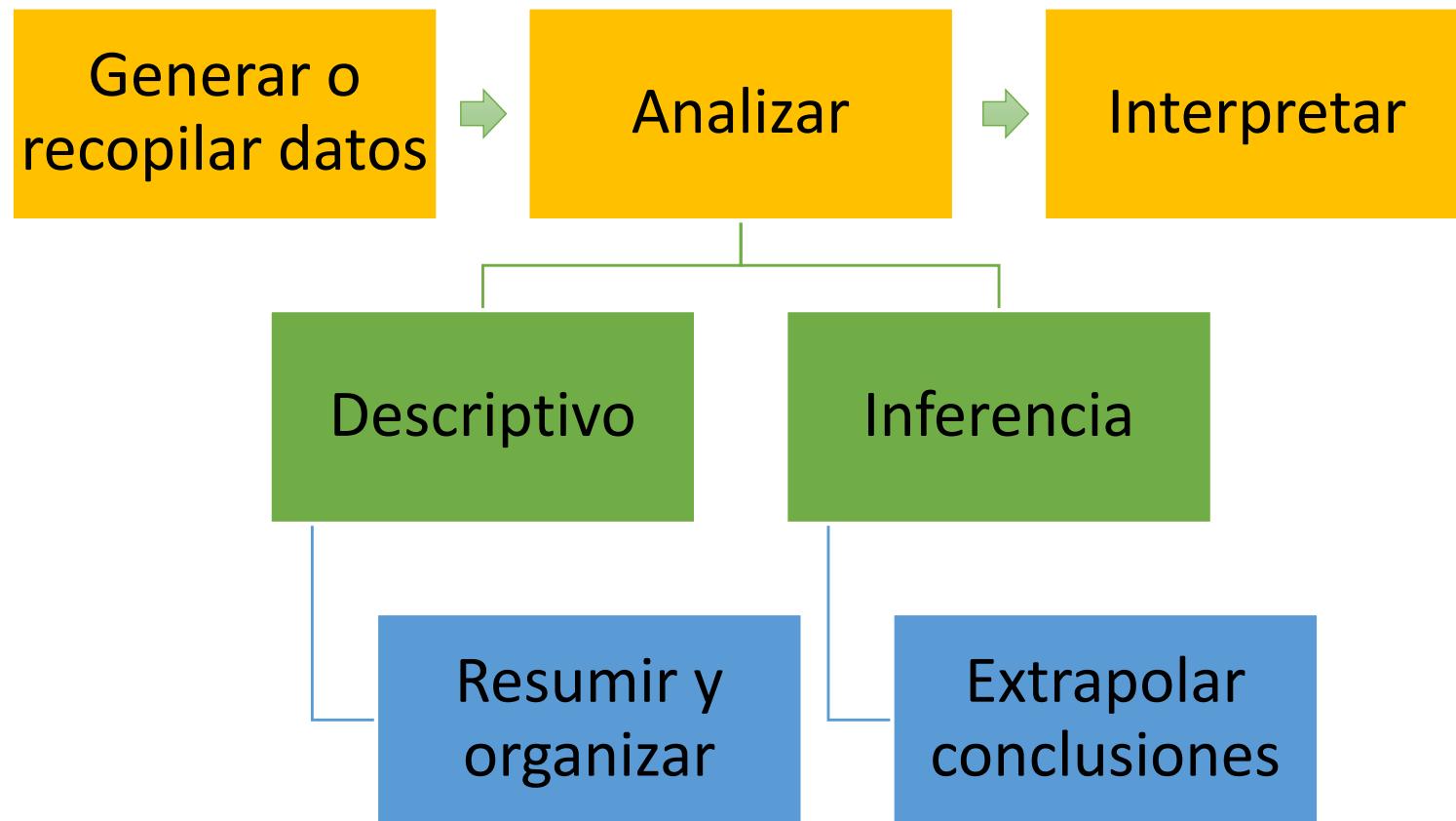
- DATOS
 - HERRAMIENTAS INFORMÁTICAS
 - MODELOS MATEMÁTICOS



Proceso de un estudio estadístico



Objetivos Estadística

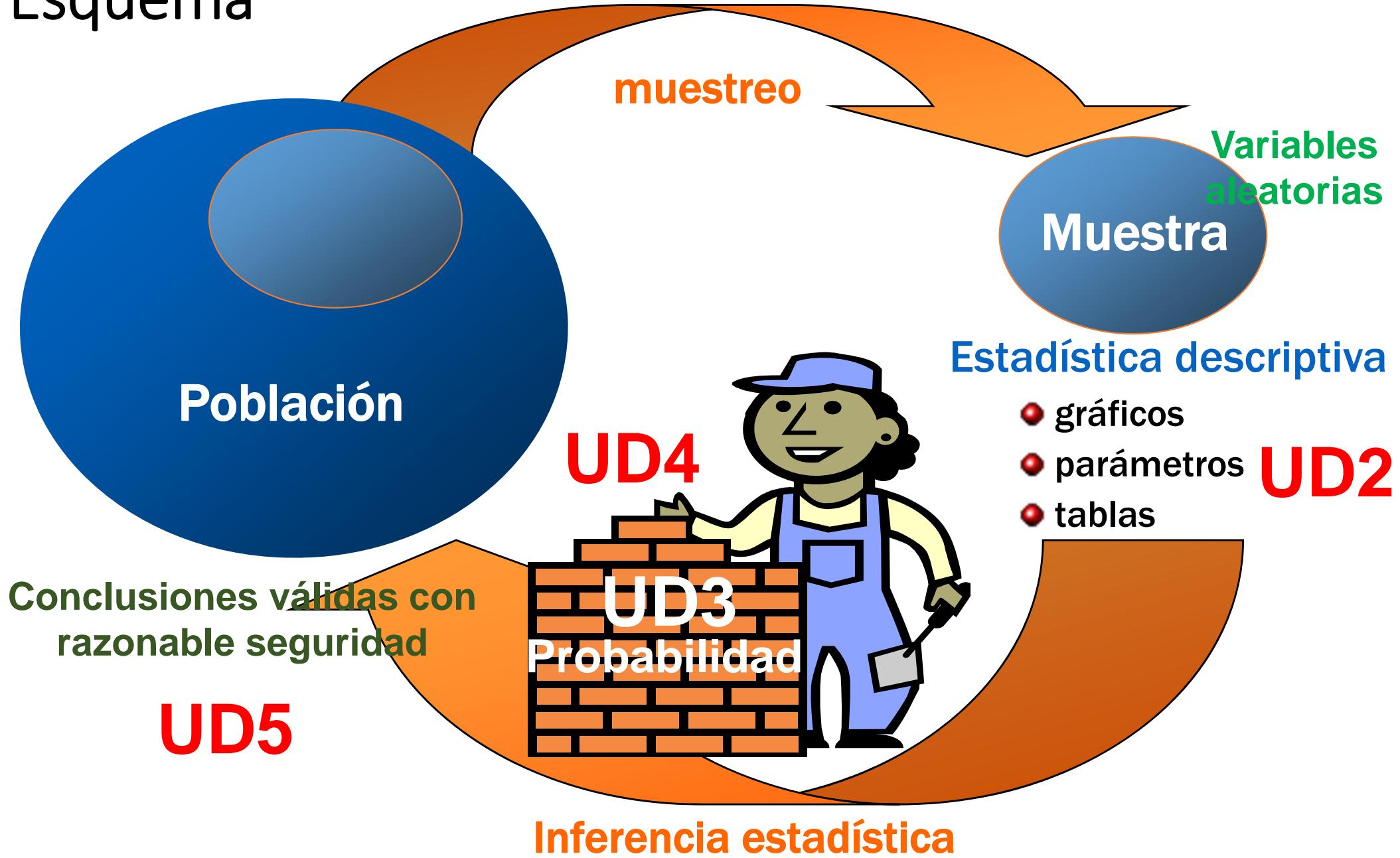


Algunas aplicaciones de la Estadística

- Control de Calidad
- Evaluación de prestaciones
- Robótica
- Comparación de algoritmos
- Fiabilidad
- Big Data
- e.t.c.



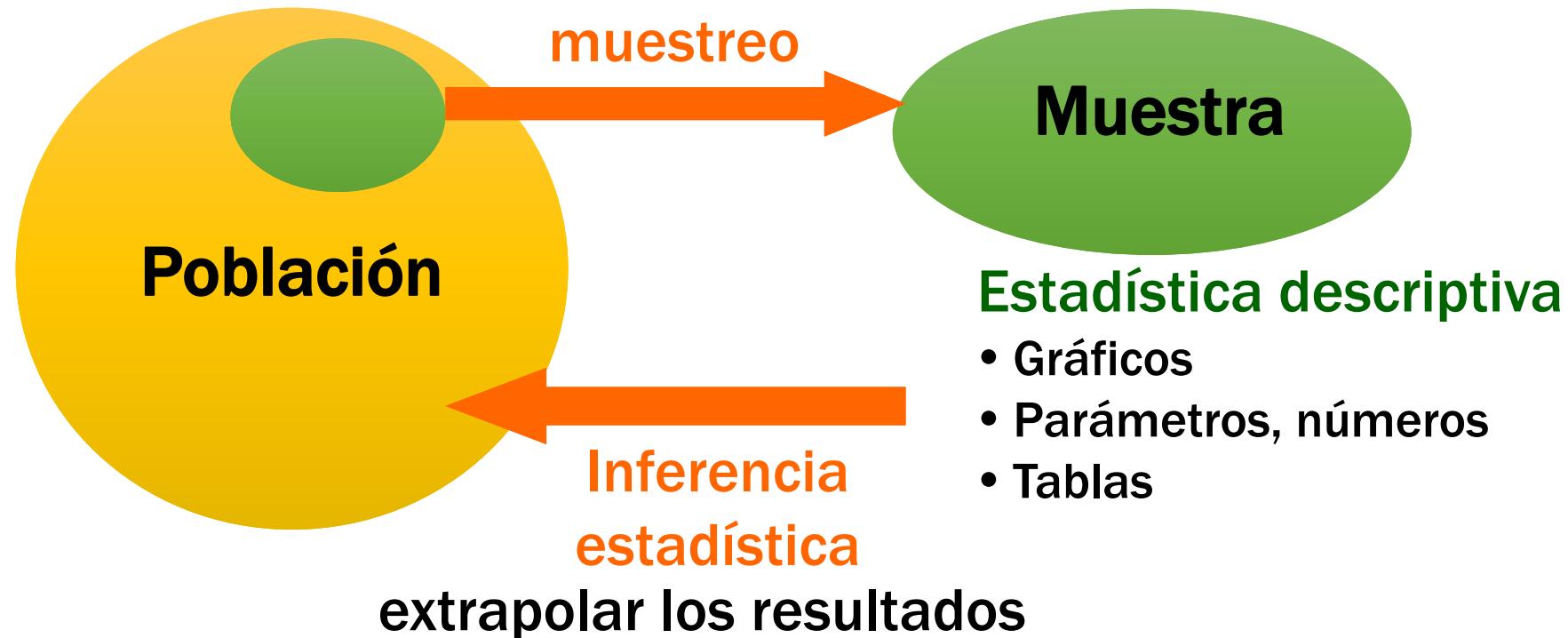
Esquema



Conceptos básicos



Población y muestra



Análisis Descriptivo datos



Primer paso de cualquier análisis

- Tratamiento descriptivo o exploratorio
- Poner de manifiesto características y regularidades de los datos:
 1. Sintetizarlas en parámetros
 2. Representaciones gráficas adecuadas

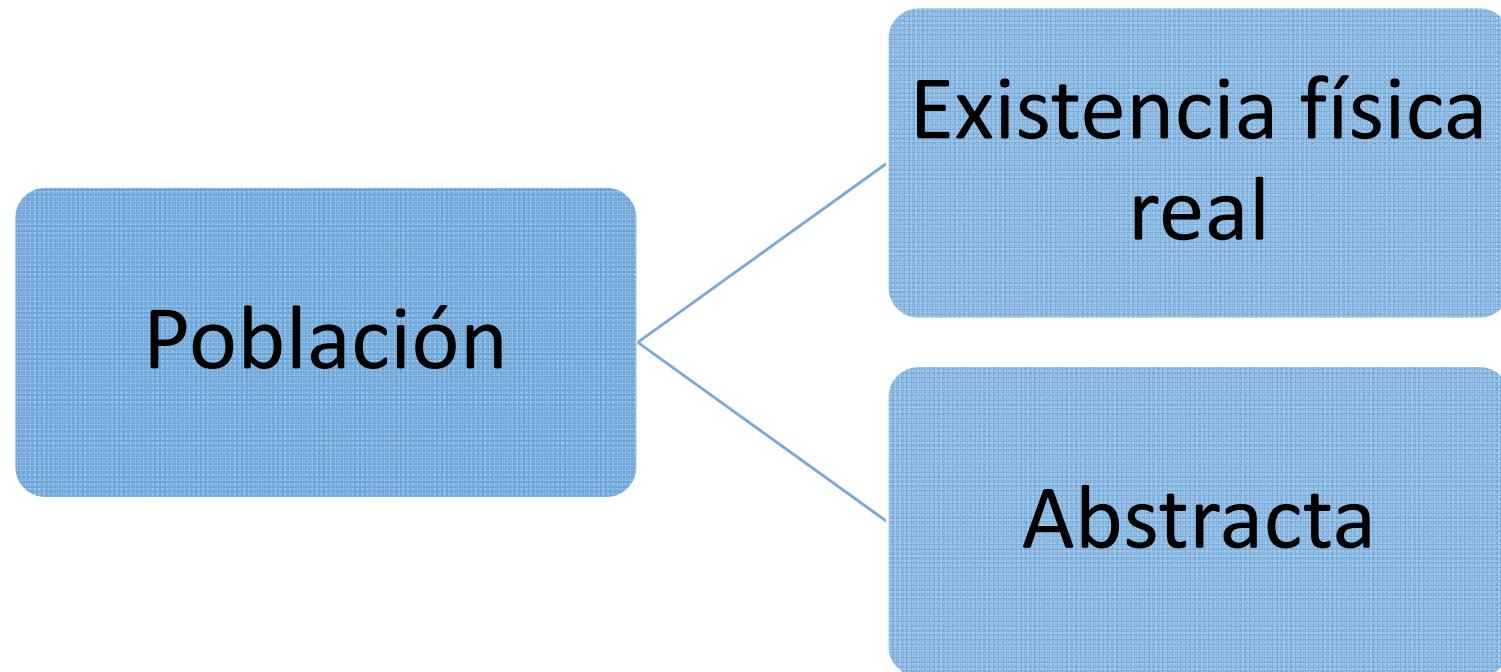
**!!No se extrapolan todavía conclusiones de
los datos a la población!!**





Poblaciones. Concepto

conjunto de todos los individuos o entes que constituyen el objeto de un determinado estudio y sobre los que se desea obtener ciertas conclusiones



Poblaciones. Tipos y Ejemplos

- Existencia física real previa a la realización del estudio:
 - Estudio intención de voto: todo español con derecho a voto
 - Estudio rendimiento en universidades españolas: todos los alumnos de la ETSINF
 - Control de calidad componentes informáticos: Todos los ventiladores de una partida
 -
- Número finito, aunque posiblemente muy elevado, de individuos.
- No es lo más frecuente





Poblaciones. Tipos y Ejemplos

- De carácter abstracto

- Estudio dado trucado o no: todos los lanzamiento de un dado
- Estudio sobre la eficiencia de diversos algoritmos de encaminamiento de mensajes entre nodos en una red de procesadores: todos los mensajes que se generan en la red de procesadores
-

Los individuos se van generando mediante un proceso →

Experimento aleatorio





VARIABLES ALEATORIAS

¡Toda población tiene **VARIABILIDAD** en sus características!

- La vida útil de varios componentes electrónicos idénticos no es la misma.
- El número de asignaturas en las que se matricula un alumno también varía
- El número de puntos que sale al lanzar un dado puede ser distinto según la tirada.
- El throughput de un sistema cambia según tecnología
- El retardo de los mensajes es diferente según la topología de la red...

Característica Aleatoria: cualquiera que puede constatarse en cada individuo de la población



Características aleatorias

- **Tipos**

- **Expresables numéricamente:**

- Tiempo hasta el fallo de un tipo de monitor
 - Número obtenido al lanzar un dado
 - Número de asignaturas matriculadas
 - Retardo de un mensaje
 - Velocidad de E de un HDD

- **De tipo cualitativo:**

- Partido votado
 - Aprobada una determinada asignatura (s/n)
 - Destino correcto del mensaje (s/n)
 - Categoría laboral



Características aleatorias



Variables

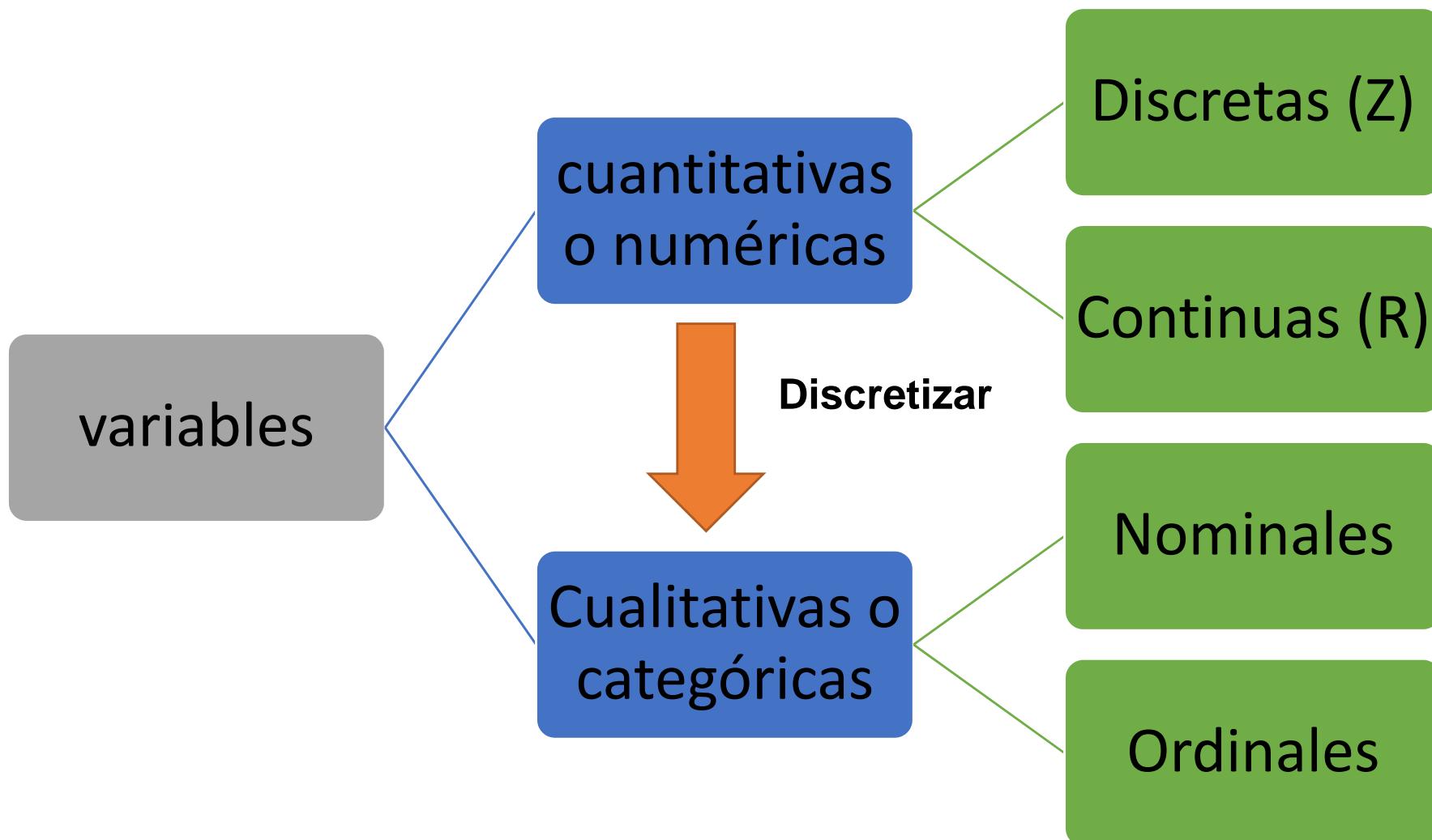
característica aleatoria
numérica o cualitativa
codificada

#	mpg	cylinders	displace	horsepower	accel	year	weight	origin	name
1	43.1	4	90	68	21.5	70	1985	2	Volkswagen Rabbit
2	36.1	4	90	66	14.4	70	1980	1	Ford Fiesta
3	32.8	4	78	52	19.4	70	1985	3	Mazda GLC D
4	39.4	4	85	70	18.6	70	2070	3	Datsun 210
5	36.1	4	91	60	16.4	70	1980	3	Honda Civic
6	19.9	8	260	110	15.5	70	3345	1	Oldsmobile Cutlass
7	19.4	8	318	140	13.2	70	3735	1	Dodge Diplomat
8	20.2	8	302	139	12.8	70	3570	1	Mercury Cougar
9	19.7	6	231	105	19.2	70	3535	1	Pontiac Phoenix
10	20.5	6	208	95	18.2	70	3355	1	Chevrolet Malibu
11	20.2	6	208	85	15.8	70	2945	1	Ford Fairlane
12	25.1	4	140	88	15.4	70	2720	1	Ford Fairlane
13	20.5	6	225	108	17.2	70	3430	1	Plymouth Volare
14	19.4	6	232	98	17.2	70	3210	1	AMC Concorde
15	20.6	6	231	105	15.8	70	3380	1	Buick Century
16	20.0	6	200	85	16.7	70	3070	1	Mercury Zephyr
17	18.6	6	225	110	16.7	70	3620	1	Dodge Aspen
18	18.1	6	258	120	15.1	70	3410	1	AMC Concorde
19	19.2	8	305	145	13.2	70	3425	1	Chevrolet Monte Carlo
20	17.7	6	231	165	13.4	70	3445	1	Buick Regal
21	18.1	8	302	139	11.2	70	3285	1	Ford Futura
22	17.5	8	318	140	11.7	70	4080	1	Dodge Monaco
23	18	4	98	68	16.5	70	2155	1	Chevrolet Chevelle
24	27.5	4	134	95	14.2	70	2560	1	Toyota Corona
25	27.2	4	119	97	14.7	70	2380	1	Datsun 510
26	18.9	4	105	75	14.5	70	2330	1	Dodge Omni
27	23.1	4	134	95	14.8	70	2515	1	Toyota Celica
28	23.2	4	156	105	16.7	70	2745	1	Plymouth Sappo
29	23.8	4	151	85	17.6	70	2855	1	Oldsmobile Starfire
30	23.9	4	119	97	14.9	70	2485	1	Datsun 280-Z
31	28.1	5	131	101	15.9	70	2810	9	Audi 5000

No tiene ningún sentido realizar operaciones matemáticas con variables aleatorias que son codificaciones de una característica aleatoria de tipo cualitativo



Variables aleatorias



Ejemplos de variables y tipos

SEXO	EDAD	MES	ESTATURA	PESO
varón	20	Enero	18,32	76
varón	21	Junio	18,54	72
varón	22	Octubre	16,57	75
varón	22	Abril	17,45	70
varón	22	Julio	17,58	
mujer	20	Noviembre	17,59	70
varón	22	Julio	17,43	65
mujer	23	Octubre	15,91	54

Dato faltante

SEXO: categórica nominal

MES: categórica ordinal

EDAD: cuantitativa discreta

ESTATURA: cuantitativa continua



Ejemplos de variables y tipos

1. Provincia con mayor índice de paro → cuantitativa y discreta
2. Temperatura de un frigorífico → cuantitativa y continua
3. Número de anuncios emitidos por una cadena de TV en un intermedio publicitario → cuantitativa y discreta
4. Tiempo necesario para fabricar una pieza → cuantitativa y continua
5. Peso neto de una botella estándar de aceite → cuantitativa y continua
6. Nivel cultural dominante entre los lectores de una revista
7. Pulsaciones por minuto de una operadora de datos → cuantitativa y discreta
8. Categoría de un hotel → atributo



Variables aleatorias

- Importante diferencia entre:
 - una variable aleatoria **K-dimensional**, en la que las K variables se miden sobre los individuos de una única población
 - y un conjunto de **K variables aleatorias unidimensionales**, definidas sobre K poblaciones distintas.



Variables aleatorias

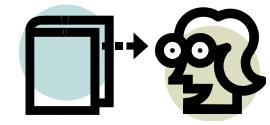
- **Ejemplo 1:**

- **Población:** Todos los archivos “mp3” que se pueden bajar de una web
- **Variable aleatoria de dimensión 4:**
(**tiempo_bajada, tamaño_archivo, nº_fuentes, nombre_tema**)

Las componentes cualitativas se suelen usar para subdividir la Población en subpoblaciones



Variables aleatorias



- **Ejemplo 2:**

¿Variable bidimensional o 2 var. Unidimensionales?

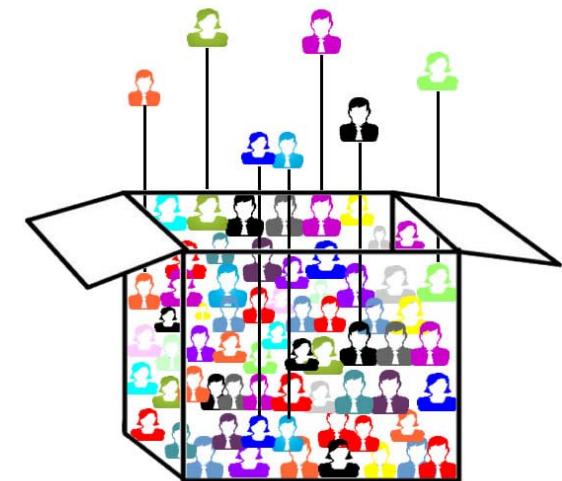
- Rendimiento de dos modelos de portátil
 - 2 var. Unidimensionales
- Rendimiento y modelo de un portátil
 - Variable bidimensional





Muestras. Datos estadísticos

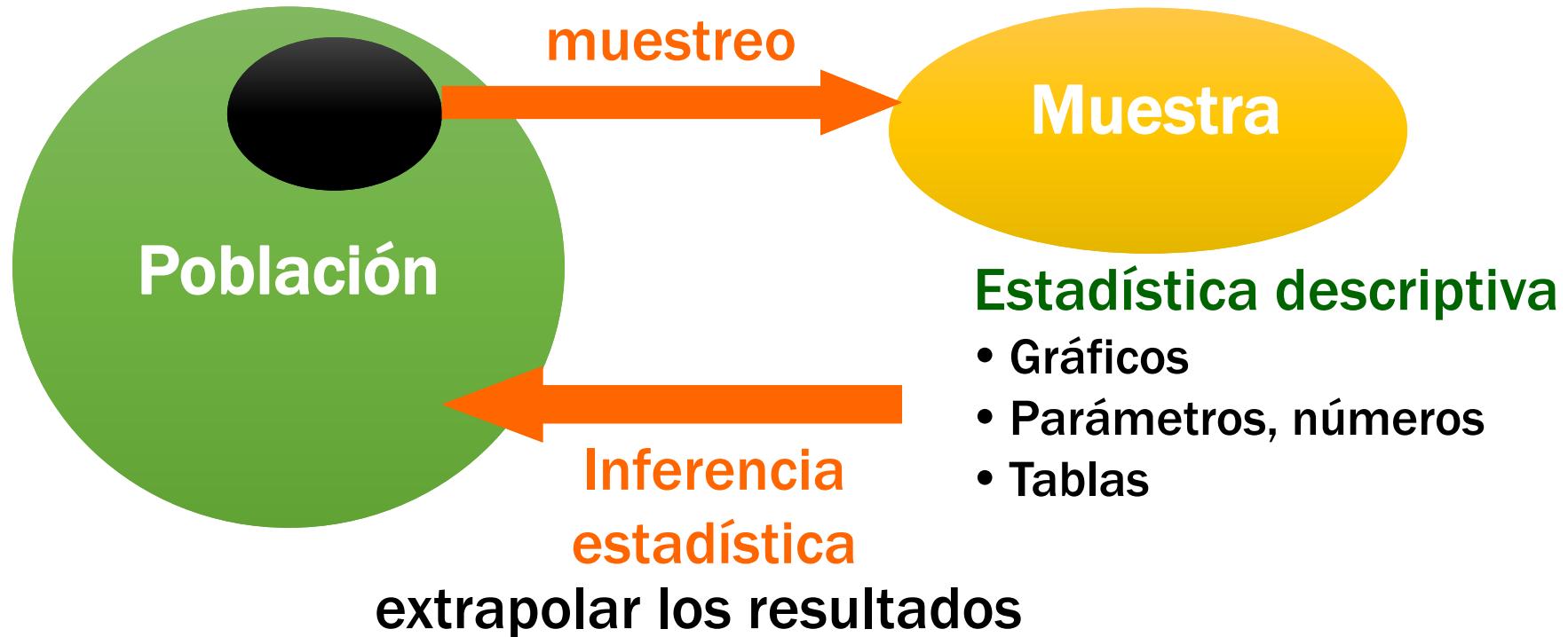
- Normalmente no se trabaja con toda la población:
 - Imposible porque es infinita
 - Porque se destruye a los individuos
 - Por razones técnicas
 - Por razones económicas
- Se analiza sólo una parte:



Muestra: subconjunto de individuos de la población sobre los que se recogen los datos a estudiar

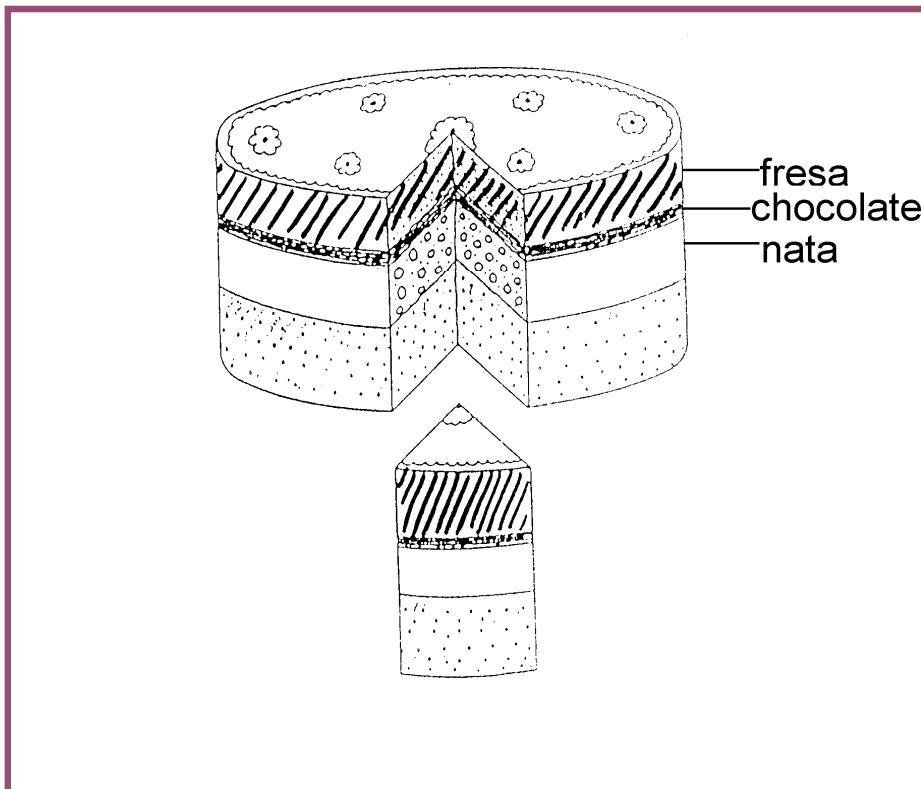


Muestras. Datos estadísticos

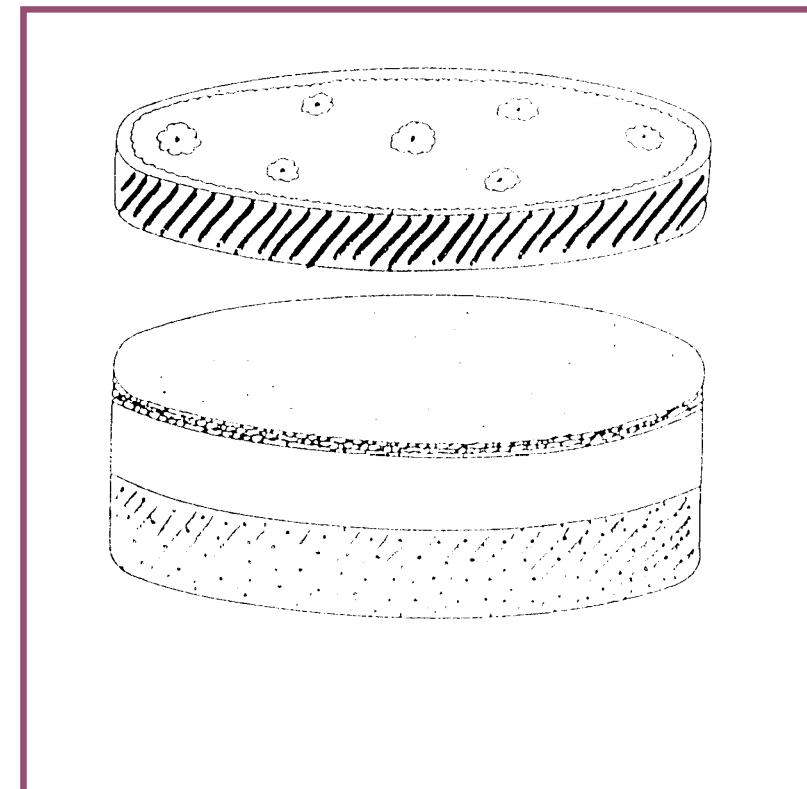


Muestras. Datos estadísticos

- Mediante Técnicas de Muestreo se consigue que la muestra sea representativa para poder inferir las conclusiones a la población a la que pertenece



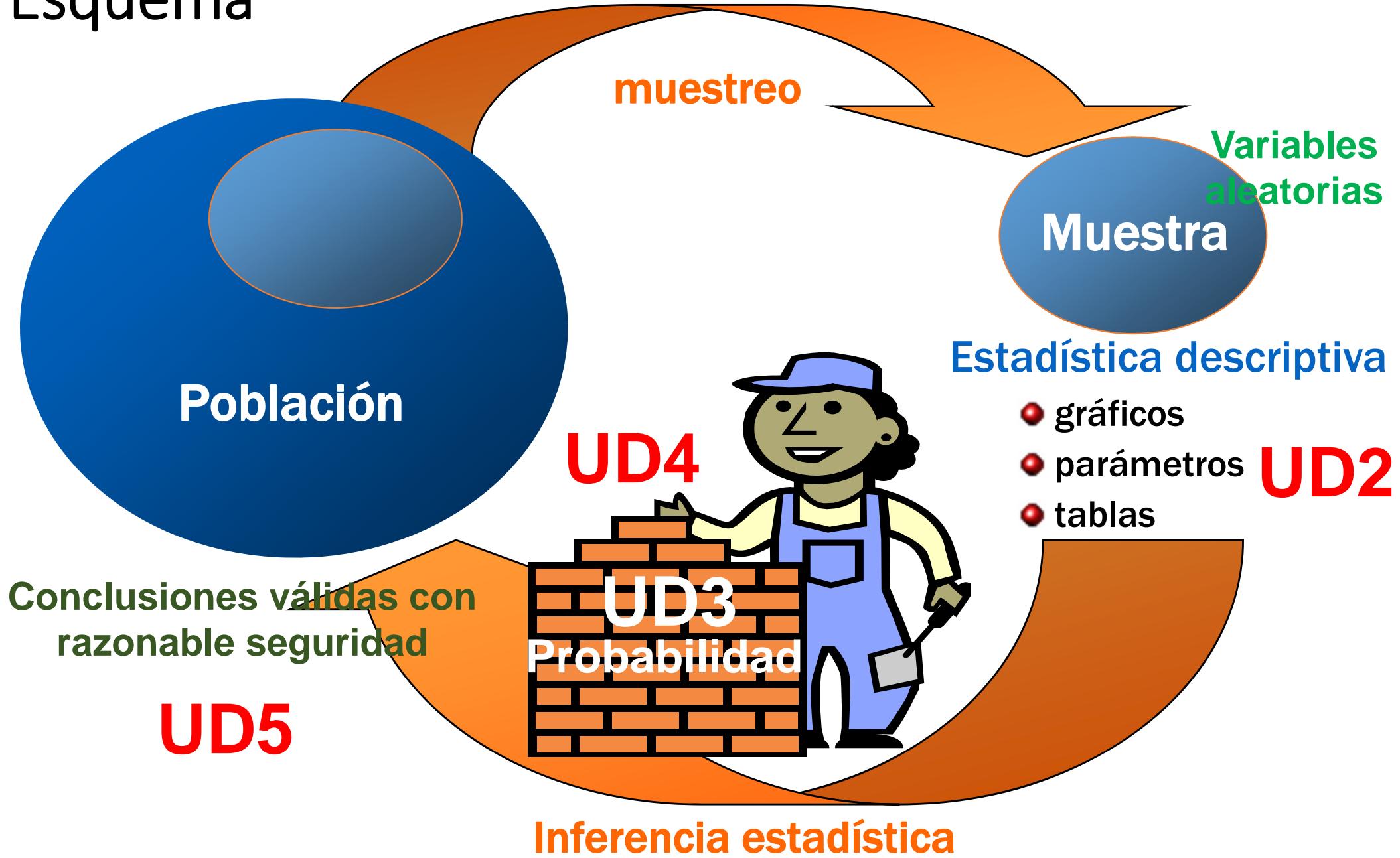
La muestra debe ser
representativa



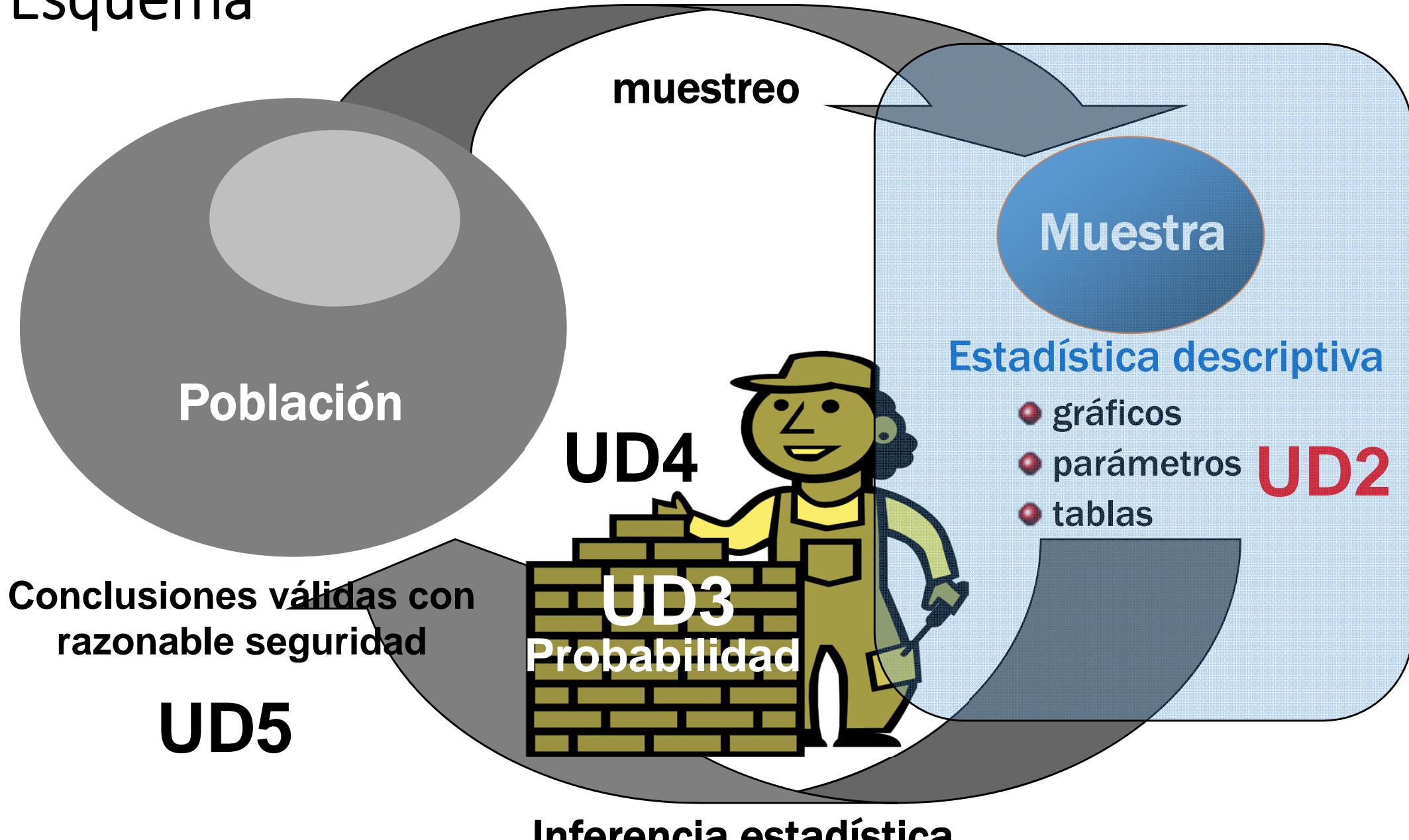
Ejemplo de un pésimo
muestreo



Esquema



Esquema



Datos estadísticos: características coches

STATGRAPHICS Plus - Sample1.sgp - [CARDATA.SF]

File Edit Plot Describe Compare Relate Special View Window Help

mpg cylinders dispel horsepower accel year weight origin make mo

	mpg	cylinders	dispel	horsepower	accel	year	weight	origin	make	mo
1	43,1	4	90	48	21,5	78	1985	2	Volkswagen	Rabbit
2	36,1	4	98	66	14,4	78	1800	1	Ford	Fiesta
3	32,8	4	78	52	19,4	78	1985	3	Mazda	GLC
4	39,4	4	85	70	18,6	78	2070	3	Datsun	B210
5	36,1	4	91	60	16,4	78	1800	3	Honda	Civic
6	19,9	8	260	110	15,5	78	3365	1	Oldsmobile	Cutlass
7	19,4	8	318	140	13,2	78	3735	1	Dodge	Diplomat
8	20,2	8	302	139	12,8	78	3570	1	Mercury	Monarch
9	19,2	6	231	105	19,2	78	3535	1	Pontiac	Phoenix
10	20,5	6	200	95	18,2	78	3155	1	Chevrolet	Malibu
11	20,2	6	200	85	15,8	78	2965	1	Ford	Fairmont
12	25,1	4	140	88	15,4	78	2720	1	Ford	Fairmont
13	20,5	6	225	100	17,2	78	3430			
14	19,4	6	232	90	17,2	78	3210			
15	20,6	6	231	105	15,8	78	3380			
16	20,8	6	200	85	16,7	78	3070	1	Mercury	Zephyr
17	18,6	6	225	110	18,7	78	3620	1	Dodge	Aspen
18	18,1	6	258	120	15,1	78	3410	1	AMC	Concord
19	19,2	8	305	145	13,2	78	3425	1	Chevrolet	Monte Carlo
20	17,7	6	231	165	13,4	78	3445	1	Buick	Regal
21	18,1	8	302	139	11,2	78	3205	1	Ford	Futura
22	17,5	8	318	140	13,7	78	4080	1	Dodge	Magnus
23	30	4	98	68	16,5	78	2155	1	Chevrolet	Chev.
24	27,5	4	134	95	14,2	78	2560	3	Toyota	Corolla
25	27,2	4	119	97	14,7	78	2300	3	Datsun	510
26	30,9	4	105	75	14,5	78	2230	1	Dodge	Omni
27	21,1	4	134	95	14,8	78	2515	3	Toyota	Celica
28	23,2	4	156	105	16,7	78	2745	1	Plymouth	Sappo
29	23,8	4	151	85	17,6	78	2855	1	Oldsmobile	Starfire
30	23,9	4	119	97	14,9	78	2405	3	Datsun	200-S
31	20,3	5	131	103	15,9	78	2830	2	Audi	5000

Individuo muestra

Variable



Datos estadísticos

¿Qué información podemos sacar de estos datos?

¿Qué podemos decir acerca del PESO de los coches?



- ✓ Hace falta sintetizar y simplificar la presentación de los datos antes de su análisis.
- ✓ Es necesario disponer los valores observados de forma clara y útil para su interpretación.



Tablas, herramientas gráficas y parámetros



Contenido

0. Introducción

1. Estadística Descriptiva Unidimensional

1.1 Conceptos básicos (población, Muestra, v.a.)

1.2 Tablas de Frecuencias

1.3 Diagramas: Barras, Sectores

1.4 Histogramas.

1.5 Parámetros de Posición

1.6 Parámetros de Dispersion

1.7 Parámetros de Asimetría y Curtosis

1.8 Diagrama Box & Whisker



Contenido

2. Estadística Descriptiva Bidimensional

2.1 Tablas de frecuencias cruzadas

2.2 Distribuciones marginales y condicionales

2.3 Diagramas de dispersión

2.4 Covarianza y Coeficiente de Correlación

Lo veremos en la UD 5



1.2 – Tablas de frecuencias

v.a. Cualitativas y Cuantitativas



Variable cualitativa o cuantitativa con pocos valores

Nº de procesadores	Nº de robots	% robots	Nº de robots acumulado	% robots acumulado
0	10	6,25%	10	6,25%
1	35	21,88%	45	28,13%
2	60	37,50%	105	65,63%
3	55	34,37%	160	100%
Total	160	100%		



Variables cualitativas

Variable cualitativa o cuantitativa con pocos valores

Valores variable	Frecuencia absoluta	Frecuencia relativa %	Frecuencia absoluta acumulada	Frecuencia relativa acumulada %
x_1	f_1	$f_{r1} = f_1/N * 100$	$F_1 = f_1$	$F_{r1} = f_{r1}$
x_2	f_2	$f_{r2} = f_2/N * 100$	$F_2 = F_1 + f_2$	$F_{r2} = F_{r1} + f_{r2}$
x_3	f_3	$F_{r2} = f_3/N * 100$	$F_3 = F_2 + f_3$	$F_{r3} = F_{r2} + f_{r3}$
...	N	100%
Total	N	100%		



Variables cualitativas

Variable Cualitativa o cuantitativa con pocos valores

Otro Ejemplo: Acceso¹ a Internet de las Viviendas - 2º semestre

2005 Copyright INE 2007

x_i	n_i	f_i
Línea telefónica convencional	1.789.513	0,20
Banda Ancha (ADSL, RDSI, Red Cable)	3.491.449	0,38
Línea ADSL	2.814.462	0,31
Línea RDSI	53.808	0,01
Red de cable	691.324	0,08
Telefonía móvil	212.785	0,02
Otras formas de conexión	68.105	0,01
1 Acceso más utilizado	$\sum n_i = 9.121.446$	$\sum f_i = 1$



Ejemplo: nº de ordenadores en casa

Variable cuantitativa discreta con pocos valores

Valores posibles	Frecuencia absoluta	Frecuencia relativa %	F absoluta acumulada	F relativa acumulada %
0				
1				
2				
3				



Ejemplo: SO móvil

Variable cualitativa

Valores posibles	Frecuencia absoluta	Frecuencia relativa %	F absoluta acumulada	F relativa acumulada %



Variables cuantitativa

- Podríamos emplear el mismo procedimiento
 - La probabilidad de encontrar valores repetidos es muy baja
 - Obtendríamos una tabla tan difícil de interpretar como los datos originales



Solución

Agrupar los datos en tramos o intervalos



Variables cuantitativa

- Resistencia de carcasas plásticas de procesadores (Nw/mm²)

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below	10.00			0	.00000	0	.00000
1	10.00	15.00	12.50	0	.00000	0	.00000
2	15.00	20.00	17.50	1	.00610	1	.00610
3	20.00	25.00	22.50	9	.05488	10	.06098
4	25.00	30.00	27.50	18	.10976	28	.17073
5	30.00	35.00	32.50	26	.15854	54	.32927
6	35.00	40.00	37.50	38	.23171	92	.56098
7	40.00	45.00	42.50	34	.20732	126	.76829
8	45.00	50.00	47.50	20	.12195	146	.89024
9	50.00	55.00	52.50	9	.05488	155	.94512
10	55.00	60.00	57.50	5	.03049	160	.97561
11	60.00	65.00	62.50	0	.00000	160	.97561
12	65.00	70.00	67.50	3	.01829	163	.99390
13	70.00	75.00	72.50	1	.00610	164	1.00000

Mean = 39.3288 Standard Deviation = 9.46009 Median = 39.1



Variables cuantitativa

	Nº de intervalo	Límite inferior	Límite superior	Valor central	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below		10.00			0						.00000
1	10.00	15.00	12.50		0						.00000
2	15.00	20.00	17.50		1						.00610
3	20.00	25.00	22.50		9						.06098
4	25.00	30.00	27.50	27.50	18				.05488		
5	30.00	35.00	32.50		26				.10976		
6	35.00	40.00	37.50		38				.15854		
7	40.00	45.00	42.50		34				.23171		
8	45.00	50.00	47.50		20				.20732		
9	50.00	55.00	52.50		9				.12195	146	.89024
10	55.00	60.00	57.50		5				.05488	155	.94512
11	60.00	65.00	62.50		0				.03049	160	.97561
12	65.00	70.00	67.50		3				.00000	160	.97561
13	70.00	75.00	72.50		1				.01829	163	.99390
									.00610	164	1.00000

Mean = 39.3288 Standard Deviation = 9.46009 Median = 39.1



Variables cuantitativa

- Se divide el campo de variabilidad en un conjunto de **K intervalos** de igual longitud, teniendo en cuenta:
 - Límites de cada intervalo ($[,]$, \leq , \geq)
 - Valor central del intervalo
 - Número de observaciones por intervalo



Problema: ¿Amplitud Óptima?

¿Número de intervalos?



Variables cuantitativa

- Número de intervalos grande

→ Tabla difícil de interpretar

- Número de intervalos pequeño

→ Se puede perder información importante



Recomendable:

!! Entre 5 y 15 intervalos !!

(dependiendo del número de observaciones de la muestra)



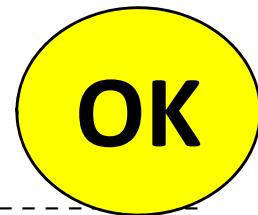
Otro ejemplo Estatura: 27 intervalos

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequen
at or below	140,0			0	0,0000	0	0,00
1	140,0	142,963	141,481	0	0,0000	0	0,00
2	142,963	145,926	144,444	0	0,0000	0	0,00
3	145,926	148,889	147,407	0	0,0000	0	0,00
4	148,889	151,852	150,37	0	0,0000	0	0,00
5	151,852	154,815	153,333	1	0,0076	1	0,00
6	154,815	157,778	156,296	3	0,0229	4	0,03
7	157,778	160,741	159,259	8	0,0611	12	0,09
8	160,741	163,704	162,222	12	0,0916	24	0,18
9	163,704	166,667	165,185	11	0,0840	35	0,26
10	166,667	169,63	168,148	10	0,0763	45	0,34
11	169,63	172,593	171,111	14	0,1069	59	0,45
12	172,593	175,556	174,074	28	0,2137	87	0,66
13	175,556	178,519	177,037	9	0,0687	96	0,73
14	178,519	181,481	180,0	14	0,1069	110	0,81
15	181,481	184,444	182,963	6	0,0458	116	0,88
16	184,444	187,407	185,926	9	0,0687	125	0,95
17	187,407	190,37	188,889	1	0,0076	126	0,96
18	190,37	193,333	191,852	2	0,0153	128	0,97
19	193,333	196,296	194,815	2	0,0153	130	0,99
20	196,296	199,259	197,778	1	0,0076	131	1,00
21	199,259	202,222	200,741	0	0,0000	131	1,00
22	202,222	205,185	203,704	0	0,0000	131	1,00
23	205,185	208,148	206,667	0	0,0000	131	1,00
24	208,148	211,111	209,63	0	0,0000	131	1,00
25	211,111	214,074	212,593	0	0,0000	131	1,00
26	214,074	217,037	215,556	0	0,0000	131	1,00
27	217,037	220,0	218,519	0	0,0000	131	1,00
above	220,0			0	0,0000	131	1,00

Muchos intervalos:
Difícil de interpretar



Otro ejemplo Estatura: 10 intervalos



Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below 1	150,0	155,0	152,5	0	0,0000	0	0,0000
2	155,0	160,0	157,5	3	0,0229	3	0,0229
3	160,0	165,0	162,5	9	0,0687	12	0,0916
4	165,0	170,0	167,5	22	0,1679	34	0,2595
5	170,0	175,0	172,5	16	0,1221	50	0,3817
6	175,0	180,0	177,5	37	0,2824	87	0,6641
7	180,0	185,0	182,5	22	0,1679	109	0,8321
8	185,0	190,0	187,5	14	0,1069	123	0,9389
9	190,0	195,0	192,5	3	0,0229	126	0,9618
10	195,0	200,0	197,5	2	0,0153	129	0,9847
above	200,0			0	0,0000	131	1,0000

Mean = 172,855 Standard deviation = 9,07585



Var. Cualitativas: Tablas de contingencia

- **Objetivo del análisis:** describir la relación existente entre las dos componentes de la v.a. bidimensional
- **Herramienta:** **Tabla de Contingencia**
 - variables **discretas**:
 - naturaleza cualitativa (codificadas)
 - cuantitativas con pocos valores

Ejemplo Tablas de contingencia

- En una empresa se realiza una encuesta relativa a diferentes características de sus empleados con el fin de llevar a cabo un estudio.
- Una de las cuestiones a estudiar es la posible relación entre el SEXO y la CATEGORÍA LABORAL de sus empleados.
- Para ello, en primer lugar se construye una TABLA DE CONTINGENCIA

Frecuencias conjuntas

	Administrativo	Seguridad	Directivo	
Hombre	157	27	74	2 - Nº de mujeres que ocupan cargo de directivo
Mujer	206	0	10	
1 - Nº de empleados encuestados Tamaño de la muestra (N)				474

Frecuencias marginales (absolutas y relativas)

	Administrativo	Seguridad	Directivo	Frecuencias Marginales De SEXO
Hombre	157	4 - Nº de mujeres encuestadas (206 + 0 + 10)	74	258 54,4%
Mujer	206	0	10	216 45,6%
Frecuencias Marginales De CATEGORÍA	363 76,6%	3 - % de empleados administrativos (363 /474*100)	5,7%	474 17,7%

Frec. Relativas condicionales de CATEGORÍA en función de SEXO

	Administrativo	Seguridad	Directivo	Frecuencias Marginales De SEXO
Hombre	60,9%	10,5%	28,7%	54,4%
Mujer	206	0	10	216
Frecuencias Marginales De CATEGORÍA	363	27	84	474
76,6%		5,7%	17,7%	
<p>5 - % de las mujeres que son directivas $(10/ 216 * 100)$</p>				

Frecuencias relativas a los totales de las filas (SEXO)

Frec. Relativas condicionales de SEXO en función de CATEGORÍA

	Administrativo	Seguridad	Directivo	Frecuencias Marginales De SEXO
Hombre	157	27	74	258
Mujer	206	100%	88,1%	54,4%
Frecuencias Marginales De CATEGORÍA	363	27	84	474
	56,7%	6 - % de las administrativos que son hombres (157/ 363 *100)	17,7%	45,6%

Frec. relativas a los totales de las columnas (CATEGORÍA)



Frecuencias relativas condicionales

- La frecuencia relativa condicional a calcular depende del objetivo de nuestro estudio.
- **EJEMPLO:** Si queremos estudiar si la proporción de mujeres (u hombres) es igual o diferente para los diferentes cargos:
 - ¿Frecuencia relativa condicional de CATEGORÍA en función de SEXO?
 - ó
 - ¿Frecuencia relativa condicional de SEXO en función de CATEGORÍA?

Frec. Relativas condicionales de SEXO en función de CATEGORÍA

	Administrativo	Seguridad	Directivo	Frecuencias Marginales De SEXO
Hombre	157 43,3%	27 100%	74 88,1%	258 54,4%
Mujer	206 56,7%	0 0%	10 11,9%	216 45,6%
Frecuencias Marginales De CATEGORÍA	363 76,6%	27 5,7%	84 17,7%	474

NO se puede deducir que en el grupo de directivos hay más hombres que mujeres, ya que en el total de la muestra hay más hombres (258 hombres frente a 216 mujeres)

Frec. Relativas condicionales de CATEGORÍA en función de SEXO

	Administrativo	Seguridad	Directivo	Frecuencias Marginales De SEXO
Hombre	157	27	74	258
Mujer	206	0	10	216
Frecuencias Marginales De CATEGORÍA	363	27	84	474
	76,6%	5,7%	17,7%	

Ahora SÍ se puede deducir que en el grupo de mujeres hay más o menos directivos, ya que las frecuencias están relativizadas con respecto a los totales de mujeres y hombres para que éstos no influyan. De las mujeres un 4,6% son directivos frente al 28,7% de los hombres que son directivos.

SEGUN ESTA ENQUISITA, EL 57%
DEL 24,3% QUE APOYA AL 31,7%
DE LOS QUE NO SABEN/NO CONTESTAN
CONSIDERAN QUE EL 82% DEPENDE LO QUE

... Y OJO, QUE ESTA
HECHA POR
LA PRESTIGIOSA
AGENCIA
PEDALSCOPIA,
PAQUI



Var. Cuantitativas: Tablas de frecuencias cruzadas

- Todo lo dicho respecto al cálculo e interpretación de las frecuencias de la Tabla de Contingencia es aplicable al caso de las v.a. cuantitativas.
- La única diferencia, como en el caso de las v.a. unidimensionales, es que previamente a la representación de la Tabla es necesario agrupar los valores de las variables en intervalos.

Tablas de frecuencias cruzadas

ESTATURA	145	155	165	175	185	Row Total
PESO	155	165	175	185	195	
40 - 55	9	17	0	0	0	
55 - 70	3	18	31	5	0	
70 - 85	0	3	24	12	3	
85 - 99	0	0	3	0	2	
Column Total						
Total						

Frecuencia conjunta: peso y estatura

Frecuencia marginal absoluta: peso y estatura

Frecuencia marginal relativa: peso y estatura

Frecuencia relativa de peso condicionada a estatura

Tablas de frecuencias cruzadas

ESTATURA PESO		145	155	165	175	185	Row Total
40	55	9	17	0	0	0	26
55	70	3	18	31	5	0	57
70	85	0	3	24	12	3	42
85	99	0	0	3	0	2	5
Column Total		12	38	58	17	5	130
		9.2	29.2	44.6	13.1	3.8	100

Frecuencia conjunta: peso y estatura

Frecuencia marginal absoluta: peso y estatura

Frecuencia marginal relativa: peso y estatura

Frecuencia relativa de peso condicionada a estatura

1.3 – Diagramas de Barras Y Sectores

- v.a. Cualitativas
- V.a. Cuantitativas discretas con pocos valores distintos
- Forma gráfica de la Tabla de frecuencias



Diagrama de Barras

- A cada alternativa de la **característica cualitativa** se le hace corresponder un barra (horizontal o vertical) cuya altura es proporcional a la **frecuencia** observada (**absoluta, relativa, acumulada**) en la muestra para ese “valor” de la característica estudiada.

Barchart for Nº de cilindros

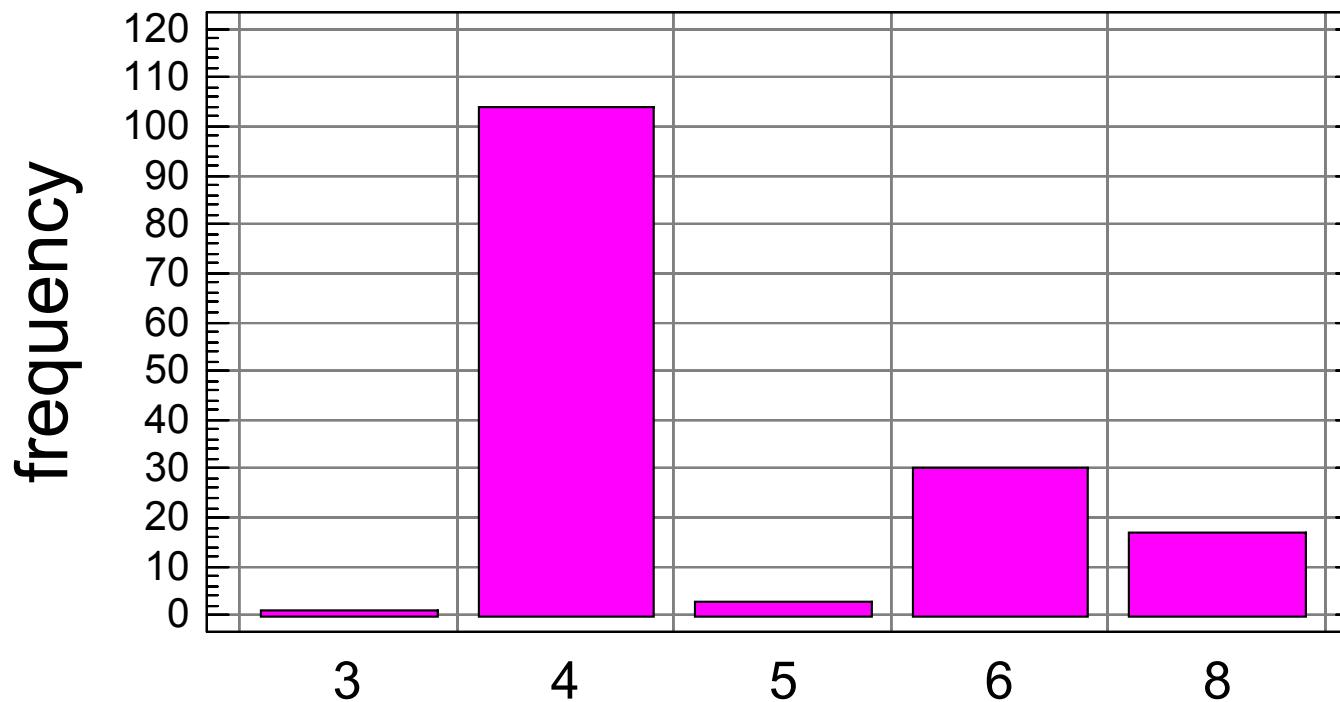
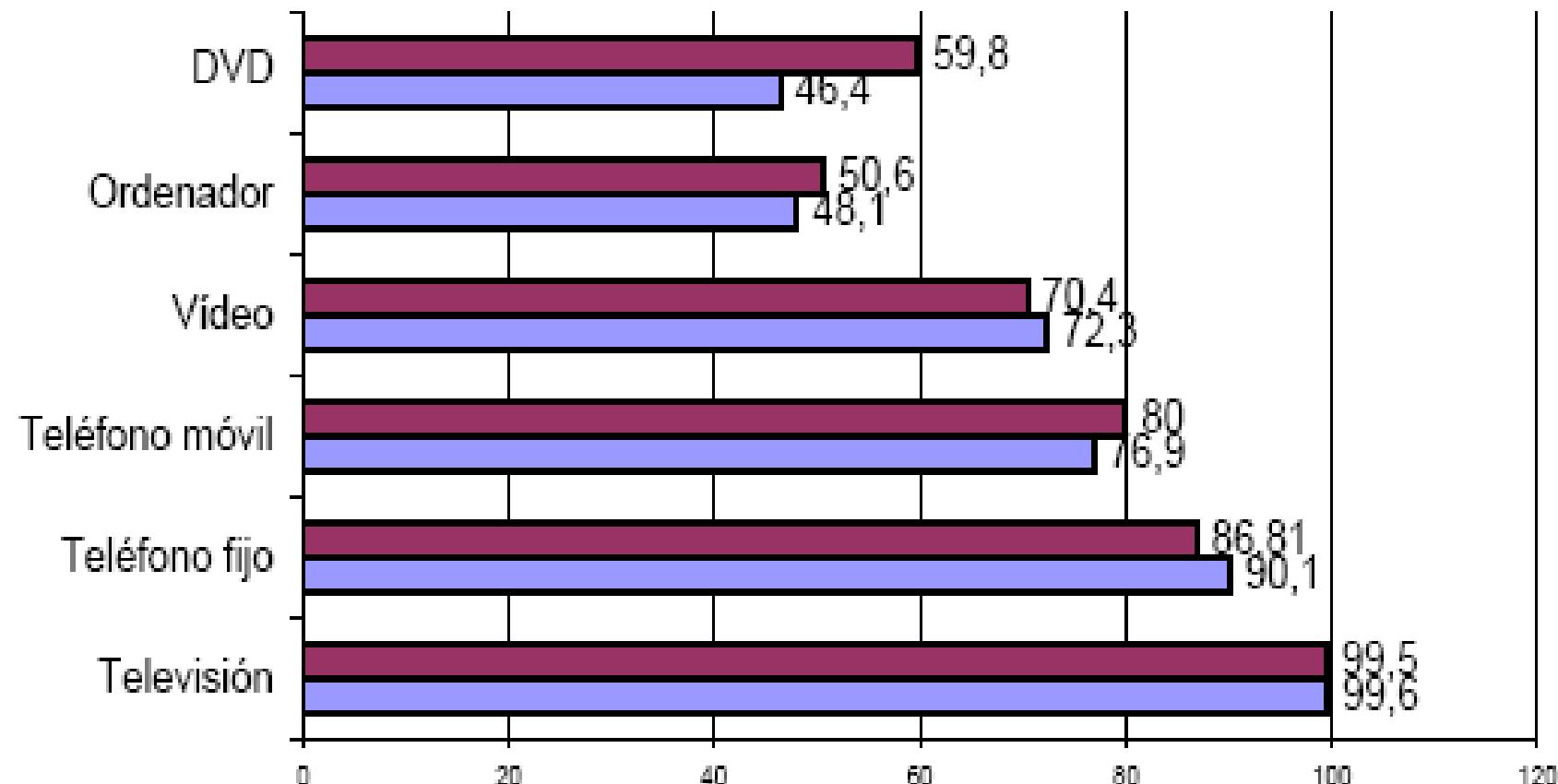


Diagrama de Barras

Equipamiento de las viviendas productos TIC (% hogares)



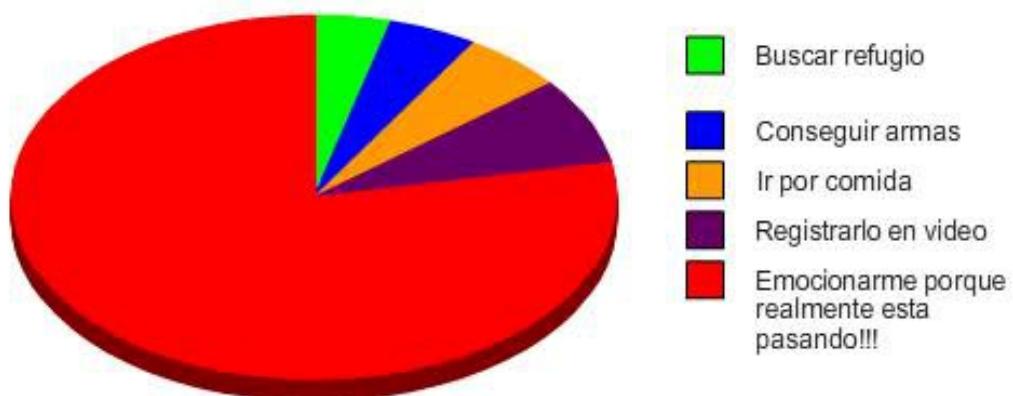
■ 2004 ■ 2005



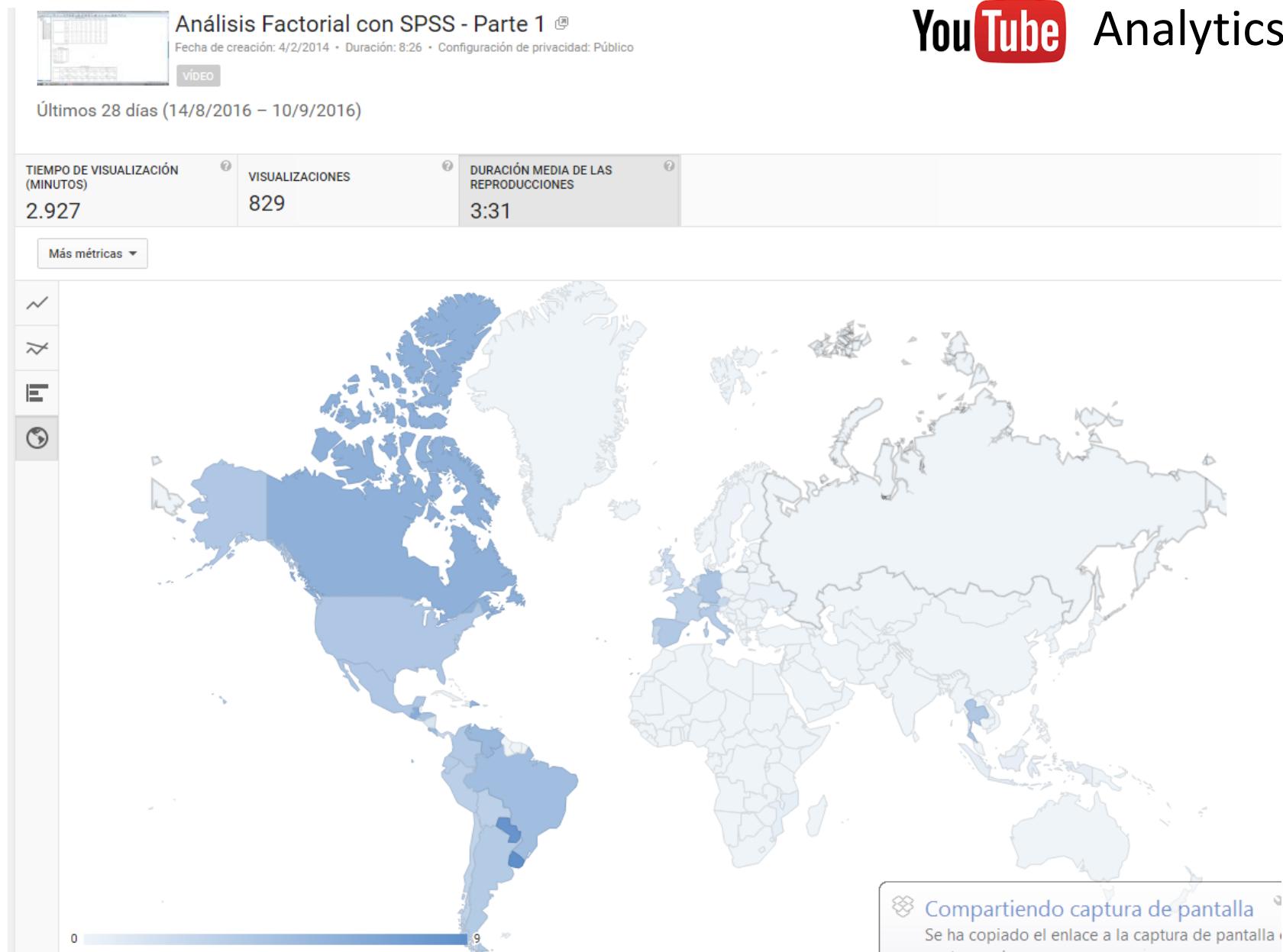
Diagrama de sectores o tarta

- La superficie total de un círculo se reparte en **sectores** cuyas áreas son proporcionales a las frecuencias observadas en la muestra para cada “valor” de la característica estudiada.
- Frecuencias absolutas o relativas

Cosas que haría durante un apocalipsis zombie



Infogramas o pictogramas



1.4 – Histogramas

v.a. Cuantitativas:

- Continuas
- Discretas con muchos valores distintos



Histograma ¿Qué es?

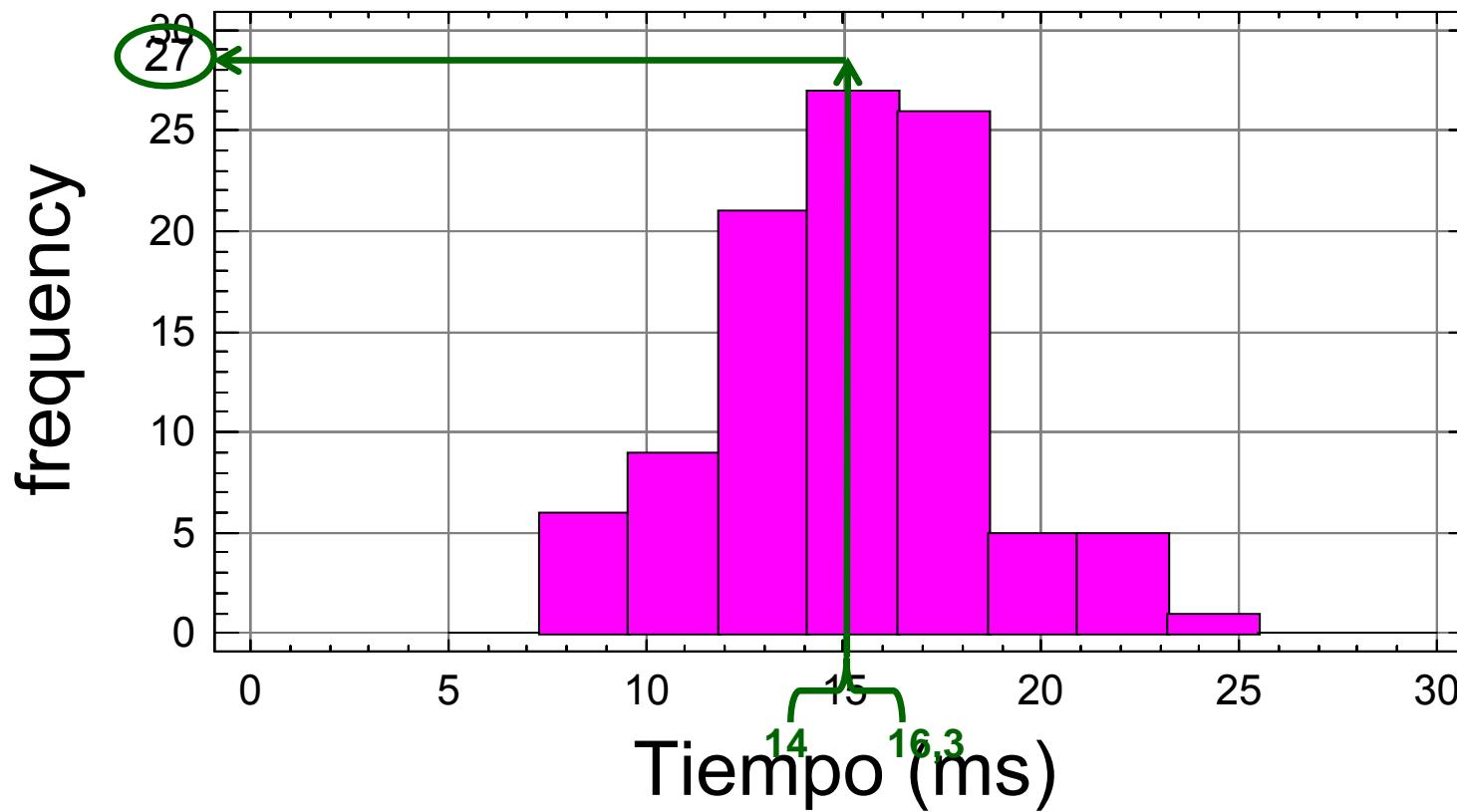
- Es un diagrama de barras para variables cuantitativas continuas o discretas con muchos valores
-  ■ Es una representación gráfica de un conjunto de datos (**mínimo 40-50 datos**)
- Para cada valor o intervalo de valores de la variable (eje de abscisas) se levanta una barra de altura proporcional a la frecuencia con que aparece dicha variable los valores del intervalo (absoluta o relativa)
- Nº de intervalos
 - regla empírica: entero cercano a \sqrt{n}
 - en general entre 15-20 intervalos



Ejemplo

v.a.: Tiempo de ejecución (ms) de 100 programas

Frecuencias absolutas (número de programas)



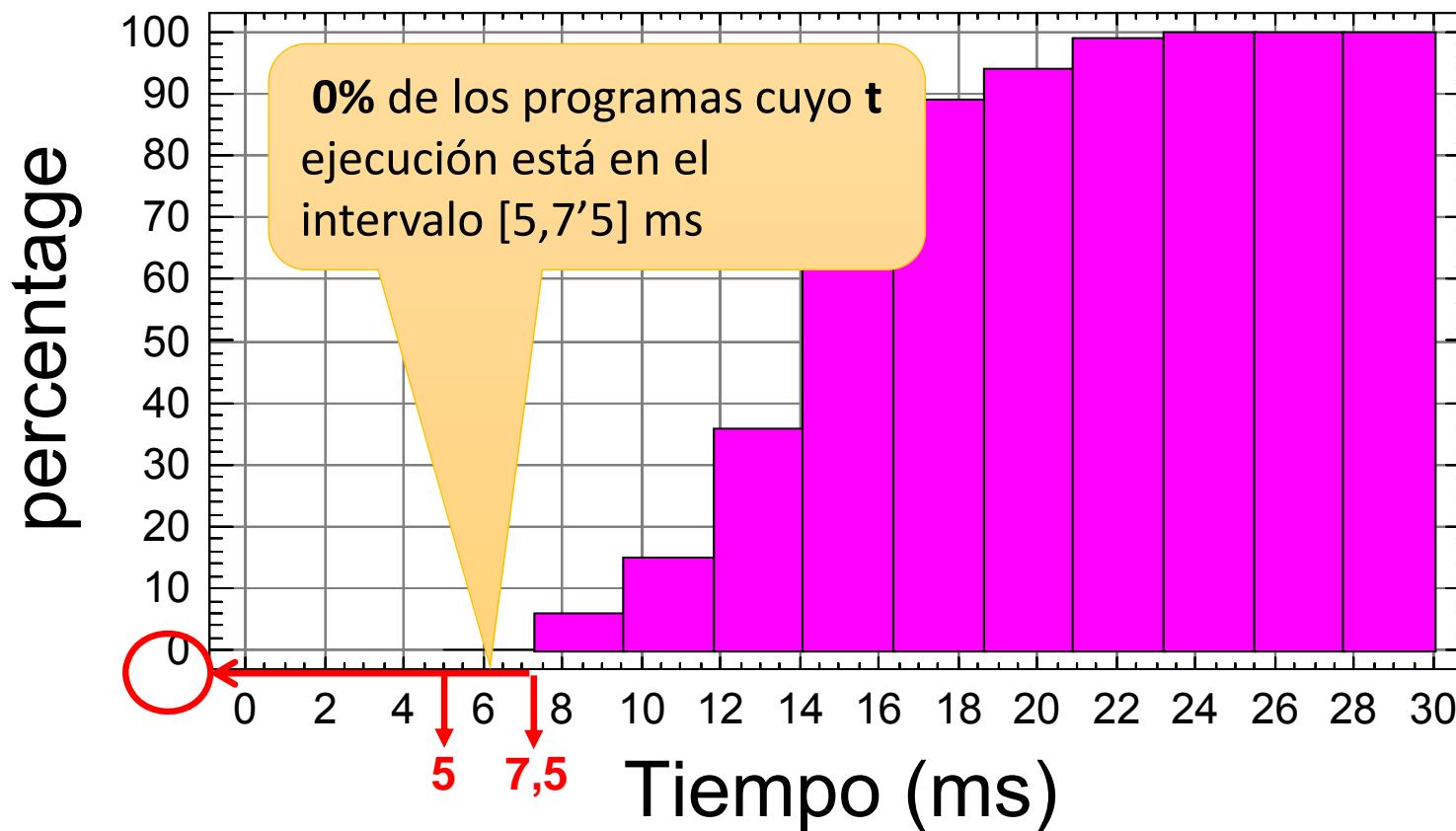
27 programas
cuyo tiempo de
ejecución ha
estado entre
14 y 16,3 ms.
(aprox.)



Ejemplo

v.a.: Tiempo de ejecución (ms) de 100 programas

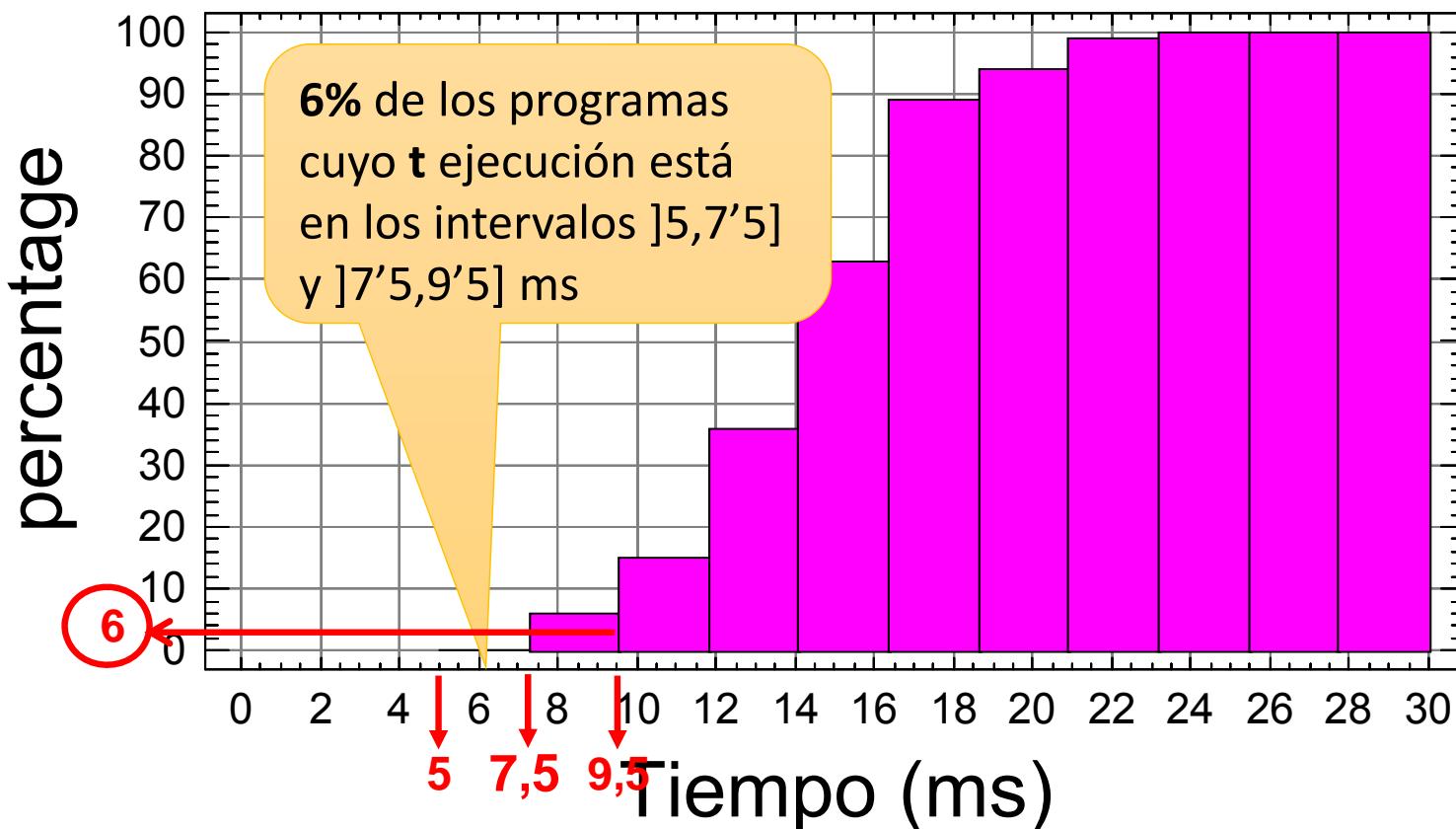
Frecuencias relativas acumuladas (% de programas)



Ejemplo

v.a.: Tiempo de ejecución (ms) de 100 programas

Frecuencias relativas acumuladas (% de programas)



El 6% de los programas tienen tiempo de ejecución menor o igual a 9,5 ms. (aprox.)



Utilidad



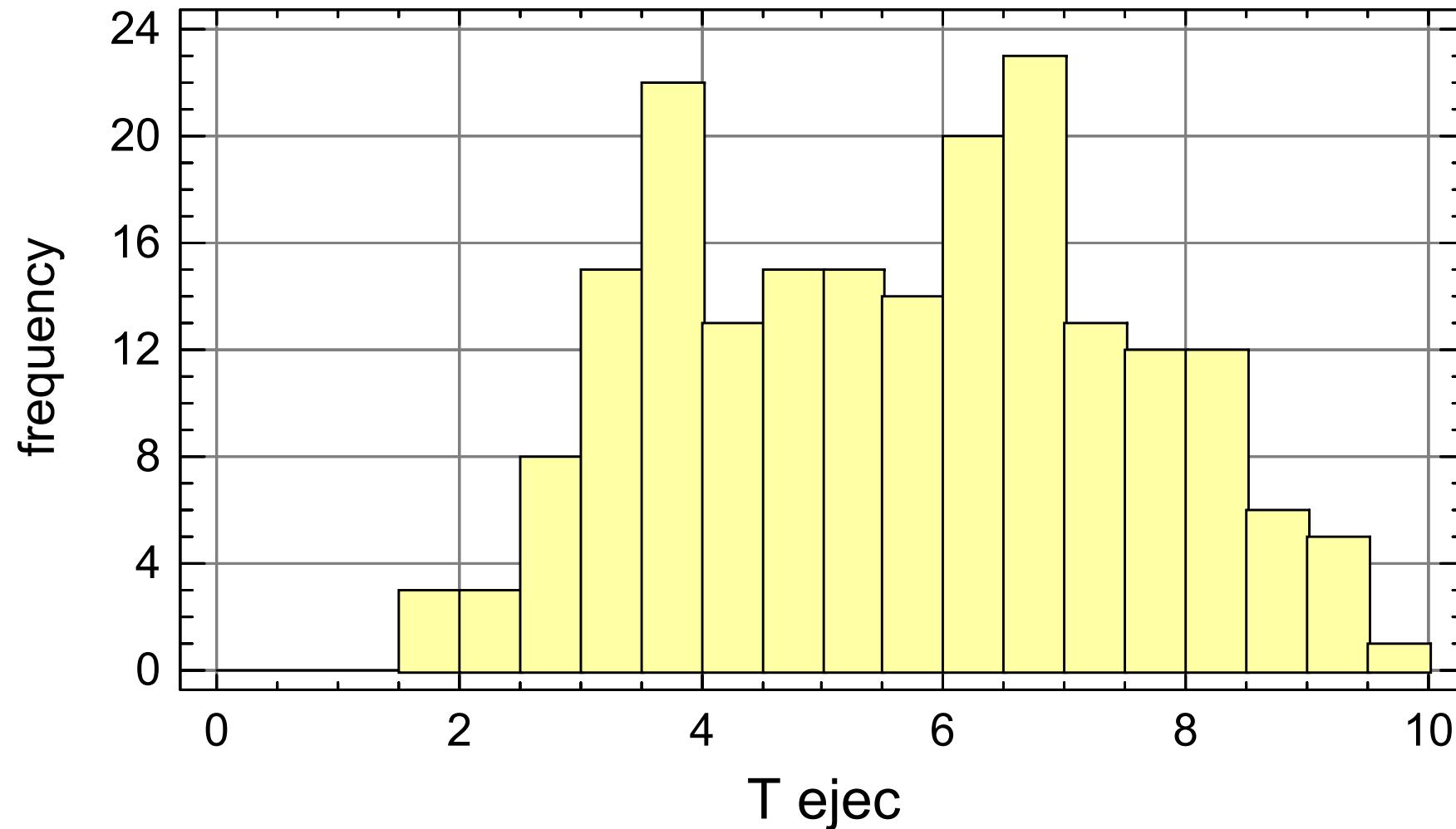
■ Podemos detectar rápidamente:

- Existencia de datos anómalos
- Mezclas de poblaciones distintas
- Datos artificialmente modificados
- No normalidad de los datos

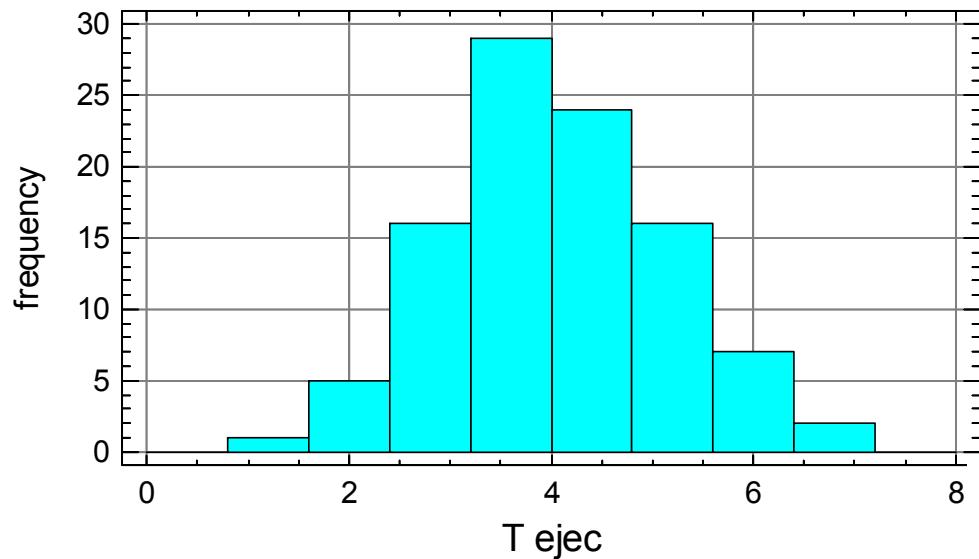


Histogramas tipo: mezcla de poblaciones

Tiempo de ejecución (ms) de 200 programas

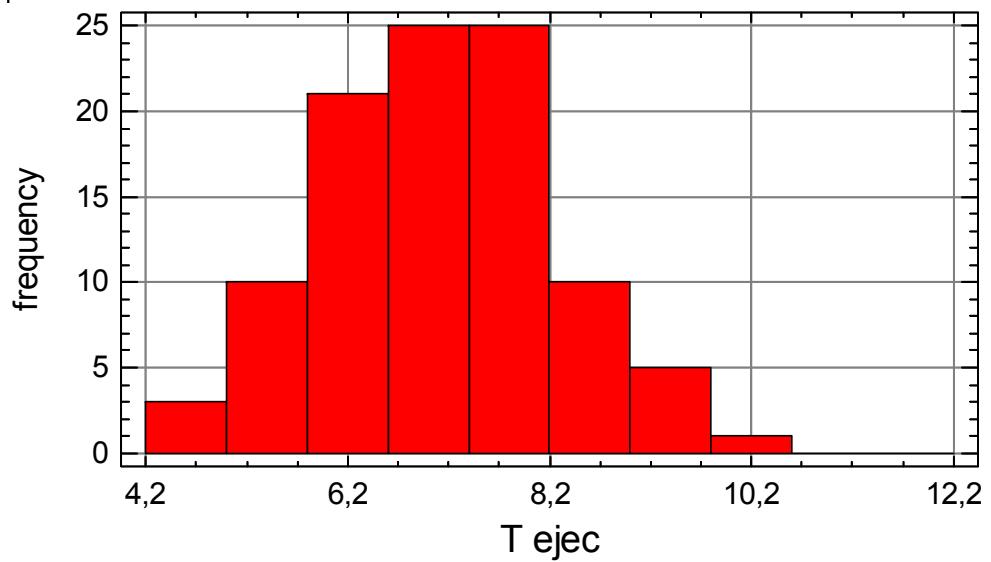


Histogramas tipo : mezcla de poblaciones

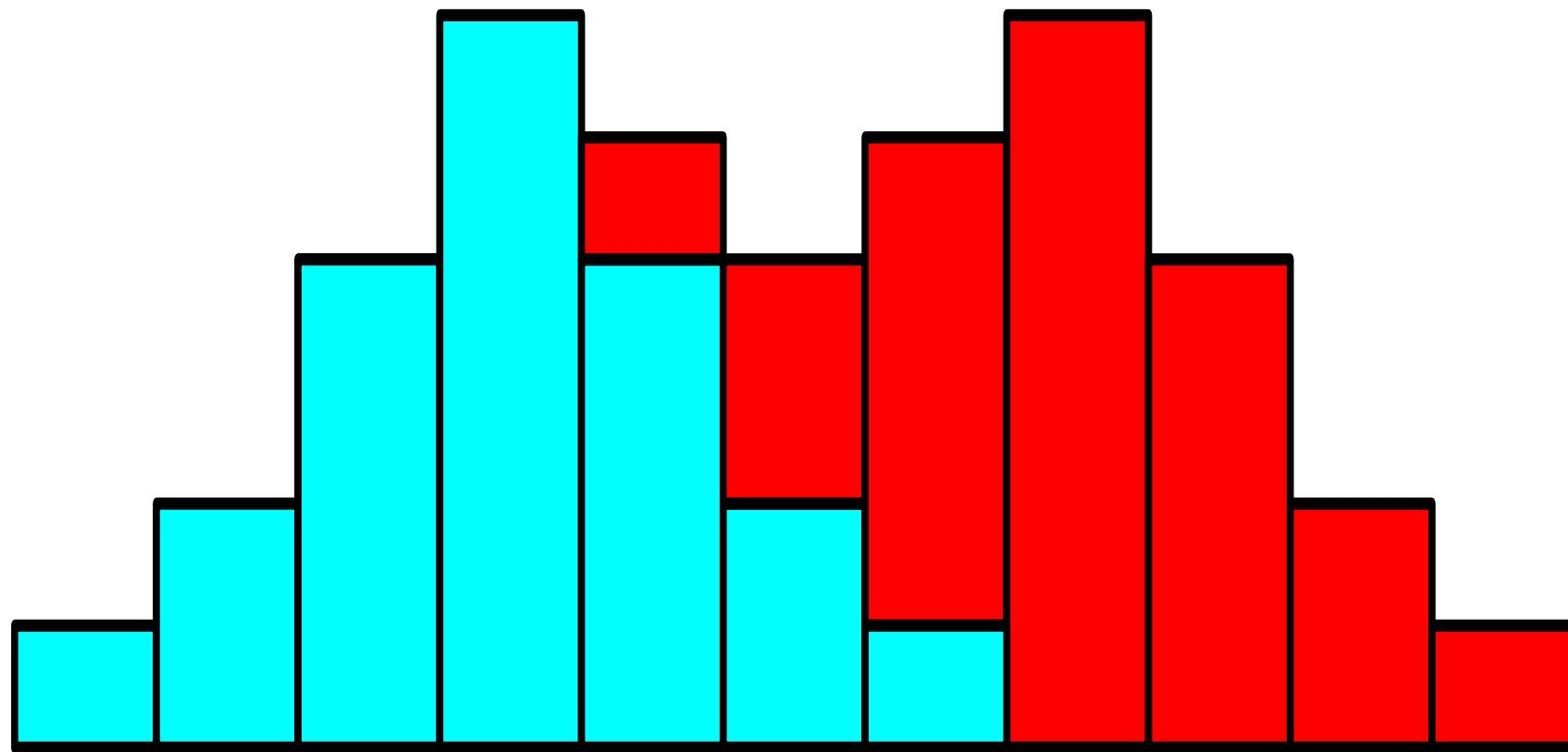


El histograma anterior
es la superposición de
estos dos

Mezcla de dos
poblaciones próximas

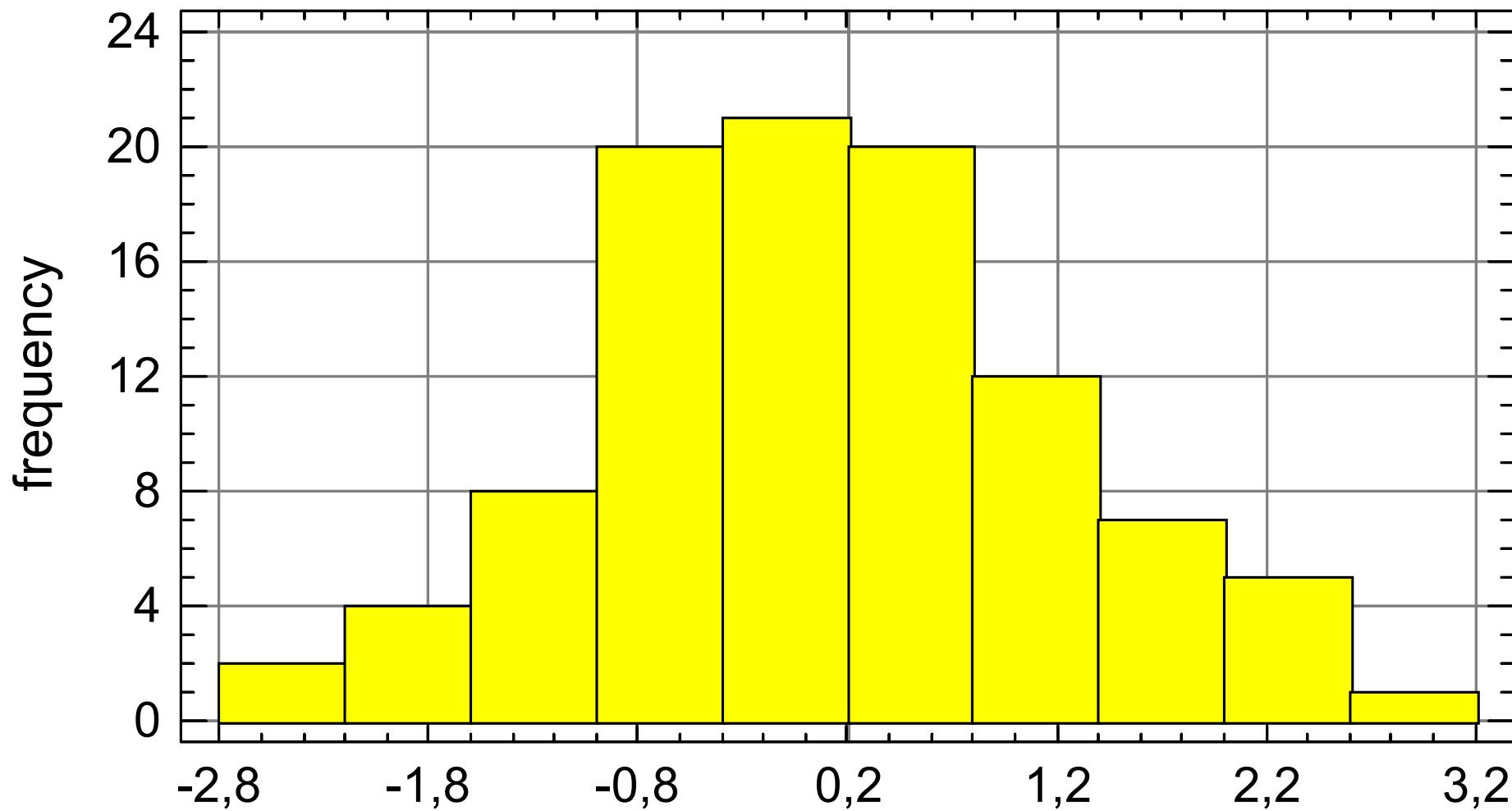


Histogramas tipo: mezcla de poblaciones

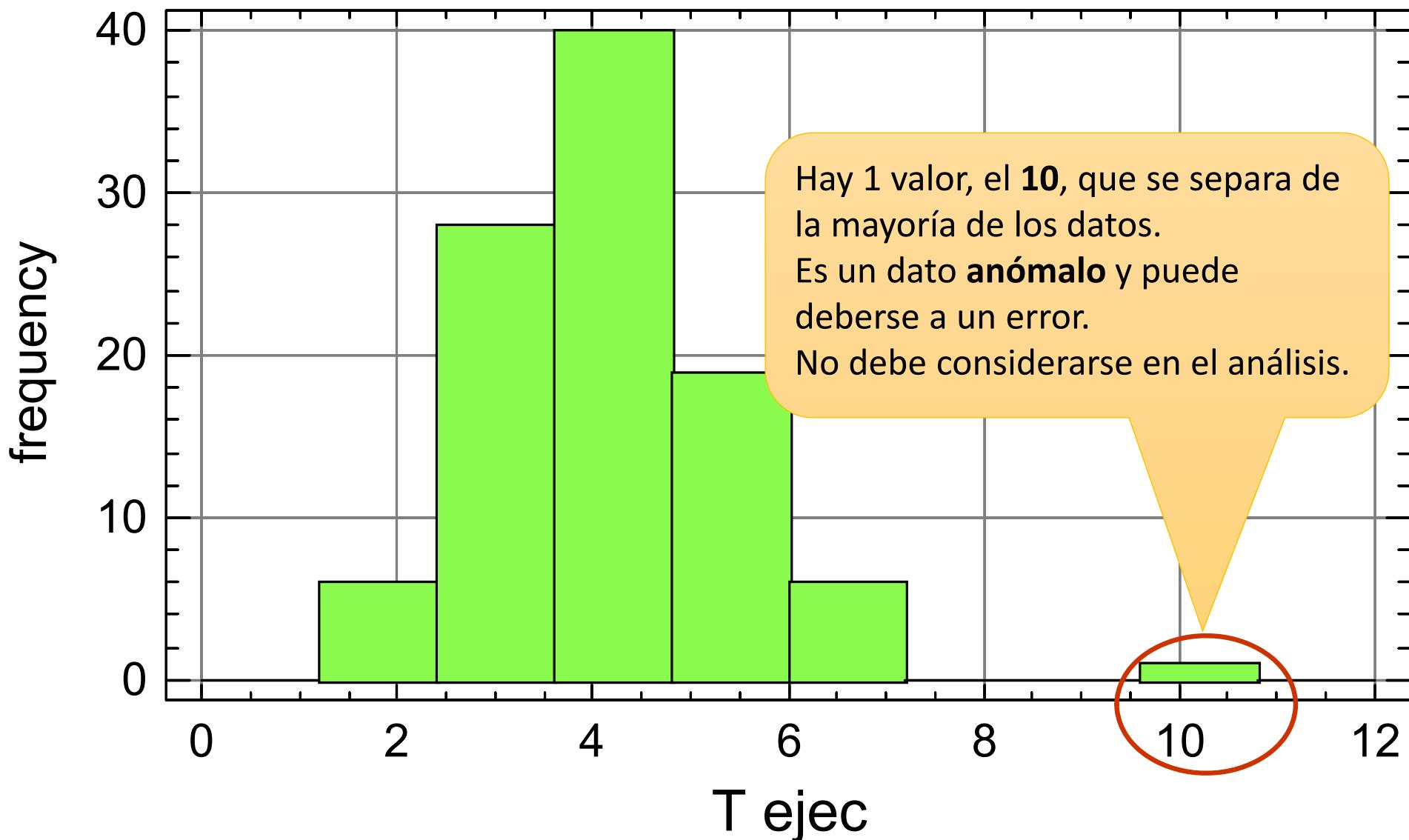


Histogramas tipo: “Normal”

v.a. : desviaciones (mm) sobre el nominal del Φ de una pieza

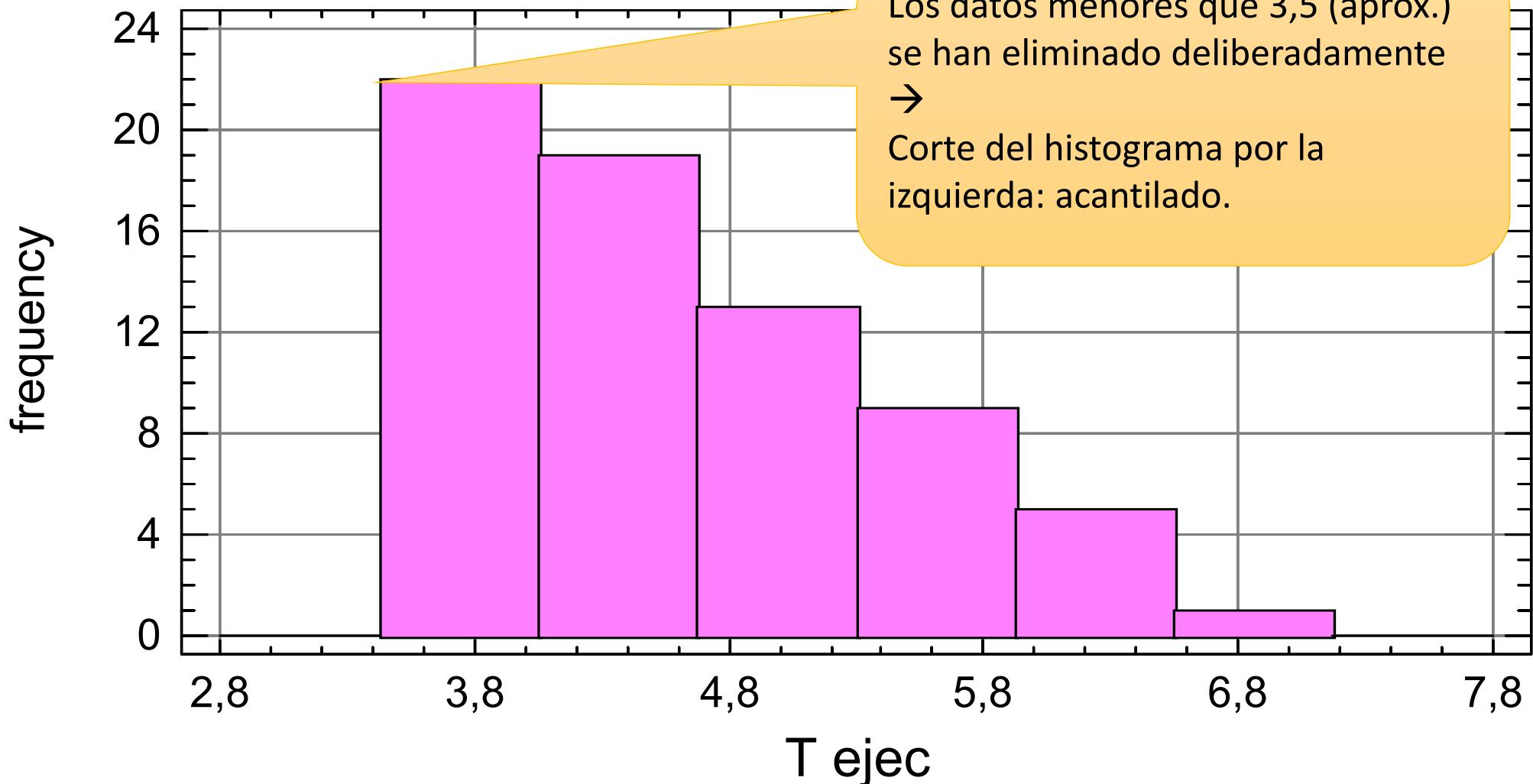


Histogramas tipo: datos anómalos



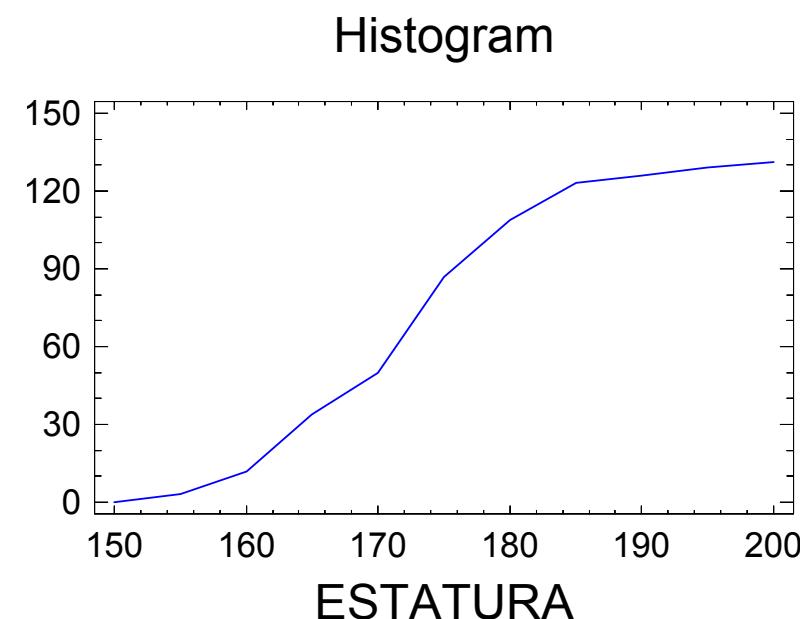
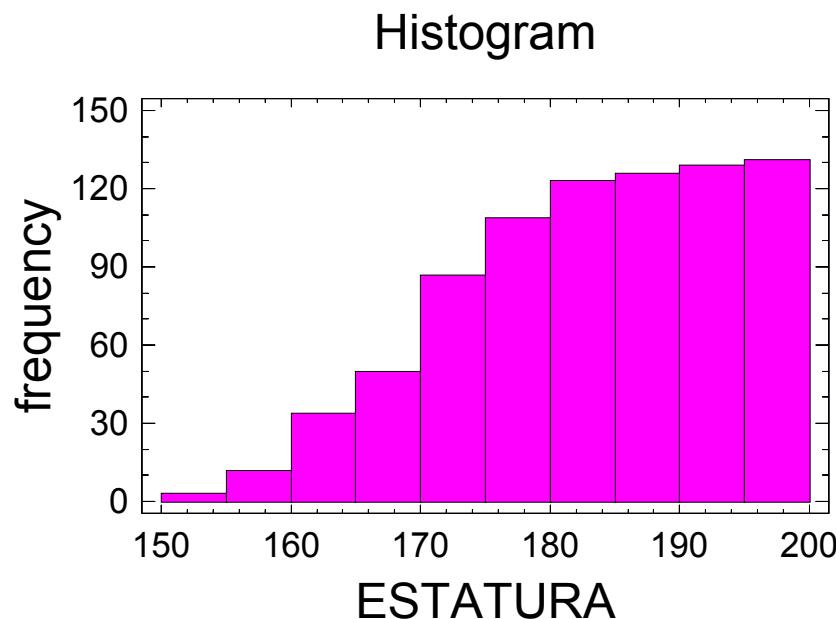
Histogramas tipo “acantilado”

Datos modificados artificialmente

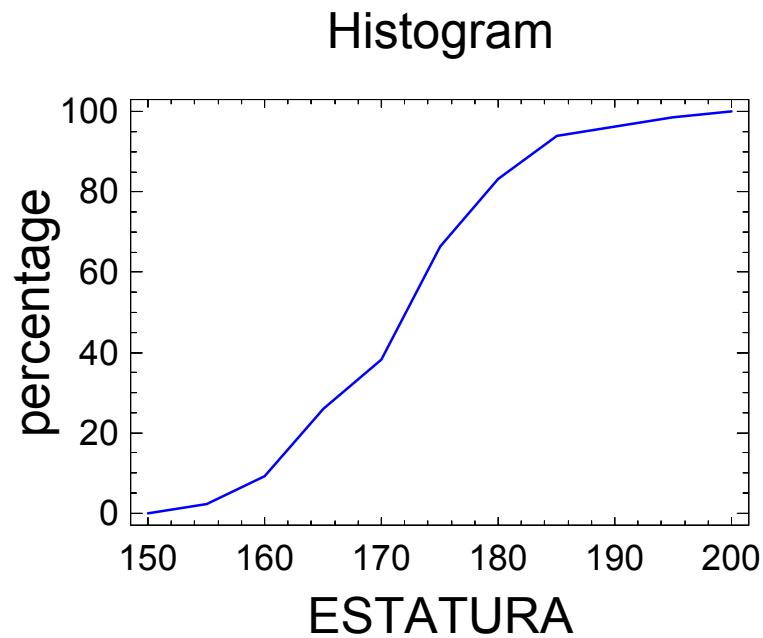
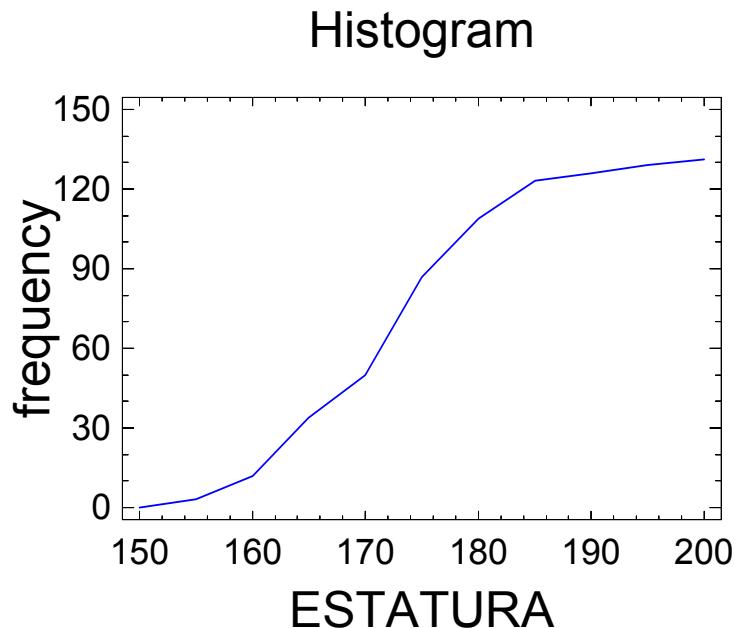


Histogramas de frecuencias acumuladas

Se puede representar un histograma a partir de las **frecuencias acumuladas** para los diferentes tramos de la variable. Si las abscisas se levantan sobre el límite de cada tramo y se unen los puntos, se obtiene una gráfica con una línea quebrada no decreciente



Polígono de frecuencias



- ▶ ¿Qué % de los alumnos miden más de 170 cm?
- ▶ ¿Qué estatura es superada por un 5% de los alumnos?
- ▶ ¿Cuántos alumnos tienen una estatura menor que 165 cm?



Parámetros muestrales

- Posición
- dispersión
- Forma



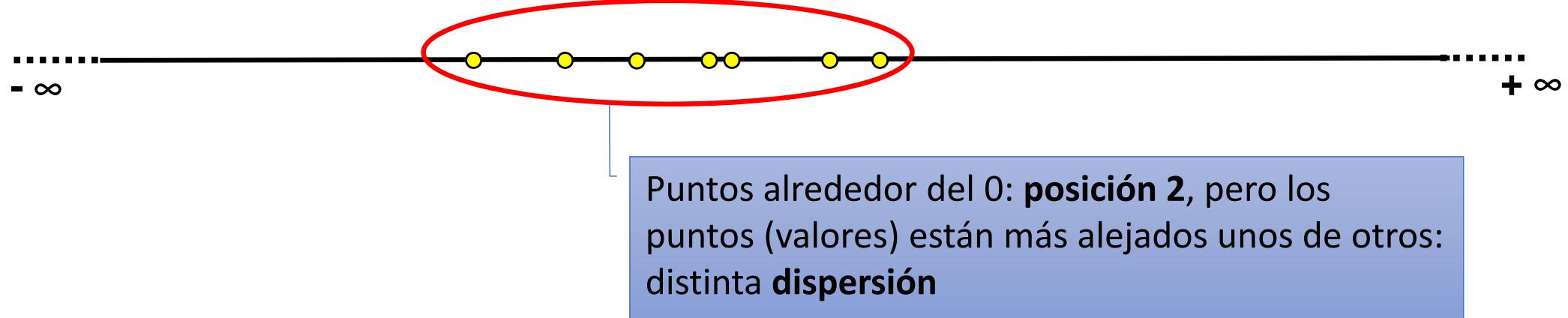
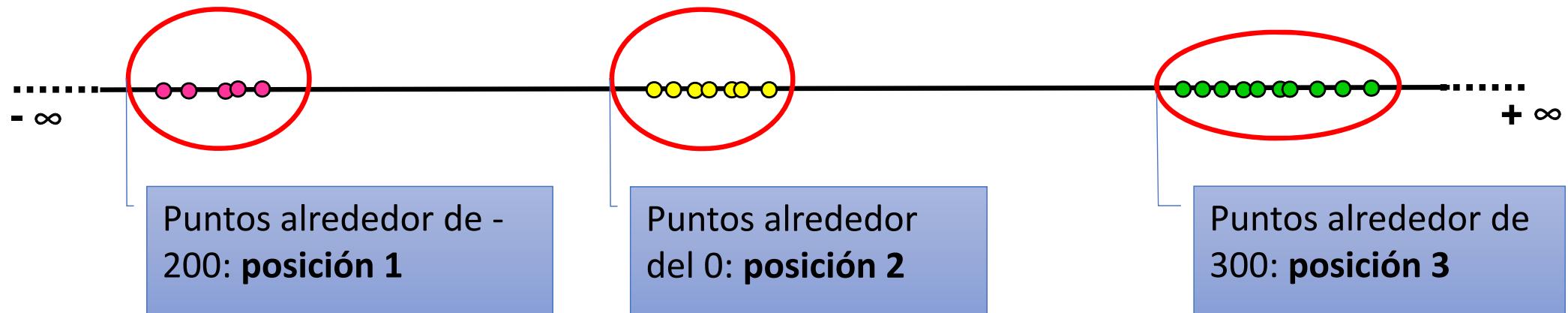
Parámetros



- La pauta de variabilidad de una variable aleatoria unidimensional se caracteriza por tres tipos de parámetros que deben definir:
 - La **Posición** de las observaciones
 - La **Dispersión** de las observaciones
 - La **Forma** de las observaciones



Posición y Dispersion



1.5 – Parámetros de Posición

- Media aritmética
- Mediana
- Moda
- Cuantiles (cuartiles, percentiles)
- ...





Parámetros de posición

- Permiten cuantificar, mediante un número, la **posición** de las observaciones
- con un número nos indican “alrededor” de qué valor están las observaciones.
- Parámetros más relevantes:
 - **media**
 - **mediana**
 - **cuartiles**
 - **moda**

¿Cuál es el más adecuado?



Media

- Es el parámetro de **posición** más utilizado
- Media aritmética de los datos
- Recoge la información existente en la totalidad de los datos:

$$\text{Media} = \bar{x} = \frac{\sum_{i=1}^N x_i}{N}$$

- La media sintetiza la información existente en la totalidad de los datos en un número que da una idea clara sobre la **posición** de los mismos



Ejercicio (Ejercicio 5 UD2)

Con el objeto de determinar la calidad de cierto componente electrónico, se ha tomado una muestra de 11 componentes, midiéndose sus tiempos de funcionamiento sin averías (horas).

Los resultados en horas son los siguientes:

$$\text{v.a. } X = \{50, 38, 45, 30, 47, 50, 48, 62, 55, 53, 52\}$$

El tiempo medio sin averías: $\bar{X} = 48,1818 \text{ h}$



Robustez

- En ocasiones la **media no** es un **buen parámetro de posición**.
- Cuando **tenemos** unos pocos **valores extremos** que pueden influir excesivamente en la media (medida engañosa)

EJERCICIO (Ejercicio 6 UD2): Al preguntar un viajero a un botones de un hotel que propina le daban normalmente, éste respondió que la media de aquel día había sido 10 €. En efecto de los 10 viajeros de aquel día 9 le habían dado 1 € y uno 100 €. La media no era evidentemente en este caso una medida adecuada de la posición de los datos. ¿Cuál considera el alumno que era una medida adecuada de la posición de los datos mencionados?



Mediana (M_e o C_2)

- Es el valor que :
 - deja a su izquierda el 50% de los datos
 - deja a su derecha el 50% de los datos
- Es un indicador **robusto** de posición:
 - si se trabaja con datos muy asimétricos o con algunos valores extremos
 - medida de posición alternativa a la media en estos casos.

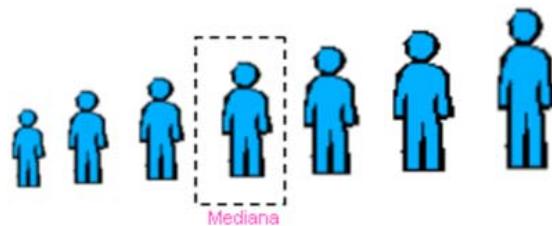


Mediana: Cálculo

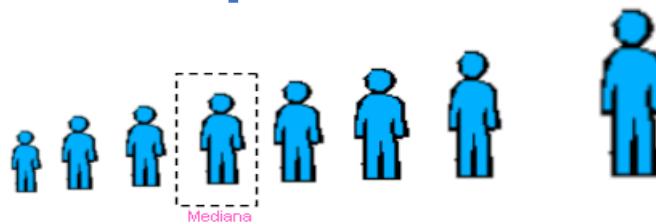
1. Se ordenan las observaciones de menor a mayor.

2. La mediana es el valor que:

→ Ocupe la posición $(N+1)/2$, si **N** es **ímpar**



→ Media entre los valores que ocupan las posiciones $N/2$ y $(N/2)+1$, si **N** es **par**



Ejercicio (Ejercicio 7 UD2)

- Siguiendo con el ejercicio del tiempo de funcionamiento sin averías (TFSA):

$$X = \{50, 38, 45, 30, 47, 50, 48, 62, 55, 53, 52\}$$

- Calcular la mediana



Ejercicio (cont.)

50, 38, 45, 30, 47, 50, 48, 62, 55, 53, 52

N es impar

$$(N+1)/2 = (11+1)/2=6$$

Nº orden	1	2	3	4	5	6	7	8	9	10	11
Tiempo	30	38	45	47	48	50	50	52	53	55	62

mediana

Mediana = M_e = 50 horas



Ejercicio (cont.)

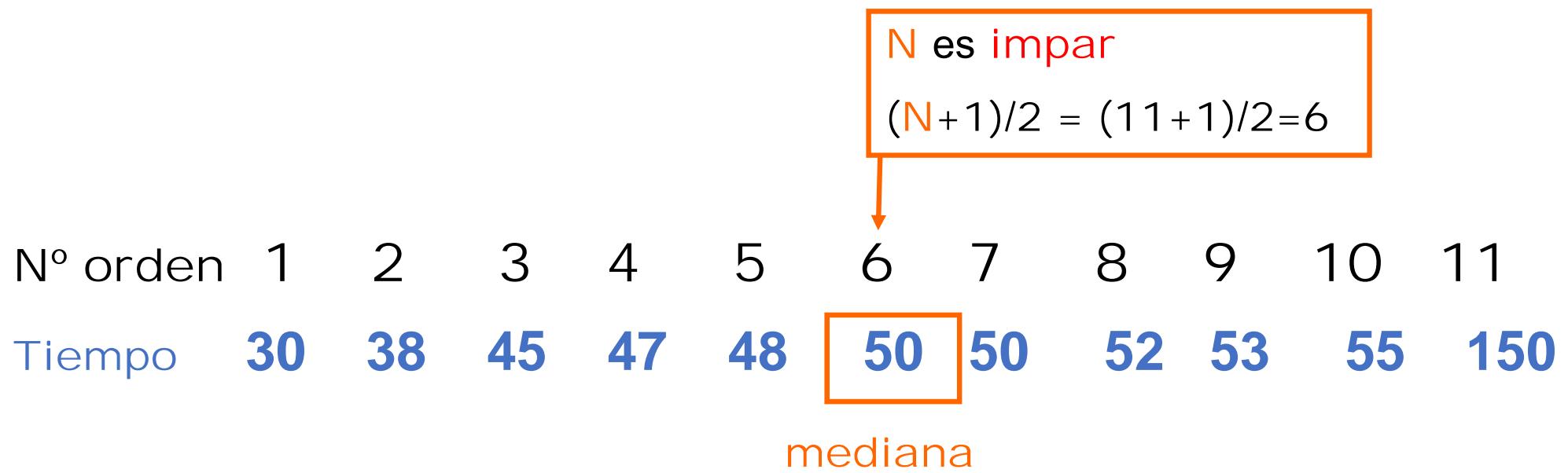
- Suponiendo que uno de los componentes hubiera tenido un TFSA de 150 en lugar de 62 horas (X'), ¿qué valor tendría ahora la media? ¿y la mediana?

$$X = \{50, 38, 45, 30, 47, 50, 48, \textcolor{red}{150}, 55, 53, 52\}$$



Ejercicio (cont.)

50, 38, 45, 30, 47, 50, 48, **150**, 55, 53, 52



Mediana = M_e = **50 horas**

¡La misma que antes!



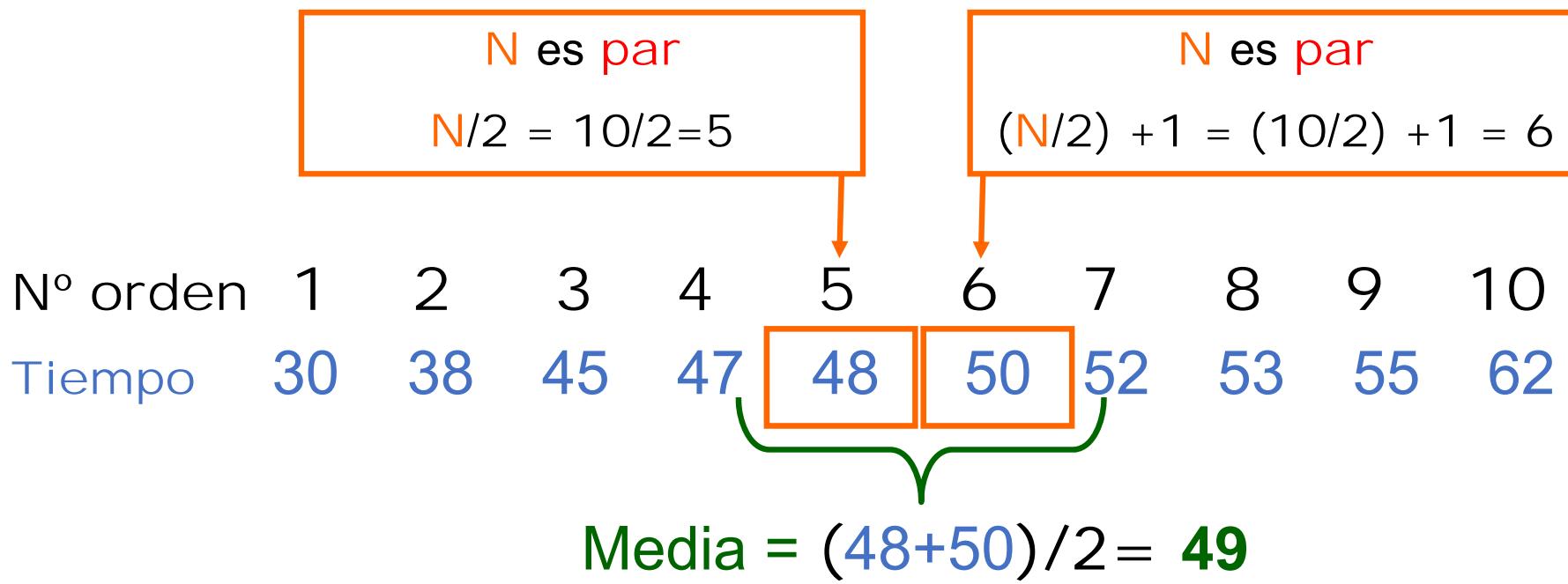
Ejercicio (cont.)

- Si se hubiera tomado una muestra de **10 componentes** para medir el TFSA, obteniendo los siguientes resultados:
38, 45, 30, 47, 48, 50, 62, 55, 53, 52
- ¿qué valor tendría ahora la mediana?



Ejercicio (cont.)

50, 38, 45, 30, 47, 48, 62, 55, 53, 52



Mediana = $M_e = 49$ horas



Otro ejemplo

Summary Statistics

	EDAD	ESTATURA	PESO	TIEMPO
Count	131	131	131	131
Average	21,0458	172,855	66,2137	26,1221
Median	21,0	174,0	66,0	20,0
Variance	2,7825	82,3711	113,569	278,031
Standard deviation	1,66808	9,07585	10,6569	16,6743
Minimum	19,0	152,0	45,0	4,0
Maximum	32,0	198,0	90,0	90,0
Range	13,0	46,0	45,0	86,0
Lower quartile	20,0	165,0	57,0	15,0
Upper quartile	22,0	179,0	74,0	35,0
Stnd. skewness	14,0474	0,916174	0,303203	5,90912
Stnd. kurtosis	34,6426	-0,194872	-1,47758	3,31496



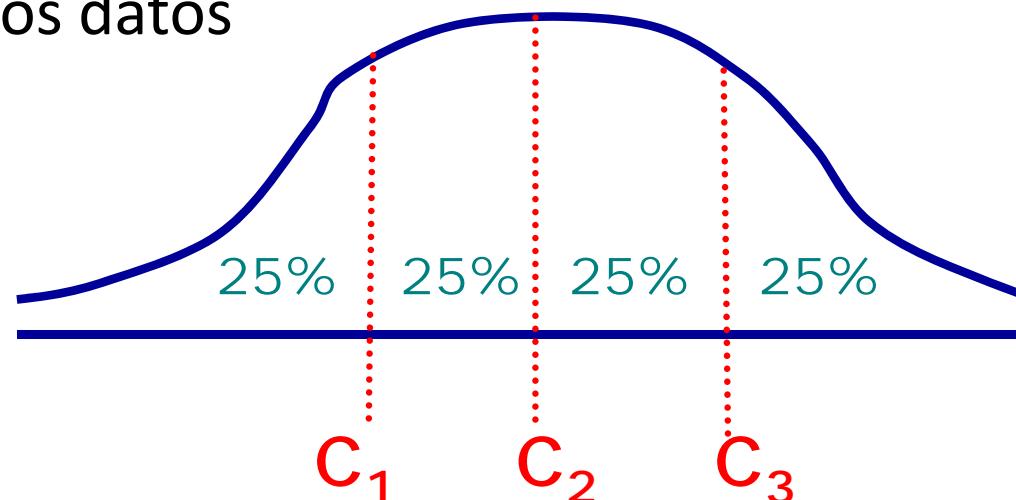
Propiedades de la Mediana

- **Ventaja:** no se ve alterada si una parte de las observaciones contiene errores grandes de medida o transcripción
- **Inconveniente:** utiliza menos información de los datos que la media
- **Conclusión:** se recomienda hallar los valores de **ambas** medidas, ambas difieren bastante si la distribución es muy asimétrica, lo que sugiere heterogeneidad en los datos.



Cuartiles: C_1 , C_2 , C_3

- Caso particular del **cuantil**
- **Curtiles:**
 - **Primer cuartil (C_1)**: el **25%** de los **datos** son **menores o iguales** a éste.
 - **Tercer cuartil (C_3)**: el **75%** de los **datos** son **menores o iguales** a éste.
 - El **segundo cuartil (C_2)** es la **mediana**.
- Entre el Primer (C_1) y el Tercer cuartil (C_3) se encuentra comprendido el **50% central** de los datos

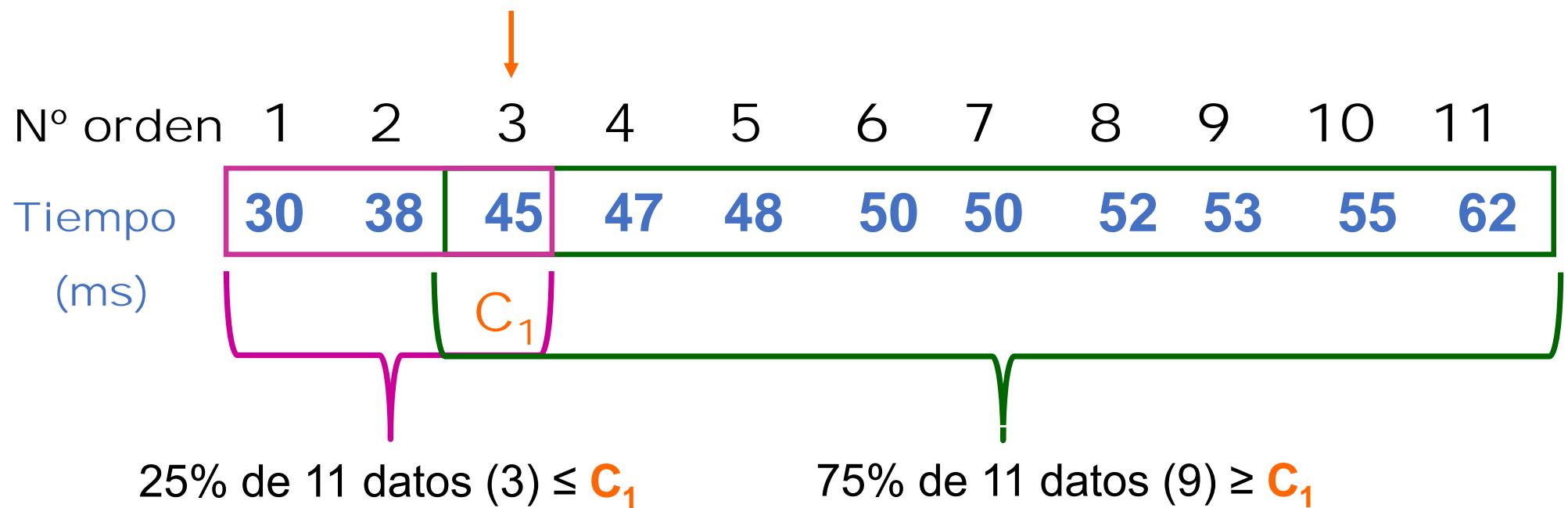


Ejercicio (Ejercicio 11 UD2)

- Siguiendo con el ejercicio del tiempo de funcionamiento sin averías:
50, 38, 45, 30, 47, 50, 48, 62, 55, 53, 52
- Calcular los cuartiles:



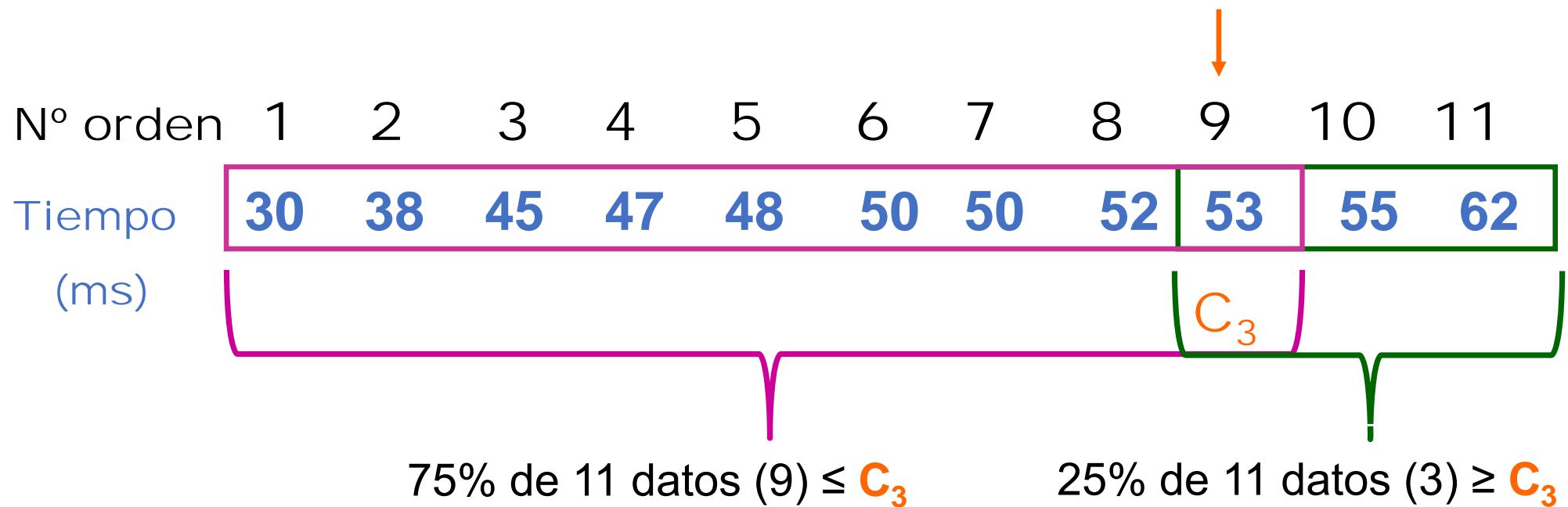
Ejercicio (cont.)



1^{er} cuartil = C_1 = 45 horas



Ejercicio (cont.)



3^{er} cuartil = C_3 = 53 horas



Percentiles

- Procedimiento a seguir para el cálculo del *percentil 85*

1) 3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

2) Calcular el índice **i**

$$i = \left(\frac{p}{100} \right) n = \left(\frac{85}{100} \right) 12 = 10,2$$

p: percentil deseado

n: nº de observaciones

3a) Como **i** no es un nº entero, habrá que redondearlo por exceso

→ **i** denota la posición del percentil → **posición 11**

El *percentil 85* es el dato de la posición 11 → **3730**

¿Qué interpretación tiene este valor?



1.6 – Parámetros de Dispersión

- Rango o Recorrido
- Rango o Recorrido Intercuartílico
- Varianza
- Desviación típica
- Coeficiente de Variación
- ...



Parámetros de dispersión



- Permiten cuantificar, mediante un número, la **dispersión** de las observaciones
- Con un número nos indican lo cerca o lejos que están unas observaciones de otras.
- Parámetros más relevantes:
 - **Recorrido o Rango**
 - **Varianza y Desviación típica**
 - **Coeficiente de Variación**
 - **Recorrido intercuartílico**



7.- Parámetros de Dispersión

- Para describir un conjunto de datos estadísticos no es suficiente con disponer de una medida de su posición → es preciso también cuantificar el grado de dispersión que hay en ellos.

Ejercicio (Ejercicio 13 UD2): Para una persona que no sabe nadar es suficiente saber que la profundidad media de un lago es 1,40 m para lanzarse al baño en el mismo? Por cierto, ¿cuál sería la población y cuál la variable aleatoria en este caso? ¿aclararía mucho la decisión el conocer además la profundidad mediana del lago?



Recorrido o Rango

- Idea intuitiva de dispersión
- **Máximo valor- Mínimo valor**
- Útil en muestras pequeñas ($N \leq 10$)
 - Control estadístico de procesos ($n=5$)
- Ignora gran parte de los datos



Ejercicio (Ejercicios 14 y 15 UD2)

$$X = \{30, 38, 45, 47, 48, 50, 50, 52, 62\}$$

$$R = \text{Max} - \text{Min} = 62 - 30 = 32 \text{ horas}$$

Representativo

$$X' = \{30, 38, 45, 47, 48, 50, 50, 52, 150\}$$

$$R = \text{Max} - \text{Min} = 150 - 30 = 120 \text{ horas}$$

NO
Representativo



Varianza (S^2)

- Dado que la media es, en general un buen parámetro de posición, parece lógico tomar como parámetro de dispersión alguno que esté relacionado con ella:

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

Suma de cuadrados



Nota: Teóricamente, al dividir por $(N-1)$ lo que obtenemos es la cuasivarianza muestral.



Varianza (S^2)

Tiempo	Diferencias	Cuadrados
50	1,82	3,31
38	-10,18	103,67
45	-3,18	10,12
30	-18,18	330,58
47	-1,18	1,40
50	1,82	3,31
48	-0,18	0,03
62	13,82	190,94
55	6,82	46,49
53	4,82	23,21
52	3,82	14,58

Media 48,18 727,64 S. de cuadrados
N 11 72,76 Varianza



Ejercicio (Ejercicio 16 UD2)

Siguiendo con el ejemplo de los tiempos de funcionamiento con y sin averías, la varianza del tiempo sin averías (S_x^2) es:

$$\begin{aligned} S_x^2 = & ((50-48,18)^2 + (38-48,18)^2 + (45-48,18)^2 + (30-48,18)^2 + \\ & + (47-48,18)^2 + (50-48,18)^2 + (48-48,18)^2 + (62-48,18)^2 + \\ & + (55-48,18)^2 + (53-48,18)^2 + (52-48,18)^2) / (11-1) = 72,76 \text{ h}^2 \end{aligned}$$

¡OJO! Las unidades de la varianza están al cuadrado



Desviación Típica o Estándar (S)



- El más utilizado
- La raíz cuadrada de la varianza
- Más fácil de interpretar puesto que viene expresada en las mismas unidades que los datos originales

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



Nota: Las propiedades anteriores NO se cumplen para S.



Ejercicio (Ejercicio 16 UD2 cont.)

Siguiendo con el ejemplo de los tiempos de funcionamiento con y sin averías, calcular:

- la desviación típica del tiempo sin averías (S_x)

$$S_x = \sqrt{72,76} = 8,53 \text{ horas}$$



Coeficiente de Variación

- Indicador de dispersión **adimensional**
- Permite comparar la variabilidad (S , S^2) de variables de naturaleza diferente

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



Coeficiente de Variación

Ejemplo:

$X = \{\text{Tamaño ficheros (Kb)}\} \approx (\text{media}_X = 20 \text{ Kb}; S_X = 10 \text{ Kb})$

$Y = \{\text{Tiempo ejecución (seg)}\} \approx (\text{media}_Y = 180 \text{ seg}; S_Y = 36 \text{ seg})$

¿ Qué variable tiene mayor VARIABILIDAD ?

$(S_Y = 36 \text{ seg}) >> (S_X = 10 \text{ Kb})$ ¡No se deben comparar!

$$[CV_Y = 36/180 = 0,2] << [CV_X = 10/20 = 0,5]$$

Mayor dispersión



Intervalo Intercuartílico (II)

- Se calcula como la diferencia entre el 3er y 1er cuartil:

$$II = C_3 - C_1$$

- En distribuciones asimétricas o con datos extremos, la S^2 no es un buen indicador de la dispersión de los datos →
- Necesitamos un indicador robusto de dispersión → II



Ejercicio (Ejercicio 17 UD2)

$$X = \{ 50, 38, 45, 30, 47, 50, 48, 62, 55, 53, 52 \}$$

$$S_x = 8,53 \text{ h} \approx II_x = C_3 - C_1 = 53 - 45 = 8 \text{ h}$$

$$X' = \{ 50, 38, 45, 30, 47, 50, 48, 150, 55, 53, 52 \}$$

$$S_{x'} = 31,94 \text{ h} \quad II_{x'} = C_3 - C_1 = 53 - 45 = 8 \text{ h}$$

$$S_x \ll S_{x'} \quad II_x = II_{x'}$$



Parámetros más adecuados

- Del mismo modo que preferimos la **mediana** a la **media** cuando los datos son **asimétricos** o presentan valores **anómalos**, el **Recorrido Intercuartílico** es más adecuado que las **desviación típica** en esas mismas situaciones.

	Datos simétricos y sin valores anómalos	Datos asimétricos o con valores anómalos
Posición	Media	Mediana
Dispersión	Desviación típica	Recorrido intercuartílico

¡¡Cuidado con las unidades!!!



Ejemplo: ESTATURA de los alumnos de la UPV

Valores en cm

Summary Statistics for ESTATURA

Count = 42

Average = 163,429

Median = 163,0

Mode =

Variance = 32,2021

Standard deviation = 5,67469

Lower quartile = 160,0

Upper quartile = 165,0

Interquartile range = 5,0

Dato en metros

Summary Statistics for ESTATURA

Count = 42

Average = 159,681

Median = 163,0

Mode =

Variance = 656,157

Standard deviation = 25,6156

Lower quartile = 160,0

Upper quartile = 165,0

Interquartile range = 5,0



1.7 – Parámetros de forma

- Asimetría
- Curtosis



7 - Introducción

- Los coeficientes de **Asimetría** y **Curtosis** son **parámetros de forma**.
- Los dos permiten comprobar si nuestros datos se asemejan suficientemente a una “campana de Gauss” (distribución Normal)



Pautas de comportamiento que se alejan sensiblemente de la Normal exigen:

- la revisión y corrección de datos anómalos, si procede
- recurrir a modelos o tratamientos estadísticos especiales.



Mediante los parámetros de Asimetría y Curtosis podemos detectar la “no normalidad” de los datos y obrar en consecuencia



Coeficiente de Asimetría (Skewness)

- Cálculo:

$$CA = \frac{\sum (X_i - \bar{X})^3 / (N - 1)}{S^3}$$

- En contextos inferenciales se utiliza el **Coeficiente de Asimetría Estandarizado (CAE)**

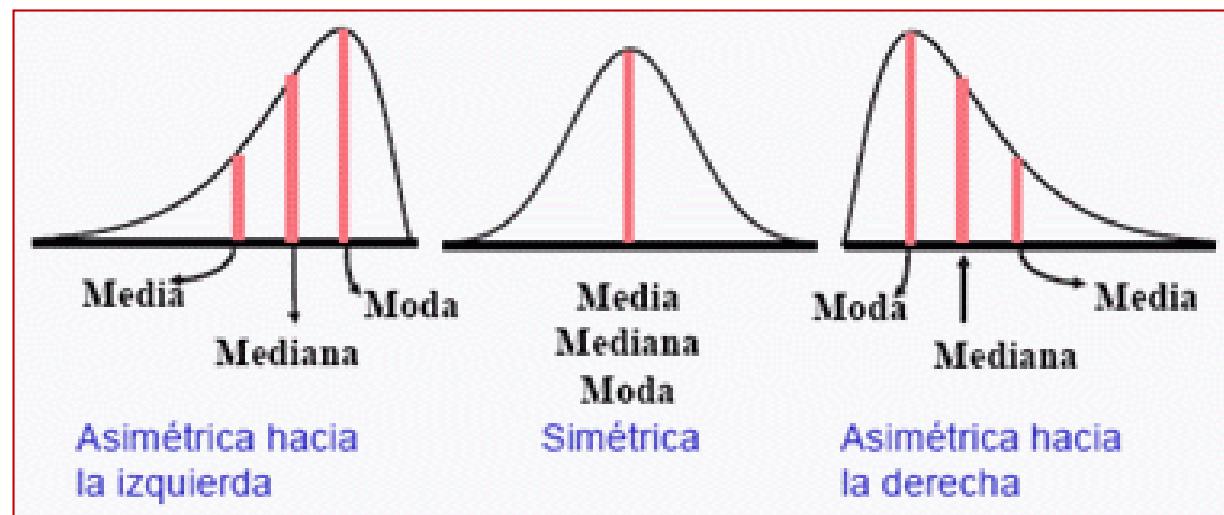
CAE $\in [-2, 2]$ \rightarrow Datos simétricos.

CAE > 2 \rightarrow Asimetría positiva (cola por la derecha)

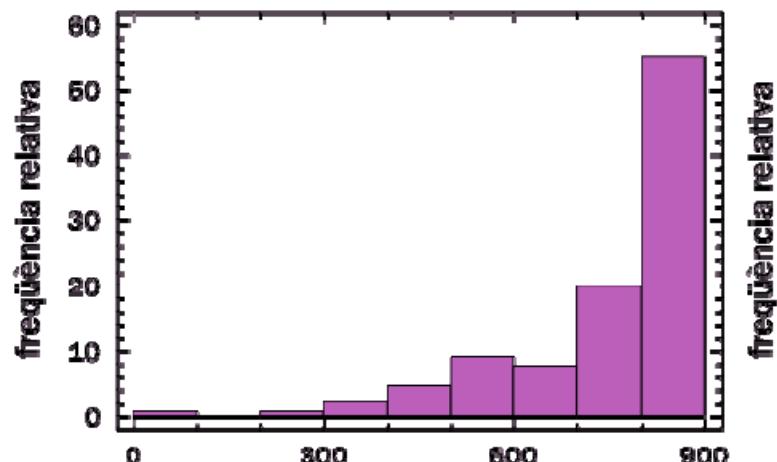
CAE < -2 \rightarrow Asimetría negativa (cola por la izquierda)



Coeficiente de Asimetría (Skewness)

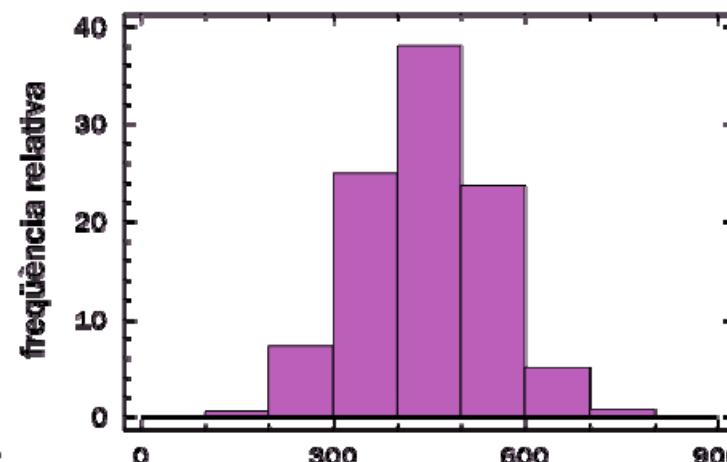


Asimetria negativa



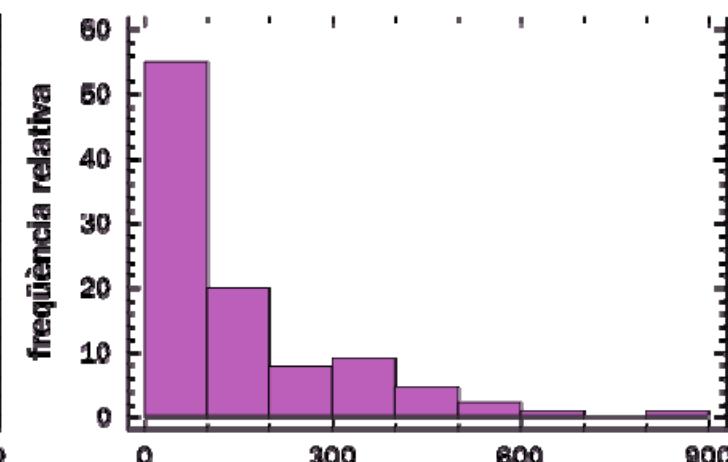
$CAE < -2$

Dades simètriques



$CAE \in [-2, +2]$

Asimetria positiva

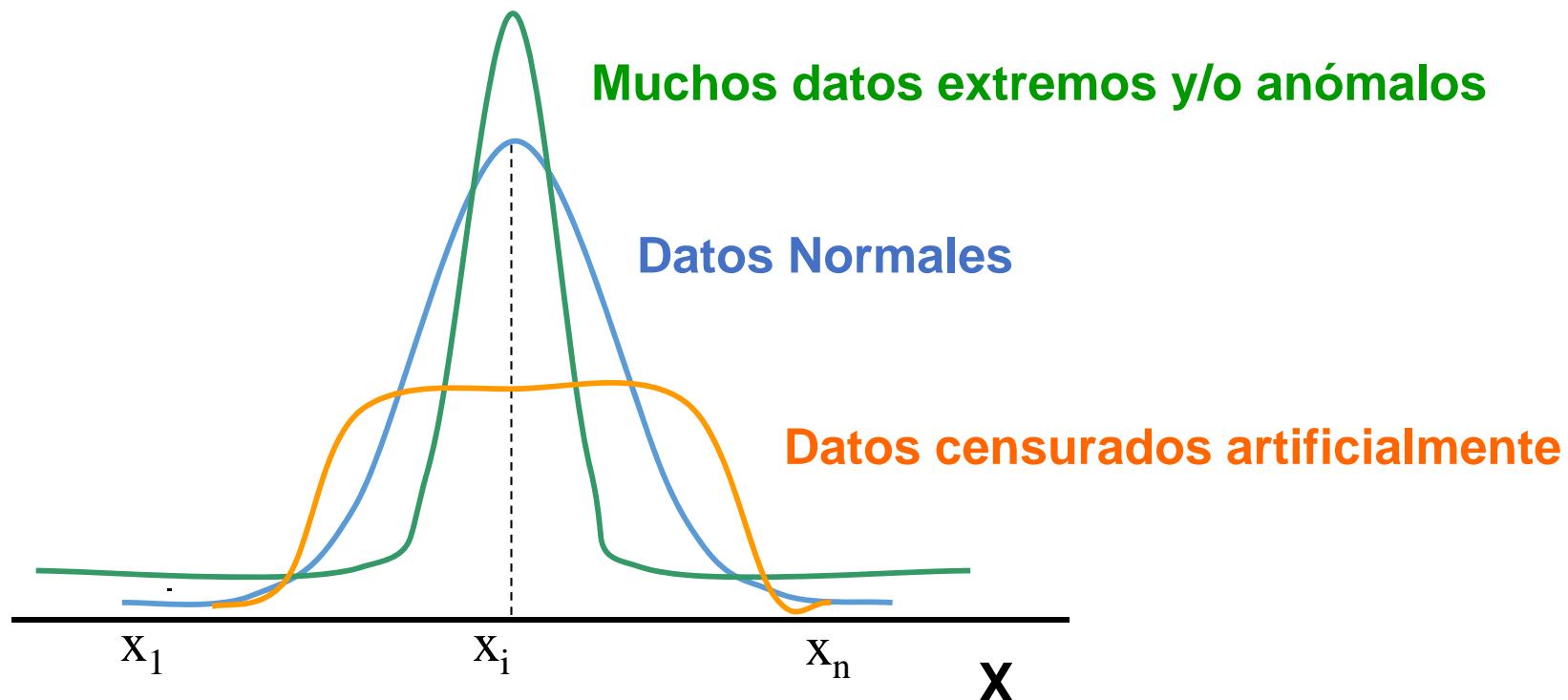


$CAE > +2$



Coeficiente de Curtosis

- El coeficiente de curtosis mide como de “puntiaguda” o “aplatanada” es la forma de la distribución.
- La referencia es la “campana de Gauss”.
- Según este coeficiente podemos tener una idea sobre los datos:



Coeficiente de Curtosis (Kurtosis)

- Cálculo:

$$C_C = \frac{\sum (X_i - \bar{X})^4 / (N - 1)}{S^4}$$

- En contextos inferenciales se utiliza el **Coeficiente de Curtosis Estandarizado (CCE)**

CCE $\in [-2, 2]$ \rightarrow Datos “normales”.

CCE > 2 \rightarrow más “apuntados” de lo normal

CCE < -2 \rightarrow datos “aplanados”



Normalidad

- Desde un punto de vista meramente descriptivo, podemos comprobar si nuestros datos se asemejan suficientemente a una “campana de Gauss” (distribución Normal) cuando

- Son simétricos → $CA = 0$ ó Std. Skewness $\in [-2, 2]$

y

- Son mesocúrticos → $CC = 3$ ó $CC = 0$ ó Std. Kurtosis $\in [-2, 2]$

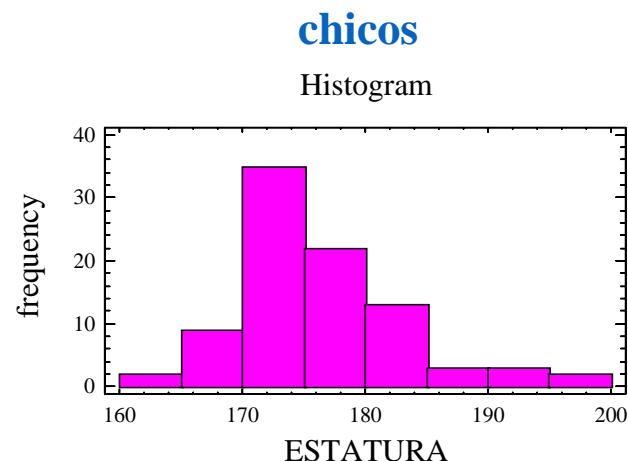


- Además podemos estudiar el histograma o el diagrama Box & Whisker y comparar los parámetros de posición y dispersión.



EJERCICIO (Ejercicio 19 UD2): Calcular los coeficientes de asimetría y curtosis de la ESTATURA en chicos y chicas y comparar los resultados obtenidos.

Obtener también dichos coeficientes para la variable TIEMPO.



$$CA = 0,878671$$

$$CA\text{-estandarizado} = 3,38412$$

$$CC = 0,979509$$

$$CC\text{-estandarizado} = 1,88624$$

CA-estandarizado > 2 Asimetría positiva

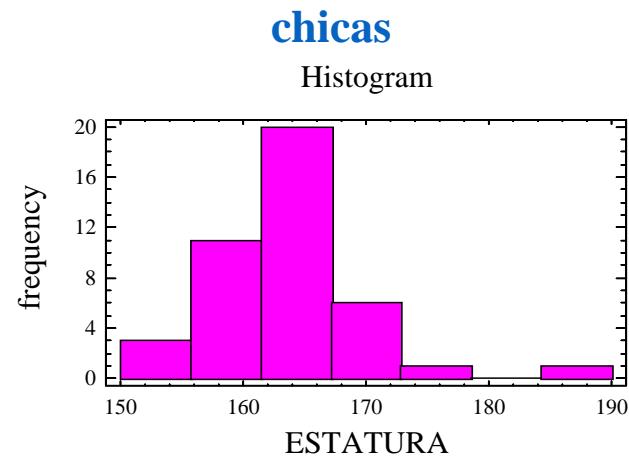
CC-estandarizado $\in [-2, 2]$ Datos normales

No Distribución Normal



EJERCICIO (Cont.): Estatura Chicas

No Distribución Normal



$$CA = 1,29119$$

$$\text{CA-estandarizado} = 3,41616$$

$$CC = 4,30461$$

$$\text{CC-estandarizado} = 5,69446$$

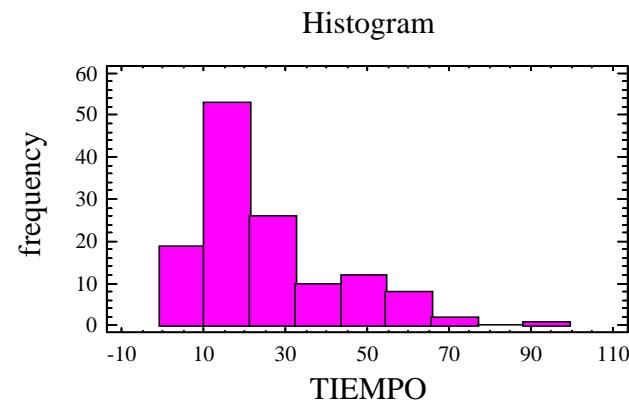
CA-estandarizado > 2 Asimetría positiva

CC-estandarizado > 2 Datos anómalos



EJERCICIO (Cont.): Tiempo en llegar UPV

No Distribución Normal



$$CA = 1,26463$$

$$\text{CA-estandarizado} = 5,90912$$

$$CC = 1,41889$$

$$\text{CC-estandarizado} = 3,31496$$

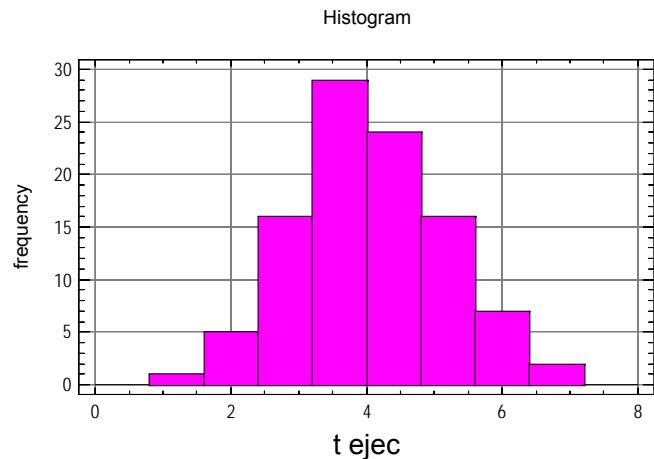
CA-estandarizado > 2 Asimetría positiva

CC-estandarizado > 2 Datos extremos



EJERCICIO (Cont.): Tiempo de ejecución

Distribución Normal



$$CA = 0,143977$$

$$\text{CA-estandarizado} = 0,587782$$

$$CC = -0,193152$$

$$\text{CC-estandarizado} = -0,39427$$

CA-estandarizado $\in [-2, 2]$ Simetría

CC-estandarizado $\in [-2, 2]$ Datos normales



1.8 – Diagrama Box & Whisker

O Diagrama de *Caja y Bigotes*

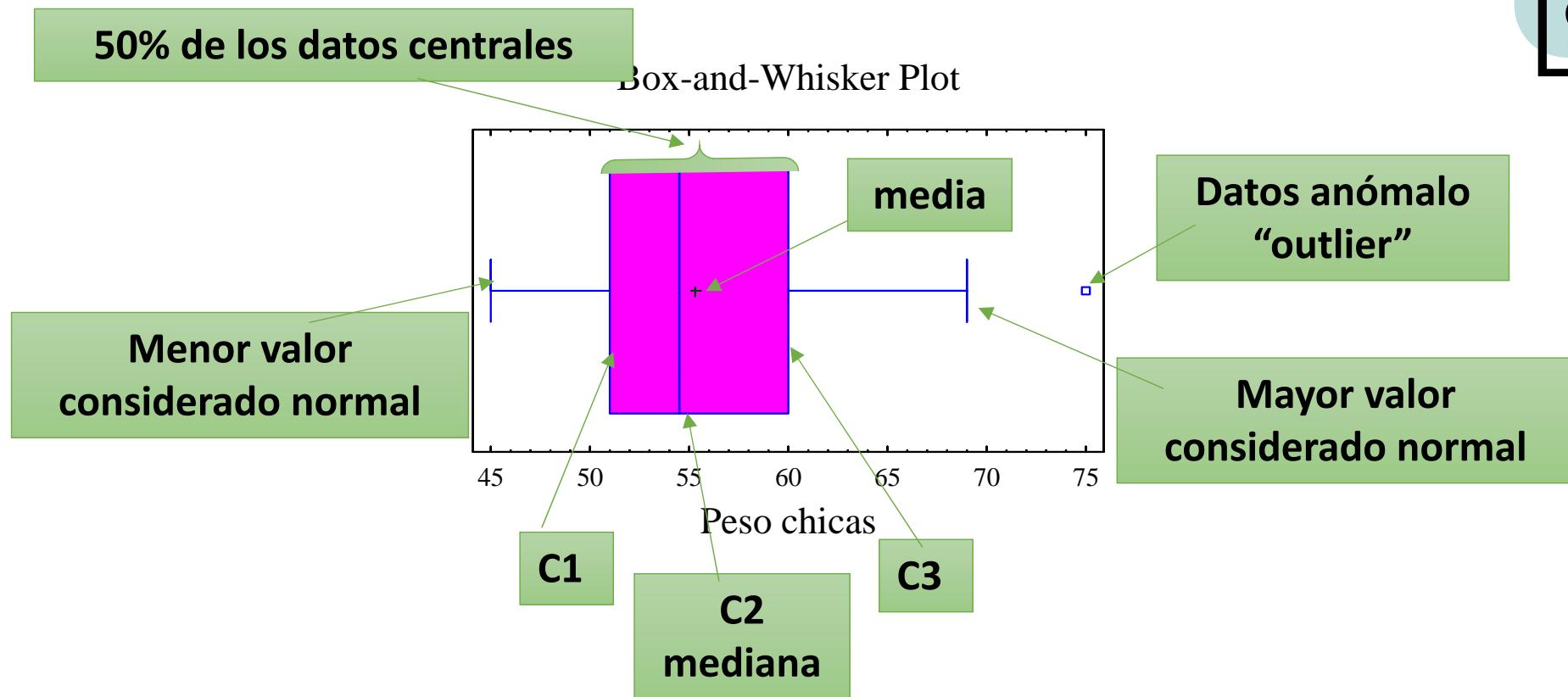


9.- Diagramas Box-Whisker

- No exige un número elevado de datos para su construcción como el Histograma
- Muy útil para comparar 2 grupos de datos y observar de forma gráfica si hay o no diferencias entre ellos.
- La “caja” comprende el 50% de los valores centrales de los datos y se extiende entre el 1º y 3º cuartil



Interpretación

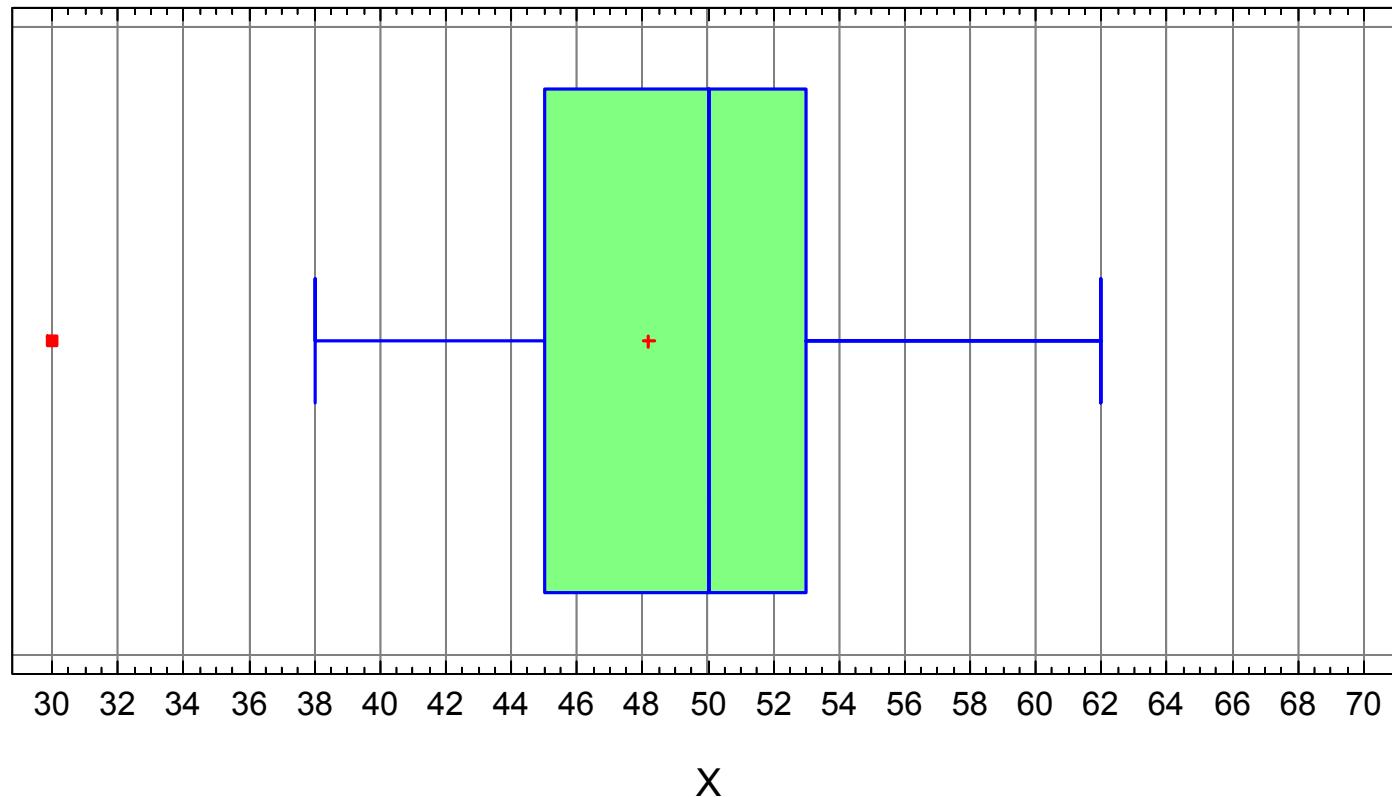


Dato anómalo (“outlier”): Valores extremos que difieren del cuartil más próximo en más de 1,5 veces el intervalo intercuartílico (C3-C1)



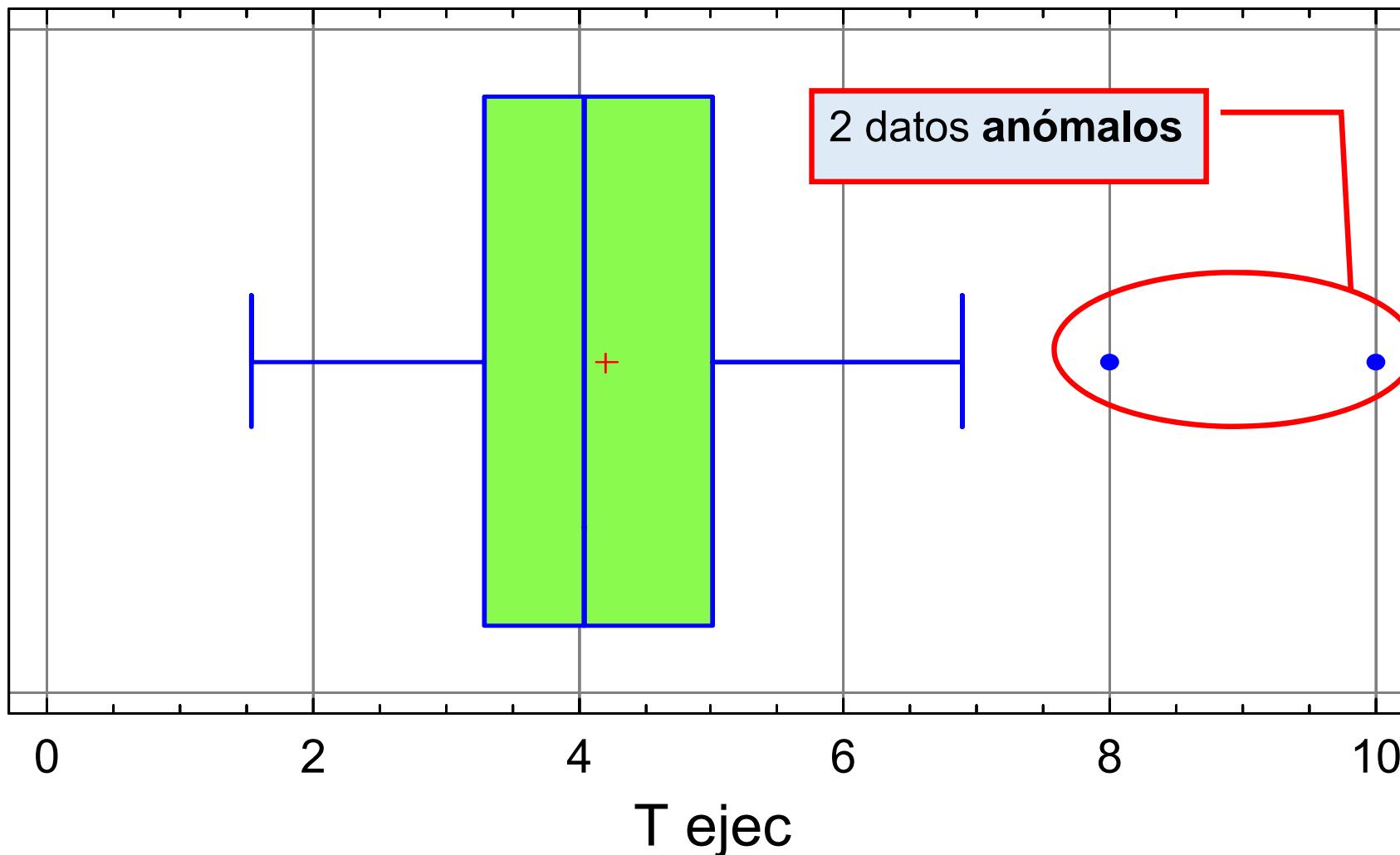
Ejercicio (UD2)

- Siguiendo con el ejercicio del tiempo de funcionamiento sin averías: X
 $= \{50, 38, 45, 30, 47, 50, 48, 62, 55, 53, 52\}$

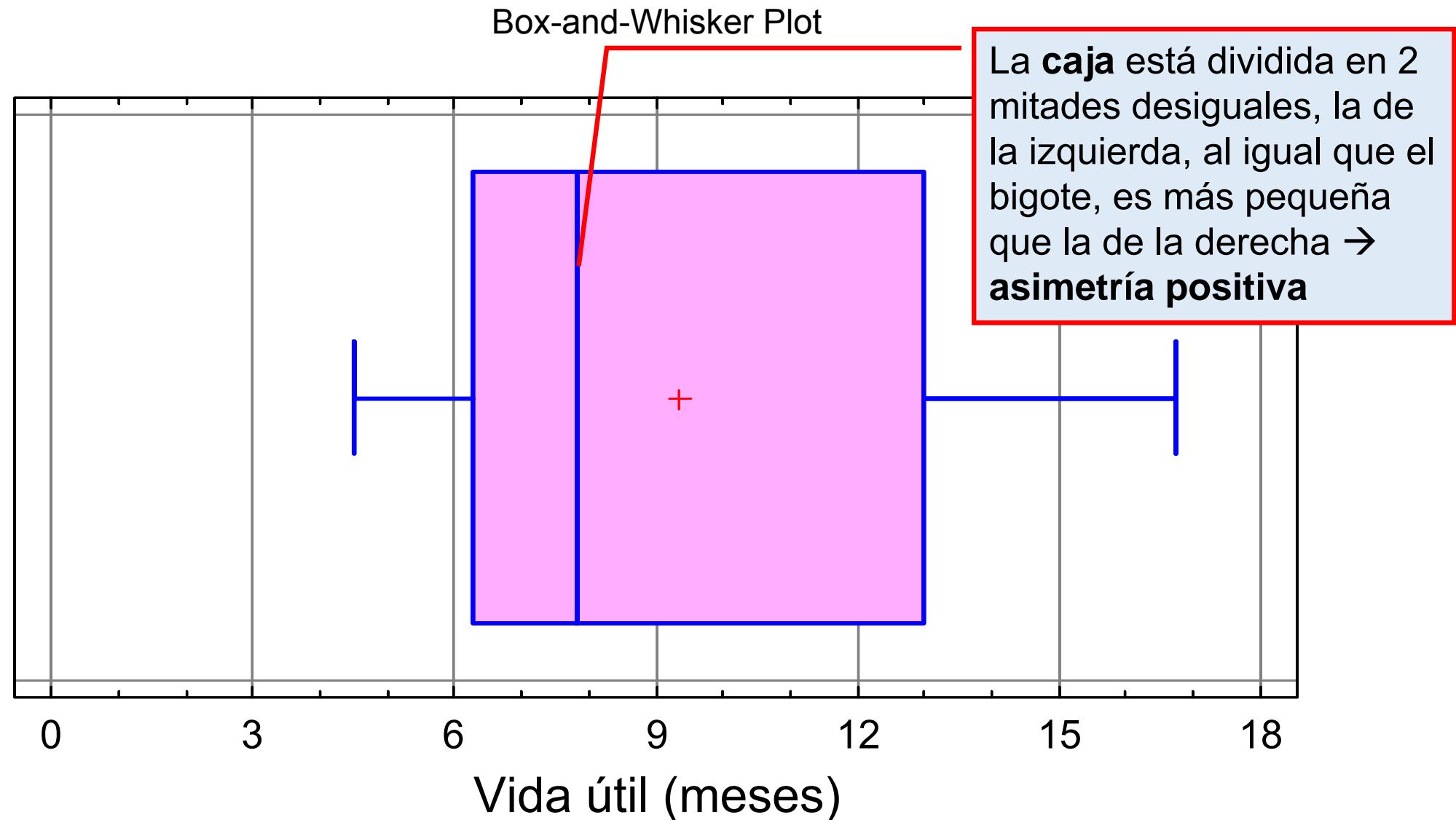


Otro ejemplo: datos anómalos

Box-and-Whisker Plot

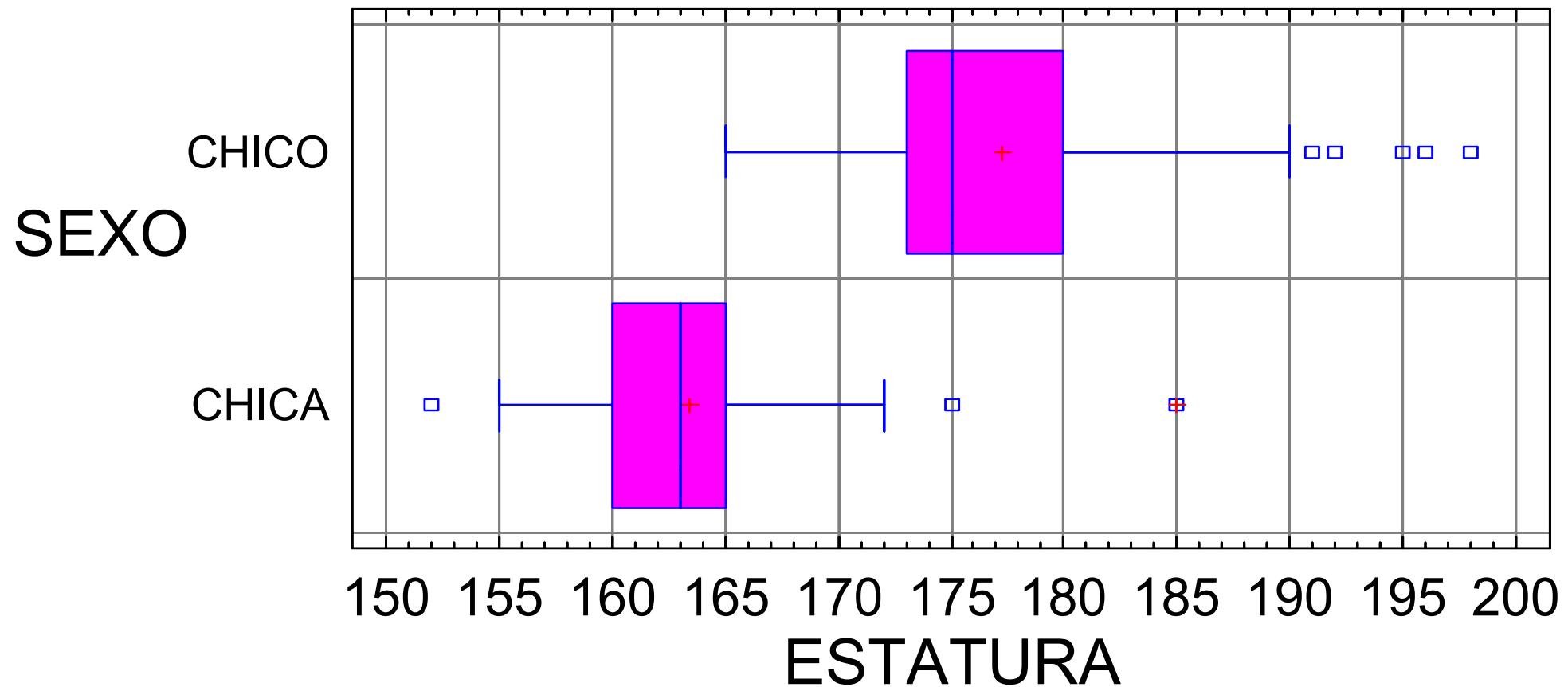


Otro ejemplo: asimetría positiva

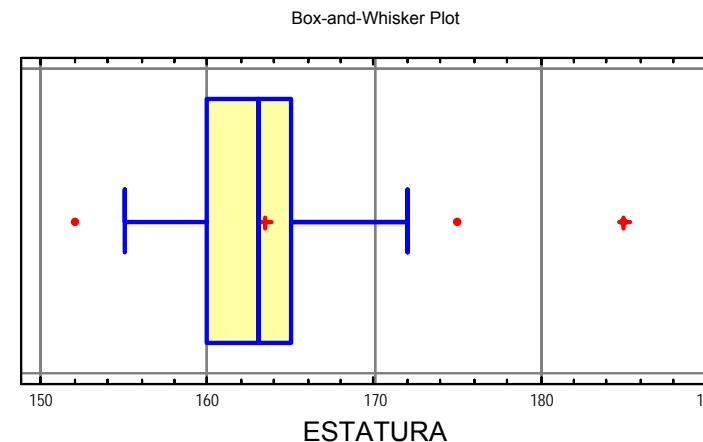
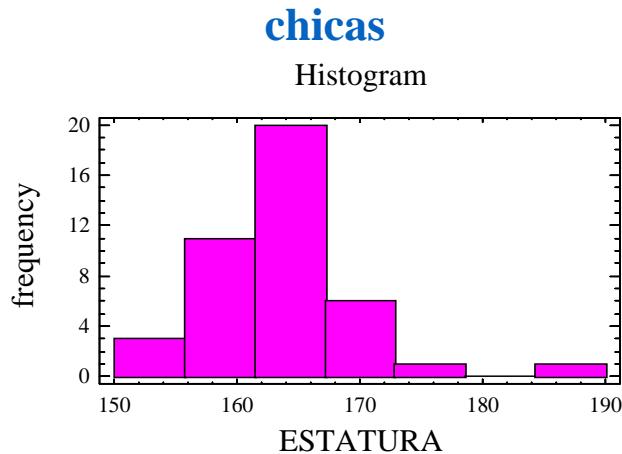
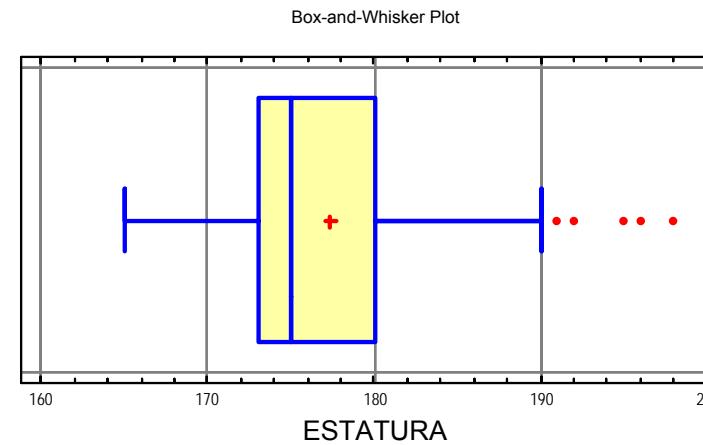
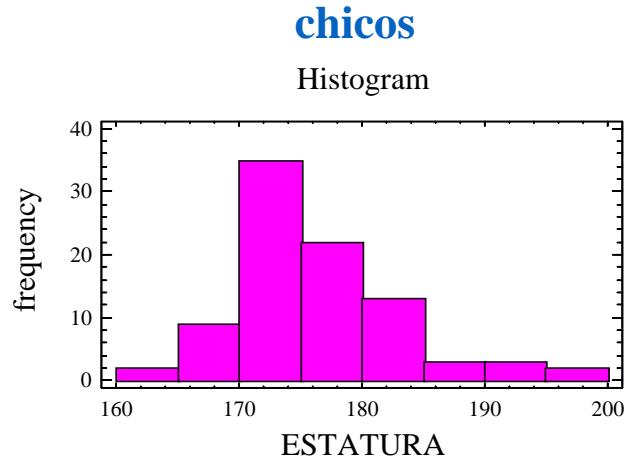


Ejercicio 20 UD2

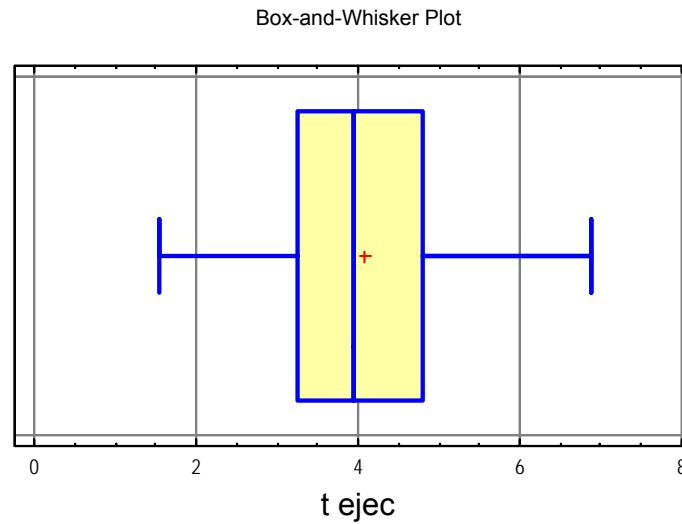
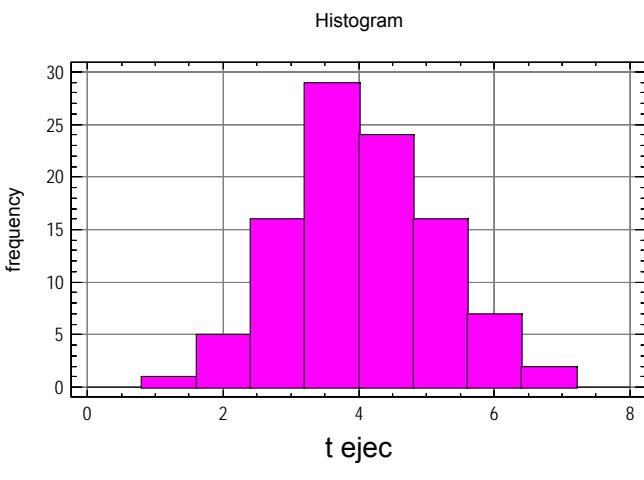
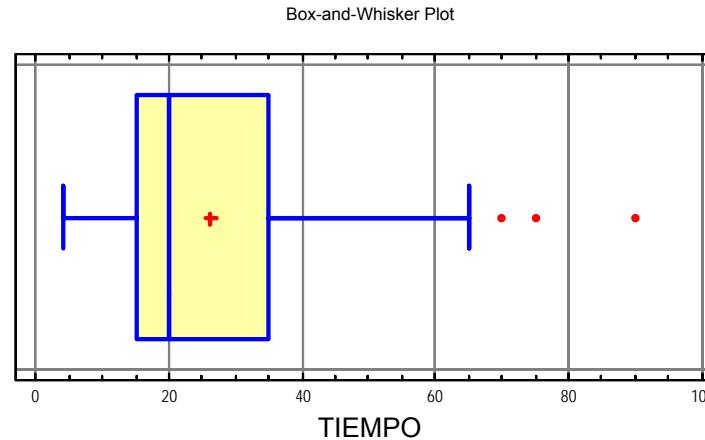
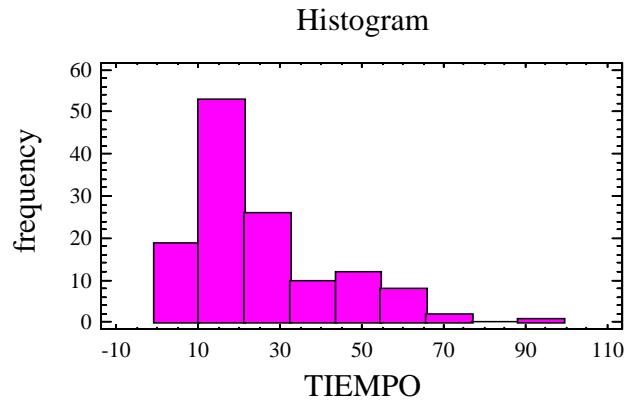
Box-and-Whisker Plot



Box&Whisker e Histogramas



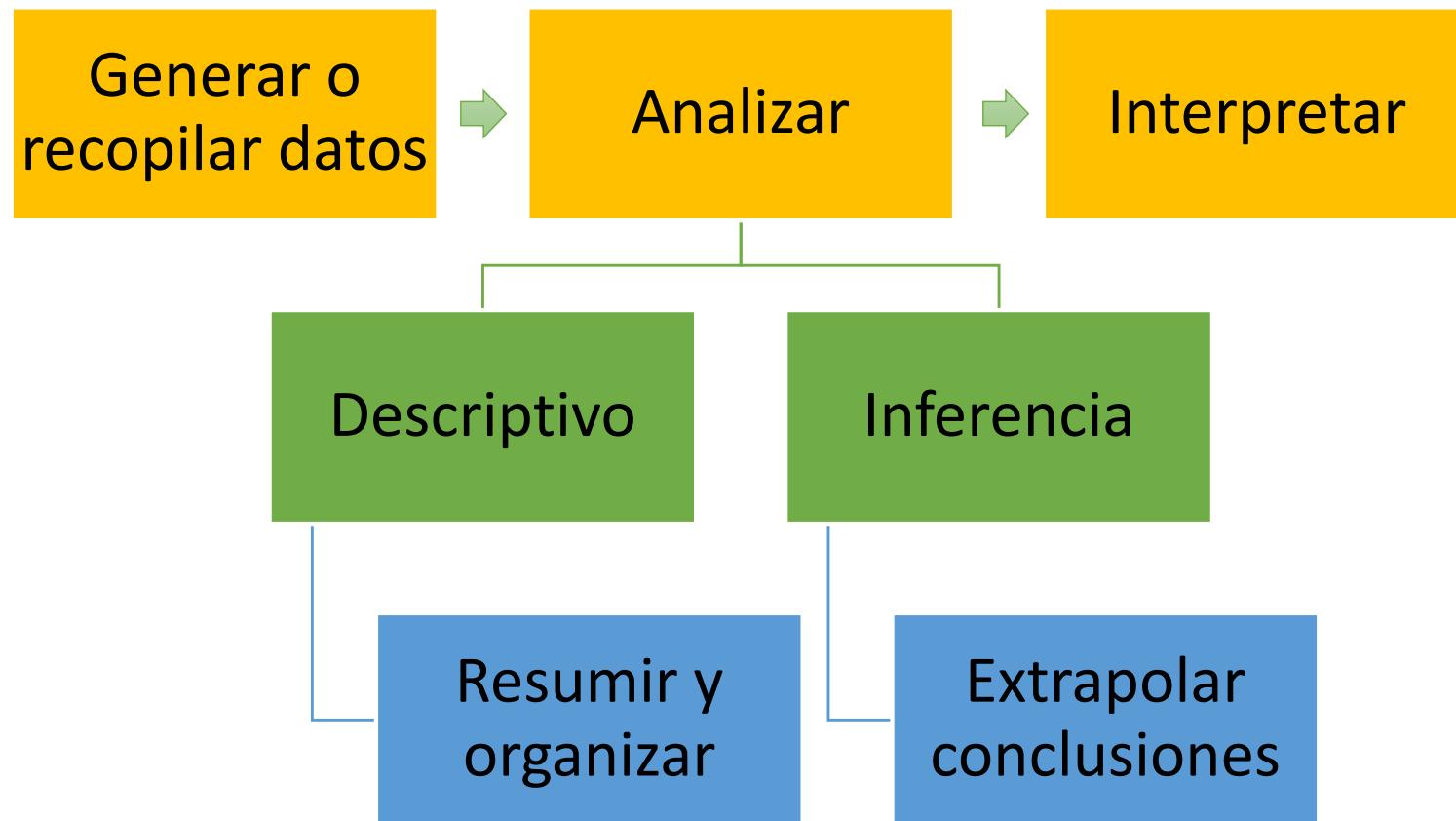
Box&Whisker e Histogramas



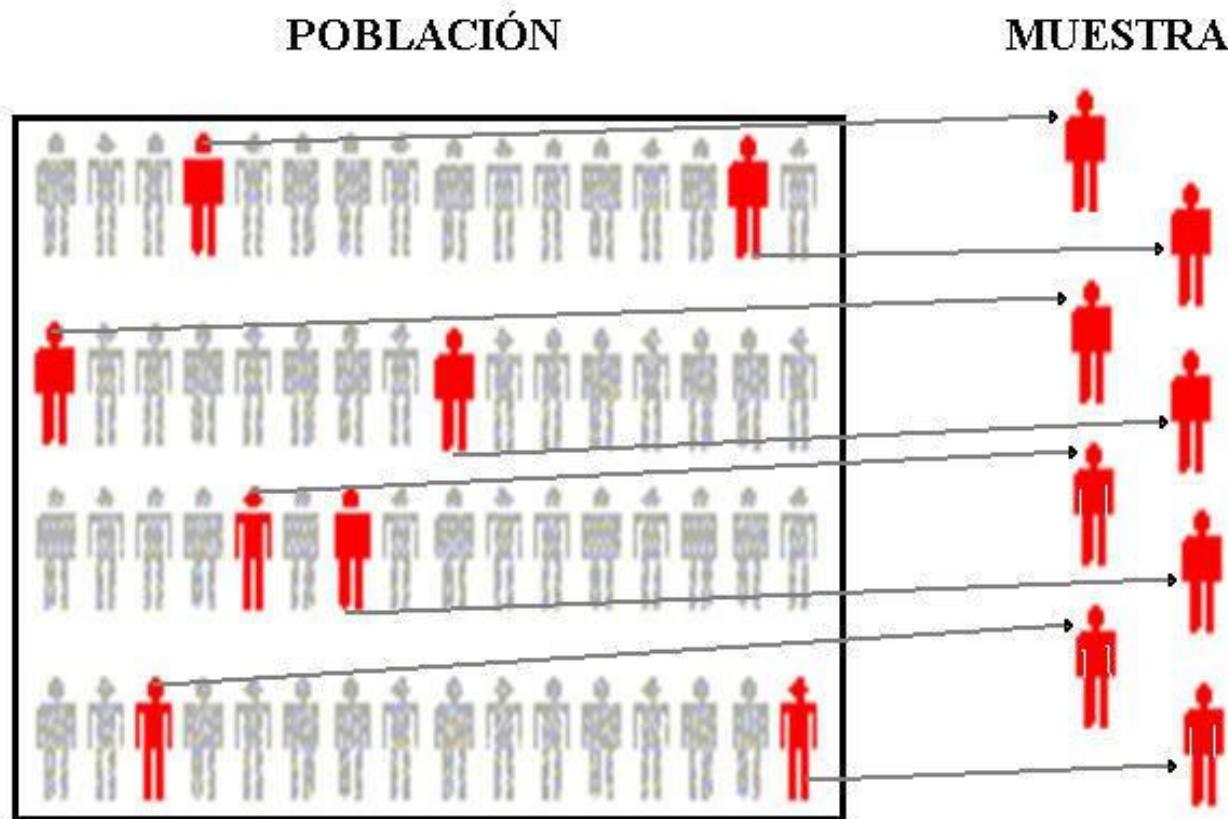
Resumen



Objetivos Estadística



Población y muestra



Datos estadísticos

STATGRAPHICS Plus - Sample1.sgp - [CARDATA.SF]

File Edit Plot Describe Compare Relate Special View Window Help

mpg cylinders displace horsepower accel year weight origin make model

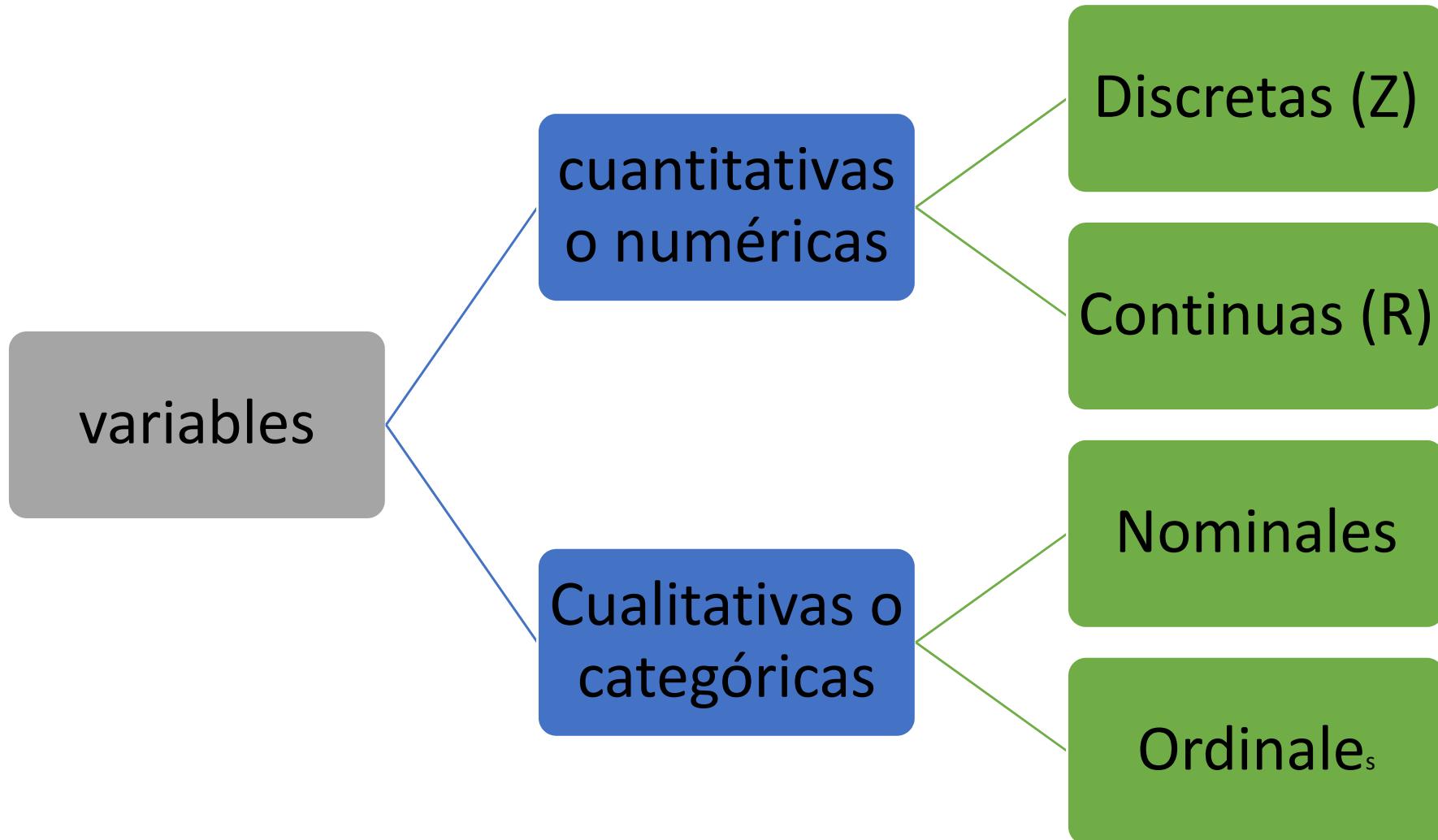
	mpg	cylinders	displace	horsepower	accel	year	weight	origin	make	model
1	43,1	4	90	48	21,5	78	1985	2	Volkswagen	Rabbit
2	36,1	4	98	66	14,4	78	1800	1	Ford	Fiesta
3	32,8	4	78	52	19,4	78	1985	3	Mazda	GLC
4	39,4	4	85	70	18,6	78	2070	3	Datsun	B210
5	36,1	4	91	60	16,4	78	1800	3	Honda	Civic
6	19,9	8	260	110	15,5	78	3365	1	Oldsmobile	Cutlass
7	19,4	8	318	140	13,2	78	3735	1	Dodge	Diplomat
8	20,2	8	302	139	12,8	78	3570	1	Mercury	Monarch
9	19,2	6	231	105	19,2	78	3535	1	Pontiac	Phoenix
10	20,5	6	200	95	18,2	78	3155	1	Chevrolet	Malibu
11	20,2	6	200	85	15,8	78	2965	1	Ford	Fairmont
12	25,1	4	140	88	15,4	78	2720	1	Ford	Fairmont
13	20,5	6	225	100	17,2	78	3430			
14	19,4	6	232	90	17,2	78	3210			
15	20,6	6	231	105	15,8	78	3380			
16	20,8	6	200	85	16,7	78	3070	1	Mercury	Zephyr
17	18,6	6	225	110	18,7	78	3620	1	Dodge	Aspen
18	18,1	6	258	120	15,1	78	3410	1	AMC	Concord
19	19,2	8	305	145	13,2	78	3425	1	Chevrolet	Monte Carlo
20	17,7	6	231	165	13,4	78	3445	1	Buick	Regal
21	18,1	8	302	139	11,2	78	3205	1	Ford	Futura
22	17,5	8	318	140	13,7	78	4080	1	Dodge	Magnus
23	30	4	98	68	16,5	78	2155	1	Chevrolet	Chev.
24	27,5	4	134	95	14,2	78	2560	3	Toyota	Corolla
25	27,2	4	119	97	14,7	78	2300	3	Datsun	510
26	30,9	4	105	75	14,5	78	2230	1	Dodge	Omni
27	21,1	4	134	95	14,8	78	2515	3	Toyota	Celica
28	23,2	4	156	105	16,7	78	2745	1	Plymouth	Sappo
29	23,8	4	151	85	17,6	78	2855	1	Oldsmobile	Starfire
30	23,9	4	119	97	14,9	78	2405	3	Datsun	200-S
31	20,3	5	131	103	15,9	78	2830	2	Audi	5000

Individuo muestra

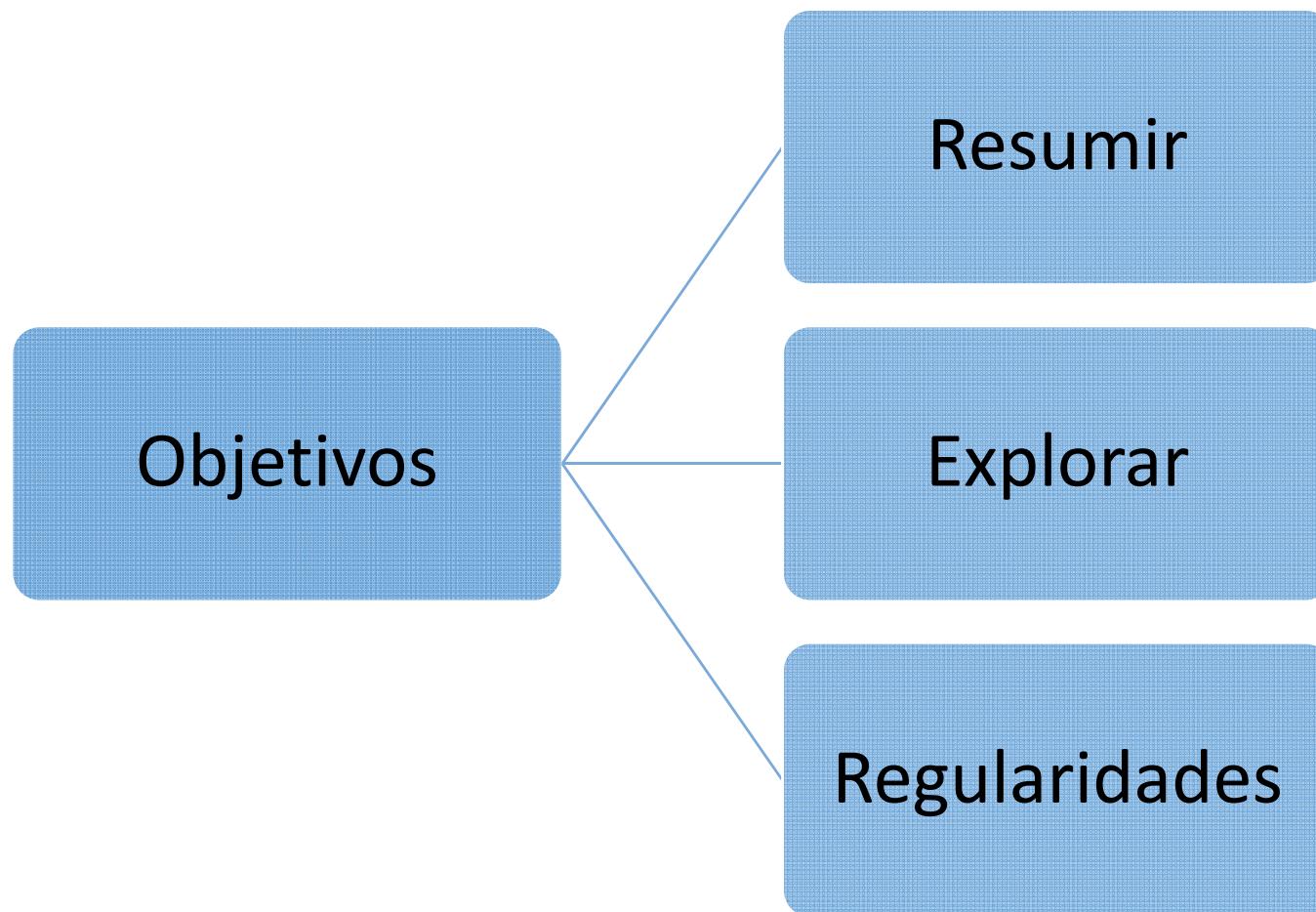
Variable



Variables aleatorias



Estadística descriptiva



Análisis descriptivo

Frecuencias:

- Absolutas
- Relativas
- Acumuladas

Tablas

Frecuencias cruzadas:

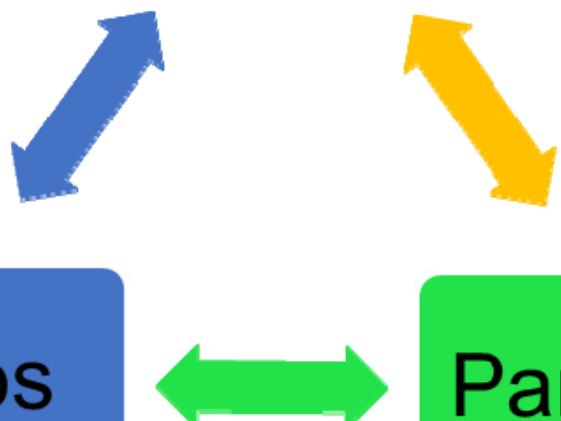
- Marginales
- condicionales

Gráficos

- Barras
- Histograma
- Sectores
- Boxplot
- ...

Parámetros

- Posición
- Dispersión
- Forma
- Relación



Unidimensional

Bidimensional

K dimensional



Técnicas básicas de estadística descriptiva unidimensional

Variable		Estadísticos	Gráficos
Tipo	Subtipo		
Cualitativa/Discreta/Categórica	Nominal	Frecuencias	Diagrama de barras Diagrama de sectores
Cualitativa/Discreta/Categórica	Ordinal	Frecuencias Moda, mediana Rango, cuartiles	Diagrama de barras Box-plot
Cuantitativa/Continua		Media, mediana, moda Varianza, desviación típica Coeficiente de variación Percentiles ...	Histograma Box-plot



Parámetros más adecuados

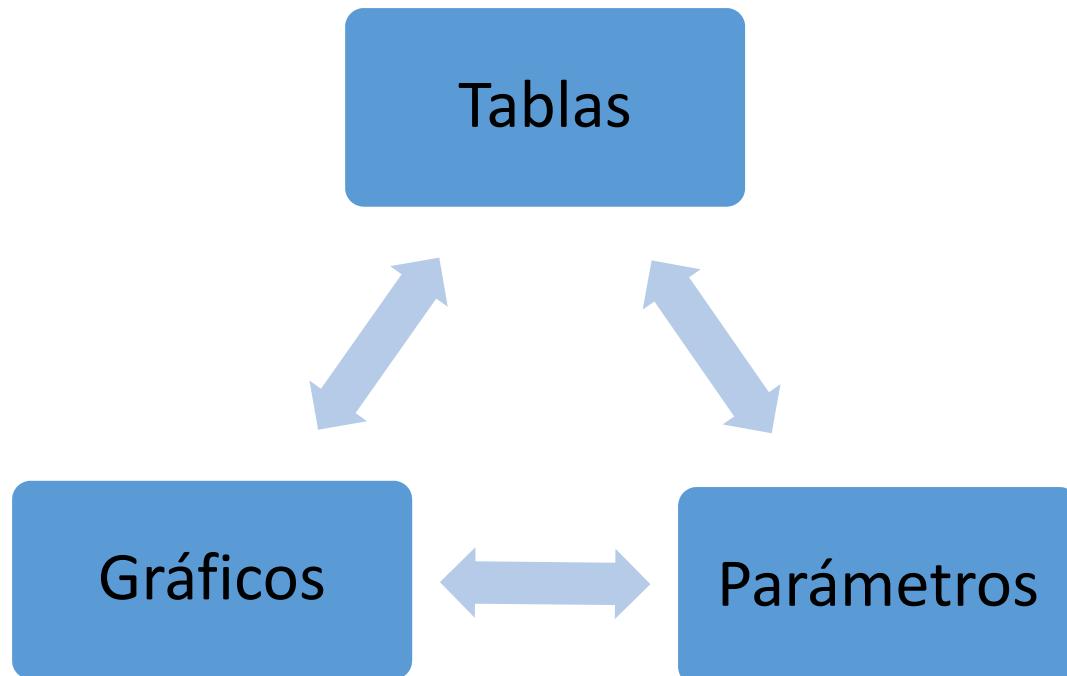
- Del mismo modo que preferimos la **mediana** a la **media** cuando los datos son **asimétricos** o presentan valores **anómalos**, el **Recorrido Intercuartílico** es más adecuado que las **desviación típica** en esas mismas situaciones.

	Datos simétricos y sin valores anómalos	Datos asimétricos o con valores anómalos
Posición	Media	Mediana
Dispersión	Desviación típica	Recorrido intercuartílico

¡¡Cuidado con las unidades!!!



Herramientas estadística descriptiva



1. ¿Qué herramientas son adecuadas?

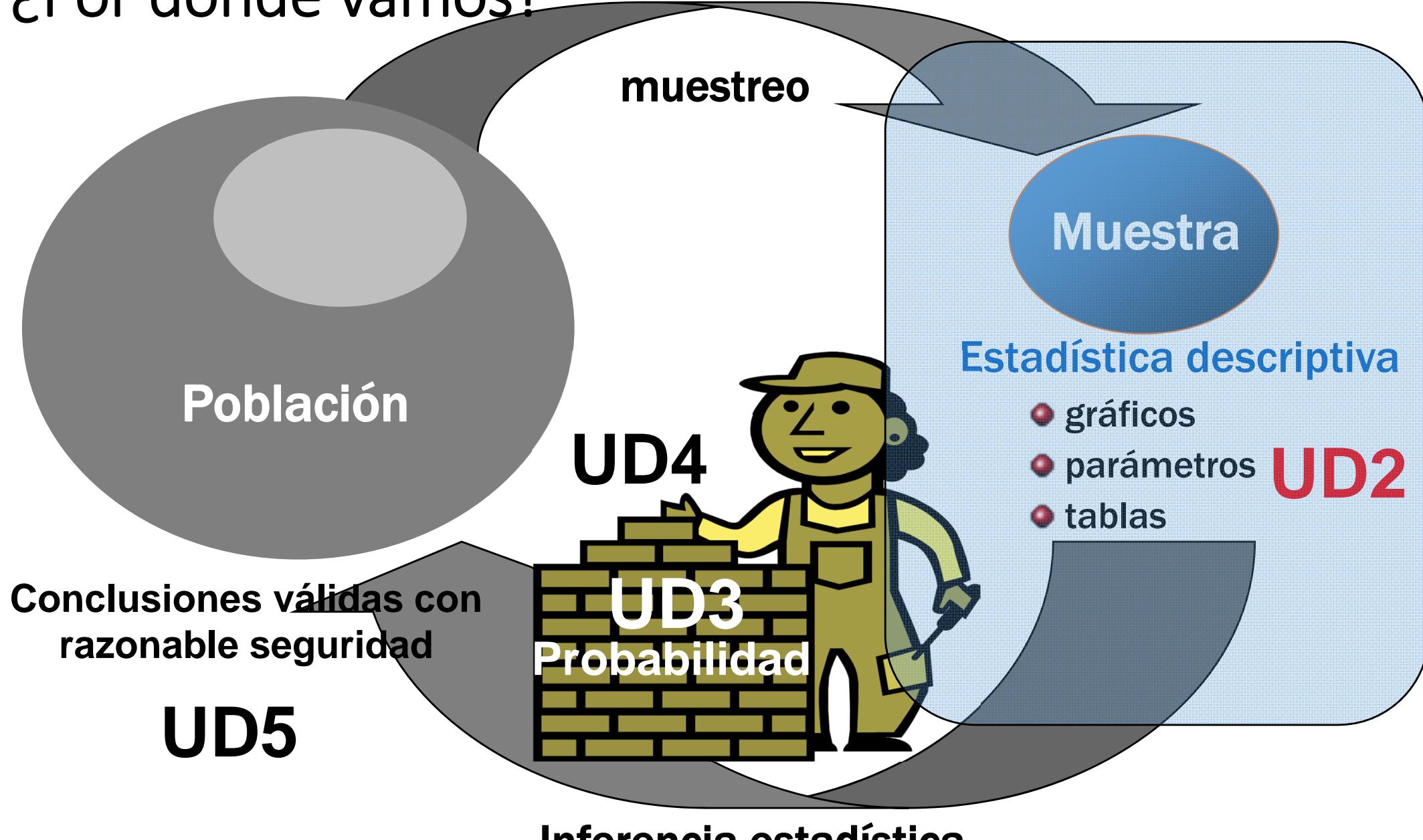


- ¿Cuál es el objetivo del estudio?
- ¿Qué tipo de variables tengo?

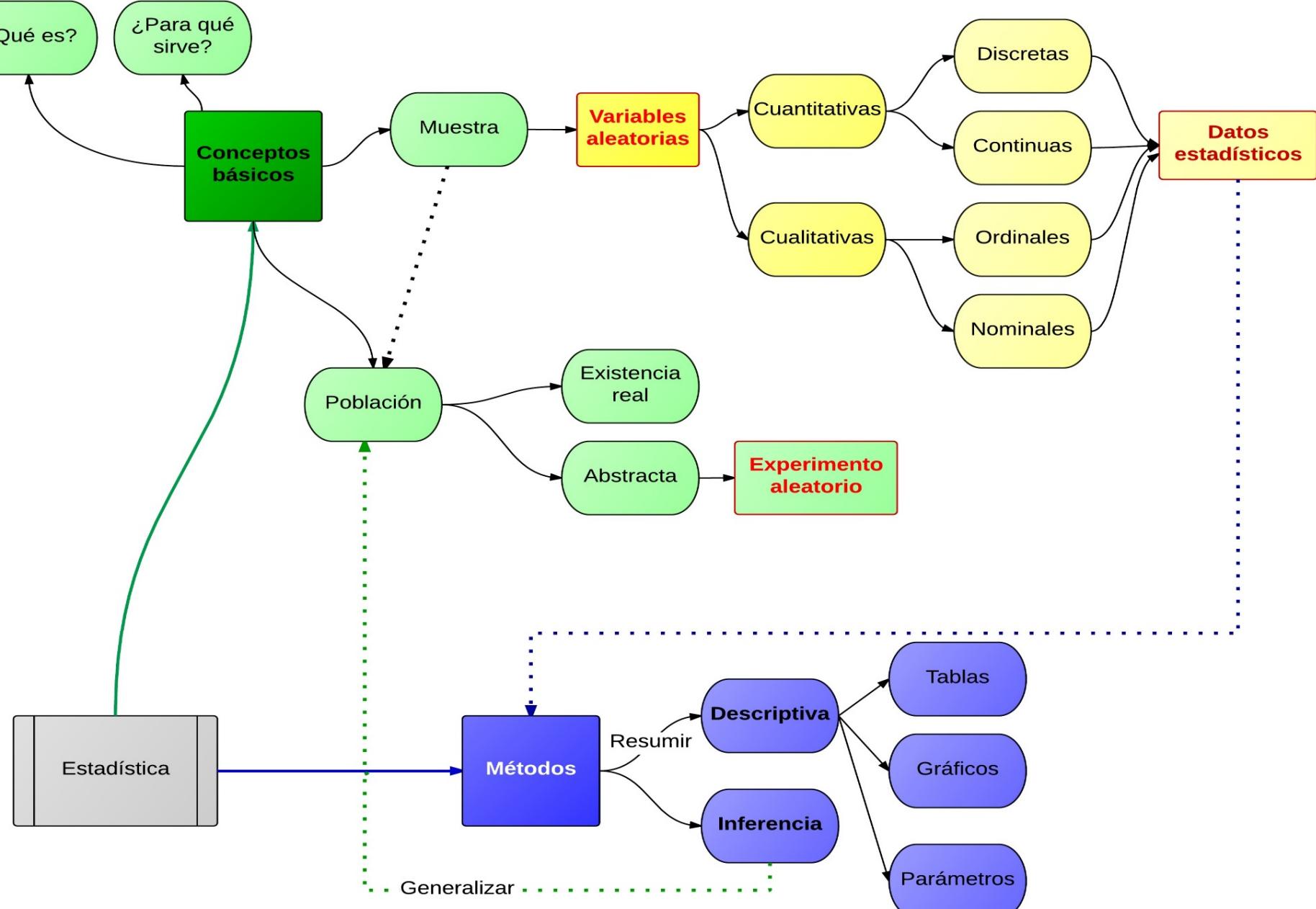
2. ¿Qué relación hay entre ellas?



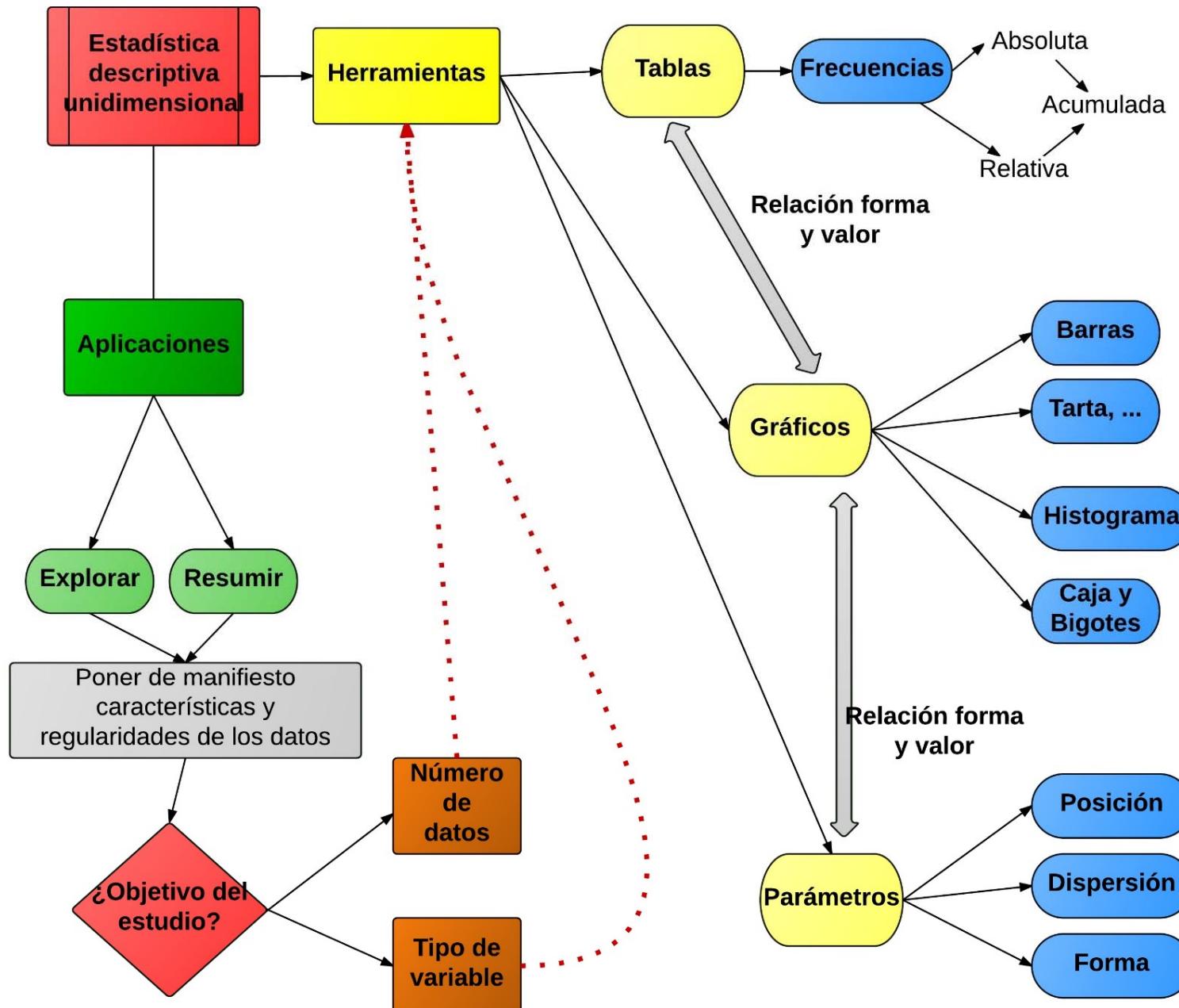
¿Por dónde vamos?



Mapa mental



Mapa mental



Glosario UD 1 y 2



Apuntamiento	Frecuencia acumulada	Parámetro muestral	Posición
Box & Whisker	Frecuencia relativa	Individuo	Recorrido o Rango
Caja y Bigotes	Histograma	Inferencia	Recorrido o Rango Intercuartílico
Campana de Gauss	Desviación típica	Media	Simetría
Característica aleatoria	Diagrama de barras	Mediana	Tabla de frecuencias
Coeficiente de asimetría	Diagrama de sectores o tarta	Moda	Tabla de frecuencias cruzadas
Coeficiente de curtosis	Dispersión	Muestra	v.a. Continua
Coeficiente de Variación	Distribución Normal	Muestreo	v.a. Cualitativa
Cuartil	Estadística	Normalidad	v.a. Cuantitativa
Dato anómalo	Experimento aleatorio	Parámetro de dispersión	v.a. Discreta
Dato o valor extremo	Frecuencia absoluta	Parámetro poblacional	Variabilidad
Datos estadísticos	Parámetro de forma	Percentil	Variable aleatoria (v.a.)
Descriptiva, Estadística	Parámetro de posición	Población	Varianza



Fuentes: Romero y Zúñica: "Métodos Estadísticos en Ingeniería" | Instituto Nacional de Estadística (INE) | Material docente de R. Alcover (DEIOAC - UPV) | Material docente de V. Giner (DEIOAC - UPV) | Material docente de S. Vidal (DEIOAC - UPV) | Material docente de A. Caldúch (DEIOAC - UPV) | Material docente previo de E. Vázquez (DEIOAC - UPV)

Estas transparencias NO son unos apuntes, son solo un guión de las explicaciones hechas en clase y algunos ejemplos adicionales.

Enlaces: [Proyecto Descartes – MEC - Est. Descriptiva Unidimensional](#)

Esta obra está bajo una licencia Reconocimiento-No comercial-Compartir bajo la misma licencia 2.5
España de Creative Commons. Para ver una copia de esta licencia, visite
<http://creativecommons.org/licenses/by-nc-sa/2.5/es/>



Fin

