UNIDAD DIDÁCTICA 2

ESTADÍSTICA DESCRIPTIVA

OBJETIVO

El objetivo de esta Unidad Didáctica es

1. introducir los conceptos más elementales de la Estadística,
2. familiarizar al alumno con algunas técnicas, sencillas pero poderosas, de **Estadística Descriptiva**.

Contenidos

Contenidos

2. Estadística descriptiva bidimensional
    2.1 Tablas de frecuencias cruzadas
    2.2 Diagrama de dispersión
    2.3 Covarianza y coeficiente de correlación

Estadística descriptiva unidimensional

- La Ciencia Estadística tiene un doble objetivo:
  - Generar y recopilar datos que contengan información relevante sobre un determinado problema.
  - Analizar los datos con el fin de extraer de ellos dicha información.

- Primer paso en el análisis: tratamiento descriptivo sencillo de los mismos para
  - Poner de manifiesto las características y regularidades existentes en los datos.
  - Sintetizarlas en un número reducido de parámetros o con representaciones gráficas.

Conceptos básicos. Poblaciones

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

Ejemplo 1: En un estudio sobre la intención de voto de los ciudadanos españoles, la población está constituida por el conjunto de los españoles con derecho a voto.

Poblaciones

Ejemplo 2: al realizar en una industria el control de calidad en recepción de una partida de piezas, la población está constituida por la totalidad de piezas de la partida.

Los 2 ejemplos anteriores tratan de poblaciones con una existencia física real, constituidas por un número finito, aunque posiblemente muy elevado, de “individuos”.

Poblaciones

- Aunque pueda parecer sorprendente no es ésta la situación más frecuente.

- En la práctica, en general las poblaciones a estudiar son de carácter abstracto, fruto del necesario proceso de conceptualización que debe preceder al estudio científico de cualquier problema.

Poblaciones

Ejemplo 3: Se desea estudiar si un dado es correcto o está trucado.

¿Qué quiere decir la afirmación de que el dado es correcto?

En la práctica que si se tira un número muy elevado de veces, los seis resultados posibles saldrán aproximadamente con la misma frecuencia.

Al abordar este problema nos referiremos a la población abstracta constituida por infinitos lanzamientos del dado en cuestión.

Sobre dicha población se desea estudiar si las frecuencias de los seis resultados posibles son idénticas.

Poblaciones

Ejemplo 4: En un estudio sobre la eficiencia de diversos algoritmos de encaminamiento de mensajes entre nudos en una red de procesadores, la población a investigar está constituida por todos los mensajes que se puedan generar en la red.

Los “individuos” que forman una población pueden corresponder a entes de naturaleza muy diversa (personas, piezas, lanzamientos de dados, mensajes, etcétera....)

Poblaciones

- En los casos de los dos primeros ejemplos dichos individuos tienen existencia real, previa a la realización del estudio.

- En casos como los de los ejemplos 3 y 4, los “individuos” que constituyen la población pueden irse generando mediante la realización de un determinado proceso (lanzar un dado, emitir un mensaje,...)

- **Experimento aleatorio**: proceso que en sucesivas realizaciones puede ir generando diferentes individuos de la población.

Conceptos básicos.
Variables aleatorias

¡En toda población real hay **VARIABILIDAD**!

Ejemplos:

Unos españoles votan a ciertos partidos y otros a otros.

Una determinada dimensión de una pieza varía algo de una pieza a otra.

El número que sale en un dado varía de unas tiradas a otras.

Unos mensajes tienen retardos más elevados que otros.

Etcétera....

Variables aleatorias

Característica aleatoria: cualquier característica que puede constatarse en cada individuo de la población

Ejemplos:

Partido a que piensan votar los individuos.

Dimensión de una pieza.

Número que sale en un dado.

Retardo de un mensaje.

Variables aleatorias

Muchas características aleatorias se expresan numéricamente (nº de puntos obtenidos al lanzar un dado, retardo de un mensaje...).

A este tipo de características aleatorias se las denomina **variables aleatorias**

Variables aleatorias

Cuando una característica aleatoria es de tipo **cualitativo o atributo** (p.ej. partido político a votar) se pueden codificar sus alternativas y tratarla como una variable aleatoria.

¡Cuidado! En este caso carecen de sentido operaciones como sumar, promediar...., aplicables con variables numéricas.

Variables aleatorias

Cuando el conjunto de valores que podría tomar una variable aleatoria es discreto (finito o infinito numerable) se dice que dicha variable es **discreta**.

Ejemplos:

Nº de puntos al lanzar un dado

Nº de errores en un programa de ordenador

Y también cualquier variable que se origine al codificar las diferentes alternativas de una característica cualitativa (sexo, partido votado, etc...)

Variables aleatorias

Cuando el conjunto de valores que podría tomar una variable aleatoria es un infinito continuo se dice que dicha variable es **continua**.

Ejemplos:

Todas las características que se miden sobre una escala de naturaleza básicamente continua (estatura, pesos, rendimiento, tiempo, resistencias, etc...)

Variables aleatorias

Cuando sobre cada individuo de la población se estudian K características se tiene **una variable aleatoria K-dimensional**.

Ejemplo:

En la población de estudiantes de la UPV se estudia el sexo (codificado como 1 ó 2), la edad, estatura y peso $\Rightarrow$ variable aleatoria de dimensión 4.

Variables aleatorias

***EJERCICIO 1**: el peso y el modelo de un Tablet PC ¿constituyen una variable aleatoria bidimensional?*

¿Y el número de líneas de código y el número de errores en los programas preparados por una empresa de software?

¿Y el contenido de leucocitos en la sangre de individuos alcohólicos y no alcohólicos?

¿Y las estaturas del marido y de la mujer en los matrimonios jóvenes de un país?

Muestra

Subconjunto de la población.

Debe ser “representativa” de ésta.

Población estudiada real: la muestra se forma seleccionando de la forma más aleatoria posible un conjunto de individuos de la misma.

Población abstracta: se obtiene la muestra realizando un cierto número de veces el experimento aleatorio que genera los individuos de la población (p.ej. lanzar varias veces un dado, generar un conjunto de mensajes en la red de multiprocesadores....)

Muestra

EJERCICIO 2: *se desea estudiar la relación que existe entre la estatura y el peso en la juventud española. El conjunto de los alumnos matriculados en Estadística en 1º del Grado en Ingeniería Informática de Valencia ¿puede considerarse una muestra representativa de la población a efectos del estudio en cuestión?*

Dicho conjunto ¿puede considerarse una muestra representativa para estudiar las tendencias políticas en la juventud española?

¿Y para estudiar el nivel cultural?

¿Y para estudiar la característica aleatoria color de los ojos?

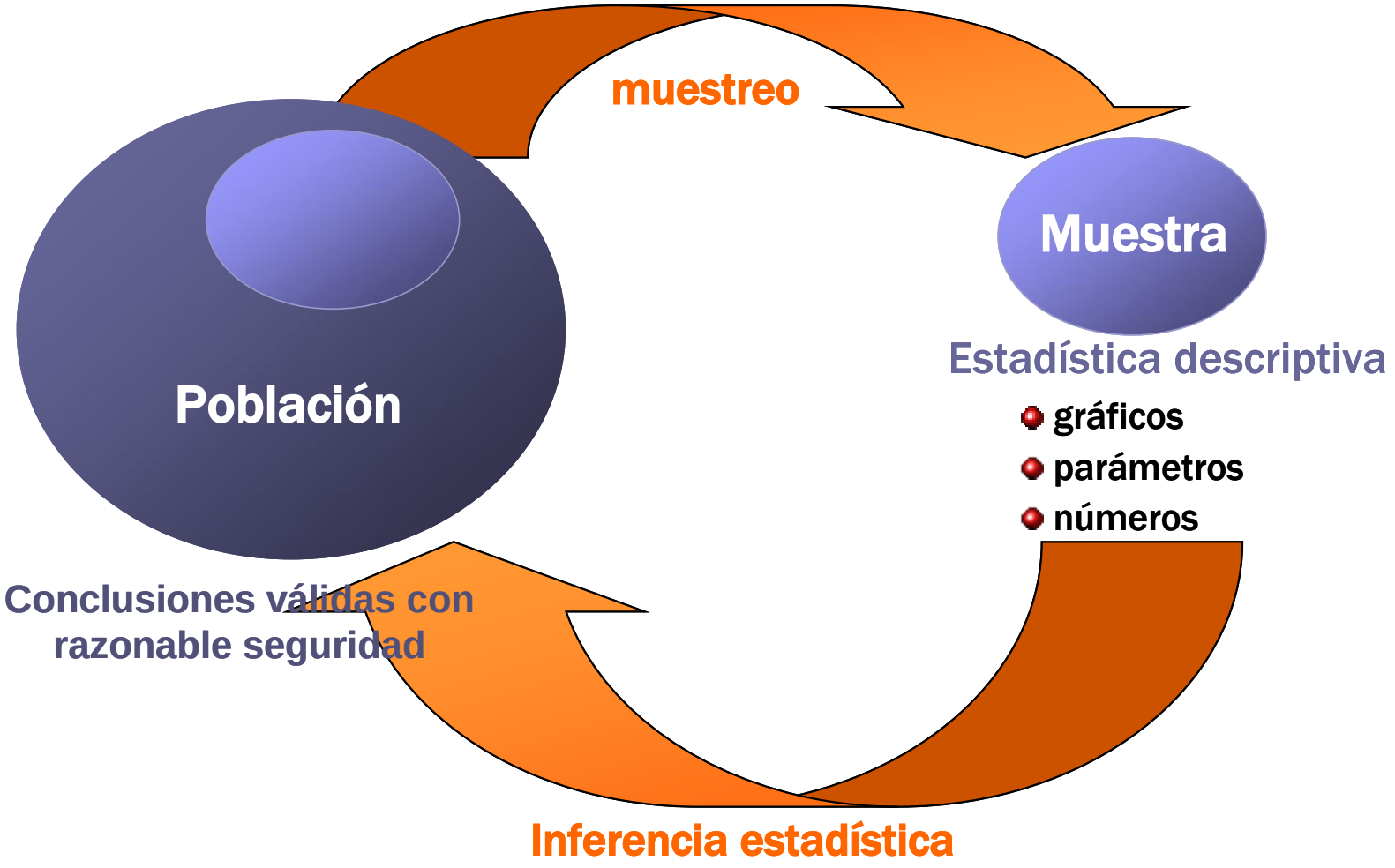
Datos estadísticos

Son los valores observados para la variable aleatoria en los individuos que forman la muestra.

El tratamiento de dichos datos con el fin de poner de manifiesto sus características más relevantes es el objeto de la Estadística Descriptiva (Unidad Didáctica 2):

- Tabulaciones
- Cálculo de parámetros
- Representaciones gráficas

Estadística Descriptiva e Inferencia Estadística



Tablas de frecuencias

Si el nº de datos (N) no es muy reducido, su interpretación se facilita presentándolos agrupados en una tabla.

Dos casos

a) La variable estudiada es cualitativa (atributo) o cuantitativa con un número reducido de valores posibles.

b) Variable continua (difícil encontrar valores repetidos).

Tablas de frecuencias

Ejemplo de a): se pretende describir el funcionamiento simultáneo de un tipo de robots multiprocesador.

Nº de procesadores funcionando (X_i)	**Frecuencia absoluta** Nº de robots **(n_i)**	**Frecuencia relativa** % robots **$f_i = n_i / N$**
0	10	6,25%
1	35	21,88%
2	60	37,50%
3	55	34,37%
Total	**N**=160	100%

Tablas de frecuencias

Caso b) variable continua:

-Como es difícil encontrar valores repetidos se agrupa la variable en K intervalos.

-Los intervalos suelen ser de la misma longitud.

-K depende del tamaño de muestra N.

-Se recomienda entre 5 y 15 intervalos

Tablas de frecuencias

Ejemplo de b): variable ESTATURA de la encuesta.

N=131

Límites del intervalo	Centro del intervalo (X_i)	Número de observaciones n_i
150-155	152.5	3
155-160	157.5	9
160-165	162.5	22
165-170	167.5	16
170-175	172.5	37
175-180	177.5	22
180-185	182.5	14
185-190	187.5	3
190-195	192.5	3
195-200	197.5	2

Diagramas de barras y de tarta

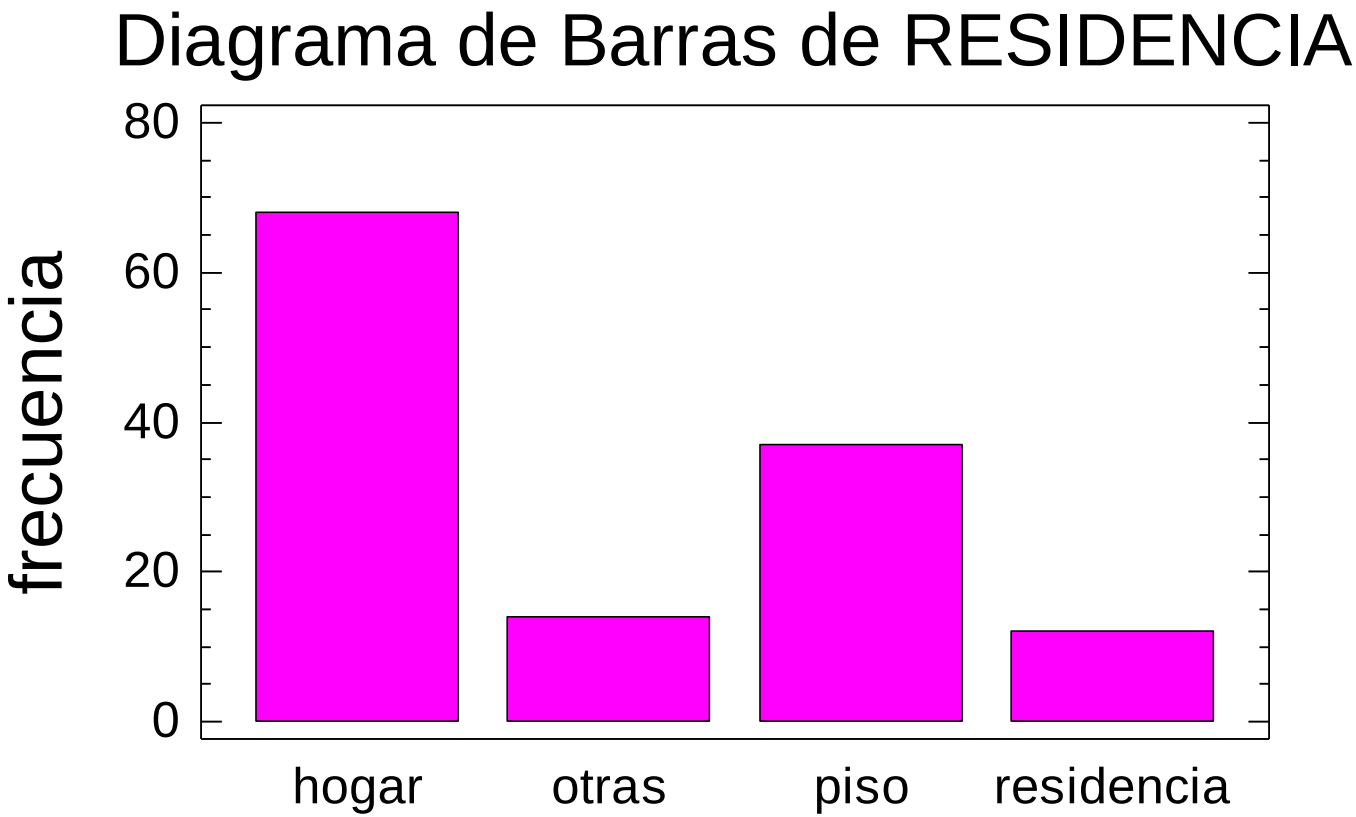
Representación gráfica de una tabla de frecuencias para una variable cualitativa

Diagrama de barras: A cada posible valor se le hace corresponder una barra de longitud proporcional a la frecuencia (absoluta o relativa)

Diagrama de tarta: La superficie de un círculo se reparte en sectores con áreas proporcionales a las frecuencias observadas para cada posible valor.

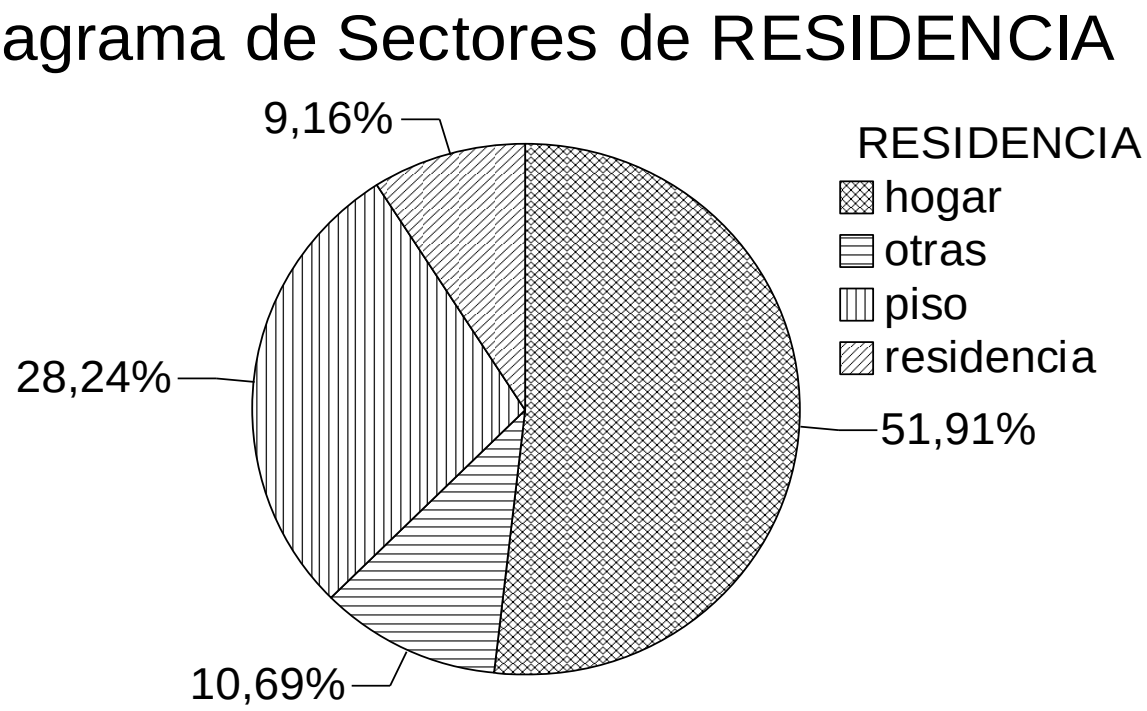
Diagramas de barras y de tarta

Ejemplo: Diagrama de barras de lugar de residencia (datos encuesta)



Diagramas de barras y de tarta

Ejemplo: Diagrama de tarta de lugar de residencia (datos encuesta)



Histograma

Representación gráfica de una tabla de frecuencias para variables cuantitativas continuas o discretas con un número elevado de valores diferentes.

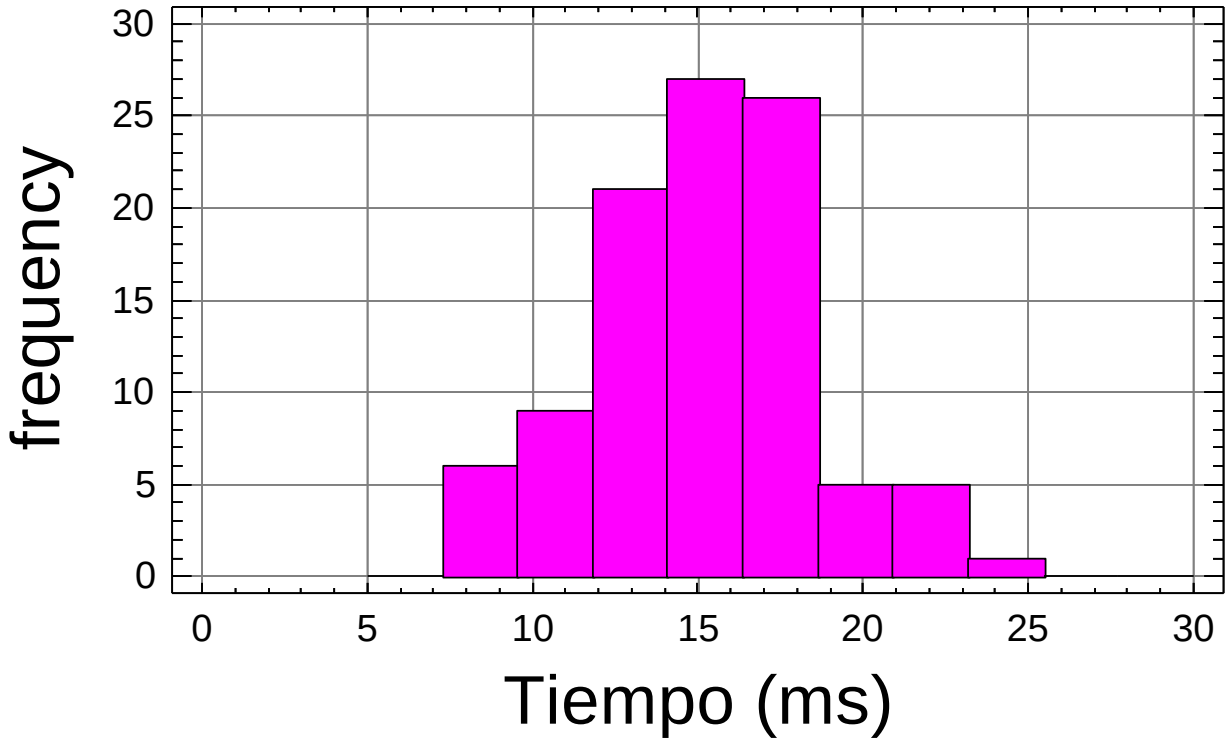
Eje horizontal: Valores de la variable agrupados en tramos.

Eje vertical: Frecuencias. Sobre cada tramo se levanta una barra de altura proporcional a la frecuencia.

Mínimo de 40 ó 50 datos. Número de tramos: entero cercano a la raíz cuadrada del tamaño de muestra. Menos de 15 ó 20 tramos.

Histograma

Ejemplo: Histograma que representa los tiempos de ejecución (ms) de una muestra de 100 programas.



Histogramas

Son útiles para detectar:

Existencia de datos anómalos

Mezclas de poblaciones

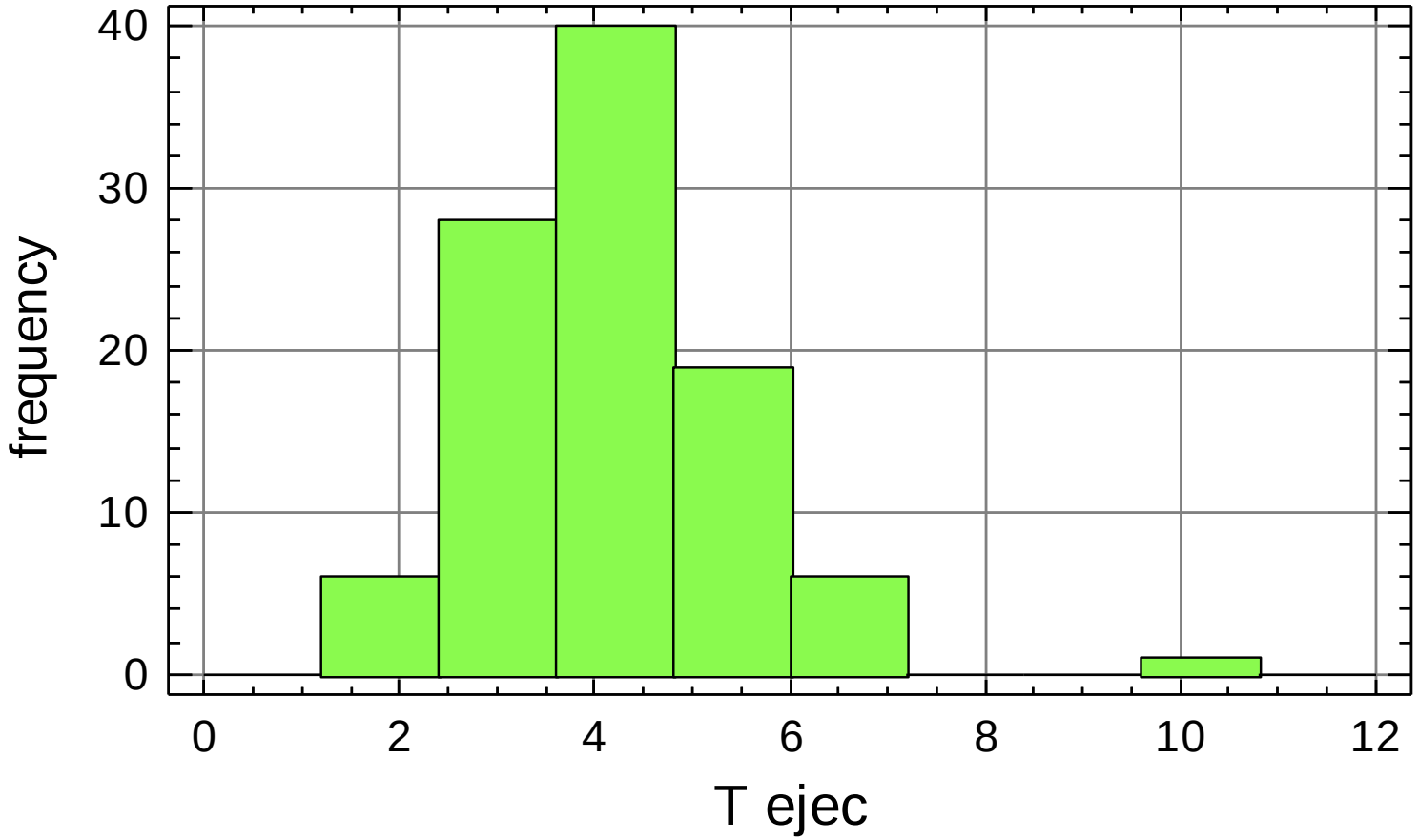
Datos artificialmente modificados

Forma de la distribución de los datos

.....

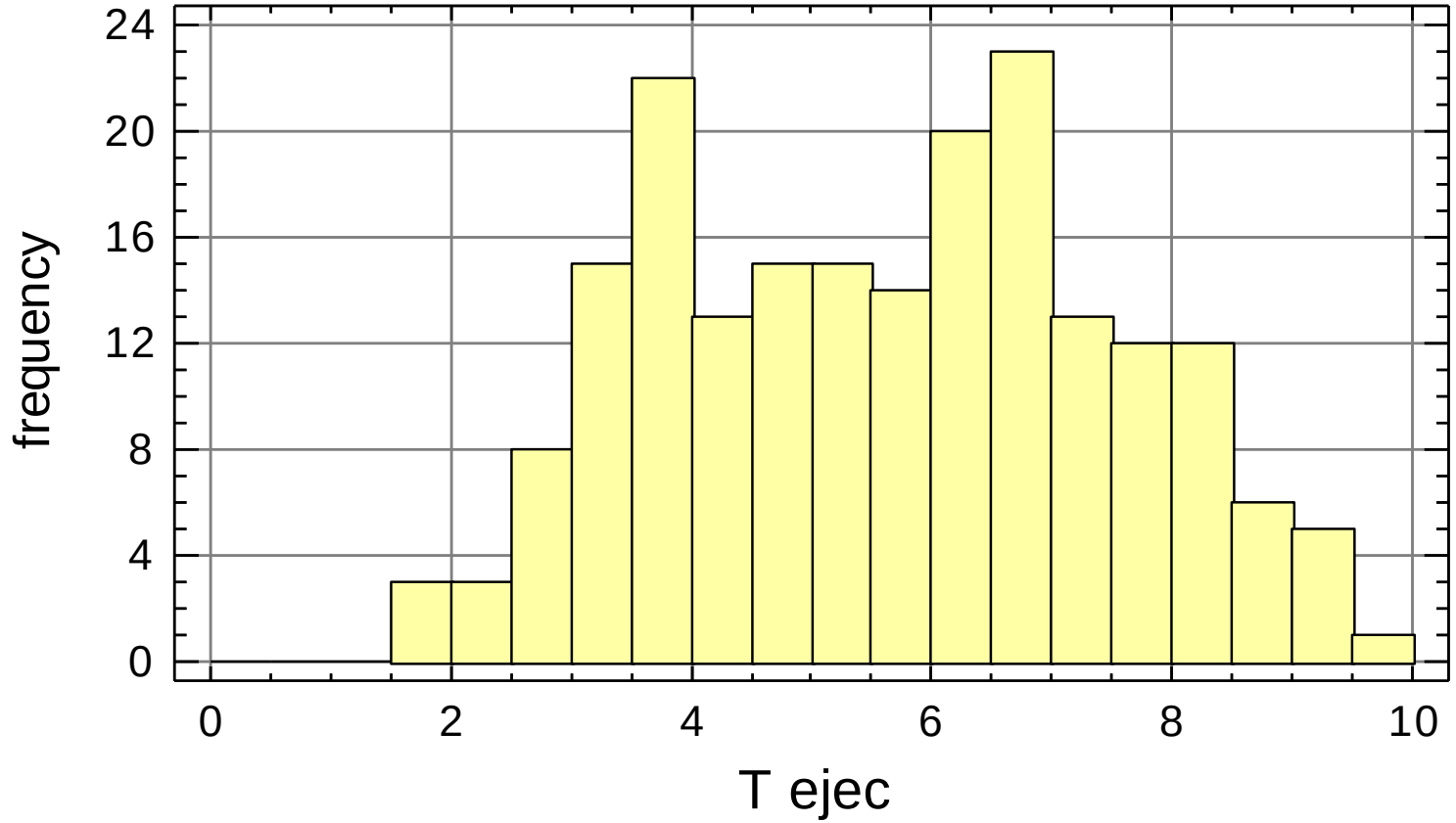
Histograma

Ejemplo: existencia de datos anómalos

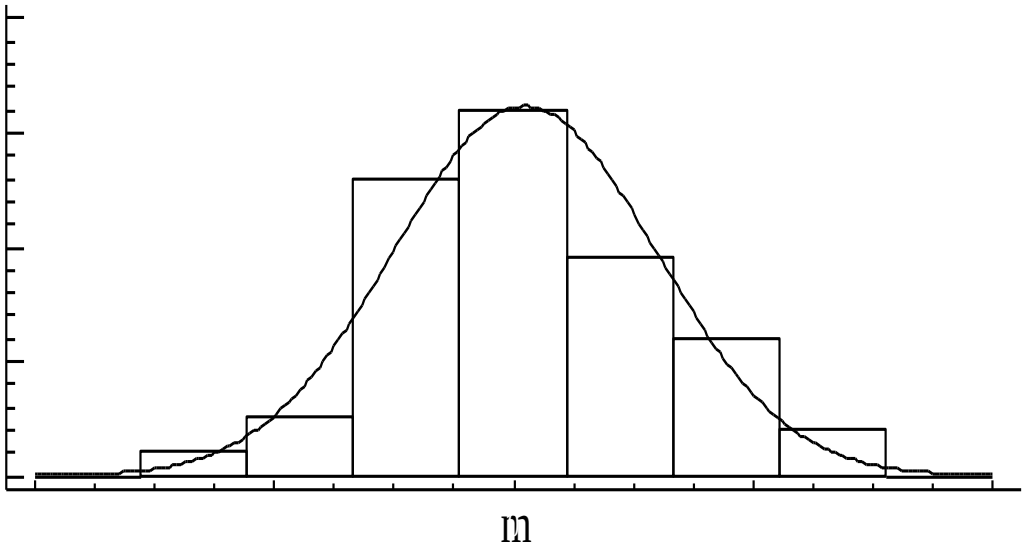


Histograma

Ejemplo: mezclas de poblaciones distintas

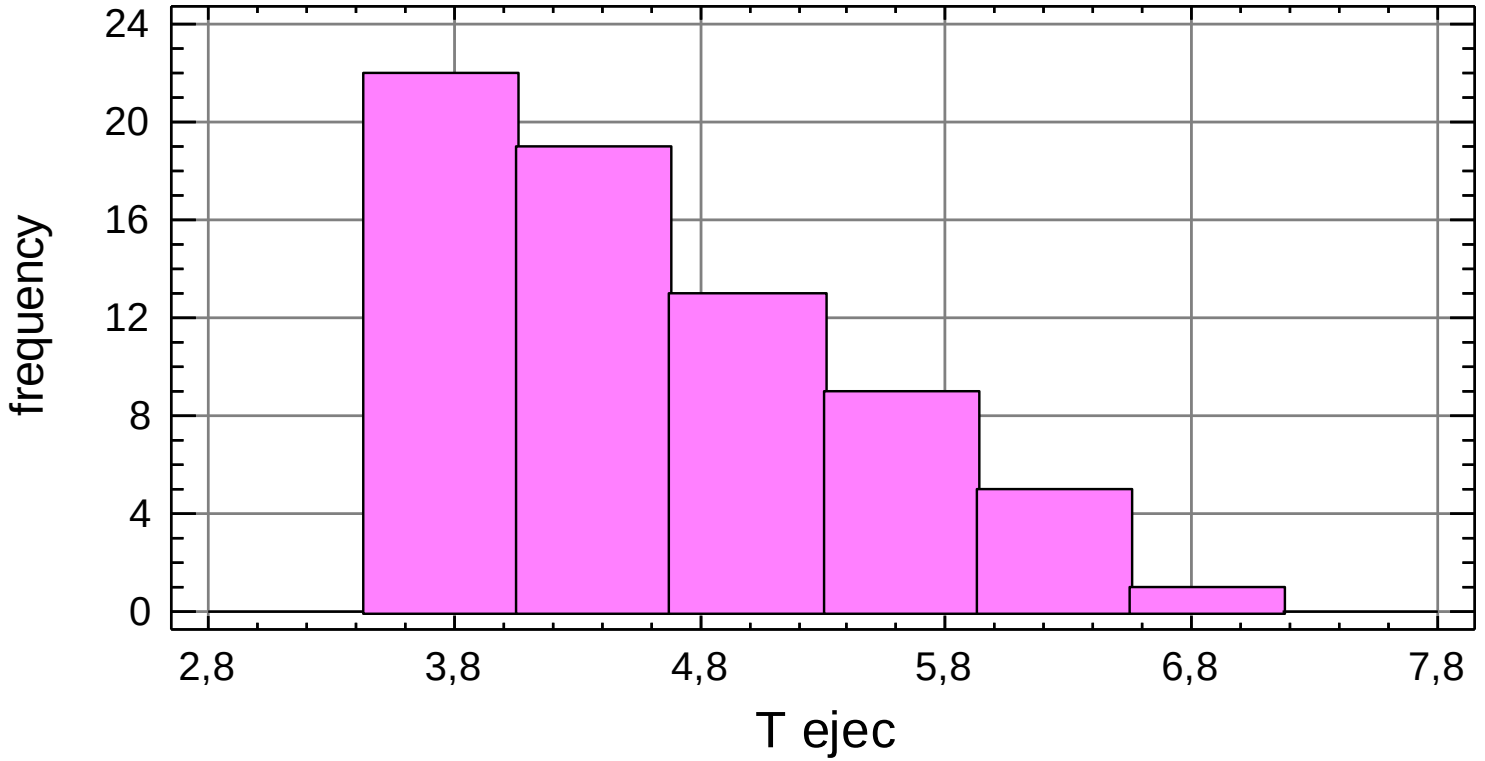


Si una muestra representativa de los datos da un histograma de este tipo:



Histograma

Pero al hacer el histograma sale esta figura, esto indica que falta la mitad de los datos, que han sido artificialmente modificados



Histograma

- Sin embargo hay variables que siendo representativa la muestra dan histogramas del tipo: se trata de variables que siguen el modelo exponencial.

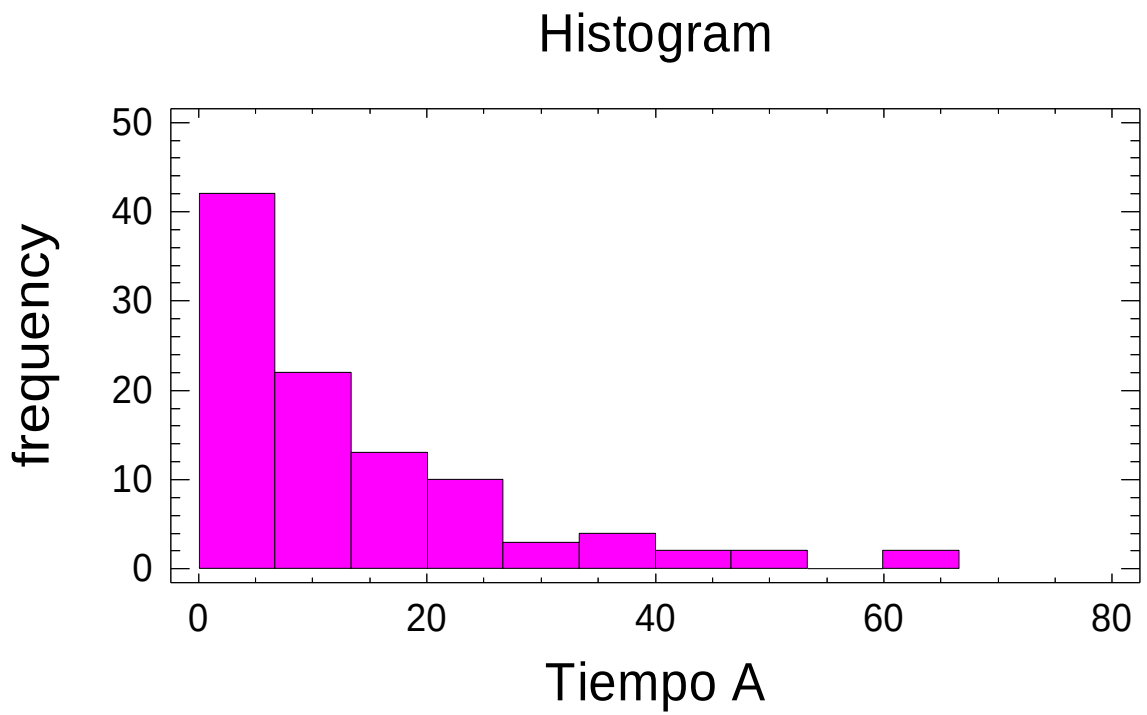


Diagrama de frecuencias acumuladas o polígono de frecuencias

Representación gráfica de las frecuencias (habitualmente relativas) acumuladas.

Eje horizontal: Valores de la variable agrupados en tramos.

Eje vertical: Sobre el límite superior de cada tramo se levanta una altura igual a la frecuencia (relativa) acumulada hasta dicho tramo.

El gráfico tiene forma de línea quebrada no decreciente. La altura final es 1 (ó 100 si las frecuencias relativas están en porcentajes)

Diagrama de frecuencias acumuladas

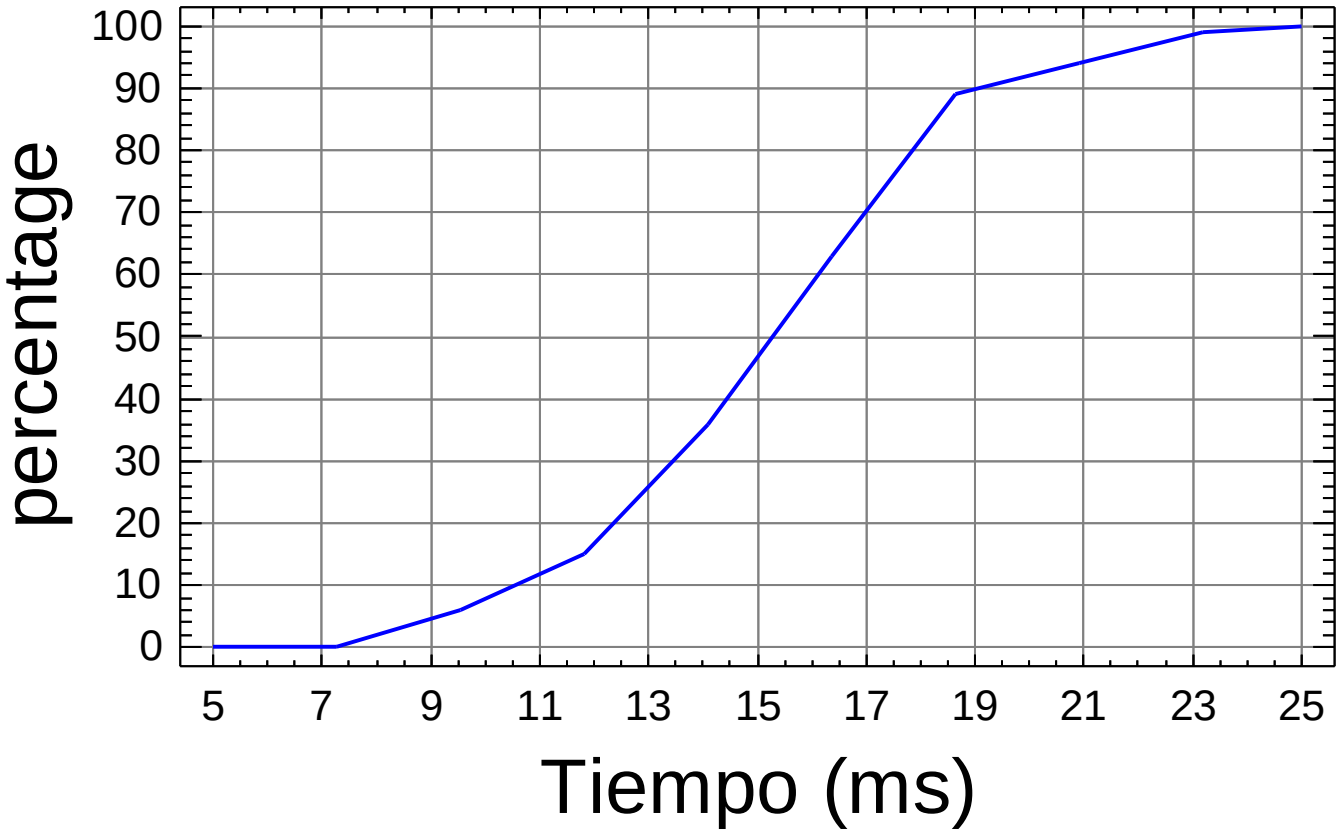


Diagrama de frecuencias acumuladas

El diagrama de frecuencias acumuladas permite responder directamente a preguntas como ¿qué porcentaje de los programas tienen un tiempo de ejecución menor o igual que 17 ms?

SOLUCIÓN:

La solución es 70%. Por tanto 17 ms es el Percentil 70 de la distribución.

Percentiles

- En general un percentil p es un valor que deja en los datos una proporción p de valores por debajo de él.
- En el ejemplo anterior Percentil 70=17ms
- ¿Cuánto valdría el percentil 50?

SOLUCIÓN:

Mirando el gráfico Percentil 50≈15,5ms

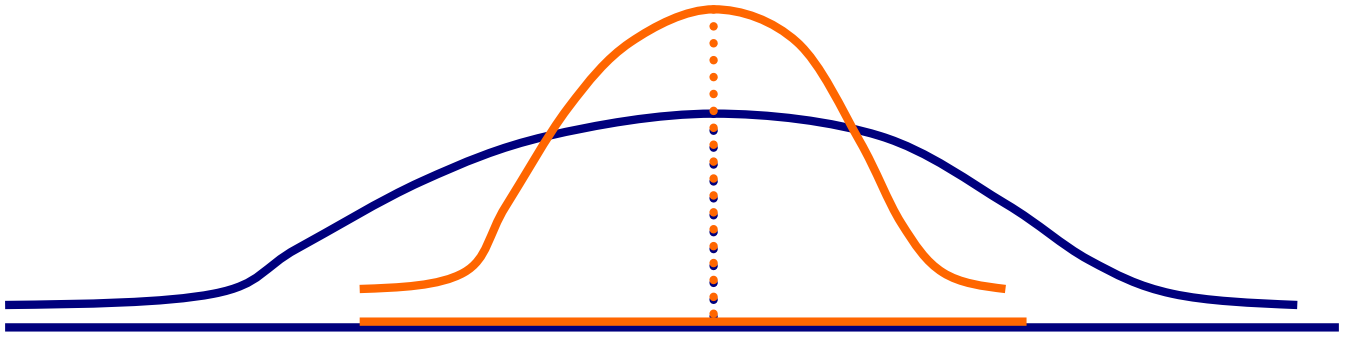
Posición y dispersión

Fundamentalmente la pauta de variabilidad constatada en un conjunto de observaciones relativas a una variable cuantitativa unidimensional puede caracterizarse por dos tipos de parámetros que definan respectivamente la **posición** y la **dispersión** de las observaciones.

En las siguientes figuras, en las que hemos sustituido por comodidad los histogramas de frecuencias por curvas continuas, se ve claramente el sentido de ambos términos.

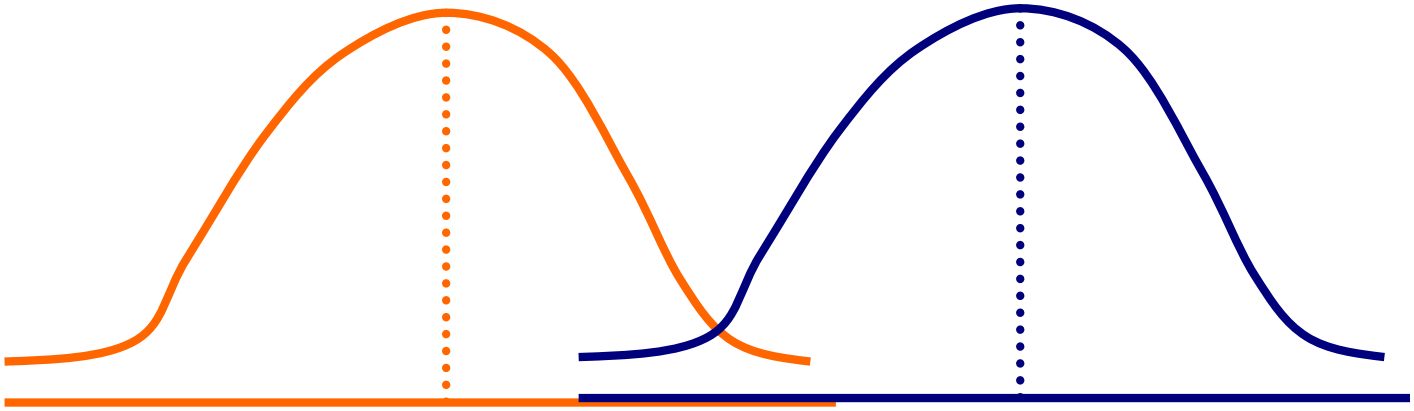
Posición y dispersión

Idéntica posición y distinta dispersión



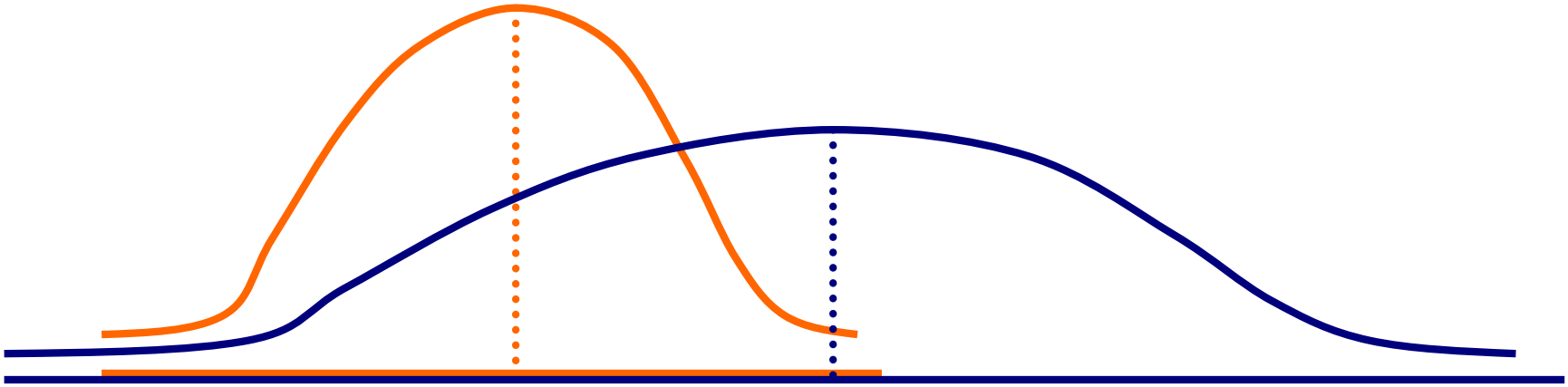
Posición y dispersión

Idéntica dispersión y distinta posición



Posición y dispersión

Distinta dispersión y distinta posición



Parámetros de posición

Media

$$\overline{X} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Parámetros de posición: media

EJERCICIO 3: *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 (meses). Los resultados en meses son 50, 38, 45, 30, 47, 50, 48, 62, 55, 53, 52. Obtener el tiempo medio sin averías.*

SOLUCIÓN:

$$\overline{X} = \frac{50 + 38 + \ldots + 52}{11} = 48{,}18 \text{ meses}$$

Parámetros de posición: media

- En algunos casos particulares la media puede resultar una medida de posición algo engañosa.
- Este es el caso en concreto con datos muy asimétricos, o en los casos en que unos pocos valores extremos (en general por la cola derecha o la izquierda del histograma) pueden influir excesivamente sobre el valor de la media.

Parámetros de posición: media

EJERCICIO 4: *Al analizar el número de visitas diarias en una página web durante una semana se ha constatado una media de 10. En efecto en los 7 días analizados las visitas han sido: 5, 3, 8, 6, 4, 4, 40.*

La media no es en este caso una medida adecuada de la posición de los datos.

Parámetros de posición: moda

- La moda de una muestra es el valor más frecuente.

Ejemplo 1: La moda de los datos del nº de visitas a la página web es 4.

Ejemplo 2: La moda de los tiempos de funcionamiento sin averías es 50.

Parámetros de posición

Mediana

Más aconsejable que la media con datos asimétricos o con valores extremos.

Es el valor en la muestra ordenada que ocupa la posición (N+1)/2 si N es impar.

Cuando N es par, se promedian los dos valores centrales, es decir los que ocupan las posiciones N/2 y (N/2)+1.

La mediana es el percentil 50 de los datos.

Parámetros de posición: mediana

EJERCICIO 5: *¿cuál sería la mediana de los datos recogidos en el ejemplo mencionado del número de visitas a la página web?*

SOLUCIÓN:

Ordenando los datos de la muestra de menor a mayor:

4 4 3 5 6 8 40

La mediana sería el dato que ocupa la posición (N+1)/2=(7+1)/2=4 por lo que sería mediana=5 más representativa que la media=10 de la posición central de los datos.

Parámetros de posición: mediana

EJERCICIO 6: *calcular la mediana de los datos del Ejercicio 3 (Tiempo de funcionamiento sin averías).*

SOLUCIÓN: *Datos ordenados N=11*

30 38 45 47 48 50 50 52 53 55 62

Mediana= dato que ocupa posición (11+1)/2=6

$\Rightarrow$ Mediana= 50 meses

Se observa que ha salido

Mediana=50<media=48,18

Parámetros de posición: mediana

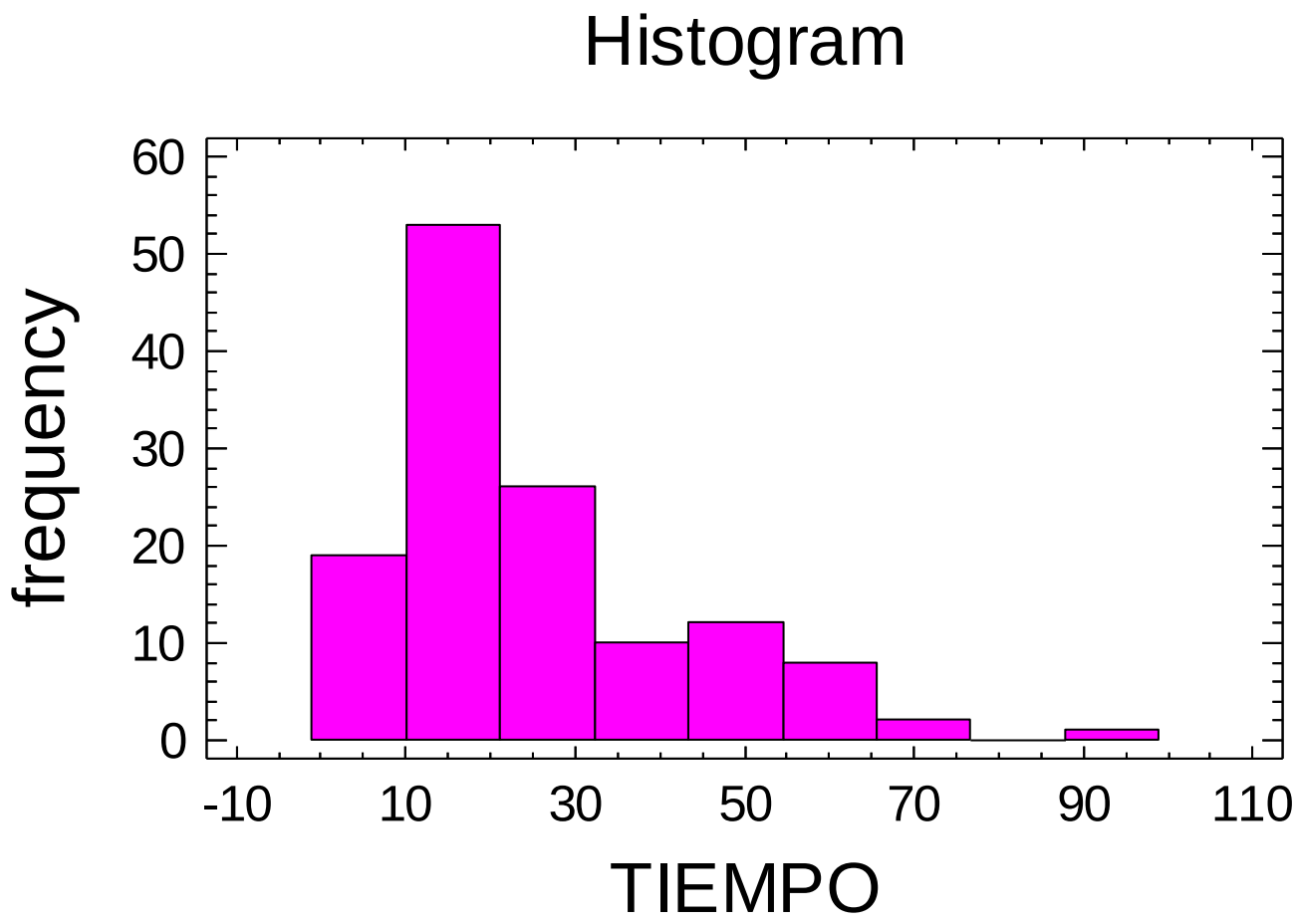
EJERCICIO 7: *Calcular las medianas de las variables EDAD, ESTATURA, PESO y TIEMPO con los datos de la encuesta y compararlos con las medias respectivas. Constatar la sensible diferencia entre ambos parámetros para la variable TIEMPO, y comprobar mediante un histograma que la distribución de esta variable es muy asimétrica.*

Parámetros de posición: mediana

SOLUCIÓN:

	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

Parámetros de posición: mediana



Parámetros de posición: mediana

- Por tanto para la variable TIEMPO

Mediana=20<media=26,1221 debido a la asimetría que presenta la distribución de frecuencias de los datos.

Parámetros de posición

1er cuartil C1

Es el valor más pequeño en la muestra ordenada tal que N/4 de las observaciones quedan por debajo, y 3N/4 quedan por encima.

El C1 es el percentil 25 de los datos.

Parámetros de posición

3er cuartil C3

Es el valor más pequeño tal que 3N/4 de las observaciones ordenadas es inferior a él y N/4 superior.

El C3 es el percentil 75 de los datos.

La mediana o percentil 50, es el 2ºcuartil C2

Parámetros de posición

C1 y C3

Una forma sencilla de calcular los cuartiles C1 y C3 es hallando las medianas de la mitad inferior y la mitad superior de los datos.

Entre los dos cuartiles C1 y C3 se encuentra el 50% central de los datos.

Parámetros de posición: C1 y C3

EJERCICIO 8: *Calcular el primer y tercer cuartil de los datos del ejemplo sobre visitas a la página web.*

SOLUCIÓN:

Datos ordenados de la muestra

4 4 3 5 6 8 40

Donde la mediana de la primera mitad de los datos es 4 entonces C1=4, y la mediana de la segunda mitad de los datos es 8, entonces C3=8.

Parámetros de posición: C1 y C3

EJERCICIO 9: *Calcular el primer y tercer cuartil de los datos del ejemplo sobre Tiempo de funcionamiento sin averías.*

SOLUCIÓN: *datos ordenados*

30 38 45 47 48 50 50 52 53 55 62

C1= 45 C3=53

Parámetros de posición: C1 y C3

EJERCICIO 10: *calcular los dos cuartiles de las variables ESTATURA y PESO con los datos de la encuesta. Repetir el cálculo por separado para los chicos y las chicas y comentar los resultados obtenidos.*

Parámetros de posición: C1 y C3

SOLUCIÓN:

	Todos		Chicos		Chicas	
	C1	C3	C1	C3	C1	C3
Estatura	165	179	173	180	160	165
Peso	57	74	66	76	51	60

Parámetros de posición: C1 y C3

- Se observa que los C1 y C3 tanto de ESTATURA como de PESO del grupo de chicos, dan mayores que en el grupo de CHICAS. Lo que indica diferente posición central en los valores de esas variables.
- Para todos los datos juntos, hay más diferencia entre C1 y C3 (más dispersión en el intervalo del 50% central) que en cada uno de los grupos por separado.

Parámetros de dispersión

Para describir la pauta de variabilidad de los datos no es suficiente con la posición.

EJERCICIO 11: *¿Para una persona que no sabe nadar es suficiente saber que la profundidad media de un lago es 1,4 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?

Parámetros de dispersión

- Intuitivamente la idea de dispersión de un conjunto de datos es bastante clara.
- El conjunto de datos 3, 3, 3, 3, y 3 tiene una dispersión nula.
- Los datos 1, 3, 5, 7 y 9 tienen dispersión, pero menos que los datos 1, 5, 10, 15 y 20.
- ¿Cómo puede precisarse esta idea intuitiva mediante un índice que cuantifique la mayor o menor dispersión de unos datos? Diferentes parámetros pueden utilizarse al respecto.

Parámetros de dispersión

Recorrido

Recorrido=Máximo–Mínimo

Inconvenientes:

- Ignora la mayor parte de la información existente en la muestra.
- Depende del tamaño de muestra (muestras grandes tienen recorridos más altos).
- Útil en muestras pequeñas (N≤5).
- Se ve afectado por los datos anómalos.

Parámetros de dispersión: R

EJERCICIO 12: *Calcular el recorrido de los datos del Tiempo de funcionamiento sin averías.*

SOLUCIÓN:

Datos del tiempo de funcionamiento sin averías:

Ordenados de menor a mayor

30 38 45 47 48 50 50 52 53 55 62

Recorrido= 62-30=32 meses

Parámetros de dispersión: R

EJERCICIO 13: *suponiendo que, por error, en los datos del Tiempo de funcionamiento sin averías se hubiera anotado el 8º dato como 150 en vez de 62 ¿cuál sería ahora el recorrido?.*

SOLUCIÓN:

Datos con anomalía

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

Recorrido= 150-30=120 meses

Sale mucho mayor debido a la anomalía y no refleja bien la dispersión.

Parámetros de dispersión

Varianza s^2

$$s^2 = \frac{\sum_{i=1}^{N} \left(x_i - \overline{X} \right)^2}{N - 1}$$

Parámetros de dispersión

Desviación típica

Preferible a la varianza, pues es más fácil de interpretar al venir expresada en las mismas unidades que los datos.

$$s = \sqrt{\frac{\sum_{i=1}^{N} \left(x_i - \overline{X} \right)^2}{N - 1}}$$

Parámetros de dispersión: s^2 y s

EJERCICIO 14: *Calcular la varianza y desviación típica de los datos del ejemplo sobre Tiempo de funcionamiento sin averías.*

SOLUCIÓN: *Datos 50, 38, 45, 30, 47, 50, 48, 62, 55, 53, 52. Media muestral* $\overline{X} = 48{,}18$ meses

$$s^2 = \frac{\sum_{i=1}^{11}(X_i - \bar{X})^2}{11-1} = \frac{(50-48{,}18)^2 + \ldots + (52-48{,}18)^2}{10} =$$

$$= 72{,}76\ meses^2$$

$$s = \sqrt{s^2} = 8{,}53\ \ meses$$

Parámetros de dispersión

Coeficiente de variación

Adimensional. Permite comparar la precisión de sistemas de medida que dan las determinaciones en escalas diferentes. Se calcula en porcentaje

$$CV = \frac{s}{\bar{X}} 100$$

Parámetros de dispersión

Intervalo o Recorrido Intercuartílico

Cuando la media no es un indicador adecuado de posición (distribuciones asimétricas), tampoco resulta la desviación típica s un parámetro adecuado de dispersión.

En esos casos: Más adecuado usar el intervalo o recorrido intercuartílico

recorrido intercuartílico=RI=C3-C1

Parámetros de dispersión: RI

EJERCICIO 15: *Obtener el recorrido intercuartílico de los datos de los ejercicios 13 y 14 y comentar el resultado.*

SOLUCIÓN:

Datos ej.13 (ordenados)

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

$C1 = 45 \quad C3 = 53 \Rightarrow RI = 53 - 45 = 8$

Datos ej.14 (ordenados)

30 38 45 47 48 50 50 52 53 55 62

$C1 = 45 \quad C3 = 53 \Rightarrow RI = 53 - 45 = 8$

Parámetros de dispersión: RI

EJERCICIO 16: *En los datos de ESTATURA de las chicas modificar un dato poniéndolo en metros en vez de en centímetros. Calcular la media, desviación típica, mediana y recorrido intercuartílico de los nuevos datos de ESTATURA de las chicas y compararlos con los valores que se obtienen tras corregir el dato erróneo. ¿Qué se observa?*

SOLUCIÓN: *La tabla siguiente recoge estos parámetros para la ESTATURA de chicas con el primer dato erróneo en metros (1,59), y con el dato correcto (159 cm).*

Se observa que la mediana y el RI no cambian aunque haya un dato erróneo. La media y desviación típica si que se ven modificadas

Parámetros de dispersión: RI

EJERCICIO 16

	ESTATURA (chicas)	
	Con dato erróneo	Con dato correcto
Media	159,68	163,43
Desviación típica	25,61	5,67
Mediana	163	163
RI	5	5

Parámetros de asimetría y curtosis

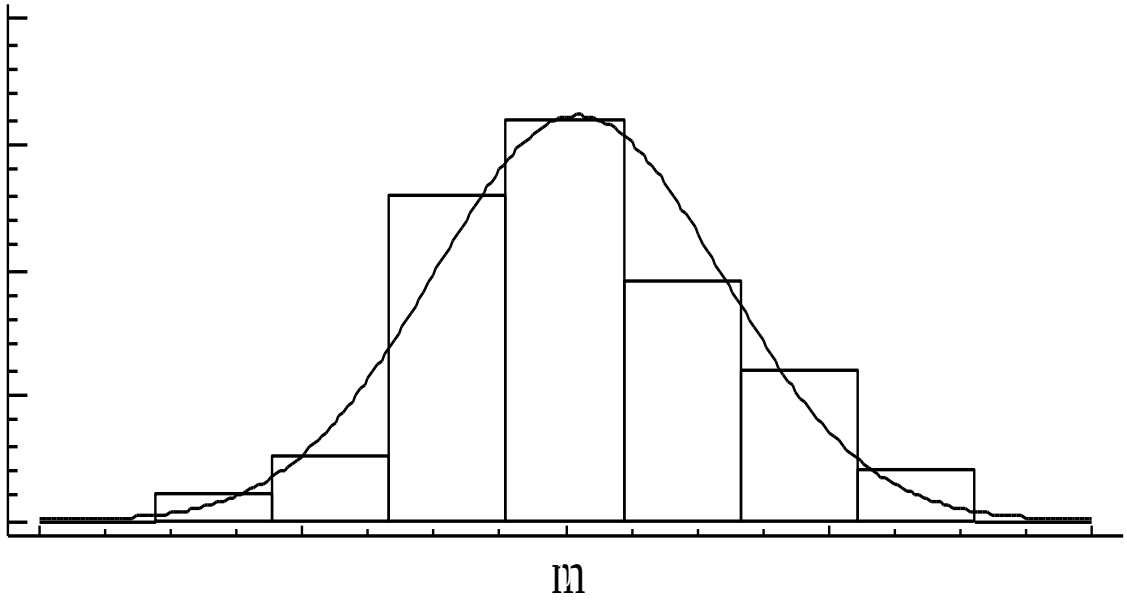
- Las variables aleatorias continuas presentan frecuentemente una pauta de variabilidad que se caracteriza por el hecho de que los datos tienden a acumularse alrededor de un valor central, decreciendo su frecuencia de forma aproximadamente simétrica a medida que se alejan por ambos lados de dicho valor.
- Ello conduce a histogramas que tienen forma de curva en campana (la famosa campana de Gauss, denominada así en honor del célebre astrónomo que estableció, junto con Laplace, la **distribución Normal** al estudiar la variabilidad en los errores de sus observaciones)

Parámetros de asimetría y curtosis

- Para estudiar este tipo de pauta de variabilidad se ha establecido un modelo matemático, la **distribución Normal**, de extraordinaria importancia en toda la Inferencia Estadística.

- Toda distribución Normal viene completamente caracterizada por su media y su desviación típica, es decir por sus parámetros de posición y de dispersión.

Histograma tipo de una distribución **normal**



Parámetros de asimetría y curtosis

- Se utilizan para estudiar la forma de la distribución de los datos.
- **Problema frecuente:** analizar si la distribución normal es adecuada para datos reales.
- Este es el objetivo de los coeficientes de asimetría y curtosis.
- Pautas de variabilidad que se alejen sensiblemente de la Normal pueden exigir el recurso a tratamientos estadísticos especiales o ser el síntoma de anomalías en los datos.

Coeficiente de asimetría

$$\text{Coeficiente de Asimetría CA} = \frac{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^3 /(N-1)}{s^3}$$

Adimensional

CA=0 datos simétricos

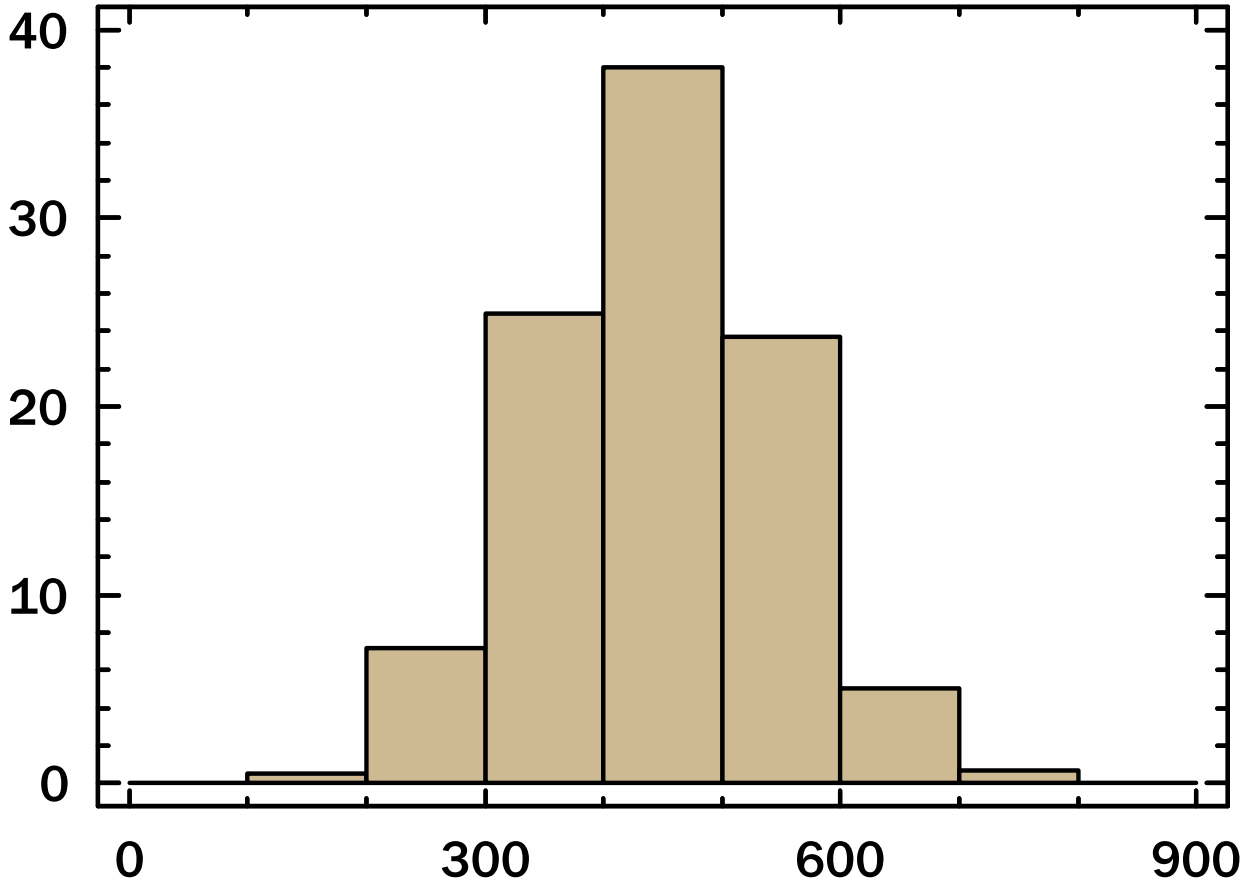
CA>0 datos con asimetría positiva

CA<0 datos con asimetría negativa

Coeficiente de asimetría

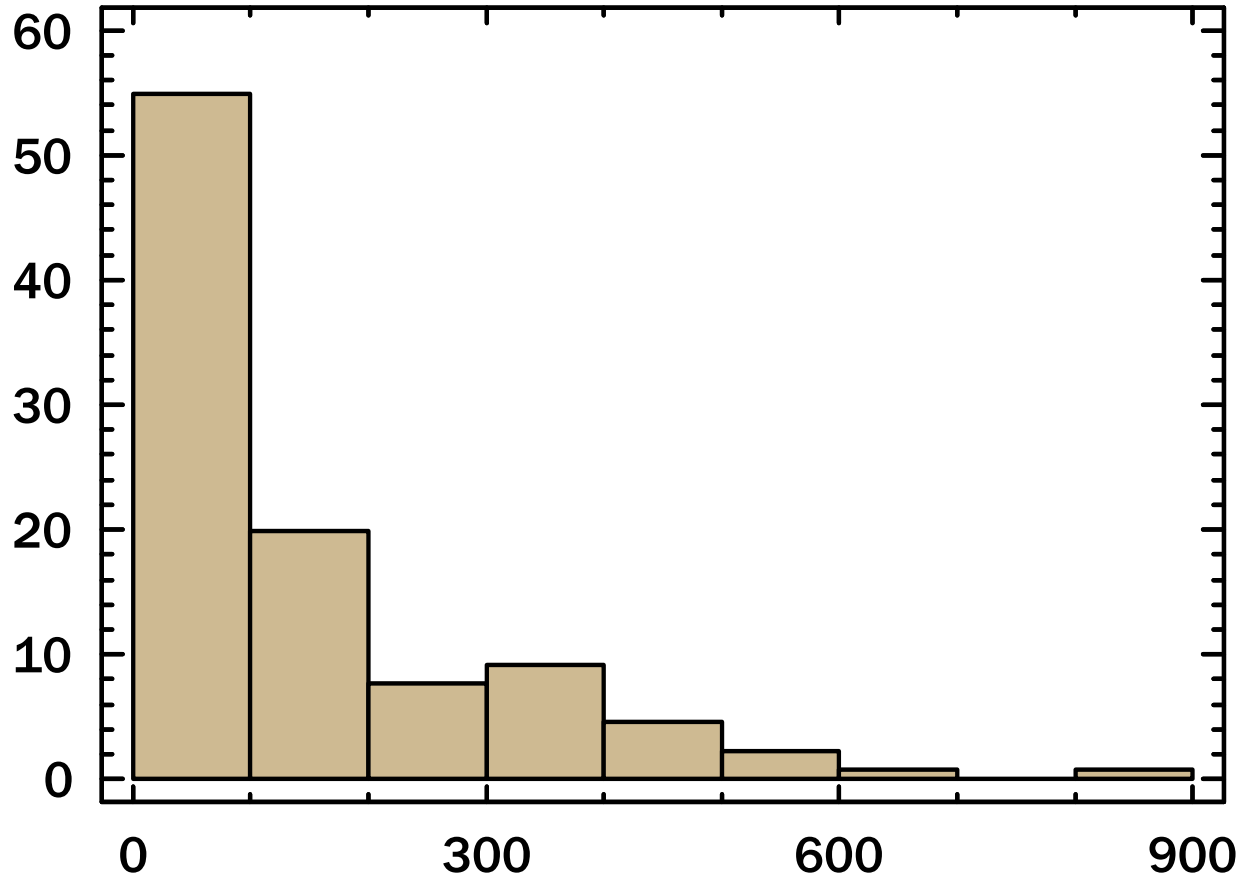
- En contextos inferenciales, cuando el objetivo es analizar hasta qué punto es verosímil que la muestra observada procede de una población en la que la variable sigue una distribución normal, se utiliza el **coeficiente de asimetría estandarizado (CAE)**.
- En muestras que proceden de poblaciones normales, el CAE está comprendido (en el 95% de los casos) entre -2 y 2.

CA≈0 CAE en el intervalo (-2,2)
Media=Mediana Datos simétricos



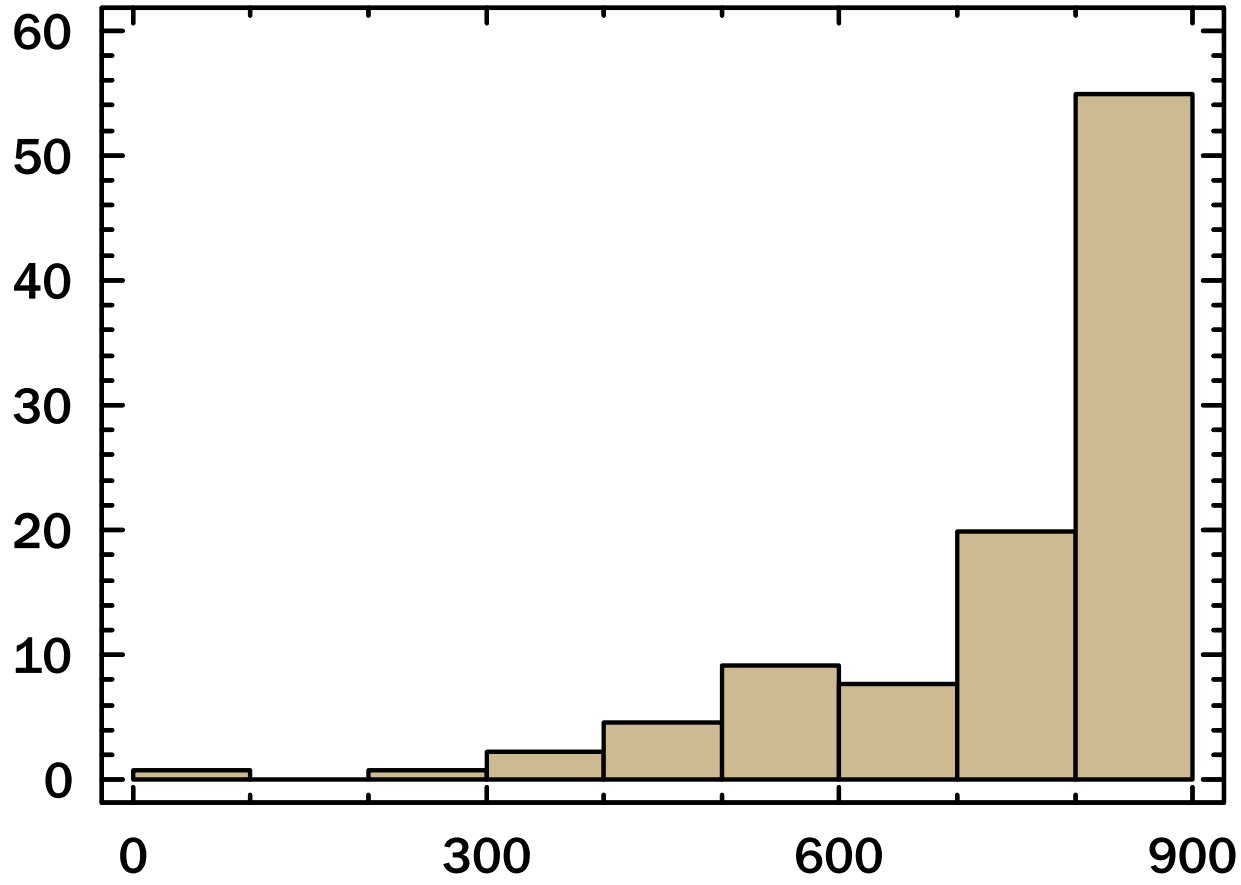
CA > 0 CAE >2 Mediana<Media

Datos con asimetría positiva



CA< 0 CAE<-2 Mediana>Media

Datos con asimetría negativa



Coeficiente de curtosis

Mide lo frecuentes que son valores alejados de la media.

Se toma como referencia la distribución normal o campana de Gauss

$$CC = \frac{\sum_{i=1}^{N} \left(x_i - \overline{X}\right)^4 / (N-1)}{s^4} - 3$$

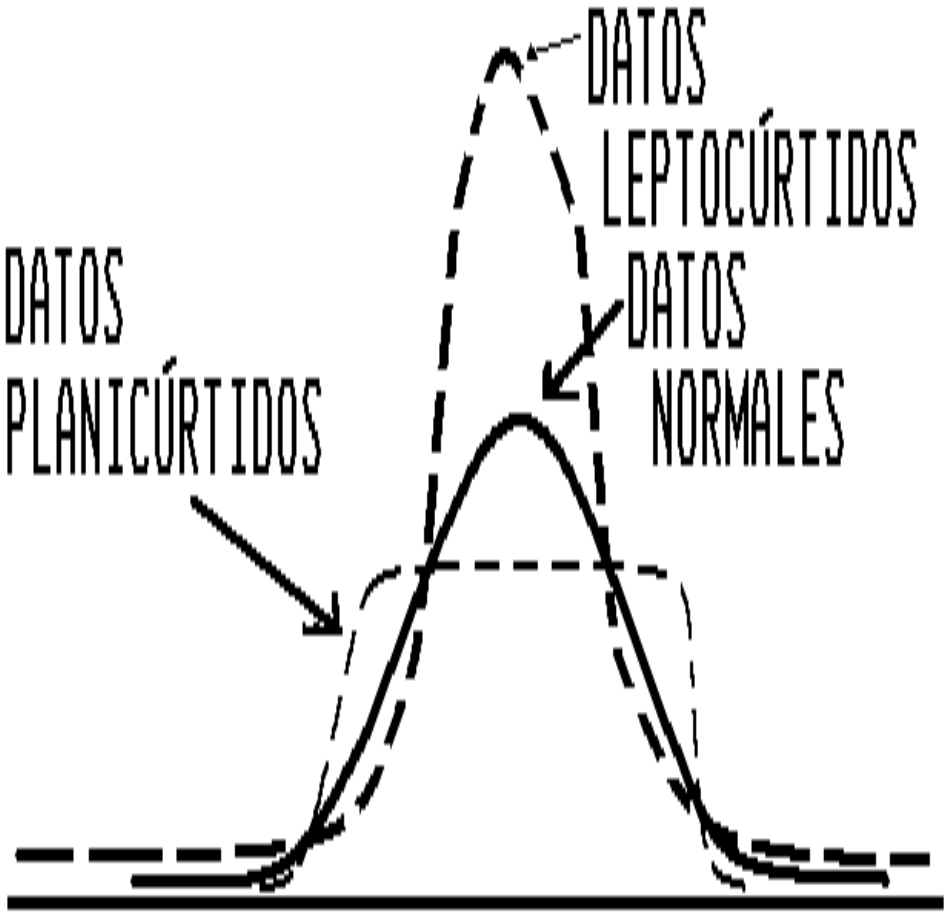
Coeficiente de curtosis

CC=0 Datos normales o mesocúrticos

CC>0 datos con mayores frecuencias en las "colas" de la distribución que una distribución normal. Leptocúrticos

CC<0 datos con menores frecuencias en las "colas" que una distribución normal. Planicúrticos

Coeficiente de curtosis



Coeficiente de curtosis

- Al igual que sucedía para el coeficiente de asimetría, en contextos inferenciales, cuando el objetivo es analizar hasta qué punto es verosímil que la muestra observada proceda de una población en la que la variable sigue una distribución normal, se utiliza el **coeficiente de curtosis estandarizado (CCE)**.
- En muestras que proceden de poblaciones normales, el CCE está comprendido (en el 95% de los casos) entre -2 y 2.

Coeficiente de curtosis

- Si CCE está en el intervalo (-2,2)= datos mesocúrticos
- Si CCE > 2 datos leptocúrticos
- Si CCE<-2 datos planicúrticos.

Parámetros de asimetría y curtosis

EJERCICIO 19: *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.

SOLUCIÓN: *La tabla siguiente recoge la asimetría y curtosis de ESTATURA para chicos y chicas. Para chicas son mayores los valores de estos coeficientes. En ambos casos indican distribución distinta de la normal.*

Parámetros de asimetría y curtosis

	ESTATURA	
CHICOS	CA=0,88 CAE=3,38>2 Asimetría positiva	CC=0,98 CCE=1,89
CHICAS	CA=1,29 CAE=3,42>2 Asimetría positiva	CC=4,3 CCE=5,69>2 Datos leptocúrticos

Parámetros de asimetría y curtosis

EJERCICIO 19: *Obtener también dichos coeficientes para la variable TIEMPO.*

SOLUCIÓN: *Para TIEMPO estos parámetros resultan:*

CA=1,26 CAE=5,91 CC=1,42 CCE=3,31

Por tanto esta variable no se distribuye normalmente: es asimétrica positiva y leptocúrtica.

Diagrama box-whisker (caja-bigote)

Representa:

- C1, Mediana y C3 para la caja

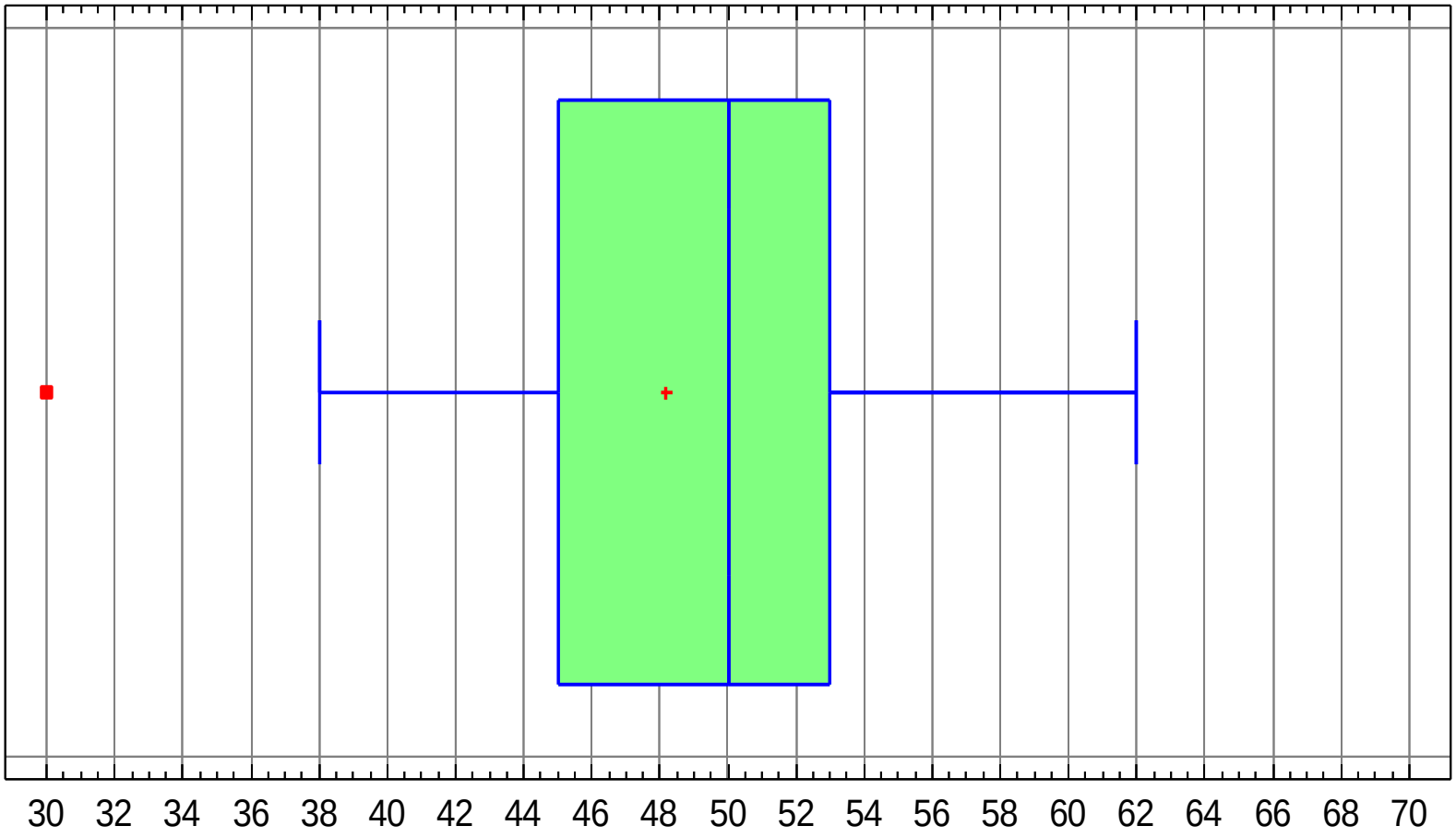
- Mínimo y máximo para los bigotes

- Los valores extremos que disten del cuartil más cercano más de 1,5 veces el intervalo intercuartílico se grafican como puntos aislados.

Diagrama box-whisker (caja-bigote)

- La figura adjunta refleja un diagrama Box-Whisker para los valores del Tiempo de funcionamiento sin averías.
- La "caja" comprende el 50% de los valores centrales de los datos, extendiéndose entre el primer cuartil y tercer cuartil (45 y 53 en la figura).
- La línea central corresponde a la mediana (50 en la figura).
- Los "bigotes" se extienden desde el menor (38) al mayor (62) de los valores observados y considerados "normales".
- Aquellos valores extremos que difieren del cuartil más próximo en más de 1,5 veces el recorrrido intercuartílico (1,5x8=12), se grafican como puntos aislados (como sucede en la figura con el valor 30<45-12) por considerar que pueden corresponder a datos anómalos ("outliers" en la terminología estadística).

Diagrama box-whisker (caja-bigote)



Tiempo de funcionamiento sin averías

Diagrama box-whisker (caja-bigote)

- Se observa que 30 es un dato anómalo en la muestra de tiempo de funcionamiento sin averías.
- Como no está compensando por ningún dato anómalo a la derecha, este dato 30 hace que la media=48,18<mediana=50

Diagrama box-whisker (caja-bigote) múltiple

Sirve para comparar la distribución de una variable cuantitativa según valores de una cualitativa

***EJERCICIO 18**: Compara la distribución de la ESTATURA (variable continua) entre chicos y chicas (variable cualitativa) mediante los diagramas Box-Whisker correspondientes.*

Diagrama box-whisker (caja-bigote) múltiple

SOLUCIÓN:

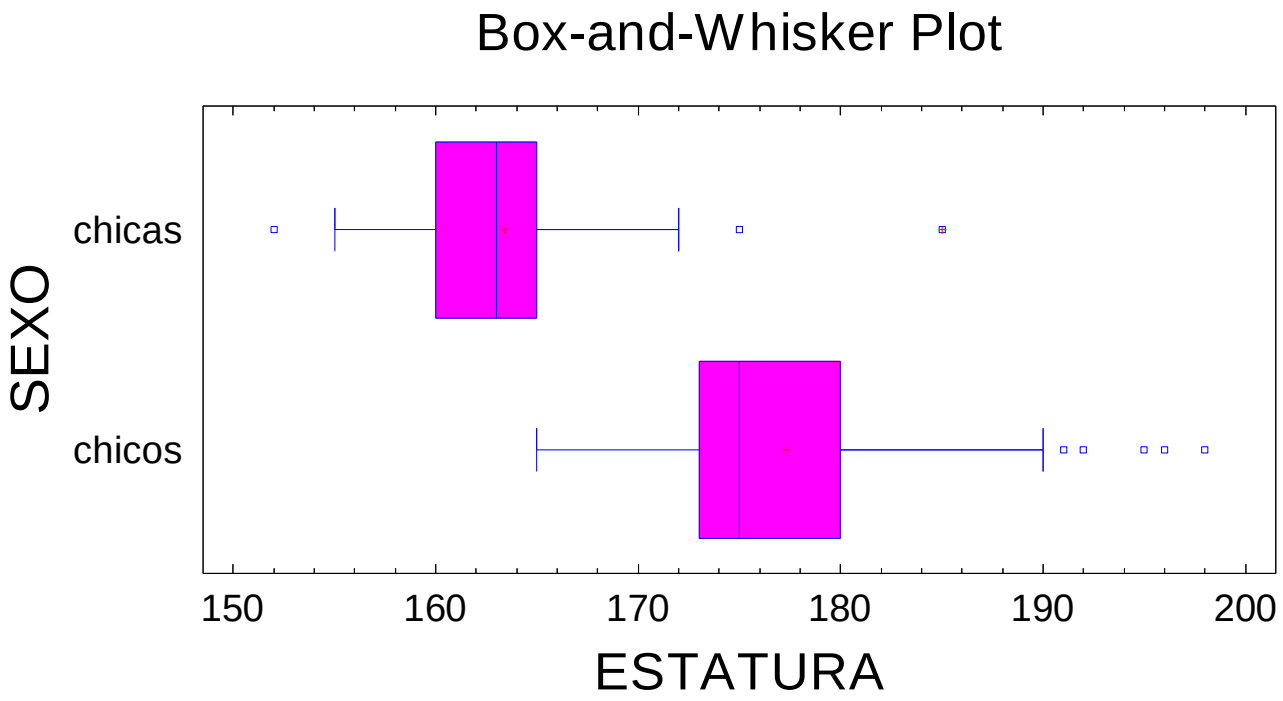


Diagrama box-whisker (caja-bigote) múltiple

- Se observa que la caja de las estaturas de chicos ocupa una posición mayor que la de las chicas (cuartiles C1, C3 y mediana). La media de estatura también es mayor en los chicos que en las chicas. Por tanto los chicos tienen más posición central de estatura que las chicas.
- La dispersión medida con el rango intercuartílico C3-C1 es también mayor en los chicos que en las chicas.

Diagrama box-whisker (caja-bigote) múltiple

- Hay asimetría positiva de la estatura tanto en chicas como en chicos, ya que en ambos casos la distancia entre mínimo y mediana es menor que la distancia entre mediana y máximo.
- En ambos grupos hay valores anómalos de estatura.