Numerical summary of data

- Measures of location: mode, median, mean, …
- Measures of spread: range, interquartile range, standard deviation, …
- Measures of form: skewness, kurtosis, …

Measures of location

There are 3 commonly used measures: the mode, the median and the mean.

The following data are the number of years spent as mayor by the last 24 mayors of Madrid (up to 2009)

3	1	1	1	1	1	2	1
7	6	13	8	3	2	1	1
2	1	1	7	3	2	12	6

The mode

… is the most frequent value

Clase	*Frecuencia*
1	10
2	4
3	3
4	0
5	0
6	2
7	2
8	1
9	0
10	0
11	0
12	1
13	1
y mayor...	0

Can we calculate the mode with qualitative data?

Does this definition make sense with continuous data?

There can be more than one mode: bimodal-trimodal-multimodal

The mode for (continuous) grouped data

Money received (millions PTAS)	Absolute frequency
≤ 30	0
(30,45]	2
(45,60]	9
(60,75]	9
(75,90]	10
(90,105]	3
(105,120]	3
> 120	0
Total	60

We have a modal class

What if the classes have different widths?

An exact formula for the mode of grouped data

The median

… is the most central datum.

5 3 11 21 7 5 2 1 3

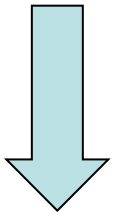
What is the value of the median?

Can we calculate the median with qualitative data?

What is the difference if N is odd or even?

The mayors

3	1	1	1	1	1	2	1
7	6	13	8	3	2	1	1
2	1	1	7	3	2	12	6



1	1	1	1	1	1	1	1
1	1	2	**2**	**2**	2	3	3
3	6	6	7	7	8	12	13

The median is ½ (2+2)=2

The median via the table of frequencies (discrete data)

Median →

x_i	n_i	N_i	f_i	F_i	
1	10	10	0,41666667	0,41666667	<0,5
2	4	14	0,16666667	0,58333333	>0,5
3	3	17	0,125	0,70833333	
4	0	17	0	0,70833333	
5	0	17	0	0,70833333	
6	2	19	0,08333333	0,79166667	
7	2	21	0,08333333	0,875	
8	1	22	0,04166667	0,91666667	
9	0	22	0	0,91666667	
10	0	22	0	0,91666667	
11	0	22	0	0,91666667	
12	1	23	0,04166667	0,95833333	
13	1	24	0,04166667	1	
y mayor...	0	24	0	1	

The median of grouped (continuous) data

Median interval →

Money received	n_i	N_i	f_i	F_i
≤ 30	0	0	0	0
(30,45]	2	2	0,05555556	0,05555556
(45,60]	9	11	0,25	0,30555556
(60,75]	9	20	0,25	0,55555556
(75,90]	10	30	0,27777778	0,83333333
(90,105]	3	33	0,08333333	0,91666667
(105,120]	3	36	0,08333333	1
> 120	0	36	0	1
Total	36		1	

An exact formula for the median of grouped data

The mean

The mean or arithmetic mean is the average of all the data.

For the mayors, the sum of the data is …

$$3 + 1 + 1 + 1 + 1 + 1 + 2 + 1$$
$$7 + 6 + 13 + 8 + 3 + 2 + 1 + 1$$
$$2 + 1 + 1 + 7 + 3 + 2 + 12 + 6$$
$$= 86$$

… and therefore, the mean is 86/24 ≈ 3,583 years.

$$\bar{x} = \frac{1}{N}(x_1 + x_2 + \cdots + x_N) = \sum_{i=1}^{N} x_i$$

Can we calculate the mean for qualitative data?

The mean using the frequency table (discrete data)

x_i	n_i	$n_i * x_i$	
1	10	10	
2	4	8	
3	3	9	
4	0	0	
5	0	0	
6	2	12	
7	2	14	
8	1	8	
9	0	0	
10	0	0	
11	0	0	
12	1	12	
13	1	13	
y mayor …	0	0	
Total	24	86	3,58333333

The formula

For data x_1, …, x_k with absolute relative frequencies n_1, …, n_k such that $n_1 + \ldots + n_k = N$:

$$
\begin{aligned}
\bar{x} &= \frac{1}{N}(n_1 \times x_1 + n_2 \times x_2 + \cdots + n_k \times x_k) \\
&= \frac{1}{N}\sum_{i=1}^{k} n_i \times x_i \\
&= \sum_{i=1}^{k} \frac{n_i}{N} \times x_i \\
&= \sum_{i=1}^{k} f_i \times x_i
\end{aligned}
$$

The mean with grouped data

Ingresos	x_i	n_i	x_i*n_i	
<= 30	22,5	0	0	
(30,45]	37,5	2	75	
(45,60]	52,5	9	472,5	
(60,75]	67,5	9	607,5	
(75,90]	82,5	10	825	
(90,105]	97,5	3	292,5	
(105,120]	112,5	3	337,5	
> 120	127,5	0	0	
Total		36	2610	72,5

This is the same formula but using the centre of each interval.

The mode, median and mean for asymmetric data

Which is most sensitive to outliers?

Other points of the distribution: minimum, maximum, quartiles and quantiles

Ordering the data, the minimum and maximum are easy to calculate.

1	1	1	1	1	1	1	1	1
1	2	2	2	2	3	3	3	6
6	7	7	8	12	**13**			

What about the quartiles?

The idea is to divide the data into quarters

Q_0 = minimum 0%

$Q_1 = x_{(n+1)/4}$ 25%

Q_2 = median 50%

$Q_3 = x_{3(n+1)/4}$ 75%

Q_4 = maximum 100%

Here, n = 24. Therefore, (n+1)/4 = 6.25.

There is no point $x_{6.25}$

We need to use interpolation.

$$x_6 = 1, x_7 = 1$$

$$x_{6.25} = x_6 + 0.25\,(x_7 - x_6) = 1$$

What about Q_3?

A more general concept is the p – quantile or 100 p % percentile. The idea is to divide the data into fractions of size p and 1-p.

This is defined as $x_{p(n+1)}$.

What is the 90% percentile?

Warning: there are many (slightly) different ways of defining quantiles.

Measures of spread

There are various measures:

- The range
- The interquartile range
- The standard deviation
- The coefficient of variation

The range and interquartile range

The range is defined as the difference between the maximum and minimum of the data.

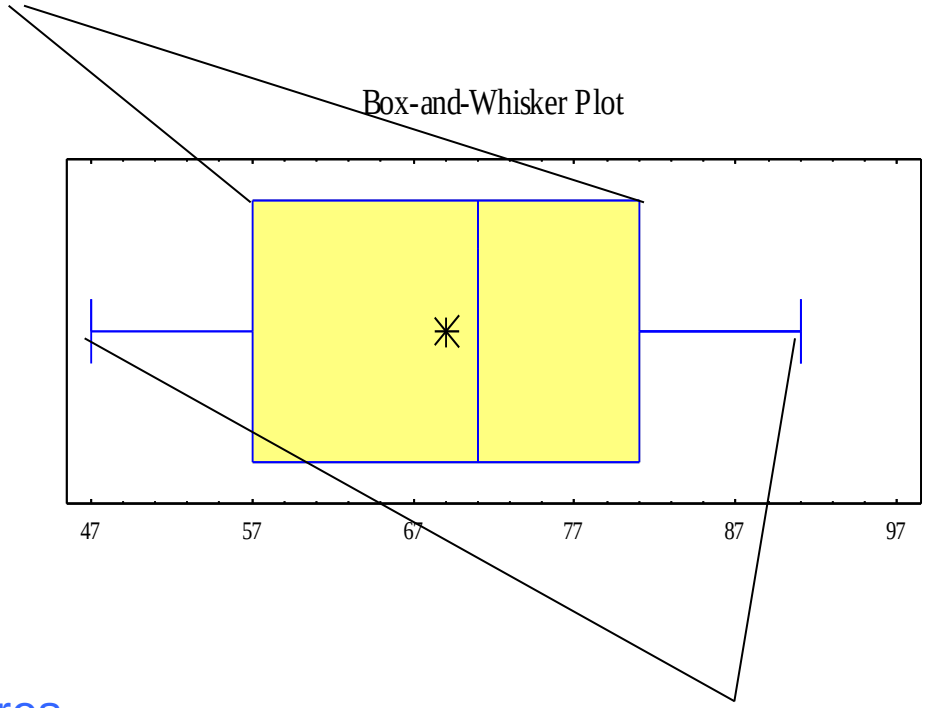
The interquartile range is Q_3-Q_1.

Calculate the range and interquartile range in the previous example.

Which of the two measures is more sensitive to outliers?

The box and whisker plot

The interquartile range



The range

Calculate the range and interquartile range in the previous examples.

Which of the two measures is more sensitive to outliers?

The variance and standard deviation

We could look at the distance of each observation from the mean

Empresa A	$x_i - \overline{X}$	Empresa B	$x_i - \overline{X}$
30700	**-2800**	27500	**-6000**
32500	**-1000**	31600	**-1900**
32900	**-600**	31700	**-1800**
33800	**300**	33800	**300**
34100	**600**	34000	**500**
34500	**1000**	35300	**1800**
36000	**2500**	40600	**7100**

What do these new columns sum to?

$$
\begin{aligned}
\sum_{i=1}^{n}(x_i-\bar{x}) &= (x_1-\bar{x})+(x_2-\bar{x})+\cdots+(x_n-\bar{x}) \\
&= x_1+x_2+\cdots+x_n-\underbrace{\bar{x}-\bar{x}-\cdots-\bar{x}}_{n \text{ veces}} \\
&= x_1+x_2+\cdots+x_n-n\bar{x} \\
&= n\bar{x}-n\bar{x} \\
&= 0 \qquad \Rightarrow \\
\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}) &= 0
\end{aligned}
$$

How can we resolve the problem?

The variance …

… is the mean squared distance

Empresa A		Empresa B	
30700	**7840000**	27500	**36000000**
32500	**1000000**	31600	**3610000**
32900	**360000**	31700	**3240000**
33800	**90000**	33800	**90000**
34100	**360000**	34000	**3240000**
34500	**1000000**	35300	**250000**
36000	**6250000**	40600	**50410000**
	16900000		**96840000**

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

What are the units of the variance? Can we change them?

The standard deviation

… is the square root of the variance. It is something like the typical distance of an observation from the mean.

Empresa A	s = 4110,9
Empresa B	s = 9840,7

Which is more sensitive to outliers. The standard deviation or the interquartile range?

What happens if we change the units of the data?

The coefficient of variation

When the mean is different to 0 we can calculate a normalized measure of spread.

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

This lets us compare two groups as it has no units.

Is it useful with a single set of data?

Exercise

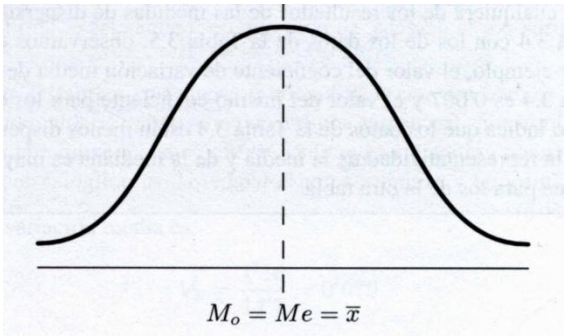
We analyzed the amount of books taken out during the exam period in 10 university libraries, and this was compared with the previous year. The % increase was:

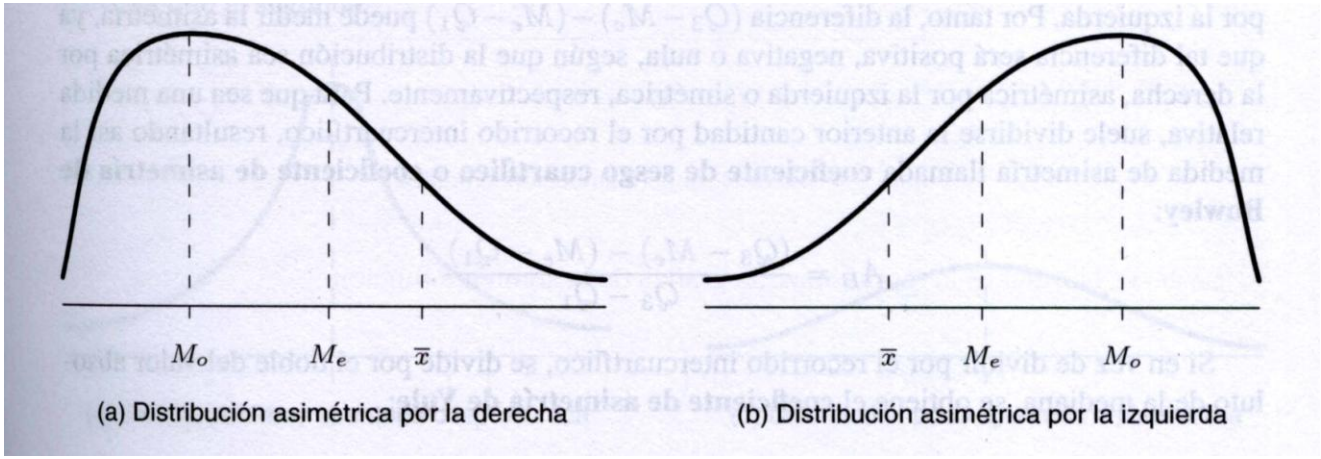
10.2	2.9	3.1	6.8	5.9
7.3	7.0	8.2	3.7	4.3

Are these data homogeneous?

Measures of form

The most commonly used measures are skewness (or asymmetry) and kurtosis.





Symmetric, right skewed and left skewed data.

Pearson's coefficient of skewness

CA=0 Symmetric
CA>0 Asymmetric to the right
CA<0 Asymmetric to the left

$$CA = \frac{\bar{x} - M_o}{s}$$

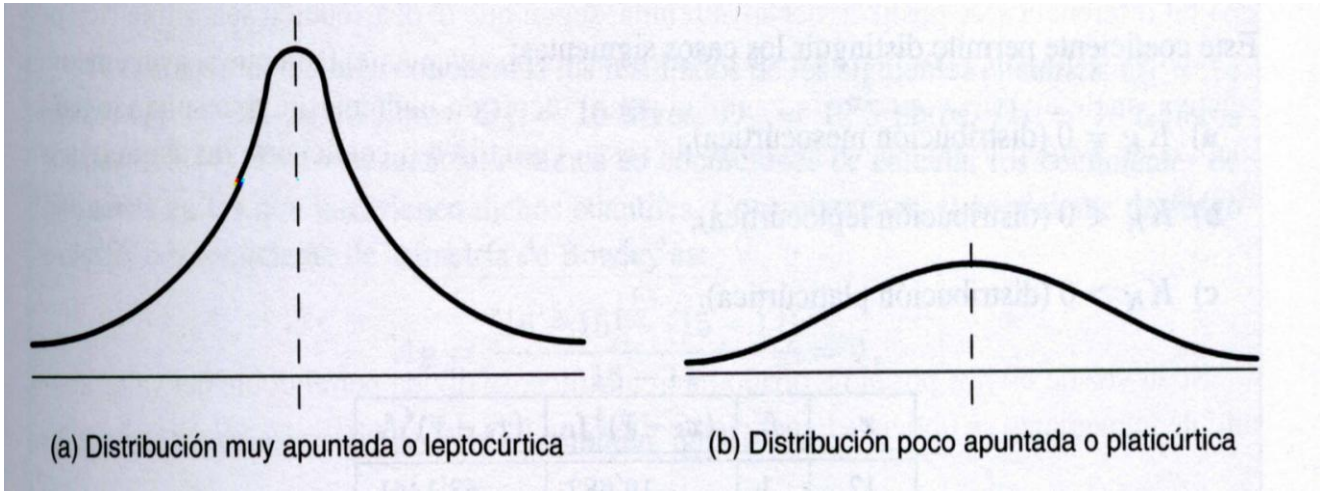
Fisher's coefficient of skewness

(used when the data are multimodal):

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

Kurtosis

We can see this graphically by comparing with a normal distribution.



Fisher's coefficient of kurtosis

$$CC = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{N s^4} - 3$$

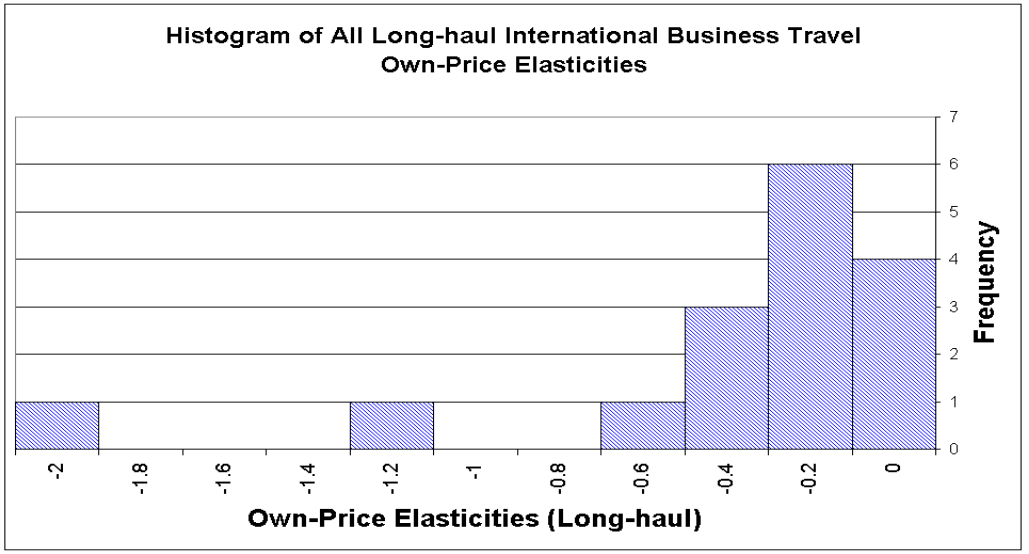
CC = 0 (mesokurtic)

CC > 0 (leptokurtic)

CC < 0 (platykurtic)

Exercise

The following histogram shows the elasticity of demand for long haul flights.



Which of the following affirmations is correct?

a) The standard deviation is 10.
b) The mean is higher than the median which is higher than the mode.
c) The mean is 1.
d) The mode is higher than the median which is higher than the mean.

Exercise

The table shows the ages and sex of different government ministers.

Name	Sex	Ministry	Age
Bibiana Aído	M	Igualdad	33
Carme Chacón	M	Defensa	38
Ángeles González-Sinde	M	Cultura	44
Cristina Garmendia	M	Ciencia e innovación	47
Trinidad Jiménez	M	Sanidad y Política Social	47
José Blanco	V	Fomento	48
Ángel Gabilondo	V	Educación	60
Elena Salgado	M	Economía y Hacienda	60

Which of the following affirmations is correct?

a) The range of ages is 33 and the absolute frequency of women is 6.
b) The mean age is 47 and the percentage of male ministers is 25%.
c) The first quartile of the ages is 39.5 and the third quartile is 57.
d) The modal age is 60 and the mean is 47.

Exercise

A simple of 10 Madrileños was taken and the sampled subjects were asked how many hours they worked every week. The results are as follows:

40	40	35	50	50	40	40	60	50	35

Select the correct solution from the following:

a) The mean and mode are 40 and the median is 44.
b) The mean and median are equal to 40 and the mode is 44.
c) The mode and median are 40 and the mean is 44.
d) None of the above is correct.

Exercise

At the end of 2009, the mean monthly wage in Spain was 1.993,15 euros. Suppose that the standard deviation was 180 euros. Given an exchange rate of 6 euros = 1000 PTAS, then:

a) The mean wage was 11959,0 thousands of PTAS and the standard deviation was 1080 thousands of PTAS.
b) The mean wage was 332,19 thousand PTAS and the standard deviation was 180 PTAS.
c) The mean wage was 1993,15 PTAS and the standard deviation was 30 thousand PTAS.
d) The mean wage was 332,19 thousand PTAS and the standard deviation was 30 thousand PTAS.