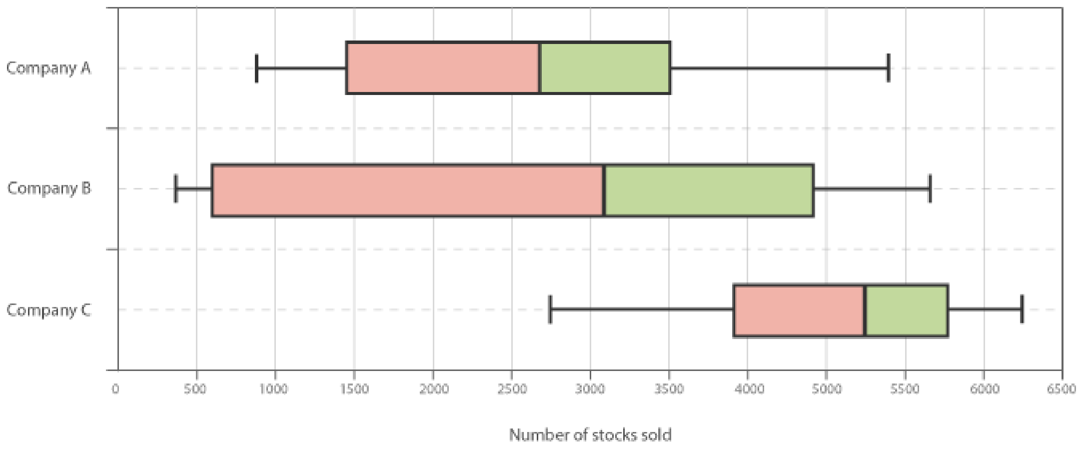
Class 5: Analysis of univariate data: measures of dispersion and other measures of shape



Measures of dispersion

Various measures of dispersión could be considered:

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

Recommended reading:

The Wikipedia page on dispersion has links to some of the relevant quantities.

The range

The range of a data set is simply the distance between the maximum and mínimum values

Times voted	Absolute frequency (n)
0	4
1	10
2	12
3	15
4	11
5	5
6	1
7	1
8	1
>8	0
Total	60

Range = 8.

What would happen if the last observation was 800 instead of 8?

Is the range a good measure of typical dispersion.?

The interquartile range

This is the distance between the first and third quartiles of a sample of data.

1 2 3 3 5 5 7 11 21

$Q_1 = 2.5$

$Q_3 = 9$

IQR = 9 – 2.5 = 6.5

Half the interquartile range is called the *semi-interquartile range*. This is sometimes used for comparison with the standard deviation (see later).

Looking for strange or outlying data using the IQR

If an observation is below the inner fence Q_1 – 1.5 IQR, (or above Q_3 + 1.5 IQR) it is called a *mild outlier*.

If an observation is below the outer fence Q_1 – 3 IQR (or above Q_3 + 3 IQR) it is an *extreme outlier.*

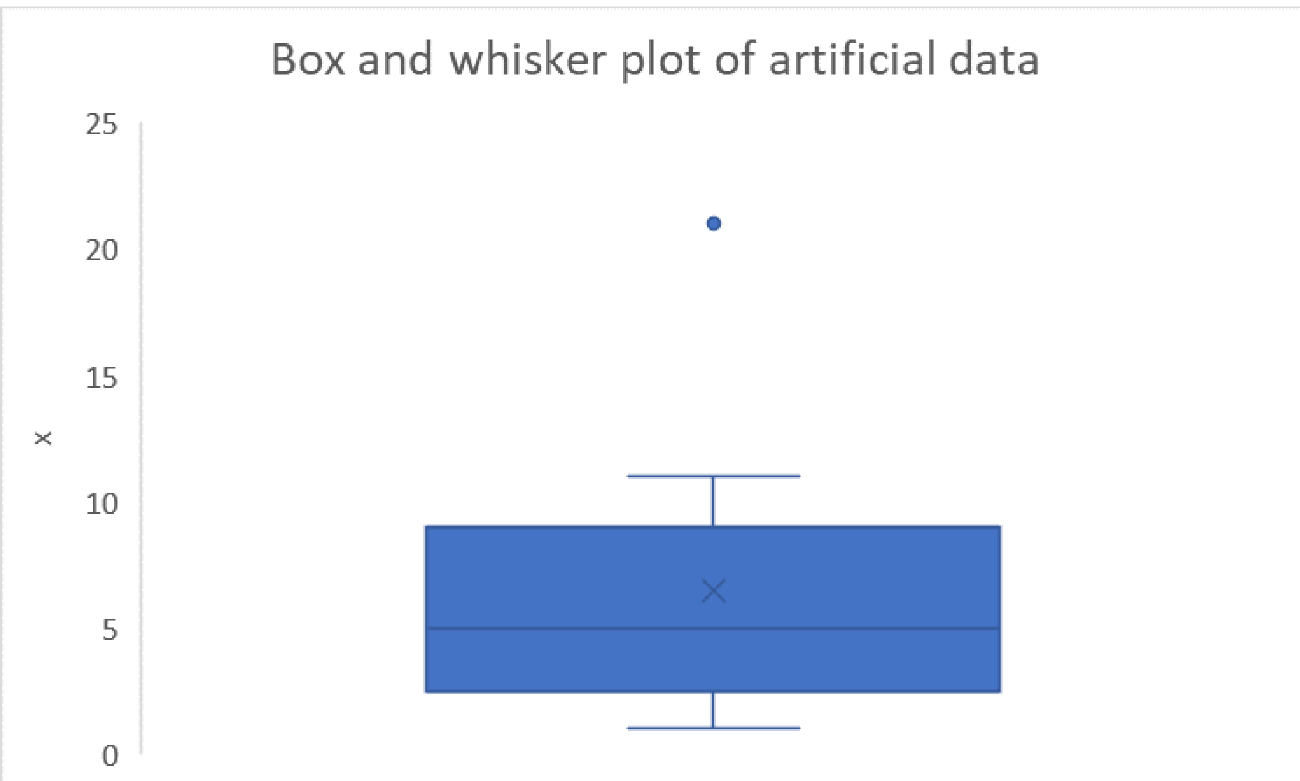
1 2 3 3 5 5 7 11 21

21 > 9 + 1.5 x 6.5 = 18.75 is the upper inner fence.
21 < 9 + 3 x 6.5 = 24.5 is the upper outer fence.

21 is a mild outlier.

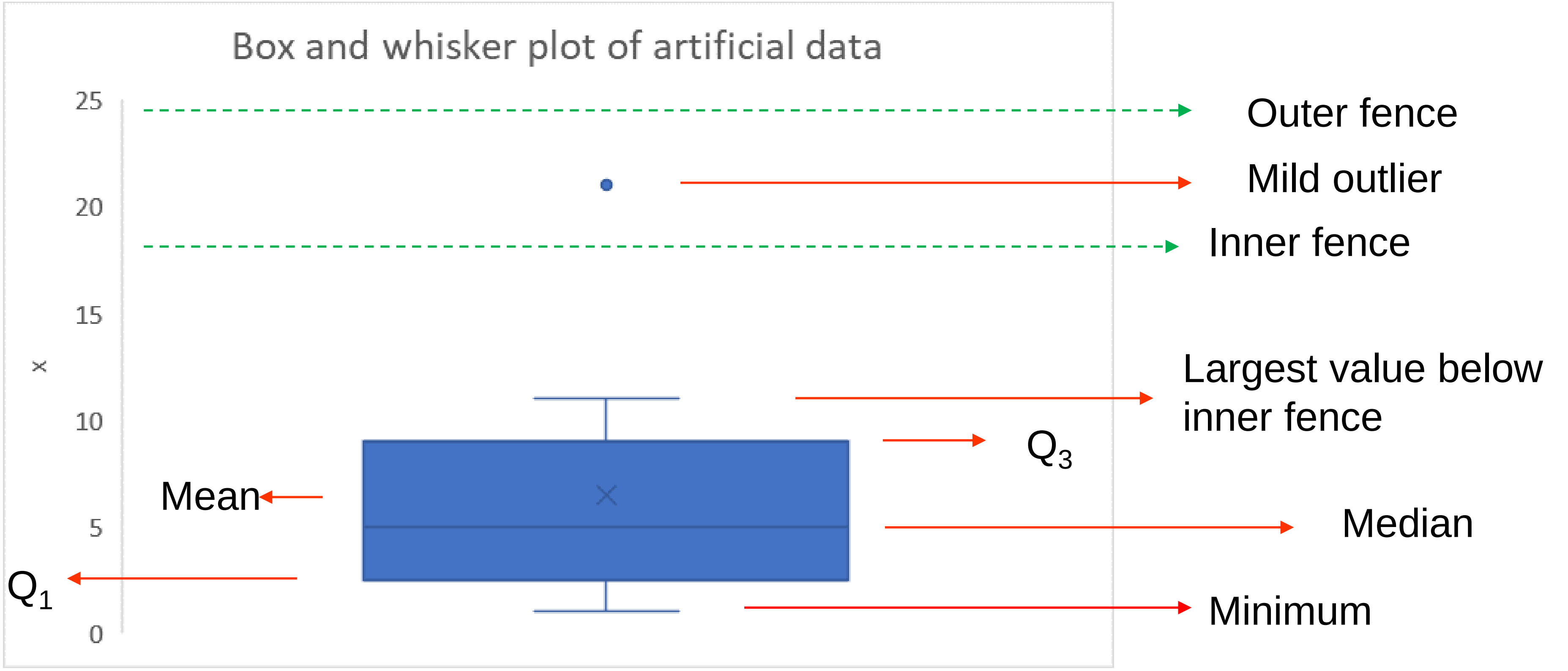
The box and whisker plot

This is a way of visualizing the shape of a data set and any outliers.

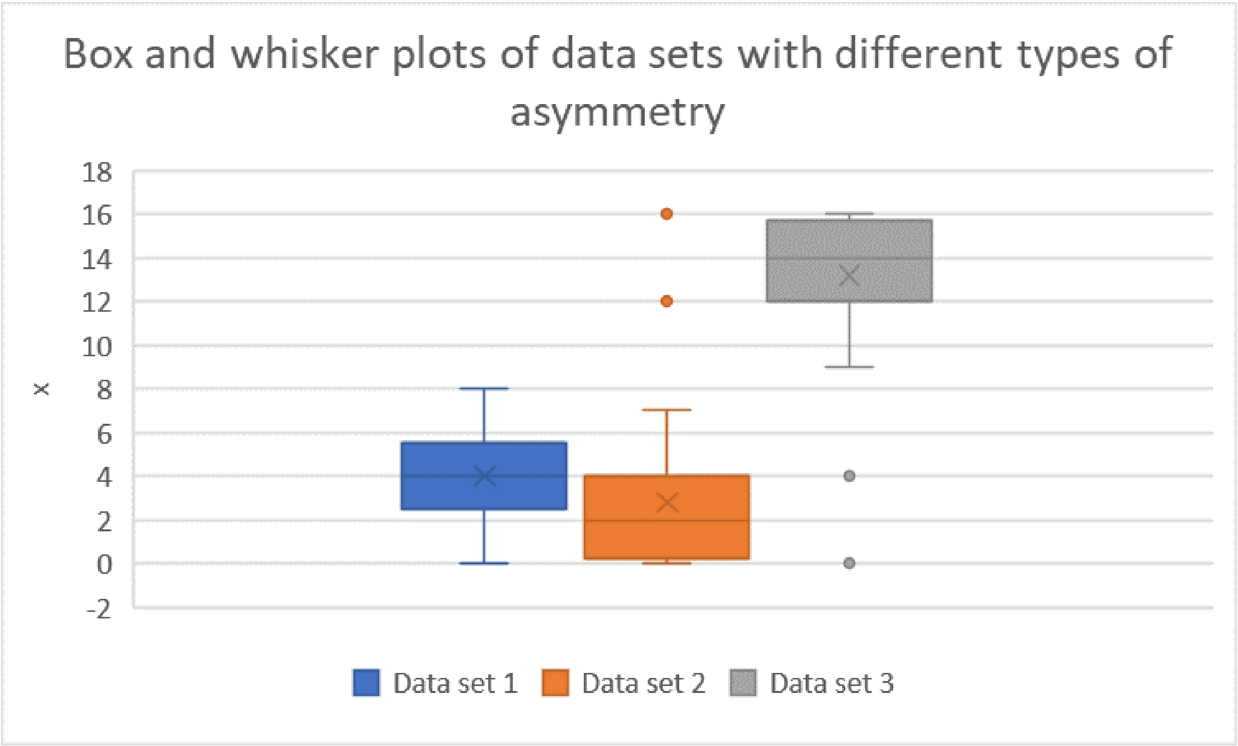


The box and whisker plot

This is a way of visualizing the shape of a data set and any outliers.

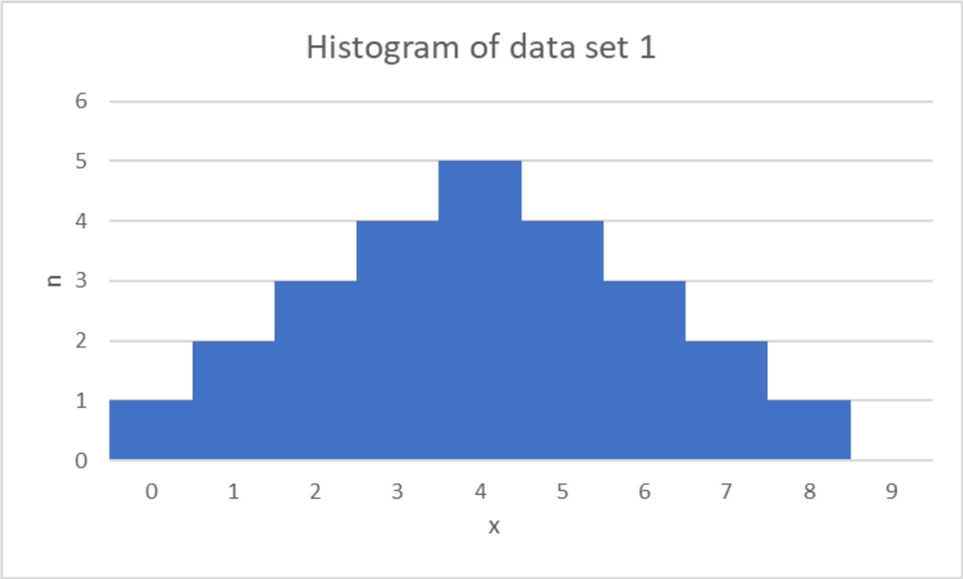


Assessing the shape of a data set with a box and whisker plot



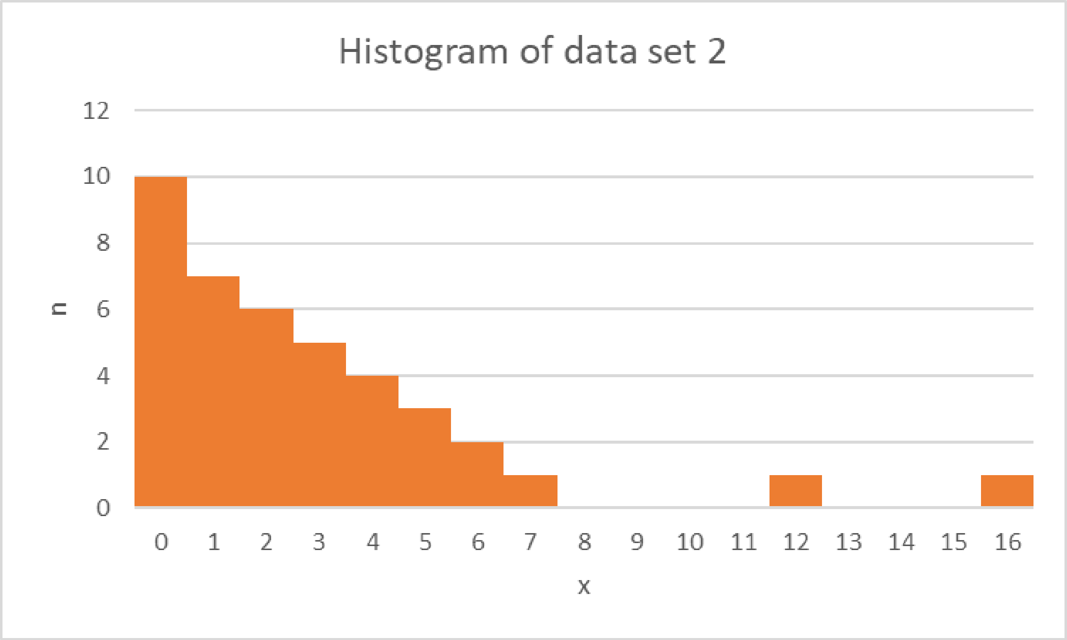
What would the histograms look like?

Assessing the shape of a data set with a box and whisker plot



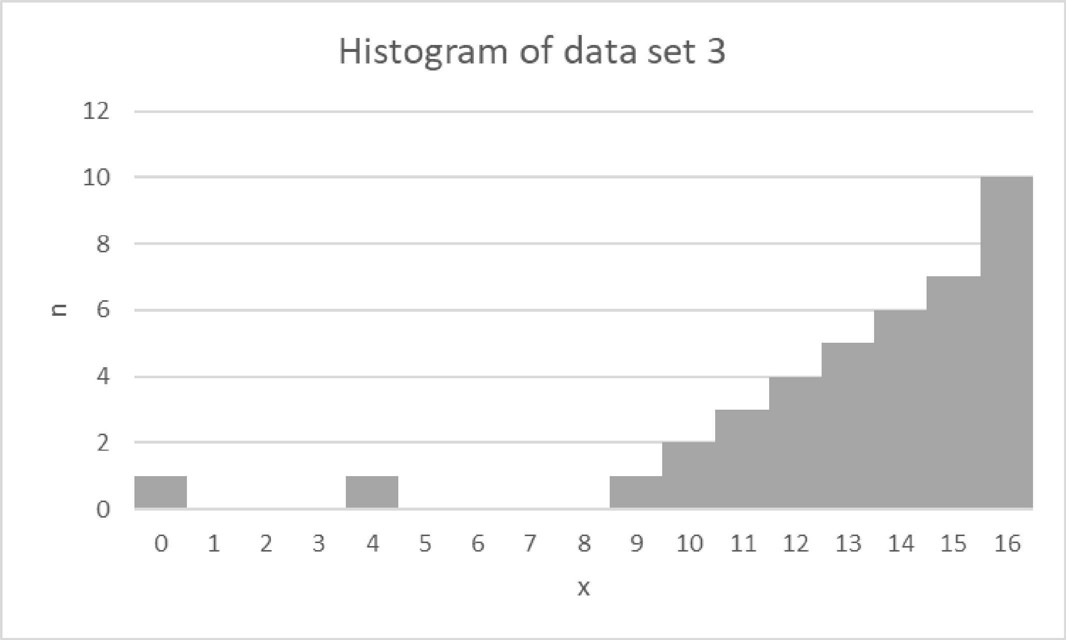
Symmetrical data.

Assessing the shape of a data set with a box and whisker plot



Data skewed to the right.

Assessing the shape of a data set with a box and whisker plot



Data skewed to the left

Measures of spread associated with the mean: variance and standard deviation

We want to find a measure of the “typical distance from the mean”.

Why not use the average distance?

$$\frac{1}{N}\{(x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_N - \bar{x})\}$$

Measures of spread associated with the mean: variance and standard deviation

We want to find a measure of the "typical distance from the mean".

Why not use the average distance?

$$\frac{1}{N}\{(x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_N - \bar{x})\}$$

1 2 3 3 5 5 7 11 21

Mean: $\bar{x}$ = 58/9 = 6.44

Differences: -5.44, -4.44, -3.44, -3.44 -1.44, -1.44, 0.56, 4.56, 14.56

Sum of differences = -5.44 -4.44 + … + 14.56 = 0.

The average distance is always 0! Not a sensible estimate of typical distance.

The variance

We need to get rid of negative distances.

One reasonable idea is: $\frac{1}{N}\{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \cdots + |x_N - \bar{x}|\}$.

This isn't used much in practice (odd statistical properties).

A more popular alternative is the variance:

$$\widehat{\sigma}^2 = \frac{1}{N}\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_N - \bar{x})^2\}$$

Differences: -5.44, -4.44, -3.44, -3.44 -1.44, -1.44, 0.56, 4.56, 14.56

$\widehat{\sigma}^2 = \frac{1}{9}\{(-5.44)^2 + (-4.44)^2 + \cdots + 14.56^2\}$ = 34.47.

How can we interpret this number? If the data are measured in cm, what are the units of the variance?

The standard deviation

The standard deviation is the square root of the variance: $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$.

$\hat{\sigma} = \sqrt{34.47} = 5.87$.

This is (approximately) the typical distance of an observation from the mean.

Which measures are more sensitive to outliers: range, IQR or standard deviation?

The quasi variance and quasi standard deviation

When we want to use the variance of a sample to estimate the variance of a population, the quasi variance s^2 is typically used:

$$s^2 = \frac{1}{N-1}\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_N - \bar{x})^2\}$$

The quasi standard deviation is s = $\sqrt{s^2}$.

$s^2 = \frac{1}{8}\{(-5.44)^2 + (-4.44)^2 + \cdots + 14.56^2)$ = 38.78.

$s = 6.23$.

The quasi variance is a little bit higher than the variance.

Looking for outlying data: Chebyshev's inequality

THEOREM

For any sample of data, then for any k > 0, more than 100 $(1-1/k^2)$ % of the data are less than k standard deviations from the mean.

More than 75% of the data are less than 2 standard deviations from the mean.
More than 88.89% of the data are less than 3 standard deviations from the mean.
More than 93.75% of the data are less than 4 standard deviations from the mean.

1 2 3 3 5 5 7 11 21

$\bar{x}$ = 6.44, $\hat{\sigma}$ = 5.87.

$\bar{x} \pm 2\hat{\sigma} =$ (-5.30, 18.19) $\bar{x} \pm 3\hat{\sigma} =$ (-11.17, 24.06)

Chebyshev's inequality is very conservative!

Looking for outlying data: a "rule of thumb"

For "normal" symmetrical data, then:

Approximately 68% of the data are less than 1 standard deviation from the mean.
Approximately 95% of the data are less than 2 standard deviations from the mean.
Approximately 99.5% of the data are less than 3 standard deviations from the mean.
Approximately 99.99% of the data are less than 4 standard deviations from the mean.

1 2 3 3 5 5 7 11 21

$\bar{x}$ = 6.44, $\hat{\sigma}$ = 5.87.

$\bar{x} \pm \hat{\sigma}$ = (0.57, 12.32) $\bar{x} \pm 2\hat{\sigma}$ = (-5.30, 18.19) $\bar{x} \pm 3\hat{\sigma}$ = (-11.17, 24.06)

Comparing the variation in samples with very different means: the coefficient of variation

Assume that we wish to compare the variability of price of two different products in order to assess competitiveness in both markets. More variability should indicate more competition.

It is not always sensible to compare markets in terms of standard deviation directly.

Average milk price: €1 per litre, standard deviation: €0.15 per litre.
Average car Price: €13000, standard deviation: €1300.

Which market is more competitive?

The coefficient of variation

We should look at the variability with respect to the average price:

coefficient of variation = $s/|\bar{x}|$

What units is this measured in?

Average milk price: €1 per litre, standard deviation: €0.15 per litre.
Average car Price: €13000, standard deviation: €1300.

Which market is more competitive?

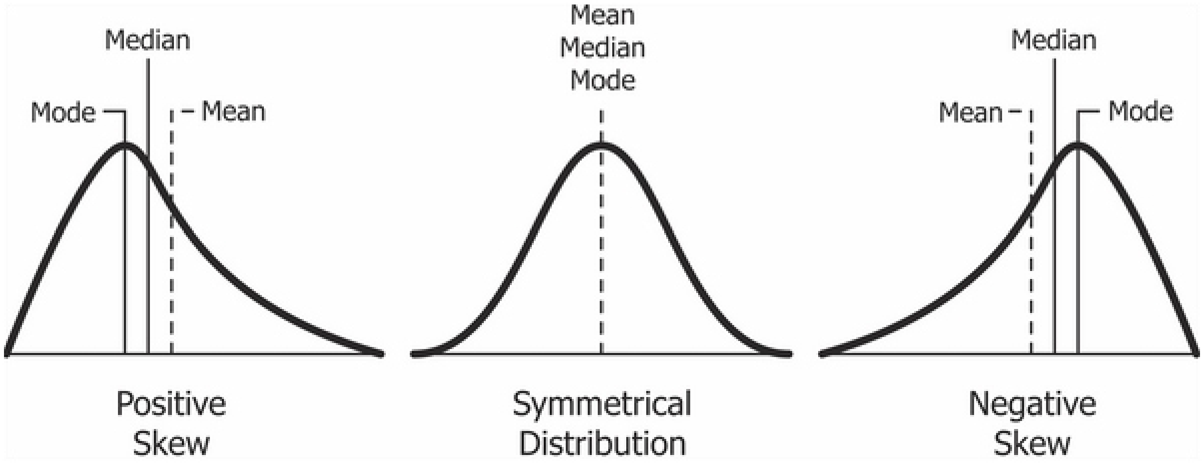
$CV_{milk} = 0.15/1 = 0.15.$

$CV_{cars} = 1300/13000 = 0.1.$

Relative to the average price, there is more competition in the milk market.

Other features of a sample of data: skewness

We are looking for a numerical measure of the asymmetry of a sample.

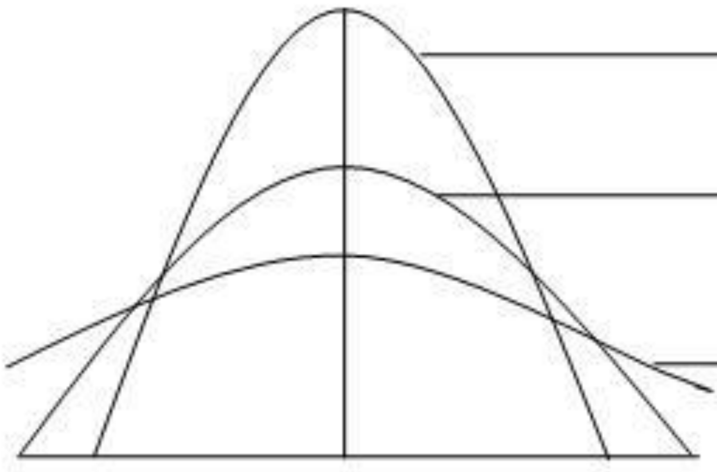


Fisher's skewness coefficient is:

$$\frac{1}{N\hat{\sigma}^3}\{(x_1-\bar{x})^3+(x_2-\bar{x})^3+\cdots+(x_N-\bar{x})^3\}$$

Other features of a sample of data: kurtosis

Kurtosis reflects the behaviour in the tails of a distribution.



"Normal", mesokurtic data has kurtosis close to 0.

The kurtosis can be calculated as:

$$\frac{1}{N\hat{\sigma}^4}\{(x_1 - \bar{x})^4 + (x_2 - \bar{x})^4 + \cdots + (x_N - \bar{x})^4\} - 3$$

Exercise (Exam)

The following tables gives the ages and gender of various ministers in the Zapatero government..

Nombre	Sexo	Ministerio	Edad
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

Mark the correct answer among the followiing?

a) The first quartile of the age distribution is 39.5 and the third quartile is 57.
b) The second quartile of the age distribution is 57 and 25% of the ministers are men.
c) The range of the age distribution is 33 and the absolute frequency of women is 6.
d) The mode of the ages is 60 y the mean is 47.

Exercise (Exam)

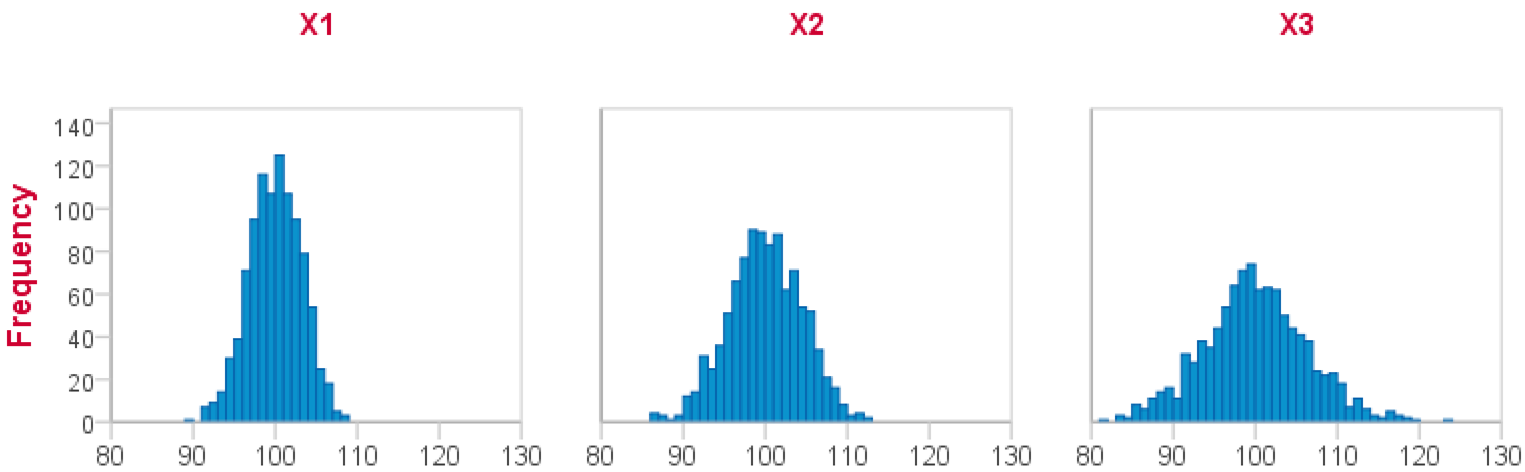
A company has two sections with 40 and 65 employees respectively. Their average weekly wages are €450 and €350. The standard deviations are €7 and €9 respectively.

(i) Which section has a larger wage bill?

(ii) Which section has larger relative variability in wages?

Exercise (Test)

The following histograms show the results of three samples of data on political leanings,



Mark the correct answer among the following.

a) The means of the three data sets are clearly different.
b) The standard deviation is lowest in data set X3.
c) The variance is lowest in data set X1.
d) None of the above.