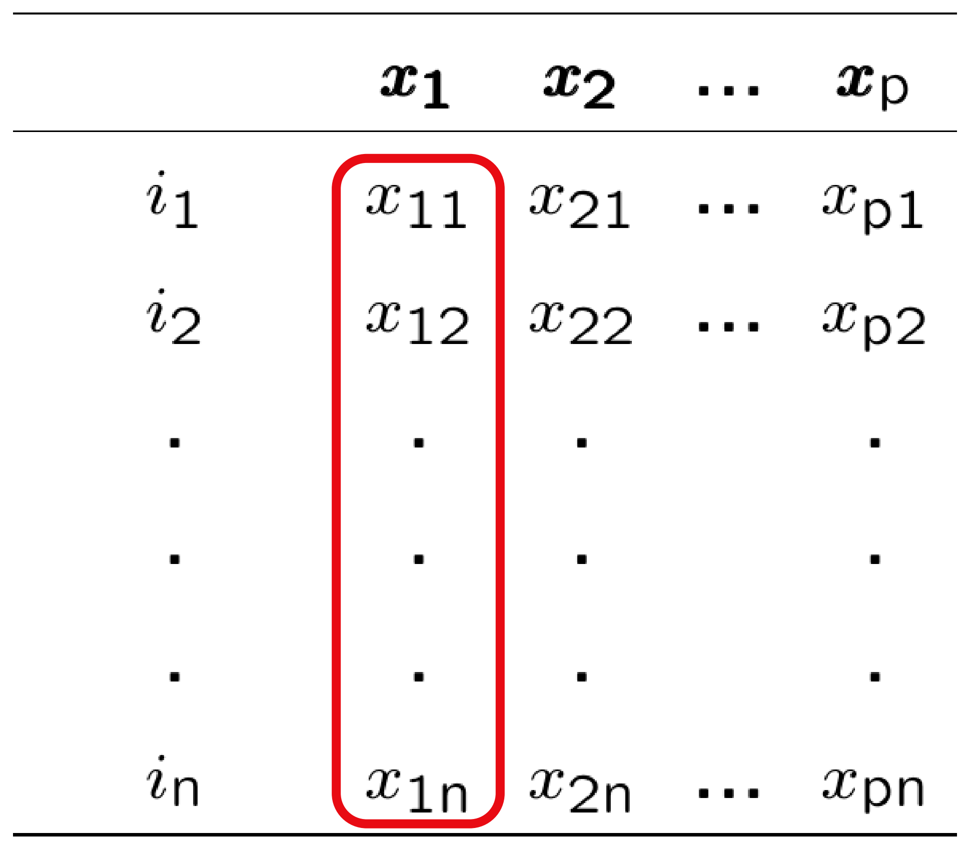
Data Analysis:
Data Description

Technological University Dublin Tallaght Campus

Department of Computing

The statistical data characterisation concepts described in this presentation all refer to a single data table column i.e. to a single variable/attribute.



	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\cdots$	$\boldsymbol{x}_p$
i_1	x_{11}	x_{21}	$\cdots$	x_{p1}
i_2	x_{12}	x_{22}	$\cdots$	x_{p2}
.	.	.		.
.	.	.		.
.	.	.		.
i_n	x_{1n}	x_{2n}	$\cdots$	x_{pn}

Tally charts and frequency distributions

For datasets collected and/or processed manually

Score	Tallies
62	𝍷
63	𝍷
64	
65	𝍷𝍷𝍷
66	𝍷
67	𝍷𝍷𝍷𝍷
68	𝍸
69	𝍸 𝍷
70	𝍷𝍷𝍷𝍷
71	𝍷𝍷
72	𝍷𝍷
73	
74	𝍷

[LEFT] Tally chart of the scores made in their final round by the 30 leading golfers in the 1992 Scottish Open

0	4, 7, 1, 0, 1, 6, 0, 1, 0
1	2, 7, 0, 3, 0, 1, 4
2	2, 6
3	9
4	
5	8

[ABOVE] Results from a low-scoring cricket match, shown in a stem and leaf diagram. The stems are chosen so that there are up to about 10 of them, for clarity.

Measures of central tendency

One of the most important ways of summarising a variable is finding some kind of centre around which its values are grouped. There are three measures of central tendency:

- **mode** - the most commonly occurring value among those observed (defined for variables with *discrete* values)
- **median** - the value that is surpassed by exactly half of the observed values (defined for variables with values that *can be ordered*)
- **mean** - the average value (defined for *numeric* variables)

HOWTO

Determining the **mode** for a set of values

Find the most commonly occurring value.

Example 1 (one mode):

`3, 4, 5, 6, 7, 7, 7, 8, 8, 9` ⟶ The mode is **7**.

Example 2 (more than one mode):

`3, 4, 5, 6, 7, 7, 7, 8, 8, 8` ⟶ The mode is **7,8** or the mean i.e. **7.5**.

HOWTO

Determining the **median** for a set of values

The **median** is the middle value in the set, when the set is ordered.

Example 1 (odd number of values):

3, 4, 5, 6, 7, **7**, 8, 8, 8, 9, 9 ⟶ The median is **7**.

Example 2 (even number of values):

3, 4, 5, 6, 7, **7**, **8**, 8, 8, 9, 9, 10 ⟶ The median is $\frac{7+8}{2} = \mathbf{7.5}$.

HOWTO

Determining the **mean** for a set of values

The **mean** is the average of the values. For a variable called x with n values it is calculated as:

$$\bar{x} = \frac{\sum_{\mathrm{i}=1}^{n} x_{\mathrm{i}}}{\mathrm{n}}$$

Example:

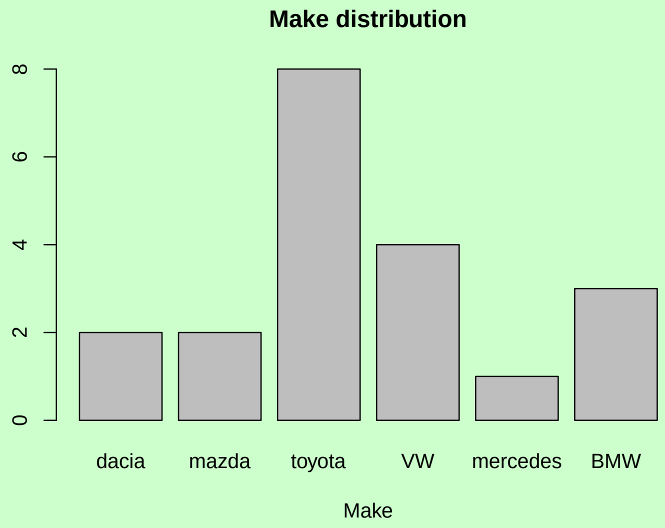
`3, 4, 5, 6, 7, 7, 8, 8, 8, 9, 9` $\longrightarrow$ The mean is $\frac{3+4+5+6+7+7+8+8+8+9+9}{11} =$ **6.73**.

Distribution visualisation

The distribution of a set of values can be visualised in several ways, each suitable for specific types of data.

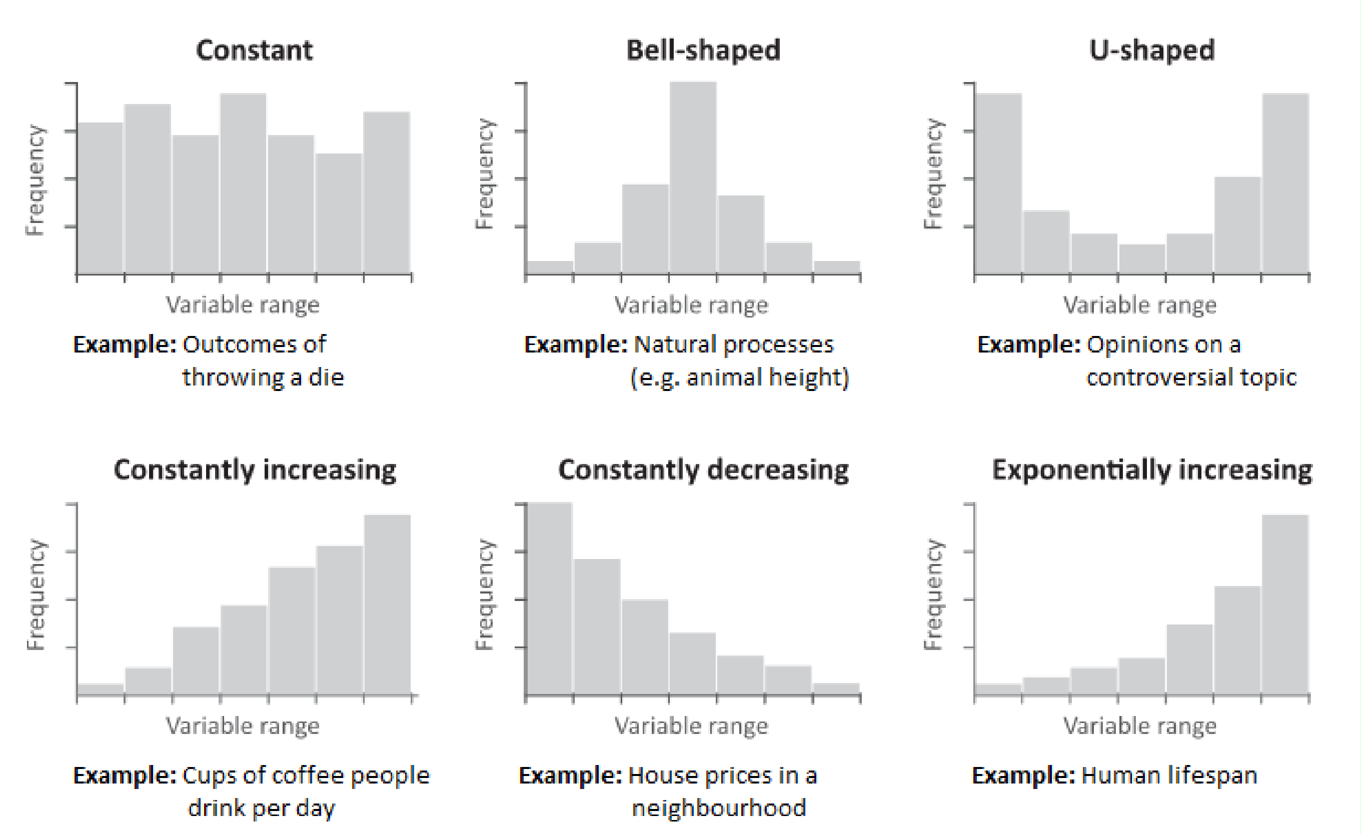
Bar chart

A bar chart shows how many time each value occurs in a set. It is used with categorical variables and discrete numeric variables of limited range.



Frequency histogram

A frequency histogram is useful for ordered variables with many values. It groups the values into ranges and gives an idea of the relative frequencies of the ranges.



Original source: [MSD]

HOWTO

Drawing a **histogram** from a frequency table manually

The x-axis should show the ranges, while the area of the box above a range should be proportional to the frequency i.e. the number of values in that range. This proportionality is achieved when the height of the boxes corresponds to a *frequency density* i.e. the number of values per some 'unit of range', which can be chosen arbitrarily.

Example:

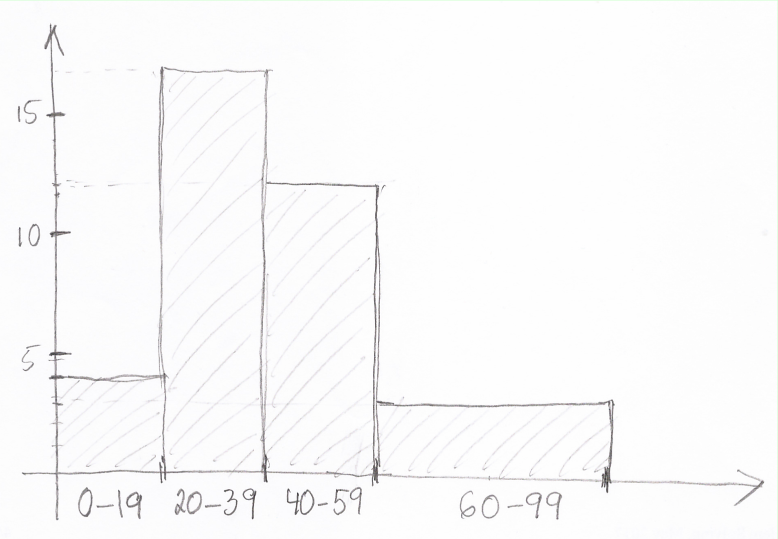
The data to be displayed in a histogram is given in the following table:

Range of values	0-19	20-39	40-59	60-99
Frequency	4	17	12	6

The frequency density value shown in a histogram for a range can be expressed as:

$$FD = \frac{\dfrac{number\ of\ values\ in\ range}{width\ of\ range}}{width\ of\ 'unit\ of\ range'}$$

For a range that has the same width as the 'unit of range' the frequency density is equal to the frequency. If we choose 20 for the 'unit of range', 3 out of the 4 ranges in the table will have 'unit of range' as their width and in those cases frequency density can be read directly from the table. For the fourth range, which has a width of 40, the frequency density value to be displayed in the histogram is $FD_4 = \frac{\frac{6}{40}}{20} = 3$. A sketch of the histogram derived in this way is shown below.



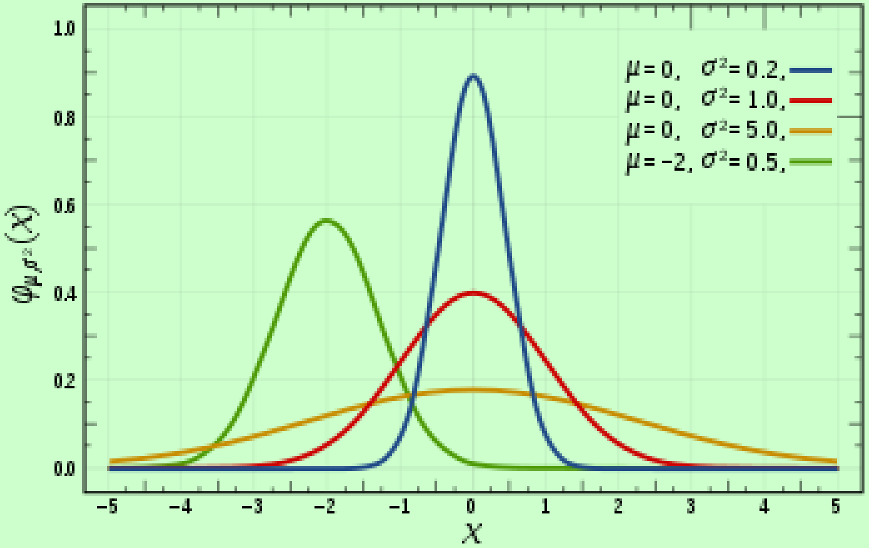
Probability density function

A probability density function (PDF) shows the idealised (based on an infinite population) distribution of a continuous variable. It could be viewed as a histogram with infinitely small bins and values normalised to a population of 1.

A very common probability density function is that for the *normal distribution*, which can be expressed with the formula:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}$$

where μ the mean and σ the standard deviation.



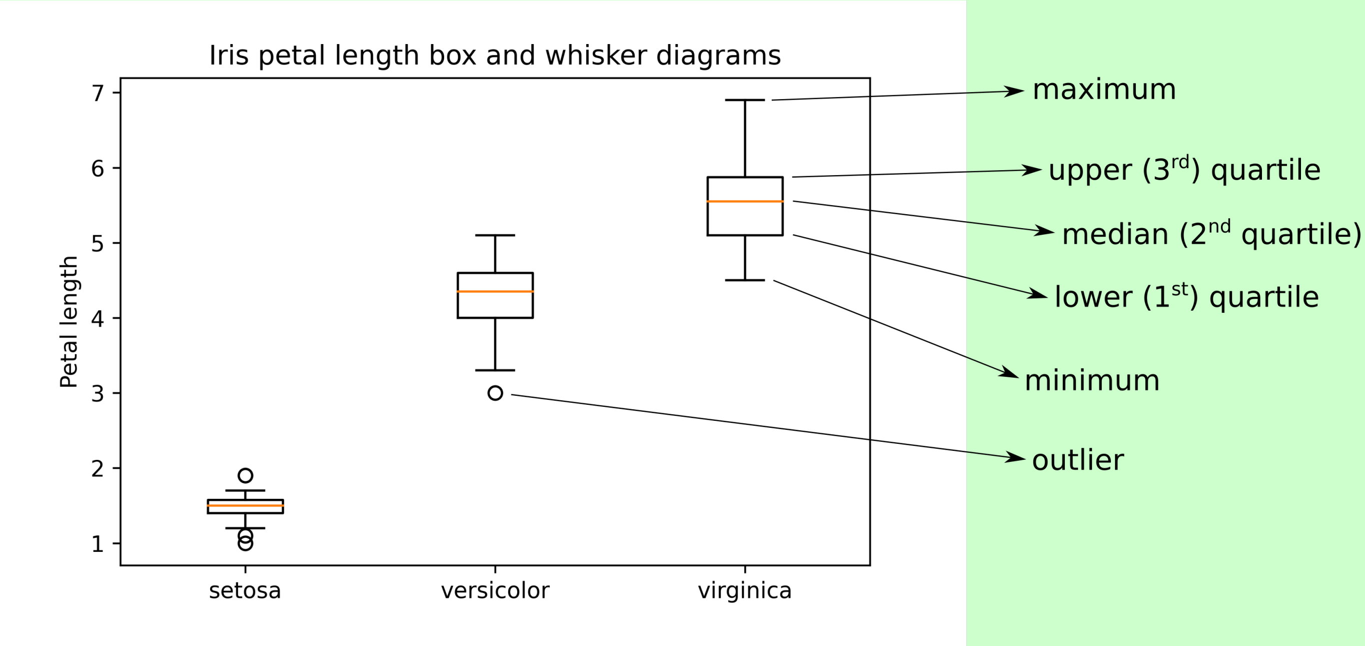
Source: Wikipedia

Box and whisker diagram

A box and whisker diagram gives a 5-point summary of a set of values. Several diagrams placed side-by-side are often used for comparison of different value sets.

The points:

- minimum
- lower quartile (a value that is greater than 25% of the values in the set and smaller than 75% of values in the set)
- median
- upper quartile (a value that is greater than 75% of the values in the set and smaller than 25% of values in the set)
- maximum



maximum

upper (3rd) quartile

median (2nd quartile)

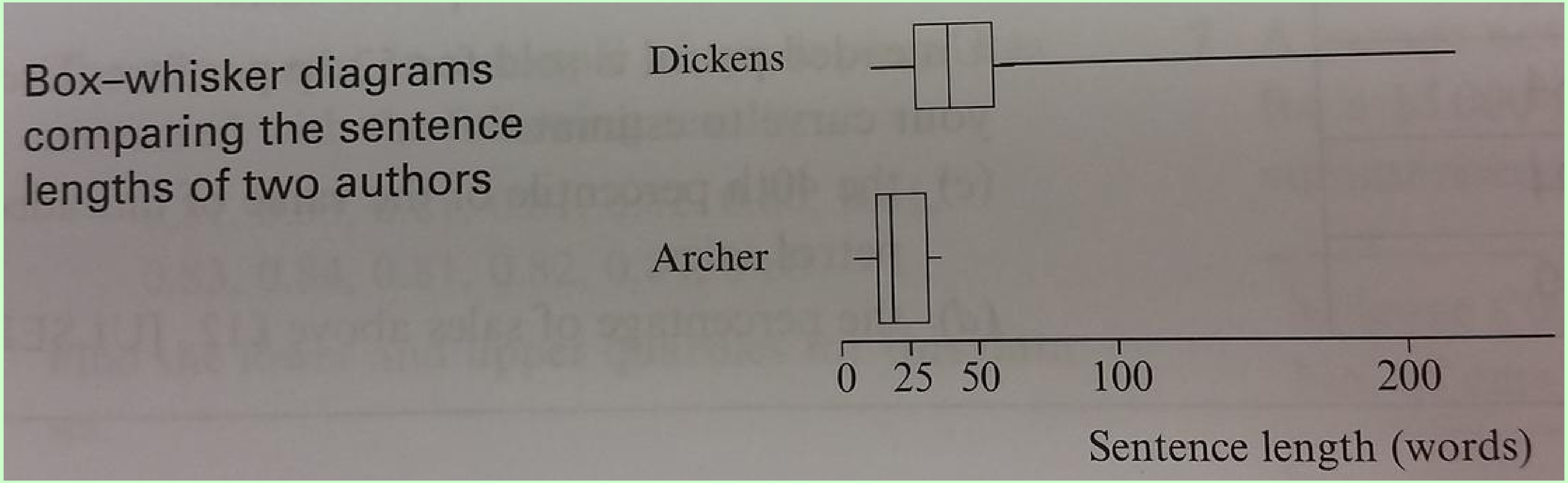
lower (1st) quartile

minimum

outlier

Box and whisker diagram example:

The box and whisker diagrams in this example show that Dickens's sentences are a lot longer (some over 200 words long!) than Archer's but also that Dickens produced a greater range of sentence lengths than Archer.



Source: [US]

Measures of spread

How much numeric data is spread out can be expressed with several measures.

Range

The *range* is the difference between the highest and the lowest value in the set.

HOWTO

Calculating the **range** of a dataset manually

1. order the data values by size

   Example: 2 3 4 4 5 5 5 7 9 10 10 11 12 12 15
2. the *range* is the difference between the largest and the smallest value

   Example: $range = 15 - 2 = 13$

Inter-Quartile Range (IQR)

The *inter-quartile range* is the difference between the upper quartile and the lower quartile i.e. the middle 50% of all the values.

HOWTO

Calculating the **inter-quartile range (IQR)** manually

1. order the data values by size

   Example 1 (odd number of data values): 2 3 4 4 5 5 5 7 9 10 10 11 12 12 15

   Example 2 (even number of data values): 2 3 4 4 5 5 5 7 9 10 10 11 12 12 15 16

2. identify the lower and upper half of the dataset (in the case of a dataset with an odd number of values, the median value is discarded)

   Example 1: (2 3 4 4 5 5 5) **7** (9 10 10 11 12 12 15)

   Example 2: (2 3 4 4 5 5 5 7) (9 10 10 11 12 12 15 16)

3. the *first quartile* or the *lower quartile* or *Q1* is equal to the median of the lower half of the dataset

   Example 1: 2 3 4 **4** 5 5 5, $Q1 = 4$

   Example 2: 2 3 4 **4** **5** 5 5 7, $Q1 = \dfrac{4+5}{2} = 4.5$

4. the *third quartile* or the *upper quartile* or *Q3* is equal to the median of the lower half of the dataset

   Example 1: 9 10 10 **11** 12 12 15, $Q3 = 11$

   Example 2: 9 10 10 **11** **12** 12 15 16, $Q3 = \dfrac{11+12}{2} = 11.5$

5. the *inter-quartile range* or *IQR* is the difference between the upper and lower quartiles

   Example 1: $IQR = Q3 - Q1 = 11 - 4 = 7$

   Example 2: $IQR = Q3 - Q1 = 11.5 - 4.5 = 7$

NOTE: The median of the data set is equal to the *second quartile* or *middle quartile* of the set.

Variance

The variance is the average squared distance of the values from the mean.

The unit of variance is the data unit squared. For example, if the data is in meters (m), then the variance will be in meters squared (m^2).

When calculated with data across an entire population, the following formula is used:

$$\sigma^2 = \frac{\sum_{\mathrm{i}=1}^{n}(x_{\mathrm{i}} - \mu)^2}{\mathrm{n}}$$

where σ^2 is the variance, n is the number of data values in the population, x_i is the i^{th} value and μ is the mean value for the population.

In the case that *sample data* is used to estimate the variance of a population, the formula is:

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

where S^2 is the estimated variance, n is the number of values in the sample, x_i is the i^{th} value and $\bar{x}$ is the mean value for the sample.

Standard deviation

The standard deviation is the square root of the variance.

The standard deviation has the same unit as the data. For example, if the data is in litres (l), then the standard deviation will also be in litres (l).

The population and sample standard deviation are each calculated as the square root of the respective variance:

$$\sigma = \sqrt{\frac{\sum_{\mathrm{i}=1}^{n} (x_{\mathrm{i}} - \mu)^2}{\mathrm{n}}} \qquad S = \sqrt{\frac{\sum_{\mathrm{i}=1}^{n} (x_{\mathrm{i}} - \bar{x})^2}{\mathrm{n} - 1}}$$

Frequency distribution shape

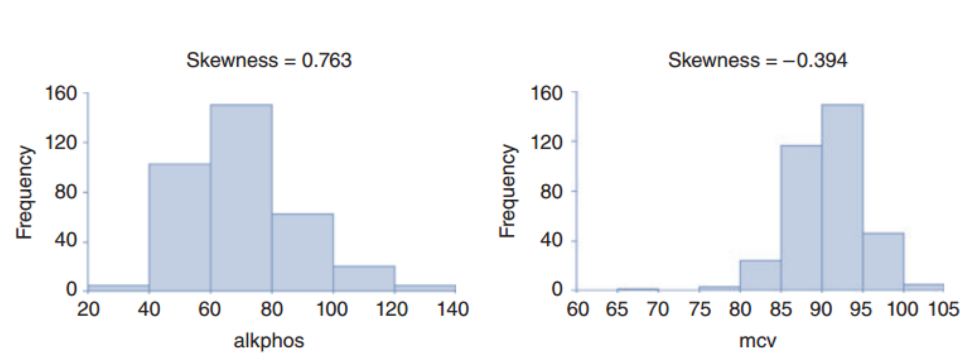
Skewness

Skewness is a measure of asymmetry of a distribution. There are various ways of calculating skewness, one of which is:

$$skewness = \left(\frac{\sqrt{n \times (n-1)}}{n-2}\right) \times \frac{{}^{1}\!/_{n} \times \sum_{i=1}^{n} (x_i - \bar{x})^3}{\left({}^{1}\!/_{n} \times \sum_{i=1}^{n} (x_i - \bar{x})^2\right)^{3/2}}$$

Source: [MSD]

Examples of distributions, one with a positive and one with a negative skewness:



Source: [MSD]

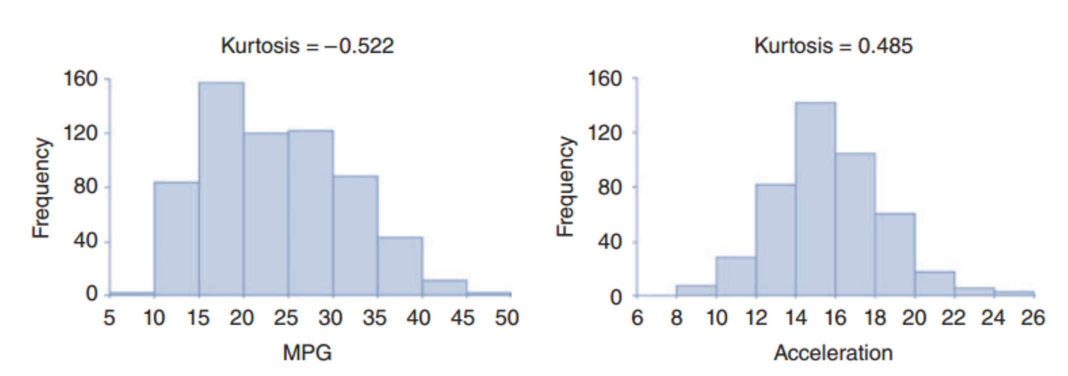
Kurtosis

Kurtosis is a measure of the 'tailedness' of a distribution i.e. of the extent and amount of outliers it includes:

$$kurtosis = \frac{n-1}{(n-2)\times(n-3)} \times \left((n+1) \times \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4 / n}{\left(\sum_{i=1}^{n} (x_i - x)^2 / n \right)^2} - 3 \right) + 6$$

Source: [MSD]

Two distributions, with different values for kurtosis:



Source: [MSD]

References The pictures in this presentation were taken from the following books. The source for each picture is cited beside it.

[DSB] *Data Science for Business: What you need to know about data mining and data-analytic thinking*, by Foster Provost and Tom Fawcett, O'Reilly Media, 2013.

[MSD] *Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*, by Glenn J. Myatt and Wayne P. Johnson, John Wiley & Sons, 2014.

[US] Understanding Statistics, by Graham Upton and Ian Cook, Oxford University Press, 1996.