COMP9033
DATA ANALYTICS

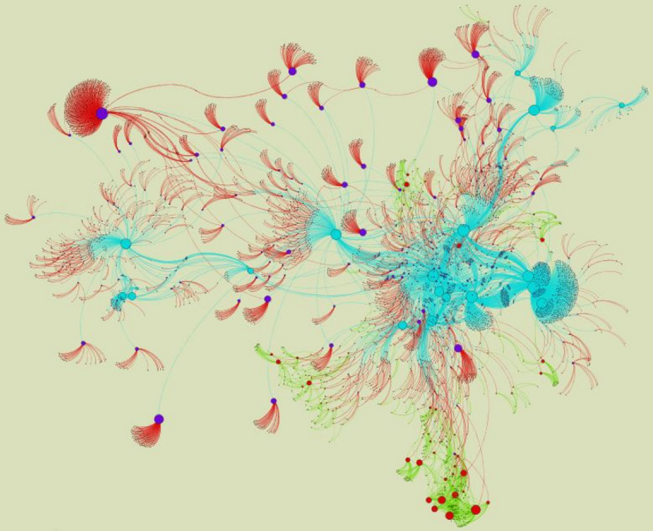
2/12

EXPLORATORY DATA ANALYSIS

DR. DONAGH HORGAN

DEPARTMENT OF COMPUTER SCIENCE
CORK INSTITUTE OF TECHNOLOGY

2017.02.08



Credit: Elijah Meeks / Stanford University

Overview

1. Introduction to data analysis:
   - What is it?
   - How does it work?
   - Real world examples.
2. Module outline:
   - Overview of topics.
   - Marking scheme.
   - Lab work.
   - Project work.
   - Contact information.
3. Data analysis processes:
   - What are they?
   - Why use them?
   - How do they work?
   - Which one to use?
4. Data sampling:
   - What is it?
   - Why is it important?
   - How to do it?

1. Data types:
    - Exploratory data analysis.
    - Types of data.
2. Visual analysis:
    - Scatter/line plots.
    - Histograms.
    - Bar charts.
    - Pie charts.
3. Summary statistics:
    - Central tendency.
    - Dispersion.
4. Detecting anomalies:
    - Outliers.
    - Robust statistics.
    - Graphical techniques.
    - Quantitative techniques.

Data types

- The term *exploratory data analysis* (EDA) refers to a commonly used approach for analysing data sets.
- Broadly, the aims of EDA are to:
  - Become familiar with the data to be analysed.
  - Uncover hidden structure in the data.
  - Determine whether the data contains important variables.
  - Detect anomalies in the data.
  - Test assumptions about the data.
- Generally, EDA is carried out *after* sampling the data but *before* cleaning and/or transforming it (recall SEMMA/CRISP-DM from Lecture 01).

- EDA techniques consist of both *quantitative* and *graphical* methods, *e.g.*
  - Categorising the type of the data, *e.g.* time series, geographic coordinates.
  - Plotting the data, *e.g.* time series plot, histogram.
  - Summarising the behaviour of variables, *e.g.* typical values, ranges.
  - Detecting outliers and anomalies, *e.g.* snow in summertime.
  - Determining whether the data follows a particular distribution, *e.g.* the normal distribution.
  - Discovering or verifying dependencies between variables, *e.g.* more sunshine $\rightarrow$ more icecream sales.
  - Finding groups or categories within the data, *e.g.* distinct species in a sample of animals.

- The *type* of data describes its content, *e.g.* whether it is numeric, categoric, a time series, GPS coordinates, *etc.*
- Defining the type of data you are working with is important as it affects the kind of techniques you can use throughout the remainder of the analysis process, *e.g.*
  - We can't compute the average of a set of categories (*e.g.* {Dog, Cat, Dog}).
  - Time series data may need to be treated sequentially in order to preserve certain chronological properties.
  - Spatial data may need to be transformed into a common coordinate reference system (*e.g.* WGS84).
- Data can have more than one type — in such cases, you should determine the type that is most relevant to the analysis you are carrying out.

Quantitative Numeric data with no inherent order or dependencies, *e.g.*

- Currency.
- Temperature.
- Population.

Categoric Data consisting of unordered groups of items, *e.g.*

- Breeds of dog.
- Car manufacturers.
- Countries with the Euro currency.

Ordinal Data with an intrinsic order (*i.e.* the sequence matters), *e.g.*

- Relative popularity of political parties (also numeric).
- Finishing positions in a race (also categoric).
- Countries with the Euro currency in GDP order (also categoric).

Spatial Data measured with respect to location, *e.g.*

- GPS coordinates (also numeric).
- Post codes (also categoric).
- Addresses (text, can be converted to coordinates or post codes).

Temporal Data measured with respect to time, *e.g.*

- Currency fluctuation over time (also numeric).
- World Cup winners (also categoric).
- GPS trace of running route (also spatial).

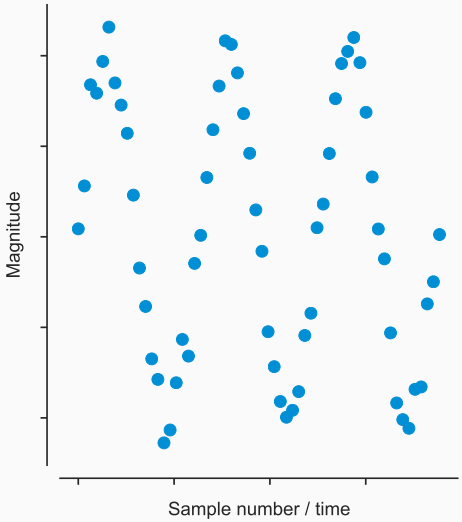
Relational Data with an inherent structure, *e.g.*

- Social network contacts.
- Organisation chart hierarchy.
- Commonly purchased groups of items.

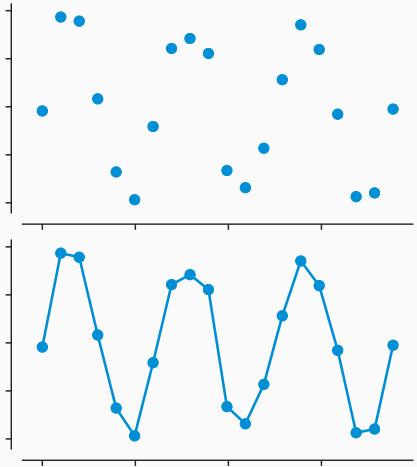
Visual analysis

- Visual analysis can give us an intuitive understanding of data quickly:
  - Quantitative techniques (*e.g.* statistics) give us precise numerical answers.
  - However, digesting large amounts of quantitative data can be overwhelming.
  - Graphical techniques allow to us to get a high level “feel” for what’s going on.
  - However, graphical techniques do not give the same level of precise numerical detail as quantitative techniques.
  - A combination of both is typically the best approach.

- Scatter plots can help us to understand trends in time series and other ordered quantitative data samples:
  - The *x*-axis measures the *sample order* (*e.g.* time) of the data points in the sample.
  - The *y*-axis measures the *magnitude* of the values in the sample.

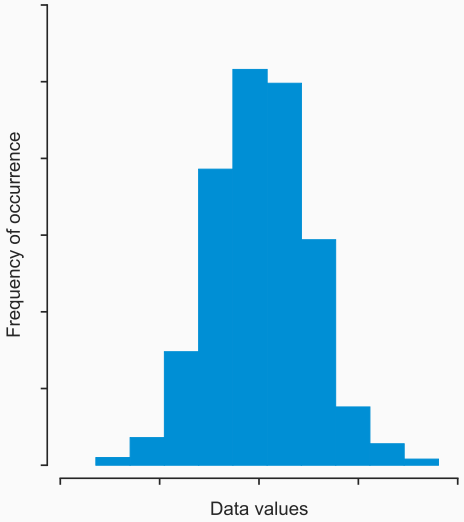


- If the number of data points in the sample is small, then it might be difficult to determine a trend visually.
- If the number of data points is not too small[1], we can use a line plot to help:
  - The axes in a line plot work the same way as in a scatter plot.
  - Consecutive points are connected by a trend line.

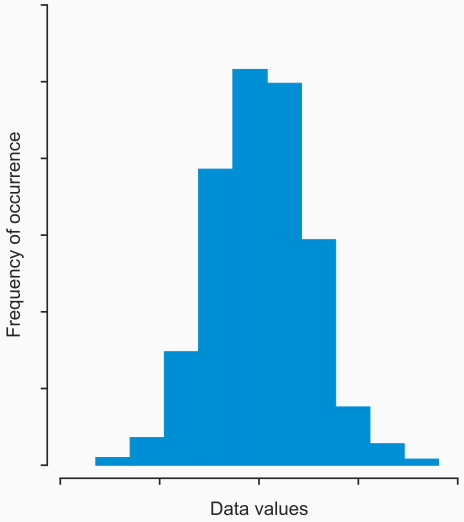


[1] For instance, smaller than the Nyquist rate. For more information, see bit.ly/2jMyzYw

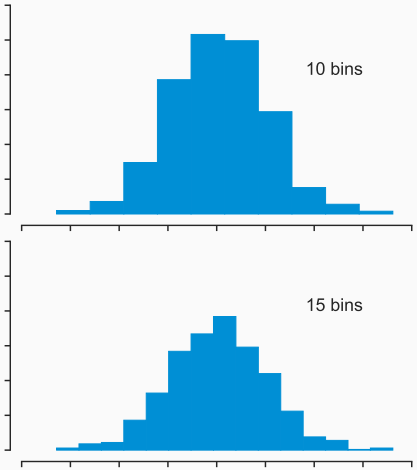
- The *histogram* is a commonly used technique for visualising the distribution of data in a sample:
  - The *x*-axis measures the *values* of the data points in the sample.
  - The *y*-axis measures the *frequency of occurrence* of the values in the sample, *i.e.* how often a given value occurs.



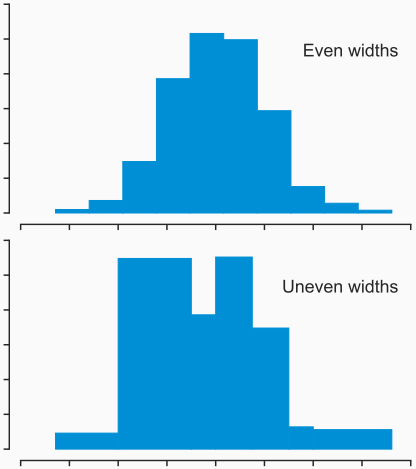
- We can create a histogram by placing the sample data points in *bins*.
- Each bin is visually represented by a vertical bar:
  - The width of the bar represents the range of the values of the sample data points contained in the bin.
  - The height of the bar represents the number of sample data points contained in the bin.



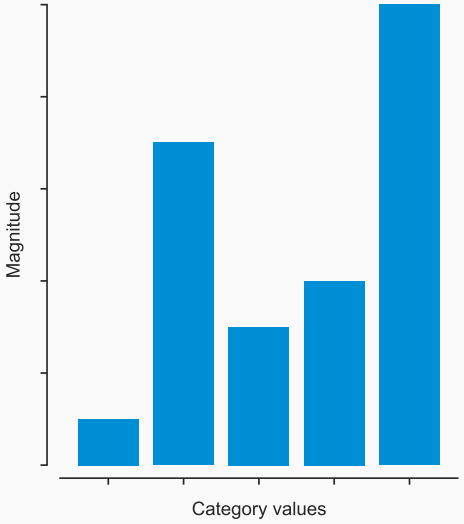
- The number of bins in a histogram is arbitrary, but the choice is important:
  - Too few bins distorts the shape of the distribution.
  - Too many bins leads to a “broken comb” look.
- The histograms on the right show the effect of varying numbers of bins on the same data sample.



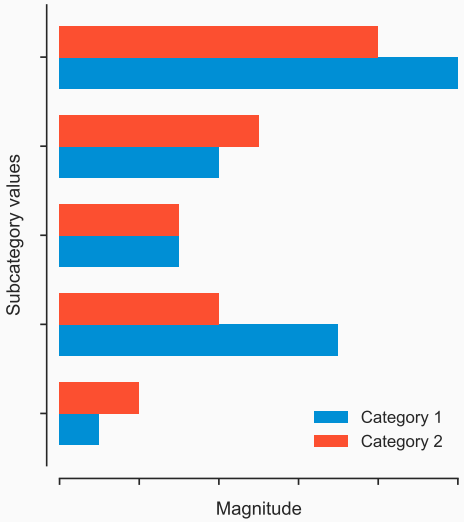
- The widths of the bins are also arbitrary, but again the choice is important:
  - Wider bins can decrease noise (spikiness) in ranges where the density of samples is low.
  - Narrower bins can increase precision in ranges where the density of data points is high.
- The histograms on the right show how uneven bin widths can distort the shape of the same distribution.



- Bar charts are a useful way to visualise the magnitude and relative proportion of quantities in a categoric sample:
  - The *x*-axis measures the *category value* of the data points in the sample.
  - The *y*-axis measures the *magnitude* of the value of the corresponding category.



- Bar charts can also be displayed horizontally — we just have to swap the *x* and *y* axes.
- Category hierarchies can be compared by using different colours (*e.g.* number of Olympic medals won by the US and UK in different events).
  - It's important to include a legend in this case, so that the meaning of the colours can be distinguished.



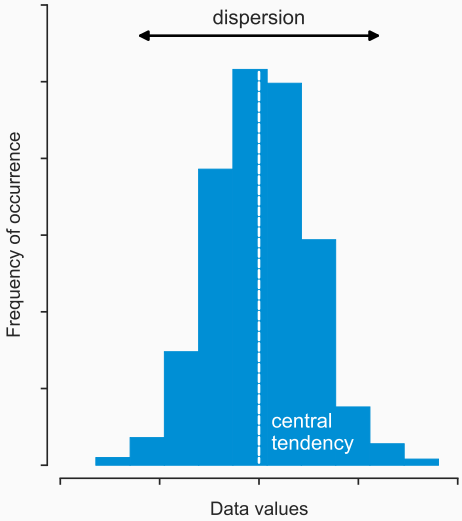
- Pie charts can also be useful when visualising proportions in a categoric sample:
  - The colours of the sections represent the *category values* of the data points in the sample.
  - The angles of the sections measure the *magnitude* of the value of the corresponding category.

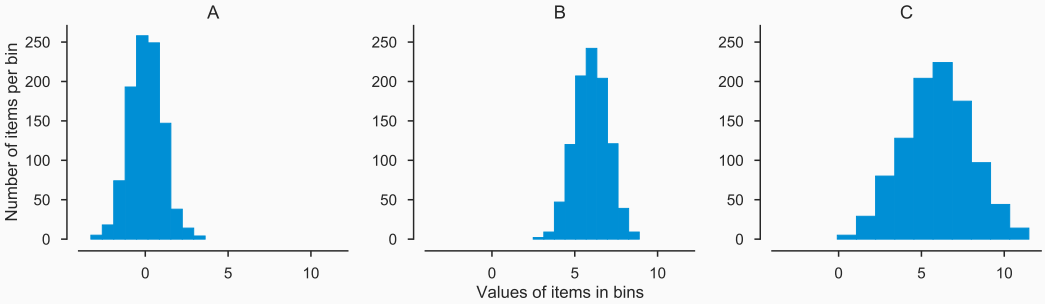
- Pie charts can be useful, but are often difficult to interpret:
  - The difference in the lengths of angles can be harder to discern than the difference in the height of bars.
  - For instance, in the image on the right, which is larger — pink or green?
- For more information, see read.bi/1MIkvcB.

Summary statistics

- One fundamental form of exploratory data analysis is the computation of *summary statistics*.
- These are statistics that *summarise* the behaviour of data in some manner, *e.g.* the temperature in Madrid in September:
  - The typical temperature is 20.5 °C.
  - The temperature generally ranges between 15 °C and 26 °C.
- Two commonly used measures are central tendency (*e.g.* the typical temperature) and dispersion (*e.g.* the range of temperatures).
- Correlation is also a useful descriptor of behaviour when you have more than one variable, *e.g.* when it is hot in Madrid, it is usually also hot in Toledo.

- Two important statistical measures of a distribution are its *central tendency* and its *dispersion*:
  - Central tendency measures the "typical" value of the data points in the sample.
  - Dispersion measures the spread or variability of the data points in the sample.





- Figure B above shows a sample with higher central tendency than Figure A, but a similar level of dispersion.
- Figure C shows a sample with the same central tendency as Figure B, but with a higher level of dispersion.

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
TEMP (°C)	6.1	7.9	10.7	12.3	16.1	21.0	24.8	24.4	20.5	14.6	9.7	7.0

Q. What is the typical temperature (*i.e.* central tendency) in Madrid throughout the year?

A. One way to compute the central tendency of the temperature is to take the average value, *i.e.*

$$\begin{aligned}\bar{T} = \frac{1}{12} \times (&6.1 + 7.9 + 10.7 + 12.3 + 16.1 + 21.0 \\ &+ 24.8 + 24.4 + 20.5 + 14.6 + 9.7 + 7.0) \\ &\approx 14.59\,^{\circ}\mathrm{C}.\end{aligned}$$

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
TEMP (°C)	6.1	7.9	10.7	12.3	16.1	21.0	24.8	24.4	20.5	14.6	9.7	7.0

Q. What is the annual temperature range (*i.e.* dispersion) in Madrid?

A. One way to compute the dispersion of the temperature is to take the range of the temperature, *i.e.* the minimum and the maximum values, [6.1 °C, 24.8 °C].

- Statistics are numerical measures that help us to characterise a data sample or the general population from which it came[2].
- They are an important part of data analysis and can provide powerful insights at an early stage.
- In the following slides, statistics are computed for a collection of n data points, X, which can be written as

$$X = \{x_1, x_2, \ldots, x_n\}, \tag{2.1}$$

where x_i is the i^{th} data point.

[2] Assuming that we have sampled correctly *and* that the data is of good quality.

- One of the most commonly used statistical measures of central tendency is the *arithmetic mean*, or *average*.
- The arithmetic mean of the sample $X = \{x_1, x_2, \ldots, x_n\}$ is denoted by $\bar{x}$ and is defined as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{2.2}$$

- Other kinds of means do exist (*e.g.* geometric, harmonic, weighted), so be careful not to confuse definitions!
- However, in general, the term *mean* can be taken to refer to the arithmetic mean.

Q. What is the mean of the sample $A = \{1, 3, -4, 5, 10\}$?

A. Using Equation 2.2, we can compute the mean as

$$\begin{aligned}\bar{a} &= \frac{1}{5} \times (1 + 3 + (-4) + 5 + 10) \\ &= \frac{15}{5} \\ &= 3.\end{aligned}$$

- The *median* of a sample is an alternative measure of its central tendency, which is less sensitive to *outliers* than the mean.
- It is defined as the middle value of the individual sample elements when arranged in *ascending* order.
- The median value of X is denoted by $\tilde{x}$ and depends on whether the total number of points, n, is odd or even:

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}}, & \text{if } n \text{ is odd,} \\ \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & \text{if } n \text{ is even.} \end{cases} \tag{2.3}$$

Q. What is the median of the sample $A = \{1, 3, -4, 5, 10\}$?

A. Before computing the median, we must first sort A in ascending order:

$$A = \{-4, 1, 3, 5, 10\}.$$

As A has an odd number of elements, we can compute the median using Equation 2.3 as

$$\tilde{a} = a_{\frac{5+1}{2}} = a_{\frac{6}{2}} = a_3 = 3.$$

Remember, sorting A in ascending order is a crucial step!

Q. What is the median of the sample $A = \{1, 3, -4, 5, 10, -1\}$?

A. Again, we must sort A in ascending order:

$$A = \{-4, -1, 1, 3, 5, 10\}.$$

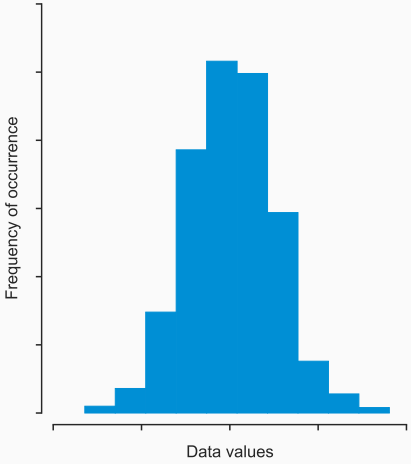
As A has an even number of elements, we can compute the median using Equation 2.3 as

$$\tilde{a} = \frac{a_{\frac{6}{2}} + a_{\frac{6}{2}+1}}{2} = \frac{a_3 + a_4}{2} = \frac{1+3}{2} = 2.$$

- The *percentile* is a statistical measure that is used to determine the value below which a given percentage of the data in an *ordered* sample falls.
- Percentiles let us make statements like *"95% of drivers in Ireland drive at or below the speed limit"*.
- Commonly used percentiles include:
  - The 0^{th} percentile, which corresponds to the minimum value of the data, *i.e.* the value below which 0% of the data falls.
  - The 50^{th} percentile, which corresponds to the median value of the data, *i.e.* the value below which 50% of the data falls.
  - The 100^{th} percentile, which corresponds to the maximum value of the data, *i.e.* the value below which 100% of the data falls.

- A related statistic is the *quartile*, which measures the three values which divide an *ordered* sample into four equally sized groups:
  - The lower quartile, Q_1, corresponds to the 25^{th} percentile.
  - The middle quartile, Q_2, corresponds to the 50^{th} percentile (*i.e.* the median).
  - The upper quartile, Q_3, corresponds to the 75^{th} percentile.

- The *mode* of a sample is a further alternative method for estimating its central tendency and is especially useful in situations where the data is not numeric.
- The mode of a sample $X = \{x_1, x_2, \ldots, x_n\}$ is defined as its most common value.
  - It is typically denoted as Mo_X.
  - It *usually* corresponds to the highest bin in a histogram[3].
- It is possible to have more than one mode, *e.g.* $X = \{0, 1, 1, 2, 3, 3\} \implies Mo_X = \{1, 3\}$.



[3]The actual modal value may differ with your choice of bin numbers or widths.

Q. What is the mode of the sample $A = \{1, 3, -4, 5, 10, 1\}$?

A. The mode is defined as the most common value of the sample, and so we can just write

$$Mo_A = 1.$$

Q. What is the mode of the sample $B = \{John, Isabelle, John, Mary\}$?

A. The mode is defined as the most common value of the sample, and so we can just write

$$Mo_B = John.$$

- One of the most common measures of dispersion in a sample is the *standard deviation*.
- The standard deviation of the sample X is denoted by σ_X and is defined as

$$\sigma_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{2.4}$$

- A related measure of dispersion is the *variance* of the sample, which is simply the square of the standard deviation, *i.e.*

$$\mathrm{Var}(X) = \sigma_X^2. \tag{2.5}$$

Q. What is the standard deviation of the sample $A = \{1, 3, -4, 5, 10\}$?

A. To compute the standard deviation, we must first compute the mean of the data. Earlier, we computed this as $\bar{a} = 3$. The standard deviation can then be computed using Equation 2.4 as

$$\begin{aligned}
\sigma_A &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{a})^2} \\
&= \sqrt{\frac{1}{4} \times ((1-3)^2 + (3-3)^2 + ((-4)-3)^2 + (5-3)^2 + (10-3)^2)} \\
&= \sqrt{\frac{1}{4} \times (4 + 0 + 49 + 4 + 49)} \\
&\approx 5.14.
\end{aligned}$$

Q. What is the variance of the sample $A = \{1, 3, -4, 5, 10\}$?

A. The variance is simply the square of the standard deviation, so we can just write

$$\mathrm{Var}(A) = \sigma_A^2 \approx 5.14 \times 5.14 \approx 26.5.$$

- The *median absolute deviation* (MAD) is an alternative measure of dispersion.
- It is less sensitive to outliers than the standard deviation (just like the median is to the mean).
- The MAD of the sample $X = \{x_1, x_2, \ldots, x_n\}$ is defined as the median of the sample $\tilde{X} = \{|x_1 - \tilde{x}|, |x_2 - \tilde{x}|, \ldots, |x_n - \tilde{x}|\}$, *i.e.*

$$MAD = median(|X - \tilde{x}|). \tag{2.6}$$

- The formulation is similar to the standard deviation, but uses the median measure instead of the mean and the absolute value instead of the square.

- The *interquartile range* (IQR) of a sample is a further alternative measure of its dispersion.
- Like the median absolute deviation, it is less sensitive to outliers than standard deviation.
- It is defined as the difference between the upper and the lower quartiles, *i.e.*

$$\text{IQR} = Q_3 - Q_1, \tag{2.7}$$

where Q_3 denotes the upper quartile and Q_1 denotes the lower quartile.

Detecting anomalies

- An *outlier* is an observation that appears to deviate markedly from the other observations in a sample.
- Typically, outliers are *anomalies*, *e.g.*
  - A snowy day in Ireland during June.
  - A person who lives to 180.
  - A slow response to a request to a server on a local network.
- Outliers can exert an extreme influence on our analysis, leading us to conclusions that do not generalise.
- Outliers can also represent interesting scientific phenomena (*e.g.* black swan events, stock market crashes), which we may want to include in our analysis.
- Whether we include them or exclude them from our analysis, it is good practice to be *aware* of them.

Q. A runner trains for a race and records track lap times with a smartphone app. On the first week of training, the times recorded by the app are as follows:

	MON	TUE	WED	THU	FRI	SAT	SUN
TIME (SECONDS)	65	68	63	61	60	62	58

After training for several weeks, the runner records lap times again:

	MON	TUE	WED	THU	FRI	SAT	SUN
TIME (SECONDS)	55	58	53	140	50	52	48

How has the runner's typical lap time changed as a result of training?

A. Let's use the arithmetic mean to calculate the typical lap time for the first week of training:

$$\bar{t}_{first} = \frac{65 + 68 + 63 + 61 + 60 + 62 + 58}{7} \approx 62.4.$$

Calculating the average lap time for the last week, we get:

$$\bar{t}_{last} = \frac{55 + 58 + 53 + 140 + 50 + 52 + 48}{7} \approx 65.1.$$

We conclude that training causes runners to become slower!

- However, looking more closely at the data for the final week of training, it's clear that the lap time recorded on the Wednesday (140 s) is very long.
- This extreme value causes the mean lap time for the final week to increase significantly, and so it appears that the runner's performance has worsened over the training period.
- However, the lap times for every other day of the last week of training are as good or significantly better than any of the lap times from the first week.
- In this case, the Wednesday lap time is an outlier caused by some error (*e.g.* forgetting to hit the stop button on the app), and has lead us to a false conclusion.

- Some statistical measures can be highly influenced by outliers, *e.g.*
  - Arithmetic mean.
  - Standard deviation / variance.
- If our data contains outliers, and we use such measures, then we risk reaching invalid conclusions!
- However, if we know that our data contains outliers, then we can use alternative measures, which are *robust* to the effects of outliers, *e.g.*
  - Median and mode for the measurement of central tendency.
  - Median absolute deviation or the interquartile range for the measurement of dispersion.
- These measures are known as *robust statistics*.

- An *estimator* is a function that calculates an estimate of a given quantity, *e.g.*
  - The arithmetic mean and median are designed to estimate the central tendency of the population that a sample was drawn from.
  - The standard deviation and interquartile range are designed to estimate the dispersion of the population that a sample was drawn from.
- The *efficiency* of an estimator is a measure of how well it estimates a given quantity, *e.g.*
  - When the sample contains outliers, the median is a more efficient measure of central tendency than the mean.
  - However, if the sample does not contain outliers, then the mean is a more efficient estimator of central tendency than the median.
- Generally, robust measures are only preferable to non-robust measures when we know the sample data we are dealing with contains outliers.

Q. Use a robust statistical measure to calculate the central tendency of the lap times from the last week of training in the earlier example.

A. The median is a robust statistical measure of central tendency. To calculate the median value, we first arrange the samples in ascending order, *i.e.*

$$T = \{48, 50, 52, 53, 55, 58, 140\}.$$

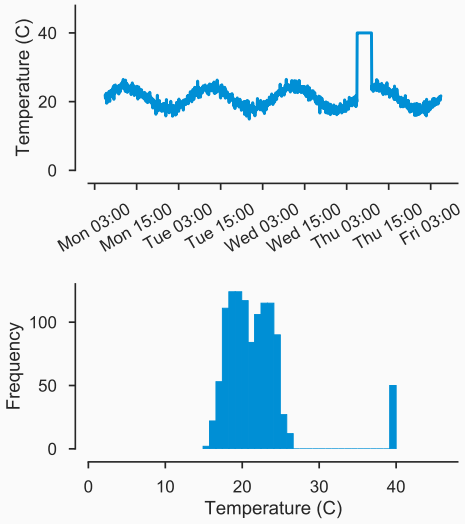
As the sample has seven values (an odd number), all we need to do is choose the middle value, *i.e.*

$$\tilde{t}_{last} = 53.$$

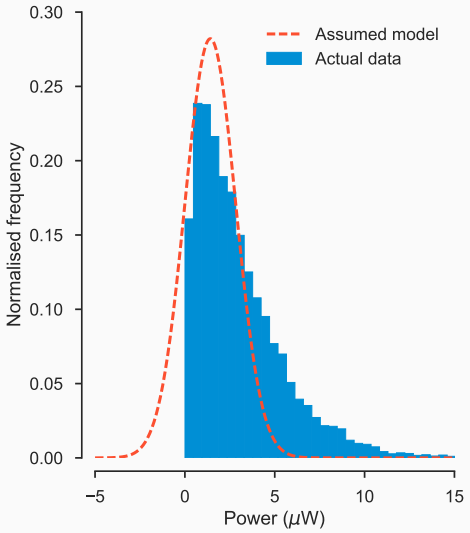
Using the robust measure has mitigated the effect of the outlier and gives us a much more accurate picture of how the data is behaving!

1. Human error:
   - Typos or mistakes during manual data entry.
2. Machine error:
   - Faulty equipment.
   - Low quality measurements.
   - Data transmission error.
3. Mismatched statistical model:
   - Assuming one kind of distribution model (*e.g.* the normal distribution) when the data follows another (*e.g.* the chi square distribution).
4. Natural/freak occurrence:
   - In many distributions, there is a small but significant chance that an extreme value may occur.
   - For instance, in the normal distribution, there is a 1% chance (approximately) that a sample has a z-score greater than 3.
   - Other examples include unseasonable weather, stock market crashes and underdog victories.

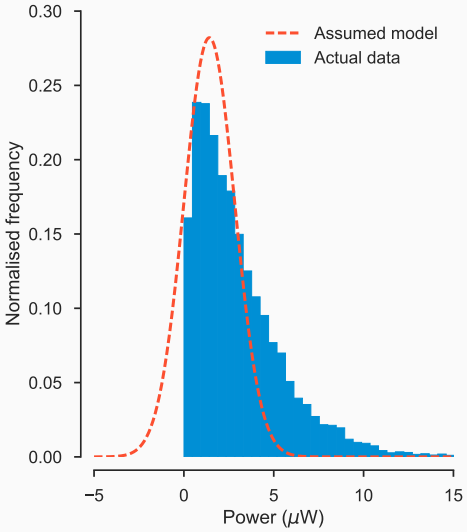
- A thermometer records the ambient temperature of a room, which oscillates in a predicable manner.
- At a certain point, the sensor undergoes a transient system fault and an incorrect temperature measurement is logged for a period of time.
- Eventually, the fault is rectified and the system returns to normal behaviour.



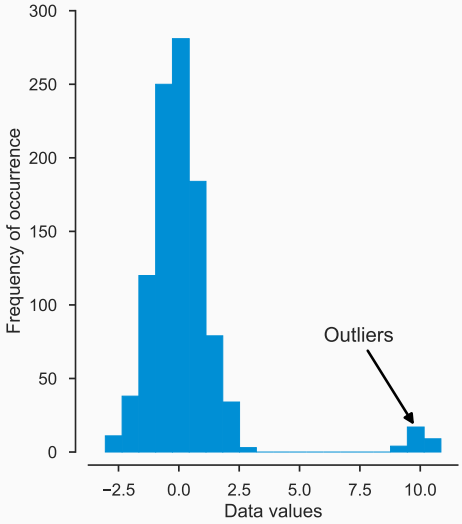
- Many data analysis techniques make distributional assumptions in order to simplify the mathematics involved:
  - For instance, data is often assumed to follow an *approximately* normal distribution.
  - However, in nature, many forms of data do *not* follow the normal distribution.
- Making an incorrect assumption can lead to normal points being misclassified as outliers as well as having other effects that affect the outcome of your analysis.



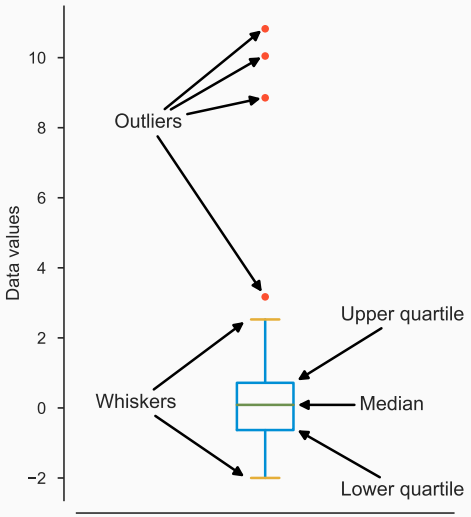
- It is a misconception to believe that the distribution of a sample *always* tends towards the normal distribution as the size of the sample grows larger:
  - While true in some cases, increasing your sample size is not *guaranteed* to make your data follow the normal distribution.
  - For more information, see the *central limit theorem* (bit.ly/2jMYbJt).



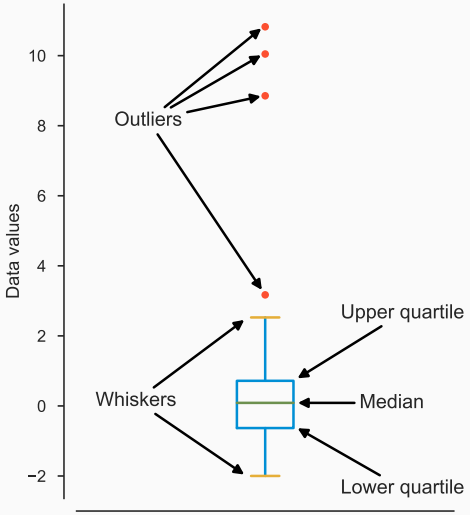
- We can visually recognise outliers as values which are markedly different from the remainder of the sample.
- For instance, in the plot on the right, the small batch of data on the far right are clearly different to majority of data on the left, and so we conclude that these are outliers.



- Box plots summarise the distribution of data using a *box* and *whiskers*:
  - The box tells us where the lower quartile (*i.e.* Q_1), middle quartile (*i.e.* the median) and upper quartile (*i.e.* Q_3) of the data lie.
  - The lower whisker is set to $Q_1 - k \times IQR$, while the upper whisker is set to $Q_3 + k \times IQR$, where k is some constant.
  - Points lying beyond the whiskers are then classified as outliers.
- Typically, $k = 1.5$, but other values are sometimes used, *e.g.* $k = 3$.



- As k is a constant, the distance from each whisker to its nearest quartile is generally equal.
- However, there can be exceptions:
  - In the plot on the right, the whisker distances are unequal because the lower whisker stops at the minimum of the data.
  - Depending on the plotting tool you use, whiskers may stop at the min/max values or continue beyond them.



- The *standard score*, or *z-score*, is a measure of the number of standard deviations away from the mean a single data point is.
  - A positive score indicates that the data point is above the mean.
  - A negative score indicates that the data point is below the mean.
  - The magnitude of the score indicates how far away from the mean the point is.
- The standard score of the data point x_i is denoted by $z(x_i)$ and defined as

$$z(x_i) = \frac{x_i - \bar{x}}{\sigma_x}. \tag{2.8}$$

- The standard score can be used to quantify how extreme a given data point is, and so can be a useful indicator that a given data point is an outlier.

Q. What is the standard score of the third data point in the sample $A = \{1, 3, -4, 5, 10\}$?

A. Earlier, we computed the mean of A as $\bar{a} = 3$ and the standard deviation as $\sigma_A \approx 5.14$. The standard score of a_3 can then be computed using Equation 2.8 as

$$z(a_3) = \frac{a_3 - \bar{a}}{\sigma_A} \approx \frac{(-4) - 3}{5.14} \approx -1.36.$$

That is, a_3 is approximately 1.36 standard deviations below the mean.

- We can use the standard score to detect outliers by specifying the maximum number of standard deviations above or below the mean we consider "normal" observations to fall in:
  - If an observation has a standard score less than or equal to the threshold, then we classify it as normal.
  - If an observation has a standard score greater than the threshold, then we classify it as anomalous, *i.e.* an outlier.
- If we denote this threshold as λ, then we can say that the point x_i is an outlier if

$$|z(x_i)| > \lambda. \tag{2.9}$$

Q. Use the standard score test to identify outliers in the following sample:

$$X = \{-7, -2, 9, -8, -10, 4, 0, -9, -5, 2\}.$$

You can assume outlying observations are those that lie more than one standard deviation from the mean.

A. Using Equation 2.2 and Equation 2.4, we can show that $\bar{x} = -2.6$ and that $\sigma_X \approx 6.29$. Then, using Equation 2.8, we can show that

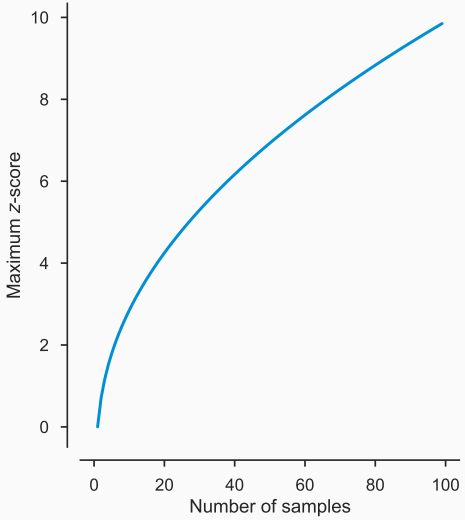
$$\begin{aligned}
&z(x_1) \approx -0.70, && z(x_2) \approx 0.10, && z(x_3) \approx 1.84, && z(x_4) \approx -0.86,\\
&z(x_5) \approx -1.18, && z(x_6) \approx 1.05, && z(x_7) \approx 0.41, && z(x_8) \approx -1.02,\\
&z(x_9) \approx -0.38, && z(x_{10}) \approx 0.73.
\end{aligned}$$

Therefore, the points $x_3 = 9$, $x_5 = -10$, $x_6 = 4$ and $x_8 = -9$ are outliers.

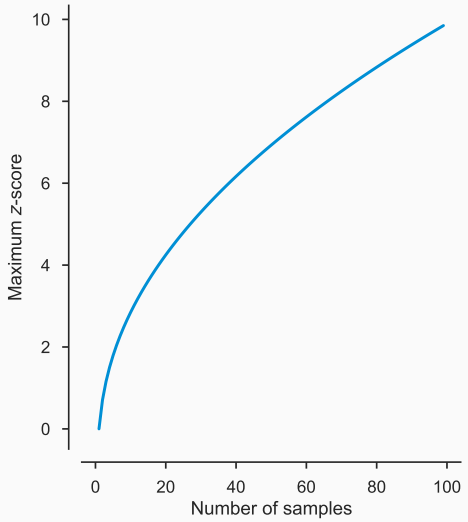
- The standard score test works well when dealing with large samples, but is not as effective when the sample size is small:
- For a sample with n values, the maximum standard score is given by

$$\max(z(x_i)) = \frac{n-1}{\sqrt{n}}, \qquad (2.10)$$

  as illustrated on the graph to the right.
- For more information, see bit.ly/2kACJqB.



- The standard score is *not* a robust statistic:
  - It is based on the mean and the standard deviation (see Equation 2.8), which are not robust.
  - Consequently, the standard score test can become unreliable when the number of outliers is large.



- The *modified standard score* is a measure of the number of median absolute deviations away from the mean a single data point is.
- The modified standard score of the data point x_i is denoted by $\tilde{z}(x_i)$ and defined as

$$\tilde{z}(x_i) = \begin{cases} \dfrac{0.6745(x_i - \tilde{x})}{MAD}, & \text{if } MAD \neq 0, \\ \dfrac{0.7979(x_i - \tilde{x})}{mean(|X - \tilde{x}|)}, & \text{if } MAD = 0. \end{cases} \tag{2.11}$$

- Because it is based on the median and MAD values, it is a more robust measure of extremity than the standard score if many outliers are present.

Summary

- Lots of maths this week! Usually, there won't be so much.
  - If you have questions, post on Blackboard!
- Lab:
  - Try it out, see how far you get.
  - If you're stuck on Python, check out the Codecademy course.
  - If you're stuck on statistics, check out the Khan Academy course.
  - If you have questions, post on Blackboard!
- Next week:
  - More exploratory data analysis.
  - How data graphics work.

1. Yau, Nathan. *Data points: Visualization that means something.* John Wiley & Sons, 2013. (bit.ly/2k8TqWR)
2. Tufte, Edward. *The Visual Display of Quantitative Information.* Graphics Press, 2001. (bit.ly/2kAU2Ic)
3. Khan Academy. *Data and statistics.* (bit.ly/1DZTQpA)
4. Shiffler, Ronald E. *Maximum Z scores and outliers.* The American Statistician 42.1 (1988): 79-80. (bit.ly/2kACJqB)