



COMP9033
DATA ANALYTICS

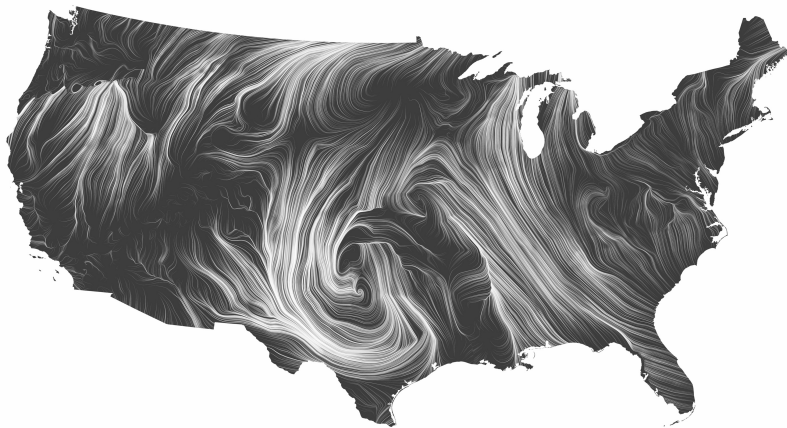
3/12

EXPLORATORY DATA
ANALYSIS II

DR. DONAGH HORGAN

DEPARTMENT OF COMPUTER SCIENCE
CORK INSTITUTE OF TECHNOLOGY

2018.02.13



Overview

1. Exploratory data analysis:

- What is exploratory data analysis?
- Why is it important?
- Data types.

2. Visual data analysis:

- Chart types.
- Visual cues — what they are, why they're important.
- The Four Pillars of Effective Visualisation.

3. Distributions:

- Histograms.
- The normal distribution.
- Other kinds of distribution.
- Visualising category distributions.

1. Summary statistics:

- Central tendency.
- Dispersion.

2. Detecting anomalies:

- Outliers.
- Robust statistics.
- Graphical techniques.
- Quantitative techniques.

3. Discovering relationships:

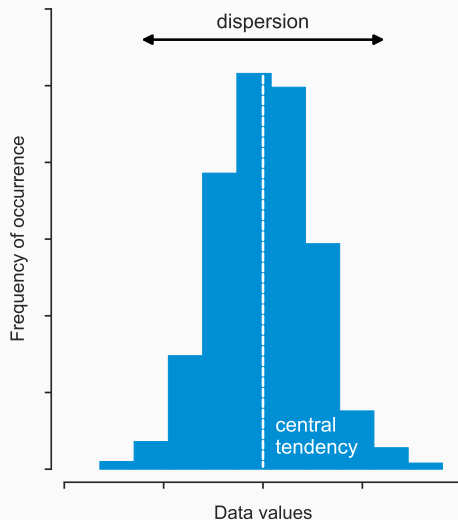
- Dependence and correlation.
- Correlation and causation.
- Quantitative measures.
- Graphical techniques.

Summary statistics

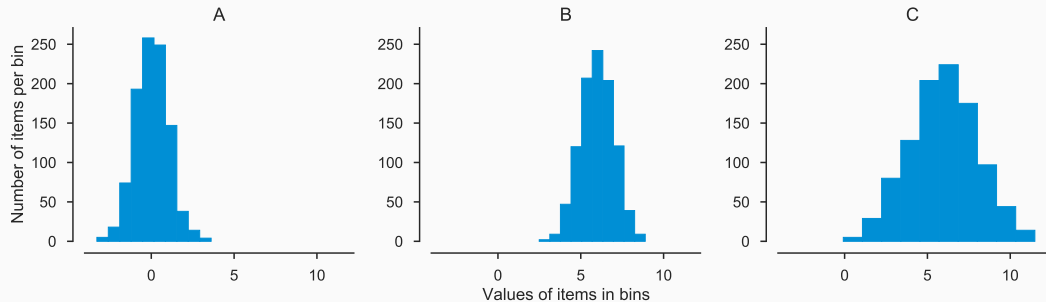
- One fundamental form of exploratory data analysis is the computation of *summary statistics*.
- These are statistics that *summarise* the behaviour of data in some manner, *e.g.* the temperature in Madrid in September:
 - The typical temperature is 20.5 °C.
 - The temperature generally ranges between 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.

1.2 / CENTRAL TENDENCY AND DISPERSION

- 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.



1.3 / CENTRAL TENDENCY AND DISPERSION



- 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.

1.4 / EXAMPLE: TEMPERATURE IN MADRID

	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.*

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

1.5 / EXAMPLE: TEMPERATURE IN MADRID

	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 came¹.
- 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

$$X = \{x_1, x_2, \dots, x_n\}, \quad (3.1)$$

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

¹Assuming that we have sampled correctly *and* that the data is of good quality.

1.7 / MEANS AND AVERAGES

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

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i. \quad (3.2)$$

- 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 3.2, 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:

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}}, & \text{if } n \text{ is odd,} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & \text{if } n \text{ is even.} \end{cases} \quad (3.3)$$

1.10 / EXAMPLE: MEDIAN OF AN ODD NUMBERED SAMPLE

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 3.3 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!

1.11 / EXAMPLE: MEDIAN OF AN EVEN NUMBERED SAMPLE

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 3.3 as

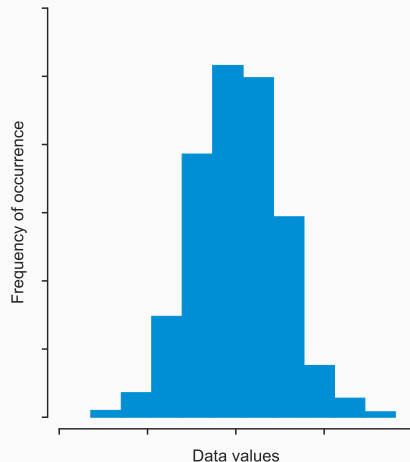
$$\tilde{a} = \frac{a_{\frac{6}{2}} + a_{\frac{6}{2}+1}}{2} = \frac{a_3 + a_4}{2} = \frac{1 + 3}{2} = 2.$$

- 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 0th percentile, which corresponds to the minimum value of the data, *i.e.* the value below which 0% of the data falls.
 - The 50th percentile, which corresponds to the median value of the data, *i.e.* the value below which 50% of the data falls.
 - The 100th 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 25th percentile.
 - The middle quartile, Q_2 , corresponds to the 50th percentile (*i.e.* the median).
 - The upper quartile, Q_3 , corresponds to the 75th percentile.

1.14 / THE MODE

- 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 $X = \{x_1, x_2, \dots, x_n\}$ is defined as its most common value and is typically denoted as Mo_x .
- It is possible to have more than one mode, e.g. $X = \{0, 1, 1, 2, 3, 3\} \implies Mo_x = \{1, 3\}$.



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

$$\sigma_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}. \quad (3.4)$$

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

$$\text{Var}(X) = \sigma_x^2. \quad (3.5)$$

1.18 / EXAMPLE: STANDARD DEVIATION

- 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 3.4 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

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

- 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 $X = \{x_1, x_2, \dots, x_n\}$ is defined as the median of the sample $\tilde{X} = \{|x_1 - \tilde{x}|, |x_2 - \tilde{x}|, \dots, |x_n - \tilde{x}|\}$, i.e.

$$MAD = \text{median}(|X - \tilde{x}|). \quad (3.6)$$

- 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.*

$$\text{IQR} = Q_3 - Q_1, \quad (3.7)$$

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

Detecting anomalies

2.1 / OUTLIERS

- 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.

2.2 / EXAMPLE: OUTLIERS AND AVERAGES

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?

2.3 / EXAMPLE: OUTLIERS AND AVERAGES

- 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!

2.4 / EXAMPLE: OUTLIERS AND AVERAGES

- 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.

2.5 / ROBUST STATISTICS

- 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*.

2.6 / WHY NOT USE ROBUST STATISTICS ALL THE TIME?

- 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.

2.7 / EXAMPLE

- 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.*

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

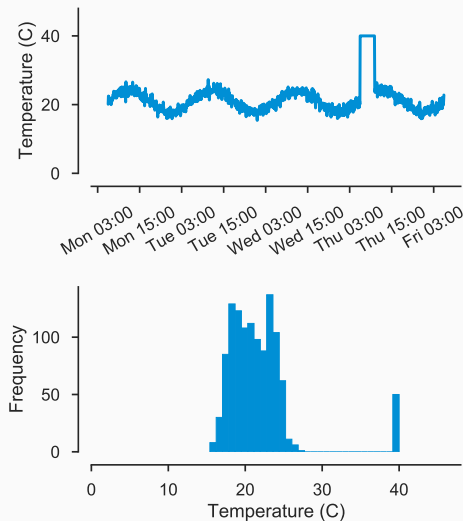
Using the robust measure has mitigated the effect of the outlier and gives us a much more accurate picture of how the data is behaving!

2.8 / CAUSES OF OUTLIERS

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.

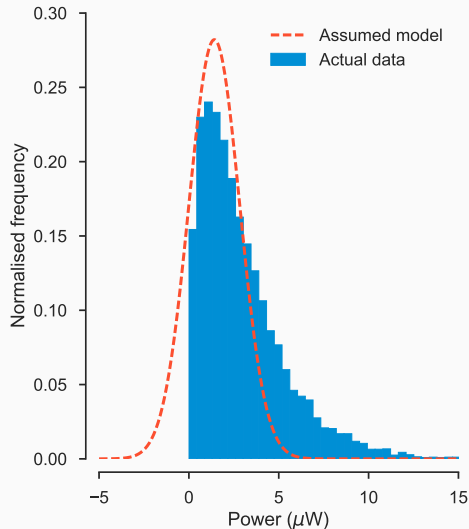
2.9 / EXAMPLE: SENSOR MEASUREMENT ERROR

- A thermometer records the ambient temperature of a room, which oscillates in a predictable 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.



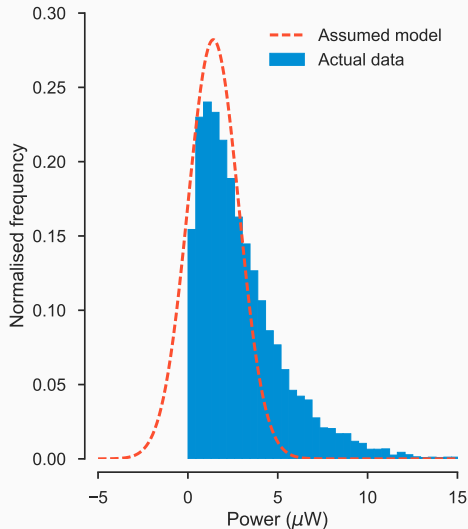
2.10 / EXAMPLE: MISMATCHED STATISTICAL MODEL

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



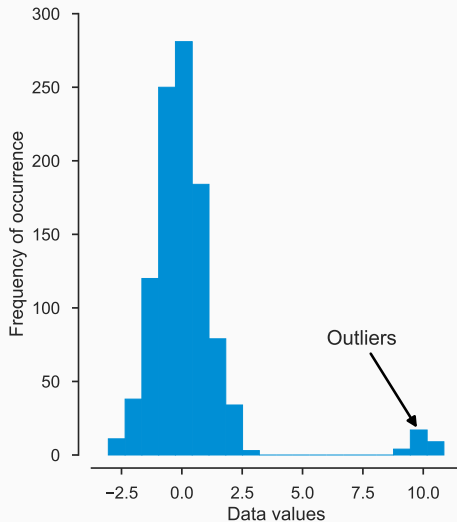
2.11 / EXAMPLE: MISMATCHED STATISTICAL MODEL

- 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).



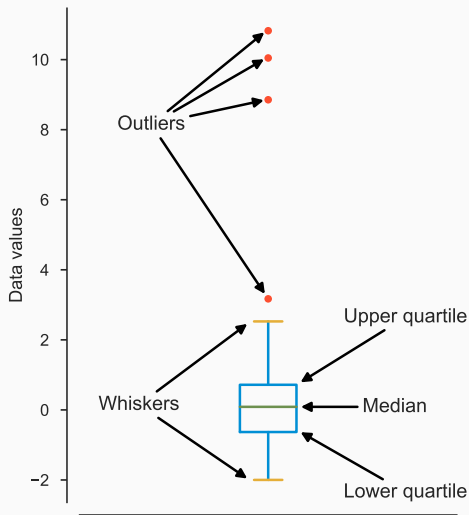
2.12 / DETECTING OUTLIERS: HISTOGRAMS

- 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.



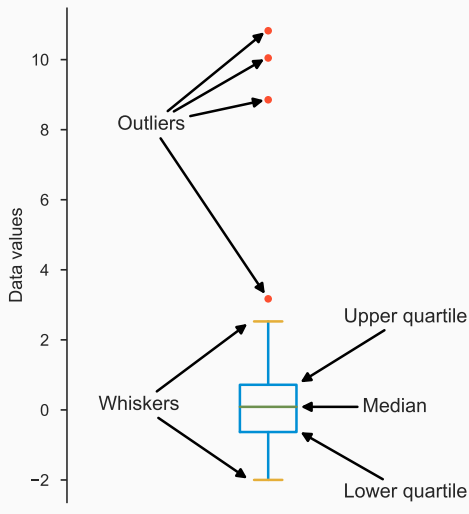
2.13 / DETECTING OUTLIERS: BOX PLOTS

- 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 is sometimes adjusted to increase/decrease sensitivity.



2.14 / DETECTING OUTLIERS: BOX PLOTS

- 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.



2.15 / DETECTING OUTLIERS: THE STANDARD SCORE

- 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

$$z(x_i) = \frac{x_i - \bar{x}}{\sigma_x}. \quad (3.8)$$

- 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 3.8 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.

2.17 / DETECTING OUTLIERS: THE STANDARD SCORE TEST

- 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

$$|z(x_i)| > \lambda. \quad (3.9)$$

2.18 / EXAMPLE

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 3.2 and Equation 3.4, we can show that $\bar{x} = -2.6$ and that $\sigma_x \approx 6.29$. Then, using Equation 3.8, 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.

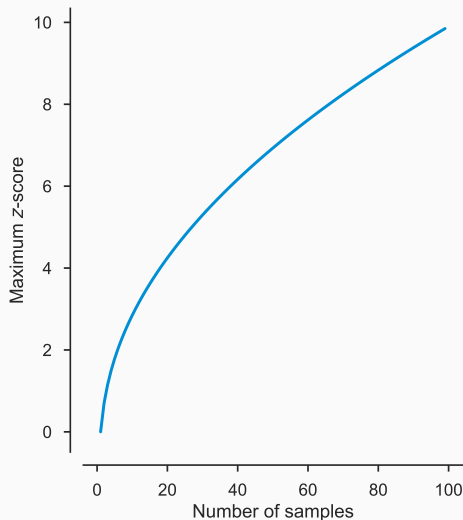
2.19 / DISADVANTAGES OF THE STANDARD SCORE TEST

- 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

$$\max(z(x_i)) = \frac{n-1}{\sqrt{n}}, \quad (3.10)$$

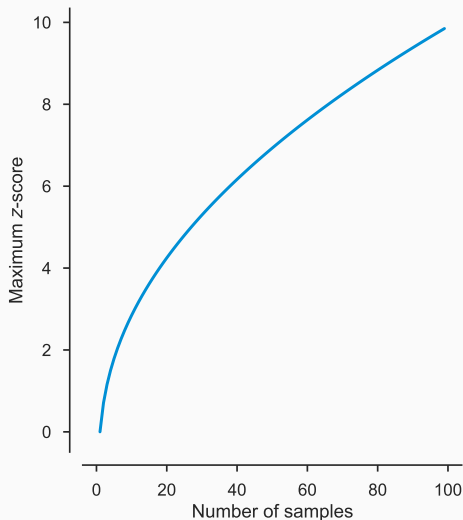
as illustrated on the graph to the right.

- For more information, see bit.ly/2kACJqB.



2.20 / DISADVANTAGES OF THE STANDARD SCORE TEST

- The standard score is *not* a robust statistic:
 - It is based on the mean and the standard deviation (see Equation 3.8), which are not robust.
 - Consequently, the standard score test can become unreliable when the number of outliers is large.



2.21 / DETECTING OUTLIERS: THE MODIFIED STANDARD SCORE

- 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

$$\tilde{z}(x_i) = \begin{cases} \frac{0.6745(x_i - \tilde{x})}{MAD}, & \text{if } MAD \neq 0, \\ \frac{0.7979(x_i - \tilde{x})}{\text{mean}(|X - \tilde{x}|)}, & \text{if } MAD = 0. \end{cases} \quad (3.11)$$

- 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.

Discovering relationships

3.1 / DEPENDENCE AND CORRELATION

- The existence of a statistical relationship between different data samples is known as *dependence*.
- Two samples are said to be *dependent* when the value of the first *depends* on the value of the second or vice-versa, *e.g.*
 - The number of people wearing shorts on a given day *depends* on the weather.
 - A team's position in a league *depends* on the number of goals they've scored.
 - A student's score on a test (ideally) *depends* on the depth of their knowledge.
- It's possible that multiple dependencies exist: in reality, a student's score on a test depends on more than just knowledge, *e.g.*
 - How well they slept the night before.
 - How much coffee they've drank.
 - The ambient temperature in the test hall.
 - ...lots of other factors!

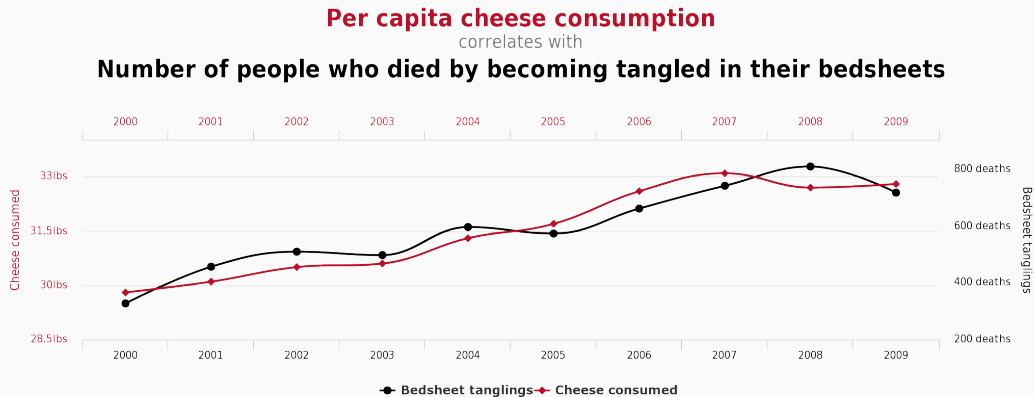
3.2 / DEPENDENCE AND CORRELATION

- *Correlation* is a measure of the dependence between two data samples.
- If two samples are *positively* correlated, then their values tend to increase or decrease together:
 - More of A leads to more of B: Sunshine → ice cream purchases.
 - Less of A leads to less of B: Eating fewer calories → losing weight.
- If two samples are *negatively* correlated, then the values in one tends to increase/decrease when the values in the other decrease/increase:
 - More of A leads to less of B: Smoking → lower life expectancy.
 - Less of A leads to more of B: Less public transport → more congestion.
- If two samples are *uncorrelated*, then there is no co-dependence:
 - More of A leads to no change in B: Higher taxes → temperature in June.
 - Less of A leads to no change in B: Population of France → days in a year.

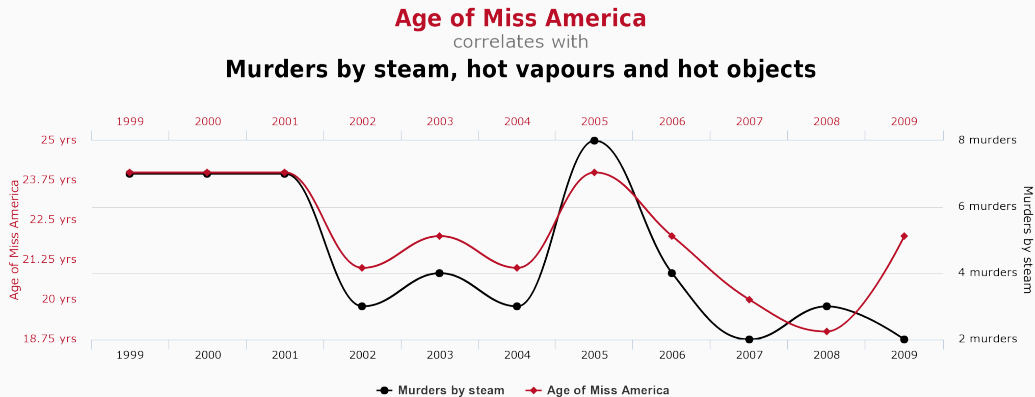
3.3 / CORRELATION IS NOT CAUSATION

- Dependence does not imply a causal relationship!
 - Buying ice creams does not make the weather sunnier.
 - Living a short life does not make you a smoker.
 - Wearing shorts does not make the weather better.
- If A and B are correlated, then there are six possible explanations:
 1. A is caused by B .
 2. B is caused by A .
 3. A causes B and B causes A .
 4. A and B are both caused by a hidden external factor, C .
 5. A causes C which, in turn, causes B .
 6. A and B are not correlated but, by random chance, they *appear* correlated.
- In most cases, it's not possible to say which of these explanations is true without deeper investigation.

3.4 / CORRELATION IS NOT CAUSATION



3.5 / CORRELATION IS NOT CAUSATION



3.6 / THE PEARSON CORRELATION COEFFICIENT

- The *Pearson correlation coefficient* is a commonly used statistical measure of correlation.
- The Pearson correlation coefficient between two data samples, X and Y , is denoted by r_{xy} and is defined as

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right) \quad (3.12)$$

$$= \frac{1}{n-1} \sum_{i=1}^n z(x_i) z(y_i). \quad (3.13)$$

- Other kinds of correlation coefficient do exist (e.g. Spearman, Kendall), so be careful not to confuse definitions!
- However, in general, the term *correlation coefficient* can be taken to mean Pearson correlation coefficient.

3.7 / THE PEARSON CORRELATION COEFFICIENT

- The Pearson correlation coefficient has a value between -1 and +1, inclusive, *i.e.* $r_{xy} \in [-1, 1]$.
- Positive correlation corresponds to a *positive* value of r_{xy} , *i.e.* $r_{xy} > 0$.
 - If $r_{xy} = 1$, then X and Y have the strongest possible level of positive correlation.
- Negative correlation corresponds to a *negative* value of r_{xy} , *i.e.* $r_{xy} < 0$.
 - If $r_{xy} = -1$, then X and Y have the strongest possible level of negative correlation.
- If $r_{xy} = 0$, then X and Y are uncorrelated.
- One handy property of the coefficient is that $r_{xy} = r_{yx}$, *i.e.* the correlation between X and Y is the same as the correlation between Y and X — the order does not matter.

3.8 / EXAMPLE

Q. What is the correlation of the samples $A = \{1, 2, 3\}$ and $B = \{4, 5, 6\}$?

A. First, let's compute the means and standard deviations of the samples using Equation 2.2 and Equation 2.4:

$$\bar{a} = 2, \sigma_A = 1.$$

$$\bar{b} = 5, \sigma_B = 1.$$

Next, we compute the standard scores of the data points in each sample using Equation 2.8:

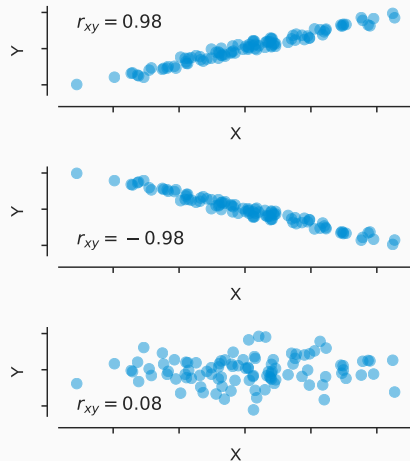
$$z(a_1) = -1, z(a_2) = 0, z(a_3) = 1. \quad z(b_1) = -1, z(b_2) = 0, z(b_3) = 1.$$

Finally, we compute the correlation coefficient using Equation 3.13:

$$r_{AB} = \frac{1}{n-1} \sum_{i=1}^n z(a_i) z(b_i) = \frac{((-1) \times (-1)) + (0 \times 0) + (1 \times 1)}{2} = 1.$$

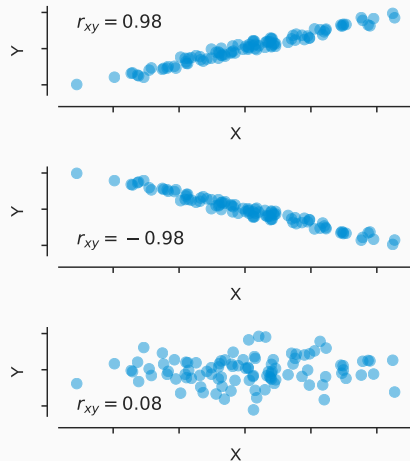
3.9 / VISUALISING CORRELATION

- The Pearson correlation coefficient is a *quantitative* technique for understanding correlation.
- However, we can also use a graphical technique — the scatter plot:
 - Positively correlated data trends linearly upwards.
 - Negatively correlated data trends linearly downwards.
 - Uncorrelated data does not have a linear relationship.

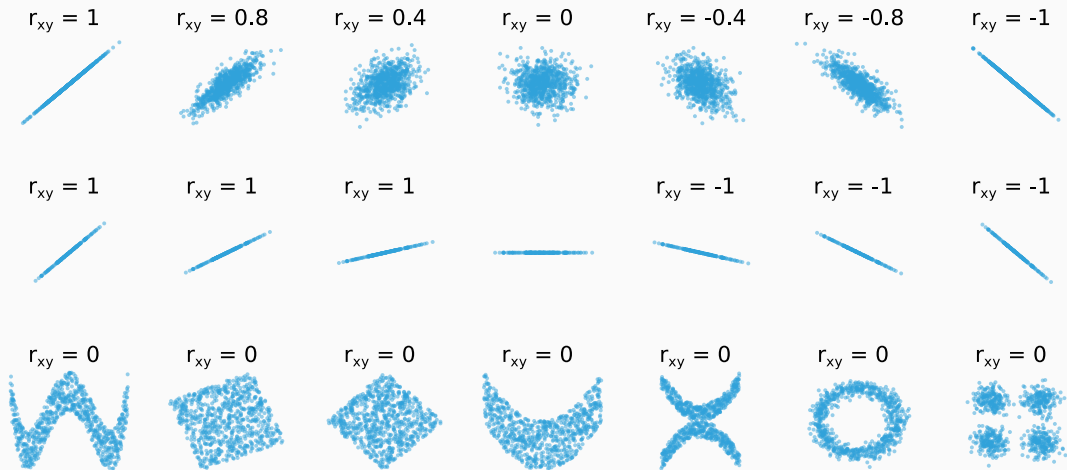


3.10 / VISUALISING CORRELATION

- Pearson correlation measures the strength of *linear* relationships between variables, *e.g.* $y = mx + c$.
- However, other relationships can exist, *e.g.* $y = \sin(x)$, $y = x^2$.
- In many such cases, Pearson correlation will *fail* to show any relationship — however, by combining graphical methods and quantitative methods, we can avoid missing relationships.

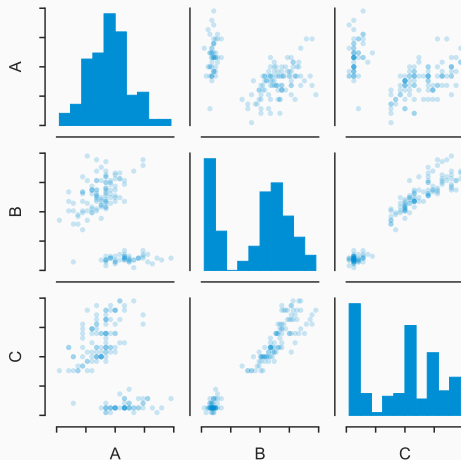


3.11 / EXAMPLE: CORRELATIONS OF VARIOUS RELATIONSHIPS



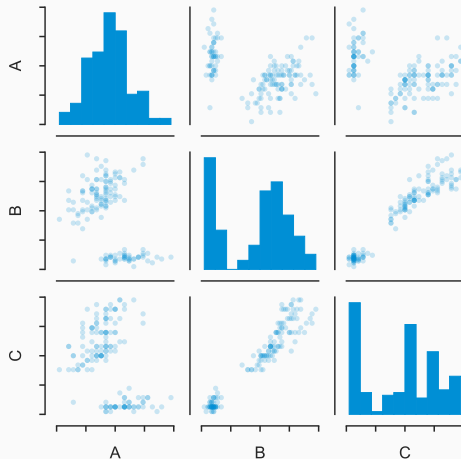
3.12 / THE SCATTER PLOT MATRIX

- The *scatter plot matrix* is a useful tool for understanding relationships between multiple variables:
 - It consists of a square grid of plots with as many rows/columns as there are variables.
 - The diagonal elements of the grid consist of histograms of single variables (they can also be blank or contain another chart type).
 - The other grid positions are scatter plots of the corresponding row and column variables.



3.13 / THE SCATTER PLOT MATRIX

- For instance, the scatter plot matrix to the right shows the relationship between the variables A, B and C:
 - If we want to see how A relates to B, we can look at the second plot in the first row or the first plot in the second row.
 - If we want to see how A relates to C, we can look at the third plot in the first row or the first plot in the third row.
- In this case, it's clear that there is a positive correlation between B and C, but no clear relationship between A and B or A and C.



Summary

- Lots of maths this week - usually, there won't be so much.
 - Statistics — summarise the behaviour of data.
 - Outliers and how to find them.
 - Dependence, correlation and causation.
 - Discovering relationships — graphical and quantitative techniques.
- This week's lab covers how to:
 - Customise histograms.
 - Create boxplots.
 - Remove data.
 - Replace data.
- Next week: cleaning and transforming data!

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