

# Principles of Data Analytics

CMSE11432 (2023/24)

Week 1


**Dr Antonia Gieschen**

The University of Edinburgh Business School



# Course introduction



# Course overview

From the course description:
*"The objective of this course is to enhance students' understanding of the importance of adopting a series of sound methodological steps in analysing data and to provide them with an artillery of data analytics techniques along with hands-on experience in using them."*

In other words, the objective of this course is to build or refresh your knowledge of **fundamental statistical principles** and how to **apply** them to data analysis problems.

The purpose of this is to give you a good foundation onto which you can build your skills further through more advanced courses throughout your studies.



# Course structure

We will spend the next five weeks covering the following topics:

- (Lecture 1) Fundamentals in Statistics and Probability
- (Lecture 2) Probability (continue)
- (Lecture 3) Hypothesis testing
- (Lecture 4) Analysis of variance
- (Lecture 5) Linear regression

The complete course structure and recommended readings for each week can be found on Learn.



# Course overview

10 lecture hours, 4 tutorial hours, 1 Q&A session
Each week will consist of a lecture to cover theoretical material and a computer lab to practice the implementation of techniques.

Lectures are every Wednesday 9-11am,
Computer labs on Wednesdays (for groups 1-3) and Thursdays (for groups 4-5).
Please check on Learn which computer lab group you are!

There is no computer lab this week (Week 1) - labs start next week!

Recommended readings are:

- Heumann, C. and Shalabh, M.S., 2016. Introduction to statistics and data analysis. Springer.
- Berenson, M., Levine, D., Szabat, K.A. and Krehbiel, T.C., 2015. Basic business statistics: Concepts and applications. Pearson.

# Course objectives and assessment

This course has five objectives which at the end students should be able to achieve:

1. Discuss the concept and methods of data analytics using the proper terminology
2. Perform data exploration through statistical and probabilistic methods and formulate data-motivated research questions
3. Analyse the data relevant to problems, critically discuss alternative data analytics approaches and methods and choose the right techniques to address research questions and to build intelligence for decision making
4. Formulate managerial guidelines from the answers to research questions and make recommendations
5. Communicate findings effectively and efficiently to a critical audience



# Course overview

We will assess the objectives through two coursework assignments:

| 70% coursework (group) | 30% coursework (individual) |
|---|---|

- Groupwork will consist of a report detailing the analysis of a dataset of your choice using the techniques learned in class
- Individual coursework will be a reflective essay to critically discuss and and reflect on what you learned during the groupwork analysis
- Further details will be shared at the end of the week via Learn

# Introducing myself


Dr Antonia Gieschen (she/her)
antonia.gieschen@ed.ac.uk
Office 3.24


- PhD from University of Edinburgh with a dissertation on spatio-temporal cluster analysis
- Postdoc at Carnegie Mellon University (Pittsburgh PA, USA) within Pittsburgh Supercomputing Center
- Now Lecturer in Predictive Analytics and teaching Principles of Data Analytics (CMSE11432) as well as Predictive Analytics and Modelling of Data (CMSE11428)



# Introducing myself

My research interests are in the area of computational social science. I'm broadly interested in modelling:

- marketing, consumer behaviour and tourism
- financial wellbeing and mental/physical population health
- local food systems/access to fresh produce
- and especially spatial inequality in all of the above.



# Recommended reading

Topics covered today:

- Introduction to this course
- (Re)introducing fundamental statistical concepts

Recommended readings:

- Heumann & Shalabh: Chapter 1
- Heumann & Shalabh: Section 3.1 – 3.3
- (Optional: Berenson et al.: Chapter 1-3)



# Introduction to business analytics



# Business analytics

The use of analytical techniques to describe, predict, answer or otherwise support business problems and decision making processes related to them. We can typically divide them into three categories depending on their goal:

- Descriptive analytics: Uses data to describe what has happened in the past or is happening in the present. We will focus on this in this lecture series.
- Predictive analytics: Uses analytical techniques to predict what will happen in the future, for example by using machine learning.
- Prescriptive analytics: Uses mathematical programming and optimisation techniques for managerial decision making.



# From business problems to research questions

**Example problem**: Alice works for a market research company. She has been hired by a soft drink company who are interested in peoples' opinion about a new soda they have developed. They want to know what people think about different aspects of their product, which is supposed to be sold nationwide in supermarkets. How should Alice proceed?



# From business problems to research questions

Business problems are often formulated in rather vague terms or using different terms than analysts are used to. The first challenge is therefore to formulate answerable research questions based on the brief.

In this case, we have to consider what the company means when they use the words "opinion" and "different aspects".

Based on this, we formulate research questions, define how to collect data to answer those questions, analyse that data, and draw conclusions based on our findings which we report back to the company.



# From business problems to research questions

# From business problems to research questions

**1) The research question**

Let's say after talking to the company again we find out that by "opinion" and "different aspects" they really mean that they want to know:

- What people think about the taste of the soda, whether they like it or not,
- what people think about the price of the soda, whether it's expensive or cheap,
- what people think about the design of the bottle.

Based on this, Alice can now start collecting data.

**2) The data collection**

There are several challenges to this problem, but let's focus on one in particular for now:

Alice can't possibly ask everyone in the country for their opinion.

If the product is supposed to be sold to the general public (the **population**), it's reasonable to ask a representative group of people from that population (a **sample**) and generalise from that.

In real life there would be a lot of other considerations of course, such as how to define the target population, how and when to collect the sample etc. But let's simplify it for now.



# Populations and samples

- A **population** is a selection of objects of interest. They can be any sort of data points, such as people, locations or objects, from which we want to gather information.
- A **sample** is a subset of that population. We can sample in different ways depending on the question we're trying to answer and data availability.



# Some examples for sampling

- A representative sample aims to select data points in such a way that they represent the population. In many cases that means points are selected randomly.
- A convenience sample is often driven by necessity when random sampling is not possible, for example when it's difficult to reach a target population. Drawing conclusions to the whole population in this case is often problematic.
- A stratified sample allows us to draw our sample in a structured way by first dividing it into groups (strata) and then randomly sampling from those. This can be helpful to represent otherwise underrepresented sub-populations.

# Some examples for sampling (illustrated)

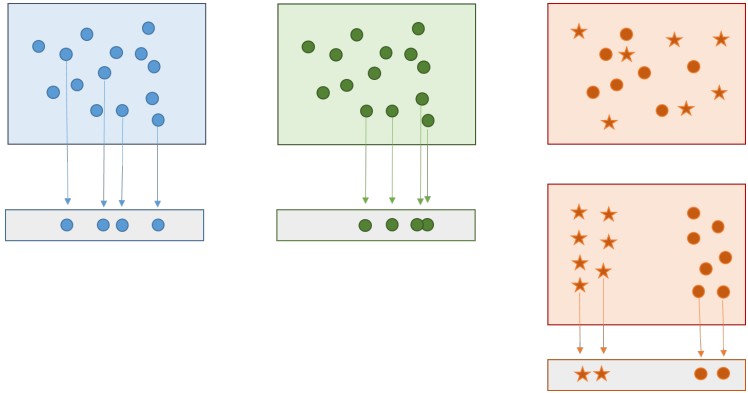


Figure: From left to right: Random sampling, convenience sampling (most convenient, here illustrated through closest points), stratified sampling



# Quantitative versus qualitative data

- Quantitative data is numeric or anything that can be converted to a numeric variable. Some examples include the price of something, a person's height or age, or the distance between two locations in kilometres.
- Qualitative data is not expressed in numeric terms. It is usually collected in the form of text, images or sounds.

Both are important and valid forms of data collection, but their analysis and the conclusions we draw from them usually differs significantly.



# Quantitative versus qualitative data

In our example, Alice could collect both quantitative or qualitative data:

- For quantitative data, she could ask for the person's opinion on the drink's taste on a scale from 1 to 7.
- For qualitative data, she could ask for the person's thoughts and feelings about the drink's taste in their own words.

What advantages do you see for Alice to collect quantitative or qualitative information in our example?



# Parameters of interest

Parameters of interest refer to a concrete and exact object or concept that we are able to measure. Deriving parameters of interest from a business problem can be very challenging.

For example, a company might ask you to find out peoples' *opinion* about a product. But the concept of opinion consists of many sub-concepts. Are they referring to the price? Availability? Product features? Brand image? Or a combination of all of those?



# Parameters of interest

It's important to be **very exact** when defining what you are measuring. A parameter is a concrete and measurable object, a number which summarises the population as a whole regarding some chosen concept.

Thinking back to Alice, this might be for example:

- The **average** opinion on the taste of the drink on a scale from 1 to 7.
- The **average** opinion on the price of the drink on a scale from 1 to 7.
- The **average** opinion on the design of the drink on a scale from 1 to 7.



# Basic descriptive statistics

**3) The data analysis**

After data collection, we start by describing the sample data through a number of descriptive statistics. This gives us a general idea of the nature of the data and its distribution.

Typically, we'll have a look at the data's

- central tendency,
- spread or variability,
- distribution.



# Measures of central tendency

Most data sets show a distinct tendency to cluster towards a central point. A descriptive value used to quantify that point is called a **measure of central tendency**.

Three such measures are:

- Mean
- Median
- Mode

# Measures of central tendency: Mean

## Arithmetic mean

### Sample mean $\bar{X}$

For a sample of $n$ measurements $X_1, X_2, ..., X_n$ the artithmetic mean is defined as

$$\bar{X} = \frac{X_1 + X_2 + ...X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

### Population mean $\mu$

Is the equivalent but calculated for the whole population $N$. The sample mean is often used as an estimator for the population mean if data on the whole population is not available or calculation is infeasible.

# Measures of central tendency: Mean

**Geometric mean**

For a sample of $n$ measurements $X_1, X_2, ..., X_n$ the geometric mean is defined as

$$\bar{X}_G = (X_1 \times X_2 \times ...X_n)^{1/n} = (\prod_{i=1}^{n} X_i)^{1/n}$$

Typically restricted to positive numbers to avoid handling negative roots. Geometric mean is the better choice if you suspect a multiplicative relationship in your data. It also tends to be less sensitive towards outliers than the arithmetic mean, can be used for data on different scales, and is often used in the analysis of stock indexes or interest rates.

# Measures of central tendency: Mean

**Harmonic mean**

For a sample of $n$ measurements $X_1, X_2, ..., X_n$ the harmonic mean is defined as

$$\bar{X}_H = \frac{n}{\frac{1}{X_1} + \frac{1}{X_2} + \cdots \frac{1}{X_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}}$$

The harmonic mean is always the smallest of the three (for a set of positive numbers) and should only be used for positive numbers. It's typically used for rates and ratios, as it's the best choice if you suspect a multiplicative and divisory relationship for data with quantities collected in different measures. A common example for this is averaging travel times over the same distance at different speeds, or in finance for price-earnings ratios.



# Measures of central tendency: Median

## Median

Imagine you have the following data and you want to know the central tendency of it:

$$1, 6, -3, 649, 10$$

The arithmetic mean of this is 132.6. That doesn't seem realistic.

# Measures of central tendency: Median

## Median

One way of approaching central tendency in data with outliers is to take the median.

The sample median is calculated as:

- If the number of measurements $n$ is odd: the median is the middle measurement in their numerical order.
- If the number of measurements $n$ is even: the median is the artihmetic mean of the two central measurements in their numerical order.

In our example, the median would be 6.

$$-3, 1, 6, 10, 649$$



# Measures of central tendency: Mode

## Mode

Another measure of central tendency with more outlier resistance is the sample mode, which is the most frequent value in a dataset.

$$4, 8, 11, 4, 7, 1, 4$$

The mode for this example is 4.

**Note:** If all values are unique, the mode does not exist.



# Measures of variability

In addition to the central tendency of your data, another important value is its variability.
Some measures for variability are:

- Range
- Variance
- Standard deviation

## Range

The range of a measurement is calculated as

$$R = X_{max} - X_{min}$$

Where $X_{max}$ is the largest value of our sample of $n$ measurements $X_1, X_2, ..., X_n$, and $X_{min}$ is the smallest.

The range gives you a general idea of how big the difference is between the highest and lowest value, but it's sensitive to outliers.

## Variance

The sample variance of a sample of $n$ measurements $X_1, X_2, ..., X_n$ is calculated as

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

where $\bar{X}$ is the arithmetic mean.
The variance corresponds roughly to the summed up squared differences of each measure from the overall mean.

As with the population mean, the population variance $\sigma^2$ can in theory be calculated as above for population $N$ but typically the sample variance is taken as an approximation.

## Standard deviation

The sample standard deviation is the square root of the variance, thus calculated as

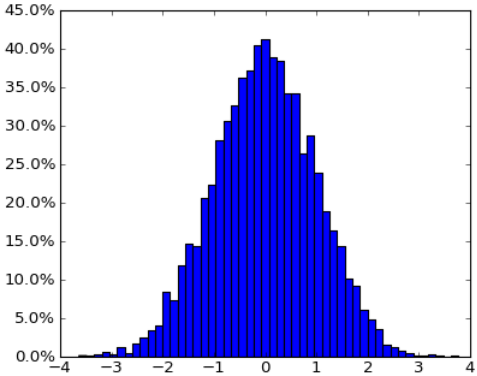
$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})}{n-1}}$$

for our sampled $n$ measurements $X_1, X_2, ..., X_n$.

The population standard deviation is denoted as $\sigma$ for a population $N$ and arithmetic mean of that population $\mu$.

# Visualising measures of variability: Histograms

**Histograms** are a common way of visualising data and are particularly helpful in identifying large differences in data variability. Here's an example from the Matplotlib documentation.[1]




[1] Accessed via the matplotlib gallery at `https://matplotlib.org/1.2.1/gallery.html`



# Visualising measures of variability: Histograms

Histograms can help us identify differences in variance between datasets if those are large enough to be spotted. We can also identify things like skewed distributions.



# Measures of location

A third category of measures to describe a dataset are what we call measures of location. These include:

- Percentiles and quartiles
- z-scores

# Measures of location: Percentiles

## Percentiles and Quartiles

Given an observed value $X$ in an *ordered* data set, $X$ is the $P$-th percentile of the data if the percentage of the data that are less than or equal to $X$ is $P$.
That means, for example, the 50th percentile is the median. 50 percent of the datapoints are less than/equal to (or, visually, left to) that percentile.

The most commonly used percentiles are the quartiles of a dataset:

- The first quartile $Q_1$ is the 25th percentile
- The second quartile $Q_2$ is the 50th percentile (the median),
- the third quartile $Q_3$ is the 75th percentile.



# Measures of location: Percentiles

## Percentiles and Quartiles

Example: What is the $Q_1$, $Q_2$ and $Q_3$ of the following numbers?

$$3, 6, 7, 10, 15, 21, 34$$

## Percentiles and Quartiles

**Example:** What is the $Q_1$, $Q_2$ and $Q_3$ of the following numbers?

$$3, 6, 7, 10, 15, 21, 34$$

Answer: First, locate the median ($Q_2$) which is the central point of an ordered list of numbers. In this case, it's 10. Then, locate the first quartile ($Q_1$) and third quartile ($Q_3$) as the midpoints between the lowest (highest) and the median point. In this case, $Q_1 = 6$ and $Q_3 = 21$.

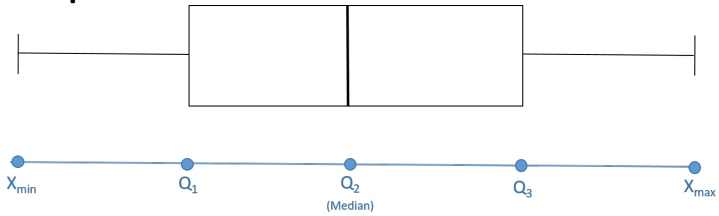
# Visualising measures of location: Box-whisker-plots

## Box-whisker-plots

Given minimum, maximum and quartiles of a dataset, we can create what we call the **five-number summary**.

$$\{X_{min}, Q_1, Q_2, Q_3, X_{max}\}$$

A popular way of visualising the five-number summary is through a **box-whisker-plot**.



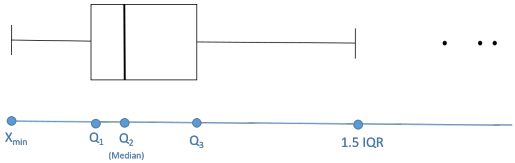
# Visualising measures of location: Box-whisker-plots

## Box-whisker-plots

Boxplots can be a great tool for checking whether your dataset is generally skewed. They can also be used to visualise outliers. Instead of drawing the whiskers to the minimum/maximum, one can draw the whiskers to the maximum length of 1.5 times the Inter-Quartile Range (IQR). Points beyond that are visualised as outliers.

$$IQR = Q_3 - Q_1$$

where $Q_3$ and $Q_1$ are the third and first quartile of your dataset respectively.





# Measures of location: z-scores

## z-scores

Z-scores (also called "standard scores") are a way of evaluating a point's relationship to the distribution of the data it's stemming from. For a population they are calculated as

$$z = \frac{X - \mu}{\sigma}$$

for a population of size $N$ with observations $X_1, X_2, ..., X_N$, their arithmetic mean $\mu$ and their standard deviation $\sigma$. Similarly, z-scores can be calculated for a sample if the population mean and standard deviation are not known. In this case, they are technically referred to as "t-statistic" instead, though literature often does not differentiate between the two.



# Measures of location: z-scores

Z-scores are very useful for understanding individual datapoints in particular, as they tell you how many standard deviations an observation is away from the distribution mean.

Another important application is in **data pre-processing**.
Converting all observations into z-scores is a process called "standardisation".
Accordingly, we may also refer to the z-score of $X$ as "standardised $X$".



# Empirical rule

The empirical rule states that given data following a standard distribution (an approximately bell- shaped relative frequency histogram), we can assume the following:

- 68% of datapoints lie within one,
- 95% of datapoints lie within two,
- 99.7% of datapoints lie within three standard deviations of the mean.

The rule is therefore also called the **"68-95-99.7 rule"**.

# Empirical rule

What is the significance of that rule?

- The rule can be used as a quick test for normality and for outliers.
- In the social science, we sometimes take the two standard deviation cut-off as an argument to set the p-value for hypothesis testing to 95%.
- Similarly, the rule can be used for risk analysis as it allows you to estimate where your data will likely be situated.

But! Remember that this only holds true if the data follows a normal distribution.

# Chebyshev's Theorem

In contrast to the empirical rule, Chebyshev's Theorem (specifically, Chebyshev's inequality[2]) can be applied to every dataset.

It states that the following holds true for most probability distributions, and we generally state that for any numerical data set with a defineable mean and variance:

- at least 3/4 (75%) of the data lie within two standard deviations of the mean,
- at least 8/9 (88.8889%) of the data lie within three standard deviations of the mean
- at least $1 - \frac{1}{k^2}$ of the data lie within $k$ standard deviations of the mean.

[2] Even more specifically "Chebyshev's theorem about the range of standard deviations around the mean" but that's a name rarely used



# Chebyshev's Theorem

Note how Chebyshev's Theorem states that **at least** so-and-so much data lies within the standard deviations. In other words, it gives us an upper bound.

**At least** $1 - \frac{1}{k^2}$ of the data lie within $k$ standard deviations of the mean

is the equivalent of saying

**No more** than $\frac{1}{k^2}$ of the data can be outside of $k$ standard deviations of the mean.



# Comparing empirical rule with Chebyshev's Theorem

Both can be very relevant depending on the application.

- Chebyshev's Theorem is more conservative; for example, for a normal distribution it would give us 75% within two standard deviations while empirical rule would give us 95%
- Empirical rule is only useful if data follows normal distribution
- If that is the case, the empirical rule often yields more accurate results
- But both can be used in a similar way to test for outliers in the data



# What comes next?

- Next week's lecture will be on probability theory
- You will also have your first computer lab next week, so please check your group number on Learn
- Assessment information will be shared at the end of the week via Learn