Introduction to Statistical Modelling

Semester 2, 2022

Dr Gentry White

Tarang Janawalkar

Contents

1 Introduction

Statistics is a field of mathematics that deals with data. It includes the study of summarising data, constructing probabilistic models, estimating parameters, and making statistical inferences.
Statistical modelling includes asking questions, obtaining data and determining a mathematical model.

1.1 Elements of Statistical Modelling

1.1.1 Data

Data is a collection of numbers that describes some characteristic that can be ranked, counted, or measured.

1.1.2 Collecting information

Statistical modelling relies upon reliably sourced data. When collecting data, we must consider

- what questions are we trying to answer,
- what information is needed to answer these questions,
- what is the best source for that information

1.1.3 Randomness

We must be aware that everything is different and that randomness introduces uncertainty in data. Random events are events whose exact outcome cannot be predicted. We can assume that all variation in the world is observed due to randomness.

1.1.4 Probability

Probability is a mathematical construct for dealing with randomness and uncertainty.

1.2 Experimental Units and Populations

Definition 1.1 (Experimental unit)**.** An **experimental unit** is an individual that generates information for the data collection process. Careful consideration of what constitutes an experimental unit must be made to ensure that it aligns with the questions of interest.

1.2.1 Sample vs. Population

Definition 1.2 (Population)**.** We might have questions about a very large collection of things called a **population**.
A dataset collected from a population is called a census.

As it is not feasible to collect data from an entire population, we must use a sample of the population.

Definition 1.3 (Sample)**.** A **sample** is a subset of a population that is representative of the population, in some cases a random sample is sufficient.

Definition 1.4 (Random sample)**.** A **random sample** is one where the sample members are selected from the population by chance.

1.3 Types of Data

1.3.1 Univariate, Bivariate, and Multivariate

Data can be described in terms of dimension, that is, how many measurements were collected from each experimental unit. By collecting multiple measurements from each experimental unit, we can ask questions about the relationship between the measurements.

- When a single measurement is collected, the resulting dataset is **univariate**.
- If two measurements are collected, the dataset is **bivariate**.
- If more than two measurements are collected, the dataset is **multivariate**.

1.3.2 Experimental vs. Observational Data

Data sets that have been collected without any specific analyses or modelling in mind are called **observational data**. By contrast, when a collection procedure is specifically designed to obtain data with a specific intent, i.e., a laboratory test, the data is called **experimental data**.
Observational data may contain biases that limit its usefulness and bias any modelling or analysis results.

1.3.3 Quantitative Data

Quantitative data is data that is expressed numerically. This data can be classified as *discrete*, *continuous*, or *ordinal*.

- Count data is classified as discrete, i.e., integer values or finite sets of real values.
- Continuous data is a measurement on a continuum or a measure that can be subdivided infinitely, i.e., time and lengths.
- Ordinal data is data where the order or ranking of values (discrete or continuous) is important.

When data is not ordinal, it is called **nominal** data.

1.3.4 Qualitative Data

Qualitative (categorical) data is data where the variable of interest is membership to a group or category.

1.4 Summarising and Describing Data

1.4.1 Tables

Tables are the most immediate way of summarising a data set. We might organise data in a table with one row for each subject and a column for each measurement.

1.5 Bar Charts

Graphical depictions of the data can also be useful but are limited in the number of variables displayed in one picture.
Bar charts are most useful for categorical data where categories are listed on the x-axis of the plot, and bars for each category are drawn with their heights corresponding to the *counts* for that category.
When the categories are **ordered** from left to right in descending order counts, the plot is called a **Pareto plot**.

1.6 Line Charts

Line charts illustrate a *trend* of change based on **two** quantitative variables. Typically line charts display trends over time (or other ordinal variables).
Often trends over time need to be aggregated by plotting the average or median per year to avoid a "busy" plot which can sometimes be difficult to read.
While the resulting chart can explain overall trends, they can obscure how much variability or "noise" is in the data and may be misleading if the overall trend is obscured by variability.

1.7 Histograms

Histograms are a special kind of bar chart that give a visual description of data by "binning" or grouping data into data ranges, then plotting bars with heights equal to the count of the bins' contents *or* the relative proportion of the bins' contents.
Histograms give us a picture of the shape of the data and help identify patterns in the distribution of values.
The binning process is performed by the computer, however in most cases we override the automatic settings and select either the number of bins, or the width of each bin.

1.8 Plots, Graphs, and Charts

- A **chart** is a visual display of data, i.e., a table, a graph, or a diagram
- A **graph** is a diagram showing the relationship between variables, each measured along orthogonal axes.
- A **plot** is used as a synonym for graph but is less precise in its definition; it also sometimes refers specifically to a graph *produced by a computer*.

1.9 Interpreting Graphical Descriptions

Graphical descriptions of data should ensure that all information about the data is expressed.

- The x and y axes should be clear in what they are measuring, including any units.
- Consider how the graph or chart was made. What choices were made and how might different options change how the graph is perceived.

- Does the graph contain any outliers that merit investigation to determine if they are accurate measurements, or if they result from either measurement or recording error.
- For Pareto charts and histograms; the y-axis should measure proportion or density rather than frequency to make comparisons easier.

1.9.1 Centrality

Histograms are a graphical representation of the distribution or density of observations. Centrality is the degree to which an observation is central to the distribution. Additionally, the data can be multi-modal if there are multiple "peaks" or "centres" in the distribution.
Altering the number of bins or bin width may reveal the centrality of the observations.

1.9.2 Skew

Another characteristic of histograms is the degree to which the distribution is skewed. Skew is the deviation from symmetry about the centre of the data. Skew is either "right" skew where the tail of the density or histogram is heavier to the right, or "left" skew if otherwise.
This can be observed by looking at how much the left/right tails are stretched in comparison to one another, i.e., the tail to the right of a right skewed chart stretches further on the x-axis than on the left.

1.9.3 Trends

Trends refer to changes in a line chart and are often described as a constant (first-derivative) pattern of increasing or decreasing values.

2 Numerical Summaries of Data

Although graphical summaries are useful for developing a general understanding data, they are limited to subjective interpretations. To form a precise understanding of the data, we need to use numerical summaries. Here we must make a distinction between sample and population summaries as measurements may vary between samples, whereas population summaries are generally constant.

2.1 Measures of Centrality

2.1.1 Mean

Given a set of n observations x_1, x_2, ..., x_n, the **arithmetic mean** or **average** is defined as

$$\frac{1}{n}\sum_{i=1}^{n} x_i \equiv \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n\right)$$

If the data is taken from a sample, the sample mean is denoted $\overline{x}$.

2.1.2 Median

A drawback to the mean is that it can be misleading when the data is skewed. The **median** is the middle value of a set of n observations when arranged from smallest to largest.
If n is odd:

$$\text{median} = x^{(\frac{n+1}{2})}$$

or the $(n+1)/2$th value of the sorted list. If n is even, the median is the :

$$\text{median} = \frac{x^{(\frac{n}{2})} + x^{(\frac{n}{2}+1)}}{2}$$

2.1.3 Mode

Given discrete data, the mode is defined as the most common value in a set of observations.

2.1.4 Population Mean

The mean of a finite population is computed in the same way as the mean of a sample, but the population mean is denoted by μ.

2.2 Measures of Dispersion

Dispersion refers to how much variation there is in a set of observations.

2.2.1 Range

Given a set of observations that are ordered such that

$$x^{(1)} \leq x^{(2)} \leq \cdots \leq x^{(n)}$$

the range is defined as

$$x^{(n)} - x^{(1)}.$$

2.2.2 Variance

The variance is the average of the squared deviations from the mean.

- Given the observations x_1, x_2, ..., x_N, from a population of size N with mean μ, the **population variance** is defined as

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2.$$

- Given the observations x_1, x_2, ..., x_n, from a sample of size n with mean $\overline{x}$, the **sample variance** is defined as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

The **population variance** is given by

2.2.3 Standard Deviation

The standard deviation is the square root of the variance. This is conceptually easier to understand as it has the same units are the data.

- The **population standard deviation** is defined as
$$\sigma = \sqrt{\sigma^2}.$$
- The **sample standard deviation** is defined as
$$s = \sqrt{s^2}.$$

Theorem 2.2.1 (Chebyshev's Theorem)**.** *Given a set of n observations, at least*
$$1 - \frac{1}{k^2}$$
of them are within k standard deviations of the mean, where $k \geq 1$*. Formally,*
$$\frac{\#\left\{x | \overline{x} - ks < x < \overline{x} + ks\right\}}{n} \geq 1 - \frac{1}{k^2}$$

Theorem 2.2.2 (Empirical Rule)**.** *If a histogram of the data is approximately unimodal and symmetric, then,*

- *68% of the data falls within **one** standard deviation of the mean*
- *95% of the data falls within **two** standard deviations of the mean*
- *99% of the data falls within **three** standard deviations of the mean*

Often the standard deviation cannot be computed directly, but can be approximated using the Empirical rule. Here we assume that
$$\text{range} \approx 4s$$
so that
$$s = \frac{\text{range}}{4}.$$

2.3 Skew

The **skew** describes the asymmetry of the distribution. For a finite population of size N, the **population skew** is defined as
$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^3$$
For a sample of size n, the **sample skew** is defined as
$$\frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^3}{\left(\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}\right)^3}$$

- When the skew is **positive**, the data is **right-skewed** and the "tail" of the distribution is **longer on the right**
- When the skew is **negative**, the data is **left-skewed** and the "tail" of the distribution is **longer on the left**

2.4 Measures of Rank

It is often useful to know the rank or *relative standing* of a value in a set of observations. This is natural for ordinal data whose ordering has implicit meaning, but it can also be useful for nominal data as a means of measuring dispersion.

2.4.1 Z-Score

The Z-score is a unitless quantity and can be used to make comparisons of relative rank between members of a population.

$$Z = \frac{x - \mu}{\sigma} \quad \text{or} \quad \frac{x - \overline{x}}{s}$$

2.4.2 Quantiles

In addition to Z-scores, quantiles can be used to make comparisons of relative ranking between populations, as well as construct intervals bounding a given proportion of the observations. For a set of n observations, x_q is the q-th quantile, if $q\%$ of the observations are less than x_q.

2.4.3 Inter-Quartile Range

The inter-quartile range (IQR) is the difference between the 75th and 25th quantiles, or the range covered by the middle 50% of data. It is a robust measure of the dispersion of the data, as it is not affected by extreme values unlike the range or variance.

2.5 Boxplots

2.5.1 Five Number Summary

The five number summary is set of measurements that indicates the

- minimum value
- 25% quartile
- median
- 75% quartile
- maximum value

A boxplot is a graphical display of the five number summary. It is a plot of the values of the data mapped to the y-axis.
Using the `ggplot2` package, the function `geom_boxplot()` draws a box encompassing the IQR with a horizontal line indicating the median. Vertical lines extend 1.5 times the IQR above and below the box. The points not within the ends of the vertical lines are also plotted to indicate outliers.

2.5.2 Outliers

Outliers are extreme observations that fall outside some interval defined either by quantiles (above 95% or below 5% quantiles) or in terms of the Empirical rule (outside two standard deviations from the mean). They should be investigated to determine if they are errors or naturally occurring extreme values.

3 Bivariate Data

Data in two dimensions is often used to describe relationships between two variables.

3.1 Bivariate Categorical Data

Bivariate categorical data is a dataset with two qualitative or categorical variables that have a relationship we want to summarise. This can be done using contingency tables or side-by-side (or stacked) bar charts.

3.1.1 Contingency Tables

Contingency tables or crosstabs are tabular representations of the frequency of occurrence of pairs of values. The categories for each variable are assigned to an axis of the table so that each cell represents the frequency of occurrence of a pair of categories, one from each variable.

3.1.2 Bar Plots

Often the data is presented more effectively as a stacked bar chart or side-by-side bar chart. Here the counts for each pair of categories are plotted on the same axis, and stacked on top of one another to display relative proportion, or side-by-side if too busy.

3.2 Bivariate Quantitative Data

Bivariate quantitative data is a dataset with one qualitative variable and one quantitative variable. This can be represented as a table, or through various charts, by comparing charts side-by-side for each category.

3.3 Scatter Plots

When both variables are quantitative, the data can be represented as a scatter plot with each variable assigned to an axis and the points plotted on the axes.

3.3.1 Covariance and Correlation Coefficients

For such data, the covariance is the measure of the linear correlation between the variables. For variables x and y,

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n-1}.$$

Note that when $x = y$, the formula simplifies to the sample variance of x. The covariance has the following characterstics:

- $s_{xy} > 0$: As x increases, y also increases.
- $s_{xy} < 0$: As x increases, y decreases.
- $s_{xy} \approx 0$: No relationship between x and y.

Although the covariance is a useful tool to measure relationships, it is only generalisable in terms of its sign. Thus, if we want to compare across data sets, we need to use the correlation coefficient.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

The **correlation coefficient** is a measure of the strength of the relationship between the variables. It is a scale-free and unitless measure bounded between -1 and 1 and has the same characteristics as the covariance.
Note that a correlation coefficient of 0 indicates **no linear relationship** between the variables, and not necessarily indicative of **no relationship**.

3.4 Regression and Least Squares

In addition to the numerical summaries above, a regression or least squares line of best fit provides both a graphical and numerical summary of the relationship between the variables. A linear relationship between two variables x and y is defined as

$$y = a + bx.$$

The least squares best fit determines the coefficients a and b that minimise the sum of the squares of the residuals (errors) between y and the line $\hat{y} = a + bx$. Mathematically,

$$\min_{a,\,b} \sum_{i=1}^{n} \left(y_i - \hat{y}\left(x_i\right)\right)^2.$$

The coefficients can be summarised by the formula

$$b = r\frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2}$$
$$a = \overline{y} - b\overline{x}.$$

4 Probability

4.1 Experiments, Events, Sample Space

Definition 4.1 (Experiment)**.** An experiment is a situation that produces some observable phenomena where the outcome is impossible to predict with certainty.

Definition 4.2 (Simple event)**.** A simple event is the outcome of a single repetition of an experiment.

Definition 4.3 (Event)**.** An event is a collection of simple events, or the outcome of multiple repetitions of an event. Events are often denoted by a capital letter.

Definition 4.4 (Mutually exclusive)**.** Events are mutually exclusive if the occurrence of one event precludes the occurrence of another. In other words, if one event occurs, the other event cannot occur.

Definition 4.5 (Sample space)**.** A sample space is the set of possible simple events, or all outcomes of an experiment.

4.2 Probability of Events

4.2.1 Probability of an Event

For a discrete finite sample space, the probability of a simple event is defined as the relative frequency of an outcome. Given the simple event A,

$$\Pr(A) = \lim_{n\to\infty} \frac{I_A}{n}$$

where I_A is a function that evaluates to 1 if A occurs and 0 otherwise.
The probabilities of events must satisfy the following conditions:

- $0 \leq \Pr(A) \leq 1$, where A is a simple event.
- The sum of the probabilities over the sample space is 1.

If an event A consists of a collection of simple events and each outcome is equally likely, then we can calculate the probability of an event as

$$\Pr(A) = \frac{\text{number of ways that } A \text{ can occur}}{\text{total number of outcomes}}$$

4.2.2 Probability of an Event in a Sample Space

Given the continuous sample space S, the event A can be defined as a subset of S, $A \subseteq S$. The definition of the probability of event A can be written as

$$\Pr(A) = \frac{\text{the area of region-}A}{\text{the area of region-}S}$$

As this probability is a ratio, it can be standardised so that the area of S is 1. Thus $\Pr(A)$ is the area of region-A.

4.2.3 Probability of the Complement

The complement of an event A is every event not in A, and is denoted as A^c or $\overline{A}$. Since the total probability for the sample space is 1, then the probability of A^c is:

$$\Pr(A^c) = 1 - \Pr(A)$$

This is true because $A \cup A^c = S$ and $\Pr(S) = 1$.

4.2.4 Probability of Subsets

If $B \subset A$, then $\Pr(B) \leq \Pr(A)$.

4.2.5 Addition Law

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B).$$

and for disjoint events:

$$\Pr(A \cup B) = \Pr(A) + \Pr(B).$$

so that the intersection $\Pr(A \cap B) = 0$.

4.3 Conditional Probability

Conditional probability is the probability of an event A given another event B occurs.
If $A \cap B = \varnothing$, then $\Pr(A \cap B) = 0$. Thus if we know that B has occurred, then we know that A cannot occur:

$$\Pr(A \mid B) = 0.$$

Therefore if $A \cap B \neq \varnothing$, then $\Pr(A \cap B) \neq 0$. If $B \neq \varnothing$, then the conditional probability of A given B is given by:

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

Note that when $\Pr(A) > \Pr(B)$, $\Pr(A \mid B) > \Pr(B \mid A)$.

4.3.1 Independence

Independence can be defined in terms of conditional probability. A and B are independent events if

$$\Pr(A \mid B) = \Pr(A).$$

This leads to the multiplication rule for independent events:

$$\Pr(A \cup B) = \Pr(A) \Pr(B).$$

4.4 Bayes' Rules

4.4.1 Law of Total Probability

By partitioning the sample space S into a collection of disjoint events $B_1,\ B_2,\ \ldots,\ B_n$, such that $\bigcup_{i=1}^{n} B_i = S$, we have

$$\Pr(A) = \sum_{i=1}^{n} \Pr(A \mid B_i) \Pr(B_i)$$

where $\Pr(A \mid B_i) \Pr(B_i) = \Pr(A \cap B_i)$. Given the probability for A given B, the probability of the reverse direction is given by

$$\Pr(B \mid A) = \frac{\Pr(A \mid B) \Pr(B)}{\Pr(A)}$$

this is known as **Bayes' Theorem**. Using the law of total probability, we can express this as

$$\Pr(B \mid A) = \frac{\Pr(A \mid B)\Pr(B)}{\sum_{i=1}^{n} \Pr(A \mid B_i)\Pr(B_i)}.$$

5 Probability Distributions

5.1 Random Variables

A random variable is a variable whose value is the result of an experiment or random trial, where the value is not known before the trial with certainty.

5.2 Discrete Random Variables

Definition 5.1 (Discrete random variables)**.** A discrete random variable takes on values in $\mathbb{N}_0$, where the random variables arises from counting processes.

5.2.1 Probability Mass Function

The probability mass function (PMF) of a discrete random variable X is a function $p(x)$ that maps values from the sample space of X onto the interval $[0, 1]$.

$$p(x) = \Pr(X = x).$$

This function is constrained by the following properties:

- $p(x) = 0$ if $x \notin X$.
- $p(x) \in [0, 1]$ if $x \in X$.
- $\sum_{\forall x \in X} p(x) = 1$.

5.2.2 Cumulative Mass Function

The cumulative mass function (CMF) is defined

$$F(x) = \Pr(X \leq x) = \sum_{-\infty}^{x} p(x)$$

and the probabilities for events can be defined using the CMF:

$$\Pr(a < X \leq b) = \sum_{x=a+1}^{b} p(x) = F(b) - F(a).$$

5.2.3 Expectation

The expectation (expected value) of a random variable X with a PMF is given by:

$$\mathrm{E}(X) = \sum_{\forall x \in X} xp(x)$$

where the expectation is often denoted μ. As the expectation is a weighted average of all possible values in X, we can extend this definition to any function of X:

$$\mathrm{E}(h(X)) = \sum_{\forall x \in X} h(x)\, p(x).$$

5.2.4 Median and Mode

The median m of a discrete random variable X is defined:

$$m \in X : \Pr(X \leq m) \geq \frac{1}{2} \wedge \Pr(X \geq m) \geq \frac{1}{2}.$$

Note that $\Pr(X \leq x) = \sum_{-\infty}^{x} p(x)$ and $\Pr(X \geq x) = \sum_{x}^{\infty} p(x)$.
The mode of a discrete random variable X is defined:

$$\max_{x \in X} p(x).$$

5.2.5 Variance

The variance of a random variable X is defined using the mean:

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}\left((X-\mu)^2\right) \\
&= \mathrm{E}(X^2) - 2\mu\,\mathrm{E}(X) + \mu^2 \\
&= \mathrm{E}(X^2) - \mathrm{E}(X)^2 \\
&= \sum_{\forall x \in X} (x-\mu)^2\, p(x).
\end{aligned}$$

5.3 Continuous Random Variables

Definition 5.2 (Continuous random variables)**.** A continuous random variable takes on values in R, where values lie on a continuum.

5.3.1 Probability Density Function

The probability density function (PDF) of a continuous random variable X is a function $f(x)$ that describes the density of possible values for a continuous random variable. Note that it does not define the probability of a specific value.
For a continuous random variable X:

- $\Pr(X = x) = 0$ and $f(x) \neq \Pr(X = x)$.
- $\forall x \in X : f(x) \geq 0$.

- $\int_{-\infty}^{\infty} f(u)\,\mathrm{d}u = 1.$

As the probability of a single value is zero, we can instead quantify the probability of a range of values:

$$\Pr(a \leq x \leq b) = \int_a^b f(x)\,\mathrm{d}x$$

where $a \leq b$.

5.3.2 Cumulative Density Function

The cumulative density function (CDF) is defined

$$F(x) = \Pr(X \leq x) = \sum_{-\infty}^{x} f(u)\,\mathrm{d}u$$

so that

$$\Pr(a \leq X \leq b) = F(b) - F(a).$$

In the continuous case, the PDF and CDF are related through the following differential equation:

$$\frac{\mathrm{d}F}{\mathrm{d}x} = f(x).$$

5.3.3 Median and Mode

The median m of a continuous random variable X is defined:

$$m : \int_{-\infty}^{m} f(u)\,\mathrm{d}u = \frac{1}{2}$$

and the mode is defined:

$$\max_{x \in X} f(x) \quad \text{or} \quad m : \frac{\mathrm{d}f(y)}{\mathrm{d}y} = 0.$$

5.3.4 Expectation

The expectation of a continuous random variable X is defined as

$$\mu = \mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$$

5.3.5 Variance

The variance of a continuous random variable X is defined:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)^2 f(x)\,\mathrm{d}x.$$

5.4 Probability Distributions

5.4.1 Bernoulli Distribution

A Bernoulli (or binary) distribution describes the probability distribution of a Boolean-valued outcome, i.e., success (1) or failure (0).
A discrete random variable X with a Bernoulli distribution is denoted

$$X \sim \text{Bernoulli}\,(p)$$

with

$$\begin{aligned}\Pr(X = x) &= \begin{cases} 1-p & x = 0 \\ p & x = 1 \end{cases} \\ &= p^x (1-p)^{1-x} \\ \Pr(X \leq x) &= \begin{cases} 0 & x < 0 \\ 1-p & 0 \leq x < 1 \\ 1 & k \geq 1 \end{cases}\end{aligned}$$

for a probability $p \in [0, 1]$ and outcomes $x \in \{0,\ 1\}$. We can also summarise the following:

$$\begin{aligned}\text{E}(X) &= p \\ \text{Var}(X) &= p(1-p)\end{aligned}$$

where $(1-p)$ is sometimes denoted as q.

5.4.2 Binomial Distribution

A binomial distribution describes the probability distribution of the number of successes for n independent trials with the same probability of success p.
A discrete random variable X with a binomial distribution is denoted

$$X \sim \text{Binomial}\,(n,\ p)$$

with

$$\begin{aligned}\Pr(X = x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ \Pr(X \leq x) &= \sum_{u=0}^{x} \binom{n}{u} p^u (1-p)^{n-u}\end{aligned}$$

for number of successes $x \in \{0,\ 1,\ \ldots,\ n\}$.
Here each individual trial is a Bernoulli trial, so that X can be written as the sum of n *independent and identically distributed* (iid) Bernoulli random variables, $Y_1,\ Y_2,\ \ldots,\ Y_n$.

$$X = Y_1 + Y_2 + \cdots + Y_n, \qquad Y_i \overset{\text{iid}}{\sim} \text{Bernoulli}\,(p) : \forall i \in \{1,\ 2,\ \ldots,\ n\}\,.$$

We can then summarise the following:

$$\begin{aligned}\text{E}(X) &= np \\ \text{Var}(X) &= np(1-p)\end{aligned}$$

5.4.3 Poisson Distribution

A Poisson distribution describes the probability distribution of the number of events N which occur over a fixed interval of time λ.
A discrete random variable N with a Poisson distribution is denoted

$$N \sim \text{Poisson}\,(\lambda)$$

with

$$\Pr\,(N = n) = \frac{\lambda^n e^{-\lambda}}{n!}$$
$$\Pr\,(N \leq n) = e^{-\lambda} \sum_{u=0}^{n} \frac{\lambda^u}{u!}$$

for number of events $n \geq 0$. We can also summarise the following:

$$\text{E}\,(N) = \lambda$$
$$\text{Var}\,(N) = \lambda$$

The Poisson PMF can be defined in terms of the Binomial PMF as $n \to \infty$ and $p \to 0$. Let $\lambda = np$, then

$$\begin{aligned} p\,(x) &= \frac{n!}{x!\,(n-x)!} p^x\,(1-p)^{n-x} \\ &= \frac{n!}{x!\,(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\ &= \frac{\lambda^x}{x!} \frac{n!}{(n-x)!} \frac{1}{n^x} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x} \end{aligned}$$

The limit of $\frac{n!}{(n-x)!}\frac{1}{n^x}$ is 1:

$$\begin{aligned} \lim_{n\to\infty} \frac{n!}{(n-x)!} \frac{1}{n^x} &= \lim_{n\to\infty} \frac{n\,(n-1)\,(n-2)\cdots(n-x+1)}{n^x} \\ &= \lim_{n\to\infty} \left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)\cdots\left(\frac{n-x+1}{n}\right) \\ &= 1 \end{aligned}$$

The term $\left(1 - \frac{\lambda}{n}\right)^n$ approaches $e^{-\lambda}$, using the substitution $u = -\frac{n}{\lambda}$:

$$\begin{aligned} \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n &= \lim_{n\to\infty} \left(1 + \frac{1}{u}\right)^{u(-\lambda)} \\ &= e^{-\lambda} \end{aligned}$$

Finally, the remaining term also evaluates to 1:

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{-x} = 1$$

Therefore by gathering the above equations, we can write the Poisson PMF as:

$$\begin{aligned} p(x) &= \lim_{n\to\infty} \frac{\lambda^x}{x!} \frac{n!}{(n-x)!} \frac{1}{n^x} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x} \\ &= \frac{\lambda^x e^{-\lambda}}{x!} \end{aligned}$$

5.4.4 Uniform Distribution

A continuous uniform distribution describes the probability distribution of an outcome within some interval, where the probability of an outcome in one interval is the same as all other intervals of the same length.
A continuous random variable X with a continuous uniform distribution is denoted

$$X \sim \text{Uniform}(a,\, b)$$

with

$$\begin{aligned} f(x) &= \frac{1}{b-a} \\ F(x) &= \frac{x-a}{b-a} \end{aligned}$$

for outcomes $a < x < b$. We can also summarise the following:

$$\begin{aligned} \text{E}(X) &= \frac{a+b}{2} \\ \text{Var}(X) &= \frac{(b-a)^2}{12} \\ m &= \frac{a+b}{2} \end{aligned}$$

5.4.5 Exponential Distribution

An exponential distribution describes the probability distribution of the time between events with rate η.
A continuous random variable T with an exponential distribution is denoted

$$T \sim \text{Exp}(\eta)$$

with

$$\begin{aligned} f(t) &= \eta e^{-\eta t} \\ F(t) &= 1 - e^{-\eta t} \end{aligned}$$

for time $t > 0$. We can also summarise the following:

$$\begin{aligned} \text{E}(T) &= \frac{1}{\eta} \\ \text{Var}(T) &= \frac{1}{\eta^2} \\ m &= \frac{\ln(2)}{\eta} \end{aligned}$$

Proof. By considering an event taking longer than t seconds, we can represent this as nothing happening over the interval $[0, t]$. Using $T \sim \text{Exp}(\eta)$ and $N \sim \text{Poisson}(\eta t)$, we have

$$\Pr(T > t) = \Pr(N = 0) = e^{-\eta t}$$

where $\lambda = \eta t$. The CDF for the exponential distribution is then

$$\begin{aligned}\Pr(T < t) &= 1 - \Pr(T > t) \\ &= 1 - e^{-\eta t}.\end{aligned}$$

□

5.4.6 Memoryless Property

In an exponential distribution with $T \sim \text{Exp}(\eta)$, the distribution of the waiting time $t + s$ until a certain event does not depend on how much time t has already passed.

$$\Pr(T > s + t \,|\, T > t) = \Pr(T > s).$$

The same property also applies in an Geometric distribution with $N \sim \text{Geometric}(p)$.

5.4.7 Normal Distribution

The normal distribution is used to represent many random situations, in particular, measurements and their errors. This distribution arises in many statistical problems and can be used to approximate other distributions under certain conditions.
A continuous random variable X with a normal distribution is denoted

$$X \sim \text{N}(\mu,\, \sigma^2)$$

with

$$\begin{aligned}f(t) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ F(t) &= \frac{1}{2}\left(1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)\end{aligned}$$

for $x \in \mathbb{R}$ where $\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, \text{d}t$ is the error function. We can also summarise the following:

$$\begin{aligned}\text{E}(X) &= \mu \\ \text{Var}(X) &= \sigma^2\end{aligned}$$

Given the complexity of the analytic expressions for the PDF and CDF of the normal distribution, we often use software to numerically determine probabilities associated with normal distributions.

5.5 Standard Normal Distribution

Given $X \sim \text{N}(\mu,\, \sigma^2)$, consider the transformation

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

so that $Z \sim \text{N}(0,\, 1)$. This distribution is called the standard normal distribution. This allows us to deal with the standard normal distribution regardless of μ and σ.

6 Sampling

This section explores the ideas of sampling and inference.

6.1 Observational and Experimental Studies

6.1.1 Observational Studies

If we are collecting or sampling data that already exists, i.e., we have no control over how we created the data, then we are conducting an observational study.

6.1.2 Experimental Studies

If data was generated in a controlled experimental environment, then the data is the result of an experimental study.

6.2 Sampling

As it is impossible to collect measurements from an entire population, we must rely on samples and sample statistics to make inferences about the population. There are several methods for this, depending on the situation.

6.2.1 Simple Random Sampling

In simple random sampling, a random subset of size n is selected from a population of size N. Simple random sampling is used in **observational studies** as data already exists. There are some caveats to this method that may lead to errors in the conclusions reached:

- Non-response bias: some members may not respond to the survey.
- Undercoverage bias: the survey may not apply to all members of the selection.
- Wording bias: the wording of the survey may lead to a biased response.

6.3 Stratified Random Sampling

Stratified random sampling is a method of sampling that divides the population into non-overlapping strata and draws random samples from each stratum.

- **Pre-stratification** is when the strata are defined before the sampling process.
- **Post-stratification** is when the strata are defined after the sampling process.

6.4 Cluster Sampling

Cluster sampling is used when there are limited resources or a lack of information about individuals in the population. It is also useful when members in each cluster are similar.

6.5 Non-random Sampling Methods

- Sequential sampling: samples are taken in a sequential manner.
- Convenience sampling: samples are self-selected from the most convenient source.
- Snowball sampling is like convenience sampling but participants are asked to recruit others.
- Quota sampling: samples are selected to balance a particular demographic.

7 Sampling Distributions

A sampling distribution is the probability distribution of a sample statistic.

7.1 Central Limit Theorem

For a sample of size n from any random probability distribution with expected value μ and variance σ^2,

$$\frac{\sqrt{n}\,(\overline{x}-\mu)}{\sigma} \xrightarrow{p} \mathrm{N}\,(0,\,1)$$

meaning that increasing the sample size will lead to a more normal distribution. In this case, a sample size of $n = 30$ is sufficient to approximate a normal distribution.

7.2 Standard Error

The standard error of a sample statistic is the standard deviation of the sampling distribution.

$$\mathrm{SE}\,(\overline{x}) = \frac{\sigma^2}{n}$$

7.3 Sample Proportion

Sometimes we are interested in estimating the population proportion from the sample proportion. For a sample of size n let x be the number of members with a particular characteristic. The sample estimate of the population proportion p is

$$\widehat{p} = \frac{x}{n}.$$

By assuming that the samples statistic x follows a binomial distribution with probability p and size n, then $\mathrm{E}\,(x) = np$ and $\mathrm{Var}\,(x) = np\,(1-p)$. Therefore the expectation is

$$\mathrm{E}\,(\widehat{p}) = p$$

and the standard error is

$$\mathrm{SE}\,(\widehat{p}) = \sqrt{\frac{p\,(1-p)}{n}}.$$

For the above to apply, we must assume that the sample proportion and size are sufficiently large. In general, if $np > 5$ and $n\,(1-p) > 5$, then we can assume that the sampling distribution of $\widehat{p}$ is approximately normal.

7.4 Assessing Normality

- Histograms: if the data is approximately normal, then the histogram will be approximately symmetric and unimodal.
- Boxplots: boxplots can be useful for showing outliers and skewness. Extreme clusters of an excessive number of outliers can be evidence of non-normality.
- Normal probability plots (q-q plots): these plots are constructed by plotting the sorted data values against their Z-scores. If the data is approximately normal, then the points will lie approximately on a straight line.

8 Large Sample Estimation

Statistical inference is the practice of using data and probabilistic models to estimate quantities and test scientific hypotheses in the face of uncertainty.

8.1 Estimation

Given some data, we can calculate sample statistics to summarise the data. The goal of statistical modelling is to rather gain insight into the population from which the data was sampled. This involves making inferences about the population parameters.

8.2 Point Estimation

In classical statistics, model parameters are unknown but assumed to be **fixed**. Parameter estimation is known as point estimation, and the resulting parameter estimators are called **point estimators**.
There are two approaches to point estimators, the method of moments and the method of maximum likelihood.

8.2.1 Method of Moments

The moments of a probability distribution are defined

$$\begin{aligned}
\mu_1 &= \mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x \\
\mu_2 &= \mathrm{E}(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x \\
&\vdots \\
\mu_n &= \mathrm{E}(X^n) = \int_{-\infty}^{\infty} x^n f(x)\,\mathrm{d}x
\end{aligned}$$

where $f(x)$ is the probability density function of the distribution. Here $\mu_1 = \mathrm{E}(X)$ and $\mathrm{Var}(X) = \mu_2 - \mu_1^2$. Sample moments are defined similarly

$$\begin{aligned} m_1 &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ m_2 &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 \\ &\vdots \\ m_n &= \frac{1}{n}\sum_{i=1}^{n} x_i^n \end{aligned}$$

where $\overline{x} = m_1$.

8.2.2 Method of Maximum Likelihood Estimation

Rather than estimating parameters using the method of moments, we can estimate parameters by maximising the likelihood function.

Definition 8.1 (Likelihood function)**.** The **likelihood function** is defined as

$$\boldsymbol{\mathcal{L}}(\theta \,|\, \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i)$$

for both continuous and discrete distributions f.

Definition 8.2 (Maximum likelihood estimator)**.** The **maximum likelihood estimator** is defined as

$$\hat{\theta} = \arg\max_{\theta} \boldsymbol{\mathcal{L}}(\theta \,|\, \boldsymbol{x}).$$

As the likelihood function is not trivial to maximise, we can instead maximise the log-likelihood function.

Definition 8.3 (Log-likelihood function)**.** The **log-likelihood function** is defined as

$$\begin{aligned} \ell(\theta \,|\, \boldsymbol{x}) &= \log\left(\boldsymbol{\mathcal{L}}(\theta \,|\, \boldsymbol{x})\right) \\ &= \sum_{i=1}^{n} \log\left(f(x_i)\right) \end{aligned}$$

Due to the monotonicity of the log function, the maximum likelihood estimator is the same as the maximum log-likelihood estimator.

Definition 8.4 (Maximum log-likelihood estimator)**.** The **maximum log-likelihood estimator** is defined as

$$\hat{\theta} = \arg\max_{\theta} \ell(\theta \,|\, \boldsymbol{x}).$$

8.3 Properties of Estimators

To assess the quality of an estimator, we can consider the following properties.

Definition 8.5 (Bias)**.** The **bias** of an estimator is defined as the difference between the expected value of the estimator $\mathrm{E}\left(\hat{\theta}\right)$ and the true value of the parameter θ_0.

$$\mathrm{Bias}\left(\hat{\theta}\right) = \mathrm{E}\left(\hat{\theta}\right) - \theta$$

An estimator $\hat{\theta}$ is **unbiased** if

$$\mathrm{E}\left(\hat{\theta}\right) = \theta$$

so that the bias is zero.

We can also compare the variance of two estimators, to assess which one is more preferable. If the variance of the estimator is small, then the estimator is more precise. Given two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$, we would choose $\hat{\theta}_1$ over $\hat{\theta}_2$ if

$$\mathrm{Var}\left(\hat{\theta}_1\right) < \mathrm{Var}\left(\hat{\theta}_2\right)$$

Definition 8.6 (Mean square error)**.** Given data x_i with variance σ^2, the estimators of $\theta = \mathrm{E}\left(X\right)$ are selected such that they minimise the **mean square error**:

$$\begin{aligned}\mathrm{MSE}\left(\hat{\theta}\right) &= \mathrm{E}\left(\left(\hat{\theta} - \theta\right)^2\right) \\ &= \mathrm{Bias}\left(\hat{\theta}\right)^2 + \mathrm{Var}\left(\hat{\theta}\right).\end{aligned}$$

This quantity is used to determine the **bias-variance trade-off** of an estimator. The **root mean square error** is defined as

$$\mathrm{RMSE}\left(\hat{\theta}\right) = \sqrt{\mathrm{MSE}\left(\hat{\theta}\right)}.$$

8.4 Confidence Intervals

In the previous section, we discuss methods to find point-estimates of parameters, that is, values that are close to the true value of the parameter. However, we may also be interested in the uncertainty of the parameter, that is, the range of values that the parameter may take.
The range of values that this parameter lies in is called a **confidence interval**. This interval ranges from the lower confidence limit (UCL) to the upper confidence limit (LCL)

$$L < \theta < U.$$

This interval has a **confidence coefficient** of $1 - \alpha$, or a **confidence level** of $(1 - \alpha)\,\%$. The confidence interval is defined as

$$CI_{1-\alpha} = \hat{\theta} \pm Z_{\alpha/2}\,\mathrm{SE}\left(\hat{\theta}\right) = \left(\hat{\theta} - Z_{\alpha/2}\,\mathrm{SE}\left(\hat{\theta}\right),\ \hat{\theta} + Z_{\alpha/2}\,\mathrm{SE}\left(\hat{\theta}\right)\right)$$

8.4.1 Confidence Interval for the Mean

Using this understanding of confidence intervals, we can now derive the confidence interval for the mean.
Recall the Z-score $Z_{\alpha/2}$, which is the value of Z such that the area to the right of the standard normal distribution is $\alpha/2$:

$$\Pr\left(Z \geq Z_{\alpha/2}\right) = \frac{\alpha}{2}.$$

Likewise, the Z-score $-Z_{\alpha/2}$ is the value of Z such that the area to the left of the standard normal distribution is also $\alpha/2$:

$$\Pr\left(Z \leq -Z_{\alpha/2}\right) = \frac{\alpha}{2}.$$

Therefore the area between $-Z_{\alpha/2}$ and $Z_{\alpha/2}$ is $1-\alpha$:

$$\begin{aligned}
\Pr\left(-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}\right) &= \Pr\left(Z \leq Z_{\alpha/2}\right) - \Pr\left(Z \leq -Z_{\alpha/2}\right) \\
&= 1 - \Pr\left(Z \geq Z_{\alpha/2}\right) - \Pr\left(Z \leq -Z_{\alpha/2}\right) \\
&= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} \\
&= 1 - \alpha.
\end{aligned}$$

Given a large sample ($n > 30$), $\overline{x}$ is the best estimator for the population mean μ and

$$\overline{x} \sim \mathrm{N}\left(\mu,\ \frac{\sigma^2}{n}\right).$$

The confidence interval for the mean is derived using the following steps:

$$\begin{array}{rcccl}
-Z_{\alpha/2} & \leq & Z & \leq & Z_{\alpha/2} \\
-Z_{\alpha/2} & \leq & \frac{\overline{x}-\mu}{\sigma/\sqrt{n}} & \leq & Z_{\alpha/2} \\
-\overline{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} & \leq & -\mu & \leq & -\overline{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\
\overline{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} & \geq & \mu & \geq & \overline{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\
\overline{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} & \leq & \mu & \leq & \overline{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}
\end{array}$$

therefore the confidence interval for the population mean is

$$CI_{1-\alpha} = \overline{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where

$$\mathrm{SE}\left(\overline{x}\right) = \frac{\sigma}{\sqrt{n}}.$$

In other words,

$$\begin{aligned}
\Pr\left(\overline{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) &= 1 - \alpha \\
\Pr\left(\overline{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu\right) &= 1 - \frac{\alpha}{2} \\
\Pr\left(\mu \leq \overline{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) &= 1 - \frac{\alpha}{2}.
\end{aligned}$$

8.4.2 Confidence Interval for the Proportion

Given the sample size n and sample proportion $\hat{p}$,

$$\hat{p} \sim \mathrm{N}\left(p,\ \frac{p\,(1-p)}{n}\right).$$

The confidence interval for the population proportion is

$$CI_{1-\alpha} = \hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$$

where the standard error is given by

$$\mathrm{SE}\,(\hat{p}) = \sqrt{\frac{p\,(1-p)}{n}}.$$

with the approximation

$$\mathrm{SE}\,(\hat{p}) = \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}.$$

Note that $n\hat{p} > 5$ and $n\,(1-\hat{p}) > 5$ are required for the approximation to be valid.

8.4.3 Confidence Interval for the Difference of Two Means

Given two population means μ_1 and μ_2, we expect the difference of the population means and the sample means to be equal. Consider the expectation of the difference of the sample means:

$$\mathrm{E}\,(\overline{x}_1 - \overline{x}_2) = \mu_1 - \mu_2$$

with standard error

$$\mathrm{SE}\,(\overline{x}_1 - \overline{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

with is estimated by

$$\mathrm{SE}\,(\overline{x}_1 - \overline{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The confidence interval for the difference of the two means is

$$CI_{1-\alpha} = \overline{x}_1 - \overline{x}_2 \pm Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

If the two populations follow a normal distribution, then the sampling distribution is exactly normal. If the two populations are not normal, then the sampling distribution is approximately normal, for $n_1 > 30$ and $n_2 > 30$.

8.4.4 Confidence Interval for the Difference of Two Proportions

Given two population proportions p_1 and p_2, we expect the difference of the population proportions and the sample proportions to be equal. Consider the expectation of the difference of the sample proportions:

$$\mathrm{E}\left(\hat{p}_1 - \hat{p}_2\right) = p_1 - p_2$$

with standard error

$$\mathrm{SE}\left(\hat{p}_1 - \hat{p}_2\right) = \sqrt{\frac{p_1\left(1-p_1\right)}{n_1} + \frac{p_2\left(1-p_2\right)}{n_2}}.$$

which is approximated by

$$\mathrm{SE}\left(\hat{p}_1 - \hat{p}_2\right) = \sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n_1} + \frac{\hat{p}_2\left(1-\hat{p}_2\right)}{n_2}}.$$

The confidence interval for the difference of the two proportions is

$$CI_{1-\alpha} = \hat{p}_1 - \hat{p}_2 \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n_1} + \frac{\hat{p}_2\left(1-\hat{p}_2\right)}{n_2}}.$$

Note that the following constraints must be satisfied:

- $n_1\hat{p}_1 > 5$
- $n_1\left(1-\hat{p}_1\right) > 5$
- $n_2\hat{p}_2 > 5$
- $n_2\left(1-\hat{p}_2\right) > 5$

9 Hypothesis Testing

Hypothesis testing is a method of statistical inference that is used to decide whether data supports a particular hypothesis.

9.1 Neyman-Pearson Lemma

The Neyman-Pearson lemma can be used to construct a hypothesis test. It requires the following:

- A test which is constructed based on the hypotheses:
  - H_0: The null hypothesis
  - H_A: The alternative hypothesis
- A test statistic $T\left(\boldsymbol{x}\right)$ defined as a function of the sample data.
- A rejection region R defined as a subset of the sample space, so that if the test statistic falls in the rejection region, then the null hypothesis is rejected.

9.1.1 Hypotheses

The hypotheses are designed such that the **null hypothesis** is *falsifiable*. The null hypothesis is typically a statement about the population parameters, and the alternative hypothesis is the complement of the null hypothesis (i.e., the null hypothesis is rejected if the alternative hypothesis is true).

9.1.2 Test Statistic

The test statistic is a function of observations modelled as random variables, and thus has its own sampling distribution. This sampling distribution can be used to identify the rejection region depending on how probable the test statistic is assuming the null hypothesis is true.

9.2 Rejection Region

The rejection region is a subset of the sample space, such that if the null hypothesis is true, the probability that the test statistic falls in the rejection region is sufficiently small.

9.3 Hypothesis Testing Procedure

Hypothesis tests can either be two-tailed (also called point null hypotheses):

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_A : \theta \neq \theta_0$$

or single-tailed, which consists of left-tail:

$$H_0 : \theta \leq \theta_0 \quad \text{vs.} \quad H_A : \theta > \theta_0$$

and right-tail tests:

$$H_0 : \theta \geq \theta_0 \quad \text{vs.} \quad H_A : \theta < \theta_0$$

for a given value of θ_0, where θ is the parameter of interest.
The rejection region R is chosen so that $\Pr\left(T\left(\boldsymbol{x} \in R\right)\right) = \alpha$, where α is the Type I error rate defined:

$$\alpha = \Pr\left(\text{reject } H_0 \,|\, H_0 \text{ is true}\right).$$

The Type II error rate, β is defined as:

$$\beta = \Pr\left(\text{failure to reject } H_0 \,|\, H_0 \text{ is false}\right)$$

where the compliment of the Type II error rate is the **power** of the test:

$$1 - \beta = \Pr\left(\text{reject } H_0 \,|\, H_0 \text{ is false}\right).$$

This can be summarised in the following table:

Decision	H_0 **is true**	H_0 **is false**
Reject H_0	α (Type I error rate)	$1 - \beta$ (Power)
Failure to reject H_0	$1 - \alpha$	β (Type II error rate)

The rejection regions are defined:

Null Hypothesis H_0	**Rejection Region** R
$\theta = \theta_0$	$\lvert T(\boldsymbol{x})\rvert > Z_{\alpha/2}$
$\theta \leq \theta_0$	$T(\boldsymbol{x}) > Z_{\alpha}$
$\theta \geq \theta_0$	$T(\boldsymbol{x}) < -Z_{\alpha}$

so that:

$$\begin{aligned}\Pr\left(Z \geq Z_{\alpha/2}\right) &= \frac{\alpha}{2} \\ \Pr\left(Z \geq Z_{\alpha}\right) &= \alpha \\ \Pr\left(Z \leq -Z_{\alpha}\right) &= \alpha.\end{aligned}$$

The Type I error rate takes the following common values:

Type I Error Rate α	**One Tail** Z_{α}	**Two-Tail** $Z_{\alpha/2}$
0.10	1.28	1.64
0.05	1.65	1.96
0.02	2.05	2.33
0.01	2.33	2.58

9.4 Hypothesis Testing for the Population Mean

Given the sample statistic $\overline{x}$,

$$\overline{x} \sim \mathrm{N}\left(\mu,\ \frac{\sigma^2}{n}\right)$$

the test statistic is defined

$$T(\boldsymbol{x}) = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,\ 1).$$

9.5 Hypothesis Testing for the Population Proportion

Given the sample statistic $\hat{p}$,

$$\hat{p} \sim \mathrm{N}\left(p,\ \frac{p(1-p)}{n}\right)$$

for $n\hat{p} > 5$ and $n(1-\hat{p}) > 5$, the test statistic is defined

$$T(\boldsymbol{x}) = \frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{p_0(1-p_0)}}.$$

9.6 Hypothesis Testing with Differences

The rejection regions for the difference between two parameters is defined:

Null Hypothesis H_0	**Rejection Region** R
$\theta_1 - \theta_2 = 0$	$\lvert T(\boldsymbol{x})\rvert > Z_{\alpha/2}$
$\theta_1 - \theta_2 \leq 0$	$T(\boldsymbol{x}) > Z_{\alpha}$
$\theta_1 - \theta_2 \geq 0$	$T(\boldsymbol{x}) < -Z_{\alpha}$

9.7 Hypothesis Testing for the Difference in Population Means

The point estimator of $\mu_1 - \mu_2$ is given by

$$\overline{x}_1 - \overline{x}_2$$

and the standard error is given by

$$\text{SE}_{\overline{x}_1 - \overline{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

The test statistic is defined

$$T(\boldsymbol{x}) = \frac{(\overline{x}_1 - \overline{x}_2) - \Delta_0}{\text{SE}_{\overline{x}_1 - \overline{x}_2}}.$$

where $\Delta_0 = \mu_1 - \mu_2$ is the hypothesised difference between the two population means.

9.8 Hypothesis Testing for the Difference in Population Proportions

The point estimator of the difference in proportions where $p_1 = p_2$ is given by

$$\hat{p}_1 - \hat{p}_2$$

and the standard error is defined

$$\text{SE}_{\hat{p}_1 - \hat{p}_2} = \sqrt{p_0(1 - p_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where

$$p_0 = \frac{x_1 + x_2}{n_1 + n_2}$$
$$p_0 = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}.$$

so that $p_0 = p_1 = p_2$. The resulting test statistic is defined:

$$T(\boldsymbol{x}) = \frac{(\hat{p}_1 - \hat{p}_2)}{\text{SE}_{\hat{p}_1 - \hat{p}_2}}.$$

When the hypothesised difference is not 0, i.e., $p_1 - p_2 = \Delta_0$, the test statistic is defined:

$$T(\boldsymbol{x}) = \frac{(\hat{p}_1 - \hat{p}_2) - \Delta_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}.$$

9.9 Power and Sample Size Selection

The power describes the probability that the test rejects the null hypothesis when the alternative hypothesis is true:

$$\begin{aligned}\text{Power} &= 1-\beta \\ &= 1-\Pr\left(|T(\boldsymbol{x})| \leq Z_{\alpha/2} \,|\, \theta=\theta^*\right) \\ &= \Pr\left(|T(\boldsymbol{x})| \geq Z_{\alpha/2} \,|\, \theta=\theta^*\right).\end{aligned}$$

Here the true value of θ is θ^* rather than θ_0.
In this equation, as n increases, the Type II error rate decreases, and hence the power increases. The sample size n can therefore be selected to achieve a desired true value of θ and a desired power.

9.10 Hypothesis Testing and Confidence Intervals

Hypothesis testing and confidence intervals both involve the probability based on the sampling distribution of a statistic. Here the rejection region is defined such that it is the **compliment** of the confidence region.
The decision to reject the null hypothesis because it falls within the rejection region is equivalent to rejecting the null hypothesis if the value θ_0 falls outside the confidence interval.

9.11 Hypothesis Testing and P-Values

Rather than constructing rejection regions based on a given Type I error rate α, we can instead measure the strength of the evidence against the null hypothesis by computing the upper tail probability of the test statistic:

$$\alpha = \Pr\left(Z \geq T(\boldsymbol{x})\right).$$

The value obtained from this calculation is called the p**-value**. The strength of the evidence against the null hypothesis increases as the p-value decreases.

9.12 Significance of Results

When interpreting the results from a test statistic, the test can only be used to reject the null hypothesis. When the strength against the null hypothesis is weak, we cannot assume that the null hypothesis is true, rather, the test is inconclusive and that there is no statistical significance.

10 Small Sample Inference

In the previous two sections, we assumed sample sizes of above 30. However, in many situations, this may be infeasible. In these situations, the following assumption may be invalid:

$$s \neq \sigma \implies \frac{\sqrt{n}\left(\overline{x}-\mu\right)}{s} \neq \frac{\sqrt{n}\left(\overline{x}-\mu\right)}{\sigma}$$

and hence we cannot use the normal distribution to approximate the sampling distribution of the test statistic. Instead, we can use **Student's t-distribution**:

$$T(\boldsymbol{x}) \sim t_\nu$$

where the degrees of freedom ν is equal to $n-1$. The t-distribution is similar to the Normal distribution, but has heavier tails for small values of n. The expectation and variance are given by

$$\begin{aligned}\mathrm{E}(X) &= 0\\ \mathrm{Var}(X) &= \frac{\nu}{\nu-2}\end{aligned}$$

for $\nu > 2$, such that the variance is always greater than 1, and converges to 1 as $\nu \to \infty$.

10.1 Inferencing

The test statistic used for inference is similar to the one used for the normal distribution:

$$T(\boldsymbol{x}) = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \sim t_{\nu,\alpha/2}.$$

Here we also consider the Type I error rate α to find the *critical value* $t_{\nu,\alpha/2}$, which is determined in the same way as for the normal distribution.

10.2 Hypothesis Testing for the Population Mean

Given a small sample, the test statistic is defined

$$T(\boldsymbol{x}) = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \sim t_{\nu,\alpha/2}.$$

10.3 Hypothesis Testing for the Difference in Population Means

Given a small sample, the property

$$\frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

will only hold if the population variances are equal, $\sigma_1^2 = \sigma_2^2$, giving us the test statistic:

$$T(\boldsymbol{x}) = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{\nu,\alpha/2}.$$

If the sample variances s_1^2 and s_2^2 are not equal, then we need to determine the common or *pooled* variance s_p^2.

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{\nu}.$$

where $\nu = n_1 + n_2 - 2$ for the two-sample t-test. This results in the following test statistic:

$$T(\boldsymbol{x}) = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

The population variances between two samples vary *greatly*, if they satisfy the following:

$$\frac{\max\left(s_1^2,\ s_2^2\right)}{\min\left(s_1^2,\ s_2^2\right)} > 3.$$

when this is the case, we must modify the test statistic to account for the different variances:

$$T\left(\boldsymbol{x}\right) = \frac{\left(\overline{x}_1 - \overline{x}_2\right) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

noting that Δ_0 is typically zero. The degrees of freedom are given by

$$\nu = \left\lfloor \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}} \right\rfloor$$

where the value is truncated (towards zero).

10.4 Paired Differences

Note that the above only applies for independent samples. In the case of two dependent samples, we must be careful about the choice of test statistic. We can instead test the mean of the differences:

$$\overline{d} = \overline{x}_1 - \overline{x}_2.$$

The test statistic is then given by:

$$T\left(\boldsymbol{x}\right) = \frac{\overline{d} - d_0}{s_d/\sqrt{n}} \sim t_{\nu,\alpha/2}$$

for the hypothesis test:

$$H_0 : \overline{d} = d_0 \quad \text{vs.} \quad H_A : \overline{d} \neq d_0$$

or the confidence interval:

$$CI_{1-\alpha} = \overline{d} \pm t_{\nu,\alpha/2}\frac{s_d}{\sqrt{n}}.$$

11 Analysis of Variance

When considering hypothesis testing, we often want to compare the effects of various *factors* that have more than one or two possible *levels*. In this case, we can use **Analysis of Variance** (ANOVA) to compare the effects of these factors.

11.1 Designing Experiments

An experiment is a trial that attempts to isolate the effects of factors of interest on specific outcomes while eliminating as much as possible, extraneous effects on outcomes. Experiments are typically designed to focus on a few factors and include some degree of repetition and randomisation to make statistical inferences.

- An **experimental unit** is the object whose outcome or response is measured and is of interest. The outcome or response measured is called the **dependent variable**.
- A **factor** is an independent variable that is controlled and varied in an experiment. The levels of a factor are the values that the factor can take on, and these levels take discrete states rather than continuous values.
- A **treatment** is a specific combination of factor levels that is applied to an experimental unit.
- A **response** is the variable measured for each experimental unit, typically a continuous numeric response.

11.2 ANOVA

ANOVA is a generalisation of the two-sample t-test. Let the outcome of an experiment for replication j of treatment i be denoted by y_{ij}.

	Treatment			
Replication	1	2	$\cdots$	I
1	y_{11}	y_{21}	$\cdots$	y_{I1}
2	y_{12}	y_{22}	$\cdots$	y_{I2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
J	y_{1J}	y_{2J}	$\cdots$	y_{IJ}

The following equation models the responses:

$$y_{ij} = \mu_i + \epsilon_{ij},$$

and the total outcome of the experiment is given by

$$\boldsymbol{y} = (y_{11},\, y_{12},\, y_{IJ})\,.$$

where there are I different treatments, and J replications of each treatment, for a total of $n = IJ$ experimental units. The total variation in experimental outcomes can be described by the **total sum of squares** (SST):

$$\text{SST} = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(y_{ij} - \overline{y}_{\cdot\cdot}\right)^2$$

where $\overline{y}_{\cdot\cdot}$ is the grand mean, or the overall average response over all treatments and repetitions. The total sum of squares can be decomposed into two parts, the **error sum of squares** (SSE), which is the pooled variation in the outcomes within treatment group i:

$$\text{SSE} = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(y_{ij} - \overline{y}_{i}\right)^2$$

and the **treatment sum of squares** (SSTr), which is the variation in the outcomes due to the different treatments:

$$\text{SSTr} = J\sum_{i=1}^{I}\left(\overline{y}_{i} - \overline{y}_{\cdot\cdot}\right)^2$$

where $\overline{y}_i$ is the mean response for treatment i given by

$$\overline{y}_i = \frac{1}{J}\sum_{j=1}^{J} y_{ij}.$$

This yields the following decomposition of the total sum of squares:

$$\text{SST} = \text{SSTr} + \text{SSE}.$$

Each source of variation is computed as a sum of squares and can be divided by the degrees of freedom, as an estimate of their contribution to the total variation. These results are presented in an **analysis of variance** (ANOVA) table.

Source	**DoF**	**SS**	**MS**	F_{test}	p
Treatment	$I-1$	SSTr	$\frac{\text{SSTr}}{I-1}$	$\frac{\text{MSTr}}{\text{MSE}}$	$\Pr\left(F_{I-1,n-I} \geq \frac{\text{MSTr}}{\text{MSE}}\right)$
Error	$n-I$	SSE	$\frac{\text{SSE}}{n-I}$		
Total	$n-1$	SST			

Note that $n = IJ$, and the mean squares are the sum of squares divided by the degrees of freedom. The statistic F is the ratio of the mean squares for the treatment and error sources of variation, or the ratio of the variation between treatments to the variation within treatments.

11.3 ANOVA Inference

Assuming that the observations y_{ij} are independent and normally distributed with mean μ_{ij} and variance σ^2, the test statistic F is distributed as an F-distribution. The F-distribution is a continuous probability distribution that describes the sampling distribution of the ratio of two sample variances. The F-distribution has two parameters, ν_1 and ν_2, which are the degrees of freedom for the numerator and denominator degrees of freedom, respectively.

$$\frac{s_1^2}{s_2^2} \sim F\left(\nu_1,\, \nu_2\right)$$

where if

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(x_i - \overline{x}\right)^2 \quad \text{and} \quad s_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}\left(y_i - \overline{y}\right)^2$$

then

$$\nu_1 = n_1 - 1 \quad \text{and} \quad \nu_2 = n_2 - 1.$$

This allows us to form a hypothesis test which tests if one treatments produces a significantly different response than another.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I \quad \text{vs.} \quad H_A : \text{at least one treatment mean is different}$$

Under these hypotheses, the null assumption means that

$$\text{SSTr} = \text{SSE}$$

or the variance between treatments was approximately equal to the variance within treatments, or that all treatment means are equal. In this case, we would reject the null hypothesis if MSTr was significantly larger than MSE, meaning that the treatment accounted for more of the total variance than the error.
This decision means that we reject the null hypothesis for large values of F_{test},

$$F_{\text{test}} > F_{\text{critical}}$$

where F_{critical} is the critical value of the F-distribution with parameters ν_1 and ν_2, based on a Type I Error rate of α.

$$F_{\text{critical}} = F_{\nu_1,\,\nu_2,\,\alpha}$$

where

$$\Pr\left(F \leq F_{\text{critical}}\right) = 1 - \alpha.$$

While the F-test for ANOVA is robust, it does not provide any information about which treatments are different.
Therefore, assuming a Type I Error rate of $\alpha = 0.05$, we can reject the null hypothesis if the column p in the ANOVA table is less than 0.05. This means that at least one treatment mean is different from the others.

11.4 Testing the Equality of the Treatment Means

Tukey's Honest Significant Difference (HSD) test is based on the distribution of

$$q_{I,\,I(J-1)} \max_{i_1,\,i_2} \frac{\left|\left(\overline{y}_{i_1} - \mu_{i_1} - \left(\overline{y}_{i_2} - \mu_{i_2}\right)\right)\right|}{s_p/\sqrt{J}}$$

where i_1 and i_2 refer to any arbitrary pairwise comparison of treatments. Using this sampling distribution, we can construct a confidence interval to perform the pairwise test of

$$H_0 : \mu_{i_1} = \mu_{i_2} \quad \text{vs.} \quad H_A : \mu_{i_1} \neq \mu_{i_2}$$

For a Type I Error rate of α we can reject H_0 if

$$\left|\overline{y}_{i_1} - \overline{y}_{i_2}\right| > q_{I,\,I(J-1),\,\alpha} \frac{s_p}{\sqrt{J}}$$

11.5 Randomised Block Design: Two-Way Classification

The single factor completely randomised designs mentioned are extensions of two-sample t-tests using F-tests for inference and adjusting the Type I Error rates to account for multiple comparisons. However, this model assumes that the only sources for observed variation in responses are the treatment effects or random effects. However, not all subjects are homogeneous, and so we must control for this source of variation. We can do this by using a **randomised block design**, as an extension to the *matched pairs design*.

11.6 Blocking

In a randomised experimental design, if there are I treatments, the experimenter chooses $J > I$ blocks each with I subjects, one for each treatment. This is done to isolate block-to-block variability that may obscure the treatment effects. The model for a randomised block design is then defined

$$y_{ij} = \alpha_i + \beta_j + \epsilon_{ij}$$

where α_i is the mean effect of the ith treatment and β_j is the mean effect of the jth block. Again using ANOVA, the total variation of the responses can be partitioned into three parts:

$$\text{SST} = \text{SSTr} + \text{SSB} + \text{SSE}$$

where SSB measures the variability between blocks:

$$\text{SSB} = I \sum_{j=1}^{J} \left(\overline{y}_j - \overline{y}_{\cdot\cdot} \right)^2$$

with

$$\overline{y}_j = \frac{1}{I} \sum_{i=1}^{I} y_{ij}$$

and

$$\text{SSE} = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(y_{ij} - \overline{y}_j - \overline{y}_i + \overline{y}_{\cdot\cdot} \right)^2 .$$

The ANOVA table is then

Source	**DoF**	**SS**	**MS**	F_{test}	p
Block	$J-1$	SSB	$\frac{\text{SSB}}{J-1}$	$\frac{\text{MSB}}{\text{MSE}}$	$\Pr\left(F_{J-1,(I-1)(J-1)} \geq \frac{\text{MSB}}{\text{MSE}}\right)$
Treatment	$I-1$	SSTr	$\frac{\text{SSTr}}{I-1}$	$\frac{\text{MSTr}}{\text{MSE}}$	$\Pr\left(F_{I-1,(I-1)(J-1)} \geq \frac{\text{MSTr}}{\text{MSE}}\right)$
Error	$(I-1)(J-1)$	SSE	$\frac{\text{SSE}}{(I-1)(J-1)}$		
Total	$n-1$	SST			

The null hypotheses are then either

H_0 : all treatment means are the same vs. H_A : at least one treatment mean is different

or

H_0 : all block means are the same vs. H_A : at least one block mean is different

Again, we can conclude that the null hypothesis is rejected if $p \leq 0.05$.

11.7 Two-Factor Randomised Block Design

The two factor randomised block design is an extension of the one factor randomised block design, where both the interaction between two different factors are of interest. The model is defined as

$$y_{ijk} = \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

where α_i is the mean effect of the first factor, β_j is the mean effect of the second factor and $(\alpha\beta)_{ij}$ is the mean effect of the interaction between the two factors.
In a two-factor experiment, factor A has I levels and factor B has J levels, and each of these factors is replicated K times. The total variation of the responses can be partitioned into four parts:

$$\text{SST} = \text{SSA} + \text{SSB} + \text{SSAB} + \text{SSE}$$

where SSA measures the variability between the means of factor A:

$$\text{SSA} = JK \sum_{i=1}^{I} \left(\overline{y}_{i} - \overline{y}_{..}\right)^2, \quad \text{with} \quad \overline{y}_{i} = \frac{1}{JK} \sum_{j=1}^{J} \sum_{k=1}^{K} y_{ijk}$$

SSB measures the variability between the means of factor B:

$$\text{SSB} = IK \sum_{j=1}^{J} \left(\overline{y}_{j} - \overline{y}_{..}\right)^2 \quad \text{with} \quad \overline{y}_{j} = \frac{1}{IK} \sum_{i=1}^{I} \sum_{k=1}^{K} y_{ijk}$$

SSAB measures the variability between the means of the interaction between the two factors:

$$\text{SSAB} = K \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\overline{y}_{ij} - \overline{y}_{i} - \overline{y}_{j} + \overline{y}_{..}\right)^2$$

and SSE measures the variability among the observations within each combination of levels for A and B:

$$\text{SSE} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \left(y_{ijk} - \overline{y}_{ij}\right)^2.$$

with

$$\overline{y}_{ij} = \frac{1}{K} \sum_{k=1}^{K} y_{ijk}$$

The ANOVA table is then

Source	**DoF**	**SS**	**MS**	F_{test}	p
A	$I-1$	SSA	$\frac{\text{SSA}}{I-1}$	$\frac{\text{MSA}}{\text{MSE}}$	$\Pr\left(F_{I-1,IJ(K-1)} \geq \frac{\text{MSA}}{\text{MSE}}\right)$
B	$J-1$	SSB	$\frac{\text{SSB}}{J-1}$	$\frac{\text{MSB}}{\text{MSE}}$	$\Pr\left(F_{J-1,IJ(K-1)} \geq \frac{\text{MSB}}{\text{MSE}}\right)$
AB	$(I-1)(J-1)$	SSAB	$\frac{\text{SSAB}}{(I-1)(J-1)}$	$\frac{\text{MSAB}}{\text{MSE}}$	$\Pr\left(F_{(I-1)(J-1),IJ(K-1)} \geq \frac{\text{MSAB}}{\text{MSE}}\right)$
Error	$IJ(K-1)$	SSE	$\frac{\text{SSE}}{IJ(K-1)}$		
Total	$IJK-1$	SST			

The null hypotheses are then either

H_0 : Factor A treatment means are the same vs. H_A : at least one treatment mean is different

or

H_0 : Factor B treatment means are the same vs. H_A : at least one treatment mean is different

or

H_0 : There is no interaction between A and B vs. H_A : A and B interact.

We can conclude that the null hypothesis is rejected if $p \leq 0.05$.

12 Linear Regression

A simple linear regression describes the relationship between a dependent variable y called the response, and an independent variable x, called the predictor. The model is given by:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where we assume that the error terms ϵ_i are independent and normally distributed with zero mean and σ^2 variance.

$$\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}\left(0,\, \sigma^2\right).$$

As y_i depends on ϵ_i, we can also show that

$$y_i \sim \mathrm{N}\left(\beta_0 + \beta_1 x_i,\, \sigma^2\right)$$

where y_i is normal for fixed values of x_i. The regression coefficients β_0 and β_1 are estimated by minimising the sum of squared residuals:

$$\left(\hat{\beta}_0,\, \hat{\beta}_1\right) = \underset{\beta_0,\, \beta_1}{\arg\min} \sum_{i=1}^{n} \epsilon_i^2 = \underset{\beta_0,\, \beta_1}{\arg\min} \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_i\right)\right)^2.$$

12.1 Estimation and Inference

By using the method of maximum likelihood for the two parameters β_0 and β_1, or using the least squares solution by minimising the squared error, we obtain the following estimators:

$$\begin{aligned}
\hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x} \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2} \\
s^2 &= \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2
\end{aligned}$$

where each estimator has a Gaussian distribution, due to them being linear combinations of x.

12.2 Hypothesis Testing

We can perform hypothesis testing on the regression coefficients β_0 and β_1. The sampling distributions for $\hat{\beta}_0$ and $\hat{\beta}_1$ are

$$\hat{\beta}_0 \sim \mathrm{N}\left(\beta_0,\, s^2_{\hat{\beta}_0}\right)$$
$$\hat{\beta}_1 \sim \mathrm{N}\left(\beta_1,\, s^2_{\hat{\beta}_1}\right)$$

where

$$s^2_{\hat{\beta}_0} = \frac{s^2 \overline{x}^2}{n \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}$$
$$s^2_{\hat{\beta}_1} = \frac{s^2}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}$$

so that the test statistics are

$$t_{\hat{\beta}_0} = \frac{\hat{\beta}_0 - \beta_0}{s_{\hat{\beta}_0}} \sim t_{n-2}$$
$$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \sim t_{n-2}$$

Based on this, we can construct confidence intervals for the regression coefficients β_0 and β_1:

$$\hat{\beta}_0 \pm t_{\alpha/2,n-2} s_{\hat{\beta}_0} \quad \text{and} \quad \hat{\beta}_1 \pm t_{\alpha/2,n-2} s_{\hat{\beta}_1}$$

where

$$\Pr\left(t < t_{\alpha/2,n-2}\right) = 1 - \alpha/2.$$

The null hypotheses are chosen such that the regression coefficients are equal to zero. As β_0 is the intercept of the linear model, it is usually not of interest, however the estimator β_1 can be tested to determine whether there is indeed a linear relationship between the predictor and the response. The hypothesis test is then

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_A : \beta_1 \neq 0$$

with the test statistic

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}.$$

12.3 Assumptions

When performing a linear regression, there are several assumptions that must be met.

1. The parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased, i.e., the expected value of $\epsilon_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ is zero.

2. The residuals ϵ_i are independent, i.e., $\mathrm{Corr}\left(\epsilon_i,\, \epsilon_j\right) = 0$ for $i \neq j$.

3. The residuals follow a Gaussian distribution, i.e., $\epsilon_i \sim \mathrm{N}\left(0,\, \sigma^2\right)$.

To test these assumptions, we can perform the following diagnostics:

1. Using a histogram of the residuals, we can check whether the residuals are unimodal and thus normally distributed.
2. Using a q-q plot, we can check whether the residuals lie on a straight line.
3. Using a plot of the residuals against the fitted values, we can check whether the residuals are independent, i.e., there are no patterns in the residuals and their variance is roughly equal or constant (they lie randomly around 0). This is known as homoscedasticity.

12.4 ANOVA for Linear Regression

By performing ANOVA, we can determine how much of the variation in y is explained by the model. The ANOVA table for linear regression is as follows:

Source	**DoF**	**SS**	**MS**
Regression	1	SSR	$\frac{\mathrm{SSR}}{1}$
Error	$n-2$	SSE	$\frac{\mathrm{SSE}}{n-2}$
Total	$n-1$	SST	

where

$$\mathrm{SSR} = \sum_{i=1}^{n} \left(\hat{y}_i - \overline{y}\right)^2$$

$$\mathrm{SSE} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

$$\mathrm{SST} = \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2$$

and $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Note that $s^2 = \frac{\mathrm{SSE}}{n-2}$.
The F statistic is

$$F = \frac{\mathrm{SSR}}{\mathrm{SSE}} \sim F_{1,n-2}$$

where the null hypothesis is

$$H_0 : \text{The regression model explains more variation in } y \text{ than the sample mean } \overline{y}$$

with the alternative hypothesis

$$H_A : \text{The regression model explains less variation in } y \text{ than the sample mean } \overline{y}.$$

where we can reject the null hypothesis if $F > F_{1-\alpha,1,n-2}$. For one independent variable x, the F statistic is equivalent to the t statistic squared, i.e., $F \equiv t^2$, and the hypothesis test is simply

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_A : \beta_1 \neq 0.$$

12.5 Coefficient of Determination

The coefficient of determination R^2 is defined as

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

which is the proportion of the total variation in y that is explained by the model. In practice, values close to 1 typically indicate "over-fitting" in the model, and sometimes small values may be acceptable. Therefore the coefficient of determination should only be used as a subjective measure.

12.6 Estimation and Prediction

Using the estimated regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$, we can estimate the response for a given predictor value x, i.e., $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. The standard error of the estimate is

$$s_{\hat{y}_i} = \sqrt{s^2 \left(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)}$$

and the confidence intervals and hypothesis testing will be based on the sampling distribution

$$\frac{\hat{y}_i - \text{E}(y_i)}{s_{\hat{y}_i}} \sim t_{n-2}$$

where the confidence interval gives us a region of values that we expect to contain the true value of y_i:

$$\hat{y}_i \pm t_{n-2,1-\alpha/2} s_{\hat{y}_i}.$$

Note that the confidence interval is narrowest when x_i is near $\overline{x}$, so that we are most confident of our estimates, for values of x close to $\overline{x}$.
If we wish to predict a value of y for a new (unobserved) value x^*, we have

$$y^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

and the standard error for this prediction is given by

$$s_{y^*} = \sqrt{s^2 \left(1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)}$$

and the prediction interval gives us a region of values that we expect to contain the true value of y^*:

$$y^* \pm t_{n-2,1-\alpha/2} s_{y^*}.$$

This interval is even wider than the confidence interval with greater confidence in predictions close to the mean. For this reason, extreme caution should be used when extrapolating outside the data domain.

12.7 ANCOVA

The ANCOVA model is a generalization of the ANOVA model that allows for the inclusion of continuous covariates that should be accounted for in our analysis. In this model, we can account for effects that could not be controlled for in the ANOVA model. Recall the ANOVA model:

$$y_{ij} = \alpha_i + \epsilon_{ij}$$

where α_i represents the ith treatment mean. The ANCOVA model accounts for covariate values x_{ij} by including the following terms:

$$y_{ij} = \alpha_i + \beta\left(x_{ij} - \overline{x}\right) + \epsilon_{ij}.$$

These terms produce similar sum of squares terms as the ANOVA model, but with the addition of the sum of the squares of the covariate (SSR) terms, with the same calculation of the F statistic. The addition of the covariate terms increases the power of the test in detecting treatment effects and is therefore useful in many situations. It should however be noted that all constraints of linear regression regarding the independence and homogeneity of residuals and their normality also apply to the covariate effects in the model.
The ANCOVA model also requires another assumption, that there is homogeneity of regression slopes, i.e., the slope, β, is approximately the same for all levels of treatment. This assumption can be visually confirmed by plotting the response variable against the covariate and fitting a least-squares line for each treatment, or by fitting an interaction term between the treatment and the covariate, and testing if this interaction is **not** statistically significant (p value greater than 0.05), which would indicate that the slopes are approximately the same for all treatments.

13 Categorical Data Analysis

In this section, data are counts of members in each category, and we are interested in the relationships between the categories.

13.1 2×2 Contingency Tables

Given two factors with two categories each, we can use a 2×2 contingency table to summarise the data. The count of the members correspond to a specific cell in the table and hence the table is a set of summary statistics displayed as cell counts.
The counts are random variables and should therefore have associated sampling distributions, however because the row sums and column sums are fixed, the sampling distribution of the counts are not independent, and we must consider a joint sampling distribution for the counts.

13.2 Hypergeometric Distribution

The hypergeometric distribution is a discrete probability distribution that describes the probability of drawing k successes from a population of size N without replacement, where the population contains a total of K successes, and we draw a sample of size n. The probability mass function is given by

$$\Pr\left(X = k\right) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

For example, if we had the data

	Factor 1	**Factor 2**	**Total Successes (K)**
Category 1	k_{11}	k_{12}	K_1
Category 2	k_{21}	k_{22}	K_2
Total Draws (n)	n_1	n_2	N

then the probability of drawing k_{11} successes from the population of size N without replacement, where the population contains a total of K_1 successes, and we draw a sample of size n_1 is given by

$$\Pr(X = k_{11}) = \frac{\binom{K_1}{k_{11}}\binom{N-K_1}{n_1-k_{11}}}{\binom{N}{n_1}}.$$

We can construct a null hypothesis that tests whether the two factors have an equal probability of being associated with a category, then we can use the hypergeometric distribution to calculate the probability of observing the data given the null hypothesis, and reject the null hypothesis if the probability (two-sided multiplied by 2) is less than 0.05. This is known as the Fisher's exact test.

13.3 Chi-Squared Distribution

When we have more than two categories, Fisher's test becomes computationally infeasible, and we must use the χ^2 distribution. The χ^2 distribution is a continuous probability distribution defined by

$$X = \sum_{i=1}^{n} Z_i^2 \sim \chi^2_\nu$$

where $Z_i \sim \mathrm{N}(0,\, 1)$ and $\nu = n - 1$.

13.3.1 Test of Homoegeneity

The χ^2 test of homogeneity tests whether the distribution of items across categories is the same for different factors.
Given the data

	Factor 1	**Factor 2**	$\cdots$	**Factor J**	**Total Successes ($n_{i.}$)**
Category 1	π_{11}	π_{12}	$\cdots$	π_{1J}	$n_{1.}$
Category 2	π_{21}	π_{22}	$\cdots$	π_{2J}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Category I	π_{I1}	π_{I2}	$\cdots$	π_{IJ}	$n_{..}$
Total Draws $n_{.j}$	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.J}$	$n_{..}$

This lets us calculate the expected counts for each cell in the table as

$$E_{ij} = \frac{n_{i.}n_{.j}}{n_{..}}.$$

We can then form the null hypothesis that the distribution of items across categories is the same for all factors,

$$H_0 : \pi_{ij} = \pi_i : \forall i, j \quad \text{vs.} \quad H_1 : \text{at least one } \pi_{ij} \text{ is different from } \pi_i.$$

The value π_i is estimated by

$$\hat{\pi}_i = \frac{n_{i.}}{n_{..}}.$$

If we assume that π_{ij} have a Poisson distribution, then the expected value and variance are equal, so that the Z-score Z_{ij} is given by

$$Z_{ij} = \frac{\pi_{ij} - E_{ij}}{\sqrt{E_{ij}}} \sim \mathrm{N}\left(0,\ 1\right).$$

The χ^2 statistic is given by

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(\pi_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi^2_\nu$$

for $\nu = (I-1)(J-1)$ degrees of freedom. We can therefore reject the null hypothesis if

$$X^2 > \chi^2_{\nu,\,1-\alpha}.$$

This is known as Pearson's χ^2 test. This test assumes that the row and column sums are fixed.

13.3.2 Test of Independence

The χ^2 test of independence tests whether the distribution of items across categories is independent of the factors. The formulation of the test statistic is the same as the test of homogeneity, and we use the following null hypothesis:

$$H_0 : \text{The categories and factors are independent}$$
$$H_A : \text{at least one category is not independent of a factor.}$$

We can reject the null hypothesis if

$$X^2 > \chi^2_{\nu,\,1-\alpha}.$$

Note that in this test we assume only the total n is fixed.