MA-M27 Probability and Statistics for Data Science

Dmitri Finkelshtein, Kristian Evans

Table of contents

Introduction **2**

1 Basic Concepts in Statistics **3**

2 Basic Concepts and Rules of Probability **14**
2.1 Introduction to Probability Theory 14
2.2 Rules of Probability 16

3 Discrete Probability Distributions **23**
3.1 Discrete Random Variables and their Characteristics 23
3.2 Bernoulli Trials 28
3.2.1 Bernoulli Distribution 28
3.2.2 Binomial Distribution 28
3.2.3 Geometric Distribution 31
3.2.4 Negative Binomial Distribution 32
3.3 Poisson Distribution 33

4 Linear Regression and Correlation, Logistic Regression **35**
4.1 The Method of Least Squares 36
Correlation 37
4.2 Simple Linear Regression 38
4.3 Multiple Linear Regression 40
4.4 (Binary) Logistic Regression 43

5 Continuous Probability Distributions **46**
5.1 Continuous Random Variables and their Characteristics 46
5.2 Main Examples 50
5.2.1 Uniform distribution 50
5.2.2 Exponential Distribution 53
5.2.3 Normal Distribution 55

6 Law of large numbers and the central limit theorem **61**
6.1 Joint behaviour of random variables 61
6.2 Law of large numbers (LLN) 66
6.3 Central limit theorem (CLT) 68

7 Hypothesis Testing: Z-tests and t-tests **71**
7.1 The Mean of n Observations from $N(\mu, \sigma^2)$ (σ^2 Known) . . . 73
7.2 The Difference between 2 Means from Normal Distributions with Known Variances 75
7.3 Large Sample Tests 76
t-tests 78
7.4 t-test: Comparing a Sample Mean 79
7.5 Paired t-test 80
7.6 Unpaired t-test 82

8 Maximum likelihood estimation **84**

9 Time series **88**
9.1 Autoregressive model $AR(1)$ 90
9.2 Autoregressive model $AR(2)$ 92
9.3 Autoregressive model $AR(p)$ 93
9.4 $ARMA(p, q)$-model 94

10 Data Reduction **95**
10.1 (Exploratory) Factor Analysis 95
Evaluating Factors 98
Rotations 99
Orthogonal Rotations 99
Varimax Rotation 100
Quartimax Rotation 100
Equamax Rotation 100
Oblique Rotation 101
10.2 Principal Component Analysis (PCA) 105
Choosing Principal Components 106

Number of Components and Rotations 107
Deciding between PCA and FA 107

Introduction

This module serves as an introduction to the concepts of Probability and Statistics required for Data Science. Both the theoretical concepts and practical examples will be explored throughout the module.

1. Basic Concepts in Statistics

This section will follow closely Chapter 3 of *Essential Math for Data Science* by T.Nield (see the Reading List on Canvas).

In simple terms, statistics is the collection, analysis and interpretation of data. Data can be qualitative (e.g. hair colour, make of car, etc.) or quantitative (numerical). Data can also be discrete or continuous, where discrete data is distinct, e.g. hair colour and continuous data takes a range of values, e.g. height.

Probability often plays a large role in statistics, as we use data to estimate how likely an event is to happen.

Statistics is the heart of many data-driven innovations. Machine learning in itself is a statistical tool, searching for possible hypotheses to correlate relationships between different variables in data.

We can easily get caught up in what the data says that we forget to ask where the data comes from. These concerns become all the more important as big data, data mining, and machine learning all accelerate the automation of statistical algorithms. Therefore, it is important to have a solid foundation in statistics and hypothesis testing so you do not treat these automations as black boxes.

Definition 1.1

Descriptive statistics involves using tools, for example calculating the mean, median, mode, and using charts, to describe data.

Note that we will recap/cover these concepts shortly.

Definition 1.2

Statistical inference tries to uncover attributes about a larger population, often based on a sample.

Descriptive statistics is the most commonly understood part of statistics and we use it to summarise data. Inferential statistics tries to uncover attributes about a larger population, often based on a sample. It is often misunderstood and less intuitive than descriptive statistics. Often we are interested in studying a group that is too large to observe, for example the average height of adults in the UK, and we have to resort to using only a few members of that group to infer conclusions about them. As you can guess, this is not easy to get right. After all, we are trying to represent a population with a sample that may not be representative.

We next consider populations, samples and bias.

Definition 1.3

A **population** is the collection of objects or people under discussion, which can be both finite and infinite.

Examples of populations could be "all Swansea University students", "all adults in the UK", or "all Golden Retrievers in Scotland".

If we are going to infer attributes about a population based on a sample, it's important the sample be as random as possible so we do not skew our conclusions, i.e. we want to avoid bias.

Definition 1.4

A **sample** is any subset of a population.

In practice it is not often possible/practical to gain information about a whole population therefore we often use a sample of the population instead. We work with samples because we want to make inferences about the population, but clearly there is a risk in coming to a false conclusion by making an

inference about the whole population using a sample. Therefore there is a need for statistics tests to ensure that similar results would be obtained if a study were to be repeated and that the results are not just due to sampling variability.

Remark 1.5

It is important to note that populations can be theoretical and not physically tangible. In these cases our population acts more like a sample from something abstract. For example, let us say that we are interested in flights that depart between 2p.m. and 3p.m. at an airport, but we lack enough flights at that time to reliably predict how often these flights are late. Therefore, we may treat this population as a sample instead from an underlying population of all theoretical flights taking off between 2p.m. and 3p.m.
Problems like this are why many researchers resort to simulations to generate data. Simulations can be useful but rarely are accurate, as simulations capture only so many variables and have assumptions built in.

Intuitively, we know that bias is when something is not evaluated in an objective way, however in statistics we have certain types of bias, see below.

Definition 1.6

A. Confirmation bias is gathering only data that supports your belief, which can even be done unknowingly. An example of this is following only social media accounts you politically agree with, reinforcing your beliefs rather than challenging them.
B. Self-selection bias is when certain types of subjects are more likely to include themselves in the experiment. For example, this could be walking onto a flight and polling the customers if they like the airline over other airlines, and using that to rank customer satisfaction among all airlines.
C. Survival bias captures only living and survived subjects, while the deceased ones are never accounted for. For example, many management consulting companies and book publishers like to identify traits of successful companies/individuals and use them as predictors for future successes. These works are pure survival bias, since these works do not account for companies/individuals that failed in obscurity, and these "success" qualities may be commonplace with failed ones as well.

We now look at some descriptive statistics in more detail, beginning with measures of location.

Definition 1.7

The **sample mean**, denoted by $\bar{x}$, of a sample of observations $x_1, x_2, \ldots, x_n$ is given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Analogously, the **population mean**, denoted by μ, of a population of observations $x_1, \ldots, x_N$ is given by

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

Example 1.8. Eight people from the general UK population were polled on the number of pets they own. The results are shown below:

$$1, 3, 2, 5, 7, 0, 2, 3.$$

These are the $x_1, \ldots, x_8$ terms as in the previous definition of the sample mean.Therefore, the sample mean is then given by:

$$\bar{x} = \frac{1+3+2+5+7+0+2+3}{8} = \frac{23}{8} = 2.875.$$

Example 1.9. We now modify the situation of the previous example to where the population is now students studying a certain Mathematics module at Swansea University. The values for the whole population are as follows:

$$2, 1, 3, 4, 2, 6, 4, 0, 1, 1, 3, 3, 4, 1, 1, 5, 5, 2, 1, 3.$$

The population mean is then given by:

$$\mu = \frac{2+1+3+4+2+6+4+0+1+1+3+3+4+1+1+5+5+2+1+3}{20}$$
$$= \frac{52}{20} = 2.6.$$

Definition 1.10

We define the **weighted mean** by

$$\frac{x_1 \cdot w_1 + x_2 \cdot w_2 + \cdots + x_n \cdot w_n}{w_1 + w_2 + \cdots + w_n},$$

where $x_1, \ldots, x_n$ denote the observations and $w_1, \ldots, w_n$ are the corresponding weights.

Example 1.11. Let us consider a module with three coursework components worth 20% each and a final exam that is worth 40%. A student scores $90, 80, 63$ and 87 respectively in these components. The weights are therefore 0.2, 0.2, 0.2 and 0.4 respectively and the weighted average is given by,

$$\frac{0.2 \cdot 90 + 0.2 \cdot 80 + 0.2 \cdot 63 + 0.4 \cdot 87}{0.2 + 0.2 + 0.2 + 0.4} = 81.4.$$

Definition 1.12

The **median** is the middle value of ranked data if n is odd and it is the mean of the two middle values if n is even, i.e.

$$\frac{\frac{1}{2}n^{\text{th}} + (\frac{1}{2}n + 1)^{\text{th}}}{2}.$$

Example 1.13. Calculate the median of the values: 5,0,1,9,7,10,14. Firstly we rank these values to obtain:

$$0, 1, 5, 7, 9, 10, 14.$$

Since n is odd (i.e. 7) we take the middle value of 7 to be the median. If we now add one value of 20 to this example, then the modified ranked data is given by:

$$0, 1, 5, 7, 9, 10, 14, 20.$$

Now we have an even number of values (i.e. 8) and hence the median is given by $\frac{7+9}{2} = 8$.

Remark 1.14

There is a concept of quantiles in descriptive statistics. The concept of quantiles is essentially the same as a median, just cutting the data in other places besides the middle. The median is actually the 50%

quantile, or the value where 50% of ordered values are behind it. Then there are the 25%, 50%, and 75% quantiles, which are known as quartiles because they cut data in 25% increments.

Definition 1.15

The **mode** is the most frequently occurring set of values. It primarily becomes useful when your data is repetitive and you want to find which values occur the most frequently.

Example 1.16. Find the mode of the values $20, 21, 19, 20, 22, 19, 20$. The most common value is 20, hence 20 is the mode of this dataset.

The mode is not necessarily unique, see the example below for an illustration.

Example 1.17. If we return to Example 1.8, we find that the mode for the number of pets is 2 and 3.

We now consider measures of variation of data. This gives us a sense of how "spread out" the data is. It is important to note that there are some calculation differences for the sample versus the population.

Definition 1.18

A. For a population of data values $x_1, \ldots, x_N$, the (population) **variance** is given by,

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

where μ is the mean of the population. Furthermore, the (population) **standard deviation** is the square root of the variance, i.e.

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

B. For a sample of data values $x_1, \ldots, x_n$, the **sample variance** is given by,

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

where $\bar{x}$ is the sample mean. Similarly, the **sample standard deviation**

is the square root of the sample variance, i.e.

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

Remark 1.19

Note that for the sample variance (and hence the sample standard deviation) we divide by $n-1$ rather than the total number of items. We do this to decrease any bias in a sample and not underestimate the variance of the population based on our sample. By counting values short of one item in our divisor, we increase the variance and therefore capture greater uncertainty in our sample.

Example 1.20. In this example we are interested in studying the number of pets owned by members of staff in a certain shop (note that this is our population, not a sample). The data are as follows:

$$0, 14, 5, 9, 7, 10, 1.$$

The mean of this sample is 6.571, hence the variance is given by

$$\begin{aligned}\sigma^2 &= \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \\ &= \frac{(0-6.571)^2 + (14-6.571)^2 + (5-6.571)^2 + (9-6.571)^2 + (7-6.571)^2 + (10-6.571)^2 + (1-6.571)^2}{7} \\ &= 21.29.\end{aligned}$$

Therefore the standard deviation is given by $\sigma = \sqrt{21.38} = 4.62$. (All to 2dp.)

Example 1.21. We now modify the previous example to the situation where the data provided are a sample of a larger population. We now calculate the sample variance and standard deviation:

$$\begin{aligned}s^2 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \\ &= \frac{(0-6.571)^2 + (14-6.571)^2 + (5-6.571)^2 + (9-6.571)^2 + (7-6.571)^2 + (10-6.571)^2 + (1-6.571)^2}{6} \\ &= 24.95.\end{aligned}$$

Therefore the sample standard deviation is given by $\sigma = \sqrt{24.95} = 4.99$. (All to 2dp.)

Notice that the sample variance and standard deviation have increased compared to the population case. This is correct as a sample could be biased and imperfect representing the population. Therefore, we increase the variance (and thus the standard deviation) to increase our estimate of how spread out the values are. A larger variance/standard deviation shows less confidence with a larger range.

Definition 1.22

Measures of characteristics of a sample are called **statistics**. (Not to be confused with the subject area of *statistics* described above.) The corresponding characteristics in the population are called **parameters**.

We work with samples because we want to make inferences about the population, but clearly there is a risk in coming to a false conclusion by making an inference about the whole population using a sample. Therefore there is a need for statistics tests to ensure that similar results would be obtained if a study were to be repeated and that the results are not just due to sampling variability.

The final topic of this chapter discusses some basic data visualisation techniques - in particular, we will consider histograms, box plots and scatter plots.

Definition 1.23

A **histogram** is a graphical display of continuous data using bars. A **bar chart** provides a graphical display of categorical data.

Note that there are no gaps between the bars of histograms and the bars can be of varying widths, i.e. they may have different sized intervals or 'bins'. Histograms can be used to help determine the distribution of the data.

Example 1.24. In this example we consider the weight of Golden Retrievers. See below for examples of histograms for this data.

This histogram does not reveal any meaningful shape to our data. The reason is because our bins are too small.

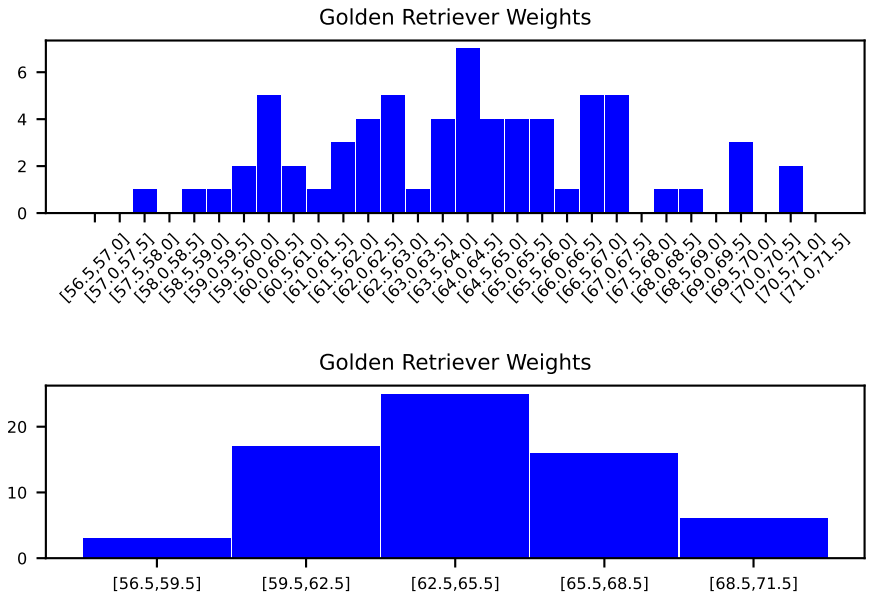


Figure 1.1

As you can see, if we get the bin sizes just right (in this case, each has a range of three pounds), we start to get a meaningful bell shape to our data.

Definition 1.25

A **box plot** or a box-and-whisker plot is a graphical technique to display data using quartiles. The box itself indicates the interquartile range, i.e. the 25% quartile to the 75% quartile. The median is indicated by a line within the box. The end of the lower (or left) whisker indicates the minimum and the top of the upper (or right) whisker denotes the maximum. Outliers are usually indicated by points.

Box plots are useful to visualise the distribution of data, in particular to check for symmetry.

Example 1.26. Let us use the data in Example 1.13 to produce the following box plot.

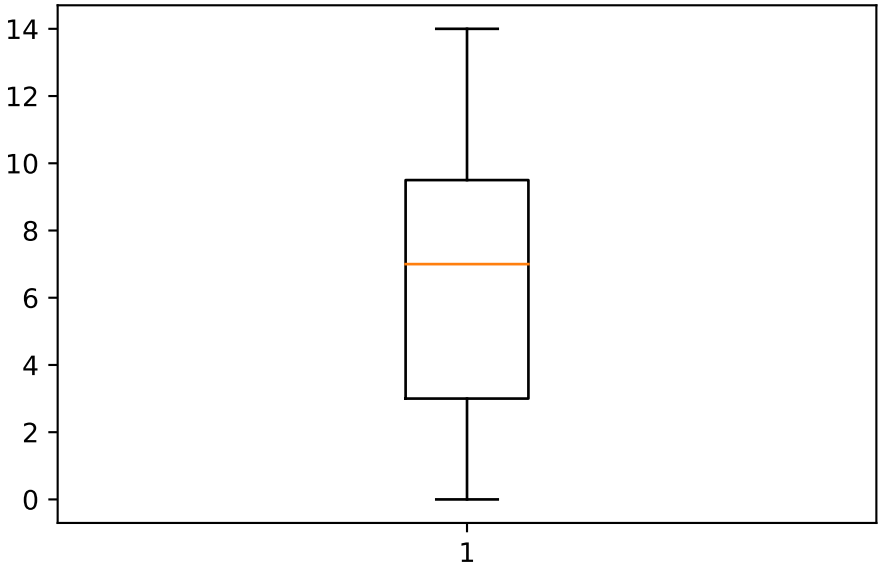


Figure 1.2

Definition 1.27

We obtain **bivariate data** when we measure two variables on each member of the population or sample.

Scatter plots can be used to plot such data, these plots can also help to visualise a relationship between the variables. One variable is plotted on the horizontal axis and the other on the vertical axis.

Example 1.28. Let us consider the data below which records exam marks for students and the corresponding time (in hours) the students spent revising for the exam.

Revision Hours	18	2	13	14	6	15	16	9	10	15
Mark	82	20	42	68	41	95	72	48	60	62

This can be represented by the following scatter plot:

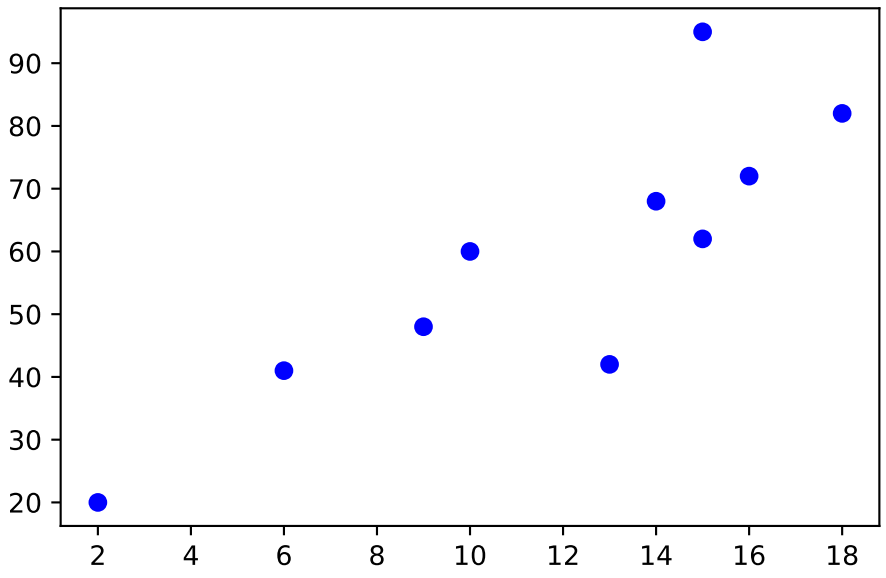


Figure 1.3

2. Basic Concepts and Rules of Probability

2.1 Introduction to Probability Theory

Probability Theory is a branch of mathematics that deals with uncertainty and randomness. It provides a framework for quantifying and analyzing uncertainty in stochastic experiments.

Probability theory plays a crucial role in data science, where we often deal with uncertain data and make predictions based on probabilities.

An **experiment** or **trial** is any procedure that can be infinitely repeated and has a well-defined set of possible outcomes.

An **outcome** (denoted by ω) is a particular result of an experiment.

A **sample space** (denoted by Ω) is the set of all possible outcomes of an experiment (i.e. $\omega \in \Omega$).

An **event** is a subset of the sample space (e.g. $A \subset \Omega$), representing a specific outcome or a collection of outcomes.

Probability (denoted by $\mathbb{P}$) is a measure of the likelihood of an event occurring. It assigns a number between 0 and 1 to an event, where 0 indicates extreme unlikelihood, and 1 indicates certainty that the event will occur. In particular,

$$0 \leq \mathbb{P}(A) \leq \mathbb{P}(\Omega) = 1.$$

Example 2.1. When we throw a coin ones, the possible outcomes are H, T (stand for 'head' and 'tail').

Therefore, $\Omega = \{H, T\}$.

There are 4 events one can consider: $\{H\}, \{T\}, \{H, T\}, \varnothing$.

Example 2.2. Consider rolling a fair six-sided dice.

The possible outcomes then are $1, 2, 3, 4, 5, 6$.

Therefore, $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Consider the event A of getting an even number: $A = \{2, 4, 6\}$.

Consider the event B of getting a prime number: $B = \{2, 3, 5\}$.

Then $\mathbb{P}(A) = \mathbb{P}(B) = \frac{1}{2}$ (think why).

Remember

The event $\varnothing$ ('empty set') describes an **impossible event** (e.g. we throw a coin or a dice with no outcome). Then

$$\mathbb{P}(\varnothing) = 0.$$

Remark. In the experiment from Example 2.2, the following $2^6 = 64$ events can be considered

$$\varnothing, \{1\}, \{2\}, \dots, \{6\},$$
$$\{1, 2\}, \{1, 3\}, \dots, \{5, 6\},$$
$$\{1, 2, 3\}, \dots, \{4, 5, 6\},$$
$$\dots$$
$$\{1, 2, \dots, 6\}.$$

Remember

If the sample space Ω contains n elements (outcomes), then the set of all events (that is the set of all subsets of Ω) is denoted by 2^Ω, and it contains 2^n events.

Memorize

We start our course with the **discrete case**, when Ω is a finite set. To calculate the probability $\mathbb{P}(A)$ of an event $A \subset \Omega$, we use the following formula:

$$\mathbb{P}(A) = \frac{\text{number of outcomes that make } A}{\text{number of all outcomes}} = \frac{\sharp(A)}{\sharp(\Omega)}$$

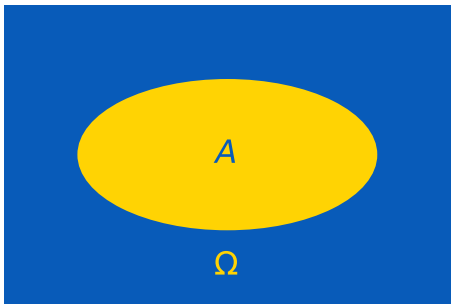


Figure 2.1: Visual representation of $\mathbb{P}(A) = \frac{\text{oval}}{\text{rectangle}}$

Example 2.3. Consider rolling twice a fair six-sided dice. Then outcomes are $\omega = (a, b)$ where $a, b \in \{1, 2, 3, 4, 5, 6\}$, i.e.

$$\Omega = \{(a, b) \mid a, b \in \{1, 2, 3, 4, 5, 6\}\}.$$

Then $\sharp(\Omega) = 6 \cdot 6 = 36$. Let A be the event of having the sum of the numbers in two rollings bigger than 10. Then

$$A = \{(5, 6), (6, 5), (6, 6)\}.$$

Therefore,

$$\mathbb{P}(A) = \frac{3}{36} = \frac{1}{12}.$$

Example 2.4. Consider drawing a card from a standard deck of 52 playing cards. The sample space Ω is the set of all 52 pairs of the form vS, where $v \in \{A, 2, 3, 4, \ldots, 10, J, Q, K\}$ is the value of a card (here J represents a Jack, Q represents a Queen, K represents a King, and A represents an Ace), and $S \in \{\clubsuit, \diamondsuit, \spadesuit, \heartsuit\}$ is the card suit (e.g. $2\diamondsuit, \ldots, A\diamondsuit$ are all diamonds). Let B be the event of drawing a red face card. Then

$$B = \{J\diamondsuit, Q\diamondsuit, K\diamondsuit, J\heartsuit, Q\heartsuit, K\heartsuit\},$$

and hence,

$$\mathbb{P}(B) = \frac{6}{52} = \frac{3}{26}.$$

2.2 Rules of Probability

Definition 2.5. The **sum** of events A and B is the event $A + B$ (also denoted $A \cup B$ or $A \vee B$) which occurs iff *either* A occurs *or* B occurs *or they both* occur.

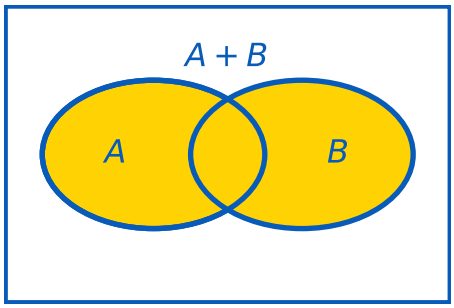


Figure 2.2: Visual representation of $A + B$

Definition 2.6. The **product** of events A and B is the event AB (also denoted $A \cap B$ or $A \wedge B$) which occurs iff *both* A *and* B occur.

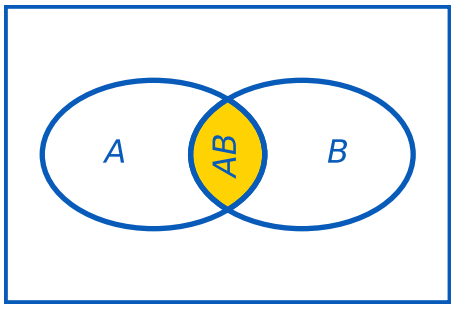


Figure 2.3: Visual representation of AB

Remember

- $A + B$ occurs under *more* outcomes than either of A or B alone:

$$\mathbb{P}(A + B) \geq \mathbb{P}(A), \quad \mathbb{P}(A + B) \geq \mathbb{P}(B).$$

- AB occurs under *less* outcomes than each of A or B alone:

$$\mathbb{P}(AB) \leq \mathbb{P}(A), \quad \mathbb{P}(AB) \leq \mathbb{P}(B).$$

Memorize

The **addition rule** states that

$$\mathbb{P}(A + B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB).$$

It can be easily interpretted using the visual representations of $\mathbb{P}(A+B)$ and $\mathbb{P}(AB)$.

Example 2.7. There is a standard deck of 52 playing cards. Find the probability of drawing either a red card or a face card (king, queen, or jack) from the deck in a single draw.

Solution: Let A be the event of drawing a red card, and B be the event of drawing a face card. Overall, there are 26 red cards, 12 face cards, and 6 red face cards. Therefore,

$$\mathbb{P}(A+B) = \frac{26}{52} + \frac{12}{52} - \frac{6}{52} = \frac{32}{52} = \frac{8}{13}.$$

Remember

Events A and B are called **mutually exclusive events** if only one of them may happen, i.e. if $AB = \varnothing$. In this case $\mathbb{P}(AB) = 0$, and the addition rule takes the form

$$\mathbb{P}(A+B) = \mathbb{P}(A) + \mathbb{P}(B).$$

Definition 2.8. The **complement** to an event A is the event A^c which occurs iff A *does not* occur. Since $\mathbb{P}(\Omega) = 1$, one has

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A).$$

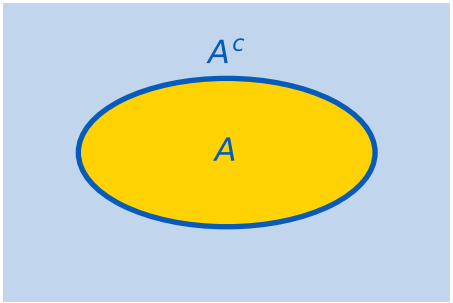


Figure 2.4: Visual representation of A^c

Example 2.9. A fair six-sided dice is rolling three times. Find the probability that the total score (the sum of three trials) will be at least 4 (event A).

Solution: The sample space Ω consists of all triples (a, b, c) with $a, b, c \in \{1, 2, \dots, 6\}$. Thus, $\sharp(\Omega) = 6^3 = 216$. The total score is 4 or more in all cases but the case $(1, 1, 1)$. Therefore, the answer is:

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c) = 1 - \frac{1}{216} = \frac{215}{216}.$$

Definition 2.10. Conditional probability $\mathbb{P}(A \mid B)$ is the probability of an event A occurring given that event B has already occurred, so we assume that $\mathbb{P}(B) \neq 0$. The formula is

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}.$$

Memorize

The **multiplication rule** follows immediately from the formula for the conditional probability:

$$\mathbb{P}(AB) = \mathbb{P}(B)\,\mathbb{P}(A \mid B) = \mathbb{P}(A)\,\mathbb{P}(B \mid A).$$

Remember

The multiplication rule can be generalised for the product of several events, e.g.

$$\mathbb{P}(ABC) = \mathbb{P}(A)\,\mathbb{P}(B \mid A)\,\mathbb{P}(C \mid AB).$$

Example 2.11. In a bag of 20 marbles, 8 are red, and 12 are green. Three marbles are drawn from the bag without replacement. What is the probability that they all are of the same color?

Solution: we need to find the probability that either $A = (r, r, r)$ or $B = (g, g, g)$ holds. Note that A and B are *mutually exclusive events*. By the multiplication rule,

$$\mathbb{P}(A) = \frac{8}{20} \cdot \frac{7}{19} \cdot \frac{6}{18},$$

and

$$\mathbb{P}(B) = \frac{12}{20} \cdot \frac{11}{19} \cdot \frac{10}{18}.$$

Therefore, by the addition rule (for mutually exclusive events),

$$\mathbb{P}(A + B) = \mathbb{P}(A) + \mathbb{P}(B) = \frac{8 \cdot 7 \cdot 6}{20 \cdot 19 \cdot 18} + \frac{12 \cdot 11 \cdot 10}{20 \cdot 19 \cdot 18} = \frac{1656}{6840} = \frac{23}{95}.$$

Definition 2.12. An event A is said to be **independent** on an event B if the occurrence of B does not affect the probability of occurrence of A. In other words, A is independent on B iff

$$\mathbb{P}(A \mid B) = \mathbb{P}(A).$$

Example 2.13. A fair coin is tossing twice. Let A: a head appeared in the first tossing, and B: a tail appeared in the second tossing. Then A and B are independent. $A = \{HH, HT\}$, $B = \{HT, TT\}$, and hence, $AB = \{HT\}$. Then $\mathbb{P}(AB) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(A)\mathbb{P}(B)$.

Remember

If A is independent on B then B is independent on A, and

$$\mathbb{P}(AB) = \mathbb{P}(A)\,\mathbb{P}(B).$$

Remark. If three (or more) events are pairwise independent: A and B are independent, the same for B and C, and for A and C, it still may be that they are not independent in total, and then, in general, $\mathbb{P}(ABC) \neq \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$ (see the multiplication rule).

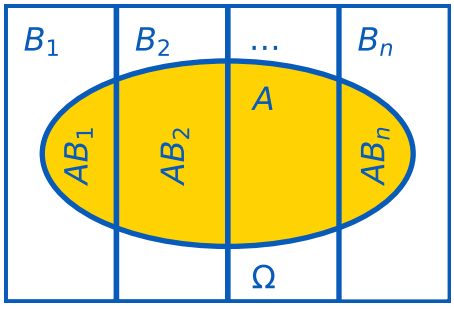


Figure 2.5: Note that $A = AB_1 + ... + AB_n$

Memorize

Let $B_1, \dots, B_n$ be pairwise exclusive events (i.e. $B_i B_j = \varnothing$ for all $i \neq j$) such that $B_1 + B_2 + ... + B_n = \Omega$ with $\mathbb{P}(B_i) \neq 0$ (it is said then that $B_1, \dots, B_n$ form a *partition* of Ω). Then the **law of total probability** holds:

$$\mathbb{P}(A) = \mathbb{P}(A \mid B_1)\,\mathbb{P}(B_1) + ... + \mathbb{P}(A \mid B_n)\,\mathbb{P}(B_n).$$

Remember

There is also a modification of the law of total probability for conditional probabilioties. If $B_1, \dots, B_n$ are as above and $\mathbb{P}(C) \neq 0$, then

$$\mathbb{P}(A \mid C) = \mathbb{P}(A \mid B_1C)\mathbb{P}(B_1 \mid C) + \dots + \mathbb{P}(A \mid B_nC)\mathbb{P}(B_n \mid C).$$

From Definition 2.10, we have that

$$\mathbb{P}(B)\mathbb{P}(A \mid B) = \mathbb{P}(AB) = \mathbb{P}(BA) = \mathbb{P}(A)\mathbb{P}(B \mid A).$$

This implies the following important statement.

Memorize

Let $\mathbb{P}(A) \neq 0$ and $\mathbb{P}(B) \neq 0$. Then **Bayes' rule** (a.k.a. Bayes' formula or Bayes' theorem) holds:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A)\mathbb{P}(B \mid A)}{\mathbb{P}(B)}.$$

It describe the *a posteriori* probability of the event A, after an experiment with the known outcome B, using the *a priori* information about the outcome B.

Remember

If $A_1, \dots, A_n$ form a *partition* of Ω (i.e. $A_1 + \dots + A_n = \Omega$ and $A_iA_j = \varnothing$ for $i \neq j$) with $\mathbb{P}(A_j) \neq 0$, then we can rewrite Bayes' rule as follows, for $\mathbb{P}(B) \neq 0$:

$$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(A_i)\mathbb{P}(B \mid A_i)}{\sum_{j=1}^{n} \mathbb{P}(B \mid A_j)\mathbb{P}(A_j)}.$$

Example 2.14. A patient has taken a test for a rare disease. The prevalence of the disease in the population is known to be very low, only 0.1%. The test correctly identifies the disease in 95% of cases when it's present. The test incorrectly indicates the presence of the disease in 3%, of cases where it's not actually present. The patient has just received a positive test result for the disease. What is the probability that he actually has the disease?

Solution: Let D denote the event of having the desease for a member of the population, then $\mathbb{P}(D) = 0.001$ (0.1%). Let T denote the event of the positive

test result. Then we know that

$$\mathbb{P}(T|D) = 0.95, \qquad \mathbb{P}(T|D^c) = 0.03.$$

By the very definition, D and D^c form a partition of Ω. Then

$$\mathbb{P}(D|T) = \frac{\mathbb{P}(T|D) \cdot \mathbb{P}(D)}{\mathbb{P}(T|D) \cdot \mathbb{P}(D) + \mathbb{P}(T|D^c) \cdot \mathbb{P}(D^c)} = \frac{0.95 \cdot 0.001}{0.95 \cdot 0.001 + 0.03 \cdot (1 - 0.001)} \approx 0.0306.$$

So, given a positive test result for the disease, the probability that the patient actually has the disease is just 3.06%.

3. Discrete Probability Distributions

3.1 Discrete Random Variables and their Characteristics

Definition 3.1. A **random variable** is a quantity which depends on random events. More rigorously, a random variable X is a function $X : \Omega \to \mathbb{R}$.

Example 3.2. Three fair six-sided dices are thrown simultaneously. Let X be the sum of scores on the dices. Then Ω consists of $\omega = (a, b, c)$ where $a, b, c \in \{1, \dots, 6\}$, and $X : \Omega \to \mathbb{R}$, $X(\omega) = a + b + c$. Thus, X may take only values from the finite set $\{3, 4, \dots, 18\}$, and hence, X is a discrete random variable.

Memorize

If a random variable X takes values only from a discrete set $\{x_1, x_2, \dots\}$, then X is called a **discrete random variable**. If X can take any values from an interval on the real line, then X is called a **continuous random variable**.

Definition 3.3. A **probability distribution** of a random variable $X : \Omega \to \mathbb{R}$ is a mapping which assigns to each interval $E \subset \mathbb{R}$ the value of $\mathbb{P}(X \in E)$.

Memorize

The **cumulative distribution function (CDF)** of a discrete random variable $X : \Omega \to \{x_1, x_2, ...\}$ is the function $F_X : \mathbb{R} \to [0, 1]$ defined by

$$F_X(x) = \sum_{x_i \leq x} \mathbb{P}(X = x_i) = \mathbb{P}(X \leq x).$$

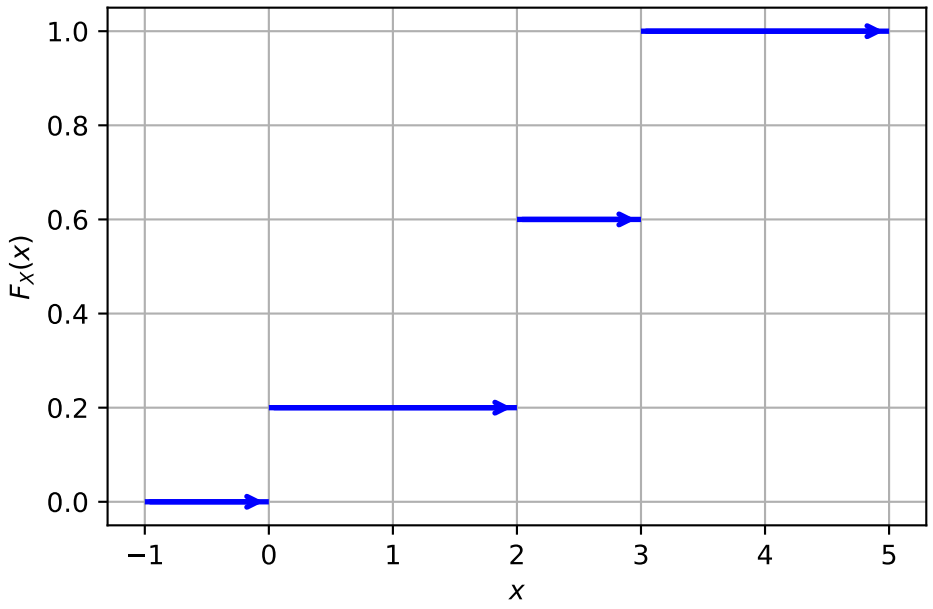


Figure 3.1: Cumulative function of the random variable $X : \Omega \to \{0, 2, 3\}$ with $\mathbb{P}(X = 0) = 0.2$, $\mathbb{P}(X = 2) = \mathbb{P}(X = 3) = 0.4$

Probability distributions provide a way to model and analyze random phenomena.

Memorize

A **discrete probability density function (discrete PDF)** a.k.a. a **probability mass function (PMF)** is a function that gives the probability that a discrete random variable is exactly equal to some value:

$$p_X(x) = \mathbb{P}(X = x).$$

Remember

- $0 \leq F_X(x) \leq 1$
- F_X is non-decreasing
- $\mathbb{P}(a<X\leq b) = F_X(b) - F_X(a)$
- $0 \leq p_X(x) \leq 1$
- If $X : \Omega \to \{x_1, x_2, ...\}$ then

$$p_X(x_1) + p_X(x_2) + ... = 1$$

Example 3.4. You throw two six-sided fair dice and calculate the sum of the numbers rolled. Let X be the random variable representing the sum. Find and sketch the CDF of X.

Solution: We have that $\Omega = \{(a, b) \mid a, b \in \{1, 2, ..., 6\}\}$, and for $\omega = (a, b)$, $X(\omega) = a + b$. Then $X \in \{2, 3, ..., 12\}$, and

$$p_X(k) = \mathbb{P}(X = k) = \frac{\sharp\{(a, b) \mid a + b = k\}}{\sharp(\Omega)}.$$

Note that $\sharp(\Omega) = 6 \cdot 6 = 36$. Next

$$2 = 1 + 1, \quad 3 = 1 + 2 = 2 + 1, \quad 4 = 1 + 3 = 2 + 2 = 3 + 1,$$
$$5 = 1 + 4 = 2 + 3 = 3 + 2 = 4 + 1, \quad 6 = 1 + 5 = 2 + 4 = 3 + 3 = 4 + 2 = 5 + 1,$$
$$7 = 1 + 6 = 2 + 5 = 3 + 4 = 4 + 3 = 5 + 2 = 6 + 1, \quad 8 = 2 + 6 = 3 + 5 = 4 + 4 = 5 + 3,$$
$$9 = 3 + 6 = 3 + 5 = 5 + 4 = 6 + 3, \quad 10 = 4 + 6 = 5 + 5 = 6 + 4,$$
$$11 = 5 + 6 = 6 + 5, \quad 12 = 6 + 6,$$

therefore,

$$p_X(2) = \frac{1}{36}, \quad p_X(3) = \frac{2}{36}, \quad p_X(4) = \frac{3}{36}, \quad p_X(5) = \frac{4}{36},$$
$$p_X(6) = \frac{5}{36}, \quad p_X(7) = \frac{6}{36}, \quad p_X(8) = \frac{5}{36}, \quad p_X(9) = \frac{4}{36},$$
$$p_X(10) = \frac{3}{36}, \quad p_X(11) = \frac{2}{36}, \quad p_X(12) = \frac{1}{36}.$$

Therefore,

$$F_X(x) = \begin{cases} 0, & \text{if } x < 2, \\ \frac{1}{36}, & \text{if } 2 \le x < 3 \\ \frac{3}{36}, & \text{if } 3 \le x < 4 \\ \frac{6}{36}, & \text{if } 4 \le x < 5 \\ \frac{10}{36}, & \text{if } 5 \le x < 6 \\ \frac{15}{36}, & \text{if } 6 \le x < 7 \\ \frac{21}{36}, & \text{if } 7 \le x < 8 \\ \frac{26}{36}, & \text{if } 8 \le x < 9 \\ \frac{30}{36}, & \text{if } 9 \le x < 10 \\ \frac{33}{36}, & \text{if } 10 \le x < 11 \\ \frac{34}{36}, & \text{if } 11 \le x < 12 \\ 1, & \text{if } x \ge 12. \end{cases}$$

Memorize

The **expected value (mean)** $\mathbb{E}(X)$ of a random variable X is the average value it takes. If $X : \Omega \to \{x_1, x_2, ...\}$, then

$$\mathbb{E}(X) := \sum_i x_i \cdot p_X(x_i) = \sum_i x_i \cdot \mathbb{P}(X = x_i).$$

Example 3.5. Calculate the expected value of the random variable X (sum of two fair six-sided dice rolls) created in Example 3.4.

Solution: We have

$$\begin{aligned} \mathbb{E}(X) &= \sum_{i=2}^{12} i \cdot p_X(i) \\ &= 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} \\ &\quad + 8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} = \frac{252}{36} = 7. \end{aligned}$$

Remember

- For any random variable $X : \Omega \to \mathbb{R}$ and any number $a \in \mathbb{R}$,

$$\mathbb{E}(aX) = a\mathbb{E}(X).$$

- For any random variables $X, Y : \Omega \to \mathbb{R}$,

$$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

Memorize

The **variance** $\mathrm{Var}(X)$ of a random variable X is a measure of the spread of its values.

$$\mathrm{Var}(X) = \mathbb{E}\Big((X - \mathbb{E}(X))^2\Big) = \mathbb{E}(X^2) - \big(\mathbb{E}(X)\big)^2 \geq 0.$$

Example 3.6. Calculate the variance of the random variable X (sum of two fair six-sided dice rolls) created in Example 3.4.

Solution: We can use the formula

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_{i=2}^{12} i^2 \cdot p_X(i) - \big(\mathbb{E}(X)\big)^2 \\
&= 2^2 \cdot \frac{1}{36} + 3^2 \cdot \frac{2}{36} + 4^2 \cdot \frac{3}{36} + 5^2 \cdot \frac{4}{36} + 6^2 \cdot \frac{5}{36} + 7^2 \cdot \frac{6}{36} \\
&\quad + 8^2 \cdot \frac{5}{36} + 9^2 \cdot \frac{4}{36} + 10^2 \cdot \frac{3}{36} + 11^2 \cdot \frac{2}{36} + 12^2 \cdot \frac{1}{36} - 7^2 \\
&= \frac{1974}{36} - 49 = \frac{35}{6} \approx 5.83.
\end{aligned}$$

Memorize

The **standard deviation** of a random variable $X : \Omega \to \mathbb{R}$ is the square root of the *variance* of X:

$$\sigma(X) := \sqrt{\mathrm{Var}(X)}.$$

Definition 3.7. Two random variables $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ are called *independent* if, for any $a, b \in \mathbb{R}$, the events

$$\{X \leq a\} := \{\omega \in \Omega \mid X(\omega) \leq a\} \qquad \text{and} \qquad \{Y \leq b\} := \{\omega \in \Omega \mid Y(\omega) < b\}$$

are independent.

Remember

- For any random variable $X : \Omega \to \mathbb{R}$ and any number $a \in \mathbb{R}$,

$$\mathrm{Var}(aX) = a^2 \,\mathrm{Var}(X).$$

- For any independent random variables $X, Y : \Omega \to \mathbb{R}$,

$$\mathbb{E}(XY) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$$

and

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

3.2 Bernoulli Trials

Definition 3.8. Conisder an experiment with two only possible outcomes: success (denoted by 1) and failure (denoted by 0). We will call such experiments **Bernoulli trials**.

3.2.1 Bernoulli Distribution

Definition 3.9. A random variable X has the **Bernoulli distribution** if X can take only two values, usually they are 1 and 0. It models, hence, a Bernoulli trial. X is fully characterized by a single parameter $p \in [0, 1]$, the probability of success, i.e. its PMF (probability mass function) is

$$p_X(1) = \mathbb{P}(X = 1) = p, \qquad p_X(0) = \mathbb{P}(X = 0) = 1 - p.$$

Remember

Since $X = X^2$, we have that

$$\mathbb{E}(X) = \mathbb{E}(X^2) = 1 \cdot p + 0 \cdot (1 - p) = p,$$
$$\mathrm{Var}(X) = p - p^2 = p(1 - p).$$

3.2.2 Binomial Distribution

Definition 3.10. The **binomial distribution** models the number of successes in a fixed number of independent and identically distributed Bernoulli trials. It is fully characterized by two parameters: n (the number of trials) and p (the probability of success in each trial). We denote this $X \sim Bin(n, p)$.

Note that the number k of successes in n trials may be any integer number between 0 (no successes at all) and n (all trials were successful).

Remark. Recall that

$$n! = 1 \cdot 2 \cdot \ldots \cdot n$$

is called the **factorial** of n.

We set

$$0! = 1,$$

and also

$$\binom{n}{0} = \binom{0}{0} = 1.$$

Memorize

The PMF of the Binomial distribution is:

$$p_X(k) = \mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

where $\binom{n}{k}$ represents the binomial coefficient, defined as

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{\overbrace{n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1)}^{k \text{ factors}}}{1 \cdot 2 \cdot 3 \cdot \ldots \cdot k}.$$

Remember

We can write $X = Y_1 + \ldots + Y_n$ where $Y_1, \ldots, Y_n$ are independent random variables with identical Bernoulli distributions with the parameter p. Then

$$\mathbb{E}(X) = np, \qquad \mathrm{Var}(X) = np(1-p).$$

Remark. Recall that the sum of **all** values of the PMF should be 1, i.e.

$$\sum_{k=0}^{n} \mathbb{P}(X = k)$$
$$= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1.$$

The latter equality is just the **binomial formual**.

Example 3.11. In a game, a player has a 20% chance of winning each round. If the player plays 5 rounds, calculate the probability of winning exactly 3 rounds.

Solution: since $n = 5$ (number of rounds) and $p = 0.2$ (probability of winning a round), we can calculate

$$\mathbb{P}(X = 3) = \binom{5}{3}(0.2)^3(1-0.2)^{5-3} = \frac{5 \cdot 4 \cdot 3}{1 \cdot 2 \cdot 3} \cdot 0.008 \cdot 0.64 = 0.0512.$$

Example 3.12. A company manufactures light bulbs, and 90% of them are of good quality, while the rest are defective. If a customer buys 50 light bulbs, what is the expected number of defective bulbs in the purchase?

Remark. Note that "success" does not need to mean that something good happened, it depends on what we are going to calculate.

Solution: Since we are interested in the number of defective bulbs, we consider a Bernoulli trial where *success* would mean that a bulb is defective. Then $p = 0.1$ and $n = 50$, therefore,

$$\mathbb{E}(X) = 50 \cdot 0.1 = 5.$$

So, the expected number of defective bulbs in the purchase is 5.

Remark. **Remember**, the following relation may be useful:

$$\binom{n}{k} = \binom{n}{n-k}.$$

For example,

$$\binom{50}{49} = \binom{50}{1} = 50,$$
$$\binom{50}{48} = \binom{50}{2} = \frac{50 \cdot 49}{2}.$$

Example 3.13. A basketball player has a free throw success rate of 70%. If she attempts 20 free throws, find the variance of the number of successful free throws.

Solution: In this case, $p = 0.7$ is the probability of making a free throw, and $n = 20$. Therefore,

$$\mathrm{Var}(X) = \sigma_X^2 = 20 \cdot 0.7 \cdot (1 - 0.7) = 4.2.$$

So, the variance of the number of successful free throws is 4.2.

3.2.3 Geometric Distribution

Definition 3.14. The **geometric distribution** models the number of trials needed to achieve the first success in a sequence of independent and identically distributed Bernoulli trials. It is characterized by a parameter p (the probability of success in each trial). We denote this $X \sim Geom(p)$.

The corresponding random variable can take *any* natural value $n = 1, 2, 3, ...$ (where n denotes the number of the first succesful trial).

Remark. We have that

$$\begin{aligned}&\sum_{n=1}^{\infty} \mathbb{P}(X = n)\\ &= \sum_{n=1}^{\infty} (1-p)^{n-1} p\\ &= \frac{p}{1-(1-p)} = 1.\end{aligned}$$

Memorize

We assume that $p \neq 0$ (otherwise, we would need infinitely many trials for success). The PMF of the geometric distribution is:

$$\mathbb{P}(X = n) = (1-p)^{n-1} p.$$

Also:

$$\mathbb{E}(X) = \frac{1}{p}, \qquad \mathrm{Var}(X) = \frac{1-p}{p^2}.$$

Example 3.15. A student is preparing for a multiple-choice exam, where each question has 4 choices, and only one is correct. If the student guesses the answers, what is the probability that the first correct answer occurs on the third guess? How many guesses the student would need to do in average to get the correct answer?

Solution: In this case, $p = \frac{1}{4}$ (probability of guessing the correct answer) and we want to find

$$\mathbb{P}(X = 3) = (1 - \frac{1}{4})^{3-1} \cdot \frac{1}{4} = \frac{9}{64}.$$

Next,

$$\mathbb{E}(X) = \frac{1}{p} = \frac{1}{\frac{1}{4}} = 4,$$

i.e. in average, the student would need to do 4 guesses to answer correctly.

3.2.4 Negative Binomial Distribution

Definition 3.16. The **negative binomial distribution** models the number of failures in a sequence of independent and identically distributed Bernoulli trials before a specified number of successes occurs. It is characterized by two parameters: r (the number of successes) and p (the probability of success in each trial). We denote this $X \sim NB(r, p)$.

If $X = k$ is the considered number of failures, the total required number of trials is $n = k + r$.

Memorize

The PMF of the negative binomial distribution is:

$$\mathbb{P}(X = k) = \binom{k+r-1}{k} p^r (1-p)^k.$$

Also:

$$\mathbb{E}(X) = \frac{r(1-p)}{p} = \frac{r}{p} - r, \qquad \mathrm{Var}(X) = \frac{r(1-p)}{p^2}$$

Example 3.17. A student is practicing basketball free throws with a success probability of 0.7. The student stops as soon as they achieves 3 successful free throws. What is the probability that by that time the student would have 2 failures (unsuccessful throws)?

Remark. The equivalent formulation of Example 3.17 is:

What is the probability that it will take the student exactly 5 trials to make 3 successful free throws?

Solution: In this problem, $p = 0.7$ (probability of success), $r = 3$ (number of desired successes), and $k = 2$ (number of failures). Then

$$\mathbb{P}(X = 2) = \binom{2+3-1}{2} \cdot (0.7)^3 \cdot (0.3)^2 = \frac{4 \cdot 3}{1 \cdot 2} \cdot 0.49 \cdot 0.09 = 0.2646.$$

Example 3.18. In a quality control process, a manufacturer wants to inspect several items to find 2 defective items. If the probability of finding a defective item is 0.1, what is the expected number of items that need to be inspected?

Solution: In this problem, $p = 0.1$, and $r = 2$. The number X of items that need to be inspected to find $r = 2$ defective items is the sum of the

number Y of proper (non-defective) items and number 2 of defective items, i.e. $X = Y + 2$, where Y has the negative binomial distribution as the number of "failures" (here "success" is to find a defective item). Then

$$\mathbb{E}(Y) = \frac{2}{0.1} - 2 = 18,$$

and hence,

$$\mathbb{E}(X) = \mathbb{E}(Y + 2) = \mathbb{E}(Y) + 2 = 20.$$

3.3 Poisson Distribution

Definition 3.19. The **Poisson distribution** models the number of independent events occurring in a fixed interval of time or space. It is characterized by a single parameter $\lambda > 0$ (the average rate of events per interval *of the same size*). We denote this by $X \sim Po(\lambda)$.

Memorize

The PMF of the Poisson distribution is:

$$\mathbb{P}(X = n) = \frac{\lambda^n}{n!} e^{-\lambda}.$$

The expected value and the variance of the Poisson random variable are equal:

$$\mathbb{E}(X) = \mathrm{Var}(X) = \lambda.$$

Remark. Note that

$$\sum_{n=0}^{\infty} \frac{\lambda^n}{n!} e^{-\lambda} = 1.$$

Remember

It is crucial to use for λ the average rate of events per interval under investigation (see Example 3.20 below).

Example 3.20. In a call center, calls arrive at an average rate of 5 calls per minute. Calculate the probability that

a) exactly 15 calls will arrive in the next 2 minutes;

b) at least 2 calls will arrive in the next 30 seconds.

Solution:

a) Since we are interested in the number of call within 2 minutes, one needs to find the average rate of calls per 2 minutes, that is $2 \cdot 5 = 10$ calls. Hence, $\lambda = 6$. Then, for $X \sim Po(10)$,

$$\mathbb{P}(X = 15) = \frac{10^{15}}{15!} e^{-10} \approx 0.0347.$$

b) We need to find

$$\mathbb{P}(X \geq 2) = \mathbb{P}(X = 2) + \mathbb{P}(X = 3) + \mathbb{P}(X = 4) + ...$$

(infinitely many). Instead, we can find the probability of compliment event:

$$\mathbb{P}(X \leq 1) = 1 - \mathbb{P}(X \geq 2).$$

To find λ we notice that the time interval is not 30 seconds, i.e. 0.5 minutes, and hence, the average rate of calls per 30 seconds is $0.5 \cdot 5 = 2.5$. Hence, $\lambda = 2.5$ and

$$\begin{aligned} \mathbb{P}(X \leq 1) &= \mathbb{P}(X = 0) + \mathbb{P}(X = 1) \\ &= \frac{2.5^0}{0!} e^{-2.5} + \frac{2.5^1}{1!} e^{-2.5} = e^{-2.5} + 2.5e^{-2.5} \\ &= 3.5e^{-2.5} \approx 0.2873. \end{aligned}$$

Therefore,

$$\mathbb{P}(X \geq 2) = 1 - \mathbb{P}(X \leq 1) \approx 1 - 0.2873 = 0.7127.$$

Remember

Often, in the problems, λ is understood is the average rate *per unit time*. Then the PMF of the distribution of events accuring in a time interval of length t (meaning "t units of time") is

$$\mathbb{P}(X = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}.$$

4. Linear Regression and Correlation, Logistic Regression

Useful resources for linear regression are *Theory and Problems of Probability and Statistics* by M.R.Spiegel and *How to Use Statistics* by S.Lakin. Furthermore, useful resources for logistic regression are *Generalised Linear Models* by P.McCullagh and J.A.Nelder, and *Using Multivariate Statistics* by B.G.Tabachnick and L.S.Fidell. The material taught in this chapter will also be met from a machine learning perspective in *MA-M28 Modelling and Machine Learning* — please see chapters 5 and 6 of *Essential Math for Data Science* by T.Nield if you would like an insight into this.

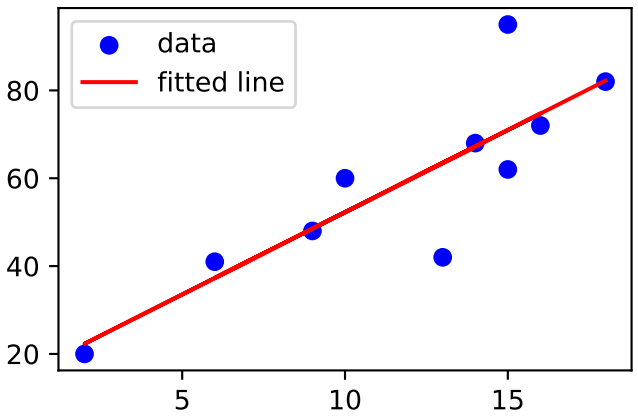


Figure 4.1

Recall in Example 1.28 we discussed bivariate data and associated scatter plot. Sometimes it is visually clear that a linear relationship exists between the variables, for example, in the scatter plot in Example 1.28 it seems that the more time is spent revising the higher the exam mark the student receives. The diagram above contains the same data, but with a line indicating the likely relationship between the variables.

A linear regression fits a straight line to observed data, attempting to demonstrate a linear relationship between variables and make predictions on new data yet to be observed. The following method will begin to address this.

4.1 The Method of Least Squares

Once a statistical model has been set up, its parameters must be estimated from the data. The method of least squares can provide good such estimates. The method minimises the sum of squared residuals, i.e. it minimises the sum of the square of the differences between an observed value and the value produced by the model. We will concentrate on linear least squares which will provide the theory behind simple linear regression.

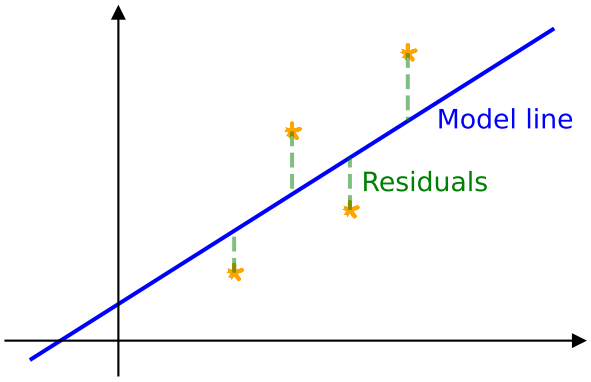


Figure 4.2

Assuming that we have data pairs $(x_1, y_1), \ldots, (x_n, y_n)$, the model line will take the form

$$\mathbb{E}(y) = \beta_0 + \beta_1 x,$$

where y is the **dependent variable (response variable)** (it depends on x!) and x is the **independent variable (explanatory/predictor variable)** (it

does not depend on another variable). In fact the variables will be related by

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where ϵ is the error. It is clear that there is some error when using our model line. By minimising the sum of the square of the residuals (using partial differentiation) we arrive at the following estimates:

Remember

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}};$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Note that

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2.$$

The least squares model is therefore

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{4.1}$$

Correlation

The strength of a linear relationship between the variables can be measured by the Pearson correlation coefficient (or just the correlation coefficient) which is given by

Remember

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}.$$

r can take values between -1 and 1 where -1 and 1 represent a perfect linear relationship. $r = 0$ means that there is no linear relationship. A positive value of r denotes a positive correlation while a negative value of r denotes a negative correlation. As a general rule of thumb we use the following criteria:

Remember

$$\begin{aligned} |r| &> 0.7 \quad \text{Strong correlation} \\ 0.7 \geq |r| &> 0.4 \quad \text{Moderate correlation} \\ |r| &\leq 0.4 \quad \text{Weak correlation} \end{aligned}$$

4.2 Simple Linear Regression

This is a process to obtain a suitable straight line to predict values of one variable (y) from the values of the other (x), where there is a linear relationship between them. The most common approach is to use the least squares model above, i.e.

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

We can use the correlation coefficient r to test the strength of the linear relationship between the variables, as above, but also r^2 (often R^2 is used) can be used. r^2 gives the proportion of the variance of y that is explained by variation in x, and the closer this value is to 1 the stronger the relationship (clearly the closer r^2 is to 0 the weaker the relationship). Essentially it measures how well the regression model fits the real data.
Now we are in a position to return to the example above and form a regression line for the data.
Note that correlation does not necessarily imply causation — variables may be related for no apparent reason.

Example 4.1. This example investigates the relationship between revision time and exam marks — see the table below for data. The calculations that we need for the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ can be found in the following table using the fact that $\bar{x} = 11.8$ and $\bar{y} = 59$:

Revision Hours (x_i)	18	2	13	14	6	15	16	9	10	15	
Mark (y_i)	82	20	42	68	41	95	72	48	60	62	*
$x_i - \bar{x}$	6.2	-9.8	1.2	2.2	-5.8	3.2	4.2	-2.8	-1.8	3.2	
$y_i - \bar{y}$	23	-39	-17	9	-18	36	13	-11	1	3	
$(x_i - \bar{x})(y_i - \bar{y})$	142.6	382.2	-20.4	19.8	104.4	115.2	54.6	30.8	-1.8	9.6	
$(x_i - \bar{x})^2$	38.44	96.04	1.44	4.84	33.64	10.24	17.64	7.84	3.24	10.24	
$(y_i - \bar{y})^2$	529	1521	289	81	324	1296	169	121	1	9	

Then

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{837}{223.6} = 3.743$$

and

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = 59 - 3.743 \times 11.8 = 14.829.$$

Therefore the equation of the regression line is given by,

$$\widehat{y} = 3.743x + 14.829.$$

The Pearson correlation coefficient is given by

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{837}{\sqrt{223.6}\sqrt{4340}} = 0.85,$$

indicating a strong positive correlation. The value of r^2 is 0.723 indicating that 72.3% of the variance of the Marks is explained by Revision. We conclude therefore that the regression model fits the data well in this case.

The regression line can then be used to estimate the value of y for a given x, for example, if we wanted to predict the exam mark obtained for 11 hours of revision we obtain,

$$\widehat{y} = 3.743 \times 11 + 14.829 \approx 56.$$

Common sense should be used when predicting using the regression model; we cannot predict outside the possible range of the x values, we would not, for example, try to predict what happens if a student were to revise for -5 hours.

Example 4.2. Calculate a regression line for the data below (the relevant conditions for linear regression may be assumed):

Classes Missed (x_i)	3	30	20	7	24	1	5	16	10	12
Mark (y_i)	82	20	42	68	41	95	72	48	60	62

where

$$\begin{aligned}
\bar{x} &= 12.8 \\
\bar{y} &= 59 \\
S_{xx} &= \sum_{i=1}^{n}(x_i - \bar{x})^2 = 821.6 \\
S_{xy} &= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = -1839 \\
S_{yy} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = 4340.
\end{aligned}$$

We now find

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-1839}{821.6} = -2.238$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 87.65.$$

Therefore the regression line is given by

$$\hat{y} = -2.238x + 87.650.$$

and the correlation coefficient by

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-1839}{\sqrt{821.6}\sqrt{4340}} = -0.97,$$

indicating a very strong negative correlation. $r^2 = 0.941$, indicating that 94.1% of the variance of Marks is explained by Classes Missed and we conclude that the regression model fits the data very well.

4.3 Multiple Linear Regression

This method is an extension of the model we met in simple linear regression, i.e.

$$y = \beta_0 + \beta_1 x + \epsilon$$

to the model

Remember

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon,$$

where we have n independent variables, $x_1, \ldots, x_n$ and the single dependent variable y (dependent on these $x_1, \ldots, x_n$).

In particular, we will concentrate on the case where we have 2 independent variables x_1 and x_2, i.e. the model,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

Obtaining the equations for the coefficients is easier if we code the variables in the following way:

$$v_i = y_i - \bar{y},$$
$$u_{1i} = x_{1i} - \bar{x}_1,$$
$$u_{2i} = x_{2i} - \bar{x}_2.$$

We now write the model as

$$V = \beta_0' + \beta_1 u_1 + \beta_2 u_2 + e.$$

The constant term changes from β_0 to β_0'. Again the method of least squares is used where the quantity to be minimised with respect to variation in β_0', β_1 and β_2 is

$$Q = \sum_{i=1}^{n} (v_i - \beta_0' - \beta_1 u_{1i} - \beta_2 u_{2i})^2.$$

Again, we will not go into the details of this process as it uses methods that are beyond the scope of this module. Essentially the process results with the equations:

$$\hat{\beta}_1 \sum_{i=1}^{n} u_{1i}^2 + \hat{\beta}_2 \sum_{i=1}^{n} u_{1i}u_{2i} = \sum_{i=1}^{n} u_{1i}v_i,$$
$$\hat{\beta}_1 \sum_{i=1}^{n} u_{1i}u_{2i} + \hat{\beta}_2 \sum_{i=1}^{n} u_{2i}^2 = \sum_{i=1}^{n} u_{2i}v_i.$$

These are sometimes called the normal equations — although no relation to the normal distribution. Using the notations

$$S_{pq} = \sum_{i=1}^{n} u_{pi}u_{qi} = \sum_{i=1}^{n} (x_{pi} - \bar{x}_p)(x_{qi} - \bar{x}_q), p, q = 1, 2,$$
$$S_{0p} = \sum_{i=1}^{n} u_{pi}v_i = \sum_{i=1}^{n} (x_{pi} - \bar{x}_p)(y_i - \bar{y}), p = 1, 2,$$

we may write the normal equations as

$$\hat{\beta}_1 S_{11} + \hat{\beta}_2 S_{12} = S_{01},$$
$$\hat{\beta}_1 S_{12} + \hat{\beta}_2 S_{22} = S_{02}.$$

Using standard techniques for solving simultaneous equations and $D = S_{11}S_{22} - S_{12}^2$ we find

$$\hat{\beta}_1 = \frac{S_{22}S_{01} - S_{12}S_{02}}{D}$$
$$\hat{\beta}_2 = \frac{S_{11}S_{02} - S_{12}S_{01}}{D}.$$

Finally,

$$\hat{\beta}_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2.$$

this gives us the following model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2.$$

If we return to the example of predicting exam marks, but now based on both Revision Hours and Coursework Marks, this would be a situation of a multiple linear regression with two predictor/independent variables.

Example 4.3. Dice were thrown to obtain ten values of each of the following: X_1 = value on a twelve-sided die, X_2 = twice the value on a six-sided die, Z =value on a six-sided die, $Y = X_1 + X_2 + Z$. The values obtained were

											Total
X_1	8	7	11	3	10	7	5	11	8	7	77
X_2	6	12	4	8	12	10	4	2	4	4	66
Y	18	23	21	17	25	18	13	19	14	14	182

We find that

$$S_{11} = \sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2 = \sum_{i=1}^{n} x_{1i}^2 - \frac{(\sum_{i=1}^{n} x_{1i})^2}{n} = 58.1$$

$$S_{22} = \sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2 = \sum_{i=1}^{n} x_{2i}^2 - \frac{(\sum_{i=1}^{n} x_{2i})^2}{n} = 120.4$$

$$S_{12} = \sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) = \sum_{i=1}^{n} x_{1i}x_{2i} - \frac{\sum_{i=1}^{n} x_{1i} \sum_{i=1}^{n} x_{2i}}{n} = -16.2$$

$$S_{01} = \sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(y_i - \bar{y}) = \sum_{i=1}^{n} x_{1i}y_i - \frac{\sum_{i=1}^{n} x_{1i} \sum_{i=1}^{n} y_i}{n} = 45.6$$

$$S_{02} = \sum_{i=1}^{n}(x_{2i} - \bar{x}_2)(y_i - \bar{y}) = \sum_{i=1}^{n} x_{2i}y_i - \frac{\sum_{i=1}^{n} x_{2i} \sum_{i=1}^{n} y_i}{n} = 84.8.$$

This gives the normal equations as follows:

$$58.1\hat{\beta}_1 - 16.2\hat{\beta}_2 = 45.6$$
$$-16.2\hat{\beta}_1 + 120.4\hat{\beta}_2 = 84.8.$$

Then

$$D = 58.1 \times 120.4 - (16.2)^2 = 6732.8$$

$$\hat{\beta}_1 = \frac{120.4 \times 45.6 + 16.2 \times 84.8}{D} = 1.019$$

$$\hat{\beta}_2 = \frac{58.1 \times 84.8 + 16.2 \times 45.6}{D} = 0.841$$

$$\hat{\beta}_0 = 18.2 - 1.019 \times 7.7 - 0.841 \times 6.6 = 4.803.$$

Therefore the regression equation is

$$\hat{y} = 4.803 + 1.019x_1 + 0.841x_2.$$

4.4 (Binary) Logistic Regression

The main use of logistic regression is to predict a binary outcome from a linear combination of independent variables. For example, we may wish to predict the probability of passing an exam from the independent variables attendance at lectures and hours of revision. This is also used in insurance to calculate the propensity to claim.

Firstly, the dependent variable $Y \sim Bin(1, p)$, and we want to use the a linear combination of the independent variables to predict p.

For this method we make use of the **logit** function, which is the **log odds**. In particular, we have

$$odds = \frac{p}{1-p}.$$

Then, the link function we use is

$$\text{logit}(p) = \ln(odds) = \ln\left(\frac{p}{1-p}\right).$$

The graph of this function is as follows:

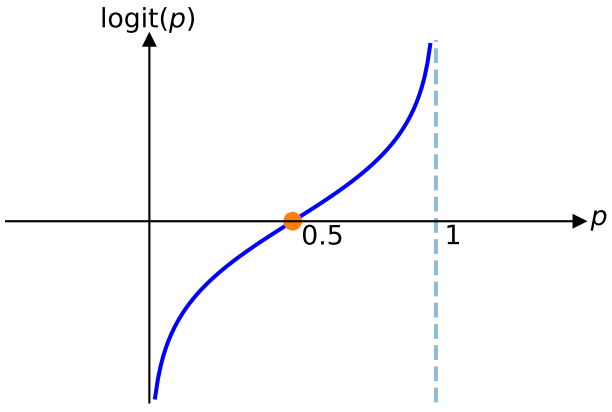


Figure 4.3

Therefore, the model we consider is the following:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{4.2}$$

for independent, or predictor variables $x_1, \ldots, x_n$ and constant coefficients $\beta_0, \ldots, \beta_n$.

As we will be seeking estimates of p, i.e. $\hat{p}$, from independent variables that could take any real value, it makes sense to next consider the inverse of the logit function. Let

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n,$$

then

$$\begin{aligned} e^y &= \frac{p}{1-p} \\ \implies p &= (1-p)e^y = e^y - pe^y \\ \implies p(1+e^y) &= e^y \\ \implies p &= \frac{e^y}{1+e^y} = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}} \end{aligned}$$

This is an example of a sigmoid function, the graph of which is as follows:

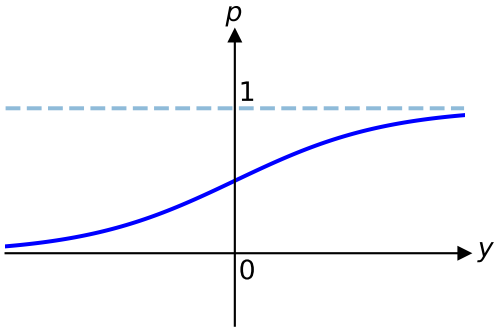


Figure 4.4

From the graph above, we can see that for any real input y, we get $0 < p < 1$ which intuitively makes sense.

In practice, we obtain an estimate $\hat{p}$ of p using maximum likelihood estimates of the coefficients $\beta_0, \beta_1, \ldots, \beta_n$, i.e.

$$\widehat{p} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\cdots+\hat{\beta}_n x_n}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\cdots+\hat{\beta}_n x_n}}.$$

5. Continuous Probability Distributions

5.1 Continuous Random Variables and their Characteristics

We recall that, see Definition 3.1, a *random variable* X is a function $X : \Omega \to \mathbb{R}$.

Reminder

If X can take any values from an interval on the real line, then X is called a **continuous random variable**.

We recall also, see Definition 3.3, that a *probability distribution* of a random variable $X : \Omega \to \mathbb{R}$ is a mapping which assigns to each interval $E \subset \mathbb{R}$ the value of $\mathbb{P}(X \in E)$.

Memorize

The **cumulative distribution function (CDF)** of a continuous random variable $X : \Omega \to \mathbb{R}$ is the *continuous* function $F_X : \mathbb{R} \to [0, 1]$ defined by

$$F_X(x) = \mathbb{P}(X \leq x).$$

As a result,

$$\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a).$$

Remark. As you can see, the same formula holds for a discrete random variables, however, for continuous r.v. it does not provide an expression to

calculate the probability. For this, we need the probaility denisty funciton defined below.

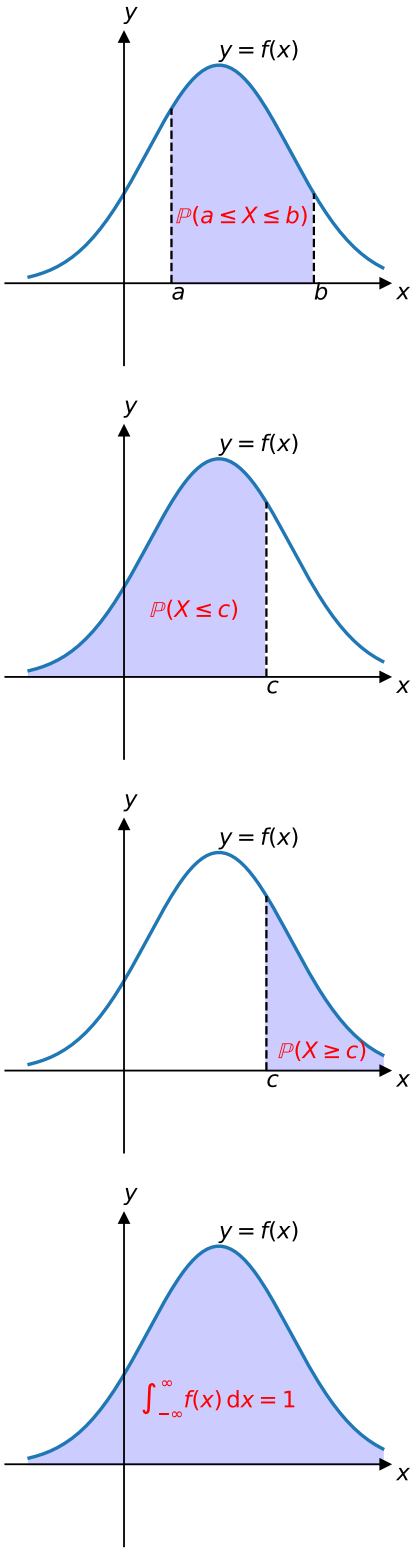


Figure 5.1: Visual representation of probabilities for a continuous random variable

Memorize

The **probability density function (PDF)** of a continuous random variable $X : \Omega \to \mathbb{R}$ is the function $f_X : \mathbb{R} \to [0, \infty)$, such that, for any $a, b \in \mathbb{R}$,

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\,\mathrm{d}x.$$

Properties:

- $f_X(x) \geq 0$
- $\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$
- $F_X(x) = \int_{-\infty}^{x} f_X(y)\,\mathrm{d}y$
- $f_X(x) = \dfrac{d}{dx} F_X(x) = F_X'(x)$

Remember

For conitnuous random variables,

$$\mathbb{P}(X = a) = 0, \qquad a \in \mathbb{R}.$$

Therefore,

$$\begin{aligned}\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X \leq b) \\ = \mathbb{P}(a < X < b) = \int_a^b f_X(x)\,\mathrm{d}x.\end{aligned}$$

Also, the following formulas may be useful:

$$\mathbb{P}(X \leq c) = \mathbb{P}(X < c) = F_X(c) = \int_{-\infty}^{c} f_X(x)\,\mathrm{d}x,$$

$$\mathbb{P}(X \geq c) = \mathbb{P}(X > c) = 1 - F_X(c) = \int_{c}^{\infty} f_X(x)\,\mathrm{d}x.$$

Recall that the *expected value (mean)* $\mathbb{E}(X)$ of a random variable X is the average value it takes.

Memorize

If $X : \Omega \to \mathbb{R}$ is a continuous random variable, then

$$\mathbb{E}(X) := \int_{-\infty}^{\infty} x \cdot f_X(x)\,\mathrm{d}x$$

(provided that the integral takes the finite value).

Remember

More generally, if $g : \mathbb{R} \to \mathbb{R}$, then

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x.$$

In particular,

$$\mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\,\mathrm{d}x.$$

Recall that the *variance* $\mathrm{Var}(X)$ of a random variable X is a measure of the spread of its values, and it is defined through the formulas

$$\mathrm{Var}(X) = \mathbb{E}\Big((X - \mathbb{E}(X))^2\Big) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \geq 0.$$

Memorize

If $X : \Omega \to \mathbb{R}$ is a continuous random variable, then

$$\begin{aligned}\mathrm{Var}\, X &= \int_{-\infty}^{\infty} (x - \mathbb{E}(X))^2 f(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} x^2 \cdot f_X(x)\,\mathrm{d}x - \left(\int_{-\infty}^{\infty} x \cdot f_X(x)\,\mathrm{d}x\right)^2.\end{aligned}$$

5.2 Main Examples

5.2.1 Uniform distribution

Definition 5.1. The **uniform distribution** is a continuous probability distribution where all outcomes within a specified interval are equally likely.

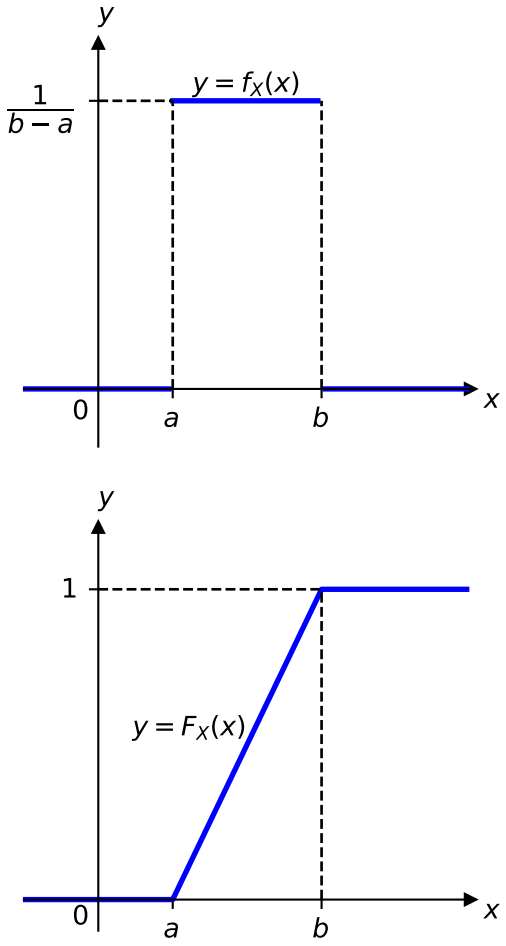


Figure 5.2: Graphs of PDF and CDF of $X \sim U(a, b)$.

Memorize

The PDF of the uniform distribution on an interval $[a, b]$ is given by

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0, & \text{otherwise.} \end{cases}$$

The CDF of the uniform distribution on an interval $[a, b]$ is given by

$$F(x) = \begin{cases} 0, & \text{if } x < a \\ \dfrac{x-a}{b-a}, & \text{if } a \le x \le b \\ 1, & \text{if } x > b. \end{cases}$$

Notation for a random variable with such distribution:

$$X \sim U(a, b).$$

The mean and variance of X are given by:

$$\mathbb{E}(X) = \frac{a+b}{2},$$

$$\mathrm{Var}(X) = \frac{(b-a)^2}{12}.$$

Example 5.1. Let $X \sim U(1, 5)$.

a) Find $\mathbb{P}(2 < X < 4)$.

b) Find $c \in [1, 5]$ such that $\mathbb{P}(3 < X < c) = \dfrac{1}{3}$.

Solution: a) Here

$$f_X(x) = \frac{1}{5-1} = \frac{1}{4}, \qquad x \in [1, 5].$$

Therefore,

$$\mathbb{P}(2 < X < 4) = \int_2^4 \frac{1}{4}\,\mathrm{d}x = \frac{1}{4} \cdot (4-2) = \frac{1}{2}.$$

b) We have

$$\frac{1}{3} = \mathbb{P}(3 < X < c) = \int_3^c \frac{1}{4}\,\mathrm{d}x = \frac{1}{4}(c-3),$$

hence,

$$c - 3 = \frac{4}{3}, \qquad c = \frac{13}{3} \in [1, 5].$$

5.2.2 Exponential Distribution

Definition 5.2. Recall that the (discrete) Poisson random variable models the number of independent events occurring in a fixed interval of time.The exponential distribution is a continuous probability distribution that models the time between these independent events. It is commonly used to model waiting times.

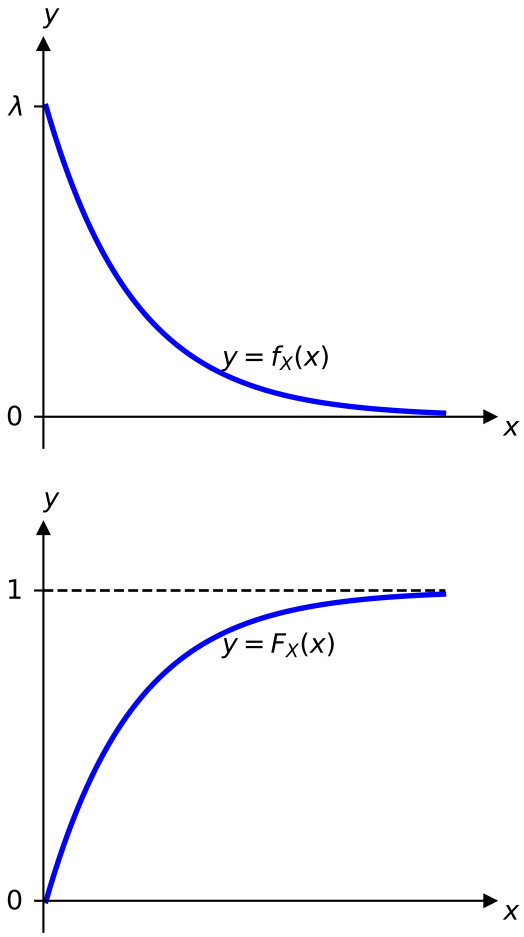


Figure 5.3: Graphs of PDF and CDF of $X \sim \text{Exp}(\lambda)$. Graphs are shown for $x \geq 0$ only. Both functions are equal to 0 for $x < 0$.

Memorize

The PDF of the exponential distribution with a parameter $\lambda > 0$ is

defined as:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

The corresponding CDF is

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

Notation for a random variable with such distribution:

$$X \sim \text{Exp}(\lambda).$$

The mean and variance of X are given by:

$$\mathbb{E}(X) = \frac{1}{\lambda},$$
$$\text{Var}(X) = \frac{1}{\lambda^2}.$$

Example 5.2. Suppose the time between arrivals at a bus stop follows an exponential distribution with a rate parameter $\lambda = 0.05$ arrivals per minute.

a) Calculate the probability that the next bus will arrive within the next 10 minutes.

b) Calculate the probability that you would need to wait at least 15 minutes until the next bus.

c) For how long on average you would need to wait for a bus?

Solution: a) Let $X \sim \text{Exp}(0.05)$ be the waiting time for the next bus. Then

$$\mathbb{P}(X \leq 10) = F_X(10) = 1 - e^{-0.05 \cdot 10} = 1 - e^{-0.5} \approx 0.3935.$$

b) We need to find

$$\mathbb{P}(X \geq 15) = 1 - \mathbb{P}(X < 15) = 1 - F_X(10) = e^{-0.05 \cdot 15} = e^{-0.75} \approx 0.4724.$$

c) Since

$$\mathbb{E}(X) = \frac{1}{0.05} = 20,$$

you would need to wait, on average, for 20 minutes.

5.2.3 Normal Distribution

Remember

- The **Normal Distribution** is also known as the **Gaussian Distribution**
- The shape of its PDF is symmetric and often called *bell-shaped*.
- The normal distribution is widely used in probability and statistics, especially, because of the central limit theorem which we will discuss later in this course. Its consequence is that the averages of large samples behave simialrly, i.e. "normally", regardless of the individual behaviour of the elements in these samples.

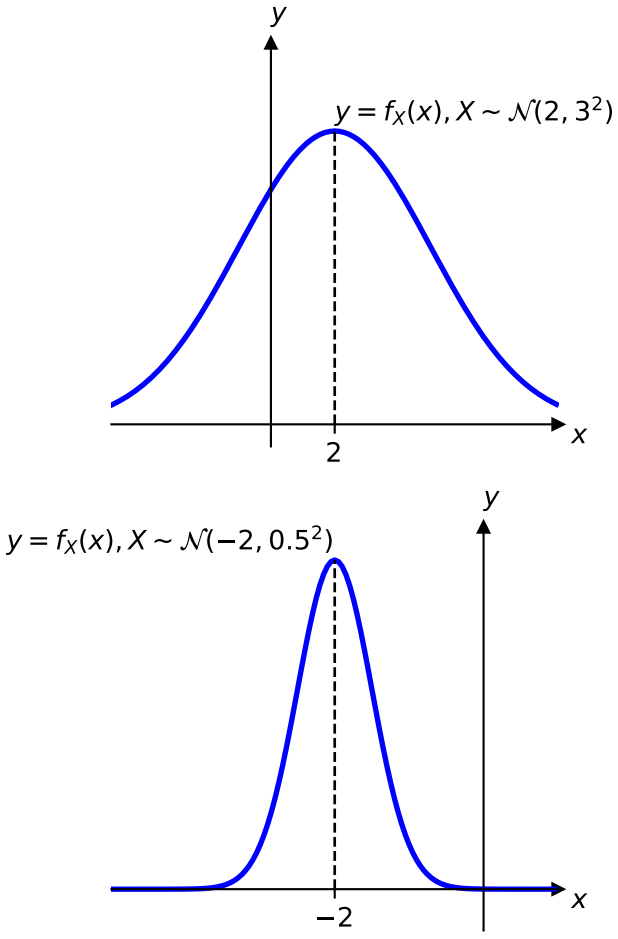


Figure 5.4: Graphs of PDF of $X \sim \mathcal{N}(2, 3^2)$ and $X \sim \mathcal{N}(-2, 0.5^2)$, respectively.

Memorize

The normal distribution with the mean $\mu \in \mathbb{R}$ and the standard deviation $\sigma > 0$ is the continuous probability distribution with the PDF given by

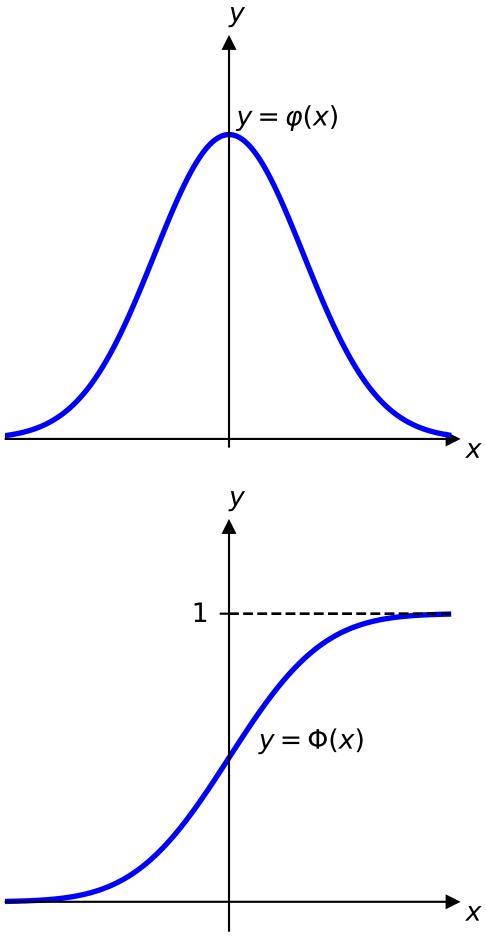
$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Notation for the random variable is

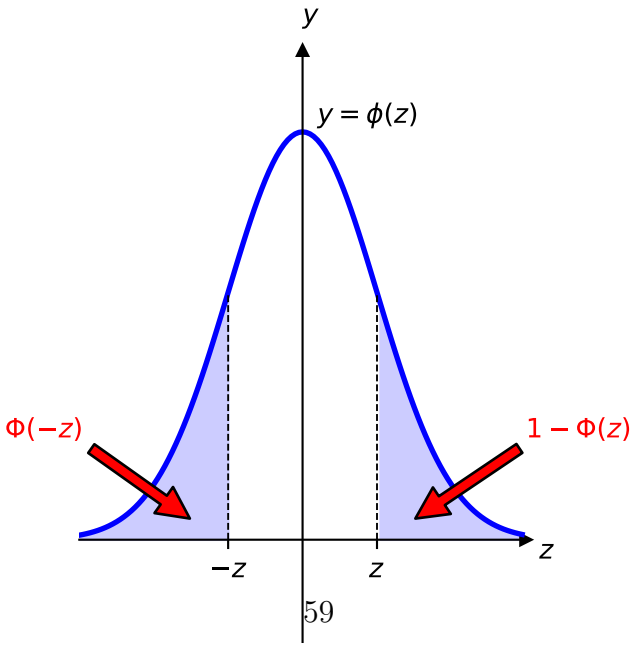
$$X \sim \mathcal{N}(\mu, \sigma^2).$$

The names for the parameters μ and σ are coming from the relations:

$$\mathbb{E}(X) = \mu,$$
$$\mathrm{Var}(X) = \sigma^2.$$



(a) Functions $\varphi(x)$ and $\Phi(x)$



(b) Illustration that $\Phi(-z) = 1 - \Phi(z)$

Figure 5.5: Graphs for the standard normal distribution

Memorize

The simplest case of a normal distribution is known as the **standard normal distribution** (or *unit normal distribution*), and it corresponds to $\mu = 0$, $\sigma = 1$. The PDF of $X \sim \mathcal{N}(0, 1)$ has special notation:

$$\varphi(x) := \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

Therefore, if $X \sim \mathcal{N}(\mu, \sigma^2)$ then

$$f_X(x) = \frac{1}{\sigma} \varphi\left(\frac{x-\mu}{\sigma}\right).$$

Similarly, for $X \sim \mathcal{N}(0, 1)$, the corresponding CDF is denoted

$$\Phi(x) := \int_{-\infty}^{x} \varphi(y)\,\mathrm{d}y = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}}\,\mathrm{d}y.$$

This function cannot be expressed in terms of elementary functions. To deal with it, one can use computer or statistical tables where its values are given for various values of x (it's called that the function Φ is *tabulated*).
It can be shown that if $X \sim \mathcal{N}(\mu, \sigma^2)$ then the corresponding CDF is

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

There is a standard notation here: $z = \frac{x-\mu}{\sigma}$, hence, we can rewrite

$$F_X(x) = \Phi(z), \qquad z = \frac{x-\mu}{\sigma}.$$

Note that

$$\Phi(-z) = 1 - \Phi(z).$$

Statistical tables usually provide values of the function $1 - \Phi(z)$ for $z \geq 0$. It gives immediately answer for $\Phi(-z) = 1 - \Phi(z)$ and for $\Phi(z) = 1 - (1 - \Phi(z))$.

$\frac{x-\mu}{\sigma}$	.00	.01	.02	.03	.04
0.0	.5000	.4960	.4920	.4880	.4840
0.1	.4602	.4562	.4522	.4483	.4443
0.2	.4207	.4168	.4129	.4090	.4052
0.3	.3821	.3783	.3745	.3707	.3669
0.4	.3446	.3409	.3372	.3336	.3300

For example, from the statistical table we can find that

$$1 - \Phi(0.23) = 0.4090.$$

Then

$$\Phi(-0.23) = 1 - \Phi(0.23) = 0.4090,$$
$$\Phi(0.23) = 1 - (1 - \Phi(0.23)) = 1 - 0.4090 = 0.5910.$$

Example 5.3. Calculate the probability that a randomly selected individual has a height between 162 cm and 169 cm, given that the population mean height is 165 cm and the standard deviation is 10 cm, and that the heights follow the normal distribution.

Solution: Let X be the random variable representing the height of an individual. It is given then that $X \sim \mathcal{N}(\mu, \sigma^2)$, where $\mu = 165$, $\sigma = 10$. We need to find

$$\mathbb{P}(162 \leq X \leq 173).$$

First step. We rewrite the required probabiligy in terms of the random variable

$$Z = \frac{X - \mu}{\sigma}, \qquad Z \sim \mathrm{N}(0, 1).$$

Namely, we have

$$\begin{aligned}
\mathbb{P}(162 \leq X \leq 169) &= \mathbb{P}(162 - 165 \leq X - 165 \leq 169 - 165) \\
&= \mathbb{P}(-3 \leq X - 165 \leq 4) \\
&= \mathbb{P}\left(-\frac{3}{10} \leq \frac{X - 165}{10} \leq \frac{4}{10}\right) \\
&= \mathbb{P}(-0.3 \leq Z \leq 0.4) \\
&= \Phi(0.4) - \Phi(-0.3).
\end{aligned}$$

From the statistical table (see above), we have that

$$1 - \Phi(0.4) = 0.3446, \qquad 1 - \Phi(0.3) = 0.3821.$$

Therefore,

$$\Phi(0.4) = 1 - 0.3446 = 0.6554, \qquad \Phi(-0.3) = 1 - \Phi(0.3) = 0.3821,$$

and hence,

$$\mathbb{P}(162 \leq X \leq 169) = 0.6554 - 0.3821 = 0.2733.$$

Surely, we can also use Python:

```
from scipy.stats import norm
norm.cdf(0.4) - norm.cdf(-0.3)
```

```
0.2733331637992768
```

6. Law of large numbers and the central limit theorem

6.1 Joint behaviour of random variables

We discussed with you discrete and continuous random variables. For a random variable X, you know now how to calculate some of its characteristics: expected value $\mathbb{E}(X)$ and variance $\mathrm{Var}(X)$. Now we consider how to characterise a pair of random variables.

Memorize

Let $X, Y : \Omega \to \mathbb{R}$ be two random variables. Their **joint cumulative distribution function** (**joint CDF**) is the function $F_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y).$$

Remark. In some cases, the joint CDF can be calculated manually from the description of the problem. However, in general, to calculate the joint CDF, we need to have additional information: joint probaility mass function in discrete case and joint probability density function in continuous case.

Memorize

Let $X : \Omega \to \{x_1, x_2, ...\}$ and $Y : \Omega \to \{y_1, y_2, ...\}$ be two *discrete random variable* with the joint CDF $F_{X,Y}$. Their **joint probability**

mass function (**joint PMF**) is the function

$$p_{X,Y}(x_i, y_j) = \mathbb{P}(X = x_i, Y = y_j)$$

(we can also say that $p_{X,Y}(x, y) = 0$ for all other x and y). Then

$$F_{X,Y}(x, y) = \sum_{x_i \le x} \sum_{y_j \le y} p_{X,Y}(x_i, y_j).$$

Memorize

Let $X, Y : \Omega \to \mathbb{R}$ be two *continuous random variable* with the joint CDF $F_{X,Y}$. Their **joint probability density function** (**joint PDF**) is the function $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ such that

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \left(\int_{-\infty}^{y} f_{X,Y}(u, v)\, \mathrm{d}v \right) \mathrm{d}u.$$

Note that

$$\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f_{X,Y}(u, v)\, \mathrm{d}v \right) \mathrm{d}u = 1.$$

Example 6.1. If joint PMF (for the discrete case) or joint PDF (for the continuous case) are not given explicitly, the joint CDF can be usually calculated only in very special cases, e.g. when one of variable is defined in terms of another one. For example, consider $X \sim U(0, 1)$ and $Y = X^2$, then $F_{X,Y}(x, y) = 0$ if $x < 0$ or $y < 0$, and for $x \ge 0, y \ge 0$, we have

$$\begin{aligned} F_{X,Y}(x, y) &= \mathbb{P}(X \le x, X^2 \le y) = \mathbb{P}(0 \le X \le x, X^2 \le y) \\ &= \mathbb{P}(0 \le X \le x, -\sqrt{y} \le X \le \sqrt{y}) \\ &= \mathbb{P}(0 \le X \le \min\{x, \sqrt{y}\}) \\ &= \begin{cases} 1, & \text{if } \min\{x, y\} \ge 1, \\ \min\{x, \sqrt{y}\}, & \text{if } \min\{x, y\} < 1. \end{cases} \end{aligned}$$

However, if we just have two random variables, e.g. $X \sim U(0, 1)$ and $Y =\sim (0, 1)$, then we can't *calculate* $F_{X,Y}$, unless we explicitly *define* the function $f_{X,Y}$.

Memorize

- In the discrete case: for any $g : \mathbb{R}^2 \to \mathbb{R}$,

$$\mathbb{E}(g(X,Y)) = \sum_i \sum_j g(x_i, y_j) p_{X,Y}(x_i, y_j),$$

in particular,

$$\mathbb{E}(XY) = \sum_i \sum_j x_i y_j p_{X,Y}(x_i, y_j).$$

- In the continuous case: for any $g : \mathbb{R}^2 \to \mathbb{R}$,

$$\mathbb{E}(g(X,Y)) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} g(x,y) \cdot f_{X,Y}(x,y)\, \mathrm{d}y \right) \mathrm{d}x,$$

in particular,

$$\mathbb{E}(XY) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x \cdot y \cdot f_{X,Y}(x,y)\, \mathrm{d}y \right) \mathrm{d}x.$$

Remember

For the given joint PDF $f_{X,Y}$ (for the continuous case), we can calculate PDFs of X and Y (so-called **marginal PDFs**):

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, \mathrm{d}y,$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, \mathrm{d}x.$$

Stress that, however, for given f_X and f_Y **one can't uniquely recover** $f_{X,Y}$.
Similarly, for the discrete case, we can define the marginal PMFs, e.g.

$$p_X(x_i) = \sum_j p_{X,Y}(x_i, y_j),$$
$$p_Y(y_j) = \sum_i p_{X,Y}(x_i, y_j).$$

Again, one can't uniquely recover $p_{X,Y}$ by the pair of p_X and p_Y.

Memorize

Recall that two random variables X and Y are *independent* if, for any $a, b \in \mathbb{R}$, the events $\{X \leq a\}$ and $\{Y \leq b\}$ are *independent*, i.e. if

$$\mathbb{P}(X \leq a, Y \leq b) = \mathbb{P}(X \leq a)\mathbb{P}(Y \leq b),$$

i.e. for all x and y,

$$F_{X,Y}(x, y) = F_X(x)F_Y(y).$$

We also have then that: in the discrete case,

$$p_{X,Y}(x_i, y_j) = p_X(x_i)p_Y(y_j),$$

and, in the continuous case,

$$f_{X,Y}(x, y) = f_X(x)f_Y(y).$$

Therefore, in both cases, we have that, for independent random variables,

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

Memorize

Let $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ be two random variables (discrete or continuous). **Covariance** $\operatorname{cov}(X, Y)$ describes the joint variability of these random variables, and it is defined by

$$\begin{aligned}\operatorname{cov}(X, Y) :&= \mathbb{E}\Big((X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))\Big) \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

Remember

For any $X, Y, V, W : \Omega \to \mathbb{R}$, $a, b, c, d \in \mathbb{R}$,

- $\operatorname{cov}(X, Y) = \operatorname{cov}(Y, X)$
- $\operatorname{cov}(X, X) = \operatorname{Var}(X) = \sigma^2(X)$
- $\operatorname{cov}(X, a) = 0$
- $\operatorname{cov}(aX, bY) = ab \operatorname{cov}(X, Y)$
- $\operatorname{cov}(X + a, Y + b) = \operatorname{cov}(X, Y)$
- $\operatorname{cov}(aX + bY, cV + dW) = ac \operatorname{cov}(X, V)$

$$+ ad\,\mathrm{cov}(X,W) + bc\,\mathrm{cov}(Y,V) + bd\,\mathrm{cov}(Y,W)$$

- $\mathrm{Var}(aX+bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{cov}(X,Y)$

Memorize

For any random variables $X, Y : \Omega \to \mathbb{R}$, we define their **correlation** as follows

$$\mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma(X)\cdot\sigma(Y)}.$$

It can be proved that

$$\big|\mathrm{corr}(X,Y)\big| \le 1,$$

i.e.

$$-1 \le \mathrm{corr}(X,Y) \le 1.$$

Memorize

Two random variables, X and Y, are called **uncorrelated** if their covariance is zero: $\mathrm{cov}(X,Y) = 0$ (and, hence, their correlation is also zero: $\mathrm{corr}(X,Y) = 0$).

Remember

For uncorrelated random variables X and Y and for any $a, b \in \mathbb{R}$,

$$\mathrm{Var}(aX+bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y).$$

Reminder

Recall, that for independent random variables X and Y, $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$, and hence $\mathrm{cov}(X,Y) = 0$. Therefore, **independent random variables are uncorrelated**. The opposite statement is wrong that is shown by the following example.

Example 6.2. Let $X \sim U(-1,1)$ and $Y = X^2$. Then $XY = X^3$, and hence

$$\mathrm{cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(X^3) - \mathbb{E}(X)\mathbb{E}(X^2).$$

We know that

$$\mathbb{E}(X) = \frac{(-1)+1}{2} = 0.$$

Next, since $f_X(x) = \frac{1}{2}$ for $x \in (-1, 1)$ and $f_X(x) = 0$ otherwise, we have

$$\mathbb{E}(X^3) = \int_{-\infty}^{\infty} x^3 f_X(x)\,\mathrm{d}x \;\; = \frac{1}{2}\int_{-1}^{1} x^3 dx = \frac{1}{2}\left[\frac{x^4}{4}\right]_{-1}^{1} = 0.$$

Therefore, $\mathrm{cov}(X, Y) = 0$ i.e. X and Y are *uncorrelated*. However, clearly, X and $Y = X^2$ are *not independent*.

6.2 Law of large numbers (LLN)

Remember

Let $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ be random variables. They are called **independent** if, for any $a_1, \ldots, a_n \in \mathbb{R}$ the events $\{X_1 \le a_1\}$, ..., $\{X_n \le a_n\}$ are independent. Or, equivalently, if their joint CDF

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) := \mathbb{P}(X_1 \le x_1, \ldots, X_n \le x_n)$$

is the product of the CDFs for each X_i:

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \ldots F_{X_n}(x_n).$$

Memorize

Random variable $X_1, X_2, \ldots, X_n, \ldots$ are called **independent and identically distributed** random variables (in brief, **i.i.d. r.v.**) if any finite group of them $X_1, \ldots, X_n$ are independent and they all have the same distribution: $F_{X_1} = F_{X_2} = \ldots = F_{X_n} = \ldots = F_X$, where X is their joint distribution; i.e. $X_1 \sim X$, $X_2 \sim X$, ...

Remember

Let $X_1, X_2, \ldots, X_n, \ldots$ be i.i.d. r.v. with $\mathbb{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2 < \infty$. Consider the sample average

$$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n}.$$

Then

$$\mathbb{E}(\bar{X}_n) = \mu, \qquad \mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Law of Large Numbers (LLN)

Let $X_1, X_2, \dots, X_n, \dots$ be i.i.d. r.v. with $\mathbb{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2 < \infty$. Then $\bar{X}_n \to \mu$ *stochastically* (or, it is also called *in probability*): namely, for each $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) = 0.$$

Remark. In other words, the bigger n you take, the smaller chances are for the event $\{|\bar{X}_n - \mu| > \varepsilon\}$. Equivalently, one can state that

$$\lim_{n\to\infty} \mathbb{P}(|\bar{X}_n - \mu| \leq \varepsilon) = 1,$$

i.e. the bigger n you take, the *larger* chances are for the event $|\bar{X}_n - \mu| \leq \varepsilon$ that is equivalent to $\mu - \varepsilon < \bar{X}_n < \mu + \varepsilon$. Thus, informally speaking, with n growing, there are good chances to find $\bar{X}_n$ around μ. We can choose ε arbitrary small, i.e. we can require that $\bar{X}_n$ must very close to μ, and the law of large numbers states that there is high probability (close to 1) to achieve this if we take n alrge enough.

Remember

Let A be a random event as a result of an experiment; let $\mathbb{P}(A) = p$. Consider the Bernoulli random variable X with $X = 1$ if A holds and $X = 0$ otherwise. Let $X_1, \dots, X_n, \dots$ be i.i.d. r.v. with $X_n \sim X$. Then

$$\mu = \mathbb{E}(X) = 1 \cdot p + 0 \cdot (1 - p) = p.$$

Next, the sample average $\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$ is the number of times when A took place when we repeated the experiment n times. (Note that $X_1 + \dots + X_n \sim Bin(n, p)$.) Therefore, $\bar{X}_n$ is the frequency of the event A took place among n trials. Then LLN states that

$$\frac{\text{number of trials when } A \text{ happened}}{\text{number } n \text{ of all trial}} \to \mathbb{P}(A)$$

in a proper sense (as $n \to \infty$). This corresponds to our "intuitive" understanding of the probability.

Example 6.3. Consider many rolls of a fair 6-sides dice. Let X_j be the score of the j-th roll, and $S_n = X_1 + \dots + X_n$ be the sum of the scores in the first

n rolls. All X_j are i.i.d. r.v. with

$$\mathbb{E}(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + ... + 6 \cdot \frac{1}{6} = \frac{7 \cdot 6}{2} \cdot \frac{1}{6} = \frac{7}{2}.$$

Therefore, by LLN, for any small $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - \frac{7}{2}\right| \leq \varepsilon\right) = 1,$$

or equivalently,

$$\lim_{n\to\infty} \mathbb{P}\left(\left(\frac{7}{2} - \varepsilon\right)n \leq S_n \leq \left(\frac{7}{2} + \varepsilon\right)n\right) = 1.$$

6.3 Central limit theorem (CLT)

As we could see, LLN states that, for i.i.d. r.v. $X_n \sim X$, $n \geq 1$,

$$\overline{X}_n = \frac{X_1 + ... + X_n}{n} \to \mathbb{E}(X)$$

stochastically (in probability) as $n \to \infty$. We have also shown that $\mathbb{E}(\overline{X}_n) = \mathbb{E}(X)$ for each n, i.e. we can reformulate LLN as follows:

$$\overline{X}_n - \mathbb{E}(\overline{X}_n) \to 0, \qquad n \to \infty.$$

Preparation

The **Central Limit Theorem** (CLT) shows *how fast* $\overline{X}_n$ converges to $\mathbb{E}(X)$. To formulate it, we recall that $\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2(X)}{n}$. Hence,

$$\sigma(\overline{X}_n) = \frac{\sigma(X)}{\sqrt{n}}.$$

We define, for $\mu := \mathbb{E}(X)$, $\sigma := \sigma(X)$,

$$Z_n := \frac{\overline{X}_n - \mathbb{E}(\overline{X}_n)}{\sigma(\overline{X}_n)} = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu).$$

Note that

$$\mathbb{E}(Z_n) = 0, \qquad \mathrm{Var}(Z_n) = 1.$$

Central Limit Theorem (CLT)

Let $X_1, \dots, X_n, \dots$ be i.i.d. r.v. with $X_n \sim X$, $\mu := \mathbb{E}(X)$, $\sigma^2 := \operatorname{Var} X < \infty$. Let Z_n be defined as above. Then

$$Z_n \to Z \sim \mathcal{N}(0,1), \quad n \to \infty,$$

where the convergence is *in distribution*; the latter means that

$$\lim_{n\to\infty} \mathbb{P}(Z_n \le z) = \Phi(z), \quad z \in \mathbb{R},$$

where $\Phi(z) = F_Z(z) = \mathbb{P}(Z \le z)$. As a corollary,

$$\lim_{n\to\infty} \mathbb{P}(a \le Z_n \le b) = \Phi(b) - \Phi(a), \quad a, b \in \mathbb{R},$$

and

$$\lim_{n\to\infty} \mathbb{P}(Z_n \ge c) = 1 - \Phi(c), \quad c \in \mathbb{R}.$$

Remark. The central limit theorem shows, in particular, that $\overline{X}_n$ fluctuates around its expected value $\mathbb{E}(\overline{X}_n) = \mu$ with the standard deviation $\sigma(\overline{X}_n) = \frac{\sigma}{\sqrt{n}}$ which is significantly less than the standard deviation σ for each of X_n around their expected value $\mathbb{E}(X_n) = \mu$. And this is tru regardless of the distribution of X_n. Consider this in an example.

Example 6.4. The average teacher's salary in New Jersey in 2023 is \$63178. Suppose that the salaries are distributed normallly with standard deviation \$7500. Hence, we have that $X \sim \mathcal{N}(63178, 7500^2)$.

Lets first find the probability that a randomly selected teacher makes less than \$60000 per year. We have

$$\begin{aligned}\mathbb{P}(X < 60000) &= \mathbb{P}\left(\frac{X - 63178}{7500} < \frac{60000 - 63178}{7500}\right) \\ &= \mathbb{P}(Z < -0.42) = \Phi(-0.42),\end{aligned}$$

where $Z = \frac{X-63178}{7500} \sim \mathcal{N}(0,1)$.

Using statistical tables (and the equality $\Phi(-0.42) = 1 - \Phi(0.42)$) or Python commands

```
from scipy.stats import norm
norm.cdf(-0.42)
```

```
0.3372427268482495
```

we conclude that

$$\mathbb{P}(X < 60000) \approx 0.337,$$

i.e. one out of three randomly picked teachers may have the salary less than \$60000.

Consider now a sample of 100 teacher salaries. The sample mean (the average salary) is then

$$\overline{X}_{100} = \frac{X_1 + ... + X_{100}}{100}$$

where all $X_j \sim X$ are i.i.d. r.v. We know that

$$\mathbb{E}(\overline{X}_{100}) = \mathbb{E}(X) = 63178$$

and

$$\sigma(\overline{X}_{100}) = \frac{\sigma(X)}{\sqrt{100}} = 750.$$

Therefore, the probability that the average salary of any sample of 100 teachers is less than \$60000 per year is

$$\begin{aligned}\mathbb{P}(\overline{X}_{100} < 60000) &= \mathbb{P}\left(\frac{\overline{X}_{100} - 63178}{750} < \frac{60000 - 63178}{750}\right) \\ &= \mathbb{P}(\overline{Z}_{100} < -4.2) \approx \Phi(-4.2)\end{aligned}$$

where the latter approximate equality is accroding to CLT. Since

```
norm.cdf(-4.2)
```

```
1.3345749015906314e-05
```

we have that

$$\mathbb{P}(\overline{X}_{100} < 60000) \approx 0.0000133,$$

i.e., informally speaking, this is very unlikely.

7. Hypothesis Testing: Z-tests and t-tests

This chapter follows chapter 17 of *A Basic Course in Statistics* by G.M.Clarke and D.Cooke. Another useful resource is *How to Use Statistics* by S.Lakin.

In many situations of uncertainty we have to make a choice between two possible alternatives, for example, given a coin we might ask whether it is fair. In this section we deal with the problem of finding a method for choosing between two possible outcomes.

Definition 7.1

A **statistical hypothesis** is an assertion concerning the probability distribution of one or more random variables.

For example, if p denotes the probability that a coin lands heads when flipped the hypotheses could be

- $p = \frac{1}{2}$
- $p \neq \frac{1}{2}$.

When we are faced with two hypotheses we call one of them the null hypothesis and denote it by H_0 and the other the alternative hypothesis and denote it by H_1. Therefore in our previous example, the null hypothesis would be $H_0 : p = \frac{1}{2}$ and the alternative hypothesis would be $H_1 : p \neq \frac{1}{2}$.

The next natural question to ask is how we choose between H_0 and H_1? We do this by obtaining a random sample from the distribution involved and

then choosing a statistic, called the **test statistic**, whose value can be used to choose between H_0 and H_1.

Therefore for the coin example we might decide **not** to reject

$$H_0 : p = \frac{1}{2}$$

if after 10 flips of the coin the test statistic (no. of heads in the sample) lies between 3 and 7, or to reject H_0 if the test statistic is at most 2 or at least 8.

Definition 7.2

The **critical region** of a statistical test is the set of values of the test statistic which lead us to the rejection of the null hypothesis H_0 (at most 2 or at least 8 in the previous example). The **acceptance region** of the test is the set of values of the test statistic which lead us to failing to reject H_0.

Remark 7.3

Some statisticians say that a hypothesis test can have one of two outcomes: you reject or accept H_0. Other statisticians do not like the phrase "accept" and prefer to to say either that you reject or fail to reject H_0.

We often specify a value for α, usually the largest value that we are prepared to tolerate and then look for a test with this value of α. The value of α is then called the **significance level** of the test. If $\alpha = 0.05$, we say that we are testing H_0 at the "5% level of significance" and, if the test rejects H_0, we say that the null hypothesis is rejected at the 5% level.

There are two types of significance tests: one-tailed; and two-tailed. We use a two-tailed test when H_1 is two sided (e.g. $H_1 : \mu \neq \mu_0$). We use a one-tailed test when H_1 is one-sided (e.g. $H_1 : \mu > \mu_0$ or $H_1 : \mu < \mu_0$).

An alternative approach to using critical regions is using **p-values**. Instead of specifying a critical region and deciding whether or not the test statistic lies within it, the probability of obtaining a value equal to, or more extreme than the test statistic is calculated. For a one-tailed test, this probability is called the p-value and it is then compared with the significance level probability α. For a two-tailed test, this probability is called the $\frac{p-\text{value}}{2}$ and it is then compared with the significance level probability $\frac{\alpha}{2}$. More on this later.

We cannot be sure of making the correct choice between H_0 and H_1. We can

make two types of incorrect decision. We can reject H_0 when it is actually true and we can accept H_0 when it is actually false.

Definition 7.4

A **type I error** occurs if the null hypothesis H_0 is rejected when it is true. A **Type II error** occurs if the null hypothesis H_0 is not rejected when it is false.

The probability of a type I error is usually denoted by α (and type II by β).

The following steps summarise how we tackle questions of this type:

1. Determine H_0, H_1 and the significance level;
2. Decide whether a one or two-tailed test is appropriate;
3. Calculate the test statistic assuming H_0 is true;
4. Compare the test statistic with the critical value(s) for the critical region or use the p-value approach;
5. Reject or do not reject H_0 as appropriate.

We now deal with the different cases.

7.1 The Mean of n Observations from $N(\mu, \sigma^2)$ (σ^2 Known)

A random sample of n observations will be collected and a decision will be made by looking at the whole sample.

$X \sim N(\mu, \sigma^2)$ then $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$. Therefore given a null hypothesis which states that a random sample $x_1, \ldots, x_n$ has been drawn from $N(\mu, \sigma^2)$ we calculate

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

and test whether it has come from $N(\mu, \frac{\sigma^2}{n})$. In this case our test statistic is

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

and we compare with the critical values or use the p-value approach in the usual way.

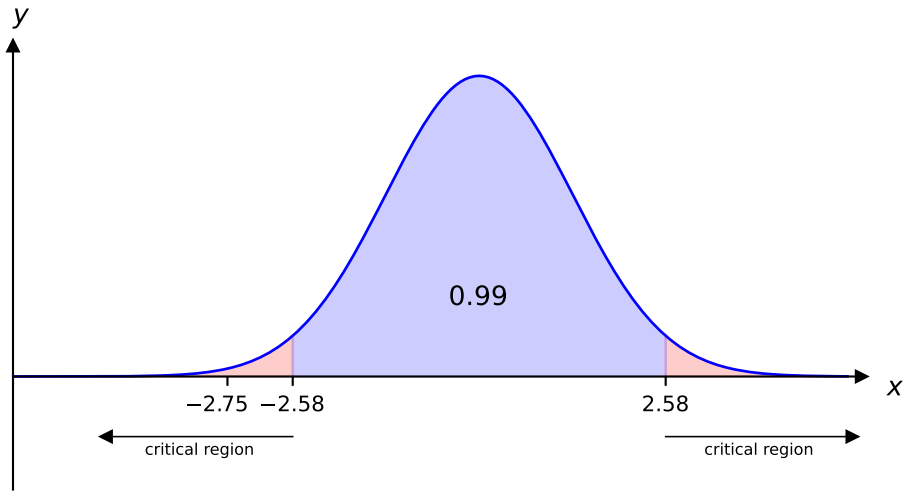
Example 7.5. A machine produces items having a nominal mass of $1kg$. The mass of a randomly selected item x follows the distribution $X \sim N(\mu, (0.02)^2)$. If $\mu \neq 1$ then the machinery should be corrected. The mean mass of a randomly sample of 25 items was found to be $0.989kg$. Test the null hypothesis that $H_0 : \mu = 1$ at the 1% significance level.

We have

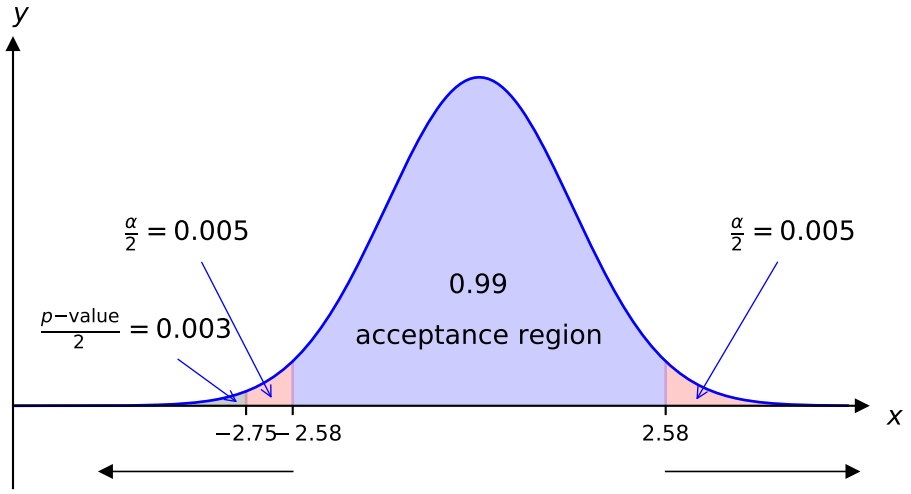
- $H_0 : \mu = 1$
- $H_1 : \mu \neq 1$
- $\alpha = 0.01$

Since $P(Z \leq 2.58) = 0.995$ the critical region is $|z| > 2.58$ and the test statistic is given by

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{0.989 - 1}{\frac{0.02}{5}} = -2.75.$$



Clearly -2.75 is in the critical region, therefore we reject H_0 at the 1% significance level and conclude that the machine settings should be corrected. Using the p-value approach we find:



In particular, $P(Z \geq -2.75) = 0.003 < 0.005(\frac{\alpha}{2})$ and therefore we reject H_0.

7.2 The Difference between 2 Means from Normal Distributions with Known Variances

If two samples are taken at random from normal distributions the first of size n_1 from $N(\mu_1, \sigma_1^2)$ and the second of size n_2 from $N(\mu_2, \sigma_2^2)$ the means of the sample may be calculated and compared. If the means are $\bar{x}_1$ and $\bar{x}_2$ respectively then the difference $(\bar{x}_1 - \bar{x}_2)$ has the distribution

$$N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}).$$

The theory then follows in the same way, with the test statistic being given by

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Example 7.6. A sample of size 25 is taken from $X \sim N(\mu_1, 66)$ and the mean $\bar{x}$ was found to be 116, then another sample of size 25 is taken from $Y \sim N(\mu_2, 66)$ and $\bar{y}$ was found to be 109. Test the following at the 5% significance level.

- $H_0 : \mu_1 - \mu_2 = 12$

- $H_1 : \mu_1 - \mu_2 \neq 12$

Since $\alpha = 0.05$ and we have a two-tailed test then the critical region is $|z| > 1.96$. The test statistic is

$$z = \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{7 - 12}{\sqrt{5.28}} = -2.17.$$

Since $-2.17 < -1.96$ then -2.17 is in the critical region and we reject the null hypothesis and conclude that the difference between the means is not 12 at the 5% significance level.

7.3 Large Sample Tests

The central limit theorem can be used to do significance tests for non-normal distributions when the sample sizes are large enough (30 or more). When a normal approximation can be used, its mean and variance will be μ and $\frac{\sigma^2}{n}$ respectively. In this way we can test hypotheses about the means of distributions which are not themselves normal, provided a large sample of observations is available. We will the same methods to test hypotheses as before except that the test statistic will be only approximately $N(0, 1)$.

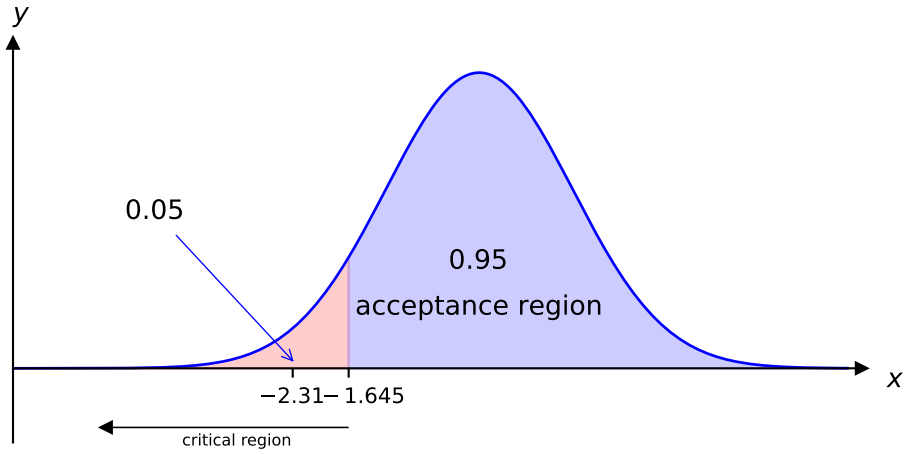
Example 7.7. The number of strokes a golfer takes to complete a round of golf has mean 84.1 and standard deviation 2.6. After lessons her mean is 83.1 in 36 subsequent rounds. At the 5% significance level test the null hypothesis that her standard of play is unaltered against the alternative hypothesis that it has improved, i.e.

- $H_0 : \mu = 84.1$
- $H_1 : \mu < 84.1$ (one-tailed)
- $\alpha = 0.05$ therefore the critical region is $z < -1.645$

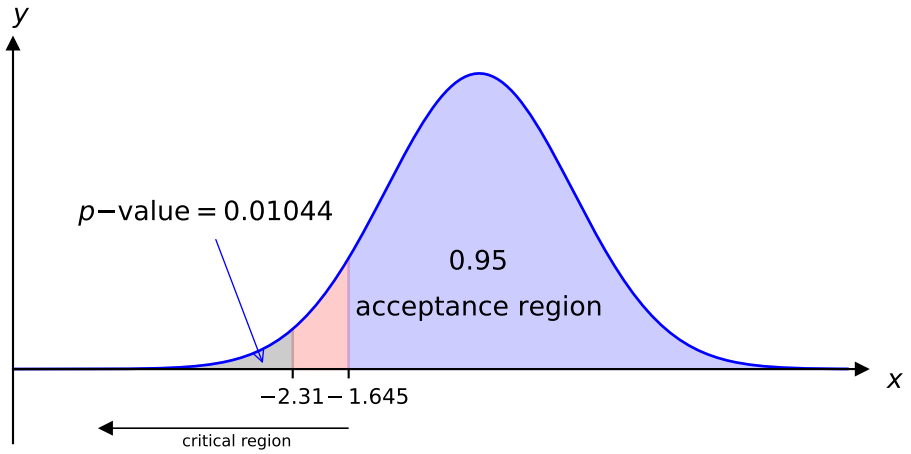
We approximate the distribution of strokes by $N(\mu, \frac{2.6^2}{36})$ and the test statistic is given by

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{83.1 - 84.1}{\frac{2.6}{6}} = -2.31.$$

This lies in the critical region therefore we reject the null hypothesis and conclude that her game seems to have improved.



Alternatively, using the p-value approach we find $P(Z \leq -2.31) = 0.01044 < 0.05$ and therefore we reject the null hypothesis.



Definition 7.8

If $x_1, \ldots, x_n$ is a family random variables of size $n \geq 2$ from the distribution $X \sim N(\mu, \sigma^2)$, the random variable

$$T = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

is said to have a t **distribution** with $(n-1)$ degrees of freedom.

As mentioned in the definition above, for the t distribution we require the degrees of freedom, this is n minus the number of samples ($(n-1)$ above due

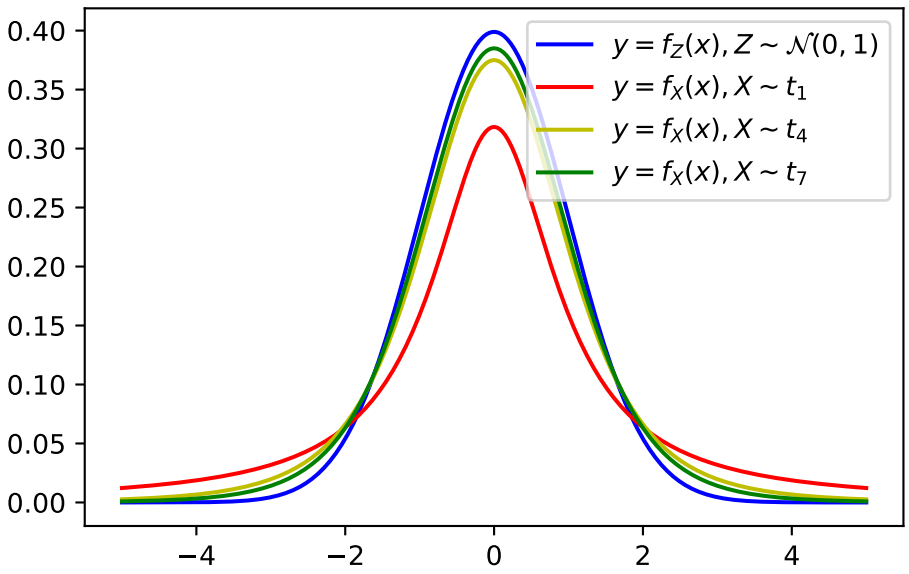
to the single sample).

The degrees of freedom come from the number of values that are free to vary. Let us suppose we have 4 numbers (a, b, c, d) and we know that the mean of these is 5. This means that

$$a + b + c + d = 20(4 \times 5).$$

Note that once we know 3 of the numbers above, then we can calculate the fourth and therefore only 3 of them are "free". In this case we would have $3 = 4 - 1 = n - 1$ degrees of freedom.

Note that as $n \to \infty$, the t distribution tends to the normal distribution. The t distribution has heavier tails than the normal distribution meaning that it is more likely to have values that fall further away from the mean.



t-tests

In the real world we often only have a random sample of data values with limited information about the underlying probability distribution. (Note that in the cases above the variance of the distribution is **known**.) The next natural question to ask is whether we can still perform hypothesis tests in these scenarios? Fortunately we can and we make use of the t distribution. t-tests may be performed on continuous data, possibly within an interval, for example, exam results as a percentage. We may also use t-tests on data which

have 7 or more ordered categories. An example of such a data set could be the outcome of questions which have answers on the following 7 point Likert scale:

- Strongly Agree
- Agree
- Agree Somewhat
- Undecided
- Disagree Somewhat
- Disagree
- Strongly Disagree

7.4 t-test: Comparing a Sample Mean

Suppose that $X \sim N(\mu, \sigma^2)$ where σ^2 is **unknown** and we wish to test the null hypothesis $H_0 : \mu = \mu_0$ against some alternative hypothesis. In this case we take a sample and we estimate σ^2 by the sample variance s^2, but the error in doing so, in particular when n is small, cannot be neglected and therefore we must use the test statistic

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

which has a t distribution with $n - 1$ degrees of freedom. The rest of the theory is similar to what we have seen with the normal distribution only that we use the t distribution to calculate the critical region.

Example 7.9. Yarn breaking strength follows a normal distribution with mean of $21N$. It is claimed that if the yarn is treated with a chemical then the mean breaking strength increases. A random sample of 9 lengths are taken, the value of the sample mean and sample standard deviation are 22.75 and 2.109 respectively. Test the following:

- $H_0 : \mu = 21$
- $H_1 : \mu > 21$ (one-tailed)
- $\alpha = 0.05$

The critical region is obtained by the t distribution:

$$t > t_{0.05,8} \quad \Longrightarrow \quad t > 1.860.$$

The test statistic is given by

$$t = \frac{22.75 - 21}{\frac{2.109}{\sqrt{9}}} = 2.5.$$

This value is clearly in the critical region therefore we reject the null hypothesis and accept the claim at the 5% significance level. Note that p-values can also be used to reject the null hypothesis or not - this will be seen in the lab class.

7.5 Paired t-test

Many statistical applications use paired data samples to draw conclusions about the difference between two population means. Data pairs occur very naturally in "before" and "after" situations, where the **same** object or item is measured before and after a treatment. Such data pairs are very common in science and business. Clearly in this situation the sample sizes will be equal. Assume we have n pairs and let X_1 and X_2 be the random variables that denote the observations made on the n pairs (the "before" and "after") with means μ_1 and μ_2 respectively. The idea is to consider the difference $D = X_1 - X_2$, assumed to be a normally distributed random variable, with mean $\mu_1 - \mu_2$ and the null hypothesis may be that μ_1 and μ_2 differ by a stated amount, say μ_0 (μ_0 is often 0, i.e. the means do not differ). The test statistic we use in this scenario is

$$t = \frac{\bar{D} - \mu_0}{\frac{s_D}{\sqrt{n}}}$$

where s_D is the sample standard deviation of D and t has a t distribution with $n - 1$ degrees of freedom.

Example 7.10. Ten joints of meat are cut in half; one half is frozen and wrapped by process A and the other is frozen and wrapped by a new process B. The halves are placed in ten freezers with halves of the same joint being put in the same freezer. The number of days to spoilage are found to be:

Joint number	1	2	3	4	5	6	7	8	9	10
Process A	63	109	82	156	161	155	47	141	92	149
Process B	129	105	76	207	253	146	62	160	90	177

Assuming the differences between these number-pairs to be normally distributed, test:

- $H_0 : \mu_D = \mu_B - \mu_A = 0$
- $H_0 : \mu_D = \mu_B - \mu_A \neq 0$ (two-tailed)
- $\alpha = 0.05$

The sample mean and variances of the two processes are given by:

$$\bar{x}_A = 115.5, \quad \bar{x}_B = 140.5, \quad s_A^2 = 1800.94, \quad s_B^2 = 3676.28.$$

We first need to calculate the difference of the days to spoilage D

Pair number	1	2	3	4	5	6	7	8	9	10	Total	Mean
D	66	−4	−6	51	92	−9	15	19	−2	28	250	25
$D - \overline{D}$	41	−29	−31	26	67	−34	−10	−6	−27	3	0	
$(D - \overline{D})^2$	1681	841	961	676	4489	1156	100	36	729	9	10678	

Then

$$s_D^2 = \frac{1}{9} \sum_{\text{all pairs}} (D - \overline{D})^2 = \frac{10678}{9} = 1186.44.$$

The test statistic is then given by

$$t = \frac{25 - 0}{\sqrt{\frac{1186.44}{10}}} = 2.3.$$

The critical region is given by

$$|t| > t_{0.025,9} \quad \Longrightarrow \quad |t| > 2.26.$$

Clearly our test statistic is in the critical region and we therefore reject H_0 and conclude that there is evidence that there is a difference in the effectiveness of the two processes.

7.6 Unpaired t-test

It is quite common to have data from two independent samples, for example not trying both drugs on every person in the sample, but trying one drug on some people in the sample and trying another drug on the rest. This would be an example of a situation where we might use an unpaired t-test. Assume that we have two samples chosen at random from normal distributions, the first of size n_1 from $X_1 \sim N(\mu_1, \sigma^2)$ and the second of size n_2 from $X_2 \sim N(\mu_2, \sigma^2)$. Note that both distributions have the **same** variance (albeit unknown) - this is important for this test. We estimate σ^2 from the samples. We consider the difference between the sample means $\overline{x}_1 - \overline{x}_2$ with the null hypothesis that $\mu_1 - \mu_2 = \mu_0$. The test statistic depends on the sample variances and this depends on whether the sample sizes are equal or not, i.e. $n_1 = n_2$ or $n_1 \neq n_2$.

If $n_1 = n_2$ then $s^2 = \frac{s_1^2 + s_2^2}{2}$. If $n_1 \neq n_2$ then

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

These are often called the **pooled estimates** of the variance σ^2. Clearly, if the sample sizes are not equal then we must give greater weight to the larger sample; the appropriate weights are the degrees of freedom corresponding to each estimate of the variance. The test statistic is given by

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

with $n_1 + n_2 - 2$ degrees of freedom. (This comes from $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$.)

Example 7.11. A trial takes place in which eight people are given only water, whereas another group of eight people are given a new energy drink. They then have to take part in an endurance task. The results of the trial are given in the following table.

	Mean	Standard deviation
Water (x_1)	12.2	2.4
Energy drink (x_2)	13.1	3.1

Note that these samples are independent; water and energy drinks are given to two different groups of people. Assuming the relevant assumptions hold,

use an unpaired t-test to decide whether people who have taken the energy drink perform better at the 5% significance level, i.e.

- $H_0 : \mu_1 - \mu_2 = 0$
- $H_1 : \mu_1 - \mu_2 < 0$ or $\mu_1 < \mu_2$ (Mean of "water" less than mean of "energy drink")

Since both samples are of the same size, the pooled estimate of the standard deviation is given by

$$s = \sqrt{\frac{2.4^2 + 3.1^2}{2}} = 2.7722,$$

and the test statistic is given by

$$t = \frac{(12.2 - 13.1) - 0}{2.7722\sqrt{\frac{2}{8}}} = -0.6493,$$

with $n_1 + n_2 - 2 = 14$ degrees of freedom. The critical region is given by

$$t < -t_{0.05,14} \quad \Longrightarrow \quad t < -1.761.$$

Since $-0.6493 > -1.761$ we do not reject the null hypothesis and conclude there is no evidence that the energy drink makes people perform better.

8. Maximum likelihood estimation

Memorize

Let X be a discrete random variable whose distribution depends on a parameter $\theta \in \mathbb{R}$. Suppose that we observe the data $x_1, \dots, x_n$ which is the output of this random variable X in course of n independent trials. In other words, we can say that we observe that i.i.d.r.v. $X_1, \dots, X_n$ with $X_i \sim X$, $1 \leq i \leq n$, take certain values: $X_1 = x_1, \dots, X_n = x_n$. The **likelihood**, or **likelihood function**, is the function $\mathcal{L}(\theta) = \mathcal{L}(\theta \mid x_1, \dots, x_n)$ of the unknown parameter θ (given the observed data $x_1, \dots, x_n$) which is equal to the probability to observe this data (given the value of the parameter θ):

$$\begin{aligned} \mathcal{L}(\theta) &= \mathcal{L}(\theta \mid x_1, \dots, x_n) \\ &:= \mathbb{P}(X_1 = x_1, \dots, X_n = x_n \mid \theta) \\ &= \mathbb{P}(X_1 = x_1 \mid \theta) \dots \mathbb{P}(X_n = x_n \mid \theta). \end{aligned}$$

Memorize

The maximum likelihood estimator θ_* of the parameter θ is the argument of the maximum of the likelihppd function:

$$\theta_* = \operatorname*{argmax}_{\theta} \mathcal{L}(\theta),$$

that means that

$$\mathcal{L}(\theta_*) = \max_\theta \mathcal{L}(\theta).$$

Remember

The standard approach to find θ_* is to consider the **log-likelihood* function

$$\begin{aligned} L(\theta) := L(\theta \mid x_1, \dots, x_n) &= \ln \mathcal{L}(\theta \mid x_1, \dots, x_n) \\ &= \ln \mathbb{P}(X_1 = x_1 \mid \theta) + \dots + \ln \mathbb{P}(X_n = x_n \mid \theta). \end{aligned}$$

Then θ_* is the point of maximum for both $\mathcal{L}$ and L:

$$\theta_* = \operatorname*{argmax}_\theta L(\theta) = \operatorname*{argmax}_\theta \mathcal{L}(\theta).$$

Remark. The reason for this is the fact that the logarigthm $y = \ln x$ is an increasing function, and then

$$\mathcal{L}(\theta) \le \mathcal{L}(\theta_*) \iff L(\theta) \le L(\theta_*).$$

Reminder

To check that θ_* is the point of maximum of $L(\theta)$, it is enough to check that

$$L'(\theta_*) = 0 \quad \text{and} \quad L''(\theta_*) < 0.$$

Example 8.1. Let $X : \Omega \to \{0, 1\}$ be a Bernouilli random variable with $\mathbb{P}(X = 1) = \theta$ and $\mathbb{P}(X = 0) = 1-\theta$, where $\theta \in [0, 1]$ is a parameter. Suppose that we are given a sample of the length n of values of X which contain k ones and $n - k$ zeros (the sample has a particular order, e.g. 010010111001 ...). Then the probability to get this particular sample, for any $\theta \in [0, 1]$, is $\theta^k(1 - \theta)^{n-k}$, i.e. the likelihood function for the given data is

$$\mathcal{L}(\theta) = \theta^k(1 - \theta)^{n-k}.$$

Hence, the log-likehood fgunction for the given data is

$$\begin{aligned} L(\theta) = \ln \mathcal{L}(\theta) &= \ln(\theta^k(1 - \theta)^{n-k}) \\ &= \ln \theta^k + \ln(1 - \theta)^{n-k} \\ &= k \ln \theta + (n - k) \ln(1 - \theta). \end{aligned}$$

Then

$$\begin{aligned} L'(\theta) &= \big(k\ln\theta + (n-k)\ln(1-\theta)\big)' \\ &= \frac{k}{\theta} - \frac{n-k}{1-\theta} \\ &= \frac{k(1-\theta)-(n-k)\theta}{\theta(1-\theta)} \\ &= \frac{k-n\theta}{\theta(1-\theta)}. \end{aligned}$$

Therefore, $L'(\theta) = 0$ iff $k - n\theta = 0$, i.e.

$$\theta = \frac{k}{n}.$$

Moreover,

$$\begin{aligned} L''(\theta) = (L'(\theta))' &= \left(\frac{k}{\theta} - \frac{n-k}{1-\theta}\right)' \\ &= -\frac{k}{\theta^2} - \frac{n-k}{(1-\theta)^2} < 0 \end{aligned}$$

for all $\theta \in [0,1]$, in particular, for $\theta_* = \dfrac{k}{n} \in [0,1]$ (as $0 \le k \le n$). Therefore, $\theta_* = \dfrac{k}{n}$ is the point of maximum of $L(\theta)$, and hence, it is the maximum likelihood estimator for the parameter θ.

Remark. Note that $S_n := X_1 + ... + X_n \sim Bin(n,\theta)$ is the binomial random variable, and k ones in n Bernoulli trials means $S_n = k$. Then the sample $\overline{X_n} = \frac{1}{n}(X_1 + ... + X_n) = S_n/n$ takes the value $\frac{k}{n}$. We have that

$$\mathbb{E}(X) = 1 \cdot \theta + 0 \cdot (1-\theta) = \theta,$$

and the law of large numbers says that (in certain sense)

$$\overline{X}_n \to \theta, \quad n \to \infty.$$

In other words, the maximum likelihood estimator converges to the theoretical value is the size of the sample converges to infinity.

Example 8.2. Let $X \sim Po(\lambda)$, i.e.

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}, \quad k \ge 0.$$

Suppose we have a sample of n values of X: $k_1, \dots, k_n$. Then

$$\begin{aligned}\mathcal{L}(\lambda) &= \mathbb{P}(X = k_1 \mid \lambda) \dots \mathbb{P}(X = k_n \mid \lambda) \\ &= \frac{\lambda^{k_1}}{k_1!} e^{-\lambda} \cdot \dots \cdot \frac{\lambda^{k_n}}{k_n!} e^{-\lambda} \\ &= \underbrace{\frac{1}{k_1! \dots k_n!}}_{=:c>0} \lambda^{k_1+\dots+k_n} e^{-\lambda n},\end{aligned}$$

and therefore,

$$\begin{aligned}L(\lambda) &= \ln \mathcal{L}(\lambda) \\ &= \ln c + (k_1 + \dots + k_n) \ln \lambda - \lambda n.\end{aligned}$$

Then

$$L'(\lambda) = \frac{k_1 + \dots + k_n}{\lambda} - n,$$

and hence, $L'(\lambda) = 0$ iff

$$\lambda = \frac{k_1 + \dots + k_n}{n}.$$

Since

$$L''(\lambda) = (L'(\lambda))' = -\frac{k_1 + \dots + k_n}{\lambda^2} < 0,$$

the found value $\lambda_* = \frac{k_1+\dots+k_n}{n}$ is the point of maximum of L, hence, it is the maximum likelihood estimator for the parameter λ.

9. Time series

Time series is an infinite sequence of random numbers parametrized (indexed) by (discrete) time:

$$X_1, X_2, X_3, \dots, X_n, \dots \in \mathbb{R}.$$

In real data, the values $X_1, X_2, \dots$ are not independent.

Memorize

A time series $\{X_n\}$ is called **(weakly) stationary** if the following conditions hold:

1) $\mathbb{E}(X_n)$ does not depend on n (i.e. is a constant in n)
2) $\mathbb{E}(X_n^2) < \infty$ for all n
3) $\operatorname{cov}(X_n, X_{n+m}) = \mathbb{E}(X_n X_{n+m}) - \mathbb{E}(X_n) \cdot \mathbb{E}(X_{n+m})$ **does not** depend on n, it depends on m only.

Remember

In other words, for stationary time series, $\operatorname{cov}(X_n, X_k)$ depend only on the time-lag $k - n$.

Example 9.1. Let Y and Z be two *uncorrelated* identically distributed random variables with zero mean and variance σ^2; i.e. $\operatorname{cov}(Y, Z) = 0$, $\mathbb{E}(Y) = \mathbb{E}(Z) = 0$, $\operatorname{Var}(Y) = \operatorname{Var}(Z) = \sigma^2$.

Let $\lambda \in [0, 2\pi]$ be a fixed number, and define

$$X_n = Y\cos(\lambda n) + Z\sin(\lambda n).$$

Then

$$\mathbb{E}(X_n) = \cos(\lambda n)\mathbb{E}(Y) + \sin(\lambda n)\mathbb{E}(Z) = 0,$$

hence,

$$\sigma^2 = \mathrm{Var}(Y) = \mathbb{E}(Y^2) - \big(\mathbb{E}(Y)\big)^2 = \mathbb{E}(Y^2),$$

and

$$0 = \mathrm{cov}(Y, Z) = \mathbb{E}(YZ) - \mathbb{E}(Y)\mathbb{E}(Z) = \mathbb{E}(YZ).$$

Then

$$\mathbb{E}(X_n^2) = \cos^2(\lambda n)\mathbb{E}(Y^2) + \sin^2(\lambda n)\mathbb{E}(Z^2) + 2\cos(\lambda n)\sin(\lambda n)\mathbb{E}(YZ) = \sigma^2 < \infty$$

and

$$\begin{aligned}
\mathrm{cov}(X_n, X_{n+m}) &= \mathbb{E}(X_n X_{n+m}) \\
&= \cos(\lambda n)\cos(\lambda(n+m))\mathbb{E}(Y^2) + \sin(\lambda n)\sin(\lambda(n+m))\mathbb{E}(Z^2) \\
&\quad + 2\cos(\lambda n)\sin(\lambda(n+m))\mathbb{E}(YZ) \\
&= \sigma^2 \cos(\lambda(n+m) - \lambda n) \\
&= \sigma^2 \cos(\lambda m),
\end{aligned}$$

hence, $\{X_n\}$ is stationary.

Remember

If $\{X_n\}$ is stationary, then

$$\mathrm{Var}(X_n) = \mathrm{cov}(X_n, X_n)$$

does not depend on n.

Memorize

A time series $\{Z_n\}$ is called a **white noise** if
1) $\mathbb{E}(Z_n) = 0$
2) for some $\sigma > 0$,

$$\mathrm{cov}(Z_n, Z_{n+m}) = \begin{cases} \sigma^2, & \text{if } m = 0, \\ 0, & \text{if } m \neq 0. \end{cases}$$

I.e. a white noise has zero mean and uncorrelated values.

Remember

The standard example of white noise is a colelction of i.i.d.r.v. $Z_n \sim \mathcal{N}(0, \sigma^2)$.

9.1 Autoregressive model $AR(1)$

We consider a time series $\{X_n\}$ which satisfies

$$X_n = \mu + \alpha(X_{n-1} - \mu) + Z_n,$$

whre $\mu \in \mathbb{R}$ is a constant, $\{Z_n\}$ is a white noise, and $\alpha \in \mathbb{R}$ is a parameter. **We will always assume that $\{Z_n\}$ is independent from $\{X_n\}$.** We denote $Y_n = X_n - \mu$, then

$$Y_n = \alpha Y_{n-1} + Z_n.$$

We are going to find conditions to have X_n stationary. In particular, one needs that $\mathbb{E}(X_n)$ and $\mathrm{Var}(X_n)$ are constants.

Since $\mathbb{E}(Y_n) = \mathbb{E}(X_n - \mu) = \mathbb{E}(X_n) - \mu$ and $\mathrm{Var}(Y_n) = \mathrm{Var}(X_n - \mu) = \mathrm{Var}(X_n)$, we must then have both $\mathbb{E}(Y_n) =: k$ and $\mathrm{Var}(Y_n) =: v$ constants. We can write

$$\mathbb{E}(Y_n) = \mathbb{E}(\alpha Y_{n-1} + Z_n) = \alpha\mathbb{E}(Y_{n-1}) + \mathbb{E}(Z_n),$$

i.e. $k = \alpha k + 0$, and hence, either $\alpha = 1$ or $k = 0$. Next, since Z_n is independent from $\{X_n\}$ (and hence, from $\{Y_n\}$), we get

$$\mathrm{Var}(Y_n) = \mathrm{Var}(\alpha Y_{n-1} + Z_n) = \alpha^2 \mathrm{Var}(Y_{n-1}) + \mathrm{Var}(Z_n),$$

i.e.

$$v(1 - \alpha^2) = \sigma^2.$$

If $1 - \alpha^2 = 0$, i.e. $\alpha = \pm 1$, then $\sigma = 0$, that is impossible. Hence $\alpha \neq \pm 1$ (and thus, $\mathbb{E}(X_n) = 0$ for all n). Moreover, for $\alpha \neq \pm 1$, we have

$$\mathrm{Var}(X_n) = v = \frac{\sigma^2}{1 - \alpha^2}.$$

Since $\mathrm{Var}(X_n) \geq 0$, we require $1 - \alpha^2 > 0$, i.e.

$$|\alpha| < 1 \iff -1 < \alpha < 1.$$

Memorize

It possible to prove that, indeed, the condition $|\alpha| < 1$ is necessary and sufficient for the stationarity of the time series $\{X_n\}$ given by $X_n = \alpha X_{n-1} + Z_n$. For the next classes of time series, however, it is more useful to rewrite this condition in term of the **characteristic equation**: for $AR(1)$

$$X_n = \alpha X_{n-1} + Z_n$$

we consider the equation

$$1 - \alpha\lambda = 0.$$

The time series is staionary if and only if $|\lambda|>1$ (note that here $\lambda = \frac{1}{\alpha}$).

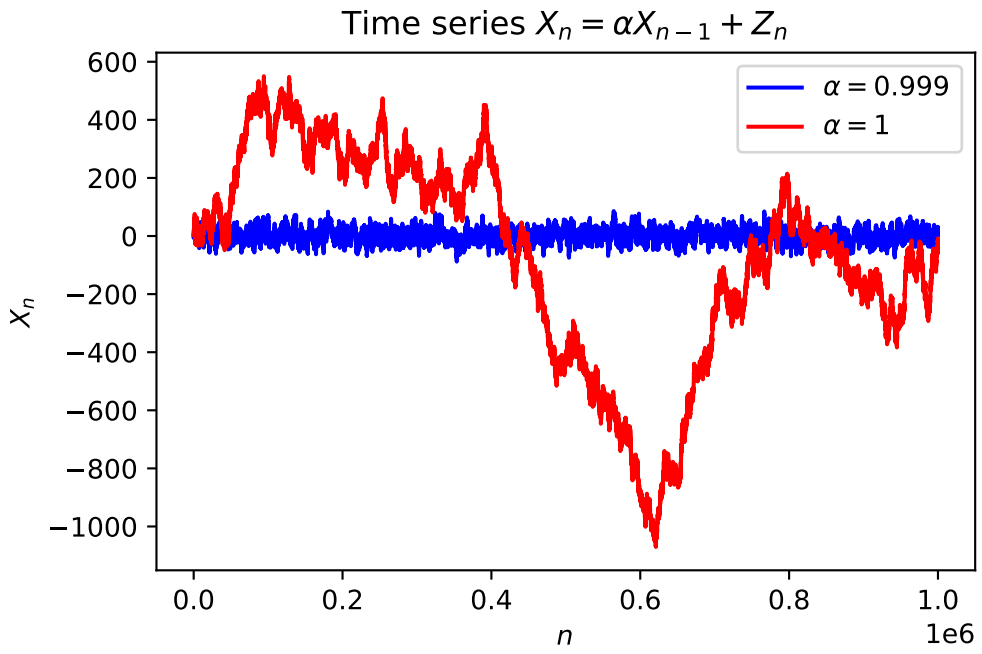


Figure 9.1: Qualitative difference between stationary (blue colour) and non-stationary (red colour) time series behaviour

9.2 Autoregressive model $AR(2)$

Consider the model

$$X_n = \mu + \alpha_1(X_{n-1} - \mu) + \alpha_2(X_{n-2} - \mu) + Z_n \tag{9.1}$$

Memorize

The **characteristic equation** of (9.1) is

$$1 - \alpha_1\lambda - \alpha_2\lambda^2 = 0.$$

This equation has two roots (they are, possibly, complex numbers): λ_1, λ_2. The time series (9.1) is *stationary* if and only if

$$|\lambda_1| > 1 \quad **\text{and}** \quad |\lambda_2| > 1.$$

Reminder

- A quadratic equation $ax^2 + bx + c = 0$ with $a \neq 0$ has two roots

$$x_1 = \frac{-b - \sqrt{D}}{2a}, \qquad x_2 = \frac{-b + \sqrt{D}}{2a},$$

  where the *discriminant* D is given by

$$D = b^2 - 4ac.$$

  If $D \leq 0$ the roots are real numbers (and they are equal if $D = 0$. If $D < 0$ the roots are complex numbers: $x_{1,2} = p \pm qi$, where $i^2 = -1$.
- Two complex numbers $p + qi$ and $p - qi$ has the same absolute value:

$$|p \pm qi| = \sqrt{p^2 + q^2}.$$

Example 9.2. Consider the time series

$$X_n = \frac{1}{12}X_{n-1} + \frac{1}{2}X_{n-2} + Z_n,$$

where $\{Z_n\}$ is a white noise. Is $\{X_n\}$ stationary?

Solution: Consoder the characteristic equation

$$1 = \frac{1}{12}\lambda + \frac{1}{2}\lambda^2,$$
$$\lambda^2 + \frac{1}{6}\lambda - 2 = 0,$$
$$D = \left(\frac{1}{6}\right)^2 - 4 \cdot (-2) = \frac{1}{36} + 8 = \frac{289}{36}, \qquad \sqrt{D} = \frac{17}{6},$$
$$\lambda_1 = \frac{-\frac{1}{6} - \frac{17}{6}}{2} = -\frac{18}{12} = -\frac{3}{2},$$
$$\lambda_2 = \frac{-\frac{1}{6} + \frac{17}{6}}{2} = \frac{16}{12} = \frac{4}{3}.$$

Since $|\lambda_1| = \frac{3}{2} > 1$ and $|\lambda_2| = \frac{4}{3} > 1$, the time series $\{X_n\}$ is stationary.

Example 9.3. Consider the time series

$$X_n = \frac{1}{3}X_{n-1} + \frac{2}{3}X_{n-2} + Z_n,$$

where $\{Z_n\}$ is a white noise. Is $\{X_n\}$ stationary?

Solution: Consider the characteristic equation

$$1 = \frac{1}{3}\lambda + \frac{2}{3}\lambda^2, \qquad 2\lambda^2 + \lambda - 3 = 0,$$
$$D = 1^2 - 4 \cdot 2 \cdot (-3) = 25,$$
$$\lambda_1 = \frac{-1-5}{2 \cdot 2} = -\frac{6}{4} = -\frac{3}{2},$$
$$\lambda_2 = \frac{-1+5}{2 \cdot 2} = \frac{4}{4} = 1.$$

Here $|\lambda_1| = \frac{3}{2} > 1$, however, $|\lambda_2| = 1$; hence, the time series $\{X_n\}$ is non-stationary.

9.3 Autoregressive model $AR(p)$

We consider a generalisation of the previous models:

$$X_n = \mu + \alpha_1(X_{n-1} - \mu) + ... + \alpha_p(X_{n-p} - \mu) + Z_n \tag{9.2}$$

Remember

The **characteristic equation** of (9.2) is

$$1 - \alpha_1 \lambda - ... - \alpha_p \lambda^p = 0.$$

This equation has p roots (if $\alpha_p \neq 0$): $\lambda_1, ..., \lambda_p$ (possibly, complex). The time series (9.2) is *stationary* iff

$$|\lambda_1| > 1, \quad ..., \quad |\lambda_p| > 1.$$

Remark. The roots of the characteristic equation may be found numerically, using e.g. Python.

9.4 $ARMA(p, q)$-model

Here "AR" stands for "autoregressive" and "MA" stands for "moving average": this model includes past white noise, namely:

$$X_n = \mu + \alpha_1(X_{n-1} - \mu) + ... + \alpha_p(X_{n-p} - \mu) + Z_n + \beta_1 Z_{n-1} + ... + \beta_q Z_{n-q}.$$

Remember

The characteristic equation and the coinditions for stationarity for $ARMA(p, q)$ coincide with such for its $AR(p)$ component.

10. Data Reduction

A useful resource for this chapter is *Using Multivariate Statistics* by B.G.Tabachnick and L.S.Fidell. The material taught in this chapter will be met from a machine learning perspective in *MA-M28 Modelling and Machine Learning* — please see chapter 4 of *Essential Math for Data Science* if you would like an insight into this.

Factor Analysis (FA) and Principal Component Analysis (PCA) are statistical techniques applied to a (large) set of variables to try to reduce them into subsets of relatively independent variables. Such subsets contain variables that are correlated with one another, but largely independent of other subsets of variables and are combined into factors (or components in PCA).

Therefore, the idea of FA and PCA is to summarise patterns of correlations among observed variables and then to use this information to reduce a large number of observed variables to a smaller number of factors. A good FA or PCA makes sense, a bad one does not, therefore a good understanding of the data is required.

10.1 (Exploratory) Factor Analysis

In Factor Analysis, the subsets of variables are unobservable **latent variables** — we cannot measure them directly. Examples of such variables could be intelligence or social class. We could try to measure such concepts indirectly, for example by measuring occupation, salary and value of home for social class.

Mathematically, the technique involves representing the original variables as a linear combination of the "hidden" factors and an error term. If $Y_1, Y_2, \dots, Y_n$ represent the n observed variables with means $\mu_1, \dots, \mu_n$, and $F_1, \dots, F_m$ represent the "hidden" m factors, then we may consider the centralised observations $X_i = Y_i - \mu_i$ as follows:

$$\begin{aligned} X_1 = Y_1 - \mu_1 &= a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + \epsilon_1 \\ X_2 = Y_2 - \mu_2 &= a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + \epsilon_2 \\ &\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ X_n = Y_n - \mu_n &= a_{n1}F_1 + a_{n2}F_2 + \cdots + a_{nm}F_m + \epsilon_n, \end{aligned} \tag{10.1}$$

where a_{ij} represents the **factor loading** of the i^{th} variable on the j^{th} factor and ϵ_i represents the error or unique specific factor. We assume that ϵ_i has 0 mean and specific variance ψ_i. In matrix notation, this can be represented as,

$$X = AF + \epsilon. \tag{10.2}$$

Consider the following illustrative example.

Example 10.1. In an experiment, 200 primary school children were psychologically tested. The children were tested on the following (the observed variables):

- Paragraph comprehension (X_1);
- Sentence completion (X_2);
- Word meaning (X_3);
- Addition (X_4);
- Counting (X_5).

A factor analysis gives the following linear combinations:

$$\begin{aligned} X_1 &= 0.81F_1 + 0.06F_2 + \epsilon_1 \\ X_2 &= 0.72F_1 + 0.08F_2 + \epsilon_2 \\ X_3 &= 0.91F_1 + 0.01F_2 + \epsilon_3 \\ X_4 &= 0.02F_1 + 0.69F_2 + \epsilon_4 \\ X_5 &= 0.11F_1 + 0.92F_2 + \epsilon_5 \end{aligned}$$

Clearly, variables X_1, X_2 and X_3 have a high factor loading with F_1 and a low factor loading with F_2. Variables X_4 and X_5 have a low factor loading

with F_1 and a high factor loading with F_2. This suggests that F_1 is the factor, or latent variable, **literacy skills** and F_2 is the factor, or latent variable, **numeracy skills**.

This gives a general insight into the method. We now consider the finer details of the procedure, in particular, we will investigate the methods of calculating factor loadings and determining factors. We first consider/recall the definition of the covariance of random variables,

$$\begin{aligned}\operatorname{cov}(X,Y) &= \mathbb{E}((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))) \\ &= \mathbb{E}(XY - \mathbb{E}(X)Y - X\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y)) \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y) \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

For matrices, this generalises to,

$$\Sigma = \mathbb{E}\left[(X-\mathbb{E}(X))(X-\mathbb{E}(X))^T\right]. \tag{10.3}$$

Since the X_i's are centralised in our calculations, $\mathbb{E}(X) = 0$ in the linear model (10.2), and we obtain Σ, the covariance matrix of the variables $X_1, \dots, X_n$, as follows:

$$\Sigma = \mathbb{E}(XX^T),$$

by (10.3) with $\mathbb{E}(X) = 0$, and by (10.2),

$$\begin{aligned}\mathbb{E}(XX^T) &= \mathbb{E}((AF+\epsilon)(AF+\epsilon)^T) \\ &= \mathbb{E}((AF+\epsilon)(F^TA^T+\epsilon^T)) \\ &= \mathbb{E}(AFF^TA^T) + A\mathbb{E}(F\epsilon^T) + \mathbb{E}(\epsilon F^T)A^T + \mathbb{E}(\epsilon\epsilon^T) \\ &= AIA^T + 0 + 0 + \Psi \\ &= AA^T + \Psi,\end{aligned}$$

where Ψ is a diagonal matrix of the specific variances ψ_i. We assume that the factors are uncorrelated with unit variance, hence $\mathbb{E}(FF^T) = I$ above. Also, the cross-multiplication terms are 0 since we assume that the factors are not correlated with the errors ϵ. Note that the factors themselves have now dropped out of the calculations. Next we set,

$$R = AA^T = \Sigma - \Psi,$$

where R is known as the adjusted covariance matrix, i.e. the variance of the observations are "adjusted" by subtracting the specific variances. Like Σ, R

is a symmetric matrix and hence using results from linear algebra we may state

$$R = VLV^T,$$

where V is a matrix of the eigenvectors of R and L a matrix of the eigenvalues of R. Furthermore,

$$\begin{aligned} R = VLV^T &= V\sqrt{L}\sqrt{L}V^T \\ &= (V\sqrt{L})(\sqrt{L}V^T) \\ &= (V\sqrt{L})(V\sqrt{L})^T \\ &= AA^T, \end{aligned}$$

where we used that L is diagonal, hence $\sqrt{L}^T = \sqrt{L}$. This implies that

$$A = V\sqrt{L}. \tag{10.4}$$

Therefore, once the eigenvectors and eigenvalues of R are known, the factor loading matrix A is easily obtained by (10.4).

Remark 10.2

Note that equation (10.4) is true if all factors, or eigenvalues, are used in the model. However, we only want to consider significant factors (i.e. we may choose to ignore certain factors) and therefore we require methods of extracting and evaluating such factors.

In Python, there are various methods of "extracting" the factors, the main one being **Principle Axis Factoring** which finds the least number of factors that account for the common variance of a set of variables.

Evaluating Factors

There are various means of evaluating and extracting the factors, including:

- **Eigenvalues:** one method of choosing factors is to consider factors with eigenvalues > 1. This is known as the **Kaiser criterion**.
- **Scree plot:** this is a plot of the eigenvalues which can indicate where there is a clear cut-off (an inflexion point) between large and small eigenvalues.
- **Communality:** this is the sum of the squared loadings for a variable across factors and it provides a percentage of variance accounted for by

the factors. The accepted proportions for communality are dependent on the sample size. The general rule is as follows:

  - If all communalities > 0.6, then this is considered very strong and we may even take relatively small samples in this scenario (< 100);
  - Communalities > 0.5 are adequate for sample of size $100 - 200$, or more;
  - Smaller communalities may be accepted for larger sample sizes.

- **Factor loadings:** we aim for factor loadings to be ≥ 0.4 for the main factor. We then study the observations with high loadings of a particular factor in order to try to identify the factor. Factor loadings appear in the **Pattern Matrix** in Python.

Rotations

In cases where the factor loading matrix A cannot be interpreted clearly, it may be rotated to try to improve interpretations. The aim is to maximise high correlations between factors and variables and to minimise low ones. This can be performed since the factor loading matrix is not uniquely defined. There are two different types of rotation, orthogonal and oblique.

Orthogonal Rotations

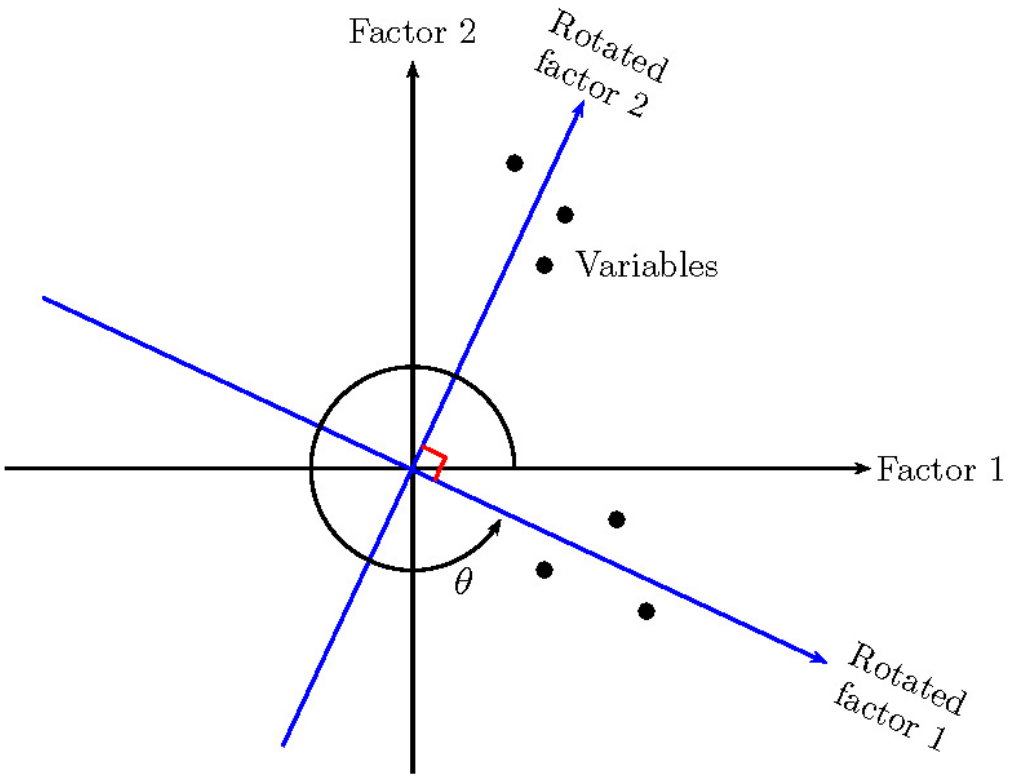
Orthogonal rotations are used when we assume that the factors are uncorrelated. There are various orthogonal rotations possible, with the most common being **Varimax**, **Quartimax** and **Equamax**. The process involves a simple matrix multiplication as follows:

$$A_{\text{rotated}} = A\Lambda, \tag{10.5}$$

where Λ is the rotation matrix,

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}. \tag{10.6}$$

For the case where we have 2 factors, a typical orthogonal rotation is illustrated as follows:



Varimax Rotation

Varimax is often the most common used rotation which involves a variance maximising procedure. The goal of varimax rotation is to maximise the variance of factor loadings by making high loadings higher and low ones lower for each factor.

Quartimax Rotation

Quartimax does for variables what varimax does for factors. It simplifies variables by increasing the dispersion of the loadings within variables, across factors.

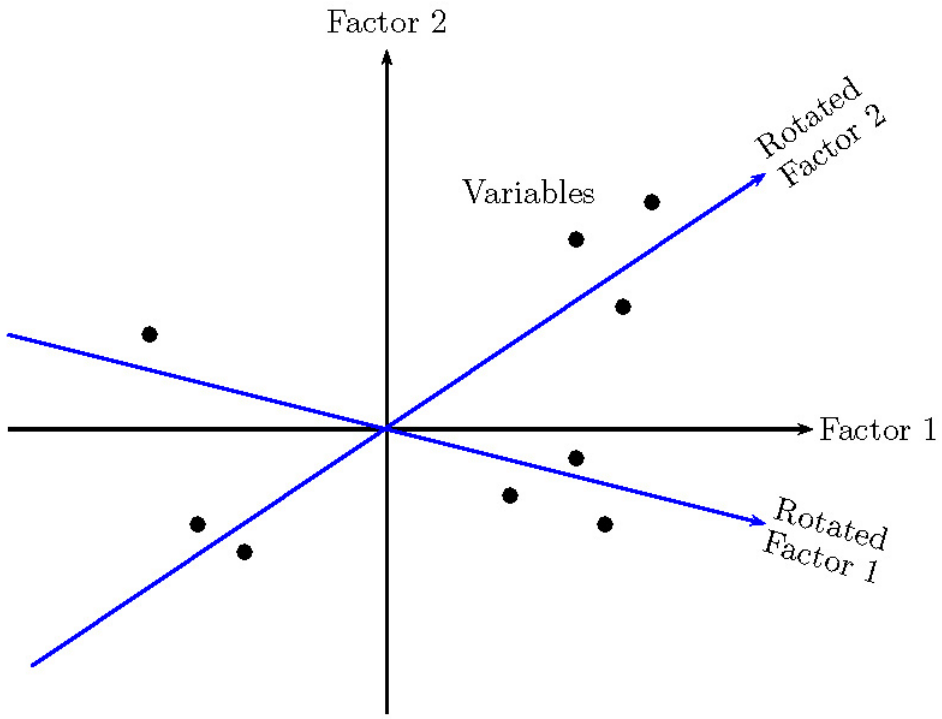
Equamax Rotation

Equamax rotation is a hybrid between varimax and quartimax that tries simultaneously to simplify the factors and the variables.

In conclusion, varimax rotation simplifies the factors, quartimax the variables and equamax both.

Oblique Rotation

Oblique rotations allow the factors to be correlated. In practice, this is a highly likely possibility. For example, if two of our factors were Achievement and Alcoholism, we would expect there to be a correlation between these factors. Oblique rotations also include orthogonal rotations, i.e. when the factors are assumed to be uncorrelated. For the case where we have 2 factors, a typical oblique rotation is illustrated as follows:



The two main types of oblique rotations are Direct Oblimin and Promax.

Direct Oblimin is the default oblique rotation we will use in Python.

The Promax method is quicker and is therefore better to use if dealing with large data sets.

It is good practice to first perform an oblique rotation and to change to an orthogonal rotation if correlation between the factors does not seem to exist. In Python, this can be checked by examining the **Factor Correlation**

Matrix Φ, which is given by,

$$\Phi = \begin{pmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1m} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2m} \\ \vdots & & \vdots & \\ \phi_{m1} & \phi_{m2} & \cdots & \phi_{mm} \end{pmatrix},$$

where m is the number of factors.

The general rule is to use an oblique rotation if $|\phi_{ij}| > 0.32$ for all $i, j = 1, \ldots, m, i \neq j$. Clearly, we do not include the diagonal terms as these will always be 1 (i.e. the correlation of a factor with itself).

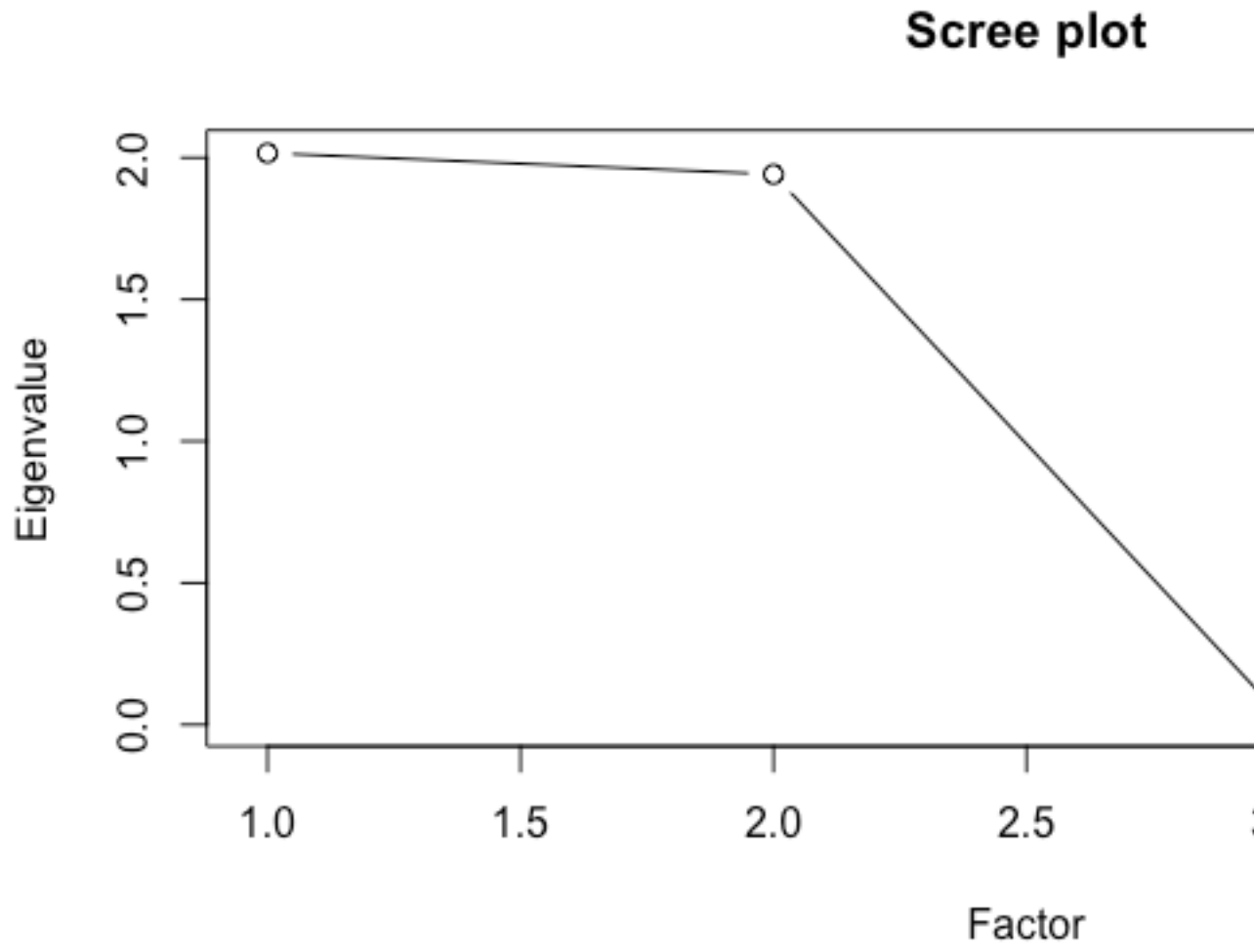
If an oblique rotation is found to be suitable, the elements of the **Pattern Matrix** are reported.

The following example is for illustrative purposes only.

Example 10.3. In an experiment, skiers were asked about their opinions on the cost of a skiing ticket (COST), the speed of the ski lifts (LIFT), the depth of the snow (DEPTH) and the moisture of snow (POWDER). Here is the raw data,

Skier	COST	LIFT	DEPTH	POWDER
S_1	32	64	65	67
S_2	61	37	62	65
S_3	59	40	45	43
S_4	36	62	34	35
S_5	62	46	43	40

When no limit is placed on the number of factors, we have 4 factors with eigenvalues 2.02, 1.94, 0.04 and 0.00. Using the Kaiser criterion and a scree plot, we keep the eigenvalues 2.02 and 1.94, and then we run the factor analysis again with these 2 factors only. Below is the scree plot that shows a clear distinction between the eigenvalues:



Once we run the analysis keeping only the 2 strong factors, we obtain the communalities under the **extraction** column in the table below:

COST	LIFT	DEPTH	POWDER
0.9704176	0.9596748	0.9880820	0.9965461

Clearly, as these communalities are close to 1, a large proportion of the variation in each variable can be accounted for by the factors. The pattern matrix is given by,

	PA1	PA2	h2	u2	com
COST	-0.40	0.90	0.97	0.0296	1.4
LIFT	0.25	-0.95	0.96	0.0403	1.1
DEPTH	0.93	0.35	0.99	0.0119	1.3
POWDER	0.96	0.29	1.00	0.0035	1.2

We see that DEPTH and POWDER have a high factor loading with Factor 1

and that COST and LIFT have a high factor loading with Factor 2. However, since the remaining factor loadings are not negligible, we will perform rotations with the aim of obtaining a clearer interpretation. Firstly, let us consider the Direct Oblimin oblique rotation. The Factor Correlation Matrix Φ is given below,

	PA1	PA2
PA1	1.00	-0.02
PA2	-0.02	1.00

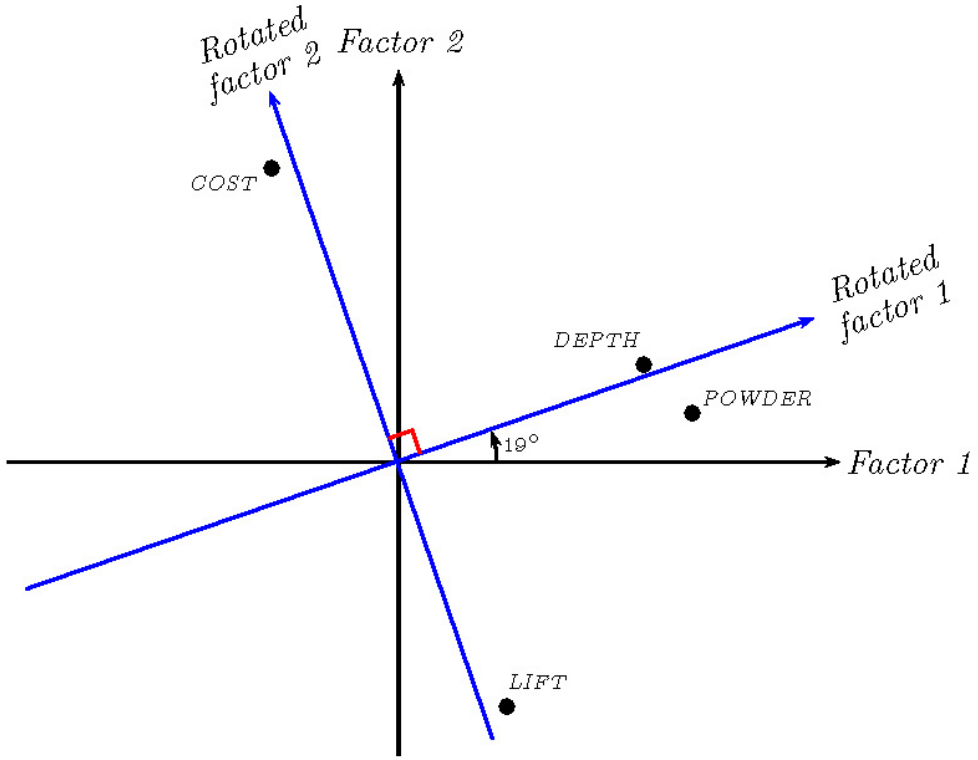
We can see that $|\phi_{ij}| \leq 0.32$ for $i \neq j$. Therefore, we conclude that an oblique rotation is not warranted and instead we use the Varimax orthogonal rotation,

	PA1	PA2	h2	u2	com
COST	-0.08	-0.98	0.97	0.0296	1
LIFT	-0.08	0.98	0.96	0.0403	1
DEPTH	0.99	-0.02	0.99	0.0119	1
POWDER	1.00	0.05	1.00	0.0035	1

This is a rotation in the sense of (10.6) by 0.33 radians (19 degrees). Using (10.5), we can confirm the rotated factors are given by:

$$
\begin{aligned}
A_{\text{rotated}} = A\Lambda &= \begin{pmatrix} -.40 & .90 \\ .25 & -.95 \\ .93 & .35 \\ .96 & .29 \end{pmatrix} \begin{pmatrix} \cos 0.33 & -\sin 0.33 \\ \sin 0.33 & \cos 0.33 \end{pmatrix} \\
&= \begin{pmatrix} -.40 & .90 \\ .25 & -.95 \\ .93 & .35 \\ .96 & .29 \end{pmatrix} \begin{pmatrix} .95 & -.33 \\ .33 & .95 \end{pmatrix} \\
&= \begin{pmatrix} -.08 & .98 \\ -.08 & -.98 \\ .99 & .02 \\ 1 & .05 \end{pmatrix}
\end{aligned}
$$

Here is an illustration of the rotation:



In this example, it is clear that the variables DEPTH and POWDER are associated with a factor concerning snow conditions. The variables COST and LIFT are associated with a factor concerning resort conditions.

10.2 Principal Component Analysis (PCA)

Principal component analysis is similar to factor analysis in that both are used for data reduction, and they often provide similar results. However, in PCA we write the components (factors in FA) as a linear combination of the variables, where as in FA, we write the variables in terms of the factors, see (10.1). This can be expressed as follows:

$$\begin{aligned} C_1 &= e_{11}X_1 + e_{12}X_2 + \cdots + e_{1n}X_n \\ C_2 &= e_{21}X_1 + e_{22}X_2 + \cdots + e_{2n}X_n \\ &\vdots \qquad\qquad \vdots \\ C_m &= e_{m1}X_1 + e_{m2}X_2 + \cdots + e_{mn}X_n, \end{aligned}$$

where $C_1, \ldots, C_m$ represent the components, $X_1, \ldots, X_n$ are the variables and e_{ij} are the regression coefficients, or weights of the variables. Similar to (10.2),

we can rewrite this system in matrix form as below,

$$C = EX.$$

In the case of factor analysis, the factor loadings were given by the eigenvectors and eigenvalues, however, in principal component analysis, the weightings of the variables are given by the eigenvectors and eigenvalues of the covariance matrix. The largest eigenvalue and associated eigenvector is applied to the first principal component, with the next largest applied to the second principal component etc.

Choosing Principal Components

The principal components are chosen in the following way:

- The first principal component,

  $$C_1 = e_{11}X_1 + e_{12}X_2 + \cdots + e_{1n}X_n,$$

  is chosen such that it accounts for as much variation in the data as possible, subject to the condition that $e_{11}^2 + e_{12}^2 + ... e_{1n}^2 = 1$.

- The second,

  $$C_2 = e_{21}X_1 + e_{22}X_2 + \cdots + e_{2n}X_n,$$

  is chosen such that the variance is as high as possible, similarly conditional on $e_{21}^2 + e_{22}^2 + ... e_{2n}^2 = 1$.

- The second principal component must be chosen such that it is **uncorrelated** with the first.

- The i^{th} principal component,

  $$C_i = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{in}X_n,$$

  again is chosen such that the variance is as high as possible, conditional on $e_{i1}^2 + e_{i2}^2 + ... e_{in}^2 = 1$ and it being uncorrelated with all other principal components.

- All principal components are uncorrelated with each other.

As previously stated, the weightings e_{ij} are obtained from eigenvectors corresponding to the i^{th} largest eigenvalue.

Remark 10.4

The process of maximising the variance uses the theory of Constrained Optimisation, which, in this case, essentially means maximising the variance for the i^{th} principal component conditional on $e_{i1}^2 + e_{i2}^2 + \ldots e_{in}^2 = 1$.

Number of Components and Rotations

We will use the same guidelines for determining the number of components as we did for factors in factor analysis, i.e. by evaluating eigenvalues, scree plots, communalities and factor loadings. Similarly, we will use rotations in the same way for principal component analysis as we did in factor analysis, i.e. if components are not clear from the unrotated results, we next perform an oblique rotation. If there is not enough correlation between the principal components to warrant the use of an oblique rotation, we then perform an orthogonal rotation.

Deciding between PCA and FA

PCA is used to simply reduce the observed variables into a smaller set of important independent composite variables (components). FA tends to be used when there are suspected latent factors (not directly measurable factors) causing the observed variables.