# Lecture Book
# Higher Mathematics II

D. T. McGuiness, Ph.D

Version: 2025.SS

Topics on Fundamental Science

# Contents

# List of Figures

# List of Tables

# Part I.

# Probability & Statistics

Chapter 1

# Theory of Probability

## Table of Contents



## 1.1. Introduction

When the data we are working are influenced by "chance," by factors whose effect we cannot predict exactly[1], we have to rely on **probability theory**. The application of this theory nowadays appears in numerous fields such as from studying a game of cards to the global financial market and allow us to model processes of chance called random experiments.

[1] This could be weather data, stock prices, life spans or ties, etc.

In such an experiment we observe a random variable $X$, that is, a function whose values in a **trial**[2] occur "by chance" according to a probability distribution which gives the individual probabilities, which possible values of $X$ may occur in the long run.

[2] a performance of an experiment.

i.e., each of the six faces of a die should occur with the same probability, 1/6.

Or we may simultaneously observe more than one random variable, for instance, height and weight of persons or hardness and tensile strength of steel. But enough about spoiling all the fun and let's begin with looking at data.

## Representing Data

Data can be represented numerically or graphically in different ways

> i.e., a news website may contain tables of stock prices and currency exchange rates, curves or bar charts illustrating economical or political developments, or pie charts showing how inflation is calculated

And there are numerous other representations of data for special purposes. In this section, we will discuss the use of standard representations of data in statistics[3].

[3] There are various software dedicated to analyse and visualise statistical data. Some of these include: R, a statistical programming language, Python, Matlab,..

### **Exercise** 1.1: Recording Data

Sample values,such as observations and measurements, should be recorded in the order in which they occur. Sorting, that is, ordering the sample values by size, is done as a first step of investigating properties of the sample and graphing it.
As an example let's look at super alloys.
Super alloys is a collective name for alloys used in jet engines and rocket motors, requiring high temperature (typically 1000řC), high strength, and excellent resistance to oxidation.
Thirty (30) specimens of Hastelloy C (nickel- based steel, investment cast) had the tensile strength (in 1000 lb>sq in.), recorded in the order obtained and rounded to integer values.

$$
\begin{matrix}
89 & 77 & 88 & 91 & 88 & 93 & 99 & 79 & 87 & 84 & 86 & 82 & 88 & 89 & 78 \\
90 & 91 & 81 & 90 & 83 & 83 & 92 & 87 & 89 & 86 & 89 & 81 & 87 & 84 & 89
\end{matrix}
\tag{1.1}
$$

Of course depending on the need the data needs to be sorted which is shown below:

$$
\begin{matrix}
77 & 78 & 79 & 81 & 81 & 82 & 83 & 83 & 84 & 84 & 86 & 86 & 87 & 87 & 87 \\
88 & 88 & 88 & 89 & 89 & 89 & 89 & 89 & 90 & 90 & 91 & 91 & 92 & 93 & 99
\end{matrix}
$$

## Graphic Representation of Data

Let's now use the data we have seen in Example 1 and see the methods we can use for graphic representations.

### **Exercise** 1.2: Leaf Plots

One of the simplest yet most useful representations of data [1]. For Eq. (1.1) it is shown in Fig. 507.

| LO | 12 12 |
|---|---|
| 7 | 789 |
| 8 | 1123344 |
| 8 | 6677788899999 |
| 9 | 001123 |
| 9 | 9 |
| HI | 172 |

**Table 1.1.:** Stem and Leaf plot of the data given in Example 1.

The numbers in Eq. (1.1) range from 78 to 99; which you can also see this in the sorted list. To visualise this data feature, we divide these numbers into five (5) groups:

75-79, 80-84, 85-89, 90-94, 95-99.

The integers in the tens position of the groups are 7, 8, 8, 9, 9. These form the stem which can be seen in **Table** 1.1. The first leaf is 789, representing 77, 78, 79. The second leaf is 1123344, representing 81, 81, 82, 83, 83, 84, 84. And so on. The number of times a value occurs is called its **absolute frequency**.
Therefore in this example, 78 has absolute frequency 1, the value 89 has absolute frequency 5, etc. The column to the extreme left in Fig. 507 shows the cumulative absolute frequencies, that is, the sum of the absolute frequencies of the values up to the line of the leaf.
Thus, the number 10 in the second line on the left shows that Eq. (1.1) has 10 values up to and including 84. The number 23 in the next line shows that there are 23 values not exceeding 89, etc. Dividing the cumulative absolute frequencies by n (= 30 in **Table** 1.1) gives the cumulative relative frequencies 0.1, 0.33, 0.76, 0.93, 1.00 respectively ■

**Exercise** 1.3: Histogram

For large sets of data, histograms are better in displaying the distribution of data than stem-and-leaf plots. The principle is explained in **Fig.** 1.1.

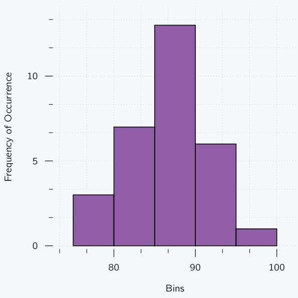


**Figure 1.1.:** The histogram of the data given in **Exercise 1**.

The bases of the rectangles in seen in **Fig.** 1.1 are the $x$-intervals[4] where there rage is:

$$74.5 - -79.5, \quad 79.5 - -84.5, \quad 84.5 - -89.5,$$
$$89.5 - -94.5, \quad 94.5 - -99.5,$$

whose midpoints, known as class marks, are

$$x = 77, 82, 87, 92, 97,$$

respectively. The height of a rectangle with class mark $x$ is the relative class frequency $f_{\text{rel}}(x)$, defined as the number of data values in that class interval, divided by $n$ (= 30 in our case). Hence the areas of the rectangles are proportional to these relative frequencies,

$$0.10,\ 0.23,\ 0.43,\ 0.17,\ 0.07,$$

so that histograms give a good impression of the distribution of data.

[4] known as class intervals.

## Mean, Standard Deviation, and Variance

Medians and quartiles are easily obtained by ordering and counting[5].

[5] This can be done without the need of calculators.

However this method does not give full information on data as you can change data values to some extent without changing the median.

The average size of the data values can be measured in a more refined way by the mean:

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_{\text{j}} = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_{\text{n}}\right). \tag{1.2}$$

This is the arithmetic mean of the data values, obtained by taking their sum and dividing by the data size ($n$). Therefore the arithmetic mean for Eq. (1.1) is:

$$\overline{x} = \frac{1}{30}(89 + 77 + \cdots + 89) = \frac{260}{3} \approx 86.7 \quad \blacksquare$$

As we can see every data value contributes, and changing one of them will change the mean. Similarly, the spread[6] of the data values can be measured in a more refined way by the **standard deviation** $s$ or by its square, the **variance**[7]

$$s^2 = \frac{1}{n-1}\sum j = 1^n \left(x_{\text{j}} - \overline{x}\right)^2 = \frac{1}{n-1}\left[\left(x_1 - \overline{x}\right)^2 + \cdots + \left(x_{\text{n}} - \overline{x}\right)^2\right] \tag{1.3}$$

[6] also known as variability.

[7] The symbol for variance is interesting as each domain have their own definition, as $s^2$, $\sigma^2$ and $\text{Var}()$ are all acceptable symbols.

Therefore, to obtain the variance of the data, take the difference (i.e., $x_{\text{j}} - \overline{x}$) of each data value from the mean, square it, take the sum of these $n$ squares, and divide it by $n - 1$.

> To get the standard deviation $s$, take the square root of $s^2$.

Returning back to our super alloy example, using $\overline{x} = 260/3$, we get for the data given in Eq. (1.1) the variance:

$$s^2 = \frac{1}{29}\left[\left(89 - \frac{260}{3}\right)^2 + \left(77 - \frac{260}{3}\right)^2 + \cdots + \left(89 - \frac{260}{3}\right)^2\right] = \frac{2006}{87} \approx 23.06 \quad \blacksquare$$

Therefore, the standard deviation is:

$$s = \sqrt{2006/87} \approx 4.802$$

> The standard deviation has the same dimension as the data values, which is an advantage. On the other hand, the variance is preferable to the standard deviation in developing statistical methods.

### Empirical Rule

For any round-shaped symmetric distribution of data the intervals:

$$\overline{x} \pm s, \quad \overline{x} \pm 2s, \quad \overline{x} \pm 3s, \qquad \text{contain about} \qquad 68\%, \quad 95\%, \quad 99.7\%.$$

respectively, of the data points.

**Exercise** 1.4: Empirical Rule, Outliers, and z-Score

For the data set given in Eq. (1.1), with $\overline{x} = 86.7$ and $s = 4.8$, the three (3) intervals in the Rule are:

$$81.9 \leq x \leq 91.5, \quad 77.1 \leq x \leq 96.3, \quad 72.3 \leq x \leq 101.1$$

and contain 73% (22 values remain, 5 are too small, and 5 too large), 93% (28 values, 1 too small, and 1 too large), and 100%, respectively.
If we reduce the sample by omitting the outlier value 99, mean and standard deviation reduce to $\overline{x}_{\text{red}} = 86.2$, and $s_{\text{red}} = 4.3$, approximately, and the percentage values become 67% (5 and 5 values outside), 93% (1 and 1 outside), and 100%.
Finally, the relative position of a value $x$ in a set of mean $\overline{x}$ and standard deviation $s$ can be measured by the **z-score**:

$$z(s) = \frac{x - \overline{x}}{s}$$

This is the distance of $x$ from the mean $\overline{x}$ measured in multiples of $s$. For instance:

$$z(s) = \frac{(83 - 86.7)}{4.8} = -0.77$$

This is negative because 83 lies below the mean. By the Empirical Rule, the extreme z-values are about -3 and 3. $\blacksquare$

## 1.2. Experiments & Outcomes

Now we have the basis covered, it is time to look at **probability theory**[8]. This theory has the purpose of providing mathematical models of situations affected or even governed by change effects,

[8] Sometimes known as probability calculus.

for instance, in weather forecasting, life insurance, quality of technical products (computers, batteries, steel sheets, etc.), traffic problems, and, of course, games of chance with cards or dice. And the accuracy of these models can be tested by suitable observations or experiments.

Let's start by defining some standard terms:

**experiment** A process of measurement or observation, in a laboratory, in a factory, ...

**randomness** Situation where absolute prediction is not possible.

**trial** A single performance of an experiment

**outcome** The result of a trial[9]

**sample Space** Defined as $S$, is the set of all possible outcomes of an experiment.

[9] also known as sample point.

**Exercise** 1.5: Sample Spaces of Random Experiments & Events

- Inspecting a lightbulb | $S = \{\text{Defective, Non-defective}\}$ .
- Rolling a die | $S = \{1, 2, 3, 4, 5, 6\}$

  events are
  - A = 1, 3, 5 ("Odd number")
  - B = 2, 4, 6 ("Even number"), etc.
- Counting daily traffic accidents in Vienna | $S = \{\text{the integers in some interval}\}$.

### 1.2.1. Unions, Intersections, and Complements of Events

In connection with basic probability laws we also need the following concepts and facts about events[10] $A, B, C, \cdots$ of a given sample space $S$.

[10] called subsets of the probability event $S$.

- The **union** $A \cup B$ of $A$ and $B$ consists of all points in $A$ or $B$ or both.
- The **intersection** $A \cap B$ of $A$ and $B$ consists of all points that are in both $A$ and $B$.

If $A$ and $B$ have no points in common, we write

$$A \cap B = \emptyset$$

where $\emptyset$ is the empty set[11]. and we call $A$ and $B$ **mutually exclusive** (or **disjoint**) as, in a trial, the occurrence of $A$ *excludes* that of $B$ (and conversely)–if your die turns up an odd number, it cannot turn up an even number in the same trial. Similarly, a coin cannot turn up *Head* and *Tail* at the same time.

[11] This means it is a set which contains nothing.

- The **Complement** of $A$ is $A^c$[12]. This is the set of all the points of $S$ *not* in $A$. Therefore,

$$A \cap A^c = \emptyset, \qquad A \cup A^c = S.$$

[12] Another notation for the complement of $A$ is $\overline{A}$ (instead of $A^c$), but we shall not use this because in set theory $\overline{A}$ is used to denote the *closure* of $A$.

**Unions and intersections** of more events are defined similarly. The **union**:

$$\bigcup_{j=1}^{m} A_j = A_1 \cup A_2 \cup \cdots \cup A_m.$$

of events $A_1, \cdots, A_m$ consists of all points that are in at least one $A_j$. Similarly for the union $A_1 \cup A_2 \cup \cdots$ of infinitely many subsets $A_1, A_2, \cdots$ of an *infinite* sample space $S$ (that is, $S$ consists of infinitely many points). The **intersection**:

$$\bigcap_{j=1}^{m} A_j = A_1 \cap A_2 \cap \cdots \cap A_m$$

of $A_1, \cdots, A_m$ consists of the points of $S$ that are in each of these events. Similarly for the intersection $A_1 \cap A_2 \cap \cdots$ of infinitely many subsets of $S$.

Working with events can be illustrated and facilitated by **Venn diagrams** for showing unions, intersections, and complements, as in **Fig.** 1.2, which are typical examples that give the idea.

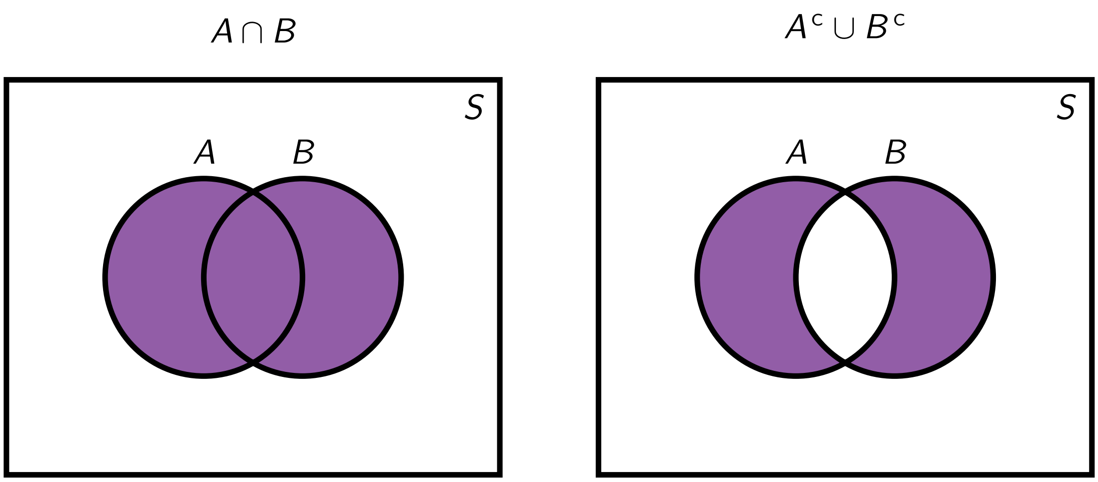


**Figure 1.2.:** Examples of Venn diagrams.

## 1.3. Probability

The probability of an event $A$ in an experiment is to measure **how frequently** $A$ is *about* to occur if we make many trials. If we flip a coin, then heads $H$ and tails $T$ will appear *about* equally often.

> we say that $H$ and $T$ are **"equally likely."**

Similarly, for a regularly shaped die of homogeneous material[13] each of the six outcomes $1, \cdots, 6$ will be equally likely. These are examples of experiments in which the sample space $S$ consists of finitely many outcomes (points) that for reasons of some symmetry can be regarded as equally likely.

[13] called a fair dice

Let's formulate this in a theory.

**Theory** 1.5: First Definition of Probability

If the sample space $S$ of an experiment consists of **finitely** many outcomes (points) that are equally likely, then the probability $P(A)$ of an event $A$ is defined to be:

$$P(A) = \frac{\text{Number of points in A}}{\text{Number of points in S}}.$$

From this definition it follows immediately that, in particular,

$$P(S) = 1.$$

**Exercise** 1.6: Fair Die

In rolling a fair die once, what is the probability $\Pr A$ of A of obtaining a 5 or a 6? The probability of B: "Even number"?

**Solution**

The six outcomes are equally likely, so that each has probability 1/6. Therefore:

$$\Pr A = \frac{2}{6} = \frac{1}{3} \qquad \text{and} \qquad \Pr B = \frac{3}{6} = \frac{1}{2} \quad \blacksquare$$

The above theory takes care of many games as well as some practical applications, but not of all experiments, as in many problems we do not have finitely many equally likely outcomes. To arrive at a more general definition of probability, we regard probability as the counterpart of **relative frequency**:

$$f_{\text{rel}}(A) = \frac{f(A)}{n} = \frac{\text{Number of times A occurs}}{\text{Number of trials}} \tag{1.4}$$

Now if A did not occur, then $f(A) = 0$. If A always occurred, then $f(A) = n$. These are of course extreme cases. Division by $n$ gives:

$$0 \leq f_{\text{rel}}(A) \leq 1 \tag{1.5}$$

In particular, for $A = S$ we have $f(S) = n$ as $S$ always occurs[14]. Division by $n$ gives:

[14] meaning that some event always occurs

$$f_{\text{rel}}(S) = 1 \tag{1.6}$$

Finally, if $A$ and $B$ are mutually exclusive, they cannot occur together. Therefore the absolute frequency of their union A = B must equal the sum of the absolute frequencies of A and B. Division by $n$ gives the same relation for the relative frequencies:

$$f_{\text{rel}}(A \cup B) = f_{\text{rel}}(A) + f_{\text{rel}}(B) \tag{1.7}$$

We can now extend the definition of probability to experiments in which equally likely outcomes are not available.

**Theory** 1.6: General Definition of Probability

Given a sample space $S$, with each event $A$ of $S$ ($A$ being a subset of $S$) there is associated a number $\Pr A$, called the **probability** of $A$, such that the following **axioms of probability** are satisfied.

- For every $A$ in $S$,
$$0 \leqq \Pr A \leqq 1. \tag{1.8}$$
- The entire sample space $S$ has the probability
$$P(S) = 1. \tag{1.9}$$
- For mutually exclusive events $A$ and $B$:
$$P(A \cup B) = P(A) + P(B) \qquad (A \cap B = \emptyset). \tag{1.10}$$

- If $S$ is infinite (has infinitely many points), Axiom 3 has to be replaced by Eq. (1.4) For mutually exclusive events $A_1, A_2, \cdots,$

$$P\left(A_1 \cup A_2 \cup \cdots\right) = P\left(A_1\right) + P\left(A_2\right) + \cdots. \tag{1.11}$$

In the infinite case the subsets of $S$ on which $P(A)$ is defined are restricted to form a so-called $\sigma$-algebra.

## Basic Theorems of Probability

We will see that the axioms of probability will enable us to build up probability theory and its application to statistics. We begin with three (3) basic theorems. The first one is useful if we can get the probability of the complement $A^c$ more easily than $\Pr A$ itself.

**Theory** 1.6: Complementation Rule

For an event $A$ and its complement $A^c$ in a sample space $S$,

$$\Pr A^c = 1 - \Pr A \tag{1.12}$$

**Exercise** 1.7: Coin Tossing

Five (5) coils are tossed simultaneously.
Find the probability of the event A:

At least one head turns up. Assume that the coins are fair.

**Solution**

As each coin can turn up either heads or tails, the sample space consists of $2^5 = 32$ outcomes. Given the coins are fair, we may assign the same probability (1/32) to each outcome. Then the event $A^c$ (No heads turn up) consists of only 1 outcome. Hence $\Pr A^c = 1/32$, and the answer is:

$$\Pr A = 1 - \Pr A^c = \frac{31}{32} \quad \blacksquare$$

**Theory** 1.7: Addition Rule for Mutually Exclusive Events

For **mutually exclusive events** $A_1, \cdots, A_m$ in a sample space $S$,

$$\Pr A_1 \cup A_2 \cup \cdots A_m = \Pr A_1 + \Pr A_2 + \cdots + \Pr A_m. \tag{1.13}$$

**Exercise** 1.8: Mutually Exclusive Events

If the probability that on any workday a garage will get 10-20, 21-30, 31-40, over 40 cars to service is 0.20, 0.35, 0.25, 0.12, respectively, what is the probability that on a given workday the garage gets at least 21 cars to service?

**Solution**

As these are mutually exclusive events, the answer is:

$$0.35 + 0.25 + 0.12 = 0.72 \quad \blacksquare$$

However, most situations, events will **not be** mutually exclusive. Then we have the following theorem

**Theory** 1.8: Addition Rule for Arbitrary Events

For events A and B in a sample space,

$$\Pr A \cup B = \Pr A + \Pr A - \Pr A \cap B. \tag{1.14}$$

For mutually exclusive events A and B we have $A \cap B = \emptyset$ by definition and, by comparing Eq. (1.14) and Eq. (1.14):

$$\Pr \emptyset = 0 \tag{1.15}$$

**Exercise** 1.9: Union of Arbitrary Events

In tossing a fair die, what is the probability of getting an odd number or a number less than 4?

**Solution**

Let A be the event "Odd number" and B the event "Number less than 4." As these event are linked we can write:

$$\Pr A \cup B = \frac{3}{6} + \frac{3}{6} - \frac{2}{6} = \frac{2}{3}$$

as $A \cup B$ = *Odd number less than 4* = $\{1, 3\}$ ■

## Conditional Probability and Independent Events

It is often required to find the probability of an event $B$ given the condition of an event $A$ occurs. This probability is called the **conditional probability** *of B given A* and is denoted by $P\left(B|A\right)$.

In this case $A$ serves as a new (reduced) sample space, and that probability is the fraction of $\Pr A$ which corresponds to $A \cap B$. Thus

$$\Pr A|B = \frac{\Pr A \cap B}{\Pr A} \qquad \text{where} \qquad \Pr A \neq 0 \tag{1.16}$$

Similarly, the *conditional probability of A given B* is

$$\Pr B|A = \frac{\Pr A \cap B}{\Pr B} \qquad \text{where} \qquad \Pr B \neq 0 \tag{1.17}$$

**Theory** 1.9: Multiplication Rule

Given $A$ and $B$ are events defined in a sample space $S$ and $P(A) \neq 0, P(B) \neq 0$, then

$$P(A \cap B) = P(A)\, P\left(B|A\right) = P(B)\, P\left(A|B\right). \tag{1.18}$$

**Exercise** 1.10: Multiplication Rule

In producing screws, let:

- $A$ mean "screw too slim",
- $B$ mean "screw too short."

Let $\Pr A = 0.1$ and let the conditional probability that a slim screw is also too short be $P\left(B|A\right) = 0.2$. What is the probability that a screw that we pick randomly from the lot produced will be both too slim and too short?

**Solution**

$$\Pr A \cap B = \Pr A \Pr B|A = 0.1 \times 0.2 = 0.02 = 2\% \quad \blacksquare$$

## Independent Events

If events $A$ and $B$ are such that

$$P(A \cap B) = P(A)\,P(B)\,, \tag{1.19}$$

they are called **independent events**. Assuming $P(A) \neq 0, P(B) \neq 0$, we see from Eq. (1.16) - Eq. (1.18):

$$\Pr A|B = \Pr A, \qquad \Pr B|A = \Pr B.$$

This means that the probability of A does not depend on the occurrence or nonoccurrence of B, and conversely. This justifies the term independent.

## Independence of m Events

Similarly, $m$ events $A_1, \cdots, A_m$ are called independent if:

$$P\left(A_1 \cap \cdots \cap A_m\right) = P\left(A_1\right) \cdots P\left(A_m\right) \tag{1.20}$$

as well as for every $k$ different events $A_{j_1}, A_{j_2}, \cdots, A_{j_k}$.

$$P\left(A_{j_1} \cap A_{j_2} \cap \cdots \cap A_{j_k}\right) = P\left(A_{j_1}\right) P\left(A_{j_2}\right) \cdots P\left(A_{j_k}\right) \tag{1.21}$$

where $k = 2, 3, \cdots, m-1$. Accordingly, three events $A$, $B$, $C$ are independent if and only if

$$P(A \cap B) = P(A)\,P(B)\,, \tag{1.22}$$

$$P(B \cap C) = P(B)\,P(C)\,, \tag{1.23}$$

$$P(C \cap A) = P(C)\,P(A)\,, \tag{1.24}$$

$$P(A \cap B \cap C) = P(A)\,P(B)\,P(C)\,. \tag{1.25}$$

## Sampling

Our next example has to do with randomly drawing objects, *one at a time*, from a given set of objects. This is called **sampling from a population**, and there are two ways of sampling, as follows.

- **In sampling with replacement**, the object that was drawn at random is placed back to the given set and the set is mixed thoroughly. Then we draw the next object at random.

- **In sampling without replacement** the object that was drawn is put aside.

**Exercise** 1.11: Sampling w/o Replacement

A box contains 10 screws, three (3) of which are defective. Two screws are drawn at random. Find the probability that neither of the two screws is defective.

**Solution**

We consider the events

**A** First drawn screw non-defective,

**B** Second drawn screw non-defective.

We can see:

$$P(A) = \frac{1}{10}$$

as 7 of the 10 screws are non-defective and we sample at random, so that each screw has the same probability ($\frac{1}{10}$) of being picked.
If we sample with replacement, the situation before the second drawing is the same as at the beginning, and $P(B) = \frac{7}{10}$. The events are independent, and the answer is

$$P(A \cap B) = P(A)\,P(B) = 0 \cdot 7 \cdot 0.7 = 0.49\%.$$

If we sample without replacement, then $P(A) = \frac{7}{10}$, as before. If $A$ has occurred, then there are 9 screws left in the box, 3 of which are defective.
Thus $P\left(B|A\right) = \frac{6}{9} = \frac{2}{3}$, therefore:

$$P(A \cap B) = \frac{7}{10} \cdot \frac{2}{3} = 47\% \quad \blacksquare$$

## 1.4. Permutations & Combinations

Permutations and combinations help in finding probabilities $\Pr A = a/k$ by systematically counting the number $a$ of points of which an event $A$ consists; here, $k$ is the number of points of the sample space $S$. The practical difficulty is that $a$ may often be surprisingly large, so that actual counting becomes hopeless. For example, if in assembling some instrument you need 10 different screws in a certain order and you want to draw them randomly from a box[15] the probability of obtaining them in the required order is only 1/3,628,800 because there are

[15] Of course, this goes without saying, there is nothing but screws in this imaginary box.

$$10! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9 \cdot 10 = 3,628,800$$

orders in which they can be drawn. Similarly, in many other situations the numbers of orders, arrangements, etc. are often incredibly large.

### 1.4.1. Permutations

A **permutation** of given things[16] is an arrangement of these things in a row in some order.

[16] such as *elements* or *objects*

> i.e., for three (3) letters $a$, $b$, $c$ there are $3! = 1 \cdot 2 \cdot 3 = 6$ permutations: $abc$, $acb$, $bca$, $cab$, $cba$

Let's write this behaviour down as a theory:

**Theory 1.11: Permutations**

**Different things**

The number of permutations of $n$ different things taken all at a time is

$$n! = 1 \cdot 2 \cdot 3, \cdots, n. \tag{1.26}$$

**Classes of Equal Things**

If $n$ given things can be divided into c classes of alike things differing from class to class, then the number of permutations of these things taken all at a time is

$$\frac{n!}{n_1!n_2!\cdots n_c!} \quad \text{where} \quad n_1 + n_2 + \cdots + n_c = n, \tag{1.27}$$

where $n_j$ is the number of things in the $j^{th}$ class.

### permutation of $n$ things taken $k$ at a time

A permutation containing only $k$ of the $n$ given things. Two such permutations consisting of the same $k$ elements, in a different order, are different, by definition.

> i.e., there are 6 different permutations of the three letters $a$, $b$, $c$, taken two letters at a time, $ab$, $ac$, $bc$, $ba$, $ca$, $cb$.

### permutation of $n$ things taken $k$ at a time with repetitions

An arrangement obtained by putting any given thing in the first position, any given thing, including a repetition of the one just used, in the second, and continuing until $k$ positions are filled.

> i.e., there are $3^2 = 9$ different such permutations of $a$, $b$, $c$ taken 2 letters at a time, namely, the preceding 6 permutations and $aa$, $bb$, $cc$.

**Theory 1.11: Permutations**

The number of different permutations of $n$ different things taken $k$ at a time **without repetitions** is

$$n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!}, \tag{1.28}$$

and with repetitions is

$$n^k. \tag{1.29}$$

**Exercise 1.12: An Encrypted Message**

In an encrypted message the letters are arranged in groups of five letters, called words. Knowing the letter can be repeated, we see that the number of different such words is

$$26^5 = 11,881,376 \quad \blacksquare$$

For the case of different such words containing each letter no more than once is

$$\frac{26!}{(26-5)!} = 26 \cdot 25 \cdot 24 \cdot 23 \cdot 22 = 7,893,600 \quad \blacksquare$$

### 1.4.2. Combinations

In a permutation, the order of the selected things is essential. In contrast, a **combination** of a given things means any selection of one or more things without regard to order. There are two (2) kinds of combinations, as follows.

The number of **combinations of $n$ different things, taken $k$ at a time, without repetitions** is the number of sets that can be made up from the $n$ given things, each set containing $k$ different things and no two sets containing exactly the same $k$ things.

The number of **combinations of $n$ different things, taken $k$ at a time, with repetitions** is the number of sets that can be made up of $k$ things chosen from the given $n$ things, each being used as often as desired.

> i.e, there are three combinations of the three letters $a$, $b$, $c$, taken two letters at a time, without repetitions, namely, $ab$, $ac$, $bc$, and six such combinations with repetitions, namely, $ab$, $ac$, $bc$, $ca$, $bb$, $cc$.

**Theory** 1.12: Combinations

The number of different combinations of n different things taken, $k$ at a time, without repetitions, is:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{1\cdot 2\cdots k}, \tag{1.30}$$

and the number of those combinations with repetitions is

$$\binom{n+k-1}{k}. \tag{1.31}$$

**Exercise** 1.13: Sampling Light-bulbs

The number of samples of five light-bulbs that can be selected from a lot of 500 bulbs is

$$\binom{500}{5} = \frac{500!}{5!495!} = \frac{500\cdot 499\cdot 498\cdot 497\cdot 476}{1\cdot 2\cdot 3\cdot 4\cdot 5} = 255,244,687,600 \quad \blacksquare$$

### 1.4.3. Factorial Function

In Eq. (1.26) - Eq. (1.31) the **factorial function** is relatively straightforward. By definition[17],

$$0! = 1.$$

Values may be computed recursively from given values by

$$(n+1)! = (n+1)n!.$$

For large $n$ the function is very large (see Table A3 in App. 5). A convenient approximation for large $n$ is the **Stirling formula**[2]

[17] This is done by convention. An intuitive way to look at it is $n!$ counts the number of ways to arrange distinct objects in a line, and there is only one way to arrange nothing

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \qquad (e = 2.718\cdots)$$

where $\sim$ is read asymptotically equal[18] and means that the ratio of the two sides of (7) approaches 1 as $n$ approaches infinity.

[18] it means the percentage difference between the vertical distances between points on the two graphs approaches 0.

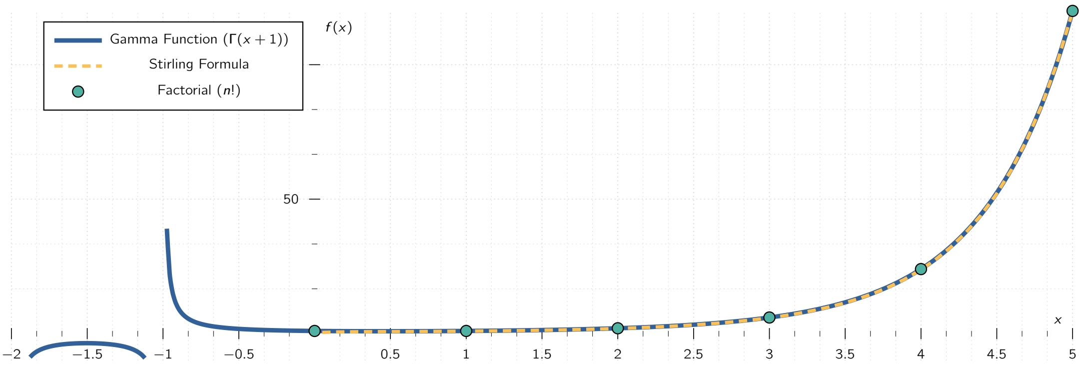


**Figure 1.3.:** A visual comparison of the Stirling formula and the actual values of the factorial function.

### 1.4.4. Binomial Coefficients

The **binomial coefficients** are defined by the following formula:

$$\binom{a}{k} = \frac{(a)(a-1)(a-2)\cdots(a-k+1)}{k!} \qquad \text{where} \qquad (k \geq 0, \text{integer}) \tag{1.32}$$

The numerator has $k$ factors. Furthermore, we define

$$\binom{a}{0} = 1, \qquad \text{in particular,} \qquad \binom{0}{0} = 1.$$

For integer $a = n$ we obtain from Eq. (1.32):

$$\binom{n}{k} = \binom{n}{n-k} \qquad (n \geqq 0, 0 \leqq k \leqq n).$$

Binomial coefficients may be computed recursively, because

$$\binom{a}{k} + \binom{a}{k+1} = \binom{a+1}{k+1} \qquad (k \geqq 0, \text{ integer}).$$

Formula Eq. (1.32) also gives:

$$\binom{-m}{k} = (-1)^k \binom{m+k-1}{k} \qquad \text{where} \qquad k \geqq 0, \text{ integer} \quad \text{and} \quad m > 0.$$

There are two (2) important relations worth mentioning:

$$\sum_{s=0}^{n-1} \binom{k+s}{k} = \binom{n+k}{k+1} \qquad (k \geq 0, n \geq 1,$$

and

$$\sum_{k=0}^{r} \binom{p}{k} \binom{q}{r-k} = \binom{p+q}{r} \qquad (r \geq 0, \text{ integer}).$$

## 1.5. Random Variables and Probability Distributions

In the beginning of this chapter we considered frequency distributions of data[19]. These distributions show the absolute or relative frequency of the data values.

[19] Remember we did a histogram and a stem-and-leaf plot.

Similarly, a **probability distribution** or, a **distribution**, shows the probabilities of events in an experiment. The quantity we observe in an experiment will be denoted by $X$ and called a **random variable**[20] as the value it will assume in the next trial depends on the **stochastic process**

[20] or **stochastic variable** if you want to be pedantic

> i.e., if you roll a die, you get one of the numbers from 1 to 6, but you don't know which one will show up next. An example would be $X$ =*Number a die turns up* which is a random variable.

If we count[21], we have a **discrete random variable and distribution**. If we *measure* (electric voltage, rainfall, hardness of steel), we have a **continuous random variable and distribution**.For both cases (discrete, discontinuous), the distribution of $X$ is determined by the **distribution function**:

[21] cars on a road, defective parts in a production, tosses until a die shows the first six (6)

$$F(x) = \Pr X \leq x \tag{1.33}$$

This is the probability that in a trial, $X$ will assume any value not exceeding $x$.

> The terminology is unfortunately not uniform across the field as $F(x)$ is sometimes also called the **cumulative distribution function**.

For Eq. (1.33) to make sense in both the discrete and the continuous case we formulate conditions as follows.

**Theory** 1.13: Random Variable

A **random variable** $X$ is a function defined on the sample space $S$ of an experiment. Its values are real numbers. For every number $a$ the probability

$$P(X = a)$$

with which $X$ assumes $a$ is defined. Similarly, for any interval $I$, the probability

$$P(X \in I)$$

with which $X$ assumes any value in $I$ is defined[22].

[22] Although this definition is very general, in practice only a very small number of distributions will occur over and over again in applications.

From Eq. (1.33) we can define the fundamental formula for the probability corresponding to an interval $a < x \leqq b$:

$$P\left(a < X \leqq b\right) = F(b) - F(a)\,. \tag{1.34}$$

This follows because $X \leqq a$ ($X$ assumes any value not exceeding $a$) and $a < X \leqq b$ ( $X$ assumes any value in the interval $a < x \leqq b$) are mutually exclusive events, so based on Eq. (1.33):

$$\begin{aligned} F(b) = P\left(X \leqq b\right) &= P\left(X \leqq a\right) + P\left(a < X \leqq b\right) \\ &= F(a) + P\left(a < X \leqq b\right) \end{aligned}$$

and subtraction of $F(a)$ on both sides gives Eq. (1.34).

### 1.5.1. Discrete Random Variables and Distributions

By definition, a random variable $X$ and its distribution are **discrete** if $X$ assumes only finitely many or at most countably many values $x_1, x_2, x_3, \cdots$, called the **possible values** of $X$, with positive probabilities,

$$p_1 = P\left(X = x_1\right), p_2 = P\left(X = x_2\right), \quad p_3 = P\left(X = x_3\right), \cdots$$

whereas the probability $P(X \in I)$ is zero for any interval $I$ containing no possible value. Clearly, the discrete distribution of X is also determined by the **probability function**$f(x)$ of $X$, defined by

$$f(x) = \begin{cases} p_{\mathrm{j}} & \text{if} \quad x = x_{\mathrm{j}} \\ 0 & \text{otherwise} \end{cases} \qquad \text{where} \qquad j = 1,\, 2, \cdots, \tag{1.35}$$

From this we get the values of the **distribution function** $F(x)$ by taking sums,

$$F(x) = \sum_{x_{\mathrm{j}} \leq x} f\left(x_{\mathrm{j}}\right) = \sum_{x_{\mathrm{j}} \leq x} p_{\mathrm{j}} \tag{1.36}$$

where for any given $x$ we sum all the probabilities $p_{\mathrm{j}}$ for which $x_{\mathrm{j}}$ is smaller than or equal to that of $x$. This is a **step function** with upward jumps of size $p_{\mathrm{j}}$ at the possible values $x_{\mathrm{j}}$ of $X$ and constant in between. The two (2) useful formulas for discrete distributions are readily obtained as follows. For the probability corresponding to intervals we have from Eq. (1.34) and Eq. (1.36):

$$P\left(a < X \leqq b\right) = F(b) - F(a) = \sum_{a < x_{\mathrm{j}} \leqq b} p_{\mathrm{j}} \tag{1.37}$$

This is the sum of all probabilities $p_{\mathrm{j}}$ for which $x_{\mathrm{j}}$ satisfies $a < x_{\mathrm{j}} \leqq b$. (Be careful about $<$ and $\leqq$ 1) From this and $P(S) = 1$ we obtain the following formula.

$$\sum_{j} p_{\mathrm{j}} = 1 \qquad \text{where} \qquad \text{(sum of all probabilities)}. \tag{1.38}$$

### 1.5.2. Continuous Random Variables and Distributions

Discrete random variables appear in experiments in which we count[23]. Continuous random variables appear in experiments in which we measure (lengths of screws, voltage in a power line, etc.). By definition, a random variable $X$ and its distribution are of *continuous type* or, briefly, **continuous**, if its distribution function $F(x)$, defined in Eq. (1.33), can be given by an integral[24]:

[23] defectives in a production, days of sunshine in Kufstein, customers in a line, etc.

[24] we write $v$ as a toss-away variable because $x$ is needed as the upper limit of the integral

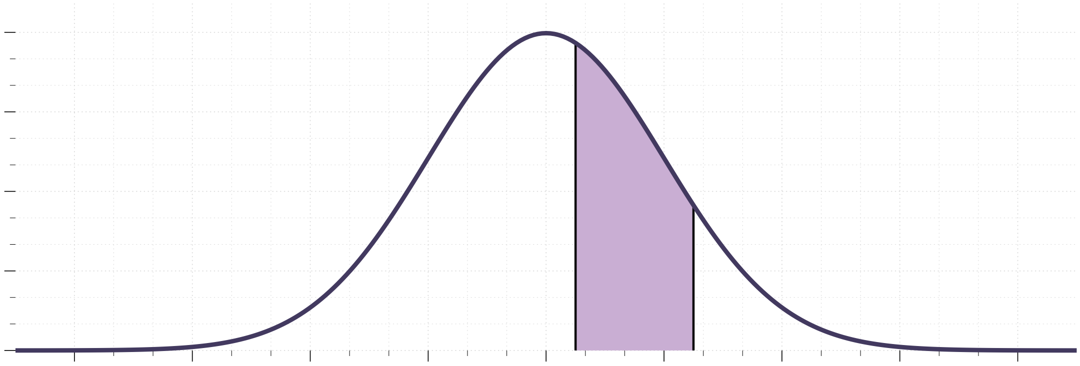

**Figure 1.4.:** A visual representation of the Eq. (1.41).

$$F(x) = \int_{-\infty}^{x} f(v)\, dv \tag{1.39}$$

whose integrand $f(x)$, called the **density** of the distribution, is non-negative, and is continuous, perhaps except for finitely many $x$-values. Differentiation gives the relation of $f$ to $F$ as

$$f(x) = F'(x) \tag{1.40}$$

for every $x$ at which $f(x)$ is continuous.

From Eq. (1.34) and Eq. (1.39) we obtain the very important formula for the probability corresponding to an interval[25]:

[25] This is an analog of Eq. (1.37)

$$P\left(a < X \leqq b\right) = F(b) - F(a) = \int_{a}^{b} f(v)\, dv \tag{1.41}$$

From Eq. (1.39) and $P(S) = 1$ we also have the analogue of Eq. (1.38):

$$\int_{-\infty}^{\infty} f(v)\, dv = 1. \tag{1.42}$$

> Continuous random variables are simple than discrete ones with respect to intervals as, in the continuous case the four probabilities corresponding to $a < X \leq b$, $a < X < b$, $a \leq X \leq b$, and $a \leq X \leq b$ with any fixed $a$ and $b$ $(> a)$ are all the same.

The next example illustrates notations and typical applications of our present formulas.

## 1.6. Mean and Variance of a Distribution

The mean $\mu$ and variance $\sigma^2$ of a random variable $X$ and of its distribution are the theoretical counterparts of the mean $\overline{x}$ and variance $s^2$ of a frequency distribution and serve a similar purpose.

The mean characterises the central location and the variance the spread (the variability) of the distribution. The **mean** $\mu$ is defined by:

$$
\begin{aligned}
&\text{(a)} \quad \mu = \sum_j x_j f\left(x_j\right) && \text{(Discrete distribution)} \\
&\text{(b)} \quad \mu = \int_{-\infty}^{\infty} x f(x)\, \mathrm{d}x && \text{(Continuous distribution)}
\end{aligned} \tag{1.43}
$$

and the **variance** $\sigma^2$ by:

$$
\begin{aligned}
&\text{(a)} \quad \sigma^2 = \sum_j (x_j - \mu)^2 f\left(x_j\right) && \text{(Discrete distribution)} \\
&\text{(b)} \quad \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\ \mathrm{d}x && \text{(Continuous distribution)}
\end{aligned} \tag{1.44}
$$

$\sigma$ (the positive square root of $\sigma^2$) is called the **standard deviation** of $X$ and its distribution. $f$ is the probability function or the density, respectively, in (a) and (b).

The mean $\mu$ is also denoted by $E(X)$ and is called the **expectation of** $X$ because it gives the average value of $X$ to be expected in many trials.

> Quantities such as $\mu$ and $\sigma^2$ that measure certain properties of a distribution are called **parameters**. $\mu$ and $\sigma^2$ are the two most important ones.

From Eq. (1.44) we see that[26]:

$$\sigma^2 > 0$$

We assume that $\mu$ and $\sigma^2$ exist[27], as is the case for practically all distributions that are useful in applications.

[26] except for a discrete "distribution" with only one possible value, so that $\sigma^2 = 0$

[27] and finite.

### Symmetry

We can obtain the mean $\mu$ without calculation if a distribution is symmetric. Indeed, we can write:

**Theory** 1.13: Mean of a Symmetric Distribution

If a distribution is **symmetric** with respect to $x = c$, that is,

$$f(c - x) = f(c + x)$$

then $\mu = c$.

## Transformation of Mean and Variance

Given a random variable $X$ with mean $\mu$ and variance $\sigma^2$, we want to calculate the mean and variance of $X^* = a_1 + a_2 X$, where $a_1$ and $a_2$ are given constants.

This problem is important in statistics, where it often appears.

**Theory** 1.13: Transformation of Mean and Variance

If a random variable X has mean $\mu$ and variance $\sigma^2$, then the random variable:

$$X^* = a_1 + a_2 X \qquad where a_2 > 0$$

has the mean $\mu^*$ and variance $\sigma^{*2}$, where

$$\mu^* = a_1 + a_2\mu \qquad \text{and} \qquad \sigma^{*2} = a_2^2\sigma^2.$$

In particular, the **standardised random variable** $Z$ corresponding to X, given by:

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

has the mean 0 and the variance 1.

## Expectation & Moments

If we recall, Eq. (1.43) defines the mean of $X$[28], written $\mu = E(X)$. More generally, if $g(x)$ is non-constant and continuous for all $x$, then $g(X)$ is a random variable. Therefore its mathematical expectation or, briefly, its expectation $E(g(X))$ is the value of $g(X)$ to be expected on the average, defined by:

[28] the value of $X$ to be expected on the average

$$E(g(X)) = \sum_j g\left(x_j\right) f\left(x_j\right) \qquad \text{or} \qquad E(g(X)) = \int_{-\infty}^{\infty} g(x) f(x) \, dx$$

In the formula on the Left Hand Side (LHS), $f$ is the probability function of the discrete random variable $X$. In the formula on the Right Hand Side (RHS), $f$ is the density of the continuous random variable $X$. Important special cases are the k$^{\text{th}}$ of $X$ (where $k = 1, 2, \cdots$)

$$E\left(X^k\right) = \sum_j x_j^k f\left(x_j\right) \qquad \text{or} \qquad \int_{-\infty}^{\infty} x^k f(x) \, dx$$

and the k$^{\text{th}}$ of $X$ ($k = 1, 2, \cdots$)

$$E\left([X - \mu]^k\right) = \sum_j \left(x_j - \mu\right)^k f\left(x_j\right) \qquad \text{or} \qquad \int_{-\infty}^{\infty} (x - \mu)^k f(x) \, dx.$$

This includes the first moment, the **mean** of $X$

$$\mu = E(X) \qquad \text{where k = 1} \tag{1.45}$$

It also includes the second central moment, the **variance** of $X$

$$\sigma^2 = E\left([X-\mu]^2\right) \qquad \text{where k = 2} \tag{1.46}$$

# Bibliography


[1] David C Hoaglin, Frederick Mosteller, and John W Tukey. *Understanding robust and exploratory data analysis*. Vol. 76. John Wiley & Sons, 2000.