Lecture Book
M.Sc Higher Mathematics II

D. T. McGuiness, Ph.D

Version: 2025.SS

Topics on Fundamental Science

Current version is 2025.SS.

This document includes the contents of Higher Mathematics II, official name being *Höhere Mathematik 2*, taught at MCI in the Mechatronik Smart Technologies. This document is the part of the module MECH-M-2-HMA-HMA-ILV taught in the M.Sc degree.

All relevant code of the document is done using *SageMath* where stated using v10.3 and Python v3.13.1.

This document was compiled with LuaTEX v1.18.0, and all editing were done using GNU Emacs v29.4 using AUCTEX and org-mode package.

This document is based on the books and resources: *G. B. Thomas, Jr. et.al* by Thomas Calculus (12th Edition) , Pearson (2010) *A. Gut* by Probability: A Graduate Course , Springer (2005) *S.L. Sobolev* by Partial Differential Equations of Mathematical Physics , Addison Wesley (2019) *W. A. Strauss* by Partial Differential Equations - An Introduction , Wiley (2008) *R. E. Walpole, et. al* by Probability and Statistics for Engineers & Scientists , Pearson (2012) *M. L. Boas* by Mathematical Methods in the Physical Sciences (3rd Edition) , Wiley (2006) *K. F. Riley, et. al* by Mathematical Methods for Physics and Engineering (3rd Edition) , Cambridge (2006) *G. F. Simmons* by Differential Equations with Applications and Historical Notes (3rd Edition) , CRC Press (2017) *E. Kreyszig* by Advanced Engineering Calculus (9th Edition) , Wiley (2011) *D. C. Montgomery* by Applied Statistics and Probability for Engineers (3rd Edition) , Wiley (2003) *J. F. James* by A Students Guide to Fourier Transform , Cambridge (2011) *J. Crank* by Mathematics of Diffusion , Oxford (1975) *A. Sommerfeld* by Partial Differential Equations in Physics , Academic Press (1949) *E. Cinlar* by Probability and Stochastics , Springer (2010) *H. C. Berg* by Random Walks in Biology , Princeton (1983)

The current maintainer of this work along with the primary lecturer
is D. T. McGuines, Ph.D. (dtm@mci4me.at).

Contents

I. Probability & Statistics 1

1. Theory of Probability 3

1.1. Introduction . . . 3
1.2. Experiments & Outcomes . . . 6
1.2.1. Unions, Intersections, and Complements of Events . . . 7
1.3. Probability . . . 8
1.4. Permutations & Combinations . . . 13
1.4.1. Permutations . . . 13
1.4.2. Combinations . . . 15
1.4.3. Factorial Function . . . 15
1.4.4. Binomial Coefficients . . . 16
1.5. Random Variables and Probability Distributions . . . 17
1.5.1. Discrete Random Variables and Distributions . . . 18
1.5.2. Continuous Random Variables and Distributions . . . 19
1.6. Mean and Variance of a Distribution . . . 21
1.7. Binomial, Poisson, and Hyper-geometric Distributions . . . 23
1.7.1. Sampling with Replacement . . . 26
1.7.2. Sampling without Replacement: Hyper-geometric Distribution . . . 26
1.8. Normal Distribution . . . 27
1.8.1. Distribution Function . . . 28
1.8.2. Numeric Values . . . 28
1.8.3. Normal Approximation of the Binomial Distribution . . . 29
1.9. Distribution of Several Random Variables . . . 30
1.9.1. Discrete Two-Dimensional Distribution . . . 30
1.9.2. Continuous Two-Dimensional Distribution . . . 31
1.9.3. Marginal Distributions of a Discrete Distribution . . . 31
1.9.4. Marginal Distributions of a Continuous Distribution . . . 32
1.9.5. Independence of Random Variables . . . 33
1.9.6. Functions of Random Variables . . . 34
1.9.7. Addition of Means . . . 34
1.9.8. Addition of Variances . . . 35

2. Statistical Methods 37

2.1. Introduction . . . 37

2.2. Point Estimation of Parameters 39
2.2.1. Maximum Likelihood Method 40
2.3. Confidence Intervals 41
2.3.1. Confidence Interval for Mean with known Variance in Normal Distribution . 42

Bibliography **47**

List of Figures

1.1. The histogram of the data given in **Exercise 1**. 5
1.2. Examples of Venn diagrams. 8
1.3. A visual comparison of the Stirling formula and the actual values of the factorial function. 16
1.4. A visual representation of the Eq. (1.42). 20
1.5. The Poisson distribution with different mean (μ) values. 25
1.6. The poster child of probability and statistics, the normal distribution. 27
1.7. A visual representation between the relationship of PDF and CDF. 28
1.8. Many samples from a bivariate normal distribution. The marginal distributions are shown on the z-axis. The marginal distribution of X is also approximated by creating a histogram of the X coordinates without consideration of the Y coordinates. . . 31

List of Tables

1.1. Stem and Leaf plot of the data given in Example 1. 4

List of Examples

1.1. Recording Data 4
1.2. Leaf Plots 4
1.3. Histogram 4
1.4. Empirical Rule, Outliers, and z-Score 6
1.5. Sample Spaces of Random Experiments & Events 7
1.6. Fair Die 9
1.7. Coin Tossing 10
1.8. Mutually Exclusive Events 10
1.9. Union of Arbitrary Events 11
1.10. Multiplication Rule 12
1.11. Sampling w/o Replacement 13
1.12. An Encrypted Message 14
1.13. Sampling Light-bulbs 15
1.14. Waiting Time Problem 19
1.15. Continuous Distribution 20
1.16. Mean and Variance 21
1.17. Binomial Distribution 24
1.18. Poisson Distribution 25
1.19. The Parking Problem 25
1.20. Marginal Distributions of a Discrete Two-Dimensional Random Variable 32
1.21. Independence and Dependence 33

2.1. Generating Random Numbers 38
2.2. Maximum Likelihood of Gaussian Distribution 41
2.3. Confidence Interval for mean with known variance in Normal Distribution 43

List of Theorems

1.1. First Definition of Probability 9
1.2. General Definition of Probability 9
1.3. Complementation Rule 10
1.4. Addition Rule for Mutually Exclusive Events 10
1.5. Addition Rule for Arbitrary Events 11
1.6. Multiplication Rule 11
1.7. Permutations 14
1.8. Permutations 14
1.9. Combinations 15
1.10. Random Variable 17
1.11. Mean of a Symmetric Distribution 22
1.12. Transformation of Mean and Variance 22
1.13. Relationship between PDF and CDF 28

1.14. Normal Probabilities for Intervals 28
1.15. Limit Theorem of De Moivre and Laplace 29
1.16. Addition of Means 35
1.17. Multiplication of Means 35
1.18. Addition of Variances 36

Part I.

Probability & Statistics

Chapter 1

Theory of Probability

Table of Contents

1.1. Introduction . . . 3
1.2. Experiments & Outcomes . . . 6
1.3. Probability . . . 8
1.4. Permutations & Combinations . . . 13
1.5. Random Variables and Probability Distributions . . . 17
1.6. Mean and Variance of a Distribution . . . 21
1.7. Binomial, Poisson, and Hyper-geometric Distributions . . . 23
1.8. Normal Distribution . . . 27
1.9. Distribution of Several Random Variables . . . 30

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 1000° C), 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 **Table** 1.1.

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. ■

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.

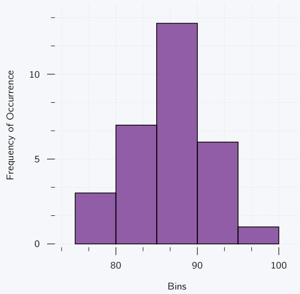

Figure 1.1.: The histogram of the data given in **Exercise 1**.

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]

[6] also known as variability.

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

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

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.

[8]which we calculated previously

Returning back to our super alloy example, using $\overline{x} = 260/3$[8], 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 calculated to be:

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

The standard deviation has the same dimension as the data values, which is an advantage, whereas, 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. This information is quite useful in doing quick calculation of statistical properties such as the quality of production which will be the focus in Chapter 2.

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

For the data set given in Example 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 of 99, mean and standard deviation reduce to $\overline{x}_{red} = 86.2$, and $s_{red} = 4.3$, 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

[9]Sometimes known as probability calculus.

Now we have the basis covered, it is time to look at **probability theory**[9]. This theory has the purpose of providing mathematical models of situations affected or even governed by change effects, 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[10]

sample space Defined as S, is the set of all possible outcomes of an experiment.

[10] 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[11] $A, B, C, \cdots$ of a given sample space S.

[11] 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[12] . 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, or a coin cannot turn up Head (H) and Tail (T) at the same time.

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

- The **Complement** of A is A^c[13] . This is the set of all the points of S *not* in A. Therefore,

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

[13] 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**:

$$\bigcup_{j=1}^{m} A_{\mathrm{j}} = A_1 \cup A_2 \cup \cdots \cup A_{\mathrm{m}}.$$

of events $A_1, \cdots, A_{\mathrm{m}}$ 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**:

$$\bigcap_{j=1}^{m} A_{\mathrm{j}} = A_1 \cap A_2 \cap \cdots \cap A_{\mathrm{m}}$$

of $A_1, \cdots, A_{\mathrm{m}}$ 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 expressing the concept covered previously.

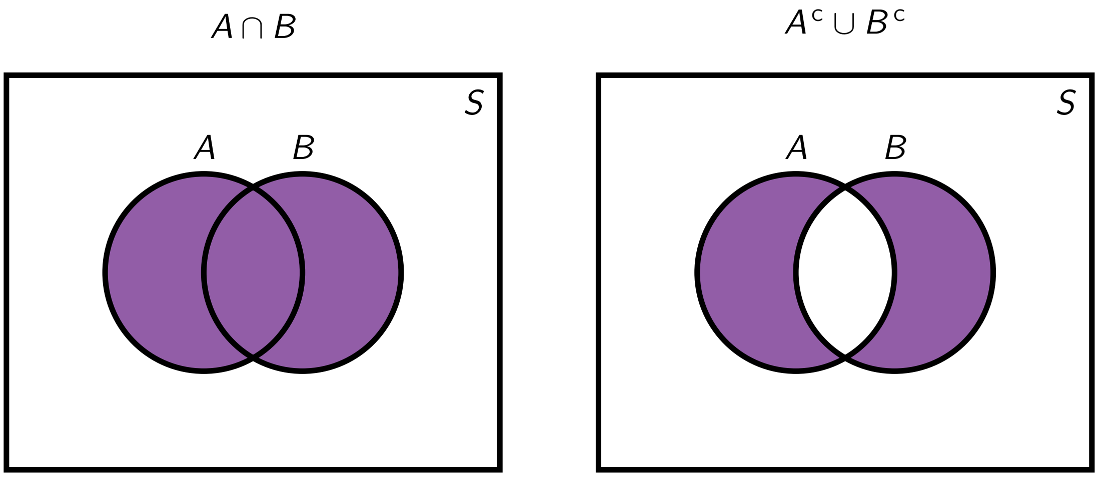

Figure 1.2.: Examples of Venn diagrams.

1.3. Probability

The probability of an event A in an experiment is to measure **how frequently** A is roughly to occur if we make many trials. If we flip a coin, then heads H and tails T will appear about equally[14] often.

[14] on the condition, the measurements are done for a long time.

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

Similarly, for a regularly shaped die of homogeneous material[15] each of the six (6) 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.

[15] called a fair dice

Let's formulate this in a theory.

Theory 1.1: First Definition of Probability

If the sample space S of an experiment consists of **finitely** many outcomes (points) being equally likely, 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, in particular, the probability of all events occurring in the sample space S is:

$$P(S) = 1.$$

Exercise 1.6: Fair Die

In rolling a fair die once:

1. What is the probability $P(A)$ of A of obtaining a 5 or a 6?
2. The probability of B: "Even number"?

Solution

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

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

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[16] . Division by n gives:

[16] 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 1.2: 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 $P(A)$, called the **probability** of A, such the following **axioms of probability** are satisfied.

- For every A in S,
$$0 \leqq P(A) \leqq 1. \tag{1.8}$$
- 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[17], the previous statement has to be replaced by Eq. (1.4), where for mutually exclusive events $A_1, A_2, \cdots$,

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

[17] i.e., has infinitely many points.

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

Theory 1.3: Complementation Rule

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

$$P\left(A^c\right) = 1 - P(A) \tag{1.12}$$

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 $P\left(A^c\right) = 1/32$, and the answer is:

$$P(A) = 1 - P\left(A^c\right) = \frac{31}{32} \quad \blacksquare$$

Theory 1.4: Addition Rule for Mutually Exclusive Events

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

$$P\left(A_1 \cup A_2 \cup \cdots A_m\right) = P\left(A_1\right) + P\left(A_2\right) + \cdots + P\left(A_m\right). \tag{1.13}$$

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 to formalise the previous statement.

Theory 1.5: Addition Rule for Arbitrary Events

For events A and B in a sample space, their union is defined as:

$$P(A \cup B) = P(A) + P(A) - P(A \cap B)\,. \tag{1.14}$$

For mutually exclusive events A and B we have $A \cap B = \emptyset$ by definition:

$$P(\emptyset) = 0 \tag{1.15}$$

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:

$$P(A \cup B) = \frac{3}{6} + \frac{3}{6} - \frac{2}{6} = \frac{2}{3}$$

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 $P(B|A)$.

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

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

Similarly, the conditional probability of A given B is:

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

Theory 1.6: Multiplication Rule

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

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

Exercise 1.10: Multiplication Rule

In producing screws, let:

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

Let $P(A) = 0.1$ 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

$$P(A \cap B) = P(A)\, P\left(B|A\right) = 0.1 \times 0.2 = 0.02 = 2\% \quad \blacksquare$$

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):

$$P\left(A|B\right) = P(A)\,, \qquad P\left(B|A\right) = P(B)\,.$$

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 $A_{j1}, A_{j2}, \cdots, A_{jk}$.

$$P\left(A_{j_1} \cap A_{j_2} \cap \cdots \cap A_{j_k}\right) = P\left(A_{j1}\right) P\left(A_{j2}\right) \cdots P\left(A_{jk}\right) \tag{1.21}$$

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 $P(A) = a/k$ by systematically counting the number a of points of which an event A consists.

where, 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[18] the probability of obtaining them in the required order is only 1/3,628,800 because there are exactly:

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

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

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

1.4.1. Permutations

A **permutation** of given things[19] is an arrangement of these things in a row in some order.

[19] 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 1.7: 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

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

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 1.8: 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 (5) 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:

1. 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 (2) sets containing exactly the same k things.

2. 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 (3) combinations of the three (3) letters a, b, c, taken two (2) letters at a time, without repetitions, namely, ab, ac, bc, and six such combinations with repetitions, namely, ab, ac, bc, ca, bb, cc.

Theory 1.9: 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 (5) 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[20] ,

$$0! = 1.$$

Values may be computed recursively from given values by

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

[20] 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.

For large n the function is very large and hard to keep track of. A convenient approximation for large n is the **Stirling formula**, defined as:

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \qquad \text{where} \qquad e = 2.718\cdots \tag{1.32}$$

where $\sim$ is read asymptotically equal[21] and means that the ratio of the two sides of Eq. (1.32) approaches 1 as n approaches infinity.

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

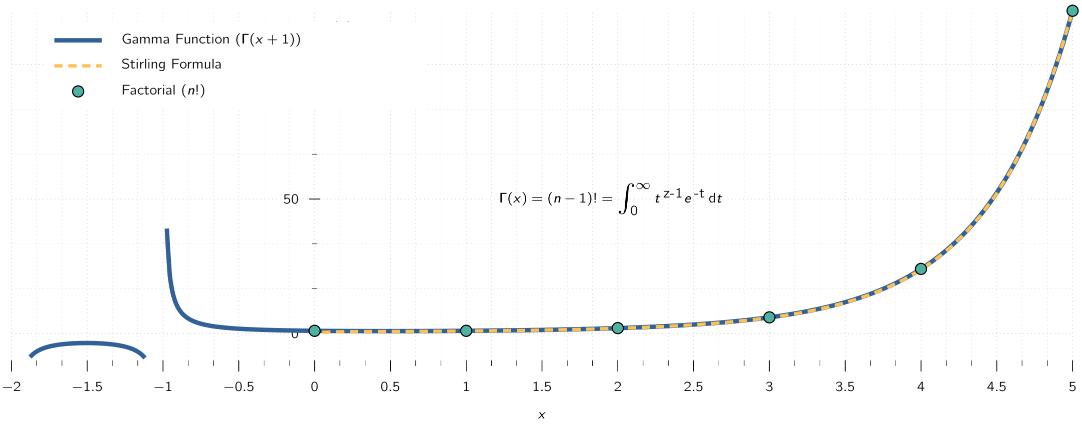

Figure 1.3.: A visual comparison of the Stirling formula and the actual values of the factorial function.

1.4.4. Binomial Coefficients

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

$$\binom{a}{k} = \frac{(a)\,(a-1)\,(a-2)\cdots(a-k+1)}{k!} \qquad \text{where} \qquad (k \geq 0, \text{integer}) \tag{1.33}$$

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. (1.33):

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

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. (1.33) 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:

$$\sum_{s=0}^{n-1} \binom{k+s}{k} = \binom{n+k}{k+1} \qquad (k \geq 0 \quad \text{and} \quad n \geq 1)$$

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[22] . These distributions show the absolute or relative frequency of the data values.

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**[23] as the value it will assume in the next trial depends on the **stochastic process**

> 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[24] , 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**:

$$F(x) = P(X \leq x) \tag{1.34}$$

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

Theory 1.10: 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[25] .

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

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

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

[25] 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. (1.34) we can define the fundamental formula for the probability corresponding to an interval $a < x \leqq b$:

$$P\left(a < X \leqq b\right) = F\left(b\right) - F\left(a\right). \tag{1.35}$$

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. (1.34):

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

and subtraction of $F\left(a\right)$ on both sides gives Eq. (1.35).

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

$$f\left(x\right) = \begin{cases} p_{\text{j}} & \text{if} \quad x = x_{\text{j}} \\ 0 & \text{otherwise} \end{cases} \qquad \text{where} \qquad j = 1, 2, \cdots, \tag{1.36}$$

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

$$F\left(x\right) = \sum_{x_{\text{j}} \leq x} f\left(x_{\text{j}}\right) = \sum_{x_{\text{j}} \leq x} p_{\text{j}} \tag{1.37}$$

where for any given x we sum all the probabilities p_{j} for which x_{j} is smaller than or equal to that of x. This is a **step function** with upward jumps of size p_{j} at the possible values x_{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. (1.35) and Eq. (1.37):

$$P\left(a < X \leqq b\right) = F\left(b\right) - F\left(a\right) = \sum_{a < x_{\text{j}} \leqq b} p_{\text{j}} \tag{1.38}$$

This is the sum of all probabilities p_{j} for which x_{j} satisfies $a < x_{\text{j}} \leqq b$[26]. From this and $P\left(S\right) = 1$ we obtain the following formula.

[26] Be careful about $<$ and $\leqq$ as the former means it is **NOT** included and the latter means it is.

$$\sum_{j} p_{\text{j}} = 1 \qquad \text{(sum of all probabilities)}. \tag{1.39}$$

Exercise 1.14: Waiting Time Problem

In tossing a fair coin, let X be the Number of trials until the first head appears. Then, by independence of events we get (where H is heads, and T is tails):

$$\begin{aligned} P(X=1) &= P(H) &&= \frac{1}{2} \\ P(X=2) &= P(TH) &&= \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4} \\ P(X=3) &= P(TTH) &&= \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{8} \end{aligned}$$

and in general, $P(X=n) = \left(\frac{1}{2}\right)^2$, $n = 1, 2, 3, \ldots$ which when all possible event are summed up will always give 1.

1.5.2. Continuous Random Variables and Distributions

Discrete random variables appear in experiments in which we count[27] . 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. (1.34), can be given by an integral[28] :

$$F(x) = \int_{-\infty}^{x} f(\upsilon)\, d\upsilon \tag{1.40}$$

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

$$f(x) = F'(x) \tag{1.41}$$

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

From Eq. (1.35) and Eq. (1.40) we obtain the very important formula for the probability corresponding to an interval[29] :

$$P(a < X \leq b) = F(b) - F(a) = \int_{a}^{b} f(\upsilon)\, d\upsilon \tag{1.42}$$

Which can be seen visually in **Fig.** 1.4. From Eq. (1.40) and $P(S) = 1$ we also have the analogue of Eq. (1.39):

$$\int_{-\infty}^{\infty} f(\upsilon)\, d\upsilon = 1. \tag{1.43}$$

Continuous random variables are simpler 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.

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

[28] we write υ as a toss-away variable because x is needed as the upper limit of the integral.

[29] This is an analog of Eq. (1.38)

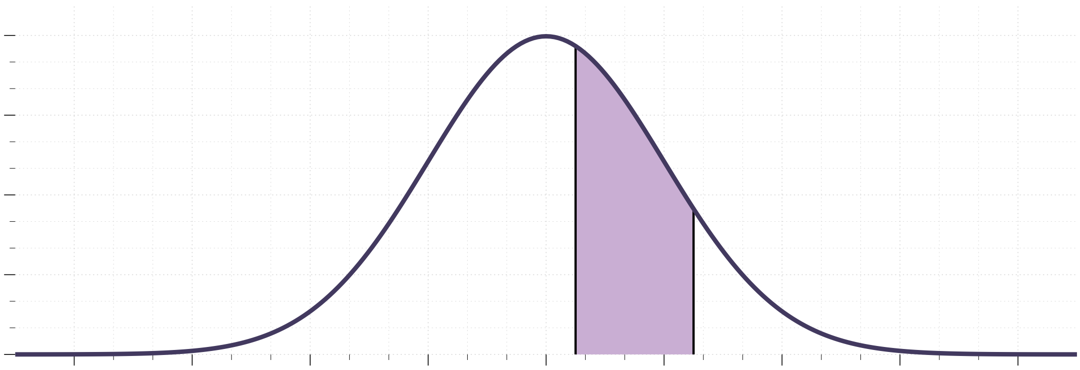

Figure 1.4.: A visual representation of the Eq. (1.42).

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

Exercise 1.15: Continuous Distribution

Let X have the density function:

$$f(x) = 0.75(1 - x^2) \qquad \text{if} \qquad -1 \leq x \leq 1,$$

and zero otherwise. Find:

1. The distribution function.
2. Find the probabilities $P\left(-\frac{1}{2} \leq X \leq \frac{1}{2}\right)$ and $P\left(\frac{1}{2} \leq X \leq 2\right)$
3. Find x such that $P(X \leq x) = 0.95$.

Solution

From Eq. (1.40), we obtain $F(x) = 0$ if $x \leqq -1$,

$$F(x) = 0.75 \int_{-1}^{x} (1 - v^2)\, dv = 0.5 + 0.75x - 0.25x^3 \qquad \text{if} \qquad -1 < x \leqq 1,$$

and $F(x) = 1$ if $x > 1$. From this and Eq. (1.42) we get:

$$P(-\tfrac{1}{2} \leq X \leq \tfrac{1}{2}) = F(\tfrac{1}{2}) - F(-\tfrac{1}{2}) = 0.75 \int_{-1/2}^{1/2} (1 - v^2)\, dv = 68.75\%$$

because $P(-\frac{1}{2} \leq X \leq \frac{1}{2}) = P(-\frac{1}{2} < X \leq \frac{1}{2})$ for a continuous distribution we can write:

$$P(\frac{1}{4} \leq X \leq 2) = F(2) - F(\frac{1}{4}) = 0.75 \int_{1/4}^{1} (1 - v^2)\, dv = 31.64\%.$$

Note that the upper limit of integration is 1, not 2. Finally,

$$P(X \leq x) = F(x) = 0.5 + 0.75x - 0.25x^2 = 0.95.$$

Algebraic simplification gives $3x - x^3 = 1.8$. A solution is $x = 0.73$, approximately ■

1.6. Mean and Variance of a Distribution

The mean μ and variance σ^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** μ is defined by:

$$\text{(a)} \quad \mu = \sum_j x_j f\left(x_j\right) \qquad \text{(Discrete distribution)} \qquad (1.44a)$$

$$\text{(b)} \quad \mu = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x \qquad \text{(Continuous distribution)} \qquad (1.44b)$$

and the **variance** σ^2 by:

$$\text{(a)} \quad \sigma^2 = \sum_j (x_j - \mu)^2 f\left(x_j\right) \qquad \text{(Discrete distribution)} \qquad (1.45a)$$

$$\text{(b)} \quad \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,\mathrm{d}x \qquad \text{(Continuous distribution)} \qquad (1.45b)$$

σ (the positive square root of σ^2) is called the **standard deviation**[30] of X and its distribution. f is the probability function or the density, respectively, in (a) and (b).

[30] Sometimes it is known as $\mathrm{Var}(x)$

The mean μ 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 μ and σ^2 that measure certain properties of a distribution are called **parameters**. μ and σ^2 are the two (2) most important ones.

From Eq. (1.45a) and Eq. (1.45b), we see that[31] :

$$\sigma^2 > 0$$

[31] except for a discrete distribution with only one possible value.

We assume that μ and σ^2 exist[32] , as is the case for practically all distributions that are useful in applications.

[32] and finite.

Exercise 1.16: Mean and Variance

The random variable X, *Number of heads in a single toss of a fair coin*, has the possible values $X = 0$ and $X = 1$ with probabilities $P(X=0) = \frac{1}{2}$ and $P(X=1) = \frac{1}{2}$. From Eq. (1.44a) we thus obtain the mean:

$$\mu = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2},$$

and Eq. (1.45a) gives the variance:

$$\sigma^2 = (0 - \frac{1}{2})^2 \cdot \frac{1}{2} + (1 - \frac{1}{2})^2 \cdot \frac{1}{2} = \frac{1}{4} \quad \blacksquare$$

Symmetry

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

Theory 1.11: 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 μ and variance σ^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 1.12: Transformation of Mean and Variance

If a random variable X has mean μ and variance σ^2, then the random variable:

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

has the mean μ^* and variance σ^{*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

[33]the value of X to be expected on the average

If we recall, Eq. (1.44a) and Eq. (1.44b) define the mean of X[33], 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:

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

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

$$\mu = E(X) \qquad \text{where} \qquad k = 1 \tag{1.46}$$

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

$$\sigma^2 = E\left([X-\mu]^2\right) \qquad \text{where} \qquad k = 2. \tag{1.47}$$

1.7. Binomial, Poisson, and Hyper-geometric Distributions

These are the three (3)most important discrete distributions, with numerous applications therefore are worth of a bit of a detailed look.

> Of course these are not the only distributions present. There are as many distributions as there are problems with some distributions used in wide variety of fields (Gaussian) whereas some are used only in a very narrow field (Nakagami).

Binomial Distribution

The **binomial distribution** occurs in problems involving of chance[34] .

What we are interested is in the number of times an event A occurs in n independent trials. In each trial, the event A has the same probability $P(A) = p$. Then in a trial, A will **NOT** occur with probability $q = 1 - p$. In n trials the random variable that interests us is:

$$X = \text{Number of times the event A occurs in n trials.} \tag{1.48}$$

X can assume the values $0, 1, \cdots, n$, and we want to determine the corresponding probabilities. Now $X = x$ means that A occurs in x trials and in $n - x$ trials it does not occur. We can write this down as follows:

$$\underbrace{A \;\; A \;\; \cdots A}_{x \text{ times}} \qquad \text{and} \qquad \underbrace{B \;\; B \;\; \cdots B}_{n \text{ - } x\text{times}} \tag{1.49}$$

Here $B = A^c$ is the complement of A, meaning that A does not occur. We now use the assumption that the trials are independent[35] . Hence Eq. (1.49) has the probability:

$$\underbrace{p \;\; p \;\; \cdots p}_{x \text{ times}} \cdot \underbrace{q \;\; q \;\; \cdots q}_{n \text{ - } \; x\text{times}} = p^x q^{n\text{-}x} \tag{1.50}$$

Now Eq. (1.49) is just one order of arranging xA's and $n - xB$'s. We will now calculate the number of permutations of n things[36] consisting of two (2)classes;

[34] rolling a dice, quality inspection (e.g., counting of the number of defectives), opinion plots (counting number of employees favouring certain schedule changes, etc.), medicine (e.g., recording the number of patterns who covered on a new medication)

[35] e.g., they do **NOT** influence each other

[36] the n outcomes of the n trials

1. class 1 containing the $n_1 = x$ A's
2. class 2 containing the $n - n_1 = n - x$ B's

This number is:

$$\frac{n!}{x!(n-x)!} = \binom{n}{x}. \tag{1.51}$$

Accordingly, Eq. (1.50), multiplied by this binomial coefficient, gives the probability $P(X = x)$ of $X = x$, that is, of obtaining A precisely x times in n trials. Hence X has the probability function:

$$f(x) = \binom{n}{x} p^x q^{n-x} \qquad (x = 0, 1, \cdots, n) \tag{1.52}$$

and $f(x) = 0$ otherwise. The distribution of X with probability function (2) is called the **binomial distribution** or *Bernoulli distribution*. The occurrence of A is called *success*[37] and the non-occurrence of A is called *failure*.

[37] regardless of what it actually is; it may mean that you miss your plane or lose your watch

The mean of the binomial distribution is:

$$\mu = np$$

and the variance is:

$$\sigma^2 = npq.$$

For the *symmetric case* of equal chance of success and failure ($p = q = \frac{1}{2}$) this gives the mean $n/2$, the variance $n/4$, and the probability function

$$f(x) = \binom{n}{x}\left(\frac{1}{2}\right) \qquad (x = 0, 1, \cdots, n).$$

Exercise 1.17: Binomial Distribution

Calculate the probability of obtaining at least two (2) "six" in rolling a fair die 4 times.

Solution

$p = P(A) = P(\text{six}) = \frac{1}{6}$, $q = \frac{5}{6}$, $n = 4$. The event "At least two (2) "six" occurs if we obtain 2 or 3 or 4 "six" Hence the answer is:

$$P = f(2) + f(3) + f(4) = \binom{4}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^2 + \binom{4}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right) + \binom{4}{4}\left(\frac{1}{6}\right)^4$$
$$= \frac{1}{6^4}(6 \cdot 25 + 4 \cdot 5 + 1) = \frac{171}{1296} = 13.2\%.$$

Poisson Distribution

The discrete distribution with infinitely many possible values and probability function:

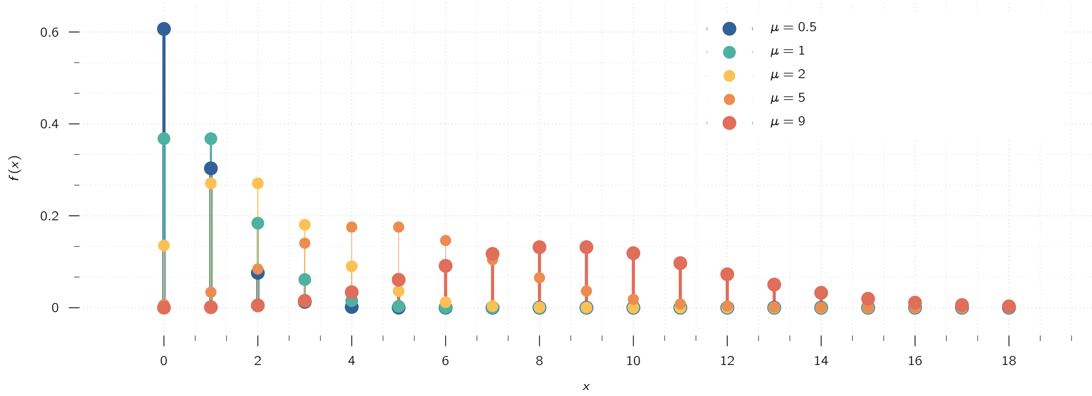

Figure 1.5.: The Poisson distribution with different mean (μ) values.

$$f(x) = \frac{\mu^x}{x!} e^{-\mu} \qquad \text{where} \qquad x = 0,\ 1, \cdots \tag{1.53}$$

is called the **Poisson distribution**, named after *S. D. Poisson*. **Fig.** 1.5 shows Eq. (1.53) for some values of μ[38] .

[38] While μ is used here, some textbook use λ

It can be proved that this distribution is obtained as a limiting case of the binomial distribution, if we let $p \to 0$ and $n \to \infty$ so that the mean $\mu = np$ approaches a finite value. The Poisson distribution has the mean μ and the variance:

$$\sigma^2 = \mu. \tag{1.54}$$

Fig. 1.5 gives the impression that, with increasing mean, the spread of the distribution increases, thereby illustrating formula Eq. (1.54), and that the distribution becomes more and more symmetric[39] .

[39] approximately

Exercise 1.18: Poisson Distribution

If the probability of producing a defective screw is $p = 0.01$, what is the probability that a lot of 100 screws will contain more than 2 defectives?

Solution

The complementary event is A^c. No more than 2 defectives. For its probability we get, from the binomial distribution with mean $\mu = np = 1$, the value.

$$P\left(A^c\right) = \binom{100}{0} 0.99^{100} + \binom{100}{1} 0.01 \cdot 0.99^{100} + \binom{100}{2} 0.01^2 \cdot 0.99^{100}.$$

Since p is very small, we can approximate this by the much more convenient Poisson distribution with mean $\mu = np = 100 \cdot 0.01 = 1$, obtaining.

$$P\left(A^c\right) = e^{-1}\left(1 + 1 + \frac{1}{2}\right) = 91.97\%.$$

Thus $P(A) = 8.03\%$. Show that the binomial distribution gives $P(A) = 7.94\%$, so that the Poisson approximation is quite good ■

Exercise 1.19: The Parking Problem

If on the average, 2 cars enter a certain parking lot per minute, what is the probability that during any given minute four (4) or more cars will enter the lot?

Solution

To understand that the Poisson distribution is a model of the situation, we imagine the minute to be divided into very many short time intervals. Let p be the (constant) probability that a car will enter the lot during any such short interval, and assume independence of the events that happen during those intervals. Then, we are dealing with a binomial distribution with very large n and very small p, which we can approximate by the Poisson distribution with

$$\mu = np = 2$$

because 2 cars enter on the average, the complementary event of the event "*4 cars or more during a given minute*" is "*3 cars or fewer enter the lot*" and has the probability

$$f(0) + f(1) + f(2) + f(3) = e^{-2}\left(\frac{2^0}{0} + \frac{2^1}{1!} + \frac{2^2}{2!} + \frac{2^3}{3!}\right) = 0.857.$$

Which means the result is 14.3% ■

1.7.1. Sampling with Replacement

This means that we draw things from a given set one by one, and after each trial we replace the thing drawn[40] before we draw the next thing. This guarantees independence of trials and leads to the **binomial distribution**. Indeed, if a box contains N things, for example, screws, M of which are defective, the probability of drawing a defective screw in a trial is $p = M/N$. Hence the probability of drawing a nondefective screw is $q = 1 - p = 1 - M/N$, and Eq. (1.52) gives the probability of drawing x defectives in n trials in the form:

[40] put it back to the given set and mix.

$$f(x) = \binom{M}{x}\left(\frac{M}{N}\right)^x\left(1 - \frac{M}{N}\right)^{n-x} \qquad (x = 0, 1, \cdots, n). \tag{1.55}$$

1.7.2. Sampling without Replacement: Hyper-geometric Distribution

Sampling without replacement means that we return no screw to the box. Then we no longer have independence of trials, and instead of Eq. (1.55) the probability of drawing x defectives in n trials is:

$$f(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \qquad \text{where} \qquad x = 1, 2, \cdots, n \tag{1.56}$$

The distribution with this probability function is called the **hyper-geometric distribution**[41] .

[41] because its moment generating function can be expressed by the hypergeometric function, which is a fact only useful to write it in a margin.

The hypergeometric distribution has the mean:

$$\mu = n\frac{M}{N},$$

and the variance

$$\sigma^2 = \frac{nM(N-M)(N-n)}{N^2(N-1)}.$$

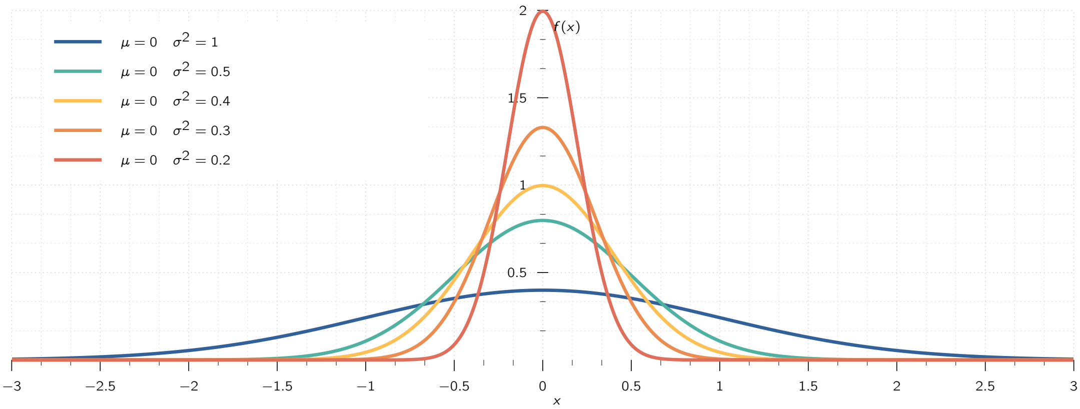

Figure 1.6.: The poster child of probability and statistics, the normal distribution.

1.8. Normal Distribution

Turning from discrete to continuous distributions, in this section we discuss the normal distribution. This is the most important continuous distribution because in applications many random variables are **normal random variables**[42] or they are approximately normal or can be transformed into normal random variables in a relatively simple fashion. Furthermore, the normal distribution is a useful approximation of more complicated distributions, and it also occurs in the proofs of various statistical tests.

[42]that is, they have a normal distribution.

The **normal distribution** or *Gauss distribution* is defined as the distribution with the density:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{1.57}$$

where exp is the exponential function with base $e = 2.718\cdots$. This is simpler than it may at first look. $f(x)$ has these features (see also **Fig.** 1.6).

1. μ is the mean, and σ the standard deviation.
2. $1/(\sigma\sqrt{2\pi})$ is a constant factor that makes the area under the curve of $f(x)$ from $-\infty$ to ∞ equal to 1, as it must be[43] .
3. The curve of $f(x)$ is symmetric with respect to $x = \mu$ because the exponent is quadratic. Hence for $\mu = 0$ it is symmetric with respect to the y-axis $x = 0$[44] .
4. The exponential function in Eq. (1.57) goes to zero very fast–the faster the smaller the standard deviation σ is, as it should be, as seen in **Fig.** 1.6.

[43]Having a probability higher than 1 does NOT make sense

[44]This distribution is also known as bell-shaped curves.

1.8.1. Distribution Function

From Eq. (1.55) and Eq. (1.57) we see that the normal distribution has the **distribution function** of the following form:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{\upsilon-\mu}{\sigma}\right)^2\right] d\upsilon. \tag{1.58}$$

Here we needed x as the upper limit of integration and wrote υ (instead of x) in the integrand.

For the corresponding **standardised normal distribution** with mean 0 and standard deviation 1 we denote $F(x)$ by $\Phi(z)$. Then we simply have from Eq. (1.58).

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2} du. \tag{1.59}$$

> This integral cannot be integrated by one of the methods of calculus.

But this is no serious handicap because its values can be obtained from standardised tables. These values are needed in working with the normal distribution. The curve of $\Phi(z)$ is S-shaped. It increases monotone from 0 to 1 and intersects the vertical axis at $\frac{1}{2}$, as shown in **Fig.** 1.7.

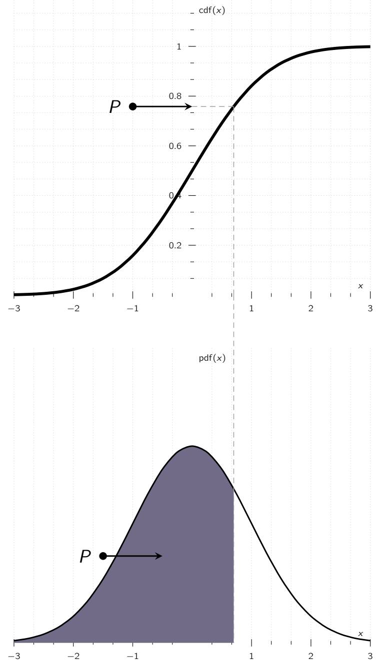

Figure 1.7.: A visual representation between the relationship of PDF and CDF.

Theory 1.13: Relationship between PDF and CDF

The distribution function $F(x)$ of the normal distribution with any μ and σ is related to the standardised distribution function $\Phi(z)$ in Eq. (1.59) by the formula

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

Theory 1.14: Normal Probabilities for Intervals

The probability a normal random variable X with mean μ and standard deviation σ assume any value in an interval $a < x \equiv b$ is:

$$P(a < X \leq b) = F(b) - F(a) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

1.8.2. Numeric Values

In practical work with the normal distribution it is good to remember that about 67% of all values of X to be observed will be between $\mu \pm \sigma$, about 95% between $\mu \pm 2\sigma$, and practically all between the **three-sigma limits** $\mu \pm 3\sigma$:

$$P(\mu - \sigma < X \leq \mu + \sigma) \approx 68\% \tag{1.60a}$$
$$P(\mu - 2\sigma < X \leq \mu + 2\sigma) \approx 95.5\% \tag{1.60b}$$
$$P(\mu - 3\sigma < X \leq \mu + 3\sigma) \approx 99.7\%. \tag{1.60c}$$

The aforementioned formulas show that a value deviating from μ by more than σ, 2σ, or 3σ will occur in one of about 3, 20, and 300 trials, respectively.

In tests[45] , we shall ask, conversely, for the intervals that correspond to certain given probabilities; practically most important use the probabilities of 95%, 99%, and 99.9%. For these,the answers are $\mu \pm 2\sigma, \mu \pm 2.6\sigma$, and $\mu \pm 3.3\sigma$, respectively.

[45] Which we shall cover in Chapter 2.

More precisely,

$$P(\mu - 1.96\sigma < X \leq \mu + 1.96\sigma) \approx 95\% \tag{1.61a}$$
$$P(\mu - 2.58\sigma < X \leq \mu + 2.58\sigma) \approx 99\% \tag{1.61b}$$
$$P(\mu - 3.29\sigma < X \leq \mu + 3.29\sigma) \approx 99.9\%. \tag{1.61c}$$

1.8.3. Normal Approximation of the Binomial Distribution

The probability function of the binomial distribution, as a reminder, is:

$$f(x) = \binom{n}{x} p^x q^{n-x} \qquad (x = 0, 1, \cdots, n). \tag{1.62}$$

If n is large, the binomial coefficients and powers become very inconvenient. It is of great practical[46] importance that, in this case, the normal distribution provides a good approximation of the binomial distribution, according to the following theorem, one of the most important theorems in all probability theory.

[46] and theoretical

Theory 1.15: Limit Theorem of De Moivre and Laplace

For large n,

$$f(x) \sim f^*(x) \qquad \text{where} \qquad x = 0, 1, \cdots n$$

Here f is given by Eq. (1.62). The function

$$f^*(\cdot) = \frac{1}{\sqrt{2\pi}\sqrt{npq}} \exp\left(-\frac{z^2}{2}\right), \qquad \text{and} \qquad z = \frac{x - np}{\sqrt{npq}}$$

is the density of the normal distribution with mean $\mu = np$ and variance $\sigma^2 = npq$ (the mean and variance of the binomial distribution). Furthermore, for any nonnegative integers a and b ($> a$):

$$P(a \leq X \leq b) = \sum_{x=a}^{b} \binom{n}{x} p^x q^{n-x} \sim \Phi(\beta) - \Phi(\alpha)$$

where,

$$\alpha = \frac{a - np - 0.5}{\sqrt{npq}} \qquad \text{and} \qquad \beta = \frac{b - np + 0.5}{\sqrt{npq}}$$

1.9. Distribution of Several Random Variables

Distributions of two (2) or more random variables are of interest for two (2) reasons:

1. They occur in experiments in which we observe several random variables, for example, carbon content X and hardness Y of steel, amount of fertiliser X and yield of corn Y, height X_1, weight X_2, and blood pressure X_3 of persons, and so on.

2. They will be needed in the mathematical justification of the methods of statistics in Chapter 2.

In this section we consider two (2) random variables X and Y or, as we also say, a **two-dimensional random variable** (X, Y). For (X, Y) the outcome of a trial is a pair of numbers $X = x$, $Y = y$, briefly $(X, Y) = (x, y)$, which we can plot as a point in the XY-plane.

The **two-dimensional probability distribution** of the random variable (X, Y) is given by the **distribution function**

$$F(x, y) = P\left(X \leqq x, Y \leqq y\right). \tag{1.63}$$

This is the probability that in a trial, X will assume any value not greater than x and in the same trial, Y will assume any value not greater than y. $F(x, y)$ determines the probability distribution uniquely, because extending the analogy we developed previously, $P(a < X \leq b) = F(b) - F(a)$, we now have for a rectangle defined using the following equation:

$$P\left(a_1 < X \leq b_1,\, a_2 < Y \leq b_2\right) = F\left(b_1, b_2\right) - F\left(a_1, b_2\right) - F\left(b_1, a_2\right) + F\left(a_1, a_2\right). \tag{1.64}$$

As before, in the two-dimensional case we shall also have discrete and continuous random variables and distributions.

1.9.1. Discrete Two-Dimensional Distribution

In analogy to the case of a single random variable, we call (X, Y) and its distribution **discrete** if (X, Y) can assume only finitely many or at most countably infinitely many pairs of values (x_1, y_1), $(x_2, y_2), \cdots$ with positive probabilities, whereas the probability for any domain containing none of those values of (X, Y) is zero.

Let (x_i, y_i) be any of those values and let $P\left(X = x_i,\, Y = y_j\right) = p_{ij}$ (where we admit that p_{ij} may be 0 for certain pairs of subscripts i). Then we define the **probability function** $f(x, y)$ of (X, Y) by:

$$f(x, y) = p_{ij} \qquad \text{if} \qquad x = x_i,\, y = y_j \qquad \text{and} \qquad f(x, y) = 0 \quad \text{otherwise;}$$

where, $i = 1, 2, \cdots$ and $j = 1, 2, \cdots$ independently. In analogy to Eq. (1.37), we now have for the distribution function the formula:

$$F(x, y) = \sum_{x_i \leq x} \sum_{y_j \leq y} f\left(x_i, y_j\right).$$

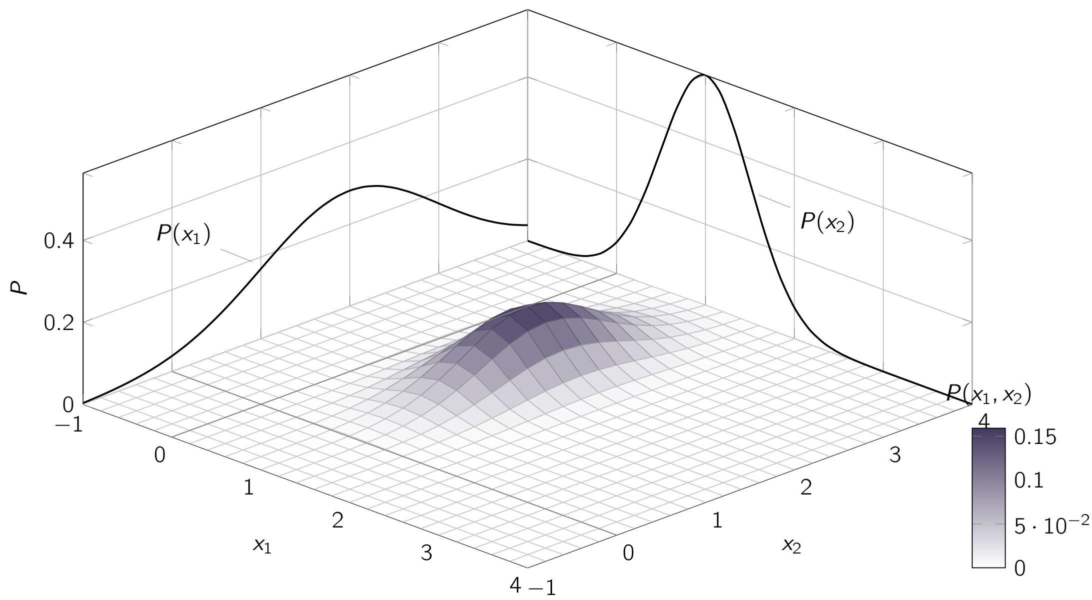

Figure 1.8.: Many samples from a bivariate normal distribution. The marginal distributions are shown on the z-axis. The marginal distribution of X is also approximated by creating a histogram of the X coordinates without consideration of the Y coordinates.

Instead of Eq. (1.39), we now have the condition:

$$\sum_i \sum_j f\left(x_\mathrm{i}, y_\mathrm{j}\right) = 1.$$

1.9.2. Continuous Two-Dimensional Distribution

In analogy to the case of a single random variable, we call (X, Y) and its distribution continuous if the corresponding distribution function $F(x, y)$ can be given by a double integral:

$$F(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f\left(x^*, y^*\right) \, \mathrm{d}x^* \, \mathrm{d}y^* \tag{1.65}$$

whose integrand f, called the **density** of (X, Y), is non-negative everywhere, and is continuous, possibly except on finitely many curves.

From Eq. (1.65) we obtain the probability that (X, Y) assume any value in a rectangle (Fig. 523) given by the formula:

$$P\left(a_1 < X \leq b_1,\, a_2 < Y \leqq b_2\right) = \int_{a_1}^{b_1} \int_{a_1}^{b_1} f(x, y) \, \mathrm{d}x \, \mathrm{d}y$$

1.9.3. Marginal Distributions of a Discrete Distribution

This is a rather natural idea, without counterpart for a single random variable.

It amounts to being interested only in one of the two variables in (X, Y), say, X, and asking for its distribution, called the **marginal distribution** of X in (X, Y). So we ask for the probability $P(X = x, Y)$ arbitrary.

Since (X, Y) is discrete, so is X. We get its probability function, call it $f_1(x)$, from the probability function $f(x, y)$ of (X, Y) by summing over y:

$$f_1(x) = P(X = x, Y, \text{arbitrary}) = \sum_y f(x, y) \tag{1.66}$$

where we sum all the values of $f(x, y)$ that are not 0 for that x.

From Eq. (1.66) we see that the distribution function of the marginal distribution of X is

$$F_1(x) = P\left(X \leqq x, Y, \text{arbitrary}\right) = \sum_{x^a \leqq x} f_1(x^*).$$

Similarly, the probability function

$$f_2(y) = P(X\text{arbitrary}, Y \equiv y) = \sum_x f(x, y)$$

determines the **marginal distribution** of Y in (X, Y). Here we sum all the values of $f(x, y)$ that are not zero for the corresponding y. The distribution function of this marginal distribution is

$$F_2(y) = P(X\text{arbitrary}, Y \equiv y) = \sum_{y^\star \equiv y} f_2(y^\star).$$

Exercise 1.20: Marginal Distributions of a Discrete Two-Dimensional Random Variable

In drawing 3 cards with replacement from a bridge deck let us consider

$$(X, Y) \quad \text{where} \quad X = \text{Number of queens} \quad \text{and} \quad Y = \text{Number of kings or aces.}$$

The deck has 52 cards. These include 4 queens, 4 kings, and 4 aces. Therefore, in a single trial a queen has probability:

$$\frac{4}{52} = \frac{1}{13}$$

and a king or ace:

$$\frac{8}{52} = \frac{2}{13}$$

This gives the probability function of (X, Y) as:

$$f(x, y) = \frac{3!}{x!y!(3-x-y)}\left(\frac{1}{13}\right)^x\left(\frac{2}{13}\right)^y\left(\frac{10}{13}\right)^{3-x-y} \quad \text{where} \quad (x + y \leq 3)$$

and $f(x, y) = 0$ otherwise.

1.9.4. Marginal Distributions of a Continuous Distribution

This is conceptually the same as for discrete distributions, with probability functions and sums replaced by densities and integrals. For a continuous random variable (X, Y) with density $f(x, y)$ we now have the marginal distribution of X in (X, Y), defined by the distribution function

$$F_1(x) = P\left(X \leqq x, -\infty < Y < \infty\right) = \int_{-\infty}^{x} f_1(x^*)\, dx^*$$

with the density f_1 of X obtained from $f(x, y)$ by integration over y,

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y)\ dy.$$

Interchanging the roles of X and Y, we obtain the **marginal distribution** of Y in (X, Y) with the distribution function

$$F_2(y) = P\left(-\infty < X < \infty, Y \leqq\right) = \int_{-\infty}^{y} f_2(y^*)\ dy^*$$

and density

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y)\ dx.$$

1.9.5. Independence of Random Variables

X and Y in a, discrete or continuous, random variable (X, Y) are said to be **independent** if

$$F(x, y) = F_1(x)F_2(y)$$

holds for all (x, y). Otherwise these random variables are said to be **dependent**. Necessary and sufficient for independence is

$$f(x, y) = f_1(x)f_2(y)$$

for all x and y. Here the f's are the above probability functions if (X, Y) is discrete or those densities if (X, Y) is continuous.

> **Exercise** 1.21: Independence and Dependence
>
> In tossing a 50 cent and a 20 cent coin, with X being the number of heads on the 50 cent, and Y number of heads on the 20 cent, we may assume the values 0 or 1 and are independent.

Extension of Independence to n-Dimensional Random Variables. This will be needed throughout Chapter 2. The distribution of such a random variable $\vec{X} = (X_1, \cdots, X_n)$ is determined by a **distribution function** of the form

$$F\left(x_1, \cdots, x_n\right) = P\left(X_1 \leqq x_1, \cdots, X_n \leqq x_n\right).$$

The random variables $X_1, \cdots, X_n$ are said to be **independent** if

$$F\left(x_1, \cdots, x_n\right) = F_1(x_1)F_2(x_2)\cdots F_n(x_n)$$

for all $(x_1, \cdots, x_n)$. Here $F_j(x_j)$ is the distribution function of the marginal distribution of X_j in $\vec{X}$, that is,

$$F_j(x_j) = P\left(X_j \leqq x_j, X_k \text{arbitrary}, k = 1, \cdots, n, k \neq j\right).$$

Otherwise these random variables are said to be **dependent**.

1.9.6. Functions of Random Variables

When $n = 2$, we write $X_1 = X$, $X_2 = Y$, $x_1 = x$, $x_2 = y$. Taking a non-constant continuous function $g(x, y)$ defined for all x, y, we obtain a random variable $Z = g(X, Y)$.

> For example, if we roll two (2) dice and X and Y are the numbers the dice turn up in a trial, then $Z = X + Y$ is the sum of those two (2) numbers.

In the case of a discrete random variable (X, Y) we may obtain the probability function $f(z)$ of $Z = g(X, Y)$ by summing all $f(x, y)$ for which $g(x, y)$ equals the value of z considered; thus

$$f(z) = P(Z = z) = \sum\sum_{g(x,y)=z} f(x, y).$$

Hence the distribution function of Z is

$$F(z) = P(Z \leq z) = \sum\sum_{g(x,y)\leq z} f(x, y),$$

where we sum all values of $f(x, y)$ for which $g(x, y) \leqq z$.

In the case of a continuous random variable (X, Y) we similarly have

$$F(z) = P\left(Z \leqq z\right) = \iint_{g(x,y)\leq z} f(x, y)\, \mathrm{d}x\, \mathrm{d}y$$

where for each z we integrate the density $f(x, y)$ of (X, Y) over the region $g(x, y) \leq z$ in the xy-plane, the boundary curve of this region being $g(x, y) = z$.

1.9.7. Addition of Means

The number

$$E\left(g(X, Y)\right) = \begin{cases} \sum_x \sum_y g(x, y) f(x, y) & \text{where } X, Y \text{ are discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) \, \mathrm{d}x \, \mathrm{d}y & \text{where } X, Y \text{ are continuous} \end{cases} \tag{1.67}$$

is called the mathematical expectation or, briefly, the **expectation of** $g(X, Y)$. Here it is assumed that the double series converges absolutely and the integral of $|g(x, y)|/(x, y)$ over the y-plane exists[47] . Since summation and integration are linear processes, we have from Eq. (1.67):

[47] meaning it is finite.

$$E\left(ag(X, Y) + bh(X, Y)\right) = aE\left(g(X, Y)\right) + bE\left(h(X, Y)\right)$$

An important special case is

$$E(X + Y) = E(X) + E(Y),$$

and by induction we have the following result.

Theory 1.16: Addition of Means

The mean (expectation) of a sum of random variables equals the sum of the means (expectations), that is,

$$E\left(X_1 + X_2 + \cdots + X_n\right) = E\left(X_1\right) + E\left(X_2\right) + \cdots + E\left(X_n\right).$$

We can also deduce the following statement:

Theory 1.17: Multiplication of Means

The mean (expectation) of the product of independent random variables equals the product of the means (expectations), that is,

$$E\left(X_1 X_2 \cdots X_n\right) = E\left(X_1\right) E\left(X_2\right) \cdots E\left(X_n\right).$$

and in the continuous case the proof of the relation is similar[48] .

[48] This is left as an exercise to the reader.

1.9.8. Addition of Variances

A final matter to cover is how we can sum up variances. Similar to before, let $Z = X + Y$ and denote the mean and variance of Z by μ and σ^2.

Then we first have:

$$\sigma^2 = E\left([Z-\mu]^2\right) = E\left(Z^2\right) - \left[E(Z)\right]^2$$

From (24) we see that the first term on the right equals

$$E\left(Z^2\right) = E\left(X^2 + 2XY + Y^2\right) = E\left(X^2\right) + 2E(XY) + E\left(Y^2\right).$$

For the second term on the right we obtain from Theorem 1

$$\left[E(Z)\right]^2 = \left[E(X) + E(Y)\right]^2 = \left[E(X)\right]^2 + 2E(X)E(Y) + \left[E(Y)\right]^2$$

By substituting these expressions into the formula for σ^2 we have

$$\begin{aligned}\sigma^2 &= E\left(X^2\right) - \left[E(X)\right]^2 + E\left(Y^2\right) - \left[E(Y)\right]^2 \\ &\quad + 2\left[E(XY) - E(X)E(Y)\right].\end{aligned}$$

the expression in the first line on the right is the sum of the variances of X and Y, which we denote by σ_1^2 and σ_2^2, respectively.

The quantity in the second line (except for the factor 2) is:

$$\sigma_{XY} = E(XY) - E(X)E(Y), \tag{1.68}$$

and is called the **covariance** of X and Y. Consequently, our result is

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 + 2\sigma_{XY}.$$

If X and Y are independent, then

$$E(XY) = E(X)E(Y);$$

hence $\sigma_{XY} = 0$, and

$$\sigma^2 = \sigma_1^2 + \sigma_2^2$$

Extension to more than two variables gives the basic

Theory 1.18: Addition of Variances

The variance of the sum of independent random variables equals the sum of the variances of these variables.

Chapter 2

Statistical Methods

Table of Contents

2.1. Introduction 37
2.2. Point Estimation of Parameters 39
2.3. Confidence Intervals 41

2.1. Introduction

Statistical methods consists of a wide range of tools for designing and evaluating random experiments to obtain information about practical problems:

> such as exploring the relation between iron content and density of iron ore, the quality of raw material or manufactured products, the efficiency of air-conditioning systems, the performance of certain cars, the effect of advertising, the reactions of consumers to a new product, etc.

Random variables occur more frequently in engineering[1] than one would think. For example, properties of mass-produced articles (screws, lightbulbs, etc.) always exhibit **random variation**, due to small (uncontrollable) differences in raw material or manufacturing processes.

[1] and of course elsewhere.

Thus the diameter of screws is a random variable X and we have nondefective screws, with diameter between given tolerance limits, and defective screws, with diameter outside those limits. We can ask for the distribution of X, for the percentage of defective screws to be expected, and for necessary improvements of the production process.

Samples are selected from populations:

> 20 screws from a lot of 1000, 100 of 5000 voters, 8 behaviours in a wildlife conservation project

as inspecting the entire population would be too expensive, time-consuming, impossible or even senseless[2] .

[2] It would be inconceivable for a company who produces over a billion lightbulbs to test all their products. That is why we have return policies.

To obtain a meaningful sense of information, samples must be **random selections**. Each of the 1000 screws must have the same chance of being sampled[3] , at least approximately. Only then will the sample mean:

[3] of being drawn when we sample.

$$\overline{x} = \frac{1}{20}\left(x_1 + \cdots x_{20}\right) \qquad \text{where} \qquad n = 20$$

will be a good approximation of the population mean μ, and the accuracy of the approximation will generally improve with increasing n, as we shall see.

> This behaviour is of course applicable to other statistical quantities such as standard deviation, variance, etc.

Independent sample values will be obtained in experiments with an infinite sample space S certainly for the normal distribution. This is also true in sampling with replacement. It is approximately true in drawing small samples from a large finite population.[4] However, if we sample without replacement from a small population, the effect of dependence of sample values may be considerable.

[4] for instance, 5 or 10 of 1000 items.

Random numbers help in obtaining samples that are in fact random selections. This is sometimes not easy to accomplish as there are numerous subtle factors which can bias sampling[5] .Random numbers can be obtained from a **random number generator**

[5] Such as by personal interviews, by poorly working machines, by the choice of nontypical observation conditions, etc.

> It is important to state that The numbers generated by a computer are not truly random, as are calculated by a tricky formula that produces numbers that do have practically all the essential features of true randomness. Because these numbers eventually repeat, they must not be used in cryptography, for example, where true randomness is required.

Exercise 2.1: Generating Random Numbers

To select a sample of size $n = 10$ from 80 given ball bearings, we number the bearings from 1 to 80. We then let the generator randomly produce 10 of the integers from 1 to 80 and include the bearings with the numbers obtained in our sample, for example,

$$44 \quad 55 \quad 57 \quad 03 \quad 61 \quad 51 \quad 68 \quad 22 \quad 34 \quad 77$$

or whichever number pops up in your head[6] .

[6] Of course in a professional setting you can't just write numbers like that as there is also a pattern when we make successive random number. Before the prevalence of computers there used to be books containing random numbers which people consulted.

Representing and processing data were considered in the previous chapter in connection with frequency distributions. These are the empirical counterparts of probability distributions and helped motivating axioms and properties in probability theory. The new aspect in this chapter is **randomness**:

> the data are samples selected **randomly** from a population.

Accordingly, we can already use the plots we have used in probability, such as stem-and-leaf plots, box plots, and histograms.

In this chapter the the mean $\overline{x}$ we defined in the previous chapter, will now be referred as **sample mean**.

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_{\mathrm{j}} = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_{\mathrm{n}}\right). \tag{2.1}$$

We call n the **sample size**, and similar to mean, the variance s^2 is called the **sample variance**:

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_{\mathrm{j}} - \overline{x}\right)^2 = \frac{1}{n-1}\left[\left(x_1 - \overline{x}\right)^2 + \cdots + \left(x_{\mathrm{n}} - \overline{x}\right)^2\right], \tag{2.2}$$

and its positive square root, s is the **sample standard deviation**.

$\overline{x}$, s^2, s are called **sample parameters**.

2.2. Point Estimation of Parameters

Before we dive deep into statistics, let's spend some time to learn the most basic practical tasks in statistics and corresponding statistical methods to accomplish them. The first is point estimation of parameters, that is, ofquantities appearing in distributions:

> such as p in the binomial distribution and μ and σ in the normal distribution.

A **point estimate** of a parameter is a number,[7] which is computed from a given sample and serves as an approximation of the unknown exact value of the parameter of the population. An interval estimate is an interval[8] obtained from a sample.

[7] which is a point on the real line.

[8] also known as confidence interval.

Estimation of parameters is of great practical importance in many applications.

As an approximation of the mean of a population we may take the mean $\overline{x}$ of a corresponding sample. This gives the estimate $\hat{\mu} = \overline{x}$ for μ, that is,

$$\hat{\mu} = \overline{x} = \frac{1}{n}\left(x_1 + \cdots + x_{\mathrm{n}}\right) \tag{2.3}$$

where n is the sample size. Similarly, an estimate $\hat{\sigma}^2$ for the variance of a population is the variance s^2 of a corresponding sample, that is:

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_{\mathrm{j}} - \overline{x}\right)^2. \tag{2.4}$$

As can be seen, Eq. (2.3) and Eq. (2.4) are estimates of parameters for distributions in which μ or σ^2 appear explicitly as parameters, such as the normal and Poisson distributions.

For the binomial distribution, $p = \mu/n$. From Eq. (2.3) we obtain for p the estimate:

$$\hat{p} = \frac{\overline{x}}{n}. \tag{2.5}$$

It is important to mention Eq. (2.3) is a special case of the so-called **method of moments**. Here, the parameters to be estimated are expressed in terms of the moments of the distribution. In the resulting formulas, those moments of the distribution are replaced by the corresponding moments of the sample, which gives the estimates. Here the k$^{\text{th}}$ moment of a sample $x_1, \cdots, x_n$ is:

$$m_k = \frac{1}{n}\sum_{j=1}^{n} x_j^k. \tag{2.6}$$

[9]Considered the father of modern statistics.

2.2.1. Maximum Likelihood Method

Another method for obtaining estimates is the so-called **maximum likelihood method** conceived by R. A. Fisher[9] . To explain it, we consider a discrete (or continuous) random variable X whose probability function (or density) $f(x)$ depends on a single parameter θ. We take a corresponding sample of n independent values $x_1, \cdots, x_n$. Then in the discrete case the probability given a sample of size n consists precisely of those n values is

$$l = f(x_1)\, f(x_2) \cdots f(x_n)\,. \tag{2.7}$$

In the continuous case the probability that the sample consists of values in the small intervals $x_j \leq x \leqq x_j + \Delta x\,(j = 1, 2, \cdots, n)$ is

$$f(x_1)\,\Delta x f(x_2)\,\Delta x \cdots f(x_n)\,\Delta x = l\,(\Delta x)^n \tag{2.8}$$

As $f\left(x_j\right)$ depends on θ, the function l in Eq. (2.8) given by Eq. (2.7) depends on $x_1, \cdots, x_n$ and θ.

> We imagine $x_1, \cdots, x_n$ to be given and fixed.

Then l is a function of θ, which is called the **likelihood function**. The basic idea of the maximum likelihood method is quite simple, as follows.

We choose an approximation for the unknown value of θ for which l is as large as possible.

[10]not at the boundary.

If l is a differentiable function of θ, a necessary condition for l to have a maximum in an interval[10] is

$$\frac{\partial l}{\partial \theta} = 0 \tag{2.9}$$

A solution of Eq. (2.9) depending on $x_1, \cdots, x_n$ is called a **maximum likelihood estimate** for θ. We may replace Eq. (2.9) by:

$$\frac{\partial \ln l}{\partial \theta} = 0 \tag{2.10}$$

as $f\left(x_j\right) > 0$, a maximum of l is in general positive, and $\ln l$ is a monotone increasing function of l. This often simplifies calculations.

Several Parameters

If the distribution of X involves r parameters $\theta_1, \cdots, \theta_r$, then instead of Eq. (2.9) we have the r conditions $\partial \ln l / \partial \theta_1, \cdot, \partial \ln l / \partial \theta_r = 0$, and instead of Eq. (2.10) we have:

$$\frac{\partial \ln l}{\partial \theta_1} = 0, \cdots, \frac{\partial \ln l}{\partial \theta_r} = 0. \tag{2.11}$$

Exercise 2.2: Maximum Likelihood of Gaussian Distribution

Find maximum likelihood estimates for $\theta_1 = \mu$ and $\theta_2 = \sigma$ in the case of the normal distribution.

Solution

We obtain the likelihood function:

$$l = \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\sigma}\right)^n e^{-h} \qquad \text{where} \qquad h = \frac{1}{2\sigma^2} \sum_{j=1}^{n} \left(x_j - \mu\right)^2 .$$

Taking logarithms, we have

$$\ln l = -n \ln \sqrt{2\pi} - n \ln \sigma - h.$$

The first equation in Eq. (2.11) is $\frac{\partial \ln l}{\partial \mu} = 0$, written out:

$$\frac{\partial \ln l}{\partial \mu} = -\frac{\partial h}{\partial \mu} = \frac{1}{\sigma^2} \sum_{j=1}^{n} \left(x_j - \mu\right) = 0, \qquad \text{therefore} \qquad \sum_{j=1}^{n} x_j - n\mu = 0.$$

The solution is the desired estimate $\hat{\mu}$ for μ: we find

$$\hat{\mu} = \frac{1}{n} \sum_{j=1}^{n} x_j = \overline{x} .$$

The second equation in Eq. (2.11) is $\frac{\partial \ln l}{\partial \sigma} = 0$, written out

$$\frac{\partial \ln l}{\partial \sigma} = -\frac{n}{\sigma} - \frac{\partial h}{\partial \sigma} = -\frac{1}{\sigma} + \frac{1}{\sigma^3} \sum_{j=1}^{n} \left(x_j - \mu\right)^2 = 0.$$

Replacing μ by $\hat{\mu}$ and solving for σ^2, we obtain the estimate:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{n} \left(x_j - \overline{x}\right)^2 \quad \blacksquare$$

2.3. Confidence Intervals

Confidence intervals for an unknown parameter θ of some distribution (e.g., $\theta = \mu$) are intervals $\theta_1 \leq \theta \leq \theta_2$ which contain θ, not with certainty but with a high probability γ, which we can choose[11] .

[11] 95% and 99% are popular

Such an interval is calculated from a sample. $\gamma = 95\%$ means probability $1-\gamma = 5\% = 1/20$ of being wrong[12] . Instead of writing $\theta_1 \leq \theta \leq \theta_2$, we denote this more **distinctly** by writing:

[12] one of about 20 such intervals will **NOT** contain θ

$$\text{CONF}_\gamma \; \{\theta_1 \leq \theta \leq \theta_2\} \tag{2.12}$$

Such a special symbol, CONF, seems worthwhile in order to avoid the misunderstanding that θ **must** lie between θ_1 and θ_2.

γ is called the **confidence level**, and θ_1 and θ_2 are called the **lower** and **upper confidence limits**, respectively and depend on the γ value. The larger we choose γ, the smaller is the error probability $1-\gamma$, but the longer is the confidence interval.

If $\gamma \to 1$, then its length goes to infinity.

The choice of γ depends on the kind of application.

> In taking no umbrella, a 5% chance of getting we's not fragile. In a medical decision of life or death, a 5% chance of being wrong may be too large and a 1% chance of being wrong ($\gamma = 99\%$) may be more desirable.

Confidence intervals are more valuable than point estimates. We can take the midpoint of Eq. (2.12) as an approximation of θ and half the length of Eq. (2.12) as an error bound[13] .

[13] not in the strict sense of numerics, but except for an error whose probability we know.

θ_1 and θ_2 in Eq. (2.12) are calculated from a sample $x_1, \cdots, x_n$. These are n observations of a random variable X. Now comes a **standard trick**.

We regard $x_1, \cdots, x_n$ as single observations of n random variables $X_1, \cdots, X_n$[14]. Then $\theta_1 = \theta_1(x_1, \cdots, x_n)$ and $\theta_2 = \theta_2(x_1, \cdots, x_n)$ in Eq. (2.12) are observed values of two random variables $\Theta_1 = \Theta_1(X_1, \cdots, X_n)$ and $\Theta_2 = \Theta_2(X_1, \cdots, X_n)$. The condition Eq. (2.12) involving γ can now be written

[14] with the same distribution, namely, that of X

$$P\left(\Theta_1 \leqq \theta \leqq \Theta_2\right) = \gamma. \tag{2.13}$$

Let us see what all this means in concrete practical cases.

In each case in this section we shall first state the steps of obtaining a confidence interval in the form of a table, then consider a typical example, and finally justify those steps theoretically.

2.3.1. Confidence Interval for Mean with known Variance in Normal Distribution

The method of tackling is this problem is as follows:

1. Choose a confidence level for γ[15] .

2. Determine the corresponding c:

γ	0.90	0.95	0.99	0.999
c	1.645	1.960	2.576	3.291

3. Compute the mean $\overline{x}$ of the sample $x_1, \cdots, x_n$.

4. Compute $k = c\sigma/\sqrt{n}$. The confidence interval for μ is

$$\text{CONF}_{\gamma}\{\overline{x} - k \leqq \mu \leqq \overline{x} + k\}. \tag{2.14}$$

[15] 95%, 99%, depending on the application

Exercise 2.3: Confidence Interval for mean with known variance in Normal Distribution

Determine 95% confidence interval for the mean of a normal distribution with variance $\sigma^2 = 9$, using a sample of $n = 100$ values with mean $\overline{x} = 5$.

Solution

1. First we define γ as 0.95.
2. Then looking at the table find the corresponding c which equals 1.960.
3. $\hat{x} = 5$ is given.
4. We need:

$$k = c\frac{\sigma}{\sqrt{n}} = 1.960\frac{3}{\sqrt{100}} = 0.588$$

Therefore

$$\hat{x} - k = 4.412 \qquad \text{and} \qquad \hat{x} + k = 5.588$$

and the confidence interval is:

$$\text{CONF}_{0.95}\{4.412 \leq \mu \leq 5.588\} \quad \blacksquare$$

Glossary

IVP Initial Value Problem. ii

LHS Left Hand Side. ii, 22

ODE Ordinary Differential Equation. ii

PDE Partial Differential Equation. ii

RHS Right Hand Side. ii, 22

Bibliography

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