Lecture Book
B.Sc Data Science I - Tutorial

D. T. McGuiness, Ph.D

Version: SS.2025

Topics on Artificial Intelligence & Machine Learning

Contents

List of Figures

List of Tables

List of Examples

List of Theorems

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 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.

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

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

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

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 [**?**]. For Eq. (1.1) it is shown in **Table ??**.
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 ??**. 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.

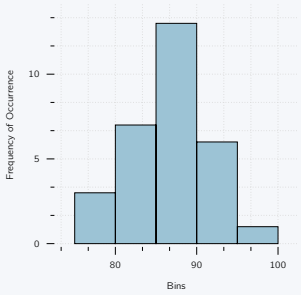


Figure 1.1: The histogram of the data given in **Exercise 1**.

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:

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

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]

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

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

[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 Var() are all acceptable symbols.

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:

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

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 $\overline{x}_{\text{red}} = 86.2$, 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[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.

[9] Sometimes known as probability calculus.

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]

[10] also known as sample point.

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

Exercise 1.5: Sample Spaces of Random Experiments & Events

- Inspecting a lightbulb | S = {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 = {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

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

same time.

[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.

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

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

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 expressing the concept covered previously.

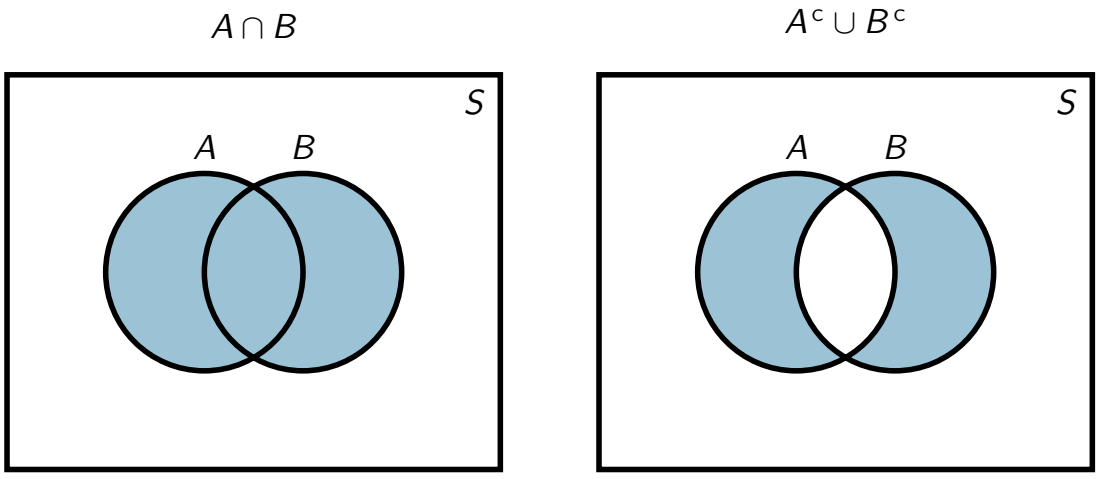


Figure 1.2: Examples of Venn diagrams.

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:

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

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:

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

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* $= \{1, 3\}$ $\blacksquare$

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 $P(A)$ which corresponds to $A \cap B$. Therefore,

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

Similarly, the conditional probability of A given B is:

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

Theory 1.6: Multiplication Rule

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

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

Exercise 1.10: Multiplication Rule

In producing screws, let:

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

Let $P\left(A\right) = 0.1$ 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

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

Independent Events

If events A and B are such that

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

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

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

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 $P(A) = a/k$ 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:

[18] 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[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

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

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 1.8: 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],

$$0! = 1.$$

[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.

Values may be computed recursively from given values by

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

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:

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

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

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.

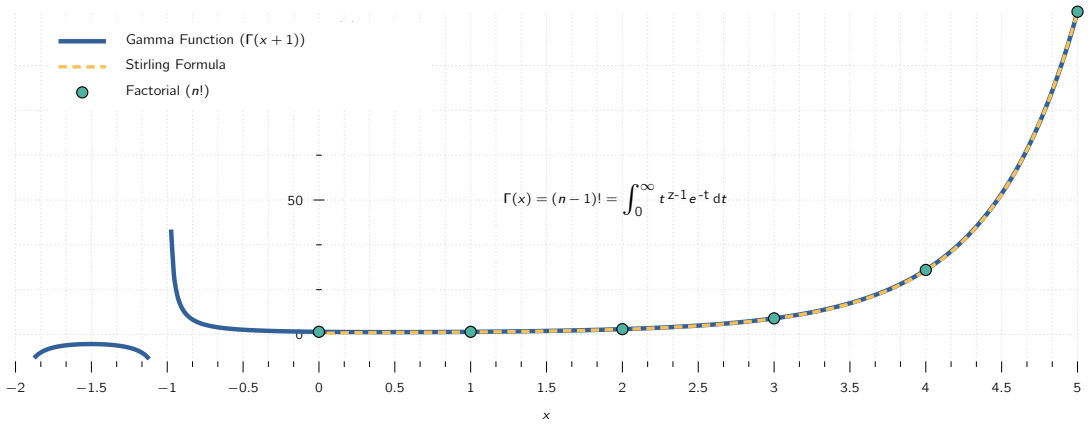


Figure 1.3: A visual comparison of the Stirling formula and the actual values of the factorial function.

1.4.4 Binomial Coefficients

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

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

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.

[22] 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[23] as the value it will assume in the next trial depends on the **stochastic process**

[23] 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[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**:

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

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

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].

[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 $a < x \leqq b$:

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

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):

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

and subtraction of $F(a)$ 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,

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

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

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

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

$$F(x) = \sum_{x_j \leq x} f\left(x_j\right) = \sum_{x_j \leq x} p_j \tag{1.37}$$

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):

$$P\left(a < X \leqq b\right) = F(b) - F(a) = \sum_{a < x_j \leqq b} p_j \tag{1.38}$$

This is the sum of all probabilities p_j for which x_j satisfies $a < x_j \leqq b$[26]. From this and $P(S) = 1$ 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.

$$\sum_j p_j = 1 \qquad \text{(sum of all probabilities)}. \tag{1.39}$$

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

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

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]:

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

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

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.

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

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

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

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):

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

Continuous random variables are simpler than discrete ones with respect to intervals as, in the continuous case the four probabilities corresponding to $a < X \leq b$, $a < X < b$, $a \leq X \leq b$,

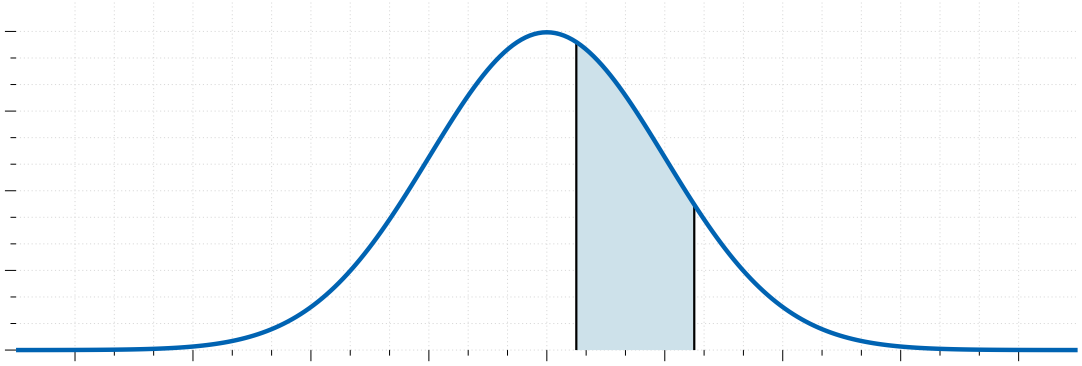


Figure 1.4: A visual representation of the Eq. (1.42).

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.

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:

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

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

and the **variance** σ^2 by:

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

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

[30] Sometimes it is known as Var (x)

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).

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.

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

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

$$\sigma^2 > 0$$

[32] and finite.

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

> **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

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 $E\left(g(X)\right)$ is the value of $g(X)$ to be expected on the average, defined by:

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

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

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

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

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

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

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].

[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)

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:

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

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

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;

[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:

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

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:

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

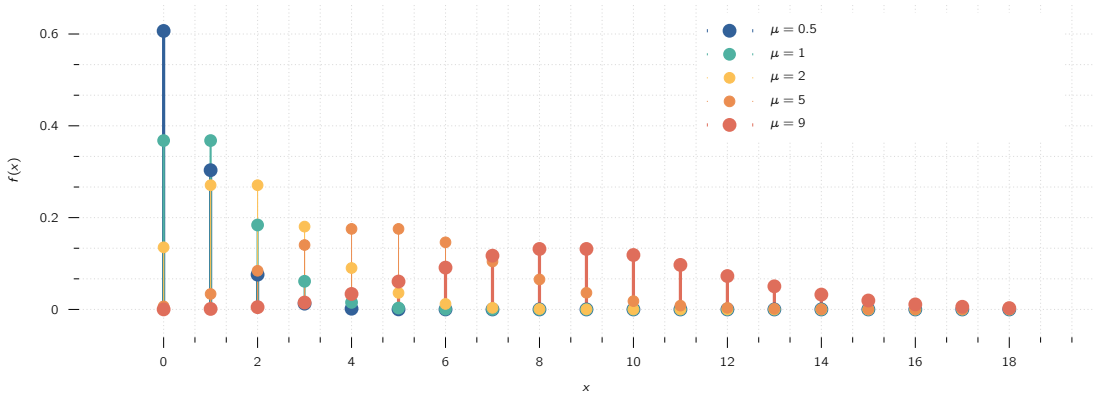


Figure 1.5: The Poisson distribution with different mean (μ) values.

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.

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

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:

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

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

The hypergeometric distribution has the mean:

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

and the variance

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

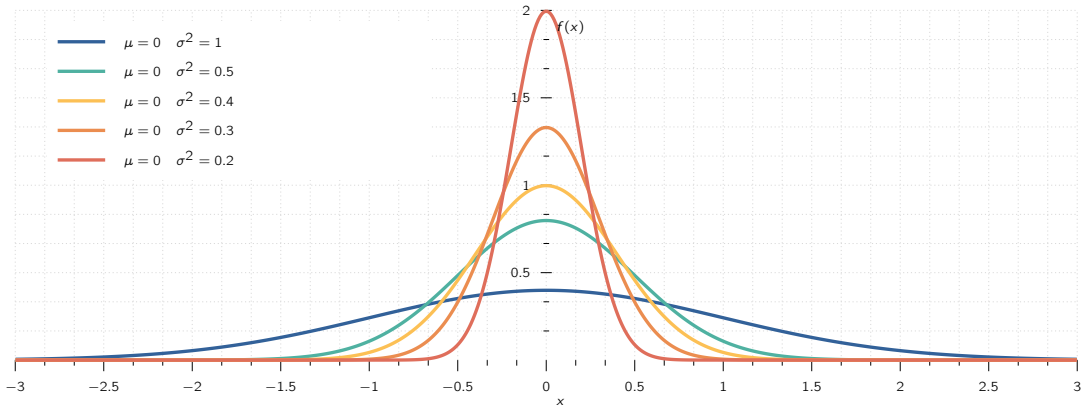


Figure 1.6: The poster child of probability and statistics, the normal distribution.

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:

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

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.

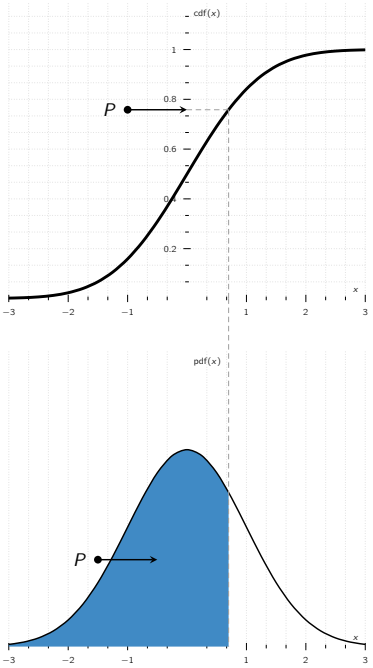


Figure 1.7: A visual representation between the relationship of PDF and CDF.

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:

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

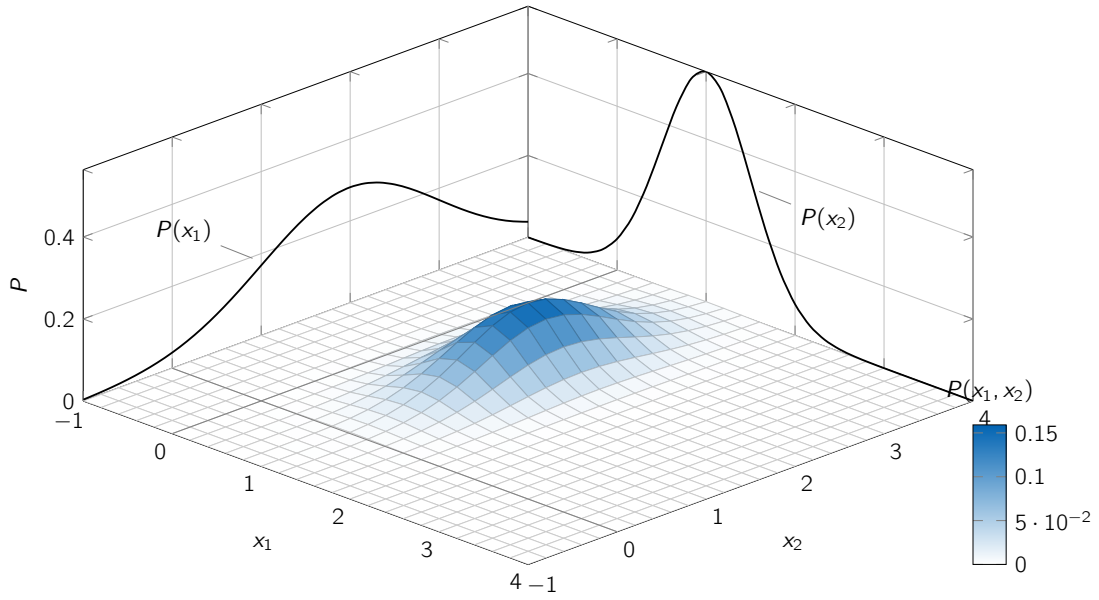


Figure 1.8: Many samples from a bivariate normal distribution. The marginal distributions are shown on the z-axis. The marginal distribution of X is also approximated by creating a histogram of the X coordinates without consideration of the Y coordinates.

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

$$\sum_i \sum_j f\left(x_i, y_j\right) = 1.$$

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

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

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

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

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 $\boldsymbol{X} = (X_1, \cdots, X_n)$ 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 $\boldsymbol{X}$, 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

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

Hence the distribution function of Z is

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

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

$$F(z) = P(Z \leqq z) = \iint_{g(x,y)\leq z} f(x, y)\, dx\, dy$$

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

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

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.

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

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:

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

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

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

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

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

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

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

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:

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

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

$$E\left(XY\right) = E\left(X\right)E\left(Y\right);$$

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

Statistical[1] methods consists of a wide range of tools for designing and evaluating random experiments to obtain information about practical problems:

[1] The word is derived from New Latin *statistica* or *statisticus* ("of the state")

> 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.

Therefore, the diameter of screws is a random variable X and we have non-defective 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 1000 screws, 100 of 5000 voters, 8 behaviours in a wildlife observation.

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

[2] It would be inconceivable for a company who produces over a billion light bulbs 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 is also 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 non-typical 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 55 57 03 61 51 68 22 34 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**:

i.e., 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 mean $\overline{x}$ we defined previously, will now be referred as **sample mean**.

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

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

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

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

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

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, of quantities 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[9] 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,

[9] to describe something which is an approximation or an educated guess, we use hat (^) notation. This is applicable for fields in statistics, machine learning or data science.

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

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:

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

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^{th} moment of a sample $x_1, \cdots, x_n$ is:

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

2.2.1 Maximum Likelihood Method

Another method for obtaining estimates is

the so-called **maximum likelihood method** conceived by 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

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

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

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

As $f\left(x_j\right)$ depends on θ, the function I 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 I 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 I is as large as possible.

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

[10] not at the boundary.

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

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:

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

as $f\left(x_j\right) > 0$, a maximum of I is in general positive, and $\ln I$ is a monotone increasing function of I. 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 $\partial \ln I/\partial \theta_1, \cdot, \partial \ln I/\partial \theta_r = 0$, and instead of Eq. (2.10) we have:

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

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:

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

Exercise 2.3: Estimating Poisson Parameters

Consider a Poisson distribution with Probability Mass Function (PMF):

$$f\left(x|\mu\right) = \frac{e^{-\mu}\mu^{x}}{x!}, \qquad \text{where} \qquad x = 1,\ 2,\ 3,\ \cdots$$

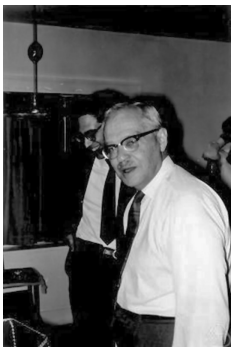
Suppose a random sample $x_1,\ x_2 \cdots x_n$ is taken from distribution. What is the maximum likelihood estimate of μ?

Solution

The likelihood function is:

$$L\left(x_1,\ x_2 \cdots x_n;\mu\right) = \prod_{i=1}^{n} f\left(x_i|\mu\right) = \frac{e^{-n\mu}\mu^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

2.3 Confidence Intervals



Confidence intervals[11] 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.[12] Such an interval is calculated from a sample. $\gamma = 95\%$ means probability $1 - \gamma = 5\% = 1/20$ of being wrong.[13] Instead of writing $\theta_1 \leq \theta \leq \theta_2$, we denote this more **distinctly** by writing:

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

> Such a special symbol, CONF, seems worthwhile 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 wet is **NOT** a problem. 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.[14]

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 $X_1, \cdots, X_n$[15]. 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

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

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.

[11] Established by Jerzy Neyman. He proposed and studied randomised experiments in 1923. Furthermore, his paper *On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection*, given at the Royal Statistical Society on 19 June 1934, was the groundbreaking event leading to modern scientific sampling. He introduced the confidence interval in his paper in 1937. Another noted contribution is the Neyman-Pearson lemma, the basis of hypothesis testing.

[12] 95% and 99% are popular

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

[14] not in the strict sense of numerical means, but except for an error whose probability we know.

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

For Mean with known Variance in Normal Distribution

The method of tackling is this problem is as follows:

[16] 95%, 99%, depending on the application.

1. Choose a confidence level for γ .[16]
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

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

Exercise 2.4: 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

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

and the confidence interval is:

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

Theory 2.19: Sum of Independent Normal Random Variables

Let $X_1, \cdots, X_n$ be independent normal random variables each of which has mean μ and variance σ^2.

Then the following holds:

a. The sum $X_1 + \cdots + X_n$ is normal with mean $n\mu$ and variance $n\sigma^2$.

b. The following random variable $\overline{X}$ is normal with mean μ and variance σ^2/n.

$$\overline{X} = \frac{1}{n}\left(X_1 + \cdots + X_n\right)$$

c. The following random variable Z is normal with mean 0 and variance 1.

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

Exercise 2.5: Sample Size Needed for a Confidence Interval of Prescribed Length

How large must n be in Example 2.3 if we want to obtain a 95% confidence interval of length $L = 0.4$?

Solution

The interval in Eq. (2.14) has the length:

$$L = 2k = 2c\sigma/\sqrt{n}$$

Solving for n, we obtain

$$n = \left(\frac{2c\sigma}{L}\right)^2$$

In the present case the answer is:

$$n = \left(\frac{2 \times 1.96 \times 3}{0.4}\right)^2 \approx 870 \quad \blacksquare$$

For Mean of the Normal Distribution with Unknown Variance

In practice σ^2 is frequently unknown. Then the method described previously does not help and the whole theory changes, although the steps of determining a confidence interval for μ remain quite similar. They will be explained in the following section. We see that k differs from previous method, namely, the sample standard deviation s has taken the place of the unknown standard deviation σ of the population. And c now depends on the sample size n and must be determined from Table 49 given in the Appendix. That table lists values z for given values of the distribution function.

$$F(z) = K_m \int_{-\infty}^{x} \left(1 + \frac{u^2}{m}\right)^{-(m+1)/2} du \tag{2.15}$$

of the t-distribution. Here, m $(= 1, 2, \cdots)$ is a parameter, called the **number of degrees of freedom** of the distribution.[17] In the present case, $m = n - 1$; see Table 25.2. The constant K_m is such that $F(\infty) = 1$. By integration it turns out that

[17] abbreviated d.f.

$$K_m = \frac{\Gamma\left(\frac{1}{2}m + \frac{1}{2}\right)}{\sqrt{m\pi}\,\Gamma\left(\frac{1}{2}m\right)},$$

where Γ is the gamma function.

The method of tackling is this problem is as follows:

1. Choose a confidence level γ .[18]

[18] 95%, 99%, or the like.

2. Determine the solution c of the equation,

$$F(c) = \frac{1}{2}(1 + \gamma)$$

   from the table of the t-distribution with $n - 1$ degrees of freedom

3. Compute the mean $\overline{x}$ and the variance s^2 of the sample $x_1, \cdots, x_n$.

4. Compute $k = cs/\sqrt{n}$. The confidence interval is:

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

This illustrates that Table 25.1 (which uses more information, namely, the known value of σ^2) yields shorter confidence intervals than Table 25.2. This is confirmed in, which also gives an idea of the gain by increasing the sample size.

Exercise 2.6: Confidence Interval for Mean of Normal Distribution with Unknown Variance

The five (5) independent measurements of flash point of Diesel oil (D-2) gave the values (in °F):

144 147 146 142 144

If we assume normality, determine a 99% confidence interval for the mean.

Solution

1. $\gamma = 0.99$ is required.
2. $F(c) = \frac{1}{2}(1+\gamma) = 0.99$ and looking at the reference table with $n-1 = 4$ d.f. giving $c = 4.60$.
3. $\overline{x} = 144.6$ and $s = 3.8$,
4. $k = \sqrt{3.8} \times 4.60/\sqrt{5} = 4.01$. Therefore the confidence interval is:

$$\text{CONF}_{0.99}\{140.5 \leq \mu \leq 148.7\} \quad \blacksquare$$

If the variance σ^2 were known and equal to the sample variance s^2, thus $\sigma^2 = 3.8$, then the Reference Table would give:

$$k = \frac{c\sigma}{\sqrt{n}} = 2.576\frac{\sqrt{3.8}}{\sqrt{3}} = 2.25$$

and

$$\text{CONF}_{0.99}\{140.5 \leq \mu \leq 148.7\}$$

We see that the present interval is almost twice as long as that with a known variance $\sigma^2 = 3.8$.

[19] William Gosset (13 June 1876 - 16 October 1937) was an English statistician, chemist and brewer who served as Head Brewer of Guinness and Head Experimental Brewer of Guinness and was a pioneer of modern statistics. He published is results under the pen name student.

Theory 2.20: Student's t-Distribution

Let $X_1, \cdots, X_n$ be independent normal random variables with the same mean μ and the same variance σ^2. Then the random variable:

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has a t-distribution[19] with $n-1$ degrees of freedom (d.f.); here $\overline{X}$ is given by (4) and

$$S^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \overline{X}^2\right).$$

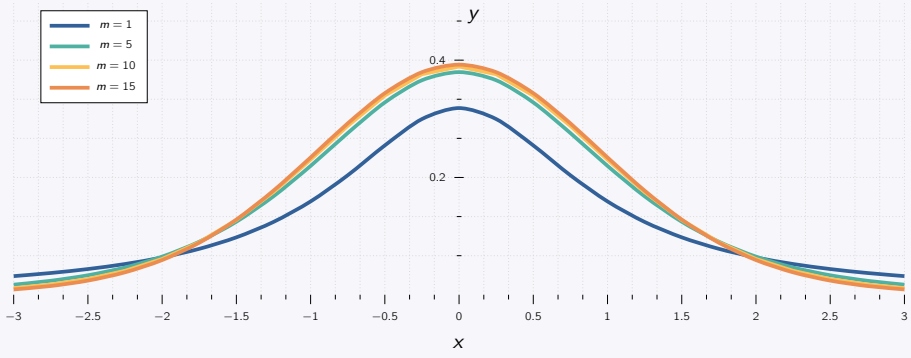


Figure 2.1: The student-t distribution with different degrees of freedom m.

For the Variance of the Normal Distribution

The method for calculating the confidence interval is similar to the previous methods, with slight change in some steps which are as follows:

1. Choose a confidence level γ.[20]

[20] as usual this can be 95%, 99%, or the like.

2. Determine solutions c_1 and c_2 of the equations:

$$F(c_1) = \frac{1}{2}(1-\gamma) \qquad \text{and} \qquad F(c_2) = \frac{1}{2}(1+\gamma).$$

where the necessary values are calculated from the table of the chi-square distribution with $n-1$ degrees of freedom.

3. Compute $(n-1)\,s^2$, where s^2 is the variance of the sample $x_1, \cdots, x_n$.

4. Compute $k_1 = (n-1)s^2/c_1$ and $k_2 = (n-1)s^2/c_2$. The confidence interval is

$$\text{CONF}_\gamma\{k_2 \cong \sigma^2 \cong k_1\}. \tag{2.16}$$

Exercise 2.7: Confidence Interval for the Variance of the Normal Distribution

Determine a 95% confidence interval Eq. (2.16) for the variance, using Table 25.3 and a sample (tensile strength of sheet steel in $\mathrm{kg\,mm^{-2}}$, rounded to integer values)

89 84 87 81 89 86 91 90 78 89 87 99 83 89

Solution

1. $\gamma = 0.95$ is required.
2. For $n-1 = 13$ we find
$$c_1 = 5.01 \qquad \text{and} \qquad c_2 = 24.74.$$
3. $13s^2 = 326.9$
4. $13s^2/c_1 = 65.25 \qquad \text{and} \qquad 13s^2/c_2 = 13.21$
5. This makes the confidence interval as:
$$\text{CONF}_{0.09}\left\{13.21 \le \sigma^2 \le 65.25\right\}.$$

This is rather large, and for obtaining a more precise result, one would need a much larger sample ■.

Theory 2.21: Chi-Square Distribution

Under the assumptions in Theorem 2 the random variable

$$Y = (n-1)\,\frac{S^2}{\sigma^2}$$

with S_2 given by (12) has a chi-square distribution with $n-1$ degrees of freedom.

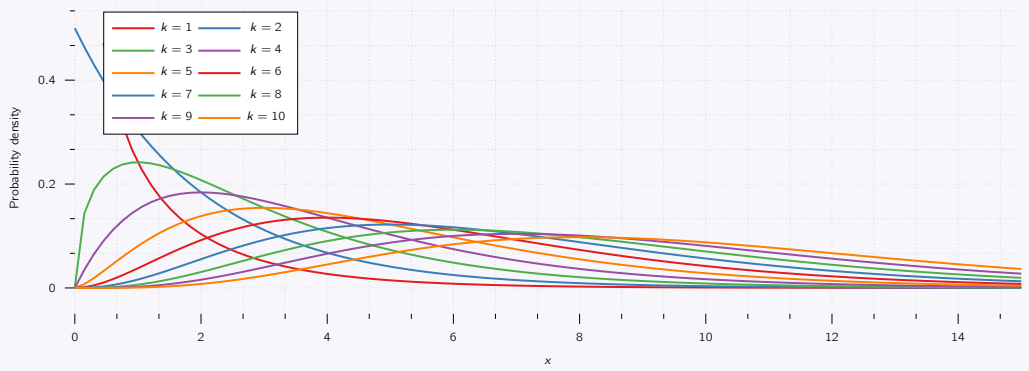


Figure 2.2: Chi-square distribution with different degrees of freedom.

The chi-squared distribution, which can be seen in **Fig.** 2.2 is used primarily in **hypothesis testing**, and to a lesser extent for confidence intervals for population variance when the underlying distribution is normal. Unlike more widely known distributions such as the normal distribution and the exponential distribution, the chi-squared distribution is not as often applied in the direct modelling of natural phenomena.

The primary reason for which the chi-squared distribution is extensively used in hypothesis testing is its relationship to the normal distribution. Many hypothesis tests use a test statistic, such as the t-statistic in a t-test. For these hypothesis tests, as the sample size n increases, the sampling distribution of the test statistic approaches the normal distribution.[21] Because the test statistic (t) is asymptotically normally distributed, provided the sample size is sufficiently large, the distribution used for hypothesis testing may be approximated by a normal distribution.

[21] This is the result of the central limit theorem.

So wherever a normal distribution could be used for a hypothesis test, a chi-squared distribution could be used.

Confidence Internals for Parameters of Other Distributions

The methods in mentioned previously for confidence intervals for μ and σ^2 are designed for the normal distribution. We will see it here that they can also be applied to other distributions if we **use large samples**.

We know that if $X_1, \cdots, X_n$ are independent random variables with the same mean μ and the same variance σ^2, then their sum $Y_n = X_1 + \cdots + X_n$ has the following properties:

- Y_n has the mean $n\mu$ and the variance $n\sigma^2$,
- If those variables are normal, then Y_n is normal.

If those random variables are **not normal**, then second property is **NOT** applicable. However, for large n the random variable Y_n is still approximately normal.

This follows from the central limit theorem, which is one of the most fundamental results in probability theory.

Theory 2.22: Central Limit Theorem

Let $X_1, \cdots, X_n$ be independent random variables having the same distribution function and therefore the same mean μ and

variance σ^2. Let $Y_n = X_1 + \cdots + X_n$, then the random variable

$$Z_n = \frac{Y_n - n\mu}{\sigma\sqrt{n}}$$

is **asymptotically normal** with mean 0 and variance 1. That is, the distribution function $F(x)$ of Z_n satisfies:

$$\lim_{n\to\infty} F(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du.$$

This theorem basically boils down to the following statement:

> Under appropriate conditions, the distribution of a normalised version of the sample mean converges to a standard normal distribution. This holds even if the original variables themselves are not normally distributed.

Therefore, when applying the previous confidence interval methods to a non-normal distribution, we must use sufficiently large samples.

As a rule of thumb, if the sample indicates that the skewness of the distribution is small, use at least $n = 20$ for the mean and at least $n = 50$ for the variance.

[22] The modern development of tests are generally attributed to Egon Sharpe Pearson and Neymar whom was mentioned previously. Egon Sharpe was one of three children of Karl Pearson and Maria, and, like his father, a British statistician. He is known throughout the world as co-author of the Neyman-Pearson theory of testing statistical hypotheses, and responsible for many important contributions to problems of statistical inference and methodology, especially in the development and use of the likelihood ratio criterion.

2.4 Testing of Hypotheses and Making Decisions

The ideas of confidence intervals and of tests[22] are the two (2) most important ideas in modern statistics. In a statistical test we make inference from sample to population through testing a **hypothesis**, resulting from experience or observations, from a theory or a quality requirement, and so on.

In many cases the result of a test is used as a basis for a **decision**:

> to buy, or not to buy a certain model of car, depending on a test of the fuel efficiency ($\mathrm{km\,L^{-1}}$), or, to apply some medication, depending on a test of its effect; to proceed with a marketing strategy, depending on a test of consumer reactions, etc.

As with most abstract mathematical concepts, it is better to explain such a test in terms of a typical example and then introduce the corresponding standard notions of statistical testing.

[23] also called null hypothesis.

[24] or alternative hypothesis

Exercise 2.8: Test of a Hypothesis

label=a Let's say we want to buy 100 coils of a certain kind of wire, provided we can verify the manufacturer's claim that the wire has a specific strength of $\mu = \mu_0 = 200\ \mathrm{kN\,m\,kg^{-1}}$, or more.

This is a test of the hypothesis:[23] $\mu = \mu_0 = 200$. We shall **NOT** buy the wire if the statistical tests shows that actually $\mu = \mu_1 < \mu_0$, the wire is weaker, the claim does **NOT** hold. μ_1 is called the **alternative** of the test.[24] We shall **accept** the hypothesis if the test suggests that it is true, except for a small error probability α, called the **significance level** of the test.

Otherwise we reject the hypothesis.

Hence α is the probability of rejecting a hypothesis although it is true. The choice of α is up to us, 5% and 1% are popular values.

For the test we need a sample. We randomly select 25 coils of the wire, cut a piece from each coil, and determine the breaking limit experimentally. Suppose that this sample of $n = 25$ values of the breaking limit has the mean $\overline{x} = 197\ \mathrm{kN\,m\,kg^{-1}}$, which is somewhat less than the claim, and the standard deviation $s = 6\ \mathrm{kN\,m\,kg^{-1}}$.

At this point we could only speculate when this difference $197-200 = -3$ is due to randomness, is a chance effect, or whether it is **significant**, due to the actual inferior quality of the wire. To continue beyond speculation requires probability theory, as follows.

We assume that the blocking limit is normally distributed. Then

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

with $\mu = \mu_0$ has a **t-distribution** with $n-1$ degrees of freedom ($n-1 = 24$ for our sample). Also $\overline{x} = 197$ and $s = 6$ are observed values of $\overline{X}$ and S to be used later. We can now choose a significance level, say, $\alpha = 95\%$ From the Reference Table, we then obtain a critical value c such that $P(T \le c) = \alpha = 5\%$. For $P(T \le \overline{c}) = 1 - \alpha = 95\%$ the table gives $\overline{c} = 1.71$, so that $c = -\overline{c} = -1.71$ because of the symmetry of the distribution shown in **Fig.** 2.3.

We now reason as follows–this is the crucial idea of the test. If the hypothesis is true, we have a chance of only α (= 5%) that we observe a value t of T (calculated from a sample) that will fall between $-\infty$ and -1.71. Hence, if we nevertheless do observe such a t, we start that the hypothesis cannot be true and we reject it.

A simple calculation gives:

$$T = \frac{(107 - 200)}{6/\sqrt{25}} = -2.5,$$

as an observed value of T. Since $-2.5 < -1.71$, we reject the hypothesis, the manufacturer's claim, and accept the alternative result of $\mu = \mu_4 < 200$, which means the wire seems to be weaker than claimed ■

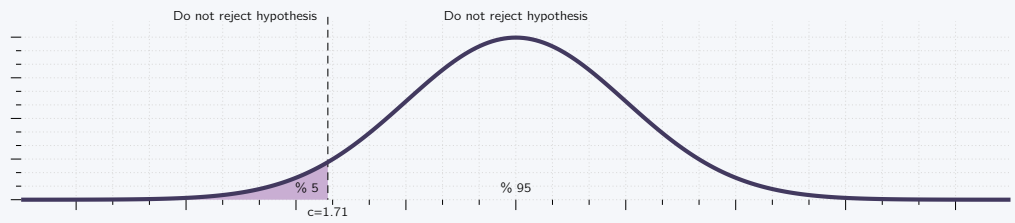


Figure 2.3: The t-distribution used in Example 2.8.

This aforementioned example perfectly captures the steps of a test:

1. Formulate the **hypothesis** $\theta = \theta_0$ to be tested. In our previous example it is $\theta_0 = \mu_0$.
2. Formulate an **alternative** $\theta = \theta_1$, which in our example is $\theta_1 = \mu_1$.
3. Choose a **significance level** α with values such as 5%, 1%, or, 0.1%.
4. Use a random variable $\hat{\Theta} = g(X_1, \cdots, X_n)$ whose distribution depends on the hypothesis and on the alternative, and this distribution is known in both cases.

   Determine a critical value c from the distribution of $\hat{\Theta}$, assuming the hypothesis to be true. In the example, $\hat{\Theta} = T$, and c is, obtained from $P(T \leq c) = \alpha$.
5. Use a sample $x_1, \cdots, x_n$ to determine an observed value $\hat{\theta} = g(x_1, \cdots, x_n)$ of $\hat{\Theta}$, where in our example it is t.
6. Accept or reject the hypothesis, depending on the size of $\hat{\theta}$ relative to c.

There are two (2) important facts require further discussion and careful attention.

1. The choice of an alternative. In the example, $\mu_1 < \mu_0$, but other applications may require $\mu_1 > \mu_0$ or $\mu_1 \neq \mu_0$.
2. Addressing errors. We know that α, the significance level of the test, is the probability of reflecting a true hypothesis. And we shall discuss the probability β of accepting a false hypothesis.

One-Sided and Two-Sided Alternatives

Let θ be an unknown parameter in a distribution, and suppose we want to test the hypothesis $\theta = \theta_0$.

Then there are three (3) main kinds of alternatives, namely,

$$\theta > \theta_0 \tag{2.17}$$

$$\theta < \theta_0 \tag{2.18}$$

$$\theta \neq \theta_0 \tag{2.19}$$

Here Eq. (2.17), and Eq. (2.18) are **one-sided alternatives**, and Eq. (2.19) is a two-sided alternative.

[25] or called the critical region

We call rejection region[25] the region such that we reject the hypothesis if the observed value in the test falls in this region. In [1] the critical c lies to the right of θ_0 because so does the alternative. Hence the rejection region extends to the right. This is called a right-sided test. In [2] the critical c lies to the left of θ_0 (as in Example 1), the rejection region extends to the left, and we have a left-sided test. These are one-sided tests. In [3]

All three kinds of alternatives occur in practical problems. For example, Eq. (2.17) may arise if θ_0 is the maximum tolerable inaccuracy of a voltmeter or some other instrument. Alternative Eq. (2.18) may occur in testing strength of material, as in Example **??**. Finally, θ_0 in Eq. (2.19) may be the diameter of axle-shafts, and shafts that are too thin or too thick are equally undesirable, so that we have to watch for deviations in both directions.

2.4.1 Errors in Tests

Tests always involve risks of making false decisions:

I Rejecting a true hypothesis (Type I error)

- α = Probability of making a Type I error.

II Accepting a false hypothesis (Type II error).

- β = Probability of making a Type II error.

Clearly, we cannot avoid these errors.

> No absolutely certain conclusions about populations can be drawn from samples.

[26] e.g., a small risk $\alpha = \%1$ is used if it is a matter of life or death.

But we show there are ways and means of choosing suitable levels of risks, that is, of values α and β. The choice of α depends on the nature of the problem.[26]

Let us discuss this systematically for a test of a hypothesis $\theta = \theta_0$ against an alternative that is a single number θ_1, for simplicity. We let $\theta_1 > \theta_0$, so that we have a right-sided test. For a left-sided or a two-sided test the discussion is quite similar.

[27] as in the upper part of Fig. 533, by methods discussed below

We choose a critical $c > \theta_0$[27]. From a given sample $x_1, \cdots, x_n$ we then compute a value:

$$\hat{\theta} = g(x_1, \cdots, x_n)$$

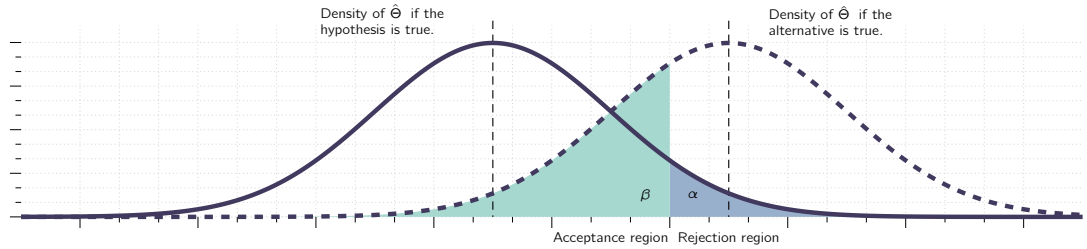


Figure 2.4: Illustration of Type I and II errors in testing a hypothesis $\theta = \theta_0$ against an alternative $\theta = \theta_0$.

with a suitable g.

> whose choice will be a main point of our further discussion; for instance, take $g = (x_1 + \cdots + x_n)/n$ in the case in which θ is the mean.

If $\hat{\theta} > c$, we reject the hypothesis. If $\hat{\theta} \leq c$, we accept it. Here, the value $\hat{\theta}$ can be regarded as an observed value of the random variable

$$\hat{\Theta} = g(X_1, \cdots, X_n)$$

because x_j may be regarded as an observed value of X_j where $j = 1, \cdots, n$. In this test there are two (2) possibilities of making an error, as follows.

Type I Error The hypothesis is true but is rejected[28] because Θ assumes a value $\hat{\theta} > c$. Obviously, the probability of making such an error equals

[28] hence the alternative is accepted.

$$P\left(\hat{\Theta} > c\right)_{\theta} = \theta_0 = \alpha. \tag{2.20}$$

α is called the **significance level** of the test, as mentioned before.

Type II Error The hypothesis is false but is accepted because $\hat{\Theta}$ assumes a value $\hat{\theta} \equiv c$. The probability of making such an error is denoted by β; thus

$$P\left(\hat{\Theta} \leqq c\right)_{\theta=\theta_0} = \beta. \tag{2.21}$$

$\eta = 1 - \beta$ is called the **power** of the test. Obviously, the power η is the probability of avoiding a Type II error.

Formulas Eq. (2.20) and Eq. (2.21) show that both α and β depend on c, and we would like to choose c so that these probabilities of making errors are as small as possible. But the important **Fig.** 2.4 shows that these are conflicting requirements because to let α decrease we must shift c to the right, but then β increases. In practice we first choose α (5%, sometimes 1%), then determine c, and finally compute β. If β is large so that the power $\eta = 1 - \beta$ is small, we should repeat the test, choosing a larger sample, for reasons that will appear shortly. If the alternative is **NOT** a single

number but is of the form Eq. (2.17)-Eq. (2.19), then β becomes a function of θ. This function $\beta(\theta)$ is called the operating characteristic (OC) of the test and its curve the OC curve. Clearly, in this case $\eta = 1 - \beta$ also depends on θ. This function $\eta(\theta)$ is called the power function of the test.

> Of course, from a test that leads to the acceptance of a certain hypothesis θ_0, it does **NOT** follow that this is the only possible hypothesis or the best possible hypothesis. Hence the terms "not reject" or "fail to reject" are perhaps better than the term "accept".

The following example explains the three (3) kinds of hypotheses.

Exercise 2.9: Test for the Mean of the Normal Distribution with Known Variance

Let X be a normal random variable with variance $\sigma^2 = 9$. Using a sample of size $n = 10$ with mean $\overline{x}$, test the hypothesis $\mu = \mu_0 = 24$ against the three (3) kinds of alternatives, namely,

$$\text{(a) } \mu > \mu_0 \qquad \text{(b) } \mu < \mu_0 \qquad \text{(c) } \mu \neq \mu_0$$

Solution

We choose the significance level $\alpha = 0.05$. An estimate of the mean will be obtained from:

$$\overline{X} = \frac{1}{n}\left(X_1 + \cdots + X_n\right).$$

If the hypothesis is true, $\hat{X}$ is normal with mean $\mu = 24$ and variance $\sigma^2/n = 0.9$. Hence we may obtain the critical value c from Table 48 in App. 5.

a. Right-Sided Test We determine c from

$$P\left(\overline{X} > c\right)_{\mu=24} = \alpha = 0.05$$

that is,

$$P\left(\hat{X} \leq c\right)_{\mu=24} = \Phi\left(\frac{c-24}{\sqrt{0.9}}\right) = 1 - \alpha = 0.95.$$

Using Table A8 in App. 5 gives $(c-24)/\sqrt{0.9} = 1.645$, and $c = 25.56$, which is greater than μ_0. If $\overline{x} \leqq 25.56$, the hypothesis is accepted. If $\overline{x} > 25.56$, it is rejected.

The power function of the test is:

$$\eta(\mu) = P(\overline{X} > 25.56)_\mu = 1 - P(\overline{X} \leq 25.56)_\mu$$
$$= 1 - \Phi\left(\frac{25.56-\mu}{\sqrt{0.9}}\right) = 1 - \Phi(26.94 - 1.05\mu)$$

b. Left-Sided Test The critical value c is obtained from the equation

$$P\left(\overline{X} \leq c\right)_{\mu-24} = \Phi\left(\frac{c-24}{\sqrt{0.9}}\right) = \alpha = 0.05.$$

Table A8 in App.5 yields $c = 24 - 1.56 = 22.44$. If $\overline{x} \cong 22.44$, we accept the hypothesis. If $\overline{x} < 22.44$, we reject it. The power function of the test is

$$\eta(\mu) = P\left(\overline{x} \cong 22.44\right)_\mu = \Phi\left(\frac{22.44-\mu}{\sqrt{0.9}}\right) = \Phi(23.65 - 1.05\mu).$$

c. Two-Sided Test Since the normal distribution is symmetric, we choose c_1 and c_2 equidistant from $\mu = 24$, say, $c_1 = 24 - k$ and $c_2 = 24 + k$, and determine k from

$$P\left(24 - k \leqq \overline{X} \leqq 24 + k\right)_{\mu=24} = \Phi\left(\frac{k}{\sqrt{0.9}}\right) - \Phi\left(-\frac{k}{\sqrt{0.9}}\right) = 1 - \alpha = 0.95.$$

Table A8 in App. 5 gives $k/\sqrt{0.9} = 1.960$, hence $k = 1.86$. This gives the values $c_1 = 24 - 1.86 = 22.14$ and $c_2 = 24 + 1.86 = 25.86$. If $\overline{x}$ is not smaller than c_1 and not greater than c_2, we accept the hypothesis. Otherwise, we reject it. The power function of the test is (Fig. 535)

$$\eta(\mu) = P\left(\overline{x} < 22.14\right)_\mu + P\left(\overline{x} > 25.86\right)_\mu = P\left(\overline{x} < 22.14\right)_\mu + 1 - P\left(\overline{x} \leq 25.86\right)_\mu$$

$$= 1 + \Phi\left(\frac{22.14 - \mu}{\sqrt{0.5}}\right) - \Phi\left(\frac{25.86 - \mu}{\sqrt{0.5}}\right)$$

$$= 1 + \Phi(23.34 - 1.05\mu) - \Phi(27.26 - 1.05\mu).$$

Consequently, the operating characteristic $\beta(\mu) = 1 - \eta(\mu)$ (see before) is (Fig. 536).

Exercise 2.10: Comparison of the Means of Two Normal Distributions

Using a sample $x_1, \cdots, x_m$ from a normal distribution with unknown mean μ_x and a sample $y_1, \cdots, y_m$ from another normal distribution with unknown mean μ_y, we want to test the hypothesis that the means are equal, $\mu_x = \mu_y$, again an alternative, say, $\mu_x > \mu_y$. The variances need not be known but are assumed to be equal.

Solution

Two cases of comparing means are of practical importance:

Case A

The sample have the same size. Furthermore, each value of the first sample corresponds to precisely one value of the other, because corresponding values result from the same person or thing (**paired comparison**)

> for example, two measurements of the same thing by two different methods or two measurements from the two cycles of the same person.

More generally, they may result from a circular individual or things, for example, the lower reason is that the lower reason is that the lower value of the second sum of the differences of corresponding values used as the previous that the population corresponds to the differences has mean 0. using the method in Example 3. If we have a choice, this method is better than the following.

Case B

The two samples are independent and not necessarily of the same size. Then we may proceed as follows. Suppose that the alternative is $\frac{1}{2}\ u^n > u_{PV}$ we choose a significance level α. Then we compute the sample means $\overline{x}$ and $\overline{y}$ as well as $(n_1 - 1)u_{PV}^2$ and $(n_2 - 1)u_{PV}^2$ where $\overline{x}^2$ and $\overline{y}^2$ are the sample variances. Using Table 9 in App.5 with $n_1 + n_2 - 2$ degrees of freedom, we now determine c from

2.5 Goodness of Fit

To test for goodness of fit[29] means that we wish to test that a certain function $F(x)$ is the distribution function of a distribution from which we have a sample $x_1, \cdots, x_n$. Then we test whether the **sample distribution function** $\widetilde{F}(x)$ defined as:

$$\widetilde{F}(x) = \text{Sum of the relative frequencies of all sample values } x_j \text{ not exceeding } x, \quad (2.22)$$

fits $\widetilde{F}(x)$ sufficiently well. If this is so, we shall accept the hypothesis that $\widetilde{F}(x)$ is the distribution function of the population; else, we shall **reject the hypothesis**.

> This test is of considerable practical importance, and it differs in character from the tests for parameters (μ, σ^2, etc.) considered thus far.

To test in that fashion, we have to know how much $\widetilde{F}(x)$ can differ from $F(x)$ if the hypothesis is **true**. Hence we must first introduce a quantity which measures the deviation of $\widetilde{F}(x)$ from $F(x)$, and we must know the probability distribution of this quantity under the assumption that the hypothesis is true.

Then we proceed as follows.

We determine a number, lets use c, such that, if the hypothesis is **true**, a deviation greater than c has a small preassigned probability. If, nevertheless, a deviation greater than c occurs, we have reason to doubt that the hypothesis is true and we reject it. On the other hand, if the deviation does not exceed c, so that $\widetilde{F}(x)$ approximates $F(x)$ sufficiently well, we accept the hypothesis. Of course, if we accept the hypothesis, this means that we have insufficient evidence to reject it, and this does not exclude the possibility that there are other functions that would not be rejected in the test.

In this respect the situation is quite similar to hypothesis testing we talked previously.

The following text-block shows a test of that type, which was introduced by *R. A. Fisher*. This test is justified by the fact that if the hypothesis is true, then χ_0^2 is an observed value of a random variable whose distribution function approaches that of the chi-square distribution with $K-1$ degrees of freedom[30] as n approaches infinity. The requirement that at least five (5) sample values lie in each interval results from the fact that for finite n that random variable has only approximately a chi-square distribution.

> If the sample is so small that the requirement cannot be satisfied, one may continue with the test, but then use the result **with caution**.

[29] In literature, this method also means χ^2-Test.

Historical Anecdote

During the 19th century, statistical analytical methods were mainly applied to biological data and it was customary for researchers to assume observations followed a normal distribution, such as Sir George Airy and Mansfield Merriman, whose works were criticized by Karl Pearson in his 1900 paper.

At the end of the 19th century, Pearson noticed the existence of significant skewness within some biological observations. To model the observations regardless of being normal or skewed, Pearson, in a series of articles published from 1893 to 1916,[3][4][5][6] devised the Pearson distribution, a family of continuous probability distributions, which includes the normal distribution and many skewed distributions, and proposed a method of statistical analysis consisting of using the Pearson distribution to model the observation and performing a test of goodness of fit to determine how well the model really fits to the observations.

[30] or $K-r-1$ degrees of freedom if r parameters are estimated.

Chi-square Test for F(x) being the Distribution Function of a Population

1. Subdivide the x-axis into n intervals $I_1, \cdots, I_n$ such that each interval contains at least five (5) values of the given sample $x_1, \cdots, x_n$.

   Determine the number b_j of sample values in the interval I_j, where $j = 1, \cdots, K$. If a sample value lies at a common boundary point of two (2) intervals, add 0.5 to each of the two (2) corresponding b_j.

2. Using $F(x)$, calculate the probability p_j that the random variable X under consideration assumes any value in the interval I_j, where $j = 1, \cdots, K$. Then, calculate

   $$e_j = np_j.$$

   This is the number of sample values **theoretically expected** in I_j if the hypothesis is true.

3. Compute the deviation:

   $$\chi_0^2 = \sum_{j=1}^{K} \frac{(b_j - e_j)^2}{e_j}.$$

4. Choose a significance level such as 5%, 1%, or the like.

5. Determine the solution c of the equation

   $$P\left(\chi^2 \leqq c\right) = 1 - \alpha.$$

   from the table of the chi-square distribution with $K - 1$ degrees of freedom (Table A10 in App. 5).

If r parameters of $F(x)$ are unknown and their maximum likelihood estimates are used, then use $K - r - 1$ degrees of freedom, instead of $K - 1$.

If $\chi_0^2 \leqq c$, accept the hypothesis. If $\chi_0^2 > c$, reject the hypothesis.

Exercise 2.11: Test of Normality

Test whether the population from which the sample in table given below was taken is normal.

320	380	340	410	380	340	360	350	320	370
350	340	350	360	370	350	380	370	300	420
370	390	390	440	330	390	330	360	400	370
320	350	360	340	340	350	350	390	380	340
400	360	350	390	400	350	360	340	370	420
420	400	350	370	330	320	390	380	400	370
390	330	360	380	350	330	360	300	360	360
360	390	350	370	370	350	390	370	370	340
370	400	360	350	380	380	360	340	330	370
340	360	390	400	370	410	360	400	340	360

Solution

The table given in the question shows the values, column by column, in the order obtained in the experiment. The next table gives the frequency distribution and Fig. 542 the histogram.

The maximum likelihood estimates for μ and σ^2 are $\hat{\mu} = \overline{x} = 364.7$ and $\overline{\sigma}^2 = 712.9$. The computation in Table 25.10 yields $\overline{x}_0^2 = 2.688$. It is very interesting that the interval $375 \cdots 385$ contributes over 50% of $\overline{x}_0^2$. From the histogram we see that the corresponding frequency looks much too small. The second largest contribution comes from $395 \cdots 405$, and the histogram shows that the frequency seems somewhat too large, which is perhaps not obvious from inspection.

Tensile Strength	Absolute Freq.	Relative Freq.	Cumulative Absolute Freq.	Cumulative Relative Freq.
300	2	0.02	2	0.02
310	0	0.00	2	0.02
320	4	0.04	6	0.06
330	6	0.06	12	0.12
340	11	0.11	23	0.23
350	14	0.14	37	0.37
360	16	0.16	53	0.53
370	15	0.15	68	0.68
380	8	0.08	76	0.76
390	10	0.10	86	0.86
400	8	0.08	94	0.94
410	2	0.02	96	0.96
420	3	0.03	99	0.99
430	0	0.00	99	0.99
440	1	0.01	100	1.00

Table 2.1: Frequency table of the sample given in the question.

We choose $\alpha = 5\%$. Since $K = 10$ and we estimated $r = 2$ parameters we have to use Table A10 in App. 5 with $K - r - 1 = 7$ degrees of freedom. We find $c = 14.07$ as the solution of $P(\chi^2 \cong c) = 95\%$. Since $\chi_0^2 < c$, we accept the hypothesis that the population is normal.

2.6 Regression and Correlation

Up to this points, we were only concerned with random experiments in which we observed a single quantity[31] and got samples whose values were single numbers. In this section we discuss experiments in which we observe or measure two (2) quantities simultaneously, so we get samples of pairs of values:

[31] In this case it is a random variable.

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

Most applications involve one of two kinds of experiments, which are as follows:

Regression one (1) of the two (2) variables, call it x, can be regarded as an ordinary variable because we can measure it without substantial error or we can even give it values we want. x is called the **independent variable**, or sometimes the **controlled variable** as we can control it.[32] The other variable, Y, is a random variable, and we are interested in the dependence of Y on x.

[32] set it at values we choose

Examples include the dependence of the blood pressure Y on the age x of a person or, as we shall now say, the regression of Y on x. The regression of the gain of weight Y of certain animals on the daily ratio of food x, the regression of the heat conductivity Y of work on the specific weight x of the rock, etc.

Correlation both quantities are random variables and we are interested in relations between them.

Examples are the relation[33] between user X and year Y of the front tires of cars, between grades X and Y of students in mathematics and in physics, respectively, between the hardness X of steel plates in the centre and the hardness Y near the edges of the plates, etc.

[33] we say correlation

2.6.1 Regression Analysis

In regression analysis the dependence of Y on x is a dependence of the mean μ of Y on x, so that $\mu = \mu(x)$ is a function in the ordinary sense. The curve of $\mu(x)$ is called the **regression curve** of Y on x.

Let's look into the simplest case, namely, that of a straight regression line:

$$\mu(x) = \kappa_0 + \kappa_1 x. \tag{2.23}$$

Then we may want to graph the sample values as n points in the xY-plane, fit a straight line through them, and use it for estimating $\mu(x)$ at values of x that interest us, so we know what values of Y we can expect for those x.

Fitting line by eye would not be good because it would be subjective as people would come up with different estimations, particularly if the points are scattered. So we need a mathematical method

which gives a unique result depending **only** on the n points. A widely used procedure is the method of least squares by Gauss and Legendre.

For our task we may formulate it as follows.

Theory 2.23: Least Square Principle

The straight line should be fitted through the given points so that the sum of the squares of the distances of those points from the straight line is **minimum**, where the distance is measured in the vertical direction, which is the y-direction.

To get uniqueness of the straight line, we need some extra condition.

To see this, lets look at a small example and take the sample $(0,\,1)\,,\,(0,-1)$. Then all the lines $y = k_1 x$ with any k_1 satisfy the principle.

The following assumption will imply uniqueness, as we shall find out.

Theory 2.24: Assumption A1

The x-values $x_1, \cdots, x_n$ in our sample $(x_1, y_1), \cdots, (x_n, y_n)$ are not all equal.

From a given sample $(x_1,\,y_1)\,,\,(x_2,\,y_2)\,,\,\cdots,\,(x_n,\,y_n)$ we shall now determine a straight line by least squares. We write the line as:

$$y = k_0 + k_1 x, \tag{2.24}$$

and call it the sample regression line as it will be the counterpart of the population regression line given in Eq. (2.23).

[34] distance measured in the y-direction

Now a sample point $\left(x_j,\,y_j\right)$ has the vertical distance[34] from Eq. (2.24) given by:

$$\left|y_j - \left(k_0 + k_1 x_j\right)\right| \tag{2.25}$$

This gives the sum of the squares of these distances as:

$$q = \sum_{j=1}^{n} \left(y_j - k_0 - k_1 x_j\right)^2 \tag{2.26}$$

In the method of least squares we now have to determine k_0 and k_1 such that q is minimum. From calculus we know that a necessary condition for this is:

$$\frac{\partial q}{\partial k_0} = 0 \qquad \text{and} \qquad \frac{\partial q}{\partial k_1} = 0. \tag{2.27}$$

We shall see that from this condition we obtain for the sample regression line the formula

$$y - y \;= k_1\left(x - \overline{x}\right). \tag{2.28}$$

Here $\overline{x}$ and $\overline{y}$ are the means of the x- and the y-values in our sample, that is,

$$\overline{x} = \frac{1}{n}\left(x_1 + \cdots + x_n\right) \qquad \text{and} \qquad \overline{y} = \frac{1}{n}\left(y_1 + \cdots + y_n\right). \tag{2.29}$$

The slope k_1 in Eq. (2.28) is called the **regression coefficient** of the sample and is given by:

$$k_1 = \frac{s_{xy}}{s_x^2}. \tag{2.30}$$

Here the **sample covariance** s_{xy} is:

$$s_{xy} = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_j - \overline{x}\right)\left(y_j - \overline{y}\right) = i\frac{1}{n-1}\left[\sum_{j=1}^{n} x_j y_j - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{j=1}^{n} y_j\right)\right] \tag{2.31}$$

and s_x^2 is given by

$$s_x^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_j - \overline{x}\right)^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n} x_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)^2\right]. \tag{2.32}$$

From Eq. (2.28) we see that the sample regression line passes through the point $(\overline{x}, \overline{y})$, by which it is determined, together with the regression coefficient Eq. (2.30). We may call s_x^2 the *variance* of the x-values, but keep in mind that x is an ordinary variable, and **NOT** a random variable.

We shall soon also need:

$$s_y^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(y_j - \overline{y}\right)^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n} y_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} y_j\right)^2\right]. \tag{2.33}$$

Now, let's try to derive Eq. (2.28) and Eq. (2.30). Differentiating Eq. (2.26) and using Eq. (2.27), we first obtain:

$$\frac{\partial q}{\partial k_0} = -2\sum\left(y_j - k_0 - k_1 x_j\right) = 0,$$
$$\frac{\partial q}{\partial k_1} = -2\sum x_j\left(y_j - k_0 - k_1 x_j\right) = 0.$$

where we sum over j from 1 to n. We now divide by 2, write each of the two sums as three sums, and take the sums containing y_j and $x_j y_j$ over to the right.

Then we get the normal equations:

$$\begin{aligned} k_0 n + k_1 \sum x_j &= \sum y_j \\ k_0 \sum x_j + k_1 \sum x_j^2 &= \sum x_j y_j. \end{aligned} \tag{2.34}$$

This is a linear system of two (2) equations with two unknowns k_0, k_1, with coefficient determinant being:

$$\begin{vmatrix} n & \sum x_j \\ \sum x_j & \sum x_j^2 \end{vmatrix} = n\sum x_j^2 - \left(\sum x_j\right)^2 = n(n-1)\,s_x^2 = n\sum\left(x_j - \overline{x}\right)^2,$$

and is **NOT** zero due to the first assumption (A1) we made prior. Hence the system has a unique solution. Dividing the first equation of Eq. (2.34) by n and using Eq. (2.29), we get $k_0 = \overline{y} - k_1\overline{x}$.

Together with $y = k_0 + k_1 x$ in Eq. (2.24) this gives Eq. (2.28). To get Eq. (2.30), we solve the system Eq. (2.34) by Cramer's rule or elimination, finding

$$k_1 = \frac{n\sum x_j y_j - \sum x_i \sum y_j}{n\,(n-1)\,s_x^2}. \tag{2.35}$$

Which completes the derivation.

Exercise 2.12: Regression Line

The decrease of volume y (in %) of leather for certain fixed values of high pressure x (atmosphere) was measured. The results are shown in the first two columns of the table below.

Given Values		Auxiliary Values	
x_j	y_j	x_j^2	$x_j y_j$
4000	2.3	16,000,000	9200
6000	4.1	36,000,000	24,600
8000	5.7	64,000,000	45,600
10,000	6.9	100,000,000	69,000
28,000	19.0	216,000,000	148,400

Table 2.2: Dataset

Find the regression line of y on x.

Solution

We see that the sample count is $n = 4$ and obtain the values

$$\overline{x} = \frac{28000}{4} = 7000 \quad \text{and} \quad \overline{y} = \frac{19.0}{4} = 4.75,$$

and from Eq. (2.32), Eq. (2.33) and Eq. (2.31)

$$s_x^2 = \frac{1}{3}\left(216,000,000 - \frac{28,000^2}{4}\right) = \frac{20,000,000}{3}$$
$$s_{xy} = \frac{1}{3}\left(148,400 - \frac{28,000 \cdot 19}{4}\right) = \frac{15,400}{3}.$$

Hence $k_1 = 15,400/20,000,000 = 0.00077$ from Eq. (2.30), and the regression line is

$$y - 4.75 = 0.00077\,(x - 7000)$$
$$y = 0.00077x - 0.64.$$

With both options being valid. Note $y\,(0) = -10.65$, which is physically meaningless, but typically indicates that a linear relation is merely an approximation valid on some restricted interval ■

2.6.2 Confidence Intervals

[35] which we have not made so far; least squares is a *geometric principle*, not involving probabilities.

If we want to get confidence intervals, we have to make assumptions about the distribution of Y.[35] We assume normality and independence in sampling:

Theory 2.25: Assumption A2

For each fixed x the random variable Y is normal with mean Eq. (2.23), that is,

$$\mu(x) = \kappa_0 + \kappa_1 x \tag{2.36}$$

and variance σ^2 **independent** of x.

Theory 2.26: Assumption A3

The n performances of the experiment by which we obtain a sample

$$(x_1, y_1), \quad (x_2, y_2), \quad \cdots, \quad (x_n, y_n) \tag{2.37}$$

are independent

κ_1 given in Eq. (2.36) is called the **regression coefficient** of the population because it can be shown that, under the assumptions given in A1 to A3, the maximum likelihood estimate of κ_1 is the sample regression coefficient k_1 given by Eq. (2.35).

Following with the assumptions from A1 to A3, we may now obtain a confidence interval for κ_1, as shown below.

Determination of Regression Coefficient under Assumptions A1 to A3

1. Choose a confidence level γ which can take values of 95%, 99%, or the like.
2. Determine the solution c of the equation,

$$F(c) = \frac{1}{2}(1+\gamma) \tag{2.38}$$

   from the table of the t-distribution with $n-2$ degrees of freedom (Table A9 in App. 5; n = sample size)

3. Using a sample $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, calculate $(n-1)\,s_x^2$ from Eq. (2.32), $(n-1)s_{xy}$ from Eq. (2.31), k_1 from Eq. (2.30),

$$(n-1)\,s_y^2 = \sum_{j=1}^{n} y_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} y_j\right)^2$$

   which was described in Eq. (2.33), and

$$q_0 = (n-1)\left(s_y^2 - k_1^2 s_x^2\right)$$

4. Calculate:

$$K = c\sqrt{\frac{q_0}{(n-2)\,(n-1)\,s_x^2}}.$$

5. The confidence interval is then defined to be:

$$\mathrm{CONF}_\gamma\left\{k_1 - K \leqq \kappa_1 \leqq k_1 + K\right\}. \tag{2.39}$$

Exercise 2.13: Confidence Interval for the Regression Coefficient

Using the sample in **Table** 2.2, determine a confidence interval for κ_1 by the method described just previously.

Solution

1. We choose $\gamma = 0.95$.
2. Eq. (2.38) takes the form $F(c) = 0.975$, and Table A9 in App. 5 with $n - 2 = 2$ degrees of freedom gives $c = 4.30$.
3. From Example 1 we have $3s_x^2 = 20{,}000{,}000$ and $k_1 = 0.00077$. From **Table** 2.2 we compute:

$$3s_y^2 = 102.0 - \frac{19^2}{4} = 11.95.$$
$$q_0 = 11.95 - 20,000,000 \cdot 0.00077^2 = 0.092.$$

4. We therefore obtain:

$$K = 4.30\sqrt{\frac{0.092}{(2 \cdot 20,000,000}}$$
$$= 0.000206$$
$$\text{CONF}_{0.95}\left\{0.00056 \leqq \kappa_1 \leqq 0.00098\right\} \quad \blacksquare$$

2.6.3 Correlation Analysis

Time to give an introduction to the basic facts in correlation analysis.

The topic of **correlation analysis** is concerned with the relation between X and Y in a two-dimensional random variable (X, Y). A sample consists of n ordered pairs of values:

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n),$$

as before. The interrelation between the x and y values in the sample is measured by the sample covariance s_{xy} in Eq. (2.31) or by the sample correlation coefficient:

$$r = \frac{s_{xy}}{s_x s_y} \tag{2.40}$$

with s_x and s_y given in Eq. (2.32) and Eq. (2.33). Here r has the advantage that it does not change under a multiplication of the x and y values by a factor.[36]

[36] Such as the unit changing from g to kg.

Theory 2.27: Sample Correlation Coefficient

The sample correlation coefficient r satisfies $-1 \leqq r \leqq 1$.

In particular, $r = \pm 1$ if and only if the sample values lie on a straight line which can be seen in **Fig.** 2.5.

The theoretical counterpart of r is the correlation coefficient ρ of X and Y,

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{2.41}$$

where:

$$\mu_X = E(X) \qquad \mu_Y = E(Y) \qquad \sigma_X^2 = E\left(\left[X - \mu_X\right]^2\right) \qquad \sigma_Y^2 = E\left(\left[Y - \mu_Y\right]^2\right)$$

which are the means and variances of the marginal distributions of X and Y. The covariance (σ_{XY}), on the other hand, is defined as:

$$\sigma_{XY} = E\left(\left[X - \mu_X\right]\left[Y - \mu_Y\right]\right) = E(XY) - E(X)\,E(Y) \tag{2.42}$$

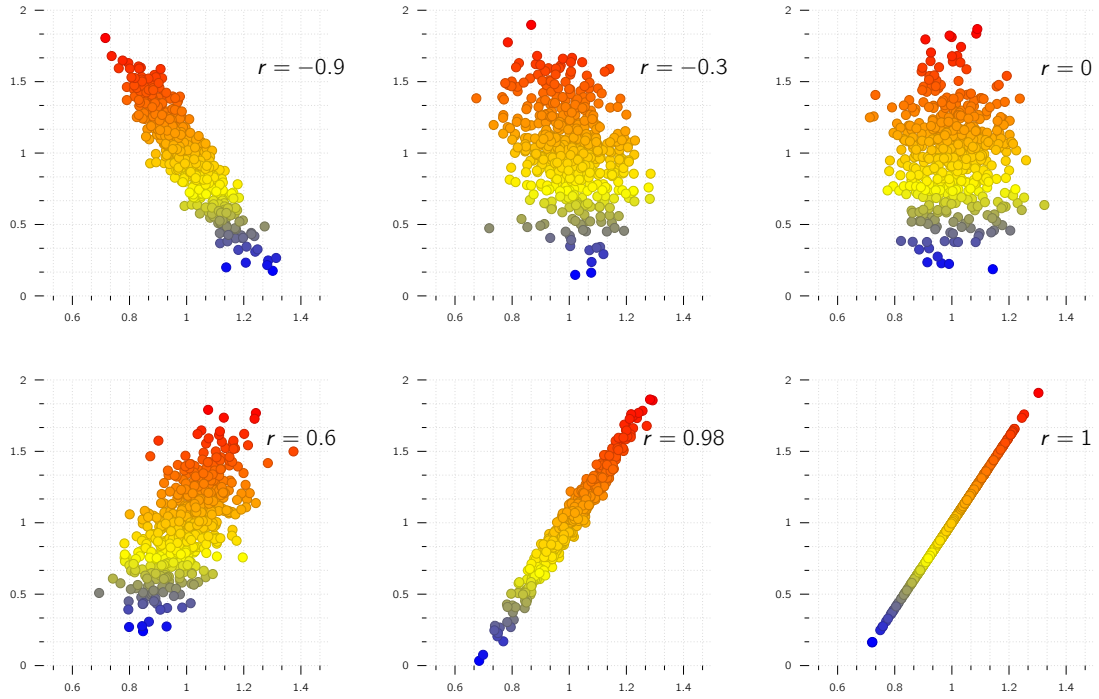


Figure 2.5: Samples with various values of the correlation coefficient r.

Theory 2.28: Correlation Coefficient

The correlation coefficient ρ satisfies $-1 \leqq \rho \leqq 1$.

In particular, $\rho = \pm 1$ if and only if X and Y are **linearly related**, that is,

$$Y = \gamma X + \delta, X = \gamma_* Y + \delta_*.$$

X and Y are **uncorrelated** if $\rho = 0$.

Theory 2.29: Independence and Relation to Normal Distribution

a. If X and Y are independent, they are uncorrelated.

b. If (X, Y) is normal, then uncorrelated X and Y are **independent**.

Here the two-dimensional normal distribution can be introduced by taking two (2) independent standardised normal random variables X^*, Y^*, whose joint distribution thus has the density:

$$f^*\left(x^*, y^*\right) = \frac{1}{2\pi} e^{-\left(x^{*2} + y^{*2}\right)/2}$$

and setting

$$X = \mu_X + \sigma_X X^*$$
$$Y = \mu_Y + \rho\sigma_Y X^* + \sqrt{1-\rho^2}\sigma_Y Y^*$$

This gives the general two-dimensional normal distribution with the density:

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-h(x, y)/2} \tag{2.43}$$

where

$$h(x, y) = \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 x\right] \tag{2.44}$$

In Theorem 3(b), normality is important, as we can see from the following example.

Exercise 2.14: Uncorrelated but Dependent Random Variables

If X assumes the value of -1, 0, 1 with probability $\frac{1}{3}$ and $Y = X^2$, then $EX = 0$ and in Eq. (2.26)

$$\sigma_{XY} = E(XY) = E\left(X^3\right) = (-1)^3 \cdot \frac{1}{3} + 0^3 \cdot \frac{1}{3} + 1^2 \cdot \frac{1}{3} = 0,$$

so that $\rho = 0$ and X and Y are uncorrelated. But they are certainly **NOT** independent since they are even functionally related.

2.6.4 Test for the Correlation Coefficient

The following text-block shows a test for ρ in the case of the two-dimensional normal distribution. t is an observed value of a random variable that has a t-distribution with $n-2$ degrees of freedom.[37]

[37] This was shown by R. A. Fisher.

Testing the Hypothesis against the Alternative in case of Two-Dimensional Normal Distribution

1. Choose a significance level α (5%, 1%, or the like).
2. Determine the solution c of the equation:

$$P\left(T \leqq c\right) = 1 - \alpha,$$

   from the t-distribution (Table A9 in App. 5) with $n-2$ degrees of freedom.
3. Calculate r from Eq. (2.40), using a sample $(x_1, y_1), \cdots, (x_n, y_n)$.
4. Calculate

$$t = r\left(\sqrt{\frac{n-2}{1-r^2}}\right).$$

5. If $t \leqq c$, accept the hypothesis. If $t > c$, reject the hypothesis.

Exercise 2.15: Test for the Correlation Coefficient

Test the hypothesis $\rho = 0$ (independence of X and Y, because of Theorem 3) against the alternative $\rho > 0$, using the data when $r = 0.6$ (normal soliding errors on 10 two-sided circuit boards done by 10 workers $x =$ front, $y =$ back of the bonds).

Solution

We choose $\alpha = 5\%$; which makes $1 - \alpha = 95\%$. Since $n = 10$, and $n - 2 = 8$, the table gives $c = 1.86$. Also,

$$t = 0.6\sqrt{\frac{8}{0.64}} = 2.12 > c.$$

We reject the hypothesis and search that there is a positive correlation. A worker making few errors on the front side also tends to make few errors on the reverse side of the bond ■