THE UNIVERSITY OF QUEENSLAND

Statistical Modelling and Analysis (STAT2004)

Professor Geoff McLachlan
School of Mathematics
https://people.smp.uq.edu.au/GeoffMcLachlan/

*The background notes for STAT2004 this semester are based on those prepared by Professor Dirk Kroese for his lectures in the course in 2016.

*There have been some minor modifications made in places such as to the notation.

*Additional material where required will be in supplementary notes to be posted during the semester.

Contents

Part I

Probability

Chapter 1

Random experiments and probability models

In Part I of the course we consider the *probability* side of statistics. In particular, we will review (or learn) how "random experiments" can be modelled and how such models allow us to derive various probabilities and other properties for those experiments.

1.1 Random experiments

The basic notion in probability is that of a **random experiment**: an experiment whose outcome cannot be determined in advance, but is nevertheless still subject to analysis.

Examples of random experiments are:

1. tossing a die,
2. measuring the amount of rainfall in Brisbane in January,
3. counting the number of calls arriving at a telephone exchange during a fixed time period,
4. selecting a random sample of fifty people and observing the number of left-handers,
5. choosing at random ten people and measuring their height.

We wish to describe these experiments via a mathematical model. This model consists of three building blocks: a *sample space*, a set of *events* and a *probability*. We will now describe each of these objects.

1.2 Sample spaces

Although we cannot predict the outcome of a random experiment with certainty we usually can specify a set of possible outcomes. This gives the first ingredient in our model for a random experiment.

Definition 1.1 The **sample space** Ω of a random experiment is the set of all possible outcomes of the experiment.

Examples of random experiments with their sample spaces are:

1. Cast two dice consecutively,
$$\Omega = \{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(6,6)\}.$$
2. The lifetime of a machine (in days),
$$\Omega = \mathbb{R}_+ = \{ \text{ positive real numbers } \}.$$
3. The number of arriving calls at an exchange during a specified time interval,
$$\Omega = \{0,1,\cdots\} = \mathbb{Z}_+ .$$
4. The heights of 10 selected people.
$$\Omega = \{(x_1,\ldots,x_{10}), x_i \geq 0, i = 1,\ldots,10\} .$$
Here $(x_1,\ldots,x_{10})$ represents the outcome that the length of the first selected person is x_1, the length of the second person is x_2, etcetera.

Notice that for modelling purposes it is often easier to take the sample space larger than necessary. For example the actual lifetime of a machine would certainly not span the entire positive real axis. And the heights of the 10 selected people would not exceed 3 metres.

1.3 Events

Often we are not interested in a single outcome but in whether or not one of a *group* of outcomes occurs. Such subsets of the sample space are called **events**. Events will be denoted by capital letters $A, B, C, \ldots$. We say that event A **occurs** if the outcome of the experiment is one of the elements in A.

Examples of events are:

1. The event that the sum of two dice is 10 or more,
$$A = \{(5,5),(5,6),(6,5),(6,6)\}.$$
2. The event that a machines lives less than 1000 days,
$$A = [0, 1000)\ .$$
3. The event that out of fifty selected people, five are left-handed,
$$A = \{5\}\ .$$

Since events are sets, we can apply the usual set operations to them:

1. the set $A \cup B$ (A **union** B) is the event that A *or* B *or* both occur,
2. the set $A \cap B$ (A **intersection** B) is the event that A *and* B both occur,
3. the event A^c (A **complement**) is the event that A does *not* occur,
4. if $A \subset B$ (A is a **subset** of B) then event A is said to *imply* event B.

Two events A and B which have no outcomes in common, that is, $A \cap B = \emptyset$, are called **disjoint** events.

Example 1.1 Suppose we cast two dice consecutively. The sample space is $\Omega = \{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(6,6)\}$. Let $A = \{(6,1),\ldots,(6,6)\}$ be the event that the first die is 6, and let $B = \{(1,6),\ldots,(1,6)\}$ be the event that the second dice is 6. Then $A \cap B = \{(6,1),\ldots,(6,6)\} \cap \{(1,6),\ldots,(6,6)\} = \{(6,6)\}$ is the event that both die are 6.

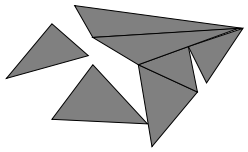
1.4 Probability

The third ingredient in the model for a random experiment is the specification of the probability of the events. It tells us how *likely* it is that a particular event will occur.

Definition 1.2 A probability P is a rule (function) which assigns a number between 0 and 1 to each event, and which satisfies the following **axioms**:

Axiom 1: $0 \leq P(A) \leq 1$.
Axiom 2: $P(\Omega) = 1$.
Axiom 3: For any sequence $A_1, A_2, \ldots$ of *disjoint* events we have

$$\boxed{P(\bigcup_i A_i) = \sum_i P(A_i)}. \tag{1.1}$$



Note that a probability rule P has exactly the same properties as the common "area measure". For example, the total area of the union of the triangles in the figure on the right is equal to the sum of the areas of the individual triangles. This is how you should interpret property (1.1). But instead of measuring areas, P measures probabilities.

As a direct consequence of the axioms we have the following properties for P. Below, A and B are events.

1. $P(A^c) = 1 - P(A)$.
2. $A \subset B \Longrightarrow P(A) \leq P(B)$.
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Exercise 1.1 Prove this.

We have now completed our model for a random experiment. It is up to the modeller to specify the sample space Ω and probability measure P which most closely describes the actual experiment. This is not always as straightforward as it looks, and sometimes it is useful to model only certain *observations* in the experiment. This is where *random variables* come into play, and we will discuss these in the next chapter.

Example 1.2 Consider the experiment where we throw a fair die. How should we define Ω and P?

Obviously, $\Omega = \{1, 2, \ldots, 6\}$; and some common sense shows that we should define P by

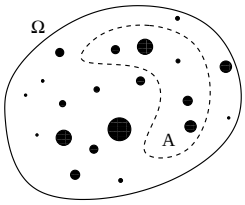
$$P(A) = \frac{|A|}{6}, \quad A \subset \Omega,$$

where $|A|$ denotes the number of elements in set A. For example, the probability of getting an even number is $P(\{2, 4, 6\}) = 3/6 = 1/2$.

In many applications the sample space is *countable*, i.e. $\Omega = \{a_1, a_2, \ldots, a_n\}$ or $\Omega = \{a_1, a_2, \ldots\}$. Such a sample space is called **discrete**.

The easiest way to specify a probability P on a discrete sample space is to specify first the probability p_i of each **elementary event** $\{a_i\}$ and then to define

$$P(A) = \sum_{i:a_i \in A} p_i \ , \quad \text{for all} \ \ A \subset \Omega.$$



This idea is graphically represented in the figure on the right. Each element a_i in the sample is assigned a probability weight p_i represented by a black dot. To find the probability of the set A we have to sum up the weights of all the elements in A.

Again, it is up to the modeller to properly specify these probabilities. Fortunately, in many applications all elementary events are *equally likely*, and thus the probability of each elementary event is equal to 1 divided by the total number of elements in Ω. E.g., in Example 1.2 each elementary event has probability 1/6.

When the sample space is not countable, for example $\Omega = \mathbb{R}_+$, it is said to be **continuous**.

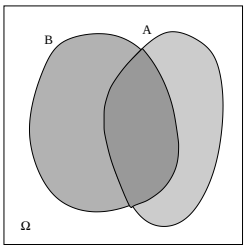
Example 1.3 We draw at random a point in the interval $[0, 1]$. Each point is equally likely to be drawn. How do we specify the model for this experiment?

The sample space is obviously $\Omega = [0, 1]$, which is a continuous sample space. We cannot define P via the elementary events $\{x\}$, $x \in [0, 1]$ because each of these events must have probability 0 (!). However we can define P as follows: For each $0 \leq a \leq b \leq 1$, let

$$P([a, b]) = b - a \ .$$

This completely specifies P. In particular, we can find the probability that the point falls into any (sufficiently nice) set A as the *length* of that set.

1.5 Conditional probability and independence



How do probabilities change when we know some event $B \subset \Omega$ has occurred? Suppose B has occurred. Thus, we know that the outcome lies in B. Then A will occur if and only if $A \cap B$ occurs, and the relative chance of A occurring is therefore

$$P(A \cap B)/P(B).$$

This leads to the definition of the **conditional probability** of A given B:

$$\boxed{P(A \,|\, B) = \frac{P(A \cap B)}{P(B)}}$$

Example 1.4 We throw two dice. Given that the sum of the eyes is 10, what is the probability that one 6 is cast?

Let B be the event that the sum is 10,

$$B = \{(4,6),(5,5),(6,4)\}.$$

Let A be the event that one 6 is cast,

$$A = \{(1,6),\ldots,(5,6),(6,1),\ldots,(6,5)\}.$$

Then, $A \cap B = \{(4,6),(6,4)\}$. And, since all elementary events are equally likely, we have

$$P(A \,|\, B) = \frac{2/36}{3/36} = \frac{2}{3}.$$

1.5.1 Chain rule

By the definition of conditional probability we have

$$P(A \cap B) = P(A)P(B \,|\, A).$$

We can generalise this to n intersections $A_1 \cap A_2 \cap \cdots \cap A_n$, which we abbreviate as $A_1 A_2 \cdots A_n$. This gives the **chain rule** of probability. The proof is by induction.

Theorem 1.1 (Chain rule) Let $A_1, \ldots, A_n$ be a sequence of events with $P(A_1 \ldots A_{n-1}) > 0$. Then,

$$\boxed{P(A_1 \cdots A_n) = P(A_1)\, P(A_2 \,|\, A_1)\, P(A_3 \,|\, A_1 A_2) \cdots P(A_n \,|\, A_1 \cdots A_{n-1}).} \quad (1.2)$$

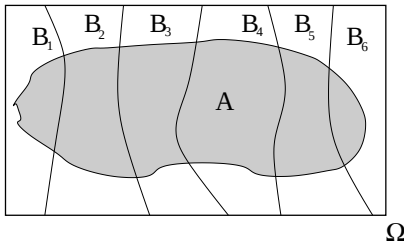
Example 1.5 We draw consecutively 3 balls from a bowl with 5 white and 5 black balls, without putting them back. What is the probability that all balls will be black?

Solution: Let A_i be the event that the ith ball is black. We wish to find the probability of $A_1 A_2 A_3$, which by (1.2) is

$$P(A_1)\, P(A_2 \,|\, A_1)\, P(A_3 \,|\, A_1 A_2) = \frac{5}{10}\,\frac{4}{9}\,\frac{3}{8} = 0.083.$$

1.5.2 Law of total probability and Bayes' rule

Suppose $B_1, B_2, \ldots, B_n$ is a **partition** of Ω. That is, $B_1, B_2, \ldots, B_n$ are disjoint and their union is Ω.



Then, by the third Axiom, $P(A) = \sum_{i=1}^{n} P(A \cap B_i)$ and hence, by the definition of conditional probability we have

$$\boxed{P(A) = \sum_{i=1}^{n} P(A|B_i)\, P(B_i)}$$

This is called the **law of total probability**.

Combining the Law of Total Probability with the definition of conditional probability gives **Bayes' Rule**:

$$\boxed{P(B_j|A) = \frac{P(A|B_j)\, P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}}$$

Example 1.6 A company has three factories (1, 2 and 3) that produce the same chip, each producing 15%, 35% and 50% of the total production. The probability of a defective chip at 1, 2, 3 is 0.01, 0.05, 0.02, respectively. Suppose someone shows us a defective chip. What is the probability that this chip comes from factory 1?

Let B_i denote the event that the chip is produced by i. The B_i's form a partition of S. Let A denote the event that the chip is faulty. By Bayes' rule,

$$P(B_1|A) = \frac{0.15 \times 0.01}{0.15 \times 0.01 + 0.35 \times 0.05 + 0.5 \times 0.02} = 0.052\ .$$

1.5.3 Independence

Independence is a very important concept in probability and statistics. Loosely speaking it models the *lack of information* between events. We say A and B are *independent* if the knowledge that A has occurred does not change the *probability* that B occurs. That is

$$A,\ B \text{ independent} \Leftrightarrow P(A|B) = P(A)$$

Since $P(A|B) = P(A \cap B)/P(B)$ an alternative definition of independence is

$$\boxed{A,\ B \text{ independent} \Leftrightarrow P(A \cap B) = P(A)P(B)}$$

This definition covers the case $B = \emptyset$ (empty set). We can extend the definition to arbitrary many events:

Definition 1.3 The events $A_1, A_2, \ldots,$ are said to be **independent** if for any k and any choice of distinct indices $i_1, \ldots, i_k$,

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}) \ .$$

Remark 1.1 In most cases independence of events is a **model assumption**. That is, we assume that there exists a P such that certain events are independent.

Example 1.7 We flip a coin n times. We can write the sample space as the set of binary n-tuples:

$$\Omega = \{(0, \ldots, 0), \ldots, (1, \ldots, 1)\} \ .$$

Here 0 represent Tails and 1 represents Heads. For example the outcome $(0, 1, 0, 1, \ldots)$ means that the first time Tails is thrown, the second time Heads, the third times Tails, the fourth time Heads, etc.

How should we define P? Let A_i denote the event of Heads during the ith throw, $i = 1, \ldots, n$. Then, P should be such that the events $A_1, \ldots, A_n$ are *independent*. And, moreover, $P(A_i)$ should be the same for all i. We don't know whether the coin is fair or not, but we can call this probability p $(0 \leq p \leq 1)$.

These two rules completely specify P. For example, the probability that the first k throws are Heads and the last $n - k$ are Tails is

$$\begin{aligned} P(\{(1, 1, \ldots, 1, 0, 0, \ldots, 0)\}) &= P(A_1) \cdots P(A_k) \cdots P(A_{k+1}^c) \cdots P(A_n^c) \\ &= p^k(1-p)^{n-k}. \end{aligned}$$

Also, let B_k be the event that there are k Heads in total. The probability of this event is the sum the probabilities of elementary events $\{(x_1, \ldots, x_n)\}$ such that $x_1 + \cdots + x_n = k$. Each of these events has probability $p^k(1-p)^{n-k}$, and there are $\binom{n}{k}$ of these. Thus,

$$P(B_k) = \binom{n}{k} p^k(1-p)^{n-k}, \quad k = 0, 1, \ldots, n \ .$$

We have thus discovered the **binomial distribution**.

Chapter 2

Some Important Continuous Distributions

Specifying a model for a random experiment via a complete description of Ω and P may not always be convenient or necessary. In practice we are only interested in various *observations* (i.e., numerical measurements) in the experiment. We include these into our modelling process via the introduction of *random variables*.

2.1 Random variables

Random variables are treated in this course in a rather "intuitive" manner. That is, we view random variables as observations of a random experiment which we have not carried out yet. In other words, a random variable is considered as a variable that can take on different values according to some random mechanism. It is then up to us to specify the probabilities that the random variable will take certain values.

We usually denote random variables with *capital* letters from the last part of the alphabet, e.g. X, $X_1, X_2, \ldots, Y, Z$. Random variables allow us to use natural and intuitive notations for certain events, such as $\{X = 10\}$, $\{X > 1000\}$, $\{\max(X, Y) \leq Z\}$, etc.

Example 2.1 We flip a coin n times. In Example 1.7 we can find a probability model for this random experiment. But suppose we are not interested in the complete outcome, e.g., HTTHT..., but only in the total number of heads. Let X be the total number of heads. X is a "random variable" in the true sense of the word: X could lie anywhere between 0 and n. What we are interested in, however, is the *probability* that X takes certain values. That is, we are interested in the **probability distribution** of X. Example 1.7 now suggests

that

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n \,. \tag{2.1}$$

This contains all the information about X that we could possibly wish to know.

We give some more examples of random variables without specifying the sample space.

1. The number of defective transistors out of 100 inspected ones,
2. the number of bugs in a computer program,
3. the amount of rain in Brisbane in June,
4. the amount of time needed for an operation.

The set of all possible values a random variable X can take is called the **range** of X. We further distinguish between discrete and continuous random variables:

Discrete random variables can only take *isolated* values.

For example: a count can only take non-negative integer values.

Continuous random variables can take values in an *interval*.

For example: rainfall measurements, lifetimes of components, lengths, ... are (at least in principle) continuous.

From a mathematical point of view the above "definition" of a random variable is not very rigorous. Now, recall that we have introduced random variables because we were not interested in a complete outcome of the experiment, but rather in a (real-valued) *function* of the outcome. This leads to the following mathematical definition of a random variable:

Definition 2.1 A **random variable** is a real-valued function on Ω.

Example 2.2 We return to Examples 1.7 and 2.1. Define X as the function that assigns to each outcome $\omega = (x_1, \ldots, x_n)$ the number $x_1 + \cdots + x_n$. Then clearly X is a random variable in terms of Definition 2.1. Moreover, the event that there are exactly k Heads in n throws can be written as

$$\{\omega \in \Omega : X(\omega) = k\}.$$

If we abbreviate this to $\{X = k\}$, and further abbreviate $P(\{X = k\})$ to $P(X = k)$, then we obtain exactly (2.1).

The example above exemplifies how random variables should be viewed mathematically: X is a function, and probabilities such as $P(X \leq x)$ should be interpreted as $P(\{\omega \in \Omega : X(\omega) \leq x\})$.

2.2 Probability distribution

Let X be a random variable. We would like to specify the probabilities of events such as $\{X = x\}$ and $\{a \leq X \leq b\}$.

> If we can specify all probabilities involving X, we say that we have specified the **probability distribution** of X.

One way to specify the probability distribution is to give the probabilities of all events of the form $\{X \leq x\}$, $x \in \mathbb{R}$. This leads to the following definition.

Definition 2.2 The **cumulative distribution function** (C.D.F.) of a random variable X is the function $F : \mathbb{R} \to [0, 1]$ defined by

$$F(x) = P(X \leq x), \ \ x \in \mathbb{R}.$$

Note that above we should have written $P(\{X \leq x\})$ instead of $P(X \leq x)$. From now on we will use this type of abbreviation throughout the course. In Figure 2.1 the graph of a C.D.F. is depicted.

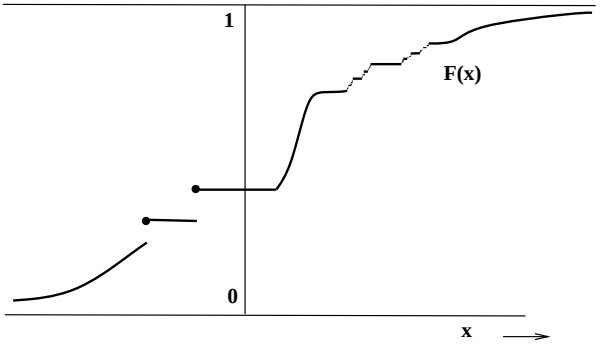


Figure 2.1: A cumulative distribution function

The following properties for F are a direct consequence of the three Axiom's for P.

1. F is right-continuous: $\lim_{h \downarrow 0} F(x + h) = F(x)$,
2. F is increasing: $x \leq y \Rightarrow F(x) \leq F(y)$,
3. $0 \leq F(x) \leq 1$.

Any such function can be used to specify the distribution of a random variable X. However, in practice we will specify the distribution in a different way, whereby we make the distinction between *discrete* and *continuous* random variables.

2.2.1 Discrete distributions

Definition 2.3 We say that X has a **discrete** distribution if X is a discrete random variable. In particular, for some finite or countable set of values $x_1, x_2, \ldots$ we have $P(X = x_i) > 0$, $i = 1, 2, \ldots$ and $\sum_i P(X = x_i) = 1$. We define the **probability mass function** (p.m.f.) of X by $f(x) = P(X = x)$. The probability mass function is also called the probability function (p.f.) $f(x)$.

The easiest way to specify the distribution of a discrete random variable is to specify its p.f.. Indeed, by the third axiom, if we know $f(x)$ for all x, then we can calculate all possible probabilities involving X. Namely,

$$\boxed{P(X \in B) = \sum_{x \in B} f(x)} \tag{2.2}$$

for any subset B of the range of X.

Example 2.3 Toss a die and let X be its face value. X is discrete with range $\{1, 2, 3, 4, 5, 6\}$. If the die is fair the probability function f(x) is given by

x	1	2	3	4	5	6	$\sum$
$f(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

Example 2.4 Toss two dice and let W be the r.v. denoting largest face value showing. The p.f. of W can be found to satisfy

w	1	2	3	4	5	6	$\sum$
$f(w)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$	1

The probability that the maximum is at least 3 is $P(X \geq 3) = \sum_{w=3}^{6} f(w) = 32/36 = 8/9$.

2.2.2 Continuous distributions

Definition 2.4 A random variable X is said to have a **continuous distribution** if X is a continuous random variable for which there exists a positive function f with total integral 1, such that for all a, b

$$\boxed{P(a \leq X \leq b) = \int_a^b f(u)\, du.} \tag{2.3}$$

The function f is called the **probability density function** (p.d.f.) of X.

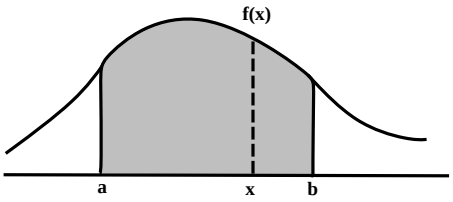


Figure 2.2: Probability density function

Note that in this case the cumulative distribution function, $F(x)$, is given by

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\, du.$$

We can *interpret* $f(x)$ as the "infinitesimal" probability that $X = x$. More precisely,

$$P(x \leq X \leq x + h) = \int_{x}^{x+h} f(u)\, du \approx h\, f(x)\ .$$

Example 2.5 Draw a random number from the interval of real numbers $[0, 2]$. Each number is equally possible. Let X represent the number. What is the probability density function $f(x)$ and the distribution function $F(x)$ of X?

Solution: Convince yourself that the density should be constant on the interval $[0, 2]$. Hence

$$f(x) = \begin{cases} 1/2 & 0 \leq x \leq 2, \\ 0 & \text{otherwise} \end{cases}$$

and

$$F(x) = \begin{cases} 0 & x < 0, \\ x/2 & 0 \leq x \leq 2, \\ 1 & x > 2. \end{cases}$$

We have modelled this random experiment using a continuous random variable and its p.d.f. (and C.D.F.). Compare this with the more "direct" model of Example 1.3.

Notice that describing an experiment via a random variable and its p.d.f., p.f., or C.D.F. seems much easier than describing the experiment by giving the probability space. In fact, we have not used a probability space in the above examples.

2.3 Expectation

Although all the probability information of a random variable is contained in its C.D.F. (or p.f. for discrete random variables and p.d.f. for continuous random variables), it is often useful to consider various numerical characteristics of that random variable. One such number is the *expectation* of a random variable; it is a sort of "weighted average" of the values that X can take. Here is a more precise definition.

Definition 2.5 Let X be a *discrete* random variable with p.f. $f(x)$. The **expectation** (or expected value) of X, denoted by $E(X)$, is defined by

$$E(X) = \sum_x x\, P(X = x) = \sum_x x\, f(x)\ .$$

The expectation of X is sometimes written as μ_X.

Example 2.6 Find $E(X)$ if X is the outcome of a toss of a fair die.

Since $P(X = 1) = \ldots = P(X = 6) = 1/6$, we have

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

Note: $E(X)$ is not necessarily a possible outcome of the random experiment as in the previous example.

For continuous random variables we can define something similar:

Definition 2.6 Let X be a *continuous* random variable with p.d.f. $f(x)$. The **expectation** (or expected value) of X, denoted by $E(X)$, is defined by

$$E(X) = \int_x x\, f(x)\, dx\ .$$

Remark 2.1 We note that the expectation need not exist; that is, the integral may diverge.

If X is a random variable, then a function of X, such as X^2 or $\sin(X)$ is also a random variable. The following theorem is not so difficult to prove, and is entirely "obvious": the expected value of a function of X is the weighted average of the values that this function can take.

Theorem 2.1 If X is *discrete* with p.f. $f(x)$, then for any real-valued function g

$$E(g(X)) = \sum_x g(x)\, f(x)\ .$$

Similarly, if X is *continuous* with p.d.f. $f(x)$, then

$$E\left(g(X)\right) = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx\ .$$

Example 2.7 Find $E(X^2)$ if X is the outcome of the toss of a fair die. We have

$$E(X^2) = 1^2\,\frac{1}{6} + 2^2\,\frac{1}{6} + 3^2\,\frac{1}{6} + \ldots + 6^2\,\frac{1}{6} = \frac{91}{6}.$$

Another useful number about (the distribution of) X is the *variance* of X. This number, sometimes written as σ_X^2, measures the *spread* or dispersion of the distribution of X.

Definition 2.7 The **variance** of a random variable X, denoted by $\mathrm{var}(X)$ is defined by

$$\mathrm{var}(X) = E(X - E(X))^2\ .$$

The square root of the variance is called the **standard deviation**. The number $E(X^r)$ is called the rth **moment** of X.

The following important properties for expectation and variance hold for discrete or continuous random variables and follow easily from the definitions of expectation and variance.

- $E(aX + b) = a\,E(X) + b$
- $\mathrm{var}(X) = E(X^2) - (E(X))^2$
- $\mathrm{var}(aX + b) = a^2\,\mathrm{var}(X)$

Many calculations and manipulations involving probability distributions are facilitated by the use of *transforms*. We discuss here only the *moment generating function*.

Definition 2.8 The **moment generating function** (MGF) of a random variable X is the function, $m : I \to [0, \infty)$, given by

$$m(t) = E\left(\mathrm{e}^{tX}\right)\ .$$

Here I is an open interval containing 0 for which the above integrals are well defined for all $t \in I$.

In particular, for a discrete random variable with p.f. $f(x)$,

$$m(t) = \sum_x \mathrm{e}^{tx}\, f(x),$$

and for a continuous random variable with p.d.f. $f(x)$,

$$m(t) = \int_x \mathrm{e}^{tx} f(x)\, dx \ .$$

It can be shown that two random variables have the same distribution if and only if they have the same moment generating function. The usefulness of this observation and of moment generating functions in general will become clear in the next chapter.

2.4 Some important discrete distributions

In this section we give a number of important discrete distributions and list some of their properties. Note that the p.f. of each of these distributions depends on one or more parameters; so in fact we are dealing with *families* of distributions.

2.4.1 Bernoulli distribution

We say that X has a **Bernoulli** distribution with success probability p if X can only assume the values 0 and 1, with the probability of a success given by

$$P(X = 1) = p = 1 - P(X = 0) \ .$$

We write $X \sim B(1, p)$; that is, a binomial distribution with $n = 1$ and probability of success p. Despite its simplicity, this is one of the most important distributions in probability! It models for example a single coin toss experiment.

2.4.2 Binomial distribution

Consider a sequence of n coin tosses. If X is the random variable which counts the total number of heads and the probability of "Heads" is p, then we say X has a **binomial** distribution with parameters n and p and write $X \sim B(n, p)$. The probability distribution of X is specified by

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n. \tag{2.4}$$

This follows from Examples 1.7 and 2.1. An example of the graph of the p.d.f. is given in Figure 2.3

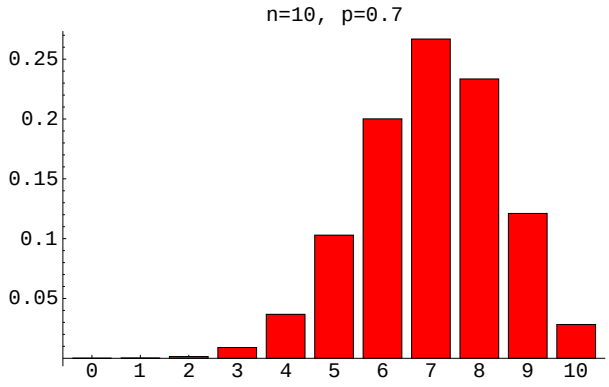


Figure 2.3: The p.d.f. of the $B(10, 0.7)$ distribution

You may verify that

$$E(X) = np \quad \text{and} \quad \text{var}(X) = np(1-p).$$

2.4.3 Hypergeometric distribution

Hypergeometric distribution A sample of size n is randomly drawn without replacement from a population of size N containing M special items (and $N-M$ non-special items). The probability that the sample contains x special items is given by

$$P\{X = x\} = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \quad x = 0, 1, \ldots, n, \tag{2.5}$$

where in (2.5) the convention that $\binom{b}{a} = 0$ if $a > b$ is adopted. This implies that for x to have a probability > 0,

$$x \leq \min(M, n) \quad \text{and} \quad x \geq \max(0, n - (N - M)).$$

The mean and variance of X are given by

$$E(X) = n\frac{M}{N}$$

and

$$\text{var}(X) = n\frac{M}{N}(1 - \frac{M}{N})\frac{N-n}{N-1},$$

respectively.

If n is small relative to N (for example, if $n/N \leq 5\%$), then one can approximate the hypergeometric distribution by the binomial distribution $B(n, p)$, where $p = (M/N)$.

2.4.4 Geometric distribution

Again we look at a sequence of coin tosses but count a different thing. Let X be the number of tosses needed before the first head occurs. Then

$$P(X = x) = (1 - p)^{x-1}p, \quad x = 1, 2, 3, \ldots \tag{2.6}$$

since the only string that has the required form is

$$\underbrace{ttt\ldots t}_{x-1}\, h$$

and this has probability $(1 - p)^{x-1}p$. Such a random variable X is said to have a **geometric** distribution with parameter p. We write $X \sim \text{Geometric}(p)$. An example of the graph of the p.f. is given in Figure 2.4

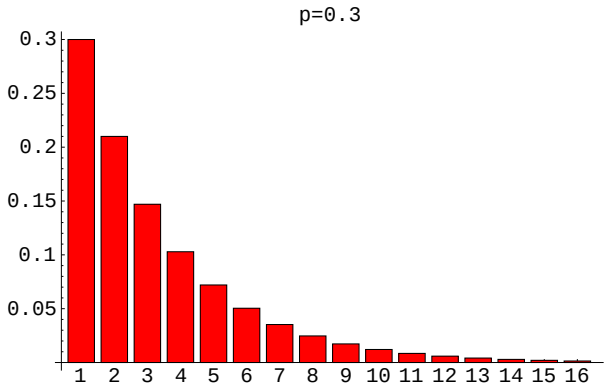


Figure 2.4: The p.f. of the Geometric (0.3) distribution

You may verify that

$$E(X) = \frac{1}{p} \quad \text{and} \quad \text{var}(X) = \frac{1 - p}{p^2}.$$

2.4.5 Negative binomial distribution

Negative binomial distribution For this distribution, the probability of x is given by

$$P\{X = x\} = \binom{x-1}{r-1} q^{x-r} p^r, \quad x = r, r+1, r+2, \ldots,$$

where $p + q = 1$.

It can be viewed as X denoting the number of tosses required to obtain r successes where p is the probability of Heads.

The special case of $r = 1$ corresponds to the geometric distribution.

2.4.6 Poisson distribution

A random variable X for which

$$P(X = x) = \frac{\lambda^x}{x!}\,\mathrm{e}^{-\lambda}, \quad x = 0, 1, 2, \ldots, \tag{2.7}$$

(for fixed $\lambda > 0$) is said to have a **Poisson** distribution. We write $X \sim$ Poisson(λ).

You may verify that

$$E\,(X) = \lambda \quad \text{and} \quad \text{var}(X) = \lambda\,.$$

The Poisson distribution is used in many probability models and may be viewed as the "limit" of the $B(n, \mu/n)$ for large n. More specifically, suppose that $X \sim B(n, p)$. If as $n \to \infty, \mu = np$ tends to a constant, say c, then

$$\binom{n}{x} p^x q^{n-x} \to \frac{\exp(-c)c^x}{x!},$$

where $q = 1 - p$.

An example of the graph of the p.f. for the Poisson distribution is given in Figure 2.5

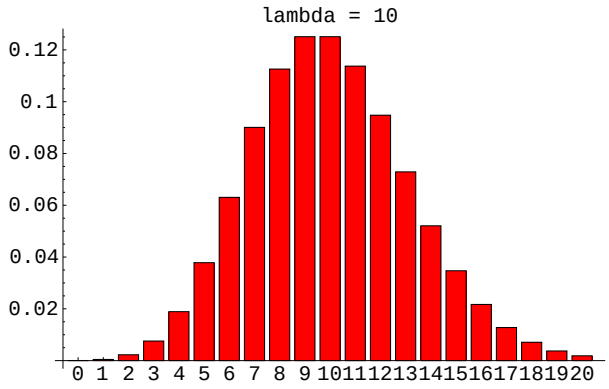


Figure 2.5: The p.f. of the Poisson (10) distribution

2.5 Some important continuous distributions

In this section we give a number of important continuous distributions and list some of their properties. Note that the p.d.f. of each of these distributions depends on one or more parameters; so, as in the discrete case discussed before, we are dealing with *families* of distributions.

At the end of this section we have included two distributions (the F-distribution and the t-distribution) that are essential to statistics, but whose role will only

become clear in Part II of this course. At *first* reading you may omit these sections.

2.5.1 Uniform distribution

We say that a random variable X has a **uniform** distribution on the interval $[a, b]$, if it has density function $f(x)$, given by

$$f(x) = \frac{1}{b-a}, \quad a \leq x \leq b\,.$$

We write $X \sim \mathrm{U}[a, b]$. X can model a randomly chosen point from the interval $[a, b]$, where each choice is equally likely. A graph of the p.d.f. is given in Figure 2.6.

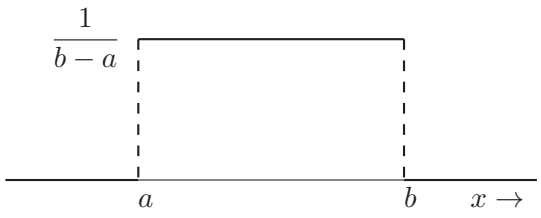


Figure 2.6: The p.d.f. of the uniform distribution on $[a, b]$

We have

$$E(X) = \int_a^b \frac{x}{b-a}\,dx = \frac{1}{b-a}\left[\frac{b^2-a^2}{2}\right] = \frac{a+b}{2}$$

and

$$\begin{aligned}\mathrm{var}(X) &= E(X^2) - (E(X))^2 = \int_a^b \frac{x^2}{b-a}\,dx - \left(\frac{a+b}{2}\right)^2 \\ &= \ldots = \frac{(a-b)^2}{12}\,.\end{aligned}$$

2.5.2 The beta distribution

A random variable is said to have a beta distribution (of the first kind) if its p.d.f. is given by

$$f(x) = \begin{cases} \frac{x^{m_1-1}(1-x)^{m_2-1}}{B(m_1,m_2)}, & 0 < x < 1, \quad m_1 > 0, m_2 > 0, \\ 0 \quad \text{elsewhere.} \end{cases}$$

We say $X \sim \beta(m_1, m_2)$ or $\beta_1(m_1, m_2)$, where the subscript 'one' of the first kind so as to distinguish it from a beta distribution of the second kind (to be defined later).

Note $B(m_1, m_2) = B(m_2, m_1)$. The Beta function $B(m_1, m_2)$ is related to the Gamma function by

$$B(m_1, m_2) = \Gamma(m_1)\Gamma(m_2)/\Gamma(m_1 + m_2).$$

Note: if $X \sim \beta_1(m_1, m_2)$, then $Y = 1 - X \sim \beta_1(m_2, m_1)$.

For $m_1 = m_2 = 1$, we obtain the $U(0, 1)$ distribution, since $\Gamma(1) = \Gamma(2) = 1$.

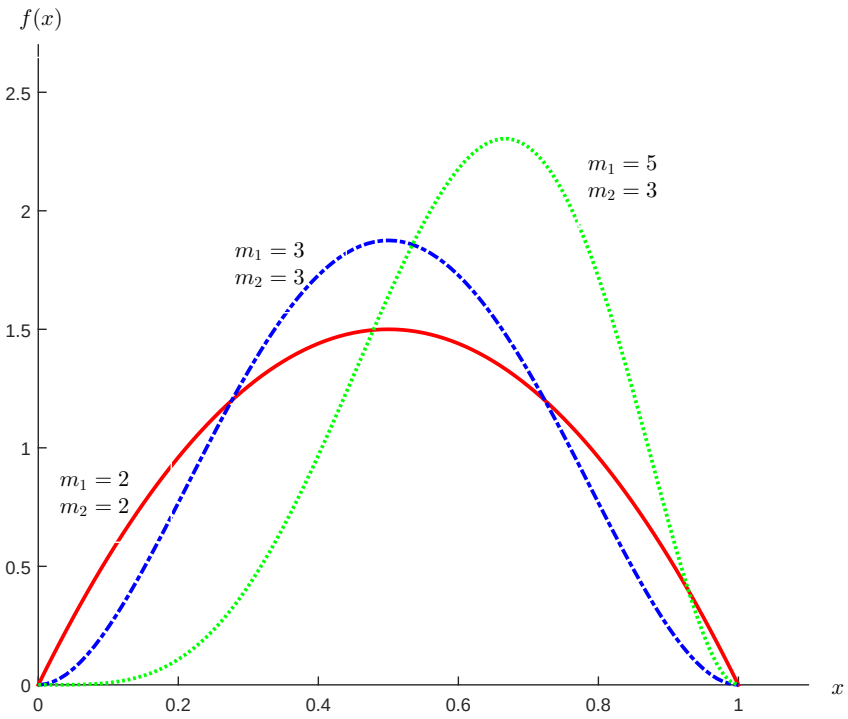


Figure 2.7: Graphs of the p.d.f. of the beta distribution for several values of m_1, m_2

2.5.3 Exponential distribution

A random variable X with probability density function f, given by

$$f(x) = \lambda \, \mathrm{e}^{-\lambda x}, \quad x \geq 0, \tag{2.8}$$

is said to have an **exponential** distribution with parameter λ. We write $X \sim$ Exponential(λ). The exponential distribution can be viewed as a continuous version of the geometric distribution. Graphs of the p.d.f. for various values of λ are given in Figure 2.8.

We have

$$E(X) = \frac{1}{\lambda} \quad \text{and} \quad \text{var}(X) = \frac{1}{\lambda^2}\,.$$

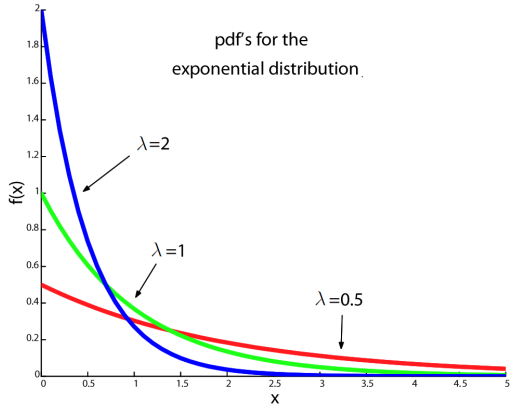


Figure 2.8: The p.d.f. of the Exponential(λ) distribution for various λ.

2.5.4 Normal (or Gaussian) distribution

The normal (or Gaussian) distribution is the most important distribution in the study of statistics. We say that a random variable has a **normal** distribution with parameters μ and σ^2 if its density function $f(x)$ is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \; x \in \mathbb{R}. \tag{2.9}$$

We write $X \sim N(\mu, \sigma^2)$. If $\mu = 0$ and $\sigma = 1$, then

$$f(x) = \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-x^2/2},$$

and the distribution is known as a **standard normal** distribution. The C.D.F. of this latter distribution is often denoted by Φ, and is tabulated in Appendix B. In Figure 2.9 the probability densities for three different normal distributions have been depicted.

We will consider the normal distribution in much more detail in Section 3.4. Here are a few important properties. We have

$$E(X) = \mu \quad \text{and} \quad \mathrm{var}(X) = \sigma^2 \, .$$

If $X \sim N(\mu, \sigma^2)$, then

$$\frac{X-\mu}{\sigma} \; \sim \; N(0,1) \, . \tag{2.10}$$

Thus by subtracting the mean and dividing by the standard deviation we obtain a standard normal distribution. This procedure is called **standardisation**. Standardisation enables us to express the C.D.F. of any normal distribution in terms of the C.D.F. of the standard normal distribution. This is the reason why only the table for the standard normal distribution is included in the appendix.

The moment generating function of $X \sim N(\mu, \sigma^2)$ is given by

$$E(\mathrm{e}^{tX}) = \mathrm{e}^{\mu t + \frac{1}{2}\sigma^2 t^2}, \quad t \in \mathbb{R}. \tag{2.11}$$

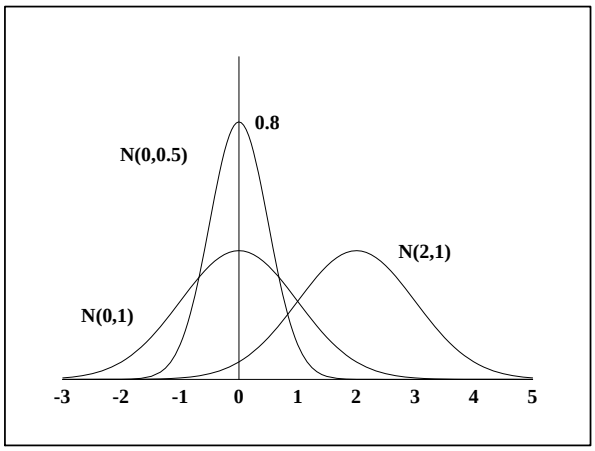


Figure 2.9: Probability density functions for various normal distributions

2.5.5 The gamma and chi-distributions

The **gamma** distribution arises frequently in Statistics. The density function is given by

$$f(x) = \frac{\lambda^\alpha x^{\alpha-1} \mathrm{e}^{-\lambda x}}{\Gamma(\alpha)}, \;\; x \geq 0, \tag{2.12}$$

where Γ is the Gamma function defined as

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1}\, \mathrm{e}^{-u}\, du, \; \alpha > 0.$$

Parameter α is called the **shape** parameter, and λ is called the **scale** parameter. We write $X \sim \text{gamma}(\alpha, \lambda)$.

Often the gamma distribution is used in reduced form where the scale parameter λ is set equal to unity. We then write

$$X \sim \gamma(\alpha)$$

for which $E(X) = \text{var}(X) = \alpha$.

That is, if X has a two-parameter gamma distribution with parameters α and λ, then $Y = \lambda X$ has a $\gamma(\alpha)$ distribution. As $E(Y) = \text{var}(Y)$, it follows that

$$\begin{aligned} E(X) &= E(Y)/\lambda \\ &= \alpha/\lambda, \end{aligned}$$

and

$$\begin{aligned} \text{var}(X) &= \text{var}(Y)/\lambda^2 \\ &= \alpha/\lambda^2. \end{aligned}$$

We mention a few properties of the Γ-function.

1. $\Gamma(a+1) = a\,\Gamma(a)$, for $a \in \mathbb{R}_+$.
2. $\Gamma(m) = (m-1)!$ for $m = 1, 2, \ldots$.
3. $\Gamma(1/2) = \sqrt{\pi}$.

The moment generating function of $X \sim \text{gamma}(\alpha, \lambda)$ is given by

$$\begin{aligned} E\left(\mathrm{e}^{tX}\right) &= \int_0^\infty \frac{\mathrm{e}^{-\lambda x}\,\lambda^\alpha\,x^{\alpha-1}}{\Gamma(\alpha)}\,\mathrm{e}^{tx}\,dx \\ &= \left(\frac{\lambda}{\lambda-t}\right)^\alpha \int_0^\infty \frac{\mathrm{e}^{-(\lambda-t)x}\,(\lambda-t)^\alpha\,x^{\alpha-1}}{\Gamma(\alpha)}\,dx \\ &= \left(\frac{\lambda}{\lambda-t}\right)^\alpha . \end{aligned} \tag{2.13}$$

2.5.6 Chi-squared distribution

A special case of the gamma distribution that is of particular importance is the chi-squared distribution.

A random variable X is said to have a chi-squared distribution with m ($\in \{1, 2, \ldots\}$) **degrees of freedom** if

$$X \sim \text{gamma}(m/2,\ 1/2);$$

that is, if $Y = X/2$ is distributed as

$$Y \sim \gamma(m/2).$$

We write $X \sim \chi^2_m$.

The latter definition of the chi-squared distribution provides an easy to calculate its mean and variance. We have that

$$\begin{aligned} E(Y) &= m/2 \\ &= E(X)/2, \end{aligned}$$

and

$$\begin{aligned} \text{var}(Y) &= m/2 \\ &= \text{var}(X)/4, \end{aligned}$$

establishing that $E(X) = m$ and $\text{var}(X) = 2m$.

A graph of the p.d.f. of the χ^2_m-distribution, for various m is given in Figure 2.10.

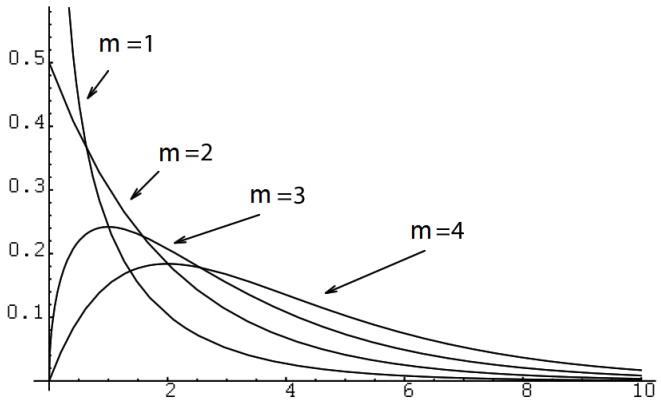


Figure 2.10: P.d.f's for the χ^2_m-distribution, for various degrees of freedom m

2.5.7 F-distribution

Let m_1 and m_2 be strictly positive integers. We say that a random variable X has an F**-distribution** (or Fisher distribution) with m_1 and m_2 **degrees of freedom** if the p.d.f. is given by

$$f(x) = \frac{\Gamma((m_1+m_2)/2)\, {m_1}^{m_1/2}\, {m_2}^{m_2/2}}{\Gamma(m_1/2)\Gamma(m_2/2)} \, \frac{x^{(m_1/2)-1}}{(m_2+m_1x)^{(m_1m_2+m_1)/2}}, \quad x>0. \tag{2.14}$$

We write $X \sim F_{m_1,m_2}$, A graph of the p.d.f. of the F_{m_1,m_2}-distribution, for various m_1 and m_2 is given in Figure 2.11.

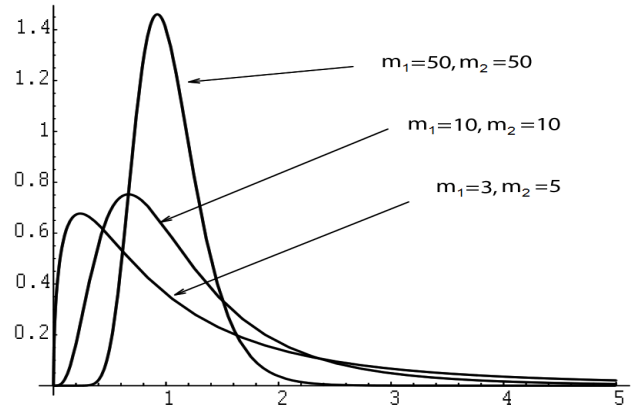


Figure 2.11: Probability density functions for the F_{m_1,m_2}-distribution, for various degrees of freedom m_1 and m_2.

For completeness we mention that if $X \sim F_{m_1,m_2}$, then

$$E(X) = \frac{m_2}{m_2-2}, \quad (m_2 \geq 3) \quad \text{and} \quad \text{var}(X) = \frac{2m_2^2(m_1+m_2-2)}{m_1(m_2-2)^2(m_2-4)}, \quad (m_2 \geq 5)\,.$$

The main reason why the F-distribution appears often in statistics is given in the following result.

Theorem 2.2 Let $U \sim \chi^2_{m_1}$ and $V \sim \chi^2_{m_2}$ be independent. Then

$$\frac{U/m_1}{V/m_2} \sim F_{m_1,m_2} \ .$$

As a corollary to this theorem, suppose

$$X = \frac{U/m_1}{V/m_2} \sim F_{m_1,m_2}.$$

Then

$$Y = 1/X \sim F_{m_2,m_1}. \tag{2.15}$$

For any $0 < \alpha < 1$, let $F_{m_1,m_2;\alpha}$ denote the α-quantile of the F_{m_1,m_2} distribution. That is, it is the number x such that $P(X \leq x) = \alpha$, where $X \sim F_{m_1,m_2}$. Using that

$$P(X < x) = P(Y > \frac{1}{x}) \ ,$$

we have that for any α

$$\alpha = P(X < F_{m_1,m_2;\alpha}) = P(Y > \frac{1}{F_{m_1,m_2;\alpha}}) \ .$$

From (2.15), it follows that

$$F_{m2,m_1;1-\alpha} = \frac{1}{F_{m_1,m_2;\alpha}} \ . \tag{2.16}$$

The tables of the F_{m_1,m_2} distribution in the appendix only list α-quantiles for $\alpha \geq 1/2$, because by (2.16) the α-quantiles with $\alpha < 1/2$ can be derived from these.

2.5.8 t-distribution

We say that a random variable X has a **t-distribution** (or Student t-distribution) with m ($\in \{1, 2, \ldots\}$) **degrees of freedom**, if the p.d.f. is given by

$$f(x) = \frac{\Gamma((m+1)/2)}{\sqrt{m\pi}\,\Gamma(m/2)} \left(1 + \frac{x^2}{m}\right)^{-(m+1)/2}, \ \ x \in \mathbb{R}.$$

We write $X \sim t_m$. A graph of the p.d.f. of the t_m-distribution, for various m is given in Figure 2.12. Note that the p.d.f. is *symmetric*. Moreover, it can be shown that the p.d.f. of the t_m-distribution converges to the p.d.f of the $N(0,1)$ distribution as $m \to \infty$.

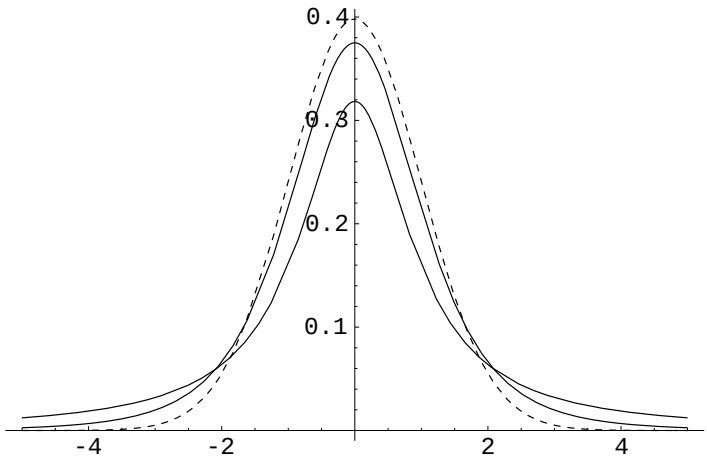


Figure 2.12: The p.d.f. of t_1, t_4, and $N(0,1)$ (dashed).

For completeness we mention that if $X \sim t_m$, then

$$E(X) = 0 \quad (m \geq 2) \qquad \text{and} \qquad \text{var}(X) = \frac{m}{m-2}, \quad (m \geq 3) \; .$$

The main reason why the t-distribution appears so often in statistics is because of the following result.

Theorem 2.3 Let $Z \sim N(0,1)$ and $V \sim \chi^2_m$ be independent. Then

$$\frac{Z}{\sqrt{V/m}} \sim t_m \; .$$

Chapter 3

Multiple Random Variables

Often a random experiment is described via more than one random variable. Examples are:

1. We select a random sample of $n = 10$ people and observe their lengths. Let $X_1, \ldots, X_n$ be the individual lengths.
2. We flip a coin repeatedly. Let $X_i = 1$ if the ith flip is "heads" and 0 else. The experiment is described by the sequence $X_1, X_2, \ldots$ of coin flips.
3. We randomly select a person from a large population and measure his/her weight X and height Y.

How can we specify the behaviour of the random variables above? We should not just specify the p.d.f. or p.f. of the individual random variables, but also say something about the "interaction" (or lack thereof) between the random variables. For example, in the third experiment above if the height Y is large, we expect that X is large as well. On the other hand, for the first and second experiment it is reasonable to assume that information about one of the random variables does not give extra information about the others. What we need to specify is the **joint distribution** of the random variables.

The theory for multiple random variables is quite similar to that of a single random variable. The most important extra feature is perhaps the concept of *independence* of random variables. Independent random variables will play a crucial role in modelling data in Part II of the course. In this chapter I have also included some results on joint normal distributions and linear transformations. This will help you understand the central role of the normal distribution in statistics.

3.1 Joint Distribution and Independence

Let $X_1, \ldots, X_n$ be random variables describing some random experiment. We can accumulate the X_i's into a row vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ or column vector $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ (here T means transposition). $\boldsymbol{X}$ is called a **random vector**.

Recall that the distribution of a *single* random variable X is completely specified by its cumulative distribution function. Analogously, the joint distribution of $X_1, \ldots, X_n$ is specified by the **joint cumulative distribution function** F, defined by

$$F(x_1, \ldots, x_n) = P(\{X_1 \leq x_1\} \cap \cdots \cap \{X_n \leq x_n\}) = P(X_1 \leq x_1, \ldots, X_n \leq x_n),$$

If we know F then we can in principle derive any probability involving the X_i's. Note the abbreviation on the right-hand side. We will henceforth use this kind of abbreviation throughout the notes.

Similar to the 1-dimensional case we distinguish between the case where the X_i are discrete and continuous. The corresponding joint distributions are again called discrete and continuous, respectively.

3.1.1 Discrete Joint Distributions

To see how things work in the discrete case, let us start with an example.

Example 3.1 In a box are three dice. Die 1 is a normal die; die 2 has no 6 face, but instead two 5 faces; die 3 has no 5 face, but instead two 6 faces. The experiment consists of selecting a die at random, followed by a toss with that die. Let X be the die number that is selected, and let Y be the face value of that die. The probabilities $P(X = x, Y = y)$ are specified below.

	y						
x	1	2	3	4	5	6	$\sum$
1	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{3}$
2	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{9}$	0	$\frac{1}{3}$
3	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	0	$\frac{1}{9}$	$\frac{1}{3}$
$\sum$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

The function $p : (x, y) \mapsto P(X = x, Y = y)$ is called the *joint* p.f. of X and Y. The following definition is just a generalisation of this.

Definition 3.1 Let $X_1, \ldots, X_p$ be *discrete* random variables. The function f defined by $p(x_1, \ldots, x_p) = P(X_1 = x_1, \ldots, X_p = x_p)$ is called the **joint probability function** (p.f.) of $X_1, \ldots, X_p$.

We sometimes write $f_{X_1,\ldots,X_p}$ instead of f to show that this is the p.f. of the random variables $X_1, \ldots, X_p$. Or, if $\boldsymbol{X}$ is the corresponding random vector, we could write $f_{\boldsymbol{X}}$ instead.

Note that, by the third Axiom, if we are given the joint p.f. of $X_1, \ldots, X_p$ we can in principle calculate *all possible probabilities* involving these random variables. For example, in the 2-dimensional case

$$P((X, Y) \in B) = \sum_{(x,y)\in B} P(X = x, Y = y) \; ,$$

for any subset B of possible values for (X, Y). In particular, we can find the p.f. of X by summing the joint p.f. over all possible values of y:

$$P(X = x) = \sum_{y} P(X = x, Y = y) \; .$$

The converse is *not* true: from the individual distributions (so-called **marginal** distribution) of X and Y we cannot in general reconstruct the joint distribution of X and Y. We are simply missing the "dependency" information. E.g., in Example 3.1 we cannot reconstruct the inside of the two-dimensional table if only given the column and row totals.

However, there is one *important exception* to this, namely when we are dealing with *independent* random variables. We have so far only defined what independence is for *events*. The following definition says that random variables $X_1, \ldots, X_p$ are independent if the events $\{X_1 \in A_1\}, \ldots, \{X_p \in A_n p\}$ are independent for any subsets $A_1, \ldots, A_p$ of $\mathbb{R}$. Intuitively, this means that any information about one of them does not affect our knowledge about the others.

Definition 3.2 The random variables $X_1, \ldots, X_p$ are called **independent** if for all $A_1, \ldots, A_p$, with $A_i \subset \mathbb{R}$, $i = 1, \ldots, p$

$$P(X_1 \in A_1, \ldots, X_n \in A_p) = P(X_1 \in A_1) \cdots P(X_p \in A_p) \; .$$

The following theorem is a direct consequence of the definition above.

Theorem 3.1 Discrete random variables $X_1, \ldots, X_p$ are independent if and only if

$$\boxed{P(X_1 = x_1, \ldots, X_p = x_p) = P(X_1 = x_1) \cdots P(X_p = x_p),} \tag{3.1}$$

for all $x_1, x_2, \ldots, x_p$.

Example 3.2 We repeat the experiment in Example 3.1 with three ordinary fair dice. What are now the joint probabilities in the table? Since the events $\{X = x\}$ and $\{Y = y\}$ are now independent, each entry in the p.f. table is $\frac{1}{3} \times \frac{1}{6}$. Clearly in the first experiment not *all* events $\{X = x\}$ and $\{Y = y\}$ are independent (why not?).

Example 3.3 Consider the experiment where we flip a coin n times. We can model this experiments in the following way. For $i = 1, \ldots, n$ let X_i be the result of the ith toss: $\{X_i = 1\}$ means Heads, $\{X_i = 0\}$ means Tails. Also, let

$$P(X_i = 1) = p = 1 - P(X_i = 0), \quad i = 1, 2, \ldots, n \, .$$

Thus, p can be interpreted as the probability of Heads, which may be known or unknown. Finally, assume that $X_1, \ldots, X_n$ are *independent*.

This completely specifies our model. In particular we can find any probability related to the X_i's. For example, let $X = X_1 + \cdots + X_n$ be the total number of Heads in n tosses. Obviously X is a random variable that takes values between 0 and n. Denote by A the set of all binary vectors $\boldsymbol{x} = (x_1, \ldots, x_n)$ such that $\sum_{i=1}^{n} x_i = k$. Note that A has $\binom{n}{k}$ elements. We now have

$$\begin{aligned} P(X = k) &= \sum_{\boldsymbol{x} \in A} P(X_1 = x_1, \ldots, X_n = x_n) \\ &= \sum_{\boldsymbol{x} \in A} P(X_1 = x_1) \cdots P(X_n = x_n) = \sum_{\boldsymbol{x} \in A} p^k (1-p)^n \\ &= \binom{n}{k} p^k (1-p)^n \, . \end{aligned}$$

In other words, $X \sim B(n, p)$. Compare this to what we did in Example 1.7.

In Figure 3.1 a typical outcome (or realisation) of the random variables $X_1, \ldots, X_n$ is depicted for the parameter values $n = 50$ and $p = 0.2$.

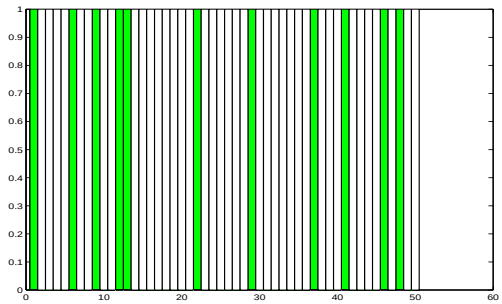


Figure 3.1: *Results of 50 coin flips with $p = 0.2$. White = 0, Grey = 1.*

Remark 3.1 If $f_{X_1,\ldots,X_n}$ denotes the joint p.f. of $X_1, \ldots, X_n$ and f_{X_i} the marginal p.f. of X_i, $i = 1, \ldots, n$, then the theorem above states that independence of the X_i's is equivalent to

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots p_{X_n}(x_n)$$

for all possible $x_1, \ldots, x_n$.

Remark 3.2 An *infinite* sequence $X_1, X_2, \ldots$ of random variables is called independent if for any finite choice of parameters $i_1, i_2, \ldots, i_n$ (none of them the same) the random variables $X_{i_1}, \ldots, X_{i_n}$ are independent.

Multinomial distribution

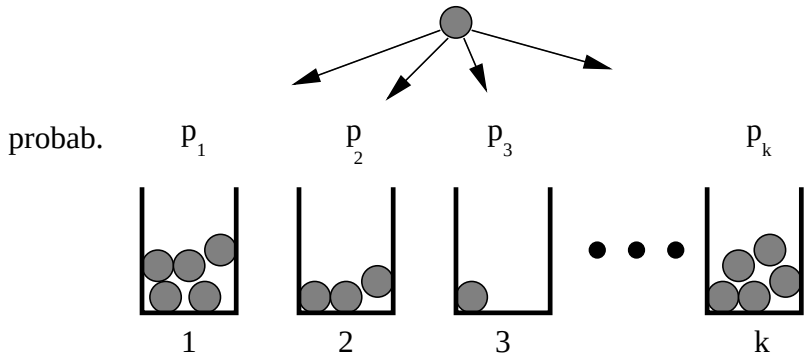
An important discrete joint distribution is the multinomial distribution. It can be viewed as a generalisation of the binomial distribution. First we give the definition, then an example how this distribution arises in applications.

Definition 3.3 We say that $(X_1, X_2, \ldots, X_k)$ has a **multinomial** distribution, with parameters n and $p_1, p_2, \ldots, p_k$, if

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1!\, x_2! \cdots x_k!}\, p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}\,, \tag{3.2}$$

for all $x_1, \ldots, x_k \in \{0, 1, \ldots, n\}$ such that $x_1 + x_2 + \cdots + x_k = n$. We write $(X_1, \ldots, X_k) \sim Mult(n, p_1, \ldots, p_k)$.

Example 3.4 We independently throw n balls into k urns, such that each ball is thrown in urn i with probability p_i, $i = 1, \ldots, k$.



Let X_i be the total number of balls in urn i, $i = 1, \ldots, k$. We show that $(X_1, \ldots, X_k) \sim Mult(n, p_1, \ldots, p_k)$. Let $x_1, \ldots, x_k$ be integers between 0 and n that sum up to n. The probability that the *first* x_1 balls fall in the first urn, the *next* x_2 balls fall in the second urn, etcetera, is

$$p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}.$$

To find the probability that there are x_1 balls in the first urn, x_2 in the second, etcetera, we have to multiply the probability above with the number of ways in which we can fill the urns with $x_1, x_2, \ldots, x_k$ balls, i.e. $n!/(x_1!\, x_2! \cdots x_k!)$. This gives (3.2).

Remark 3.3 Note that for the *binomial* distribution there are only *two* possible urns. Also, note that for each $i = 1, \ldots, k$, $X_i \sim B(n, p_i)$.

3.1.2 Continuous joint distributions

Joint distributions for continuous random variables are usually defined via the joint p.d.f. The results are very similar to the discrete case discussed in Section 3.1.1. Compare this section also with the 1-dimensional case in Section 2.2.2.

Definition 3.4 We say that the continuous random variables $X_1, \ldots, X_n$ have a **joint probability density function** (p.d.f) f if

$$P(a_1 \leq X_1 \leq b_1, \ldots, a_n \leq X_n \leq b_n) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n$$

for all $a_1, \ldots, b_n$.

We sometimes write $f_{X_1,\ldots,X_n}$ instead of f to show that this is the p.d.f. of the random variables $X_1, \ldots, X_n$. Or, if $\boldsymbol{X}$ is the corresponding random vector, we could write $f_{\boldsymbol{X}}$ instead.

We can interpret $f(x_1, \ldots, x_n)$ as a continuous analogue of a p.f., or as the infinitesimal probability that $X_1 = x_1$, $X_2 = x_2$, ..., and $X_n = x_n$. For example in the 2-dimensional case:

$$\begin{aligned} &P(x \leq X \leq x + h,\ y \leq Y \leq y + h) \\ &= \int_x^{x+h} \int_y^{y+h} f(u, v)\, du\, dv \approx h^2\, f(x, y)\ . \end{aligned}$$

Note that if the joint p.d.f. is given, then in principle we can calculate *all probabilities*. Specifically, in the 2-dimensional case we have

$$P((X, Y) \in B) = \int\int_{(x,y)\in B} f(x, y)\, dx\, dy\ , \tag{3.3}$$

for any subset B of possible values for $\mathbb{R}^2$. Thus, the calculation of probabilities is reduced to *integration*.

Similarly to the discrete case, if $X_1, \ldots, X_n$ have joint p.d.f. f, then the (individual, or marginal) p.d.f. of each X_i can be found by integrating f over all other variables. For example, in the two-dimensional case

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

However, we usually cannot reconstruct the joint p.d.f. from the marginal p.d.f's unless we assume that the the random variables are *independent*. The definition of independence is exactly the same as for discrete random variables, see Definition 3.2. But, more importantly, we have the following analogue of Theorem 3.1.

Theorem 3.2 Let $X_1, \ldots, X_n$ be continuous random variables with joint p.d.f. f and marginal p.d.f's $f_{X_1}, \ldots, f_{X_n}$. The random variables $X_1, \ldots, X_n$ are independent if and only if

$$\boxed{f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n),} \tag{3.4}$$

for all $x_1, \ldots, x_n$.

Example 3.5 Consider the experiment were we select randomly and independently n points from the interval [0,1]. We can carry this experiment out using a calculator or computer, using the *random generator*. On your calculator this means pushing the RAN# or Rand button. Here is a possible outcome, or **realisation**, of the experiment, for $n = 12$.

0.9451226800	0.2920864820	0.0019900900	0.8842189383	0.8096459523
0.3503489150	0.9660027079	0.1024852543	0.7511286891	0.9528386400
0.2923353821	0.0837952423			

A model for this experiment is: Let $X_1, \ldots, X_n$ be independent random variables, each with a uniform distribution on [0,1]. The joint p.d.f. of $X_1, \ldots, X_n$ is very simple, namely

$$f(x_1, \ldots, x_n) = 1, \quad 0 \le x_1 \le 1, \ \ldots, \quad 0 \le x_n \le 1 \ ,$$

(and 0 else). In principle we can now calculate any probability involving the X_i's. For example, for the case $n = 2$, what is the probability

$$P\left(\frac{X_1 + X_2^2}{X_1 X_2} > \sin(X_1^2 - X_2) \right) ?$$

The answer, by (3.3), is

$$\iint_A 1 \, dx_1 \, dx_2 = \text{ Area } (A),$$

where

$$A = \left\{ (x_1, x_2) \in [0,1]^2 \ : \ \frac{x_1 + x_2^2}{x_1 x_2} > \sin(x_1^2 - x_2) \right\} \ .$$

(Here $[0,1]^2$ is the unit square in $\mathbb{R}^2$).

Remark 3.4 The type of model used in the previous example, i.e., $X_1, \ldots, X_n$ are independent and all have the same distribution, is the most widely used

model in statistics. We say that $X_1, \ldots, X_n$ is a **random sample** of **size** n, from some given distribution. In Example 3.5 $X_1, \ldots, X_n$ is a random sample from a $U[0,1]$ distribution. In Example 3.3 we also had a random sample, this time from a $B(1,p)$ distribution. The common distribution of a random sample is sometimes called the **sampling distribution**.

Using the computer we can generate the outcomes of random samples from many (sampling) distributions. In Figure 3.2 the outcomes of a two random samples, both of size 100 are depicted in a **dotplot**. The first sample is from the $U[0,1]$ distribution, and the second sample is from the $N(1/2, 1/12)$ distribution. Note that the true expectation and variance of the distributions are the same. However, the "density" of points in the two samples is clearly different, and follows that shape of the corresponding p.d.f's.

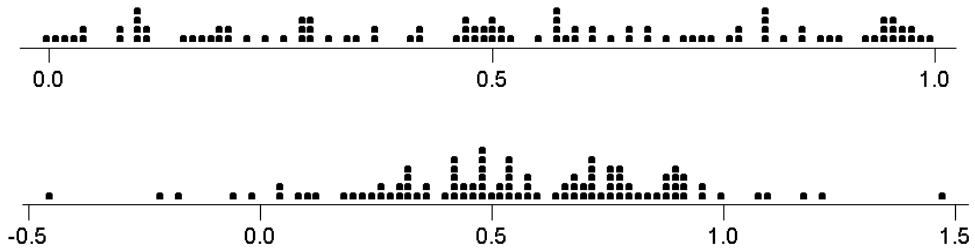


Figure 3.2: *A dotplot of a random sample of size 100 from the* $U[0,1]$ *distribution (above) and the* $N(1/2, 1/12)$ *distribution (below).*

3.2 Expectation

Similar to the 1-dimensional case, the expected value of any real-valued function of $X_1, \ldots, X_n$ is the weighted average of all values that this function can take. Specifically, if $Z = g(X_1, \ldots, X_n)$ then in the discrete case

$$E(Z) = \sum_{x_1} \cdots \sum_{x_n} g(x_1, \ldots, x_n)\, p(x_1, \ldots, x_n) \ ,$$

where p is the joint p.f.; and in the continuous case

$$E(Z) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n \ ,$$

where f is the joint p.d.f.

Example 3.6 Let X and Y be continuous, possibly *dependent*, r.v.'s with joint

p.d.f. f. Then,

$$\begin{aligned} E(X+Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f(x,y)\,dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf(x,y)\,dxdy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf(x,y)\,dxdy \\ &= \int_{-\infty}^{\infty}xf_X(x)\,dx + \int_{-\infty}^{\infty}yf_Y(y)\,dy \\ &= E(X)+E(Y)\,. \end{aligned}$$

The previous example is easily generalised to the following result:

Theorem 3.3 Suppose X_1, X_2, ..., X_n are discrete or continuous random variables with means $\mu_1, \mu_2, \ldots, \mu_n$. Let

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n$$

where a, b_1, b_2, ..., b_n are constants. Then

$$\boxed{\begin{aligned} E(Y) &= a + b_1E(X_1) + \cdots + b_nE(X_n) \\ &= a + b_1\mu_1 + \cdots + b_n\mu_n \end{aligned}}$$

Another important result is the following. Prove this yourself for the continuous case with $n = 2$.

Theorem 3.4 If $X_1, \ldots, X_n$ are *independent*, then

$$\boxed{E(X_1X_2\cdots X_n) = E(X_1)E(X_2)\cdots E(X_n)\,.}$$

The **covariance** of two random variables X and Y is defined as the number

$$\mathrm{cov}(X,Y) = E(X - E(X))(Y - E(Y)).$$

It is a measure for the amount of linear dependency between the variables. If small values of X (smaller than the expected value of X) go together with small values of Y, and at the same time large values of x go together with large values of Y, then $\mathrm{cov}(X,Y)$ will be *positive*. If on the other hand small values of X go together with large values of Y, and large values of X go together with small values of Y, then $\mathrm{cov}(X,Y)$ will be *negative*. A scaled version of the covariance is given by the **correlation coefficient**.

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{varX}}\sqrt{\mathrm{varY}}}.$$

It can be shown that the correlation coefficient always lies between -1 and 1. In Figure 3.3 an illustration of the correlation coefficient is given. Each figure

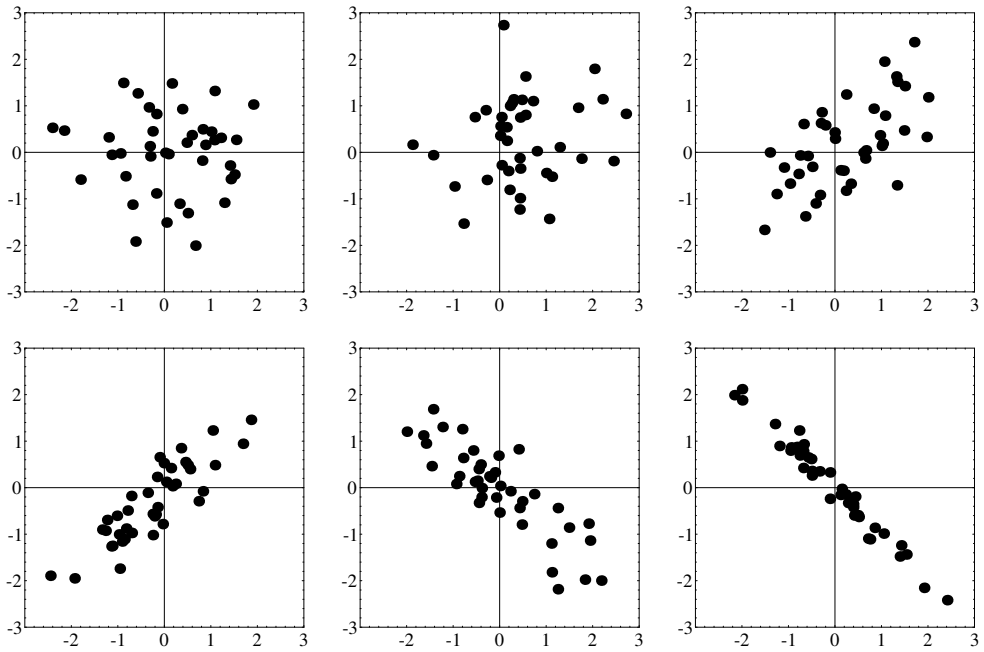


Figure 3.3: *Illustration of correlation coefficient. Above:* $\rho = 0$, $\rho = 0.4$, $\rho = 0.7$. *Below:* $\rho = 0.9$, $\rho = -0.8$, $\rho = -0.98$.

corresponds to samples of size 40 from a different 2-dimensional distribution. In each case $E(X) = E(Y) = 0$ and $\mathrm{var}X = \mathrm{var}Y = 1$.

For easy reference we list some important properties of the variance and covariance. The proofs follow directly from the definitions of covariance and variance and the properties of the expectation.

1	$\mathrm{var(X)} = \mathrm{E(X^2)} - (\mathrm{E(X)})^2$.
2	$\mathrm{var(aX + b)} = \mathrm{a^2 var(X)}$.
3	$\mathrm{cov}(X, Y) = E(XY) - E(X)E(Y)$.
4	$\mathrm{cov}(X, Y) = \mathrm{cov}(Y, X)$.
5	$\mathrm{cov}(aX + bY, Z) = a\,\mathrm{cov}(X, Z) + b\,\mathrm{cov}(Y, Z)$.
6	$\mathrm{cov}(X, X) = \mathrm{var(X)}$.
7	$\mathrm{var(X + Y)} = \mathrm{var(X)} + \mathrm{var(Y)} + 2\,\mathrm{cov(X, Y)}$.
8	X and Y indep. $\Longrightarrow \mathrm{cov}(X, Y) = 0$.

Table 3.1: Properties of variance and covariance

As a consequence of properties 2 and 7, we have

Theorem 3.5 Suppose X_1, X_2, ..., X_n are discrete or continuous *independent* random variables with variances $\sigma_1^2, \sigma^2, \ldots, \sigma_n^2$. Let

$$Y = a + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n$$

where a, b_1, b_2, ..., b_n are constants. Then

$$\begin{array}{rcl} \operatorname{var}(Y) & = & b_1^2 \operatorname{var}(X_1) + \cdots + b_n^2 \operatorname{var}(X_n) \\ & = & b_1^2\, \sigma_1^2 + \cdots + b_n^2\, \sigma_n^2 \end{array}$$

Expectation Vector and Covariance Matrix

Let $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ be a random vector. Sometimes it is convenient to write the expectations and covariances in vector notation.

Definition 3.5 For any random vector $\boldsymbol{X}$ we define the **expectation vector** as the vector of expectations

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T = (E(X_1), \ldots, E(X_n))^T \; .$$

The **covariance matrix** $\boldsymbol{\Sigma}$ is defined as the matrix whose (i, j)th element is

$$\operatorname{cov}(X_i, X_j) = E\{(X_i - \mu_i)(X_j - \mu_j)\} \; .$$

If we define the expectation of a vector (matrix) to be the vector (matrix) of expectations, then we can write:

$$\boldsymbol{\mu} = E(\boldsymbol{X})$$

and

$$\boldsymbol{\Sigma} = E\{(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T\} \; .$$

Note that $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ take the same role as μ and σ^2 in the 1-dimensional case. We sometimes write $\boldsymbol{\mu_X}$ and $\boldsymbol{\Sigma_X}$ if we wish to emphasise that $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ belong to the vector $\boldsymbol{X}$.

Remark 3.5 Note that any covariance matrix $\boldsymbol{\Sigma}$ is a *symmetric* matrix. In fact, it is *positive semi-definite*, i.e., for any (column) vector $\boldsymbol{u}$, we have

$$\boldsymbol{u}^T\, \boldsymbol{\Sigma}\, \boldsymbol{u} \geq 0 \; .$$

To see this, suppose $\boldsymbol{\Sigma}$ is the covariance matrix of some random vector $\boldsymbol{X}$ with expectation vector $\boldsymbol{\mu}$. Write $\boldsymbol{Y} = \boldsymbol{X} - \boldsymbol{\mu}$. Then

$$\begin{array}{rcl} \boldsymbol{u}^T\, \boldsymbol{\Sigma}\, \boldsymbol{u} & = & \boldsymbol{u}^T\, E(\boldsymbol{Y}\boldsymbol{Y}^T)\, \boldsymbol{u} = E\left(\boldsymbol{u}^T\, \boldsymbol{Y}\boldsymbol{Y}^T\, \boldsymbol{u}\right) \\ & = & E\left(\boldsymbol{Y}^T\boldsymbol{u}\right)^T\, \boldsymbol{Y}^T\boldsymbol{u} = E\left(\boldsymbol{Y}^T\boldsymbol{u}\right)^2 \geq 0. \end{array}$$

Note that $\boldsymbol{Y}^T\boldsymbol{u}$ is a random variable.

3.3 Functions of random variables

Suppose $X_1, \ldots, X_n$ are the measurements on a random experiment. Often we are interested in certain *functions* of the measurements only, rather than all measurements themselves. For example, if $X_1, \ldots, X_n$ are the repeated measurements of the strength of a certain type of fishing line, then what we are really interested in is not the individual values for $X_1, \ldots, X_n$ but rather quantities such as the average strength $(X_1 + \cdots + X_n)/n$, the minimum strength $\min(X_1, \ldots, X_n)$ and the maximum strength $\max(X_1, \ldots, X_n)$. Note that these quantities are again random variables. The distribution of these random variables can in principle be derived from the joint distribution of the X_i's. We give a number of examples.

Example 3.7 Let X be a continuous random variable with p.d.f. f_X, and let $Y = aX + b$, where $a \neq 0$. We wish to determine the p.d.f. f_Y of Y. We first express the C.D.F. of Y into the C.D.F. of Y. Suppose first that $a > 0$. We have for any y

$$F_Y(y) = P(Y \leq y) = P(X \leq (y-b)/a) = F_X((y-b)/a) \; .$$

Differentiating this with respect to y gives $f_Y(y) = f_X((y-b)/a)\,/a$. For $a < 0$ we get similarly $f_Y(y) = f_X((y-b)/a)\,/(-a)$. Thus in general

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right) \; . \tag{3.5}$$

Example 3.8 Let $X \sim N(0,1)$. We wish to determine the distribution of $Y = X^2$. We can use the same technique as in the example above, but note first that Y can only take values in $[0, \infty)$. For $y > 0$ we have

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) = 2F_X(\sqrt{y}) - 1 \; . \end{aligned}$$

Differentiating this with respect to y gives

$$\begin{aligned} f_Y(y) &= 2\, f_X(\sqrt{y})\, \frac{1}{2\sqrt{y}} = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\sqrt{y})^2\right) \frac{1}{\sqrt{y}} \\ &= \frac{(1/2)^{1/2}\, y^{-1/2}\, \mathrm{e}^{-y/2}}{\Gamma(1/2)}. \end{aligned}$$

This is exactly the formula for the p.d.f. of a χ^2_1-distribution. Thus $Y \sim \chi^2_1$.

As a corollary to the last example, we give the following theorem which will be of importance in part II of this course.

Theorem 3.6 Let $X_1, \ldots, X_n$, i.i.d. $N(0,1)$, then

$$X_1^2 + \cdots + X_n^2 \sim \chi^2_n \; .$$

PROOF. Let m_{X_i} be the moment generating function of X_i^2, $i = 1, \ldots, n$. Thus, by Example 3.8 and (2.13)

$$\begin{aligned} m_{X_i}(t) &= \left(\frac{\frac{1}{2}}{\frac{1}{2} - t}\right)^{\frac{1}{2}} \\ &= (1-2t)^{-1/2} \quad (i = 1, \ldots, n). \end{aligned}$$

Let m be the moment generating function of $X_1^2 + \cdots + X_n^2$. We have

$$\begin{aligned} m(t) &= E(\exp\left(t(X_1^2 + \cdots X_n^2)\right)) = E(\exp\left(tX_1^2\right) \cdots \exp\left(tX_n^2\right)) \\ &= E(\exp\left(tX_1^2\right)) \cdots E(\exp\left(tX_n^2\right)) = M_{X_1}(t) \cdots M_{X_n}(t) \\ &= (1-2t)^{-n/2}. \end{aligned}$$

In the third equation we have used the fact that since the X_i are independent, the random variables $\exp\left(tX_1^2\right)$, $\ldots$, $\exp\left(tX_n^2\right)$ are also independent. Hence, the expectation of their product is equal to the product of their expectations, see Theorem 3.4. The proof is now completed by observing that $m(t)$ is the generating function of the χ_n^2-distribution. ■

Exercise 3.1 Let $U \sim U(0,1)$. Let F be continuous and strictly increasing C.D.F.. Show that $Y = F^{-1}(U)$ is a random variable having C.D.F. F. Explain, using the above, how we can computer generate samples from an Exponential(λ) distribution by using the standard random generator of the computer, which generates random numbers between 0 and 1.

Exercise 3.2 Let X and Y be independent random variables, with $Y > 0$. Show that the random variable $U = X/Y$ has density

$$f_U(u) = \int_0^\infty f_X(uy)\, y\, f_Y(y)\, dy, \quad u \in \mathbb{R}. \tag{3.6}$$

Linear Transformations

$X_1, \ldots, X_n$. Let $\boldsymbol{z} = (z_1, \ldots, z_n)^T$ be a (column) vector in $\mathbb{R}^n$ and A an $(n \times m)$-matrix. The mapping $\boldsymbol{x} \mapsto \boldsymbol{x}$, with

$$\boldsymbol{x} = A\boldsymbol{z}$$

is called a **linear transformation**. Now consider a *random* vector $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^T$, and let

$$\boldsymbol{X} = A\boldsymbol{Z}\,.$$

Then $\boldsymbol{X}$ is a random vector in $\mathbb{R}^m$. Again, in principle, if we know the distribution of $\boldsymbol{Z}$ then we can derive the distribution of $\boldsymbol{X}$. Let us first see how the expectation vector and covariance matrix are transformed.

Theorem 3.7 If $\boldsymbol{Z}$ has expectation vector $\boldsymbol{\mu_Z}$ and covariance matrix $\boldsymbol{\Sigma_Z}$, then the expectation vector and covariance matrix of $\boldsymbol{Z} = \boldsymbol{AX}$ are respectively given by

$$\boldsymbol{\mu_X} = \boldsymbol{A}\boldsymbol{\mu_Z} \tag{3.7}$$

and

$$\boldsymbol{\Sigma_X} = \boldsymbol{A}\ \boldsymbol{\Sigma_Z}\ \boldsymbol{A}^T\ . \tag{3.8}$$

PROOF. We have $\boldsymbol{\mu_X} = E(\boldsymbol{X}) = E(\boldsymbol{AZ}) = \boldsymbol{A}E(\boldsymbol{Z}) = \boldsymbol{A}\boldsymbol{\mu_Z}$ and

$$\begin{aligned}\boldsymbol{\Sigma_X} &= E(\boldsymbol{X}-\boldsymbol{\mu_X})(\boldsymbol{X}-\boldsymbol{\mu_X})^T = E(\boldsymbol{A}(\boldsymbol{Z}-\boldsymbol{\mu_Z})(\boldsymbol{A}(\boldsymbol{Z}-\boldsymbol{\mu_Z}))^T)\\ &= \boldsymbol{A}E((\boldsymbol{Z}-\boldsymbol{\mu_Z}))(\boldsymbol{Z}-\boldsymbol{\mu_Z})^T\boldsymbol{A}^T\\ &= \boldsymbol{A}\ \boldsymbol{\Sigma_Z}\ \boldsymbol{A}^T\end{aligned}$$

which completes the proof. ■

From now on assume $\boldsymbol{A}$ is a nonsingular (invertible) $(n\times n)-matrix$. If $\boldsymbol{Z}$ has density $f_{\boldsymbol{Z}}(\boldsymbol{z})$, what is the density $f_{\boldsymbol{X}}$ of $\boldsymbol{X}(\boldsymbol{x})$?

Consider Figure 3.4. For any fixed $\boldsymbol{z}$, let $\boldsymbol{x} = \boldsymbol{Az}$. Hence, $\boldsymbol{z} = \boldsymbol{A}^{-1}\boldsymbol{x}$. Consider the n-dimensional cube $C = [x_1, x_1+h]\times\cdots\times[x_n, x_n+h]$. Let D be the *image* of C under $\boldsymbol{A}^{-1}$, i.e., the parallelepiped of all points $\boldsymbol{z}$ such that $\boldsymbol{Az}\in C$. Then,

$$P(\boldsymbol{X}\in C)\approx h^n\, f_{\boldsymbol{X}}(\boldsymbol{x})\ .$$

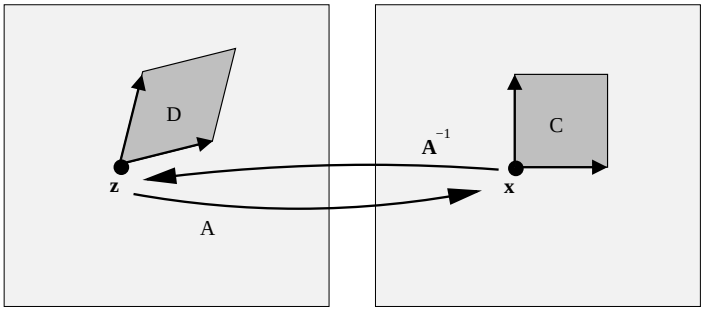


Figure 3.4: Linear transformation

Now recall from linear algebra that any n-dimensional rectangle with "volume" V is transformed into a n-dimensional parallelepiped with volume $V\,|\boldsymbol{A}|$, where $|A| = |\det(A)|$. Thus,

$$P(\boldsymbol{X}\in C) = P(\boldsymbol{Z}\in D)\approx h^n|\boldsymbol{A}^{-1}|\, f_{\boldsymbol{Z}}(\boldsymbol{z}) = h^n|\boldsymbol{A}|^{-1}\, f_{\boldsymbol{Z}}(\boldsymbol{z})$$

Letting h go to 0 we conclude that

$$\boxed{f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{f_{\boldsymbol{Z}}(A^{-1}\boldsymbol{x})}{|A|},\ \ \boldsymbol{x}\in\mathbb{R}^n.} \tag{3.9}$$

3.4 Jointly normal random variables

In this section we have a closer look at normally distributed random variables and their properties. Also, we will introduce normally distributed random *vectors*.

It is helpful to view normally distributed random variables as simple transformations of standard normal random variables. For example, let $Z \sim N(0,1)$. Then, Z has density f_Z given by

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{z^2}{2}}.$$

Now consider the transformation

$$X = \mu + \sigma Z.$$

Then, by (3.5) X has density

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

In other words, $X \sim N(\mu, \sigma^2)$. We could also write this as follows, if $X \sim N(\mu, \sigma^2)$, then $(X-\mu)/\sigma \sim N(0,1)$. This **standardisation** procedure was already mentioned in Section 2.5.4.

Let us generalise this to p dimensions. Let $Z_1, \ldots, Z_p$ be independent and standard normal random variables. The p.d.f. of $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^T$ is given by

$$f_{\mathbf{Z}}(\boldsymbol{z}) = (2\pi)^{-p/2} \mathrm{e}^{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{z}}, \quad \boldsymbol{z} \in \mathbb{R}^n. \tag{3.10}$$

Consider the transformation

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{B}\,\boldsymbol{Z}, \tag{3.11}$$

for some $(m \times p)$ matrix $\boldsymbol{B}$. Note that by Theorem 3.7 $\boldsymbol{X}$ has expectation vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^T$. Any random vector of the form (3.11) is said to have a **normal** (multivariate normal) distribution. We write $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Suppose $\boldsymbol{B}$ is a nonsingular $(p \times p)$-matrix. Then, by (3.9) the density of $\boldsymbol{Y} = \boldsymbol{X} - \boldsymbol{\mu}$ is given by

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{|\boldsymbol{B}|\sqrt{(2\pi)^p}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{B}^{-1}\boldsymbol{y})^T\boldsymbol{B}^{-1}\boldsymbol{y}} = \frac{1}{|\boldsymbol{B}|\sqrt{(2\pi)^p}} \mathrm{e}^{-\frac{1}{2}\boldsymbol{y}^T(\boldsymbol{B}^{-1})^T\boldsymbol{B}^{-1}\boldsymbol{y}}.$$

We have $|\boldsymbol{B}| = \sqrt{|\boldsymbol{\Sigma}|}$ and $(\boldsymbol{B}^{-1})^T\boldsymbol{B}^{-1} = (\boldsymbol{B}^T)^{-1}\boldsymbol{B}^{-1} = (\boldsymbol{B}\boldsymbol{B}^T)^{-1} = \boldsymbol{\Sigma}^{-1}$, so that

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{\sqrt{(2\pi)^p\,|\boldsymbol{\Sigma}|}} \mathrm{e}^{-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}}.$$

Because $\boldsymbol{X}$ is obtained from $\boldsymbol{Y}$ by simply adding a constant vector $\boldsymbol{\mu}$, we have $f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{\boldsymbol{Y}}(\boldsymbol{x} - \boldsymbol{\mu})$, and therefore

$$\boxed{f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^p \, |\boldsymbol{\Sigma}|}} \, \mathrm{e}^{-\frac{1}{2} \, (\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}, \quad \boldsymbol{x} \in \mathbb{R}^p \, .} \tag{3.12}$$

Note that this formula is very similar to the 1-dimensional case.

Example 3.9 Consider the 2-dimensional case with $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$, and

$$B = \begin{pmatrix} \sigma_1 & 0 \\ \sigma_2 \rho & \sigma_2 \sqrt{1-\rho^2} \end{pmatrix} \, . \tag{3.13}$$

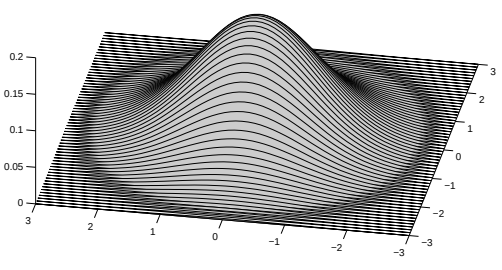
The covariance matrix is now

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix} \, . \tag{3.14}$$

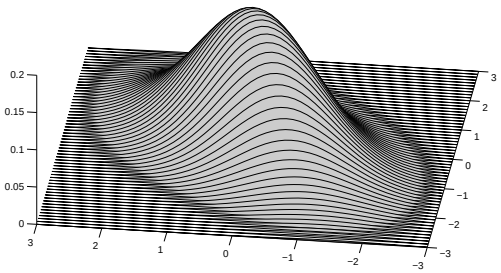
Therefore, the density is

$$f_{\mathbf{X}}(\boldsymbol{x}) = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1-\rho^2}} \, \exp \Bigg\{ - \frac{1}{2(1-\rho^2)} \left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1 \sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} \right) \Bigg\}. \tag{3.15}$$

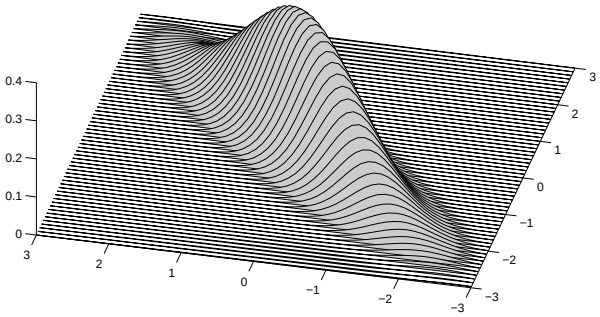
Here are some pictures of the density, for $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1$, and for various ρ.



$\rho = 0$



$\rho = 0.5$



$\rho = 0.9$

We say that $(X_1, X_2)^T$ has a **bivariate normal** distribution. Note that in this example $E(X_i) = \mu_i, i = 1, 2$. Moreover, since we have chosen $\boldsymbol{B}$ such that the covariance matrix has the form (3.14), we have $\text{var}(X_i) = \sigma_i^2, i = 1, 2$, and $\rho(X_1, X_2) = \rho$. We will see shortly that X_1 and X_2 both have univariate normal distributions.

Compare the following with property 8 of Table 3.1.

Theorem 3.8 If X_1 and X_2 have a bivariate normal distribution then

$$\text{cov}(X_1, X_2) = 0 \quad \Longrightarrow \quad X_1 \text{ and } X_2 \text{ are independent.}$$

PROOF. If $\text{cov}(X_1, X_2) = 0$, then $\boldsymbol{B}$ in (3.13) is a diagonal matrix. Thus, trivially $Y_1 = \sigma_1 X_1$ and $Y_2 = \sigma_2 X_2$ are independent. ■

The following theorem is of importance for the proofs of many theorems in part II.

Theorem 3.9 (Fisher's lemma) Let $Z_1, \ldots, Z_n$ be independent and standard normal, and let $(X_1, \ldots, X_p)^T = \boldsymbol{B}(Z_1, \ldots, Z_p)^T$. If $\boldsymbol{B}$ is **orthogonal**, then $X_1, \ldots, X_p$ are independent and standard normal as well.

PROOF. The proof is very simple. If $\boldsymbol{B}$ is orthogonal, then $\boldsymbol{B}\boldsymbol{B}^T = I$, the identity matrix. Hence the density of $\boldsymbol{X}$ in (3.12) is of the form (3.10), which is the density corresponding to a vector of i.i.d. $N(0, 1)$ random variables. ■

One of the most (if not the most) important properties of the normal distribution is that linear combinations of independent normal random variables are normally distributed. Here is a more precise formulation.

Theorem 3.10 If $X_i \sim N(\mu_i, \sigma_i^2)$, independently, for $i = 1, 2, \ldots, p$, then

$$\boxed{Y = a + \sum_{i=1}^{p} b_i X_i \sim N\left(a + \sum_{i=1}^{n} b_i \mu_i, \sum_{i=1}^{p} b_i^2 \sigma_i^2\right) .} \tag{3.16}$$

PROOF. The easiest way to prove this is by using moment generating functions, similar to the proof of Theorem 3.6. First, recall that the MGF of a $N(\mu, \sigma^2)$ distributed random variable X is given by

$$m_X(t) = \mathrm{e}^{\mu t + \frac{1}{2}\sigma^2 t^2} .$$

Let M_Y be the moment generating function of Y. Since $X_1, \ldots, X_n$ are inde-

pendent, we have

$$\begin{aligned} m_Y(t) &= E(\exp\{at + \sum_{i=1}^{n} b_i X_i t\}) \\ &= \mathrm{e}^{at} \prod_{i=1}^{n} m_{X_i}(b_i t) \\ &= \mathrm{e}^{at} \prod_{i=1}^{n} \exp\{\mu_i (b_i t) + \frac{1}{2}\sigma_i^2 (b_i t)^2\} \\ &= \exp\{ta + t\sum_{i=1}^{n} b_i \mu_i + \frac{1}{2}\sum_{i=1}^{n} b_i^2 \sigma_i^2 t^2\}, \end{aligned}$$

which is the MGF of a normal distribution of the form (3.16). ■

Remark 3.6 Note that from Theorems 3.3 and 3.5 we had already established the expectation and variance of Y in (3.16). But we have now found that the *distribution* is normal.

Example 3.10 I promised at the end of Example 3.9 to show that Z_1 and Z_2 both have a normal distribution. This follows now from Theorem 3.10 and the fact that both Z_1 and Z_2 are linear combinations of independent normally distributed random variables.

Example 3.11 A machine produces ball bearings with a $N(1, 0.01)$ diameter (cm).The balls are placed on a sieve with a $N(1.1, 0.04)$ diameter. The diameter of the balls and the sieve are assumed to be independent of each other.
Question: What is the probability that a ball will fall through?

Answer: Let $X \sim N(1, 0.01)$ and $Y \sim N(1.1, 0.04)$. We need to calculate $P(Y > X) = P(Y - X > 0)$. But, $Z = Y - X \sim N(0.1, 0.05)$. Hence

$$P(Z > 0) = P\left(\frac{Z - 0.1}{\sqrt{0.05}} > \frac{-0.1}{\sqrt{0.05}}\right) = \Phi(0.447) \approx 0.67 \,,$$

where Φ is the C.D.F. of the $N(0, 1)$ distribution.

3.5 Limit Theorems

In this section we briefly discuss two of the main results in probability: the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT). Both are about sums of independent random variables.

Let $X_1, X_2, \ldots$ be independent and identically distributed random variables. For each n let

$$S_n = X_1 + \cdots + X_n \,.$$

Suppose $E(X_i) = \mu$ and $\mathrm{var(X_i)} = \sigma^2$. By the rules for expectation and variance we know that

$$E(S_n) = n\,E(X_1) = n\mu$$

and

$$\mathrm{var}(S_n) = n\,\mathrm{var}(X_1) = n\sigma^2.$$

Moreover, if we know the p.d.f. or p.f. of X_i, then we can (in principle) determine the p.d.f. or p.f. of S_n.

The law of large numbers roughly states that S_n/n is close to μ, for large n. Here is a more precise statement.

Theorem 3.11 ((Weak) law of large numbers) If $X_1, \ldots, X_n$ are independent and identically distributed with expectation μ, then for all $\epsilon > 0$

$$\boxed{\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) = 0\,.}$$

The Central Limit Theorem says something about the approximate *distribution* of S_n (or S_n/n). Roughly it says this:

> *The sum of a large number of i.i.d. random variables has approximately a* **normal** *distribution*

Here is a more precise statement.

Theorem 3.12 (Central Limit Theorem) If $X_1, \ldots, X_n$ are independent and identically distributed with expectation μ and variance $\sigma^2 < \infty$, then for all $x \in \mathbb{R}$,

$$\boxed{\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x\right) = \Phi(x)},$$

where Φ is the C.D.F. of the standard normal distribution.

In other words, S_n has approximately a normal distribution with expectation $n\mu$ and variance $n\sigma^2$.

To see the CLT in action consider Figure 3.5. The first picture shows the p.d.f's of $S_1, \ldots, S_4$ for the case where the X_i have a U$[0, 1]$ distribution. The second show the same, but now for an Exponential(1) distribution. We clearly see convergence to a bell shaped curve.

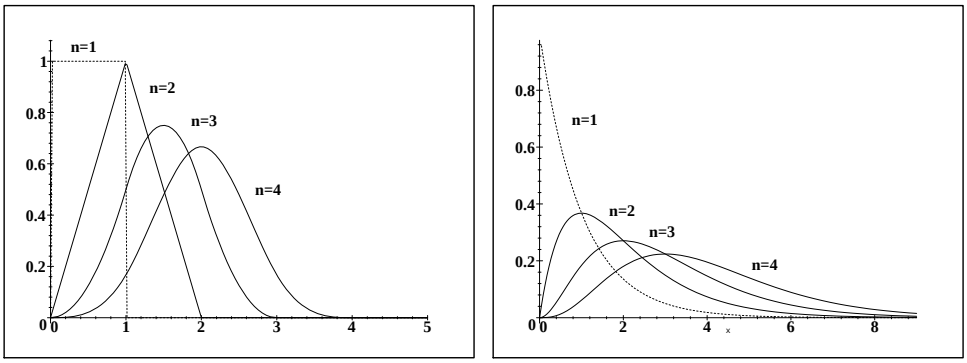


Figure 3.5: *Illustration of the CLT for the uniform and exponential distribution*

The CLT is not restricted to continuous distributions. For example, Figure 3.6 shows the C.D.F. of S_{30} in the case where the X_i have a Bernoulli distribution with success probability 1/2. Note that $S_{30} \sim B(30, 1/2)$, see Example 3.3.

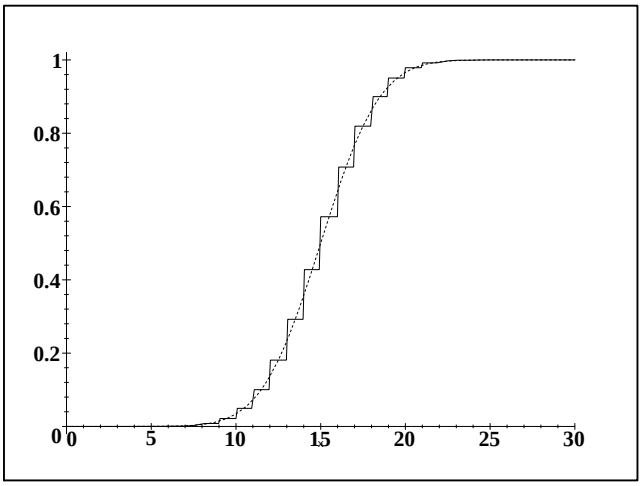


Figure 3.6: *The C.D.F. of a* $B(20, 1/2)$ *distribution and its normal approximation.*

In general we have:

Theorem 3.13 Let $X \sim B(n, p)$. For large n we have

$$P(X \leq k) \approx P(Y \leq k),$$

where $Y \sim N(np, np(1-p))$.

As a rule of thumb, the approximation is accurate if both np and $n(1-p)$ are larger than 5.

There is also a CLT for random *vectors*.

The multidimensional version is as follows:

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independent identically distributed random vectors with expectation vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then for large n

$$\boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$$

has approximately a *multivariate normal* distribution with expectation vector $n\boldsymbol{\mu}$ and covariance matrix $n\boldsymbol{\Sigma}$.

Here is an application of multidimensional CLT.

Theorem 3.14 Let $\boldsymbol{U} = (U_1, \ldots, U_k)^T \sim Mult(n, p_1, \ldots, p_k)$. The random vector has for large n approximately a multivariate normal distribution with expectation vector

$$\boldsymbol{\mu}_{\boldsymbol{U}} = n\,(p_1, \ldots, p_k)^T \tag{3.17}$$

and covariance matrix

$$\boldsymbol{\Sigma}_{\boldsymbol{U}} = n \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_k \\ -p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_k \\ \vdots & \vdots & \ddots & \vdots \\ -p_1p_k & -p_2p_k & \cdots & p_k(1-p_k) \end{pmatrix}. \tag{3.18}$$

Proof. Think of U_j as the total number of balls in urn j, if we throw n balls independently into k urns, with probabilities $p_1, \ldots, p_k$. Let $X_{ij} = 1$ if at the ith throw urn j is chosen, and $X_{ij} = 0$ if at the ith throw another urn is chosen. For $i = 1, \ldots, n$ define the vector $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ik})^T$. Hence, each $\boldsymbol{X}_i$ is a random vector of $k-1$ zeros and and a single 1. The vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent and identically distributed and, moreover, we have

$$\boldsymbol{U} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n.$$

Applying the CLT for random vectors, we have that $\boldsymbol{U}$ has for large n approximately a multivariate normal distribution. To complete the proof it only remains to be shown that the expectation vector and covariance matrix of any vector $\boldsymbol{X}_\ell$ are given by the right-hand side of (3.17) and (3.18), *without* the factor n. Let us write $\boldsymbol{X}_\ell = (I_1, \ldots, I_k)^T$. Note that each I_j is a Bernoulli random variable with success parameter p_j. Hence, we have $E(I_j) = p_j$ and $\mathrm{cov}(I_j, I_j) = \mathrm{var}(I_j) = p_j(1-p_j)$, $j = 1, \ldots, k$. Moreover, since exactly *one* of the random variables $I_1, \ldots, I_k$ is 1 and the rest are zero, we have for $i \neq j$, $\mathrm{cov}(I_i, I_j) = E(I_iI_j) - E(I_i)E(I_j) = 0 - p_ip_j = -p_ip_j$, as required. ■

Part II

Statistics

Chapter 4

Statistical Inference

Statistics is about gathering, summarising, analysing and interpreting **data**. There are various branches of statistics, the main three are:

1. **Data analysis.** The focus is here on *summarising* the data so that the main structures and relationships become apparent.
2. **Classical statistics.** Here the data are viewed as outcomes of some random experiment which can be modelled via one or more **parameters**. On the basis of both this model and of the observed data, we wish to draw conclusions and make decisions about reality. This usually involves estimating the unknown parameters or testing whether they belong to some set or not.
3. **Bayesian statistics.** In this approach the model parameters themselves are random and summarise in some way the information we have about reality. The observed data are used to update/change this information by applying Bayes' formula.

In these course notes the focus will be mainly on **classical statistics**. However, we will start with a little bit of data analysis. You will also get to practise this in the computer labs.

4.1 Data analysis

Example 4.1 (Milk cartons) Consider a machine that fills 1-litre cartons of milk. The factory is required by law to put at least 1 litre of milk in each packet. Since there is always some "variability" in filling process the machine is set to a slightly higher *target* of 1.10 litres. Note that if the machine was set to a target of 1 litre, then about half of milk cartons would be filled with less than 1 litre of milk!

To see if the machine is still "on target", we measure the volumes of 10 randomly selected packets of milk. Suppose the result (in litres) is given below.

1.035	1.019	1.068	1.084	1.077	0.998	1.041	1.037	1.051	1.102

Do these numbers show that the machine is still working properly? In the chapters to come we will come back to this example, and learn how to answer this question using a precise statistical framework.

There are various ways to *summarise* the data numerically. Suppose $x_1, x_2, \ldots, x_n$ are measurement data. In Example 4.1 we have $n = 10$. The **median of the data**, or **sample median**, is the observation "in the middle" of the data. That is, we sort the data from smallest to biggest and take the middle value. If n is even we take the average of the two middle observations. The **first quartile** (Q1) of the data is the median of the ordered list of observations below the location of the median; The **third quartile** (Q3) of the data is the median of the ordered list of observations above the location of the median. The **sample median** and the **sample mean**, defined as

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}\,, \tag{4.1}$$

are measures of the *central tendency* of the data. In the next chapter we will see that the (bias-corrected) **sample variance**

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{4.2}$$

is a useful measure for the *spread* of the data. $s = \sqrt{s^2}$ is called the **standard error (se)**. Another measure of the spread is the **range** of the data, which is defined as the **maximum** minus the **minimum** of the data:

$$r = \max\{x_1, \ldots, x_n\} - \min\{x_1, \ldots, x_n\}.$$

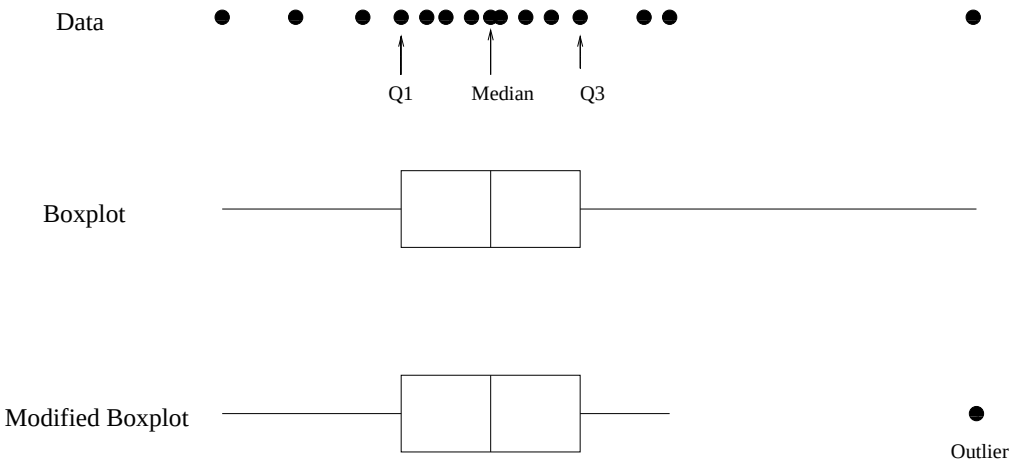
R gives the following descriptive statistics for our milk example:

```
Variable             N       Mean     Median      se
litre               10     1.0512     1.0460  0.0317

Variable       Minimum    Maximum         Q1         Q3
milk            0.9980     1.1020     1.0310     1.0787
```

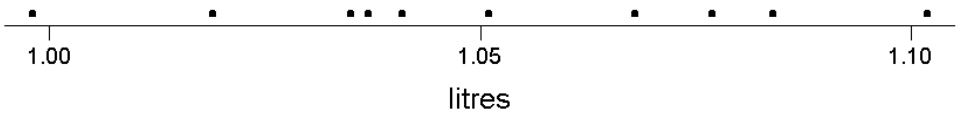
Other ways of summarising data include **dotplots** and **boxplots**. A boxplot summaries 5 numbers:

1. Minimum of the data
2. Maximum of the data
3. Medium of the data
4. First quartile of the data (Q1)
5. Third quartile of the data (Q3)

In a **modified boxplot**, the "whiskers" (the lines that are sticking out of the box) can extend a maximum of 1.5(Q3-Q1). **Outliers** are plotted separately. The figure below illustrates the concepts.



Here is a dotplot for the data in Example 4.1:



Note that for this specific example it does not make much sense to draw a boxplot, since we only have 10 observations. If we had more data, we could, in addition to a boxplot, draw a **histogram** of the data. That is, we count how many observations there are in each of various preselected "classes" and plot the results in a bar chart. Of course a different choice of the classes will give a (slightly) different picture.

In Figure 4.1 a histogram is given for the volume (in litres) of 100 milk cartons. We note that the measurements were taken on a different day than the 10 measurements given in Example 4.1.

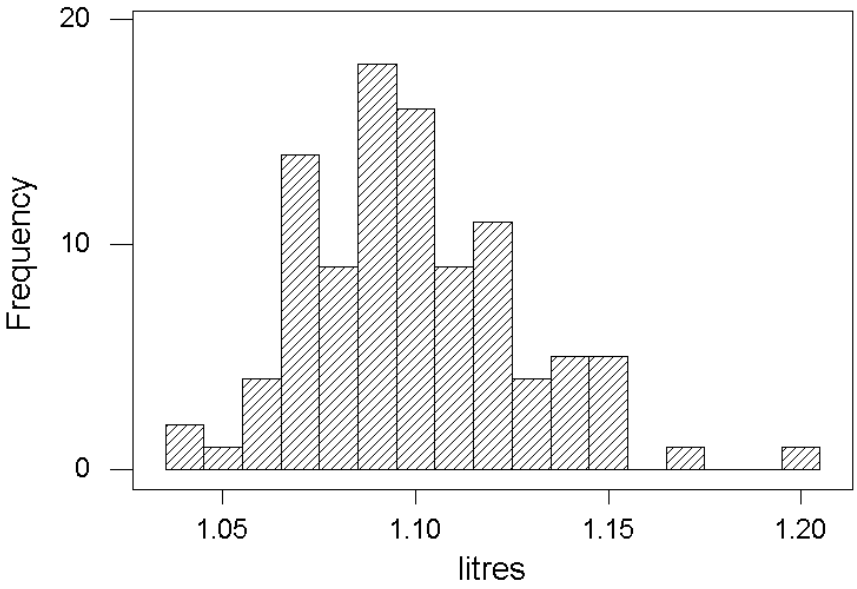


Figure 4.1: Histogram for the volume of 100 milk cartons

4.2 Modelling data

The mathematical approach to statistical inference is illustrated in Figure 4.2. The starting point is some real-life situation/problem (*Reality*) and a corresponding set of experimental *Data*. On the basis of that data we wish to say something about the real-life situation. The second step consists of finding a mathematical *Model* for the data. This model contains what we know about the Reality and how the data was obtained. *Within* the model we carry out our calculations and analysis. This leads to conclusions about model. Finally, the conclusions about the model are translated into conclusions about the Reality. In order to completely understand (classical) mathematical statistics it is imperative that you understand this way of thinking and make it your own.

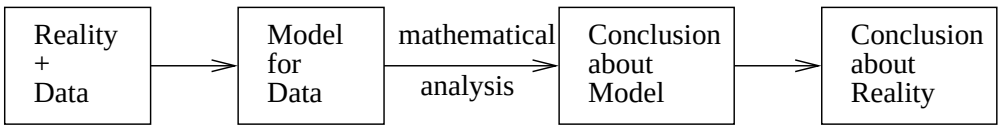


Figure 4.2: The modelling process in Statistical Inference

Example 4.2 Let us see how the procedure in Figure 4.2 works for the situation in Example 4.1. The main question here (the "Reality" part) is whether the machine is correctly set (with a target of 1.100 litres) or not. The 10 observations (the Data) should give us an idea whether this is true or not. What do we know about the reality and the data? We expect that if we would take another 10 observations, then the values would be different. In fact, we can view the observations as outcomes of a *random experiment*. Namely, the exper-

iment that consists of measuring the volume of 10 cartons of milk. A reasonable model for this random experiment, would be:

Let $X_1, \ldots, X_n$ ($n = 10$) be the milk volumes. We assume that

1. $X_1, \ldots, X_n$ are all **independent** of each other,
2. $X_1, \ldots, X_n$ all have the **same** distribution.

In other words, $X_1, X_2, \ldots, X_n$ is a *random sample* from some sampling distribution (see also Remark 3.4). The histogram in Figure 4.1 suggests that a *normal* sampling distribution may be appropriate. What we're saying is that if we generated outcomes of 10 independent normally distributed random variables, then the results would be very similar to what we obtained in real life. What about the parameters of the normal distribution; should we take for example μ equal to $\bar{x} = 1.0512$? No! We shouldn't because if we would repeat the experiment, we would get a different mean. So the true mean μ is not equal to the sample mean. The trick is *not* to specify μ and σ^2, and to formulate the model as above with *unknown* parameters μ and σ. We write our model for the milk data as

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2),$$

where μ and σ^2 are unknown. This completes the second step of Figure 4.2. Using the above model, we can identify two types of statistical inference:

1. **Estimation** – estimate the unknown parameters μ and σ^2 from the data.
2. **Hypothesis testing** – decide if the parameter μ (or σ^2) lies in predescribed set or not. E.g., $\mu = 1.100$ or not.

In *general*, suppose $x_1, \ldots, x_n$ are the observed data from some experiment. We view $x_1, \ldots, x_n$ as outcomes of random variables $X_1, \ldots, X_n$. Think of these as the observations that we would make if we carried the experiment out "tomorrow". Here is what we mean by a "model for the data".

> The statistical model for the random variables $X_1, \ldots, X_n$ forms the **model** for the data.

Most often we assume that the random variables $X_1, \ldots, X_n$ have a joint distribution which is completely specified up to a parameter (or parameter vector) θ. For example, in Example 4.2 the joint distribution of $X_1, \ldots, X_n$ is completely specified apart from the unknown parameters μ and σ^2 (or, equivalently, the unknown parameter vector $\boldsymbol{\theta} = (\mu, \sigma^2)^T$.

Remark 4.1 Suppose the distribution of $X_1, \ldots, X_n$ depends on an unknown parameter (vector) θ. Then any probability involving these random variables,

such as $P(X_1 \leq 3, X_2 + X_3 > 5)$, depends on θ. To emphasise this dependence we will sometimes write P_θ instead of P for the probability measure. For example, the probability above would be written as $P_\theta(X_1 \leq 3, X_2 + X_3 > 5)$. A similar notation is used for expectations, i.e., E_θ instead of E.

We can formalise this by describing the model via a *family* of probability measures $\{P_\theta,\, \theta \in \mathbf{\Theta}\}$, where $\mathbf{\Theta}$ is the set of all possible parameter values, usually a subset of $\mathbb{R}$ or $\mathbb{R}^d$, where d is the number of unknown parameters.

Chapter 5

Estimation

5.1 Estimate and Estimator

Example 5.1 We continue the milk example, see Examples 4.1 and 4.2. Our model for the data was formulated in terms of independent random variables $X_1, \ldots, X_n$ each with a $N(\mu, \sigma^2)$ distribution. Here μ is the current target value. The important thing to observe is that μ is unknown, and will never be known! But we "estimate" μ from the data $x_1, \ldots, x_n$. Namely, it is plausible that μ will be "close" to the *sample mean* $\bar{x} = (x_1 + \cdots + x_n)/n$. Hence we could estimate μ by $\bar{x} = 1.0512$. Another candidate could be the sample *median*, in this case 1.0460. Which one of the two is a better estimate of the "true mean" μ?

In general the situation is as follows. We have a model for the data $x_1, \ldots, x_n$ in terms of random variables $X_1, \ldots, X_n$, whose distribution is completely specified up to an unknown parameter (or parameter vector) θ. We wish to estimate θ on the basis of $x_1, \ldots, x_n$ only. Specifically, we wish to find a function T of the data such that number $t = T(x_1, \ldots, x_n)$ is close to the unknown θ. We call t an **estimate** of θ. The corresponding *random variable* $T(X_1, \ldots, X_n)$ is called an **estimator** of θ. A function such as T above that only depends on the data *but not on any unknown parameter* is called a **statistic**.

Example 5.2 We randomly select 10 thermometers from a large number of thermometers, and test their quality. We wish to say something about the unknown overall proportion p of thermometers that are compliant with the quality standards. The outcomes for the 10 selected thermometers are 0,1,1,0,1,0,0,1,0,0; here 0 means "faulty" and 1 means "OK".

We see the data as outcomes of random variables $X_1, \ldots, X_{10}$ such that

$$X_i = \begin{cases} 1 & \text{if the } i\text{th thermometer works properly} \\ 0 & \text{if the } i\text{th thermometer is defect.} \end{cases}$$

Moreover, we assume that $X_1, \ldots, X_n$ are independent with $P(X_i = 1) = p$, $i = 1, \ldots, 10$ (why is this valid?).

A "common sense" estimate for p is

$$T(x_1, \ldots, x_{10}) = \frac{1}{10} \sum_{i=1}^{10} x_i = 0.4 \; .$$

The corresponding estimator is the random variable $\frac{1}{10} \sum_{i=1}^{10} X_i$, which prescribes how we would obtain our estimate if we would do the experiment "tomorrow".

Remark 5.1 Note that our definition of estimator is very broad. Any function of the data is in principle an estimator. Of course some estimators are "better" than others. In Section 5.4 we give various criteria for comparing estimates and estimators.

Another issue is how to find or construct good estimators, rather than just using "common sense". In Sections 5.2 and 5.3 we will consider two systematic methods for finding estimators: the Method of Moments and the Method of Maximum Likelihood.

Remark 5.2 It is important to note that an estimator should be such that we can always *evaluate* it on the basis of the observations only. For example, in the example above $(X_1 + X_4)/2$ is an estimator of p, but $(X_1 + E(X_4))/2$ is not, because $E(X_4) = p$ is unknown!

5.2 Method of Moments

Suppose $x_1, \ldots, x_n$ are outcomes from a random sample $X_1, \ldots, X_n$ with a sampling distribution which is completely known up to an unknown parameter (vector) $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. It may not be easy to directly estimate $\theta_1, \ldots, \theta_k$, but the *moments* of the sampling distribution can be easily estimated. Namely, if X is distributed according to the sampling distribution, then the r-th moment of X (if it exists), i.e., $E(X^r)$, can be sensibly estimated using the **sample moment**

$$\frac{1}{n} \sum_{i=1}^{n} X_i^r \; .$$

The theoretical moments are a function of $\boldsymbol{\theta}$; see also Remark 4.1). Thus, a possible way to estimate $\boldsymbol{\theta}$ is to solve the k nonlinear equations

$$E(X^r) = \frac{1}{n} \sum_{i=1}^{n} x_i^r, \quad r = 1, 2, \ldots, k$$

with respect to θ. The solution of these equations is taken as the estimate of $\boldsymbol{\theta}$. Often the equations have to be solved numerically.

Example 5.3 Suppose $X_1, \ldots, X_n$ are a random sample from a general distribution with mean μ and variance σ^2. How can we estimate these parameters? The Method of Moments suggests that we should estimate μ using the sample mean

$$\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n) \,, \tag{5.1}$$

see also (4.1). The corresponding random variable, written as $\bar{X}$, is also called the **sample mean**. Moreover, since for all i

$$E(X_i^2) = \mu^2 + \sigma^2$$

the Method of Moments suggests the following estimate for σ^2:

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \,. \tag{5.2}$$

Note that this can be written as $s^2\,(n-1)/n$, where s^2 is the (bias-corrected) sample variance defined in (4.2). The random variable corresponding to (4.2) is also called the (bias-corrected) **sample variance**, and is usually still written as s^2, and not S^2. The reason why s^2 is often preferred over the maximum likelihood estimate $\hat{\sigma}^2 = s^2\,(n-1)/n$ as an estimator for σ^2 is that s^2 is *unbiased*. That is, the expectation of s^2 is equal to σ^2. However, there are also valid arguments for preferring $\hat{\sigma}^2$ over s^2. More about this in Section 5.4.

Example 5.4 Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} B(1, p)$. Since $E(X_i) = p$, the Method of Moments gives $\bar{x}$ as an estimate for p. This is what we used intuitively in Example 5.2.

Example 5.5 (Sample correlation coefficient) Suppose that instead of 1-dimensional data, we have *2-dimensional* data, i.e., $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are outcomes of a random sample of *vectors* $(X_1, Y_1), \ldots, (X_n, Y_n)$ with a two-dimensional sampling distribution. Think for example of a bivariate normal sampling distribution. Suppose ρ is the true *correlation coefficient* of X and Y, where (X, Y) are distributed according to our two-dimensional sampling distribution. We can estimate ρ from the data, by using the same "moment matching" ideas as in the 1-dimensional case. In particular, write

$$\rho = \rho(X, Y) = \frac{E(XY) - \mu_X\,\mu_Y}{\sigma_X\,\sigma_Y}, \tag{5.3}$$

where μ_X and μ_Y are the expectations of X and Y, respectively, and σ_X and σ_Y are the standard deviations of X and Y, respectively. We can estimate these parameters as before from (5.1) and (5.2). Moreover, we can estimate $E(XY)$ by

$$\frac{1}{n}\sum_{i=1}^{n} X_i Y_i \,,$$

and hence we can estimate the numerator of (5.3) by

$$\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\,\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})\ .$$

This leads to the following estimator of ρ :

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\,\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}, \tag{5.4}$$

which is called the **sample correlation coefficient**.

5.3 Maximum Likelihood Method

The Maximum Likelihood Method is a method for constructing "good" estimates/estimators by considering how likely the observation are, as a function of the parameter θ.

Example 5.6 As part a quality control process we randomly select 10 spark plugs from a large batch. Two of the spark plugs are defective. How would you estimate on the basis of this the proportion of defective plugs in the batch?

The answer should obviously be $2/10 = 0.2$. Let's look at a systematic method for deriving this. We first formulate our model. Let X be the number of defective spark plugs in the sample. We assume $X \sim B(10, p)$, with $0 \le p \le 1$ unknown. We let $f(x; p)$ denote the probability function of X given by

$$\begin{aligned} f(x;p) &= P\{X = x\} \\ &= \binom{n}{x} p^x (1-p)^{n-x}. \end{aligned}$$

We have observed the event $\{X = 2\}$. Suppose that $p = 0.4$. How likely would the event $\{X = 2\}$ be? The answer is

$$f(2; 0.4) = \binom{10}{2}(0.4)^2(0.6)^8 = 0.121\ .$$

Similarly,

$$f(3; 0.3) = \binom{10}{2}(0.3)^2(0.7)^8 = 0.233\ ,$$

which is almost twice as likely. The idea is now to choose as the estimate of p that value for which $f(2; p)$ is *maximal*. Hence, we have the following elementary calculus problem:

Determine $p \in [0, 1]$ for which

$$f(2; p) = \binom{10}{2} p^2 (1-p)^8$$

is maximal. Differentiating $p^2(1-p)^8$, or equivalently, $\log f(x;p)$, with respect to p gives

$$2/p - 8/(1-p) = 0\ ,$$

which shows that $p^2(1-p)^8$ is maximal at $p = 2/10$. We have thus found the *Maximum Likelihood Estimate* of p.

For continuous random variables we can do something similar. We just replace the probabilities with probability *densities*.

Example 5.7 Consider a production line in a chemical factory. It is known from previous studies that the production follows a $N(\mu, 50)$ distribution. The value of μ is however unknown. Suppose the daily production over 50 days gave a sample mean of 871 tons. As our model we take: $X_1, \ldots, X_{50} \overset{\text{i.i.d.}}{\sim} N(\mu, 50)$. We have observed the event $\{\bar{X} = 871\}$. How likely is this event for a given μ? The probability of $\{\bar{X} = 871\}$ is of course 0, for any μ. However, the probability density is $f_{\bar{X}}(871;\mu)$, where $f_{\bar{X}}(\bar{x};\ \mu)$, is the p.d.f. of $\bar{X}$. Note that we have included μ into the notation, to emphasise that this p.d.f. depends on the unknown parameter μ.

Since $\bar{X} \sim N(\mu, 1)$ (check yourself), we have

$$f_{\bar{X}}(871;\mu) = \frac{1}{\sqrt{2\pi}}\ \mathrm{e}^{-\frac{1}{2}(871-\mu)^2}\ .$$

As in the previous example, the idea is to choose as the estimate of μ that value for which the above density is maximal. Obviously $f_{\bar{X}}(871;\mu)$ *maximal* if and only if $(871-\mu)^2$ is *minimal*, which is the case when $\mu = 871$. Thus, the Maximum Likelihood Estimate of μ is found to be 871.

We now describe the Maximum Likelihood Method in more generality. First, we define the likelihood function.

Definition 5.1 Let $X_1, \ldots, X_n$ be discrete/continuous random variables with a joint p.f./p.d.f., $f_{X_1,\ldots,X_n}(x_1,\ \ldots,\ x_n;\ \theta)$, which depends on a parameter (vector) $\theta \in \mathbf{\Omega}$. Let $x_1, \ldots, x_n$ be the observed values of $X_1, \ldots, X_n$. We define the **likelihood function** $L(\theta)$ function for θ as

$$L(\theta;\ x_1,\ \ldots,\ x_n) = f_{X_1,\ldots,X_n}(x_1,\ \ldots,\ x_n;\ \theta). \tag{5.5}$$

Remark 5.3 It is important to note that the likelihood function is considered as a function of $\boldsymbol{\theta}$, for fixed parameter values $x_1, \ldots, x_n$, whereas the joint p.f. and p.d.f. are considered as functions of $x_1, \ldots, x_n$, for a fixed parameter value θ. However, the formulas are exactly the same.

Remark 5.4 (Notation) To simplify the notation, we sometimes write $L(\theta)$ instead of $L(\theta;\ x_1,\ \ldots,\ x_n)$.

Example 5.8 In Example 5.6 we have $n = 1$, $X_1 = X$, $\theta = p$ and $x_1 = 2$.

$$L(\theta; 2) = f(2; 2) = \binom{10}{2} p^2 (1-p)^8 \ .$$

Note that $L(p; 2)$ indicates how **likely** it is that $\bar{x}$ occurs as outcome of $\bar{X}$.

Example 5.9 In Example 5.7 it is tempting to consider the likelihood function corresponding to $X_1, \ldots, X_{50}$. But since we do not know the individual values $x_1, \ldots, x_{50}$, we cannot evaluate the likelihood function. The only information we have is $\bar{x}$. So, in fact in terms of the definition we have $n = 1$, $X_1 = \bar{X}$, $\theta = \mu$ and $x_1 = 871$. Thus,

$$L(\mu; 871) = f_{\bar{X}}(871;\, \mu) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(871-\mu)^2} \ .$$

Note that we shall see later with the concept of sufficiency that for the maximum likelihood (ML) estimation of μ of a normal distribution, we do not need to know the individual values $x_1, \ldots, x_n$, as the statistic $\sum_{j=1}^{n} X_j$ is sufficient for μ.

do not know the individual

The Method of Maximum Likelihood can simply be stated as:

Choose the estimate $\hat{\theta}$ such that $L(\hat{\theta})$ is **maximal**.

See Figure 5.1 for an illustration. A more precise definition is given in Definition 5.2.

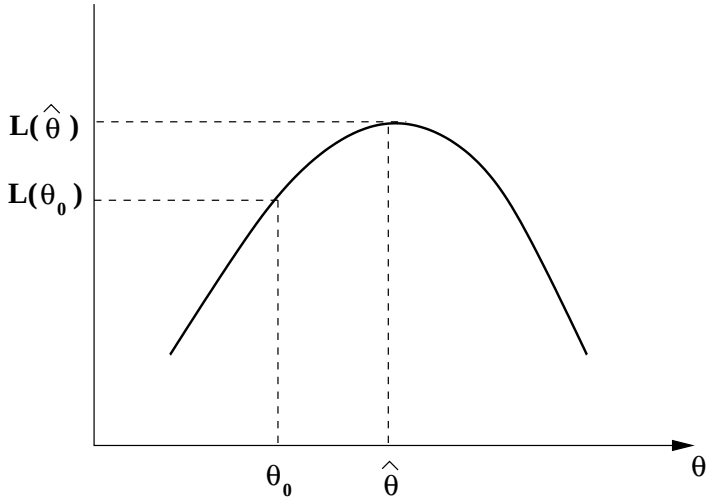


Figure 5.1: $\hat{\theta}$ is such that $L(\hat{\theta})$ is maximal.

Definition 5.2 Let $L(\theta; x_1, \ldots, x_n)$ be the likelihood function, based on the observations $x_1, \ldots, x_n$ of the random variables $X_1, \ldots, X_n$. If $\hat{\theta}$ is that value

of $\theta \in \Omega$ for which the function L is maximal, then $\hat{\theta}$ is called the **Maximum Likelihood Estimate** of θ.

Remark 5.5 Note that the Maximum Likelihood Estimate $\hat{\theta}$ is a function of the data. We can write this as $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$. The corresponding random variable $\hat{\theta}(X_1, \ldots, X_n)$ is called the **Maximum Likelihood Estimator** of θ. It is customary to denote both the estimate and the estimator of θ by the symbol $\hat{\theta}$. It should be clear from the context whether $\hat{\theta}$ should be interpreted as a random variable or the outcome of a random variable (a number). In the same spirit, we will use the acronym MLE for both Maximum Likelihood Estimator and Maximum Likelihood Estimate.

Before we look at more examples, let us consider the special case of interest where the random variables $X_1, \ldots, X_n$ form a *random sample* from a sampling distribution with p.f. or p.d.f. $f(\cdot; \theta)$, depending on whether we are dealing with the discrete or continuous case. Then the likelihood function is simply

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\, \theta).$$

Because L and $\log L$ attain their maximum at the same point ($\log L$ is a monotone increasing function), we can instead maximise the **log-likelihood function**

$$\log L(\theta) = \sum_{i=1}^{n} \log f(x_i; \theta) \ .$$

This is often easier because now we have to maximise over a *sum of terms* rather than over a *product of factors*. In particular, if θ is a *scalar*, then (often) the ML estimate can be found as the solution to

$$\frac{\partial L(\theta)}{\partial \theta} = 0 \ ,$$

or, equivalently, (because log is an increasing function) as the solution to

$$\frac{\partial \log L(\theta)}{\partial \theta} = 0 \ ,$$

If $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is a *vector*, then we need to evaluate

$$\frac{\partial \log L(\boldsymbol{\theta})}{\partial \theta_i} = 0 \quad (i = 1, \ \ldots, \ k) \ .$$

The equations above are called the **likelihood equations** or just the likelihood equation.

Finally, we give an important generalisation of Definition 5.2.

Definition 5.3 Let $\hat{\theta}$ be the MLE of θ. Then, for any function g the MLE of $g(\theta)$ is defined by $g(\hat{\theta})$.

Remark 5.6 There is a possible clash between Definitions 5.2 and 5.3. For example, suppose that g is a monotone function with inverse g^{-1}. Then, we can re-parametrise the likelihood function in terms of $\nu = g(\theta)$. Call the re-parametrised likelihood function $\tilde{L}$, thus ,

$$\tilde{L}(\nu) = L(g^{-1}(\nu)) \; . \tag{5.6}$$

The MLE of ν is by Definition 5.2 that number $\hat{\nu}$ for which $\tilde{L}(\hat{\nu})$ is maximal. Since L is maximal for $\theta = \hat{\theta}$, we can maximise the right-hand side of (5.6) by taking $g^{-1}(\nu)$ equal to $\hat{\theta}$. In other words, by taking ν equal to $g(\hat{\theta})$ we maximise $\tilde{L}(\nu)$. Consequently, $\hat{\nu} = g(\hat{\theta})$; thus in this case Definition 5.2 coincides with Definition 5.3. In a similar way it can be shown that no clash between the two definitions exist for general functions g.

5.3.1 MLE for the Binomial Distribution

Let $X \sim B(n, p)$. We wish to determine the MLE of p. The likelihood function for a observed value x of X is given by

$$L(p; x) = \binom{n}{x} p^x (1-p)^{n-x} \; .$$

Hence the log-likelihood is

$$\log L(p; x) = x \log p + (n-x) \log(1-p) + \text{ constant } \; .$$

Differentiation with respect to p gives the log-likelihood equation

$$\frac{x}{p} - \frac{n-x}{1-p} = 0 \; .$$

Solving this for p shows that ML estimate of p is $\hat{p} = x/n$, and the ML estimator of p is

$$\hat{p} = \frac{X}{n} \; .$$

Thus the ML method gives the "common sense" estimator for p.

5.3.2 MLE for the Normal Distribution

Consider the classical model for data,

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2).$$

Let f be the p.d.f. of each X_i, i.e.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \; x \in \mathbb{R}.$$

The p.d.f. depends on the (unknown) parameter vector $\boldsymbol{\theta} = (\mu, \sigma^2)^T$.

For a given outcome $x_1, \ldots, x_n$, the likelihood function is given by

$$L(\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[\sum_{i=1}^{n} -\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right] ,$$

for all $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Hence,

$$\log L(\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{n}{2}\{\log(2\pi) + \log\sigma^2)\} .$$

The likelihood equations are:

$$\frac{\partial \log L}{\partial \mu} = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0, \tag{5.7}$$

and

$$\frac{\partial \log L}{\partial \sigma^2} = \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{n}{2}\frac{1}{\sigma^2} = 0 . \tag{5.8}$$

From (5.7), it follows that for any value of σ^2 the MLE of μ is given by

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

Moreover, from (5.8), we have that

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

We can check that the solutions for $\hat{\mu}$ and $\hat{\sigma}^2$ correspond to a maximum. Thus the maximum likelihood estimators of μ and σ^2 are respectively

$$\hat{\mu} = \bar{X} \tag{5.9}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 . \tag{5.10}$$

Thus, in this case the ML method yields the same estimators as the Method of Moments in Example 5.3.

Remark 5.7 In the derivation (5.8), we have worked with σ^2 rather than σ.

5.4 Comparison of Estimators

Suppose we have various estimators of a parameter θ. How can we decide which is the best, or which is better than another? To assess this we could look at the *properties* of the (random) estimators.

For example, an often desired property is that the expectation of the estimator is equal to θ. Intuitively this means that "on average" the estimator gives the right value. We say that the estimator is **unbiased**.

Example 5.10 Let $X_1, \ldots, X_n$ be a random sample from a distribution with expectation μ and variance σ^2. Then the sample mean $\bar{X}$ is an unbiased estimator of μ. Namely,

$$E(\bar{X}) = E\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n}\,(E(X_1) + \cdots + E(X_n)) = \frac{1}{n}\,n\mu = \mu\ .$$

The (bias-corrected) sample variance s^2 is an unbiased estimator of σ^2, so the expected value of the MLE of σ^2, given by $\hat{\sigma}^2 = (n-1)s^2/n$, is equal to $\{(n-1)/n\}\sigma^2$. Its bias is thus equal to $-\sigma^2/n$.

An unbiased estimator does not always have to be better than a biased one. What *is* important is that an estimator is in some sense "close" to the parameter it wishes to estimate. An often used measure of closeness is the **mean squared error** (MSE) of the estimator, which is defined as

$$\mathrm{MSE}(T) = E(T-\theta)^2.$$

The MSE of T can be decomposed into two components, the variance of T and the square of the bias of T. To see this, we have that

$$\begin{aligned}
\mathrm{MSE}(T) &= E(T-\theta)^2 \\
&= E\{T - E(T) + E(T) - \theta)\}^2 \\
&= E[\{T-E(T)\}^2 + \{E(T)-\theta\}^2 + 2\{E(T)-\theta\}E\{T-E(T)\}] \\
&= \mathrm{var}(T) + \{\mathrm{bias}(T)\}^2.
\end{aligned}$$

Often in practice, reducing the bias of an estimator can increase its variance, sometimes more than the decrease in the square of the bias. We call this the variance-bias tradeoff. It is a very important concept in Statistics.

If an estimator has a smaller MSE than another for all values of the parameter to be estimated, then we say it is more efficient. The concept of efficiency is to be discussed in supplementary notes to be considered in the future.

Chapter 6

Confidence Intervals

In this chapter we cover confidence intervals.

Example 6.1 To prevent accidents at work a company has issued new safety measures. In the 50 months since these new measure were introduced the amount of lost working hours due to accidents was on average 91 hours per month. In the 50 months before the new measures were introduced the amount of lost working hours was on average 108 hours per month. In addition, the (bias-corrected) sample standard deviations are 14.1 and 14.3 (month). As an estimate for the difference in average working hours per month we take (of course) $91 - 108 = -17$ hours. What can we say about the *accuracy* of this estimate?

First we translate the above "reality" into a statistical model. Let $X_{11}, \ldots, X_{1,50}$ be the amount of lost hours in the 1st, 2nd, ..., 50th month *after* the new measures took effect, and similarly, let $X_{21}, \ldots, X_{2,50}$ denote the lost hours *before* the new measure took effect. The outcomes of the following statistics are known

$$\bar{X}_1 = \tfrac{1}{50}\sum_{j=1}^{50} X_{1j}, \qquad s_1^2 = \tfrac{1}{49}\sum_{j=1}^{50}(X_{1j} - \bar{X}_1)^2,$$

$$\bar{X}_2 = \tfrac{1}{50}\sum_{j=1}^{50} X_{2j}, \qquad s_2^2 = \tfrac{1}{49}\sum_{j=1}^{50}(X_{2j} - \bar{X}_2)^2.$$

Suppose now that $X_{11}, \ldots, X_{1,50}, X_{21}, \ldots, X_{2,50}$ are independent r.v.s, where $X_{11}, \ldots, X_{1,50}$ have the same distribution with mean μ_1 and variance σ_1^2 and $X_{21}, \ldots, X_{2,50}$ have the same distribution with mean μ_2 and *also* variance σ_2^2. The above mentioned "difference" that we wish to estimate is simply $\mu_1 - \mu_2$. As estimator we take $\bar{X}_1 - \bar{X}_2$. This is an unbiased estimator of $\mu_1 - \mu_2$. The *estimate* is the corresponding observed value of $\bar{X}_1 - \bar{X}_2$, which is therefore $91 - 108 = -17$, as above.

To say something about the accuracy we determine the variance of the estima-

tor, which is

$$\text{var}(\bar{X}_1 - \bar{X}_2) = \text{var}(\bar{X}_1) + \text{var}(\bar{X}_2) = \frac{\sigma^2}{50} + \frac{\sigma^2}{50} = \frac{\sigma^2}{25} \, .$$

Note that this number is unknown, since we do not know σ^2. However, we can estimate σ^2 by pooling the estimators s_1^2 and s_2^2 to give

$$\begin{aligned} s_p^2 &= \{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\}/(n_1 + n_2 - 2) \\ &= (14.1 + 14.3)/2 \\ &= 14.2 \end{aligned}$$

The square root $s_p^2/25$ is 2.84. The combined information above is often reported as

$$\mu_1 - \mu_2 = -17 \pm 2.84$$

or

$$\mu_1 - \mu_2 = -17 \pm 2 \times 2.84$$

However, we still would like to say that in some way the unknown quantity lies between two bounds. The following example shows a possible approach.

Example 6.2 Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, 1)$. We estimate μ with $\bar{X}$. The variance of $\bar{X}$ is $\text{var}(\bar{\text{X}}) = 1/\text{n}$, and therefore the standard deviation is $n^{-1/2}$. In the reporting style mentioned in the previous example, this could be written as $\bar{X} \pm n^{-1/2}$.

What is the probability that the interval $(\bar{X} - n^{-1/2}, \bar{X} + n^{-1/2})$ indeed contains the value μ? This probability can be calculated as

$$\begin{aligned} &P(\bar{X} - n^{-1/2} < \mu < \bar{X} + n^{-1/2}) = P(-n^{-1/2} < \bar{X} - \mu < n^{-1/2}) \\ &= P(-1 < n^{1/2}(\bar{X} - \mu) < 1) = \Phi(1) - \Phi(-1) = 2\Phi(1) - 1 = 0.68 \, , \end{aligned}$$

because $n^{1/2}(\bar{X} - \mu)$ has a standard normal distribution. Thus, if we would repeat the experiment many times, and get many outcomes of the interval $(\bar{X} - n^{-1/2}, \bar{X} + n^{-1/2})$, in only 68% of the cases our true μ would be contained in these intervals!

However, if we take our the interval above a bit wider, such as $(\bar{X} - 2n^{-1/2}, \bar{X} + 2n^{-1/2})$, then the probability that μ is contained in such an interval is much bigger. Can you calculate how much?

This leads to the following concept.

Definition 6.1 Let $X_1, \ldots, X_n$ be random variables with a joint distribution depending on a parameter $\theta \in \Theta$. Let $T_1 < T_2$ be statistics[1] such that

$$P_\theta(T_1 < \theta < T_2) = 1 - \alpha, \quad \text{for all } \theta \in \Theta. \tag{6.1}$$

[1] Thus, $T_i = T_i(X_1, \ldots, X_n)$, $i = 1, 2$ are functions of the data, but not of θ

Then the interval (T_1, T_2) is called a **confidence interval** for θ with confidence coefficient $(1-\alpha)$ or with confidence $100(1-\alpha)$ per cent. If t_1 and t_2 are the observed values of T_1 and T_2, then this realized or obserevd interval (t_1, t_2) is still called the **confidence interval** for θ with confidence coefficient $1-\alpha$.

Remark 6.1 The smaller we choose α the wider our confidence interval will be, and a very wide interval is not very useful. Common choices for α are $0.01, 0.05$ and 0.1.

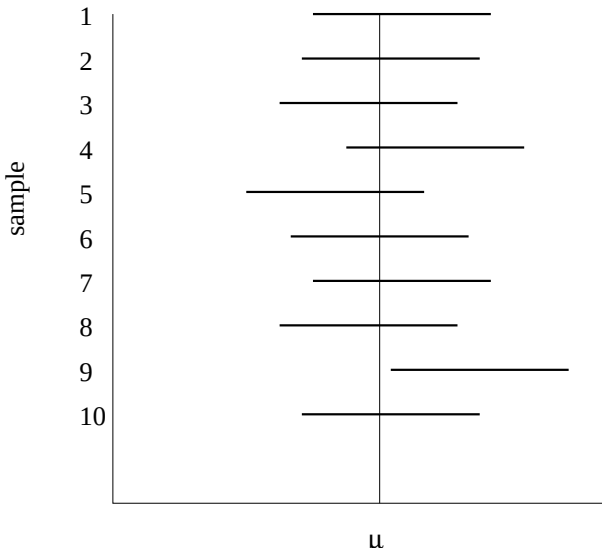
Remark 6.2 Instead of a confidence interval with confidence 95% we simply speak of a 95% confidence interval.

In general, we form a confidence interval for a parameter (for example, a mean or variance) or for a function of parameters (for example, a ratio of two variances) by constructing a pivot, $h(\boldsymbol{T}; \boldsymbol{\theta})$, which is a function of a vector T of statistics and of a parameter vector $\boldsymbol{\theta}$, but its distribution does not depend on any unknown parameters. Assuming we know the distributon of the pivot, we can form the statement

$$h_{\frac{1}{2}\alpha} < h(\boldsymbol{T}; \boldsymbol{\theta}) < h_{1-\frac{1}{2}\alpha}, \tag{6.2}$$

where h_α denotes the quantile of order α of the distribution of the pivot. We proceed to rearrange the probability statement (6.2)to give a confidence interval for the parameter of interest.

One has to be careful in stating what is actually meant by a confidence interval. Suppose in Example 6.2 we find a 90% confidence interval (9.5,10.5) for μ. Does this mean that $P(9.5 < \mu < 10.5)$? No! Since μ is a number the probability $P(9.5 < \mu < 10.5)$ is either 0 or 1, and we don't know which one, because we don't know μ. To find the meaning we have to go back to the definition of a confidence interval. There we see that the interval (9.5,10.5) is an *outcome* of a random confidence interval (T_1, T_2), say, such that $P(T_1 < \mu < T_2) = 0.9$. Note that μ is constant, but the interval bounds T_1 and T_2 are random. If we would repeat this experiment many times, then we would get many confidence intervals, as illustrated in the figure below.



Only in (on average) 9 out of 10 cases would these intervals contain our unknown μ. To put it in another way: Consider an urn with 90 white and 10 black balls. We pick at random a ball from the urn *but we do not open our hand to see what colour ball we have*. Then we are pretty confident that the ball we have in our hand is white. This is how confident you should be that the unknown μ lies in the interval $(9.5, 10.5)$.

The remainder of this chapter is about the construction of confidence intervals for a number of standard situations.

6.1 Normal distribution: one sample

Example 6.3 An oil company wishes to investigate how much on average each household in Melbourne spends on on petrol and heating oil per year. The company randomly select 51 households from Melbourne, and finds that these spent on average \$1136 on petrol and heating oil, with a (bias-corrected) sample standard deviation of \$178. We wish to construct a 95% confidence interval for the expected amount of money per year that the households in Melbourne spend on petrol and heating oil.

How do we do this? First, we need a model. Suppose we do this experiment (selecting 51 households) *tomorrow*. Let $X_1, \ldots, X_{51}$ be the amount the 51 households spend. We assume that the random variables $X_1, \ldots, X_{51}$ are independent and have the same distribution, because we select the households completely at random from the larger population. We wish to find a confidence interval for the expectation μ say of each X_j. Notice that we know nothing about the distribution of the X_j (the sampling distribution) and that our only data are the outcomes of the sample mean $\bar{X} = (X_1 + \cdots + X_{51})/51$ and the (bias-corrected) sample standard deviation $s = \sqrt{\frac{1}{50}\sum_{j=1}^{51}(X_j - \bar{X})^2}$. However, an often used model in these situations is to assume that the sampling

distribution is *normal* with mean μ and variance σ^2 (both unknown). We will show how confidence intervals can be constructed in this situation.

6.1.1 Confidence interval for μ

Consider $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$. We have seen that $\bar{X}$ is an unbiased estimator for μ. How can we find a confidence interval for μ?

We have seen that the probability distribution of $\bar{X}$ is $N(\mu, \sigma^2/n)$. Standardizing gives

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \; .$$

However, usually σ^2 is **not** known, and so then we replace Z with

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}},$$

i.e., replacing σ with the (bias-corrected) sample standard deviation: $s = \sqrt{s^2}$. What is the distribution of T? It is no longer standard normal, but we can imagine that for large n it is close to a standard normal distribution, because for large n s will be close to σ. In fact, T has a *t-distribution* (or **Student**-distribution, after W.S. Gosset, who discovered this distribution, and published under the pseudonym "Student"). The definition and various properties of the t-distribution are given in Section 2.5.8.

We now give a precise statement of what we indicated above.

Theorem 6.1 Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$. Then,

$$\frac{\bar{X} - \mu}{s/\sqrt{n}},$$

with $\bar{X} = n^{-1} \sum_{j=1}^{n} X_j$ and $s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{j=1}^{n} (X_j - \bar{X})^2}$, has a *Student* t-distribution with $n-1$ degrees of freedom.

PROOF. Write $T = \frac{Z}{\sqrt{V/(n-1)}}$, where $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ and $V = \frac{(n-1)s^2}{\sigma^2}$. The result now follows directly from Theorem 2.3, *provided* we can show that (a) $V \sim \chi^2_{n-1}$, and (b) Z and V are independent, or equivalently, that $\bar{X}$ and V are independent. We have proved the latter in Supplement 5. Appendix **??**. ■

For $X \sim t_n$ and $0 < \alpha < 1$ let $t_{m;\alpha}$ be the number such that

$$P(X \leq t_{m;\alpha}) = \alpha \; .$$

$t_{m;\alpha}$ is called the quantile of order α of the t_m distribution.

Now consider $T = \frac{\bar{X}-\mu}{s/\sqrt{n}}$. By Theorem 6.1, $T \sim t_{n-1}$. Thus,

$$P\left(-t_{n-1;1-\alpha/2} < \frac{\bar{X}-\mu}{s/\sqrt{n}} < t_{n-1;1-\alpha/2}\right) = 1-\alpha\ .$$

Rearranging gives

$$P\left(\bar{X} - t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}\right) = 1-\alpha\ .$$

Hence a $100\%(1-\alpha)$ confidence interval for μ is

$$\boxed{\left(\bar{X} - t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}},\ \bar{X} + t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}\right)\ .}$$

6.1.2 Confidence interval for σ^2

Next, we construct a confidence interval for σ^2. As before let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$. We have seen (Example 5.10) that the sample variance

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2$$

is an unbiased estimator for σ^2. What is the *distribution* of s^2? It turns out that s^2 times $(n-1)/\sigma^2$ has a χ^2 distribution.

Here is the precise statement of what was announced above.

Theorem 6.2 Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$. Then

$$\frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{j=1}^{n}(X_j - \bar{X})^2}{\sigma^2} = \sum_{j=1}^{n}\left(\frac{X_j - \bar{X}}{\sigma}\right)^2$$

has a chi-squared distribution with $n-1$ degrees of freedom. Notation: χ^2_{n-1}.

Proof. See Supplement 5. ■

For $W \sim \chi^2_m$ and $0 < \alpha < 1$, let $\chi^2_{m;\alpha}$ be the quantile of order α of the χ^2_m distribution, i.e., the number such that

$$P(W \le \chi^2_{m;\alpha}) = \alpha.$$

From Theorem 6.2 we know $\frac{(n-1)}{\sigma^2}s^2 \sim \chi^2_{n-1}$. Hence,

$$P\left(\chi^2_{n-1;\alpha/2} < \frac{(n-1)}{\sigma^2}s^2 < \chi^2_{n-1;1-\alpha/2}\right) = 1-\alpha\ .$$

Rearranging:

$$P\left(\frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}}\right) = 1-\alpha\ .$$

Hence a $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\boxed{\left(\frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}},\ \frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}}\right)\ .}$$

Example 6.4 On the label of a certain packet of aspirin it is written that the standard deviation of the tablet weight is 1.0 mg. To investigate if this is true we take a sample of 25 tablets and discover that the sample standard deviation is 1.3mg. A 95% confidence interval for σ^2 is

$$\left(\frac{24\times 1.3^2}{39.4},\ \frac{24\times 1.3^2}{12.4}\right) = (1.04, 3.27)\ ,$$

where we have used the table for the χ^2_{24} in the back to find $\chi^2_{24;0.025} = 12.4$ and $\chi^2_{24;0.975} = 39.4$. A 95% confidence interval for σ^2 is found by taking square roots,

$$(1.02, 1.81)\ .$$

Note that this CI does not contain the asserted weight of 1.0 mg. We therefore have some doubt whether the "true" standard deviation is indeed equal to 1.0 mg.

6.2 Normal distribution: two samples

Consider the situation of Example 6.1. We have *two* independent samples $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$ from, respectively, a $N(\mu_1, \sigma_2^2)$ and $N(\mu_2, \sigma_2^2)$ distribution. We wish to provide confidence intervals for $\mu_1 - \mu_2$ and σ_1^2/σ_2^2. The difference $\mu_1 - \mu_2$ tells us how the two *means* relate to each other, and σ_1^2/σ_2^2 gives an indication how the *variances* relate to each other.

6.2.1 Confidence interval for $\mu_1 - \mu_2$

Constructing a confidence interval for $\mu_1 - \mu_2$ is very similar to the one-sample case *provided* that we assume the **extra model assumption** that

the variances of the two samples are the same.

That is, we assume that $\sigma_1^2 = \sigma_2^2 = \sigma^2$, for some unknown σ^2. The analysis now proceeds as follows. First, check that the ML estimator for $\mu_1 - \mu_2$ is $\bar{X}_1 - \bar{X}_2$.

Next, observe that

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_2 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim N(0,1) \; .$$

However, if σ^2 is unknown, we should replace it with an appropriate estimator. For this we will use the pooled sample variance, s_p^2, which is defined as

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n - 1)s_2^2}{n_1 + n_2 - 2} \; , \tag{6.3}$$

where s_1^2 and s_2^2 are the sample variances for the X_{1j}'s and X_{2j}'s, respectively.

Similar to Theorem 6.1 we have

Theorem 6.3 We have

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has a t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

PROOF. See Supplement 5. ■

Using the pivot, $\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, we find (completely analogous to the one-sample case) that

$$\boxed{\bar{X}_1 - \bar{X}_2 \pm t_{n_1+n_2-2;1-\alpha/2}\, s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

is a $100\%(1-\alpha)$ confidence interval for $\mu_1 - \mu_2$.

Example 6.5 Let us return to Example 6.1. Here we have $n_1 = n_2 = 50$, and the outcomes for $\bar{X}_1 - \bar{X}_2$ and s_p are respectively $91 - 108 = -17$ and

$$s_p = \sqrt{\frac{49 \times 14.1^2 + 49 \times 14.3^2}{98}} = 14.20 \; ,$$

so that a 95% confidence interval for $\mu_1 - \mu_2$ is given by

$$(-17 - 1.98 \times 14.20/5,\; -17 + 1.98 \times 14.20/5) = (-19.84,\; -14.16),$$

where we used the table of the t distribution at the end of these notes to find $t_{98;0.975} \approx t_{100;0.975} = 1.98$. Note that in this table the 0.975 quantile of the t_{98} distribution is not provided. But we can see that it must be very close to that of the t_{100} distribution.

6.2.2 Confidence interval for σ_1^2/σ_2^2

It is important that you realize that the construction of the confidence interval for $\mu_1 - \mu_2$ only "works" when the variances σ_1^2 and σ_2^2 are assumed to be *equal*. To check this, we could compare the outcomes of s_1^2 and s_2^2. More precisely, we could construct a confidence interval for σ_1^2/σ_2^2, on the basis of the outcomes of the statistic s_1^2/s_2^2 and see if it contains the number 1. The distribution of this statistic is called the **F-distribution**, after R.A. Fisher, one of the founders of modern Statistics. The definition and various properties of the F-distribution are given in Section 2.5.7.

Theorem 6.4 Let $X_{11}, \ldots, X_{1n_1} \overset{\text{i.i.d.}}{\sim} N(\mu_1, \sigma_1^2)$, $X_{21}, \ldots, X_{2n_2} \overset{\text{i.i.d.}}{\sim} N(\mu_2, \sigma_2^2)$, and $X_{11}, \ldots, X_{1n_1}$, $X_{21}, \ldots, X_{2n_2}$ independent. Then

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

has an F_{n_1-1,n_2-1}-distribution, where $s_1^2 = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(X_{1j} - \bar{X}_1)^2$ and $s_2^2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(X_{2j} - \bar{X}_2)^2$.

Proof. Let $U = (n_1-1)s_1^2/\sigma_1^2$ and $V = (n_2-1)s_2^2/\sigma_2^2$. Then, by Theorem 6.2, $U \sim \chi^2_{n_1-1}$ and $V \sim \chi^2_{n_2-1}$. Moreover, U and V are independent. Now, if we write

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{U/(n_1-1)}{V/(n_2-1)},$$

then the result follows from Theorem 2.2. ■

For $X \sim F_{m_1,m_2}$ and $0 < \alpha < 1$, let $F_{m_1,m_2;\alpha}$ denote the number such that

$$P(X \leq F_{m_1,m_2;\alpha}) = \alpha\ ,$$

$F_{m_1,m_2;\alpha}$ is called the quantile of order α of the F_{m_1,m_2}-distribution. Recall (see (2.16)) that

$$\frac{1}{F_{m_1,m_2;1-\alpha}} = F_{m_2,m_1;\alpha}\ . \tag{6.4}$$

Using the pivot $\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$, which has an F_{n_1-1,n_2-1} distribution, we have

$$P\left(F_{n_1-1,n_2-1;\alpha/2} < \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} < F_{n_1-1,n_2-1;1-\alpha/2}\right) = 1-\alpha\ .$$

Rearranging:

$$P\left(\frac{1}{F_{n_1-1,n_2-1;1-\alpha/2}}\,\frac{s_1^2}{s_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{F_{n_1-1,n_2-1;\alpha/2}}\,\frac{s_1^2}{s_2^2}\right) = 1-\alpha\ .$$

So that by (6.4) we have

$$P\left(\frac{1}{F_{n_1-1,n_2-1;1-\alpha/2}}\frac{s_1^2}{s_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < F_{n_2-1,n_1-1;1-\alpha/2}\frac{s_1^2}{s_2^2}\right) = 1-\alpha\ .$$

It follows that

$$\boxed{\left(\frac{1}{F_{n_1-1,n_2-1;1-\alpha/2}}\frac{s_1^2}{s_2^2},\ F_{n_2-1,n_1-1;1-\alpha/2}\frac{s_1^2}{s_2^2}\right)}$$

is a $100(1-\alpha)\%$ confidence interval for σ_1^2/σ_2^2 .

Example 6.6 In Example 6.5 we assumed the variances σ_1^2 and σ_2^2 of the two samples were the same. To *check* this, we could construct a confidence interval for σ_1^2/σ_2^2 and see if it contains the value 1. We have $s_1^2/s_2^2 = 14.1^2/14.2^2$ and $F_{49,49;0.975} = 1.7622$ so that a 95% CI for σ_1^2/σ_2^2 is given by

$$\left(\frac{1}{1.7622}\frac{14.1^2}{14.2^2},\ 1.7622\frac{14.1^2}{14.2^2}\right) = (0.56, 1.74),$$

which clearly contains 1, so that there is on the basis of this no ground at the 5% level to suspect that the true variances are different. The value of 1.7622 was found via R. One could also use the table in the back of the notes, and "interpolate" via $((1.88+1.74)/2+(1.80+1.67)/2)/2 \approx 1.77$.

6.3 Binomial distribution: one sample

Example 6.7 In an opinion poll of 1000 registered voters, 227 voters say they will vote for the Greens. Give a 95% confidence interval for the proportion p of Green voters of the total population.

A systematic way to proceed is to view the data, 227, as the outcome of a random variable X (the number of green voters under 1000 registered voters) with a $B(1000, p)$ distribution. In other words, we view X as the total number of "Heads" (= votes green) in a coin flip experiment with some unknown probability p of getting Heads. Note that this is only a *model* for the data. In practice is not always possible to truly randomly select 1000 people from the population and find their true party preference. For example a randomly selected person may not wish to participate or could deliberately give the "wrong answer".

Now, let us proceed to make a confidence interval for p, in the general situation that we have an outcome of some random variable X with a $B(n, p)$ distribution. It is not so easy to find an exact confidence interval for p that satisfies (6.1) in Definition 6.1. Instead, when n is large we rely on the Central Limit Theorem to construct an *approximate* confidence interval. The reasoning is as follows:

For large n, X has approximately a $N(np, np(1-p))$ distribution. Let $\hat{p} = X/n$ denote the estimator of p. Then $\hat{p}$ has approximately a $N(p, p(1-p)/n)$ distribution. For any $0 < \alpha < 1$. Let z_α be the α-quantile of the standard normal distribution. Thus, with Φ the cdf of the $N(0,1)$ distribution, we have

$$\Phi(z_\alpha) = \alpha \; .$$

Then, using the pivot

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

we have

$$P\left(-z_{1-\alpha/2} < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < z_{1-\alpha/2} \right) \approx 1 - \alpha \; .$$

Rearranging gives:

$$P\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \right) \approx 1 - \alpha \; .$$

This would suggest that we take $\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$ as an (approximate) $(1-\alpha)$ confidence interval for p, were it not for the fact that the bounds still contain the unknown p! However, for large n the estimator $\hat{p}$ is close to the real p, so that we have

$$P\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) \approx 1 - \alpha \; .$$

Hence, an *approximate* $(1-\alpha)$-confidence interval for p is

$$\boxed{\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

Example 6.8 For Example 6.7 we have $\hat{p} = 227/1000 = 0.227$, and $z_{0.975} = 1.960$, so that an approximate 95% CI for p is given by

$$(0.227 - 1.960 \times 0.0132, 0.227 + 1.960 \times 0.0132) = (0.20, 0.25) \; .$$

6.4 Binomial distribution: two samples

Example 6.9 Two groups of men and women are asked whether they experience nightmares "often" (at least once a month) or "seldom" (less than once a month). The results are given below.

	Men	Women	Total
Often	55	60	115
Seldom	105	132	237
Total	160	192	

The observed proportions of frequent nightmares by men and women are 34.4% and 31.3%. Is this difference statistically significant, or due to chance? To assess this we could make a confidence interval for the difference of the true proportions p_1 and p_2.

The general model is as follows. Let X be the number of "successes" in group 1; $X \sim B(n_1, p_1)$. (p_1 unknown.) Let Y be the number of "successes" in group 2; $Y \sim B(n_2, p_2)$. (p_2 unknown.) Assume X and Y are independent.

We wish to compare the two proportions via a $(1-\alpha)$-confidence interval for $p_1 - p_2$.

The easiest way is to again rely on the CLT. We assume from now on that n_1 and n_2 are sufficiently large ($n_1 p_1$ and $n_1(1-p_1) > 5$, $n_2 p_2$ and $n_2(1-p_2) > 5$), so that the normal approximation the binomial distribution can be applied.

Let $\hat{p}_1 = X/n_1$ and $\hat{p}_2 = Y/n_2$. By the CLT,

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

has approximately a $N(0,1)$ distribution. Hence, with z_α the α-quantile of the $N(0,1)$ distribution (as in Section 6.3), we have

$$P\left(-z_{1-\alpha/2} \leq \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \leq z_{1-\alpha/2}\right) \approx 1 - \alpha\ .$$

Rewriting, this gives

$$\begin{aligned} P\Bigg(&\hat{p}_1 - \hat{p}_2 - z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \leq p_1 - p_2 \\ &\leq \hat{p}_1 - \hat{p}_2 + z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\Bigg) \\ &\approx 1 - \alpha. \end{aligned}$$

As in the one-sample case of Section 6.3, the same is *approximately* true, We now have bounds which only depend on the data.

Hence, an *approximate* $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$ is

$$\boxed{\hat{p}_1 - \hat{p}_2 \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

Example 6.10 We continue Example 6.9. We have $\hat{p}_1 = 55/160$, $\hat{p}_2 = 60/192$ and $z_{0.975} = 1.96$, so that a 95% CI for $p_1 - p_2$ is given by

$$(0.031 - 0.099, 0.031 + 0.099) = (-0.07, 0.13) \; .$$

This interval contains 0, so there is no evidence at the 5% level that men and women are different in their experience of nightmares.

Chapter 7

Hypothesis testing

Hypothesis testing is about making *decisions* about certain hypotheses on the basis of the observed data. In many cases we have to decide whether the observations are due to "chance" or due to an "effect".

Example 7.1 (Diabetic Blood pressure) The mean systolic blood pressure for white males aged 35-44 is 127 and the standard deviation is 7. A paper in "public health" considers a sample of 101 diabetic males and reports a sample mean of 130. Is this good evidence that diabetics differ from the general population?

To assess this, we could ask the question how likely it would be, *if diabetics were similar to the general population*, that a sample of 101 diabetics would have a mean blood pressure this far from 127?

We can actually calculate this, assuming normality of the sample mean, which is plausible fro the Central Limit Theorem. Specifically, let $X_1, \ldots, X_{101}$ be the blood pressure for the 1st, 2nd, etcetera, diabetic person. If diabetics are similar to the general population, then a good model would be $X_1, \ldots, X_{101} \sim N(127, 7^2)$ (and independent). Thus, for the sample mean, we would have $\bar{X} \sim N(127, 49/101)$. And the probability that the sample mean would take a value as extreme or even more extreme than 130 is

$$P(\bar{X} \geq 130) = P\left(\frac{\bar{X} - 127}{\sqrt{49/101}}\right) > \frac{130 - 127}{\sqrt{49/101}} = P(Z > 4.31) = 8.1610^{-6},$$

(where $Z \sim N(0, 1)$). So it is extremely unlikely that the event $\{\bar{X} \geq 130\}$ occurs if the two groups are the same with regard to blood pressure. However, the event *has* occurred. Therefore, there is is *strong* evidence that the blood pressure of diabetics differs from the general public.

Example 7.2 (Tennis serves) In competition last year 60% of a player's first serves went in. The player has received some coaching, and this year 13 out

of 15 first serves have been in. Is this evidence that her first serve rate has improved?

We ask ourselves the same type of question as in the previous example: Suppose the player's serve is in fact unchanged. What is the probability that out of 15 serves 13 or more would have been in. To calculate this, let X be the number of servers in (out of 15). If the player's serve is unchanged, then $X \sim B(15, 0.6)$. The probability of interest is

$$P(X \geq 13) = \sum_{k=13}^{15} \binom{15}{k} 0.6^k \, 0.4^{15-k} \approx 0.0271.$$

This is quite small. Hence, we have *reasonable* evidence that the first serve has improved.

The examples above indicate the main idea behind (classical) hypothesis testing: we wish to show that a "change" or an "effect" has occurred. The way we do this is by showing that an event has occurred which would be highly unlikely if not impossible if that change of effect did *not* take place.

7.1 Mathematical formulation

In this section we investigate how we can we further develop the ideas mentioned in the introduction. Again our starting point is to formulate a *model* for the data which depends on one or more unknown parameters.

Example 7.3 Consider the blood pressure example. Our model could be as follows: Let $X_1, \ldots, X_{101}$ be the blood pressure for the 101 diabetics. We assume that the X_i are independent and all have a $N(\mu, 7^2)$ distribution. Note that we do not say that $\mu = 127$; that is precisely what we wish to test! (However, *if* diabetics are similar to the general population then $X_i \sim N(127, 7^2)$.) Using this model, we have

$$\bar{X} \sim N(\mu, 49/101) \, , \tag{7.1}$$

μ unknown.

Example 7.4 In the tennis serves example, the model is: let X be the number of serves out of 15, we assume

$$X \sim B(15, p),$$

with p unknown. Again, we should not put $p = 0.6$, because that's what we wish to test. Thus our model for the data depends on the unknown parameter p.

7.1.1 Hypotheses

Given a model for the distribution of the data which depends on an unknown parameter (vector) θ, we can formulate our **hypotheses** as statements about the parameter θ. Usually we consider two competing hypotheses, generally denoted by H_0 and H_1. The **null** hypothesis H_0 contains the statement that is tested. The **alternative** hypothesis H_1 contains the statement that we hope or suspect is true, instead of H_0.

Example 7.5 In the blood pressure example, we have the hypotheses:

$$\begin{aligned} H_0 &: \mu = 127, \\ versus & \\ H_1 &: \mu \neq 127\ . \end{aligned}$$

In the tennis serves example, we have the hypotheses:

$$\begin{aligned} H_0 &: p = 0.6, \\ versus & \\ H_1 &: p > 0.6\ . \end{aligned}$$

7.1.2 Test statistic

We wish to test the hypothesis H_0 against the hypothesis H_1 on the basis of (the outcomes of) $X_1, \ldots, X_n$. In other words, our decision whether we accept or reject H_0 depends purely on the data. Moreover, in many cases the decision depends only on a certain (real) *function* of the data. Any such function is called a **test statistic**.

Example 7.6 Suppose in the blood pressure example all individual readings $x_1, \ldots, x_n$ are given. As a model we assume that they come from a random sample from a $N(\mu, 49)$ distribution. As a good test statistic we could use the sample mean $\bar{X}$, Other possible test statistics are $\bar{X} - 127$ or $X_1 + \cdots + X_n$.

Note that, in principle, any real function of $X_1, \ldots, X_n$ can serve as a test statistic, but that only a few of those are useful. Any "decent" test statistic should be such that we can decide (on the basis of its outcome) whether to accept H_0 or not.

We will discuss later *how* to find/choose sensible test statistics.

Note that a test statistic must not depend on any unknown model parameters!

7.1.3 Critical region

Suppose we are given a test statistic $T = T(X_1, \ldots, X_n)$. We wish to make a decision to accept H_0 or not, on the basis of the outcome of T. We will use the following decision rule:

Decision Rule I: Reject H_0 if T falls in the **critical region**.

Here the critical region is any appropriately chosen region in $\mathbb{R}$. In practice a critical region is one of the following:

- **one-sided**
  - to the left: $(-\infty, c]$; we call the test a **left one-sided test**
  - to the right: $[c, \infty)$; we call the test a **right one-sided test**
- **two-sided**: $(-\infty, c_1] \cup [c_2, \infty)$; we call the test a **two-sided test**

Hence, for a critical region which is one-sided to the left, we reject H_0 if the outcome of the test statistic is relatively small.

How should we choose a critical region? Of course the choice of the test statistic and the critical region go hand-in-hand.

Example 7.7 Consider again the blood pressure example. As test statistic we choose $\bar{X}$, because $\bar{X}$ will give us an idea of how large actually μ is. Now, if $\bar{X}$ is around 127, then we see no reason to reject H_0. However, if $\bar{X}$ attains a value which is much larger or much smaller than 127, we would like to reject H_0. This type of reasoning shows us that for our test statistic $\bar{X}$ we should use a *two-sided* critical region $(-\infty, c_1] \cup [c_2, \infty)$. However, we have not decided upon the constants c_1 and c_2 yet. More about this later.

7.1.4 p-value

Another way to decide whether to accept H_0 or not is to look at the p-value. The **p-value** associated with an observed value of the test statistic is the probability that *under* H_0 the (random) test statistic takes a value as extreme or more extreme value than the one observed.

In practice this means the following. Suppose we wish to test $H_0 : \theta = \theta_0$ against some alternative. Let T be a test statistic whose observed outcome is t. Let us write P_{H_0} for the probability measure when we wish to emphasise the fact that H_0 holds true. For a *left one-sided* test the p value is defined as

$$p = P_{H_0}(T \leq t);$$

for a *right one-sided*

$$p = P_{H_0}(T \geq t) \ ;$$

and for a *two-sided* test

$$p = \min\{2P_{H_0}(T \leq t),\ 2P_{H_0}(T \geq t)\}\ .$$

Our second decision rule is now as follows:

Decision Rule II: Reject H_0 if p is smaller than some p_0.

Note that the *smaller* the p-value, the *bigger* the strength of the *evidence* against H_0 provided by the data. As a rule of thumb we will use the following conventions:

$p < 0.10$	suggestive
$p < 0.05$	reasonable evidence
$p < 0.01$	strong evidence

How much evidence is required will of course depend on the purpose of the test. To sentence someone to death requires a lot more evidence than to sentence someone to prison for a week.

Example 7.8 Consider again Example 7.3. If we use $\bar{X} \sim N(\mu, 49/101)$ as our test statistic, then under $H_0 : \mu = 127$ we have

$$\bar{X} \sim N(127, 49/101).$$

Since $H_1 : \mu \neq 127$ is two-sided, the p-value associated with the observed value 130 is

$$2\,P_{H_0}(\bar{X} \geq 130) = 2\,P(Z \geq 4.31) = 1.610^{-5}.$$

(Above, $Z \sim N(0,1)$). Note that this is *twice* the probability that we had in the intuitive/naive approach in Example 7.3. Why this difference? In the naive approach we only calculated the probability of the "extreme" event $\{X \geq 130\}$. This is exactly the p-value corresponding to the right one-sided test of the hypothesis $H_0 : \mu = 127$ against $H_1 : \mu > 127$. In other words, this is the p-value we would use if we wanted to show that diabetics have a *higher* mean blood pressure than the average person. However, if we only want to show that there is a *difference* in mean blood pressure, then our alternative hypothesis should indeed be $H_1 : \mu \neq 127$. In this case the event "the test statistic attains a value which is as extreme or more extreme than observed" is really $\{X \geq 130\} \cup \{X \leq 124\}$ and not just $\{X \geq 130\}$. In any case, we would reject the zero hypothesis.

7.1.5 Type I and Type II Errors

Whether we use Decision Rule I or II to decide between H_0 and H_1, we can make two types of mistakes:

	True state of nature	
Decision	H_0 **is true**	H_1 **is true**
Accept H_0	Correct	Type II Error
Reject H_0	Type I Error	Correct

Ideally we would like construct tests which make both types of errors (let's call then E_I and E_{II}) as small as possible. Unfortunately, this is not possible, because the two errors "compete" with each other. For example, if we make the critical region larger, we decrease E_{II} but at the same time increase E_I. On the other hand, if we make the critical region smaller, we decrease E_I but at the same time increase E_{II}.

Now, in classical statistics the null hypothesis and alternative hypothesis do not play equivalent roles. We are only prepared to reject the zero hypothesis if the observed value of the test statistic is very "unlikely" under the zero hypothesis. Only if this evidence is strong do we wish to reject H_0. In other words, we certainly do not wish to make a large Type I error.

Suppose we wish to test the hypothesis using decision rule I. That is, our test is based on the combination of test statistic and critical region, rather than test statistic and p-value. How should we in this case choose the critical region? Since the error of first kind (= Type I error) is more serious, Neyman and Pearson suggested the following approach:

Neyman-Pearson approach: Choose the critical region such that the Type II error is as small as possible, while keeping the Type I error below a predetermined **significance level** α.

Example 7.9 Consider again Example 7.2. Let X, the number of serves "in" out of 15, be the test statistic. Because we reject H_0 in favour of H_1 for large values of X, the critical region C is one-sided to the right, i.e. of the form

$$C = \{c, \ldots, 15\} .$$

Suppose we wish to keep the error of first kind below $\alpha = 0.05$. It follows that we choose the **critical value** c such that

$$P_{H_0}(X \geq c) \leq 0.05 .$$

For $X \sim B(15, 0.6)$ we have $P(X \geq 13) = 0.0271$ and $P(X \geq 12) = 0.0905$. It follows that we should choose $c = 13$. Note that the *actual* Type I error is 0.0271.

Remark 7.1 Note that Decision Rule I and II are equivalent in the following sense:

Reject H_0 if T falls in the critical region, at significance level α.
Reject H_0 if the p-value is $\leq$ significance level α.

In other words, the p-value of the test is the smallest level of significance that would lead to the rejection of H_0.

Exercise 7.1 Compute the critical region for the blood pressure example, using the test statistic

$$T = \frac{\bar{X} - 127}{\sqrt{7/10}} \ .$$

Take significance level $\alpha = 0.05$.

7.1.6 The 8 steps in a statistical test

The following table summarises the ideas above. It can be used for *any* statistical test. In fact, I urge you to always use the 8 steps outlined below.

1. Formulate an appropriate statistical model for the data.
2. Give the null and alternative hypotheses.
3. Determine the test statistic.
4. Give the distribution of the test statistic under H_0.
5. Calculate the outcome of the test statistic.
6. Calculate the p-value **or** calculate the critical region, given a pre-selected significance level α
7. Accept or reject the null hypothesis.
8. Formulate the conclusion in your own words.

Example 7.10 It is known that a certain gene occurs in 15% of a fish population. An ecologist wishes to determine whether there is any difference in the occurrence of this characteristic within a particular lake. A sample of 100 fish is taken, giving 10 with the gene. Is there any evidence that this population is different from the usual?

Let us go through the eight steps:

1. Let X be the number of fish (out of 100) with the gene. Assume
$$X \sim B(100, p) \ .$$
2. Hypotheses are statements concerning the *parameters* of the model, in this case p. The alternative hypothesis is always about the "strong statement",

in our case the statement that the fish population is *different*, as opposed to the "weak statement" that there is no difference.

This translates into $H_0 : p = 0.15$ and $H_1 : p \neq 0.15$.

3. As test statistic we could choose X. This is a trivial function of the data (X) and moreover, gives us a good idea about p.

4. Under H_0, X has a $B(100, 0.15)$ distribution.

5. The outcome of X is $x = 10$.

6. The p-value is (two-sided test)

$$2\, P_{H_0}(X \leq 10) \; .$$

We could calculate this numerically. Alternatively, we can use the CLT and approximate this with $2\, P(Y \leq 10)$, where $Y \sim N(100 \times 0.15, 100 \times 0.15 \times 0.85)$. Hence, the p value is approximately

$$2\, \Phi\left(\frac{10 - 15}{\sqrt{12.75}}\right) = 2\Phi(-1.40) = 0.162 \; .$$

7. As the p-value is large (> 0.1) we accept H_0.

8. Hence, there is no evidence of a difference in the proportion of fish with the gene.

7.1.7 Power

We have not discussed yet how to *choose* a test statistic. This is somewhat akin to choosing an estimator for an unknown parameter. In fact in many cases the test statistics can be obtained from the estimator(s) of parameter(s). There are *systematic* procedures to construct test statistics, such as the use of the likelihood ratio test; see Supplement 6 on supplementray notes for this chapter.

Given a number of possible test statistics, the question is which of these will constitute a "good" test statistic, or which one will be the best? To answer this question we have to define what "good" and "best" means in this context. The question is somewhat similar to asking which estimator is the best.

One way way to assess whether a test statistic is good is to look at the **power** of the test.

Definition 7.1 Consider a test with test statistic T, alternative hypothesis $H_1 : \theta \in \Omega_1$ and critical region C. The *power* of the test at θ ($\in \Omega_1$) is defined as the probability that H_0 is rejected (as it should). That is,

$$\text{Power}(\theta) = P_\theta(T \in C) \; .$$

In other words, the power is equal to $1-E_{II}$, where E_{II} is the error of the *second* kind (which depends on the alternative θ). The function of θ, Power(θ), with $\theta \in \Omega_1$, is called the **power curve**.

We now have: test statistic T_1 is better than T_2 if the power curve of T_1 lies completely above that of T_2.

Exercise 7.2 For the tennis serve example calculate and plot the power curve of Y, using critical region $\{13, 14, 15\}$.

Next, we will construct tests for a number standard situations. The general rule here will be that we will choose a "common sense" test statistics and corresponding to this an appropriate critical region.

7.2 Normal distribution; one sample

7.2.1 Test for μ: one-sample t-test

Example 7.11 One of the statements in a research article is that the amount of cafeine in regular cola is "19 mg per 6-oz serving". In a second study the cafeine content was determined for a sample of size $n = 40$ of a different brand of cola. The sample mean and standard deviation were 19.57 (mg) and 1.40 (mg). Should we conclude that the expected amount of cofeine in this brand is more than "19 mg per 6-oz serving"?

To answer this question, we again consider an appropriate model for this situation. We represent the observations by $X_1, \ldots, X_n$, and assume that they form a random sample from a $N(\mu, \sigma^2)$ distribution, where μ and σ are *unknown*. The hypotheses can now be formulated as: $H_0 : \mu = 19$ against $H_1 : \mu > 19$. Choose as significance level $\alpha = 0.01$.

Which test statistic should we choose? Since we wish to make a statement about μ the test statistic should reflect this. We could take $\bar{X}$ as our test static and reject H_0 for large values of $\bar{X}$. However, this leads to a complication. It looks like our null hypothesis only contains one parameter value, but in fact it contains *many*, because we should have written

$$H_0 : \mu = 19, \quad 0 < \sigma^2 < \infty \, .$$

It is the unknown σ^2 that leads to the complication in choosing $\bar{X}$ as our test statistic. To see this, consider the following two cases. First consider the case where σ^2 is very small. In that case, $\bar{X}$ is under H_0 very much concentrated around 19, and therefore any deviation from 19, such as 19.57 would be most unlikely under H_0. We would therefore reject H_0. On the other hand, if σ^2 is

very large, then a value of 19.57 could very well be possible under H_0, so we would not reject it.

This shows that $\bar{X}$ is not a good test statistic, but that we should "weigh" it with the standard deviation. That is, we should measure our deviation from 19 in units of σ rather than in units of 1. However, we do not know σ. But this is easily fixed by replacing σ with an appropriate estimator. This leads to the test statistic

$$T = \frac{\bar{X} - 19}{s/\sqrt{40}} .$$

The factor $\sqrt{40}$ is a "normalising" constant which enables us to utilise Theorem 6.1. Namely, under H_0 the random variable T has a $t_{n-1} = t_{39}$ distribution. Note that this is true for *any* value of σ^2. From the table of the student distribution with 39 degrees of freedom we find

$$P_{H_0}(T \geq 2.43) = 0.01$$

The observed outcome of T is

$$\frac{19.57 - 19}{1.40}\sqrt{40} = 2.57 .$$

Since this falls in the critical region $[2.43, \infty)$ we reject H_0. Therefore, there is significant evidence that the average caffeine content is more for the non-regular brand. However, it is important to note that the "statistical significance" of this statement – the difference is not due to chance – is something completely different from its "practical significance" – the difference is important practically. First of all, the true value of μ could very well still be close to 19. To see this, verify that a 99% CI for μ is $(19.03, 20.11)$. Secondly, even if μ would be 20 or even 20.5, it is not at all clear that such a deviation is "practically" important. For example, is it unhealthy to have 1 mg more caffeine per once in your cola? Thus although there is strong evidence of the difference, the difference is so small that it is likely to be irrelevant in practice.

Analogously to the previous example, we obtain the following general formulation for the so-called **one-sample t-test**: Let α be the significance level of the test. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$. Let μ_0 be a given number. We wish to test the hypothesis $H_0 : \mu = \mu_0$ against various alternatives by using the test statistic

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

with $\bar{X} = \frac{1}{n}\sum_{i-1}^{n} X_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Under H_0 we have $T \sim t_{n-1}$. Reject H_0 according to the following table.

H_1	**Reject H_0 if**
$\mu > \mu_0$	$T \geq t_{n-1;1-\alpha}$
$\mu < \mu_0$	$T \leq -t_{n-1;1-\alpha}$
$\mu \neq \mu_0$	$T \leq -t_{n-1;1-\alpha/2}$ or $T \geq t_{n-1;1-\alpha/2}$

Here, as always, $t_{n-1;\alpha}$ denotes the α-quantile of the t_{n-1}-distribution: if $T \sim t_{n-1}$ then $P(T \leq t_{n-1;\alpha}) = \alpha$.

Remark 7.2 Note that the last row of the table above indicates a *two-sided* test. It is simply a combination of a left and a right one-sided test. By choosing for each of these a significance level of $\alpha/2$, we get an overall significance level of α.

Remark 7.3 (σ^2 known: z-test) Suppose in the situation above σ^2 is *known*. In that case we (obviously) do not have to estimate σ^2 via s^2 and can take as our test statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{40}} .$$

Under $H_0 : \mu = \mu_0$ has a $N(0,1)$ distribution (check!). We can summarise the test, called the **z-test**, as follows:

H_1	**Reject H_0 if**
$\mu > \mu_0$	$Z \geq z_{1-\alpha}$
$\mu < \mu_0$	$Z \leq -z_{1-\alpha}$
$\mu \neq \mu_0$	$Z \leq -z_{1-\alpha/2}$ or $Z \geq z_{1-\alpha/2}$

Here z_α is the α-quantile of the standard normal distribution.

7.2.2 Test for σ^2

Example 7.12 A factory produces tuning forks with a frequency which is $N(\mu, 1.1)$ distributed. A new manufacturing process is introduced to improve the "variability" of the process. A batch of 7 forks is tested, giving a sample variance of 0.488. Is the new process better?

To answer this, let's go through the 8 steps of a statistical test: First the model: Let X_i be the frequency of the ith fork, $i = 1, \ldots, 7$. We assume

$$X_1, \ldots, X_7 \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2),$$

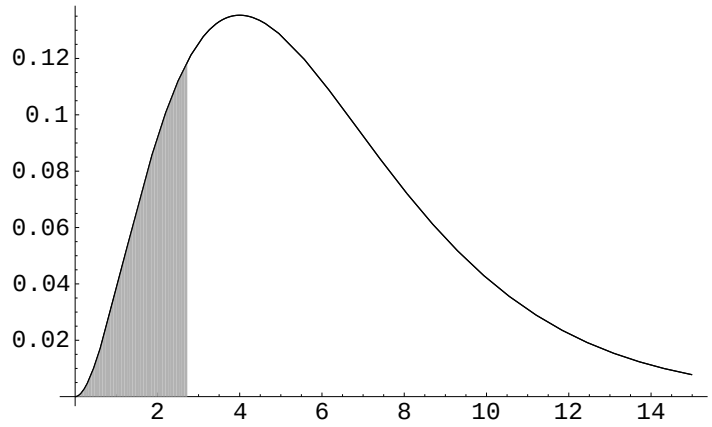
for some unknown μ and σ^2. The hypotheses are

$$\begin{aligned} H_0 &: \sigma^2 = 1.1 \\ H_1 &: \sigma^2 < 1.1 \end{aligned}$$

A sensible test statistic would be based on the sample variance s^2. Theorem 6.2 suggests that we take as test statistic

$$\frac{6s^2}{1.1} .$$

because under H_0 we know its distribution. Namely, $\frac{6s^2}{1.1} \sim \chi^2_6$. The outcome of $\frac{6s^2}{1.1}$ is 2.66, and the corresponding p-value is $P_{H_0}(6s^2/1.1 \leq 2.66) \approx 0.15$ (see figure and the table for the χ^2_6-distribution), which is quite large, hence we accept H_0. Hence, there is not enough evidence to claim that the new process is better.



Alternatively, suppose we choose a significance level $\alpha = 0.05$. The critical region is one-sided to the left: $(0, c]$, where c is such that for $V \sim \chi^2_6$,

$$P(V \leq c) = 0.05 \, .$$

From the table, it follows that $c = 1.635$. Since the outcome of the test statistic does not lie in the critical region, we accept H_0. Again, there is no evidence that the new process is more accurate.

In general we can summarise the test as follows:

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$. Let σ_0^2 be a given number. We wish to test the hypothesis $H_0 : \sigma^2 = \sigma_0^2$ against various alternatives by using the test statistic:

$$\frac{(n-1)s^2}{\sigma_0^2} \, ,$$

where $s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ and $\bar{X} = \frac{1}{n} \sum_{i-1}^n X_i$. Under H_0 we have $s^2(n-1)/\sigma_0^2 \sim \chi^2_{n-1}$. Reject H_0 according to the following table.

H_1	**Reject H_0 if**
$\sigma^2 > \sigma_0^2$	$\frac{(n-1)s^2}{\sigma_0^2} \geq \chi^2_{n-1;1-\alpha}$
$\sigma^2 < \sigma_0^2$	$\frac{(n-1)s^2}{\sigma_0^2} \leq \chi^2_{n-1;\alpha}$
$\sigma^2 \neq \sigma_0^2$	$\frac{(n-1)s^2}{\sigma_0^2} \geq \chi^2_{n-1;1-\alpha/2}$ or $\frac{(n-1)s^2}{\sigma_0^2} \leq \chi^2_{n-1;\alpha/2}$

Here, as before, $\chi^2_{n-1;\alpha}$ is the α-quantile of the χ^2_{n-1} distribution; it is the number at which the cdf of the χ^2_{n-1} distribution attains the value α.

7.3 Normal distribution: two samples

Example 7.13 A human movement student has a theory that the mean weight of 3rd year students is greater than that of 1st years. To test this theory, random samples are taken from each of the two groups. A sample of 10 3rd years has a mean of 71.5kg and a standard deviation of 12kg, while the sample of 15 1st years has a mean of 62.0kg and a standard deviation of 15kg. Does this show that 3rd year students are indeed (on average) heavier? Another question that we could ask is: is the *variance* in weight within the two groups the same?

To answer these questions via statistics, we first need to make a model for the data. A standard model is as follows:

Let X_{1j} be the jth 1st year student, $j = 1, \ldots, n_1$. Let X_{2j} be the jth reading of 3rd year students, $j = 1, \ldots, n_2$. (Here $n_1 = 15$ and $n_2 = 10$). We assume that

- $X_{11}, \ldots, X_{1n_1} \overset{\text{i.i.d.}}{\sim} N(\mu_1, \sigma_1^2)$.
- $X_{21}, \ldots, X_{2n_2} \overset{\text{i.i.d.}}{\sim} N(\mu_2, \sigma_2^2)$.
- $X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{1n_2}$ are *independent*,

where μ_1, μ_2, σ_1^2 and σ_2^2 are unknown parameters.

We can now test two sets of hypotheses. The first one is about equality of the *means* of the two samples; specifically we wish to test $H_0 : \mu_1 = \mu_2$, versus $H_1 : \mu_1 < \mu_2$. The second one is about equality of the *variances* of the two samples; here we wish to test $H_0 : \sigma_1^2 = \sigma_2^2$, versus $H_1 : \sigma_1^2 \neq \sigma_2^2$.

7.3.1 Test for $\mu_1 - \mu_2$: two-sample t-test

Example 7.14 Consider Example 7.13. Suppose that the variances of the two samples are the *same*: $\sigma_1^2 = \sigma_2^2 = \sigma^2$, for some unknown σ^2. In analogy with the one-sample t-test and in view of Theorem 6.3 a good test statistic is

$$T = \frac{\bar{1} - \bar{2}}{s_p \sqrt{1/n_1 + 1/n_2}} ,$$

where s_p is the pooled sample variance in (6.3). We reject the null hypothesis for small values of T. By Theorem 6.3 if $\mu_1 = \mu_2$, then T has a t_{23} distribution (the number of degrees of freedom is $n_1 + n_2 - 2 = 15 + 10 - 2 = 23$). This

enables us to find the exact critical region or the p-value. In particular, from our sample we have the following outcome of s_p^2:

$$s_p^2 = \frac{14 \times (15^2) + 9 \times (12^2)}{23} = 193.304$$

So we have an observed value of T of

$$t = \frac{62.0 - 71.5}{\sqrt{193.304\left(\frac{1}{15} + \frac{1}{10}\right)}} = -1.674\ .$$

From the tables of the student distribution we have $t_{23;0.9} = 1.319$ and $t_{23;0.95} = 1.7139$. The p-value for the observation -1.674 is

$$P_{H_0}(T \leq -1.674)\ .$$

It follows by the symmetry of the student distribution (make a picture) that the p value lies between 0.05 and 0.1. Hence, there is a very small amount of evidence to support the student's theory that the 3rd year students are on average heavier.

In general we have the following situation. Consider the model given by the dotpoints in Example 7.13. But also assume $\sigma_1^2 = \sigma_2^2 = \sigma^2$. We wish to test the hypothesis $H_0 : \mu_1 - \mu_2 = \delta_0$ against various alternatives. The **two-sample t-test** is described as follows. As test statistic we take

$$T = \frac{\bar{X}_1 - \bar{X}_2 - \delta_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\ ,$$

where s_p^2 is the pooled sample variance:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{\sum_{j=1}^{n_1}(X_{1j} - \bar{X}_1)^2 + \sum_{j=1}^{n_2}(X_{2j} - \bar{X}_2)^2}{n_1 + n_2 - 2},$$

$\bar{X}_1 = {n_1}^{-1}\sum_{j=1}^{n_1} X_{1j}$ and $\bar{X}_2 = {n_2}^{-1}\sum_{j=1}^{n_2} X_{2j}$.

Under H_0 we have $T \sim t_{n_1+n_2-2}$. Reject H_0 according to the following table:

H_1	**Reject H_0 if**
$\mu_1 - \mu_2 > \delta_0$	$T \geq t_{n_1+n_2-2;1-\alpha}$
$\mu_1 - \mu_2 < \delta_0$	$T \leq -t_{n_1+n_2-2;1-\alpha}$
$\mu_1 - \mu_2 \neq \delta_0$	$T \leq -t_{n_1+n_2-2;1-\alpha/2}$ or $T \geq t_{n_1+n_2-2;1-\alpha/2}$

Remark 7.4 Note that if σ^2 is *known*, then we can use instead the test statistic

$$\frac{\bar{1} - \bar{2} - \delta_0}{\sigma\sqrt{1/n_1 + 1/n_2}} ,$$

which under H_0 has a $N(0,1)$-distribution. Check this yourself.

7.3.2 Test for σ_1^2/σ_2^2; F-test

For the two-sample t test it is necessary that the variances are equal. Can we test for this?

We wish to test $H_0 : \sigma_1^2 = \sigma_2^2$ against various alternatives (in particular $H_0 : \sigma_1^2 \neq \sigma_2^2$). An "obvious" choice for the test statistic is

$$F = \frac{s_1^2}{s_2^2} .$$

Theorem 2.2 tells us that under H_0, $F \sim F_{n_1-1,n_2-1}$. Then we can summarise the test (the so-called F-test) as follows:

H_1	**Reject H_0 if**
$\sigma_1^2 > \sigma_2^2$	$F \geq F_{n_1-1,n_2-1;1-\alpha}$
$\sigma_1^2 < \sigma_2^2$	$F \leq F_{n_1-1,n_2-1;\alpha} = 1/F_{n_1-1;n_2-1;1-\alpha}$
$\sigma_1^2 \neq \sigma_2^2$	$F \geq F_{n_1-1,n_2-1;1-\alpha/2}$ or $F \leq F_{n_1-1,n_2-1;\alpha/2} = 1/F_{n_1-1,n_2-1;1-\alpha/2}$

Here, as before $F_{n_1-1,n_2-1;\alpha}$ is the α-quantile of the F_{n_1-1,n_2-1}-distribution.

Example 7.15 Consider Example 7.13. The outcome of F is $\frac{15^2}{12^2} = 1.56$. Under H_0 F has an $F_{14,9}$-distribution. Suppose we test at a significance level of $\alpha = 0.05$. From the table of the F-distribution we find $F_{14,9;0.975} = 3.798$ and $F_{14,9;0.025} = 1/F_{9,14;0.975} = 1/3.209 = 0.312$. Hence, the critical region for the test is is $[0, 0.312] \cup [3.798, \infty)$. Since, the outcome of F does not fall in the critical region, we do accept the null hypothesis. So our assumption of equal variances in Example 7.14 seems justified.

7.4 Paired samples

Consider again Examples 7.13 and 7.14 Did this experiment show that a student's weight increases between 1st and 3rd year?

A better way to set up an experiment to test such a theory would be to weigh the *same students* in 1st year and again in 3rd year.

Suppose such an experiment weighed 10 students and gave the following results:

Person	1	2	3	4	5	6	7	8	9	10
1st year:	71.6	69.7	50.0	56.9	60.9	77.5	53.9	49.6	32.3	61.3
3rd year:	83.4	77.7	57.9	67.0	69.2	87.4	63.7	59.0	39.7	71.1

Can we analyse this data in the same way as before, that is by using a two-sample t-test? The answer is *no*, since the samples are no longer independent! However, we can proceed as follows:

Let $D_i = X_{2i} - X_{1i}$ be the difference between the 3rd year and 1st year weight for person i, $i = 1, \ldots, 10$. Now suppose, as a model we assume that $D_1, \ldots, D_{10}$ are iid $N(\mu, \sigma^2)$ distributed for some unknown μ and σ^2. To assess whether 3rd year students are on average heavier we need to test the hypothesis $H_0 : \mu = 0$ against $H_1 : \mu > 0$. But this we can simply do via the one-sample t-test!

Thus, although conceptually the problem above seems a two-sample problem, it is in fact a one-sample problem. Beware of paired observations!

For our problem, the outcomes of $D_1, \ldots, D_{10}$ are given in the table below.

Person	1	2	3	4	5	6	7	8	9	10
Difference	11.8	8.0	7.9	10.1	8.3	9.9	9.8	9.4	7.6	9.8

This gives on observed test statistic of 22.47, which has an extremely small p-value. So there is overwhelming evidence that a student's weight increases between 1st and 3rd year.

7.5 Binomial test; one sample

Example 7.16 In a certain market research study we wish to investigate whether people would prefer a new type of sweetner in a certain brand of yoghurt. Ten people were given two packets of yoghurt, one with the old sweetner and one with the new sweetner. Eight of the ten people preferred the yoghurt with the new sweetner and two preferred the old yoghurt. Is there enough evidence that the new style of yoghurt is preferred?

First we formulate the model. Let $X_1, \ldots, X_{10}$ be such that

$$X_i = \begin{cases} 1 & \text{if person } i \text{ prefers the new yoghurt,} \\ 0 & \text{if person } i \text{ prefers the old yoghurt,} \end{cases}$$

$i = 1, \ldots, 10$. We assume that $X_1, \ldots, X_{10}$ are independent and that for all i, $P(X_i = 1) = p = 1 - P(X_i = 0)$, for some unknown p (between 0 and 1). We wish to test

$$H_0 : p = 0.5 \quad \text{against} \quad H_1 : p > 0.5 \, .$$

Suppose we wish to test at significance level $\alpha = 0.05$. As test statistic we could use $X = \sum_{i=1}^{10} X_i$, and we would reject H_0 for large values of X. Under H_0 the test statistic has a $B(10, 1/2)$ distribution. We wish to choose the critical region of the form $\{c, \ldots, 10\}$ as large as possible, such that the error of the first kind remains below α. Since, under H_0, $P(X \geq 8) = 0.055$ and $P(X \geq 8) = 0.01 < 0.05$ we take our critical region as $\{9, 10\}$. Since the outcome of X is 8, which does not belong to the critical region, we do not reject H_0 at the 0.05 level of significance. But, note that the p-value is quite small (0.055), and thus there is some doubt about H_0.

Remark 7.5 Our model above is in a sense over-specific. We assume that we observe the preference X_i for each individual. But in fact, we only observe the total number of preferences $X = X_1 + \cdots + X_n$ for the new yoghurt. An alternative and simpler model would suffice here, namely: let X be the total number of preferences for the new type of yoghurt, we assume $X \sim B(n, p)$, for some unknown p. The test now proceeds in exactly the same way as before.

Remark 7.6 The example above shows again that it is better to report the finding of a test using p-values than using a preselected significance level. The choice of the significance level is somewhat arbitrary.

We now describe the general situation for the **one-sample binomial test**. Suppose that $X_1, \ldots, X_n$ are the results of n independent Bernoulli trials with success parameter p. That is the X_i's are independent and

$$P(X_i = 1) = p = 1 - P(X_i = 0) \; .$$

Then, $X = X_1 + \cdots + X_n \sim B(n, p)$. We wish to test $H_0 : p = p_0$ against some alternative.

As test statistic we can use X, which under H_0 has a $B(n, p_0)$ distribution. We reject H_0 at the α level of significance according to the following table.

H_1	**Reject H_0 if**	
$p > p_0$	$X \geq c$	where c is the smallest integer such that $P_{H_0}(X \geq c) \leq \alpha$,
$p < p_0$	$X \leq c$	where c is the largest integer such that $P_{H_0}(X \leq c) \leq \alpha$,
$p \neq p_0$	$X \leq c_1$ or $X \geq c_2$	where c_1 is the largest integer such that $P_{H_0}(X \leq c_1) \leq \alpha/2$ and c_2 is the smallest integer such that $P_{H_0}(X \geq c_2) \leq \alpha/2$

To determine the critical region we can use the exact probabilities of the bino-

mial distribution, or we can for large n use the Central Limit Theorem. In fact, for large n it is convenient to consider the test statistic

$$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} .$$

which under H_0 has approximately a $N(0,1)$-distribution. For large n the approximate test can be summarised as follows:

H_1	**Reject H_0 if**
$p > p_0$	$Z \geq z_{1-\alpha}$
$p < p_0$	$Z \leq -z_{1-\alpha}$
$p \neq p_0$	$Z \leq -z_{1-\alpha/2}$ or $Z \geq z_{1-\alpha/2}$

Here z_α is the α-quantile of the standard normal distribution.

7.6 Binomial distribution; two samples

Example 7.17 A politician believes that audience members of the ABC news are in general more left wing than audience members of a commercial news broadcast. A poll of two party preferences is taken. Of seventy ABC viewers, 40 claim left wing allegiance, while of 100 commercial station viewers, 50 claim left wing allegiance. Is there any evidence to support the politician's claim?

Our model is as follows. Let X_1 be the number of left-wing ABC viewers out of $n_1 = 70$, and let X_2 be the number of left-wing "commercial" viewers out of $n_2 = 100$. We assume that X_1 and X_2 are independent, with $X_1 \sim B(n_1, p_1)$ and $X_2 \sim B(n_2, p_2)$, for some unknown p_1 and p_2. We wish to test $H_0 : p_1 = p_2$ against $p_1 > p_2$.

Since m and n are fairly large here, we proceed by using the CLT, analogously to Sections 6.3 and 6.4. Let $\hat{p}_1 = X_1/n_1$ and $\hat{p}_2 = X_2/n_2$ be the empirical proportions. By the CLT $\hat{p}_1$ has approximately a $N(p_1, p_1(1-p_1)/n_1)$ distribution, and $\hat{p}_2$ has approximately a $N(p_2, p_2(1-p_2)/n_2)$ distribution. It follows that

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

has approximately a $N(0,1)$ distribution. Now, under H_0, $p_1 = p_2 = p$, say, and hence under H_0

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}}$$

has approximately a $N(0,1)$ distribution. As we don't know what p is, we need to estimate it. If H_0 is true, then $X_1 + X_2 \sim B(n_1 + n_2, p)$, and thus the MLE of p is given by the *pooled* success proportion

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} \,. \tag{7.2}$$

Concluding, we take as our test statistic:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \tag{7.3}$$

which under H_0 has approximately a $N(0,1)$ distribution.

Our general formulation for the **two-sample binomial test** (also called the **test for proportions**) is as follows. Let X be the number of "successes" in group 1; $X_1 \sim B(n_1, p_1)$. (p_1 unknown.) Let X_2 be the number of "successes" in group 2; $X_2 \sim B(n_2, p_2)$. (p_2 unknown.) Assume X_1 and X_2 are independent. We wish to test $H_0 : p_1 = p_2$ against various alternatives. As test statistic we use Z given in (7.3).

We reject H_0 according to the following table

H_1	**Reject H_0 if**
$p_1 > p_2$	$Z \geq z_{1-\alpha}$
$p_1 < p_2$	$Z \leq -z_{1-\alpha}$
$p_1 \neq p_2$	$Z \geq z_{1-\alpha/2}$ or $z \leq -z_{1-\alpha/2}$

Again, z_α is the α-quantile of the $N(0,1)$ distribution.

Example 7.18 Returning to Example 7.17, from our data we have the estimates $\hat{p_1} = \frac{40}{100}$, $\hat{p_2} = \frac{50}{100}$, and

$$\hat{p} = \frac{40 + 50}{70 + 100} = \frac{90}{170} \,.$$

Thus, the outcome of the test statistic is

$$\frac{\frac{40}{70} - \frac{50}{100}}{\sqrt{\frac{90}{170} \times \frac{80}{170}\left(\frac{1}{70} + \frac{1}{100}\right)}} = 0.9183 \,.$$

This gives a p-value of 0.18, so there is no evidence to support the politician's claim.

7.7 Sign test

We conclude this chapter with an example of a **non-parametric** test. Throughout these notes we have been dealing with models for the data in which distribution is known in advance, except for a number of unknown parameters. For non-parametric test no specific form for the underlying distribution of the data is assumed.

Example 7.19 To test if two examiners are equally strict on their students, the head of the department gives them 12 exams, which are graded (on a scale from 0 to 10) by both examiners. The results are listed below.

Exam	1	2	3	4	5	6
Grade 1	3.5	5	8	6	7	6
Grade 2	4	4.5	9	6.5	6.5	5.5

Exam	7	8	9	10	11	12
Grade 1	5	7	5.5	6.5	7	8
Grade 2	5.5	8	6.5	6	8	9

Do these figure suggest that one of the lecturers is stricter than the other?

If the data were normally distributed, this would call for a paired t-test. (Why not a two-sample t-test?) But, we do not need to assume normality. A simpler, non-parametric model is the following.

Model: The paired data $(x_1, y_1), \ldots, (x_n, y_n)$ are the outcomes of i.i.d. vectors $(X_1, Y_1), \ldots, (X_n, Y_n)$. (Here $n = 12$.)

What we're interested in is whether $p = P(Y_i > X_i)$ is equal to $1/2$ or not. We may easily test this, i.e., $H_0 : p = 1/2$ against $H_1 : p \neq 1/2$, by looking at the number of observations for which the y-value is larger than the x-value. That is, we simply use a one-sample binomial test, where the test statistic N is the total number of indices i for which $Y_i > X_i$.

Thus, in general $N \sim B(n, p)$, and in particular under H_0, $N \sim B(n, 1/2)$.

Example 7.20 (Examiners, continued) For the examiners, the outcome of N is 8. The p-value for the binomial test is

$$\min\{2P_{H_0}(N \geq 8),\ 2P_{H_0}(N \leq 8)\},$$

Hence the p-value is $2 \times 0.19 = 0.38$ and we do not reject H_0. Question: what would the p-value be when $H_1 : p > 1/2$?

Chapter 8

Chi-Squared Goodness-of-Fit Tests

In this chapter we discuss a class of statistical tests that are known under the name *goodness of fit* (GoF) tests. Although there are various types of goodness of fit tests, the ones we will consider here all rely, in some way or the other, on the properties of the *multinomial distribution*; and the corresponding test statistics have asymptotically a χ^2-distribution.

Recall the definition of multinomial distribution. We write $(X_1, \ldots, X_k) \sim \text{Mult}(n, p_1, \ldots, p_k)$ to denote that the random variables $X_1, \ldots, X_k$ have a **multinomial** distribution, i.e.,

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1!\, x_2! \cdots x_k!}\, p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \,,$$

for all $x_1, \ldots, x_k \in \{0, 1, \ldots, n\}$ for which $x_1 + x_2 + \cdots + x_k = n$.

The multinomial distribution appears in probability and statistics if we categorise data into different categories.

Example 8.1 Suppose the IQ of army recruits is $N(100, 16^2)$ distributed. Army recruits are classified as

$$\begin{array}{lcl} \text{Class 1} & : & \text{IQ} \leq 90 \\ \text{Class 2} & : & 90 < \text{IQ} \leq 110 \\ \text{Class 3} & : & \text{IQ} > 110 \end{array}$$

The *proportion* p_1, p_2 and p_3 of army recruits in the three categories are given by $P(Y \leq 90) = p_1$, $P(90 < Y \leq 110) = p_2$ and $P(Y > 110) = p_3$, where $Y \sim N(100, 16^2)$. It follows that we have the following proportions:

$$\begin{array}{lcl} \text{Class 1} & : & p_1 = 0.266 \\ \text{Class 2} & : & p_2 = 0.468 \\ \text{Class 3} & : & p_3 = 0.266 \end{array}$$

Now suppose we have 7 new recruits. What is the probability that of these 7 new recruits, two are Class 1; four are Class 2 and one is Class 3?

To answer this, let X_i be the number in class $i, i = 1, 2, 3$. Then, $(X_1, X_2, X_3) \sim \text{Mult}(7, p_1, p_2, p_3)$. Thus, it follows immediately that

$$P(X_1 = 2, X_2 = 4, X_3 = 1) = \frac{7!}{2!\,4!\,1!}\, p_1^2\, p_2^4\, p_3^1 \approx 0.0957.$$

A good way to think of random variables $X_1, \ldots, X_k$ with a $\text{Mult}(n, p_1, \ldots, p_k)$ distribution is to view X_i as the total number of balls in the ith urn, if we throw randomly n balls into urns $1, \ldots, k$ with probability $p_1, \ldots, p_k$.

For GoF tests the following theorem is of utmost importance. A proof can be found in Appendix **??**. The idea of the proof is to write the vector $\boldsymbol{Z} = (Z_1, \ldots, Z_k)^T$, with $Z_i = (X_i - n\,p_i)/(n\,p_i)$ as a sum of i.i.d. vectors, and then to apply the (multivariate) Central Limit Theorem.

Theorem 8.1 Let $(X_1, \ldots, X_k) \sim \text{Mult}(n, p_1, \ldots, p_k)$, then the random variable

$$\sum_{i=1}^{k} \frac{(X_i - n\,p_i)^2}{n\,p_i}$$

has approximately a χ^2_{k-1} distribution, for large n.

Remark 8.1 As a rule of thumb, we can use the approximation above provided that

$$np_i \geq 5, \qquad \text{for all } i.$$

8.1 GoF test with known parameters

Example 8.2 The phenomenon of *complete dominance* predicts that progeny whose genetic component is (F,F) will be extremely frizzled, progeny with (F,f) slightly frizzled and (f,f) will be normal. According to the theory (genetics) the proportions FF : Ff : ff should be 1 : 2 : 1. Out of 93 randomly selected chickens the observed frequencies phenotypes are 23 (extremely frizzled), 50 (slightly frizzled) and 20 (normal). Is this in accordance with the theory?

We can test this with a **goodness-of-fit**-test:

First of all, our model is: Let X_1, X_2, X_3 be the total number of FF, Ff and ff chickens out of 93. We have $(X_1, X_2, X_3) \sim \text{Mult}(93, p_1, p_2, p_3)$.

We want to test: $H_0 : p_1 = \frac{1}{4}, p_2 = \frac{1}{2}, p_3 = \frac{1}{4}$ against the alternative hypothesis that H_0 is not true.

As test statistic we use

$$T := \frac{(X_1 - 93/4)^2}{93/4} + \frac{(X_2 - 93/2)^2}{93/2} + \frac{(X_3 - 93/4)^2}{93/4} .$$

Under H_0 this has approximately a χ^2_2 distribution, see Theorem 8.1. We reject H_0 at significance level $\alpha = 0.05$ if the outcome of T is larger than $\chi^2_{2;0.95} = 5.991$.

The outcome of T is 0.71, which does not fall in the critical region, so we accept H_0. Thus, there is no evidence to reject the theory.

We now describe the general situation.

Suppose $(X_1, \ldots, X_k) \sim \text{Mult}(n, p_1, \ldots, p_k)$, we can test $H_0 : p_1 = \pi_1, \ldots, p_k = \pi_k$ against the alternative hypothesis that H_0 is not true by using the test statistic

$$T = \sum_{i=1}^{k} \frac{(X_i - n\,\pi_i)^2}{n\,\pi_i},$$

which under H_0 has a χ^2_{k-1} distribution. We reject H_0 at the α level of significance if

$$T \geq \chi^2_{k-1;1-\alpha},$$

where $\chi^2_{k-1;1-\alpha}$ is the $(1-\alpha)$-quantile of the χ^2_{k-1} distribution.

Remark 8.2 We can symbolically write the test statistic as

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

where O_i is the *observed* number of observations in class i and E_i is the *expected* number of observations in class i. This form for the test statistic is found in any goodness of fit test.

8.2 GoF test with unknown parameters

Now consider the case where

$$(X_1, \ldots, X_k) \sim \text{Mult}(n, p_1, \ldots, p_k),$$

but now the $p_i = p_i(\theta)$ depend on an *unknown* parameter vector $\theta = (\theta_1, \ldots, \theta_r)$.

Let $\hat{\theta}$ be the MLE of θ and $\hat{p}_i = p_i(\hat{\theta})$ the MLE of $p_i(\theta)$. Similar to Theorem 8.1 we have:

Theorem 8.2 Let $(X_1, \ldots, X_k) \sim \text{Mult}(n, p_1, \ldots, p_k)$, where the $p_i = p_i(\theta)$ depend on an unknown parameter vector $\theta = (\theta_1, \ldots, \theta_r)$. Let $\hat{\theta}$ be the MLE of θ and $\hat{p}_i = p_i(\hat{\theta})$ the MLE of $p_i(\theta)$. Then, the random variable

$$\sum_{i=1}^{k} \frac{(X_i - n\,\hat{p}_i)^2}{n\,\hat{p}_i}$$

has approximately a χ^2_{k-1-r} distribution, for large n. ($n\hat{p}_i \geq 5, \forall i$).

The proof of this theorem is much more difficult than that of Theorem 8.1. See for example: P.K. Sen & J.M. Singer (1993): Large sample methods in statistics.

Remark 8.3 Comparing this with Theorem 8.1 we see that we apparently "loose" r degrees of freedom if we have to estimate r parameters.

Example 8.3 (Village housing) Events which occur randomly in time or space often follow a Poisson distribution. The position of 911 houses in a village west of Tokyo was analysed to determine whether or not the positioning was random. A grid was superimposed over the 3×4km village dividing its 12 square kilometres into 1200 plots and the number of houses in each plot was counted.

The results are given in the first and second column of Table 8.1. Test if the distribution of the number houses per plot has a Poisson distribution.

Our model is as follows: Let Y_i be the number of plots with i houses on them, $i = 0, 1, \ldots, 1200$. We wish to test that the probability p_i that there are i houses on a plot is given by

$$p_i = \mathrm{e}^{-\mu}\,\frac{\mu^i}{i!}, \quad i = 0, 1, \ldots,$$

for some (unknown) μ. The MLE for μ is the average number of houses per plot, i.e.,

$$\hat{\mu} = \frac{1}{1200}\sum_{i=0}^{1200} i\,y_i = \frac{911}{1200} = 0.7592\ .$$

Hence we have

$$\hat{p}_i = \mathrm{e}^{-\hat{\mu}}\,\frac{\hat{\mu}^i}{i!},$$

and the (ML) estimate of the *expected* number of plots with i houses on them is

$$E_i = 1200\,\hat{p}_i \quad i = 0, \ldots, 1200.$$

Table 8.1 gives the observed (y_i) and the expected (E_i) number of plots with i houses on them. We wish to use a test statistic of the form in Theorem 8.2, but in order to do that, we must determine the "classes" such that the expected number of observations is 5 or more. To achieve this we we choose the classes

$\{0\}, \{1\}, \{2\}, \{3\}, \{4, 5, \ldots\}$, and let $X_0, \ldots, X_4$ be the number of occurrences. For the X_i we have

$$(X_1, \ldots, X_5) \sim \text{Mult}(1200, p_0, \ldots, p_3, 1 - p_0 - \ldots - p_3) \; .$$

The actual hypothesis that we wish to test is now H_0: $p_0 = \mathrm{e}^{-\mu}, \ldots, p_3 = \mathrm{e}^{-\mu}\mu^3/3!$ and $p_4 = 1 - p_0 - p_1 - p_2 - p_3$, for some μ.

No. Houses per plot	Observed: y_i	Expected: E_i
0	584	561.65
1	398	426.40
2	168	161.86
3	35	40.962
4	9	7.77
5	4	1.18
6	0	0.15
7	1	0.02
8	0	0.00
9	1	0.00

Table 8.1: Distribution of number of houses per plot

As our test statistic we choose

$$T = \sum_{i=0}^{4} \frac{(X_i - E_i)^2}{E_i} \; ,$$

which under H_0 has approximately a χ^2_3 distribution (because $k - 1 - r = 5 - 1 - 1 = 3$). The outcome of T is

$$\frac{(584 - 561.65)^2}{561.65} + \cdots + \frac{(15 - 9.12)^2}{9.12} \approx 7.67 \; .$$

From the tables we see that the p-value is close to 5% (for example, $\chi^2_{3,0.95} = 7.815$), thus there seems to be slight evidence that the distribution of houses is *not* Poisson. However, we would fail to reject H_0 at the 5% level.

8.3 Contingency tables

Contingency tables are used to test the **independence** of data.

Example 8.4 (ESP belief) We wish to examine whether artists differ from non-artists in Extra-Sensory Perception (ESP) belief. Table 8.2 lists the amount of belief in ESP for a group of 114 Artists and a group of 344 Non-artists. We wish to investigate whether being an artist or not is "independent" of the ESP belief (strong, moderate or not).

	ESP belief			
	Strong	Moderate	Not	total
Artists	67	41	6	114
Non-artists	129	183	32	344
	196	224	38	458

Table 8.2: ESP belief

To see that this is a type of goodness of fit situation, we need to properly formulate a model for the data and express the null and alternative hypotheses in terms of the parameters in the model.

If we ignore the row and column totals, we have a table with $r = 2$ rows and $c = 3$ columns. We can imagine the table to be filled in the following way: We randomly select 458 people and ask whether they are an artist or not and what their ESP belief is. Let (U_k, V_k) denote the response for the kth selected person, where $U_k \in \{1, 2\}$, where (1 = artist, 2=non-artist), and $V_k \in \{1, 2, 3\}$, where, (1 = strong belief , 2 = medium belief, 3= no belief). We assume that $(U_1, V_1), \ldots, (U_n, V_n)$ are independent and distributed as a random vector (U, V) that can take values $(1, 1), (1, 2), (1, 3), (2, 1), (2, 2)$ and $(2, 3)$ with probabilities $p_{(1,1)}, p_{(1,2)}, \ldots, p_{(2,3)}$.

Now, instead of recording all (U_k, V_k), we could instead *count* how many people are artist with a strong ESP belief, artist with a Moderate ESP belief, etc. Let $X_{(i,j)}$ be the *count* in row i and column j. That is, the total number of observations out of $n = 458$ that fall in "cell" (i, j). For example, the outcome of $X_{(2,2)}$ is 183. From the model above we have

$$(X_{(1,1)}, \ldots, X_{(2,3)}) \sim \text{Mult}(n, p_{(1,1)}, \ldots, p_{(2,3)}) \, .$$

We wish to test whether null hypothesis that the random variables U and V are *independent*.

> **Notation:** For notational simplicity let us write from now on X_{ij} and p_{ij} for $X_{(i,j)}$ and $p_{(i,j)}$

In terms of the parameters of the model the null hypothesis can be written as

$$H_0 : \quad p_{ij} = p_i \, q_j, \qquad \text{for all } i, j,$$

where p_1, p_2, q_1, q_2 and q_3 are unknown probabilities. Using Theorem 8.2, we can test the null hypothesis against the alternative hypothesis that $p_{ij} \neq p_i \, q_j$ for some i and j, by using the test statistic

$$T = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(X_{ij} - E_{ij})^2}{E_{ij}},$$

where E_{ij} is the MLE of EX_{ij}, the expected number of observations in cell (i, j). Now, $EX_{ij} = np_{ij}$, and under H_0, $EX_{ij} = np_i\, q_j$. The MLEs for p_i and q_j are

$$\hat{p_i} = \frac{\sum_{j=1}^{3} X_{ij}}{458} \quad \text{and} \quad \hat{q_j} = \frac{\sum_{i=1}^{2} X_{ij}}{458};$$

and hence the MLE for np_iq_j is $n\hat{p}_i\hat{q}_j$.

By Theorem 8.2 the test statistic T has under H_0 approximately a χ^2 distribution with 2 degrees of freedom. To see this note that the total number of parameters to be estimated is $1+2=3$. Using Remark 8.3 we have to subtract this from the total number of classes (cells) minus 1, i.e., $6-1=5$. Thus, the number of degrees of freedom is $5-3=2$.

We reject the null hypothesis for large values of T. The various outcomes and estimates are given in the table below.

i,j	x_{ij}	$n\hat{p}_i\hat{q}_i$	$\frac{(x_{ij}-n\hat{p}_i\hat{q}_i)^2}{n\hat{p}_i\hat{q}_i}$
1,1	67	48.8	6.79
1,2	41	55.8	3.93
1,3	6	9.46	1.27
2,1	129	147	2.20
2,2	183	168	1.34
2,3	32	28.5	0.43

(e.g. $n\hat{p}_1\hat{q}_1 = 114 \times 196/458 \approx 48.8$.)

It follows that the outcome of T is $t = 6.79+3.93+1.27+2.20+1.34+0.43 = 15.96$. The p-value is 0.00034. Hence, we strongly reject H_0. Artists indeed seem to differ from non-artists in ESP belief.

For the *general* case consider a random vector (U, V) taking values in $\{1,\ldots,r\}\times\{1,\ldots,c\}$. We wish to test for *independence* of U and V.

We take hereto a random sample of size n from the distribution of (U, V). Let X_{ij} be the total number of observations (out of n) that fall in *cell* (i, j) (i.e. in the ith row and jth column). We have

$$(X_{11},\ldots,X_{rc}) \sim \text{Mult}(n, p_{11},\ldots,p_{rc})\ .$$

U and V are independent if and only if

$$p_{ij} = p_i\, q_j, \quad \forall i, j,$$

for some (unknown) $p_1,\ldots,p_r$ and $q_1,\ldots,q_c$. Call this H_0. We can test H_0 by using the test statistic

$$T = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(X_{ij} - n\,\hat{p}_i\hat{q}_j)^2}{n\,\hat{p}_i\hat{q}_j},$$

where

$$\hat{p}_i = \frac{\sum_{j=1}^{c} X_{ij}}{n},$$

(average number in row i), and

$$\hat{q}_i = \frac{\sum_{i=1}^{r} X_{ij}}{n}$$

(average number in column j).

Under H_0, T has a $\chi^2_{(r-1)(c-1)}$ distribution. Check: the degree of freedom is

$$rc - 1 - (r-1) - (c-1) = (r-1)(c-1) .$$

And we reject H_0 for large values of T.

Chapter 9

Regression

The term *regression* was introduced by Galton. Galton observed in an article in 1889 that the height of adult offspring must on the hole, be more "mediocre" than the height of their parents. That is, if a father is 5% taller than average, than a child will be (on the whole) *less* than 5% taller than average. Galton interpreted this as a degenerative phenomenon, hence the name "regression", i.e., regression to mediocracy. An illustration is given in Figure 9.1.

Table 9.1: Regression from father to son. (from Freeman, Pisani and Purves: Statistics).

Regression analysis is about finding relationships between a number of variables. In general there is a *response* variable which we would like to "explain" by one or more *explanatory* variables. For example, Galton was interested in the

relationship between the height of the father (the explanatory variable) and the height of his offspring (the response variable).

Around the turn of the century, Karl Pearson conducted a comprehensive study comparing various relationships between members of the same family. One of the things he did was measure the lenghts of 1078 fathers and their adult sons (one son per father). The results are given in Figure 9.1.

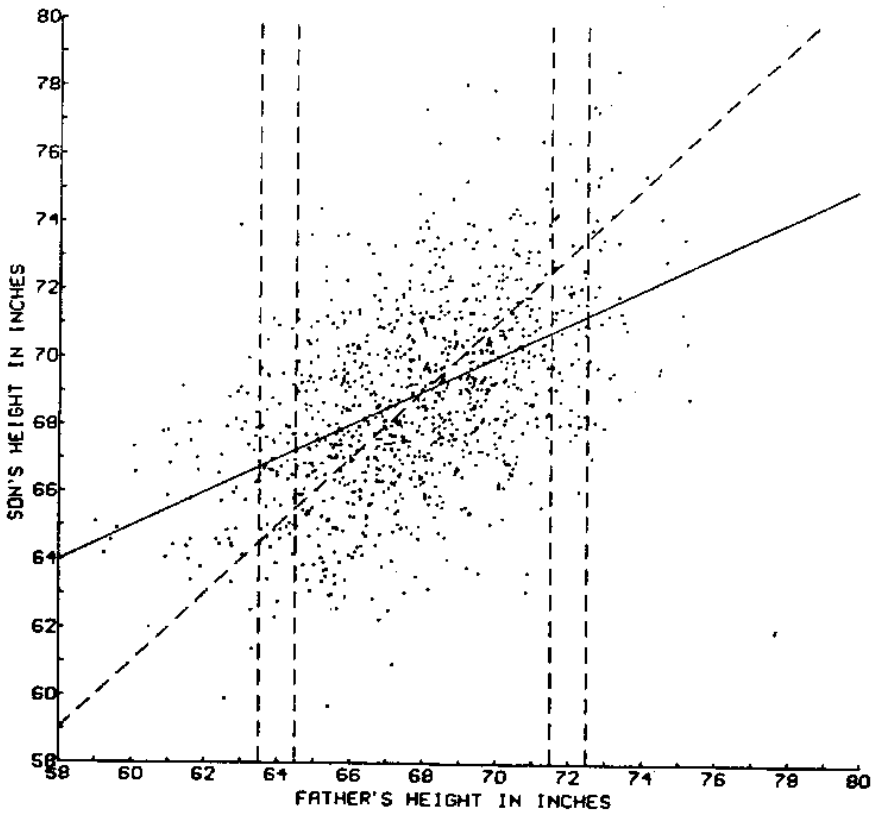


Figure 9.1: A scatter plot of heights. (Freeman, Pisani and Purves: Statistics).

The average length of the fathers was 67 inches and of the sons was 68 inches. Because sons are on average 1 inch longer than the fathers, we could try to "explain" the length of the son by taking the length of his father and adding 1 inch. The predicition line $y = x + 1$ is given in Figure 9.1 as a dashed line. Now consider the strip of points near 64. This corresponds to father that are around 64 inches tall. However, most points in this strip fall *above* the dashed line. In the strip around 72, it's just the other way around: most points lie *below* the dashed line. We observe here the regression effect. Apparently, the line $y = x + 1$ is not so good to predict the height of the sons. The solid line in Figure 9.1 seems to fit better. This line has slope takes the regression effect into account, and has therefore a slope less than 1. We find this line by using the Method of Least Squares.

9.1 Method of Least Squares

The method of least squares is purely a *data analysis* method. It is (can be) used in the following situation: Given a set of data points $(x_1, y_1), \ldots, (x_n, y_n)$ which lie approximately on a straight line. How should we choose the constants b_0 and b_1 such that the line

$$y = b_0 + b_1 x$$

optimally "fits" the data?

Example 9.1 We wish to relate the monthly sales figures of a company to the monthly advertising expenditure. We have the following data

Month	1	2	3	4	5	6	7	8	9	10
Cost x	1.2	0.8	1.0	1.3	0.7	0.8	1.0	0.6	0.9	1.1
Sales y	101	92	110	120	90	82	93	75	91	105

Table 9.2: Sales volume against advertisement expenditure (times $10,000).

Let x_i be the advertising expenditure in the ith month, and let y_i denote the sales in the ith month, $i = 1, \ldots, 10$. A **scatter plot** of the data $(x_1, y_1), (x_2, y_2)$, $\ldots$, (x_{10}, y_{10}) is given in Figure 9.2.

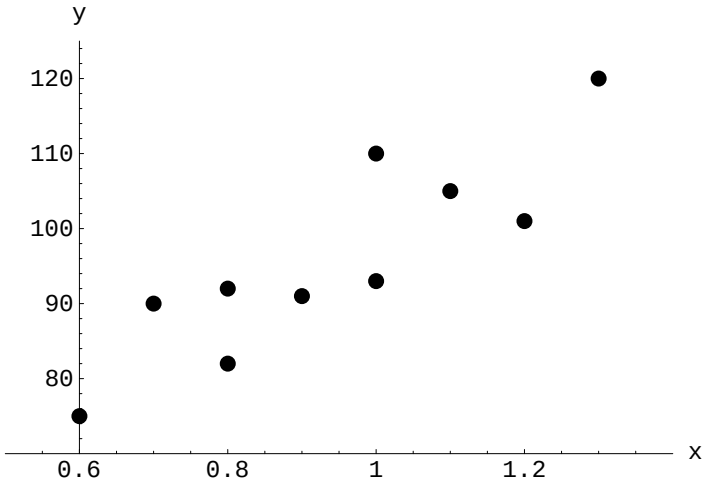


Figure 9.2: Sales volume against advertisement expenditure (times $10,000).

Note that we have here two types of variables: the **response** variable y (sales volume) which we wish to explain/predict via the **explanatory** variable x (cost of ad). The scatter plot shows that the relationship between y and x can be reasonably well described via a straight line $y = b_0 + b_1 x$.

Not all the lines $y = b_0 + b_1 x$ are equally good. When *does* a straight line fit the data well? One often used criterion is the following:

> **Least Square Criterion**: *Choose the line of fit such that the sum of squared deviations from the line is minimal.*

Given a set of data points $(x_1, y_2), \ldots, (x_n, y_n)$, the Least Square Criterion says that we should choose b_0 and b_1 such that

$$L(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2,$$

is minimal. It is not difficult to show that the b_0 and b_1 that minimise L satisfy

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} ,$$

and

$$b_0 = \bar{y} - b_1 \bar{x},$$

where, as usual, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Exercise 9.1 Prove this.

Example 9.2 For our example the LS method yields the line of best fit $y = 46.5 + 52.6\, x$.

Remark 9.1 When speaking of the line of *best* fit, it is important to realize that the line is only "best" in the sense that the sum of squared deviations is minimal. Other criteria could be used to fit a line to the data; and they would lead to *different* lines. Here are some examples.

- We could minimise *absolute* deviations rather than squared deviations. One of the reason that squared deviations are used instead of absolute deviations is that it is mathematically easier to calculate b_0 and b_1 via differentiation of the function L.
- We could minimise "horizontal" squared deviations, for example when we wish to explain x via y.
- We could use *probabilistic* models to find the line of best fit. That is, we view the data as outcomes of random variables (vectors) whose distribution is known up to a few unknown parameters. We then wish to estimate the parameters from the data, and in that way establish the form of the relationship. This will be the subject of Sections 9.2 and 9.3.

Remark 9.2 Sometimes we need to fit a line $y = b\,x$ to the data, i.e. a line which is known to go through the origin. In that case the Least Squares Criterion leads to the formula

$$b = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \ .$$

9.1.1 Non-linear relationships

The response and explanatory variables are not always *linearly* related. There are two ways to deal with this. The first way is to try to fit the data not to a straight line but to another curve, such as a polynomial curve.

Example 9.3 In table below the rate of melanoma per 100,000 white males in the USA is plotted against the north latitude.

Location	Latitude	Melanoma rate
1	32.8	9.0
2	33.9	5.9
3	34.1	6.6
4	37.9	5.8
5	40.0	5.5
6	40.8	3.0
7	41.7	3.4
8	42.2	3.1
9	45.0	3.8

Table 9.3: Rate of melanoma against the latitude.

A linear fit yields: $y = 20.6 - 0.399\,x$.

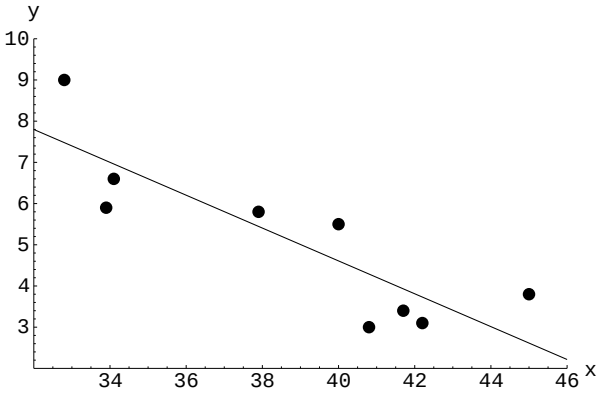


Figure 9.3: A linear fit of the melanoma data.

However, this line does not seem to fit the data very well. Apparently there is a non-linear between x and y. We could try to fit a **polynomial** to the data:

$$y = b_0 + b_1\,x + b_2\,x^2 + \cdots + b_n\,x^k \ .$$

Using the Least Squares Criterion, the optimal $b_0, \ldots, b_k$ are found by minimising

$$L(b_0, \ldots, b_k) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_1 - \cdots - b_k x_i^k)^2$$

with respect to the $k+1$ parameters $b_0, \ldots, b_k$. This is done by solving

$$\frac{\partial L}{\partial b_0} = 0, \ldots, \frac{\partial L}{\partial b_k} = 0,$$

which leads to a set of $k+1$ linear equations for $b_0, \ldots, b_k$. For the example above, a quadratic fit is given by $y = 66.4 - 2.81\,x + 0.0314\,x^2$.

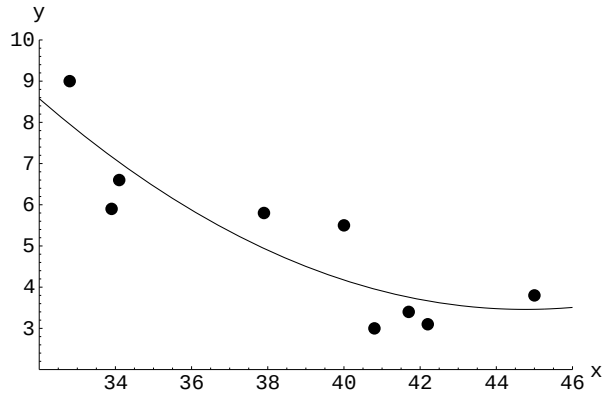


Figure 9.4: A quadratic fit to the melanoma data.

The second way to deal with non-linearities is to *transform* the data in order to achieve a linear relationship. This is particularly useful when the relationships follows from certain theoretical considerations. For example, suppose that from some physical law it follows that $y = a\,\mathrm{e}^{bx}$, for some unknown a and b. Then, $\log y = \log a + b\,x$. Hence, there is a *linear* relationship between x and $\log y$. Thus for some given data $(x_1, y_1), \ldots, (x_n, y_n)$ if we plot $(x_1, \log y_1), \ldots, (x_n, \log y_n)$, these points should approximately lie on a straight line, and we can find the line of fit by applying the formula for the linear case.

Examples of non-linear relationships which can be easily transformed into linear ones:

$$\begin{aligned}
y &= a\,\mathrm{e}^{bx}\,. \\
y &= a\,x^b\,. \\
y &= \frac{L}{1+\mathrm{e}^{a+bx}} \quad \text{(logistic equation)}. \\
y &= 1-\mathrm{e}^{-\frac{x^b}{a}} \quad \text{(in reliability analysis)}\,.
\end{aligned}$$

Which transformations should we use here? For our example, a log-transformation of y gives $\log y = 4.51 - 0.0761\,x$.

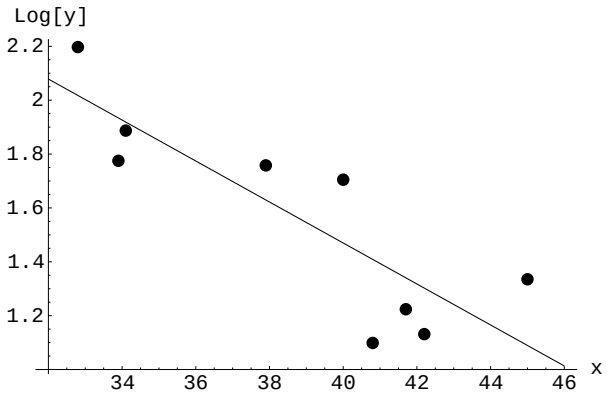


Figure 9.5: A linear fit to a log-transformation of y, for the melanoma data

9.2 A linear regression model

Suppose we have some data $(x_1, y_1), \ldots, (x_n, y_n)$ which is scattered around a straight line. In regression analysis one often views the data as being the outcomes of vectors $(x_1, Y_1), \ldots, (x_n, Y_n)$, where for each *fixed* explanatory variable x_i, the response variable Y_i is a *random* variable with

$$EY_i = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, n, \tag{9.1}$$

for certain *unknown* parameters β_0 and β_1. The line

$$y = \beta_0 + \beta_1 \, x \tag{9.2}$$

is called the **regression line**. This line gives the linear relationship between the expected value of Y and the response variable x. Note that we *cannot draw* this line in a scatter plot because both β_0 and β_1 are *unknown*.

An equivalent formulation of (9.1) is

$$Y_i = \beta_0 + \beta_1 \, x_i + \epsilon_i, \quad \text{with } E\epsilon_i = 0, \quad i = 1, \ldots, n \, . \tag{9.3}$$

This formulation makes it even more clear that we view the responses as random variables which would lie exactly on the regression line, would it not be for some "disturbance" or "error" term (represented by the ϵ_i). Because we wish to describe the behaviour of Y on the basis of x, we consider x as non-random (deterministic) and Y as random. If we wish for example to predict a new value of Y on the basis of x, then x is of course known and Y not. This is expressed by the fact that we use capital letters for Y and lower case letters for x.

The Least Squares Method suggests the following *estimators* for β_0 and β_1.

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \, \bar{x} \tag{9.4}$$

and

$$\hat{\beta}_1 = \frac{S_{xY}}{S_{xx}}, \tag{9.5}$$

where we have used the abbreviations

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{and} \quad S_{xY} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) \; . \tag{9.6}$$

The *observed values* of $\hat{\beta}_0$ and $\hat{\beta}_1$ are also denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$. Usually this does not lead to confusion, but you should be aware of this. If $\hat{\beta}_0$ and $\hat{\beta}_1$ are the observed values of the random variables $\hat{\beta}_0$ and $\hat{\beta}_1$, then the straight line

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

is called the **estimated regression line**. It is important to know that this is not the real regression line.

Here is a nice property of the LS estimators:

Theorem 9.1 $\hat{\beta}_0$ and $\hat{\beta}_1$ are *unbiased* estimators of β_0 and β_1

PROOF. First note that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ (why?), and therefore also $\sum_{i=1}^{n}(x_i - \bar{x})\bar{Y} = 0$. Consequently, we can write S_{xY} also as $\sum_{i=1}^{n}(x_i - \bar{x})Y_i$, and $\hat{\beta}_1$ as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{S_{xx}}. \tag{9.7}$$

Again using $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ we have $\sum_{i=1}^{n}(x_i - \bar{x})\beta_0 = 0$ and $\sum_{i=1}^{n}(x_i - \bar{x})\bar{x} = 0$. As a consequence, we can write the expectation of the numerator (9.7) as

$$\begin{aligned} E\sum_{i=1}^{n}(x_i - \bar{x})Y_i &= \sum_{i=1}^{n}(x_i - \bar{x})EY_i = \sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i) \\ &= \beta_1 \sum_{i=1}^{n}(x_i - \bar{x})x_i = \beta_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 \; . \end{aligned}$$

And thus $E\hat{\beta}_1 = \beta_1$, which shows that $\hat{\beta}_1$ is an unbiased estimator or β_1. To show that $\hat{\beta}_0$ is unbiased as well, observe that

$$E\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} EY_i = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1 \bar{x} \; ,$$

so that

$$E\hat{\beta}_0 = E(\bar{Y} - \hat{\beta}_1 \bar{x}) = E\bar{Y} - \bar{x}\beta_1 = \beta_0 + \beta_1 \bar{x} - \bar{x}\beta_1 = \beta_0 \; .$$

■

For a further analysis of the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ and of the linear regression model in general, we need to specify our model a bit more. In particular, we have to say something about the (random) error terms $\epsilon_1, \ldots, \epsilon_n$.

From now on we assume:

$$\epsilon_1, \ldots, \epsilon_n \text{ are independent and } \epsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n, \tag{9.8}$$

for some unknown parameter σ^2. This model describes exactly how we think the error terms behave. A graphical representation of the densities of the random variables Y_i is given in Figure 9.6.

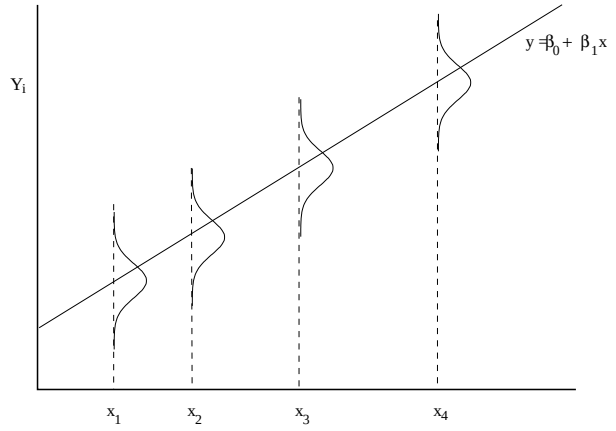


Figure 9.6: A graphical representation of the linear regression model

Combination of (9.3) and (9.8) gives that

$$Y_1, \ldots, Y_n \text{ indep. with } \quad Y_i \sim N(\beta_0 + \beta_1 x_i,\, \sigma^2), \quad i = 1, \ldots, n. \tag{9.9}$$

Thus the distribution of $Y_1, \ldots, Y_n$ is completely known, up to three unknown parameters β_0, β_1 and σ^2. Instead of estimating β_0 and β_1 via the Least Squares Method, we could instead use the Maximum Likelihood Method. Suppose the observed values are $y_1, \ldots, y_n$. Then likelihood function is given by

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2; y_1, \ldots, y_n) &= \prod_{i=1}^{n} \frac{\mathrm{e}^{-\frac{1}{2}(y_i - \beta_0 - \beta_1 x_i)^2/\sigma^2}}{\sqrt{2\pi\sigma^2}} \\ &= \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right\} / (2\pi\sigma^2)^{n/2} \, . \end{aligned}$$

For any given value of σ^2, the function L is maximal if $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ is minimal. Thus the MLEs of β_0 and β_1 are exactly the same as the ones we already found with the Method of Least Squares. It remains to find the MLE of σ^2. Simply copy the derivation of σ^2 in (5.10) to find that the MLE of σ^2 is given by $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1\, x_i)^2$.

Resuming, we have the following ML Estimators for our linear regression model:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{S_{xY}}{S_{xx}}\,, && (9.10)\\
\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1\,\bar{x}\,, && (9.11)\\
\widehat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1\,x_i)^2\,. && (9.12)
\end{aligned}$$

Here S_{xx} and S_{xY} are defined in (9.6).

9.2.1 Properties of the ML Estimators

The following theorem summarises the important properties of the estimators $\hat{\beta}_0$, $\hat{\beta}_1$ and $\widehat{\sigma^2}$.

Theorem 9.2

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2\sum_{i=1}^n x_i^2}{n\,S_{xx}}\right). \tag{9.13}$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right). \tag{9.14}$$

$$\frac{n\,\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-2}\,. \tag{9.15}$$

Also, $\hat{\beta}_1, \bar{Y}$ and $\widehat{\sigma^2}$ are (mutually) independent.

Proof. Recall from the proof of Theorem 9.1 that $\hat{\beta}_1$ can be written as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{S_{xx}}\,. \tag{9.16}$$

In other words, $\hat{\beta}_1$ is a *linear combination of independent normally distributed random variables*, and hence, see Theorem 3.10, it must have a normal distribution. It remains to determine the expectation and variance of this normal distribution. We already saw from Theorem 9.1 that the expectation of $\hat{\beta}_1$ is β_1. As for the variance, we have by the independence of the Y_i,

$$\mathrm{var}(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \mathrm{var} Y_i}{S_{xx}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2\sigma^2}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}},$$

which proves (9.14). Next, we show that $\hat{\beta}_1$ and $\bar{Y}$ are independent. First, observe that $\hat{\beta}_1$ and $\bar{Y}$ have a *joint* (bi-variate) normal distribution, since they both come from a linear combination of Y_i's. Hence, to prove independence, it suffices to show that $\mathrm{cov}(\hat{\beta}_1, \bar{Y}) = 0$. From (9.16) we have that this covariance is equal to

$$\frac{1}{S_{xx}\,n}\mathrm{cov}\Big(\sum_{i=1}^n (x_i - \bar{x})Y_i,\ \sum_{i=1}^n Y_i\Big)\,.$$

Now, since the Y_i's are independent, and the covariance is linear in both arguments, we have

$$\operatorname{cov}(\hat{\beta}_1, \bar{Y}) = \frac{1}{S_{xx}\, n} \sum_{i=1}^{n} (x_i - \bar{x}) \operatorname{cov}(Y_i, Y_i) = \frac{1}{S_{xx}\, n} \sum_{i=1}^{n} (x_i - \bar{x}) \sigma^2 = 0 \,,$$

because $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$.

We can now prove (9.14). First, since $\hat{\beta}_0$ is a linear combination of independent normally distributed random variables it must have a normal distribution; and by Theorem 9.1 the expectation is β_0. Moreover, by the independence of $\hat{\beta}_1$ and $\bar{Y}$ we have

$$\operatorname{var}(\hat{\beta}_0) = \operatorname{var}(\bar{Y}) + \bar{x}^2 \operatorname{var}(\hat{\beta}_1) = \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{xx}} \,.$$

Since

$$n\bar{x}^2 + S_{xx} = \sum_{i=1}^{n} x_i^2,$$

the result (9.14) follows. For the remainder of the proof we refer to the appendix, Section **??**. ■

Theorem 9.2 has a number of important consequences for estimation and hypothesis testing. First, observe that the MLE of σ^2, denoted by $\widehat{\sigma^2}$ is not unbiased. Namely, if we temporarily introduce a random variable $X \sim \chi^2_{n-2}$, then by (9.15) we can write

$$E\widehat{\sigma^2} = \frac{\sigma^2}{n} EX = \frac{\sigma^2}{n}(n-2) \,.$$

However,

$$S^2 := \frac{n}{n-2} \widehat{\sigma^2}$$

is an unbiased estimator of σ^2. Observe that

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2} \,. \tag{9.17}$$

Note that here $\boxed{S^2 \text{ does } \textit{not} \text{ stand for the sample variance,}}$ although it plays a similar role. Unbiased estimators for $\operatorname{var}(\hat{\beta}_0)$ and $\operatorname{var}(\hat{\beta}_1)$ are

$$S_0^2 := \frac{S^2 \sum_{i=1}^{n} x_i^2}{n S_{xx}} \,,$$

and

$$S_1^2 := \frac{S^2}{S_{xx}} \,.$$

The following theorem will enable us to construct a confidence interval for β_1.

Theorem 9.3 The random variable

$$\frac{\hat{\beta}_1 - \beta_1}{S_1}$$

has a t-distribution with $n-2$ degrees of freedom.

PROOF. Write

$$\frac{\hat{\beta}_1 - \beta}{S_1} = \frac{(\hat{\beta}_1 - \beta_1)/\sqrt{\frac{\sigma^2}{S_{xx}}}}{\sqrt{\frac{(n-2)S^2/\sigma^2}{n-2}}} .$$

This is of the form

$$\frac{Z}{\sqrt{V/(n-2)}}$$

with $Z \sim N(0,1)$ and $V \sim \chi^2_{n-2}$ independent, by Theorem 9.2. The result now follows from Theorem 2.3. ■

Corollary 9.1 A $100(1-\alpha)\%$ (stochastic) *confidence interval* for β_1 is given by

$$\left(\hat{\beta}_1 - t_{n-2;1-\alpha/2}\, S_1,\ \hat{\beta}_1 + t_{n-2;1-\alpha/2}\, S_1\right).$$

Now suppose we wish to test the hypothesis $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. It seems sensible to take as test statistic

$$T = \frac{\hat{\beta}_1}{S_1} .$$

Under H_0, has a t_{n-2} distribution. Hence, we reject H_0 at the α level of significance, if

$$T \leq -t_{n-2;1-\alpha/2} \quad \text{or} \quad T \geq t_{n-2;1-\alpha/2}.$$

Example 9.4 The heart rate and the oxygen intake of a certain individual are measured simultaneously. The results are given in Table 9.4.

HR	VO_2	HR	VO_2
94	.473	108	1.403
96	.753	110	1.499
95	.929	113	1.529
95	.939	113	1.599
94	.832	118	1.749
95	.983	115	1.746
94	1.049	121	1.897
104	1.178	127	2.040
104	1.176	131	2.231
106	1.292		

Table 9.4: *Heart rate versus oxygen intake.*

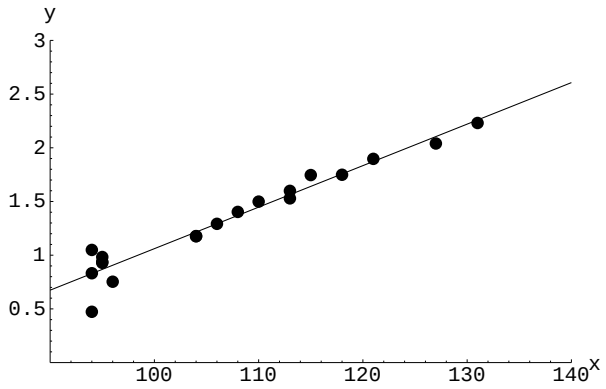


Figure 9.7: *Heart rate versus oxygen intake*

A plot of the data is given in Figure 9.7.

The estimated regression line is

$$y = -2.804 + 0.03865\,x\ .$$

An unbiased estimate of σ^2 is $s^2 = (0.12046)^2$. Moreover, $s_0 = 0.2583$, $s_1 = 0.00256$. With $t_{17;0.975} = 2.1098$, a 95% confidence interval for β_1 is $(0.033, 0.044)$. Note that this does not contain 0, hence we would reject the $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at the $\alpha = 0.05$ level of significance.

Alternatively, The value for the T statistic is $t = \hat{\beta}_1/s_1 = 16.10 > 2.1098$. Thus, reject H_0.

9.2.2 Residuals and fitted values

Note that if $\beta_1 = 0$ our linear regression model (9.9) reduces to the standard model for data that $Y_1, \ldots, Y_n$ are independent and normally distributed with mean β_0 and variance σ^2. Hence, it is important to verify that the introduction of extra "explanatory" variables is justified. This verification can be done as we have seen by examining the confidence interval for β_1 or by testing that β_1 is 0. A third alternative is to compare the "residual variation" in the more complicated model with that of the simple model.

Specifically, if we do not include β_1 into our model, then our model is: $Y_1, \ldots, Y_n$ independent and $N(\beta_0, \sigma^2)$-distributed. The MLE of σ^2

$$\frac{S_{YY}}{n} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

gives an indication how much "residual variance" there is in the model.

If we instead use the linear regression model, then the amount residual variance is (again the MLE of σ^2)

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1\, x_i)^2 \ .$$

The values $\hat{y}_i := \hat{\beta}_0 + \hat{\beta}_1\, x_i$ are called the **fitted values**. The values $y_i - \hat{y}_i$ are called the **residuals**. σ is also called the **residual standard error**.

The amount of variation that is "explained" by the more complicated model is

$$\frac{S_{YY}}{n} - \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1\, x_i)^2 \ .$$

It can be shown (see Section **??**) that this is equal to $\hat{\beta}_1{}^2\, S_{xx}/n$. The quantity

$$R^2 = \frac{\hat{\beta}_1{}^2\, S_{xx}}{S_{YY}} \tag{9.18}$$

is called the **proportion of explained variation**. This quantity lies in the interval [0,1] (why?). The closer R^2 is to 1, the better the regression model fits the data. From (9.5) and (9.18) it follows that we can also write

$$R^2 = \frac{S_{xY}^2}{S_{xx}\, S_{YY}} \ .$$

Note that the quantity

$$R = \frac{S_{xY}}{\sqrt{S_{xx}\, S_{YY}}}$$

looks like a sort of *sample correlation coefficient*, see equation (5.4); indeed we have the same formula!

9.2.3 Confidence interval for $\beta_0 + \beta_1\, x$

Linear regression is most useful when we wish to *predict* how a new response variable will behave, on the basis of a new explanatory variable x. For example, it may be difficult to measure the response variable, but by knowing the estimated regression line and the value for x, we will have a reasonably good idea what Y or the expected value of Y is going to be.

Thus, consider a new x and assume $Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$, First we're going to look at the *expected* value of Y, that is $EY = \beta_0 + \beta_1 x$. Since we do not know β_0 and β_1, we do not know (and will never know) the expected response EY. However, we can *estimate* EY via the MLE

$$\hat{Y} := \hat{\beta}_0 + \hat{\beta}_1\, x = \bar{Y} + \hat{\beta}_1(x - \bar{x}),$$

where $\bar{Y}$ and $\hat{\beta}_1$ are independent and normally distributed. It follows that $\hat{Y}$ has a normal distribution with mean $\beta_0 + \beta_1 x$ and variance

$$\frac{\sigma^2}{n} + \frac{\sigma^2 (x - \bar{x})^2}{S_{xx}} .$$

We can estimate this variance with

$$S^2 \left\{ \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right\} .$$

Then, similar to Theorem 9.3 we can show that

$$\frac{\hat{Y} - (\beta_0 + \beta_1 x)}{S\sqrt{\left\{ \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}} \right\}}} \sim t_{n-2}.$$

Consequently, a $100(1 - \alpha)\%$ (stochastic) *confidence interval* for $\beta_0 + \beta_1 x$ is given by

$$\boxed{\left(\hat{Y} - t_{n-2;1-\alpha/2}\, S\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}},\ \hat{Y} + t_{n-2;1-\alpha/2}\, S\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \right) .}$$

Example 9.5 A 95% confidence interval for the expected oxygen intake at a heart rate of 140 is

$$2.60694 \pm 2.1098 \times 0.0839009 = (2.43, 2.78) .$$

9.2.4 Prediction interval for Y

If we wish to *predict* the value of Y for a given value of x, then we have *two* sources of variation:

1. Y itself is a random variable, which is normally distributed with variance σ^2,
2. we don't know the expectation $\beta_0 + \beta_1 x$ of Y. Estimating this number on the basis of previous observations $Y_1, \ldots, Y_n$ brings another source of variation.

Thus instead of a confidence interval for $\beta_0 + \beta_1 x$ we need a *prediction interval* for a new response $Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$, where Y is independent of every Y_i.

It is not so difficult to show that $Y - (\hat{\beta}_0 + \hat{\beta}_1 x)$ has a normal distribution with mean 0 and variance

$$\sigma^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) .$$

Let S^2_{pred} be as above with σ^2 replaced by S^2. It follows in the usual way that

$$\frac{Y - (\hat{\beta}_0 + \hat{\beta}_1 x)}{S_{\text{pred}}} \sim t_{n-2}.$$

Thus, Y lies with probability $1 - \alpha$ in the **prediction interval**

$$\boxed{\left(\hat{Y} - t_{n-2;1-\alpha/2}\, S_{\text{pred}},\ \hat{Y} + t_{n-2;1-\alpha/2}\, S_{\text{pred}}\right).}$$

Example 9.6 (Oxygen/heartbeat, continued)

A 95% prediction interval for the same heart rate is

$$2.60694 \pm 2.1098 \times 0.146799 = (2.30, 2.92)\ .$$

The latter one is much wider, because now there are two sources of variation.

9.3 Linear regression via the bi-variate normal distribution

Consider the scatter plot of in Figure 9.1. In Section 9.2 we viewed the data $(x_1, y_1), \ldots, (x_n, y_n)$ as outcomes of vectors $(x_1, Y_1), \ldots, (x_n, Y_n)$ where the x_i are fixed and the Y_i are random. What if we also take the x_i random? In particular, suppose we view the data $(x_1, y_1), \ldots, (x_n, y_n)$ as outcomes of a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from some two-dimensional distribution. It is customary to take this sampling distribution to be bi-variate normal.

To make the notation more in line with Section 3.4 let us view the data as outcomes of *column* vectors. Thus our model for the data is that $(X_1, Y_1)^T, \ldots, (X_n, Y_n)^T$ are i.i.d. vectors with a bi-variate normal distribution with unknown mean vector $\boldsymbol{\mu} = (\mu_X, \mu_Y)^T$ and unknown covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\ .$$

Hence, the model depends on 5 unknown parameters. From Example 3.9 we know that

$$X_i \sim N(\mu_X, \sigma_X^2), \quad Y_i \sim N(\mu_Y, \sigma_Y^2) \quad \text{and} \quad \rho(X_i, Y_i) = \rho\ . \tag{9.19}$$

The density of each $(X_i, Y_i)^T$ is given in (3.15) (of course with the appropriate substitutions $\mu_1 = \mu_X$, etc). In particular, the *contourlines* of the density are *ellipses*. This suggests that a scatter plot of the outcomes of $(X_1, Y_1)^T, \ldots, (X_n, Y_n)^T$ could look something like Figure 9.1 where it seems that the data points are scattered in an "elipsoid" fashion.

In fact, we can easily *generate* outcomes of $(X_1, Y_1)^T, \ldots, (X_n, Y_n)^T$ via the computer. The procedure is based on the transformation (3.11). Let U and V be independent and standard normally distributed random variables. Then, by Example 3.9 the random vector

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} + \begin{pmatrix} \sigma_X & 0 \\ \sigma_Y \rho & \sigma_Y \sqrt{1-\rho^2} \end{pmatrix} \begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} \mu_X + \sigma_X U \\ \mu_Y + \sigma_Y \rho U + \sigma_Y \sqrt{1-\rho^2}\, V \end{pmatrix}$$

has a bivariate normal distribution with the same expectation vector $\boldsymbol{\mu}$ and covariance matrix Σ as above. This shows how to generate the $(x_i, y_i)^T$. First generate $(u_i, v_i)^T$ and then apply the transformation above. Alternatively, we can write (check yourself)

$$Y = \mu_Y - \mu_X \frac{\sigma_Y}{\sigma_X} \rho + \frac{\sigma_Y}{\sigma_X} \rho X + \sigma_Y \sqrt{1-\rho^2}\, V \ .$$

Thus, for a *given* value $X = x$ the random variable Y has a normal distribution with mean

$$\mu_Y - \mu_X \frac{\sigma_Y}{\sigma_X} \rho + \frac{\sigma_Y}{\sigma_X} \rho x \tag{9.20}$$

and variance $\sigma_Y^2(1-\rho^2)$. As a consequence, we could draw the $(x_i, y_i)^T$ by first generating x_i from a $N(\mu_X, \sigma_X^2)$ distribution, and then (independently) generate y_i from a normal distribution with expectation as in (9.20) (with $x = x_i$) and variance $\sigma_Y^2(1-\rho^2)$.

Let us now consider some of the statistical aspects of this model. First, what would be sensible estimators for the five unknown parameters $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$ and ρ. The method of moments (see Section 5.2) and equation (9.19) give respectively the estimators: $\bar{X}$, $\bar{Y}$, S_{XX}/n, S_{YY}/n, and

$$R := \frac{S_{XY}}{\sqrt{S_{XX}\, S_{YY}}} \ ,$$

where we have used the abbreviations

$$S_{XX} = \sum_{i=1}^{n} (X_i - \bar{X})^2, \quad S_{XY} = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}), \quad S_{YY} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \ .$$

Note that R is exactly the sample correlation coefficient in (5.4). By the usual method – taking partial derivatives of the likelihood function with respect to the five parameters – it can also be shown that the above estimators are also the *Maximum Likelihood Estimators.*

We may test $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$ using the test statistic

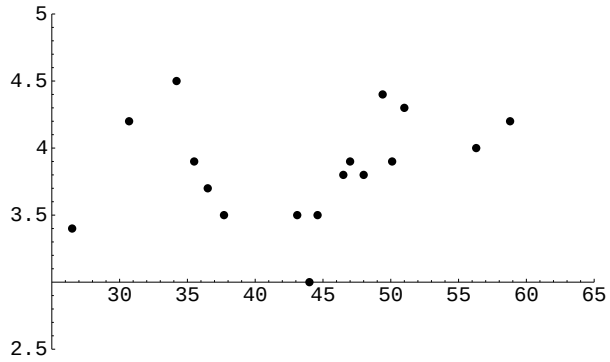
$$\frac{R\sqrt{n-2}}{\sqrt{1-R^2}},$$

which under H_0 has a t_{n-2} distribution. We omit the proof of this result.

Example 9.7 The following set of data is computer generated from a bi-variate normal distribution.

i	x_i	y_i
1	56.3	4.0
2	19.9	3.0
3	26.5	3.4
4	44.6	3.5
5	34.2	4.5
6	50.1	3.9
7	48.0	3.8
8	51.0	4.3
9	36.5	3.7
10	58.8	4.2

i	x_i	y_i
11	47.0	3.9
12	49.4	4.4
13	39.9	5.2
14	44.0	3.0
15	35.5	3.9
16	30.7	4.2
17	79.8	4.7
18	37.7	3.5
19	46.5	3.8
20	43.1	3.5



The maximum likelihood estimates are: $\widehat{\mu_X} = 43.975$, $\widehat{\mu_Y} = 3.92$, $\widehat{\sigma_X^2} = 157.237$, $\widehat{\sigma_Y^2} = 0.2846$, and $\hat{\rho} = 0.44$.

The statistic $\sqrt{18}\, R/\sqrt{1-R^2}$ has value 2.06. Do we reject $H_0 : \rho = 0$ in favour of $H_1 : \rho \neq 0$?

The **true** parameters used in the computer program were: $\mu_X = 43$, $\mu_Y = 4$, $\sigma_X^2 = 169$, $\sigma_Y^2 = 0.25$ and $\rho = 0.4$.

Chapter 10

Analysis of Variance

Analysis of variance (ANOVA) is a statistical technique for comparing the *means* of different populations by relating the variance *between* the populations to the variance *within* the populations.

Example 10.1 Four groups of people (categorised by their age) are asked to do a physical test. The *increase* in heart rate is recorded below.

Do the data provide sufficient evidence to indicate a difference in mean increase in heart rate?

A (10-19)	B (20-39)	C (40-59)	D (60-69)
29	24	37	28
33	27	25	29
26	33	22	34
27	31	33	36
39	21	28	21
35	28	26	20
33	24	30	25
29	34	34	24
36	21	27	33
22	32	33	32
309	275	295	282

Table 10.1: *Increase in heart rate*

The boxplots in Figure 10.1 show that the variability *within* each sample is larger than the variability *between* the samples. This supports the case that the population means are identical. Or that *age* is not a "factor" that explains the increase in heart rate.

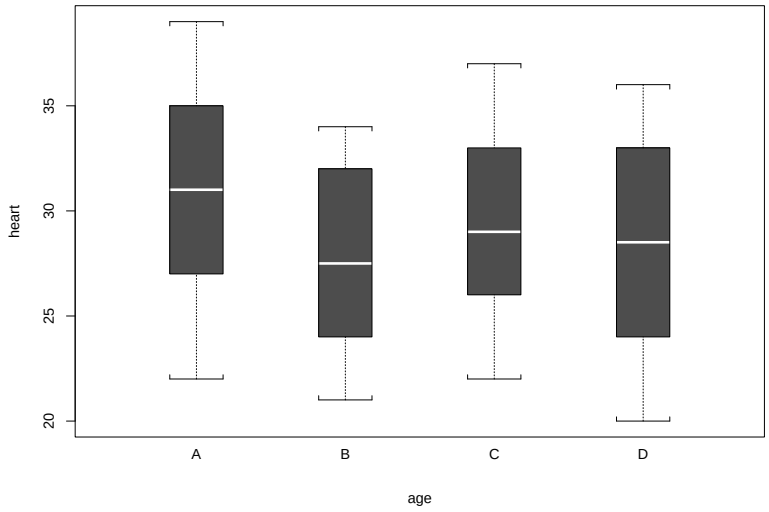


Figure 10.1: *Boxplots of the increase in heart rate for the four different age categories*

10.1 Completely randomised design

The situation in Example 10.1 is very similar to the one in the 2-sample t-test. In the 2-sample t-test we had *two* independent samples and we wanted to test whether the two sample means were the same or not. Here we have *four* independent samples and want to test whether the (four) sample means are the same or not. However, there is no such thing as a 4-sample t-test (can you see why?). Instead we base our test statistic on the proportion of the variation between samples and the variation within samples, as indicated in the beginning of this chapter. But first, we need a proper model for the data.

10.1.1 Model

We wish to analyse how a **response variable** depends on one **factor** which is tested at k **levels**. In Example 10.1 the "increase in heart rate" is the response variable and "age" is the factor, which is tested at 4 levels. The different levels are also called **treatments**.

Denote by $Y_{1j}, Y_{2j}, \ldots, Y_{n_j j}$ the responses at level $j = 1, \ldots, k$. The simplest model is that

$$\boxed{Y_{ij} = \mu_j + \epsilon_{ij},}$$

where ϵ_{ij} are independent and $N(0, \sigma^2)$ distributed. This type of model, or *design*, is called a **completely randomised, one-factor design**.

An equivalent model is: $\{Y_{ij}\}$ are independent and

$$Y_{ij} \sim N(\mu_j, \sigma^2), \quad i = 1, \ldots, n_j, \quad j = 1, \ldots, k \, .$$

Note that this model is just a k-sample generalisation of the one used for the 2-sample t test.

The parameters in the model are interpreted as follows:

- μ_j is the true mean within the jth level,
- the *error terms* ϵ_{ij} give the deviations from the true mean.

Note that the parameters $\mu_1, \ldots, \mu_k$ and σ^2 are unknown.

10.1.2 Estimation

How can we estimate the parameters in the model above? Note that there are $k + 1$ unknown parameters. Let $n = n_1 + \cdots + n_k$, and define

$$\begin{aligned} \bar{y}_{\bullet\bullet} &:= \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} y_{ij} \quad \text{(overall sample average)} \\ \bar{y}_{\bullet j} &:= \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} \quad \text{(average within level } j\text{)} \end{aligned}$$

By the usual method, it can be shown that that MLEs of the parameters are given by

$$\hat{\mu}_j := \bar{y}_{\bullet j}, \quad j = 1, \ldots, k,$$

and

$$\widehat{\sigma^2} := \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{\bullet j})^2 \, .$$

Example 10.2 Estimate the parameters for the heart rate example.

10.1.3 Hypothesis testing

We wish to test

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = 0$$

versus H_1: not all μ_i are the same. We hereto define

Total Sum of Squares:

$$\mathrm{SS}_{\mathrm{total}} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\bullet\bullet})^2 \, .$$

Treatment Sum of Squares:

$$\mathrm{SS}_{\mathrm{treatment}} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})^2 \ .$$

Error Sum of Squares:

$$\mathrm{SS}_{\mathrm{error}} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\bullet j})^2 \ .$$

Check that $\mathrm{SS}_{\mathrm{treatment}}$ measures the variability **between** the groups; and $\mathrm{SS}_{\mathrm{error}}$ measures the variability **within** the groups. It can also be shown that

$$\mathrm{SS}_{\mathrm{total}} = \mathrm{SS}_{\mathrm{treatment}} + \mathrm{SS}_{\mathrm{error}} \ .$$

Namely, write $\mathrm{SS}_{\mathrm{total}}$ as

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} \left[(\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet}) + (Y_{ij} - \bar{Y}_{\bullet j}) \right]^2 \ , \tag{10.1}$$

and note that sum of the "cross product" vanishes, i.e.,

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} (\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})(Y_{ij} - \bar{Y}_{\bullet j}) = \sum_{j=1}^{k} (\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet}) \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\bullet j}) = 0 \ .$$

The result follows now by expanding the square root in (10.1).

Based on our reasoning in Example 10.1 and the definitions above, it seems reasonable to define (a function of) $\mathrm{SS}_{\mathrm{treatment}}/\mathrm{SS}_{\mathrm{error}}$ to be our (ad hoc) test statistic. Namely, if the means are different, then variability between groups will be comparatively large to the variability within groups, and hence the quantity above will be large. On the other hand, if the means are the same then it is unlikely that this quantity will be large.

Before give the final test statistic and its distribution under H_0 we consider some properties of the sum of squares random variables above.

Theorem 10.1 We have

$$\frac{\mathrm{SS}_{\mathrm{error}}}{\sigma^2} \sim \chi^2_{n-k} \ .$$

and under H_0

$$\frac{\mathrm{SS}_{\mathrm{treatment}}}{\sigma^2} \sim \chi^2_{k-1} \ .$$

Moreover, $\mathrm{SS}_{\mathrm{treatment}}$ and $\mathrm{SS}_{\mathrm{error}}$ are independent.

PROOF. A proof of this is a bit beyond the scope of this course, and can be found in Section **??** of the appendix. ■

As a direct consequence of the above theorem and Theorem 2.2 we have the following result.

Corollary 10.1 Under H_0,

$$F := \frac{\mathrm{SS}_{\text{treatment}}/(k-1)}{\mathrm{SS}_{\text{error}}/(n-k)} \sim F^{k-1}_{n-k}. \tag{10.2}$$

Remark 10.1 The numerator of (10.2) is also written as $\mathrm{MS}_{\text{treatment}}$ (**Mean Square treatment**) and the denominator as $\mathrm{MS}_{\text{error}}$ (**Mean Square Error**). In some statistical packages "treatment" is replaced by "factor".

We thus have the following test. As test statistic we take

$$F = \frac{\mathrm{MS}_{\text{treatment}}}{\mathrm{MS}_{\text{error}}}.$$

And for significance level α we reject H_0 if $F \geq F^{k-1}_{n-k;1-\alpha}$, where $F^{k-1}_{n-k;1-\alpha}$ is the $(1-\alpha)$-quantile of the F^{k-1}_{n-k}-distribution.

Remark 10.2 When there are only **two** samples, we can use the 2-sample t-test to test equality of the means.

Exercise 10.1 For the heart rate data verify the following: Sum of squares of treatment: 67.475 . Sum of squares of error: 935.500 . Number of levels/treatments $k = 4$. Number of data $n = 40$. The outcome of F is $\frac{67.475/3}{935.500/36} = 0.8655265$. For $\alpha = 0.05$, $F^3_{36;0.95} \approx 2.87$. Hence, do not reject H_0. (The p-value of the test is 0.4678498).

10.1.4 Using the computer

We can use statistical packages to do the ANOVA for us. For ANOVA with one factor, data is often entered as a **1-way table**:

response	factor
y_{11}	1
$\vdots$	$\vdots$
$y_{n_1 1}$	1
$\vdots$	$\vdots$
$\vdots$	$\vdots$
y_{1k}	k
$\vdots$	$\vdots$
$y_{n_k k}$	k

The output is an **ANOVA table**, e.g.

```
           Df  Sum of Sq  Mean Sq  F Value  Pr(F)
Factor     3   67.475     22.49    0.86552  0.467
Residuals 36   935.500    25.98
```

Example 10.3 (Corn example) Five varieties of corn are planted in three plots in a large field. The following 1-way table lists the yields.

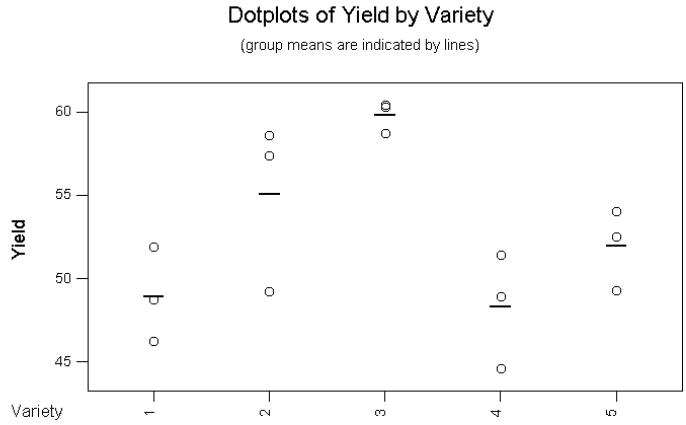
Yield	Variety
46.2	1
51.9	1
48.7	1
49.2	2
58.6	2
57.4	2
60.3	3
58.7	3
60.4	3
48.9	4
51.4	4
44.6	4
52.5	5
54.0	5
49.3	5

In Minitab, `Stat>ANOVA>One-way` yields

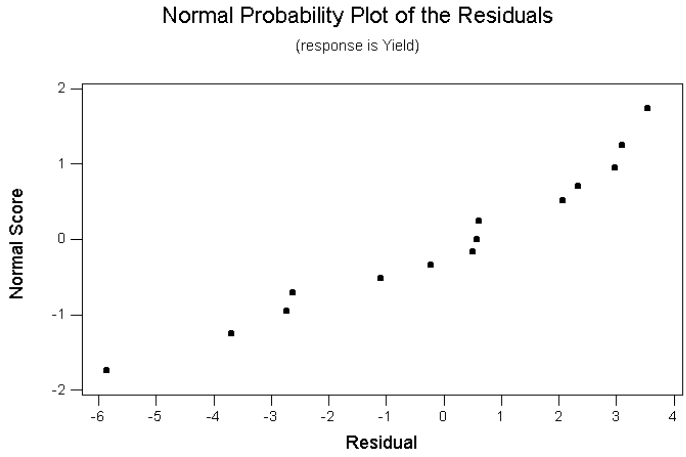
```
One-way ANOVA: Yield versus Variety

Analysis of Variance for Yield
Source     DF        SS     MS      F      P
Variety     4     270.3   67.6   6.39  0.008
Error      10     105.7   10.6
Total      14     375.9
```

```
Level       N      Mean  StDev
1           3    48.933  2.857
2           3    55.067  5.116
3           3    59.800  0.954
4           3    48.300  3.439
5           3    51.933  2.401

Pooled StDev =    3.251

Individual 95% CIs For Mean
Based on Pooled StDev
-------+---------+---------+---------
  (------*------)
           (------*------)
                  (------*------)
 (-----*------)
       (------*------)
-------+---------+---------+---------
    48.0      54.0      60.0
```

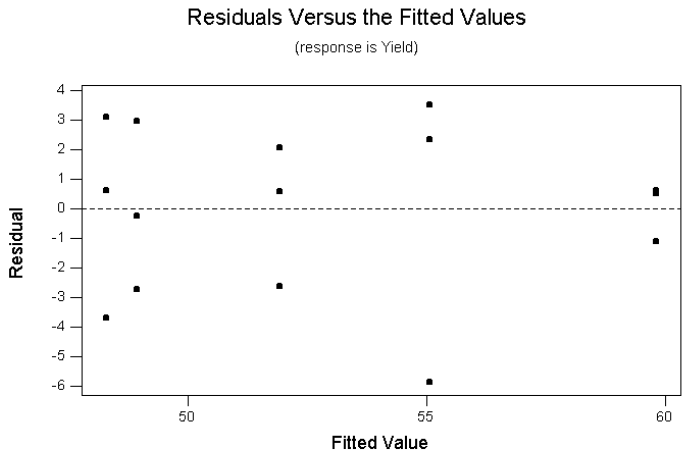
Minitab also gives various plots. For example, a **dotplot** of the data:



A **normal plot** of the residuals:



Residuals vs Fits for Yield



10.1.5 Contrasts

Contrasts are used to test sub-hypotheses which can be formulated **before** the data is taken.

Example 10.4 In the corn example, we had five varieties of corn. Suppose varieties 1, 2, 4 and 5 are mutations of a common variety, unlike variety 3. We may wish to test if the mean yield of variety 5 is the same as the average mean yield of the other varieties. In other words, test

$$H_0' : \frac{\mu_1 + \mu_2 + \mu_4 + \mu_5}{4} = \mu_3 \ .$$

A **contrast** is a linear combination of μ_i's such that the coefficients sum up to 0.

Let $C = \sum_{j=1}^{k} c_j \mu_j$ be a contrast. An unbiased estimator for C is

$$\hat{C} := \sum_{j=1}^{k} c_j \bar{Y}_{\bullet j}.$$

The variance of $\hat{C}$ is

$$\sigma^2 \sum_{j=1}^{k} c_j^2 / n_j.$$

Define

$$\mathrm{SS}_C := \frac{\hat{C}^2}{\sum_{j=1}^{k} c_j^2 / n_j}$$

Then under $H_0 : C = 0$ we can show that $\mathrm{SS}_C / \sigma^2 \sim \chi_1^2$, independent of $\mathrm{SS}_{\mathrm{error}}$. Consequently, to test H_0 against $H_1 : C \neq 0$ we can use the test statistic

$$F := \frac{\mathrm{SS}_C}{\mathrm{MS}_{\mathrm{error}}},$$

which under H_0 has a F^1_{n-k} distribution.

Example 10.5 For the corn example, the hypothesis

$$H_0' : \frac{\mu_1 + \mu_2 + \mu_4 + \mu_5}{4} = \mu_3 \ ,$$

corresponds with the contrast

$$C = \frac{\mu_1 + \mu_2 + \mu_4 + \mu_5}{4} - \mu_3 \ .$$

An unbiased estimate for C is

$$\frac{48.933 + 55.067 + 48.300 + 51.933}{4} - 59.800 = -8.742 \ .$$

Also,

$$\sum_{j=1}^{5} c_j^2/n_j = \frac{1}{3}\left(\frac{4}{16} + 1\right) = 0.41667 \ .$$

Hence, the value for the F-statistic is

$$\frac{(-8.742)^2/0.41667}{10.6} \approx 17 \ .$$

From the table of the F^1_{10} distribution we see that the p value is smaller than 0.005. Hence, we reject the hypothesis $H_0' : C = 0$ against $H_1' : C \neq 0$.

10.1.6 Multiple comparisons

Consider a completely randomized, one-factor design with k levels and $n_j = r$ (constant) for all $j = 1, \ldots, k$.

The probability is $1-\alpha$ that all $\binom{k}{2}$ pairwise contrast $\mu_i - \mu_j$ will simultaneously satisfy

$$\bar{Y}_{\bullet i} - \bar{Y}_{\bullet j} - D\sqrt{\mathrm{MS}_{\mathrm{error}}} < \mu_i - \mu_j < \bar{Y}_{\bullet i} - \bar{Y}_{\bullet j} + D\sqrt{\mathrm{MS}_{\mathrm{error}}},$$

where $D = Q_{k,r;1-\alpha}/\sqrt{r}$ and $Q_{k,r;1-\alpha}$ is the $(1-\alpha)$-quantile of the *studentised range* (with parameters k and r). The (stochastic) intervals above are called *Tukey intervals.*

If for a given i and j, zero is not contained in the above inequality, then $H_0' : \mu_i = \mu_j$ can be rejected in favour of $H_1' : \mu_i \neq \mu_j$ at the α level of significance.

Example 10.6 (Corn example, continued) We wish to construct the 10 confidence intervals for $\mu_i - \mu_j$ at a significance level of 0.05.

Use Minitab, `Stat>ANOVA>One-way`. Under `Comparisons...` select and specify `Tukey's, Family error rate`: 0.05. This gives

```
          1          2          3          4

2     -14.861
        2.594

3     -19.594    -13.461
       -2.139      3.994

4      -8.094     -1.961      2.773
        9.361     15.494     20.227

5     -11.727     -5.594     -0.861    -12.361
        5.727     11.861     16.594      5.094
```

At this level of significance we reject two sub-hypotheses: $\mu_1 = \mu_3$ and $\mu_3 = \mu_4$.

Exercise 10.2 Verify the Tukey interval for $\mu_3 - \mu_4$.

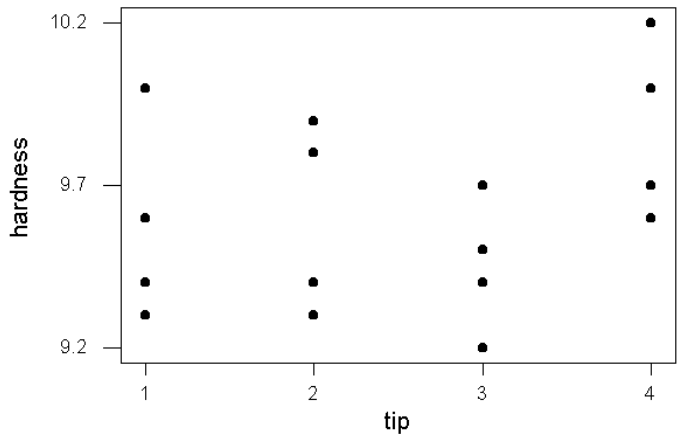
10.2 Randomized Block Design

Randomized block design is a statistical method in which data is collected in blocks, in order to reduce variation in the experiment.

Example 10.7 The hardness of metals is measured by pressing the tip of a "hardness tester" onto a coupon of metal. In the table below the hardness readings are given for four different types of tips, tested on four different coupons of metal. Is there any difference in hardness reading between the different tips?

	Tip			
Coupon	1	2	3	4
1	9.3	9.4	9.2	9.7
2	9.4	9.3	9.4	9.6
3	9.6	9.8	9.5	10.0
4	10.0	9.9	9.7	10.2

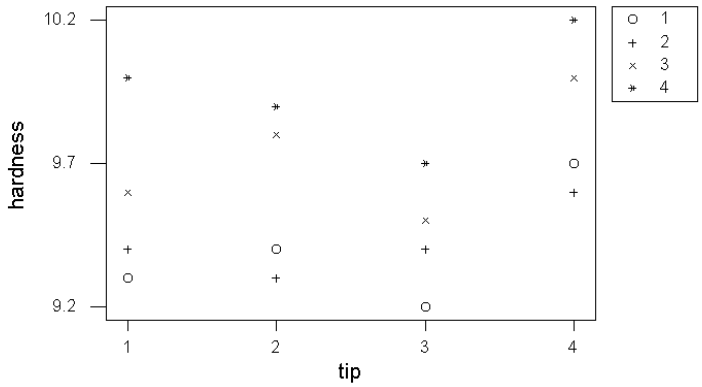
First let us look at a plot of the data.



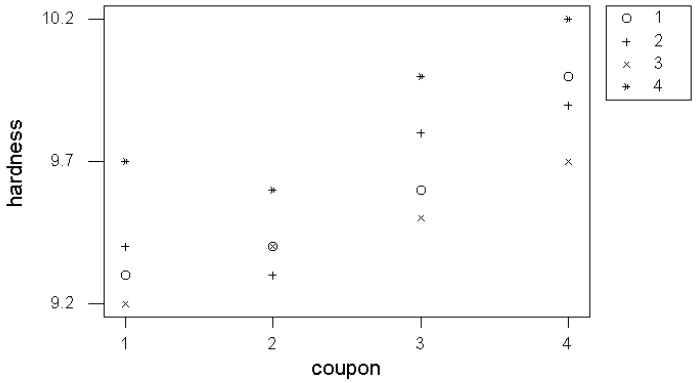
This picture does not suggest any difference between the mean hardness readings. Let's do an F-test for the completely randomized one-factor design, where the factor is the tip type.

```
Analysis of Variance for hardness

Source     DF        SS        MS        F        P
tip         3    0.3850    0.1283     1.70    0.220
Error      12    0.9050    0.0754
Total      15    1.2900
```

We see that the p-value is quite large. The variability between groups is small compared to the variability within groups (of the same type). No reason to doubt that the tip type has an effect on the readings. But wait! Maybe the variability in the readings is due to the different types of coupons. If we specify the coupon types in the above plot, we get the following picture.



Coupon type is important in explaining the hardness readings. For example, for each tip type, coupon 4 is giving a higher reading than coupon 1. We should really be looking at the effect of tip type *within* each coupon:



This picture indicates that there *is* a difference between tip types. (Why?) How can we quantify this? We need a model.

10.2.1 Model

The response data depends on b **blocks** and t **treatments**. In the example, the blocks are the coupon types, and the treatments are the tip types.

Let y_{ij} be the response for block i and treatment j. The model for the **randomized block design** is that each y_{ij} is the outcome of a random variable

$$\boxed{Y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij},}$$

where the ϵ_{ij} are mutually independent and $N(0, \sigma^2)$ distributed, and $\sum_i \beta_i = 0$ and $\sum_j \tau_j = 0$ (fixed effects).

- μ is the *overall (true) mean*,
- β_i is the *differential effect* of the ith block,
- τ_j is the *differential effect* of the jth treatment,
- the ϵ_{ij} are the *error terms*.

10.2.2 Hypothesis testing

We wish to test

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_t = 0$$

versus H_1: not all τ_i are 0. Similar to the completely randomized design, we define

Total Sum of Squares:

$$\mathrm{SS}_{\mathrm{total}} = \sum_{i=1}^{b}\sum_{j=1}^{t}(Y_{ij} - \bar{Y}_{\bullet\bullet})^2 \ .$$

Blocks Sum of Squares:

$$\mathrm{SS}_{\mathrm{blocks}} = \sum_{i=1}^{b}\sum_{j=1}^{t}(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 \ .$$

Treatment Sum of Squares

$$\mathrm{SS}_{\mathrm{treatment}} = \sum_{i=1}^{b}\sum_{j=1}^{t}(\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})^2 \ .$$

Error Sum of Squares:

$$\mathrm{SS}_{\mathrm{error}} = \sum_{i=1}^{b}\sum_{j=1}^{t}(Y_{ij} - \bar{Y}_{i\bullet} - \bar{Y}_{\bullet j} + \bar{Y}_{\bullet\bullet})^2 \ .$$

The total "variability" can be decomposed into three parts:

$$\mathrm{SS}_{\mathrm{total}} = \mathrm{SS}_{\mathrm{treatment}} + \mathrm{SS}_{\mathrm{blocks}} + \mathrm{SS}_{\mathrm{error}} \ .$$

Theorem 10.2 It can be shown that

$$\frac{\mathrm{SS}_{\mathrm{error}}}{\sigma^2} \sim \chi^2_{(b-1)(t-1)} \ .$$

And, when $\tau_1 = \ldots = \tau_t = 0$,

$$\frac{\mathrm{SS}_{\mathrm{treatment}}}{\sigma^2} \sim \chi^2_{t-1} \ .$$

And, when $\beta_1 = \ldots = \beta_b = 0$,

$$\frac{\mathrm{SS}_{\mathrm{blocks}}}{\sigma^2} \sim \chi^2_{b-1} \ .$$

Moreover, $\mathrm{SS}_{\mathrm{treatment}}$, $\mathrm{SS}_{\mathrm{blocks}}$ and $\mathrm{SS}_{\mathrm{error}}$ are independent.

Corollary 10.2 Hence, when $\tau_1 = \ldots = \tau_t = 0$,

$$\frac{\mathrm{MS}_{\mathrm{treatment}}}{\mathrm{MS}_{\mathrm{error}}} \sim F^{t-1}_{(b-1)(t-1)}.$$

And, when $\beta_1 = \ldots = \beta_b = 0$,

$$\frac{\mathrm{MS}_{\mathrm{blocks}}}{\mathrm{MS}_{\mathrm{error}}} \sim F^{b-1}_{(b-1)(t-1)}.$$

Test $H_0 : \tau_1 = \ldots = \tau_t = 0$, against H_1 : not all τ_i are 0. Reject H_0 at the α level of significance if

$$\frac{\text{MS}_{\text{treatment}}}{\text{MS}_{\text{error}}} > F^{t-1}_{(b-1)(t-1);1-\alpha} .$$

Test $H_0 : \beta_1 = \ldots = \beta_b = 0$, against H_1 : not all β_j are 0. Reject H_0 at the α level of significance if

$$\frac{\text{MS}_{\text{blocks}}}{\text{MS}_{\text{error}}} > F^{b-1}_{(b-1)(t-1);1-\alpha} .$$

Exercise 10.3 Determine the outcomes of the F statistics for the hardness-tip example. What are your conclusions?

10.2.3 Using the computer

We may analyse the data with a 2-way ANOVA table, in which the data could be entered as

response	block	treatment
y_{11}	1	1
$\vdots$	$\vdots$	$\vdots$
y_{b1}	b	1
$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$
y_{1t}	1	t
$\vdots$	$\vdots$	$\vdots$
y_{bt}	b	t

Remark 10.3 A *cell* is a particular combination of the levels of the two factors (blocks and treatment). There are $b \times t$ cells. Each cell has the same number of observations (only 1). This is called a **balanced design**.

Example 10.8 Let us examine the hardness vs tip-type example. Minitab gives the following results

```
Analysis of Variance for hardness

Source  DF       SS        MS        F        P
coupon   3  0.82500   0.27500    30.94    0.000
tip      3  0.38500   0.12833    14.44    0.001
Error    9  0.08000   0.00889
Total   15  1.29000
```

We see now a very significant difference between the tip types, because the F-test yields a p-value of 0.001.

The large value for the coupon SS indicates that block type is an important contributor to the variability of the responses, indicating it was sensible to use the block design here.

Remark 10.4 The same approach may be used for **repeated measurements** (more than one observation per cell), as long as the design remains balanced.

10.2.4 Contrasts and Tukey intervals

We may test sub-hypotheses in the same way as for the completely randomized block design. For the contrast C the appropriate test statistic is again

$$\frac{\mathrm{SS}_C}{\mathrm{MS}_{\mathrm{error}}},$$

which under $H_0 : C = 0$ has a $F^1_{(b-1)(t-1)}$-distribution. Tukey's intervals for all $\mu_i - \mu_j$ are the same as before, except that

$$D = Q_{t,(b-1)(t-1);1-\alpha}/\sqrt{b}\,.$$

(Note: $\mu_i = \mu + \tau_i$ is the true mean for the ith treatment.)

Question. What is the true mean for the jth block? How can we estimate it?

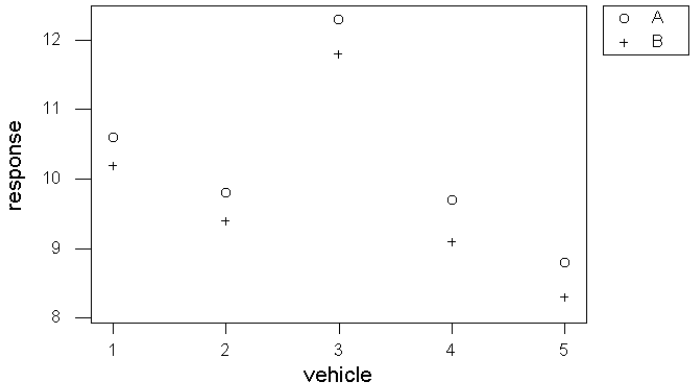
10.2.5 Paired t-test

Example 10.9 Consider two types of tyres (A and B), which are tested on 5 cars. Each car has one tyre of type A and B. To see if there is a difference in quality between the types, the difference in tyre thickness after a driving distance of 10000 km is measured. The results are listed below.

Car	1	2	3	4	5
Type A	10.6	9.8	12.3	9.7	8.8
Type B	10.2	9.4	11.8	9.1	8.3

Table 10.2: Tyre thickness

The wear on each car clearly depends on the car and driver. This calls for a Randomized Block Design with 5 blocks and 2 treatments.



We may perform an F-test. However, notice that under the RBD model assumptions, the **difference**

$$D_i = Y_{i1} - Y_{i2}$$

has a $N(\mu_D, \sigma^2)$ distribution, with $\mu_D = \tau_1 - \tau_2$ and σ^2 unknown. Moreover, the D_i are independent.

Hence, to test $H_0 : \tau_1 = \tau_2$ we may simply use the one-sample t-test on the differences

Car	1	2	3	4	5
Difference	0.4	0.4	0.5	0.6	0.5

We find a sample mean of 0.48 and a standard deviation of 0.0837. The corresponding t-value is $0.48/(0.0837/4) \approx 23$ which gives a very small p-value. Hence we reject the hypothesis that the mean tyre wear is the same.

Remark 10.5 The completely randomized design is a k-sample generalisation of the model for the 2-sample t-test. Similarly, the randomized block design in a b-block generalisation of the model for the paired t-test.

Exercise 10.4 Proof that for $t = 2$ the paired t-test and the F-test for the randomized block design are *equivalent*.

Index