Lecture notes

Probability Theory & Statistics

Jørgen Larsen

2006

English version 2008/9

Roskilde University

This document was typeset whith Kp-Fonts using KOMA-Script and LaTeX. The drawings are produced with METAPOST.

October 2010.

Contents

Preface

Probability theory and statistics are two subject fields that either can be studied on their own conditions, or can be dealt with as ancillary subjects, and the way these subjects are taught should reflect which of the two cases that is actually the case. The present set of lecture notes is directed at students who have to do a certain amount of probability theory and statistics as part of their general mathematics education, and accordingly it is directed neither at students specialising in probability and/or statistics nor at students in need of statistics as an ancillary subject. — Over a period of years several draft versions of these notes have been in use at the mathematics education at Roskilde University.

When preparing a course in probability theory it seems to be an ever unresolved issue how much (or rather how little) emphasis to put on a general axiomatic presentation. If the course is meant to be a general introduction for students who cannot be supposed to have any great interests (or qualifications) in the subject, then there is no doubt: don't mention Kolmogorov's axioms at all (or possibly in a discreet footnote). Things are different with a course that is part of a general mathematics education; here it would be perfectly meaningful to study the mathematical formalism at play when trying to establish a set of building blocks useful for certain modelling problems (the modelling of random phenomena), and hence it would indeed be appropriate to study the basis of the mathematics concerned.

Probability theory is definitely a proper mathematical discipline, and statistics certainly makes use of a number of different sub-areas of mathematics. Both fields are, however, as for their mathematical content, organised in a rather different way from "normal" mathematics (or at least from the way it is presented in math education), because they primarily are designed to be able to *do* certain things, e.g. prove the Law of Large Numbers and the Central Limit Theorem, and only to a lesser degree to conform to the mathematical community's current opinions about how to present mathematics; and if, for instance, you believe probability theory (the measure theoretic probability theory) simply to be a special case of measure theory in general, you will miss several essential points as to what probability theory is and is about.

Moreover, anyone who is attempting to acquaint himself with probability

theory and statistics soon realises that that it is a project that heavily involves concepts, methods and results from a number of rather different "mainstream" mathematical topics, which may be one reason why probability theory and statistics often are considered quite difficult or even incomprehensible.

Following the traditional pattern the presentation is divided into two parts, a probability part and a statistics part. The two parts are rather different as regards style and composition.

The probability part introduces the fundamental concepts and ideas, illustrated by the usual examples, the curriculum being kept on a tight rein throughout. Firstly, a thorough presentation of axiomatic probability theory on finite sample spaces, using Kolmogorov's axioms restricted to the finite case; by doing so the amount of mathematical complications can be kept to a minimum without sacrificing the possibility of proving the various results stated. Next, the theory is extended to countable sample spaces, specifically the sample space $\mathbb{N}_0$ equipped with the σ algebra consisting of all subsets of the sample space; it is still possible to prove "all" theorems without too much complicated mathematics (you will need a working knowledge of infinite series), and yet you will get some idea of the problems related to infinite sample spaces. However, the formal axiomatic approach is not continued in the chapter about continuous distributions on $\mathbb{R}$ and $\mathbb{R}^n$, i.e. distributions having a density function, and this chapter states many theorems and propositions without proofs (for one thing, the theorem about transformation of densities, the proof of which rightly is to be found in a mathematical analysis course). Part I finishes with two chapters of a somewhat different kind, one chapter about generating functions (including a few pages about branching processes), and one chapter containing some hints about how and why probabilists want to deal with probability measures on general sample spaces.

Part II introduces the classical likelihood-based approach to mathematical statistics: statistical models are made from the standard distributions and have a modest number of parameters, estimators are usually maximum likelihood estimators, and hypotheses are tested using the likelihood ratio test statistic. The presentation is arranged with a chapter about the notion of a statistical model, a chapter about estimation and a chapter about hypothesis testing; these chapters are provided with a number of examples showing how the theory performs when applied to specific models or classes of models. Then follows three extensive examples, showing the general theory in use. Part II is completed with an introduction to the theory of linear normal models in a linear algebra formulation, in good keeping with tradition in Danish mathematical statistics.

JL

Part I

Probability Theory

Introduction

Probability theory is a discipline that deals with mathematical formalisations of the everyday concepts of chance and randomness and related concepts. At first one might wonder how it could be possible at all to make mathematical formalisations of anything like chance and randomness, is it not a fundamental characteristic of these concepts that you cannot give any kinds of exact description relating to them? Not quite so. It seems to be an empirical fact that at least some kinds of random experiments and random phenomena will display a considerable degree of regularity when repeated a large number of times, this applies to experiments such as tossing coins or dice, roulette and other kinds of chance games. In order to treat such matters in a rigorous way we have to introduce some mathematical ideas and notations, which at first may seem a little vague, but later on will be given a precise mathematical meaning (hopefully not too far from the common usage of the same terms).

When run, the random experiment will produce a result of some kind. The rolling of a die, for instance, will result in the die falling with a certain, yet unpredictable, number of pips uppermost. The result of the random experiment is called an *outcome*. The set of all possible outcomes is called the *sample space*.

Probabilities are real numbers expressing certain quantitative characteristics of the random experiment. A simple example of a probability statement could be "the probability that the die will fall with the number 5 uppermost is $1/6$"; another example could be "the probability of a white Christmas is $1/20$". What is the meaning of such statements? Some people will claim that probability statements are to be interpreted as statements about predictions about the outcome of some future phenomenon (such as a white Christmas). Others will claim that probability statements describe the relative frequency of the occurrence of the outcome in question, when the random experiment (e.g. the rolling of the die) is repeated over and over. The way probability theory is formalised and axiomatised is very much inspired by an interpretation of probability as *relative frequency in the long run*, but the formalisation is not tied down to this single interpretation.

Probability theory makes use of notation from simple set theory, although with slightly different names, see the overview on the next page. A *probability*, or to

Some notions from set theory together with their counterparts in probabilility theory.

Typical notation	Probability theory	Set theory
Ω	sample space; the certain event	universal set
$\varnothing$	the impossible event	the empty set
ω	outcome	member of Ω
A	event	subset of Ω
$A \cap B$	both A and B	intersection of A and B
$A \cup B$	either A or B	union of A and B
$A \setminus B$	A but not B	set difference
A^{c}	the opposite event of A	the complement of A, i.e. $\Omega \setminus A$

be precise, a *probability measure* will be defined as a map from a certain domain into the set of reals. What this domain should be may not be quite obvious; one might suggest that it should simply be the entire sample space, but it turns out that this will cause severe problems in the case of an uncountable sample space (such as the set of reals). It has turned out that the proper way to build a useful theory, is to assign probabilities to *events*, that is, to certain kinds of subsets of the sample space. Accordingly, a probability measure will be a special kind of map assigning real numbers to certain kinds of subsets of the sample space.

1 Finite Probability Spaces

This chapter treats probabilities on finite sample spaces. The definitions and the theory are simplified versions of the general setup, thus hopefully giving a good idea also of the general theory without too many mathematical complications.

1.1 Basic definitions

Definition 1.1: Probability space over a finite set

A probability space over a finite set is a triple $(\Omega, \mathscr{F}, \mathrm{P})$ *consisting of*

1. *a sample space* Ω *which is a non-empty finite set,*
2. *the set* $\mathscr{F}$ *of all subsets of* Ω*,*
3. *a probability measure on* $(\Omega, \mathscr{F})$*, that is, a map* $\mathrm{P} : \mathscr{F} \to \mathbb{R}$ *which is*
   - *positive:* $\mathrm{P}(A) \geq 0$ *for all* $A \in \mathscr{F}$*,*
   - *normed:* $\mathrm{P}(\Omega) = 1$*, and*
   - *additive: if* $A_1, A_2, \ldots, A_n \in \mathscr{F}$ *are mutually disjoint, then*

$$\mathrm{P}\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} \mathrm{P}(A_i).$$

The elements of Ω *are called outcomes, the elements of* $\mathscr{F}$ *are called events.*

Here are two simple examples. Incidentally, they also serve to demonstrate that there indeed exist mathematical objects satisfying Definition 1.1:

Example 1.1: Uniform distribution

Let Ω be a finite set with n elements, $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, let $\mathscr{F}$ be the set of subsets of Ω, and let P be given by $\mathrm{P}(A) = \frac{1}{n}\#A$, where $\#A$ means "the number of elements of A". Then $(\Omega, \mathscr{F}, \mathrm{P})$ satisfies the conditions for being a probability space (the additivity follows from the additivity of the number function). The probability measure P is called the *uniform distribution* on Ω, since it distributes the "probability mass" uniformly over the sample space.

Example 1.2: One-point distribution

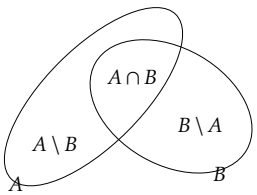
Let Ω be a finite set, and let $\omega_0 \in \Omega$ be a selected point. Let $\mathscr{F}$ be the set of subsets of Ω, and define P by $\mathrm{P}(A) = 1$ if $\omega_0 \in A$ and $\mathrm{P}(A) = 0$ otherwise. Then $(\Omega, \mathscr{F}, \mathrm{P})$ satisfies the conditions for being a probability space. The probability measure P is called the *one-point distribution* at ω_0, since it assigns all of the probability mass to this single point.

Then we can prove a number of results:

Lemma 1.1
For any events A and B of the probability space $(\Omega, \mathscr{F}, \mathrm{P})$ we have

1. $\mathrm{P}(A) + \mathrm{P}(A^{\mathrm{c}}) = 1$.
2. $\mathrm{P}(\varnothing) = 0$.
3. *If $A \subseteq B$, then $\mathrm{P}(B \setminus A) = \mathrm{P}(B) - \mathrm{P}(A)$ and hence $\mathrm{P}(A) \leq \mathrm{P}(B)$ (and so P is increasing).*
4. $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$.

Proof
Re. 1: The two events A and A^{c} are disjoint and their union is Ω; then by the axiom of additivity $\mathrm{P}(A) + \mathrm{P}(A^{\mathrm{c}}) = \mathrm{P}(\Omega)$, and $\mathrm{P}(\Omega)$ equals 1.
Re. 2: Apply what has just been proved to the event $A = \Omega$ to get $\mathrm{P}(\varnothing) = 1 - \mathrm{P}(\Omega) = 1 - 1 = 0$.
Re. 3: The events A and $B \setminus A$ are disjoint with union B, and therefore $\mathrm{P}(B) = \mathrm{P}(A) + \mathrm{P}(B \setminus A)$; since $\mathrm{P}(B \setminus A) \geq 0$, we see that $\mathrm{P}(A) \leq \mathrm{P}(B)$.
Re. 4: The three events $A \setminus B$, $B \setminus A$ and $A \cap B$ are mutually disjoint with union $A \cup B$. Therefore



$$\begin{aligned}\mathrm{P}(A \cup B) &= \mathrm{P}(A \setminus B) + \mathrm{P}(B \setminus A) + \mathrm{P}(A \cap B) \\ &= \Big(\mathrm{P}(A \setminus B) + \mathrm{P}(A \cap B)\Big) + \Big(\mathrm{P}(B \setminus A) + \mathrm{P}(A \cap B)\Big) - \mathrm{P}(A \cap B) \\ &= \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B).\end{aligned}$$

□

All presentations of probability theory use coin-tossing and die-rolling as examples of chance phenomena:

Example 1.3: Coin-tossing
Suppose we toss a coin once and observe whether a head or a tail turns up.

The sample space is the two-point set $\Omega = \{\text{head}, \text{tail}\}$. The set of events is $\mathscr{F} = \{\Omega, \{\text{head}\}, \{\text{tail}\}, \varnothing\}$. The probability measure corresponding to a fair coin, i.e. a coin where heads and tails have an equal probability of occurring, is the uniform distribution on Ω:

$$\mathrm{P}(\Omega) = 1, \quad \mathrm{P}(\{\text{head}\}) = 1/2, \quad \mathrm{P}(\{\text{tail}\}) = 1/2, \quad \mathrm{P}(\varnothing) = 0.$$

Example 1.4: Rolling a die
Suppose we roll a die once and observe the number turning up.

The sample space is the set $\Omega = \{1, 2, 3, 4, 5, 6\}$. The set $\mathscr{F}$ of events is the set of subsets of Ω (giving a total of $2^6 = 64$ different events). The probability measure P corresponding to a fair die is the uniform distribution on Ω.

This implies for instance that the probability of the event $\{3,6\}$ (the number of pips is a multiple of 3) is $\mathrm{P}(\{3,6\}) = 2/6$, since this event consists of two outcomes out of a total of six possible outcomes.

Example 1.5: Simple random sampling without replacement
From a box (or urn) containing b black and w white balls we draw a k-sample, that is, a subset of k elements (k is assumed to be not greater than $b+w$). This is done as *simple random sampling without replacement*, which means that all the $\binom{b+w}{k}$ different subsets with exactly k elements (cf. the definition of binomial coefficientes page 30) have the same probability of being selected. — Thus we have a uniform distribution on the set Ω consisting of all these subsets.

So the probability of the event "exactly x black balls" equals the number of k-samples having x black and $k-x$ white balls divided by the total number of k-samples, that is $\binom{b}{x}\binom{w}{k-x}\Big/\binom{b+w}{k}$. See also page 31 and Proposition 1.13.

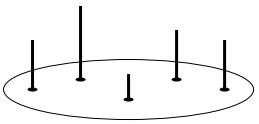
Point probabilities

The reader may wonder why the concept of probability is mathematified in this rather complicated way, why not just simply have a function that maps every single outcome to the probability for this outcome? As long as we are working with finite (or countable) sample spaces, that would indeed be a feasible approach, but with an uncountable sample space it would never work (since an uncountable infinity of positive numbers cannot add up to a finite number).

However, as long as we are dealing with the finite case, the following definition and theorem are of interest.

Definition 1.2: Point probabilities
Let $(\Omega, \mathscr{F}, \mathrm{P})$ *be a probability space over a finite set* Ω. *The function*

$$\begin{aligned} p : \Omega &\longrightarrow [0\,;1] \\ \omega &\longmapsto \mathrm{P}(\{\omega\}) \end{aligned}$$

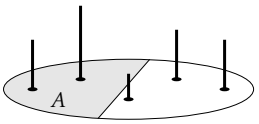


is the point probability function (or the probability mass function) of P.

Point probabilities are often visualised as "probability bars".

Proposition 1.2
If p *is the point probability function of the probability measure* P, *then for any event* A *we have* $\mathrm{P}(A) = \sum_{\omega\in A} p(\omega)$.

Proof
Write A as a disjoint union of its singletons (one-point sets) and use additivity:
$\mathrm{P}(A) = \mathrm{P}\Big(\bigcup_{\omega\in A}\{\omega\}\Big) = \sum_{\omega\in A}\mathrm{P}(\{\omega\}) = \sum_{\omega\in A} p(\omega)$. □



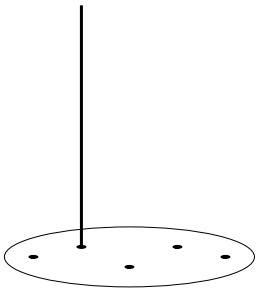
Remarks: It follows from the theorem that different probability measures must have different point probability functions. It also follows that p adds up to unity, i.e. $\sum_{\omega\in\Omega} p(\omega) = 1$; this is seen by letting $A = \Omega$.

Proposition 1.3
If $p : \Omega \to [0\,;1]$ adds up to unity, i.e. $\sum_{\omega\in\Omega} p(\omega) = 1$, then there is exactly one probability measure P on Ω which has p as its point probability function.

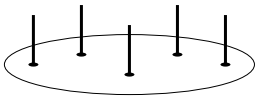
Proof
We can define a function $\mathrm{P} : \mathscr{F} \to [0\,;+\infty[$ as $\mathrm{P}(A) = \sum_{\omega\in A} p(\omega)$. This function is positive since $p \geq 0$, and it is normed since p adds to unity. Furthermore it is additive: for disjoint events $A_1, A_2, \ldots, A_n$ we have

$$\begin{aligned}\mathrm{P}(A_1 \cup A_2 \cup \ldots \cup A_n) &= \sum_{\omega\in A_1\cup A_2\cup\ldots\cup A_n} p(\omega) \\ &= \sum_{\omega\in A_1} p(\omega) + \sum_{\omega\in A_2} p(\omega) + \ldots + \sum_{\omega\in A_n} p(\omega) \\ &= \mathrm{P}(A_1) + \mathrm{P}(A_2) + \ldots + \mathrm{P}(A_n),\end{aligned}$$

where the second equality follows from the associative law for simple addition. Thus P meets the conditions for being a probability measure. By construction p is the point probability function of P, and by the remark to Proposition 1.2 only one probability measure can have p as its point probability function. □


A one-point distribution

Example 1.6
The point probability function of the one-point distribution at ω_0 (cf. Example 1.2) is given by $p(\omega_0) = 1$, and $p(\omega) = 0$ when $\omega \neq \omega_0$.


A uniform distribution

Example 1.7
The point probability function of the uniform distribution on $\{\omega_1, \omega_2, \ldots, \omega_n\}$ (cf. Example 1.1) is given by $p(\omega_i) = 1/n$, $i = 1, 2, \ldots, n$.

Conditional probabilities and distributions

Often one wants to calculate the probability that some event A occurs *given* that some other event B is known to occur — this probability is known as the *conditional probability* of A given B.

Definition 1.3: Conditional probability
Let $(\Omega, \mathscr{F}, \mathrm{P})$ be a probability space, let A and B be events, and suppose that $\mathrm{P}(B) > 0$. The number

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$$

is called the conditional probability of A given B.

Example 1.8
Two coins, a 10-krone piece and a 20-krone piece, are tossed simultaneously. What is the probability that the 10-krone piece falls heads, given that at least one of the two coins fall heads?

The standard model for this chance experiment tells us that there are four possible outcomes, and, recording an outcome as

("how the 10-krone falls", "how the 20-krone falls"),

the sample space is

$$\Omega = \{(\text{tails}, \text{tails}), (\text{tails}, \text{heads}), (\text{heads}, \text{tails}), (\text{heads}, \text{heads})\}.$$

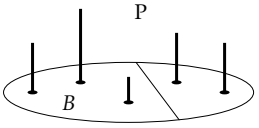
Each of these four outcomes is supposed to have probability $1/4$. The conditioning event B (at least one heads) and the event in question A (the 10-krone piece falls heads) are

$$B = \{(\text{tails}, \text{heads}), (\text{heads}, \text{tails}), (\text{heads}, \text{heads})\} \quad \text{and}$$
$$A = \{(\text{heads}, \text{tails}), (\text{heads}, \text{heads})\},$$

so the conditional probability of A given B is

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{2/4}{3/4} = 2/3.$$

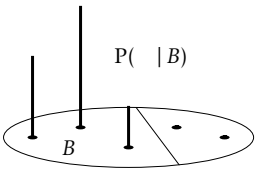
From Definition 1.3 we immediately get



Proposition 1.4
If A and B are events, and $\mathrm{P}(B) > 0$, *then* $\mathrm{P}(A \cap B) = \mathrm{P}(A \mid B)\,\mathrm{P}(B)$.

Definition 1.4: Conditional distribution
Let $(\Omega, \mathscr{F}, \mathrm{P})$ *be a probability space, and let B be an event such that* $\mathrm{P}(B) > 0$. *The function*

$$\begin{aligned} \mathrm{P}(\quad \mid B) : \mathscr{F} &\longrightarrow [0\,;1] \\ A &\longmapsto \mathrm{P}(A \mid B) \end{aligned}$$



is called the conditional distribution given B.

Bayes' formula

Let us assume that Ω can be writtes as a disjoint union of k events $B_1, B_2, \ldots, B_k$ (in other words, $B_1, B_2, \ldots, B_k$ is a *partition* of Ω). We also have an event A. Further it is assumed that we beforehand (or *a priori*) know the probabilities $\mathrm{P}(B_1), \mathrm{P}(B_2), \ldots, \mathrm{P}(B_k)$ for each of the events in the partition, as well as all conditional probabilities $\mathrm{P}(A \mid B_j)$ for A given a B_j. The problem is find the so-called

posterior probabilities, that is, the probabilities $\mathrm{P}(B_j \mid A)$ for the B_j, given that the event A is known to have occurred.

(An example of such a situation could be a doctor attempting to make a medical diagnosis: A would be the set of symptoms observed on a patient, and the Bs would be the various diseases that might cause the symptoms. The doctor knows the (approximate) frequencies of the diseases, and also the (approximate) probability of a patient with disease B_i exhibiting just symptoms A, $i = 1, 2, \ldots, k$. The doctor would like to know the conditional probabilities of a patient exhibiting symptoms A to have disease B_i, $i = 1, 2, \ldots, k$.)

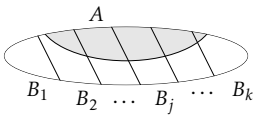
Thomas Bayes
English mathematician and theologian (1702-1761).
"Bayes' formula" (from Bayes (1763)) nowadays has a vital role in the so-called Bayesian statistics and in Bayesian networks.

Since $A = \bigcup_{i=1}^{k} A \cap B_i$ where the sets $A \cap B_i$ are disjoint, we have

$$\mathrm{P}(A) = \sum_{i=1}^{k} \mathrm{P}(A \cap B_i) = \sum_{i=1}^{k} \mathrm{P}(A \mid B_i)\,\mathrm{P}(B_i),$$

and since $\mathrm{P}(B_j \mid A) = \mathrm{P}(A \cap B_j)/\mathrm{P}(A) = \mathrm{P}(A \mid B_j)\,\mathrm{P}(B_j)/\mathrm{P}(A)$, we obtain

$$\mathrm{P}(B_j \mid A) = \frac{\mathrm{P}(A \mid B_j)\,\mathrm{P}(B_j)}{\sum_{i=1}^{k} \mathrm{P}(A \mid B_i)\,\mathrm{P}(B_i)}. \tag{1.1}$$



This formula is known as *Bayes' formula*; it shows how to calculate the posterior probabilities $\mathrm{P}(B_j \mid A)$ from the prior probabilities $\mathrm{P}(B_j)$ and the conditional probabilities $\mathrm{P}(A \mid B_j)$.

Independence

Events are independent if our probability statements pertaining to any sub-collection of them do not change when we learn about the occurrence or non-occurrence of events not in the sub-collection.

One might consider defining independence of events A and B to mean that $\mathrm{P}(A \mid B) = \mathrm{P}(A)$, which by the definition of conditional probability implies that $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$. This last formula is meaningful also when $\mathrm{P}(B) = 0$, and furthermore A and B enter in a symmetrical way. Hence we define independence of *two* events A and B to mean that $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$. — To do things properly, however, we need a definition that covers several events. The general definition is

Definition 1.5: Independent events
The events $A_1, A_2, \ldots, A_k$ are said to be independent, if it is true that for any subset $\{A_{i_1}, A_{i_2}, \ldots, A_{i_m}\}$ of these events, $\mathrm{P}\Big(\bigcap_{j=1}^{m} A_{i_j}\Big) = \prod_{j=1}^{m} \mathrm{P}(A_{i_j})$.

Note: When checking out for independence of events, it does not suffice to check whether they are pairwise independent, cf. Example 1.9. Nor does it suffice to check whether the probability for the intersection of *all* of the events equals the product of the individual events, cf. Example 1.10.

Example 1.9
Let $\Omega = \{a, b, c, d\}$ and let P be the uniform distribution on Ω. Each of the three events $A = \{a, b\}$, $B = \{a, c\}$ and $C = \{a, d\}$ has a probability of $1/2$. The events A, B and C are pairwise independent. To verify, for instance, that $\mathrm{P}(B \cap C) = \mathrm{P}(B)\mathrm{P}(C)$, note that $\mathrm{P}(B \cap C) = \mathrm{P}(\{a\}) = 1/4$ and $\mathrm{P}(B)\mathrm{P}(C) = 1/2 \cdot 1/2 = 1/4$.

On the other hand, the three events are *not* independent: $\mathrm{P}(A \cap B \cap C) \neq \mathrm{P}(A)\mathrm{P}(B)\mathrm{P}(C)$, since $\mathrm{P}(A \cap B \cap C) = \mathrm{P}(\{a\}) = 1/4$ and $\mathrm{P}(A)\mathrm{P}(B)\mathrm{P}(C) = 1/8$.

Example 1.10
Let $\Omega = \{a, b, c, d, e, f, g, h\}$ and let P be the uniform distribution on Ω. The three events $A = \{a, b, c, d\}$, $B = \{a, e, f, g\}$ and $C = \{a, b, c, e\}$ then all have probability $1/2$. Since $A \cap B \cap C = \{a\}$, it is true that $\mathrm{P}(A \cap B \cap C) = \mathrm{P}(\{a\}) = 1/8 = \mathrm{P}(A)\mathrm{P}(B)\mathrm{P}(C)$, but the events A, B and C are *not* independent; we have, for instance, that $\mathrm{P}(A \cap B) \neq \mathrm{P}(A)\mathrm{P}(B)$ (since $\mathrm{P}(A \cap B) = \mathrm{P}(\{a\}) = 1/8$ and $\mathrm{P}(A)\mathrm{P}(B) = 1/2 \cdot 1/2 = 1/4$).

Independent experiments; product space

Very often one wants to model chance experiments that can be thought of as compound experiments made up of a number of separate sub-experiments (the experiment "rolling five dice once", for example, may be thought of as made up of five copies of the experiment "rolling one die once"). If we are given models for the sub-experiments, and if the sub-experiments are assumed to be independent (in a sense to be made precise), it is fairly straightforward to write up a model for the compound experiment.

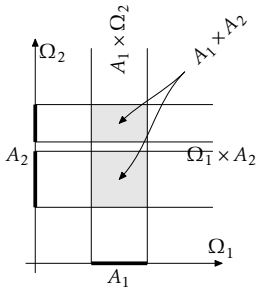
For convenience consider initially a compound experiment consisting of just *two* sub-experiment I and II, and let us assume that they can be modelled by the probability spaces $(\Omega_1, \mathscr{F}_1, \mathrm{P}_1)$ and $(\Omega_2, \mathscr{F}_2, \mathrm{P}_2)$, respectively.

We seek a probability space $(\Omega, \mathscr{F}, \mathrm{P})$ that models the compound experiment of carrying out both I and II. It seems reasonable to write the outcomes of the compound experiment as (ω_1, ω_2) where $\omega_1 \in \Omega_1$ and $\omega_2 \in \Omega_2$, that is, Ω is the product set $\Omega_1 \times \Omega_2$, and $\mathscr{F}$ could then be the set of all subsets of Ω. But which P should we use?

Consider an I-event $A_1 \in \mathscr{F}_1$. Then the compound experiment event $A_1 \times \Omega_2 \in \mathscr{F}$ corresponds to the situation where the I-part of the compound experiment gives an outcome in A_1 and the II-part gives just anything, that is, you don't care about the outcome of experiment II. The two events $A_1 \times \Omega_2$ and A_1 therefore model the same real phenomenon, only in two different probability spaces, and the probability measure P that we are searching, should therefore have

the property that $\mathrm{P}(A_1 \times \Omega_2) = \mathrm{P}_1(A_1)$. Simililarly, we require that $\mathrm{P}(\Omega_1 \times A_2) = \mathrm{P}_2(A_2)$ for any event $A_2 \in \mathscr{F}_2$.

If the compound experiment is going to have the property that its components are independent, then the two events $A_1 \times \Omega_2$ and $\Omega_1 \times A_2$ have to be independent (since they relate to separate sub-experiments), and since their intersection is $A_1 \times A_2$, we must have



$$\begin{aligned}\mathrm{P}(A_1 \times A_2) &= \mathrm{P}\Big((A_1 \times \Omega_2) \cap (\Omega_1 \times A_2)\Big) \\ &= \mathrm{P}(A_1 \times \Omega_2)\,\mathrm{P}(\Omega_1 \times A_2) = \mathrm{P}_1(A_1)\,\mathrm{P}_2(A_2).\end{aligned}$$

This is a condition on P that relates to all product sets in $\mathscr{F}$. Since all singletons are product sets ($\{(\omega_1, \omega_2)\} = \{\omega_1\} \times \{\omega_2\}$), we get in particular a condition on the point probability function p for P; in an obvious notation this condition is that $p(\omega_1, \omega_2) = p_1(\omega_1)\,p_2(\omega_2)$ for all $(\omega_1, \omega_2) \in \Omega$.

Inspired by this analysis of the problem we will proceed as follows:

1. Let p_1 and p_2 be the point probability functions for P_1 and P_2.
2. Then define a function $p : \Omega \to [0\,;1]$ by $p(\omega_1, \omega_2) = p_1(\omega_1)\,p_2(\omega_2)$ for $(\omega_1, \omega_2) \in \Omega$.
3. This function p adds up to unity:

$$\begin{aligned}\sum_{\omega \in \Omega} p(\omega) &= \sum_{(\omega_1,\omega_2) \in \Omega_1 \times \Omega_2} p_1(\omega_1)\,p_2(\omega_2) \\ &= \sum_{\omega_1 \in \Omega_1} p_1(\omega_1) \sum_{\omega_2 \in \Omega_2} p_2(\omega_2) \\ &= 1 \cdot 1 = 1.\end{aligned}$$

4. According to Proposition 1.3 there is a unique probability measure P on Ω with p as its point probability function.
5. The probability measure P satisfies the requirement that $\mathrm{P}(A_1 \times A_2) = \mathrm{P}_1(A_1)\,\mathrm{P}_2(A_2)$ for all $A_1 \in \mathscr{F}_1$ and $A_2 \in \mathscr{F}_2$, since

$$\begin{aligned}\mathrm{P}(A_1 \times A_2) &= \sum_{(\omega_1,\omega_2) \in A_1 \times A_2} p_1(\omega_1)\,p_2(\omega_2) \\ &= \sum_{\omega_1 \in A_1} p_1(\omega_1) \sum_{\omega_2 \in A_2} p_2(\omega_2) = \mathrm{P}_1(A_1)\,\mathrm{P}_2(A_2).\end{aligned}$$

This solves the specified problem. — The probability measure P is called the *product* of P_1 and P_2.

One can extend the above considerations to cover situations with an arbitrary number n of sub-experiments. In a compound experiment consisting of n independent sub-experiments with point probability functions $p_1, p_2, \ldots, p_n$, the

joint point probability function is given by

$$p(\omega_1, \omega_2, \ldots, \omega_n) = p_1(\omega_1)\, p_2(\omega_2) \ldots p_n(\omega_n),$$

and for the corresponding probability measures we have

$$\mathrm{P}(A_1 \times A_2 \times \ldots \times A_n) = \mathrm{P}_1(A_1)\, \mathrm{P}_2(A_2) \ldots \mathrm{P}_n(A_n).$$

In this context p and P are called the *joint point probability function* and the *joint distribution*, and the individual p_is and P_is are called the *marginal point probability functions* and the *marginal distributions*.

Example 1.11
In the experiment where a coin is tossed once and a die is rolled once, the probability of the outcome (heads, 5 pips) is $1/2 \cdot 1/6 = 1/12$, and the probability of the event "heads and at least five pips" is $1/2 \cdot 2/6 = 1/6$. — If a coin is tossed 100 times, the probability that the last 10 tosses all give the result heads, is $(1/2)^{10} \approx 0.001$.

1.2 Random variables

One reason for mathematics being so successful is undoubtedly that it to such a great extent makes use of symbols, and a successful translation of our everyday speech about chance phenomena into the language of mathematics will therefore—among other things—rely on using an appropriate notation. In probability theory and statistics it would be extremely useful to be able to work with symbols representing "the numeric outcome that the chance experiment will provide when carried out". Such symbols are *random variables*; random variables are most frequently denoted by capital letters (such as X, Y, Z). Random variables appear in mathematical expressions just as if they were ordinary mathematical entities, e.g. $X + Y = 5$ and $Z \in B$.

Example: To describe the outcome of rolling two dice once let us introduce random variables X_1 and X_2 representing the number of pips shown by die no. 1 and die no. 2, respectively. Then $X_1 = X_2$ means that the two dice show the same number of pips, and $X_1 + X_2 \geq 10$ means that the sum of the pips is at least 10.

Even if the above may suggest the intended *meaning* of the notion of a random variable, it is certainly not a proper definition of a mathematical object. And conversely, the following definition does not convey that much of a meaning.

Definition 1.6: Random variable
Let $(\Omega, \mathscr{F}, \mathrm{P})$ *be a probability space over a finite set. A random variable* $(\Omega, \mathscr{F}, \mathrm{P})$ *is a map* X *from* Ω *to* $\mathbb{R}$, *the set of reals.*

More general, an n-dimensional random variable on $(\Omega, \mathscr{F}, \mathrm{P})$ *is a map* $\boldsymbol{X}$ *from* Ω *to* $\mathbb{R}^n$.

Below we shall study random variables in the mathematical sense. First, let us clarify how random variables enter into plain mathematical parlance: Let X be a random variable on the sample space in question. If u is a proposition about real numbers so that for every $x \in \mathbb{R}$, $u(x)$ is either true or false, then $u(X)$ is to be read as a meaningful proposition about X, and it is true if and only if the event $\{\omega \in \Omega : u(X(\omega))\}$ occurs. Thereby we can talk of the probability $\mathrm{P}(u(X))$ of $u(X)$ being true; this probability is per definition $\mathrm{P}(u(X)) = \mathrm{P}(\{\omega \in \Omega : u(X(\omega))\})$.

To write $\{\omega \in \Omega : u(X(\omega))\}$ is a precise, but rather lengthy, and often unduly detailed, specification of the event that $u(X)$ is true, and that event is therefore almost always written as $\{u(X)\}$.

Example: The proposition $X \geq 3$ corresponds to the event $\{X \geq 3\}$, or more detailed to $\{\omega \in \Omega : X(\omega) \geq 3\}$, and we write $\mathrm{P}(X \geq 3)$ which per definition is $\mathrm{P}(\{X \geq 3\})$ or more detailed $\mathrm{P}(\{\omega \in \Omega : X(\omega) \geq 3\})$.—This is extended in an obvious way to cases with several random variables.

For a subset B of $\mathbb{R}$, the proposition $X \in B$ corresponds to the event $\{X \in B\} = \{\omega \in \Omega : X(\omega) \in B\}$. The probability of the proposition (or the event) is $\mathrm{P}(X \in B)$. Considered as a function of B, $\mathrm{P}(X \in B)$ satisfies the conditions for being a probability measure on $\mathbb{R}$; statisticians and probabilists denote this measure the *distribution of* X, whereas mathematicians will speak of a *transformed measure*.—Actually, the preceding is not entirely correct: at present we cannot talk about probability measures on $\mathbb{R}$, since we have reached only probabilities on finite sets. Therefore we have to think of that probability measure with the two names as a probability measure on the finite set $X(\Omega) \subset \mathbb{R}$.

Denote for a while the transformed probability measure $X(\mathrm{P})$; then $X(\mathrm{P})(B) = \mathrm{P}(X \in B)$ where B is a subset of $\mathbb{R}$. This innocuous-looking formula shows that all probability statements regarding X can be rewritten as probability statements solely involving (subsets of) $\mathbb{R}$ and the probability measure $X(\mathrm{P})$ on (the finite subset $X(\Omega)$ of) $\mathbb{R}$. Thus we may forget all about the original probability space Ω.

Example 1.12
Let $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$, and let P be the probability measure on Ω given by the point probabilities $p(\omega_1) = 0.3$, $p(\omega_2) = 0.1$, $p(\omega_3) = 0.4$, $p(\omega_4) = 0$, and $p(\omega_5) = 0.2$. We can define a random variable $X : \Omega \to \mathbb{R}$ by the specification $X(\omega_1) = 3$, $X(\omega_2) = -0.8$, $X(\omega_3) = 4.1$, $X(\omega_4) = -0.8$, and $X(\omega_5) = -0.8$.

The range of the map X is $X(\Omega) = \{-0.8, 3, 4.1\}$, and the distribution of X is the probability measure on $X(\Omega)$ given by the point probabilities

$$\begin{aligned}
\mathrm{P}(X = -0.8) &= \mathrm{P}(\{\omega_2, \omega_4, \omega_5\}) = p(\omega_2) + p(\omega_4) + p(\omega_5) = 0.3,\\
\mathrm{P}(X = 3) &= \mathrm{P}(\{\omega_1\}) = p(\omega_1) = 0.3,\\
\mathrm{P}(X = 4.1) &= \mathrm{P}(\{\omega_3\}) = p(\omega_3) = 0.4.
\end{aligned}$$

Example 1.13: Continuation of Example 1.12
Now introduce two more random variables Y and Z, leading to the following situation:

ω	$p(\omega)$	$X(\omega)$	$Y(\omega)$	$Z(\omega)$
ω_1	0.3	3	−0.8	3
ω_2	0.1	−0.8	3	−0.8
ω_3	0.4	4.1	4.1	4.1
ω_4	0	−0.8	4.1	4.1
ω_5	0.2	−0.8	3	−0.8

It appears that X, Y and Z are different, since for example $X(\omega_1) \neq Y(\omega_1)$, $X(\omega_4) \neq Z(\omega_4)$ and $Y(\omega_1) \neq Z(\omega_1)$, although X, Y and Z do have the same range $\{-0.8, 3, 4.1\}$. Straightforward calculations show that

$$\begin{aligned} \mathrm{P}(X = -0.8) &= \mathrm{P}(Y = -0.8) = \mathrm{P}(Z = -0.8) = 0.3, \\ \mathrm{P}(X = 3) &= \mathrm{P}(Y = 3) = \mathrm{P}(Z = 3) = 0.3, \\ \mathrm{P}(X = 4.1) &= \mathrm{P}(Y = 4.1) = \mathrm{P}(Z = 4.1) = 0.4, \end{aligned}$$

that is, X, Y and Z have the *same distribution*. It is seen that $\mathrm{P}(X \neq Z) = \mathrm{P}(\{\omega_4\}) = 0$, so $X = Z$ with probability 1, and $\mathrm{P}(X = Y) = \mathrm{P}(\{\omega_3\}) = 0.4$.

Being a distribution on a finite set, the distribution of a random variable X can be described (and illustrated) by its point probabilities. And since the distribution "lives" on a subset of the reals, we can also describe it by its distribution function.

Definition 1.7: Distribution function
The distribution function F of a random variable X is the function

$$\begin{aligned} F : \mathbb{R} &\longrightarrow [0\,;1] \\ x &\longmapsto \mathrm{P}(X \leq x). \end{aligned}$$

A distribution function has certain properties:

Lemma 1.5
If the random variable X has distribution function F, then

$$\begin{aligned} \mathrm{P}(X \leq x) &= F(x), \\ \mathrm{P}(X > x) &= 1 - F(x), \\ \mathrm{P}(a < X \leq b) &= F(b) - F(a), \end{aligned}$$

for any real numbers x and $a < b$.

Proof
The first equation is simply the definitionen of F. The other two equations follow from items 1 and 3 in Lemma 1.1 (page 14). □

Increasing functions
Let $F : \mathbb{R} \to \mathbb{R}$ be an increasing function, that is, $F(x) \le F(y)$ whenever $x \le y$.
Then at each point x, F has a *limit from the left*

$$F(x-) = \lim_{h \searrow 0} F(x-h)$$

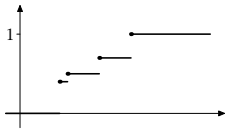
and a *limit from the right*

$$F(x+) = \lim_{h \searrow 0} F(x+h),$$

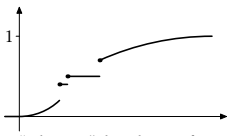
and

$$F(x-) \le F(x) \le F(x+).$$

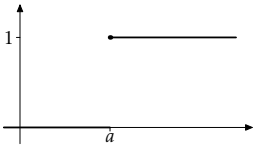
If $F(x-) \ne F(x+)$, x is a point of discontinuity of F (otherwise, x is a point of continuity of F).



"arbitrary" distribution function, finite sample space



an "arbitrary" distribution function



Proposition 1.6
The distribution function F of a random variable X has the following properties:

1. *It is non-decreasing, i.e. if $x \le y$, then $F(x) \le F(y)$.*
2. $\lim_{x \to -\infty} F(x) = 0$ *and* $\lim_{x \to +\infty} F(x) = 1$.
3. *It is continuous from the right, i.e. $F(x+) = F(x)$ for all x.*
4. $\mathrm{P}(X = x) = F(x) - F(x-)$ *at each point x.*
5. *A point x is a discontinuity point of F, if and only if* $\mathrm{P}(X = x) > 0$.

Proof
1: If $x \le y$, then $F(y) - F(x) = \mathrm{P}(x < X \le y)$ (Lemma 1.5), and since probabilities are always non-negative, this implies that $F(x) \le F(y)$.
2: Since X takes only a finite number of values, there exist two numbers $x_{\min}$ and $x_{\max}$ such that $x_{\min} < X(\omega) < x_{\max}$ for all ω. Then $F(x) = 0$ for all $x < x_{\min}$, and $F(x) = 1$ for all $x > x_{\max}$.
3: Since X takes only a finite number of values, then for each x it is true that for all sufficiently small numbers $h > 0$, X does not take values in the interval $]x\,;x+h]$; this means that the events $\{X \le x\}$ and $\{X \le x+h\}$ are identical, and so $F(x) = F(x+h)$.
4: For a given number x consider the part of range of X that is to the left of x, that is, the set $X(\Omega) \cap]-\infty\,;x[$. If this set is empty, take $a = -\infty$, and otherwise take a to be the largest element in $X(\Omega) \cap]-\infty\,;x[$.

Since $X(\omega)$ is outside of $]a\,;x[$ for all ω, it is true that for every $x^* \in]a\,;x[$, the events $\{X = x\}$ and $\{x^* < X \le x\}$ are identical, so that $\mathrm{P}(X = x) = \mathrm{P}(x^* < X \le x) = \mathrm{P}(X \le x) - \mathrm{P}(X \le x^*) = F(x) - F(x^*)$. For $x^* \nearrow x$ we get the desired result.
5: Follows from 4 and 3. □

Random variables will appear again and again, and the reader will soon be handling them quite easily.—Here are some simple (but not unimportant) examples of random variables.

Example 1.14: Constant random variable
The simplest type of random variables is the random variables that always take the same value, that is, random variables of the form $X(\omega) = a$ for all $\omega \in \Omega$; here a is some real number. The random variable $X = a$ has distribution function

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

Actually, we could replace the condition "$X(\omega) = a$ for all $\omega \in \Omega$" with the condition "$X = a$ with probability 1" or $\mathrm{P}(X = a) = 1$; the distribution function would remain unchanged.

Example 1.15: 01*-variable*
The second simplest type of random variables must be random variables that takes only *two* different values. Such variables occur in models for binary experiments, i.e. experiments with two possible outcomes (such as Success/Failure or Head/Tails). To simplify matters we shall assume that the two values are 0 and 1, whence the name 01*-variables* for such random variables (another name is *Bernoulli variables*).

The distribution of a 01-variable can be expressed using a parameter p which is the probability of the value 1:

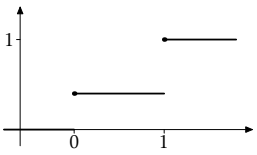
$$\mathrm{P}(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

or

$$\mathrm{P}(X = x) = p^x (1-p)^{1-x}, \quad x = 0, 1. \tag{1.2}$$

The distribution function of a 01-variable with parameter p is

$$F(x) = \begin{cases} 1 & \text{if } 1 \le x \\ 1 - p & \text{if } 0 \le x < 1 \\ 0 & \text{if } x < 0. \end{cases}$$



Example 1.16: Indicator function
If A is an event (a subset of Ω), then its *indicator function* is the function

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \in A^{\mathrm{c}}. \end{cases}$$

This indicator function is a 01-variable with parameter $p = \mathrm{P}(A)$.

Conversely, if X is a 01-variable, then X is the indicator funtion of the event $A = X^{-1}(\{1\}) = \{\omega : X(\omega) = 1\}$.

Example 1.17: Uniform random variable
If $x_1, x_2, \ldots, x_n$ are n different real number and X a random variable that takes each of these numbers with the same probability, i.e.

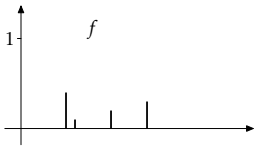
$$\mathrm{P}(X = x_i) = \frac{1}{n}, \quad i = 1, 2, \ldots, n,$$

then X is said to be uniformly distributed on the set $\{x_1, x_2, \ldots, x_n\}$.

The distribution funtion is not particularly well suited for giving an informative visualisation of the distribution of a random variable, as you may have realised from the above examples. A far better tool is the probability funtion. The probability funtion for the random variable X is the function $f : x \mapsto \mathrm{P}(X = x)$, considered as a function defined on the range of X (or some reasonable superset of it). — In situations that are modelled using a finite probability space, the random variables of interest are often variables taking values in the non-negative integers; in that case the probability function can be considered as defined either on the actual range of X, or on the set $\mathbb{N}_0$ of non-negative integers.

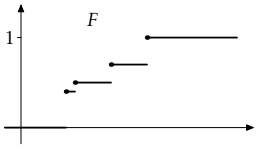
Definition 1.8: Probability function
The probability function of a random variablel X is the function

$$f : x \mapsto \mathrm{P}(X = x).$$



Proposition 1.7
For a given distribution the distribution function F and the probability function f are related as follows:

$$f(x) = F(x) - F(x-),$$
$$F(x) = \sum_{z:z\leq x} f(z).$$



Proof
The expression for f is a reformulation of item 4 in Proposition 1.6. The expression for F follows from Proposition 1.2 page 15. □

Independent random variables

Often one needs to study more than one random variable at a time.

Let $X_1, X_2, \ldots, X_n$ be random variables on the same finite probability space. Their *joint probability function* is the function

$$f(x_1, x_2, \ldots, x_n) = \mathrm{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$

For each j the function

$$f_j(x_j) = \mathrm{P}(X_j = x_j)$$

is called the *marginal probability function* of X_j; more generally, if we have a non-trivial set of indices $\{i_1, i_2, \ldots, i_k\} \subset \{1, 2, \ldots, n\}$, then

$$f_{i_1,i_2,\ldots,i_k}(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) = \mathrm{P}(X_{i_1} = x_{i_1}, X_{i_2} = x_{i_2}, \ldots, X_{i_k} = x_{i_k})$$

is the marginal probability function of $X_{i_1}, X_{i_2}, \ldots, X_{i_k}$.

Previously we defined independence of events (page 18f); now we can define independence of random variables:

Definition 1.9: Independent random variables
Let $(\Omega, \mathscr{F}, \mathrm{P})$ be a probability space over a finite set. Then the random variables $X_1, X_2, \ldots, X_n$ on $(\Omega, \mathscr{F}, \mathrm{P})$ are said to be independent, if for any choice of subsets $B_1, B_2, \ldots, B_n$ of $\mathbb{R}$, the events $\{X_1 \in B_1\}, \{X_2 \in B_2\}, \ldots, \{X_n \in B_n\}$ are independent.

A simple and clear criterion for independence is

Proposition 1.8
The random variables $X_1, X_2, \ldots, X_n$ are independent, if and only if

$$\mathrm{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} \mathrm{P}(X_i = x_i) \tag{1.3}$$

for all n-tuples $x_1, x_2, \ldots, x_n$ of real numbers such that x_i belongs to the range of X_i, $i = 1, 2, \ldots, n$.

Corollary 1.9
The random variables $X_1, X_2, \ldots, X_n$ are independent, if and only if their joint probability function equals the product of the marginal probability functions:

$$f_{12\ldots n}(x_1, x_2, \ldots, x_n) = f_1(x_1)\, f_2(x_2) \ldots f_n(x_n).$$

Proof of Proposition 1.8
To show the "only if" part, let us assume that $X_1, X_2, \ldots, X_n$ are independent according to Definition 1.9. We then have to show that (1.3) holds. Let $B_i = \{x_i\}$, then $\{X_i \in B_i\} = \{X_i = x_i\}$ $(i = 1, 2, \ldots, n)$, and

$$\mathrm{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \mathrm{P}\Big(\bigcap_{i=1}^{n} \{X_i = x_i\}\Big) = \prod_{i=1}^{n} \mathrm{P}(X_i = x_i);$$

since $x_1, x_2, \ldots, x_n$ are arbitrary, this proves (1.3).

Next we shall show the "if" part, that is, under the assumption that equation (1.3) holds, we have to show that for arbitrary subsets $B_1, B_2, \ldots, B_n$ of $\mathbb{R}$, the events $\{X_1 \in B_1\}, \{X_2 \in B_2\}, \ldots, \{X_n \in B_n\}$ are independent. In general, for an arbitrary subset B of $\mathbb{R}^n$,

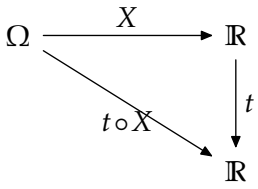
$$\mathrm{P}\Big((X_1, X_2, \ldots, X_n) \in B\Big) = \sum_{(x_1, x_2, \ldots, x_n) \in B} \mathrm{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$

Then apply this to $B = B_1 \times B_2 \times \cdots \times B_n$ and make use of (1.3):

$$\begin{aligned}
\mathrm{P}\Big(\bigcap_{i=1}^{n} \{X_i \in B_i\}\Big) &= \mathrm{P}\Big((X_1, X_2, \ldots, X_n) \in B\Big) \\
&= \sum_{x_1 \in B_1} \sum_{x_2 \in B_2} \cdots \sum_{x_n \in B_n} \mathrm{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\
&= \sum_{x_1 \in B_1} \sum_{x_2 \in B_2} \cdots \sum_{x_n \in B_n} \prod_{i=1}^{n} \mathrm{P}(X_i = x_i) = \prod_{i=1}^{n} \sum_{x_i \in B_i} \mathrm{P}(X_i = x_i) \\
&= \prod_{i=1}^{n} \mathrm{P}(X_i \in B_i),
\end{aligned}$$

showing that $\{X_1 \in B_1\}, \{X_2 \in B_2\}, \ldots, \{X_n \in B_n\}$ are independent events. □

If the individual Xs have identical probability functions and thus identical probability distributions, they are said to be *identically distributed*, and if they furthermore are independent, we will often describe the situation by saying that the Xs are "*independent identically distributed* random variables" (abbreviated to i.i.d. r.v.s).



Functions of random variables

Let X be a random variable on the finite probability space $(\Omega, \mathscr{F}, \mathrm{P})$. If t a function mapping the range of X into the reals, then the function $t \circ X$ is also a random variable on $(\Omega, \mathscr{F}, \mathrm{P})$; usually we write $t(X)$ instead of $t \circ X$. The distribution of $t(X)$ is easily found, in principle at least, since

$$\mathrm{P}(t(X) = y) = \mathrm{P}\Big(X \in \{x : t(x) = y\}\Big) = \sum_{x\,:\,t(x)=y} f(x) \tag{1.4}$$

where f denotes the probability function of X.

Similarly, we write $t(X_1, X_2, \ldots, X_n)$ where t is a function of n variables (so t maps $\mathbb{R}^n$ into $\mathbb{R}$), and $X_1, X_2, \ldots, X_n$ are n random variables. The function t can be as simple as $+$. Here is a result that should not be surprising:

Proposition 1.10
Let $X_1, X_2, \ldots, X_m, X_{m+1}, X_{m+2}, \ldots, X_{m+n}$ be independent random variables, and let t_1 and t_2 be functions of m and n variables, respectively. Then the random variables $Y_1 = t_1(X_1, X_2, \ldots, X_m)$ and $Y_2 = t_2(X_{m+1}, X_{m+2}, \ldots, X_{m+n})$ are independent

The proof is left as an exercise (Exercise 1.22).

The distribution of a sum

Let X_1 and X_2 be random variables on $(\Omega, \mathscr{F}, \mathrm{P})$, and let $f_{12}(x_1, x_2)$ denote their joint probability function. Then $t(X_1, X_2) = X_1 + X_2$ is also a random variable on $(\Omega, \mathscr{F}, \mathrm{P})$, and the general formula (1.4) gives

$$\begin{aligned}
\mathrm{P}(X_1 + X_2 = y) &= \sum_{x_2 \in X_2(\Omega)} \mathrm{P}(X_1 + X_2 = y \text{ and } X_2 = x_2) \\
&= \sum_{x_2 \in X_2(\Omega)} \mathrm{P}(X_1 = y - x_2 \text{ and } X_2 = x_2) \\
&= \sum_{x_2 \in X_2(\Omega)} f_{12}(y - x_2, x_2),
\end{aligned}$$

that is, the probability function of $X_1 + X_2$ is obtained by adding $f_{12}(x_1, x_2)$ over all pairs (x_1, x_2) with $x_1 + x_2 = y$.—This has a straightforward generalisation to sums of more than two random variables.

If X_1 and X_2 are independent, then their joint probability function is the product of their marginal probability functions, so

Proposition 1.11
If X_1 and X_2 are independent random variables with probability functions f_1 and f_2, then the probability function of $Y = X_1 + X_2$ is

$$f(y) = \sum_{x_2 \in X_2(\Omega)} f_1(y - x_2) f_2(x_2).$$

Example 1.18
If X_1 and X_2 are independent identically distributed 01-variables with parameter p, then what is the distribution of $X_1 + X_2$?

The joint distribution function f_{12} of (X_1, X_2) is (cf. (1.2) page 25)

$$\begin{aligned} f_{12}(x_1, x_2) &= f_1(x_1) f_2(x_2) \\ &= p^{x_1}(1-p)^{1-x_1} \cdot p^{x_2}(1-p)^{1-x_2} \\ &= p^{x_1+x_2}(1-p)^{2-(x_1+x_2)} \end{aligned}$$

for $(x_1, x_2) \in \{0,1\}^2$, and 0 otherwise. Thus the probability function of $X_1 + X_2$ is

$$\begin{aligned} f(0) &= f_{12}(0,0) &&= (1-p)^2, \\ f(1) &= f_{12}(1,0) + f_{12}(0,1) &&= 2p(1-p), \\ f(2) &= f_{12}(1,1) &&= p^2. \end{aligned}$$

We check that the three probabilities add to 1:

$$f(0) + f(1) + f(2) = (1-p)^2 + 2p(1-p) + p^2 = \Big((1-p) + p\Big)^2 = 1.$$

(This example can be generalised in several ways.)

1.3 Examples

The probability function of a 01-variable with parameter p is (cf. equation (1.2) on page 25)

$$f(x) = p^x(1-p)^{1-x}, \quad x = 0, 1.$$

If $X_1, X_2, \ldots, X_n$ are independent 01-variables with parameter p, then their joint probability function is (for $(x_1, x_2, \ldots, x_n) \in \{0,1\}^n$)

$$f_{12\ldots n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^s(1-p)^{n-s} \tag{1.5}$$

where $s = x_1 + x_2 + \ldots + x_n$ (cf. Corollary 1.9).

Probability function for the binomial distribution with $n = 7$ and $p = 0.618$

Definition 1.10: Binomial distribution
The distribution of the sum of n independent and identically distributed 01*-variables with parameter p is called the binomial distribution with parameters n and p (*$n \in \mathbb{N}$ *and* $p \in [0\,;1]$*).*

To find the probability function of a binomial distribution, consider n independent 01-variables $X_1, X_2, \ldots, X_n$, each with parameter p. Then $Y = X_1 + X_2 + \cdots + X_n$ follows a binomial distribution with parameters n and p. For y an integer between 0 and n we then have, using that the joint probability function of $X_1, X_2, \ldots, X_n$ is given by (1.5),

$$\begin{aligned} \mathrm{P}(Y = y) &= \sum_{x_1+x_2+\ldots+x_n=y} f_{12\ldots n}(x_1, x_2, \ldots, x_n) \\ &= \sum_{x_1+x_2+\ldots+x_n=y} p^y (1-p)^{n-y} \\ &= \left(\sum_{x_1+x_2+\ldots+x_n=y} 1 \right) p^y (1-p)^{n-y} \\ &= \binom{n}{y} p^y (1-p)^{n-y} \end{aligned}$$

Binomial coefficients
The number of different k-subsets of an n-set, i.e. the number of different subsets with exactly k elements that can be selected from a set with n elements, is denoted $\binom{n}{k}$, and is called a *binomial coefficient*. Here k and n are non-negative integers, and $k \le n$.

• We have $\binom{n}{0} = 1$, and $\binom{n}{k} = \binom{n}{n-k}$.

since the number of different n-tuples $x_1, x_2, \ldots, x_n$ with y 1s and $n - y$ 0s is $\binom{n}{y}$. Hence, the probability function of the binomial distribution with parameters n and p is

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 0, 1, 2, \ldots, n. \tag{1.6}$$

Theorem 1.12
If Y_1 *and* Y_2 *are independent binomial variables such that* Y_1 *has parameters* n_1 *and* p*, and* Y_2 *has parameters* n_2 *and* p*, then the distribution of* $Y_1 + Y_2$ *is binomial with parameters* $n_1 + n_2$ *and* p*.*

Proof
One can think of (at least) two ways of proving the theorem, a troublesome way based on Proposition 1.11, and a cleverer way:

Let us consider $n_1 + n_2$ independent and identically distributed 01-variables $X_1, X_2, \ldots, X_{n_1}, X_{n_1+1}, \ldots, X_{n_1+n_2}$. Per definition, $X_1 + X_2 + \ldots + X_{n_1}$ has the same distribution as Y_1, namely a binomial distribution with parameters n_1 and p, and similarly $X_{n_1+1} + X_{n_1+2} + \ldots + X_{n_1+n_2}$ follows the same distribution as Y_2. From Proposition 1.10 we have that Y_1 and Y_2 are independent. Thus all the premises of Theorem 1.12 are met for these variables Y_1 and Y_2, and since $Y_1 + Y_2$

is a sum of $n_1 + n_2$ independent and identically distributed 01-variables with parameter p, Definition 1.10 tells us that $Y_1 + Y_2$ is binomial with parameters $n_1 + n_2$ and p. This proves the theorem!

Perhaps the reader is left with a feeling that this proof was a little bit odd, and is it really a proof? Is it more than a very special situation with a Y_1 and a Y_2 constructed in a way that makes the conclusion almost self-evident? Should the proof not be based on "arbitrary" binomial variables Y_1 and Y_2?

No, not necessarily. The point is that the theorem is not a theorem about random variables as functions on a probability space, but a theorem about how the map $(y_1, y_2) \mapsto y_1 + y_2$ transforms certain probability distributions, and in this connection it is of no importance how these distributions are provided. The random variables only act as symbols that are used for writing things in a lucid way. (See also page 22.) □

One application of the binomial distribution is for modelling the number of favourable outcomes in a series of n independent repetitions of an experiment with only two possible outcomes, favourable and unfavourable.

Think of a series of experiments that extends over two days. Then we can model the number of favourable outcomes on each day and add the results, or we can model the total number. The final result should be the same in both cases, and Proposition 1.12 tells us that this is in fact so.

Then one might consider problems such as: Suppose we made n_1 repetitions on day 1 and n_2 repetitions on day 2. If the total number of favourable outcomes is known to be s, what can then be said about the number of favourable outcomes on day 1? That is what the next result is about.

Proposition 1.13
If Y_1 and Y_2 are independent random variables such that Y_1 is binomial with parameters n_1 and p, and Y_2 is binomial with parameters n_2 and p, then the conditional distribution of Y_1 given that $Y_1 + Y_2 = s$ has probability function

$$\mathrm{P}\big(Y_1 = y \mid Y_1 + Y_2 = s\big) = \frac{\binom{n_1}{y}\binom{n_2}{s-y}}{\binom{n_1+n_2}{s}}, \tag{1.7}$$

which is non-zero when $\max\{s - n_2, 0\} \le y \le \min\{n_1, s\}$.

Note that the conditional distribution does not depend on p.—The proof of Proposition 1.13 is left as an exercise (Exercise 1.17).

The probability distribution with probability function (1.7) is a *hypergeometric distribution*. The hypergeometric distribution already appeared in Example 1.5

• We can deduce a recursion formula for binomial coefficients: Let $G \neq \emptyset$ be a set with n elements and let $g_0 \in G$; now there are two kinds of k-subsets of G: 1) those k-subsets that do not contain g_0 and therefore can be seen as k-subsets of $G \setminus \{g_0\}$, and 2) those k-subsets that do contain g_0 and therefore can be seen as the disjoint union of a $(k-1)$-subset of $G \setminus \{g_0\}$ and $\{g_0\}$. The total number of k-subsets is therefore

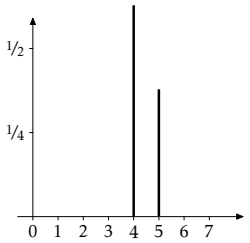
$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}.$$

This is true for $0 < k < n$.

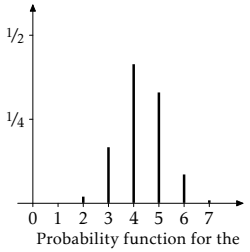
• The formula

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

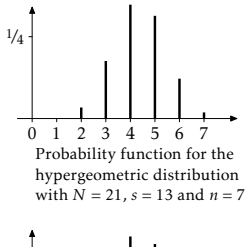
can be verified by checking that both sides of the equality sign satisfy the same recursion formula $f(n,k) = f(n-1,k) + f(n-1,k-1)$, and that they agree for $k = n$ and $k = 0$.



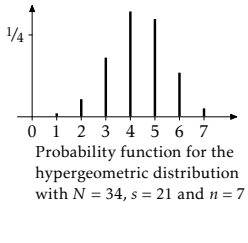
Probability function for the hypergeometric distribution with $N = 8$, $s = 5$ and $n = 7$



Probability function for the hypergeometric distribution with $N = 13$, $s = 8$ and $n = 7$



Probability function for the hypergeometric distribution with $N = 21$, $s = 13$ and $n = 7$



Probability function for the hypergeometric distribution with $N = 34$, $s = 21$ and $n = 7$

on page 15 in connection with simple random sampling *without* replacement. Simple random sampling *with* replacement, however, leads to the binomial distribution, as we shall see below.

When taking a small sample from a large population, it can hardly be of any importance whether the sampling is done with or without replacement. This is spelled out in Theorem 1.14; to better grasp the meaning of the theorem, consider the following situation (cf. Example 1.5): From a box containing N balls, b of which are black and $N - b$ white, a sample of size n is taken without replacement; what is the probability that the sample contains exactly y black balls, and how does this probability behave for large values of N and b?

THEOREM 1.14

When $N \to \infty$ and $b \to \infty$ in such a way that $\dfrac{b}{N} \to p \in [0\,;1]$,

$$\frac{\binom{b}{y}\binom{N-b}{n-y}}{\binom{N}{n}} \longrightarrow \binom{n}{y} p^y (1-p)^{n-y}$$

for all $y \in \{0, 1, 2, \ldots, n\}$.

PROOF

Standard calculations give that

$$\begin{aligned}
\frac{\binom{b}{y}\binom{N-b}{n-y}}{\binom{N}{n}} &= \binom{n}{y} \frac{(N-n)!}{(b-y)!\,(N-n-(b-y))!} \frac{b!\,(N-b)!}{N!} \\
&= \binom{n}{y} \frac{\dfrac{b!}{(b-y)!} \cdot \dfrac{(N-b)!}{(N-b-(n-y))!}}{\dfrac{N!}{(N-n)!}} \\
&= \binom{n}{y} \frac{\overbrace{b(b-1)(b-2)\ldots(b-y+1)}^{y \text{ factors}} \cdot \overbrace{(N-b)(N-b-1)(N-b-2)\ldots(N-b-(n-y)+1)}^{n-y \text{ factors}}}{\underbrace{N(N-1)(N-2)\ldots(N-n+1)}_{n \text{ factors}}} .
\end{aligned}$$

The long fraction has n factors both in the nominator and in the denominator, so with a suitable matching we can write it as a product of n short fractions, each of which has a limiting value: there will be y fractions of the form $\frac{b-\text{something}}{N-\text{something}}$, each converging to p; and there will be $n - y$ fractions of the form $\frac{N-b-\text{something}}{N-\text{something}}$, each converging to $1 - p$. This proves the theorem. □

1.4 Expectation

The expectation — or mean value — of a real-valued function is a weighted average of the values that the function takes.

Definition 1.11: Expectation / expected value / mean value
The expectation, or expected value, or mean value, of a real random variable X on a finite probability space $(\Omega, \mathscr{F}, \mathrm{P})$ is the real number

$$\mathrm{E}(X) = \sum_{\omega \in \Omega} X(\omega)\, \mathrm{P}(\{\omega\}).$$

We often simply write $\mathrm{E}\, X$ instead of $\mathrm{E}(X)$.

If t is a function defined on the range of X, then, using Definition 1.11, the expected value of the random variable $t(X)$ is

$$\mathrm{E}\Big(t(X)\Big) = \sum_{\omega \in \Omega} t\Big(X(\omega)\Big)\, \mathrm{P}(\{\omega\}). \tag{1.8}$$

Proposition 1.15
The expectation of $t(X)$ can be expressed as

$$\mathrm{E}\Big(t(X)\Big) = \sum_{x} t(x)\, \mathrm{P}(X = x) = \sum_{x} t(x) f(x),$$

where the summation is over all x in the range $X(\Omega)$ of X, and where f denotes the probability function of X.

In particular, if t is the identity mapping, we obtain an alternative formula for the expected value of X:

$$\mathrm{E}(X) = \sum_{x} x\, \mathrm{P}(X = x).$$

Proof
The proof is the following series of calculations :

$$\begin{aligned}
\sum_{x} t(x)\, \mathrm{P}(X = x) &\overset{1}{=} \sum_{x} t(x)\, \mathrm{P}\Big(\{\omega : X(\omega) = x\}\Big) \\
&\overset{2}{=} \sum_{x} t(x) \sum_{\omega \in \{X = x\}} \mathrm{P}(\{\omega\}) \\
&\overset{3}{=} \sum_{x} \sum_{\omega \in \{X = x\}} t(x)\, \mathrm{P}(\{\omega\}) \\
&\overset{4}{=} \sum_{x} \sum_{\omega \in \{X = x\}} t(X(\omega))\, \mathrm{P}(\{\omega\})
\end{aligned}$$

$$\begin{aligned}
&\overset{5}{=} \sum_{\omega \in \bigcup_x \{X=x\}} t(X(\omega))\, \mathrm{P}(\{\omega\}) \\
&\overset{6}{=} \sum_{\omega \in \Omega} t(X(\omega))\, \mathrm{P}(\{\omega\}) \\
&\overset{7}{=} \mathrm{E}(t(X)).
\end{aligned}$$

Comments:
Equality 1: making precise the meaning of $\mathrm{P}(X = x)$.
Equality 2: follows from Proposition 1.2 on page 15.
Equality 3: multiply $t(x)$ into the inner sum.
Equality 4: if $\omega \in \{X = x\}$, then $t(X(\omega)) = t(x)$.
Equality 5: the sets $\{X = x\}, \omega \in \Omega$, are disjoint.
Equality 6: $\bigcup_x \{X = x\}$ equals Ω.
Equality 7: equation (1.8). □

Remarks:

1. Proposition 1.15 is of interest because it provides us with three different expressions for the expected value of $Y = t(X)$:

$$\mathrm{E}(t(X)) = \sum_{\omega \in \Omega} t(X(\omega))\, \mathrm{P}(\{\omega\}), \tag{1.9}$$

$$\mathrm{E}(t(X)) = \sum_{x} t(x)\, \mathrm{P}(X = x), \tag{1.10}$$

$$\mathrm{E}(t(X)) = \sum_{y} y\, \mathrm{P}(Y = y). \tag{1.11}$$

   The point is that the calculation can be carried out either on Ω and using P (equation (1.9)), or on (the copy of $\mathbb{R}$ containing) the range of X and using the distribution of X (equation (1.10)), or on (the copy of $\mathbb{R}$ containing) the range of $Y = t(X)$ and using the distribution of Y (equation (1.11)).
2. In Proposition 1.15 it is implied that the symbol X represents an ordinary one-dimensional random variable, and that t is a function from $\mathbb{R}$ to $\mathbb{R}$. But it is entirely possible to interpret X as an n-dimensional random variable, $X = (X_1, X_2, \ldots, X_n)$, and t as a map from $\mathbb{R}^n$ to $\mathbb{R}$. The proposition and its proof remains unchanged.
3. A mathematician will term $\mathrm{E}(X)$ the *integral* of the function X with respect to P and write $\mathrm{E}(X) = \int_\Omega X(\omega)\, \mathrm{P}(d\omega)$, briefly $\mathrm{E}(X) = \int_\Omega X d\mathrm{P}$.

We can think of the expectation operator as a map $X \mapsto \mathrm{E}(X)$ from the set of random variables (on the probability space in question) into the reals.

Theorem 1.16
The map $X \mapsto \mathrm{E}(X)$ *is a linear map from the vector space of random variables on* $(\Omega, \mathscr{F}, \mathrm{P})$ *into the set of reals.*

The proof is left as an exercise (Exercise 1.18 on page 44).

In particular, therefore, the expected value of a sum always equals the sum of the expected values. On the other hand, the expected value of a product is not in general equal to the product of the expected values:

Proposition 1.17
If X *and* Y *are independent random variables on the probability space* $(\Omega, \mathscr{F}, \mathrm{P})$, *then* $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$.

Proof
Since X and Y are independent, $\mathrm{P}(X = x, Y = y) = \mathrm{P}(X = x)\,\mathrm{P}(Y = y)$ for all x and y, and so

$$\begin{aligned}\mathrm{E}(XY) &= \sum_{x,y} xy\,\mathrm{P}(X = x, Y = y)\\ &= \sum_{x,y} xy\,\mathrm{P}(X = x)\mathrm{P}(Y = y)\\ &= \sum_{x}\sum_{y} xy\,\mathrm{P}(X = x)\mathrm{P}(Y = y)\\ &= \sum_{x} x\,\mathrm{P}(X = x)\sum_{y} y\,\mathrm{P}(Y = y) = \mathrm{E}(X)\,\mathrm{E}(Y).\end{aligned}$$

□

Proposition 1.18
If $X(\omega) = a$ *for all* ω, *then* $\mathrm{E}(X) = a$, *in brief* $\mathrm{E}(a) = a$, *and in particular,* $\mathrm{E}(1) = 1$.

Proof
Use Definition 1.11. □

Proposition 1.19
If A *is an event and* $\mathbf{1}_A$ *its indicator function, then* $\mathrm{E}(\mathbf{1}_A) = \mathrm{P}(A)$.

Proof
Use Definition 1.11. (Indicator functions were introduced in Example 1.16 on page 25.) □

In the following, we shall deduce various results about the expectation operator, results of the form: if we know this of X (and $Y, Z, \ldots$), then we can say that

about $\mathrm{E}(X)$ (and $\mathrm{E}(Y), \mathrm{E}(Z), \ldots$). Often, what is said about X is of the form "$u(X)$ with probability 1", that is, it is not required that $u(X(\omega))$ be true for *all* ω, only that the event $\{\omega : u(X(\omega))\}$ has probability 1.

Lemma 1.20
If A is an event with $\mathrm{P}(A) = 0$, and X a random variable, then $\mathrm{E}(X\mathbf{1}_A) = 0$.

Proof
Let ω be a point in A. Then $\{\omega\} \subseteq A$, and since P is increasing (Lemma 1.1), it follows that $\mathrm{P}(\{\omega\}) \leq \mathrm{P}(A) = 0$. We also know that $\mathrm{P}(\{\omega\}) \geq 0$, so we conclude that $\mathrm{P}(\{\omega\}) = 0$, and the first one of the sums below vanishes:

$$\mathrm{E}(X\mathbf{1}_A) = \sum_{\omega \in A} X(\omega)\,\mathbf{1}_A(\omega)\,\mathrm{P}(\{\omega\}) + \sum_{\omega \in \Omega \setminus A} X(\omega)\,\mathbf{1}_A(\omega)\,\mathrm{P}(\{\omega\}).$$

The second sum vanishes because $\mathbf{1}_A(\omega) = 0$ for all $\omega \in \Omega \setminus A$. □

Proposition 1.21
The expectation operator is positive, i.e., if $X \geq 0$ with probability 1, then $\mathrm{E}(X) \geq 0$.

Proof
Let $B = \{\omega : X(\omega) \geq 0\}$ and $A = \{\omega : X(\omega) < 0\}$. Then $1 = \mathbf{1}_B(\omega) + \mathbf{1}_A(\omega)$ for all ω, and so $X = X\mathbf{1}_B + X\mathbf{1}_A$ and hence $\mathrm{E}(X) = \mathrm{E}(X\mathbf{1}_B) + \mathrm{E}(X\mathbf{1}_A)$. Being a sum of non-negative terms, $\mathrm{E}(X\mathbf{1}_B)$ is non-negative, and Lemma 1.20 shows that $\mathrm{E}(X\mathbf{1}_A) = 0$. □

Corollary 1.22
The expectation operator is increasing in the sense that if $X \leq Y$ with probability 1, then $\mathrm{E}(X) \leq \mathrm{E}(Y)$. In particular, $|\mathrm{E}(X)| \leq \mathrm{E}(|X|)$.

Proof
If $X \leq Y$ with probability 1, then $\mathrm{E}(Y) - \mathrm{E}(X) = \mathrm{E}(Y - X) \geq 0$ according to Proposition 1.21.

Using $|X|$ as Y, we then get $\mathrm{E}(X) \leq \mathrm{E}(|X|)$ and $-\mathrm{E}(X) = \mathrm{E}(-X) \leq \mathrm{E}(|X|)$, so $|\mathrm{E}(X)| \leq \mathrm{E}(|X|)$. □

Corollary 1.23
If $X = a$ with probability 1, then $\mathrm{E}(X) = a$.

Proof
With probability 1 we have $a \leq X \leq a$, so from Corollary 1.22 and Proposition 1.18 we get $a \leq \mathrm{E}(X) \leq a$, i.e. $\mathrm{E}(X) = a$. □

Lemma 1.24: Markov inequality
If X is a non-negative random variable and c a positive number, then
$\mathrm{P}(X \geq c) \leq \dfrac{1}{c}\,\mathrm{E}(X)$.

Andrei Markov
Russian mathematician (1856-1922).

Proof
Since $c\,\mathbf{1}_{\{X\geq c\}}(\omega) \leq X(\omega)$ for all ω, we have $c\,\mathrm{E}(\mathbf{1}_{\{X\geq c\}}) \leq \mathrm{E}(X)$, that is, $\mathrm{E}(\mathbf{1}_{\{X\geq c\}}) \leq \frac{1}{c}\mathrm{E}(X)$. Since $\mathrm{E}(\mathbf{1}_{\{X\geq c\}}) = \mathrm{P}(X \geq c)$ (Proposition 1.19), the Lemma is proved. □

Corollary 1.25: Chebyshev inequality
If a is a positive number and X a random variable, then $\mathrm{P}(|X| \geq a) \leq \dfrac{1}{a^2}\,\mathrm{E}(X^2)$.

Panuftii Chebyshev
Russian mathematician (1821-94).

Proof
$\mathrm{P}(|X| \geq a) = \mathrm{P}(X^2 \geq a^2) \leq \frac{1}{a^2}\,\mathrm{E}(X^2)$, using the Markov inequality. □

Corollary 1.26
If $\mathrm{E}|X| = 0$, *then* $X = 0$ *with probability* 1.

Proof
We will show that $\mathrm{P}(|X| > 0) = 0$. From the Markov inequality we know that $\mathrm{P}(|X| \geq \varepsilon) \leq \frac{1}{\varepsilon}\,\mathrm{E}(|X|) = 0$ for all $\varepsilon > 0$. Now chose a positive ε less than the smallest non-negative number in the range of $|X|$ (the range is finite, so such an ε exists). Then the event $\{|X| \geq \varepsilon\}$ equals the event $\{|X| > 0\}$, and hence $\mathrm{P}(|X| > 0) = 0$. □

Augustin Louis Cauchy
French mathematician (1789-1857).

Proposition 1.27: Cauchy-Schwarz inequality
For random variables X and Y on a finite probability space

$$\Big(\mathrm{E}(XY)\Big)^2 \leq \mathrm{E}(X^2)\,\mathrm{E}(Y^2) \tag{1.12}$$

with equality if and only if there exists $(a,b) \neq (0,0)$ such that $aX + bY = 0$ with probability 1.

Hermann Amandus Schwarz
German mathematician (1843-1921).

Proof
If X and Y are both 0 with probability 1, then the inequality is fulfilled (with equality).

Suppose that $Y \neq 0$ with positive probability. For all $t \in \mathbb{R}$

$$0 \leq \mathrm{E}\Big((X + tY)^2\Big) = \mathrm{E}(X^2) + 2t\,\mathrm{E}(XY) + t^2\,\mathrm{E}(Y^2). \tag{1.13}$$

The right-hand side is a quadratic in t (note that $\mathrm{E}(Y^2) > 0$ since Y is non-zero with probability 1), and being always non-negative, it cannot have two different real roots, and therefore the discriminant is non-positive, that is, $(2\,\mathrm{E}(XY))^2 - 4\mathrm{E}(X^2)\mathrm{E}(Y^2) \leq 0$, which is equivalent to (1.12). The discriminant equals zero if and only if the quadratic has one real root, i.e. if and only if there exists a t such that $\mathrm{E}(X + tY)^2 = 0$, and Corollary 1.26 shows that this is equivalent to saying that $X + tY$ is 0 with probability 1.

The case $X \neq 0$ with positive probability is treated in the same way. □

Variance and covariance

If you have to describe the distribution of X using one single number, then $\mathrm{E}(X)$ is a good candidate. If you are allowed to use two numbers, then it would be very sensible to take the second number to be some kind of measure of the random variation of X about its mean. Usually one uses the so-called variance.

Definition 1.12: Variance and standard deviation
The variance of the random variable X is the real number

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

The standard deviation of X is the real number $\sqrt{\mathrm{Var}(X)}$.

From the first of the two formulas for $\mathrm{Var}(X)$ we see that $\mathrm{Var}(X)$ is always non-negative. (Exercise 1.20 is about showing that $\mathrm{E}\Big((X - \mathrm{E}\,X)^2\Big)$ is the same as $\mathrm{E}(X^2) - (\mathrm{E}\,X)^2$.)

Proposition 1.28
$\mathrm{Var}(X) = 0$ *if and only if X is constant with probability* 1, *and if so, the constant equals* $\mathrm{E}\,X$.

Proof
If $X = c$ with probability 1, then $\mathrm{E}\,X = c$; therefore $(X - \mathrm{E}\,X)^2 = 0$ with probability 1, and so $\mathrm{Var}(X) = \mathrm{E}((X - \mathrm{E}\,X)^2) = 0$.

Conversely, if $\mathrm{Var}(X) = 0$ then Corollary 1.26 shows that $(X - \mathrm{E}\,X)^2 = 0$ with probability 1, that is, X equals $\mathrm{E}\,X$ with probability 1. □

The expectation operator is linear, which means that we always have $\mathrm{E}(aX) = a\,\mathrm{E}\,X$ and $\mathrm{E}(X + Y) = \mathrm{E}\,X + \mathrm{E}\,Y$. For the variance operator the algebra is different.

Proposition 1.29
For any random variable X and any real number a, $\mathrm{Var}(aX) = a^2\,\mathrm{Var}(X)$.

Proof
Use the definition of Var and the linearity of E. □

Definition 1.13: Covariance
The covariance between two random variables X and Y on the same probability space is the real number $\mathrm{Cov}(X, Y) = \mathrm{E}\Big((X - \mathrm{E}\,X)(Y - \mathrm{E}\,Y)\Big)$.

The following rules are easily shown:

Proposition 1.30
For any random variables X, Y, U, V and any real numbers a, b, c, d

$$\begin{aligned}
\operatorname{Cov}(X, X) &= \operatorname{Var}(X), \\
\operatorname{Cov}(X, Y) &= \operatorname{Cov}(Y, X), \\
\operatorname{Cov}(X, a) &= 0, \\
\operatorname{Cov}(aX + bY, cU + dV) &= ac \operatorname{Cov}(X, U) + ad \operatorname{Cov}(X, V) \\
&\quad + bc \operatorname{Cov}(Y, U) + bd \operatorname{Cov}(Y, V).
\end{aligned}$$

Moreover:

Proposition 1.31
If X and Y are independent random variables, then $\operatorname{Cov}(X, Y) = 0$.

Proof
If X and Y are independent, then the same is true of $X - \mathrm{E}\, X$ and $Y - \mathrm{E}\, Y$ (Proposition 1.10 page 28), and using Proposition 1.17 page 35 we see that $\mathrm{E}\Big((X - \mathrm{E}\, X)(Y - \mathrm{E}\, Y)\Big) = \mathrm{E}(X - \mathrm{E}\, X)\mathrm{E}(Y - \mathrm{E}\, Y) = 0 \cdot 0 = 0$. □

Proposition 1.32
For any random variables X and Y, $|\operatorname{Cov}(X, Y)| \leq \sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}$. The equality sign holds if and only if there exists $(a, b) \neq (0, 0)$ such that $aX + bY$ is constant with probability 1.

Proof
Apply Proposition 1.27 to the random variables $X - \mathrm{E}\, X$ and $Y - \mathrm{E}\, Y$. □

Definition 1.14: Correlation
The correlation between two non-constant random variables X and Y is the number

$$\operatorname{corr}(X, Y) = \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}}.$$

Corollary 1.33
For any non-constant random variables X and Y

1. $-1 \leq \operatorname{corr}(X, Y) \leq +1$,
2. $\operatorname{corr}(X, Y) = +1$ *if and only if there exists (a, b) with $ab < 0$ such that $aX + bY$ is constant with probability* 1,
3. $\operatorname{corr}(X, Y) = -1$ *if and only if there exists (a, b) with $ab > 0$ such that $aX + bY$ is constant with probability* 1.

Proof
Everything follows from Proposition 1.32, except the link between the sign of

the correlation and the sign of ab, and that follows from the fact that $\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X,Y)$ cannot be zero unless the last term is negative. □

Proposition 1.34
For any random variables X and Y on the same probability space

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + 2\,\mathrm{Cov}(X,Y) + \mathrm{Var}(Y).$$

If X and Y are independent, then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Proof
The general formula is proved using simple calculations. First

$$\Big(X + Y - \mathrm{E}(X+Y)\Big)^2 = (X - \mathrm{E}\,X)^2 + (Y - \mathrm{E}\,Y)^2 + 2(X - \mathrm{E}\,X)(Y - \mathrm{E}\,Y).$$

Taking expectation on both sides then gives

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{E}\Big((X - \mathrm{E}\,X)(Y - \mathrm{E}\,Y)\Big)\\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y).\end{aligned}$$

Finally, if X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$. □

Corollary 1.35
Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with expectation μ and variance σ^2, and let $\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$ denote the average of $X_1, X_2, \ldots, X_n$. Then $\mathrm{E}\,\overline{X}_n = \mu$ and $\mathrm{Var}\,\overline{X}_n = \sigma^2/n$.

This corollary shows how the random variation of an average of n values decreases with n.

Examples

Example 1.19: 01*-variables*
Let X be a 01-variable with $\mathrm{P}(X=1) = p$ and $\mathrm{P}(X=0) = 1-p$. Then the expected value of X is $\mathrm{E}\,X = 0\cdot\mathrm{P}(X=0) + 1\cdot\mathrm{P}(X=1) = p$. The variance of X is, using Definition 1.12, $\mathrm{Var}(X) = \mathrm{E}(X^2) - p^2 = 0^2\cdot\mathrm{P}(X=0) + 1^2\cdot\mathrm{P}(X=1) - p^2 = p(1-p)$.

Example 1.20: Binomial distribution
The binomial distribution with parameters n and p is the distribution of a sum of n independent 01-variables with parameter p (Definition 1.10 page 30). Using Example 1.19, Theorem 1.16 and Proposition 1.31, we find that the expected value of this binomial distribution is np and the variance is $np(1-p)$.

It is perfectly possible, of course, to find the expected value as $\sum xf(x)$ where f is the probability function of the binomial distribution (equation (1.6) page 30), and similarly one can find the variance as $\sum x^2 f(x) - \Big(\sum xf(x)\Big)^2$.

In a later chapter we will find the expected value and the variance of the binomial distribution in an entirely different way, see Example 4.3 on page 78.

Law of large numbers

A main field in the theory of probability is the study of the asymptotic behaviour of sequences of random variables. Here is a simple result:

Theorem 1.36: Weak Law of Large Numbers
Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of independent and identically distributed random variables with expected value μ and variance σ^2.

Then the average $\overline{X}_n = (X_1 + X_2 + \ldots + X_n)/n$ converges to μ in the sense that

$$\lim_{n\to\infty} \mathrm{P}\Big(\big|\overline{X}_n - \mu\big| < \varepsilon\Big) = 1$$

for all $\varepsilon > 0$.

Strong Law of Large Numbers
There is also a *Strong Law of Large Numbers*, stating that under the same assumptions as in the Weak Law of Large Numbers,

$$\mathrm{P}\Big(\lim_{n\to\infty} \overline{X}_n = \mu\Big) = 1,$$

i.e. $\overline{X}_n$ converges to μ with probability 1.

Proof
From Corollary 1.35 we have $\mathrm{Var}(\overline{X}_n - \mu) = \sigma^2/n$. Then the Chebyshev inequality (Corollary 1.25 page 37) gives

$$\mathrm{P}\Big(\big|\overline{X}_n - \mu\big| < \varepsilon\Big) = 1 - \mathrm{P}\Big(\big|\overline{X}_n - \mu\big| \geq \varepsilon\Big) \geq 1 - \frac{1}{\varepsilon^2}\frac{\sigma^2}{n},$$

and the last expression converges to 1 as n tends to ∞. □

Comment: The reader might think that the theorem and the proof involves infinitely many independent and identically distributed random variables, and is that really allowed (at this point of the exposition), and can one have infinite sequences of random variables at all? But everything is in fact quite unproblematic. We consider limits of sequences of real number only; we are working with a finite number n of random variables at a time, and that is quite unproblematic, we can for instance use a product probability space, cf. page 19f.

The Law of Large Numbers shows that when calculating the average of a large number of independent and identically distributed variables, the result will be close to the expected value with a probability close to 1.

We can apply the Law of Large Numbers to a sequence of independent and identically distributed 01-variables $X_1, X_2, \ldots$ that are 1 with probability p; their distribution has expected value p (cf. Example 1.19). The average $\overline{X}_n$ is the relative frequency of the outcome 1 in the finite sequence $X_1, X_2, \ldots, X_n$. The Law of Large Numbers shows that the relative frequency of the outcome 1 with a probability close to 1 is close to p, that is, the relative frequency of 1 is close to the probability of 1. In other words, within the mathematical universe we can deduce that (mathematical) probability is consistent with the notion of relative frequency. This can be taken as evidence of the appropriateness of the mathematical discipline Theory of Probability.

The Law of Large Numbers thus makes it possible to interpret probability as relative frequency in the long run, and similarly to interpret expected value as the average of a very large number of observations.

1.5 Exercises

Exercise 1.1
Write down a probability space that models the experiment "toss a coin three times". Specify the three events "at least one Heads", "at most one Heads", and "not the same result in two successive tosses". What is the probability of each of these events?

Exercise 1.2
Write down an example (a mathematics example, not necessarily a "real wold" example) of a probability space where the sample space has three elements.

Exercise 1.3
Lemma 1.1 on page 14 presents a formula for the probability of the occurrence of at least one of two events A and B. Deduce a formula for the probability that exactly one of the two events occurs.

Exercise 1.4
Give reasons for defining conditional probability the way it is done in Definition 1.3 on page 16, assuming that probability has an interpretation as relative frequency in the long run.

Exercise 1.5
Show that the function $\mathrm{P}(\ \ |\ B)$ in Definition 1.4 on page 17 satisfies the conditions for being a probability measure (Definition 1.1 on page 13). Write down the point probabilities of $\mathrm{P}(\ \ |\ B)$ in terms of those of P.

Exercise 1.6
Two fair dice are thrown and the number of pips of each is recorded. Find the probability that die number 1 turns up a prime number of pips, given that the total number of pips is 8?

Exercise 1.7
A box contains five red balls and three blue balls. In the half-dark in the evening, when all colours look grey, Pedro takes four balls at random from the box, and afterwards Antonia takes one ball at random from the box.

1. Find the probability that Antonia takes a blue ball.
2. Given that Antonia takes a blue ball, what is the probability that Pedro took four red balls?

Exercise 1.8
Assume that children in a certain familily will have blue eyes with probability $1/4$, and that the eye colour of each child is independent of the eye colour of the other children. This family has five children.

1. If it is known that the youngest child has blue eyes, what is probability that at least three of the children have blue eyes?
2. If it is known that at least one of the five children has blue eyes, what is probability that at least three of the children have blue eyes?

Exercise 1.9
Often one is presented to problems of the form: "In a clinical investigation it was found that in a group of lung cancer patients, 80% were smokers, and in a comparable control group 40% were smokers. In the light af this, what can be said about the impact of smoking on lung cancer?"
[A control group is (in the present situation) a group of persons chosen such that it resembles the patient group in "every" respect except that persons in the control group are not diagnosed with lung cancer. The patient group and the control group should be comparable with respect to distribution of age, sex, social stratum, etc.]

Disregarding discussions of the exact meaning of a "comparable" control group, as well as problems relating to statistical and biological variability, then what quantitative statements of any interest can be made, relating to the impact of smoking on lung cancer?

Hint: The *odds* of an event A is the number $\mathrm{P}(A)/\mathrm{P}(A^{\mathrm{c}})$. Deduce formulas for the odds for a lung cancer patient having lung cancer, and the odds for a non-smoker having lung cancer.

Exercise 1.10
From time to time proposals are put forward about general *screenings* of the population in order to detect specific diseases an an early stage.

Suppose one is going to test every single person in a given section of the population for a fairly infrequent disease. The test procedure is not 100% reliable (test procedures seldom are), so there is a small probability of a "false positive", i.e. of erroneously saying that the test person has the disease, and there is also a small probability of a "false negative", i.e. of erroneously saying that the person does not have the disease.

Write down a mathematical model for such a situation. Which parameters would it be sensible to have in such a model?

For the individual person two questions are of the utmost importance: If the test turns out positive, what is the probability of having the disease, and if the test turns out negative, what is the probability of having the disease?

Make use of the mathematical model to answer these questions. You might also enter numerical values for the parameters, and calculate the probabilities.

Exercise 1.11
Show that if the events A and B are independent, then so are A and B^{c}.

Exercise 1.12
Write down a set of necessary and sufficient conditions that a function f is a probability function.

Exercise 1.13
Sketch the probability function of the binomial distribution for various choices of the

parameters n and p. What happens when $p \to 0$ or $p \to 1$? What happens if p is relaced by $1 - p$?

Exercise 1.14
Assume that X is uniformly distributed on $\{1, 2, 3, \ldots, 10\}$ (cf. Definition 1.17 on page 25). Write down the probability function and the distribution function of X, and make sketches of both functions. Find $\mathrm{E}\, X$ and $\mathrm{Var}\, X$.

Exercise 1.15
Consider random variables X and Y on the same probability space $(\Omega, \mathscr{F}, \mathrm{P})$. What connections are there between the statements "$X = Y$ with probability 1" and "X and Y have the same distribution"? – Hint: Example 1.13 on page 22 might be of some use.

Exercise 1.16
Let X and Y be random variables on the same probability space $(\Omega, \mathscr{F}, \mathrm{P})$. Show that if X is constant (i.e. if there exists a number c such that $X(\omega) = c$ for all $\omega \in \Omega$), then X and Y are independent.

Next, show that if X is constant with probability 1 (i.e. if there is a number c such that $\mathrm{P}(X = c) = 1$), then X are Y independent.

Exercise 1.17
Prove Proposition 1.13 on page 31. — Use the definition of conditional probability and the fact that $Y_1 = y$ and $Y_1 + Y_2 = s$ is equivalent to $Y_1 = y$ and $Y_2 = s - y$; then make use of the independence of Y_1 and Y_2, and finally use that we know the probability function for each of the variables Y_1, Y_2 and $Y_1 + Y_2$.

Exercise 1.18
Show that the set of random variables on a finite probability space is a vector space over the set of reals.

Show that the map $X \mapsto \mathrm{E}(X)$ is a linear map from this vector space into the vector space $\mathbb{R}$.

Exercise 1.19
Let the random variable X have a distribution given by $\mathrm{P}(X = 1) = \mathrm{P}(X = -1) = 1/2$. Find $\mathrm{E}\, X$ and $\mathrm{Var}\, X$.

Let the random variable Y have a distribution given by $\mathrm{P}(Y = 100) = \mathrm{P}(Y = -100) = 1/2$, and find $\mathrm{E}\, Y$ and $\mathrm{Var}\, Y$.

Exercise 1.20
Definition 1.12 on page 38 presents two apparently different expressions for the variance of a random variable X. Show that it is in fact true that $\mathrm{E}\big((X - \mathrm{E}\, X)^2\big) = \mathrm{E}(X^2) - (\mathrm{E}\, X)^2$.

Exercise 1.21
Let X be a random variable such that

$$\begin{aligned} \mathrm{P}(X = 2) &= 0.1 & \mathrm{P}(X = -2) &= 0.1 \\ \mathrm{P}(X = 1) &= 0.4 & \mathrm{P}(X = -1) &= 0.4, \end{aligned}$$

and let $Y = t(X)$ where t is the function from $\mathbb{R}$ to $\mathbb{R}$ given by

$$t(x) = \begin{cases} -x & \text{if } |x| \leq 1 \\ x & \text{otherwise.} \end{cases}$$

Show that X and Y have the same distribution. Find the expected value of X and the expected value of Y. Find the variance of X and the variance of Y. Find the covariance between X and Y. Are X and Y independent?

Johan Ludwig William Valdemar Jensen (1859-1925), Danish mathematician and engineer at Kjøbenhavns Telefon Aktieselskab (the Copenhagen Telephone Company). Among other things, known for Jensen's inequality (see Exercise 1.23).

Exercise 1.22
Give a proof of Proposition 1.10 (page 28): In the notation of the proposition, we have to show that

$$\mathrm{P}(Y_1 = y_1 \text{ and } Y_2 = y_2) = \mathrm{P}(Y_1 = y_1)\,\mathrm{P}(Y_2 = y_2)$$

for all y_1 and y_2. Write the first probability as a sum over a certain set of $(m+n)$-tuples $(x_1, x_2, \ldots, x_{m+n})$; this set is a product of a set of m-tuples $(x_1, x_2, \ldots, x_m)$ and a set of n-tuples $(x_{m+1}, x_{m+2}, \ldots, x_{m+n})$.

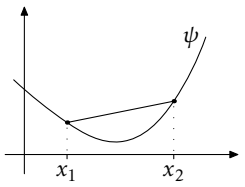
Exercise 1.23: Jensen's inequality
Let X be a random variable and ψ a convex function.Show Jensen's inequality

$$\mathrm{E}\,\psi(X) \geq \psi(\mathrm{E}\,X).$$

Convex functions
A real-valued function ψ defined on an interval $I \subseteq \mathbb{R}$ is said to be *convex* if the line segment connecting any two points $(x_1, \psi(x_1))$ and $(x_2, \psi(x_2))$ of the graph of ψ lies entirely above or on the graph, that is, for any $x_1, x_2 \in I$,

$$\lambda\psi(x_1) + (1-\lambda)\psi(x_2) \geq \psi\big(\lambda x_1 + (1-\lambda)x_2\big)$$

for all $\lambda \in [0\,;1]$.



If ψ possesses a continuous second derivative, then ψ is convex if and only if $\psi'' \geq 0$.

Exercise 1.24
Let $x_1, x_2, \ldots, x_n$ be n positive numbers. The number $G = (x_1\, x_2 \ldots x_n)^{1/n}$ is called the geometric mean of the xs, and $A = (x_1 + x_2 + \cdots + x_n)/n$ is the arithmetic mean of the xs. Show that $G \leq A$.

Hint: Apply Jensen's inequality to the convex function $\psi = -\ln$ and the random variable X given by $\mathrm{P}(X = x_i) = \frac{1}{n}$, $i = 1, 2, \ldots, n$.

2 Countable Sample Spaces

This chapter presents elements of the theory of probability distributions on countable sample spaces. In many respects it is an entirely unproblematic generalisation of the theory covering the finite case; there will, however, be a couple of significant exceptions.

Recall that in the finite case we were dealing with the set $\mathscr{F}$ of all subsets of the finite sample space Ω, and that each element of $\mathscr{F}$, that is, each subset of Ω (each event) was assigned a probability. In the countable case too, we will assign a probability to each subset of the sample space. This will meet the demands specified by almost any modelling problem on a countable sample space, although it is somewhat of a simplification from the formal point of view. — The interested reader is referred to Definition 5.2 on page 89 to see a general definition of probability space.

2.1 Basic definitions

Definition 2.1: Probability space over countable sample space
A probability space over a countable set is a triple $(\Omega, \mathscr{F}, \mathrm{P})$ *consisting of*

1. *a sample space* Ω*, which is a non-empty countable set,*
2. *the set* $\mathscr{F}$ *of all subsets of* Ω*,*
3. *a probability measure on* $(\Omega, \mathscr{F})$*, that is, a map* $\mathrm{P} : \mathscr{F} \to \mathbb{R}$ *that is*
   - *positive:* $\mathrm{P}(A) \geq 0$ *for all* $A \in \mathscr{F}$*,*
   - *normed:* $\mathrm{P}(\Omega) = 1$*, and*
   - σ*-additive: if* $A_1, A_2, \ldots$ *is a sequence in* $\mathscr{F}$ *of disjoint events, then*

$$\mathrm{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathrm{P}(A_i). \tag{2.1}$$

A countably infinite set has a non-countable infinity of subsets, but, as you can see, the condition for being a probability measure involves only a countable infinity of events at a time.

Lemma 2.1
Let $(\Omega, \mathscr{F}, \mathrm{P})$ *be a probability space over the countable sample space* Ω*. Then*

i. $\mathrm{P}(\varnothing) = 0$.

ii. *If $A_1, A_2, \ldots, A_n$ are disjoint events from $\mathscr{F}$, then*

$$\mathrm{P}\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} \mathrm{P}(A_i),$$

that is, P *is also finitely additive.*

iii. $\mathrm{P}(A) + \mathrm{P}(A^{\mathrm{c}}) = 1$ *for all events* A.

iv. *If* $A \subseteq B$, *then* $\mathrm{P}(B \setminus A) = \mathrm{P}(B) - \mathrm{P}(A)$, *an* $\mathrm{P}(A) \leq \mathrm{P}(B)$, *i.e.* P *is increasing.*

v. $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$.

Proof

If we in equation (2.1) let $A_1 = \Omega$, and take all other As to be $\varnothing$, we see that $\mathrm{P}(\varnothing)$ must equal 0. Then it is obvious that σ-additivity implies finite additivity.

The rest of the lemma is proved in the same way as in the finite case, see the proof of Lemma 1.1 on page 14. □

The next lemma shows that even in a countably infinite sample space almost all of the probability mass will be located on a finite set.

Lemma 2.2

Let $(\Omega, \mathscr{F}, \mathrm{P})$ *be a probability space over the countable sample space* Ω. *For each* $\varepsilon > 0$ *there is a finite subset* Ω_0 *of* Ω *such that* $\mathrm{P}(\Omega_0) \geq 1 - \varepsilon$.

Proof

If $\omega_1, \omega_2, \omega_3, \ldots$ is an enumeration of the elements of Ω, then we can write $\bigcup_{i=1}^{\infty} \{\omega_i\} = \Omega$, and hence $\sum_{i=1}^{\infty} \mathrm{P}(\{\omega_i\}) = \mathrm{P}(\Omega) = 1$ by the σ-additivity of P. Since the series converges to 1, the partial sums eventually exceed $1 - \varepsilon$, i.e. there is an n such that $\sum_{i=1}^{n} \mathrm{P}(\{\omega_i\}) \geq 1 - \varepsilon$. Now take Ω_0 to be the set $\{\omega_1, \omega_2, \ldots, \omega_n\}$. □

Point probabilities

Point probabilities are defined the same way as in the finite case, cf. Definition 1.2 on page 15:

Definition 2.2: Point probabilities

Let $(\Omega, \mathscr{F}, \mathrm{P})$ *be a probability space over the countable sample space* Ω. *The function*

$$\begin{aligned} p : \Omega &\longrightarrow [0\,;1] \\ \omega &\longmapsto \mathrm{P}(\{\omega\}) \end{aligned}$$

is the point probability function of P.

The proposition below is proved in the same way as in the finite case (Proposition 1.2 on page 15), but note that in the present case the sum can have infinitely many terms.

Proposition 2.3
If p is the point probability function of the probability measure P, *then* $\mathrm{P}(A) = \sum_{\omega \in A} p(\omega)$ *for any event* $A \in \mathscr{F}$.

Remarks: It follows from the proposition that different probability measures cannot have the same point probability function. It also follows (by taking $A = \Omega$) that p adds up to 1, i.e. $\sum_{\omega \in \Omega} p(\omega) = 1$.

The next result is proved the same way as in the finite case (Proposition 1.3 on page 16); note that in sums with non-negative terms, you can change the order of the terms freely without changing the value of the sum.

Proposition 2.4
If $p : \Omega \to [0\,;1]$ *adds up to* 1, *i.e.* $\sum_{\omega \in \Omega} p(\omega) = 1$, *then there exists exactly one probability measure on* Ω *with* p *as its point probability function.*

Conditioning; independence

Notions such as conditional probability, conditional distribution and independence are defined exactly as in the finite case, see page 16 and 18.

Note that Bayes' formula (see page 17) has the following generalisation: If $B_1, B_2, \ldots$ is a partition of Ω, i.e. the Bs are disjoint and $\bigcup_{i=1}^{\infty} B_i = \Omega$, then

$$\mathrm{P}(B_j \mid A) = \frac{\mathrm{P}(A \mid B_j)\,\mathrm{P}(B_j)}{\sum_{i=1}^{\infty} \mathrm{P}(A \mid B_i)\,\mathrm{P}(B_i)}$$

for any event A.

Random variables

Random variables are defined the same way as in the finite case, except that now they can take infinitely many values.

Here are some examples of distributions living on finite subsets of the reals. First some distributions that can be of some use in actual model building, as we shall see later (in Section 2.3):

- The *Poisson distribution* with parameter $\mu > 0$ has point probability function
$$f(x) = \frac{\mu^x}{x!}\exp(-\mu), \quad x = 0, 1, 2, 3, \ldots$$
- The *geometric distribution* with probability parameter $p \in]0\,;1[$ has point probability function
$$f(x) = p\,(1-p)^x, \quad x = 0, 1, 2, 3, \ldots$$
- The *negative binomial distribution* with probability parameter $p \in]0\,;1[$ and shape parameter $k > 0$ has point probability function
$$f(x) = \binom{x+k-1}{x} p^k (1-p)^x, \quad x = 0, 1, 2, 3, \ldots$$
- The *logarithmic distribution* with parameter $p \in]0\,;1[$ has point probability function
$$f(x) = \frac{1}{-\ln p}\,\frac{(1-p)^x}{x}, \quad x = 1, 2, 3, \ldots$$

Here is a different kind of example, showing what the mathematic formalism also is about.

- The range of a random variable can of course be all kinds of subsets of the reals. A not too weird example is a random variable X taking the values $\pm 1, \pm\frac{1}{2}, \pm\frac{1}{3}, \pm\frac{1}{4}, \ldots, 0$ with these probabilies
$$\begin{aligned} \mathrm{P}\Big(X = \tfrac{1}{n}\Big) &= \tfrac{1}{3\cdot 2^n}, \qquad n = 1, 2, 3, \ldots \\ \mathrm{P}\Big(X = 0\Big) &= \tfrac{1}{3}, \\ \mathrm{P}\Big(X = -\tfrac{1}{n}\Big) &= \tfrac{1}{3\cdot 2^n}, \qquad n = 1, 2, 3, \ldots \end{aligned}$$
Thus X takes infinitely many values, all lying in a finite interval.

Distribution function and probability function are defined in the same way in the countable case as in the finite case (Definition 1.7 on page 23 and Definition 1.8 on page 26), and Lemma 1.5 and Proposition 1.6 on page 24 are unchanged. The proof of Proposition 1.6 calls for a modification: either utilise that a countable set is "almost finite" (Lemma 2.2), or use the proof of the general case (Proposition 5.3).

The criterion for independence of random variables (Proposition 1.8 or Corollary 1.9 on page 27) remain unchanged. The formula for the point probability function for a sum of random variables is unchanged (Proposition 1.11 on page 29).

2.2 Expectation

Let $(\Omega, \mathscr{F}, \mathrm{P})$ be a probability space over a countable set, and let X be a random variable on this space. One might consider defining the expectation of X in the same way as in the finite case (page 33), i.e. as the real number $\mathrm{E}\,X = \sum_{\omega\in\Omega} X(\omega)\,\mathrm{P}(\{\omega\})$ or $\mathrm{E}\,X = \sum_x x\,f(x)$, where f is the probability function of X. In a way this is an excellent idea, except that the sums now may involve infinitely many terms, which may give problems of convergence.

Example 2.1

It is assumed to be known that for $\alpha > 1$, $c(\alpha) = \sum_{x=1}^{\infty} x^{-\alpha}$ is a finite positive number; therefore we can define a random variable X having probability function $f_X(x) = \frac{1}{c(\alpha)}\,x^{-\alpha}$, $x = 1, 2, 3, \ldots$ However, the series $\mathrm{E}\,X = \sum_{x=1}^{\infty} x f_X(x) = \frac{1}{c(\alpha)} \sum_{x=1}^{\infty} x^{-\alpha+1}$ converges to a well-defined limit only if $\alpha > 2$. For $\alpha \in]1\,;\,2]$, the probability function converges too slowly towards 0 for X to have an expected value. So not all random variables can be assigned an expected value in a meaningful way.

One might suggest that we introduce $+\infty$ and $-\infty$ as permitted values for $\mathrm{E}\,X$, but that does not solve all problems. Consider for example a random variable Y with range $\mathbb{N} \cup -\mathbb{N}$ and probability function $f_Y(y) = \frac{1}{2c(\alpha)}\,|y|^{-\alpha}$, $y = \pm1, \pm2, \pm3, \ldots$; in this case it is not so easy to say what $\mathrm{E}\,Y$ should really mean.

Definition 2.3: Expectation / expected value / mean value
Let X be a random variable on $(\Omega, \mathscr{F}, \mathrm{P})$. If the sum $\sum_{\omega\in\Omega} |X(\omega)|\,\mathrm{P}(\{\omega\})$ converges to a finite value, then X is said to have an expected value, and the expected value of X is then the real number $\mathrm{E}\,X = \sum_{\omega\in\Omega} X(\omega)\,\mathrm{P}(\{\omega\})$.

Proposition 2.5
Let X be a real random variable on $(\Omega, \mathscr{F}, \mathrm{P})$, and let f denote the probability function of X. Then X has an expected value if and only if the sum $\sum_x |x| f(x)$ is finite, and if so, $\mathrm{E}\,X = \sum_x x f(x)$.

This can be proved in almost the same way as Proposition 1.15 on page 33, except that we now have to deal with infinite series.

Example 2.1 shows that not all random variables have an expected value. — We introduce a special notation for the set of random variables that do have an expected value.

Definition 2.4
The set of random variables on $(\Omega, \mathscr{F}, \mathrm{P})$ *which have an expected value, is denoted* $\mathscr{L}^1(\Omega, \mathscr{F}, \mathrm{P})$ *or* $\mathscr{L}^1(\mathrm{P})$ *or* $\mathscr{L}^1$.

We have (cf. Theorem 1.16 on page 35)

Theorem 2.6
The set $\mathscr{L}^1(\Omega, \mathscr{F}, \mathrm{P})$ *is a vector space* (*over* $\mathbb{R}$), *and the map* $X \mapsto \mathrm{E}\, X$ *is a linear map of this vector space into* $\mathbb{R}$.

This is proved using standard results from the calculus of infinite series.

The results from the finite case about expected values are easily generalised to the countable case, but note that Proposition 1.17 on page 35 now becomes

Proposition 2.7
If X *and* Y *are* $independent$ *random variables and both have an expected value, then* XY *also has an expected value, and* $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$.

Lemma 1.20 on page 36 now becomes

Lemma 2.8
If A *is an event with* $\mathrm{P}(A) = 0$, *and* X *is a random variable, then the random variable* $X\mathbf{1}_A$ *has an expected value, and* $\mathrm{E}(X\mathbf{1}_A) = 0$.

The next result shows that you can change the values of a random variable on a set of probability 0 without changing its expected value.

Proposition 2.9
Let X *and* Y *be random variables. If* $X = Y$ *with probability* 1, *and if* $X \in \mathscr{L}^1$, *then also* $Y \in \mathscr{L}^1$, *and* $\mathrm{E}\, Y = \mathrm{E}\, X$.

Proof
Since $Y = X + (Y - X)$, then (using Theorem 2.6) it suffices to show that $Y - X \in \mathscr{L}^1$ and $\mathrm{E}(Y - X) = 0$. Let $A = \{\omega : X(\omega) \neq Y(\omega)\}$. Now we have $Y(\omega) - X(\omega) = (Y(\omega) - X(\omega))\mathbf{1}_A(\omega)$ for all ω, and the result follows from Lemma 2.8. □

An application of the comparison test for infinite series gives

Lemma 2.10
Let Y *be a random variable. If there is a non-negative random variable* $X \in \mathscr{L}^1$ *such that* $|Y| \leq X$, *then* $Y \in \mathscr{L}^1$, *and* $\mathrm{E}\, Y \leq \mathrm{E}\, X$.

Variance and covariance

In brief the variance of a random variable X is defined as $\mathrm{Var}(X) = \mathrm{E}\big((X - \mathrm{E}\,X)^2\big)$ whenever this is a finite number.

Definition 2.5
The set of random variables X on $(\Omega, \mathscr{F}, \mathrm{P})$ such that $\mathrm{E}(X^2)$ exists, is denoted $\mathscr{L}^2(\Omega, \mathscr{F}, \mathrm{P})$ or $\mathscr{L}^2(\mathrm{P})$ or $\mathscr{L}^2$.

Note that $X \in \mathscr{L}^2(\mathrm{P})$ if and only if $X^2 \in \mathscr{L}^1(\mathrm{P})$.

Lemma 2.11
If $X \in \mathscr{L}^2$ and $Y \in \mathscr{L}^2$, then $XY \in \mathscr{L}^1$.

Proof
Use the fact that $|xy| \leq x^2 + y^2$ for any real numbers x and y, together with Lemma 2.10. □

Proposition 2.12
The set $\mathscr{L}^2(\Omega, \mathscr{F}, \mathrm{P})$ is a linear subspace of $\mathscr{L}^1(\Omega, \mathscr{F}, \mathrm{P})$, and this subspace contains all constant random variables.

Proof
To see that $\mathscr{L}^2 \subseteq \mathscr{L}^1$, use Lemma 2.11 with $Y = X$.

It follows from the definition of $\mathscr{L}^2$ that if $X \in \mathscr{L}^2$ and a is a constant, then $aX \in \mathscr{L}^2$, i.e. $\mathscr{L}^2$ is closed under scalar multiplication.

If $X, Y \in \mathscr{L}^2$, then X^2, Y^2 and XY all belongs to $\mathscr{L}^1$ according to Lemma 2.11. Hence $(X+Y)^2 = X^2 + 2XY + Y^2 \in \mathscr{L}^1$, i.e. $X + Y \in \mathscr{L}^2$. Thus $\mathscr{L}^2$ is also closed under addition. □

Definition 2.6: Variance
If $\mathrm{E}(X^2) < +\infty$, then X is said to have a variance, and the variance of X is the real number

$$\mathrm{Var}(X) = \mathrm{E}\big((X - \mathrm{E}\,X)^2\big) = \mathrm{E}(X^2) - (\mathrm{E}\,X)^2.$$

Definition 2.7: Covariance
If X and Y both have a variance, then their covariance is the real number

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

The rules of calculus for variances and covariances are as in the finite case (page 38ff).

Proposition 2.13: Cauchy-Schwarz inequality
If X and Y both belong to $\mathscr{L}^2(\mathrm{P})$, then

$$\big(\mathrm{E}|XY|\big)^2 \le \mathrm{E}(X^2)\,\mathrm{E}(Y^2). \tag{2.2}$$

Proof
From Lemma 2.11 it is known that $XY \in \mathscr{L}^1(\mathrm{P})$. Then proceed along the same lines as in the finite case (Proposition 1.27 on page 37). □

If we replace X with $X - \mathrm{E}\,X$ and Y with $Y - \mathrm{E}\,Y$ in equation (2.2), we get

$$\mathrm{Cov}(X,Y)^2 \le \mathrm{Var}(X)\,\mathrm{Var}(Y).$$

2.3 Examples

Geometric distribution

Generalised binomial coefficiens
For $r \in \mathbb{R}$ and $k \in \mathbb{N}$ we define $\binom{r}{k}$ as the number $\frac{r(r-1)\ldots(r-k+1)}{1\cdot 2\cdot 3\cdot \ldots \cdot k}$.
(When $r \in \{0,1,2,\ldots,n\}$ this agrees with the combinatoric definition of binomial coefficient (page 30).)

Consider a random experiment with two possible outcomes 0 and 1 (or Failure and Success), and let the probability of outcome 1 be p. Repeat the experiment until outcome 1 occurs for the first time. How many times do we have to repeat the experiment before the first 1? Clearly, in principle there is no upper bound for the number of repetitions needed, so this seems to be an example where we cannot use a finite sample space. Moreover, we cannot be sure that the outcome 1 will ever occur.

To make the problem more precise, let us assume that the elementary experiment is carried out at time $t = 1, 2, 3, \ldots$, and what is wanted is the waiting time to the first occurrence of a 1. Define

$$\begin{aligned} T_1 &= \text{ time of first 1,} \\ V_1 = T_1 - 1 &= \text{ number of 0s before the first 1.} \end{aligned}$$

Strictly speaking we cannot be sure that we shall ever see the outcome 1; we let $T_1 = \infty$ and $V_1 = \infty$ if the outcome 1 never occurs. Hence the possible values of T_1 are $1, 2, 3, \ldots, \infty$, and the possible values of V_1 are $0, 1, 2, 3, \ldots, \infty$.

Let $t \in \mathbb{N}_0$. The event $\{V_1 = t\}$ (or $\{T_1 = t+1\}$) means that the experiment produces outcome 0 at time $1, 2, \ldots, t$ and outcome 1 at time $t+1$, and therefore $\mathrm{P}(V_1 = t) = p\,(1-p)^t$, $t = 0, 1, 2, \ldots$ Using the formula for the sum of a geometric series we see that

Geometric series
For $|t| < 1$, $\frac{1}{1-t} = \sum_{n=0}^{\infty} t^n$.

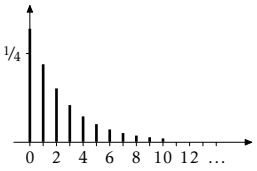
$$\mathrm{P}(V_1 \in \mathbb{N}_0) = \sum_{t=0}^{\infty} \mathrm{P}(V_1 = t) = \sum_{t=0}^{\infty} p\,(1-p)^t = 1,$$

that is, V_1 (and T_1) is finite with probability 1. In other words, with probability 1 the outcome 1 will occur sooner or later.

The distribution of V_1 is a geometric distribution:

Definition 2.8: Geometric distribution
The geometric distribution with parameter p is the distribution on $\mathbb{N}_0$ with probability function

$$f(t) = p\,(1-p)^t, \quad t \in \mathbb{N}_0.$$



Geometric distribution with $p = 0.316$

We can find the expected value of the geometric distribution:

$$\begin{aligned}\mathrm{E}\,V_1 &= \sum_{t=0}^{\infty} tp(1-p)^t = p(1-p)\sum_{t=1}^{\infty} t\,(1-p)^{t-1} \\ &= p(1-p)\sum_{t=0}^{\infty}(t+1)(1-p)^t = \frac{1-p}{p}\end{aligned}$$

(cf. the formula for the sum of a binomial series). So on the average there will be $(1-p)/p$ 0s before the first 1, and the mean waiting time to the first 1 is $\mathrm{E}\,T_1 = \mathrm{E}\,V_1 + 1 = 1/p$. — The variance of the geometric distribution can be found in a similar way, and it turns out to be $(1-p)/p^2$.

Binomial series
1. For $n \in \mathbb{N}$ and $t \in \mathbb{R}$,

$$(1+t)^n = \sum_{k=0}^{n}\binom{n}{k}t^k.$$

2. For $\alpha \in \mathbb{R}$ and $|t| < 1$,

$$(1+t)^\alpha = \sum_{k=0}^{\infty}\binom{\alpha}{k}t^k.$$

3. For $\alpha \in \mathbb{R}$ and $|t| < 1$,

$$\begin{aligned}&(1-t)^{-\alpha} \\ &\quad= \sum_{k=0}^{\infty}\binom{-\alpha}{k}(-t)^k \\ &\quad= \sum_{k=0}^{\infty}\binom{\alpha+k-1}{k}t^k.\end{aligned}$$

The probability that the waiting time for the first 1 exceeds t is

$$\mathrm{P}(T_1 > t) = \sum_{k=t+1}^{\infty}\mathrm{P}(T_1 = k) = (1-p)^t.$$

The probability that we will have to wait more than $s+t$, given that we have waited already the time s, is

$$\mathrm{P}(T_1 > s+t \mid T_1 > s) = \frac{\mathrm{P}(T_1 > s+t)}{\mathrm{P}(T_1 > s)} = (1-p)^t = \mathrm{P}(T_1 > t),$$

showing that the remaining waiting time does not depend on how long we have been waiting until now — this is the *lack of memory* property of the distribution of T_1 (or of the mechanism generating the 0s and 1s). (The continuous time analogue of this distribution is the exponential distribution, see page 69.)

The "shifted" geometric distribution is the only distribution on $\mathbb{N}_0$ with this property: If T is a random variable with range $\mathbb{N}$ and satisfying $\mathrm{P}(T > s+t) = \mathrm{P}(T > s)\,\mathrm{P}(T > t)$ for all $s, t \in \mathbb{N}$, then

$$\mathrm{P}(T > t) = \mathrm{P}(T > 1+1+\ldots+1) = \Big(\mathrm{P}(T > 1)\Big)^t = (1-p)^t$$

where $p = 1 - \mathrm{P}(T > 1) = \mathrm{P}(T = 1)$.

Negative binomial distribution

In continuation of the previous section consider the variables

$$\begin{aligned} T_k &= \text{ time of } k\text{th 1} \\ V_k &= T_k - k = \text{ number of 0s before } k\text{th 1}. \end{aligned}$$

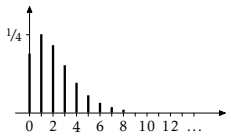
The event $\{V_k = t\}$ (or $\{T_k = t+k\}$) corresponds to getting the result 1 at time $t+k$, and getting $k-1$ 1s and t 0s before time $t+k$. The probability of obtaining a 1 at time $t+k$ equals p, and the probability of obtaining exactly t 0s among the $t+k-1$ outcomes before time $t+k$ is the binomial probability $\binom{t+k-1}{t} p^{k-1}(1-p)^t$, and hence

$$\mathrm{P}(V_k = t) = \binom{t+k-1}{t} p^k (1-p)^t, \quad t \in \mathbb{N}_0.$$

This distribution of V_k is a negative binomial distribution:

Definition 2.9: Negative binomial distribution
The negative binomial distribution with probability parameter $p \in]0\,;1[$ and shape parameter $k > 0$ is the distribution on $\mathbb{N}_0$ that has probability function

$$f(t) = \binom{t+k-1}{t} p^k (1-p)^t, \quad t \in \mathbb{N}_0. \tag{2.3}$$



Negative binomial distribution with $p = 0.618$ and $k = 3.5$

Remarks: 1) For $k = 1$ this is the geometric distribution. 2) For obvious reasons, k is an integer in the above derivations, but actually the expression (2.3) is well-defined and a probability function for any positive real number k.

Proposition 2.14
If X and Y are independent and have negative binomial distributions with the same probability parameter p and with shape parameters j and k, respectively, then $X + Y$ has a negative binomial distribution with probability parameter p and shape parameter $j + k$.

Proof
We have

$$\begin{aligned} \mathrm{P}(X+Y=s) &= \sum_{x=0}^{s} \mathrm{P}(X=x)\,\mathrm{P}(Y=s-x) \\ &= \sum_{x=0}^{s} \binom{x+j-1}{x} p^j (1-p)^x \cdot \binom{s-x+k-1}{s-x} p^k (1-p)^{s-x} \\ &= p^{j+k} A(s) (1-p)^s \end{aligned}$$

where $A(s) = \sum_{x=0}^{s} \binom{x+j-1}{x}\binom{s-x+k-1}{s-x}$. Now one approach would be to do some lengthy rewritings of the $A(s)$ expression. Or we can reason as follows: The total probability has to be 1, i.e. $\sum_{s=0}^{\infty} p^{j+k} A(s)(1-p)^s = 1$, and hence $p^{-(j+k)} = \sum_{s=0}^{\infty} A(s)(1-p)^s$. We also have the identity $p^{-(j+k)} = \sum_{s=0}^{\infty} \binom{s+j+k-1}{s}(1-p)^s$, either from the formula for the sum of a binomial series, or from the fact that the sum of the point probabilities of the negative binomial distribution with parameters p and $j+k$ is 1. By the uniqueness of power series expansions it now follows that $A(s) = \binom{s+j+k-1}{s}$, which means that $X+Y$ has the distribution stated in the proposition. □

It may be added that within the theory of stochastic processes the following informal reasoning can be expanded to a proper proof of Proposition 2.14: The number of 0s occurring before the $(j+k)$th 1 equals the number of 0s before the jth 1 plus the number of 0s between the jth and the $(j+k)$th 1; since values at different times are independent, and the rule determining the end of the first period is independent of the number of 0s in that period, then the number of 0s in the second period is independent of the number of 0s in the first period, and both follow negative binomial distributions.

The expected value and the variance in the negative binomial distribution are $k(1-p)/p$ and $k(1-p)/p^2$, respectively, cf. Example 4.5 on page 79. Therefore, the expected number of 0s before the kth 1 is $\mathrm{E}\, V_k = k(1-p)/p$, and the expected waiting time to the kth 1 is $\mathrm{E}\, T_k = k + \mathrm{E}\, V_k = k/p$.

Poisson distribution

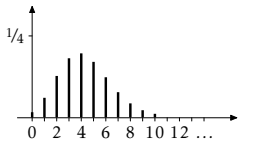
Suppose that instances of a certain kind of phenomena occur "entirely at random" along some "time axis"; examples could be earthquakes, deaths from a non-epidemic disease, road accidents in a certain crossroads, particles of the cosmic radiation, α particles emitted from a radioactive source, etc. etc. In such connection you could ask various questions, such as: what can be said about the number of occurrences in a given time interval, and what can be said about the waiting time from one occurrence to the next. — Here we shall discuss the number of occurrences.

Consider an interval $]a\,;b]$ on the time axis. We shall be assuming that the occurrences take place at points that are selected entirely at random on the time axis, and in such a way that the intensity (or rate of occurrence) remains

constant throughout the period of interest (this assumption will be written out more clearly later on). We can make an approximation to the "entirely random" placement in the following way: divide the interval $]a\,;b]$ into a large number of very short subintervals, say n subintervals of length $\Delta t = (b-a)/n$, and then for each subinterval have a random number generator determine whether there be an occurrence or not; the probability that a given subinterval is assigned an occurrence is $p(\Delta t)$, and different intervals are treated independently of each other. The probability $p(\Delta t)$ is the same for all subintervals of the given length Δt. In this way the total number of occurrences in the big interval $]a\,;b]$ follows a binomial distribution with parameters n and $p(\Delta t)$ — the total number is the quantity that we are trying to model.

Siméon-Denis Poisson
French mathematician and physicist (1781-1840).

It seems reasonable to believe that as n increases, the above approximation approaches the "correct" model. Therefore we will let n tend to infinity; at the same time $p(\Delta t)$ must somehow tend to 0; it seems reasonable to let $p(\Delta t)$ tend to 0 in such a way that the expected value of the binomial distribution is (or tends to) a fixed finite number $\lambda\,(b-a)$. The expected value of our binomial distribution is $np(\Delta t) = \big((b-a)/\Delta t\big)p(\Delta t) = \frac{p(\Delta t)}{\Delta t}\,(b-a)$, so the proposal is that $p(\Delta t)$ should tend to 0 in such a way that $\dfrac{p(\Delta t)}{\Delta t} \to \lambda$ where λ is a positive constant.



Poisson distribution, $\mu = 2.163$

The parameter λ is the rate or intensity of the occurrences in question, and its dimension is number pr. time. In demography and actuarial mathematics the picturesque term *force of mortality* is sometimes used for such a parameter.

According to Theorem 2.15, when going to the limit in the way just described, the binomial probabilities converge to Poisson probabilities.

Definition 2.10: Poisson distribution
The Poisson distribution with parameter $\mu > 0$ is the distribution on the non-negative integers $\mathbb{N}_0$ given by the probability function

$$f(x) = \frac{\mu^x}{x!}\exp(-\mu), \quad x \in \mathbb{N}_0.$$

(That f adds up to unity follows easily from the power series expansion of the exponential function.)

Power series expansion of the exponential function
For all (real and complex) t, $\exp(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!}$.

The expected value of the Poisson distribution with parameter μ is μ, and the variance is μ as well, so $\operatorname{Var} X = \operatorname{E} X$ for a Poisson random variable X. — Previously we saw that for a binomial random variable X with parameters n and p, $\operatorname{Var} X = (1-p)\operatorname{E} X$, i.e. $\operatorname{Var} X < \operatorname{E} X$, and for a negative binomial random variable X with parameters k and p, $p \operatorname{Var} X = \operatorname{E} X$, i.e. $\operatorname{Var} X > \operatorname{E} X$.

THEOREM 2.15
When $n \to \infty$ and $p \to 0$ simultaneously in such a way that np converges to $\mu > 0$, the binomial distribution with parameters n and p converges to the Poisson distribution with parameter μ in the sense that

$$\binom{n}{x} p^x (1-p)^{n-x} \longrightarrow \frac{\mu^x}{x!} \exp(-\mu)$$

for all $x \in \mathbb{N}_0$.

PROOF
Consider a fixed $x \in \mathbb{N}_0$. Then

$$\begin{aligned}\binom{n}{x} p^x (1-p)^{n-x} &= \frac{n(n-1)\ldots(n-x+1)}{x!}\, p^x\, (1-p)^{-x}\, (1-p)^n \\ &= \frac{1(1-\frac{1}{n})\ldots(1-\frac{x-1}{n})}{x!}\, (np)^x\, (1-p)^{-x}\, (1-p)^n\end{aligned}$$

where the long fraction converges to $1/x!$, $(np)^x$ converges to μ^x, and $(1-p)^{-x}$ converges to 1. The limit of $(1-p)^n$ is most easily found by passing to logarithms; we get

$$\ln\Big((1-p)^n\Big) = n\ln(1-p) = -np \cdot \frac{\ln(1-p) - \ln(1)}{-p} \longrightarrow -\mu$$

since the fraction converges to the derivative of $\ln(t)$ evaluated at $t = 1$, which is 1. Hence $(1-p)^n$ converges to $\exp(-\mu)$, and the theorem is proved. □

PROPOSITION 2.16
If X_1 and X_2 are independent Poisson random variables with parameters μ_1 and μ_2, then $X_1 + X_2$ follows a Poisson distribution with parameter $\mu_1 + \mu_2$.

PROOF
Let $Y = X_1 + X_2$. Then using Proposition 1.11 on page 29

$$\begin{aligned}\mathrm{P}(Y = y) &= \sum_{x=0}^{y} \mathrm{P}(X_1 = y - x)\, \mathrm{P}(X_2 = x) \\ &= \sum_{x=0}^{y} \frac{\mu_1^{y-x}}{(y-x)!} \exp(-\mu_1)\, \frac{\mu_2^x}{x!} \exp(-\mu_2) \\ &= \frac{1}{y!} \exp\Big(-(\mu_1+\mu_2)\Big) \sum_{x=0}^{y} \binom{y}{x} \mu_1^{y-x} \mu_2^x \\ &= \frac{1}{y!} \exp\Big(-(\mu_1+\mu_2)\Big) (\mu_1+\mu_2)^y.\end{aligned}$$

□

2.4 Exercises

Exercise 2.1
Give a proof of Proposition 2.4 on page 49.

Exercise 2.2
In the definition of variance of a distribution on a countable probability space (Definition 2.6 page 53), it is tacitly assumed that $\mathrm{E}\big((X - \mathrm{E}\,X)^2\big) = \mathrm{E}(X^2) - (\mathrm{E}\,X)^2$. Prove that this is in fact true.

Exercise 2.3
Prove that $\operatorname{Var} X = \mathrm{E}(X(X-1)) - \xi(\xi - 1)$, where $\xi = \mathrm{E}\,X$.

Exercise 2.4
Sketch the probability function of the Poisson distribution for different values of the parameter μ.

Exercise 2.5
Find the expected value and the variance of the Poisson distribution with parameter μ. (In doing so, Exercise 2.3 may be of some use.)

Exercise 2.6
Show that if X_1 and X_2 are independent Poisson random variables with parameters μ_1 and μ_2, then the conditional distribution of X_1 given $X_1 + X_2$ is a binomial distribution.

Exercise 2.7
Sketch the probability function of the negative binomial distribution for different values of the parameters k and p.

Exercise 2.8
Consider the negative binomial distribution with parameters k and p, where $k > 0$ and $p \in]0\,;1[$. As mentioned in the text, the expected value is $k(1-p)/p$ and the variance is $k(1-p)/p^2$.

Explain that the parameters k and p can go to a limit in such a way that the expected value is constant and the variance converges towards the expected value, and show that in this case the probability function of the negative binomial distribution converges to the probability function of a Poisson distribution.

Exercise 2.9
Show that the definition of covariance (Definition 2.7 on page 53) is meaningful, i.e. show that when X and Y both are known to have a variance, then the expectation $\mathrm{E}\big((X - \mathrm{E}\,X)(Y - \mathrm{E}\,Y)\big)$ exists.

Exercise 2.10
Assume that on given day the number of persons going by the S-train without a valid ticket, follows a Poisson distribution with parameter μ. Suppose there is a probability of p that a fare dodger is caught by the ticket controller, and that different dodgers are caught independently of each other. In view of this, what can be said about the number of fare dodgers caught?

In order to make statements about the actual number of passengers without a ticket when knowing the number of passengers caught without at ticket, what more information will be needed?

Exercise 2.11: St. Petersburg paradox
A gambler is going to join a game where in each play he has an even chance of doubling the bet and of losing the bet, that is, if he bets k € he will win $2k$ € with probability $1/2$ and lose his bet with probability $1/2$. Our gambler has a clever strategy: in the first play he bets 1 €; if he loses in any given play, then he will bet the double amount in the next play; if he wins in any given play, he will collect his winnings and go home.

How big will his net winnings be?

How many plays will he be playing before he can get home?

The strategy is certain to yield a profit, but surely the player will need some working capital. How much money will he need to get through to the win?

Exercise 2.12
Give a proof of Proposition 2.7 on page 52.

Exercise 2.13
The expected value of the product of two independent random variables can be expressed in a simple way by means of the expected values of the individual variables (Proposition 2.7).

Deduce a similar result about the variance of two independent random variables. — Is independence a crucial assumption?

3 Continuous Distributions

In the previous chapters we met random variables and/or probability distributions that are actually "living" on a finite or countable subset of the real numbers; such variables and/or distributions are often labelled *discrete*. There are other kinds of distributions, however, distributions with the property that all non-empty open subsets of a given fixed interval have strictly positive probability, whereas every single point in the interval has probability 0.

Example: When playing "spin the bottle" you rotate a bottle to pick a random direction, and the person pointed at by the bottle neck must do whatever is agreed on as part of the play. We shall spoil the fun straight away and model the spinning bottle as "a point on the unit circle, picked at random from a uniform distribution"; for the time being, the exact meaning of this statement is not entirely well-defined, but for one thing it must mean that all arcs of a given length b are assigned the same amount of probability, no matter where on the circle the arc is placed, and it seems to be a fair guess to claim that this probability must be $b/2\pi$. Every point on the circle is contained in arcs of arbitrarily small length, and hence the point itself must have probability 0.

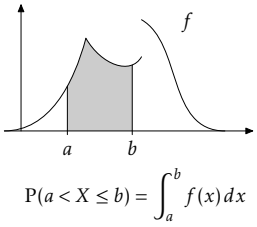
One can prove that there is a unique corresponcence between on the one hand probability measures on $\mathbb{R}$ and on the other hand distribution functions, i.e. functions that are increasing and right-continuous and with limits 0 and 1 at $-\infty$ and $+\infty$ respectively, cf. Proposition 1.6 and/or Proposition 5.3. In brief, the corresponcence is that $\mathrm{P}(]a\,;b]) = F(b) - F(a)$ for every half-open interval $]a\,;b]$. — Now there exist very strange functions meeting the conditions for being a distribution function, but never fear, in this chapter we shall be dealing with a class of very nice distributions, the so-called continuous distributions, distributions with a density function.

3.1 Basic definitions

Definition 3.1: Probability density function
A probability density function on $\mathbb{R}$ *is an integrable function* $f : \mathbb{R} \to [0\,;+\infty[$ *with integral* 1, *i.e.* $\int_{\mathbb{R}} f(x)\,dx = 1$.

A probability density function on $\mathbb{R}^d$ is an integrable function $f : \mathbb{R}^d \to [0\,;+\infty[$ with integral 1, i.e. $\int_{\mathbb{R}^d} f(x)\,dx = 1$.

Definition 3.2: Continuous distribution
A continuous distribution is a probability distribution with a probability density function: If f is a probability density function on $\mathbb{R}$, then the continuous function $F(x) = \int_{-\infty}^{x} f(u)\,du$ is the distribution function for a continuous distribution on $\mathbb{R}$.



We say that the random variable X has density function f if the distribution function $F(x) = P(X \le x)$ of X has the form $F(x) = \int_{-\infty}^{x} f(u)\,du$. Recall that in this case $F'(x) = f(x)$ at all points of continuity x of f.

Proposition 3.1
Consider a random variable X with density function f. Then for all $a < b$

1. $P(a < X \le b) = \int_a^b f(x)\,dx$.
2. $P(X = x) = 0$ *for all* $x \in \mathbb{R}$.
3. $P(a < X < b) = P(a < X \le b) = P(a \le X < b) = P(a \le X \le b)$.
4. *If f is continuous at x, then $f(x)\,dx$ can be understood as the probability that X belongs to an infinitesimal interval of length dx around x.*

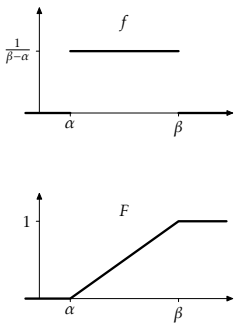
Proof
Item 1 follows from the definitions of distribution function and density function. Item 2 follows from the fact that

$$0 \le P(X = x) \le P(x - \tfrac{1}{n} < X \le x) = \int_{x-1/n}^{x} f(u)\,du,$$

where the integral tends to 0 as n tends to infinity. Item 3 follows from items 1 and 2. A proper mathematical formulation of item 4 is that if x is a point of continuity of f, then as $\varepsilon \to 0$,

$$\frac{1}{2\varepsilon} P(x - \varepsilon < X < x + \varepsilon) = \frac{1}{2\varepsilon} \int_{x-\varepsilon}^{x+\varepsilon} f(u)\,du \longrightarrow f(x).$$

□



Density function f and distribution function F for the uniform distribution on the interval from α to β.

Example 3.1: Uniform distribution
The uniform distribution on the interval from α to β (where $-\infty < \alpha < \beta < +\infty$) is the distribution with density function

$$f(x) = \begin{cases} \dfrac{1}{\beta - \alpha} & \text{for } \alpha < x < \beta, \\ 0 & \text{otherwise.} \end{cases}$$

Its distribution function is the function

$$F(x) = \begin{cases} 0 & \text{for } x \le \alpha \\ \dfrac{x-\alpha}{\beta-\alpha} & \text{for } x_0 < x \le \beta \\ 1 & \text{for } x \ge \beta. \end{cases}$$

A uniform distribution on an interval can, by the way, be used to construct the above-mentioned uniform distribution on the unit circle: simply move the uniform distribution on $]0\,;2\pi[$ onto the unit circle.

Definition 3.3: Multivariate continuous distribution
The d-dimensional random variable $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)$ is said to have density function f, if f is a d-dimensional density function such that for all intervals $K_i =]a_i\,;b_i]$, $i = 1, 2, \ldots, d$,

$$\mathrm{P}\left(\bigcap_{i=1}^{d}\{X_i \in K_i\}\right) = \int\limits_{K_1\times K_2\times\ldots\times K_d} f(x_1, x_2, \ldots, x_d)\,dx_1\,dx_2\ldots dx_d.$$

The function f is called the joint density function for $X_1, X_2, \ldots, X_n$.

A number of definitions, propositions and theorems are carried over more or less immediately from the discrete to the continuous case, formally by replacing probability functions with density functions and sums with integrals. Examples:

The counterpart to Proposition 1.8/Corollary 1.9 is

Proposition 3.2
The random variables $X_1, X_2, \ldots, X_n$ are independent, if and only if their joint density function is a product of density functions for the individual X_is.

The counterpart to Proposition 1.11 is

Proposition 3.3
If the random variables X_1 and X_2 are independent and with density functions f_1 and f_2, then the density function of $Y = X_1 + X_2$ is

$$f(y) = \int_{\mathbb{R}} f_1(x)\,f_2(y-x)\,dx, \quad y \in \mathbb{R}.$$

(See Exercise 3.6.)

Transformation of distributions

Transformation of distributions is still about deducing the distribution of $Y = t(X)$ when the distribution of X is known and t is a function defined on the range of X. The basic formula $\mathrm{P}(t(X) \le y) = \mathrm{P}\big(X \in t^{-1}(]-\infty\,;y])\big)$ always applies.

Example 3.2
We wish to find the distribution of $Y = X^2$, assuming that X is uniformly distributed on $]-1\,;1[$.

For $y \in [0\,;1]$ we have $\mathrm{P}(X^2 \le y) = \mathrm{P}\big(X \in [-\sqrt{y}\,;\sqrt{y}]\big) = \displaystyle\int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{2}\,dx = \sqrt{y}$, so the distribution function of Y is

$$F_Y(y) = \begin{cases} 0 & \text{for } y \le 0, \\ \sqrt{y} & \text{for } 0 < y \le 1, \\ 1 & \text{for } y > 1. \end{cases}$$

A density function f for Y can be found by differentiating F (only at points where F is differentiable):

$$f(y) = \begin{cases} \frac{1}{2}y^{-1/2} & \text{for } 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

In case of a differentiable function t, we can sometimes write down an explicit formula relating the density function of X and the density function of $Y = t(X)$, simply by applying a standard result about substitution in integrals:

Theorem 3.4
Suppose that X is a real random variable, I an open subset of $\mathbb{R}$ such that $\mathrm{P}(X \in I) = 1$, and $t : I \to \mathbb{R}$ a C^1-function defined on I and with $t' \neq 0$ everywhere on I. If X has density function f_X, then the density function of $Y = t(X)$ is

$$f_Y(y) = f_X(x)\left|t'(x)\right|^{-1} = f_X(x)\left|(t^{-1})'(y)\right|$$

where $x = t^{-1}(y)$ and $y \in t(I)$. The formula is sometimes written as

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|.$$

The result also comes in a multivariate version:

Theorem 3.5
Suppose that $\boldsymbol{X}$ is a d-dimensional random variable, I an open subset of $\mathbb{R}^d$ such that $\mathrm{P}(\boldsymbol{X} \in I) = 1$, and $\boldsymbol{t} : I \to \mathbb{R}^d$ a C^1-function defined on I and with a Jacobian $D\boldsymbol{t} = \dfrac{d\boldsymbol{y}}{d\boldsymbol{x}}$ which is regular everywhere on I. If $\boldsymbol{X}$ has density function $f_{\boldsymbol{X}}$, then the density function of $\boldsymbol{Y} = \boldsymbol{t}(\boldsymbol{X})$ is

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(\boldsymbol{x})\left|\det D\boldsymbol{t}(\boldsymbol{x})\right|^{-1} = f_{\boldsymbol{X}}(\boldsymbol{x})\left|\det D\boldsymbol{t}^{-1}(\boldsymbol{y})\right|$$

where $\boldsymbol{x} = \boldsymbol{t}^{-1}(\boldsymbol{y})$ *and* $\boldsymbol{y} \in \boldsymbol{t}(I)$. *The formula is sometimes written as*

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(\boldsymbol{x}) \left| \det \frac{d\boldsymbol{x}}{d\boldsymbol{y}} \right|.$$

Example 3.3
If X is uniformly distributed on $]0\,;1[$, then the density function of $Y = -\ln X$ is

$$f_Y(y) = 1 \cdot \left| \frac{dx}{dy} \right| = \exp(-y)$$

when $0 < x < 1$, i.e. when $0 < y < +\infty$, and $f_Y(y) = 0$ otherwise. — Thus, the distribution of Y is an exponential distribution, see Section 3.3.

Proof of Theorem 3.4
The theorem is a rephrasing of a result from Mathematical Analysis about substitution in integrals. The derivative of the function t is continuous and everywhere non-zero, and hence t is either strictly increasing or strictly decreasing; let us assume it to be strictly increasing. Then $Y \le b \Leftrightarrow X \le t^{-1}(b)$, and we obtain the following expression for the distribution function F_Y of Y:

$$\begin{aligned} F_Y(b) &= \mathrm{P}(Y \le b) \\ &= \mathrm{P}\Big(X \le t^{-1}(b)\Big) \\ &= \int_{-\infty}^{t^{-1}(b)} f_X(x)\,dx \qquad [\text{ substitute } x = t^{-1}(y) \text{ }] \\ &= \int_{-\infty}^{b} f_X\Big(t^{-1}(y)\Big)(t^{-1})'(y)\,dy. \end{aligned}$$

This shows that the function $f_Y(y) = f_X\Big(t^{-1}(y)\Big)(t^{-1})'(y)$ has the property that $F_Y(y) = \displaystyle\int_{-\infty}^{y} f_Y(u)\,du$, showing that Y has density function f_Y. □

Conditioning

When searching for the conditional distribution of one continuous random variable X given another continuous random variable Y, one cannot take for granted that an expression such as $\mathrm{P}(X = x \mid Y = y)$ is well-defined and equal to $\mathrm{P}(X = x, Y = y)/\mathrm{P}(Y = y)$; on the contrary, the denominator $\mathrm{P}(Y = y)$ always equals zero, and so does the nominator in most cases.

Instead, one could try to condition on a suitable event with non-zero probability, for instance $y - \varepsilon < Y < y + \varepsilon$, and then examine whether the (now well-defined) conditional probability $\mathrm{P}(X \le x \mid y - \varepsilon < Y < y + \varepsilon)$ has a limiting value as $\varepsilon \to 0$, in which case the limiting value would be a candidate for $\mathrm{P}(X \le x \mid Y = y)$.

Another, more heuristic approach, is as follows: let f_{XY} denote the joint density function of X and Y and let f_Y denote the density function of Y. Then the probability that (X, Y) lies in an infinitesimal rectangle with sides dx and dy around (x, y) is $f_{XY}(x, y)\,dx\,dy$, and the probability that Y lies in an infinitesimal interval of length dy around y is $f_Y(y)\,dy$; hence the conditional probability that X lies in an infinitesimal interval of length dx around x, given that Y lies in an infinitesimal interval of length dy around y, is $\dfrac{f_{XY}(x,y)\,dx\,dy}{f_Y(y)\,dy} = \dfrac{f_{XY}(x,y)}{f_Y(y)}\,dx$, so a guess is that the conditional density is

$$f(x \mid y) = \frac{f_{XY}(x,y)}{f_Y(y)}.$$

This is in fact true in many cases.

The above discussion of conditional distributions is of course very careless indeed. It is however possible, in a proper mathematical setting, to give an exact and thorough treatment of conditional distributions, but we shall not do so in this presentation.

3.2 Expectation

The definition of expected value of a random variable with density function resembles the definition pertaining to random variables on a countable sample space, the sum is simply replaced by an integral:

Definition 3.4: Expected value
Consider a random variable X with density function f.
If $\displaystyle\int_{-\infty}^{+\infty} |x| f(x)\,dx < +\infty$, then X is said to have an expectation, and the expected value of X is defined to be the real number $\mathrm{E}\,X = \displaystyle\int_{-\infty}^{+\infty} x f(x)\,dx$.

Theorems, propositions and rules of arithmetic for expected value and variance/covariance are unchanged from the countable case.

3.3 Examples

Exponential distribution

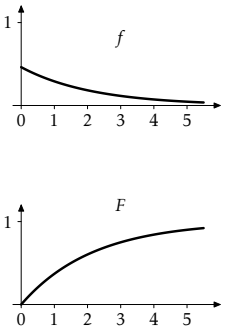
Definition 3.5: Exponential distribution
The exponential distribution with scale parameter $\beta > 0$ is the distribution with

density function

$$f(x) = \begin{cases} \frac{1}{\beta}\exp(-x/\beta) & \text{for } x > 0 \\ 0 & \text{for } x \le 0. \end{cases}$$

Its distribution function is

$$F(x) = \begin{cases} 1 - \exp(-x/\beta) & \text{for } x > 0 \\ 0 & \text{for } x \le 0. \end{cases}$$



Density function and distribution function of the exponential distribution with $\beta = 2.163$.

The reason why β is a *scale* parameter is evident from

Proposition 3.6
If X is exponentially distributed with scale parameter β, and a is a positive constant, then aX is exponentially distributed with scale parameter $a\beta$.

Proof
Applying Theorem 3.4, we find that the density function of $Y = aX$ is the function f_Y defined as

$$f_Y(y) = \begin{cases} \frac{1}{\beta}\exp\bigl(-(y/a)/\beta\bigr)\cdot\frac{1}{a} = \frac{1}{a\beta}\exp\bigl(-y/(a\beta)\bigr) & \text{for } y > 0 \\ 0 & \text{for } y \le 0. \end{cases}$$

□

Proposition 3.7
If X is exponentially distributed with scale parameter β, then $\mathrm{E}\,X = \beta$ and $\mathrm{Var}\,X = \beta^2$.

Proof
As β is a scale parameter, it suffices to show the assertion in the case $\beta = 1$, and this is done using straightforward calculations (the last formula in the side-note (on page 70) about the gamma function may be useful). □

The exponential distribution is often used when modelling waiting times between events occurring "entirely at random". The exponential distribution has a "lack of memory" in the sense that the probability of having to wait more than $s + t$, given that you have already waited t, is the same as the probability of waiting more that t; in other words, for an exponential random variable X,

$$\mathrm{P}(X > s + t \mid X > s) = \mathrm{P}(X > t), \qquad s, t > 0. \tag{3.1}$$

Proof of equation (3.1)
Since $X > s + t \Rightarrow X > s$, we have $\{X > s + t\} \cap \{X > s\} = \{X > s + t\}$, and hence

$$\mathrm{P}(X > s + t \mid X > s) = \frac{\mathrm{P}(X > s + t)}{\mathrm{P}(X > s)} = \frac{1 - F(s + t)}{1 - F(s)}$$

$$= \frac{\exp(-(s+t)/\beta)}{\exp(-s/\beta)} = \exp(-t/\beta) = \mathrm{P}(X > t)$$

for any $s, t > 0$. □

In this sense the exponential distribution is the continuous counterpart to the geometric distribution, see page 55.

Furthermore, the exponential distribution is related to the Poisson distribution. Suppose that instances of a certain kind of phenomenon occur at random time points as described in the section about the Poisson distribution (page 57). Saying that you have to wait more that t for the first instance to occur, is the same as saying that there is no occurrences in the time interval from 0 to t; the latter event is an event in the Poisson model, and the probability of that event is $\frac{(\lambda t)^0}{0!}\exp(-\lambda t) = \exp(-\lambda t)$, showing that the waiting time to the occurrence of the first instance is exponentially distributed with scale parameter $1/\lambda$.

Gamma distribution

THE GAMMA FUNCTION
The gamma function is the function

$$\Gamma(t) = \int_0^{+\infty} x^{t-1}\exp(-x)\,dx$$

where $t > 0$. (Actually, it is possible to define $\Gamma(t)$ for all $t \in \mathbb{C} \setminus \{0, -1, -2, -3, \ldots\}$.)

The following holds:

- $\Gamma(1) = 1$,
- $\Gamma(t+1) = t\,\Gamma(t)$ for all t,
- $\Gamma(n) = (n-1)!$ for $n = 1, 2, 3, \ldots$.
- $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.
  (See also page 74.)

A standard integral substitution leads to the formula

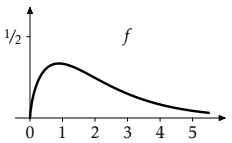
$$\Gamma(k)\,\beta^k = \int_0^{+\infty} x^{k-1}\exp(-x/\beta)\,dx$$

for $k > 0$ and $\beta > 0$.

DEFINITION 3.6: GAMMA DISTRIBUTION
The gamma distribution with shape parameter $k > 0$ and scale parameter $\beta > 0$ is the distribution with density function

$$f(x) = \begin{cases} \dfrac{1}{\Gamma(k)\,\beta^k}\,x^{k-1}\exp(-x/\beta) & \text{for } x > 0 \\ 0 & \text{for } x \le 0. \end{cases}$$

For $k = 1$ this is the exponential distribution.



Gamma distribution with $k = 1.7$ and $\beta = 1.27$

PROPOSITION 3.8
If X is gamma distributed with shape parameter $k > 0$ and scale parameter $\beta > 0$, and if $a > 0$, then $Y = aX$ is gamma distributed with shape parameter k and scale parameter $a\beta$.

PROOF
Similar to the proof of Proposition 3.6, i.e. using Theorem 3.4. □

PROPOSITION 3.9
If X_1 and X_2 are independent gamma variables with the same scale parameter β and with shape parameters k_1 and k_2 respectively, then $X_1 + X_2$ is gamma distributed with scale parameter β and shape parameter $k_1 + k_2$.

Proof
From Proposition 3.3 the density function of $Y = X_1 + X_2$ is (for $y > 0$)

$$f(y) = \int_0^y \frac{1}{\Gamma(k_1)\beta^{k_1}} x^{k_1-1} \exp(-x/\beta) \cdot \frac{1}{\Gamma(k_2)\beta^{k_2}} (y-x)^{k_2-1} \exp(-(y-x)/\beta)\, dx,$$

and the substitution $u = x/y$ transforms this into

$$f(y) = \left(\frac{1}{\Gamma(k_1)\Gamma(k_2)\,\beta^{k_1+k_2}} \int_0^1 u^{k_1-1}(1-u)^{k_2-1}\, du \right) y^{k_1+k_2-1} \exp(-y/\beta),$$

so the density is of the form $f(y) = \text{const} \cdot y^{k_1+k_2-1} \exp(-y/\beta)$ where the constant makes f integrate to 1; the last formula in the side note about the gamma function gives that the constant must be $\dfrac{1}{\Gamma(k_1+k_2)\,\beta^{k_1+k_2}}$. This completes the proof. □

A consequence of Proposition 3.9 is that the expected value and the variance of a gamma distribution must be linear functions of the shape parameter, and also that the expected value must be linear in the scale parameter and the variance must be quadratic in the scale parameter. Actually, one can prove that

Proposition 3.10
If X is gamma distributed with shape parameter k and scale parameter β, then $\mathrm{E}\,X = k\beta$ *and* $\mathrm{Var}\,X = k\beta^2$.

In mathematical statistics as well as in applied statistics you will encounter a special kind of gamma distribution, the so-called χ^2 distribution.

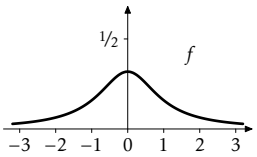
Definition 3.7: χ^2 distribution
A χ^2 distribution with n degrees of freedom is a gamma distribution with shape parameter $n/2$ and scale parameter 2, i.e. with density

$$f(x) = \frac{1}{\Gamma(n/2)\,2^{n/2}} x^{n/2-1} \exp(-\tfrac{1}{2}x), \quad x > 0.$$

Cauchy distribution

Definition 3.8: Cauchy distribution
The Cauchy distribution with scale parameter $\beta > 0$ is the distribution with density function

$$f(x) = \frac{1}{\pi\beta} \frac{1}{1 + (x/\beta)^2}, \quad x \in \mathbb{R}.$$



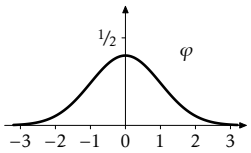
Cauchy distribution with $\beta = 1$

This distribution is often used as a counterexample (!) — it is, for one thing, an example of a distribution that does not have an expected value (since the function $x \mapsto |x|/(1+x^2)$ is not integrable). Yet there are "real world" applications of the Cauchy distribution, one such is given in Exercise 3.4.

Normal distribution

Definition 3.9: Standard normal distribution
The standard normal distribution is the distribution with density function

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}x^2\right), \quad x \in \mathbb{R}.$$



Standard normal distribution

The distribution function of the standard normal distribution is

$$\Phi(x) = \int_{-\infty}^{x} \varphi(u)\,du, \quad x \in \mathbb{R}.$$

Definition 3.10: Normal distribution
The normal distribution (or Gaussian distribution) with location parameter μ and quadratic scale parameter $\sigma^2 > 0$ is the distribution with density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right), \qquad x \in \mathbb{R}.$$

Carl Friedrich Gauss
German mathematician (1777-1855). Known among other things for his work in geometry and mathematical statistics (including the method of least squares) He also dealt with practical applications of mathematics, e.g. geodecy. German 10 DM bank notes from the decades prededing the shift to euro (in 2002), had a portrait of Gauss, the graph of the normal density, and a sketch of his triangulation of a region in northern Germany.

The terms location parameter and quadratic scale parameter are justified by Proposition 3.11, which can be proved using Theorem 3.4:

Proposition 3.11
If X follows a normal distribution with location parameter μ and quadratic scale parameter σ^2, and if a and b are real numbers and $a \neq 0$, then $Y = aX + b$ follows a normal distribution with location parameter $a\mu + b$ and quadratic scale parameter $a^2\sigma^2$.

Proposition 3.12
If X_1 and X_2 are independent normal variables with parameters μ_1, σ_1^2 and μ_2, σ_2^2 respectively, then $Y = X_1 + X_2$ is normally distributed with location parameter $\mu_1 + \mu_2$ and quadratic scale parameter $\sigma_1^2 + \sigma_2^2$.

Proof
Due to Proposition 3.11 it suffices to show that if X_1 has a standard normal distribution, and if X_2 is normal with location parameter 0 and quadratic scale parameter σ^2, then $X_1 + X_2$ is normal with location parameter 0 and quadratic scale parameter $1 + \sigma^2$. That is shown using Proposition 3.3. □

Proof that φ is a density function

We need to show that φ is in fact a probability density function, that is, that $\int_{\mathbb{R}} \varphi(x)\,dx = 1$, or in other words, if c is defined to be the constant that makes $f(x) = c\exp(-\frac{1}{2}x^2)$ a probability density, then we have to show that $c = 1/\sqrt{2\pi}$.

The clever trick is to consider two independent random variables X_1 and X_2, each with density f. Their joint density function is

$$f(x_1)\,f(x_2) = c^2 \exp\Big(-\tfrac{1}{2}(x_1^2 + x_2^2)\Big).$$

Now consider a function $(y_1, y_2) = t(x_1, x_2)$ where the relation between xs and ys is given as $x_1 = \sqrt{y_2}\cos y_1$ and $x_2 = \sqrt{y_2}\sin y_1$ for $(x_1, x_2) \in \mathbb{R}^2$ and $(y_1, y_2) \in \,]0\,;2\pi[\,\times\,]0\,;+\infty[$. Theorem 3.5 shows how to write down an expression for the density function of the two-dimensional random variable $(Y_1, Y_2) = t(X_1, X_2)$; from standard calculations this density is $\frac{1}{2}c^2\exp(-\frac{1}{2}y_2)$ when $0 < y_1 < 2\pi$ and $y_2 > 0$, and 0 otherwise. Being a probability density function, it integrates to 1:

$$\begin{aligned} 1 &= \int_0^{+\infty}\int_0^{2\pi} \tfrac{1}{2}c^2\exp(-\tfrac{1}{2}y_2)\,dy_1\,dy_2 \\ &= 2\pi\,c^2 \int_0^{+\infty} \tfrac{1}{2}\exp(-\tfrac{1}{2}y_2)\,dy_2 \\ &= 2\pi\,c^2, \end{aligned}$$

and hence $c = 1/\sqrt{2\pi}$. □

Proposition 3.13

If $X_1, X_2, \ldots, X_n$ are independent standard normal random variables, then $Y = X_1^2 + X_2^2 + \ldots + X_n^2$ follows a χ^2 distribution with n degrees of freedom.

Proof

The χ^2 distribution is a gamma distribution, so it suffices to prove the claim in the case $n = 1$ and then refer to Proposition 3.9. The case $n = 1$ is treated as follows: For $x > 0$,

$$\begin{aligned} \mathrm{P}(X_1^2 \le x) &= \mathrm{P}(-\sqrt{x} < X_1 \le \sqrt{x}) \\ &= \int_{-\sqrt{x}}^{\sqrt{x}} \varphi(u)\,du \qquad [\ \varphi \text{ is an even function }] \\ &= 2\int_0^{\sqrt{x}} \varphi(u)\,du \qquad [\ \text{substitute } t = u^2\ ] \\ &= \int_0^x \frac{1}{\sqrt{2\pi}}\,t^{-1/2}\exp(-\tfrac{1}{2}t)\,dt, \end{aligned}$$

and differentiation with respect to t gives this expression for the density function of X_1^2:

$$\frac{1}{\sqrt{2\pi}}\, x^{-1/2} \exp(-\tfrac{1}{2}x), \quad x > 0.$$

The density function of the χ^2 distribution with 1 degree of freedom is (Definition 3.7)

$$\frac{1}{\Gamma(1/2)\, 2^{1/2}}\, x^{-1/2} \exp(-\tfrac{1}{2}x), \quad x > 0.$$

Since the two density functions are proportional, they must be equal, and that concludes the proof. — As a by-product we see that $\Gamma(1/2) = \sqrt{\pi}$. □

Proposition 3.14
If X is normal with location parameter μ and quadratic scale parameter σ^2, then $\mathrm{E}\, X = \mu$ and $\mathrm{Var}\, X = \sigma^2$.

Proof
Referring to Proposition 3.11 and the rules of calculus for expected values and variances, it suffices to consider the case $\mu = 0$, $\sigma^2 = 1$, so assume that $\mu = 0$ and $\sigma^2 = 1$. By symmetri we must have $\mathrm{E}\, X = 0$, and then $\mathrm{Var}\, X = \mathrm{E}((X - \mathrm{E}\, X)^2) = \mathrm{E}(X^2)$; X^2 is known to be χ^2 with 1 degree of freedom, i.e. it follows a gamma distribution with parameters $1/2$ and 2, so $\mathrm{E}(X^2) = 1$ (Proposition 3.10). □

The normal distribution is widely used in statistical models. One reason for this is the Central Limit Theorem. (A quite different argument for and derivation of the normal distribution is given on pages 197ff.)

Theorem 3.15: Central Limit Theorem
Consider a sequence $X_1, X_2, X_3, \ldots$ of independent identically distributed random variables with expected value μ and variance $\sigma^2 > 0$, and let S_n denote the sum of the first n variables, $S_n = X_1 + X_2 + \ldots + X_n$.

Then as $n \to \infty$, $\dfrac{S_n - n\mu}{\sqrt{n\sigma^2}}$ is asymptotically standard normal in the sense that

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \le x\right) = \Phi(x)$$

for all $x \in \mathbb{R}$.

Remarks:

- The quantity $\dfrac{S_n - n\mu}{\sqrt{n\sigma^2}} = \dfrac{\frac{1}{n}S_n - \mu}{\sqrt{\sigma^2/n}}$ is the sum and the average of the first n variables minus the expected value of this, divided by the standard deviation; so the quantity has expected value 0 and variance 1.

- The Law of Large Numbers (page 41) shows that $\frac{1}{n}S_n - \mu$ becomes very small (with very high probability) as n increases. The Central Limit Theorem shows how $\frac{1}{n}S_n - \mu$ becomes very small, namely in such a way that $\frac{1}{n}S_n - \mu$ divided by $\sqrt{\sigma^2/n}$ is a standard normal variable.
- The theorem is very general as it assumes nothing about the distribution of the Xs, except that it has an expected value and a variance.

We omit the proof of the Central Limit Theorem.

3.4 Exercises

Exercise 3.1
Sketch the density function of the gamma distribution for various values of the shape parameter and the scale parameter. (Investigate among other things how the behaviour near $x = 0$ depends on the parameters; you may wish to consider $\lim_{x\to 0} f(x)$ and $\lim_{x\to 0} f'(x)$.)

Exercise 3.2
Sketch the normal density for various values of μ and σ^2.

Exercise 3.3
Consider a continuous and strictly increasing function F defined on an open interval $I \subseteq \mathbb{R}$, and assume that the limiting values of F at the two endpoints of the interval are 0 and 1, respectively. (This means that F, possibly after an extension to all of the real line, is a distribution function.)

1. Let U be a random variable having a uniform distribution on the interval $]0\,;1[$. Show that the distribution function of $X = F^{-1}(U)$ is F.
2. Let Y be a random variable with distribution function F. Show that $V = F(Y)$ is uniform on $]0\,;1[$.

Remark: Most computer programs come with a function that returns random numbers (or at least pseudo-random numbers) from the uniform distribution on $]0\,;1[$. This exercise indicates one way to transform uniform random numbers into random numbers from any given continuous distribution.

Exercise 3.4
A two-dimensional cowboy fires his gun in a random direction. There is a fence at some distance from the cowboy; the fence is very long (infinitely long?).

1. Find the probability that the bullet hits the fence.
2. Given that the bullet hits the fence, what can be said of distribution of the point of hitting?

Exercise 3.5
Consider a two-dimensional random variable (X_1, X_2) with density function $f(x_1, x_2)$, cf. Definition 3.3. Show that the function $f_1(x_1) = \int_{\mathbb{R}} f(x_1, x_2)\,dx_2$ is a density function of X_1.

Exercise 3.6
Let X_1 and X_2 be independent random variables with density functions f_1 and f_2, and let $Y_1 = X_1 + X_2$ and $Y_2 = X_1$.

1. Find the density function of (Y_1, Y_2). (Utilise Theorem 3.5.)
2. Find the distribution of Y_1. (You may use Exercise 3.5.)
3. Explain why the above is a proof of Proposition 3.3.

Exercise 3.7
There is a claim on page 71 that it is a consequence of Proposition 3.9 that the expected value and the variance of a gamma distribution must be linear functions of the shape parameter, and that the expected value must be linear in the scale parameter and the variance quadratic in the scale parameter. — Explain why this is so.

Exercise 3.8
The mercury content in swordfish sold in some parts of the US is known to be normally distributed with mean 1.1 ppm and variance $0.25\,\text{ppm}^2$ (see e.g. Lee and Krutchkoff (1980) and/or exercises 5.2 and 5.3 in Larsen (2006)), but according to the health authorities the average level of mercury in fish for consumption should not exceed 1 ppm. Fish sold through authorised retailers are always controlled by the FDA, and if the mercury content is too high, the lot is discarded. However, about 25% of the fish caught is sold on the black market, that is, 25% of the fish does not go through the control. Therefore the rule "discard if the mercury content exceeds 1 ppm" is not sufficient.

How should the "discard limit" be chosen to ensure that the average level of mercury in the fish actually bought by the consumers is as small as possible?

4 Generating Functions

In mathematics you will find several instances of constructions of one-to-one correspondences, representations, between one mathematical structure and another. Such constructions can be of great use in argumentations, proofs, calculations etc. A classic and familiar example of this is the logarithm function, which of course turns multiplication problems into addition problems, or stated in another way, it is an isomorphism between $(\mathbb{R}_+, \cdot)$ and $(\mathbb{R}, +)$.

In this chapter we shall deal with a representation of the set of $\mathbb{N}_0$-valued random variables (or more correctly: the set of probability distributions on $\mathbb{N}_0$) using the so-called generating functions.

Note: All random variables in this chapter are random variables with values in $\mathbb{N}_0$.

4.1 Basic properties

Definition 4.1: Generating function
Let X be a $\mathbb{N}_0$-valued random variable. The generating function of X (or rather: the generating function of the distribution of X) is the function

$$G(s) = \mathrm{E}\, s^X = \sum_{x=0}^{\infty} s^x \,\mathrm{P}(X = x), \quad s \in [0\,;1].$$

The theory of generating functions draws to a large extent on the theory of power series. Among other things, the following holds for the generating function G of the random variable X:

1. G is continuous and takes values in $[0\,;1]$; moreover $G(0) = \mathrm{P}(X = 0)$ and $G(1) = 1$.
2. G has derivatives of all orders, and the kth derivative is

$$\begin{aligned} G^{(k)}(s) &= \sum_{x=k}^{\infty} x(x-1)(x-2)\ldots(x-k+1)\, s^{x-k}\, \mathrm{P}(X = x) \\ &= \mathrm{E}\Big(X(X-1)(X-2)\ldots(X-k+1)\, s^{X-k} \Big). \end{aligned}$$

In particular, $G^{(k)}(s) \geq 0$ for all $s \in]0\,;1[$.

3. If we let $s = 0$ in the expression for $G^{(k)}(s)$, we obtain $G^{(k)}(0) = k!\,\mathrm{P}(X = k)$; this shows how to find the probability distribution from the generating function.
   Hence, different probability distributions cannot have the same generating function, that is to say, the map that takes a probability distributions into its generating function is injective.
4. Letting $s \to 1$ in the expression for $G^{(k)}(s)$ gives

$$G^{(k)}(1) = \mathrm{E}\Big(X(X-1)(X-2)\dots(X-k+1)\Big); \tag{4.1}$$

   one can prove that $\lim_{s\to 1} G^{(k)}(s)$ is a finite number if and only if X^k has an expected value, and if so equation (4.1) holds true.
   Note in particular that

$$\mathrm{E}\,X = G'(1) \tag{4.2}$$

   and (using Exercise 2.3)

$$\begin{aligned} \operatorname{Var} X &= G''(1) - G'(1)\Big(G'(1) - 1\Big) \\ &= G''(1) - \Big(G'(1)\Big)^2 + G'(1). \end{aligned} \tag{4.3}$$

An important and useful result is

THEOREM 4.1
If X and Y are independent random variables with generating functions G_X and G_Y, then the generating function of the sum of X and Y is the product of the generating functions: $G_{X+Y} = G_X\,G_Y$.

PROOF
For $|s| < 1$, $G_{X+Y}(s) = \mathrm{E}\,s^{X+Y} = \mathrm{E}\big(s^X s^Y\big) = \mathrm{E}\,s^X\,\mathrm{E}\,s^Y = G_X(s)G_Y(s)$ where we have used a well-known property of exponential functions and then Proposition 1.17/2.7 (on page 35/52). □

Example 4.1: One-point distribution
If $X = a$ with probability 1, then its generating function is $G(s) = s^a$.

Example 4.2: 01-variable
If X is a 01-variable with $\mathrm{P}(X = 1) = p$, then its generating function is $G(s) = 1 - p + sp$.

Example 4.3: Binomial distribution
The binomial distribution with parameters n and p is the distribution of a sum of n independent and identically distributed 01-variables with parameter p (Definition 1.10 on page 30). Using Theorem 4.1 and Example 4.2 we see that the generating function of a binomial random variable Y with parameters n and p therefore is $G(s) = (1 - p + sp)^n$.

In Example 1.20 on page 40 we found the expected value and the variance of the binomial distribution to be np and $np(1-p)$, respectively. Now let us derive these quantities using generating functions: We have

$$G'(s) = np(1-p+sp)^{n-1} \qquad \text{and} \qquad G''(s) = n(n-1)p^2(1-p+sp)^{n-2},$$

so

$$G'(1) = \mathrm{E}\, Y = np \qquad \text{and} \qquad G''(1) = \mathrm{E}(Y(Y-1)) = n(n-1)p^2.$$

Then, according to equations (4.2) and (4.3)

$$\mathrm{E}\, Y = np \qquad \text{and} \qquad \mathrm{Var}\, Y = n(n-1)p^2 - np(np-1) = np(1-p).$$

We can also give a new proof of Theorem 1.12 (page 30): Using Theorem 4.1, the generating function of the sum of two independent binomial variables with parameters n_1, p and n_2, p, respectively, is

$$(1-p+sp)^{n_1} \cdot (1-p+sp)^{n_2} = (1-p+sp)^{n_1+n_2}$$

where the right-hand side is recognised as the generating function of the binomial distribution with parameters $n_1 + n_2$ and p.

Example 4.4: Poisson distribution
The generating function of the Poisson distribution with parameter μ (Definition 2.10 page 58) is

$$G(s) = \sum_{x=0}^{\infty} s^x \frac{\mu^x}{x!} \exp(-\mu) = \exp(-\mu) \sum_{x=0}^{\infty} \frac{(\mu s)^x}{x!} = \exp\Big(\mu(s-1)\Big).$$

Just like in the case of the binomial distribution, the use of generating functions greatly simplifies proofs of important results such as Proposition 2.16 (page 59) about the distribution of the sum of independent Poisson variables.

Example 4.5: Negative binomial distribution
The generating function of a negative binomial variable X (Definition 2.9 page 56) with probability parameter $p \in]0\,;1[$ and shape parameter $k > 0$ is

$$G(s) = \sum_{x=0}^{\infty} \binom{x+k-1}{x} p^k (1-p)^x s^x = p^k \sum_{x=0}^{\infty} \binom{x+k-1}{x} \Big((1-p)s\Big)^x = \left(\frac{1}{p} - \frac{1-p}{p}\, s\right)^{-k}.$$

(We have used the third of the binomial series on page 55 to obtain the last equality.) Then

$$G'(s) = k\frac{1-p}{p}\left(\frac{1}{p} - \frac{1-p}{p}\, s\right)^{-k-1} \quad \text{and} \quad G''(s) = k(k+1)\left(\frac{1-p}{p}\right)^2 \left(\frac{1}{p} - \frac{1-p}{p}\, s\right)^{-k-2}.$$

This is entered into (4.2) and (4.3) and we get

$$\mathrm{E}\, X = k\frac{1-p}{p} \qquad \text{and} \qquad \mathrm{Var}\, X = k\frac{1-p}{p^2},$$

as claimed on page 57.

By using generating functions it is also quite simple to prove Proposition 2.14 (on page 56) concerning the distribution of a sum of negative binomial variables.

As these examples might suggest, it is often the case that once you have written down the generating function of a distribution, then many problems are easily solved; you can, for instance, find the expected value and the variance simply by differentiating twice and do a simple calculation. The drawback is of course, that it can be quite cumbersome to find a usable expression for the generating function.

Earlier in this presentation we saw examples of convergence of probability distributions: the binomial distribution converges in certain cases to a Poisson distribution (Theorem 2.15 on page 59), and so does the negative binomial distribution (Exercise 2.8 on page 60). Convergence of probability distributions on $\mathbb{N}_0$ is closely related to convergence of generating functions:

THEOREM 4.2: CONTINUITY THEOREM
Suppose that for each $n \in \{1, 2, 3, \ldots, \infty\}$ we have a random variable X_n with generating function G_n and probability function f_n. Then $\lim_{n\to\infty} f_n(x) = f_\infty(x)$ for all $x \in \mathbb{N}_0$, if and only if $\lim_{n\to\infty} G_n(s) = G_\infty(s)$ for all $s \in [0\,;1[$.

The proof is omitted (it is an exercise in mathematical analysis rather than in probability theory).

We shall now proceed with some new problems, that is, problems that have not been treated earlier in this presentation.

4.2 The sum of a random number of random variables

THEOREM 4.3
Assume that $N, X_1, X_2, \ldots$ are independent random variables, and that all the Xs have the same distribution. Then the generating function G_Y of $Y = X_1 + X_2 + \ldots + X_N$ is

$$G_Y = G_N \circ G_X\,, \tag{4.4}$$

where G_N is the generating function of N, and G_X is the generating function of the distribution of the Xs.

PROOF
For an outcome $\omega \in \Omega$ whith $N(\omega) = n$, we have $Y(\omega) = X_1(\omega) + X_2(\omega) + \cdots + X_n(\omega)$, i.e., $\{Y = y \text{ and } N = n\} = \{X_1 + X_2 + \cdots + X_n = y \text{ and } N = n\}$, and so

$$G_Y(s) = \sum_{y=0}^{\infty} s^y \,\mathrm{P}(Y = y) = \sum_{y=0}^{\infty} s^y \sum_{n=0}^{\infty} \mathrm{P}(Y = y \text{ and } N = n)$$

y	number of leaves with y mites
0	70
1	38
2	17
3	10
4	9
5	3
6	2
7	1
8+	0
	150

Table 4.1
Mites on apple leaves

$$= \sum_{y=0}^{\infty} s^y \sum_{n=0}^{\infty} \mathrm{P}\Big(X_1 + X_2 + \cdots + X_n = y\Big)\, \mathrm{P}(N = n)$$
$$= \sum_{n=0}^{\infty} \Bigg(\sum_{y=0}^{\infty} s^y\, \mathrm{P}\Big(X_1 + X_2 + \cdots + X_n = y\Big) \Bigg) \mathrm{P}(N = n),$$

where we have made use of the fact that N is independent of $X_1 + X_2 + \cdots + X_n$. The expression in large brackets is nothing but the generating function of $X_1 + X_2 + \cdots + X_n$, which by Theorem 4.1 is $\Big(G_X(s)\Big)^n$, and hence

$$G_Y(s) = \sum_{n=0}^{\infty} \Big(G_X(s)\Big)^n \mathrm{P}(N = n).$$

Here the right-hand side is in fact the generating function of N, evaluated at the point $G_X(s)$, so we have for all $s \in]0\,;1[$

$$G_Y(s) = G_N\Big(G_X(s)\Big),$$

and the theorem is proved. □

Example 4.6: Mites on apple leaves
On the 18th of July 1951, 25 leaves were selected at random from each of six McIntosh apple trees in Connecticut, and for each leaf the number of red mites (female adults) was recorded. The result is shown in Table 4.1 (Bliss and Fisher, 1953).

If mites were scattered over the trees in a haphazard manner, we might expect the number of mites per leave to follow a Poisson distribution. Statisticians have ways to test whether a Poisson distribution gives a reasonable description of a given set of data, and in the present case the answer is that the Poisson distribution does *not* apply. So we must try something else.

One might argue that mites usually are found in clusters, and that this fact should somehow be reflected in the model. We could, for example, consider a model where the

number of mites on a given leaf is determined using a two-stage random mechanism: first determine the number of clusters on the leaf, and then for each cluster determine the number of mites in that cluster. So let N denote a random variable that gives the number of clusters on the leaf, and let $X_1, X_2, \ldots$ be random variables giving the number of individuals in cluster no. $1, 2, \ldots$ What is observed, is 150 values of $Y = X_1 + X_2 + \ldots + X_N$ (i.e., Y is a sum of a random number of random variables).

It is always advisable to try simple models before venturing into the more complex ones, so let us assume that the random variables $N, X_1, X_2, \ldots$ are independent, and that all the Xs have the same distribution. This leaves the distribution of N and the distribution of the Xs as the only unresolved elements of the model.

The Poisson distribution did not apply to the number of *mites* per leaf, but it might apply to the number of *clusters* per leaf. Let us assume that N is a Poisson variable with parameter $\mu > 0$. For the number of individuals per cluster we will use the *logarithmic distribution* with parameter $\alpha \in]0\,;1[$, that is, the distribution with point probabilities

$$f(x) = \frac{1}{-\ln(1-\alpha)} \frac{\alpha^x}{x}, \qquad x = 1, 2, 3, \ldots$$

(a cluster necessarily has a non-zero number of individuals). It is not entirely obvious why to use this particular distribution, except that the final result is quite nice, as we shall see right away.

Theorem 4.3 gives the generating function of Y in terms of the generating functions of N and X. The generating function of N is $G_N(s) = \exp\bigl(\mu(s-1)\bigr)$, see Example 4.4. The generating function of the distribution of X is

$$G_X(s) = \sum_{x=1}^{\infty} s^x f(x) = \frac{1}{-\ln(1-\alpha)} \sum_{x=1}^{\infty} \frac{(\alpha s)^x}{x} = \frac{-\ln(1-\alpha s)}{-\ln(1-\alpha)}.$$

From Theorem 4.3, the generating function of Y is $G_Y(s) = G_N\bigl(G_X(s)\bigr)$, so

$$\begin{aligned} G_Y(s) &= \exp\left(\mu\left(\frac{-\ln(1-\alpha s)}{-\ln(1-\alpha)} - 1\right)\right) \\ &= \exp\left(\frac{\mu}{\ln(1-\alpha)} \ln\frac{1-\alpha s}{1-\alpha}\right) = \left(\frac{1-\alpha s}{1-\alpha}\right)^{\mu/\ln(1-\alpha)} \\ &= \left(\frac{1}{1-\alpha} - \frac{\alpha}{1-\alpha} s\right)^{\mu/\ln(1-\alpha)} = \left(\frac{1}{p} - \frac{1-p}{p} s\right)^{-k} \end{aligned}$$

where $p = 1 - \alpha$ and $k = -\mu/\ln p$. This function is nothing but the generating function of the negative binomial distribution with probability parameter p and shape parameter k (Example 4.5). Hence the number Y of mites per leaf follows a negative binomial distribution.

Corollary 4.4
In the situation described in Theorem 4.3,

$$\begin{aligned} \mathrm{E}\, Y &= \mathrm{E}\, N\, \mathrm{E}\, X \\ \mathrm{Var}\, Y &= \mathrm{E}\, N\, \mathrm{Var}\, X + \mathrm{Var}\, N\, (\mathrm{E}\, X)^2. \end{aligned}$$

Proof
Equations (4.2) and (4.3) express the expected value and the variance of a distribution in terms of the generating function and its derivatives evaluated at 1. Differentiating (4.4) gives $G'_Y = (G'_N \circ G_X)\, G'_X$, which evaluated at 1 gives $\mathrm{E}\, Y = \mathrm{E}\, N\, \mathrm{E}\, X$. Similarly, evaluate G''_Y at 1 and do a little arithmetic to obtain the second of the results claimed. □

4.3 Branching processes

A branching process is a special type of stochastic process.In general, a stochastic process is a family $(X(t) : t \in T)$ of random variables, indexed by a quantity t ("time") varying in a set T that often is $\mathbb{R}$ or $[0\,;+\infty[$ ("continuous time"), or $\mathbb{Z}$ or $\mathbb{N}_0$ ("discret time"). The range of the random variables would often be a subset of either $\mathbb{Z}$ ("discret state space" or $\mathbb{R}$ ("continuous state space").

About the notation
In the first chapters we made a point of random variables being functions defined on a sample space Ω; but gradually both the ωs and the Ω seem to have vanished. So, does the t of "$Y(t)$" in the present chapter replace the ω? No, definitely not. An Ω is still implied, and it would certainly be more correct to write $Y(\omega, t)$ instead of just $Y(t)$. The meaning of $Y(\omega, t)$ is that for each t there is a usual random variable $\omega \mapsto Y(\omega, t)$, and for each ω there is a function $t \mapsto Y(\omega, t)$ of "time" t.

We can introduce a branching process with discrete time $\mathbb{N}_0$ and discrete state space $\mathbb{Z}$ in the following way. Assume that we are dealing with a special kind of individuals that live exactly one time unit: at each time $t = 0, 1, 2, \ldots$, each individual either dies or is split into a random number of new individuals (descendants); all individuals of all generations behave the same way and independently of each other, and the numbers of descendants are always selected from the same probability distribution. If $Y(t)$ denotes the total number of individuals at time t, calculated when all deaths and splittings at time t have taken place, then $(Y(t) : t \in \mathbb{N}_0)$ is an example of a branching process.

Some questions you would ask regarding branching processes are: Given that $Y(0) = y_0$, what can be said of the distribution of $Y(t)$ and of the expected value and the variance of $Y(t)$? What is the probability that $Y(t) = 0$, i.e., that the population is extinct at time t? and how long will it take the population to become extinct?

We shall now state a mathematical model for a branching process and then answer some of the questions.

The so-called *offspring distribution* plays a decisive role in the model. The offspring distribution is the probability distribution that models the number of descendants of each single individual in one time unit (generation). The offspring distribution is a distribution on $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$; the outcome 0 means that the individual (dies and) has no descendants, the value 1 means that the individual (dies and) has exactly one descendant, the value 2 means that the individual (dies and) has exactly two descendants, etc.

Let G be the generating function of the offspring distribution. The initial number of individuals is 1, so $Y(0) = 1$. At time $t = 1$ this individual becomes

$Y(1)$ new individuals, and the generating function of $Y(1)$ is

$$s \mapsto \mathrm{E}\left(s^{Y(1)}\right) = G(s).$$

At time $t = 2$ each of the $Y(1)$ individuals becomes a random number of new ones, so $Y(2)$ is a sum of $Y(1)$ independent random variables, each with generating function G. Using Theorem 4.3, the generating function of $Y(2)$ is

$$s \mapsto \mathrm{E}\left(s^{Y(2)}\right) = (G \circ G)(s).$$

At time $t = 3$ each of the $Y(2)$ individuals becomes a random number of new ones, so $Y(3)$ is a sum of $Y(2)$ independent random variables, each with generating function G. Using Theorem 4.3, the generating function of $Y(3)$ is

$$s \mapsto \mathrm{E}\left(s^{Y(3)}\right) = (G \circ G \circ G)(s).$$

Continuing this reasoning, we see that the generating function of $Y(t)$ is

$$s \mapsto \mathrm{E}\left(s^{Y(t)}\right) = (\underbrace{G \circ G \circ \ldots \circ G}_{t})(s). \tag{4.5}$$

(Note by the way that if the initial number of individuals is y_0, then the generating function becomes $\left((G \circ G \circ \ldots \circ G)(s)\right)^{y_0}$; this is a consequence of Theorem 4.1.)

Let μ denote the expected value of the offspring distribution; recall that $\mu = G'(1)$ (equation (4.2)). Now, since equation (4.5) is a special case of Theorem 4.3, we can use the corollary of that theorem and obtain the hardly surprising result $\mathrm{E}\, Y(t) = \mu^t$, that is, the expected population size either increases or decreases exponentially. Of more interest is

THEOREM 4.5
Assume that the offspring distribution is not degenerate at 1, *and assume that* $Y(0) = 1$.

- *If* $\mu \leq 1$, *then the population will become extinct with probability* 1, *more specifically is*

$$\lim_{t \to \infty} \mathrm{P}(Y(t) = y) = \begin{cases} 1 & \text{for } y = 0 \\ 0 & \text{for } y = 1, 2, 3, \ldots \end{cases}$$

- *If* $\mu > 1$, *then the population will become extinct with probability* $s^\star$, *and will grow indefinitely with probability* $1 - s^\star$, *more specifically is*

$$\lim_{t \to \infty} \mathrm{P}(Y(t) = y) = \begin{cases} s^\star & \text{for } y = 0 \\ 0 & \text{for } y = 1, 2, 3, \ldots \end{cases}$$

Here $s^\star$ *is the solution of* $G(s) = s$, $s < 1$.

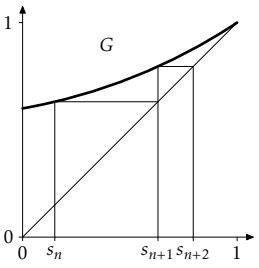
Proof

We can prove the theorem by studying convergence of the generating function of $Y(t)$: for each s_0 we will find the limiting value of $\mathrm{E}\, s_0^{Y(t)}$ as $t \to \infty$. Equation (4.5) shows that the number sequence $\left(\mathrm{E}\, s_0^{Y(t)}\right)_{t\in\mathbb{N}}$ is the same as sequence $(s_n)_{n\in\mathbb{N}}$ determined by

$$s_{n+1} = G(s_n), \quad n = 0, 1, 2, 3, \ldots$$

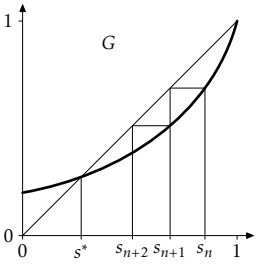
The behaviour of this sequence depends on G and on the initial value s_0. Recall that G and all its derivatives are non-negative on the interval $]0\,;1[$, and that $G(1) = 1$ and $G'(1) = \mu$. By assumption we have precluded the possibility that $G(s) = s$ for all s (which corresponds to the offspring distribution being degenerate at 1), so we may conclude that

- If $\mu \le 1$, then $G(s) > s$ for all $s \in]0\,;1[$.
  The sequence (s_n) is then increasing and thus convergent. Let the limiting value be $\bar{s}$; then $\bar{s} = \lim\limits_{n\to\infty} s_n = \lim\limits_{n\to\infty} G(s_{n+1}) = G(\bar{s})$, where the centre equality follows from the definition of (s_n) and the last equality from G being continuous at $\bar{s}$. The only point $\bar{s} \in [0\,;1]$ such that $G(\bar{s}) = \bar{s}$, is $\bar{s} = 1$, so we have now proved that the sequence (s_n) is increasing and converges to 1.
- If $\mu > 1$, then there is exactly one number $s^\star \in [0\,;1[$ such that $G(s^\star) = s^\star$.
  - If $s^\star < s < 1$, then $s^\star \le G(s) < s$.
    Therefore, if $s^\star < s_0 < 1$, then the sequence (s_n) decreases and hence has a limiting value $\bar{s} \ge s^\star$. As in the case $\mu \le 1$ it follows that $G(\bar{s}) = \bar{s}$, so that $s^\star$.
  - If $0 < s < s^\star$, then $s < G(s) \le s^\star$.
    Therefore, if $0 < s_0 < s^\star$, then the sequence (s_n) increases, and (as above) the limiting value must be $s^\star$.
  - If s_0 is either $s^\star$ or 1, then the sequence (s_n) is identically equal to s_0.



The case $\mu \le 1$

In the case $\mu \le 1$ the above considerations show that the generating function of $Y(t)$ converges to the generating function for the one-point distribution at 0 (viz. the constant function 1), so the probability of ultimate extinction is 1.



The case $\mu > 1$

In the case $\mu > 1$ the above considerations show that the generating function of $Y(t)$ converges to a function that has the value $s^\star$ throughout the interval $[0\,;1[$ and the value 1 at $s = 1$. This function is *not* the generating function of any ordinary probability distribution (the function is not continuous at 1), but one might argue that it could be the generating function of a probability distribution that assigns probability $s^\star$ to the point 0 and probability $1 - s^\star$ to an infinitely distant point.

It is in fact possible to extend the theory of generating functions to cover distributions that place part of the probability mass at infinity, but we shall not

do so in this exposition. Here, we restrict ourselves to showing that

$$\lim_{t\to\infty} \mathrm{P}(Y(t)=0) = s^\star \tag{4.6}$$

$$\lim_{t\to\infty} \mathrm{P}(Y(t)\le c) = s^\star \quad \text{for all } c>0. \tag{4.7}$$

Since $\mathrm{P}(Y(t)=0)$ equals the value of the generating function of $Y(t)$ evaluated at $s=0$, equation (4.6) follows immediately from the preceeding. To prove equation (4.7) we use the following inequality which is obtained from the Markov inequality (Lemma 1.24 on page 36):

$$\mathrm{P}(Y(t)\le c) = \mathrm{P}\left(s^{Y(t)} \ge s^c\right) \le s^{-c}\,\mathrm{E}\,s^{Y(t)}$$

where $0<s<1$. From what is proved earlier, $s^{-c}\,\mathrm{E}\,s^{Y(t)}$ converges to $s^{-c}s^\star$ as $t\to\infty$. By chosing s close enough to 1, we can make $s^{-c}s^\star$ be as close to $s^\star$ as we wish, and therefore $\limsup_{t\to\infty} \mathrm{P}(Y(t)\le c) \le s^\star$. On the other hand, $\mathrm{P}(Y(t)=0) \le \mathrm{P}(Y(t)\le c)$ and $\lim_{t\to\infty} \mathrm{P}(Y(t)=0) = s^\star$, and hence $s^\star \le \liminf_{t\to\infty} \mathrm{P}(Y(t)\le c)$. We may conclude that $\lim_{t\to\infty} \mathrm{P}(Y(t)\le c)$ exists and equals $s^\star$.

Equations (4.6) and (4.7) show that no matter how large the interval $[0\,;c]$ is, in the limit as $t\to\infty$ it will contain no other probability mass than the mass at $y=0$. □

4.4 Exercises

Exercise 4.1
Cosmic radiation is, at least in this exercise, supposed to be particles arriving "totally at random" at some sort of instrument that counts them. We will interpret "totally at random" to mean that the number of particles (of a given energy) arriving in a time interval of length t is Poisson distributed with parameter λt, where λ is the intensity of the radiation.

Suppose that the counter is malfunctioning in such a way that each particle is registered only with a certain probability (which is assumed to be constant in time). What can now be said of the distribution of particles registered?

Exercise 4.2
Use Theorem 4.2 to show that when $n\to\infty$ and $p\to 0$ in such a way that $np\to\mu$, then the binomial distribution with parameters n and p converges to the Poisson distribution with parameter μ. (In Theorem 2.15 on page 59 this result was proved in another way.)

HINT:
If (z_n) is a sequence of (real or complex) numbers converging to the finite number z, then

$$\lim_{n\to\infty}\left(1+\frac{z_n}{n}\right)^n = \exp(z).$$

Exercise 4.3
Use Theorem 4.2 to show that when $k\to\infty$ and $p\to 1$ in such a way that $k(1-p)/p\to\mu$, then the negative binomial distribution with parameters k and p converges to the Poisson distribution with parameter μ. (See also Exercise 2.8 on page 60.)

Exercise 4.4
In Theorem 4.5 it is assumed that the process starts with one single individual ($Y(0) = 1$). Reformulate the theorem to cover the case with a general initial value ($Y(0) = y_0$).

Exercise 4.5
A branching process has an offspring distribution that assigns probabilities p_0, p_1 and p_2 to the integers 0, 1 and 2 ($p_0 + p_1 + p_2 = 1$), that is, after each time step each individual becomes 0, 1 or 2 individuals with the probabilities mentioned. How does the probability of ultimate extinction depend on p_0, p_1 and p_2?

5 General Theory

This chapter gives a tiny glimpse of the theory of probability measures on general sample spaces.

Andrei Kolmogorov
Russian mathematician (1903-87).
In 1933 he published the book *Grundbegriffe der Wahrscheinlichkeitsrechnung*, giving for the first time a fully satisfactory axiomatic presentation of probability theory.

Definition 5.1: σ-algebra
Let Ω be a non-empty set. A set $\mathscr{F}$ of subsets of Ω is called a σ-algebra on Ω if

1. $\Omega \in \mathscr{F}$,
2. *if $A \in \mathscr{F}$ then $A^c \in \mathscr{F}$,*
3. *$\mathscr{F}$ is closed under countable unions, i.e. if $A_1, A_2, \ldots$ is a sequence in $\mathscr{F}$ then the union $\bigcup_{n=1}^{\infty} A_n$ is in $\mathscr{F}$.*

Remarks: Let $\mathscr{F}$ be a σ-algebra. Since $\varnothing = \Omega^c$, it follows from 1 and 2 that $\varnothing \in \mathscr{F}$. Since $\bigcap_{n=1}^{\infty} A_n = \left(\bigcup_{n=1}^{\infty} A_n^c\right)^c$, it follows from 2 and 3 that $\mathscr{F}$ is closed under countable intersections as well.

Émile Borel
French mathematician (1871-1956).

On $\mathbb{R}$, the set of reals, there is a special σ-algebra $\mathscr{B}$, the *Borel σ-algebra*, which is defined as the smallest σ-algebra containing all open subsets of $\mathbb{R}$; $\mathscr{B}$ is also the smallest σ-algebra on $\mathbb{R}$ containing all intervals.

Definition 5.2: Probability space
A probability space is a tripel $(\Omega, \mathscr{F}, \mathrm{P})$ consisting of

1. *a sample space Ω, which is a non-empty set,*
2. *a σ-algebra $\mathscr{F}$ of subsets of Ω,*
3. *a probability measure on $(\Omega, \mathscr{F})$, i.e. a map $\mathrm{P} : \mathscr{F} \to \mathbb{R}$ which is*
   - *positive: $\mathrm{P}(A) \geq 0$ for all $A \in \mathscr{F}$,*
   - *normed: $\mathrm{P}(\Omega) = 1$, and*
   - *σ-additive: for any sequence $A_1, A_2, \ldots$ of disjoint events from $\mathscr{F}$, $\mathrm{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathrm{P}(A_i)$.*

Proposition 5.1
Let $(\Omega, \mathscr{F}, \mathrm{P})$ be a probability space. Then P is continuous from below in the sense that if $A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$ is an increasing sequence in $\mathscr{F}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathscr{F}$ and

$\lim_{n\to\infty} \mathrm{P}(A_n) = \mathrm{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big)$ *; furthermore,* P *is continuous from above in the sense that if* $B_1 \supseteq B_2 \supseteq B_3 \supseteq \ldots$ *is a decreasing sequence in* $\mathscr{F}$*, then* $\bigcap_{n=1}^{\infty} B_n \in \mathscr{F}$ *and* $\lim_{n\to\infty} \mathrm{P}(B_n) = \mathrm{P}\Big(\bigcap_{n=1}^{\infty} B_n\Big)$.

Proof

Due to the finite additivty,

$$\begin{aligned}\mathrm{P}(A_n) &= \mathrm{P}\Big((A_n \setminus A_{n-1}) \cup (A_{n-1} \setminus A_{n-2}) \cup \ldots \cup (A_2 \setminus A_1) \cup (A_1 \setminus \varnothing)\Big) \\ &= \mathrm{P}(A_n \setminus A_{n-1}) + \mathrm{P}(A_{n-1} \setminus A_{n-2}) + \ldots + \mathrm{P}(A_2 \setminus A_1) + \mathrm{P}(A_1 \setminus \varnothing),\end{aligned}$$

and so (with $A_0 = \varnothing$)

$$\lim_{n\to\infty} \mathrm{P}(A_n) = \sum_{n=1}^{\infty} \mathrm{P}(A_n \setminus A_{n-1}) = \mathrm{P}\Big(\bigcup_{n=1}^{\infty} (A_n \setminus A_{n-1})\Big) = \mathrm{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big),$$

where we have used that $\mathscr{F}$ is a σ-algebra and P is σ-additive. This proves the first part (continuity from below).

To prove the second part, apply "continuity from below" to the increasing sequence $A_n = B_n^{\mathrm{c}}$. □

Random variables are defined in this way (cf. Definition 1.6 page 21):

Definition 5.3: Random variable

A random variable on $(\Omega, \mathscr{F}, \mathrm{P})$ *is a function* $X : \Omega \to \mathbb{R}$ *with the property that* $\{X \in B\} \in \mathscr{F}$ *for each* $B \in \mathscr{B}$.

Remarks:

1. The condition $\{X \in B\} \in \mathscr{F}$ ensures that $\mathrm{P}(X \in B)$ is meaningful.
2. Sometimes we write $X^{-1}(B)$ instead of $\{X \in B\}$ (which is short for $\{\omega \in \Omega : X(\omega) \in B\}$), and we may think of X^{-1} as a map that takes subsets of $\mathbb{R}$ into subsets of Ω.
   Hence a random variable is a function $X : \Omega \to \mathbb{R}$ such that $X^{-1} : \mathscr{B} \to \mathscr{F}$.
   We say that $X : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$ is a *measurable map*.

The definition of distribution function is that same as Definition 1.7 (page 23):

Definition 5.4: Distribution function

The distribution function of a random variable X *is the function*

$$\begin{aligned} F : \mathbb{R} &\longrightarrow [0\,;1] \\ x &\longmapsto \mathrm{P}(X \le x)\end{aligned}$$

We obtain the same results as earlier:

Lemma 5.2
If the random variable X has distribution function F, then

$$\begin{aligned} \mathrm{P}(X \le x) &= F(x), \\ \mathrm{P}(X > x) &= 1 - F(x), \\ \mathrm{P}(a < X \le b) &= F(b) - F(a), \end{aligned}$$

for all real numbers x and $a < b$.

Proposition 5.3
The distribution function F of a random variable X has the following properties:

1. *It is non-decreasing: if $x \le y$ then $F(x) \le F(y)$.*
2. $\lim_{x\to-\infty} F(x) = 0$ *and* $\lim_{x\to+\infty} F(x) = 1$.
3. *It is right-continuous, i.e. $F(x+) = F(x)$ for all x.*
4. *At each point x, $\mathrm{P}(X = x) = F(x) - F(x-)$.*
5. *A point x is a point of discontinuity of F, if and only if $\mathrm{P}(X = x) > 0$.*

Proof
These results are proved as follows:

1. As in the proof of Proposition 1.6.
2. The sets $A_n = \{X \in]-n\,;n]\}$ increase to Ω, so using Proposition 5.1 we get $\lim_{n\to\infty}\big(F(n) - F(-n)\big) = \lim_{n\to\infty} \mathrm{P}(A_n) = \mathrm{P}(X \in \Omega) = 1$. Since F is non-decreasing and takes values between 0 and 1, this proves item 2.
3. The sets $\{X \le x + \frac{1}{n}\}$ decrease to $\{X \le x\}$, so using Proposition 5.1 we get $\lim_{n\to\infty} F(x + \frac{1}{n}) = \lim_{n\to\infty} \mathrm{P}(X \le x + \frac{1}{n}) = \mathrm{P}\Big(\bigcap_{n=1}^{\infty}\{X \le x + \frac{1}{n}\}\Big) = \mathrm{P}(X \le x) = F(x)$.
4. Again we use Proposition 5.1:

$$\begin{aligned} \mathrm{P}(X = x) &= \mathrm{P}(X \le x) - \mathrm{P}(X < x) \\ &= F(x) - \mathrm{P}\Big(\bigcup_{n=1}^{\infty}\{X \le x - \tfrac{1}{n}\}\Big) \\ &= F(x) - \lim_{n\to\infty} \mathrm{P}(X \le x - \tfrac{1}{n}) \\ &= F(x) - \lim_{n\to\infty} F(x - \tfrac{1}{n}) \\ &= F(x) - F(x-). \end{aligned}$$

5. Follows from the preceeding.

□

Proposition 5.3 has a "converse", which we state without proof:

Proposition 5.4
If the function $F : \mathbb{R} \to [0\,;1]$ is non-decreasing and right-continuous, and if $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to+\infty} F(x) = 1$, then F is the distribution function of a random variable.

5.1 Why a general axiomatic approach?

1. One reason to try to develop a unified axiomatic theory of probability is that from a mathematical-aesthetic point of view it is quite unsatisfactory to have to treat discrete and continuous probability distributions separately (as is done in these notes) — and what to do with distributions that are "mixtures" of discrete and continuous distributions?

Example 5.1
Suppose we want to model the lifetime T of some kind of device that either breaks down immediately, i.e. $T = 0$, or after a waiting time that is assumed to follow an exponential distribution. Hence the distribution of T can be specified by saying that $\mathrm{P}(T = 0) = p$ and $\mathrm{P}(T > t \mid T > 0) = \exp(-t/\beta)$, $t > 0$; here the parameter p is the probability of an immediate breakdown, and the parameter β is the scale parameter of the exponential distribution.

The distribution of T has a discrete component (the probability mass p at 0) and a continuous component (the probability mass $1 - p$ smeared out on the positive reals according to an exponential distribution).

2. It is desirable to have a mathematical setting that allows a proper definition of *convergence of probability distributions*. It is, for example, "obvious" that the discrete uniform distribution on the set $\{0, \frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \ldots, \frac{n-1}{n}, 1\}$ converges to the continuous uniform distribution on $[0\,;1]$ as $n \to \infty$, and also that the normal distribution with parameters μ and $\sigma^2 > 0$ converges to the one-point distribution at μ as $\sigma^2 \to 0$. Likewise, the Central Limit Theorem (page 74) tells that the distribution of sums of properly scaled independent random variables converges to a normal distribution. The theory should be able to place these convergences within a wider mathematical context.

3. We have seen simple examples of how to model compound experiments with independent components (see e.g. page 19f and page 26). How can this be generalised, and how to handle infinitely many components?

Example 5.2
The Strong Law of Large Numbers (cf. page 41) states that if (X_n) is a sequence of independent identically distributed random variables with expected value μ, then

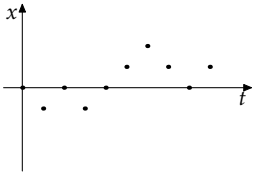
$\mathrm{P}\big(\lim_{n\to\infty} \overline{X}_n = \mu\big) = 1$ where $\overline{X}_n$ denotes the average of $X_1, X_2, \ldots, X_n$. The event in question here, $\{\lim_{n\to\infty} \overline{X}_n = \mu\}$, involves infinitely many Xs, and one might wonder whether it is at all possible to have a probability space that allows infinitely many random variables.

Example 5.3
In the presentations of the geometric and negative binomial distributions (pages 54 and 56), we studied how many times to repeat a 01 experiment before the first (or kth) 1 occurs, and a random variable T_k was introduced as the number of repetitions needed in order that the outcome 1 occurs for the kth time.

It can be of some interest to know the probability that sooner or later a total of k 1s have appeared, that is, $\mathrm{P}(T_k < \infty)$. If $X_1, X_2, X_3, \ldots$ are random variables representing outcome no. 1, 2, 3, ... of the 01-experiment, then T_k is a straightforward function of the Xs: $T_k = \inf\{t \in \mathbb{N} : X_t = k\}$. — How can we build a probability space that allows us to discuss events relating to infinite sequences of random variables?

Example 5.4
Let $(U(t) : t = 1, 2, 3\ldots)$ be a sequence of independent identically distributed uniform random variables on the two-element set $\{-1, 1\}$, and define a new sequence of random variables $(X(t) : t = 0, 1, 2, \ldots)$ as

$$\begin{aligned} X(0) &= 0 \\ X(t) &= X(t-1) + U(t), \quad t = 1, 2, 3, \ldots \end{aligned}$$

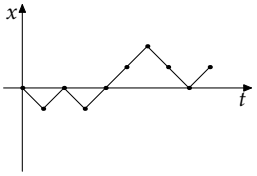


that is, $X(t) = U(1) + U(2) + \cdots + U(t)$. This new stochastic process $(X(t)$ is known as a *simple random walk* in one dimension. Stochastic processes are often graphed in a coordinate system with the abscissa axis as "time" (t) and the ordinate axis as "state" (x). Accordingly, we say that the simple random walk $(X(t) : t \in \mathbb{N}_0)$ is in the state $x = 0$ at time $t = 0$, and for each time step thereafter, it moves one unit up or down in the state space.

The steps need not be of length 1. We can easily define a random walk that moves in steps of length Δx, and moves at times $\Delta t, 2\Delta t, 3\Delta t, \ldots$ (here, $\Delta x \neq 0$ and $\Delta t > 0$). To do this, begin with an i.i.d. sequence $(U(t) : t = \Delta t, 2\Delta t, 3\Delta t, \ldots)$ (where the Us have the same distribution as the previous Us), and define

$$\begin{aligned} X(0) &= 0 \\ X(t) &= X(t-\Delta t) + U(t)\Delta x, \quad t = \Delta t, 2\Delta t, 3\Delta t, \ldots \end{aligned}$$

that is, $X(n\Delta t) = U(\Delta t)\Delta x + U(2\Delta t)\Delta x + \cdots + U(n\Delta t)\Delta x$. This defines $X(t)$ for non-negative multipla af Δt, i.e. for $t \in \mathbb{N}_0\Delta t$; then use linear interpolation in each subinterval $]k\Delta t\,;(k+1)\Delta t[$ to define $X(t)$ for all $t \geq 0$. The resulting function $t \mapsto X(t)$, $t \geq 0$, is continuous and piecewise linear.



It is easily seen that $\mathrm{E}\big(X(n\Delta t)\big) = 0$ and $\mathrm{Var}\big(X(n\Delta t)\big) = n(\Delta x)^2$, so if $t = n\Delta t$ (and hence $n = t/\Delta t$), then $\mathrm{Var}\, X(t) = \big((\Delta x)^2/\Delta t\big)t$. This suggests that if we are going to let Δt and Δx tend to 0, it might be advisable to do it in such a way that $(\Delta x)^2/\Delta t$ has a limiting value $\sigma^2 > 0$, so that $\mathrm{Var}\, X(t) \to \sigma^2 t$.

The problem is to develop a mathematical setting in which the "limiting process" just outlined is meaningful; this includes an elucidation of the mathematical objects involved, and of the way they converge.

The solution is to work with probability distributions on spaces of continuous functions (the interpolated random walks are continuous functions of t, and the limiting process should also be continuous). For convenience, assume that $\sigma^2 = 1$; then the limiting process is the so-called standard Wiener process $(W(t) : t \geq 0)$. The sample paths $t \mapsto W(t)$ are continuous and have a number of remarkable properties:

- If $0 \leq s_1 < t_1 \leq s_2 < t_2$, the increments $W(t_1) - W(s_1)$ and $W(t_2) - W(s_2)$ are independent normal random variables with expected value 0 and variance $(t_1 - s_1)$ and $(t_2 - s_2)$, respectively.
- With probability 1 the function $t \mapsto W(t)$ is nowhere differentiable.
- With probability 1 the process will visit all states, i.e. with probability 1 the range of $t \mapsto W(t)$ is all of $\mathbb{R}$.
- For each $x \in \mathbb{R}$ the process visits x infinitely often with probability 1, i.e. for each $x \in \mathbb{R}$ the set $\{t > 0 : W(t) = x\}$ is infinite with probability 1.

Pioneering works in this field are Bachelier (1900) and the papers by Norbert Wiener from the early 1920s, see Wiener (1976, side 435-519).

Part II

Statistics

Introduction

While probability theory deals with the formulation and analysis of probability models for random phenomena (and with the creation of an adequate conceptual framework), *mathematical statistics* is a discipline that basically is about developing and investigating methods to extract information from data sets burdened with such kinds of uncertainty that we are prepared to describe with suitable probability distributions, and *applied statistics* is a discipline that deals with these methods in specific types of modelling situations.

Here is a brief outline of the kind of issues that we are going to discuss: We are given a data set, a set of numbers $x_1, x_2, \ldots, x_n$, referred to as the *observations* (a simple example: the values of 115 measuments of the concentration of mercury i swordfish), and we are going to set up a *statistical model* stating that the observations are observed values of random variables $X_1, X_2, \ldots, X_n$ whose joint distribution is known except for a few unknown *parameters*. (In the swordfish example, the random variables could be indepedent and identically distributed with the two unknown parameters μ and σ^2). Then we are left with three main problems:

1. *Estimation*. Given a data set and a statistical model relating to a given subject matter, then how do we calculate an estimate of the values of the unknown parameters? The estimate has to be optimal, of course, in a sense to be made precise.
2. *Testing hypotheses*. Usually, the subject matter provides a number of interesting hypotheses, which the statistician then translates into *statistical hypotheses* that can be *tested*. A statistical hypothesis is a statement that the parameter structure can be simplified relative to the original model (for example that some of the parameters are equal, or have a known value).
3. *Validation of the model*. Ususally, statistical models are certainly not realistic in the sense that they claim to imitate the "real" mechanisms that generated the observations. Therefore, the job of adjusting and checking the model and assessing its usefulness gets a different content and a different importance from what is the case with many other types of mathematical models.

Ronald Aylmer Fisher
English statistician and geneticist (1890-1962). The founder of mathematical statistics (or at least, founder of mathematical statistics as it is presented in these notes).

This is how Fisher in 1922 explained the object of statistics (Fisher (1922), here quoted from Kotz and Johnson (1992)):

> *In order to arrive at a distinct formulation of statistical problems, it is necessary to define the task which the statistician sets himself: briefly, and in its most concrete form, the object of statistical methods is the reduction of data. A quantity of data, which usually by its mere bulk is incapabel of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data.*

6 The Statistical Model

We begin with a fairly abstract presentation of the notion of a statistical model and related concepts, afterwards there will be a number of illustrative examples.

There is an *observation* $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ which is assumed to be an observed value of a random variable $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ taking values in the *observation space* $\mathfrak{X}$; in many cases, the set $\mathfrak{X}$ is $\mathbb{R}^n$ or $\mathbb{N}_0^n$.

For each value of a *parameter* $\boldsymbol{\theta}$ from the *parameter space* Θ there is a probability measure $\mathrm{P}_{\boldsymbol{\theta}}$ on $\mathfrak{X}$. All of these $\mathrm{P}_{\boldsymbol{\theta}}$s are candidates for being the distribution of $\boldsymbol{X}$, and for (at least) one value $\boldsymbol{\theta}$ is it true that the distribution of $\boldsymbol{X}$ is $\mathrm{P}_{\boldsymbol{\theta}}$.

The statement "$\boldsymbol{x}$ is an observered value of the random variable $\boldsymbol{X}$, and for at least one value of $\boldsymbol{\theta} \in \Theta$ is it true that the distribution of $\boldsymbol{X}$ is $\mathrm{P}_{\boldsymbol{\theta}}$" is one way of expressing the *statistical model*.

The *model function* is a function $f : \mathfrak{X} \times \Theta \to [0\,; +\infty[$ such that for each fixed $\boldsymbol{\theta} \in \Theta$ the function $\boldsymbol{x} \mapsto f(\boldsymbol{x}, \boldsymbol{\theta})$ is the probability (density) function pertaining to $\boldsymbol{X}$ when $\boldsymbol{\theta}$ is the true parameter value.

The *likelihood function* corresponding to the observation $\boldsymbol{x}$ is the function $L : \Theta \to [0\,; +\infty[$ given as $L(\boldsymbol{\theta}) = f(\boldsymbol{x}, \boldsymbol{\theta})$.—The likelihood function plays an essential role in the theory of estimation of parameters and test of hypotheses.

If the likelihood function can be written as $L(\boldsymbol{\theta}) = g(\boldsymbol{x})\, h(t(\boldsymbol{x}), \boldsymbol{\theta})$ for suitable functions g, h and t, then t is said to be a *sufficient statistic* (or to give a *sufficient reduction of data*).—Sufficient reductions are of interest mainly in stiuations where the reduction is to a very small dimension (such as one or two).

Remarks:

1. Parameters are ususally denoted by Greek letters.
2. The dimension of the parameter space Θ is ususally much smaller than that of the observation space $\mathfrak{X}$.
3. In most cases, an *injective parametrisation* is wanted, that is, the map $\boldsymbol{\theta} \mapsto \mathrm{P}_{\boldsymbol{\theta}}$ should be injective.
4. The likelihood function does *not* have to sum or integrate to 1.
5. Often, the *log-likelihood function*, i.e. the logarithm of likelihood function, is easier to work with than the likelihood function itself; note that "logarithm" always means "natural logarithm".

Some more notation:

1. When we need to emphasise that L is the likelihood function corresponding to a certain observation $\boldsymbol{x}$, we will write $L(\boldsymbol{\theta};\boldsymbol{x})$ instead of $L(\boldsymbol{\theta})$.
2. The symbols $\mathrm{E}_{\boldsymbol{\theta}}$ and $\mathrm{Var}_{\boldsymbol{\theta}}$ are sometimes used when we wish to emphasise that the expectation and the variance are evaluated with respect to the probability distribution corresponding to the parameter value $\boldsymbol{\theta}$, i.e. with respect to $\mathrm{P}_{\boldsymbol{\theta}}$.

6.1 Examples

The one-sample problem with 01-variables

THE DOT NOTATION
Given a set of indexed quantities, for example $a_1, a_2, \ldots, a_n$, then a shorthand notation for the sum of these quantities is same symbol, now with a dot as index:

$$a_\cdot = \sum_{i=1}^{n} a_i$$

Similarly with two or more indices, e.g.:

$$b_{i\cdot} = \sum_j b_{ij}$$

and

$$b_{\cdot j} = \sum_i b_{ij}$$

In the general formulation of the one-sample problem with 01-variables, $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is assumed to be an observation of an n-dimensional random variable $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ taking values in $\mathfrak{X} = \{0,1\}^n$. The X_is are assumed to be independent indentically distributed 01-variables, and $\mathrm{P}(X_i = 1) = \theta$ where θ is the unknown parameter. The parameter space is $\Theta = [0\,;1]$. The model function is (cf. (1.5) on page 29)

$$f(\boldsymbol{x},\theta) = \theta^{x_\cdot}(1-\theta)^{n-x_\cdot}, \quad (\boldsymbol{x},\theta) \in \mathfrak{X} \times \Theta,$$

and the likelihood function is

$$L(\theta) = \theta^{x_\cdot}(1-\theta)^{n-x_\cdot}, \quad \theta \in \Theta.$$

We see that the likelihood function depends on $\boldsymbol{x}$ only through $x_\cdot$, that is, $x_\cdot$ is sufficient, or rather: the function that maps $\boldsymbol{x}$ into $x_\cdot$ is sufficient.

▷ [To be continued on page 115.]

Example 6.1
Suppose we have made seven replicates of an experiment with two possible outcomes 0 and 1 and obtained the results $1,1,0,1,1,0,0$. We are going to write down a statistical model for this situation.

The observation $\boldsymbol{x} = (1,1,0,1,1,0,0)$ is assumed to be an observation of the 7-dimensional random variable $\boldsymbol{X} = (X_1, X_2, \ldots, X_7)$ taking values in $\mathfrak{X} = \{0,1\}^7$. The X_is are assumed to be independent identically distributed 01-variables, and $\mathrm{P}(X_i = 1) = \theta$ where θ is the unknown parameter. The parameter space is $\Theta = [0\,;1]$. The model function is

$$f(\boldsymbol{x},\theta) = \prod_{i=1}^{7} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{x_\cdot}(1-\theta)^{7-x_\cdot}.$$

The likelihood function corresponding to the observation $\boldsymbol{x} = (1,1,0,1,1,0,0)$ is

$$L(\theta) = \theta^4(1-\theta)^3, \quad \theta \in [0\,;1].$$

The simple binomial model

The binomial distribution is the distribution of a sum of independent 01-variables, so it is hardly surprising that statistical analysis of binomial variables resembles that of 01-variables very much.

If Y is binomial with parameters n (known) and θ (unknown, $\theta \in [0\,;1]$), then the model function is

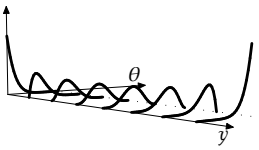
$$f(y,\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

where $(y,\theta) \in \mathfrak{X} \times \Theta = \{0,1,2,\ldots,n\} \times [0\,;1]$, and the likelihood function is

$$L(\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, \quad \theta \in [0\,;1].$$

▷ [To be continued on page 116.]

Example 6.2
If we in Example 6.1 consider the total number of 1s only, not the outcomes of the individual experiments, then we have an observation $y = 4$ of a binomial random variable Y with parameters $n = 7$ and unknown θ. The observation space is $\mathfrak{X} = \{0,1,2,3,4,5,6,7\}$ and the parameter space is $\Theta = [0\,;1]$. The model function is er



$f(y,\theta) = \binom{7}{y}\theta^y(1-\theta)^{7-y}$

$$f(y,\theta) = \binom{7}{y}\theta^y\,(1-\theta)^{7-y}, \quad (y,\theta) \in \{0,1,2,3,4,5,6,7\} \times [0\,;1].$$

The likelihood function corresponding to the observation $y = 4$ is

$$L(\theta) = \binom{7}{4}\theta^4\,(1-\theta)^3, \quad \theta \in [0\,;1].$$

Example 6.3: Flour beetles I
As part of an example to be discussed in detail in Section 9.1, 144 flour beetles (*Tribolium castaneum*) are exposed to the insecticide Pyrethrum in a certain concentration, and during the period of observation 43 of the beetles die. If we assume that all beetles are "identical", and that they die independently of each other, then we may assume that the value $y = 43$ is an observation of a binomial random variable Y with parameters $n = 144$ and θ (unknown). The statistical model is then given by means of the model function

$$f(y,\theta) = \binom{144}{y}\theta^y(1-\theta)^{144-y}, \quad (y,\theta) \in \{0,1,2,\ldots,144\} \times [0\,;1].$$

The likelihood function corresponding to the observation $y = 43$ is

$$L(\theta) = \binom{144}{43}\theta^{43}(1-\theta)^{144-43}, \quad \theta \in [0\,;1].$$

▷ [To be continued in Example 7.1 page 116.]

Table 6.1
Comparison of binomial distributions

	group no.				
	1	2	3	...	s
number of successes	y_1	y_2	y_3	...	y_s
number of failures	$n_1 - y_1$	$n_2 - y_2$	$n_3 - y_3$	...	$n_s - y_s$
total	n_1	n_2	n_3	...	n_s

Comparison of binomial distributions

We have observations $y_1, y_2, \ldots, y_s$ of independent binomial random variables $Y_1, Y_2, \ldots, Y_s$, where Y_j has parameters n_j (known) and $\theta_j \in [0\,;1]$. You might think of the observations as being arranged in a table like the one in Table 6.1. The model function is

$$f(\boldsymbol{y}, \boldsymbol{\theta}) = \prod_{j=1}^{s} \binom{n_j}{y_j} \theta_j^{y_j} (1-\theta_j)^{n_j - y_j}$$

where the parameter variable $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_s)$ varies in $\Theta = [0\,;1]^s$, and the observation variable $\boldsymbol{y} = (y_1, y_2, \ldots, y_s)$ varies in $\mathfrak{X} = \prod_{j=1}^{s} \{0, 1, \ldots, n_j\}$. The likelihood function and the log-likelihood function corresponding to $\boldsymbol{y}$ are

$$L(\boldsymbol{\theta}) = \text{const}_1 \cdot \prod_{j=1}^{s} \theta_j^{y_j} (1-\theta_j)^{n_j - y_j},$$

$$\ln L(\boldsymbol{\theta}) = \text{const}_2 + \sum_{j=1}^{s} \Big(y_j \ln \theta_j + (n_j - y_j) \ln(1-\theta_j) \Big);$$

here const_1 is the product of the s binomial coefficients, and $\text{const}_2 = \ln(\text{const}_1)$.

▷ [To be continued on page 116.]

Example 6.4: Flour beetles II
Flour beetles are exposed to an insecticide which is administered in four different concentrations, 0.20, 0.32, 0.50 and 0.80 mg/cm^2, and after 13 days the numbers of dead beetles are registered, see Table 6.2. (The poison is sprinkled on the floor, so the concentration is given as weight per area.) One wants to examine whether there is a difference between the effects of the various concentrations, so we have to set up a statistical model that will make that possible.

Just like in Example 6.3, we will assume that for each concentration the number of dead beetles is an observation of a binomial random variable. For concentration no. j, $j = 1, 2, 3, 4$, we let y_j denote the number of dead beetles and n_j the total number of beetles. Then the statistical model is that $\boldsymbol{y} = (y_1, y_2, y_3, y_4) = (43, 50, 47, 48)$ is an observation of four-dimensional random variable $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4)$ where Y_1, Y_2, Y_3

	concentration			
	0.20	0.32	0.50	0.80
dead	43	50	47	48
alive	101	19	7	2
total	144	69	54	50

Table 6.2
Survival of flour beetles at four different doses of poison.

and Y_4 are independent binomial random with parameters $n_1 = 144$, $n_2 = 69$, $n_3 = 54$, $n_4 = 50$ and $\theta_1, \theta_2, \theta_3, \theta_4$. The model function is

$$f(y_1,y_2,y_3,y_4;\theta_1,\theta_2,\theta_3,\theta_4) = \binom{144}{y_1}\theta_1^{y_1}(1-\theta_1)^{144-y_1}\binom{69}{y_2}\theta_2^{y_2}(1-\theta_2)^{69-y_2}\times \binom{54}{y_3}\theta_3^{y_3}(1-\theta_3)^{54-y_3}\binom{50}{y_4}\theta_4^{y_4}(1-\theta_4)^{50-y_4}\,.$$

The log-likelihood function corresponding to the observation $\boldsymbol{y}$ is

$$\begin{aligned}\ln L(\theta_1,\theta_2,\theta_3,\theta_4) = \text{const}\\ &+ 43\ln\theta_1 + 101\ln(1-\theta_1) + 50\ln\theta_2 + 19\ln(1-\theta_2)\\ &+ 47\ln\theta_3 + \;7\ln(1-\theta_3) + 48\ln\theta_4 + \;2\ln(1-\theta_4).\end{aligned}$$

▷ [To be continued in Example 7.2 on page 117.]

The multinomial distribution

The multinomial distribution is a generalisation of the binomial distribution: In a situation with n independent replicates of a trial with two possible outcomes, the number of "successes" follows a binomial distribution, cf. Definition 1.10 on page 30. In a situation with n independent replicates of a trial with r possible outcomes $\omega_1, \omega_2, \ldots, \omega_r$, let the random variable Y_i denote the number of times outcome ω_i is observed, $i = 1, 2, \ldots, r$; then the r-dimensional random variable $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_r)$ follows a multinomial distribution.

In the circumstances just described, the distribution of $\boldsymbol{Y}$ has the form

$$\mathrm{P}(\boldsymbol{Y} = \boldsymbol{y}) = \binom{n}{y_1\ y_2\ \ldots\ y_r}\prod_{i=1}^{r}\theta_i^{y_i} \tag{6.1}$$

if $\boldsymbol{y} = (y_1, y_2, \ldots, y_r)$ is a set of non-negative integers with sum n, and $\mathrm{P}(\boldsymbol{Y} = \boldsymbol{y}) = 0$ otherwise. The parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_r)$ is a set of r non-negative real numbers adding up to 1; here, θ_i is the probability of the outcome ω_i in a single trial. The quantity

$$\binom{n}{y_1\ y_2\ \ldots\ y_r} = \frac{n!}{\prod_{i=1}^{r} y_i!}$$

Table 6.3
Distribution of haemoglobin genotype in Baltic cod.

	Lolland	Bornholm	Åland
AA	27	14	0
Aa	30	20	5
aa	12	52	75
total	69	86	80

is a so-called *multinomial coefficient*, and it is by definition the number of ways that a set with n elements can be partitioned into r subsets such that subset no. i contains exactly y_i elements, $i = 1, 2, \ldots, r$.

The probability distribution with probability function (6.1) is known as the multinomial distribution with r classes (or categories) and with parameters n (known) and $\boldsymbol{\theta}$.

▷ [To be continued on page 117.]

Example 6.5: Baltic cod
On 6 March 1961 a group of marine biologists caught 69 cod in the waters at Lolland, and for each fish the haemoglobin in the blood was examined. Later that year, a similar experiment was carried out at Bornholm and at the Åland Islands (Sick, 1965).

It is believed that the type of haemoglobin is determined by a single gene, and in fact the biologists studied the genotype as to this gene. The gene is found in two versions A and a, and the possible genotypes are then AA, Aa and aa. The empirical distribution of genotypes for each location is displayed in Table 6.3.

At each of the three places a number of fish are classified into three classes, so we have three multinomial situations. (With three classes, it would often be named a *tri*nomial situation.) Our basic model will therefore be the model stating that each of the three observed triples

$$\boldsymbol{y}_{\mathrm{L}} = \begin{pmatrix} y_{1\mathrm{L}} \\ y_{2\mathrm{L}} \\ y_{3\mathrm{L}} \end{pmatrix} = \begin{pmatrix} 27 \\ 30 \\ 12 \end{pmatrix}, \quad \boldsymbol{y}_{\mathrm{B}} = \begin{pmatrix} y_{1\mathrm{B}} \\ y_{2\mathrm{B}} \\ y_{3\mathrm{B}} \end{pmatrix} = \begin{pmatrix} 14 \\ 20 \\ 52 \end{pmatrix}, \quad \boldsymbol{y}_{\text{Å}} = \begin{pmatrix} y_{1\text{Å}} \\ y_{2\text{Å}} \\ y_{3\text{Å}} \end{pmatrix} = \begin{pmatrix} 0 \\ 5 \\ 75 \end{pmatrix}$$

originates from its own multinomial distribution; these distributions have parameters $n_{\mathrm{L}} = 69$, $n_{\mathrm{B}} = 86$, $n = 80$, and

$$\boldsymbol{\theta}_{\mathrm{L}} = \begin{pmatrix} \theta_{1\mathrm{L}} \\ \theta_{2\mathrm{L}} \\ \theta_{3\mathrm{L}} \end{pmatrix}, \quad \boldsymbol{\theta}_{\mathrm{B}} = \begin{pmatrix} \theta_{1\mathrm{B}} \\ \theta_{2\mathrm{B}} \\ \theta_{3\mathrm{B}} \end{pmatrix}, \quad \boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix}.$$

▷ [To be continued in Example 7.3 on page 118.]

The one-sample problem with Poisson variables

The simplest situation is as follows: We have observations $y_1, y_2, \ldots, y_n$ of independent and identically distributed Poisson random variables $Y_1, Y_2, \ldots, Y_n$ with

no. of deaths y	no. of regiment-years with y deaths
0	109
1	65
2	22
3	3
4	1
	200

Table 6.4
Number of deaths by horse-kicks in the Prussian army.

parameter μ. The model function is

$$f(\boldsymbol{y},\mu) = \prod_{j=1}^{n} \frac{\mu^{y_j}}{y_j!} \exp(-\mu) = \frac{\mu^{y_\cdot}}{\prod_{j=1}^{n} y_j!} \exp(-n\mu),$$

where $\mu \geq 0$ and $\boldsymbol{y} \in \mathbb{N}_0^n$.

The likelihood function is $L(\mu) = \text{const}\, \mu^{y_\cdot} \exp(-n\mu)$, and the log-likelihood function is $\ln L(\mu) = \text{const} + y_\cdot \ln \mu - n\mu$.

▷ [To be continued on page 118.]

Example 6.6: Horse-kicks
For each of the 20 years from 1875 to 1894 and each of 10 cavalry regiments of the Prussian army, the number of soldiers kicked to death by a horse is recorded (Bortkiewicz, 1898). This means that for each of 200 "regiment-years" the number of deaths by horse-kick is known.

We can summarise the data in a table showing the number of regiment-years with 0, 1, 2, ... deaths, i.e. we classify the regiment-years by number of deaths; it turns out that the maximum number of deaths per year in a regiment-year was 4, so there will be five classes, corresponding to 0, 1, 2, 3 and 4 deaths per regiment-year, see Table 6.4.

It seems reasonable to believe that to a great extent these death accidents happened truely at random. Therefore there will be a random number of deaths in a given regiment in a given year. If we assume that the individual deaths are independent of one another and occur with constant intensity throughout the "regiment-years", then the conditions for a Poisson model are met. We will therefore try a statistical model stating that the 200 observations $y_1, y_2, \ldots, y_{200}$ are observed values of independent and identically distributed Poisson variables $Y_1, Y_2, \ldots, Y_{200}$ with parameter μ.

▷ [To be continued in Example 7.4 on page 118.]

Uniform distribution on an interval

This example is not, as far as is known, of any great importance, but it can be useful to try out the theory.

Consider observations $x_1, x_2, \ldots, x_n$ of independent and identically distributed random variables $X_1, X_2, \ldots, X_n$ with a uniform distribution on the interval $]0\,;\theta[$, where $\theta > 0$ is the unknown parameter. The density function of X_i is

$$f(x,\theta) = \begin{cases} 1/\theta & \text{when } 0 < x < \theta \\ 0 & \text{otherwise,} \end{cases}$$

so the model function is

$$f(x_1, x_2, \ldots, x_n, \theta) = \begin{cases} 1/\theta^n & \text{when } 0 < x_{\min} \text{ and } x_{\max} < \theta \\ 0 & \text{otherwise.} \end{cases}$$

Here $x_{\min} = \min\{x_1, x_2, \ldots, x_n\}$ and $x_{\max} = \max\{x_1, x_2, \ldots, x_n\}$, i.e., the smallest and the largest observation.

▷ [To be continued on page 119.]

The one-sample problem with normal variables

We have observations $y_1, y_2, \ldots, y_n$ of independent and identically distributed normal variables with expected value μ and variance σ^2, that is, from the distribution with density function $y \mapsto \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2}\dfrac{(y-\mu)^2}{\sigma^2}\Big)$. The model function is

$$\begin{aligned} f(\boldsymbol{y}, \mu, \sigma^2) &= \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\tfrac{1}{2}\frac{(y_j-\mu)^2}{\sigma^2}\Big) \\ &= (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\mu)^2\Big) \end{aligned}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$, $\mu \in \mathbb{R}$, and $\sigma^2 > 0$. Standard calculations lead to

$$\begin{aligned} \sum_{j=1}^{n}(y_j-\mu)^2 &= \sum_{j=1}^{n}\Big((y_j-\overline{y}) + (\overline{y}-\mu)\Big)^2 \\ &= \sum_{j=1}^{n}(y_j-\overline{y})^2 + 2(\overline{y}-\mu)\sum_{j=1}^{n}(y_j-\overline{y}) + n(\overline{y}-\mu)^2 \\ &= \sum_{j=1}^{n}(y_j-\overline{y})^2 + n(\overline{y}-\mu)^2, \end{aligned}$$

so

$$\sum_{j=1}^{n}(y_j-\mu)^2 = \sum_{j=1}^{n}(y_j-\overline{y})^2 + n(\overline{y}-\mu)^2. \tag{6.2}$$

28	26	33	24	34	−44
27	16	40	−2	29	22
24	21	25	30	23	29
31	19	24	20	36	32
36	28	25	21	28	29
37	25	28	26	30	32
36	26	30	22	36	23
27	27	28	27	31	27
26	33	26	32	32	24
39	28	24	25	32	25
29	27	28	29	16	23

Table 6.5
Newcomb's measurements of the passage time of light over a distance of 7442 m. The entries of the table multiplied by 10^{-3} + 24.8 are the passage times in 10^{-6} sec.

Using this we see that the likelihood function is

$$\begin{aligned}\ln L(\mu,\sigma^2) &= \text{const} - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\mu)^2 \\ &= \text{const} - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\overline{y})^2 - \frac{n(\overline{y}-\mu)^2}{2\sigma^2}.\end{aligned} \tag{6.3}$$

We can make use of equation (6.2) once more: for $\mu = 0$ we obtain

$$\sum_{j=1}^{n}(y_j-\overline{y})^2 = \sum_{j=1}^{n}y_j^2 - n\overline{y}^2 = \sum_{j=1}^{n}y_j^2 - \frac{1}{n}y_\bullet^2\ ,$$

showing that the sum of squared deviations of the ys from $\overline{y}$ can be calculated from the sum of the ys and the sum of squares of the ys. Thus, to find the likelihood function we need not know the individual observations, just their sum and sum of squares (in other words, $t(\boldsymbol{y}) = (\sum y, \sum y^2)$ is a *sufficient statistic*, cf. page 99).

▷ [To be continued on page 119.]

Example 6.7: The speed of light
With a series of experiments that were quite precise for their time, the American physicist A.A. Michelson and the American mathematician and astronomer S. Newcomb were attempting to determine the velocity of light in air (Newcomb, 1891). They were employing the idea of Foucault of letting a light beam go from a fast rotating mirror to a distant fixed mirror and then back to the rotating mirror where the angular displacement is measured. Knowing the speed of revolution and the distance between the mirrors, you can then easily calculate the velocity of the light.

Table 6.5 (from Stigler (1977)) shows the results of 66 measurements made by Newcomb in the period from 24th July to 5th September in Washington, D.C. In Newcomb's set-up there was a distance of 3721 m between the rotating mirror placed in Fort Myer

on the west bank of the Potomac river and the fixed mirror placed on the base of the George Washington monument. The values reported by Newcomb are the passage times of the light, that is, the time used to travel the distance in question.

Two of the 66 values in the table stand out from the rest, –44 and –2. They should probably be regarded as *outliers*, values that in some sense are "too far away" from the majority of observations. In the subsequent analysis of the data, we will ignore those two observations, which leaves us with a total of 64 observations.

▷ [To be continued in Example 7.5 on page 120.]

The two-sample problem with normal variables

In two groups of individuals the value of a certain variable Y is determined for each individual. Individuals in one group have no relation to individuals in the other group, they are "unpaired", and there need not be the same number of individuals in each group. We will depict the general situation like this:

	observations					
group 1	y_{11}	y_{12}	$\cdots$	y_{1j}	$\cdots$	y_{1n_1}
group 2	y_{21}	y_{22}	$\cdots$	y_{2j}	$\cdots$	y_{2n_2}

Here y_{ij} is observation no. j in group no. i, $i = 1, 2$, and n_1 and n_2 the number of observations in the two groups. We will assume that the variation between observations within a single group is random, whereas there is a *systematic variation* between the two groups (that is indeed why the grouping is made!). Finally, we assume that the y_{ij}s are observed values of independent normal random variables Y_{ij}, all having the same variance σ^2, and with expected values $\mathrm{E}\, Y_{ij} = \mu_i$, $j = 1, 2, \ldots, n_i$, $i = 1, 2$. In this way the two mean value parameters μ_1 and μ_2 describe the *systematic variation*, i.e. the two group levels, and the variance parameter σ^2 (and the normal distribution) describe the *random variation*, which is the same in the two groups (an assumption that may be tested, see Exercise 8.2 in page 136). The model function is

$$f(\boldsymbol{y}, \mu_1, \mu_2, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \mu_i)^2\right)$$

where $\boldsymbol{y} = (y_{11}, y_{12}, y_{13}, \ldots, y_{1n_1}, y_{21}, y_{22}, \ldots, y_{2n_2}) \in \mathbb{R}^n$, $(\mu_1, \mu_2) \in \mathbb{R}^2$ and $\sigma^2 > 0$; here we have put $n = n_1 + n_2$. The sum of squares can be partitioned in the same way as equation (6.2):

$$\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \mu_i)^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2 + \sum_{i=1}^{2} n_i(\overline{y}_i - \mu_i)^2;$$

here $\overline{y}_i = \dfrac{1}{n}\sum_{j=1}^{n_i} y_{ij}$ is the ith group average.
The log-likelihood function is

$$\begin{aligned}&\ln L(\mu_1, \mu_2, \sigma^2) \\ &= \text{const} - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\Big(\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2 + \sum_{i=1}^{2} n_i(\overline{y}_i - \mu_i)^2\Big).\end{aligned} \tag{6.4}$$

▷ [To be continued on page 120.]

Example 6.8: Vitamin C
Vitamin C (ascorbic acid) is a well-defined chemical substance, and one would think that industrially produced Vitamin C is as effective as "natural" vitamin C. To check whether this is indeed the fact, a small experiment with guinea pigs was conducted.

In this experiment 20 almost "identical" guinea pigs were divided into two groups. Over a period of six weeks the animals in group 1 got a certain amount of orange juice, and the animals in group 2 got an equivalent amount of synthetic vitamin C, and at the end of the period, the length of the odontoblasts was measured for each animal. The resulting observations (put in order in each group) were:

orange juice:	8.2	9.4	9.6	9.7	10.0	14.5	15.2	16.1	17.6	21.5
synthetic vitamin C:	4.2	5.2	5.8	6.4	7.0	7.3	10.1	11.2	11.3	11.5

This must be a kind of two-sample problem. The nature of the observations seems to make it reasonable to try a model based on the normal distribution, so why not suggest that this is an instance of the "two-sample problem with normal variables". Later on, we shall analyse the data using this model, and we shall in fact test whether the mean increase of the odontoblasts is the same in the two groups.

▷ [To be continued in Example 7.6 on page 121.]

Simple linear regression

Regression analysis, a large branch of statistics, is about modelling the mean value structure of the random variables, using a small number of quantitative covariates (explanatory variables). Here we consider the simplest case.

There is a number of pairs (x_i, y_i), $i = 1, 2, \ldots, n$, where the ys are regarded as observed values of random variables $Y_1, Y_2, \ldots, Y_n$, and the xs are non-random so-called *covariates* or *explanatory variables*.—In regression models covariates are always non-random.

In the simple linear regression model the Ys are independent normal random variables with the same variance σ^2 and with a mean value structure of the form $\mathrm{E}\, Y_i = \alpha + \beta x_i$, or more precisely: for certain constants α and β, $\mathrm{E}\, Y_i = \alpha + \beta x_i$ for

boiling point	barometric pressure	boiling point	barometric pressure
194.5	20.79	201.3	24.01
194.3	20.79	203.6	25.14
197.9	22.40	204.6	26.57
198.4	22.67	209.5	28.49
199.4	23.15	208.6	27.76
199.9	23.35	210.7	29.04
200.9	23.89	211.9	29.88
201.1	23.99	212.2	30.06
201.4	24.02		

Table 6.6
Forbes' data.—Boiling point in degrees F, barometric pressure in inches of Mercury.

all i. Hence the model has three unknown parameters, α, β and σ^2. The model function is

$$f(\boldsymbol{y}, \alpha, \beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{\big(y_i - (\alpha + \beta x_i)\big)^2}{\sigma^2}\right)$$
$$= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\big(y_i - (\alpha + \beta x_i)\big)^2\right)$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$, $\alpha, \beta \in \mathbb{R}$ and $\sigma^2 > 0$. The log-likelihood function is

$$\ln L(\alpha, \beta, \sigma^2) = \text{const} - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\big(y_i - (\alpha + \beta x_i)\big)^2. \tag{6.5}$$

▷ [To be continued on page 122.]

Example 6.9: Forbes' data on boiling points
The atmospheric pressure decreases with height above sea level, and hence you can predict altitude from measurements of barometric pressure. Since the boiling point of water decreases with barometric pressure, you can in fact calculate the altitude from measurements of the boiling point of water.—In 1840s and 1850s the Scottish physicist James D. Forbes performed a series of measurements at 17 different locations in the Swiss alps and in Scotland. He determined the boiling point of water and the barometric pressure (converted to pressure at a standard temperature of the air) (Forbes, 1857). The results are shown in Table 6.6 (from Weisberg (1980)).

A plot of barometric pressure against boiling point shows an evident correlation (Figure 6.1, bottom). Therefore a linear regression model with pressure as y and boiling point as x might be considered. If we ask a physicist, however, we learn that we should rather expect a linear relation between boiling point and the logarithm of the pressure, and that is indeed affirmed by the plot in Figure 6.1, top. In the following we will therefore try to fit a regression model where the logarithm of the barometric pressure is the y variable and the boiling point the x variable.

▷ [To be continued in Example 7.7 on page 123.]

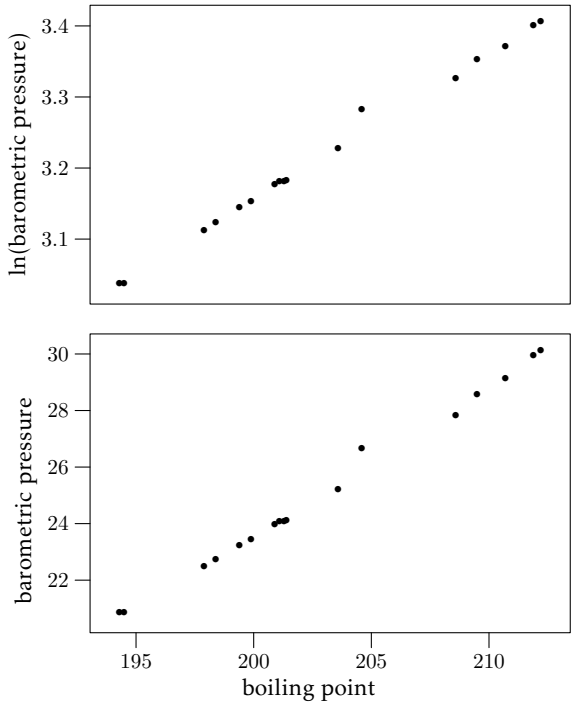


Figure 6.1
Forbes' data.
Top: logarithm of barometric pressure against boiling point.
Bottom: barometric pressure against boiling point.
Barometric pressure in inches of Mercury, boiling point in degrees F.

6.2 Exercises

Exercise 6.1
Explain how the binomial distribution is in fact an instance of the multinomial distribution.

Exercise 6.2
The binomial distribution was defined as the distribution of a sum of independent and identically distributed 01-variables (Definition 1.10 on page 30).

Could you think of a way to generalise that definition to a definition of the multinomial distribution as the distribution of a sum of independent and identically distributed random variables?

7 Estimation

The statistical model asserts that we may view the given set of data as an observation from a certain probability distribution which is completely specified except for a few unknown parameters. In this chapter we shall be dealing with the *problem of estimation*, that is, how to calculate, on the basis of model plus observations, an *estimate* of the unknown parameters of the model.

You cannot from purely mathematical reasoning reach a solution to the problem of estimation, somewhere you will have to introduce one or more external principles. Dependent on the principles chosen, you will get different methods of estimation. We shall now present a method which is the "usual" method in this country (and in many other countries). First some terminology:

- A *statistic* is a function defined on the observation space $\mathfrak{X}$ (and taking values in $\mathbb{R}$ or $\mathbb{R}^n$).
  If t is a statistic, then $t(\boldsymbol{X})$ will be a random variable; often one does not worry too much about distinguishing between t and $t(\boldsymbol{X})$.
- An *estimator* of the parameter θ is a statistic (or a random variable) taking values in the parameter space Θ. An estimator of $g(\theta)$ (where g is a function defined on Θ) is a statistic (or a random variable) taking values in $g(\Theta)$.
  It is more or less implied that the estimator should be an educated guess of the true value of the quantity it estimates.
- An *estimate* is the value of the estimator, so if t is an estimator (more correctly: if $t(\boldsymbol{X})$ is an estimator), then $t(\boldsymbol{x})$ is an estimate.
- An *unbiased* estimator of $g(\boldsymbol{\theta})$ is an estimator t with the property that for all $\boldsymbol{\theta}$, $\mathrm{E}_{\boldsymbol{\theta}}(t(\boldsymbol{X})) = g(\boldsymbol{\theta})$, that is, an estimator that on the average is right.

7.1 The maximum likelihood estimator

Consider a situation where we have an observation $\boldsymbol{x}$ which is assumed to be described by a statistical model given by the model function $f(\boldsymbol{x}, \boldsymbol{\theta})$. To find a suitable estimator for θ we will use a criterion based on the likelihood function $L(\boldsymbol{\theta}) = f(\boldsymbol{x}, \boldsymbol{\theta})$: it seems reasonable to say that if $L(\boldsymbol{\theta}_1) > L(\boldsymbol{\theta}_2)$, then $\boldsymbol{\theta}_1$ is a better estimate of θ than $\boldsymbol{\theta}_2$ is, and hence the best estimate of θ must be the value $\widehat{\boldsymbol{\theta}}$ that maximises L.

Definition 7.1
The maximum likelihood estimator is the function that for each observation $\boldsymbol{x} \in \mathfrak{X}$ returns the point of maximum $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ of the likelihood function corresponding to $\boldsymbol{x}$.

The maximum likelihood estimate is the value taken by the maximum likelihood estimator.

This definition is rather sloppy and incomplete: it is not obvious that the likelihood function has a single point of maximum, there may be several, or none at all. Here is a better definition:

Definition 7.2
A maximum likelihood estimate corresponding to the observation $\boldsymbol{x}$ is a point of maximum $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ of the likelihood function corresponding to $\boldsymbol{x}$.

A maximum likelihood estimator is a (not necessarily everywhere defined) function from $\mathfrak{X}$ to Θ, that maps an observation $\boldsymbol{x}$ into a corresponding maximum likelihood estimate.

This method of maximum likelihood is a proposal of a general method for calculating estimators. To judge whether it is a sensible method, let us ask a few questions and see how they are answered.

1. How easy is it to use the method in a concrete model?
   In practise what you have to do is to find point(s) of maximum of a real-valued function (the likelihood function), and that is a common and well-understood mathematical issue that can be addressed and solved using standard methods.
   Often it is technically simpler to find $\widehat{\boldsymbol{\theta}}$ as the point of maximum of the *log-likelihood function* $\ln L$. If $\ln L$ has a continuous derivative $D \ln L$, then the points of maximum in the interior of Θ are solutions of $D \ln L(\boldsymbol{\theta}) = \mathbf{0}$, where the differential operator D denotes differentiation with respect to $\boldsymbol{\theta}$.
2. Are there any general results about properties of the maximum likelihood estimator, for example about existence and uniqueness, and about how close $\widehat{\boldsymbol{\theta}}$ is to $\boldsymbol{\theta}$?
   Yes. There are theorems telling that, under certain conditions, then as the number of observations increases, the probability that there exists a unique maximum likelihood estimator tends to 1, and $\mathrm{P}_{\boldsymbol{\theta}}\big(|\widehat{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{\theta}| < \varepsilon\big)$ tends to 1 (for all $\varepsilon > 0$).
   Under stronger conditions, such as Θ be an open set, and the first three derivatives of $\ln L$ exist and satisfy some conditions of regularity, then as $n \to \infty$, the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}(\boldsymbol{X})$ is asymptotically normal with asymptotic mean $\boldsymbol{\theta}$ and an asymptotic variance which is the inverse of $\mathrm{E}_{\boldsymbol{\theta}}\big(-D^2 \ln L(\boldsymbol{\theta}; \boldsymbol{X})\big)$; moreover, $\mathrm{E}_{\boldsymbol{\theta}}\big(-D^2 \ln L(\boldsymbol{\theta}; \boldsymbol{X})\big) = \mathrm{Var}_{\boldsymbol{\theta}}\big(D \ln L(\boldsymbol{\theta}; \boldsymbol{X})\big)$.

(According to the so-called Cramér-Rao inequality this is the lower bound of the variance of a central estimator, so in this sense the maximum likelihood estimator is asymptotically optimal.)
In situations with a one-dimensional parameter θ, the asymptotic normality of $\widehat{\theta}$ means that the random variable

$$U = \frac{\widehat{\theta} - \theta}{\sqrt{\mathrm{E}_\theta\big(-D^2 \ln L(\theta; \boldsymbol{X})\big)}}$$

is asymptotically standard normal as $n \to \infty$, which is to say that $\lim_{n\to\infty} \mathrm{P}(U \leq u) = \Phi(u)$ for all $u \in \mathbb{R}$. (Φ is the distribution function of the standard normal distribution, cf. Definition 3.9 on page 72.)

3. Does the method produce estimators that seem sensible? In rare cases you can argue for an "obvious" common sense solution to the estimation problem, and our new method should not be too incompatible to common sense.
   This has to be settled through examples.

7.2 Examples

The one-sample problem with 01-variables

◁ [Continued from page 100.]
In this model the log-likelihood function and its first two derivatives are

$$\begin{aligned} \ln L(\theta) &= x_\bullet \ln\theta + (n - x_\bullet)\ln(1-\theta), \\ D\ln L(\theta) &= \frac{x_\bullet - n\theta}{\theta(1-\theta)}, \\ D^2 \ln L(\theta) &= -\frac{x_\bullet}{\theta^2} - \frac{n - x_\bullet}{(1-\theta)^2} \end{aligned}$$

for $0 < \theta < 1$. When $0 < x_\bullet < n$, the equation $D\ln L(\theta) = 0$ has the unique solution $\widehat{\theta} = x_\bullet/n$, and with a negative second derivative this must be the unique point of maximum. If $x_\bullet = n$, then L and $\ln L$ are strictly increasing, and if $x_\bullet = 0$, then L and $\ln L$ are strictly decreasing, so even in these cases $\widehat{\theta} = x_\bullet/n$ is the unique point of maximum. Hence the maximum likelihood estimate of θ is the relative frequency of 1s—which seems quite sensible.

The expected value of the estimator $\widehat{\theta} = \widehat{\theta}(\boldsymbol{X})$ is $\mathrm{E}_\theta\,\widehat{\theta} = \theta$, and its variance is $\mathrm{Var}_\theta\,\widehat{\theta} = \theta(1-\theta)/n$, cf. Example 1.19 on page 40 and the rules of calculus for expectations and variances.

The general theory (see above) tells that for large values of n, $\mathrm{E}_\theta\,\widehat{\theta} \approx \theta$ and $\mathrm{Var}_\theta\,\widehat{\theta} \approx \Big(\mathrm{E}_\theta\big(-D^2 \ln L(\theta, \boldsymbol{X})\big)\Big)^{-1} = \left(\mathrm{E}_\theta\left(\frac{X_\bullet}{\theta^2} + \frac{n - X_\bullet}{(1-\theta)^2}\right)\right)^{-1} = \theta(1-\theta)/n.$

▷ [To be continued on page 127.]

The simple binomial model

◁ [Continued from page 101.]

In the simple binomial model the likelihood function is

$$L(\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, \quad \theta \in [0\,;1]$$

and the log-likelihood function

$$\ln L(\theta) = \ln\binom{n}{y} + y\ln\theta + (n-y)\ln(1-\theta), \quad \theta \in [0\,;1].$$

Apart from a constant this function is identical to the log-likelihood function in the one-sample problem with 01-variables, so we get at once that the maximum likelihood estimator is $\widehat{\theta} = Y/n$. Since Y and $X_\bullet$ have the same distribution, the maximum likelihood estimators in the two models also have the same distribution, in particular we have $\mathrm{E}_\theta\,\widehat{\theta} = \theta$ and $\mathrm{Var}_\theta\,\widehat{\theta} = \theta(1-\theta)/n$.

▷ [To be continued on page 127.]

Example 7.1: Flour beetles I

◁ [Fortsat fra Example 6.3 page 101.]

In the practical example with flour beetles, $\widehat{\theta} = 43/144 = 0.30$. The estimated standard deviation is $\sqrt{\widehat{\theta}(1-\widehat{\theta})/144} = 0.04$.

▷ [To be continued in Example 8.1 on page 127.]

Comparison of binomial distributions

◁ [Continued from page 102.]

The log-likelihood function corresponding to $\boldsymbol{y}$ is

$$\ln L(\boldsymbol{\theta}) = \text{const} + \sum_{j=1}^{s}\Big(y_j \ln\theta_j + (n_j - y_j)\ln(1-\theta_j)\Big). \tag{7.1}$$

We see that $\ln L$ is a sum of terms, each of which (apart from a constant) is a log-likelihood function in a simple binomial model, and the parameter θ_j appears in the jth term only. Therefore we can write down the maximum likelihood estimator right away:

$$\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_s) = \left(\frac{Y_1}{n_1}, \frac{Y_2}{n_2}, \ldots, \frac{Y_s}{n_s}\right).$$

Since $Y_1, Y_2, \ldots, Y_s$ are independent, the estimators $\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_s$ are also independent, and from the simple binomial model we know that $\mathrm{E}\widehat{\theta}_j = \theta_j$ and $\mathrm{Var}\,\widehat{\theta}_j = \theta_j(1-\theta_j)/n_j$, $j = 1, 2, \ldots, s$.

▷ [To be continued on page 128.]

Example 7.2: Flour beetles II

◁ [Continued from Example 6.4 on page 102.]

In the flour beetle example where each group (or concentration) has its own probability parameter θ_j, this parameter is estimated as the fraction of dead in the actual group, i.e. $(\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3, \widehat{\theta}_4) = (0.30, 0.72, 0.87, 0.96)$.

The estimated standard deviations are $\sqrt{\widehat{\theta}_j(1-\widehat{\theta}_j)/n_j}$, i.e. 0.04, 0.05, 0.05 and 0.03.

▷ [To be continued in Example 8.2 on page 129.]

The multinomial distribution

◁ [Continued from page 104.]

If $\boldsymbol{y} = (y_1, y_2, \ldots, y_r)$ is an observation from a multinomial distribution with r classes and parameters n and $\boldsymbol{\theta}$, then the log-likelihood function is

$$\ln L(\boldsymbol{\theta}) = \text{const} + \sum_{i=1}^{r} y_i \ln \theta_i.$$

The parameter $\boldsymbol{\theta}$ is estimated as the point of maximum $\widehat{\boldsymbol{\theta}}$ (in Θ) of $\ln L$; the parameter space Θ is the set of r-tuples $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_r)$ satisfying $\theta_i \geq 0$, $i = 1, 2, \ldots, r$, and $\theta_1 + \theta_2 + \cdots + \theta_r = 1$. An obvious guess is that θ_i is estimated as the relative frequency y_i/n, and that is indeed the correct answer; but how to prove it?

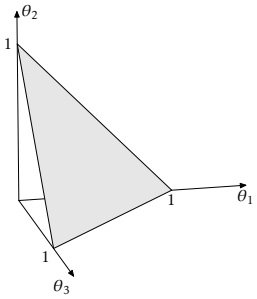
One approach would be to employ a general method for determining extrema subject to constraints. Another possibility is prove that the relative frequencies indeed maximise the likelihood function, that is, to show that if we let $\widehat{\theta}_i = y_i/n$, $i = 1, 2, \ldots, r$, and $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_r)$, then $\ln L(\boldsymbol{\theta}) \leq \ln L(\widehat{\boldsymbol{\theta}})$ for all $\boldsymbol{\theta} \in \Theta$:

The clever trick is to note that $\ln t \leq t - 1$ for all t (and with equality if and only if $t = 1$). Therefore

$$\begin{aligned}\ln L(\boldsymbol{\theta}) - \ln L(\widehat{\boldsymbol{\theta}}) &= \sum_{i=1}^{r} y_i \ln \frac{\theta_i}{\widehat{\theta}_i} \leq \sum_{i=1}^{r} y_i \left(\frac{\theta_i}{\widehat{\theta}_i} - 1 \right) \\ &= \sum_{i=1}^{r} \left(y_i \frac{\theta_i}{y_i/n} - y_i \right) = \sum_{i=1}^{r} (n\theta_i - y_i) = n - n = 0.\end{aligned}$$

The inequality is strict unless $\theta_i = \widehat{\theta}_i$ for all $i = 1, 2, \ldots, r$.

▷ [To be continued on page 130.]



The probability simplex in the three-dimensional space, i.e., the set of non-negative triples $(\theta_1, \theta_2, \theta_3)$ with $\theta_1 + \theta_2 + \theta_3 = 1$.

Example 7.3: Baltic cod

◁ [Continued from Example 6.5 on page 104.]

If we restrict our study to cod in the waters at Lolland, the job is to determine the point $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_1)$ in the three-dimensional probability simplex that maximises the log-likelihood function

$$\ln L(\boldsymbol{\theta}) = \text{const} + 27 \ln \theta_1 + 30 \ln \theta_2 + 12 \ln \theta_3.$$

From the previous we know that $\widehat{\theta_1} = 27/69 = 0.39$, $\widehat{\theta_2} = 30/69 = 0.43$ and $\widehat{\theta_3} = 12/69 = 0.17$.

▷ [To be continued in Example 8.3 on page 130.]

The one-sample problem with Poisson variables

◁ [Continued from page 105.]

The log-likelihood function and its first two derivatives are for $\mu > 0$

$$\begin{aligned} \ln L(\mu) &= \text{const} + y_\bullet \ln \mu - n\mu, \\ D \ln L(\mu) &= \frac{y_\bullet}{\mu} - n, \\ D^2 \ln L(\mu) &= -\frac{y_\bullet}{\mu^2}. \end{aligned}$$

When $y_\bullet > 0$, the equation $D \ln L(\mu) = 0$ has the unique solution $\widehat{\mu} = y_\bullet / n$, and since $D^2 \ln L$ is negative, this must be the unique point of maximum. It is easily seen that the formula $\widehat{\mu} = y_\bullet / n$ also gives the point of maximum in the case $y_\bullet = 0$.

In the Poisson distribution the variance equals the expected value, so by the usual rules of calculus, the expected value and the variance of the estimator $\widehat{\mu} = Y_\bullet / n$ are

$$\mathrm{E}_\mu \widehat{\mu} = \mu, \qquad \mathrm{Var}_\mu \widehat{\mu} = \mu / n. \tag{7.2}$$

The general theory (cf. page 114) can tell us that for large values of n, $\mathrm{E}_\mu \widehat{\mu} \approx \mu$ and $\mathrm{Var}_\mu \widehat{\mu} \approx \left(\mathrm{E}_\mu \left(-D^2 \ln L(\mu, \boldsymbol{Y}) \right) \right)^{-1} = \left(\mathrm{E}_\mu \left(\frac{Y_\bullet}{\mu^2} \right) \right)^{-1} = \mu / n$.

Example 7.4: Horse-kicks

◁ [Continued from Example 6.6 on page 105.]

In the horse-kicks example, $y_\bullet = 0 \cdot 109 + 1 \cdot 65 + 2 \cdot 22 + 3 \cdot 3 + 4 \cdot 1 = 122$, so $\widehat{\mu} = 122/200 = 0.61$.

Hence, the number of soldiers in a given regiment and a given year that are kicked to death by a horse is (according to the model) Poisson distributed with a parameter that is estimated to 0.61. The estimated standard deviation of the estimate is $\sqrt{\widehat{\mu}/n} = 0.06$, cf. equation (7.2).

Uniform distribution on an interval

◁ [Continued from page 106.]

The likelihood function is

$$L(\theta) = \begin{cases} 1/\theta^n & \text{if } x_{\max} < \theta \\ 0 & \text{otherwise.} \end{cases}$$

This function does not attain its maximum. Nevertheless it seems tempting to name $\widehat{\theta} = x_{\max}$ as the maximum likelihood estimate. (If instead we were considering the uniform distribution on the closed interval from 0 to θ, the estimation problem would of course have a nicer solution.)

The likelihood function is not differentiable in its entire domain (which is $]0\,;+\infty[$), so the regularity conditions that ensure the asymptotic normality of the maximum likelihood estimator (page 114) are not met, and the distribution of $\widehat{\theta} = X_{\max}$ is indeed not asymptotically normal (see Exercise 7.4).

The one-sample problem with normal variables

◁ [Continued from fra page 107.]

By solving the equation $D \ln L = \mathbf{0}$, where $\ln L$ is the log-likelihood function (6.3) on page 107, we obtain the maximum likelihood estimates for μ and σ^2:

$$\widehat{\mu} = \overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j, \qquad \widehat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n} (y_j - \overline{y})^2.$$

Usually, however, we prefer another estimate for the variance parameter σ^2, namely

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n} (y_j - \overline{y})^2.$$

One reason for this is that s^2 is an *unbiased* estimator; this is seen from Proposition 7.1 below, which is a special case of Theorem 11.1 on page 169, see also Section 11.2.

Proposition 7.1
Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed normal random variables with mean μ and variance σ^2. Then it is true that

1. *The random variable $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ follows a normal distribution with mean μ and variance σ^2/n.*

2. *The random variable* $s^2 = \dfrac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2$ *follows a gamma distribution with shape parameter* $f/2$ *and scale parameter* $2\sigma^2/f$ *where* $f = n-1$*, or put in another way,* $(f/\sigma^2)s^2$ *follows a* χ^2 *distribution with* f *degrees of freedom. From this follows among other things that* $\mathrm{E}\, s^2 = \sigma^2$.
3. *The two random variables* $\overline{X}$ *and* s^2 *are independent.*

Remarks: As a rule of thumb, the number of degrees of freedom of the variance estimate in a normal model can be calculated as the number of observations minus the number of free mean value parameters estimated. The degrees of freedom contains information about the precision of the variance estimate, see Exercise 7.3.

▷ [To be continued on page 131.]

Example 7.5: The speed of light

◁ [Continued from Example 6.7 on page 107.]

Assuming that the 64 positive values in Table 6.5 on page 107 are observations from a single normal distribution, the estimate of the mean of this normal distribution is $\overline{y} = 27.75$ and the estimate of the variance is $s^2 = 25.8$ with 63 degrees of freedom. Consequently, the estimated mean of the passage time is $(27.75 \times 10^{-3} + 24.8) \times 10^{-6}\text{sec} = 24.828 \times 10^{-6}\text{sec}$, and the estimated variance of the passage time is $25.8 \times (10^{-3} \times 10^{-6}\text{sec})^2 = 25.8 \times 10^{-6}(10^{-6}\text{sec})^2$ with 63 degrees of freedom, i.e., the estimated standard deviation is $\sqrt{25.8 \times 10^{-6}}10^{-6}\text{sec} = 0.005 \times 10^{-6}\text{sec}$.

▷ [To be continued in Example 8.4 on page 133.]

The two-sample problem with normal variables

◁ [Continued from page 109.]

The log-likelihood function in this model is given in (6.4) on page 109. It attains its maximum at the point $(\overline{y}_1, \overline{y}_2, \widehat{\sigma}^2)$, where $\overline{y}_1$ and $\overline{y}_2$ are the group averages, and $\widehat{\sigma}^2 = \dfrac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2$ (here $n = n_1 + n_2$). Often one does not use $\widehat{\sigma}^2$ as an estimate of σ^2, but rather the central estimator

$$s_0^2 = \frac{1}{n-2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2.$$

The denominator $n-2$, the number of *degrees of freedom*, causes the estimator to become unbiased; this is seen from the propositon below, which is a special case of Theorem 11.1 on page 169, see also Section 11.3.

	n	S	$\overline{y}$	f	SS	s^2
orange juice	10	131.8	13.18	9	177.236	19.69
synthetic vitamin C	10	80.0	8.00	9	68.960	7.66
sum	20	211.8		18	246.196	
average			10.59			13.68

Table 7.1
Vitamin C-example, calculations.
n stands for number of observations y, S for Sum of ys, $\overline{y}$ for average of ys, f for degrees of freedom, SS for Sum of Squared deviations, and s^2 for estimated variance ($s^2 = SS/f$).

Proposition 7.2
For independent normal random variables X_{ij} with mean values $\mathrm{E}\,X_{ij} = \mu_i$, $j = 1, 2, \ldots, n_i$, $i = 1, 2$, and all with the same variance σ^2, the following holds:

1. *The random variables $\overline{X}_i = \dfrac{1}{n_i}\displaystyle\sum_{j=1}^{n_i} X_{ij}$, $i = 1, 2$, follow normal distributions with mean μ_i and variance σ^2/n_i, $i = 1, 2$.*
2. *The random variable $s_0^2 = \dfrac{1}{n-2}\displaystyle\sum_{i=1}^{2}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2$ follows a gamma distribution with shape parameter $f/2$ and scale parameter $2\sigma^2/f$ where $f = n - 2$ and $n = n_1 + n_2$, or put in another way, $(f/\sigma^2)s_0^2$ follows a χ^2 distribution with f degrees of freedom.*
   From this follows among other things that $\mathrm{E}\,s_0^2 = \sigma^2$.
3. *The three random variables $\overline{X}_1$, $\overline{X}_2$ and s_0^2 are independent.*

Remarks:

- In general the number of degrees of freedom of a variance estimate is the number of observations minus the number of free mean value parameters estimated.
- A quantity such as $y_{ij} - \overline{y}_i$, which is the difference between an actual observed value and the best "fit" provided by the actual model, is sometimes called a *residual*. The quantity $\displaystyle\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2$ could therefore be called a *residual sum of squares*.

▷ [To be continued on page 133.]

Example 7.6: Vitamin C

◁ [Continued from Example 6.8 on page 109.]

Table 7.1 gives the parameter estimates and some intermediate results. The estimated mean values are 13.18 (orange juice group) and 8.00 (synthetic vitamin C). The estimate

of the common variance is 13.68 on 18 degrees of freedom, and since both groups have 10 observations, the estimated standard deviation of each mean is $\sqrt{13.68/10} = 1.17$.

▷ [To be continued in Example 8.5 on page 135.]

Simple linear regression

◁ [Continued from page 109.]

We shall estimate the parameters α, β and σ^2 in the linear regression model. The log-likelihood function is given in equation (6.5) on page 110. The sum of squares can be partioned like this:

$$\begin{aligned}\sum_{i=1}^{n}\Big(y_i-(\alpha+\beta x_i)\Big)^2 &= \sum_{i=1}^{n}\Big((y_i-\overline{y})+(\overline{y}-\alpha-\beta\overline{x})-\beta(x_i-\overline{x})\Big)^2\\ &= \sum_{i=1}^{n}(y_i-\overline{y})^2+\beta^2\sum_{i=1}^{n}(x_i-\overline{x})^2\\ &\quad -2\beta\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})+n(\overline{y}-\alpha-\beta\overline{x})^2,\end{aligned}$$

since the other two double products from squaring the trinomial both sum to 0. In the final expression for the sum of squares, the unknown α enters in the last term only; the minimum value of that term is 0, which is attained if and only if $\alpha = \overline{y} - \beta\overline{x}$. The remaining terms constitute a quadratic in β, and this quadratic has the single point of minimum $\beta = \dfrac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2}$. Hence the maximum likelihood estimates are

$$\widehat{\beta} = \frac{\displaystyle\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\displaystyle\sum_{i=1}^{n}(x_i-\overline{x})^2} \qquad \text{and} \qquad \widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x}.$$

(It is tacitly assumed that $\sum(x_i-\overline{x})^2$ is non-zero, i.e., the xs are not all equal.—It makes no sense to estimate a regression line in the case where all xs are equal.)

The *estimated regression line* is the line whose equation is $y = \widehat{\alpha} + \widehat{\beta}x$.

The maximum likelihood estimate of σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\Big(y_i-(\widehat{\alpha}+\widehat{\beta}x_i)\Big)^2,$$

but usually one uses the *unbiased* estimator

$$s_{02}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\Big(y_i-(\widehat{\alpha}+\widehat{\beta}x_i)\Big)^2.$$

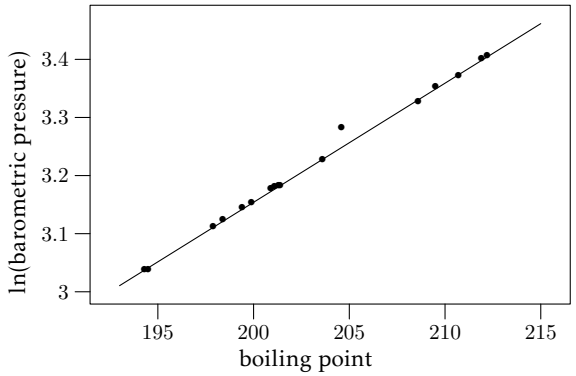


Figure 7.1
Forbes' data: Data points and estimated regression line.

with $n-2$ degrees of freedom. With the notation $SS_x = \sum_{i=1}^{n}(x_i - \overline{x})^2$ we have (cf. Theorem 11.1 on page 169 and Section 11.6):

Proposition 7.3
The estimators $\widehat{\alpha}$, $\widehat{\beta}$ and s_{02}^2 in the linear regression model have the following properties:

1. *$\widehat{\beta}$ follows a normal distribution with mean β and variance σ^2/SS_x.*
2. *a) $\widehat{\alpha}$ follows a normal distribution with mean α and with variance $\sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{SS_x}\right)$.*
   b) $\widehat{\alpha}$ and $\widehat{\beta}$ are correlated with correlation $-1\Big/\sqrt{1 + \frac{SS_x}{n\overline{x}^2}}$.
3. *a) $\widehat{\alpha} + \widehat{\beta}\overline{x}$ follows a normal distribution with mean $\alpha + \beta\overline{x}$ and variance σ^2/n.*
   b) $\widehat{\alpha} + \widehat{\beta}\overline{x}$ and $\widehat{\beta}$ are independent.
4. *The variance estimator s_{02}^2 is independent of the estimators of the mean value parameters, and it follows a gamma distribution with shape parameter $f/2$ and scale parameter $2\sigma^2/f$ where $f = n-2$, or put in another way, $(f/\sigma^2)s_{02}^2$ follows a χ^2 distribution with f degrees of freedom.*
   From this follows among other things that $\mathrm{E}\, s_{02}^2 = \sigma^2$.

Example 7.7: Forbes' data on boiling points
◁ [Continued from Example 6.9 on page 110.]
As is seen from Figure 6.1, there is a single data point that deviates rather much from the ordinary pattern. We choose to discard this point, which leaves us with 16 data points.

The estimated regression line is

$$\ln(\text{pressure}) = 0.0205 \cdot \text{boiling point} - 0.95$$

and the estimated variance $s_{02}^2 = 6.84 \times 10^{-6}$ on 14 degrees of freedom. Figure 7.1 shows

the data points together with the estimated line which seems to give a good description of the points.

To make any practical use of such boiling point measurements you need to know the relation between altitude and barometric pressure. With altitudes of the size of a few kilometres, the pressure decreases exponentially with altitude; if p_0 denotes the barometric pressure at sea level (e.g. 1013.25 hPa) and p_h the pressure at an altitude h, then $h \approx 8150\,\mathrm{m} \cdot (\ln p_0 - \ln p_h)$.

7.3 Exercises

Exercise 7.1
In Example 4.6 on page 81 it is argued that the number of mites on apple leaves follows a negative binomial distribution. Write down the model function and the likelihood function, and estimate the parameters.

Exercise 7.2
Find the expected value of the maximum likelihood estimator $\widehat{\sigma}^2$ for the variance parameter σ^2 in the one-sample problem with normal variables.

Exercise 7.3
Find the variance of the variance estimator s^2 in the one-sample problem with normal variables.

Exercise 7.4
In the continuous uniform distribution on $]0\,;\theta[$, the maximum likelihood estimator was found (on page 119) to be $\widehat{\theta} = X_{\max}$ where $X_{\max} = \max\{X_1, X_2, \ldots, X_n\}$.

1. Show that the probability density function of $X_{\max}$ is

$$f(x) = \begin{cases} \dfrac{nx^{n-1}}{\theta^n} & \text{for } 0 < x < \theta \\ 0 & \text{otherwise.} \end{cases}$$

   Hint: first find $\mathrm{P}(X_{\max} \leq x)$.
2. Find the expected value and the variance of $\widehat{\theta} = X_{\max}$, and show that for large n, $\mathrm{E}\,\widehat{\theta} \approx \theta$ and $\mathrm{Var}\,\widehat{\theta} \approx \theta^2/n^2$.
3. We are now in a position to find the asymptotic distribution of $\dfrac{\widehat{\theta} - \theta}{\sqrt{\mathrm{Var}\,\widehat{\theta}}}$. For large n we have, using item 2, that $\dfrac{\widehat{\theta} - \theta}{\sqrt{\mathrm{Var}\,\widehat{\theta}}} \approx -Y$ where $Y = \dfrac{\theta - X_{\max}}{\theta/n} = n\left(1 - \dfrac{X_{\max}}{\theta}\right)$.
   Prove that $\mathrm{P}(Y > y) = \left(1 - \dfrac{y}{n}\right)^n$, and conclude that the distribution of Y asymptotically is an exponential distribution with scale parameter 1.

8 Hypothesis Testing

Consider a general statistical model with model function $f(\boldsymbol{x}, \boldsymbol{\theta})$, where $\boldsymbol{x} \in \mathfrak{X}$ and $\boldsymbol{\theta} \in \Theta$. A *statistical hypothesis* H_0 is an assertion that the true value of the parameter $\boldsymbol{\theta}$ actually belongs to a certain subset Θ_0 of Θ, formally

$$H_0 : \boldsymbol{\theta} \in \Theta_0$$

where $\Theta_0 \subset \Theta$. Thus, the statistical hypothesis claims that we may use a simpler model.

A *test* of the hypothesis is an assessment of the degree of concordance between the hypothesis and the observations actually available. The test is based on a suitably selected one-dimensional statistic t, the *test statistic*, which is designed to measure the deviation between observations and hypothesis. The *test probability* is the probability of getting a value of $\boldsymbol{X}$ which is less concordant (as measured by the test statistic t) with the hypothesis than the actual observation $\boldsymbol{x}$ is, calculated "under the hypothesis" [i.e. the test probability is calculated under the assumption that the hypothesis is true]. If there is very little concordance between the model and the observations, then the hypothesis is *rejected*.

8.1 The likelihood ratio test

Just as we in Chapter 7 constructed an estimation procedure based on the likelihood function, we can now construct a procedure of hypothesis testing based on the likelihood function. The general form of this procedure is as follows:

1. Determine the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ in the full model and the maximum likelihood estimator $\widehat{\widehat{\boldsymbol{\theta}}}$ under the hypothesis, i.e. $\widehat{\boldsymbol{\theta}}$ is a point of maximum of $\ln L$ in Θ, and $\widehat{\widehat{\boldsymbol{\theta}}}$ is a point of maximum of $\ln L$ in Θ_0.
2. To test the hypothesis we compare the maximal value of the likelihood function under the hypothesis to the maximal value in the full model, i.e. we compare the best description of $\boldsymbol{x}$ obtainable under the hypothesis to the best description obtainable with the full model. This is done using the likelihood ratio test statistic

$$Q = \frac{L(\widehat{\widehat{\boldsymbol{\theta}}})}{L(\widehat{\boldsymbol{\theta}})} = \frac{L(\widehat{\widehat{\boldsymbol{\theta}}}(\boldsymbol{x}), \boldsymbol{x})}{L(\widehat{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{x})}.$$

Per definition Q is always between 0 and 1, and the closer Q is to 0, the less agreement between the observation $\boldsymbol{x}$ and the hypothesis. It is often more convenient to use $-2\ln Q$ instead of Q; the test statistic $-2\ln Q$ is always non-negative, and the larger its value, the less agreement between the observation $\boldsymbol{x}$ and the hypothesis.

3. The test probability ε is the probability (calculated under the hypothesis) of getting a value of $\boldsymbol{X}$ that agrees less with the hypothesis than the actual observation $\boldsymbol{x}$ does, in mathematics:

$$\varepsilon = \mathrm{P}_0\Big(Q(\boldsymbol{X}) \le Q(\boldsymbol{x})\Big) = \mathrm{P}_0\Big(-2\ln Q(\boldsymbol{X}) \ge -2\ln Q(\boldsymbol{x})\Big)$$

or simply

$$\varepsilon = \mathrm{P}_0(Q \le Q_{\mathrm{obs}}) = \mathrm{P}_0\Big(-2\ln Q \ge -2\ln Q_{\mathrm{obs}}\Big).$$

(The subscript 0 on P indicates that the probability is to be calculated under the assumption that the hypothesis is true.)

4. If the test probability is small, then the hypothesis is rejected.
Often one adopts a strategy of rejecting a hypothesis if and only if the test probability is below a certain *level of significance* α which has been fixed in advance. Typical values of the significance level are 5%, 2.5% and 1%. We can interprete α as the probability of rejecting the hypothesis when it is true.

5. In order to calculate the test probability we need to know the distribution of the test statistic under the hypothesis.
In some situations, including normal models (see page 131ff and Chapter 11), exact test probabilities can be obtained by means of known and tabulated standard distributions (t, χ^2 and F distributions).
In other situations, general results can tell that in certain circumstances (under assumptions that assure the asymptotic normality of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\widehat{\boldsymbol{\theta}}}$, cf. page 114), the asymptotic distribution of $-2\ln Q$ is a χ^2 distribution with $\dim\Theta - \dim\Theta_0$ degrees of freedom (this requires that we extend the definition of dimension to cover cases where Θ is something like a differentiable surface). In most cases however, the number of degrees of freedom equals "the number of free parameters in the full model minus the number of free parameters under the hypothesis".

8.2 Examples

The one-sample problem with 01-variables

◁ [Continued from page 116.]

Assume that in the actual problem, the parameter θ is supposed to be equal to some known value θ_0, so that it is of interest to test the statistical hypothesis $H_0 : \theta = \theta_0$ (or written out in more details: the statistical hypothesis $H_0 : \theta \in \Theta_0$ where $\Theta_0 = \{\theta_0\}$.)

Since $\widehat{\theta} = x_\bullet/n$, the likelihood ratio test statistic is

$$Q = \frac{L(\theta_0)}{L(\widehat{\theta})} = \frac{\theta_0^{x_\bullet}(1-\theta_0)^{n-x_\bullet}}{\widehat{\theta}^{x_\bullet}(1-\widehat{\theta})^{n-x_\bullet}} = \left(\frac{n\theta_0}{x_\bullet}\right)^{x_\bullet}\left(\frac{n-n\theta_0}{n-x_\bullet}\right)^{n-x_\bullet},$$

and

$$-2\ln Q = 2\left(x_\bullet \ln\frac{x_\bullet}{n\theta_0} + (n-x_\bullet)\ln\frac{n-x_\bullet}{n-n\theta_0}\right).$$

The exact test probability is

$$\varepsilon = \mathrm{P}_{\theta_0}\left(-2\ln Q \geq -2\ln Q_{\mathrm{obs}}\right),$$

and an approximation to this can be calculated as the probability of getting values exceeding $-2\ln Q_{\mathrm{obs}}$ in the χ^2 distribution with $1-0=1$ degree of freedom; this approximation is adequate when the "expected" numbers $n\theta_0$ and $n-n\theta_0$ both are at least 5.

The simple binomial model

◁ [Continued from page 116.]

Assume that we want to test a statistical hypothesis $H_0 : \theta = \theta_0$ in the simple binomial model. Apart from a constant factor the likelihood function in the simple binomial model is identical to the likelihood function in the one-sample problem with 01-variables, and therefore the two models have the same test statistics Q and $-2\ln Q$, so

$$-2\ln Q = 2\left(y\ln\frac{y}{n\theta_0} + (n-y)\ln\frac{n-y}{n-n\theta_0}\right),$$

and the test probabilities are found in the same way in both cases.

Example 8.1: Flour beetles I

◁ [Continued from Example 7.1 on page 116.]

Assume that there is a certain standard version of the poison which is known to kill 23% of the beetles. [Actually, that is not so. This part of the example is pure imagination.]

It would then be of interest to judge whether the poison actually used has the same power as the standard poison. In the language of the statistical model this means that we should test the statistical hypothesis $H_0 : \theta = 0.23$.

When $n = 144$ and $\theta_0 = 0.23$, then $n\theta_0 = 33.12$ and $n - n\theta_0 = 110.88$, so $-2\ln Q$ considered as a function of y is

$$-2\ln Q(y) = 2\Big(y \ln \frac{y}{33.12} + (144 - y) \ln \frac{144 - y}{110.88}\Big)$$

and $-2\ln Q_{\text{obs}} = -2\ln Q(43) = 3.60$. The exact test probability is

$$\varepsilon = \mathrm{P}_0\big(-2\ln Q(Y) \geq 3.60\big) = \sum_{y\,:\,-2\ln Q(y) \geq 3.60} \binom{144}{y} 0.23^y\, 0.77^{144-y}.$$

Standard calculations show that the inequality $-2\ln Q(y) \geq 3.60$ holds for the y-values $0, 1, 2, \ldots, 23$ and $43, 44, 45, \ldots, 144$. Furthermore, $\mathrm{P}_0(Y \leq 23) = 0.0249$ and $\mathrm{P}_0(Y \geq 43) = 0.0344$, so that $\varepsilon = 0.0249 + 0.0344 = 0.0593$.

To avoid a good deal of calculations one can instead use the χ^2 approximation of $-2\ln Q$ to obtain the test probability. Since the full model has one unknown parameter and the hypothesis leaves us with zero unknown parameters, the χ^2 distribution in question is the one with $1 - 0 = 1$ degree of freedom. A table of quantiles of the χ^2 distribution with 1 degree of freedom (see for example page 208) reveals that the 90% quantile is 2.71 and the 95% quantile 3.84, so that the test probability is somewhere between 5% and 10%; standard statistical computer software has functions that calculate distribution functions of most standard distributions, and using such software we will see that in the χ^2 distribution with 1 degree of freedom, the probability of getting values larger that 3.60 is 5.78%, which is quite close to the exact value of 5.93%.

The test probability exceeds 5%, so with the common rule of thumb of a 5% level of significance, we cannot reject the hypothesis; hence the actual observations do not disagree significantly with the hypothesis that the actual poison has the same power as the standard poison.

Comparison of binomial distributions

◁ [Continued from page 116.]

This is the problem of investigating whether s different binomial distributions actually have the same probability parameter. In terms of the statistical model this becomes the statistical hypothesis $H_0 : \theta_1 = \theta_2 = \cdots = \theta_s$, or more accurately, $H_0 : \boldsymbol{\theta} \in \Theta_0$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_s)$ and where the parameter space is

$$\Theta_0 = \{\boldsymbol{\theta} \in [0\,;1]^s : \theta_1 = \theta_2 = \cdots = \theta_s\}.$$

In order to test H_0 we need the maximum likelihood estimator under H_0, i.e. we must find the point of maximum of $\ln L$ restricted to Θ_0. The log-likelihood

function is given as equation (7.1) on page 116, and when all the θs are equal,

$$\begin{aligned}\ln L(\theta,\theta,\ldots,\theta) &= \text{const} + \sum_{j=1}^{s}\Big(y_j \ln\theta + (n_j - y_j)\ln(1-\theta)\Big)\\ &= \text{const} + y_\bullet \ln\theta + (n_\bullet - y_\bullet)\ln(1-\theta).\end{aligned}$$

This restricted log-likelihood function attains its maximum at $\widehat{\theta} = y_\bullet/n_\bullet$; here $y_\bullet = y_1 + y_2 + \ldots + y_s$ and $n_\bullet = n_1 + n_2 + \ldots + n_s$. Hence the test statistic is

$$\begin{aligned}-2\ln Q &= -2\ln\frac{L(\widehat{\theta},\widehat{\theta},\ldots,\widehat{\theta})}{L(\widehat{\theta}_1,\widehat{\theta}_2,\ldots,\widehat{\theta}_s)}\\ &= 2\sum_{j=1}^{s}\left(y_j \ln\frac{\widehat{\theta}_j}{\widehat{\theta}} + (n_j - y_j)\ln\frac{1-\widehat{\theta}_j}{1-\widehat{\theta}}\right)\\ &= 2\sum_{j=1}^{s}\left(y_j \ln\frac{y_j}{\widehat{y}_j} + (n_j - y_j)\ln\frac{n_j - y_j}{n_j - \widehat{y}_j}\right)\end{aligned} \tag{8.1}$$

where $\widehat{y}_j = n_j\widehat{\theta}$ is the "expected number" of successes in group j under H_0, that is, when the groups have the same probability parameter. The last expression for $-2\ln Q$ shows how the test statistic compares the observed numbers of successes and failures (the y_js and $(n_j - y_j)$s) to the "expected" numbers of successes and failures (the $\widehat{y}_j$s and $(n_j - \widehat{y}_j)$s). The test probability is

$$\varepsilon = \mathrm{P}_0(-2\ln Q \geq -2\ln Q_{\text{obs}})$$

where the subscript 0 on P indicates that the probability is to be calculated under the assumption that the hypothesis is true.* If so, the asymptotic distribution of $-2\ln Q$ is a χ^2 distribution with $s-1$ degrees of freedom. As a rule of thumb, the χ^2 approximation is adequate when all of the $2s$ "expected" numbers exceed 5.

Example 8.2: Flour beetles II

◁ [Continued from Example 7.2 on page 117.]

The purpose of the investigation is to see whether the effect of the poison varies with the dose given. The way the statistical analysis is carried out in such a situation is to see whether the observations are consistent with an assumption of *no* dose effect, or in model terms: to test the statistical hypothesis

$$H_0 : \theta_1 = \theta_2 = \theta_3 = \theta_4.$$

Under H_0 the θs have a common value which is estimated as $\widehat{\theta} = y_\bullet/n_\bullet = 188/317 = 0.59$. Then we can calculate the "expected" numbers $\widehat{y}_j = n_j\widehat{\theta}$ and $n_j - \widehat{y}_j$, see Table 8.1. The

* This less innocuous that it may appear. Even when the hypothesis is true, there is still an unknown parameter, the common value of the θs.

Table 8.1
Survival of flour beetles at four different doses of poison: expected numbers under the hypothesis of no dose effect.

	concentration			
	0.20	0.32	0.50	0.80
dead	85.4	40.9	32.0	29.7
alive	58.6	28.1	22.0	20.3
total	144	69	54	50

likelihood ratio test statistic $-2\ln Q$ (eq. (8.1)) compares these numbers to the observed numbers from Table 6.2 on page 103:

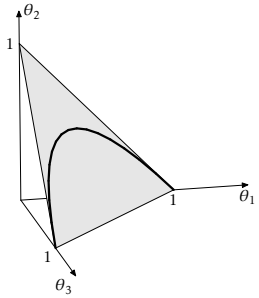
$$-2\ln Q_{\text{obs}} = 2\Big(43\ln\frac{43}{85.4} + 50\ln\frac{50}{40.9} + 47\ln\frac{47}{32.0} + 48\ln\frac{48}{29.7} \\ 101\ln\frac{101}{58.6} + 19\ln\frac{19}{28.1} + 7\ln\frac{7}{22.0} + 2\ln\frac{2}{20.3}\Big) = 113.1.$$

The full model has four unknown parameters, and under H_0 there is just a single unknown parameter, so we shall compare the $-2\ln Q$ value to the χ^2 distribution with $4-1=3$ degrees of freedom. In this distribution the 99.9% quantile is 16.27 (see for example the table on page 208), so the probability of getting a larger, i.e. more significant, value of $-2\ln Q$ is far below 0.01%, if the hypothesis is true. This leads us to rejecting the hypothesis, and we may conclude that the four doses have significantly different effects.

The multinomial distribution

◁ [Continued from page 117.]

In certain situations the probability parameter $\boldsymbol{\theta}$ is known to vary in some subset Θ_0 of the parameter space. If Θ_0 is a differentiable curve or surface then the problem is a "nice" problem, from a mathematically point of view.



The shaded area is the probability simplex, i.e. the set of triples $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ of non-negative numbers adding to 1. The curve is the set Θ_0, the values of $\boldsymbol{\theta}$ corresponding to Hardy-Weinberg equilibrium.

Example 8.3: Baltic cod

◁ [Continued from Example 7.3 on page 118.]

Simple reasoning, which is left out here, will show that in a closed population with random mating, the three genotypes occur, in the equilibrium state, with probabilities $\theta_1 = \beta^2$, $\theta_2 = 2\beta(1-\beta)$ and $\theta_3 = (1-\beta)^2$ where β is the fraction of A-genes in the population (β is the probability that a random gene is A).—This situation is know as a Hardy-Weinberg equilibrium.

For each population one can test whether there actually is a Hardy-Weinberg equilibrium. Here we will consider the Lolland population. On the basis of the observation $\boldsymbol{y}_L = (27, 30, 12)$ from a trinomial distribution with probability parameter $\boldsymbol{\theta}_L$ we shall test the statistical hypothesis $H_0 : \boldsymbol{\theta}_L \in \Theta_0$ where Θ_0 is the range of the map $\beta \mapsto \left(\beta^2, 2\beta(1-\beta), (1-\beta)^2\right)$ from $[0\,;1]$ into the probability simplex,

$$\Theta_0 = \left\{ \left(\beta^2, 2\beta(1-\beta), (1-\beta)^2\right) \,:\, \beta \in [0\,;1] \right\}.$$

Before we can test the hypothesis, we need an estimate of β. Under H_0 the log-likelihood function is

$$\begin{aligned}\ln L_0(\beta) &= \ln L\Big(\beta^2, 2\beta(1-\beta), (1-\beta)^2\Big)\\ &= \text{const} + 2\cdot 27\ln\beta + 30\ln\beta + 30\ln(1-\beta) + 2\cdot 12\ln(1-\beta)\\ &= \text{const} + (2\cdot 27 + 30)\ln\beta + (30 + 2\cdot 12)\ln(1-\beta)\end{aligned}$$

which attains its maximum at $\widehat{\beta} = \dfrac{2\cdot 27 + 30}{2\cdot 69} = \dfrac{84}{138} = 0.609$ (that is, β is estimated as twice the observed numbers of A genes divided by the total number of genes). The likelihood ratio test statistic is

$$-2\ln Q = -2\ln\frac{L\Big(\widehat{\beta}^2, 2\widehat{\beta}(1-\widehat{\beta}), (1-\widehat{\beta})^2\Big)}{L(\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3)} = 2\sum_{i=1}^{3} y_i \ln\frac{y_i}{\widehat{y}_i}$$

where $\widehat{y}_1 = 69\cdot\widehat{\beta}^2 = 25.6$, $\widehat{y}_2 = 69\cdot 2\widehat{\beta}(1-\widehat{\beta}) = 32.9$ and $\widehat{y}_3 = 69\cdot(1-\widehat{\beta})^2 = 10.6$ are the "expected" numbers under H_0.

The full model has two free parameters (there are three θs, but they add to 1), and under H_0 we have one single parameter (β), hence $2 - 1 = 1$ degree of freedom for $-2\ln Q$.

We get the value $-2\ln Q = 0.52$, and with 1 degree of freedom this corresponds to a test probability of about 47%, so we can easily assume a Hardy-Weinberg equilibrium in the Lolland population.

The one-sample problem with normal variables

◁ [Continued from page 120.]

Suppose that one wants to test a hypothesis that the mean value of a normal distribution has a given value; in the formal mathematical language we are going to test a hypothesis $H_0 : \mu = \mu_0$. Under H_0 there is still an unknown parameter, namely σ^2. The maximum likelihood estimate of σ^2 under H_0 is the point of maximum of the log-likelihood function (6.3) on page 107, now considered as a function of σ^2 alone (since μ has the fixed value μ_0), that is, the function

$$\sigma^2 \mapsto \ln L(\mu_0, \sigma^2) = \text{const} - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j - \mu_0)^2$$

which attains its maximum when σ^2 equals $\widehat{\widehat{\sigma}}^2 = \dfrac{1}{n}\displaystyle\sum_{j=1}^{n}(y_j - \mu_0)^2$. The likelihood ratio test statistic is $Q = \dfrac{L(\mu_0, \widehat{\widehat{\sigma}}^2)}{L(\widehat{\mu}, \widehat{\sigma}^2)}$. Standard caclulations lead to

$$-2\ln Q = 2\Big(\ln L(\overline{y}, \widehat{\sigma}^2) - \ln L(\mu_0, \widehat{\widehat{\sigma}}^2)\Big) = n\ln\Big(1 + \frac{t^2}{n-1}\Big)$$

where

$$t = \frac{\overline{y} - \mu_0}{\sqrt{s^2/n}}. \tag{8.2}$$

The hypothesis is rejected for large values of $-2\ln Q$, i.e. for large values of $|t|$. The standard deviation of $\overline{Y}$ equals $\sqrt{\sigma^2/n}$ (Proposition 7.1 page 119), so the t test statistic measures the difference between $\overline{y}$ and μ_0 in units of estimated standard deviations of $\overline{Y}$. The likelihood ratio test is therefore equivalent to a test based on the more directly understable t test statistic.

The test probability is $\varepsilon = \mathrm{P}_0(|t| > |t_{\mathrm{obs}}|)$. The distribution of t under H_0 is known, as the next proposition shows.

t distribution.
The t distribution with f degrees of freedom is the continuous distribution with density function

$$\frac{\Gamma(\frac{f+1}{2})}{\sqrt{\pi f}\,\Gamma(\frac{f}{2})}\left(1+\frac{x^2}{f}\right)^{-\frac{f+1}{2}}$$

where $x \in \mathbb{R}$.

Proposition 8.1
Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed normal random variables with mean value μ_0 and variance σ^2, and let $\overline{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$ and $s^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$. Then the random variable $t = \dfrac{\overline{X} - \mu_0}{\sqrt{s^2/n}}$ follows a t distribution with $n-1$ degrees of freedom.

Additional remarks:

- It is a major point that the distribution of t under H_0 does not depend on the unknown parameter σ^2 (and neither does it depend on μ_0). Furthermore, t is a dimensionless quantity (test statistics should always be dimensionless). See also Exercise 8.1.
- The t distribution is symmetric about 0, and consequently the test probability can be calculated as $\mathrm{P}(|t| > |t_{\mathrm{obs}}|) = 2\,\mathrm{P}(t > |t_{\mathrm{obs}}|)$.
  In certain situations it can be argued that the hypothesis should be rejected for large positive values of t only (i.e. when $\overline{y}$ is larger than μ_0); this could be the case if it is known in advance that the alternative to $\mu = \mu_0$ is $\mu > \mu_0$. In such a situation the test probability is calculated as $\varepsilon = \mathrm{P}(t > t_{\mathrm{obs}})$.
  Similarly, if the alternative to $\mu = \mu_0$ is $\mu < \mu_0$, the test probability is calculated as $\varepsilon = \mathrm{P}(t < t_{\mathrm{obs}})$.
  A t test that rejects the hypothesis for large values of $|t|$, is a *two-sided* t test, and a t test that rejects the hypothesis when t is large and positive [or large and negative], is a *one-sided* test.
- Quantiles of the t distribution are obtained from the computer or from a collection of statistical tables. (On page 214 you will find a short table of the t distribution.)
  Since the distribution is symmetric about 0, quantiles usually are tabulated for probabilities larger that 0.5 only.

William Sealy Gosset (1876-1937). Gosset worked at Guinness Brewery in Dublin, with a keen interest in the then very young science of statistics. He developed the t test during his work on issues of quality control related to the brewing process, and found the distribution of the t statistic.

- The t statistic is also known as *Student's t* in honour of W.S. Gosset, who in 1908 published the first article about this test, and who wrote under the pseudonym 'Student'.
- See also Section 11.2 on page 171.

Example 8.4: The speed of light

◁ [Continued from Example 7.5 on page 120.]

Nowadays, one metre is defined as the distance that the light travels in vacuum in 1/299 792 458 of a second, so light has the known speed of 299 792 458 metres per second, and therefore it will use $\tau_0 = 2.48238 \times 10^{-5}$ seconds to travel a distance of 7442 metres. The value of τ_0 must be converted to the same scale as the values in Table 6.5 (page 107): $((\tau_0 \times 10^6) - 24.8) \times 10^3 = 23.8$. It is now of some interest to see whether the data in Table 6.5 are consistent with the hypothesis that the mean is 23.8, so we will test the statistical hypothesis $H_0 : \mu = 23.8$.

From a previous example, $\overline{y} = 27.75$ and $s^2 = 25.8$, so the t statistic is

$$t = \frac{27.75 - 23.8}{\sqrt{25.8/64}} = 6.2 \, .$$

The test probability is the probability of getting a value of t larger than 6.2 or smaller that -6.2. From a table of quantiles in the t distribution (for example the table on page 214) we see that with 63 degrees of freedom the 99.95% quantile is slightly larger than 3.4, so there is a probability of less than 0.05% of getting a t larger than 6.2, and the test probability is therefore less than $2 \times 0.05\% = 0.1\%$. This means that we should reject the hypothesis. Thus, Newcomb's measurements of the passage time of the light are not in accordance with the today's knowledge and standards.

The two-sample problem with normal variables

◁ [Continued from page 121.]

We are going to test the hypothesis $H_0 : \mu_1 = \mu_2$ that the two samples come from the same normal distribution (recall that our basic model assumes that the two variances are equal). The estimates of the parameters of the basic model were found on page 120. Under H_0 there are two unknown parameters, the common mean μ and the common variance σ^2, and since the situation under H_0 is nothing but a one-sample problem, we can easily write down the maximum likelihood estimates: the estimate of μ is the grand mean $\overline{y} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{2} \sum_{j=1}^{n_i} y_{ij}$, and the estimate of σ^2 is $\widehat{\widehat{\sigma}}^2 = \dfrac{1}{n} \displaystyle\sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} - \overline{y})^2$. The likelihood ratio test statistic

for H_0 is $Q = \dfrac{L(\overline{y}, \overline{y}, \widehat{\widehat{\sigma}}^2)}{L(\overline{y}_1, \overline{y}_2, \widehat{\sigma}^2)}$. Standard calculations show that

$$-2\ln Q = 2\Big(\ln L(\overline{y}_1, \overline{y}_2, \widehat{\sigma}^2) - \ln L(\overline{y}, \overline{y}, \widehat{\widehat{\sigma}}^2)\Big) = n\ln\Big(1 + \frac{t^2}{n-2}\Big)$$

where

$$t = \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{s_0^2\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)}}. \tag{8.3}$$

The hypothesis is rejected for large values of $-2\ln Q$, that is, for large values of $|t|$.

From the rules of calculus for variances, $\operatorname{Var}(\overline{Y}_1 - \overline{Y}_2) = \sigma^2\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)$, so the t statistic measures the difference between the two estimated means relative to the estimated standard deviation of the difference. Thus the likelihood ratio test ($-2\ln Q$) is equivalent to a test based on an immediately appealing test statistic (t).

The test probability is $\varepsilon = \mathrm{P}_0(|t| > |t_{\mathrm{obs}}|)$. The distribution of t under H_0 is known, as the next proposition shows.

Proposition 8.2
Consider independent normal random variables X_{ij}, $j = 1, 2, \ldots, n_i$, $i = 1, 2$, all with a common mean μ and a common variance σ^2. Let

$$\overline{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}, \quad i = 1, 2, \qquad \textit{and}$$

$$s_0^2 = \frac{1}{n-2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2$$

Then the random variable

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_0^2\,(\frac{1}{n_1} + \frac{1}{n_2})}}$$

follows a t distribution with $n-2$ degrees of freedom ($n = n_1 + n_2$).

Additional remarks:

- If the hypothesis H_0 is accepted, then the two samples are not really different, and they can be pooled together to form a single sample of size $n_1 + n_2$. Consequently the proper estimates must be calculated using the one-sample formulas

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i} y_{ij} \qquad \text{and} \qquad s_{01}^2 = \frac{1}{n-1}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2.$$

- Under H_0 the distribution of t does not depend on the two unknown parameters (σ^2 and the common μ).
- The t distribution is symmetric about 0, and therefore the test probability can be calculated as $P(|t| > |t_{obs}|) = 2\,P(t > |t_{obs}|)$.
  In certain situations it can be argued that the hypothesis should be rejected for large positive values of t only (i.e. when $\overline{y}_1$ is larger than $\overline{y}_2$); this could be the case if it is known in advance that the alternative to $\mu_1 = \mu_2$ is that $\mu_1 > \mu_2$. In such a situation the test probability is calculated as $\varepsilon = P(t > t_{obs})$.
  Similarly, if the alternative to $\mu_1 = \mu_2$ is that $\mu_1 < \mu_2$, the test probability is calculated as $\varepsilon = P(t < t_{obs})$.
- Quantiles of the t distribution are obtained from the computer or from a collection of statistical tables, possibly page 214.
  Since the distribution is symmetric about 0, quantiles are usually tabulated for probabilities larger that 0.5 only.
- See also Section 11.3 page 172.

Example 8.5: Vitamin C

◁ [Continued from Example 7.6 on page 121.]

The t test for equality of the two means requires variance homogeneity (i.e. that the groups have the same variance), so one might want to perform a test of variance homogeneity (see Exercise 8.2). This test would be based on the variance ratio

$$R = \frac{s^2_{\text{orange juice}}}{s^2_{\text{synthetic}}} = \frac{19.69}{7.66} = 2.57\,.$$

The R value is to be compared to the F distribution with $(9, 9)$ degrees of freedom in a two-sided test. Using the tables on page 210ff we see that in this distribution the 95% quantile is 3.18 and the 90% quantile is 2.44, so the test probability is somewhere between 10 and 20 percent, and we cannot reject the hypothesis of variance homogeneity.

Then we can safely carry on with the original problem, that of testing equality of the two means. The t test statistic is

$$t = \frac{13.18 - 8.00}{\sqrt{13.68\,(\frac{1}{10} + \frac{1}{10})}} = \frac{5.18}{1.65} = 3.13.$$

The t value is to be compared to the t distribution with 18 degrees of freedom; in this distribution the 99.5% quantile is 2.878, so the chance of getting a value of $|t|$ larger than 3.13 is less than 1%, and we may conclude that there is a highly significant difference between the means of the two groups.—Looking at the actual observation, we see that the growth in the "synthetic" group is less than in the "orange juice" group.

8.3 Exercises

Exercise 8.1
Let $x_1, x_2, \ldots, x_n$ be a sample from the normal distribution with mean ξ and variance σ^2. Then the hypothesis $H_{0x} : \xi = \xi_0$ can be tested using a t test as described on page 131. But we could also proceed as follows: Let a and $b \neq 0$ be two constants, and define

$$\begin{aligned} \mu &= a + b\xi, \\ \mu_0 &= a + b\xi_0, \\ y_i &= a + bx_i, \quad i = 1, 2, \ldots, n \end{aligned}$$

If the xs are a sample from the normal distribution with parameters ξ and σ^2, then the ys are a sample from the normal distribution with parameters μ and $b^2\sigma^2$ (Proposition 3.11 page 72), and the hypothesis $H_{0x} : \xi = \xi_0$ is equivalent to the hypothesis $H_{0y} : \mu = \mu_0$. Show that the t test statistic for testing H_{0x} is the same as the t test statistic for testing H_{0y} (that is to say, the t test is invariant under affine transformations).

Exercise 8.2
In the discussion of the two-sample problem with normal variables (page 108ff) it is assumed that the two groups have the same variance (there is "variance homogeneity"). It is perfectly possible to test that assumption. To do that, we extend the model slightly to the following: $y_{11}, y_{12}, \ldots, y_{1n_1}$ is a sample from the normal distribution with parameters μ_1 and σ_1^2, and $y_{21}, y_{22}, \ldots, y_{2n_2}$ is a sample form the normal distribution with parameters μ_2 and σ_2^2. In this model we can test the hypothesis $H : \sigma_1^2 = \sigma_2^2$.

Do that! Write down the likelihood function, find the maximum likelihood estimators, and find the likelihood ratio test statistic Q.

Show that Q depends on the observations only through the quantity $R = \dfrac{s_1^2}{s_2^2}$, where $s_i^2 = \dfrac{1}{n_i - 1} \displaystyle\sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2$ is the unbiased variance estimate in group i, $i = 1, 2$.

It can be proved that when the hypothesis is true, then the distribution of R is an F distribution (see page 171) with $(n_1 - 1, n_2 - 1)$ degrees of freedom.

9 Examples

9.1 Flour beetles

This example will show, for one thing, how it is possible to extend a binomial model in such a way that the probability parameter may depend on a number of covariates, using a so-called logistic regression model.

In a study of the reaction of insects to the insecticide Pyrethrum, a number of flour beetles, *Tribolium castaneum*, were exposed to this insecticide in various concentrations, and after a period of 13 days the number of dead beetles was recorded for each concentration (Pack and Morgan, 1990). The experiment was carried out using four different concentrations, and on male and female beetles separately. Table 9.1 shows (parts of) the results.

The basic model

The study comprises a total of 144+69+54+...+47 = 641 beetles. Each individual is classified by sex (two groups) and by dose (four groups), and the response is either death or survival.

An important part of the modelling process is to understand the status of the various quantities entering into the model:

- Dose and sex are covariates used to classify the beetles into $2 \times 4 = 8$ groups—the idea being that dose and sex have some influence on the probability of survival.
- The group totals $144, 69, 54, 50, 152, 81, 44, 47$ are known constants.
- The number of dead in each group $(43, 50, 47, 48, 26, 34, 27, 43)$ are observed values of random variables.

Initially, we make a few simple calculations and plots, such as Table 9.1 and Figure 9.1.

It seems reasonable to describe, for each group, the number of dead as an observation from a binomial distribution whose parameter n is the total number of beetles in the group, and whose unknown probability parameter has an interpretation as the probability that a beetle of the actual sex will die when given the insecticide in the actual dose. Although it could be of some interest to know whether there is a significant difference between the groups, it would be

Table 9.1
Flour beetles:
For each combination of dose and sex the table gives "the number of dead" / "group total" = "observed death frequency".
Dose is given as mg/cm^2.

dose	M	F
0.20	43/144 = 0.30	26/152 = 0.17
0.32	50/ 69 = 0.72	34/ 81 = 0.42
0.50	47/ 54 = 0.87	27/ 44 = 0.61
0.80	48/ 50 = 0.96	43/ 47 = 0.91

much more interesting if we could also give a more detailed description of the relation between the dose and death probability, and if we could tell whether the insecticide has the same effect on males and females. Let us introduce some notation in order to specify the *basic model*:

1. The group corresponding to dose d and sex k consists of n_{dk} beetles, of which y_{dk} dead; here $k \in \{M, F\}$ and $d \in \{0.20, 0.32, 0.50, 0.80\}$.
2. It is assumed that y_{dk} is an observation of a binomial random variable Y_{dk} with parameters n_{dk} (known) and $p_{dk} \in \,]0\,;1[$ (unknown).
3. Furthermore, it is assumed that all of the random variables Y_{dk} are independent.

The likelihood function corresponding to this model is

$$\begin{aligned} L &= \prod_k \prod_d \binom{n_{dk}}{y_{dk}} p_{dk}^{y_{dk}} (1-p_{dk})^{n_{dk}-y_{dk}} \\ &= \prod_k \prod_d \binom{n_{dk}}{y_{dk}} \cdot \prod_k \prod_d \Big(\frac{p_{dk}}{1-p_{dk}}\Big)^{y_{dk}} \cdot \prod_k \prod_d (1-p_{dk})^{n_{dk}} \\ &= \text{const} \cdot \prod_k \prod_d \Big(\frac{p_{dk}}{1-p_{dk}}\Big)^{y_{dk}} \cdot \prod_k \prod_d (1-p_{dk})^{n_{dk}}, \end{aligned}$$

and the log-likelihood function is

$$\begin{aligned} \ln L &= \text{const} + \sum_k \sum_d y_{dk} \ln \frac{p_{dk}}{1-p_{dk}} + \sum_k \sum_d n_{dk} \ln(1-p_{dk}) \\ &= \text{const} + \sum_k \sum_d y_{dk} \operatorname{logit}(p_{dk}) + \sum_k \sum_d n_{dk} \ln(1-p_{dk}). \end{aligned}$$

The eight parameters p_{dk} vary independently of each other, and $\widehat{p}_{dk} = y_{dk}/n_{dk}$. We are now going to model the depencence of p on d and k.

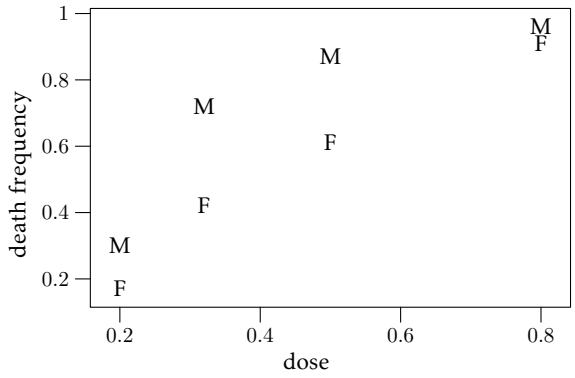


Figure 9.1
Flour beetles: Observed death frequency plotted against dose, for each sex.

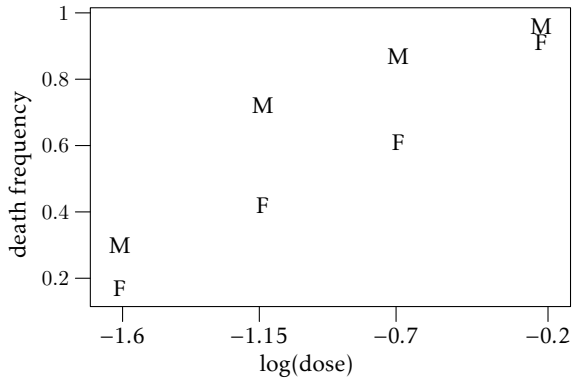


Figure 9.2
Flour beetles: Observed death frequency plotted against the logarithm of the dose, for each sex.

A dose-response model

How is the relation between the dose d and the probability p_d that a beetle will die when given the insecticide in dose d? If we knew almost everything about this very insecticide and its effects on the beetle organism, we might be able to deduce a mathematical formula for the relation between d and p_d. However, the designer of statistical models will approach the problem in a much more pragmatic and down-to-earth manner, as we shall see.

In the actual study the set of dose values seems to be rather odd, 0.20, 0.32, 0.50 and 0.80, but on closer inspection you will see that it is (almost) a geometric progression: the ratio between one value and the next is approximately constant, namely 1.6. The statistician will take this as an indication that dose should be measured on a logarithmic scale, so we should try to model the relation between $\ln(d)$ and p_d. Hence, we make another plot, Figure 9.2.

The simplest relationship between two numerical quantities is probably a linear (or affine) relation, but it is ususally a bad idea to suggest a linear relation between $\ln(d)$ and p_d (that is, $p_d = \alpha + \beta \ln d$ for suitably values of the constants α and β), since this would be incompatible with the requirement that probabilities

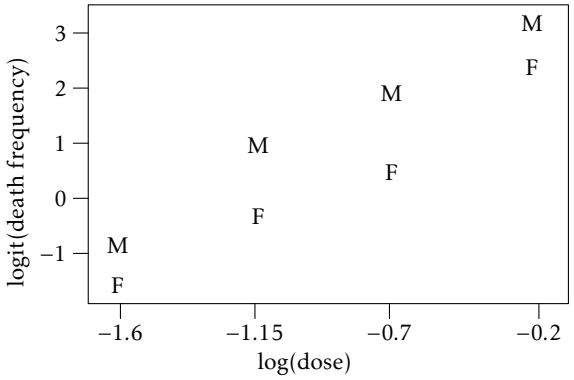


Figure 9.3
Flour beetles: Logit of observed death frequency plotted against logarithm of dose, for each sex.

must always be numbers between 0 and 1. Often, one would now convert the probabilities to a new scale and claim a linear relation between "probabilities on the new scale" and the logarithm of the dose. A frequently used conversion function is the logit function.

The logit function. The function

$$\operatorname{logit}(p) = \ln \frac{p}{1-p}.$$

is a bijection from $]0,1[$ to the real axis $\mathbb{R}$. Its inverse function is

$$p = \frac{\exp(z)}{1+\exp(z)}.$$

If p is the probability of a given event, then $p/(1-p)$ is the ratio between the probability of the event and the probability of the opposite event, the so-called *odds* of the event. Hence, the logit function calculates the logarithm of the odds.

We will therefore suggest/claim the following often used model for the relation between dose and probability of dying: For each sex, $\operatorname{logit}(p_d)$ is a linear (or more correctly: an affine) function of $x = \ln d$, that is, for suitable values of the constants α_M, β_M and α_F, β_F,

$$\operatorname{logit}(p_{dM}) = \alpha_M + \beta_M \ln d$$
$$\operatorname{logit}(p_{dF}) = \alpha_F + \beta_F \ln d.$$

This is a *logistic regression model*.

In Figure 9.3 the logit of the death frequencies is plotted against the logarithm of the dose; if the model is applicable, then each set of points should lie approximately on a straight line, and that seems to be the case, although a closer examination is needed in order to determine whether the model actually gives an adequate description of the data.

In the subsequent sections we shall see how to estimate the unknown parameters, how to check the adequacy of the model, and how to compare the effects of the poison on male and female beetles.

Estimation

The likelihood function L_0 in the logistic model is obtained from the basic likelihood function by considering the ps as functions of the αs and βs. This gives the log-likelihood function

$$\ln L_0(\alpha_M, \alpha_F, \beta_M, \beta_F)$$

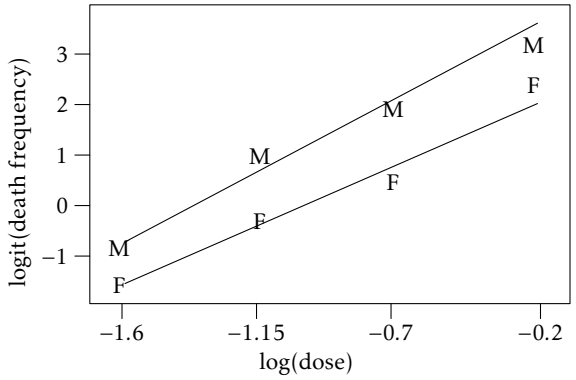


Figure 9.4
Flour beetles: Logit of observed death frequency plotted against logarithm of dose, for each sex, and the estimated regression lines.

$$= \text{const} + \sum_k \sum_d y_{dk}(\alpha_k + \beta_k \ln d) + \sum_k \sum_d n_{dk} \ln(1 - p_{dk})$$
$$= \text{const} + \sum_k \Big(\alpha_k y_{\bullet k} + \beta_k \sum_d y_{dk} \ln d + \sum_d n_{dk} \ln(1 - p_{dk}) \Big),$$

which consists of a contribution from the male beetles plus a contribution from the female beetles. The parameter space is $\mathbb{R}^4$.

To find the stationary point (which hopefully is a point of maximum) we equate each of the four partial derivatives of $\ln L$ to zero; this leads (after some calculations) to the four equations

$$\sum_d y_{dM} = \sum_d n_{dM} p_{dM}, \qquad \sum_d y_{dF} = \sum_d n_{dF} p_{dF},$$
$$\sum_d y_{dM} \ln d = \sum_d n_{dM} p_{dM} \ln d, \qquad \sum_d y_{dF} \ln d = \sum_d n_{dF} p_{dF} \ln d,$$

where $p_{dk} = \dfrac{\exp(\alpha_k + \beta_k \ln d)}{1 + \exp(\alpha_k + \beta_k \ln d)}$. It is not possible to write down an explicit solution to these equations, so we have to accept a numerical solution.

Standard statistical computer software will easily find the estimates in the present example: $\widehat{\alpha}_M = 4.27$ (with a standard deviation of 0.53) and $\widehat{\beta}_M = 3.14$ (standard deviation 0.39), and $\widehat{\alpha}_F = 2.56$ (standard deviation 0.38) and $\widehat{\beta}_F = 2.58$ (standard deviation 0.30).

We can add the estimated regression lines to the plot Figure 9.3; this results in Figure 9.4.

Model validation

We have estimated the parameters in the model where

$$\text{logit}(p_{dk}) = \alpha_k + \beta_k \ln d$$

or

$$p_{dk} = \frac{\exp(\alpha_k + \beta_k \ln d)}{1 + \exp(\alpha_k + \beta_k \ln d)}.$$

An obvious thing to do next is to plot the graphs of the two functions

$$x \mapsto \frac{\exp(\alpha_M + \beta_M x)}{1 + \exp(\alpha_M + \beta_M x)} \qquad \text{and} \qquad x \mapsto \frac{\exp(\alpha_F + \beta_F x)}{1 + \exp(\alpha_F + \beta_F x)}$$

into Figure 9.2, giving Figure 9.5. This shows that our model is not entirely erroneous. We can also perform a numerical test based on the likelihood ratio test statistic

$$Q = \frac{L(\widehat{\alpha}_M, \widehat{\alpha}_F, \widehat{\beta}_M, \widehat{\beta}_F)}{L_{\max}} \tag{9.1}$$

where $L_{\max}$ denotes the maximum value of the basic likelihood function (given on page 138). Using the notation $\widehat{p}_{dk} = \text{logit}^{-1}(\widehat{\alpha}_k + \widehat{\beta}_k \ln d)$ and $\widehat{y}_{dk} = n_{dk}\widehat{p}_{dk}$, we get

$$\begin{aligned} Q &= \frac{\prod_k \prod_d \binom{n_{dk}}{y_{dk}} \widehat{p}_{dk}^{\,y_{dk}} (1 - \widehat{p}_{dk})^{n_{dk} - y_{dk}}}{\prod_k \prod_d \binom{n_{dk}}{y_{dk}} \left(\frac{y_{dk}}{n_{dk}}\right)^{y_{dk}} \left(1 - \frac{y_{dk}}{n_{dk}}\right)^{n_{dk} - y_{dk}}} \\ &= \prod_k \prod_d \left(\frac{\widehat{y}_{dk}}{y_{dk}}\right)^{y_{dk}} \left(\frac{n_{dk} - \widehat{y}_{dk}}{n_{dk} - y_{dk}}\right)^{n_{dk} - y_{dk}} \end{aligned}$$

and

$$-2 \ln Q = 2 \sum_k \sum_d \left(y_{dk} \ln \frac{y_{dk}}{\widehat{y}_{dk}} + (n_{dk} - y_{dk}) \ln \frac{n_{dk} - y_{dk}}{n_{dk} - \widehat{y}_{dk}} \right).$$

Large values of $-2 \ln Q$ (or small values of Q) mean large discrepancy between the observed numbers (y_{kd} and $n_{kd} - y_{kd}$) and the predicted numbers ($\widehat{y}_{kd}$ and $n_{kd} - \widehat{y}_{kd}$), thus indicating that the model is inadequate. In the present example we get $-2 \ln Q_{\text{obs}} = 3.36$, and the corresponding test probability, i.e. the probability (when the model is correct) of getting values of $-2 \ln Q$ that are larger than $-2 \ln Q_{\text{obs}}$, can be found approximately using the fact that $-2 \ln Q$ is asymptotically χ^2 distributed with $8 - 4 = 4$ degrees of freedom; the (approximate) test probability is therefore about 0.50. (The number of degrees of freedom is the number of free parameters in the basic model minus the number of free parameters in the model tested.)

We can therefore conclude that the model seems to be adequate.

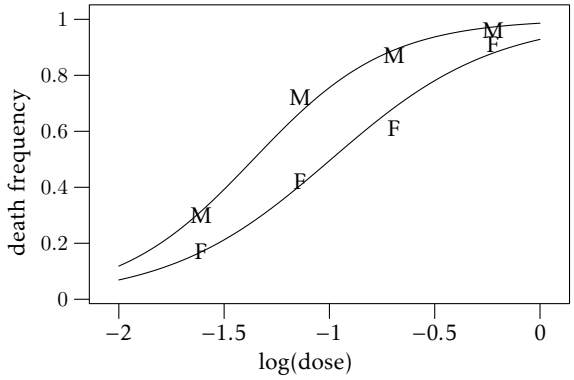


Figure 9.5
Flour beetles: Two different curves, and the observed death frequencies.

Hypotheses about the parameters

The current model with four parameters seems to be quite good, but maybe it can be simplified. We could for example test whether the two curves are parallel, and if they are parallel, we could test whether they are identical. Consider therefore these two statistical hypotheses:

1. The hypothesis of parallel curves: $H_1 : \beta_M = \beta_F$, or more detailed: for suitable values of the constants α_M, α_F and β

$$\operatorname{logit}(p_{dM}) = \alpha_M + \beta \ln d \qquad \text{and} \qquad \operatorname{logit}(p_{dF}) = \alpha_F + \beta \ln d$$

   for all d.
2. The hypothesis of identical curves: $H_2 : \alpha_M = \alpha_F$ and $\beta_M = \beta_F$ or more detailed: for suitable values of the constants α and β

$$\operatorname{logit}(p_{dM}) = \alpha + \beta \ln d \qquad \text{and} \qquad \operatorname{logit}(p_{dF}) = \alpha + \beta \ln d$$

   for all d.

The hypothesis H_1 of parallel curves is tested as the first one. The maximum likelihood estimates are $\widehat{\widehat{\alpha}}_M = 3.84$ (with a standard deviation of 0.34), $\widehat{\widehat{\alpha}}_F = 2.83$ (standard deviation 0.31) and $\widehat{\widehat{\beta}} = 2.81$ (standard deviation 0.24). We use the standard likelihood ratio test and compare the maximal likelihood under H_1 to the maximal likelihood under the current model:

$$\begin{aligned} -2\ln Q &= -2\ln \frac{L(\widehat{\widehat{\alpha}}_M, \widehat{\widehat{\alpha}}_F, \widehat{\widehat{\beta}}, \widehat{\widehat{\beta}})}{L(\widehat{\alpha}_M, \widehat{\alpha}_F, \widehat{\beta}_M, \widehat{\beta}_F)} \\ &= 2\sum_k \sum_d \left(y_{dk} \ln \frac{\widehat{y}_{dk}}{\widehat{\widehat{y}}_{dk}} + (n_{dk} - y_{dk}) \ln \frac{n_{dk} - \widehat{y}_{dk}}{n_{dk} - \widehat{\widehat{y}}_{dk}} \right) \end{aligned}$$

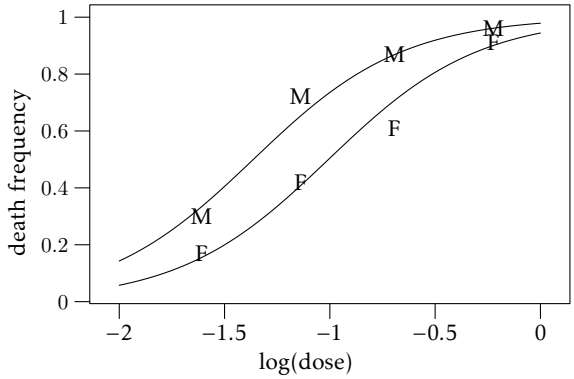


Figure 9.6
Flour beetles: The final model.

where $\widehat{\widehat{y}}_{dk} = n_{dk}\dfrac{\exp(\widehat{\widehat{\alpha}}_k + \widehat{\widehat{\beta}}\ln d)}{1+\exp(\widehat{\widehat{\alpha}}_k + \widehat{\widehat{\beta}}\ln d)}$. We find that $-2\ln Q_{\text{obs}} = 1.31$, to be compared to the χ^2 distribution with $4-3=1$ degrees of freedom (the change in the number of free parameters). The test probability (i.e. the probability of getting $-2\ln Q$ values larger than 1.31) is about 25%, so the value $-2\ln Q_{\text{obs}} = 1.31$ is not significantly large. Thus the model with parallel curves is not significantly inferior to the current model.

Having accepted the hypothesis H_1, we can then test the hypothesis H_2 of identical curves. (If H_1 had been rejected, we would not test H_2.) Testing H_2 against H_1 gives $-2\ln Q_{\text{obs}} = 27.50$, to be compared to the χ^2 distribution with $3-2=1$ degrees of freedom; the probability of values larger than 27.50 is extremely small, so the model with identical curves is very much inferior to the model with parallel curves. The hypothesis H_2 is rejected.

The conclusion is therefore that the relation between the dose d and the probability p of dying can be described using a model that makes $\text{logit}\,d$ be a linear function of $\ln d$, for each sex; the two curves are parallel but not identical. The estimated curves are

$$\text{logit}(p_{dM}) = 3.84 + 2.81\ln d \qquad \text{and} \qquad \text{logit}(p_{dF}) = 2.83 + 2.81\ln d\,,$$

corresponding to

$$p_{dM} = \frac{\exp(3.84+2.81\ln d)}{1+\exp(3.84+2.81\ln d)} \quad \text{and} \quad p_{dF} = \frac{\exp(2.83+2.81\ln d)}{1+\exp(2.83+2.81\ln d)}.$$

Figure 9.6 illustrates the situation.

9.2 Lung cancer in Fredericia

This is an example of a so-called multiplicative Poisson model. It also turns out to be an example of a modelling situation where small changes in the way the

age group	Fredericia	Horsens	Kolding	Vejle	total
40-54	11	13	4	5	33
55-59	11	6	8	7	32
60-64	11	15	7	10	43
65-69	10	10	11	14	45
70-74	11	12	9	8	40
75+	10	2	12	7	31
total	64	58	51	51	224

Table 9.2
Number of lung cancer cases in six age groups and four cities (from Andersen, 1977).

model is analysed lead to quite contradictory conclusions.

The situation

In the mid 1970s there was some debate about whether the citizens of the Danish city Fredericia had a higher risk of getting lung cancer than citizens in the three comparable neighbouring cities Horsens, Kolding and Vejle, since Fredericia, as opposed to the three other cities, had a considerable amount of polluting industries located in the mid-town. To shed light on the problem, data were collected about lung cancer incidence in the four cities in the years 1968 to 1971

Lung cancer is often the long-term result of a tiny daily impact of harmful substances. An increased lung cancer risk in Fredericia might therefore result in lung cancer patients being younger in Fredericia than in the three control cities; and in any case, lung cancer incidence is generally known to vary with age. Therefore we need to know, for each city, the number of cancer cases in different age groups, see Table 9.2, and we need to know the number of inhabitants in the same groups, see Table 9.3.

Now the statistician's job is to find a way to describe the data in Table 9.2 by means of a statistical model that gives an appropriate description of the risk of lung cancer for a person in a given age group and in a given city. Furthermore, it would certainly be expedient if we could separate out three or four parameters that could be said to describe a "city effect" (i.e. differences between cities) after having allowed for differences between age groups.

Specification of the model

We are not going to model the variation in the number of inhabitants in the various cities and age groups (Table 9.3), so these numbers are assumed to be known constants. The numbers in Table 9.2, the number of lung cancer cases, are the numbers that will be assumed to be observed values of random variables, and the statistical model should specify the joint distribution of these random

Table 9.3
Number of inhabitants in six age groups and four cities (from Andersen, 1977).

age group	Fredericia	Horsens	Kolding	Vejle	total
40-54	3059	2879	3142	2520	11600
55-59	800	1083	1050	878	3811
60-64	710	923	895	839	3367
65-69	581	834	702	631	2748
70-74	509	634	535	539	2217
75+	605	782	659	619	2665
total	6264	7135	6983	6026	26408

variables. Let us introduce some notation:

$$y_{ij} = \text{number of cases in age group } i \text{ in city } j,$$
$$r_{ij} = \text{number of persons in age group } i \text{ in city } j,$$

where $i = 1,2,3,4,5,6$ numbers the age groups, and $j = 1,2,3,4$ numbers the cities. The observations y_{ij} are considered to be observed values of random variables Y_{ij}.

Which distribution for the Ys should we use? At first one might suggest that Y_{ij} could be a binomial random variable with some small p and with $n = r_{ij}$; since the probability of getting lung cancer fortunately is rather small, another suggestion would be, with reference to Theorem 2.15 on page 59, that Y_{ij} follows a Poisson distribution with mean μ_{ij}. Let us write μ_{ij} as $\mu_{ij} = \lambda_{ij} r_{ij}$ (recall that r_{ij} is a known number), then the intensity λ_{ij} has an interpretation as "number of lung cancer cases per person in age group i in city j during the four year period", so λ is the *age- and city-specific cancer incidence*. Furthermore we will assume independence between the Y_{ij}s (lung cancer is believed to be non-contagious). Hence our basic model is:

> The Y_{ij}s are independent Poisson random variables with $\mathrm{E}\, Y_{ij} = \lambda_{ij} r_{ij}$ where the λ_{ij}s are unknown positive real numbers.

The parameters of the basic model are estimated in the straightforward manner as $\widehat{\lambda}_{ij} = y_{ij}/r_{ij}$, so that for example the estimate of the intensity λ_{21} for 55-59–year-old inhabitants of Fredericia is $11/800 = 0.014$ (there are 0.014 cases per person per four years).

Now, the idea was that we should be able to compare the cities after having allowed for differences between age groups, but the basic model does not in any obvious way allow such comparisons. We will therefore specify and test a hypothesis that we can do with a slightly simpler model (i.e. a model with fewer parameters) that does allow such comparisons, namely a model in which λ_{ij} is

written as a product $\alpha_i\beta_j$ of an *age effect* α_i and a *city effect* β_j. If so, we are in a good position, because then we can compare the cities simply by comparing the city parameters β_j. Therefore our first statistical hypothesis will be

$$H_0 : \lambda_{ij} = \alpha_i\beta_j$$

where $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \beta_1, \beta_2, \beta_3, \beta_4$ are unknown parameters. More specifically, the hypothesis claims that there exist values of the 10 unknown parameters $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \beta_1, \beta_2, \beta_3, \beta_4 \in \mathbb{R}_+$ such that lung cancer incidence λ_{ij} in city j and age group i is $\lambda_{ij} = \alpha_i\beta_j$. The model specified by H_0 is a *multiplicative* model since city parameters and age parameters enter multiplicatively.

A detail pertaining to the parametrisation

It is a special feature of the parametrisation under H_0 that it is not injective: the mapping

$$(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \beta_1, \beta_2, \beta_3, \beta_4) \mapsto (\alpha_i\,\beta_j)_{i=1,\ldots,6;\ j=1,\ldots,4}$$

from $\mathbb{R}_+^{10}$ to $\mathbb{R}^{24}$ is not injective; in fact the range of the map is a 9-dimensional surface. (It is easily seen that $\alpha_i\,\beta_j = \alpha_i^*\,\beta_j^*$ for all i and j, if and only if there exists a $c > 0$ such that $\alpha_i = c\,\alpha_i^*$ for all i and $\beta_j^* = c\,\beta_j$ for all j.)

Therefore we must impose one linear constraint on the parameters in order to obtain an injective parametrisation; examples of such a constraint are: $\alpha_1 = 1$, or $\alpha_1 + \alpha_2 + \cdots + \alpha_6 = 1$, or $\alpha_1\alpha_2\ldots\alpha_6 = 1$, or a similar constraint on the βs, etc. In the actual example we will chose the constraint $\beta_1 = 1$, that is, we fix the Fredericia parameter to the value 1. Henceforth the model H_0 has only nine independent parameters

Estimation in the multiplicative model

In the multiplicative model you cannot write down explicit expressions for the estimates, but most statistical software will, after a few instructions, calculate the actual values for any given set of data. The likelihood function of the basic model is

$$L = \prod_{i=1}^{6}\prod_{j=1}^{4} \frac{(\lambda_{ij}r_{ij})^{y_{ij}}}{y_{ij}!}\exp(-\lambda_{ij}r_{ij}) = \text{const}\cdot\prod_{i=1}^{6}\prod_{j=1}^{4}\lambda_{ij}^{y_{ij}}\exp(-\lambda_{ij}r_{ij})$$

as a function of the λ_{ij}s. As noted earlier, this function attains its maximum when $\widehat{\lambda}_{ij} = y_{ij}/r_{ij}$ for all i and j.

If we substitute $\alpha_i\beta_j$ for λ_{ij} in the expression for L, we obtain the likelihood function L_0 under H_0:

$$\begin{aligned} L_0 &= \text{const} \cdot \prod_{i=1}^{6}\prod_{j=1}^{4} \alpha_i^{y_{ij}} \beta_j^{y_{ij}} \exp(-\alpha_i\beta_j r_{ij}) \\ &= \text{const} \cdot \Big(\prod_{i=1}^{6} \alpha_i^{y_{i\cdot}}\Big)\Big(\prod_{j=1}^{4} \beta_j^{y_{\cdot j}}\Big)\exp\Big(-\sum_{i=1}^{6}\sum_{j=1}^{4} \alpha_i\beta_j r_{ij}\Big), \end{aligned}$$

and the log-likelihood function is

$$\ln L_0 = \text{const} + \Big(\sum_{i=1}^{6} y_{i\cdot} \ln \alpha_i\Big) + \Big(\sum_{j=1}^{4} y_{\cdot j} \ln \beta_j\Big) - \sum_{i=1}^{6}\sum_{j=1}^{4} \alpha_i\beta_j r_{ij}.$$

We are seeking the set of parameter values $(\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\alpha}_3, \widehat{\alpha}_4, \widehat{\alpha}_5, \widehat{\alpha}_6, 1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4)$ that maximises $\ln L_0$. The partial derivatives are

$$\begin{aligned} \frac{\partial \ln L_0}{\partial \alpha_i} &= \frac{y_{i\cdot}}{\alpha_i} - \sum_j \beta_j r_{ij}, \quad i = 1,2,3,4,5,6 \\ \frac{\partial \ln L_0}{\partial \beta_j} &= \frac{y_{\cdot j}}{\beta_j} - \sum_i \alpha_i r_{ij}, \quad j = 1,2,3,4, \end{aligned}$$

and these partial derivatives are equal to zero if and only if

$$y_{i\cdot} = \sum_j \alpha_i\beta_j r_{ij} \quad \text{for all } i \qquad \text{and} \qquad y_{\cdot j} = \sum_i \alpha_i\beta_j r_{ij} \quad \text{for all } j.$$

We note for later use that this implies that

$$y_{\cdot\cdot} = \sum_i \sum_j \widehat{\alpha}_i\widehat{\beta}_j r_{ij}. \tag{9.2}$$

The maximum likelihood estimates in the multiplicative model are found to be

$$\begin{aligned} \widehat{\alpha}_1 &= 0.0036 & \widehat{\beta}_1 &= 1 \\ \widehat{\alpha}_2 &= 0.0108 & \widehat{\beta}_2 &= 0.719 \\ \widehat{\alpha}_3 &= 0.0164 & \widehat{\beta}_3 &= 0.690 \\ \widehat{\alpha}_4 &= 0.0210 & \widehat{\beta}_4 &= 0.762 \\ \widehat{\alpha}_5 &= 0.0229 & & \\ \widehat{\alpha}_6 &= 0.0148. & & \end{aligned}$$

age group	Fredericia	Horsens	Kolding	Vejle
40-54	3.6	2.6	2.5	2.7
55-59	10.8	7.8	7.5	8.2
60-64	16.4	11.8	11.3	12.5
65-69	21.0	15.1	14.5	16.0
70-74	22.9	16.5	15.8	17.4
75+	14.8	10.6	10.2	11.3

Table 9.4
Estimated age- and city-specific lung cancer intensities in the years 1968-71, assuming a multiplicative Poisson model. The values are number of cases per 1000 inhabitants per 4 years.

How does the multiplicative model describe the data?

We are going to investigate how well the multiplicative model describes the actual data. This can be done by testing the multiplicative hypothesis H_0 against the basic model, using the standard log likelihood ratio test statistic

$$-2\ln Q = -2\ln \frac{L_0(\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\alpha}_3, \widehat{\alpha}_4, \widehat{\alpha}_5, \widehat{\alpha}_6, 1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4)}{L(\widehat{\lambda}_{11}, \widehat{\lambda}_{12}, \ldots, \widehat{\lambda}_{63}, \widehat{\lambda}_{64})}.$$

Large values of $-2\ln Q$ are significant, i.e. indicate that H_0 does not provide an adequate description of data. To tell if $-2\ln Q_{\text{obs}}$ is significant, we can use the test probability $\varepsilon = \mathrm{P}_0\big(-2\ln Q \geq -2\ln Q_{\text{obs}}\big)$, that is, the probability of getting a larger value of $-2\ln Q$ than the one actually observed, provided that the true model is H_0. When H_0 is true, $-2\ln Q$ is approximately a χ^2-variable with $f = 24 - 9 = 15$ degrees of freedom (provided that all the expected values are 5 or above), so that the test probability can be found approximately as $\varepsilon = \mathrm{P}\big(\chi^2_{15} \geq -2\ln Q_{\text{obs}}\big)$.

Standard calculations (including an application of (9.2)) lead to

$$Q = \prod_{i=1}^{6}\prod_{j=1}^{4}\left(\frac{\widehat{y}_{ij}}{y_{ij}}\right)^{y_{ij}},$$

and thus

$$-2\ln Q = 2\sum_{i=1}^{6}\sum_{j=1}^{4} y_{ij}\ln\frac{y_{ij}}{\widehat{y}_{ij}}.$$

where $\widehat{y}_{ij} = \widehat{\alpha}_i\widehat{\beta}_j r_{ij}$ is the expected number of cases in age group i in city j.

The calculated values $1000\,\widehat{\alpha}_i\widehat{\beta}_j$, the expected number of cases per 100 inhabitants, can be found in Table 9.4, and the actual value of $-2\ln Q$ is $-2\ln Q_{\text{obs}} = 22.6$. In the χ^2 distribution with $f = 24 - 9 = 15$ degrees of freedom the 90% quantile is 22.3 and the 95% quantile is 25.0; the value $-2\ln Q_{\text{obs}} = 22.6$ therefore corresponds to a test probability above 5%, so there seems to be no serious evidence against the multiplicative model. Hence we will assume multiplicativity, that is, the lung cancer risk depends multiplicatively on age group and city.

Table 9.5
Expected number of lung cancer cases $\widehat{y_{ij}}$ under the multiplicative Poisson model.

age group	Fredericia	Horsens	Kolding	Vejle	total
40-54	11.01	7.45	7.80	6.91	33.17
55-59	8.64	8.41	7.82	7.23	32.10
60-64	11.64	10.88	10.13	10.48	43.13
65-69	12.20	12.59	10.17	10.10	45.06
70-74	11.66	10.44	8.45	9.41	39.96
75+	8.95	8.32	6.73	6.98	30.98
total	64.10	58.09	51.10	51.11	224.40

We have arrived at a statistical model that describes the data using a number of city-parameters (the βs) and a number of age-parameters (the αs), but with no interaction between cities and age groups. This means that any difference between the cities is the same for all age groups, and any difference between the age groups is the same in all cities. In particular, a comparison of the cities can be based entirely on the βs.

Identical cities?

The purpose of all this is to see whether there is a significant difference between the cities. If there is no difference, then the city parameters are equal, $\beta_1 = \beta_2 = \beta_3 = \beta_4$, and since $\beta_1 = 1$, the common value is necessarily 1. So we shall test the statistical hypothesis

$$H_1 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 1.$$

The hypothesis has to be tested against the current basic model H_0, leading to the test statistic

$$-2\ln Q = -2\ln \frac{L_1(\widehat{\widehat{\alpha}}_1, \widehat{\widehat{\alpha}}_2, \widehat{\widehat{\alpha}}_3, \widehat{\widehat{\alpha}}_4, \widehat{\widehat{\alpha}}_5, \widehat{\widehat{\alpha}}_6)}{L_0(\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\alpha}_3, \widehat{\alpha}_4, \widehat{\alpha}_5, \widehat{\alpha}_6, 1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4)}$$

where $L_1(\alpha_1, \alpha_2, \ldots, \alpha_6) = L_0(\alpha_1, \alpha_2, \ldots, \alpha_6, 1, 1, 1, 1)$ is the likelihood function under H_1, and $\widehat{\widehat{\alpha}}_1, \widehat{\widehat{\alpha}}_2, \ldots, \widehat{\widehat{\alpha}}_6$ are the estimates under H_1, that is, $(\widehat{\widehat{\alpha}}_1, \widehat{\widehat{\alpha}}_2, \ldots, \widehat{\widehat{\alpha}}_6)$ must be the point of maximum of L_1.

The function L_1 is a product of six functions, each a function of a single α:

$$\begin{aligned} L_1(\alpha_1, \alpha_2, \ldots, \alpha_6) &= \text{const} \cdot \prod_{i=1}^{6} \prod_{j=1}^{4} \alpha_i^{y_{ij}} \exp(-\alpha_i r_{ij}) \\ &= \text{const} \cdot \prod_{i=1}^{6} \alpha_i^{y_{i\cdot}} \exp(-\alpha_i r_{i\cdot}) \,. \end{aligned}$$

age group	Fredericia	Horsens	Kolding	Vejle	total
40-54	8.70	8.19	8.94	7.17	33.00
55-59	6.72	9.10	8.82	7.38	32.02
60-64	9.09	11.81	11.46	10.74	43.10
65-69	9.53	13.68	11.51	10.35	45.07
70-74	9.16	11.41	9.63	9.70	39.90
75+	7.02	9.07	7.64	7.18	30.91
total	50.22	63.26	58.00	52.52	224.00

Table 9.6
Expected number of lung cancer cases $\widehat{\widehat{y}}_{ij}$ under the hypothesis H_1 of no difference between cities.

Therefore, the maximum likelihood estimates are $\widehat{\widehat{\alpha}}_i = \dfrac{y_{i\bullet}}{r_{i\bullet}}$, $i = 1, 2, \ldots, 6$, and the actual values are

$$\begin{aligned} \widehat{\widehat{\alpha}}_1 &= 33/11600 = 0.002845 & \widehat{\widehat{\alpha}}_4 &= 45/2748 = 0.0164 \\ \widehat{\widehat{\alpha}}_2 &= 32/3811 \;\; = 0.00840 & \widehat{\widehat{\alpha}}_5 &= 40/2217 = 0.0180 \\ \widehat{\widehat{\alpha}}_3 &= 43/3367 \;\; = 0.0128 & \widehat{\widehat{\alpha}}_6 &= 31/2665 = 0.0116. \end{aligned}$$

Standard calculations show that

$$Q = \frac{L_1(\widehat{\widehat{\alpha}}_1, \widehat{\widehat{\alpha}}_2, \widehat{\widehat{\alpha}}_3, \widehat{\widehat{\alpha}}_4, \widehat{\widehat{\alpha}}_5, \widehat{\widehat{\alpha}}_6)}{L_0(\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\alpha}_3, \widehat{\alpha}_4, \widehat{\alpha}_5, \widehat{\alpha}_6, 1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4)} = \prod_{i=1}^{6} \prod_{j=1}^{4} \left(\frac{\widehat{\widehat{y}}_{ij}}{\widehat{y}_{ij}} \right)^{y_{ij}}$$

and hence

$$-2 \ln Q = 2 \sum_{i=1}^{6} \sum_{j=1}^{4} y_{ij} \ln \frac{\widehat{y}_{ij}}{\widehat{\widehat{y}}_{ij}} \; ;$$

here $\widehat{\widehat{y}}_{ij} = \widehat{\widehat{\alpha}}_i r_{ij}$, and $\widehat{y}_{ij} = \widehat{\alpha}_i \widehat{\beta}_j r_{ij}$ as earlier. The expected numbers $\widehat{\widehat{y}}_{ij}$ are shown in Table 9.6.

As always, large values of $-2 \ln Q$ are significant. Under H_1 the distribution of $-2 \ln Q$ is approximately a χ^2 distribution with $f = 9 - 6 = 3$ degrees of freedom.

We get $-2 \ln Q_{\text{obs}} = 5.67$. In the χ^2 distribution with $f = 9 - 6 = 3$ degrees of freedom, the 80% quantile is 4.64 and the 90% quantile is 6.25, so the test probability is almost 20%, and therefore there seems to be a good agreement between the data and the hypothesis H_1 of no difference between cities. Put in another way, *there is no significant difference between the cities.*

Another approach

Only in rare instances will there be a definite course of action to be followed in a statistical analysis of a given problem. In the present case, we have investigated the question of an increased lung cancer risk in Fredericia by testing the

hypothesis H_1 of identical city parameters. It turned out that we could accept H_1, that is, there seems to be no significant difference between the cities.

But we could chose to investigate the problem in a different way. One might argue that since the question is whether Fredericia has a higher risk than the other cities, then it seems to be implied that the three other cities are more or less identical—an assumption which can be tested. We should therefore adopt this strategy for testing hypotheses:

1. The basic model is still the multiplicative Poisson model H_0.
2. Initially we test whether the three cities Horsens, Kolding and Vejle are identical, i.e. we test the hypothesis
$$H_2 : \beta_2 = \beta_3 = \beta_4.$$
3. If H_2 is accepted, it means that there is a common level β for the three "control cities". We can then compare Fredericia to this common level by testing whether $\beta_1 = \beta$. Since β_1 per definition equals 1, the hypothesis to test is
$$H_3 : \beta = 1.$$

Comparing the three control cities

We are going to test the hypothesis $H_2 : \beta_2 = \beta_3 = \beta_4$ of identical control cities relative to the multiplicative model H_0. This is done using the test statistic

$$Q = \frac{L_2(\widetilde{\alpha}_1, \widetilde{\alpha}_2, \ldots, \widetilde{\alpha}_6, \widetilde{\beta})}{L_0(\widehat{\alpha}_1, \widehat{\alpha}_2, \ldots, \widehat{\alpha}_6, 1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4)}$$

where $L_2(\alpha_1, \alpha_2, \ldots, \alpha_6, \beta) = L_0(\alpha_1, \alpha_2, \ldots, \alpha_6, 1, \beta, \beta, \beta)$ is the likelihood function under H_2, and $\widetilde{\alpha}_1, \widetilde{\alpha}_2, \ldots, \widetilde{\alpha}_6, \widetilde{\beta}$ are the maximum likelihood estimates under H_2.

The distribution of $-2\ln Q$ under H_2 can be approximated by a χ^2 distribution with $f = 9 - 7 = 2$ degrees of freedom.

The model H_2 is a multiplicative Poisson model similar to the previous one, now with two cities (Fredericia and the others) and six age groups, and therefore estimation in this model does not raise any new problems. We find that

$\widetilde{\alpha}_1 = 0.00358$	$\widetilde{\alpha}_4 = 0.0210$	$\widetilde{\beta}_1 = 1$
$\widetilde{\alpha}_2 = 0.0108$	$\widetilde{\alpha}_5 = 0.0230$	$\widetilde{\beta} = 0.7220$
$\widetilde{\alpha}_3 = 0.0164$	$\widetilde{\alpha}_6 = 0.0148$	

Furthermore,

$$-2\ln Q = 2\sum_{i=1}^{6}\sum_{j=1}^{4} y_{ij} \ln \frac{\widehat{y}_{ij}}{\widetilde{y}_{ij}}$$

age group	Fredericia	Horsens	Kolding	Vejle	total
40-54	10.95	7.44	8.12	6.51	33.02
55-59	8.64	8.44	8.19	6.85	32.12
60-64	11.64	10.93	10.60	9.93	43.10
65-69	12.20	12.65	10.64	9.57	45.06
70-74	11.71	10.53	8.88	8.95	40.07
75+	8.95	8.36	7.04	6.61	30.96
total	64.09	58.35	53.47	48.42	224.33

Table 9.7
Expected number of lung cancer cases $\widetilde{y}_{ij}$ *under* H_2.

where $\widehat{y}_{ij} = \widehat{\alpha}_i \widehat{\beta}_j r_{ij}$, see Table 9.5, and

$$\begin{aligned} \widetilde{y}_{i1} &= \widetilde{\alpha}_i \, r_{i1} \\ \widetilde{y}_{ij} &= \widetilde{\alpha}_i \widetilde{\beta} \, r_{ij}, \quad j = 2, 3, 4. \end{aligned}$$

The expected numbers $\widetilde{y}_{ij}$ are shown in Table 9.7. We find that $-2\ln Q_{\text{obs}} = 0.40$, to be compared to the χ^2 distribution with $f = 9 - 7 = 2$ degrees of freedom. In this distribution the 20% quantile is 0.446, which means that the test probability is well above 80%, so H_2 is perfectly compatible with the available data, and we can easily assume that there is no significant difference between the three cities.

Then we can test H_3, the model with six age groups with parameters $\alpha_1, \alpha_2, \ldots, \alpha_6$ and with identical cities. Assuming H_2, H_3 is identical to the previous hypothesis H_1, and hence the estimates of the age parameters are $\widehat{\widehat{\alpha}}_1, \widehat{\widehat{\alpha}}_2, \ldots, \widehat{\widehat{\alpha}}_6$ on page 151.

This time we have to test H_3 (= H_1) relative to the current basic model H_2. The test statistic is $-2\ln Q$ where

$$Q = \frac{L_1(\widehat{\widehat{\alpha}}_1, \widehat{\widehat{\alpha}}_2, \ldots, \widehat{\widehat{\alpha}}_6)}{L_2(\widetilde{\alpha}_1, \widetilde{\alpha}_2, \ldots, \widetilde{\alpha}_6, \widetilde{\beta})} = \frac{L_0(\widehat{\widehat{\alpha}}_1, \widehat{\widehat{\alpha}}_2, \ldots, \widehat{\widehat{\alpha}}_6, 1, 1, 1, 1)}{L_0(\widetilde{\alpha}_1, \widetilde{\alpha}_2, \ldots, \widetilde{\alpha}_6, 1, \widetilde{\beta}, \widetilde{\beta}, \widetilde{\beta})}$$

which can be rewritten as

$$Q = \prod_{i=1}^{6} \prod_{j=1}^{4} \left(\frac{\widehat{\widehat{y}}_{ij}}{\widetilde{y}_{ij}} \right)^{y_{ij}}$$

so that

$$-2\ln Q = 2 \sum_{i=1}^{6} \sum_{j=1}^{4} y_{ij} \ln \frac{\widetilde{y}_{ij}}{\widehat{\widehat{y}}_{ij}}.$$

Large values of $-2\ln Q$ are significant. The distribution of $-2\ln Q$ under H_3 can be approximated by a χ^2 distribution with $f = 7 - 6 = 1$ degree of freedom (provided that all the expected numbers are at least 5).

Overview of approach no. 1.

Model/Hypotesis		$-2\ln Q$	f	ε
M: arbitrary parameters H: multiplicativity		22.65	$24-9=15$	a good 5%
M: multiplicativity H: four identical cities		5.67	$9-6=3$	about 20%

Overview of approach no. 2.

Model/Hypotesis		$-2\ln Q$	f	ε
M: arbitrary parameters H: multiplicativity		22.65	$24-9=15$	a good 5%
M: multiplicativity H: identical control cities		0.40	$9-7=2$	a good 80%
M: identical control cities H: four identical cities		5.27	$7-6=1$	about 2%

Using values from Tables 9.2, 9.6 and 9.7 we find that $-2\ln Q_{\text{obs}} = 5.27$. In the χ^2 distribution with 1 degree of freedom the 97.5% quantile is 5.02 and the 99% quantile is 6.63, so the test probability is about 2%. On this basis the usual practice is to reject the hypothesis H_3 $(= H_1)$. So the conclusion is that *as regards lung cancer risk there is no significant difference between the three cities Horsens, Kolding and Vejle, whereas Fredericia has a lung cancer risk that is significantly different from those three cities*.

The lung cancer incidence of the three control cities relative to that of Fredericia is estimated as $\widetilde{\beta} = 0.7$, so our conclusion can be sharpened: the lung cancer incidence of Fredericia is significantly larger than that of the three cities.

Now we have reached a nice and clear conclusion, which unfortunately is quite contrary to the previous conclusion on page 151!

Comparing the two approaches

We have been using two just slightly different approaches, which nevertheless lead to quite opposite conclusions. Both approaches follow this scheme:

1. Decide on a suitable basic model.
2. State a simplifying hypothesis.
3. Test the hypothesis relative to the current basic model.
4. a) If the hypothesis can be accepted, then we have a new basic model, namely the previous basic model with the simplifications implied by the hypothesis. Continue from step 2.
   b) If the hypothesis is rejected, then keep the current basic model and exit.

The two approaches take their starting point in the same Poisson model, they differ only in the choice of hypotheses (in item 2 above). The first approach goes from the multiplicative model to "four identical cities" in a single step, giving a test statistic of 5.67, and with 3 degrees of freedom this is not significant. The second approach uses two steps, "multiplicativity → three identical" and "three identical → four identical", and it turns out that the 5.67 with 3 degrees of freedom is divided into 0.40 with 2 degrees of freedom and 5.27 with 1 degrees of freedom, and the latter is highly significant.

Sometimes it may be desirable to do such a stepwise testing, but you should never just blindly state and test as many hypotheses as possible. Instead you should consider only hypotheses that are reasonable within the actual context.

About test statistics

You may have noted some common traits of the $-2\ln Q$ statistics occuring in sections 9.1 and 9.2: They all seem to be of the form

$$-2\ln Q = 2\sum \text{obs.number} \cdot \ln \frac{\text{expected number in current basic model}}{\text{expected number under hypothesis}}$$

and they all have an approximate χ^2 distribution with a number of degrees of freedom which is found as "number of free parameters in current basic model" minus "number of free parameters under the hypothesis". This is in fact a general result about testing hypotheses in Poisson models (only such hypotheses where the sum of the expected numbers equals the sum of the observed numbers).

9.3 Accidents in a weapon factory

This example seems to be a typical case for a Poisson model, but it turns out that it does not give an adequate fit. So we have to find another model.

The situation

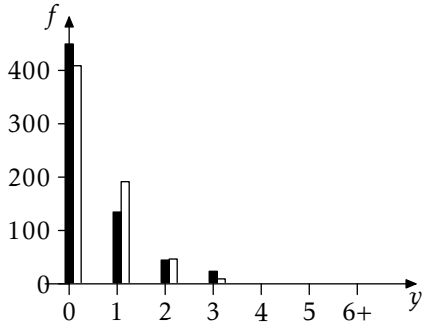
The data is the number of accidents in a period of five weeks for each woman working on 6-inch HE shells (high-explosive shells) at a weapon factory in England during World War I. Table 9.10 gives the distribution of $n = 647$ women by number of accidents over the period in question. We are looking for a statistical model that can describe this data set. (The example originates from Greenwood and Yule (1920) and is known through its occurence in Hald (1948, 1968) (English version: Hald (1952)), which for decades was a leading statistics textbook (at least in Denmark).)

Table 9.10
Distribution of 647 women by number y of accidents over a period of five weeks (from Greenwood and Yule (1920)).

y	number of women having y accidents
0	447
1	132
2	42
3	21
4	3
5	2
6+	0
	647

Table 9.11
Model 1: Table and graph showing the observed numbers f_y (black bars) and the expected numbers $\widehat{f_y}$ (white bars).

y	f_y	$\widehat{f_y}$
0	447	406.3
1	132	189.0
2	42	44.0
3	21	6.8
4	3	0.8
5	2	0.1
6+	0	0.0
	647	647.0



Let y_i be the number of accidents that come upon woman no. i, and let f_y denote the number of women with exactly y accidents, so that in the present case, $f_0 = 447$, $f_1 = 132$, etc. The total number of accidents then is $0f_0 + 1f + 2f_2 + \ldots = \sum_{y=0}^{\infty} y f_y = 301$.

We will assume that y_i is an observed value of a random variable Y_i, $i = 1, 2, \ldots, n$, and we will assume the random variables $Y_1, Y_2, \ldots, Y_n$ to be independent, even if this may be a questionable assumption.

The first model

The first model to try is the model stating that $Y_1, Y_2, \ldots, Y_n$ are independent and identically Poisson distributed with a common parameter μ. This model is plausible if we assume the accidents to happen "entirely at random", and the parameter μ then describes the women's "accident proneness".

The Poisson parameter μ is estimated as $\widehat{\mu} = \overline{y} = 301/647 = 0.465$ (there has been 0.465 accidents per woman per five weeks). The expected number of

women hit by y accidents is $\widehat{f_y} = n\dfrac{\widehat{\mu}^y}{y!}\exp(-\widehat{\mu})$; the actual values are shown in Table 9.11. The agreement between the observed and the expected numbers is not very impressive. Further evidence of the inadequacy of the model is obtained by calculating the central variance estimate $s^2 = \frac{1}{n-1}\sum(y_i - \overline{y})^2 = 0.692$ which is almost 50% larger than $\overline{y}$—and in the Poisson distribution the mean and the variance are equal. Therefore we should probably consider another model.

The second model

Model 1 can be extended in the following way:

- We still assume $Y_1, Y_2, \ldots, Y_n$ to be independent Poisson variables, but this time each Y_i is allowed to have its own mean value, so that Y_i now is a Poisson random variable with parameter μ_i, for all $i = 1, 2, \ldots, n$.
  If the modelling process came to a halt at this stage, there would be one parameter per person, allowing a perfect fit (with $\widehat{\mu_i} = y_i$, $i = 1, 2, \ldots, n$), but that would certainly contradict Fisher's maxim, that "the object of statistical methods is the reduction of data" (cf. page 98). But there is a further step in modelling process:
- We will further assume that $\mu_1, \mu_2, \ldots, \mu_n$ are independent observations from a single probability distribution, a continuous distribution on the positive real numbers. It turns out that in this context a gamma distribution is a convenient choice, so assume that the μs come from a gamma distribution with shape parameter κ and scale parameter β, that is, with density function

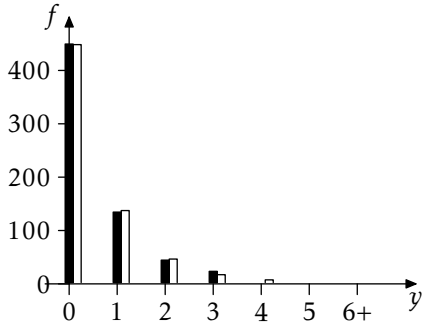
$$g(\mu) = \frac{1}{\Gamma(\kappa)\beta^\kappa}\,\mu^{\kappa-1}\exp(-\mu/\beta), \qquad \mu > 0.$$

- The conditional probability, for a given value of μ, that a woman is hit by exactly y accidents, is $\dfrac{\mu^y}{y!}\exp(-\mu)$, and the unconditional probability is then obtained by mixing the conditional probabilities with respect to the distribution of μ:

$$\begin{aligned}
\mathrm{P}(Y = y) &= \int_0^{+\infty} \frac{\mu^y}{y!}\exp(-\mu)\cdot g(\mu)\,d\mu \\
&= \int_0^{+\infty} \frac{\mu^y}{y!}\exp(-\mu)\,\frac{1}{\Gamma(\kappa)\beta^\kappa}\,\mu^{\kappa-1}\exp(-\mu/\beta)\,d\mu \\
&= \frac{\Gamma(y+\kappa)}{y!\,\Gamma(\kappa)}\left(\frac{1}{\beta+1}\right)^\kappa\left(\frac{\beta}{\beta+1}\right)^y,
\end{aligned}$$

y	f_y	$\widehat{\widehat{f}}_y$
0	447	445.9
1	132	134.9
2	42	44.0
3	21	14.7
4	3	5.0
5	2	1.7
6+	0	0.9
	647	647.1

Table 9.12
Model 2: Table and graph of the observed numbers f_y (black bars) and expected numbers $\widehat{\widehat{f}}_y$ (white bars).



where the last equality follows from the definition of the gamma function (use for example the formula at the end of the side note on page 70). If κ is a natural number, $\Gamma(\kappa) = (\kappa-1)!$, and then

$$\binom{y+\kappa-1}{y} = \frac{\Gamma(y+\kappa)}{y!\,\Gamma(\kappa)}.$$

In this equation, the right-hand side is actually defined for all $\kappa > 0$, so we can see the equation as a way of defining the symbol on the left-hand side for all $\kappa > 0$ and all $y \in \{0, 1, 2, \ldots\}$. If we let $p = 1/(\beta+1)$, the probability of y accidents can now be written as

$$\mathrm{P}(Y = y) = \binom{y+\kappa-1}{y} p^{\kappa}(1-p)^{y}, \qquad y = 0, 1, 2, \ldots \tag{9.3}$$

It is seen that Y has a *negative binomial distribution* with shape parameter κ and probability parameter $p = 1/(\beta+1)$ (cf. Definition 2.9 on page 56).

The negative binomial distribution has two parameters, so one would expect Model 2 to give a better fit than the one-parameter model Model 1.

In the new model we have (cf. Example 4.5 on page 79)

$$\mathrm{E}(Y) = \kappa(1-p)/p = \kappa\beta,$$
$$\mathrm{Var}(Y) = \kappa(1-p)/p^2 = \kappa\beta(\beta+1).$$

It is seen that the variance is $(\beta+1)$ times larger than the mean. In the present data set, the variance is actually larger than the mean, so for the present Model 2 cannot be rejected.

Estimation of the parameters in Model 2

As always we will be using the maximum likelihood estimator of the unknown parameters. The likelihood function is a product of probabilities of the form

(9.3):

$$\begin{aligned} L(\kappa, p) &= \prod_{i=1}^{n} \binom{y_i + \kappa - 1}{y_i} p^{\kappa} (1-p)^{y_i} \\ &= p^{n\kappa} (1-p)^{y_1 + y_2 + \cdots + y_n} \prod_{i=1}^{n} \binom{y_i + \kappa - 1}{y_i} \\ &= \text{const} \cdot p^{n\kappa} (1-p)^{y_\cdot} \cdot \prod_{k=1}^{\infty} (\kappa + k - 1)^{f_k + f_{k+1} + f_{k+2} + \ldots} \end{aligned}$$

where f_k still denotes the number of observations equal to k. The logarithm of the likelihood function is then (except for a constant term)

$$\ln L(\kappa, p) = n\kappa \ln p + y_\cdot \ln(1-p) + \sum_{k=1}^{\infty} \Big(\sum_{j=k}^{\infty} f_j \Big) \ln(\kappa + k - 1),$$

and more innocuous-looking with the actual data entered:

$$\begin{aligned} \ln L(\kappa, p) = {} & 647\kappa \ln p + 301 \ln(1-p) \\ & + 200 \ln \kappa + 68 \ln(\kappa + 1) + 26 \ln(\kappa + 2) \\ & + 5 \ln(\kappa + 3) + 2 \ln(\kappa + 4). \end{aligned}$$

We have to find the point(s) of maximum for this function. It is not clear whether an analytic solution can be found, but it is fairly simple to find a numerical solution, for example by using a simplex method (which does not use the derivatives of the function to be maximised). The simplex method is an iterative method and requires an initial value of (κ, p), say $(\widetilde{\kappa}, \widetilde{p})$. One way to determine such an initial value is to solve the equations

$$\begin{aligned} \text{theoretical mean} &= \text{empirical mean} \\ \text{theoretical variance} &= \text{empirical variance}. \end{aligned}$$

In the present case these equations become $\kappa\beta = 0.465$ and $\kappa\beta(\beta + 1) = 0.692$, where $\beta = (1-p)/p$. The solution to these two equations is $(\widetilde{\kappa}, \widetilde{p}) = (0.953, 0.672)$ (and hence $\widetilde{\beta} = 0.488$). These values can then be used as starting values in an iterative procedure that will lead to the point of maximum of the likelihood function, which is found to be

$$\begin{aligned} &\widehat{\kappa} = 0.8651 \\ &\widehat{p} = 0.6503 \quad \text{(and hence } \widehat{\beta} = 0.5378\text{)}. \end{aligned}$$

Table 9.12 shows the corresponding expected numbers

$$\widehat{\widehat{f}}_y = n\binom{y+\widehat{\kappa}-1}{y}\widehat{p}^{\widehat{\kappa}}(1-\widehat{p})^y$$

calculated from the estimated negative binomial distribution. There seems to be a very good agreement between the observed and the expected numbers, and we may therefore conclude that our Model 2 gives an adequate description of the observations.

10 The Multivariate Normal Distribution

Linear normal models can, as we shall see in Chapter 11, be expressed in a very clear and elegant way using the language of linear algebra, and there seems to be a considerable advantage in discussing estimation and hypothesis testing in linear models in a linear algebra setting.

Initially, however, we have to clarify a few concepts relating to multivariate random variables (Section 10.1), and to define the multivariate normal distribution, which turns out to be rather complicated (Section 10.2). The reader may want to take a look at the short summary of notation and results from linear algebra (page 201ff).

10.1 Multivariate random variables

An n-dimensional random variable $\boldsymbol{X}$ can be regarded as a set of n one-dimensional random variables, or as a single random vector (in the vector space $V = \mathbb{R}^n$) whose coordinates relative to the standard coordinate system are n one-dimensional random variables. The *mean value* of the n-dimensional random variabel $\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$ is the vector $\mathrm{E}\,\boldsymbol{X} = \begin{bmatrix} \mathrm{E}\,X_1 \\ \mathrm{E}\,X_2 \\ \vdots \\ \mathrm{E}\,X_n \end{bmatrix}$, that is, the set of mean values of the coordinates of $\boldsymbol{X}$—here we assume that all the one-dimensional random variables have a mean value. The *variance* of $\boldsymbol{X}$ is the symmetric and positive semidefinite $n \times n$ matrix $\mathrm{Var}\,\boldsymbol{X}$ whose (i,j)th entry is $\mathrm{Cov}(X_i, X_j)$:

$$\begin{aligned} \mathrm{Var}\,\boldsymbol{X} &= \begin{bmatrix} \mathrm{Var}\,X_1 & \mathrm{Cov}(X_1,X_2) & \cdots & \mathrm{Cov}(X_1,X_n) \\ \mathrm{Cov}(X_2,X_1) & \mathrm{Var}\,X_2 & \cdots & \mathrm{Cov}(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n,X_1) & \mathrm{Cov}(X_n,X_2) & \cdots & \mathrm{Var}\,X_n \end{bmatrix} \\ &= \mathrm{E}\Big((\boldsymbol{X} - \mathrm{E}\,\boldsymbol{X})(\boldsymbol{X} - \mathrm{E}\,\boldsymbol{X})'\Big), \end{aligned}$$

if all the one-dimensional random variables have a variance.

Using these definitions the following proposition is easily shown:

Proposition 10.1
Let $\boldsymbol{X}$ be an n-dimensional random variable that has a mean and a variance. If A is a linear map from $\mathbb{R}^n$ to $\mathbb{R}^p$ and $\boldsymbol{b}$ a constant vector in $\mathbb{R}^p$ [or A is a $p \times n$ matrix and $\boldsymbol{b}$ constant $p \times 1$ matrix], then

$$\mathrm{E}(A\boldsymbol{X} + \boldsymbol{b}) = A(\mathrm{E}\,\boldsymbol{X}) + \boldsymbol{b} \tag{10.1}$$

$$\mathrm{Var}(A\boldsymbol{X} + \boldsymbol{b}) = A(\mathrm{Var}\,\boldsymbol{X})A'. \tag{10.2}$$

Example 10.1: The multinomial distribution
The multinomial distribution (page 103) is an example of a multivariate distribution. If $\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_r \end{bmatrix}$ is multinomial with parameters n and $\boldsymbol{p} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_r \end{bmatrix}$, then $\mathrm{E}\,\boldsymbol{X} = \begin{bmatrix} np_1 \\ np_2 \\ \vdots \\ np_r \end{bmatrix}$ and

$$\mathrm{Var}\,\boldsymbol{X} = \begin{bmatrix} np_1(1-p_1) & -np_1p_2 & \cdots & -np_1p_r \\ -np_2p_1 & np_2(1-p_2) & \cdots & -np_2p_r \\ \vdots & \vdots & \ddots & \vdots \\ -np_rp_1 & -np_rp_2 & \cdots & np_r(1-p_r) \end{bmatrix}.$$

The Xs always add up to n, that is, $A\boldsymbol{X} = n$ where A is the linear map

$$A : \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_r \end{bmatrix} \longmapsto \begin{bmatrix} 1 & 1 & \ldots & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_r \end{bmatrix} = x_1 + x_2 + \cdots + x_r.$$

Therefore $\mathrm{Var}(A\boldsymbol{X}) = 0$. Since $\mathrm{Var}(A\boldsymbol{X}) = A(\mathrm{Var}\,\boldsymbol{X})A'$ (equation (10.2)), the variance matrix $\mathrm{Var}\,\boldsymbol{X}$ is positive semidefinite (as always), but not positive definite.

10.2 Definition and properties

In this section we shall define the multivariate normal distribution and prove the multivariate version of Proposition 3.11 on page 72, namely that if $\boldsymbol{X}$ is n-dimensional normal with parameters $\boldsymbol{\mu}$ and Σ, and if A is a $p \times n$ matrix and $\boldsymbol{b}$ a $p \times 1$ matrix, then $A\boldsymbol{X} + \boldsymbol{b}$ is p-dimensional normal with parameters $A\boldsymbol{\mu} + \boldsymbol{b}$ and $A\Sigma A'$. Unfortunately, it is not so easy to define the n-dimensional normal distribution with parameters $\boldsymbol{\mu}$ and Σ when Σ is singular. We proceed step by step:

Definition 10.1: n-dimensional standard normal distribution
The n-dimensional standard normal distribution is the n-dimensional continuous

distribution with density function

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}} \exp\Big(-\tfrac{1}{2}\|\boldsymbol{x}\|^2\Big), \quad \boldsymbol{x} \in \mathbb{R}^n. \tag{10.3}$$

Remarks:

1. For $n = 1$ this is our earlier definition of standard normal distribution (Definition 3.9 on page 72).
2. The n-dimensional random variable $\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$ is n-dimensional standard normal if and only if $X_1, X_2, \ldots, X_n$ are independent and one-dimensional standard normal. This follows from the identity
$$\frac{1}{(2\pi)^{n/2}} \exp\Big(-\tfrac{1}{2}\|\boldsymbol{x}\|^2\Big) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}x_i^2)$$
(cf. Proposition 3.2 on page 65).
3. If $\boldsymbol{X}$ is standard normal, then $\mathrm{E}\,\boldsymbol{X} = \boldsymbol{0}$ and $\operatorname{Var}\boldsymbol{X} = \mathrm{I}$; this follows from item 2. (Here I is the identity map of $\mathbb{R}^n$ onto itself [or the $n \times n$ unit matrix].)

Proposition 10.2
If A is an isometric linear mapping of $\mathbb{R}^n$ onto itself, and $\boldsymbol{X}$ is n-dimensional standard normal, then $A\boldsymbol{X}$ is also n-dimensional standard normal.

Proof
From Theorem 3.5 (on page 66) about transformation of densities, the density function of $\boldsymbol{Y} = A\boldsymbol{X}$ is $f(A^{-1}\boldsymbol{y})|\det A^{-1}|$ where f is given by (10.3). Since f depends on $\boldsymbol{x}$ only though $\|\boldsymbol{x}\|$, and since $\|A^{-1}\boldsymbol{y}\| = \|\boldsymbol{y}\|$ because A is an isometry, we have that $f(A^{-1}\boldsymbol{y}) = f(\boldsymbol{y})$; and also because A is an isometry, $\det A^{-1} = 1$. Hence $f(A^{-1}\boldsymbol{y})|\det A^{-1}| = f(\boldsymbol{y})$. □

Corollary 10.3
If $\boldsymbol{X}$ is n-dimensional standard normal and $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ an orthonormal basis for $\mathbb{R}^n$, then the coordinates $X_1, X_2, \ldots, X_n$ of $\boldsymbol{X}$ in this basis are independent and one-dimensional standard normal.

Recall that the coordinates of $\boldsymbol{X}$ in the basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ can be expressed as $X_i = \langle \boldsymbol{X}, \boldsymbol{e}_i \rangle$, $i = 1, 2, \ldots, n$,

Proof
Since the coordinate transformation matrix is an isometry, this follows from Proposition 10.2 and remark 2 to Definition 10.1. □

Definition 10.2: Regular normal distribution
Suppose that $\boldsymbol{\mu} \in \mathbb{R}^n$ and Σ is a positive definite linear map of $\mathbb{R}^n$ onto itself [*or that $\boldsymbol{\mu}$ is a n-dimensional column vector and Σ a positive definite $n \times n$ matrix*].

The n-dimensional regular normal distribution with mean $\boldsymbol{\mu}$ and variance Σ is the n-dimensional continuous distribution with density function

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\det \Sigma|^{1/2}} \exp\Big(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big), \quad \boldsymbol{x} \in \mathbb{R}^n.$$

Remarks:

1. For $n = 1$ we get the ususal (one-dimensional) normal distribution with mean μ and variance σ^2 (= the single element of Σ).
2. If $\Sigma = \sigma^2 \mathrm{I}$, the density function becomes

$$f(\boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\tfrac{1}{2}\frac{\|\boldsymbol{x}-\boldsymbol{\mu}\|^2}{\sigma^2}\right), \quad \boldsymbol{x} \in \mathbb{R}^n.$$

   In other words, $\boldsymbol{X}$ is n-dimensional regular normal with parameters $\boldsymbol{\mu}$ and $\sigma^2\mathrm{I}$, if and only if $X_1, X_2, \ldots, X_n$ are independent and one-dimensional normal and the parameters of X_i are μ_i and σ^2.
3. The definition refers to the parameters $\boldsymbol{\mu}$ and Σ simply as mean and variance, but actually we must prove that $\boldsymbol{\mu}$ and Σ are in fact the mean and the variance of the distribution:
   In the case $\boldsymbol{\mu} = \mathbf{0}$ and $\Sigma = \mathrm{I}$ this follows from remark 3 to Definition 10.1. In the general case consider $\boldsymbol{X} = \boldsymbol{\mu} + \Sigma^{1/2}\boldsymbol{U}$ where $\boldsymbol{U}$ is n-dimensional standard normal and $\Sigma^{1/2}$ is as in Proposition B.5 on page 203 (with $A = \Sigma$). Then using Proposition 10.4 below, $\boldsymbol{X}$ is regular normal with parameters $\boldsymbol{\mu}$ and Σ, and Proposition 10.1 now gives $\mathrm{E}\,\boldsymbol{X} = \boldsymbol{\mu}$ and $\operatorname{Var}\boldsymbol{X} = \Sigma$.

Proposition 10.4
If $\boldsymbol{X}$ has an n-dimensional regular normal distribution with parameters $\boldsymbol{\mu}$ and Σ, if A is a bijective linear mapping of $\mathbb{R}^n$ onto itself, and if $\boldsymbol{b} \in \mathbb{R}^n$ is a constant vector, then $\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{b}$ has an n-dimensional regular normal distribution with paramters $A\boldsymbol{\mu} + \boldsymbol{b}$ and $A\Sigma A'$.

Proof
From Theorem 3.5 on page 66 the density function of $\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{b}$ is $f_{\boldsymbol{Y}}(\boldsymbol{y}) = f(A^{-1}(\boldsymbol{y}-\boldsymbol{b}))|\det A^{-1}|$ where f is as in Definition 10.2. Standard calculations give

$$\begin{aligned} &f_{\boldsymbol{Y}}(\boldsymbol{y}) \\ &= \frac{1}{(2\pi)^{n/2}\,|\det\Sigma|^{1/2}\,|\det A|} \exp\Big(-\tfrac{1}{2}\Big(A^{-1}(\boldsymbol{y}-\boldsymbol{b})-\boldsymbol{\mu}\Big)'\Sigma^{-1}\Big(A^{-1}(\boldsymbol{y}-\boldsymbol{b})-\boldsymbol{\mu}\Big)\Big) \end{aligned}$$

$$= \frac{1}{(2\pi)^{n/2}\,|\det(A\Sigma A')|^{1/2}} \exp\Big(-\tfrac{1}{2}\big(\boldsymbol{y}-(A\boldsymbol{\mu}+\boldsymbol{b})\big)'(A\Sigma A')^{-1}\big(\boldsymbol{y}-(A\boldsymbol{\mu}+\boldsymbol{b})\big)\Big)$$

as requested. □

Example 10.2

Suppose that $\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ has a two-dimensional regular normal distribution with parameters $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ and $\Sigma = \sigma^2 \mathrm{I} = \sigma^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Then let $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} X_1 + X_2 \\ X_1 - X_2 \end{bmatrix}$, that is, $\boldsymbol{Y} = A\boldsymbol{X}$ where $A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$.

Proposition 10.4 shows that $\boldsymbol{Y}$ has a two-dimensional regular normal distribution with mean $\mathrm{E}\,\boldsymbol{Y} = \begin{bmatrix} \mu_1 + \mu_2 \\ \mu_1 - \mu_2 \end{bmatrix}$ and variance $\sigma^2 AA' = 2\sigma^2 \mathrm{I}$, so $X_1 + X_2$ and $X_1 - X_2$ are independent one-dimensional normal variables with means $\mu_1 + \mu_2$ and $\mu_1 - \mu_2$, respectively, and whith identical variances.

Now we can define the general n-dimensional normal distribution:

Definition 10.3: The n-dimensional normal distribution

Let $\boldsymbol{\mu} \in \mathbb{R}^n$ and let Σ be a positive semidefinit linear map of $\mathbb{R}^n$ into itself [*or $\boldsymbol{\mu}$ is an n-dimensional column vector and Σ a positive semidefinit $n \times n$ matrix*]. *Let p denote the rank of Σ.*

The n-dimensional normal distribution with mean $\boldsymbol{\mu}$ and variance Σ is the distribution of $\boldsymbol{\mu} + B\boldsymbol{U}$ where $\boldsymbol{U}$ is p-dimensional standard normal and B an injective linear map of $\mathbb{R}^p$ into $\mathbb{R}^n$ such that $BB' = \Sigma$.

Remarks:

1. It is known from Proposition B.6 on page 203 that there exists a B with the required properties, and that B is unique up to isometries of $\mathbb{R}^p$. Since the standard normal distribution is invariant under isometries (Proposition 10.2), it follows that the distribution of $\boldsymbol{\mu} + B\boldsymbol{U}$ does not depend on the choice of B, as long as B is injective and $BB' = \Sigma$.
2. It follows from Proposition 10.4 that Definition 10.3 generalises Definition 10.2.

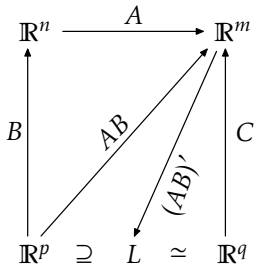
Proposition 10.5

If $\boldsymbol{X}$ has an n-dimensional normal distribution with parameters $\boldsymbol{\mu}$ and Σ, and if A is a linear map of $\mathbb{R}^n$ into $\mathbb{R}^m$, then $\boldsymbol{Y} = A\boldsymbol{X}$ has an m-dimensional normal distribution with parameters $A\boldsymbol{\mu}$ and $A\Sigma A'$.

Proof

We may assume that $\boldsymbol{X}$ is of the form $\boldsymbol{X} = \boldsymbol{\mu} + B\boldsymbol{U}$ where $\boldsymbol{U}$ is p-dimensional

standard normal and B is an injective linear map of $\mathbb{R}^p$ into $\mathbb{R}^n$. Then $\boldsymbol{Y} = A\boldsymbol{\mu} + AB\boldsymbol{U}$. We have to show that $AB\boldsymbol{U}$ has the same distribution as $C\boldsymbol{V}$, where C is an appropriate injective linear map $\mathbb{R}^q$ into $\mathbb{R}^m$, and $\boldsymbol{V}$ is q-dimensional standard normal; here q is the rank of AB.



Let L be the range of $(AB)'$, $L = \mathcal{R}((AB)')$; then $q = \dim L$. Now the idea is that we as C can use the restriction of AB to $L \simeq \mathbb{R}^q$. According to Proposition B.1 (page 202), L is the orthogonal complement of $\mathcal{N}(AB)$, the null space of AB, and from this follows that the restriction of AB to L is injective. We can therefore choose an orthonormal basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p$ for $\mathbb{R}^p$ such that $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_q$ is a basis for L, and the remaining $\boldsymbol{e}$s is a basis for $\mathcal{N}(AB)$. Let $U_1, U_2, \ldots, U_p$ be the coordinates of $\boldsymbol{U}$ in this basis, that is, $U_i = \langle \boldsymbol{U}, \boldsymbol{e}_i \rangle$. Since $\boldsymbol{e}_{q+1}, \boldsymbol{e}_{q+2}, \ldots, \boldsymbol{e}_p \in \mathcal{N}(AB)$,

$$AB\boldsymbol{U} = AB \sum_{i=1}^{p} U_i \boldsymbol{e}_i = \sum_{i=1}^{q} U_i AB\boldsymbol{e}_i.$$

According to Corollary 10.3 $U_1, U_2, \ldots, U_p$ are independent one-dimensional standard normal variables, and therefore the same is true of $U_1, U_2, \ldots, U_q$, so if we define $\boldsymbol{V}$ as the q-dimensional random variable whose entries are $U_1, U_2, \ldots, U_q$, then $\boldsymbol{V}$ is q-dimensional standard normal. Finally define a linear map C of $\mathbb{R}^q$ into $\mathbb{R}^n$ as the map that maps $\boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_q \end{bmatrix}$ to $\sum_{i=1}^{q} v_i AB\boldsymbol{e}_i$. Then $\boldsymbol{V}$ and C have the desired properties. □

Theorem 10.6: Partitioning the sum of squares
Let $\boldsymbol{X}$ have an n-dimensional normal distribution with mean $\boldsymbol{0}$ and variance $\sigma^2\mathrm{I}$. If $\mathbb{R}^n = L_1 \oplus L_2 \oplus \cdots \oplus L_k$, where the linear subspaces $L_1, L_2, \ldots, L_k$ are orthogonal, and if $\mathfrak{p}_1, \mathfrak{p}_2, \ldots, \mathfrak{p}_k$ are the projections onto $L_1, L_2, \ldots, L_k$, then the random variables $\mathfrak{p}_1\boldsymbol{X}, \mathfrak{p}_2\boldsymbol{X}, \ldots, \mathfrak{p}_k\boldsymbol{X}$ are independent; $\mathfrak{p}_j\boldsymbol{X}$ has an n-dimensional normal distribution with mean $\boldsymbol{0}$ and variance $\sigma^2\mathfrak{p}_j$, and $\|\mathfrak{p}_j\boldsymbol{X}\|^2$ has a χ^2 distribution with scale parameter σ^2 and $f_j = \dim L_j$ degrees of freedom.

Proof
It suffices to consider the case $\sigma^2 = 1$.

We can choose an orthonormal basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ such that each basis vector belongs to one of the subspaces L_i. The random variables $X_i = \langle \boldsymbol{X}, \boldsymbol{e}_i \rangle$, $i = 1, 2, \ldots, n$, are independent one-dimensional standard normal according to Corollary 10.3. The projection $\mathfrak{p}_j\boldsymbol{X}$ of $\boldsymbol{X}$ onto L_j is the sum of the f_j terms $X_i\boldsymbol{e}_i$ for which $\boldsymbol{e}_i \in L_j$. Therefore the distribution of $\mathfrak{p}_j\boldsymbol{X}$ is as stated in the Theorem, cf. Definition 10.3.

The individual $\mathrm{p}_j\boldsymbol{X}$s are caclulated from non-overlapping sets of X_is, and therefore they are independent. Finally, $\|\mathrm{p}_j\boldsymbol{X}\|^2$ is the sum of the f_i X_i^2s for which $\boldsymbol{e}_i \in L_j$, and therefore it has a χ^2 distribution with f_i degrees of freedom, cf. Proposition 3.13 (page 73). □

10.3 Exercises

Exercise 10.1
On page 161 it is claimed that $\operatorname{Var}\boldsymbol{X}$ is positive semidefinite. Show that this is indeed the case.—Hint: Use the rules of calculus for covariances (Proposition 1.30 page 39; that proposition is actually true for all kinds of real random variables with variance) to show that $\operatorname{Var}(\boldsymbol{a}'\boldsymbol{X}\boldsymbol{a}) = \boldsymbol{a}'(\operatorname{Var}\boldsymbol{X})\boldsymbol{a}$ where $\boldsymbol{a}$ is an $n \times 1$ matrix (i.e. a column vector).

Exercise 10.2
Fill in the details in the reasoning in remark 2 to the definition of the n-dimensional standard normal distribution.

Exercise 10.3
Let $\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ have a two-dimensional normal distribution. Show that X_1 and X_2 are independent if and only if their covariance is 0. (Compare this to Proposition 1.31 (page 39) which is true for all kinds of random variables with variance.)

Exercise 10.4
Let t be the function (from $\mathbb{R}$ to $\mathbb{R}$) given by $t(x) = -x$ for $|x| < a$ and $t(x) = x$ otherwise; here a is a positive constant. Let X_1 be a one-dimensional standard normal variable, and define $X_2 = t(X_1)$.

Are X_1 and X_2 independent? Show that a can be chosen in such a way that their covariance is zero. Explain why this result is consistent with Exercise 10.3. (Exercise 1.21 dealt with a similar problem.)

11 Linear Normal Models

This chapter presents the classical linear normal models, such as analysis of variance and linear regression, in a linear algebra formulation.

11.1 Estimation and test in the general linear model

In the general linear normal model we consider an observation $\boldsymbol{y}$ of an n-dimensional normal random variable $\boldsymbol{Y}$ with mean vector $\boldsymbol{\mu}$ and variance matrix $\sigma^2\mathrm{I}$; the parameter $\boldsymbol{\mu}$ is assumed to be a point in the linear subspace L of $V = \mathbb{R}^n$, and $\sigma^2 > 0$; hence the parameter space is $L \times]0\,;+\infty[$. The model is a *linear* normal model because the mean $\boldsymbol{\mu}$ is an element of a linear subspace L.

The likelihood function corresponding to the observation $\boldsymbol{y}$ is

$$L(\boldsymbol{\mu},\sigma^2) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \exp\left(-\tfrac{1}{2}\,\frac{\|\boldsymbol{y}-\boldsymbol{\mu}\|^2}{\sigma^2}\right), \qquad \boldsymbol{\mu} \in L,\ \ \sigma^2 > 0.$$

Estimation

Let $\mathfrak{p}$ be the orthogonal projection of $V = \mathbb{R}^n$ onto L. Then for any $\boldsymbol{z} \in V$ we have that $\boldsymbol{z} = (\boldsymbol{z}-\mathfrak{p}\boldsymbol{z})+\mathfrak{p}\boldsymbol{z}$ where $\boldsymbol{z}-\mathfrak{p}\boldsymbol{z} \perp \mathfrak{p}\boldsymbol{z}$, and so $\|\boldsymbol{z}\|^2 = \|\boldsymbol{z}-\mathfrak{p}\boldsymbol{z}\|^2+\|\mathfrak{p}\boldsymbol{z}\|^2$ (Pythagoras); applied to $\boldsymbol{z} = \boldsymbol{y}-\boldsymbol{\mu}$ this gives

$$\|\boldsymbol{y}-\boldsymbol{\mu}\|^2 = \|\boldsymbol{y}-\mathfrak{p}\boldsymbol{y}\|^2 + \|\mathfrak{p}\boldsymbol{y}-\boldsymbol{\mu}\|^2,$$

from which it follows that $L(\boldsymbol{\mu},\sigma^2) \le L(\mathfrak{p}\boldsymbol{y},\sigma^2)$ for each σ^2, so the maximum likelihood estimator of $\boldsymbol{\mu}$ is $\mathfrak{p}\boldsymbol{y}$. We can then use standard methods to see that $L(\mathfrak{p}\boldsymbol{y},\sigma^2)$ is maximised with respect to σ^2 when σ^2 equals $\frac{1}{n}\|\boldsymbol{y}-\mathfrak{p}\boldsymbol{y}\|^2$, and standard calculations show that the maximum value $L(\mathfrak{p}\boldsymbol{y},\frac{1}{n}\|\boldsymbol{y}-\mathfrak{p}\boldsymbol{y}\|^2)$ can be expressed as $a_n\,(\|\boldsymbol{y}-\mathfrak{p}\boldsymbol{y}\|^2)^{-n/2}$ where a_n is a number that depends on n, but not on $\boldsymbol{y}$.

We can now apply Theorem 10.6 to $\boldsymbol{X} = \boldsymbol{Y}-\boldsymbol{\mu}$ and the orthogonal decomposition $\mathbb{R} = L \oplus L^\perp$ to get the following

Theorem 11.1
In the general linear model,

- *the mean vector $\boldsymbol{\mu}$ is estimated as $\widehat{\boldsymbol{\mu}} = \mathfrak{p}\boldsymbol{y}$, the orthogonal projection of $\boldsymbol{y}$ onto L,*

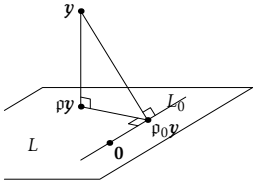
- *the estimator* $\boldsymbol{\mu} = \mathfrak{p}\boldsymbol{Y}$ *is n-dimensional normal with mean* $\boldsymbol{\mu}$ *and variance* $\sigma^2\mathfrak{p}$,
- *an unbiased estimator of the variance* σ^2 *is* $s^2 = \frac{1}{n-\dim L}\|\boldsymbol{y} - \mathfrak{p}\boldsymbol{y}\|^2$, *and the maximum likelihood estimator of* σ^2 *is* $\widehat{\sigma}^2 = \frac{1}{n}\|\boldsymbol{y} - \mathfrak{p}\boldsymbol{y}\|^2$,
- *the distribution of the estimator* s^2 *is a* χ^2 *distribution with scale parameter* $\sigma^2/(n - \dim L)$ *and* $n - \dim L$ *degrees of freedom,*
- *the two estimators* $\widehat{\boldsymbol{\mu}}$ *and* s^2 *are independent (and so are the two estimators* $\widehat{\boldsymbol{\mu}}$ *and* $\widehat{\sigma}^2$*).*

The vector $\boldsymbol{y} - \mathfrak{p}\boldsymbol{y}$ is the *residual vector*, and $\|\boldsymbol{y} - \mathfrak{p}\boldsymbol{y}\|^2$ is the *residual sum of squares*. The quantity $(n - \dim L)$ is the *number of degrees of freedom* of the variance estimator and/or the residual sum of squares.

You can find $\widehat{\boldsymbol{\mu}}$ from the relation $\boldsymbol{y} - \widehat{\boldsymbol{\mu}} \perp L$, which in fact is nothing but a set of $\dim L$ linear equations with as many unknowns; this set of equations is known as the *normal equations* (they are expressing the fact that $\boldsymbol{y} - \widehat{\boldsymbol{\mu}}$ is a normal of L).

Testing hypotheses about the mean

Suppose that we want to test a hypothesis of the form $H_0 : \boldsymbol{\mu} \in L_0$ where L_0 is a linear subspace of L. According to Theorem 11.1 the maximum likelihood estimators of $\boldsymbol{\mu}$ and σ^2 under H_0 are $\mathfrak{p}_0\boldsymbol{y}$ and $\frac{1}{n}\|\boldsymbol{y} - \mathfrak{p}_0\boldsymbol{y}\|^2$. Therefore the likelihood ratio test statistic is



$$Q = \frac{L(\mathfrak{p}_0\boldsymbol{y}, \frac{1}{n}\|\boldsymbol{y} - \mathfrak{p}_0\boldsymbol{y}\|^2)}{L(\mathfrak{p}\boldsymbol{y}, \frac{1}{n}\|\boldsymbol{y} - \mathfrak{p}\boldsymbol{y}\|^2)} = \left(\frac{\|\boldsymbol{y} - \mathfrak{p}\boldsymbol{y}\|^2}{\|\boldsymbol{y} - \mathfrak{p}_0\boldsymbol{y}\|^2}\right)^{n/2}.$$

Since $L_0 \subseteq L$, we have $\mathfrak{p}\boldsymbol{y} - \mathfrak{p}_0\boldsymbol{y} \in L$ and therefore $\boldsymbol{y} - \mathfrak{p}\boldsymbol{y} \perp \mathfrak{p}\boldsymbol{y} - \mathfrak{p}_0\boldsymbol{y}$ so that $\|\boldsymbol{y} - \mathfrak{p}_0\boldsymbol{y}\|^2 = \|\boldsymbol{y} - \mathfrak{p}\boldsymbol{y}\|^2 + \|\mathfrak{p}\boldsymbol{y} - \mathfrak{p}_0\boldsymbol{y}\|^2$ (Pythagoras). Using this we can rewrite Q further:

$$\begin{aligned} Q &= \left(\frac{\|\boldsymbol{y} - \mathfrak{p}\boldsymbol{y}\|^2}{\|\boldsymbol{y} - \mathfrak{p}\boldsymbol{y}\|^2 + \|\mathfrak{p}\boldsymbol{y} - \mathfrak{p}_0\boldsymbol{y}\|^2}\right)^{n/2} \\ &= \left(1 + \frac{\|\mathfrak{p}\boldsymbol{y} - \mathfrak{p}_0\boldsymbol{y}\|^2}{\|\boldsymbol{y} - \mathfrak{p}\boldsymbol{y}\|^2}\right)^{-n/2} = \left(1 + \frac{\dim L - \dim L_0}{n - \dim L} F\right)^{-n/2} \end{aligned}$$

where

$$F = \frac{\frac{1}{\dim L - \dim L_0}\|\mathfrak{p}\boldsymbol{y} - \mathfrak{p}_0\boldsymbol{y}\|^2}{\frac{1}{n - \dim L}\|\boldsymbol{y} - \mathfrak{p}\boldsymbol{y}\|^2}$$

is the test statistic commonly used.

Since Q is a decreasing function of F, the hypothesis is rejected for large values of F. It follows from Theorem 10.6 that under H_0 the nominator and the denominator of the F statistic are independent; the nominator has a χ^2 distribution with scale parameter $\sigma^2/(\dim L - \dim L_0)$ and $(\dim L - \dim L_0)$ degrees

of freedom, and the denominator has a χ^2 distribution with scale parameter $\sigma^2/(n-\dim L)$ and $(n-\dim L)$ degrees of freedom. The distribution of the test statistic is an F distribution with $(\dim L - \dim L_0)$ and $(n-\dim L)$ degrees of freedom.

F distribution
The F distribution with $f_t = f_{\text{top}}$ and $f_b = f_{\text{bottom}}$ degrees of freedom is the continuous distribution on $]0\,;+\infty[$ given by the density function

$$C\,\frac{x^{(f_t-2)/2}}{(f_b+f_t x)^{(f_t+f_b)/2}}$$

where C is the number

$$\frac{\Gamma\left(\frac{f_t+f_b}{2}\right)}{\Gamma\left(\frac{f_t}{2}\right)\Gamma\left(\frac{f_b}{2}\right)} f_t^{f_t/2} f_b^{f_b/2}.$$

All problems of estimation and hypothesis testing are now solved, at least in principle (and Theorems 7.1 and 7.2 on page 119/121 are proved). In the next sections we shall apply the general results to a number of more specific models.

11.2 The one-sample problem

We have observed values $y_1, y_2, \ldots, y_n$ of independent and identically distributed (one-dimensional) normal variables $Y_1, Y_2, \ldots, Y_n$ with mean μ and variance σ^2, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. The problem is to estimate the parameters μ and σ^2, and to test the hypothesis $H_0 : \mu = 0$.

We will think of the y_is as being arranged into an n-dimensional vector $\boldsymbol{y}$ which is regarded as an observation of an n-dimensional normal random variable $\boldsymbol{Y}$ with mean $\boldsymbol{\mu}$ and variance $\sigma^2\mathrm{I}$, and where the model claims that $\boldsymbol{\mu}$ is a point in the one-dimensional subspace $L = \{\mu\mathbf{1} : \mu \in \mathbb{R}\}$ consisting of vectors having the same value μ in all the places. (Here $\mathbf{1}$ is the vector with the value 1 in all places.)

According to Theorem 11.1 the maximum likelihood estimator $\widehat{\boldsymbol{\mu}}$ is found by projecting $\boldsymbol{y}$ orthogonally onto L. The residual vector $\boldsymbol{y} - \widehat{\boldsymbol{\mu}}$ will then be orthogonal to L and in particular orthogonal to the vector $\mathbf{1} \in L$, so

$$0 = \langle \boldsymbol{y} - \widehat{\boldsymbol{\mu}}, \mathbf{1} \rangle = \sum_{i=1}^{n} (y_i - \widehat{\mu}) = y_\bullet - n\widehat{\mu}$$

where as usual $y_\bullet$ denotes the sum of the y_is. Hence $\widehat{\boldsymbol{\mu}} = \widehat{\mu}\mathbf{1}$ where $\widehat{\mu} = \overline{y}$, that is, μ is estimated as the average of the observations. Next, we find that the maximum likelihood estimate of σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n}\|\boldsymbol{y} - \widehat{\boldsymbol{\mu}}\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2,$$

and that an unbiased estimate of σ^2 is

$$s^2 = \frac{1}{n-\dim L}\|\boldsymbol{y} - \widehat{\boldsymbol{\mu}}\|^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2.$$

These estimates were found by other means on page 119.

Next, we will test the hypothesis $H_0 : \mu = 0$, or rather $\boldsymbol{\mu} \in L_0$ where $L_0 = \{\mathbf{0}\}$. Under H_0, $\boldsymbol{\mu}$ is estimated as the projection of $\boldsymbol{y}$ onto L_0, which is $\mathbf{0}$. Therefore the F test statistic for testing H_0 is

$$F = \frac{\frac{1}{1-0}\|\widehat{\boldsymbol{\mu}} - \mathbf{0}\|^2}{\frac{1}{n-1}\|\boldsymbol{y} - \widehat{\boldsymbol{\mu}}\|^2} = \frac{n\widehat{\mu}^2}{s^2} = \left(\frac{\overline{y}}{\sqrt{s^2/n}}\right)^2,$$

so $F = t^2$ where t is the usual t test statistic, cf. page 132. Hence, the hypothesis H_0 can be tested either using the F statistic (with 1 and $n-1$ degrees of freedom) or using the t statistic (with $n-1$ degrees of freedom).

The hypothesis $\mu = 0$ is the only hypothesis of the form $\boldsymbol{\mu} \in L_0$ where L_0 is a linear subspace of L; then what to do if the hypothesis of interest is of the form $\mu = \mu_0$ where μ_0 is non-zero? Answer: Subtract μ_0 from all the ys and then apply the method just described. The resulting F or t statistic will have $\overline{y} - \mu_0$ instead of $\overline{y}$ in the denominator, everything else remains unchanged.

11.3 One-way analysis of variance

We have observations that are classified into k groups; the observations are named y_{ij} where i is the group number ($i = 1, 2, \ldots, k$), and j is the number of the observations within the group ($j = 1, 2, \ldots, n_i$). Schematically it looks like this:

	observations					
group 1	y_{11}	y_{12}	$\cdots$	y_{1j}	$\cdots$	y_{1n_1}
group 2	y_{21}	y_{22}	$\cdots$	y_{2j}	$\cdots$	y_{2n_2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
group i	y_{i1}	y_{i2}	$\cdots$	y_{ij}	$\cdots$	y_{in_i}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
group k	y_{k1}	y_{k2}	$\cdots$	y_{kj}	$\cdots$	y_{kn_k}

The variation between observations within a single group is assumed to be random, whereas there is a systematic variation between the groups. We also assume that the y_{ij}s are observed values of independent random variables Y_{ij}, and that the random variation can be described using the normal distribution. More specifically, the random variables Y_{ij} are assumed to be independent and normal, with mean $\mathrm{E}\, Y_{ij} = \mu_i$ and variance $\mathrm{Var}\, Y_{ij} = \sigma^2$. Spelled out in more details: there exist real numbers $\mu_1, \mu_2, \ldots, \mu_k$ and a positive number σ^2 such that Y_{ij} is normal with mean μ_i and variance σ^2, $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, k$, and all Y_{ij}s are independent. In this way the mean value parameters $\mu_1, \mu_2, \ldots, \mu_k$ describe the *systematic variation*, that is, the levels of the individual groups,

while the variance parameter σ^2 (and the normal distribution) describes the *random variation* within the groups. The random variation is supposed to be the same in all groups; this assumption can sometimes be tested, see Section 11.4 (page 176).

We are going to estimate the parameters $\mu_1, \mu_2, \ldots, \mu_k$ and σ^2, and to test the hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ of no difference between groups.

We will be thinking of the y_{ij}s as arranged as a single vector $\boldsymbol{y} \in V = \mathbb{R}^n$, where $n = n_\bullet$ is the number of observations. Then the basic model claims that $\boldsymbol{y}$ is an observed value of an n-dimensional normal random variable $\boldsymbol{Y}$ with mean $\boldsymbol{\mu}$ and variance $\sigma^2\mathrm{I}$, where $\sigma^2 > 0$, and where $\boldsymbol{\mu}$ belongs to the linear subspace that is written briefly as

$$L = \{\boldsymbol{\xi} \in V : \xi_{ij} = \mu_i\};$$

in more details, L is the set of vectors $\boldsymbol{\xi} \in V$ for which there exists a k-tuple $(\mu_1, \mu_2, \ldots, \mu_k) \in \mathbb{R}^k$ such that $\xi_{ij} = \mu_i$ for all i and j. The dimension of L is k (provided that $n_i > 0$ for all i).

According to Theorem 11.1 $\boldsymbol{\mu}$ is estimated as $\widehat{\boldsymbol{\mu}} = \mathfrak{p}\boldsymbol{y}$ where $\mathfrak{p}$ is the orthogonal projection of V onto L. To obtain a more usable expression for the estimator, we make use of the fact that $\widehat{\boldsymbol{\mu}} \in L$ and $\boldsymbol{y} - \widehat{\boldsymbol{\mu}} \perp L$, and in particular is $\boldsymbol{y} - \widehat{\boldsymbol{\mu}}$ orthogonal to each of the k vectors $\boldsymbol{e}^1, \boldsymbol{e}^2, \ldots, \boldsymbol{e}^k \in L$ defined in such a way that the (i,j)th entry of $\boldsymbol{e}^s$ is $(\boldsymbol{e}^s)_{ij} = \delta_{is}$, that is,

Kronecker's δ is the symbol $\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$

$$0 = \langle \boldsymbol{y} - \widehat{\boldsymbol{\mu}}, \boldsymbol{e}^s \rangle = \sum_{j=1}^{n_s} (y_{sj} - \widehat{\mu}_s) = y_{s\bullet} - n_s \widehat{\mu}_s, \qquad s = 1, 2, \ldots, k,$$

from which we get $\widehat{\mu}_i = y_{i\bullet}/n_i = \overline{y}_i$, so $\widehat{\mu}_i$ is the average of the observations in group i.

The variance parameter σ^2 is estimated as

$$s_0^2 = \frac{1}{n - \dim L} \|\boldsymbol{y} - \mathfrak{p}\boldsymbol{y}\|^2 = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2$$

with $n - k$ degrees of freedom. The two estimators $\widehat{\boldsymbol{\mu}}$ and s_0^2 are independent.

The hypothesis H_0 of identical groups can be expressed as $H_0 : \boldsymbol{\mu} \in L_0$, where

$$L_0 = \{\boldsymbol{\xi} \in V : \xi_{ij} = \mu\}$$

is the one-dimensional subspace of vectors having the same value in all n entries. From Section 11.2 we know that the projection of $\boldsymbol{y}$ onto L_0 is the vector $\mathfrak{p}_0\boldsymbol{y}$ that

has the grand average $\overline{y} = y_{\cdot\cdot}/n_{\cdot}$ in all entries. To test the hypothesis H_0 we use the test statistic

$$F = \frac{\frac{1}{\dim L - \dim L_0}\|\mathfrak{p}\boldsymbol{y} - \mathfrak{p}_0\boldsymbol{y}\|^2}{\frac{1}{n-\dim L}\|\boldsymbol{y} - \mathfrak{p}\boldsymbol{y}\|^2} = \frac{s_1^2}{s_0^2}$$

where

$$\begin{aligned} s_1^2 &= \frac{1}{\dim L - \dim L_0}\|\mathfrak{p}\boldsymbol{y} - \mathfrak{p}_0\boldsymbol{y}\|^2 \\ &= \frac{1}{k-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{y}_i - \overline{y})^2 = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2. \end{aligned}$$

The distribution of F under H_0 is the F distribution with $k-1$ and $n-k$ degrees of freedom. We say that s_0^2 describes the *variation within groups*, and s_1^2 describes the *variation between groups*. Accordingly, the F statistic measures the variation between groups relative to the variation within groups—hence the name of this method: one-way *analysis of variance*.

If the hypothesis is accepted, the mean vector $\boldsymbol{\mu}$ is estimated as the vector $\mathfrak{p}_0\boldsymbol{y}$ which has $\overline{y}$ in all entries, and the variance σ^2 is estimated as

$$s_{01}^2 = \frac{1}{n - \dim L_0}\|\boldsymbol{y} - \mathfrak{p}_0\boldsymbol{y}\|^2 = \frac{\|\boldsymbol{y} - \mathfrak{p}\boldsymbol{y}\|^2 + \|\mathfrak{p}\boldsymbol{y} - \mathfrak{p}_0\boldsymbol{y}\|^2}{(n - \dim L) + (\dim L - \dim L_0)},$$

that is,

$$s_{01}^2 = \frac{1}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2 + \sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2\Big).$$

Example 11.1: Strangling of dogs
In an investigation of the relations between the duration of hypoxia and the concentration of hypoxanthine in the cerebrospinal fluid (described in more details in Example 11.2 on page 192), an experiment was conducted in which a number of dogs were exposed to hypoxia and the concentration of hypoxanthine was measured at four different times, see Table 11.7 on page 192. In the present example we will examine whether there is a significant difference between the four groups corresponding to the four times.

First, we calculate estimates of the unknown parameters, see Table 11.1 which also gives a few intermediate results. The estimated mean values are 1.46, 5.50, 7.48 and 12.81, and the estimated variance is $s_0^2 = 4.82$ with 21 degrees of freedom. The estimated variance between groups is $s_1^2 = 465.5/3 = 155.2$ with 3 degrees of freedom, so the F statistic for testing homogeneity between groups is $F = 155.2/4.82 = 32.2$ which is highly significant. Hence there is a highly significant difference between the means of the four groups.

i	n_i	$\sum_{j=1}^{n_i} y_{ij}$	$\overline{y}_i$	f_i	$\sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2$	s_i^2
1	7	10.2	1.46	6	7.64	1.27
2	6	33.0	5.50	5	14.94	2.99
3	5	37.4	7.48	4	30.51	7.63
4	7	89.7	12.81	6	48.23	8.04
sum	25	170.3		21	101.32	
average			6.81			4.82

Table 11.1
Strangling of dogs: intermediate results.

	f	SS	s^2	test
variation within groups	21	101.32	4.82	
variation between groups	3	465.47	155.16	155.16/4.82=32.2
total variation	24	566.79	23.62	

Table 11.2
Strangling of dogs: Analysis of variance table.
f is the number of degrees of freedom, SS is the sum of squared deviations, and $s^2 = SS/f$.

Traditionally, calculations and results are summarised in an *analysis of variance table* (ANOVA table), see Table 11.2.

The analysis of variance model assumes homogeneity of variances, and we can test for that, for example using Bartlett's test (Section 11.4). Using the s^2 values from Table 11.1 we find that the value of the test statistic B (equation (11.1) on page 177) is $B = -\left(6 \ln \frac{1.27}{4.82} + 5 \ln \frac{2.99}{4.82} + 4 \ln \frac{7.63}{4.82} + 6 \ln \frac{8.04}{4.82}\right) = 5.5$. This value is well below the 10% quantile of the χ^2 distribution with $k - 1 = 3$ degrees of freedom, and hence it is not significant. We may therefore assume homogeneity of variances.

▷ [See also Example 11.2 on page 192.]

The two-sample problem with unpaired observations

The case $k = 2$ is often called the two-sample problem (no surprise!). In fact, many textbooks treat the two-sample problem as a separate topic, and sometimes even before the k-sample problem. The only interesting thing to be added about the case $k = 2$ seems to be the fact that the F statistic equals t^2, where

$$t = \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s_0^2}},$$

cf. page 134. Under H_0 the distribution of t is the t distribution with $n - 2$ degrees of freedom.

11.4 Bartlett's test of homogeneity of variances

Normal models often require the observations to come from normal distributions all with the same variance (or a variance which is known except for an unknown factor). This section describes a test procedure that can be used for testing whether a number of groups of observations of normal variables can be assumed to have equal variances, that is, for testing *homogeneity of variances* (or *homoscedasticity*). All groups must contain more than one observation, and in order to apply the usual χ^2 approximation to the distribution of the $-2\ln Q$ statistic, each group has to have at least six observations, or rather: in each group the variance estimate has to have at least five degrees of freedom.

The general situation is as in the k-sample problem (the one-way analysis of variance situation), see page 172, and we want to perform a test of the assumption that the groups have the same variance σ^2. This can be done as follows.

Assume that each of the k groups has its own mean value parameter and its own variance parameter: the parameters associated with group no. i are μ_i and σ_i^2. From earlier results we know that the usual unbiased estimate of σ_i^2 is $s_i^2 = \dfrac{1}{f_i}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2$, where $f_i = n_i - 1$ is the number of degrees of freedom of the estimate, and n_i is the number of observations in group i. According to Proposition 7.1 (page 119) the estimator s_i^2 has a gamma distribution with shape parameter $f_i/2$ and scale parameter $2\sigma_i^2/f_i$. Hence we will consider the following statistical problem: We are given independent observations $s_1^2, s_2^2, \ldots, s_k^2$ such that s_i^2 is an observation from a gamma distribution with shape parameter $f_i/2$ and scale parameter $2\sigma_i^2/f_i$ where f_i is a known number; in this model we want to test the hypothesis

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2.$$

The likelihood function (when having observed $s_1^2, s_2^2, \ldots, s_k^2$) is

$$\begin{aligned} L(\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2) &= \prod_{i=1}^{k} \frac{1}{\Gamma(\frac{f_i}{2})\left(\frac{2\sigma_i^2}{f_i}\right)^{f_i/2}} \left(s_i^2\right)^{f_i/2-1} \exp\left(-s_i^2 \Big/ \frac{2\sigma_i^2}{f_i}\right) \\ &= \text{const} \cdot \prod_{i=1}^{k} \left(\sigma_i^2\right)^{-f_i/2} \exp\left(-\frac{f_i}{2}\frac{s_i^2}{\sigma_i^2}\right) \\ &= \text{const} \cdot \left(\prod_{i=1}^{k} (\sigma_i^2)^{-f_i/2}\right) \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{k} f_i \frac{s_i^2}{\sigma_i^2}\right). \end{aligned}$$

The maximum likelihood estimates of $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$ are $s_1^2, s_2^2, \ldots, s_k^2$. The maximum likelihood estimate of the common value σ^2 under H_0 is the point of maximum of the function $\sigma^2 \mapsto L(\sigma^2, \sigma^2, \ldots, \sigma^2)$, and that is $s_0^2 = \frac{1}{f_\bullet} \sum_{i=1}^{k} f_i s_i^2$, where $f_\bullet = \sum_{i=1}^{k} f_i$. We see that s_0^2 is a weighted average of the s_i^2s, using the numbers of degrees of freedom as weights. The likelihood ratio test statistic is $Q = L(s_0^2, s_0^2, \ldots, s_0^2) \big/ L(s_1^2, s_2^2, \ldots, s_k^2)$. Normally one would use the test statistic $-2 \ln Q$, and $-2 \ln Q$ would then be denoted B, the *Bartlett test statistic* for testing homogeneity of variances; some straightforward algebra leads to

$$B = -\sum_{i=1}^{k} f_i \ln \frac{s_i^2}{s_0^2}. \tag{11.1}$$

The B statistic is always non-negative, and large values are significant, meaning that the hypothesis H_0 of homogeneity of variances is wrong. Under H_0, B has an approximate χ^2 distribution with $k-1$ degrees of freedom, so an approximate test probability is $\varepsilon = \mathrm{P}(\chi_{k-1}^2 \geq B_{\mathrm{obs}})$. The χ^2 approximation is valid when all f_is are large; a rule of thumb is that they all have to be at least 5.

With only two groups (i.e. $k = 2$) another possibility is to test the hypothesis of variance homogeneity using the ratio between the two variance estimates as a test statistic. This was discussed in connection with the two-sample problem with normal variables, see Exercise 8.2 page 136. (That two-sample test does not rely on any χ^2 approximation and has no restrictions on the number of degrees of freedom.)

11.5 Two-way analysis of variance

We have a number of observations y arranged as a two-way table:

	1	2	…	j	…	s
1	y_{11k} $k=1,2,\ldots,n_{11}$	y_{12k} $k=1,2,\ldots,n_{12}$	⋯	y_{1jk} $k=1,2,\ldots,n_{1j}$	⋯	y_{1sk} $k=1,2,\ldots,n_{1s}$
2	y_{21k} $k=1,2,\ldots,n_{21}$	y_{22k} $k=1,2,\ldots,n_{22}$	⋯	y_{2jk} $k=1,2,\ldots,n_{2j}$	⋯	y_{2sk} $k=1,2,\ldots,n_{2s}$
⋮	⋮	⋮		⋮		⋮
i	y_{i1k} $k=1,2,\ldots,n_{i1}$	y_{i2k} $k=1,2,\ldots,n_{i2}$	⋯	y_{ijk} $k=1,2,\ldots,n_{ij}$	⋯	y_{isk} $k=1,2,\ldots,n_{is}$
⋮	⋮	⋮		⋮	⋱	⋮
r	y_{r1k} $k=1,2,\ldots,n_{r1}$	y_{r2k} $k=1,2,\ldots,n_{r2}$	⋯	y_{rjk} $k=1,2,\ldots,n_{rj}$	⋯	y_{rsk} $k=1,2,\ldots,n_{rs}$

Here y_{ijk} is observation no. k in the (i,j)th *cell*, which contains a total of n_{ij} observations. The (i,j)th cell is located in the ith *row* and jth *column* of the table, and the table has a total of r rows and s columns.

We shall now formulate a statistical model in which the y_{ijk}s are observations of independent normal variables Y_{ijk}, all with the same variance σ^2, and with a mean value structure that reflects the way the observations are classified. Here follows the basic model and a number of possible hypotheses about the mean value parameters (the variance is always the same unknown σ^2):

- The *basic model* G states that Ys belonging to the same cell have the same mean value. More detailed this model states that there exist numbers η_{ij}, $i = 1,2,\ldots,r$, $j = 1,2,\ldots,s$, such that $\mathrm{E}\,Y_{ijk} = \eta_{ij}$ for all i, j and k. A short formulation of the model is

$$G : \mathrm{E}\,Y_{ijk} = \eta_{ij}\,.$$

- The hypothesis of *additivity*, also known as the hypothesis of *no interaction*, states that there is no interaction between rows and columns, but that row effects and column effects are additive. More detailed the hypothesis is that there exist numbers $\alpha_1, \alpha_2, \ldots, \alpha_r$ and $\beta_1, \beta_2, \ldots, \beta_s$ such that $\mathrm{E}\,Y_{ijk} = \alpha_i + \beta_j$ for all i, j and k. A short formulation of this hypothesis is

$$H_0 : \mathrm{E}\,Y_{ijk} = \alpha_i + \beta_j\,.$$

- The hypothesis of *no column effects* states that there is no difference between the columns. More detailed the hypothesis is that there exist numbers $\alpha_1, \alpha_2, \ldots, \alpha_r$ such that $\mathrm{E}\, Y_{ijk} = \alpha_i$ for all i, j and k. A short formulation of this hypothesis is
$$H_1 : \mathrm{E}\, Y_{ijk} = \alpha_i \,.$$
- The hypothesis of *no row effects* states that there is no difference between the rows. More detailed the hypothesis is that there exist numbers $\beta_1, \beta_2, \ldots, \beta_s$ such that $\mathrm{E}\, Y_{ijk} = \beta_j$ for all i, j and k. A short formulation of this hypothesis is
$$H_2 : \mathrm{E}\, Y_{ijk} = \beta_j \,.$$
- The hypothesis of *total homogeneity* states that there is no difference at all between the cells, more detailed the hypothesis is that there exists a number γ such that $\mathrm{E}\, Y_{ijk} = \gamma$ for all i, j and k. A short formulation of this hypothesis is
$$H_3 : \mathrm{E}\, Y_{ijk} = \gamma \,.$$

We will express these hypotheses in linear algebra terms, so we will think of the observations as forming a single vector $\boldsymbol{y} \in V = \mathbb{R}^n$, where $n = n_{\cdot\cdot}$ is the number of (one-dimensional) observations; we shall be thinking of vectors in V as being arranged as two-way tables like the one on the preceeding page. In this set-up, the basic model and the four hypotheses can be rewritten as statements that the mean vector $\boldsymbol{\mu} = \mathrm{E}\, \boldsymbol{Y}$ belongs to a certain linear subspace:

$$\begin{aligned}
G &: \boldsymbol{\mu} \in L && \text{where } L = \{\boldsymbol{\xi} : \xi_{ijk} = \eta_{ij}\},\\
H_0 &: \boldsymbol{\mu} \in L_0 && \text{where } L_0 = \{\boldsymbol{\xi} : \xi_{ijk} = \alpha_i + \beta_j\},\\
H_1 &: \boldsymbol{\mu} \in L_1 && \text{where } L_1 = \{\boldsymbol{\xi} : \xi_{ijk} = \alpha_i\},\\
H_2 &: \boldsymbol{\mu} \in L_2 && \text{where } L_2 = \{\boldsymbol{\xi} : \xi_{ijk} = \beta_j\},\\
H_3 &: \boldsymbol{\mu} \in L_3 && \text{where } L_3 = \{\boldsymbol{\xi} : \xi_{ijk} = \gamma\}.
\end{aligned}$$

There are certain relations between the hypotheses/subspaces:

$$G \Leftarrow H_0 \begin{array}{c} \Leftarrow \\ \Leftarrow \\ \Leftarrow \end{array} \begin{array}{c} H_1 \\ \\ H_2 \end{array} \begin{array}{c} \Leftarrow \\ \\ \Leftarrow \end{array} H_3 \qquad\qquad L \supseteq L_0 \begin{array}{c} \supsetneq \\ \\ \supsetneq \end{array} \begin{array}{c} L_1 \\ \\ L_2 \end{array} \begin{array}{c} \supsetneq \\ \\ \supsetneq \end{array} L_3$$

The basic model and the models corresponding to H_1 and H_2 are instances of k sample problems (with k equal to rs, r, and s, respectively), and the model corresponding to H_3 is a one-sample problem, so it is straightforward to write

down the estimates in these four models:

$$\begin{aligned}
&\text{under } G,\ \widehat{\eta_{ij}} = \overline{y}_{ij} && \text{i.e., the average in cell (i, j),}\\
&\text{under } H_1,\ \widehat{\alpha_i} = \overline{y}_{i\bullet} && \text{i.e., the average in row } i,\\
&\text{under } H_2,\ \widehat{\beta_j} = \overline{y}_{\bullet j} && \text{i.e., the average in column } j,\\
&\text{under } H_3,\ \widehat{\gamma} = \overline{y}_{\bullet\bullet} && \text{i.e., the overall average.}
\end{aligned}$$

On the other hand, it can be quite intricate to estimate the parameters under the hypothesis H_0 of no interaction.

First we shall introduce the notion of a connected model.

Connected models

What would be a suitable requirement for the numbers of observations in the various cells of $\boldsymbol{y}$, that is, requirements for the table

$$\boldsymbol{n} = \begin{array}{|c|c|c|c|}\hline n_{11} & n_{12} & \cdots & n_{1s}\\ \hline n_{21} & n_{22} & \cdots & n_{2s}\\ \hline \vdots & \vdots & \ddots & \vdots\\ \hline n_{r1} & n_{r2} & \cdots & n_{rs}\\ \hline\end{array}$$

We should not allow rows (or columns) consisting entirely of 0s, but do all the n_{ij}s have to be non-zero?

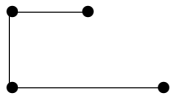
To clarify the problems consider a simple example with $r = 2$, $s = 2$, and

$$\boldsymbol{n} = \begin{array}{|c|c|}\hline 0 & 9\\ \hline 9 & 0\\ \hline\end{array}$$

In this case the subspace L corresponding to the additive model has dimension 2, since the only parameters are η_{12} and η_{21}, and in fact $L = L_0 = L_1 = L_2$. In particular, $L_1 \cap L_2 \neq L_3$, and hence it is not true in general that $L_1 \cap L_2 = L_3$. The fact is, that the present model consists of two separate sub-models (one for cell $(1,2)$ and one for cell $(2,1)$), and therefore $\dim(L_1 \cap L_2) > 1$, or equivalently, $\dim(L_0) < r + s - 1$ (to see this, apply the general formula $\dim(L_1 + L_2) = \dim L_1 + \dim L_2 - \dim(L_1 \cap L_2)$ and the fact that $L_0 = L_1 + L_2$).

If the model cannot be split into separate sub-models, we say that the model is *connected*. The notion of a connected model can be specified in the following way: From the table $\boldsymbol{n}$ form a graph whose vertices are those integer pairs (i, j) for which $n_{ij} > 0$, and whose edges are connecting pairs of adjacent vertices, i.e., pairs of vertices with a common i or a common j. Example:

$$\begin{array}{|c|c|c|}\hline 5 & 8 & 0\\ \hline 1 & 0 & 4\\ \hline\end{array} \quad \longrightarrow$$



Then the model is said to be connected if the graph is connected (in the usual graph-theoretic sense).

It is easy to verify that the model is connected, if and only if the null space of the linear map that takes the $(r+s)$-dimensional vector $(\alpha_1, \alpha_2, \ldots, \alpha_r, \beta_1, \beta_2, \ldots, \beta_s)$ into the "corresponding" vector in L_0, is one-dimensional, that is, if and only if $L_3 = L_1 \cap L_2$. In plain language, a connected model is a model for which it is true that "if all row parameters are equal and all column parameters are equal, then there is total homogeneity".

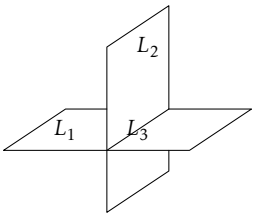
From now on, all two-way ANOVA models are assumed to be connected.

The projection onto L_0

The general theory tells that under H_0 the mean vector $\boldsymbol{\mu}$ is estimated as the projection $\mathfrak{p}_0\boldsymbol{y}$ of $\boldsymbol{y}$ onto L_0. In certain cases this projection can be calculated very easily. – First recall that $L_0 = L_1 + L_2$ and (since the model is connected) $L_3 = L_1 \cap L_2$.

Now *suppose* that

$$(L_1 \cap L_3^\perp) \perp (L_2 \cap L_3^\perp). \tag{11.2}$$



$(L_1 \cap L_3^\perp) \perp (L_2 \cap L_3^\perp)$

Then

$$L_0 = (L_1 \cap L_3^\perp) \oplus (L_2 \cap L_3^\perp) \oplus L_3$$

and hence

$$\mathfrak{p}_0 = (\mathfrak{p}_1 - \mathfrak{p}_3) + (\mathfrak{p}_2 - \mathfrak{p}_3) + \mathfrak{p}_3,$$

that is,

$$\begin{aligned}\widehat{\alpha_i + \beta_j} &= (\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot\cdot}) + (\overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot\cdot}) + \overline{y}_{\cdot\cdot\cdot} \\ &= \overline{y}_{i\cdot\cdot} + \overline{y}_{\cdot j\cdot} - \overline{y}_{\cdot\cdot\cdot}\end{aligned}$$

This is indeed a very simple and nice formula for (the coordinates of) the projection of $\boldsymbol{y}$ onto L_0; we only need to know when the premise is true. A necessary and sufficient condition for (11.2) is that

$$\langle \mathfrak{p}_1\boldsymbol{e} - \mathfrak{p}_3\boldsymbol{e}, \mathfrak{p}_2\boldsymbol{f} - \mathfrak{p}_3\boldsymbol{f} \rangle = 0 \tag{11.3}$$

for all $\boldsymbol{e}, \boldsymbol{f} \in \mathfrak{B}$, where $\mathfrak{B}$ is a basis for L. As $\mathfrak{B}$ we can use the vectors $\boldsymbol{e}^{\rho\sigma}$ where $(\boldsymbol{e}^{\rho\sigma})_{ijk} = 1$ if $(i,j) = (\rho,\sigma)$ and 0 otherwise. Inserting such vectors in (11.3) leads to this necessary and sufficient condition for (11.2):

$$n_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}, \qquad i = 1, 2, \ldots, r;\ j = 1, 2, \ldots, s. \tag{11.4}$$

When this condition holds, we say that the model is *balanced*. Note that if all the cells of $\boldsymbol{y}$ have the same number of observations, then the model is always balanced.

In conclusion, in a balanced model, that is, if (11.4) is true, the estimates under the hypothesis of additivity can be calculated as

$$\widehat{\alpha_i + \beta_j} = \overline{y}_{i\bullet\bullet} + \overline{y}_{\bullet j\bullet} - \overline{y}_{\bullet\bullet\bullet}. \tag{11.5}$$

Testing the hypotheses

The various hypotheses about the mean vector can be tested using an F statistic. As always, the denominator of the test statistic is the estimate of the variance parameter calculated under the current basic model, and the nominator of the test statistic is an estimate of "the variation of the hypothesis about the current basic model". The F statistic inherits its two numbers of degrees of freedom from the nominator and the denominator variance estimates.

1. For H_0, the hypothesis of *additivity*, the test statistic is $F = s_1^2 \big/ s_0^2$ where

$$s_0^2 = \frac{1}{n - \dim L} \|\boldsymbol{y} - \mathfrak{p}\boldsymbol{y}\|^2 = \frac{1}{n-g} \sum_{i=1}^{r} \sum_{j=1}^{s} \sum_{k=1}^{n_{ij}} \left(y_{ijk} - \overline{y}_{ij}\right)^2 \tag{11.6}$$

   describes the *within groups variation* ($g = \dim L$ is the number of non-empty cells), and

$$\begin{aligned} s_1^2 &= \frac{1}{\dim L - \dim L_0} \|\mathfrak{p}\boldsymbol{y} - \mathfrak{p}_0\boldsymbol{y}\|^2 \\ &= \frac{1}{g - (r+s-1)} \sum_{i=1}^{r} \sum_{j=1}^{s} \sum_{k=1}^{n_{ij}} \left(\overline{y}_{ij} - (\widehat{\alpha_i + \beta_j})\right)^2 \end{aligned}$$

   describes the *interaction variation*. In a balanced model

$$s_1^2 = \frac{1}{g - (r+s-1)} \sum_{i=1}^{r} \sum_{j=1}^{s} \sum_{k=1}^{n_{ij}} \left(\overline{y}_{ij} - \overline{y}_{i\bullet} - \overline{y}_{\bullet j} + \overline{y}_{\bullet\bullet\bullet}\right)^2.$$

   This requires that $n > g$, that is, there must be some cells with more than one observation.

2. Once H_0 is accepted, the additive model is named the current basic model, and an updated variance estimate is calculated:

$$\begin{aligned} s_{01}^2 &= \frac{1}{n - \dim L_0} \|\boldsymbol{y} - \mathfrak{p}_0\boldsymbol{y}\|^2 \\ &= \frac{1}{n - \dim L_0} \left(\|\boldsymbol{y} - \mathfrak{p}\boldsymbol{y}\|^2 + \|\mathfrak{p}\boldsymbol{y} - \mathfrak{p}_0\boldsymbol{y}\|^2 \right). \end{aligned}$$

   Dependent on the actual circumstances, the next hypothesis to be tested is either H_1 or H_2 or both.

- To test the hypothesis H_1 of *no column effects*, use the test statistic $F = s_1^2 / s_{01}^2$ where

$$s_1^2 = \frac{1}{\dim L_0 - \dim L_1} \|\mathfrak{p}_0 \boldsymbol{y} - \mathfrak{p}_1 \boldsymbol{y}\|^2$$
$$= \frac{1}{s-1} \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij} \left(\widehat{\alpha_i + \beta_j} - \overline{y}_{i\bullet} \right)^2.$$

  In a balanced model,

$$s_1^2 = \frac{1}{s-1} \sum_{j=1}^{s} n_{\bullet j} \left(\overline{y}_{\bullet j} - \overline{y}_{\bullet\bullet} \right)^2.$$

- To test the hypothesis H_2 of *no row effects*, use the test statistic $F = s_2^2 / s_{01}^2$ where

$$s_2^2 = \frac{1}{\dim L_0 - \dim L_2} \|\mathfrak{p}_0 \boldsymbol{y} - \mathfrak{p}_2 \boldsymbol{y}\|^2$$
$$= \frac{1}{r-1} \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij} \left(\widehat{\alpha_i + \beta_j} - \overline{y}_{\bullet j} \right)^2.$$

  In a balanced model,

$$s_2^2 = \frac{1}{r-1} \sum_{i=1}^{r} n_{i\bullet} \left(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet} \right)^2.$$

Note that H_1 and H_2 are tested side by side—both are tested against H_0 and the variance estimate s_{01}^2; in a balanced model, the nominators s_1^2 and s_2^2 of the F statistics are independent (since $\mathfrak{p}_0 \boldsymbol{y} - \mathfrak{p}_1 \boldsymbol{y}$ and $\mathfrak{p}_0 \boldsymbol{y} - \mathfrak{p}_2 \boldsymbol{y}$ are orthogonal).

An example (growing of potatoes)

To obtain optimal growth conditions, plants need a number of nutrients administered in correct proportions. This example deals with the problem of finding an appropriate proportion between the amount of nitrogen and the amount of phosphate in the fertilisers used when growing potatoes.

Potatoes were grown in six different ways, corresponding to six different combinations of amount of phosphate supplied (0, 1 or 2 units) and amount of nitrogen supplied (0 or 1 unit). This results in a set of observations, the yields, that are classified according to two criteria, amount of phosphate and amount of nitrogen.

Table 11.3
Growing of potatoes: Yield of each of 36 parcels. The entries of the table are $1000 \times (\log(\textit{yield in lbs}) - 3)$.

		nitrogen					
		0			1		
phosphate	0	591	450	584	619	618	524
		509	636	413	651	655	564
	1	722	689	625	801	688	682
		584	614	513	703	774	623
	2	702	677	684	814	757	810
		643	668	699	792	790	703

Table 11.4
Growing of potatoes: Estimated mean $\overline{y}$ *(top) and variance* s^2 *(bottom) in each of the six groups, cf. Table 11.3.*

		nitrogen	
		0	1
phosphate	0	530.50	605.17
		7680.30	2644.57
	1	624.50	711.83
		5569.90	4248.57
	2	678.83	777.67
		474.97	1745.07

Table 11.3 records some of the results from one such experiment, conducted in 1932 at Ely.* We want to investigate the effects of the two factors nitrogen and phosphate separately and jointly, and to be able to answer questions such as: does the effect of adding one unit of nitrogen depend on whether we add phosphate or not? We will try a two-way analysis of variance.

The two-way analysis of variance model assumes homogeneity of the variances, and we are in a position to perform a test of this assumption. We calculate the empirical mean and variance for each of the six groups, see Table 11.4. The total sum of squares is found to be 111817, so that the estimate of the within-groups variance is $s_0^2 = 111817/(36-6) = 3727.2$ (cf. equation (11.6)). The Bartlett test statistic is $B_{\text{obs}} = 9.5$, which is only slightly above the 10% quantile of the χ^2 distribution with $6-1=5$ degrees of freedom, so there is no significant evidence against the assumption of homogeneity of variances.

Then we can test whether the data can be described adequately using an additive model in which the effects of the two factors "phosphate" and "ni-

* Note that Table 11.3 does not give the actual yields, but the yields transformed as indicated.
The reason for taking the logarithm of the yields is that experience shows that the distribution of the yield of potatoes is not particularly normal, whereas the distribution of the logarithm of the yield looks more like a normal distribution. After having passed to logarithms, it turned out that all the numbers were of the form 3-point-something, and that is the reason for subtracting 3 and multiplying by 1000 (to get rid of the decimal point).

		nitrogen	
		0	1
phosphate	0	530.50	605.17
		524.36	611.30
	1	624.50	711.83
		624.70	711.64
	2	678.83	777.67
		684.78	771.72

Table 11.5
Growing of potatoes: Estimated group means in the basic model (top) and in the additive model (bottom).

trogen" enter in an additive way. Let y_{ijk} denote the kth observation in row no. i and column no. j. According to the basic model, the y_{ijk}s are understood as observations of independent normal random variables Y_{ijk}, where Y_{ijk} has mean μ_{ij} and variance σ^2. The hypothesis of additivity can be formulated as $H_0 : \mathrm{E}\, Y_{ijk} = \alpha_i + \beta_j$. Since the model is balanced (equation (11.4) is satisfied), the estimates $\widehat{\alpha_i + \beta_j}$ can be calculated using equation (11.5). First, calculate some auxiliary quantities

$$\begin{array}{lll} \overline{y}_{\cdot\cdot} = 654.75 & \overline{y}_{1\cdot} - \overline{y}_{\cdot\cdot} = -86.92 & \overline{y}_{\cdot 1} - \overline{y}_{\cdot\cdot} = -43.47 \\ & \overline{y}_{2\cdot} - \overline{y}_{\cdot\cdot} = 13.42 & \overline{y}_{\cdot 2} - \overline{y}_{\cdot\cdot} = 43.47 \\ & \overline{y}_{3\cdot} - \overline{y}_{\cdot\cdot} = 73.50 & \end{array} .$$

Then we can calculate the estimated group means $\widehat{\alpha_i + \beta_j}$ under the hypothesis of additivity, see Table 11.5. The variance estimate to be used under this hypothesis is $s_{01}^2 = \dfrac{112693.5}{36-(3+2-1)} = 3521.7$ with $36-(3+2-1) = 32$ degrees of freedom.

The interaction variance is

$$s_1^2 = \frac{877}{6-(3+2-1)} = \frac{877}{2} = 438.5\,,$$

and we have already found that $s_0^2 = 3727.2$, so the F statistic for testing the hypothesis of additivity is

$$F = \frac{s_1^2}{s_0^2} = \frac{438.5}{3727.2} = 0.12\,.$$

This value is to be compared to the F distribution with 2 and 30 degrees of freedom, and we find the test probability to be almost 90%, so there is no doubt that we can accept the hypothesis of additivity.

Having accepted the additive model, we should from now on use the improved variance estimate

$$s_{01}^2 = \frac{112694}{36-(3+2-1)} = 3521.7$$

Table 11.6
Growing of potatoes: Analysis of variance table.
f is the number of degrees of freedom, SS is the Sum of squared deviations, $s^2 = SS/f$.

variation	f	SS	s^2	test
within groups	30	111817	3727	
interaction	2	877	439	439/3727=0.12
the additive model	32	112694	3522	
between N-levels	1	68034	68034	68034/3522=19
between P-levels	2	157641	78820	78820/3522=22
total	35	338369		

with $36 - (3 + 2 - 1) = 32$ degrees of freedom.

Knowing that the additive model gives an adequate description of the data, it makes sense to talk of a nitrogen effect and a phosphate effect, and it may be of some interest to test whether these effects are significant.

We can test the hypothesis H_1 of no nitrogen effect (column effect): The variance between nitrogen levels is

$$s_2^2 = \frac{68034}{2-1} = 68034\,,$$

and the variance estimate in the additive model is $s_{01}^2 = 3522$ with 32 degrees of freedom, so the F statistic is

$$F_{\text{obs}} = \frac{s_2^2}{s_{02}^2} = \frac{68034}{3522} = 19$$

to be compared to the $F_{1,32}$ distribution. The value $F_{\text{obs}} = 19$ is undoubtedly significant, so the hypothesis of no nitrogen effect must be rejected, that is, nitrogen has definitely an effect.

We can test for the effect of phosphate in a similar way, and it turns out that phosphate, too, has a highly significant effect.

The analysis of variance table (Table 11.6) gives an overview of the analysis.

11.6 Regression analysis

Regression analysis is a technique for modelling and studying the dependence of one quantity on one or more other quantities.

Suppose that for each of a number of items (test subjects, test animals, single laboratory measurements, etc.) a number of quantities (variables) have been measured. One of these quantities/variables has a special position, because we wish to "describe" or "explain" this variable in terms of the others. The generic

symbol for the variable to be described is usually y, and the generic symbols for those quantities that are used to describe y, are $x_1, x_2, \ldots, x_p$. Often used names for the x and y quantities are

y	$x_1, x_2, \ldots, x_p$
dependent variable	independent variables
explained variable	explanatory variables
response variable	covariate

We briefly outline a few examples:

1. The doctor observes the survival time y of patients treated for a certain disease, but the doctor also records some extra information, for example sex, age and weight of the patient. Some of these "covariates" may contain information that could turn out to be useful in a model for the survival time.
2. In a number of fairly similar industrialised countries the prevalence of lung cancer, cigarette consumption, the consumption of fossil fuels was recorded and the results given as per capita values. Here, one might name the lung cancer prevalence the y variable and then try to "explain" it using the other two variables as explanatory variables.
3. In order to determine the toxicity of a certain drug, a number of test animals are exposed to various doses of the drug, and the number of animals dying is recorded. Here, the dose x is an independent variable determined as part of the design of the experiment, and the number y of dead animals is the dependent variable.

Regression analysis is about developing statistical models that attempt to describe a variable y using a known and fairly simple function of a few number of covariates $x_1, x_2, \ldots, x_p$ and parameters $\beta_1, \beta_2, \ldots, \beta_p$. The parameters are the same for all data points $(y, x_1, x_2, \ldots, x_p)$.

Obviously, a statistical model cannot be expected to give a perfect description of the data, partly because the model hardly is totally correct, and partly because the very idea of a statistical model is to model only the main features of the data while leaving out the finer details. So there will almost always be a certain difference between an observed value y and the corresponding "fitted" value $\widehat{y}$, i.e. the value that the model predicts from the covariates and the estimated parameters. This difference is known as the *residual*, often denoted by e, and we can write

$$y = \widehat{y} + e$$
$$\text{observed value} = \text{fitted value} + \text{residual}.$$

The residuals are what the model does not describe, so it would seem natural

to describe them as being *random*, that is, as random numbers drawn from a certain probability distribution.

-oOo-

Vital conditions for using regression analysis are that

1. the ys (and the residuals) are subject to random variation, wheras the xs are assumed to be known constants.
2. the individual ys are *independent*: the random causes that act on one measurement of y (after taking account for the covariates) are assumed to be independent of those acting on other measurements of y.

The simplest examples of regression analysis use only a single covariate x, and the task then is to describe y using a known, simple function of x. The simplest non-trivial function would be of the form $y = \alpha + x\beta$ where α and β are two parameters, that is, we assume y to be an affine function of x. Thus we have arrived at the *simple linear regression* model, cf. page 109.

A more advanced example is the *multiple linear regression* model, which has p covariates $x_1, x_2, \ldots, x_p$ and attempts to model y as $y = \sum_{j=1}^{p} x_j \beta_j$.

Specification of the model

We shall now discuss the multiple linear regression model. In order to have a genuine statistical model, we need to specify the probability distribution describing the random variation of the ys. Throughout this chapter we assume this probability distribution to be a normal distribution with variance σ^2 (which is the same for all observations).

We can specify the model as follows: The available data consists of n measurements of the dependent variable y, and for each measurement the corresponding values of the p covariates $x_1, x_2, \ldots, x_p$ are known. The ith set of values is $\big(y_i, (x_{i1}, x_{i2}, \ldots, x_{ip})\big)$. It is assumed that $y_1, y_2, \ldots, y_n$ are observed values of independent normal random variables $Y_1, Y_2, \ldots, Y_n$, all having the same variance σ^2, and with

$$\mathrm{E}\, Y_i = \sum_{j=1}^{p} x_{ij}\beta_j, \quad i = 1, 2, \ldots, n, \tag{11.7}$$

where $\beta_1, \beta_2, \ldots, \beta_p$ are unknown parameters. Often, one of the covariates will be the constant 1, that is, the covariate has the value 1 for all i.

Using matrix notation, equation (11.7) can be formulated as $\mathrm{E}\,\boldsymbol{Y} = X\boldsymbol{\beta}$, where $\boldsymbol{Y}$ is the column vector containing $Y_1, Y_2, \ldots, Y_n$, X is a known $n \times p$ matrix (known as the *design matrix*) whose entries are the x_{ij}s, and $\boldsymbol{\beta}$ is the column

vector consisting of the p unknown βs. This could also be written in terms of linear subspaces: $\mathrm{E}\,\boldsymbol{Y} \in L_1$ where $L_1 = \{X\boldsymbol{\beta} \in \mathbb{R}^n : \boldsymbol{\beta} \in \mathbb{R}^p\}$.

The name *linear* regression is due to the fact that $\mathrm{E}\,\boldsymbol{Y}$ is a linear function of $\boldsymbol{\beta}$.

The above model can be generalised in different ways. Instead of assuming the observations to have the same variance, we might assume them to have a variance which is known apart from a constant factor, that is, $\operatorname{Var}\boldsymbol{Y} = \sigma^2\,\Sigma$, where $\Sigma > 0$ is a known matrix and σ^2 an unknown parameter; this is a so-called *weighted* linear regression model

Or we could substitute the normal distribution with another distribution, such as a binomial, Poisson or gamma distribution, and at the same time generalise (11.7) to

$$g(\mathrm{E}\,Y_i) = \sum_{j=1}^{p} x_{ij}\beta_j, \quad i = 1, 2, \ldots, n$$

for a suitable function g; then we get a *generalised* linear regression. Logistic regression (see Section 9.1, in particular page 139ff) is an example of generalised linear regression.

This chapter treats ordinary linear regression only.

Estimating the parameters

According to the general theory the mean vector $X\boldsymbol{\beta}$ is estimated as the orthogonal projection or $\boldsymbol{y}$ onto $L_1 = \{X\boldsymbol{\beta} \in \mathbb{R}^n : \boldsymbol{\beta} \in \mathbb{R}^p\}$. This means that the estimate of $\boldsymbol{\beta}$ is a vector $\widehat{\boldsymbol{\beta}}$ such that $\boldsymbol{y} - X\widehat{\boldsymbol{\beta}} \perp L_1$. Now, $\boldsymbol{y} - X\widehat{\boldsymbol{\beta}} \perp L_1$ is equivalent to $\left\langle \boldsymbol{y} - X\widehat{\boldsymbol{\beta}}, X\boldsymbol{\beta} \right\rangle = 0$ for all $\boldsymbol{\beta} \in \mathbb{R}^p$, which is equivalent to $\left\langle X'\boldsymbol{y} - X'X\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta} \right\rangle = 0$ for all $\boldsymbol{\beta} \in \mathbb{R}^p$, which finally is equivalent to $X'X\widehat{\boldsymbol{\beta}} = X'\boldsymbol{y}$. The matrix equation $X'X\widehat{\boldsymbol{\beta}} = X'\boldsymbol{y}$ can be written out as p linear equations in p unknowns, the *normal equations* (see also page 170). If the $p \times p$ matrix $X'X$ is regular, there is a unique solution, which in matrix notation is

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}\,.$$

(The condition that $X'X$ is regular, can be formulated in several equivalent ways: the dimension of L_1 is p; the rank of X is p; the rank of $X'X$ is p; the columns of X are linearly independent; the parametrisation is injective.)

The variance parameter is estimated as

$$s^2 = \frac{\|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}\|^2}{n - \dim L_1}.$$

Using the rules of calculus for variance matrices we get

$$\begin{aligned}\operatorname{Var}\widehat{\boldsymbol{\beta}} &= \operatorname{Var}\Big((X'X)^{-1}X'\boldsymbol{Y}\Big)\\ &= \Big((X'X)^{-1}X'\Big)\operatorname{Var}\boldsymbol{Y}\Big((X'X)^{-1}X'\Big)'\\ &= \Big((X'X)^{-1}X'\Big)\sigma^2\mathrm{I}\Big((X'X)^{-1}X'\Big)'\\ &= \sigma^2(X'X)^{-1}\end{aligned} \tag{11.8}$$

which is estimated as $s^2(X'X)^{-1}$. The square root of the diagonal elements in this matrix are estimates of the *standard error* (the standard deviation) of the corresponding $\widehat{\beta}$s.

Standard statistical software comes with functions that can solve the normal equations and return the estimated βs and their standard errors (and sometimes even all of the estimated $\operatorname{Var}\widehat{\boldsymbol{\beta}}$)).

Simple linear regression

In simple linear regression (see for example page 109),

$$X = \begin{bmatrix}1 & x_1\\ 1 & x_2\\ \vdots & \\ 1 & x_n\end{bmatrix}, \quad X'X = \begin{bmatrix}n & \sum_{i=1}^n x_i\\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2\end{bmatrix} \quad \text{and} \quad X'\boldsymbol{y} = \begin{bmatrix}\sum_{i=1}^n y_i\\ \sum_{i=1}^n x_i y_i\end{bmatrix},$$

so the normal equations are

$$\begin{aligned}\alpha n + \beta\sum_{i=1}^n x_i &= \sum_{i=1}^n y_i\\ \alpha\sum_{i=1}^n x_i + \beta\sum_{i=1}^n x_i^2 &= \sum_{i=1}^n x_i y_i\end{aligned}$$

which can be solved using standard methods. However, it is probably easier simply to calculate the projection $\mathfrak{p}\boldsymbol{y}$ of $\boldsymbol{y}$ onto L_1. It is easily seen that the two vectors $\boldsymbol{u} = \begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix}$ and $\boldsymbol{v} = \begin{bmatrix}x_1-\overline{x}\\x_2-\overline{x}\\\vdots\\x_n-\overline{x}\end{bmatrix}$ are orthogonal and span L_1, and therefore

$$\mathfrak{p}\boldsymbol{y} = \frac{\langle\boldsymbol{y},\boldsymbol{u}\rangle}{\|\boldsymbol{u}\|^2}\boldsymbol{u} + \frac{\langle\boldsymbol{y},\boldsymbol{v}\rangle}{\|\boldsymbol{v}\|^2}\boldsymbol{v} = \frac{\sum_{i=1}^n y_i}{n}\boldsymbol{u} + \frac{\sum_{i=1}^n (x_i-\overline{x})y_i}{\sum_{i=1}^n (x_i-\overline{x})^2}\boldsymbol{v},$$

which shows that the jth coordinate of $\mathfrak{p}\boldsymbol{y}$ is $(\mathfrak{p}\boldsymbol{y})_j = \overline{y} + \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}(x_j - \overline{x})$, cf. page 122. By calculating the variance matrix (11.8) we obtain the variances and correlations postulated in Proposition 7.3 on page 123.

Testing hypotheses

Hypotheses of the form $H_0 : \mathrm{E}\,\boldsymbol{Y} \in L_0$ where L_0 is a linear subspace of L_1, are tested in the standard way using an F-test.

Usually such a hypothesis will be of the form $H : \beta_j = 0$, which means that the jth covariate x_j is of no significance and hence can be omitted. Such a hypothesis can be tested either using an F test, or using the t test statistic

$$t = \frac{\widehat{\beta_j}}{\text{estimated standard error of } \widehat{\beta_j}}.$$

Factors

There are two different kinds of covariates. The covariates we have met so far all indicate some numerical quantity, that is, they are *quantitative* covariates. But covariates may also be qualitative, thus indicating membership of a certain group as a result of a classification. Such covariates are known as *factors*.

Example: In one-way analysis of variance observations y are classified into a number of groups, see for example page 172; one way to think of the data is as a set of pairs (y, f) of an observation y and a factor f which simply tells the group that the actual y belongs to. This can be formulated as a regression model: Assume there are k levels of f, denoted by $1, 2, \ldots, k$ (i.e. there are k groups), and introduce k artificial (and quantitative) covariates $x_1, x_2, \ldots, x_k$ such that $x_i = 1$ if $f = i$, and $x_i = 0$ otherwise. In this way, we can replace the set of data points (y, f) with "points" $\big(y, (x_1, x_2, \ldots, x_p)\big)$ where all xs except one equal zero, and the single non-zero x (which equals 1) identifies the group that y belongs to. In this notation the one-way analysis of variance model can be written as

$$\mathrm{E}\,Y = \sum_{j=1}^{p} x_j \beta_j$$

where β_j corresponds to μ_j in the original formulation of the model.

Table 11.7
Strangling of dogs: The concentration of hypoxanthine at four different times. In each group the observations are arranged according to size.

time (min)	concentration (μmol/l)						
0	0.0	0.0	1.2	1.8	2.1	2.1	3.0
6	3.0	4.9	5.1	5.1	7.0	7.9	
12	4.9	6.0	6.5	8.0	12.0		
18	9.5	10.1	12.0	12.0	13.0	16.0	17.1

By combining factors and quantitative covariates one can build rather complicated models, e.g. with different levels of groupings, or with different groups having different linear relationships.

An example

Example 11.2: Strangling of dogs
Seven dogs (subjected to anaesthesia) were exposed to hypoxia by compression of the trachea, and the concentration of hypoxantine was measured after 0, 6, 12 and 18 minutes. For various reasons it was not possible to carry out measurements on all of the dogs at all of the four times, and furthermore the measurements no longer can be related to the individual dogs. Table 11.7 shows the available data.

We can think of the data as $n = 25$ pairs of a concentration and a time, the times being know quantities (part of the plan of the experiment), and the concentrations being observed values of random variables: the variation within the four groups can be ascribed to biological variation and measurement errors, and it seems reasonable to model ths variation as a random variation. We will therefore apply a simple linear regression model with concentration as the dependent variable and time as a covariate. Of course, we cannot know in advance whether this model will give an appropriate description of the data. Maybe it will turn out that we can achieve a much better fit using simple linear regression with the square root of the time as covariate, but in that case we would still have a linear regression model, now with the square root of the time as covariate.

We will use a linear regression model with time as the independent variable (x) and the concentration of hypoxantine as the dependent variable (y).

So let x_1, x_2, x_3 and x_4 denote the four times, and let y_{ij} denote the jth concentration mesured at time x_i. With this notation the suggested statistical model is that the observations y_{ij} are observed values of independent random variables Y_{ij}, where Y_{ij} is normal with $\mathrm{E}\, Y_{ij} = \alpha + \beta x_i$ and variance σ^2.

Estimated values of α and β can be found, for example using the formulas on page 122, and we get $\widehat{\beta} = 0.61\ \mu\mathrm{mol\,l^{-1}\,min^{-1}}$ and $\widehat{\alpha} = 1.4\ \mu\mathrm{mol\,l^{-1}}$. The estimated variance is $s_{02}^2 = 4.85\ \mu\mathrm{mol^2\,l^{-2}}$ with $25 - 2 = 23$ degrees of freedom.

The standard error of $\widehat{\beta}$ is $\sqrt{s_{02}^2/SS_x} = 0.06\ \mu\mathrm{mol\,l^{-1}\,min^{-1}}$, and the standard error of $\widehat{\alpha}$ is $\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right)s_{02}^2} = 0.7\ \mu\mathrm{mol\,l^{-1}}$ (Proposition 7.3). The orders of magnitude of these two standard errors suggest that $\widehat{\beta}$ should be written using two decimal places, and $\widehat{\alpha}$

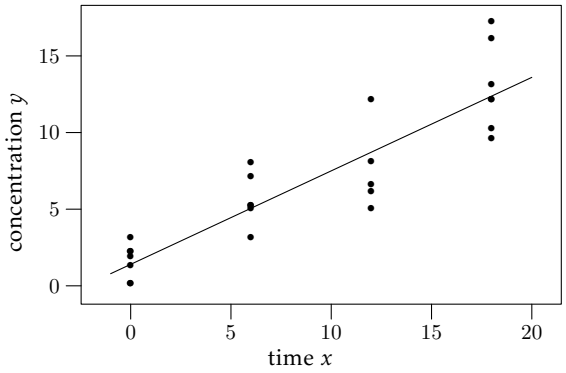


Figure 11.1
Strangling of dogs: The concentration of hypoxantine plotted against time, and the estimated regression line.

using one decimal place, so that the estimated regression line is to be written as

$$y = 1.4\,\mu\text{mol}\,\text{l}^{-1} + 0.61\,\mu\text{mol}\,\text{l}^{-1}\,\text{min}^{-1}\,x\,.$$

As a kind of model validation we will carry out a numerical test of the hypothesis that the concentration in fact does depend on time in a linear (or rather: an affine) way. We therefore embed the regression model into the larger k-sample model with $k = 4$ groups corresponding to the four different values of time, cf. Example 11.1 on page 174.

We will state the method in linear algebra terms. Let $L_1 = \{\boldsymbol{\xi} : \xi_{ij} = \alpha + \beta x_i\}$ be the subspace corresponding to the linear regression model, and let $L = \{\boldsymbol{\xi} : \xi = \mu_i\}$ be the subspace corresponding to the k-sample model. We have $L_1 \subset L$, and from Section 11.1 we know that the test statistic for testing L_1 against L is

$$F = \frac{\frac{1}{\dim L - \dim L_1}\|\mathrm{p}\boldsymbol{y} - \mathrm{p}_1\boldsymbol{y}\|^2}{\frac{1}{n - \dim L}\|\boldsymbol{y} - \mathrm{p}\boldsymbol{y}\|^2}$$

where p and p_1 are the projections onto L and L_1. We know that $\dim L - \dim L_1 = 4 - 2 = 2$ and $n - \dim L = 25 - 4 = 21$. In an earlier treatment of the data (page 174) it was found that $\|\boldsymbol{y} - \mathrm{p}\boldsymbol{y}\|^2$ is 101.32; $\|\mathrm{p}\boldsymbol{y} - \mathrm{p}_1\boldsymbol{y}\|^2$ can be calculated from quantities already known: $\|\boldsymbol{y} - \mathrm{p}_1\boldsymbol{y}\|^2 - \|\boldsymbol{y} - \mathrm{p}\boldsymbol{y}\|^2 = 111.50 - 101.32 = 10.18$. Hence the test statistic is $F = 5.09/4.82 = 1.06$. Using the F distribution with 2 and 21 degrees of freedom, we see that the test probability is above 30%, implying that we may accept the hypothesis, that is, we may assume that there is a linear relation between time and concentration of hypoxantine. This conclusion is confirmed by Figure 11.1.

11.7 Exercises

Exercise 11.1: Peruvian Indians
Changes in somebody's lifestyle may have long-term physiological effects, such as changes in blood pressure.

Anthropologists measured the blood pressure of a number of Peruvian Indians who had been moved from their original primitive environment high in the Andes mountains

y	x_1	x_2	y	x_1	x_2
170	0.048	71.0	114	0.474	59.5
120	0.273	56.5	136	0.289	61.0
125	0.208	56.0	126	0.289	57.0
148	0.042	61.0	124	0.538	57.5
140	0.040	65.0	128	0.615	74.0
106	0.704	62.0	134	0.359	72.0
120	0.179	53.0	112	0.610	62.5
108	0.893	53.0	128	0.780	68.0
124	0.194	65.0	134	0.122	63.4
134	0.406	57.0	128	0.286	68.0
116	0.394	66.5	140	0.581	69.0
114	0.303	59.1	138	0.605	73.0
130	0.441	64.0	118	0.233	64.0
118	0.514	69.5	110	0.432	65.0
138	0.057	64.0	142	0.409	71.0
134	0.333	56.5	134	0.222	60.2
120	0.417	57.0	116	0.021	55.0
120	0.432	55.0	132	0.860	70.0
114	0.459	57.0	152	0.741	87.0
124	0.263	58.0			

Table 11.8
Peruvian Indians: Values of y: systolic blood pressure (mm Hg), x_1: fraction of lifetime in urban area, and x_2: weight (kg).

into the so-called civilisation, the modern city life, which also is at a much lower altitude (Davin (1975), here from Ryan et al. (1976)). The anthropologists examined a sample consisting of 39 males over 21 years, who had been subjected to such a migration into civilisation. On each person the systolic and diastolic blood pressure was measured, as well as a number of additional variables such as age, number of years since migration, height, weight and pulse. Besides, yet another variable was calculated, the "fraction of lifetime in urban area", i.e. the number of years since migration divided by current age.

This exercise treats only part of the data set, namely the systolic *blood pressure* which will be used as the dependent variable (y), and the *fraction of lifetime in urban area* and *weight* which will be used as independent variables (x_1 and x_2). The values of y, x_1 and x_2 are given in Table 11.1 (from Ryan et al. (1976)).

1. The anthropologists believed x_1, the fraction of lifetime in urban area, to be a good indicator of how long the person had been living in the urban area, so it would be interesting to see whether x_1 could indeed explain the variation in blood pressure y. The first thing to do would therefore be to fit a simple linear regression model with x_1 as the independent variable. Do that!
2. From a plot of y against x_1, however, it is not particularly obvious that there should be a linear relationship between y and x_1, so maybe we should include one or two of the remaining explanatory variables.
   Since a person's weight is known to influence the blood pressure, we might try a regression model with both x_1 and x_2 as explanatory variables.

a) Estimate the parameters in this model. What happens to the variance estimate?
b) Examine the residuals in order to assess the quality of the model.

3. How is the final model to be interpreted relative to the Peruvian Indians?

Exercise 11.2
This is a renowned two-way analysis of variance problem from University of Copenhagen.

Five days a week, a student rides his bike from his home to the *H.C. Ørsted Institutet* [the building where the Mathematics department at University of Copenhagen is situated] in the morning, and back again in the afternoon. He can chose between two different routes, one which he usually uses, and another one which he thinks might be a shortcut. To find out whether the second route really is a shortcut, he sometimes measures the time used for his journey. His results are given in the table below, which shows the travel times minus 9 minutes, in units of 10 seconds.

	shortcut			the usual route				
morning	4	−1	3	13	8	11	5	7
afternoon	11	6		16	18	17	21	19

As we all know [!] it can be difficult to get away from the H.C. Ørsted Institutet by bike, so the afternoon journey will on the average take a little longer than the morning journey.

Is the shortcut in fact the shorter route?

Hint: This is obviously a two-sample problem, but the model is not a balanced one, and we cannot apply the ususal formula for the projection onto the subspace corresponding to additivity.

Exercise 11.3
Consider the simple linear regression model $\mathrm{E}\,Y = \alpha + x\beta$ for a set of data points (y_i, x_i), $i = 1, 2, \ldots, n$.

What does the design matrix look like? Write down the normal equations, and solve them.

Find formulas for the standard errors (i.e. the standard deviations) of $\widehat{\alpha}$ and $\widehat{\beta}$, and a formula for the correlation between the two estimators. Hint: use formula (11.8) on page 190.

In some situations the designer of the experiment can, within limits, decide the x values. What would be an optimal choice?

A A Derivation of the Normal Distribution

The normal distribution is used very frequently in statistical models. Sometimes you can argue for using it by referring to the Central Limit Theorem, and sometimes it seems to be used mostly for the sake of mathematical convenience (or by habit). In certain modelling situations, however, you may also make arguments of the form "if you are going to analyse data in such-and-such a way, then you are implicitly assuming such-and-such a distribution". We shall now present an argumentation of this kind leading to the normal distribution; the main idea stems from Gauss (1809, book II, section 3).

Suppose that we want to find a type of probability distributions that may be used to describe the random scattering of observations around some unknown value μ. We make a number of assumptions:

Assumption 1: The distributions are continuous distributions on $\mathbb{R}$, i.e. they possess a continuous density function.

Assumption 2: The parameter μ may be any real number, and μ is a *location parameter*; hence the model function is of the form

$$f(x-\mu), \quad (x,\mu) \in \mathbb{R}^2,$$

for a suitable function f defined on all of $\mathbb{R}$. By assumption 1 f must be continuous.

Assumption 3: The function f has a continuous derivative.

Assumption 4: The maximum likelihood estimate of μ is the arithmetic mean of the observations, that is, for any sample $x_1, x_2, \ldots, x_n$ from the distribution with density function $x \mapsto f(x-\mu)$, the maximum likelihood estimate must be the arithmetic mean $\overline{x}$.

Now we can make a number of deductions:

The log-likelihood function corresponding to the observations $x_1, x_2, \ldots, x_n$ is

$$\ln L(\mu) = \sum_{i=1}^{n} \ln f(x_i - \mu).$$

This function is differentiable with

$$(\ln L)'(\mu) = \sum_{i=1}^{n} g(x_i - \mu)$$

where $g = -(\ln f)'$.

Since f is a probability density function, we have $\liminf_{x\to\pm\infty} f(x) = 0$, and so $\liminf_{x\to\pm\infty} \ln L(\mu) = -\infty$; this means that $\ln L$ attains its maximum value at (at least) one point $\widehat{\mu}$, and that this is a stationary point, i.e. $(\ln L)'(\widehat{\mu}) = 0$. Since we require that $\widehat{\mu} = \overline{x}$, we may conclude that

$$\sum_{i=1}^{n} g(x_i - \overline{x}) = 0 \tag{A.1}$$

for all samples $x_1, x_2, \ldots, x_n$.

Conside a sample with two elements x and $-x$; the mean of this sample is 0, so equation (A.1) becomes $g(x) + g(-x) = 0$. Hence $g(-x) = -g(x)$ for all x. In particular, $g(0) = 0$.

Next, consider a sample consisting of the three elements x, y and $-(x+y)$; again the sample mean is 0, and equation (A.1) now becomes $g(x)+g(y)+g(-(x+y)) = 0$; we just showed that g is an odd function, so it follows that $g(x+y) = g(x) + g(y)$, and this is true for all x and y. From Proposition A.1 below, such a funtion must be of the form $g(x) = cx$ for some constant c. Per definition, $g = -(\ln f)'$, so $\ln f(x) = b - \frac{1}{2}cx^2$ for some constant of integration b, and $f(x) = a\exp(-\frac{1}{2}cx^2)$ where $a = e^b$. Since f has to be non-negative and integrate to 1, the constant c is positive and we can rename it to $1/\sigma^2$, and the constant a is uniquely determined (in fact, $a = 1/\sqrt{2\pi\sigma^2}$, cf. page 73). The function f is therefore necessarily the density function of the normal distribution with parameters 0 and σ^2, and the type of distributions searched for is the normal distributions with parameters μ and σ^2 where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.

Exercise A.1

The above investigation deals with a situation with observations from a continuous distribution on the real line, and such that the maximum likelihood estimator of the *location* parameter is equal to the arithmetic mean of the observations.

Now assume that we instead had observations from a continuous distribution on the set of *positive* real numbers, that this distribution is parametrised by a *scale* parameter, and that the maximum likelihood estimator of the *scale* parameter is equal to the arithmetic mean of the observations. — What can then be said about the distribuion of the observations?

Cauchy's functional equation

Here we prove a general result which was used above, and which certainly should be part of general mathematical education.

Proposition A.1
Let f be a real-valued function such that

$$f(x+y) = f(x) + f(y) \tag{A.2}$$

for all $x, y \in \mathbb{R}$. If f is known to be continuous at a point $x_0 \neq 0$, then for some real number c, $f(x) = cx$ for all $x \in \mathbb{R}$.

Proof
First let us deduce some consequences of the requirement $f(x+y) = f(x) + f(y)$.

1. Letting $x = y = 0$ gives $f(0) = f(0) + f(0)$, i.e. $f(0) = 0$.
2. Letting $y = -x$ shows that $f(x) + f(-x) = f(0) = 0$, that is, $f(-x) = -f(x)$ for all $x \in \mathbb{R}$.
3. Repeated applications of (A.2) gives that for any real numbers $x_1, x_2, \ldots, x_n$,

$$f(x_1 + x_2 + \ldots + x_n) = f(x_1) + f(x_2) + \ldots + f(x_n). \tag{A.3}$$

Consider now a real number $x \neq 0$ and positive integers p and q, and let $a = p/q$. Since

$$f(\underbrace{ax + ax + \ldots + ax}_{q \text{ tems}}) = f(\underbrace{x + x + \ldots + x}_{p \text{ terms}}),$$

equation (A.3) gives that $qf(ax) = pf(x)$ or $f(ax) = af(x)$. We may conclude that

$$f(ax) = af(x). \tag{A.4}$$

for all rational numbers a and all real numbers x

Until now we did not make use of the continuity of f at x_0, but we shall do that now. Let x be a non-zero real number. There exists a sequence (a_n) of non-zero rational numbers converging to x_0/x. Then the sequence $(a_n x)$ converges to x_0, and since f is continuous at this point,

$$\frac{f(a_n x)}{a_n} \longrightarrow \frac{f(x_0)}{x_0/x} = \frac{f(x_0)}{x_0}\, x.$$

Now, according to equation (A.4), $\dfrac{f(a_n x)}{a_n} = f(x)$ for all n, and hence $f(x) = \dfrac{f(x_0)}{x_0}\, x$, that is, $f(x) = c\,x$ where $c = \dfrac{f(x_0)}{x_0}$. Since x was arbitrary (and c does not depend on x), the proof is finished. □

Remark: We can easily give examples of functions satisfying (A.2) and with no points of continuity (except $x = 0$); here is one:

$$f(x) = \begin{cases} x & \text{if } x \text{ is rational,} \\ 2x & \text{if } x/\sqrt{2} \text{ is rational,} \\ 0 & \text{otherwise.} \end{cases}$$

B Some Results from Linear Algebra

These pages present a few results from linear algebra, more precisely from the theory of finite-dimensional real vector spaces with an inner product.

Notation and definitions

The generic *vector space* is denoted V. *Subspaces*, i.e. linear subspaces, are denoted $L, L_1, L_2, \ldots$ *Vectors* are usually denoted by boldface letters ($\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u}, \boldsymbol{v}$ etc.). The zero vector is $\mathbf{0}$. The set of coordinates of a vector relative to a given basis is written as a *column matrix*.

V
L $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}$
$\mathbf{0}$

Linear transformations [and their matrices] are often denoted by letters as A and B; the *transposed* of A is denoted A'. The *null space* (or kernel) of A is denoted $\mathcal{N}(A)$, and the *range* is denoted $\mathcal{R}(A)$; the *rank* of A is the number $\dim \mathcal{R}(A)$. The identity transformation [the unit matrix] is denoted I.

A A'
$\mathcal{N}(A)$ $\mathcal{R}(A)$
I

The *inner product* of $\boldsymbol{u}$ and $\boldsymbol{v}$ is written as $\langle \boldsymbol{u}, \boldsymbol{v} \rangle$ and in matrix notation $\boldsymbol{u}'\boldsymbol{v}$. The *length* of $\boldsymbol{u}$ is $\|\boldsymbol{u}\|$.

$\langle \boldsymbol{u}, \boldsymbol{v} \rangle$ $\boldsymbol{u}'\boldsymbol{v}$
$\|\boldsymbol{u}\|$

Two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are *orthogonal*, $\boldsymbol{u} \perp \boldsymbol{v}$, if $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = 0$. Two subspaces L_1 and L_2 are orthogonal, $L_1 \perp L_2$, if every vector in L_1 is orthogonal to every vector in L_2, and if so, the orthogonal direct sum of L_1 and L_2 is defined as the subspace

$\boldsymbol{u} \perp \boldsymbol{v}$
$L_1 \perp L_2$
$L_1 \oplus L_2$

$$L_1 \oplus L_2 = \{\boldsymbol{u}_1 + \boldsymbol{u}_2 : \boldsymbol{u}_1 \in L_1, \boldsymbol{u}_2 \in L_2\}.$$

If $\boldsymbol{v} \in L_1 \oplus L_2$, then $\boldsymbol{v}$ has a unique decomposition as $\boldsymbol{v} = \boldsymbol{u}_1 + \boldsymbol{u}_2$ where $\boldsymbol{u}_1 \in L_1$ and $\boldsymbol{u}_2 \in L_2$. The dimension of the space is $\dim L_1 \oplus L_2 = \dim L_1 + \dim L_2$.

The *orthogonal complement* of the subspace L is denoted $L^\perp$. We have $V = L \oplus L^\perp$. The *orthogonal projection* of V onto the subspace L is the linear transformation $\mathfrak{p} : V \to V$ such that $\mathfrak{p}\boldsymbol{x} \in L$ and $\boldsymbol{x} - \mathfrak{p}\boldsymbol{x} \in L^\perp$ for all $\boldsymbol{x} \in V$.

$L^\perp$
$\mathfrak{p}$

A symmetric linear transformation A of V [a symmetric matrix A] is *positive semidefinite* if $\langle \boldsymbol{x}, A\boldsymbol{x} \rangle \geq 0$ [or $\boldsymbol{x}'A\boldsymbol{x} \geq 0$] for all $\boldsymbol{x}$; it is *positive definite* if there is a strict inequality for all $\boldsymbol{x} \neq \mathbf{0}$. Sometimes we write $A \geq 0$ or $A > 0$ to tell that A is positive semidefinite or positive definite.

$A \geq 0$ $A > 0$

Various results

This proposition is probably well-known from linear algebra:

Proposition B.1
Let A be a linear transformation of $\mathbb{R}^p$ into $\mathbb{R}^n$. Then $\mathcal{R}(A)$ is the orthogonal complement of $\mathcal{N}(A')$ (in $\mathbb{R}^n$), and $\mathcal{N}(A')$ is the orthogonal complement of $\mathcal{R}(A)$ (in $\mathbb{R}^n$), in brief, $\mathbb{R}^n = \mathcal{R}(A) \oplus \mathcal{N}(A')$.

Corollary B.2
Let A be a symmetric linear transformation of $\mathbb{R}^n$. Then $\mathcal{R}(A)$ and $\mathcal{N}(A)$ are orthogonal complements of one another: $\mathbb{R}^n = \mathcal{R}(A) \oplus \mathcal{N}(A)$.

Corollary B.3
$\mathcal{R}(A) = \mathcal{R}(AA')$.

Proof of Corollary B.3
According to Proposition B.1 it suffices to show that $\mathcal{N}(A') = \mathcal{N}(AA')$, that is, to show that $A'\boldsymbol{u} = \mathbf{0} \Leftrightarrow AA'\boldsymbol{u} = \mathbf{0}$.

Clearly $A'\mathbf{0} \Rightarrow AA'\boldsymbol{u} = \mathbf{0}$, so it remains to show that $AA'\boldsymbol{u} = \mathbf{0} \Rightarrow A'\boldsymbol{u} = \mathbf{0}$. Suppose that $AA'\boldsymbol{u} = \mathbf{0}$. Since $A(A'\boldsymbol{u}) = \mathbf{0}$, $A'\boldsymbol{u} \in \mathcal{N}(A) = \mathcal{R}(A')^\perp$, and since we always have $A'\boldsymbol{u} \in \mathcal{R}(A')$, we have $A'\boldsymbol{u} \in \mathcal{R}(A')^\perp \cap \mathcal{R}(A') = \{\mathbf{0}\}$. □

Proposition B.4
Suppose that A and B are injective linear transformations of $\mathbb{R}^p$ into $\mathbb{R}^n$ such that $AA' = BB'$. Then there exists an isometry C of $\mathbb{R}^p$ onto itself such that $A = BC$.

Proof
Let $L = \mathcal{R}(A)$. From the hypothesis and Corollary B.3 we have $L = \mathcal{R}(A) = \mathcal{R}(AA') = \mathcal{R}(BB') = \mathcal{R}(B)$.

Since A is injective, $\mathcal{N}(A) = \{\mathbf{0}\}$; Proposition B.1 then gives $\mathcal{R}(A') = \{\mathbf{0}\}^\perp = \mathbb{R}^p$, that is, for each $\boldsymbol{u} \in \mathbb{R}^p$ the equation $A'\boldsymbol{x} = \boldsymbol{u}$ has at least one solution $\boldsymbol{x} \in \mathbb{R}^n$, and since $\mathcal{N}(A') = \mathcal{R}(A)^\perp = L^\perp$, there will be exactly one solution $\boldsymbol{x}(\boldsymbol{u})$ in L to $A'\boldsymbol{x} = \boldsymbol{u}$.

We will show that this mapping $\boldsymbol{u} \mapsto \boldsymbol{x}(\boldsymbol{u})$ of $\mathbb{R}^p$ into L is linear. Let $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ be two vectors in $\mathbb{R}^p$, and let α_1 and α_2 be two scalars. Then

$$\begin{aligned} A'\big(\boldsymbol{x}(\alpha_1\boldsymbol{u}_1 + \alpha_2\boldsymbol{u}_2)\big) &= \alpha_1\boldsymbol{u}_1 + \alpha_2\boldsymbol{u}_2 \\ &= \alpha_1 A'\boldsymbol{x}(\boldsymbol{u}_1) + \alpha_2 A'\boldsymbol{x}(\boldsymbol{u}_1) \\ &= A'\big(\alpha_1\boldsymbol{x}(\boldsymbol{u}_1) + \alpha_2\boldsymbol{x}(\boldsymbol{u}_2)\big), \end{aligned}$$

and since the equation $A'\boldsymbol{x} = \boldsymbol{u}$ as mentioned has a unique solution $\boldsymbol{x} \in L$, this implies that $\boldsymbol{x}(\alpha_1\boldsymbol{u}_1 + \alpha_2\boldsymbol{u}_2) = \alpha_1\boldsymbol{x}(\boldsymbol{u}_1) + \alpha_2\boldsymbol{x}(\boldsymbol{u}_2)$. Hence $\boldsymbol{u} \mapsto \boldsymbol{x}(\boldsymbol{u})$ is linear.

Now define the linear transformation $C : \mathbb{R}^p \to \mathbb{R}^p$ as $C\boldsymbol{u} = B'\boldsymbol{x}(\boldsymbol{u})$, $\boldsymbol{u} \in \mathbb{R}^p$. Then $BC\boldsymbol{u} = BB'\boldsymbol{x}(\boldsymbol{u}) = AA'\boldsymbol{x}(\boldsymbol{u}) = A\boldsymbol{u}$ for all $\boldsymbol{u} \in \mathbb{R}^p$, showing that $A = BC$.

Finally let us show that C preserves inner products (an hence is an isometry): For $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathbb{R}^p$ we have $\langle C\boldsymbol{u}_1, C\boldsymbol{u}_2\rangle = \langle B'\boldsymbol{x}(\boldsymbol{u}_1), B'\boldsymbol{x}(\boldsymbol{u}_2)\rangle = \langle \boldsymbol{x}(\boldsymbol{u}_1), BB'\boldsymbol{x}(\boldsymbol{u}_2)\rangle = \langle \boldsymbol{x}(\boldsymbol{u}_1), AA'\boldsymbol{x}(\boldsymbol{u}_2)\rangle = \langle A'\boldsymbol{x}(\boldsymbol{u}_1), A'\boldsymbol{x}(\boldsymbol{u}_2)\rangle = \langle \boldsymbol{u}_1, \boldsymbol{u}_2\rangle$. □

Proposition B.5
Let A be a symmetric and positive semidefinite linear transformation of $\mathbb{R}^n$ into itself. Then there exists a unique symmetric and positive semidefinite linear transformation $A^{1/2}$ of $\mathbb{R}^n$ into itself such that $A^{1/2}A^{1/2} = A$.

Proof
Let $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ be an orthonormal basis of eigenvectors of A, and let the corresponding eigenvalues be $\lambda_1, \lambda_2, \ldots, \lambda_n \geq 0$. If we define $A^{1/2}$ to be the linear transformation that takes basis vector $\boldsymbol{e}_j$ into i $\lambda_j^{1/2}\boldsymbol{e}_j$, $j = 1, 2, \ldots, n$, then we have one mapping with the requested properties.

Now suppose that $A^\star$ is any other solution, that is, $A^\star$ is symmetric and positive semidefinite, and $A^\star A^\star = A$. There exists an orthonormal basis $\boldsymbol{e}_1^\star, \boldsymbol{e}_2^\star, \ldots, \boldsymbol{e}_n^\star$ of eigenvectors of $A^\star$; let the corresponding eigenvalues be $\mu_1, \mu_2, \ldots, \mu_n \geq 0$. Since $A\boldsymbol{e}_j^\star = A^\star A^\star \boldsymbol{e}_j^\star = \mu_j^2 \boldsymbol{e}_j^\star$, we see that $\boldsymbol{e}_j^\star$ is an A-eigenvector with eigenvalue μ_j^2. Hence $A^\star$ and A have the same eigenspaces, and for each eigenspace the $A^\star$-eigenvalue is the non-negative squareroot of the A-eigenvalue, and so $A^\star = A^{1/2}$. □

Proposition B.6
Assume that A is a symmetric and positive semidefinite linear transformation of $\mathbb{R}^n$ into itself, and let $p = \dim \mathcal{R}(A)$.

Then there exists an injective linear transformation B of $\mathbb{R}^p$ into $\mathbb{R}^n$ such that $BB' = A$. B is unique up to isometry.

Proof
Let $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ be an orthonormal basis of eigenvectors of A, with the corresponding eigenvectors $\lambda_1, \lambda_2, \ldots, \lambda_n$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ and $\lambda_{p+1} = \lambda_{p+2} = \ldots = \lambda_n = 0$.

As B we can use the linear transformation whose matrix representation relative to the standard basis of $\mathbb{R}^p$ and the basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ of $\mathbb{R}^n$ is the $n \times p$-matrix whose (i, j)th entry is $\lambda_i^{1/2}$ if hvis $i = j \leq p$ and 0 otherwise. The uniqueness modulo isometry follows from Proposition B.4. □

C Statistical Tables

In the pre-computer era a collection of statistical tables was an indispensable tool for the working statistician. The following pages contain some examples of such statistical tables, namely tables of quantiles of a few distributions occuring in hypothesis testing.

Recall that for a probability distribution with a strictly increasing and continuous distribution function F, the *α-quantile* x_α is defined as the solution of the equation $F(x) = \alpha$; here $0 < \alpha < 1$. (If F is not known to be strictly increasing and continuous, we may define *the set of α-quantiles* as the closed interval whose endpoints are $\sup\{x : F(x) < \alpha\}$ and $\inf\{x : F(x) > \alpha\}$.)

The distribution function $\Phi(u)$ of the standard normal distribution

u	$u+5$	$\Phi(u)$	u	$u+5$	$\Phi(u)$
			0.00	5.00	0.5000
−3.75	1.25	0.0001	0.10	5.10	0.5398
−3.50	1.50	0.0002	0.20	5.20	0.5793
−3.25	1.75	0.0006	0.30	5.30	0.6179
−3.00	2.00	0.0013	0.40	5.40	0.6554
−2.80	2.20	0.0026	0.50	5.50	0.6915
−2.60	2.40	0.0047	0.60	5.60	0.7257
−2.40	2.60	0.0082	0.70	5.70	0.7580
−2.20	2.80	0.0139	0.80	5.80	0.7881
−2.00	3.00	0.0228	0.90	5.90	0.8159
−1.90	3.10	0.0287	1.00	6.00	0.8413
−1.80	3.20	0.0359	1.10	6.10	0.8643
−1.70	3.30	0.0446	1.20	6.20	0.8849
−1.60	3.40	0.0548	1.30	6.30	0.9032
−1.50	3.50	0.0668	1.40	6.40	0.9192
−1.40	3.60	0.0808	1.50	6.50	0.9332
−1.30	3.70	0.0968	1.60	6.60	0.9452
−1.20	3.80	0.1151	1.70	6.70	0.9554
−1.10	3.90	0.1357	1.80	6.80	0.9641
−1.00	4.00	0.1587	1.90	6.90	0.9713
−0.90	4.10	0.1841	2.00	7.00	0.9772
−0.80	4.20	0.2119	2.20	7.20	0.9861
−0.70	4.30	0.2420	2.40	7.40	0.9918
−0.60	4.40	0.2743	2.60	7.60	0.9953
−0.50	4.50	0.3085	2.80	7.80	0.9974
−0.40	4.60	0.3446	3.00	8.00	0.9987
−0.30	4.70	0.3821	3.25	8.25	0.9994
−0.20	4.80	0.4207	3.50	8.50	0.9998
−0.10	4.90	0.4602	3.75	8.75	0.9999

Quantiles of the standard normal distribution

α	$u_\alpha = \Phi^{-1}(\alpha)$	$u_\alpha + 5$
0.001	−3.090	1.910
0.002	−2.878	2.122
0.005	−2.576	2.424
0.010	−2.326	2.674
0.015	−2.170	2.830
0.020	−2.054	2.946
0.025	−1.960	3.040
0.030	−1.881	3.119
0.035	−1.812	3.188
0.040	−1.751	3.249
0.045	−1.695	3.305
0.050	−1.645	3.355
0.055	−1.598	3.402
0.060	−1.555	3.445
0.070	−1.476	3.524
0.080	−1.405	3.595
0.090	−1.341	3.659
0.100	−1.282	3.718
0.125	−1.150	3.850
0.150	−1.036	3.964
0.175	−0.935	4.065
0.200	−0.842	4.158
0.220	−0.772	4.228
0.240	−0.706	4.294
0.250	−0.674	4.326
0.260	−0.643	4.357
0.280	−0.583	4.417
0.300	−0.524	4.476
0.320	−0.468	4.532
0.340	−0.412	4.588
0.360	−0.358	4.642
0.380	−0.305	4.695
0.400	−0.253	4.747
0.420	−0.202	4.798
0.440	−0.151	4.849
0.460	−0.100	4.900
0.480	−0.050	4.950
0.500	0.000	5.000
0.520	0.050	5.050
0.540	0.100	5.100
0.560	0.151	5.151
0.580	0.202	5.202
0.600	0.253	5.253
0.620	0.305	5.305
0.640	0.358	5.358
0.660	0.412	5.412
0.680	0.468	5.468
0.700	0.524	5.524
0.720	0.583	5.583
0.740	0.643	5.643
0.750	0.674	5.674
0.760	0.706	5.706
0.780	0.772	5.772
0.800	0.842	5.842
0.825	0.935	5.935
0.850	1.036	6.036
0.875	1.150	6.150
0.900	1.282	6.282
0.910	1.341	6.341
0.920	1.405	6.405
0.930	1.476	6.476
0.940	1.555	6.555
0.945	1.598	6.598
0.950	1.645	6.645
0.955	1.695	6.695
0.960	1.751	6.751
0.965	1.812	6.812
0.970	1.881	6.881
0.975	1.960	6.960
0.980	2.054	7.054
0.985	2.170	7.170
0.990	2.326	7.326
0.995	2.576	7.576
0.998	2.878	7.878
0.999	3.090	8.090

Quantiles of the χ^2 distribution with f degrees of freedom

	Probability in percent						
f	50	90	95	97.5	99	99.5	99.9
1	0.45	2.71	3.84	5.02	6.63	7.88	10.83
2	1.39	4.61	5.99	7.38	9.21	10.60	13.82
3	2.37	6.25	7.81	9.35	11.34	12.84	16.27
4	3.36	7.78	9.49	11.14	13.28	14.86	18.47
5	4.35	9.24	11.07	12.83	15.09	16.75	20.52
6	5.35	10.64	12.59	14.45	16.81	18.55	22.46
7	6.35	12.02	14.07	16.01	18.48	20.28	24.32
8	7.34	13.36	15.51	17.53	20.09	21.95	26.12
9	8.34	14.68	16.92	19.02	21.67	23.59	27.88
10	9.34	15.99	18.31	20.48	23.21	25.19	29.59
11	10.34	17.28	19.68	21.92	24.72	26.76	31.26
12	11.34	18.55	21.03	23.34	26.22	28.30	32.91
13	12.34	19.81	22.36	24.74	27.69	29.82	34.53
14	13.34	21.06	23.68	26.12	29.14	31.32	36.12
15	14.34	22.31	25.00	27.49	30.58	32.80	37.70
16	15.34	23.54	26.30	28.85	32.00	34.27	39.25
17	16.34	24.77	27.59	30.19	33.41	35.72	40.79
18	17.34	25.99	28.87	31.53	34.81	37.16	42.31
19	18.34	27.20	30.14	32.85	36.19	38.58	43.82
20	19.34	28.41	31.41	34.17	37.57	40.00	45.31
21	20.34	29.62	32.67	35.48	38.93	41.40	46.80
22	21.34	30.81	33.92	36.78	40.29	42.80	48.27
23	22.34	32.01	35.17	38.08	41.64	44.18	49.73
24	23.34	33.20	36.42	39.36	42.98	45.56	51.18
25	24.34	34.38	37.65	40.65	44.31	46.93	52.62
26	25.34	35.56	38.89	41.92	45.64	48.29	54.05
27	26.34	36.74	40.11	43.19	46.96	49.64	55.48
28	27.34	37.92	41.34	44.46	48.28	50.99	56.89
29	28.34	39.09	42.56	45.72	49.59	52.34	58.30
30	29.34	40.26	43.77	46.98	50.89	53.67	59.70
31	30.34	41.42	44.99	48.23	52.19	55.00	61.10
32	31.34	42.58	46.19	49.48	53.49	56.33	62.49
33	32.34	43.75	47.40	50.73	54.78	57.65	63.87
34	33.34	44.90	48.60	51.97	56.06	58.96	65.25
35	34.34	46.06	49.80	53.20	57.34	60.27	66.62
36	35.34	47.21	51.00	54.44	58.62	61.58	67.99
37	36.34	48.36	52.19	55.67	59.89	62.88	69.35
38	37.34	49.51	53.38	56.90	61.16	64.18	70.70
39	38.34	50.66	54.57	58.12	62.43	65.48	72.05
40	39.34	51.81	55.76	59.34	63.69	66.77	73.40

Quantiles of the χ^2 distribution with f degrees of freedom

	Probability in percent						
f	50	90	95	97.5	99	99.5	99.9
41	40.34	52.95	56.94	60.56	64.95	68.05	74.74
42	41.34	54.09	58.12	61.78	66.21	69.34	76.08
43	42.34	55.23	59.30	62.99	67.46	70.62	77.42
44	43.34	56.37	60.48	64.20	68.71	71.89	78.75
45	44.34	57.51	61.66	65.41	69.96	73.17	80.08
46	45.34	58.64	62.83	66.62	71.20	74.44	81.40
47	46.34	59.77	64.00	67.82	72.44	75.70	82.72
48	47.34	60.91	65.17	69.02	73.68	76.97	84.04
49	48.33	62.04	66.34	70.22	74.92	78.23	85.35
50	49.33	63.17	67.50	71.42	76.15	79.49	86.66
51	50.33	64.30	68.67	72.62	77.39	80.75	87.97
52	51.33	65.42	69.83	73.81	78.62	82.00	89.27
53	52.33	66.55	70.99	75.00	79.84	83.25	90.57
54	53.33	67.67	72.15	76.19	81.07	84.50	91.87
55	54.33	68.80	73.31	77.38	82.29	85.75	93.17
56	55.33	69.92	74.47	78.57	83.51	86.99	94.46
57	56.33	71.04	75.62	79.75	84.73	88.24	95.75
58	57.33	72.16	76.78	80.94	85.95	89.48	97.04
59	58.33	73.28	77.93	82.12	87.17	90.72	98.32
60	59.33	74.40	79.08	83.30	88.38	91.95	99.61
61	60.33	75.51	80.23	84.48	89.59	93.19	100.89
62	61.33	76.63	81.38	85.65	90.80	94.42	102.17
63	62.33	77.75	82.53	86.83	92.01	95.65	103.44
64	63.33	78.86	83.68	88.00	93.22	96.88	104.72
65	64.33	79.97	84.82	89.18	94.42	98.11	105.99
66	65.33	81.09	85.96	90.35	95.63	99.33	107.26
67	66.33	82.20	87.11	91.52	96.83	100.55	108.53
68	67.33	83.31	88.25	92.69	98.03	101.78	109.79
69	68.33	84.42	89.39	93.86	99.23	103.00	111.06
70	69.33	85.53	90.53	95.02	100.43	104.21	112.32
71	70.33	86.64	91.67	96.19	101.62	105.43	113.58
72	71.33	87.74	92.81	97.35	102.82	106.65	114.84
73	72.33	88.85	93.95	98.52	104.01	107.86	116.09
74	73.33	89.96	95.08	99.68	105.20	109.07	117.35
75	74.33	91.06	96.22	100.84	106.39	110.29	118.60
76	75.33	92.17	97.35	102.00	107.58	111.50	119.85
77	76.33	93.27	98.48	103.16	108.77	112.70	121.10
78	77.33	94.37	99.62	104.32	109.96	113.91	122.35
79	78.33	95.48	100.75	105.47	111.14	115.12	123.59
80	79.33	96.58	101.88	106.63	112.33	116.32	124.84

90% quantiles of the F distribution

f_1 is the number of degrees of freedom of the nominator,
f_2 is the number of degrees of freedom of the denominator.

	f_1										
f_2	1	2	3	4	5	6	7	8	9	10	15
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	61.22
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.42
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.20
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.87
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.24
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.87
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.63
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.46
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.34
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.24
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.17
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.10
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.05
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.01
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	1.97
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.94
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.91
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.89
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.86
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.84
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.81
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.78
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.76
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.74
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.72
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.66
50	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73	1.63
75	2.77	2.37	2.16	2.02	1.93	1.85	1.80	1.75	1.72	1.69	1.58
100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69	1.66	1.56
150	2.74	2.34	2.12	1.98	1.89	1.81	1.76	1.71	1.67	1.64	1.53
200	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66	1.63	1.52
300	2.72	2.32	2.10	1.96	1.87	1.79	1.74	1.69	1.65	1.62	1.51
400	2.72	2.32	2.10	1.96	1.86	1.79	1.73	1.69	1.65	1.61	1.50
500	2.72	2.31	2.09	1.96	1.86	1.79	1.73	1.68	1.64	1.61	1.50

95% quantiles of the F distribution

f_1 is the number of degrees of freedom of the nominator,
f_2 is the number of degrees of freedom of the denominator.

	f_1										
f_2	1	2	3	4	5	6	7	8	9	10	15
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	245.95
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.11
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.07
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.04
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87
75	3.97	3.12	2.73	2.49	2.34	2.22	2.13	2.06	2.01	1.96	1.80
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.77
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.73
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.72
300	3.87	3.03	2.63	2.40	2.24	2.13	2.04	1.97	1.91	1.86	1.70
400	3.86	3.02	2.63	2.39	2.24	2.12	2.03	1.96	1.90	1.85	1.69
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.69

97.5% quantiles of the F distribution

f_1 is the number of degrees of freedom of the nominator,
f_2 is the number of degrees of freedom of the denominator.

	f_1										
f_2	1	2	3	4	5	6	7	8	9	10	15
1	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28	968.63	984.87
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.43
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.25
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.66
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.43
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.27
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.57
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.10
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.77
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.52
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.33
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.18
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.05
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	2.95
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.86
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.79
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.72
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.67
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.62
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.57
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.50
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.44
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.39
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.34
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.31
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.18
50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.11
75	5.23	3.88	3.30	2.96	2.74	2.58	2.46	2.37	2.29	2.22	2.01
100	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	1.97
150	5.13	3.78	3.20	2.87	2.65	2.49	2.37	2.28	2.20	2.13	1.92
200	5.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18	2.11	1.90
300	5.07	3.73	3.16	2.83	2.61	2.45	2.33	2.23	2.16	2.09	1.88
400	5.06	3.72	3.15	2.82	2.60	2.44	2.32	2.22	2.15	2.08	1.87
500	5.05	3.72	3.14	2.81	2.59	2.43	2.31	2.22	2.14	2.07	1.86

99% quantiles of the F distribution

f_1 is the number of degrees of freedom of the nominator,
f_2 is the number of degrees of freedom of the denominator.

	f_1										
f_2	1	2	3	4	5	6	7	8	9	10	15
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6157.28
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.43
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	26.87
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.20
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.72
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.56
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.31
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.52
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	4.96
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.56
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.25
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.01
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.82
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.66
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.52
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.41
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.31
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.23
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.15
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.09
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	2.98
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	2.89
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.81
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.75
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.70
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.52
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.42
75	6.99	4.90	4.05	3.58	3.27	3.05	2.89	2.76	2.65	2.57	2.29
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.22
150	6.81	4.75	3.91	3.45	3.14	2.92	2.76	2.63	2.53	2.44	2.16
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.13
300	6.72	4.68	3.85	3.38	3.08	2.86	2.70	2.57	2.47	2.38	2.10
400	6.70	4.66	3.83	3.37	3.06	2.85	2.68	2.56	2.45	2.37	2.08
500	6.69	4.65	3.82	3.36	3.05	2.84	2.68	2.55	2.44	2.36	2.07

Quantiles of the t distribution with f degrees of freedom

	Probability in percent						
f	90	95	97.5	99	99.5	99.9	f
1	3.078	6.314	12.706	31.821	63.657	318.309	1
2	1.886	2.920	4.303	6.965	9.925	22.327	2
3	1.638	2.353	3.182	4.541	5.841	10.215	3
4	1.533	2.132	2.776	3.747	4.604	7.173	4
5	1.476	2.015	2.571	3.365	4.032	5.893	5
6	1.440	1.943	2.447	3.143	3.707	5.208	6
7	1.415	1.895	2.365	2.998	3.499	4.785	7
8	1.397	1.860	2.306	2.896	3.355	4.501	8
9	1.383	1.833	2.262	2.821	3.250	4.297	9
10	1.372	1.812	2.228	2.764	3.169	4.144	10
11	1.363	1.796	2.201	2.718	3.106	4.025	11
12	1.356	1.782	2.179	2.681	3.055	3.930	12
13	1.350	1.771	2.160	2.650	3.012	3.852	13
14	1.345	1.761	2.145	2.624	2.977	3.787	14
15	1.341	1.753	2.131	2.602	2.947	3.733	15
16	1.337	1.746	2.120	2.583	2.921	3.686	16
17	1.333	1.740	2.110	2.567	2.898	3.646	17
18	1.330	1.734	2.101	2.552	2.878	3.610	18
19	1.328	1.729	2.093	2.539	2.861	3.579	19
20	1.325	1.725	2.086	2.528	2.845	3.552	20
21	1.323	1.721	2.080	2.518	2.831	3.527	21
22	1.321	1.717	2.074	2.508	2.819	3.505	22
23	1.319	1.714	2.069	2.500	2.807	3.485	23
24	1.318	1.711	2.064	2.492	2.797	3.467	24
25	1.316	1.708	2.060	2.485	2.787	3.450	25
30	1.310	1.697	2.042	2.457	2.750	3.385	30
50	1.299	1.676	2.009	2.403	2.678	3.261	50
75	1.293	1.665	1.992	2.377	2.643	3.202	75
100	1.290	1.660	1.984	2.364	2.626	3.174	100
150	1.287	1.655	1.976	2.351	2.609	3.145	150
200	1.286	1.653	1.972	2.345	2.601	3.131	200
400	1.284	1.649	1.966	2.336	2.588	3.111	400

D Dictionaries

Danish-English

afbildning map
afkomstfordeling offspring distribution

betinget sandsynlighed conditional probability
binomialfordeling binomial distribution

Cauchyfordeling Cauchy distribution
central estimator unbiased estimator
Central Grænseværdisætning Central Limit Theorem

delmængde subset
Den Centrale Grænseværdisætning The Central Limit Theorem
disjunkt disjoint
diskret discrete

eksponentialfordeling exponential distribution
endelig finite
ensidet variansanalyse one-way analysis of variance
enstikprøveproblem one-sample problem
estimat estimate
estimation estimation
estimator estimator
etpunktsfordeling one-point distribution
etpunktsmængde singleton

flerdimensional fordeling multivariate distribution
fordeling distribution
fordelingsfunktion distribution function
foreningsmængde union
forgreningsproces branching process
forklarende variabel explanatory variable
formparameter shape parameter
forventet værdi expected value
fraktil quantile
frembringende funktion generating function
frihedsgrader degrees of freedom
fællesmængde intersection

gammafordeling gamma distribution
gennemsnit average, mean
gentagelse repetition, replicate
geometrisk fordeling geometric distribution

hypergeometrisk fordeling hypergeometric distribution
hypotese hypothesis
hypoteseprøvning hypothesis testing
hyppighed frequency
hændelse event

indikatorfunktion indicator function
khi i anden chi squared
klassedeling partition
kontinuert continuous
korrelation correlation
kovarians covariance
kvadratsum sum of squares
kvotienttesstørrelse likelihood ratio test statistic

ligefordeling uniform distribution
likelihood likelihood
likelihoodfunktion likelihood function

maksimaliseringsestimat maximum likelihood estimate

maksimaliseringsestimator maximum likelihood estimator
marginal fordeling marginal distribution
middelværdi expected value; mean
multinomialfordeling multinomial distribution
mængde set

niveau level
normalfordeling normal distribution
nævner (i brøk) nominator

odds odds

parameter parameter
plat eller krone heads or tails
poissonfordeling Poisson distribution
produktrum product space
punktsandsynlighed point probability

regression regression
relativ hyppighed relative frequency
residual residual
residualkvadratsum residual sum of squares
række row

sandsynlighed probability
sandsynlighedsfunktion probability function
sandsynlighedsmål probability measure
sandsynlighedsregning probability theory
sandsynlighedsrum probability space
sandsynlighedstæthedsfunktion probability density function
signifikans significance
signifikansniveau level of significance
signifikant significant
simultan fordeling joint distribution
simultan tæthed joint density
skalaparameter scale parameter
skøn estimate
statistik statistics
statistisk model statistical model
stikprøve sample, random sample
stikprøvefunktion statistic
stikprøveudtagning sampling
stokastisk uafhængighed stochastic independence
stokastisk variabel random variable
Store Tals Lov Law of Large Numbers
Store Tals Stærke Lov Strong Law of Large Numbers
Store Tals Svage Lov Weak Law of Large Numbers
støtte support
sufficiens sufficiency
sufficient sufficient
systematisk variation systematic variation
søjle column

test test
teste test
testsandsynlighed test probability
teststørrelse test statistic
tilbagelægning (med/uden) replacement (with/without)
tilfældig variation random variation
tosidet variansanalyse two-way analysis of variance
tostikprøveproblem two-sample problem
totalgennemsnit grand mean
tæller (i brøk) denominator
tæthed density
tæthedsfunktion density function

uafhængig independent
uafhængige identisk fordelte independent identically distributed; i.i.d.
uafhængighed independence
udfald outcome
udfaldsrum sample space
udtage en stikprøve sample
uendelig infinite

variansanalyse analysis of variance
variation variation
vekselvirkning interaction

English-Danish

analysis of variance variansanalyse
average gennemsnit

binomial distribution binomialfordeling
branching process forgreningsproces

Cauchy distribution Cauchyfordeling
Central Limit Theorem Den Centrale Grænseværdisætning
chi squared khi i anden
column søjle
conditional probability betinget sandsynlighed
continuous kontinuert
correlation korrelation
covariance kovarians

degrees of freedom frihedsgrader
denominator tæller
density tæthed
density function tæthedsfunktion
discrete diskret
disjoint disjunkt
distribution fordeling
distribution function fordelingsfunktion
estimate estimat, skøn
estimation estimation
estimator estimator
event hændelse
expected value middelværdi; forventet værdi
explanatory variable forklarende variabel
exponential distribution eksponentialfordeling

finite endelig
frequency hyppighed

gamma distribution gammafordeling
generating function frembringende funktion
geometric distribution geometrisk fordeling
grand mean totalgennemsnit

heads or tails plat eller krone
hypergeometric distribution hypergeometrisk fordeling
hypothesis hypotese
hypothesis testing hypoteseprøvning

i.i.d. = independent identically distributed
independence uafhængighed
independent uafhængig
indicator function indikatorfunktion
infinite uendelig
interaction vekselvirkning
intersection fællesmængde

joint density simultan tæthed
joint distribution simultan fordeling

Law of Large Numbers Store Tals Lov
level niveau
level of significance signifikansniveau
likelihood likelihood
likelihood function likelihoodfunktion
likelihood ratio test statistic kvotientteststørrelse

map afbildning
marginal distribution marginal fordeling
maximum likelihood estimate maksimaliseringsestimat
maximum likelihood estimator maksimaliseringsestimator
mean gennemsnit, middelværdi
multinomial distribution multinomialfordeling
multivariate distribution flerdimensional fordeling
nominator nævner
normal distribution normalfordeling

odds odds
offspring distribution afkomstfordeling
one-point distribution etpunktsfordeling
one-sample problem enstikprøveproblem
one-way analysis of variance ensidet variansanalyse
outcome udfald

parameter parameter
partition klassedeling
point probability punktsandsynlighed
Poisson distribution poissonfordeling
probability sandsynlighed
probability density function sandsynlighedstæthedsfunktion
probability function sandsynlighedsfunktion
probability measure sandsynlighedsmål
probability space sandsynlighedsrum
probability theory sandsynlighedsregning
product space produktrum

quantile fraktil

r.v. = random variable
random sample stikprøve
random variable stokastisk variabel
random variation tilfældig variation
regression regression
relative frequency relativ hyppighed
replacement (with/without) tilbagelægning (med/uden)
residual residual
residual sum of squares residualkvadratsum
row række

sample stikprøve; at udtage en stikprøve
sample space udfaldsrum
sampling stikprøveudtagning
scale parameter skalaparameter
set mængde
shape parameter formparameter
significance signifikans
significant signifikant
singleton etpunktsmængde
statistic stikprøvefunktion
statistical model statistisk model
statistics statistik
Strong Law of Large Numbers Store Tals Stærke Lov
subset delmængde
sufficiency sufficiens
sufficient sufficient
sum of squares kvadratsum
support støtte
systematic variation systematisk variation

test at teste; et test
test probability testsandsynlighed
test statistic teststørrelse
two-sample problem tostikprøveproblem
two-way analysis of variance tosidet variansanalyse

unbiased estimator central estimator
uniform distribution ligefordeling
union foreningsmængde
variation variation

Weak Law of Large Numbers Store Tals Svage Lov

Bibliography

Andersen, E. B. (1977). Multiplicative poisson models with unequal cell rates, *Scandinavian Journal of Statistics* **4**: 153–8.

Bachelier, L. (1900). Théorie de la spéculation, *Annales scientifiques de l'École Normale Supérieure, 3^e série* **17**: 21–86.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances, *Philosophical Transactions of the Royal Society of London* **53**: 370–418.

Bliss, C. I. and Fisher, R. A. (1953). Fitting the negative binomial distribution to biological data and Note on the efficient fitting of the negative binomial, *Biometrics* **9**: 176–200.

Bortkiewicz, L. (1898). *Das Gesetz der kleinen Zahlen*, Teubner, Leipzig.

Davin, E. P. (1975). *Blood pressure among residents of the Tambo Valley*, Master's thesis, The Pennsylvania State University.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics, *Philosophical Transactions of the Royal Society of London, Series A* **222**: 309–68.

Forbes, J. D. (1857). Further experiments and remarks on the measurement of heights by the boiling point of water, *Transactions of the Royal Society of Edinburgh* **21**: 135–43.

Gauss, K. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientum*, F. Perthes und I.H. Besser, Hamburg.

Greenwood, M. and Yule, G. U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents, *Journal of the Royal Statistical Society* **83**: 255–79.

Hald, A. (1948, 1968). *Statistiske Metoder*, Akademisk Forlag, København.

Hald, A. (1952). *Statistical Theory with Engineering Applications*, John Wiley, London.

Kolmogorov, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer, Berlin.

Kotz, S. and Johnson, N. L. (eds) (1992). *Breakthroughs in Statistics*, Vol. 1, Springer-Verlag, New York.

Larsen, J. (2006). *Basisstatistik, 2. udgave*, IMFUFA tekst nr 435, Roskilde Universitetscenter.

Lee, L. and Krutchkoff, R. G. (1980). Mean and variance of partially-truncated distributions, *Biometrics* **36**: 531–6.

Newcomb, S. (1891). Measures of the velocity of light made under the direction of the Secretary of the Navy during the years 1880-1882, *Astronomical Papers* **2**: 107–230.

Pack, S. E. and Morgan, B. J. T. (1990). A mixture model for interval-censored time-to-response quantal assay data, *Biometrics* **46**: 749–57.

Ryan, T. A., Joiner, B. L. and Ryan, B. F. (1976). *MINITAB Student Handbook*, Duxbury Press, North Scituate, Massachusetts.

Sick, K. (1965). Haemoglobin polymorphism of cod in the Baltic and the Danish Belt Sea, *Hereditas* **54**: 19–48.

Stigler, S. M. (1977). Do robust estimators work with *real* data?, *The Annals of Statistics* **5**: 1055–98.

Weisberg, S. (1980). *Applied Linear Regression*, Wiley series in Probability and Mathematical Statistics, John Wiley & Sons.

Wiener, N. (1976). *Collected Works*, Vol. 1, MIT Press, Cambridge, MA. Edited by P. Masani.

Index