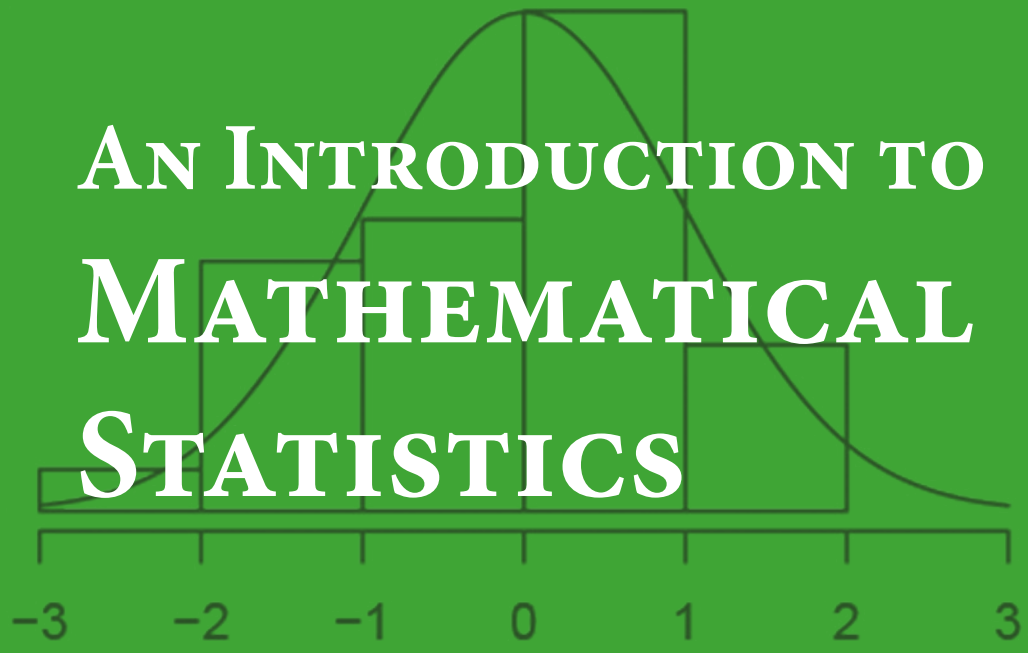
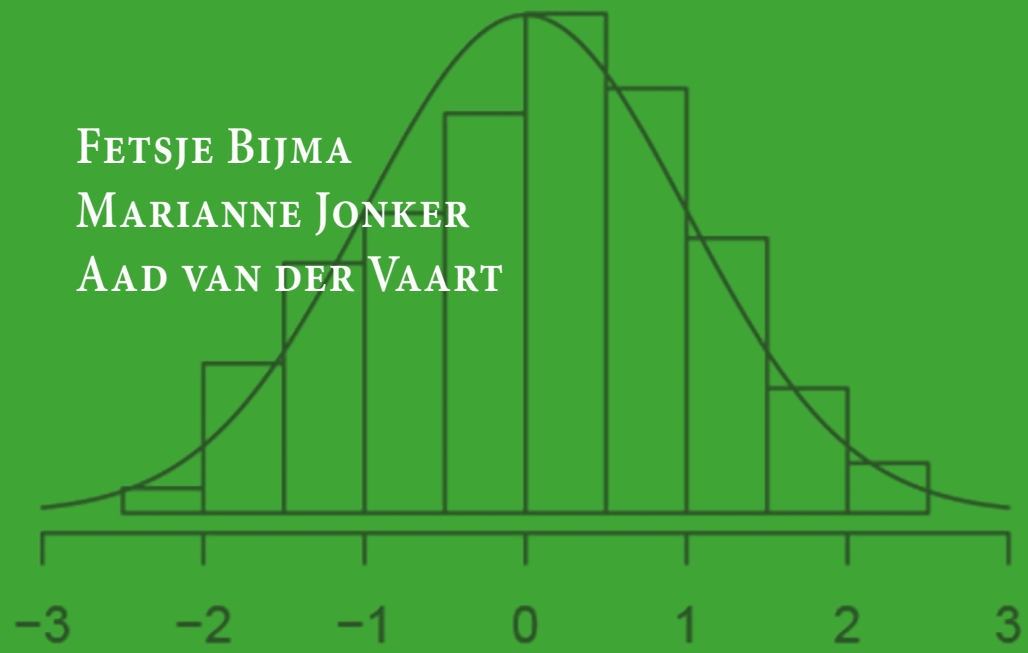
An Introduction to Mathematical Statistics



Fetsje Bijma
Marianne Jonker
Aad van der Vaart



Amsterdam
University
Press

An Introduction to Mathematical Statistics

An Introduction to Mathematical Statistics

Fetsje Bijma, Marianne Jonker, Aad van der Vaart

Amsterdam University Press

Original publication: Fetsje Bijma, Marianne Jonker, Aad van der Vaart, *Inleiding in de statistiek*. Epsilon Uitgaven, 2016 [ISBN 978-90-5041-135-6]

Translated by: Reinie Erné

Cover design: V3-Services, Baarn

Amsterdam University Press English-language titles are distributed in the US and Canada by the University of Chicago Press.

ISBN 978 94 6298 510 0
e-ISBN 978 90 4853 611 5 (pdf)
NUR 916
DOI 10.5117/9789462985100

PREFACE

This book gives an introduction into mathematical statistics. It was written for bachelor students in (business) mathematics, econometrics, or any other subject with a solid mathematical component. We assume that the student already has solid knowledge of probability theory to the extent of a semester course at the same level.

In Chapter 1, we give the definition and several examples of a statistical model, the foundation of every statistical procedure. Some techniques from descriptive statistics that can assist in setting up and validating statistical models are discussed in Chapter 2. The following chapters discuss the three main topics in mathematical statistics: estimating, testing, and constructing confidence regions. These subjects are discussed in Chapters 3, 4, and 5, respectively. Next, Chapter 6 provides deeper theoretical insight, in particular into the question under what circumstances and in what sense certain statistical models are mathematically optimal. In Chapter 7, we describe several regression models that are commonly used in practice. The theory from the previous chapters is applied to estimate and test unknown model parameters and give confidence regions for them. Finally, in Chapter 8, we discuss model selection. In that chapter, various criteria are presented that can be used to find the best-fitting model from a collection of (regression) models. Sections and examples marked with a * are more difficult and do not belong to the basic subject matter of mathematical statistics. Every chapter concludes with a summary.

In Appendix A, we recall elements from probability theory that are relevant for understanding the subject matter of this book. In Appendix B, we discuss properties of the multivariate normal distribution, which is used in several sections. Appendix C contains tables with values of distribution and quantile functions of several distributions to which we refer in the text. These are meant to be used at home or during problem sessions. In "real life," these tables are no longer used: the computer is faster, more accurate, and easier to use. The statistical package R, for example, contains standard functions for the distribution function, the density function, and the quantile function of all standard distributions.

The mathematical style of this book is more informal than that of many mathematics books. Theorems and lemmas are not always proved or may be formulated in an informal manner. The reason is that a pure mathematical treatment is only possible using measure theory, of which we do not assume any knowledge. On the other hand, the relevance and motivation of the theorems are also clear without going into all the details.

Each chapter concludes with a case study. It often contains a statistical problem that is answered as well as possible based on the collected data, using the statistical techniques and methods available at that point in the book. The R-code and data of these applications, as well as the data of several case studies described in the book, are available and can be downloaded from the book's webpage at `http://www.aup.nl`.

Though this book includes examples, practice is indispensable to gain insight into the subject matter. The exercises at the end of each chapter include both theoretical and more practically oriented problems. Appendix D contains short answers to most

exercises. Solutions that consist of a proof are not included.

The book has taken form over a period of 20 years. It was originally written in Dutch and used yearly for the course "Algemene Statistiek" (General Statistics) for (business) mathematics and econometrics students given by the mathematics department of VU University Amsterdam. The various lecturers of the course contributed to the book to a greater or lesser extent. One of them is Bas Kleijn. We want to thank him for his contribution to the appendix on probability theory. More than 2000 students have studied the book. Their questions on the subject and advice on the presentation have helped give the book its present form. They have our thanks. The starting point of the book was the syllabus "Algemene Statistiek" (General Statistics) of J. Oosterhoff, professor of mathematical statistics at VU University Amsterdam until the mid-'90s. We dedicate this book to him.

In 2013, the first edition of this book was published in Dutch, and three years later, in 2016, the second Dutch edition came out. This second edition has been translated into English, with some minor changes. We thank Reinie Erné for translation.

Amsterdam and Leiden, March 2017

FURTHER READING

Reference [1] is an introduction to many aspects of statistics, somewhat comparable to *An Introduction to Mathematical Statistics*. References [3] and [4] are standard books that focus more on mathematical theory, and estimation and tests, respectively. Reference [6] describes the use of asymptotic methods in statistics, on a higher mathematical level, and gives several proofs left out in *An Introduction to Mathematical Statistics*. Reference [5] is a good starting point for whoever wants to delve further into the Bayesian thought process, and reference [7] provides the same for nonparametric methods, which are mentioned in *An Introduction to Mathematical Statistics* but perhaps less prominently than in current practice. Reference [2] elaborates on the relevance of modeling using regression models, for example to draw causal conclusions in economic or social sciences.

[1] Davison, A.C., (2003). *Statistical models*. Cambridge University Press.

[2] Freedman, D., (2005). *Statistical models: theory and applications*. Cambridge University Press.

[3] Lehmann, E.L. and Casella, G., (1998). *Theory of point estimation*. Springer.

[4] Lehmann, E.L. and Romano, J.P., (2005). *Testing statistical hypotheses*. Springer.

[5] Robert, C.P., (2001). *The Bayesian choice*. Springer-Verlag.

[6] van der Vaart, A.W., (1998). *Asymptotic statistics*. Cambridge University Press.

[7] Wasserman, L., (2005). *All of nonparametric statistics*. Springer.

TABLE OF CONTENTS

1 Introduction

1.1 What Is Statistics?

Statistics is the art of modeling (describing mathematically) situations in which probability plays a role and drawing conclusions based on data observed in such situations.

Here are some typical research questions that can be answered using statistics:

(i) What is the probability that the river the Meuse will overflow its banks this year?
(ii) Is the new medical treatment significantly better than the old one?
(iii) What is the margin of uncertainty in the prediction of the number of representatives for political party A?

Answering such questions is not easy. The three questions above correspond to the three basic concepts in mathematical statistics: *estimation*, *testing*, and *confidence regions*, which we will deal with extensively in this book. Mathematical statistics develops and studies methods for analyzing observations based on probability models, with the aim to answer research questions as above. We discuss a few more examples of research questions, observed data, and corresponding statistical models in Section 1.2.

In contrast to mathematical statistics, *descriptive statistics* is concerned with summarizing data in an insightful manner by averaging, tabulating, making graphical representations, and processing them in other ways. Descriptive methods are only discussed briefly in this book, as are methods for collecting data and the modeling of data.

1.2 Statistical Models

In a sense, the direction of statistics is precisely the opposite of that of probability theory. In probability theory, we use a given probability distribution to compute the probabilities of certain events. In contrast, in statistics, we observe the results of an experiment, but the underlying probability distribution is (partly) unknown and must be derived from the results. Of course, the experimental situation is not entirely unknown. All known information is used to construct the best possible statistical model. A formal definition of a "statistical model" is as follows.

Definition 1.1 Statistical model

A statistical model is a collection of probability distribution on a given sample space.

The interpretation of a statistical model is: the collection of all possible probability distributions of the observation X. Usually, this observation is made up of "subobservations," and $X = (X_1, \ldots, X_n)$ is a random vector. When the variables $X_1, \ldots, X_n$ correspond to independent replicates of the same experiment, we speak of a *sample*. The variables $X_1, \ldots, X_n$ are then independent, identically distributed, and their joint distribution is entirely determined by the marginal distribution, which is the same for all X_i. In that case, the statistical model for $X = (X_1, \ldots, X_n)$ can be described by a collection of (marginal) probability densities for the subobservations $X_1, \ldots, X_n$.

The concept of "statistical model" only truly becomes clear through examples. As simply as the mathematical notion of "statistical model" is expressed in the definition above, so complicated is the process of the statistical modeling of a given practical situation. The result of a statistical study depends on the construction of a good model.

Example 1.2 Sample

In a large population consisting of N persons, a proportion p has a certain characteristic A; we want to "estimate" this proportion p. It is too much work to examine everyone in the population for characteristic A. Instead, we randomly choose n persons from the population, with replacement. We observe (a realization of) the random variables $X_1, \ldots, X_n$, where

$$X_i = \begin{cases} 0 & \text{if the } i\text{th person does not have } A, \\ 1 & \text{if the } i\text{th person has } A. \end{cases}$$

Because of the set-up of the experiment (sampling with replacement), we know beforehand that $X_1, \ldots, X_n$ are independent and Bernoulli-distributed. The latter means that

$$\mathrm{P}(X_i = 1) = 1 - \mathrm{P}(X_i = 0) = p$$

for $i = 1, \ldots, n$. There is no prior knowledge concerning the parameter p, other then $0 \leq p \leq 1$. The observation is the vector $X = (X_1, \ldots, X_n)$. The statistical model for X consists of all possible (joint) probability distributions of X whose coordinates $X_1, \ldots, X_n$ are independent and have a Bernoulli distribution. For every possible value of p, the statistical model contains exactly one probability distribution for X.

It seems natural to "estimate" the unknown p by the proportion of the persons with property A, that is, by $n^{-1}\sum_{i=1}^{n} x_i$, where x_i is equal to 1 or 0 according to whether the person has property A or not. In Chapter 3, we give a more precise definition of "estimating." In Chapter 5, we use the model we just described to quantify the difference between this estimator and p, using a "confidence region." The population and sample proportions will almost never be exactly equal. A confidence region gives a precise meaning to the "margin of errors" that is often mentioned with the results of an opinion poll. We will also determine how large that margin is when we, for example, study 1000 persons from a population, a common number in polls under the Dutch population.

Example 1.3 Measurement errors

If a physicist uses an experiment to determine the value of a constant μ repeatedly, he will not always find the same value. See, for example, Figure 1.1, which shows the 23 determinations of the speed of light by Michelson in 1882. The question is how to "estimate" the unknown constant μ from the observations, a sequence of numbers $x_1, \ldots, x_n$. For the observations in Figure 1.1, this estimate will lie in the range 700–900, but we do not know where. A statistical model provides support for answering this question. Probability models were first applied in this context at the end of the 18th century, and the normal distribution was "discovered" by Gauss around 1810 for the exact purpose of obtaining insight into the situation described here.

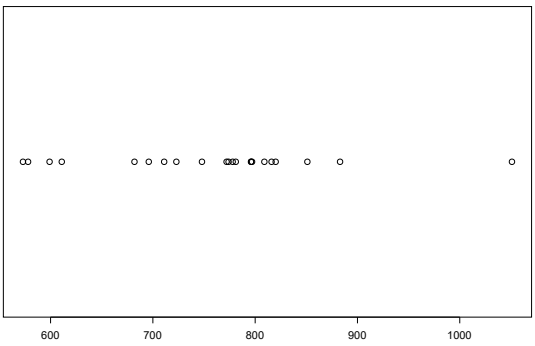


Figure 1.1. The results of the 23 measurements of the speed of light by Michelson in 1882.The scale along the horizontal axis gives the measured speed of light (in km/s) minus 299000 km/s.

If the measurements are all carried out under the same circumstances, independently of the past, then it is reasonable to include in the model that these numbers are realizations of independent, identically distributed random variables $X_1, \ldots, X_n$. The measurement errors $e_i = X_i - \mu$ are then also random variables. A common assumption is that the expected measurement error is equal to 0, in other words, $\mathrm{E}e_i = 0$, in which case $\mathrm{E}X_i = \mathrm{E}(e_i + \mu) = \mu$. Since we have assumed that $X_1, \ldots, X_n$ are independent random variables and all have the same probability distribution, the model for $X = (X_1, \ldots, X_n)$ is fixed by the choice of a statistical model for X_i. For X_i, we propose the following model: all probability distributions with finite expectation μ. The statistical model for X is then: all possible probability distributions of $X = (X_1, \ldots, X_n)$ such that the coordinates $X_1, \ldots, X_n$ are independent and identically distributed with expectation μ.

Physicists often believe that they have more prior information and make more assumptions on the model. For example, they assume that the measurement errors are normally distributed with expectation 0 and variance σ^2, in other words, that the observations $X_1, \ldots, X_n$ are normally distributed with expectation μ and variance σ^2. The statistical model is then: all probability distributions of $X = (X_1, \ldots, X_n)$ such that the coordinates are independent and $N(\mu, \sigma^2)$-distributed.

The final goal is to say something about μ. In the second model, we know more, so we should be able to say something about μ with more "certainty." On the other hand, there is a higher "probability" that the second model is incorrect, in which case the gain in certainty is an illusory one. In practice, measurement errors are often, but not always, approximately normally distributed. Such normality can be justified using the central limit theorem (see Theorem A.28) if a measurement error can be viewed as the sum of a large number of small independent measurement errors (with finite variances), but cannot be proved theoretically. In Chapter 2, we discuss methods to study normality on the data itself.

The importance of a precisely described model is, among other things, that it allows us to determine what is a meaningful way to "estimate" μ from the observations. An obvious choice is to take the average of $x_1, \ldots, x_n$. In Chapter 6, we will see that this is the best choice (according to a particular criterion) if the measurement errors indeed have a normal distribution with expectation 0. If, on the other hand, the measurement errors are Cauchy-distributed, then taking the average is disastrous. This can be seen in Figure 1.2. It shows the average $n^{-1}\sum_{i=1}^{n} x_i$, for $n = 1, 2, \ldots, 1000$, of the first n realizations $x_1, \ldots, x_{1000}$ of a sample from a standard Cauchy distribution. The behavior of the averages is very chaotic, and they do not converge to 0. This can be explained by the remarkable theoretic result that the average $n^{-1}\sum_{i=1}^{n} X_i$ of independent standard Cauchy-distributed random variables $X_1, \ldots, X_n$ also has a standard Cauchy distribution. So taking the averages changes nothing!

Example 1.4 Poisson stocks

A certain product is sold in numbers that vary for different retailers and fluctuate over time. To estimate the total number of items needed, the central distribution center

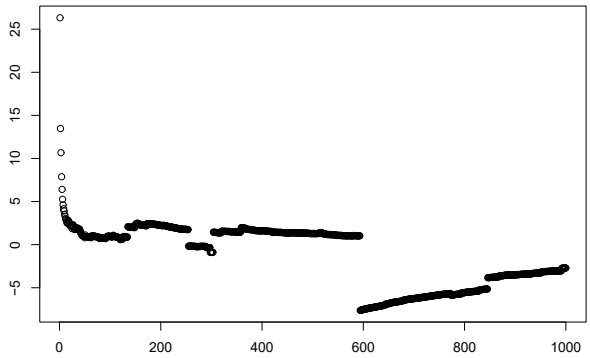


Figure 1.2. Cumulative averages (vertical axis) of $n = 1, 2, \ldots, 1000$ (horizontal axis) realizations from the standard Cauchy distribution.

registers the total number of items sold per week and retailer for several weeks. They observe $x = (x_{1,1}, x_{1,2}, \ldots, x_{I,J})$, where $x_{i,j}$ is the number of items sold by retailer i in week j. The observation is therefore a vector of length the product IJ of the number of retailers and the number of weeks, with integral coordinates. The observations can be seen as realizations of the random vector $X = (X_{1,1}, X_{1,2}, \ldots, X_{I,J})$. Many different statistical models for X are possible and meaningful in given situations. A common (because often reasonably fitting) model states:

- Every $X_{i,j}$ is Poisson-distributed with unknown parameter $\mu_{i,j}$.
- The $X_{1,1}, \ldots, X_{I,J}$ are independent.

This fixes the probability distribution of X up to the expectations $\mu_{i,j} = \mathrm{E}X_{i,j}$. It is these expectations that the distribution center is interested in. The total expected demand in week j, for example, is $\sum_i \mu_{i,j}$. Using the Poisson-character of the demand $\sum_i X_{i,j}$, the distribution center can choose a stock size that gives a certain (high) probability that there is sufficient stock.

The goal of the statistical analysis is to deduce $\mu_{i,j}$ from the data. Up to now, we have left the $\mu_{i,j}$ completely "free." This makes it difficult to estimate them from the data, because only one observation, $x_{i,j}$, is available for each $\mu_{i,j}$. It seems reasonable to reduce the statistical model by including prior assumptions on $\mu_{i,j}$. We could, for example, postulate that $\mu_{i,j} = \mu_i$ does not depend on j. The expected number of items sold then depends on the retailer but is constant over time. We are then left with I unknowns, which can be "estimated" reasonable well from the data provided that the number of weeks J is sufficiently large. More flexible, alternative models are $\mu_{i,j} = \mu_i + \beta_i j$ and $\mu_{i,j} = \mu_i + \beta\mu_i j$, with, respectively, $2I$ and $I+1$ parameters. Both models correspond to a linear dependence of the expected demand on time.

Example 1.5 Regression

Tall parents in general have tall children, and short parents, short children. The heights of the parents have a high predictive value for the final (adult) length of their children, their heights once they stop growing. More factors influence it. The gender of the

child, of course, plays an important role. Environmental factors such as healthy eating habits and hygiene are also important. Through improved nutrition and increased hygiene in the past 150 years, factors that hinder growth like infectious diseases and malnutrition have decreased in most Western countries. Consequently, the average height has increased, and each generation of children is taller.

The target height of a child is the height that can be expected based on the heights of the parents, the gender of the child, and the increase of height over generations. The question is how the target height depends on these factors.

Let Y be the height a child will reach, let x_1 and x_2 be the heights of the biological father and mother, respectively, and let x_3 be an indicator for the gender ($x_3 = -1$ for a girl and $x_3 = 1$ for a boy). The target height $\mathrm{E}Y$ is modeled using a so-called linear regression model

$$\mathrm{E}Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

where β_0 is the increase in average height per generation, β_1 and β_2 are the extent to which the heights of the parents influence the target height of their offspring, and β_3 is the deviation of the target height from the average final height that is caused by the gender of the child. Since men are, on average, taller than women, β_3 will be positive.

The model described above does not say anything about individual heights, only about the heights of the offspring of parents of a certain height. Two brothers have the same target height, since they have the same biological parents, the same gender, and belong to the same generation. The actual final height Y can be described as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e,$$

where $e = Y - \mathrm{E}Y$ is the deviation of the actual final height Y from the target height $\mathrm{E}Y$. The observation Y is also called the dependent variable, and the variables x_1, x_2, and x_3 the independent or predictor variables. The deviation e is commonly assumed to have a normal distribution with expectation 0 and unknown variance σ^2. The final height Y then has a normal distribution with expectation $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ and variance σ^2.

In the Netherlands, the increase in the height of youth is periodically recorded. In 1997, the Fourth National Growth Study took place. Part of the study was to determine the correlation between the final height of the children and the heights of their parents. To determine this correlation, data were collected on adolescents and their parents. This resulted in the following observations: $(y_1, x_{1,1}, x_{1,2}, x_{1,3})$, $\ldots$, $(y_n, x_{n,1}, x_{n,2}, x_{n,3})$, where y_i is the height of the ith adolescent, $x_{i,1}$ and $x_{i,2}$ are the heights of the biological parents, and $x_{i,3}$ is an indicator for the gender of the ith adolescent. Suppose that the observations are independent replicates of linear regression model given above; in other words, given $x_{i,1}, x_{i,2}$, and $x_{i,3}$, the variable Y_i has expectation $\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3}$ and variance σ^2. The parameters $(\beta_0, \beta_1, \beta_2, \beta_3)$ are unknown and can be estimated from the observations. For a simple interpretation of the model, we choose $\beta_1 = \beta_2 = 1/2$, so that the target height is equal to the average height of the parents corrected for the gender of the child and the influence of time. The parameters β_0 and β_3 are equal to the increase in height in the

previous generation and half the average height difference between men and women. These parameters are estimated using the least-squares method (see Example 3.44). The parameter β_0 is estimated to be 4.5 centimeters, and β_3 is estimated to be 6.5 centimeters.[†] The estimated regression model is then equal to

$$Y = 4.5 + \frac{1}{2}(x_1 + x_2) + 6.5x_3 + e. \tag{1.1}$$

Figure 1.3 shows the heights of 44 young men (on the left) and 67 young women (on the right) set out against the average heights of their parents.[‡] The line is the estimated regression line found in the Fourth National Growth Study.

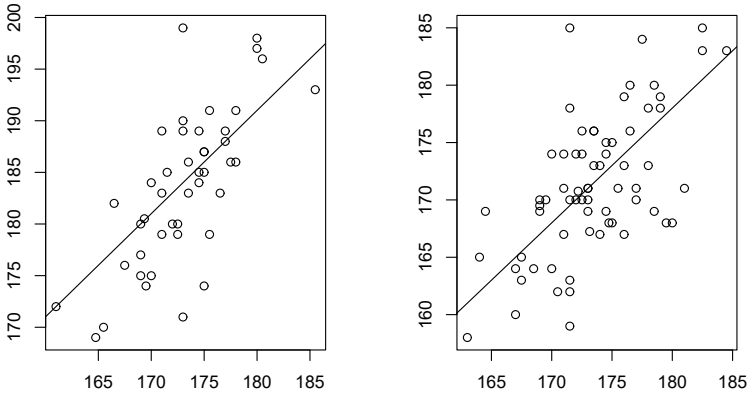


Figure 1.3. Heights (in cm) of sons (left) and daughters (right) set out against the average height of their parents. The line is the regression line found in the Fourth National Growth Study.

We can use the estimated regression model found in the Fourth National Growth Study to predict the final heights of children born now. We must then assume that the height increase in the next generation is again 4.5 centimeters and that the average height difference between men and women remains 13 centimeters. Based on the model presented above, the target heights of sons and daughters of a man of height 180 cm (≈ 71 in or 5'9") and a woman of height 172 cm are $4.5 + (180 + 172)/2 + 6.5 = 187$ cm and $4.5 + (180 + 172)/2 - 6.5 = 174$ cm, respectively.

Other European countries use other models. In Switzerland, for example, the target height is

$$\mathrm{E}Y = 51.1 + 0.718\,\frac{x_1 + x_2}{2} + 6.5x_3.$$

[†] An inch is approximately 2.54 cm, so 4.5 cm corresponds to $4.5/2.54 \approx 1.8$ in and 6.5 cm ≈ 2.6 in.

[‡] Source: The data were gathered by the department of Biological Psychology of VU University Amsterdam during a study on health, lifestyle, and personality. The data can be found on the book's webpage at `http://www.aup.nl` under `heightdata`.

The target heights of sons and daughters of parents of the same heights as above are now 184 and 171 centimeters, respectively.

In the example above, there is a linear correlation between the response Y and the unknown parameters $\beta_0, \ldots, \beta_3$. In that case, we speak of a linear regression model. The simplest linear regression model is that where there is only one predictor variable:

$$Y = \beta_0 + \beta_1 x + e;$$

this is called a simple linear regression model (in contrast to the multiple linear regression model when there are more predictor variables).

In general, we speak of a regression model when there is a specific correlation between the response Y and the observations $x_1, \ldots, x_p$:

$$Y = f_\theta(x_1, \ldots, x_p) + e,$$

where f_θ describes the correlation between the observations $x_1, \ldots, x_p$ and the response Y, and the random variable e is an unobservable measurement error with expectation 0 and variance σ^2. If the function f_θ is known up to the finite-dimensional parameter θ, we speak of a parameterized model. The linear regression model is an example of this; in this model, we have $\theta = (\beta_0, \ldots, \beta_p) \in \mathbb{R}^{p+1}$ and $f_\theta(x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$. The regression model is then fixed if we know the values of θ and σ^2. The function f_θ can, however, also be known up to the finite-dimensional parameter θ and an infinite-dimensional parameter. We then speak of a semiparametric model. An example of a semiparametric model is the Cox regression model. This model is described at the end of this chapter, after the exercises. In Chapter 7, we discuss several regression models in detail, including the linear regression model and the Cox regression model.

Example 1.6 Water levels

In the 20th century (between 1910 and 2000), extreme water levels were measured 70 times in the river the Meuse near the town of Borgharen (Netherlands). Here, "extreme" is defined by Rijkswaterstaat (the Dutch government agency responsible for the management of waterways) as "more than 1250 m^3/s." The maximal water flows during those 70 periods are shown in chronological order in Figure 1.4.[♭] The problem is predicting the future. Rijkswaterstaat is particularly interested in how high the dikes must be to experience flooding at most once every 10 000 years. We can use a hydraulic model to compute the height of the water from the water flow.

[♭] The data can be found on the book's webpage at `http://www.aup.nl` under `maxflows` and `flows1965`.

Since the maximal water flows $x_1, \ldots, x_{70}$ were measured in (mostly) different years, and the water level of the Meuse depends mainly on the weather in the Ardennes and further upstream, it is not unreasonable to view these numbers as realizations of independent random variables $X_1, \ldots, X_{70}$. The assumption that these parameters are also identically distributed is somewhat questionable because the course of the Meuse (and also the climate) has gradually changed during of the last century, but this assumption is usually made anyway. We can then view $X_1, \ldots, X_{70}$ as independent copies of one variable X and use the measured values $x_1, \ldots, x_{70}$ to answer the question.

Let E be the event that flooding takes place in an (arbitrary) year. The probability of event E is approximately equal to the expected number $\mathrm{E}N$ of extreme periods in a year, times the probability that there is a flood in an extreme period, that is, $\mathrm{P}(E) \approx \mathrm{E}N\ \mathrm{P}(X > h)$ for X a maximal water flow in a period of extreme water flow, h the maximal water flow so that there is no flood, and N the number of times we have extremely high water levels in an arbitrary year. For this computation, we use that the probability of flooding in an extreme period $\mathrm{P}(X > h)$ is small. The probability distribution of N is unknown, but it is reasonable to assume that the expectation of N is approximately equal to the average number of periods of extreme water flow per year in the past 90 years, so $\mathrm{E}N \approx 70/90$. The question is now: for which number h do we have $\mathrm{P}(X > h) = 1/10000 \cdot 90/70 = 0.00013$?

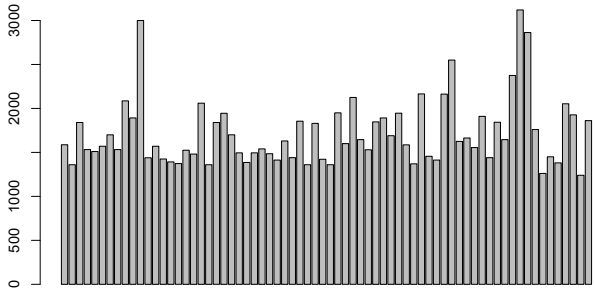


Figure 1.4. Maximal water flows in m^3/s (vertical axis) in the Meuse near Borgharen in the 20th century in chronological order (horizontal axis).

This question cannot easily be answered. If the observed maxima for a period of 100 000 years (or more) were available, then we could determine h with a reasonable accuracy, for example as the value of the 10%th highest measured water level ($10\% = 10\,000/100\,000$). Unfortunately, we dispose over only 70 observations, and must therefore extrapolate far into the future to a (probably) much more extreme situation than ever measured. If we can determine a good model for the distribution of X, then this is not a problem. If we, for example, knew that X has the standard exponential

distribution, then we could determine h from the equation $0.00013 = \mathrm{P}(X > h) = e^{-h}$. This is not, however, a realistic assumption.

An alternative is given by fitting an *extreme value distribution* to the data. These are probability distributions that are commonly used for modeling variables X that can be viewed as a maximum $X = \max(Y_1, \ldots, Y_m)$ of a large number of independent variables $Y_1, \ldots, Y_m$. Given the interpretation of X as a maximal water flow in a period, such distributions seem reasonable. Of the three types of extreme value distributions, one type proves to fit the data reasonably well. This is the *Fréchet family*, where the distribution function is given by

$$F(x) = \begin{cases} e^{-((x-a)/b)^{-\alpha}} & \text{if } x \geq a, \\ 0 & \text{if } x < a. \end{cases}$$

The Fréchet family has three parameters: $a \in \mathbb{R}$, $b > 0$, and $\alpha > 0$. If we are convinced of the usefulness of the resulting model, we can estimate these parameters from the 70 data points and then answer the question through a simple computation. In Chapter 3, we discuss suitable estimation methods and in the application after Chapter 6, we further work out the data of the water flows.

Example 1.7 Survival analysis

In survival analysis, we study the probability distribution of time spans. You can think of the life span of a light bulb, but also of the time before the next bug occurs in a computer program ("reliability analysis") and, in particular, of the remaining time until death or until the occurrence of a disease in medical statistics. Below is an example.

In persons with a leaking heart valve, the heart valve is often replaced by a biological or mechanical heart valve. A disadvantage of the biological over the mechanical heart valve is the relatively short life span (10 to 15 years). To study the distribution function F of the life span of a biological heart valve, n persons with such a valve are followed from the operation up to the moment that the valve must be replaced. At the end of the study, we have measured the life spans $t_1, \ldots, t_n$ of all of the n heart valves. We view these numbers as realizations of independent random variables $T_1, \ldots, T_n$ with distribution function F. The probability $F(t)$ that a biological heart valve must be replaced within t years can be estimated by the proportion of heart valves in the sample that is replaced within t years.

A special aspect of survival analysis is that, often, not all life spans are observed. At the moment that we want to draw conclusions from the data, for example, not all heart valves have needed replacement or a patient may have died with a heart valve that was still good. In those cases, too, we only observe a lower bound for the life spans, the time until the end of the study or until the death of the patient. We know that the heart valve still worked when the study was ended or the patient died. We then speak of *censored data*.

Long life spans are more frequently censored than short ones because the probability that the patient dies is greater during a long period of time than during a short one (and the same holds for the study ending). It would therefore be wrong to ignore censored data and estimate the distribution function F based on the uncensored data. This would lead to an overestimate of the distribution function of the life span and an underestimate of the expected life span because relatively longer life spans would be ignored. A correct approach is to use a statistical model for all observations, both censored and uncensored.

The statistical model becomes even more complex if we suspect that there are factors that could influence the life span of the heart valve, for example the age, weight, or gender of the patient. In such a case, the life span can be modeled using, for example, the Cox regression model. This model is studied at the end of this chapter (after the exercises) and in Chapter 7.

Example 1.8 Selection bias

To correctly answer a research question, it is important that this question, the collected data, and the statistical model are correctly aligned. This is illustrated below.

The Dutch Railways (Nederlandse Spoorwegen or NS for short) regularly receive complaints about crowding in the trains during rush hour. A study is set up to investigate whether these complaints are justified. There are two research questions. The first is what percentage of the passengers does not have a seat during rush hour. The second is what percentage of rush hour trains is too crowded. Note that these are two fundamentally different questions. The first question concerns people, a percentage of passengers, while the second question concerns trains. A passenger is probably only interested in the first research question, while the NS also attach importance to the answer of the second. They have to identify on which trains there are problems, and where measures must be taken.

To answer the first research question, a sample of size 50 is taken from train passengers that have just got off. Each person is asked whether they could sit. We observe the sequence $x_1, \ldots, x_{50}$, where x_i equals 1 if the ith person did not have a seat and x_i is equal to 0 if the ith person did have a seat. Then $x_1, \ldots, x_{50}$ are realizations of independent random variables $X_1, \ldots, X_{50}$ with a Bernoulli distribution with parameter p, where $p = \mathrm{P}(X_i = 1)$ is the proportion of passengers that could not be seated. As in Example 1.2, we can estimate the proportion p using the sample mean $50^{-1} \sum_{i=1}^{50} x_i$. This is a correct way to answer the research question.

Answering the second research question is more difficult, because it concerns trains and not persons. To carry out this study, during rush hour, 50 head conductors are randomly chosen and asked whether the train they were just on was overcrowded. We observe the sequence $y_1, \ldots, y_{50}$, where y_i is equal to 1 if the ith head conductor indicates that the train was overcrowded and y_i is equal to 0 if this was not the case. We can again view $y_1, \ldots, y_{50}$ as realizations of $Y_1, \ldots, Y_{50}$, which are independent Bernoulli variables with probability $q = \mathrm{P}(Y_i = 1)$. If we assume that there is only one head conductor on each train, the probability q equals the proportion of rush hour

trains that were overcrowded. We can see $Y_1, \ldots, Y_{50}$ as a sample from the trains that just pulled in. The proportion q can be estimated using the sample mean $50^{-1} \sum_{i=1}^{50} y_i$.

It is simpler to also ask the sample of travelers we gathered to answer the first research question whether the train they were in was overcrowded. In that case, we observe a sequence of realization of the independent Bernoulli variables $Z_1, \ldots, Z_{50}$ with $r = \mathrm{P}(Z_i = 1)$. Here, Z_i is defined analogously to Y_i. Since a train carries more than one passenger, not every train passenger will correspond to a unique train. Since there are more persons in crowded trains than in quiet ones, the percentage "people from crowded trains" in the population of train passengers will be much higher than the percentage of "crowded trains" in the population of trains. In other words, r will be greater than q. It is difficult to give a correlation between r and q without making additional assumptions. That is why the second research question could not easily be answered based on a sample from the passengers, while the first research question could.

In most of the examples given above, the statistical model is *parameterized* by a parameter, for example p, (μ, σ^2), $(\beta_0, \beta_1, \beta_2, \beta_3)$, or (a, b, α). Many statistical models are known up to a parameter. In this book, we often denote that parameter by θ ("theta"). The statistical model can then be denoted by $\{P_\theta : \theta \in \Theta\}$, where P_θ is the probability distribution of the observation X and Θ is the set of possible parameters. There is a tacit assumption that exactly one of the parameter values (or exactly one element of the model) gives the "true" distribution of X. The purpose of statistics is to find that value. What makes statistics difficult, is that we never fully succeed and that statements about the true parameter value always contain a certain element of uncertainty (by definition).

Exercises

1. Suppose that n persons are chosen randomly from a population and asked their political affiliation. Denote by X the number of persons from the sample whose affiliation is with political party A. The proportion of individuals in the population affiliated with party A is the unknown probability p. Describe a corresponding statistical model. Give an intuitively reasonable "estimate" of p.

2. Suppose that $m + n$ patients with high blood pressure are chosen randomly and divided arbitrarily into two groups of sizes m and n. The first group, the "treatment group," is given a particular blood-pressure-lowering drug; the second group, the "control group," is given a placebo. The blood pressure of each patient is measured before and one week after administering the drug or placebo, and the difference in blood pressure is determined. This gives observations $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$.

(i) Formulate a suitable statistical model.

(ii) Give an intuitively reasonable "estimate" of the effect of the drug on the height of the blood pressure, based on the observations (several answers are possible!).

3. We want to estimate the number of fish, say N, in a pond. We proceed as follows. We catch r fish and mark them. We then set them free. After some time, we catch n fish (without putting them back). Of these, X are marked. Consider r and n as constants we choose ourselves, and let X be the observation.
(i) Formulate a suitable statistical model.
(ii) Give an intuitively reasonable "estimate" of N based on the observation.
(iii) Answer the previous questions if, the second time we catch fish, they are put back directly after catching them (sampling with replacement).

4. When assessing a batch of goods, we continue until 3 items are rejected.
(i) Formulate a suitable statistical model.
(ii) The third rejected item is the 50th we assess. Give an estimate of the percentage of defect items in the batch. Justify your choice.

5. The number of customers in the post office seems to depend on the day of the week (weekday or Saturday) and half-day (morning or afternoon). On workdays, the post office is open in the morning and in the afternoon, and on Saturday, is it open only in the morning. To determine how many employees are required to provide prompt service, the number of customers is registered over a period of ten weeks. Every day, the number of customers in the post office in the morning (on weekdays and Saturdays) and in the afternoon (on weekdays only) is noted.
(i) Formulate a suitable statistical model.
(ii) Give an intuitively reasonable "estimate" of the number of clients on a Monday afternoon. Justify your choice.
(iii) The biggest difference in numbers of customers is between the half-days during the workweek (Monday through Friday, mornings and afternoons) and the Saturday morning. It was therefore decided to only take into account this difference in the staff planning. Reformulate the statistical model and give a new estimate.

6. The yearly demand for water in the African city of Masvingo is greater than the amount that can be recovered from the precipitation in one year. Therefore, water is supplied from a nearby lake according to the need. The amount of water that needs to be supplied per year depends on the precipitation in that year and on the size of the population of Masvingo. Moreover, rich people use more water than poor people. Describe a linear regression model with "amount of water to be supplied" as dependent variable and "population size," "precipitation," and "average income" as predictor variables. Indicate for each of the parameters whether you expect them to be positive or negative.

7. A linear correlation is suspected between the income of a person and their age and level of education (low, middle, high).
(i) Describe a linear regression model with "income" as dependent variable and "age" and "education" as predictor variables. Think carefully about how to include the variable "education" in the model.
(ii) We want to study whether the gender of a person has an influence on the income. Adapt the linear regression model so that this can be studied.

8. We want to estimate the average length of wool fibers in a large bin. The bin is first shaken well, after which we take a predefined number of fibers from the bin, one by one and with closed eyes. We estimate the average length of the wool fibers in the bin to be the average length of the wool fibers in the sample. Is the estimated length systematically too long, systematically too short, or just right?

9. At a call center, we want to estimate how long a customer must wait before being helped. For one day, we register how long each customer must wait. If the customer looses patience and hangs up, their waiting time up to that moment is noted. Afterward, we calculate the average waiting time by taking the average of the noted times. This average is used as an estimate of the waiting time of a new customer. What do you think of this method?