Statistics

An introductory course for the departments
D-UWIS, D-ERDW & D-AGRL
Spring semester 2009

(based on lecture notes by H.-R. Künsch)

Peter Bühlmann
Seminar für Statistik
ETH Zürich

Contents

1 Introduction (Stahel, ch. 1) 7

2 Models for count data 9

2.1 Introduction (Stahel, ch. 4.1) 9

2.2 Discrete probabilities (Stahel, ch. 4.2, 4.6) 9

2.2.1 The probabilistic concept of independence 11

2.3 Random variables (Stahel, ch. 4.3, 4.4) 11

2.4 The binomial distribution (Stahel, ch. 5.1) 13

2.5 Characteristic numbers of a distribution (Stahel, ch. 5.3) 14

2.5.1 Cumulative probability distributions 16

2.6 The Poisson distribution (Stahel, ch. 5.2) 16

2.6.1 Poisson approximation of the binomial distribution 17

3 Statistics for count data 19

3.1 Three key questions in statistics (Stahel, Kap. 7.1) 19

3.2 Estimation, statistical testing and confidence intervals for the binomial distribution (Stahel, ch. 7.2, 8.2, 9.1, 9.2) 19

3.2.1 (Point) Estimation 20

3.2.2 Statistical testing 20

3.2.3 Confidence intervals 24

3.3 Estimation, statistical testing and confidence intervals for the Poisson distribution (Stahel, ch. 7.2, 8.1, 9.1) 24

3.3.1 (Point) Estimation 25

3.3.2 Statistical testing 25

3.3.3 Confidence intervals 25

4 Models and statistics for continuous data 27

4.1 Introduction 27

4.2 Descriptive statistics (Stahel, ch. 2 and 3.1, 3.2) 28

4.2.1 Characteristic Numbers 28

4.2.2 Graphical methods 29

4.3 Continuous random variables and distributions (Stahel, ch. 6.1 – 6.4, 11.2) . . 31

4.3.1 (Probability) Densities 31

4.4 Important continuous distributions (Stahel, ch. 6.2, 6.4, 6.5, 11.2) 33

4.4.1 The uniform distribution 33

4.4.2 The exponential distribution 34

4.4.3 The normal (or Gauss) distribution 34

4.4.4 Transformations 36

4.4.5 Analogies between models and data 37

4.4.6 Checking assumptions of normality 38
4.5 Functions of random variables and propagation of errors (Stahel, ch. 6.8 – 6.11) 39
4.6 Single-sample statistics (Stahel, ch. 8.3 – 8.5, 9.3) 41
4.6.1 (Point) Estimates 41
4.6.2 Testing μ 41
4.6.3 Confidence intervals for μ 45
4.6.4 Testing μ for non-normal data 45
4.7 Testing with two independent samples (Stahel, ch. 8.8) 47
4.7.1 Paired and unpaired samples 47
4.7.2 Paired tests 48
4.7.3 Unpaired tests 48
4.7.4 Two-sample t-test for equal variances 48
4.7.5 Further two-sample tests 49

5 Regression 51
5.1 Correlation und empirical correlation 51
5.1.1 Empirical correlation 51
5.2 Simple linear regression 52
5.2.1 The simple linear regression model 52
5.2.2 Parameter estimation 53
5.2.3 Tests and confidence intervals 54
5.2.4 R^2, the coefficient of determination 56
5.2.5 General procedure for simple linear regression 57
5.2.6 Analysis of residuals 57
5.3 Multiple linear regression 60
5.3.1 The multiple linear regression model 60
5.3.2 Parameter estimation and the t-test 61
5.3.3 The F-test 62
5.3.4 R^2, the coefficient of determination 62
5.3.5 Analysis of residuals 62
5.3.6 Strategies for data analysis: a closing example 62

Structure of this course

In the first part of this course we treat the basics of probability theory and of statistics in discrete cases, where the variables involved have values e.g. in $\{0, 1\}$, $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$ or $\mathbb{Z} = \{\ldots, -1, 0, 1, \ldots\}$.

We will subsequently transfer concepts from the discrete to the continuous case, where variables have ranges such as $\mathbb{R}$ or $[0, 1]$. This makes for a slightly repetitive structure of the course, but it has so far proven to be a successful approach.

In the final part of this course we shall look into more complex models, such as those given by multiple regression.

Chapter 1

Introduction (Stahel, ch. 1)

For many areas of science the significance of statistics lies in its ability to

draw general conclusions about future data or entire populations using samples of data.

In particular, the fact that

all data are subject to certain variations

is taken into account. To quantify this,

models and laws of probability

are used.

Chance is subject to certain laws from probability theory, all of which are as reliable as the other laws of mature. Whether the world really is random – or randomness is simply a term to describe all unfathomable deterministic factors – is secondary to our considerations here.

Chapter 2

Models for count data

2.1 Introduction (Stahel, ch. 4.1)

A **probability model** describes the possible outcomes of an experiment and the chances these outcomes have of being realized. In this chapter we shall treat discrete probability models, whose outcomes are finite or "countable" (e.g. natural numbers). A probability model allows the simulation of further data and thus admits insights into the plausibility of particular variations in the data.

All experiments described here are to be understood as **random experiments**:

random experiment =
experiment whose outcome cannot be predicted, even by an oracle

2.2 Discrete probabilities (Stahel, ch. 4.2, 4.6)

To describe random experiments, we shall utilize a probability model. This consists of the following parts:

- **An underlying space** Ω
- **Elementary events** ω
- **A probability** P

The underlying space and the elementary events relate as follows:

$$\Omega = \{\underbrace{\text{possible elementary events } \omega}_{\text{potential outcomes}}\}$$

Exampe: Tossing a coin twice
$\Omega = \{KK, KZ, ZK, ZZ\}$, where K denotes "heads" and Z denotes "tails".
Elementary event: e.g. $\omega = KZ$

An **event** A is a subset of Ω:

$$\text{Event A} \subset \Omega$$

Example (cont.): $A = \{\text{exactly 1 head}\} = \{KZ, ZK\}$.

Set-theoretical operations (taking complements, unions, intersections) have a natural interpretation in the language of events.

$$\begin{aligned} A \cup B &\Leftrightarrow \text{A } \textbf{or} \text{ B, where "or" is non-exclusive ("and/or")} \\ A \cap B &\Leftrightarrow \text{A } \textbf{and} \text{ B} \\ A^c &\Leftrightarrow \textbf{not} \text{ A} \end{aligned}$$

Example: $A =$ the sun will shine tomorrow, $B =$ it will rain tomorrow.
$A \cup B$ means: tomorrow the sun will shine or it will rain (or possibly both); $A \cap B$ works out as: tomorrow the sun will shine and it will also rain; A^c means: tomorrow the sun will not shine.

A **probability measure** assigns a probability $P(A)$ to each event A, such that the following three basic assertations (Kolmogorov axioms) hold:

1. All probabilities are non-negative: $P(A) \geq 0$
2. The certain event (the full underlying space) has a probability equal to 1: $P(\Omega) = 1$
3. $P(A \cup B) = P(A) + P(B)$ whenever $A \cap B = \emptyset$, i.e. for events that cannot occur simultaneously.

From these axioms, further rules can be derived, e.g.

$$\begin{aligned} &P(A^c) = 1 - P(A), \\ &P(A \cup B) = P(A) + P(B) - P(A \cap B). \end{aligned}$$

In the discrete case, all probabilities are determined by the probabilities of the elementary events $P(\{\omega\})$:

$$P(A) = \sum_{\omega \in A} P(\{\omega\}).$$

Example (cont.) The probability of $A =$ {exactly 1 head} $= \{KZ, ZK\}$ is
$P(A) = P(KZ) + P(ZK) = 1/4 + 1/4 = 1/2$.

Probability theory essentially fixes the probabilities of certain events A (based on plausibility or symmetry arguments, scientific theories, expert knowledge and data) and uses the rules given above to derive the probabilities of certain other events B.
(Statistics takes the reverse approach: data – i.e. information about the occurence of certain events – is used to draw conclusions about an unknown probability model and the probabilities within it.)

Possible interpretions of a probability
• Idealized value of a relative frequency from many independent repetitions of the same thing (**frequentist**)
• Measure of the belief that an event will occur (**Bayesian**)

2.2.1 The probabilistic concept of independence

In general, knowledge of the individual probabilities $P(A)$ and $P(B)$ does not allow us to compute $P(A \cap B)$.

If there is no causal link between the events A and B (i.e. neither common causes nor preclusion), we define the following:

$$\text{A and B are (stochastically) independent } \Leftrightarrow P(A \cap B) = P(A)P(B).$$

The independence of events simplifies many situations; in particular, it permits the computation of $P(A \cap B)$ from $P(A)$ and $P(B)$. In practice plausibility arguments are often used to **declare** the independence of two events.

Independence of multiple events $A_1, \ldots A_n$ means e.g. that

$$\begin{aligned} &P(A_1 \cap A_2) = P(A_1)P(A_2), \\ &P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3). \end{aligned}$$

The general defining property of independent events is the following:

$P(A_{i_1} \cap \ldots \cap A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k})$ for each $k \leq n$ and each $1 \leq i_1 < \ldots < i_k \leq n$.

2.3 Random variables (Stahel, ch. 4.3, 4.4)

Often random experiments have associated numerical values, i.e. for each elementary event (outcome) ω there is a number $X(\omega) = x$.

Example: Random draw of a playing card
Define the value function X by:

$$\begin{array}{rcl} \omega = \text{Ace} & \mapsto & X(\omega) = 11 \\ \omega = \text{King} & \mapsto & X(\omega) = 4 \\ \omega = \text{Queen} & \mapsto & X(\omega) = 3 \\ \omega = \text{Jack} & \mapsto & X(\omega) = 2 \\ \omega = \text{Ten} & \mapsto & X(\omega) = 10 \\ \omega = \text{Nine} & \mapsto & X(\omega) = 0 \\ \vdots & & \vdots \\ \omega = \text{Six} & \mapsto & X(\omega) = 0 \end{array}$$

Thus in the example above, $X(\cdot)$ is a function. In general we define:

A **random variable** X is a **function**:

$$\begin{aligned} X: \quad & \Omega \to \mathbb{R} \\ & \omega \mapsto X(\omega) \end{aligned}$$

The function $X(\cdot)$ is not random, but its argument ω is.

While it is rather unusual to denote a function by X (or $Y, Z, \ldots$), we shall see that random variables sometimes admit calculations like those with ordinary variables such as x (or $y, z, \ldots$).

The outcomes of the random experiment (i.e. of ω) yield different possible values of $x = X(\omega)$: the value of x is a **realization** of the random variable X. Thus a realization of a random variable is the result of a random experiment (which may be described by a number).

We call a random variable *discrete* if its range $W = W_X$ (the set of potential values of X) is discrete, i.e. countable (its potential values can be numbered). For example: $W = \{0, 1, \ldots, 10\}$ is finite and thus discrete, while $W = \mathbb{N}_0 = \{0, 1, 2, \ldots\}$ is infinite, but still discrete; $W = \mathbb{R}$ is not discrete (but continuous). In this chapter we shall only treat discrete random variables.

The distribution of a random variable

The random variable X takes its values (its potential realizations) with certain probabilities. These are defined as follows:

$$\begin{aligned}
&\text{Probability of } X \text{ taking the value } x\\
&= P(X = x) = P(\{\omega;\ X(\omega) = x\})\\
&= \sum_{\omega; X(\omega)=x} P(\omega).
\end{aligned}$$

Example (cont.): $X =$ Value of a playing card drawn at random

$$\begin{aligned}
&\text{Probability of } 4 = P(X = 4)\\
= \quad & P(\{\omega;\ \omega = \text{ a king}\})\\
= \quad & P(\text{King of diamonds}) + P(\text{King of hearts}) + P(\text{King of clubs}) + P(\text{King of spades})\\
= \quad & 4/36 = 1/9.
\end{aligned}$$

The "list" of probabilities $P(X = x)$ for all possible values of x is called the (discrete) **(probability) distribution** of the (discrete) random variable X. Each random variable X has a corresponding (probability) distribution, and vice versa:

$$\text{Random variable } X \ \Leftrightarrow\ \text{(probability) distribution}$$

Each (discrete) probability distribution satisfies the equality

$$\sum_{\text{all } x \text{ possible}} P(X = x) = 1.$$

Example (cont.): $X =$ Value of a playing card drawn at random
The probability distribution of X is

$$P(X = 11) \quad = \quad 1/9$$

$$\begin{aligned} P(X = 10) &= 1/9 \\ P(X = 4) &= 1/9 \\ P(X = 3) &= 1/9 \\ P(X = 2) &= 1/9 \\ P(X = 0) &= 4/9 \end{aligned}$$

If our only interest lies in the random variable X, we can ignore the underlying space Ω – as it then suffices to know the distribution of X.

2.4 The binomial distribution (Stahel, ch. 5.1)

Regard the situation where the quantity of interest is the number of successes (or failures) at something. Examples of this include quality control, success or failure of (medical or biological) treatments, or gambling.

Example: Coin toss
A coin is tossed and randomly comes up heads (K) or tails (Z).
Regard the random variable X with values in $W = \{0, 1\}$ describing the following:

$$\begin{aligned} X &= 0 \quad \text{if the outcome is tails,} \\ X &= 1 \quad \text{if the outcome is heads.} \end{aligned}$$

The probability distribution of X can be described by a single parameter π:

$$P(X = 1) = \pi, \quad P(X = 0) = 1 - \pi, \quad 0 \leq \pi \leq 1.$$

A fair coin has the parameter $\pi = 1/2$.

Bernoulli(π) distribution:

A random variable X with range $W = \{0, 1\}$ has a Bernoulli(π) distribution if
$P(X = 1) = \pi, \quad P(X = 0) = 1 - \pi, \quad 0 \leq \pi \leq 1.$

The Bernoulli distribution is a trivial mathematical description of the (non-)occurence of an event.

Example (cont.): n-fold coin toss
Regard $X =$ Number of heads from n independent coin tosses. Obviously the range of X is the set $W = \{0, 1, \ldots, n\}$. X can also be written as the sum of independent Bernoulli-distributed random variables:

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

$$X_i = \begin{cases} 1 & \text{i-th toss comes up heads} \\ 0 & \text{i-th toss comes up tails.} \end{cases}$$

In the example above, the distribution of X can be computed analytically. If $X_1, \ldots, X_n$ are all independent and each follows a Bernoulli(π) distribution, then we know e.g. that

$$\begin{aligned}
P(X = 0) &= P(\text{all } X_1 = \ldots = X_n = 0) = (1-\pi)^n, \\
P(X = 1) &= P(\text{one } X_i = 1 \text{ and all other } X_j = 0) = \binom{n}{1}\pi(1-\pi)^{n-1}.
\end{aligned}$$

In general cases the binomial formula applies.

Binomial(n, π) distribution:

A random variable X with range $W = \{0, 1, \ldots, n\}$ has a Binomial(n, π) distribution if

$$P(X = x) = \binom{n}{x}\pi^x(1-\pi)^{n-x}, \; x = 0, 1, \ldots, n$$

where $0 \leq \pi \leq 1$ is the success rate associated to the distribution.

(Here $\binom{n}{x}$ is the binomial coefficient, which denotes the number of possible arrangements of x successes and $n - x$ failures).

As in the previous example, X denotes the number of successes/failures (occurence of a particular event) out of n **independent** experiments. The independence of these experiments is crucial if the binomial distribution is to apply.

Example: Sperm sexing (Tages-Anzeiger 6.12.2000)
The gender of calves can be influenced by a technique called sperm sexing, with a view to breeding female calves. In an experiment 12 cows were inseminated with sperm that had previously been sorted according to whether or not a Y chromosome was visible (i.e. sperm sexing was applied). As this technique does not guarantee any outcome with 100% certainty, we can consider it to be a random experiment. Let X be the number of female calves bred by this method. A reasonable model for X is given by

$$X \sim \text{ Binomial}(12, \pi),$$

where π is an unknown parameter. In the experiment, $x = 11$ female calves were observed: in other words, $X = x = 11$ was the actual **realization**.

Properties of the binomial distribution (cf. Fig. 2.1): $P(X = x)$ attains its maximum when x is equal to the integer part of $(n+1)\pi$, and on both sides of this, the probabilities decrease monotonically. When $n\pi(1-\pi)$ is not too small, the distribution has a bell shape.

2.5 Characteristic numbers of a distribution (Stahel, ch. 5.3)

An arbitrary (discrete) distribution can be summarized by 2 characteristic numbers, its **mean** $\mathcal{E}(X)$ and its **variance** $\text{Var}(X)$ (or its **standard deviation** $\sigma(X) = \sqrt{\text{Var}(X)}$).

The mean of a distribution describes its average location and its defined as

$$\mathcal{E}(X) = \sum_{x \in W_x} xP(X = x), \quad W_x = \text{ range of } X.$$

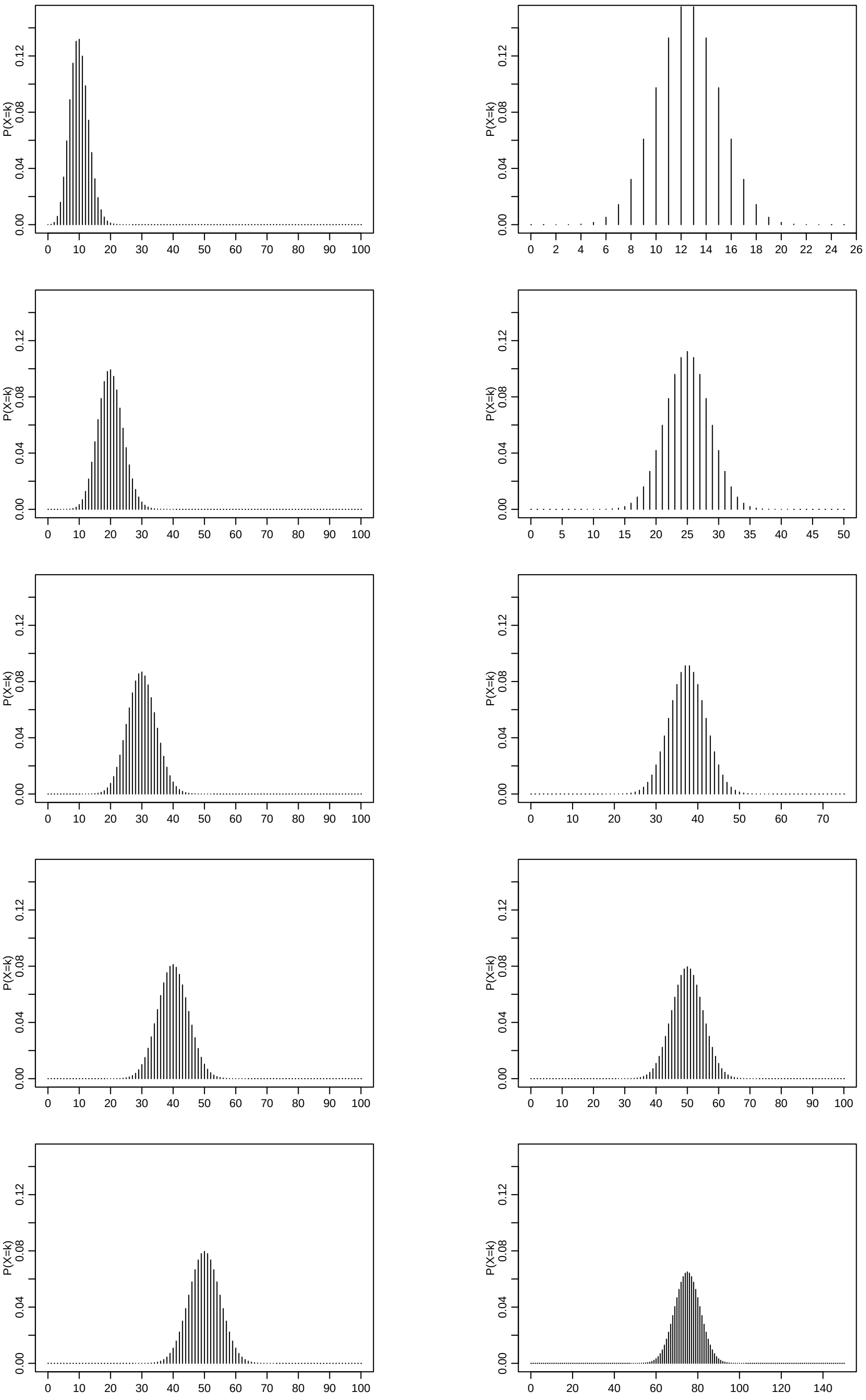

Figure 2.1: The binomial probabilities $P(X = x)$ as a function of x for various choices of n and π. On the left, $n = 100$ and $\pi = 0.1, 0.2, 0.3, 0.4, 0.5$, and on the right, $\pi = 0.5$ and $n = 25, 50, 75, 100, 150$.

The variance and standard deviation of a distribution describe its variability:

$$\begin{aligned} &\mathrm{Var}(X) \sum_{x \in W_x} (x - \mathcal{E}(X))^2 P(X = x) \\ &\sigma(X) = \sqrt{\mathrm{Var}(X)}. \end{aligned}$$

A standard deviation has the same unit of measurement as the data it describes: if X is measured e.g. in metres (m), then the unit of $\mathrm{Var}(X)$ is the square metre (m^2), and that of $\sigma(X)$ the metre once again.

Example: Let $X \sim$ Bernoulli(π).
Then:

$$\begin{aligned} \mathcal{E}(X) &= 0 \cdot P(X = 0) + 1 \cdot P(X = 1) = \pi, \\ \mathrm{Var}(X) &= (0 - \mathcal{E}(X))^2 P(X = 0) + (1 - \mathcal{E}(X))^2 P(X = 1) = \pi^2(1 - \pi) + (1 - \pi)^2 \pi \\ &= \pi(1 - \pi), \\ \sigma(X) &= \sqrt{\pi(1 - \pi)}. \end{aligned}$$

The binomial distribution has the following general properties (note that Bernoulli(π) = Binomial(1,π)):

$$\begin{aligned} &X \sim \text{ Binomial}(n, \pi), \\ &\mathcal{E}(X) = n\pi, \quad \mathrm{Var}(X) = n\pi(1 - \pi), \quad \sigma(X) = \sqrt{n\pi(1 - \pi)}. \end{aligned}$$

2.5.1 Cumulative probability distributions

In some situations it is more convenient to express a distribution by the so-called **cumulative distribution function** (**CDF**), rather than by the "list" of values of $P(X = x)$ for all x:

$$F(x) = P(X \le x) = \sum_{k \le x} P(X = k).$$

The function $F(\cdot)$ is monotone and increasing (not strictly, though), and it has the following limit behaviour:

$$F(-\infty) = 0, \quad F(+\infty) = 1.$$

See Figure 2.2. Knowing the "list" of values $P(X = x)$ (for all x) is the same as knowing the CDF $F(\cdot)$, as either of them can be derived from the other. If for example X has the range $W_X = \{0, 1, \dots, n\}$, then $P(X = x) = F(x) - F(x - 1)$ ($x = 1, 2 \dots, n$) and $P(X = 0) = F(0)$.

2.6 The Poisson distribution (Stahel, ch. 5.2)

The range of the Binomial(n, π) distribution is $W = \{0, 1, \dots, n\}$. If the range of a random variable cannot be restricted a priori to a bounded set, then the Poisson distribution is an option, at least for count data.

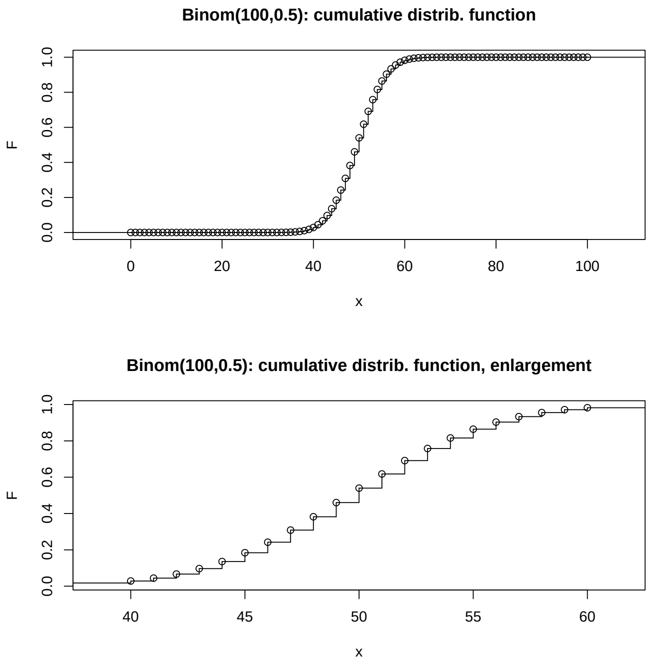

Figure 2.2: Cumulative distribution function $F(\cdot)$ for $X \sim$ Binomial(100,0.5). Below: enlargement for $x \in [40, 60]$.

A random variable X with range $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$ has a Poisson(λ) distribution if

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

using $\lambda > 0$ as a parameter. The Poisson distribution is the standard distribution for unbounded **count data**.

Examples: The Poisson(λ) distribution can used to model the distribution of a random variable X in the following cases:
$X =$ Number of claims by an insuree within a year
$X =$ Number of spontaneous events in a nerve cell within a second
(by the release of transmitters at a synapse)

The characteristic numbers of the Poisson distribution are:

$$\mathcal{E}(X) = \lambda, \ \ \mathrm{Var}(X) = \lambda, \ \ \sigma(X) = \sqrt{\lambda}.$$

2.6.1 Poisson approximation of the binomial distribution

Regard $X \sim$ Binomial(n, π) and $Y \sim$ Poisson(λ). If n is large, π is small and $\lambda = n\pi$, then we can do the following approximation:

$$P(X = x) = \binom{n}{x}\pi^x(1-\pi)^{n-x} \approx P(Y = x) = \exp(-\lambda)\frac{\lambda^x}{x!} \ (x = 0, 1, \ldots, n).$$

This means that for large n and small values of π, Binomial$(n, \pi) \approx$ Poisson(λ) for $\lambda = n\pi$. In other words, the Poisson distribution can be interpreted as the distribution of **rare, independent events** (rare in each individual case, but possibly occuring in large numbers otherwise.

Chapter 3

Statistics for count data

3.1 Three key questions in statistics (Stahel, Kap. 7.1)

A basic concern of statistics and statisticians is the use of one or more observations to make inferences about one (or more) parameters in a probability model.

Example (cont.): Let $x = 11$ be the actual number of female calves bred by the sperm sexing method (cf. chapter 2.4). We regard $x = 11$ as a **realization** of the random variable $X \sim$ Binom$(12, \pi)$, and would like to draw conclusions about the unknown parameter π based on the observation $x = 11$.

First key question: Which parameter value is the most plausible given the observations? The answer to this question is **(point) estimation**.

Second key question: Are the observations (statistically) compatible with a given parameter value? The answer to this second question is **statistical testing**.

Third key question: Which parameter values are (statistically) compatible with the observations? The answer to this question takes the form of a **confidence interval**. Confidence intervals are more general and informative than statistical tests.

Example (cont.): In the sperm sexing example, the three key questions could be formulated as follows:
1. Which is the most plausible value for the parameter π (given the observation $x = 11$)?
2. Is the observation $x = 11$ compatible with the parameter $\pi = 0.7$?
3. Which set (interval) for the parameter π is compatible with the observation $x = 11$?

3.2 Estimation, statistical testing and confidence intervals for the binomial distribution (Stahel, ch. 7.2, 8.2, 9.1, 9.2)

Regard the following situation: we have an observation x given as a realization of $X \sim$ Binomial(n, π). From this, we would like to make inferences about the unknown parameter π.

3.2.1 (Point) Estimation

There is a very practical way of finding an estimate of π. As $\mathcal{E}(X) = n\pi$ (see chapter 2.5), we can express π as $\mathcal{E}(X)/n$. The value of n can be assumed to be known (as it is the number of independent replications of the measurement), and thus the only unknown quantity is $\mathcal{E}(X)$. One pragmatic estimate is then given by $\widehat{\mathcal{E}(X)} = x(=$ observation), i.e. by equating the (single) observation and its expectation. We thus obtain:

$$\hat{\pi} = x/n.$$

3.2.2 Statistical testing

Example: We toss a coin 100 times.
Regard the random variable $X =$ number of heads (K) from 100 tosses of a coin. A reasonable model for this is $X \sim$ Binomial$(100, \pi)$. We observe (the realization) $x = 58$ and would like to test whether the coin is fair, i.e. whether $\pi = 1/2$.

Motivation

To design a test, we can reason as follows: Assume the coin is fair, i.e. $\pi = 1/2$, and compute the probability of “implausible” events of the form $\{X \geq c\}$ for “large” values of c. The aim is to quantify whether or not the observation $x = 58$ belongs to an “implausible” event (whence the conclusion might be drawn that the coin is unfair, that is: $\pi > 1/2$). The following table gives us the relevant probabilities for $X \sim$ Binomial$(100, 1/2)$:

	$c = 52$	$c = 53$	$c = 54$	$c = 55$	$c = 56$	$c = 57$	$c = 58$	$c = 59$	$c = 60$
$P(X \geq c)$	0.382	0.309	0.242	0.184	0.136	0.097	0.067	0.044	0.028

Usually an event is declared to be “implausible” if its probability does not exceed 5%. In our example we can see that the event $X \geq 58$, which just about contains our observation $x = 58$, has probability 6.7% and is thus still plausible. In other words, the observation $x = 58$ can be regarded as plausible when a fair coin is used. However, if we had obtained 59 heads, we would no longer have seen this as a plausible event, as the corresponding probability is 4.4%, which is insufficient. Of course the probability level 5% is an arbitrary cutoff point, and we shall later characterize this by the so-called p-value.

Formal procedure

A statistical test of the parameter π in the model $X \sim$ Binomial(n, π) has the following structure:

1. Specify the so-called **null hypothesis** H_0:

$$H_0: \ \pi = \pi_0,$$

and (taking the exact formulation of the underlying problem into account) a so-called **alternative hypothesis** H_A:

$$\begin{aligned} H_A: \quad & \pi \neq \pi_0 \text{ (two-sided)} \\ & \pi > \pi_0 \text{ (one-sided and open above)} \\ & \pi < \pi_0 \text{ (one-sided and open below).} \end{aligned}$$

Example(cont.): We toss a coin 100 times.
Regard the random variable $X =$ number of heads from 100 tosses, for which we have the model $X \sim$ Binomial$(100, \pi)$. We would like to investigate the fairness of the coin. Is it biased towards turning up heads too often? In the language of statistical testing, this means: $H_0 : \pi = \pi_0 = 1/2$ and $H_A : \pi > \pi_0 = 1/2$. In particualar, the choice of the alternative must come from the question under investigation.

2. Fix the **significance level** α. Typical choices would be $\alpha = 0.05$ (5%) or $\alpha = 0.01$ (1%).

3. Determine the **rejection region** K. Qualitatively speaking, K should be directed towards the alternative hypothesis:

$$\begin{aligned} K = [0, c_u] \cup [c_o, n] & \qquad \text{if } H_A : \pi \neq \pi_0, \\ K = [c, n] & \qquad \text{if } H_A : \pi > \pi_0, \\ K = [0, c] & \qquad \text{if } H_A : \pi < \pi_0. \end{aligned}$$

At a quantitative level K is chosen to satisfy the equation

$$P_{H_0}(X \in K) = \underbrace{P_{\pi_0}}_{\text{from Binomial}(n,\pi_0)} (X \in K) \overset{\approx}{\leq} \alpha. \tag{3.1}$$

Example (cont.): In the example of 100 coin tosses we had $H_0 : \pi = 1/2$ (i.e. $\pi_0 = 1/2$) and $H_A : \pi > 1/2$. For a test at level $\alpha = 0.05$ we have already seen in the above table that $K = [59, 100]$.

4. Now – and only now – look at whether the observation x lies in the rejection region:
if it does: reject H_0 (which means that the alternative hypothesis is "significant")
if it does not: stay with H_0 (which does not imply that H_0 has been proven statistically).
This kind of testing is founded on the principle of contradiction; a statistical proof is only possible when the null hypothesis can be rejected. This type of scientific inductive reasoning was already propagated by Aristotle in the classical era.

Example (cont.): In the example of 100 coin tosses we observed $x = 58$ and will therefore not be able to reject H_0. This means that (at the significance level $\alpha = 0.05$) there is no evidence the coin is biased towards turning up heads (K).

Example (cont.): When using sperm sexing (cf. chapter 2.4), $x = 11$ out of $n = 12$ calves bred were female. Those marketing this method claim that the probability of success exceeds 70%. The test of this can be carried out as follows:
Model: $X \sim$ Binomial$(12, \pi)$
$H_0 : \pi = \pi_0 = 0.7$
$H_A : \pi > \pi_0 = 0.7$
Significance level: take $\alpha = 0.05$
Rejection region: $P_{\pi=0.7}(X \in K) \overset{\approx}{\leq} 0.05 \rightsquigarrow K = \{12\}$
Conclusion: Keep H_0, i.e. the marketing claim is not significant in the light of the evidence.

Type I and Type II errors

A statistical test can be susceptible to two types of errors.
Type I errors: Mistakenly rejecting H_0 despite H_0 being true.
Type II errors: Mistakenly keeping H_0 despite the alternative being true.
A type I error is considered "worse" than a Type II one and is directly controlled by the construction of the test: we have (cf. formula (3.1)):

$$P(\text{Type I error}) = P_{H_0}(X \in K) \overset{\approx}{\leq} \alpha.$$

Thus the significance level controls the probability of a Type I error. We also know that:

$$P(\text{Type II error}) \text{ increases if } \alpha \text{ is made smaller.}$$

So the choice of α involves a compromise between Type I and Type II errors. As the primary aim is to avoid Type I errors, it is α that is kept especially small, e.g. at $\alpha = 0.05$.

Example (cont.): In the sperm sexing setup, assume now that the true value of the parameter π is $0.8 \in H_A$ (the test above was specified for $H_0: \ \pi = 0.7$, $H_A: \ \pi > 0.7$ and $\alpha = 0.05$). As the rejection region for this test is $K = \{12\}$ (see above), we then have

$$P(\text{test keeps } H_0 \text{ even though } \pi = 0.8) = P_{\pi=0.8}(X \leq 11) = 1 - P_{\pi=0.8}(X = 12) = 0.93.$$

This means that a Type II error has a very high probability of occurring (assuming that $\pi = 0.8$). While this is naturally very disappointing, it is fairly inevitable for small samples sizes such as 12. Note, though, that the probability of a type I error nonetheless is $\overset{\approx}{\leq} 0.05$.

P-values

The decision whether to "reject" or "keep" the null hypothesis H_0 depends on the somewhat arbitrary choice of the significance level α. On a mathematical level, this means that the rejection region $K = K(\alpha)$ is dependent on the choice of α.

The following qualitative conclusion is fairly clear:

The rejection region $K = K(\alpha)$ becomes smaller when α decreases$-$

as a small value for α signifies a small probability for a type I error (and this is the case when rejecting the null hypothesis H_0 is difficult, i.e. the rejection region is small). Conversely, $K = K(\alpha)$ becomes ever larger as α increases. This implies the existence of a significance level at which the null hypothesis H_0 is "just about" rejected.

The p-value of a statistical test is defined a the smallest significance level
at which the null hypothesis H_0 is (still) rejected

We can compute the p-value of a test by postulating that the observation $X = x$ (which we know) lies on the boundary of the rejection region $K = K(\text{p-value})$, where the significance level = the p-value; cf. Figure 3.1.

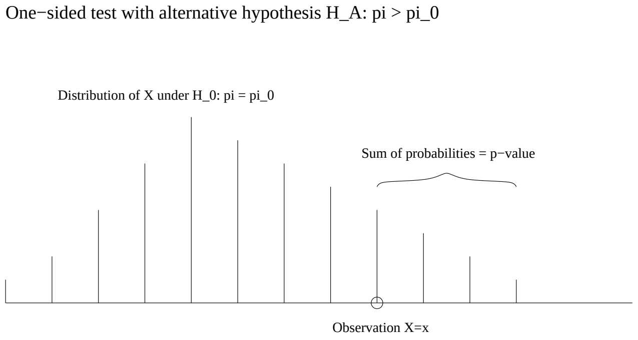
One−sided test with alternative hypothesis H_A: pi > pi_0
Distribution of X under H_0: pi = pi_0
Sum of probabilities = p−value
Observation X=x

Figure 3.1: The p-value of a test with one-sided alternative hypothesis $H_A: \ \pi > \pi_0$.

A p-value is more informative than the mere decision taken at some pre-specified significance level α (e.g. $\alpha = 0.05$). In particular the definition of the p-value implies that we

$$\begin{aligned} &\text{reject } H_0 \text{ if p-value} \leq \alpha \\ &\text{keep } H_0 \text{ if p-value} > \alpha. \end{aligned}$$

In addition to this simple decision rule, a p-value quantifies how significant an alternative hypothesis is (i.e. how much evidence there is for the rejection of H_0). This sometimes put into words in the following way:

$$\begin{aligned} \text{p-value } &\approx 0.05: \text{ weakly significant} \\ \text{p-value } &\approx 0.01: \text{ significant} \\ \text{p-value } &\approx 0.001: \text{ highly significant} \\ \text{p-value } &\leq 10^{-4}: \text{ extremely significant} \end{aligned}$$

Example (cont.): In the sperm sexing example we use the null hypothesis $\pi = 0.7$ and the alternative hypothesis $\pi > 0.7$. For the observation $x = 11$, a realization of the random variable $X \sim$ Binomial$(12, \pi)$, we have the p-value

$$P_{\pi=0.7}(X \geq 11) = P_{\pi=0.7}(X = 11) + P_{\pi=0.7}(X = 12) = 0.085.$$

As previously seen, this does not lead to the rejection of H_0 at the significance level $\alpha = 0.05$. (If - for whatever reason - the significance level $\alpha = 0.09$ had been fixed in advance, H_0 would now be rejected at this significance level $\alpha = 0.09$).

3.2.3 Confidence intervals

One instrument which is more informative than a statistical test is a so-called confidence interval. It suggests an answer to the 3rd basic question of Chapter 3.1: Which values of π are (statistically) compatible with the observation x?

A confidence interval I at level $1 - \alpha$ consists of all parameter values that are compatible with the observation in terms of the statistical test at level α (in general, the two-sided test is used here). Mathematically speaking, this is expressed as:

$$I = \{\pi_0;\ \text{null hypothesis } H_0:\ \pi = \pi_0 \text{ is kept}\}. \tag{3.2}$$

This constitutes a kind of duality between tests and confidence intervals.

The computation of a confidence interval can be performed graphically or by using a table. If the sample size n is "large", a so-called normal approximation (cf. Chap. 4.5) may be used. The latter yields the following confidence interval I at level $1 - \alpha = 0.95$ for the unknown parameter π:

$$I \approx \frac{x}{n} \pm 1.96\sqrt{\frac{x}{n}(1 - \frac{x}{n})\frac{1}{n}} \tag{3.3}$$

The confidence interval $I = I(x)$ depends on the observed value x. If the corresponding random variable X is plugged in instead of the observation, $I(X)$ is random and has the following property:

$$P(\pi \in I(X)) \overset{\approx}{>} 1 - \alpha.$$

This can be interpreted in the following way: the true parameter value π has probability $1 - \alpha$ of being contained in the confidence interval I.

Example (Cont.): For the sperm sexing procedure, the two-sided confidence interval obtained at level $1 - \alpha = 0.95$ by reference to a table or by using a computer to compute (3.2) is

$$I = (0.615, 0.998)\ .$$

In other words, the true "breeding" parameter π has a probability of 95% of being contained in I. Thus due to the small sample size, there is still substantial uncertainty as to the long-term success of the procedure. The approximating formula in (3.3) is not suited to this example, as the sample size $n = 12$ is rather small. If it were used nonetheless, it would yield the "confidence interval"

$$I \approx (0.760, 1.073)\ .$$

Here the right endpoint is obviously too large, as the parameter π cannot be greater than 1.

3.3 Estimation, statistical testing and confidence intervals for the Poisson distribution (Stahel, ch. 7.2, 8.1, 9.1)

Consider the following situation: let an observation x be given and understood to be a realization of the random variable $X \sim$ Poisson(λ). We would like to make inferences about the unknown parameter λ.

3.3.1 (Point) Estimation

As $E(X) = \lambda$ (cf. Chap. 2.6), we can use the pragmatic estimation of $E(X)$ by the observation x to obtain the following estimate of λ:

$$\hat{\lambda} = x.$$

3.3.2 Statistical testing

A statistical test for the parameter λ in the model $X \sim$ Poisson(λ) can be carried out in complete analogy to the one for the binomial distribution in Chapter 3.2.2.

1. Specify the **null hypothesis** H_0:

$$H_0: \ \lambda = \lambda_0,$$

and an **alternative hypothesis** H_A that suits the problem at hand:

$$\begin{aligned} H_A: \quad & \lambda \neq \lambda_0 \text{ (two-sided)} \\ & \lambda > \lambda_0 \text{ (one-sided and open above)} \\ & \lambda < \lambda_0 \text{ (one-sided and open below).} \end{aligned}$$

2. Fix the **significance level** α, e.g. $\alpha = 0.05$.

3. Determine the **rejection region** K. Qualitatively speaking, K should be directed towards the alternative hypothesis:

$$\begin{aligned} K = [0, c_u] \cup [c_o, \infty) \quad & \text{if } H_A: \ \lambda \neq \lambda_0, \\ K = [c, \infty) \quad & \text{if } H_A: \ \lambda > \lambda_0, \\ K = [0, c] \quad & \text{if } H_A: \ \lambda < \lambda_0. \end{aligned}$$

At a quantitative level K is chosen to satisfy the equation

$$P_{H_0}(X \in K) = \underbrace{P_{\lambda_0}}_{\text{from Poisson}(\lambda_0)} (X \in K) \overset{\approx}{\leq} \alpha.$$

4. Now – and only now – look at whether the observation x lies in the rejection region:
if it does: reject H_0;
if it does not: stay with H_0.

The concepts of Type I and Type II errors are identical to those in Chapter 3.2.2.

3.3.3 Confidence intervals

The confidence interval I at level $1 - \alpha$ (cf. Chapter 3.2.3) consists of all values of the parameter λ that are accepted by the corresponding statistical test. Sometimes the following approximate two-sided confidence interval at level $1 - \alpha = 0.95$ may also be of use:

$$I = I(x) \approx x \pm 1.96\sqrt{x}.$$

Example: During the year 1992, there were $x = 554$ deaths at road accidents in Switzerland. We can take this sum to be a realization of $X \sim$ Poisson(λ). The estimate for λ is then $\hat{\lambda} = 554$, and the corresponding confidence interval is $I = I(x) \approx (507.9, 600.1)$.

Chapter 4

Models and statistics for continuous data

4.1 Introduction

Many applications do not involve count data but measured data, which in principle have continuous values. As an illustration we consider two sets of data. In the first set, we compare two methods of determining the latent heat of melting ice. Repeated measurements of energy released during the transition of ice at −0.72° C to liquid water at 0° C have yielded the following data (in cal/g):

Method A	79.98	80.04	80.02	80.04	80.03	80.03	80.04	79.97	80.05	80.03
Method A	80.02	80.00	80.02							
Method B	80.02	79.94	79.98	79.97	79.97	80.03	79.95	79.97		

Although these measurements have been performed with the utmost care, and all confounding influences have been removed, the measurements nonetheless exhibit a certain degree of variance. We shall model this variance within the series of measurements as random, i.e. as realizations of random variables. Afterwards we will be able to answer the question as to whether the measurements are random – or whether it is more plausible that a systematic difference exists between these methods, one that would be visible in the whole population and thus in any further measurements. In the latter case, we shall quantify this systematic difference.

In the second example the aggregation of platelets in the blood was measured in 11 individuals before and after smoking a cigarette. The following data quantify the proportion (in percent) of blood platelets aggregated after stimulation.

Individual	1	2	3	4	5	6	7	8	9	10	11
Before	25	25	27	44	30	67	53	53	52	60	28
After	27	29	37	56	46	82	57	80	61	59	43

Once more, the data fluctuate in an unpredictable manner. This time, however, the variation is not so much due to measurement errors, but rather to variation between individuals (presumably there would also be a certain amount of variation for each individual, if repeated measurements were carried out). Most – but not all – of the people in the study exhibit higher rates of aggregation after smoking, and the main question is whether this effect is random (and thus specific to the sample) or systematic (thus applying to the wider population). In the latter case we would again like to quantify the average increase as far as possible.

4.2 Descriptive statistics (Stahel, ch. 2 and 3.1, 3.2)

For statistical analyses, it is important not to merely fit a model blindly or apply a statistical technique without thinking. The data should always be displayed graphically in a suitable way, as only this permits the discovery of unknown structures and peculiarities. Certain characteristic numbers can also give rough characteristics of a dataset.

In what follows, the data will generally be referred to as $x_1, \ldots, x_n$.

4.2.1 Characteristic Numbers

Often the aim is to provide a numerical summary of the distribution of the data. For this at least two characteristic numbers are needed: one for location and one for spread. The best-known characteristic numbers of these types are the *arithmetic mean*

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

to describe the location and the *empirical standard deviation*

$$s_x = \sqrt{\text{var}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

to describe the spread. (The denominator $n-1$ is taken instead of n for mathematical reasons, to stop the estimate from having a "systematic" error.)

Some alternative characteristic numbers are the *median* as a measure of location and the *interquartile range* as a measure of spread. These two are defined using quantiles.

Quantiles

The *empirical α quantile* is the value which $\alpha \times 100\%$ of the data are smaller than and $(1-\alpha) \times 100\%$ of the data greater than.

For its formal definition we shall need the ordered values:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}.$$

Now the empirical α quantile can be defined as

$$\begin{array}{ll} \frac{1}{2}(x_{(\alpha n)} + x_{(\alpha n+1)}) & \text{if } \alpha \cdot n \text{ is an integer,} \\ x_{(k)} \text{ where } k = \text{next integer greater than } \alpha \cdot n; & \text{if } \alpha \cdot n \text{ is not an integer.} \end{array}$$

The (empirical) median is the empirical 50% quantile, i.e. it marks the "middle" observation and is a measure of the location of the data.

The interquartile range is defined as

$$\text{empirical 75\% quantile} - \text{empirical 25\% quantile} \ ,$$

and is a measure of the spread of the data.

The advantage of the median and interquartile range lies in their robustness: they are less susceptible to the influence of extreme observations than the arithmetic mean and standard deviation.

Example: Measuring the latent heat of melting ice by Method A
On the basis of $n = 13$ measurements we obtain the arithmetic mean $\overline{x} = 80.02$ and the standard deviation $s_x = 0.024$. Furthermore, we have $0.25n = 3.25$, $0.5n = 6.5$ and $0.75n = 9.75$ – which means that the 25% quantile is $x_{(4)} = 80.02$, the median $x_{(7)} = 80.03$ and the 75% quantile $x_{(10)} = 80.04$.

Standardizing

By shifting and scaling their values, we can make two or more sets of data have the same location and spread. In particular, we can standardize a dataset in such a way that its mean becomes zero and its standard deviation 1. This is achieved by the linear transformation

$$z_i = \frac{x_i - \overline{x}}{s_x} \ (i = 1, \ldots, n) \ .$$

All those properties of a distribution that are invariant under shifts and scaling constitute the shape of a distribution. One of these properties is the skewness (asymmetry) of a distribution, which can also be quantified by characteristic numbers.

4.2.2 Graphical methods

One method of gaining an overview of the data is the *histogram*. To plot a histogram, we form classes $(c_{k-1}, c_k]$ and ascertain the frquency h_k of data lying in each of these intervals. Then for each class, we plot a bar whose area is *proportional* to its corresponding h_k.

A *boxplot* is a rectangle bounded by the 25% and 75% quantiles, which additional lines reaching out to the smallest and greatest "normal" values, respectively (where by definition the a normal value is inside the box or outside it by at most 1.5 times the interquartile range). In addition to this, outliers (any data other than the "normal" values) are represented by stars, and the median by a line. The boxplot is especially suited to comparisons of a variable that appears in different groups (which generally represent different types of experimental conditions); cf. Figure 4.1.

The *empirical cumulative distribution function* $F_n(\cdot)$ is a step function that is zero below $x_{(1)}$ and jumps by $\frac{1}{n}$ at each $x_{(i)}$ (values that appear multiple times lead to jumps that are multiples of $\frac{1}{n}$). In other words:

$$F_n(x) = \frac{1}{n} \cdot \text{number of} \{i \mid x_i \leq x\}.$$

Figure 4.2 shows the empirical cumulative distribution function for the measurements of the latent heat of melting ice when using Method A.

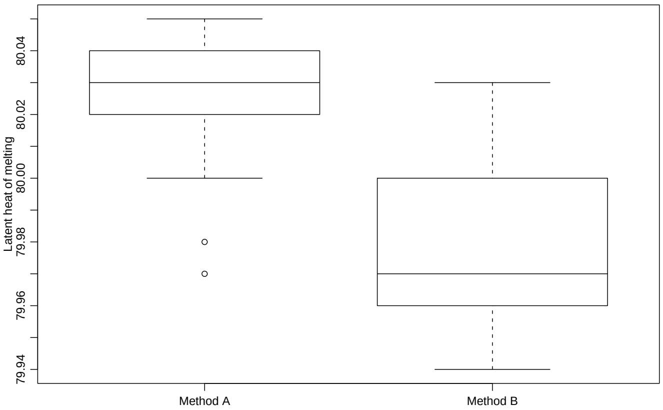

Figure 4.1: Boxplots for the two methods of determining the latent heat of melting ice.

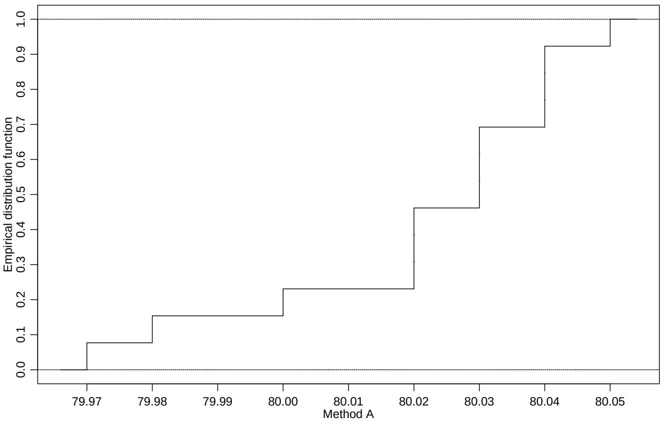

Figure 4.2: Empirical cumulative distribution function for the measurements of the latent heat of melting ice when using method A.

Several variables

When we measure two different quantities, i.e. we have data of the form $(x_1, y_1), \dots (x_n, y_n)$, our main interest lies in the connections and dependencies between these variables. These can be observed in the *scatter plot*, which displays the data as points in the plane: the i-th observation corresponds to the point with the coordinates (x_i, y_i). Figure 4.3 shows a scatter plot for the values "before" and "after" in the blood platelet aggregation study. There evidently is a clear monotone dependency; thus individuals can tend to strong or weak aggregation regardless of their smoking habit.

The most common numerical summary of dependence is given by the **empirical correlation** r (also denoted by $\hat{\rho}$):

$$r = \frac{s_{xy}}{s_x s_y}, \quad s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}.$$

The empirical correlation is a scalar in the interval $[-1, +1]$. Its sign indicates the direction of the linear dependence of x and y, and its absolute value measures the strength of this dependence. In the example on aggregation of platelets in the blood, the empirical correlation is 0.9, which matches the impression we get from the scatterplot. When computing

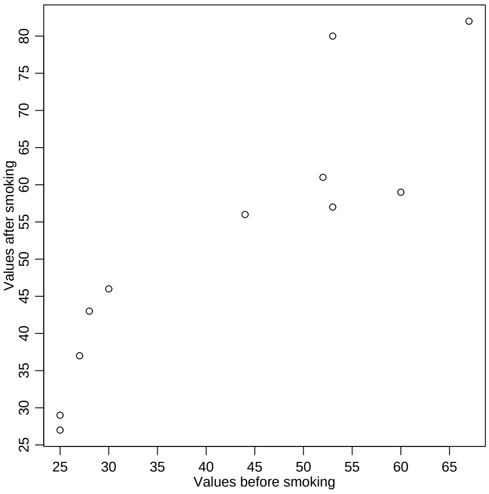

Figure 4.3: Scatter plot of blood platelet aggregation before and after smoking a cigarette.

r, it is important to take the scatterplot into account too, as very different structures in the data can yield the same value of r. More on this can be found in Chapter 5.1.

4.3 Continuous random variables and distributions (Stahel, ch. 6.1 – 6.4, 11.2)

A random variable X is *continuous* if its range W_X is continuous, e.g. $W_x = \mathbb{R}$, $\mathbb{R}^+$ or $[0, 1]$.

In Chapter 2.3 we learned how to describe the probability distribution of a discrete random variable by means of the "point" probabilities $P(X = x)$ for all x in its range. For any continuous random variable X, however, we know that

$$P(X = x) = 0 \text{ for all } x \in W_X \ .$$

This makes it impossible to describe the distribution of X using point probabilities.

We can however describe the distribution of a continuous random variable X by the probabilities of all intervals $(a, b]$ $(a < b)$:

$$P(X \in (a, b]) = P(a < X \leq b) \ .$$

This information is also contained in the cumulative distribution function $F(x) = P(X \leq x)$:

$$P(a < X \leq b) = F(b) - F(a) \ .$$

Altogether this means that we can describe the distribution of a continuous random variable X by means of its cumulative distribution function.

4.3.1 (Probability) Densities

At an infinitesimal level, the concept of a "point" probability $P(X = x)$ can also be described for continuous random variables.

The (probability) density $f(\cdot)$ is defined as the derivative of the cumulative distribution function:

$$f(x) = F'(x) \ .$$

This leads us to the following interpretation:

$$P(x < X \leq x + h) \approx hf(x) \ \text{ if } h \text{ is small.}$$

This is due to an approximation stemming from the definition of the dervative:

$$P(x < X \leq x + h)/h = (F(x + h) - F(x))/h \approx f(x) \ .$$

As $F(x) = \int_{-\infty}^{x} f(y)dy$ is its integral, we obtain the following properties of the density:

1. $f(x) \geq 0$ for all x (as $F(\cdot)$ is an increasing function)
2. $P(a < X \leq b) = \int_a^b f(x)dx$
3. $\int_{-\infty}^{\infty} f(x)dx = 1$ (due to 2.)

Characteristic Numbers of continuous distributions

For a continuous random variable X, we can also define a mean $\mathcal{E}(X)$ and standard deviation σ_X. Their meaning is the same as in the discrete case in Chapter 2.5; only their computation differs. We have:

$$\begin{aligned}\mathcal{E}(X) &= \int_{-\infty}^{\infty} xf(x)dx \ , \\ \mathrm{Var}(X) &= \int_{-\infty}^{\infty} (x - \mathcal{E}(X))^2 f(x)dx, \quad \sigma_X = \sqrt{\mathrm{Var}(X)} \ .\end{aligned}$$

The frequentist interpretation of the mean is an ideal value for the arithmetic mean of a random variable (for a large number of samples).

If we transform X by means of a function $g: \mathbb{R} \to \mathbb{R}$ to obtain a new random variable $Y = g(X)$, we get:

$$\mathcal{E}(Y) = \mathcal{E}(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx \ .$$

This tells us that

$$\mathrm{Var}(X) = \mathcal{E}\left((X - \mathcal{E}(X))^2\right) \ ,$$

which incidentally also holds for discrete random variables. The following **computational rules** (valid also for discrete random variables) have proven useful: for any $a, b \in \mathbb{R}$,

$$\begin{aligned}&\mathcal{E}[a + bX] = a + b\,\mathcal{E}[X] \ , \\ &Var(X) = \mathcal{E}[X^2] - (\mathcal{E}[X])^2 \ , \\ &Var(a + bX) = b^2 Var(X) \ .\end{aligned}$$

The **quantiles** (of the distribution of X) $q(\alpha)$ $(0 < \alpha < 1)$ are defined as follows:

$$P(X \leq q(\alpha)) = \alpha \ .$$

This means that

$$F(q(\alpha)) = \alpha \Leftrightarrow q(\alpha) = F^{-1}(\alpha) \ .$$

We can also interpret this by saying that $q(\alpha)$ is the point at which the area under the density curve $f(\cdot)$ from $-\infty$ to $q(\alpha)$ is equal to α. For this cf. Figure 4.4. The 50% quantile is the **median**.

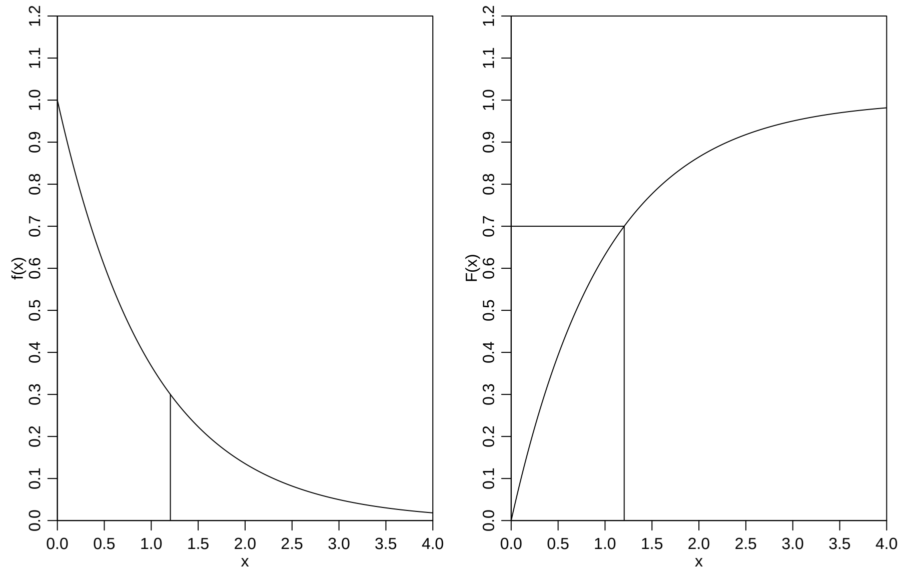

Figure 4.4: Illustration of the 70% quantile $q(0.7)$. On the left we see the density curve, for which the area under the curve from $-\infty$ (or 0) to $q(0.7)$ is equal to 0.7. On the right we see the cumulative distribution function, for which $q(0.7)$ is the vlaue of the inverse at 0.7.

4.4 Important continuous distributions (Stahel, ch. 6.2, 6.4, 6.5, 11.2)

In Chapter 4.3 we saw that we can characterize the distribution of a continuous random variable by its cumulative distribution function $F(\cdot)$ or by its density $f(\cdot)$.

4.4.1 The uniform distribution

The uniform distribution appears when describing rounding errors and when formalizing complete "ignorance".

A random variable X with range $W_X = [a, b]$ has a Uniform($[a, b]$) distribution if

$$f(x) = \begin{cases} 1/(b-a) & \text{if } a \leq x \leq b \\ 0 & \text{else} \end{cases}$$

Here the density is constant on the whole range $W_X = [a, b]$ of X – thence the name "uniform".

The corresponding cumulative distribution function is

$$F(x) = \begin{cases} 0 & \text{if } x < a \\ (x-a)/(b-a) & \text{if } a \leq x \leq b \\ 1 & \text{if } x > b \end{cases}$$

The key characteristic numbers of $X \sim$ Uniform($[a, b]$) are as follows:

$$\begin{aligned} &\mathcal{E}(X) = (a+b)/2 \ , \\ &\text{Var}(X) = (b-a)^2/12, \quad \sigma_X = \sqrt{Var(X)} \ . \end{aligned}$$

4.4.2 The exponential distribution

The exponential distribution is the simplest model of waiting times (e.g. until failure).

Example: Ion channels
Within muscle and nerve cell membranes there are many channels that in an open state permit the circulation of ions. Simple kinetic models indicate that the opening times of such channels can be modelled with an exponential distribution.

A random variable X with range $W_X = \mathbb{R}^+ = [0, \infty)$ has an exponential distribution with parameter $\lambda \in \mathbb{R}^+$ (Exp(λ)) if

$$f(x) = \begin{cases} \lambda \exp(-\lambda x), & \text{if } x \geq 0 \\ 0 & \text{else} \end{cases}$$

The corresponding cumulative distribution function is

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

The density and cumulative distribution function for $\lambda = 1$ can be seen in Figure 4.4.

The key characteristic numbers of a random variable $X \sim$ Exp(λ) are as follows:

$$\begin{aligned} &\mathcal{E}(X) = 1/\lambda \ , \\ &\text{Var}(X) = 1/\lambda^2, \quad \sigma_X = \sqrt{Var(X)} \ . \end{aligned}$$

One connection between the exponential and the Poisson distribution is the following: if the times between successive failures of a system are distributed as Exp(λ), then the number of failures in an interval of length t is distributed according to Poisson(λt).

4.4.3 The normal (or Gauss) distribution

The normal distribution (also known as the Gaussian distribution) is the most common distribution taken on by measured values.

Examples: Measurements of light emissions from a "white dwarf" star can be modelled as realizations of normally distributed random variables.

A random variable X with range $W_X = \mathbb{R}$ has a normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$ ($\mathcal{N}(\mu, \sigma^2)$) if

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

The corresponding cumulative distribution function $F(\cdot)$ does not have a closed-form expression, but can only be written by the integral $F(x) = \int_{-\infty}^{x} f(y)dy$.

The key characteristic numbers for $X \sim \mathcal{N}(\mu, \sigma^2)$ are:

$$\mathcal{E}(X) = \mu,$$
$$\text{Var}(X) = \sigma^2, \quad \sigma_X = \sqrt{Var(X)} .$$

Thus the parameters μ and σ^2 possess a natural interpretation as mean and variance of the distribution. Normal distributions with three different combinations of these parameters are shown in Figure 4.5.

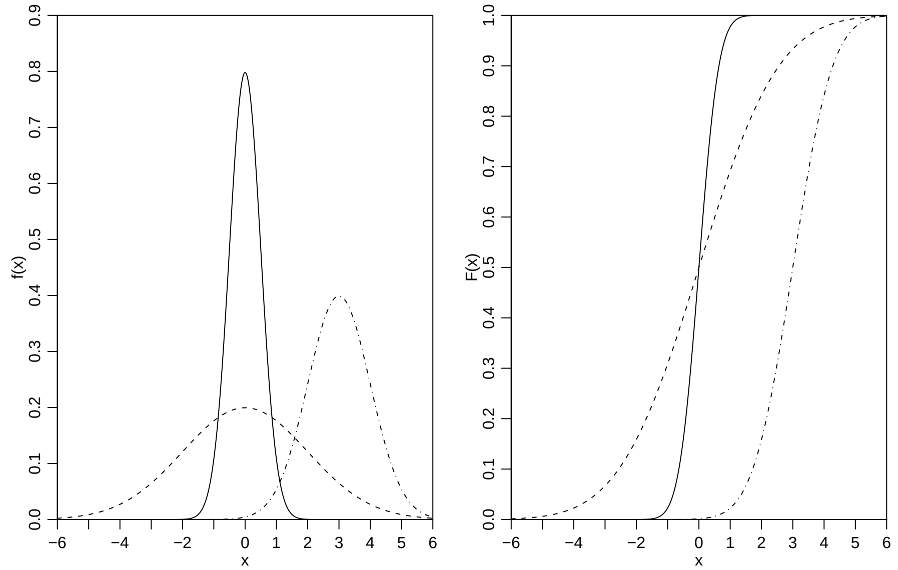

Figure 4.5: Densities (left) and cumulative distribution functions (right) of the normal distributions with parameters $\mu = 0, \sigma = 0.5$ (—), $\mu = 0, \sigma = 2$ (- - - -) and $\mu = 3, \sigma = 1$ (- · - ·).

The standard normal distribution

The normal distribution with parameters $\mu = 0$ and $\sigma^2 = 1$ is known as the starndard normal distribution. Its density and cumulative distribution function have their own notation:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right),$$
$$\Phi(x) = \int_{-\infty}^{x} \varphi(y)dy.$$

The values of the function $\Phi(\cdot)$ are tabulated. We shall see below that any normal distribution $\mathcal{N}(\mu, \sigma^2)$ can be transformed into a standard normal distribution. Thus the values of $\Phi(\cdot)$ will suffice for the calculation of probabilities and quantiles from any general normal distribution $\mathcal{N}(\mu, \sigma^2)$.

4.4.4 Transformations

Sometimes it can be useful to transform a continuous random variable X:

$$Y = g(X) \ ,$$

where $g : \mathbb{R} \to \mathbb{R}$ is a transformation.

Linear transformations

Regard the linear transformation

$$g(x) = a + bx \quad (a,\ b \in \mathbb{R}) \ .$$

The mean and variance of the transformed random variable $Y = g(X)$ are then (cf. chapter 4.3.1):

$$\begin{aligned} &\mathcal{E}(Y) = \mathcal{E}(a + bX) = a + b\,\mathcal{E}(X), \\ &\mathrm{Var}(Y) = \mathrm{Var}(a + bX) = b^2\,\mathrm{Var}(X), \quad \sigma_Y = b\sigma_X \ . \end{aligned} \tag{4.1}$$

Furthermore, the following inequalities hold for $b > 0$:

$$\begin{aligned} &\alpha \text{ quantile of } Y = q_Y(\alpha) = a + bq_X(\alpha), \\ &f_Y(y) = \frac{1}{b} f_X\left(\frac{y-a}{b}\right) \ . \end{aligned} \tag{4.2}$$

Standardizing a random variable

Regard a continuous random variable X. We can always apply a suitable linear transformation to X such that the transformed random variable has mean 0 and variance 1. This can be done in the following manner: take the linear transformation

$$g(x) = -\frac{\mathcal{E}(X)}{\sigma_X} + \frac{1}{\sigma_X} x$$

and define the transformed random variable

$$Z = g(X) = \frac{X - \mathcal{E}(X)}{\sigma_X} \ .$$

Using the rules in (4.1), we then get: $\mathcal{E}(Z) = 0$, $\mathrm{Var}(Z) = 1$.

If $X \sim \mathcal{N}(\mu, \sigma)$, the standardized random variable has a standard normal distribution:

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1) \ .$$

This follows from the calculation rules on linear transformations of densities (4.2). (More generally, we know that any linear transformation of a normal distribution yields another normal distribution. This property of linear transformations keeping the distribution within the same family is a special property which the normal distribution has – for other families of distributions, it does not always hold).

Example: Computing probabilities for $\mathcal{N}(\mu, \sigma^2)$.
We regard $X \sim \mathcal{N}(2, 4)$, and would like to compute $P(X \leq 5)$. To this end, we proceed in the following way:

$$P(X \leq 5) = P(\frac{X-2}{\sqrt{4}} \leq \frac{5-2}{\sqrt{4}}) = P(Z \leq 3/2) \ ,$$

where $Z \sim \mathcal{N}(0, 1)$. Thus

$$P(X \leq 5) = P(Z < 3/2) = 0.933 \ ,$$

a numerical value easily looked up in a table or using a computer.

Applying Rule (4.2) to the computation of the 95% quantile when $a = -\mu/\sigma = -2/2 = -1$ and $b = 1/\sigma = 1/2$, we get:

$$q_X(0.95) = (q_Z(0.95) + 1) \cdot 2 = (\Phi^{-1}(0.95) + 1) \cdot 2 = 5.290 \ .$$

Non-linear transformations

If $g : \mathbb{R} \to \mathbb{R}$ is an arbitrary transformation, many of the properties discussed above become more complicated. However, one formula can be quite useful. For a continuous random variable X with density $f_X(\cdot)$, we have

$$\mathcal{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx \ .$$

For more see chapter 4.3.1.

One frequently-used distribution is the **lognormal distribution**. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Y = \exp(X)$ has a lognormal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$. The lognormal distribution is no longer symmetric, and has mean $\mathcal{E}(Y) = \exp(\mu + \sigma^2/2)$.

4.4.5 Analogies between models and data

Random variables and distributions describe an entire population, i.e. what could happen and with what probability. We interpret data $x_1, \ldots, x_n$ as realizations of random variables $X_1, \ldots, X_n$ (we could look at the n data as n realizations of of a single random variable X), but the notation using several random variables has some advantages (cf. Section 4.5).

We can use data to draw conclusions about the underlying distribution. In particular, all quantities defined for random variables have a counterpart for finite data sets. The following table lists these related quantities, for which the empirical expressions are estimates of the theoretical quantities. As the sample size n increases, these estimates get ever closer to their theoretical counterparts.

Data	Population (Model)
Histogram	Density
Empirical cumulative distribution function	Theoretical cumulative distribution function
Empirical quantile	Theoretical quantile
Arithmetic mean	Mean
Empirical standard deviation	Theoretical standard deviation

4.4.6 Checking assumptions of normality

We often would like to establish whether a given distribution constitutes a useful model for a set of data. In other words, we want to check if the data $x_1, \ldots, x_n$ can be considered as realizations of a random variable X (e.g. with a cumulative distribution function $F(\cdot)$) with this distribution.

We could in principle compare the histogram of the empirical data with the density of the model distribution. However, deviations and similarities are often better seen when looking at quantiles.

Q-Q plots

The idea of a Q-Q (quantile-quantile) plot is to plot the empirical quantiles against the theoretical ones. Specifically: let α run through the sequence $0.5/n, 1.5/n, \ldots, (n-0.5)/n$ and plot the theoretical quantiles $q(\alpha)$ of the model distribution on the x-axis, and the empirical quantiles corresponding to the ordered observations $x_{[1]} < x_{[2]} < \ldots < x_{[n]}$ on the y-axis. If the observations really do stem from the distribution according to the model, these plotted points should lie roughly on the diagonal line $y = x$.

Normal plots

We usually do not want to check just one distribution, but a whole class of them – for instance the class of normal distributions with arbitrary parameters μ and σ.

A Q-Q plot using the standard normal distribution $\mathcal{N}(0,1)$ as a model
is called a normal plot.

If the data are realizations of the random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, then the quantiles of X are:

$$q(\alpha) = \mu + \sigma\Phi^{-1}(\alpha) \ .$$

See also (4.2). Thus for such a distribution, the points in any normal plot should lie on the line $\mu + \sigma \cdot x$. Figure 4.6 contains two normal plots: one generated by a normal distribution, and one case where the data stems from a heavy-tailed distribution. Further illustrated examples can be found in Figure 11.2 of Stahel's book.

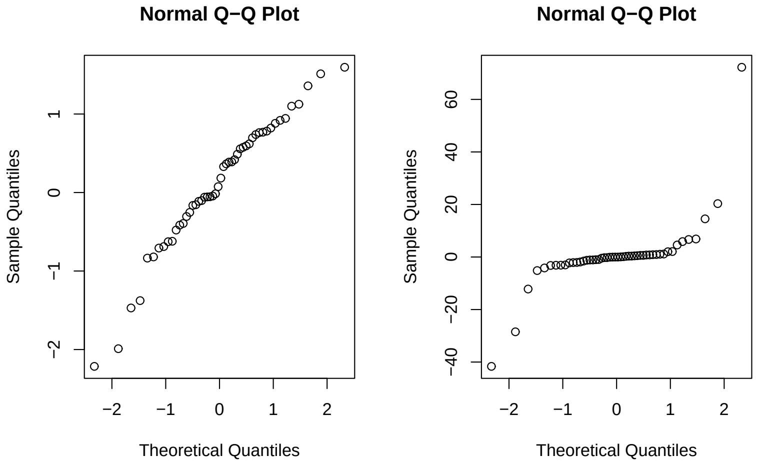

Figure 4.6: Left: Normal plot for 50 realizations of $\mathcal{N}(0,1)$. Right: Normal plot for 50 realizations of the Cauchy distribution (very heavy-tailed).

4.5 Functions of random variables and propagation of errors

(Stahel, ch. 6.8 – 6.11)

Most applications contain not just one, but several random variables. Typically, the same quantity is measured several times (either by having several individuals, or by repeating the measurements).

We consider the measurements $x_1, x_2, \ldots, x_n$ to be realizations of the random variables $X_1, \ldots, X_n$. This notation is often more convenient than interpreting the measurements as n independent realizations of one random variable X. Often the quantity of interest is some function of $X_1, \ldots, X_n$:

$$Y = g(X_1, \ldots, X_n) \ ,$$

where $g : \mathbb{R}^n \to \mathbb{R}$ is a map and Y a further random variable. The main example we have in mind is the map

$$\overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i \ .$$

Consider the following connection: if the x_i are realizations of the random variables X_i, then their arithmetic mean $\overline{x}_n = n^{-1} \sum_{i=1}^{n} x_i$ is a realization of the random variable $\overline{X}_n$.

Our interest lies in the distribution of the random variables $\overline{X}_n$ (knowing this distribution is an important tool for later using statistics based on the arithmetic mean of data). To this end, the following assumption is typically made:

The i.i.d. assumption

We often assume that the random variables $X_1, \ldots, X_n$ are **independent** and **identically distributed**. This is denoted by the following abbreviation:

$$X_1, \ldots, X_n \text{ i.i.d. } .$$

Example:

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

means that the X_i's are independent and all have the same normal distribution $\mathcal{N}(\mu, \sigma^2)$.

Characteristic Numbers and distribution of $\overline{X}_n$

In this part, we shall assume that

$$X_1, \ldots, X_n \text{ i.i.d. } \sim \text{ cumulative distribution function } F.$$

Due to the second "i" in i.i.d. , each X_i has the same distribution and thus the same mean and variance: $\mathcal{E}(X_i) = \mu$, $\mathrm{Var}(X_i) = \sigma_X^2$.
For $\overline{X}_n$ we then have:

$$\begin{aligned} \mathcal{E}(\overline{X}_n) &= \mu \ , \\ \mathrm{Var}(\overline{X}_n) &= \frac{\sigma_X^2}{n} \ . \end{aligned}$$

Thus the mean of $\overline{X}_n$ is the same as the mean of each individual X_i, while the *variance decreases in n*. This leads to the following rule:

Law of large numbers: if $X_1, \ldots, X_n$ i.i.d. , then

$$\overline{X}_n \longrightarrow \mu \ (n \to \infty) \ .$$

Example: Throwing a die n times
Regard $X_i =$ outcome of the i-th dice throw. Then

$$\begin{aligned} \overline{X}_n &= \text{ average outcome of } n \text{ throws} \\ &\longrightarrow \mu = \mathcal{E}(X_i) = (1+2+3+4+5+6)/6 = 3.5 \ (n \to \infty). \end{aligned}$$

In other words, the average outcome of a large number of dice throws is close to 3.5.

It generally is rather difficult to write down the distribution of $\overline{X}_n$. One special case is given by the normal distribution:

$$\overline{X}_n \sim \mathcal{N}(\mu, \sigma_X^2/n) \text{ if } X_1, \ldots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu, \sigma_X^2) \ .$$

Surprisingly, the above formula for the distribution still is approximately true even if the individual X_i's are not normally distributed. This we know from the following famous theorem:

Central limit theorem: if $X_1, \ldots, X_n$ i.i.d. , then

$$\overline{X}_n \approx \mathcal{N}(\mu, \sigma_X^2/n) ,$$

an approximation that gets better as n gets larger. Furthermore the approximation is better, the closer the distribution of X_i is to the normal distribution $\mathcal{N}(\mu, \sigma_X^2)$.

The standardized random variable

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma_X}$$

also has the distribution $\mathcal{N}(0, 1)$.

4.6 Single-sample statistics (Stahel, ch. 8.3 – 8.5, 9.3)

We regard data $x_1, \ldots, x_n$, which we consider to be realizations of the random variables $X_1, \ldots, X_n$ i.i.d. . Two characteristic numbers of the variables X_i are $\mathcal{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma_X^2$. Typically these are unknown, as are other key quantities, and we wish to make inferences about them based on the data.

Example: Aggregation of blood platelets (cf. Section 4.1)
Aggregation of blood platelets is an example of a so-called *paired comparison*, where each test object is measured under two different sets of conditions. We are looking to see whether aggregation is significantly different before and after smoking a cigarette. To investigate this, we compute the *differences* x_i = aggregation "after smoking" - aggregation "before smoking" $(i = 1, \ldots, 11)$, and we obtain the sample we are interested in.

4.6.1 (Point) Estimates

The (point) estimates for mean and variance are:

$$\hat{\mu} = n^{-1} \sum_{i=1}^{n} X_i,$$
$$\hat{\sigma}_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 .$$

Note that the estimates are written here as functions of the random variables $X_1, \ldots, X_n$. In particular, $\hat{\mu}$ and $\hat{\sigma}_X^2$ are themselves random variables (we already discussed the distributional properties of $\hat{\mu}$ in Section 4.5. As we consider the data x_i to be realizations of the random variables X_i, the estimates are precisely the arithmetic mean and empirical variance of the data.

4.6.2 Testing μ

Example: Aggregation of blood platelets (cont.)
We would like to test whether there is a systematic difference in aggregation before and

after smoking. As x_i is precisely the difference in aggregation "before" and "after", we can regard the following test setup:

$$H_0: \ \mu = 0, \quad H_A: \ \mu \neq 0 \ .$$

To test the hypothesis about the parameter μ, we first assume that

$$X_1, \ldots, X_n \text{ i.i.d. } \mathcal{N}(\mu, \sigma_X^2) \ . \tag{4.3}$$

Later on, we will consider weakening this assumption.

The z-test

We assume the data $x_1, \ldots, x_n$ to be realizations of (4.3). Furthermore we assume that the variance σ_X^2 is known.

Then the z-test of the parameter μ is as follows:

1. Specify the null hypothesis $H_0: \ \mu = \mu_0$
and the alternative $H_A: \ \mu \neq \mu_0$ (or "<", or ">").

2. Fix a significance level α for the test (e.g. $\alpha = 0.05$).

3. Consider the **test statistic** $\overline{X}_n$. If the null hypothesis is true, then we know that (cf. section 4.5):

$$\overline{X}_n \ \sim \ \mathcal{N}(\mu_0, \sigma_X^2/n) \ .$$

The rejection region for the test statistic $\overline{X}_n$ when using the two-sided alternative $H_A: \mu \neq \mu_0$ then is

$$K = (-\infty, \mu_0 - \Phi^{-1}(1-\alpha/2)\sigma_X/\sqrt{n}] \cup [\mu_0 + \Phi^{-1}(1-\alpha/2)\sigma_X/\sqrt{n}, \infty) \ .$$

Thus some simple algebra shows that

$$P_{H_0}[\overline{X}_n \in K] = P_{\mu_0}[|\overline{X}_n - \mu_0| > \frac{\sigma_X}{\sqrt{n}}\Phi^{-1}(1-\frac{\alpha}{2})] = \alpha \ ,$$

that is: the probability of a Type I error is exactly the significance level α.

4. Reject H_0 if the arithmetic mean $\overline{x}_n$ lies in K (if not, keep H_0).

In summary, the z-test is as follows:

$$\begin{array}{lll} \text{reject } H_0 \text{ if} & |\frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\sigma_X}| > \Phi^{-1}(1-\alpha/2) & \text{using } H_A: \ \mu \neq \mu_0 \ , \\ & \frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\sigma_X} < -\Phi^{-1}(1-\alpha) & \text{using } H_A: \ \mu < \mu_0 \ , \\ & \frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\sigma_X} > \Phi^{-1}(1-\alpha) & \text{using } H_A: \ \mu > \mu_0 \ . \end{array}$$

Unlike the tests in chapter 3.2.2, the z-test uses *several* observations. However, these are summarized by the realization $\overline{x}_n$ of a test statistic – or even in the standardized form

$$z = \frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\sigma_X}$$

(which is a function of the data). Apart from this minor difference, all concepts are the same as in chapter 3.2.2. In particular, we need the distribution of the random variables under the null hypothesis $H_0: \ \mu = \mu_0$ in order to determine the rejection region K:

$$Z = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\sigma_X} \sim \mathcal{N}(0,1) \ .$$

The t-test

As before, we assume the data to be realizations of (4.3). In practice, though, the assumption that we know the variance σ is often unrealistic. We can however use the estimate

$$\hat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \ .$$

This then leads to further uncertainty which must be taken into account.

The t-test uses the test statistic

$$t = \frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\hat{\sigma}_X} \ ,$$

and the distribution of this statistic under the null hypothesis $H_0: \ \mu = \mu_0$ is

$$T = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\hat{\sigma}_X} \sim t_{n-1} \ ,$$

where t_{n-1} is a so-called t distribution with $n-1$ degrees of freedom.

The t_ν distribution is a symmetric distribution around 0 which has heavier tails than the standard normal distribution $\mathcal{N}(0,1)$. For $T \sim t_\nu$, we have:

$$\begin{aligned} \mathcal{E}(T) &= 0 \\ \mathrm{Var}(T) &= \frac{\nu}{\nu - 2} \ . \end{aligned}$$

For large values of ν, t_ν is very similar to $\mathcal{N}(0,1)$. In particular, the t_ν distribution converges to the standard normal distribution $\mathcal{N}(0,1)$ when $\nu \to \infty$. In Figure 4.7, the density of a t_5 distribution can be seen.

In summary, the t-test is as follows:

$$\begin{aligned} \text{reject } H_0: \ \mu = \mu_0 \text{ if} \quad & |t| = |\frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\hat{\sigma}_X}| > t_{n-1,1-\alpha/2} && \text{using } H_A: \ \mu \neq \mu_0 \ , \\ & t = \frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\hat{\sigma}_X} < -t_{n-1,1-\alpha} && \text{using } H_A: \ \mu < \mu_0 \ , \\ & t = \frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\hat{\sigma}_X} > t_{n-1,1-\alpha} && \text{using } H_A: \ \mu > \mu_0 \ , \end{aligned}$$

where $t_{n-1;\alpha}$ denotes the α quantile of the t_{n-1} distribution. This quantile can be found in tables (see e.g. Stahel, Table 8.5.g, p. 187) or can be calculated using a computer. It is somewhat larger than the α quantile of the standard normal distribution, and thus leads to a slightly smaller rejection region. For large sample sizes n, however, the difference is quite minimal (as $t_{n-1} \approx \mathcal{N}(0,1)$ when n is large).

Figure 4.7 illustrates the rejection region of the t-test using $n = 6$ observations. The p-value for the two-sided alternative $H_A: \ \mu \neq \mu_0$ is computed as:

$$p - value = 2\left(1 - F_{t_{n-1}}\left(\frac{\sqrt{n}|\overline{x}_n - \mu_0|}{\hat{\sigma}_X}\right)\right) \ ,$$

where $F_{t_{n-1}}$ denotes the cumulative distribution function of the t distribution with $n-1$ degrees of freedom.

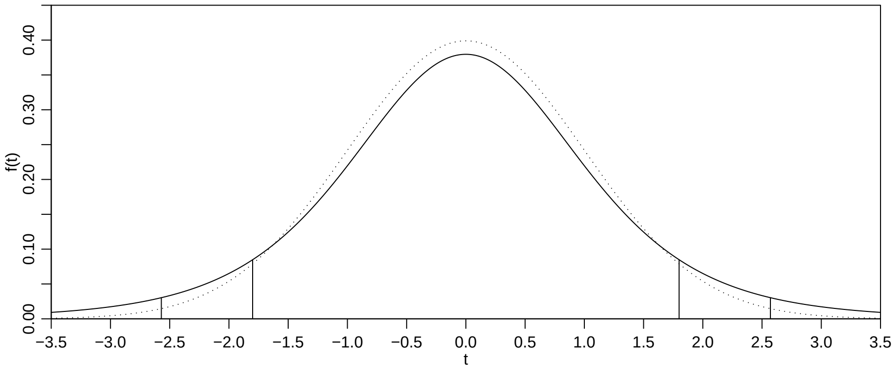

Figure 4.7: Density of the t distribution with 5 degrees of freedom (solid line) and of the standard normal distribution (dashed line). The inner vertical lines mark the inner boundaries of areas equal to half the p-value for hypothetical data with $\sqrt{6}|\overline{x}_6 - \mu_0|/\hat{\sigma}_X = 1.8$; the outer vertical lines mark areas equal to 2.5% each.

Example (cont.) Aggregation of blood platelets (cf. Section 4.1)
Regard the differences x_i = aggregation "before" - aggregation "after" ($i = 1, \ldots, 11$), and consider these to be realizations of $\mathcal{N}(\mu, \sigma_X^2)$. The null and alternative hypotheses of interest are $H_0: \ \mu = \mu_0 = 0$ and $H_A: \ \mu > \mu_0 = 0$. The value of the test statistic is

$$\frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\hat{\sigma}_X} = 4.27 \ ,$$

and the relevant quantile for $\alpha = 0.05$ is $t_{10;0.95} = 1.812$. Thus the outcome of the test is to reject H_0 at a level of 5%. The corresponding p-value is

$$P_{H_0}[T > 4.27] = 1 - F_{10}(4.27) = 0.00082 \ .$$

This means that the influence of smoking cigarettes on the aggregation of blood platelets is highly significant.
If the one-sided alternative were replaced by the two-sided alternative $H_A: \ \mu \neq \mu_0 = 0$, the result would be as follows: the relevant quantile for $\alpha = 0.05$ is $t_{10;0.975} = 2.23$. The outcome of the test would remain the same: reject H_0 at a significance level of 5%. Here the p-value is

$$P_{H_0}[|T| > 4.27] = 2(1 - F_{10}(4.27)) = 0.0016 \ .$$

4.6.3 Confidence intervals for μ

Like with count data as seen in chapter 3.2.3, the confidence interval consists of those values of μ for which the corresponding test does not lead to the rejection of the null hypothesis.

Once again we assume that the data are realizations of (4.3). This then leads to the following two-sided confidence intervals (for which the corresponding tests are two-sided, using the alternative hypothesis $H_A : \mu \neq \mu_0$) at level $1-\alpha$:

$$\begin{array}{ll} \overline{x}_n \pm \Phi^{-1}(1-\alpha/2)\,\dfrac{\sigma_X}{\sqrt{n}} & \text{if } \sigma_X \text{ is known },\\ \overline{x}_n \pm t_{n-1,1-\alpha/2}\,\dfrac{\hat{\sigma}_X}{\sqrt{n}} & \text{if } \sigma_X \text{ is unknown }. \end{array}$$

Example (cont.): Aggregation of blood platelets
We have 10 degrees of freedom and $t_{10,0.975} = 2.23$. The two-sided confidence interval for the increase in aggregation of blood platelets after smoking a cigarette is thus (in terms of increased percentage)

$$I = 10.27 \pm 2.23 \cdot 7.9761/\sqrt{11} = [4.91, 15.63] \ .$$

In particular, the interval I does not contain zero: this means that the value $\mu = 0$ is incompatible with the data (as we already found by means of the t-test).

4.6.4 Testing μ for non-normal data

The z- and t-tests are optimal if the data are realizations of normal random variables, as in (4.3). By optimality we mean that these tests have the best power (see below).

Now we will look at the more general situation in which the data are realizations of

$$X_1, \dots, X_n \text{ i.i.d. } , \tag{4.4}$$

where X_i has an arbitrary distribution. Denote a location parameter of the ditribution by μ (e.g. μ = median of the distribution of X_i). Then the null hypothesis takes on the form $H_0: \ \mu = \mu_0$.

The power of a test

In chapter 3.2.2 we saw that there are two types of errors a test can make:

$$\text{Type I error} = \text{erroneously rejecting } H_0\text{, despite } H_0 \text{ being true,}$$

and

$$\begin{array}{rcl} \text{Type II error } (\mu) & = & \text{(erroneously) keeping } H_0\text{, despite } \mu (\in H_A) \text{ being} \\ & & \text{the correct parameter value.} \end{array}$$

The probability of a Type I error is precisely α; instead of the Type II error (μ), we often look at the power:

$$\text{Power } (\mu) = 1 - P(\text{Type II error } (\mu)) = P(\text{rejecting } H_0 \text{ when } \mu \text{ is true}).$$

For any $\mu \in H_A$, we can interpret the power(μ) as the probability of correctly discovering H_A when $\mu \in H_A$ is the truth. For a test statistic T and its corresponding rejection region K, the following then holds true:

$$\begin{aligned} &P_{\mu_0}(T \in K) = \alpha \ , \\ &\text{Power}(\mu) = P_\mu(T \in K) \ . \end{aligned}$$

The sign test

Regard the situation where the data are realizations of (4.4), but the individual X_i do not follow a normal distribution. Here we denote $\mu =$ median of the distribution of X_i; for a symmetric distribution $\mu = \mathcal{E}(X_i)$ holds.

The sign test uses the following test statistic:

$$V = \text{ number of } X_i \text{ for which } (X_i > \mu_0) \ .$$

Note that $V =$ number of positive signs of $(X_i - \mu_0)$, which gives the test its name.

Regard the null hypothesis $H_0: \ p = P(X_i > \mu_0) = 1/2$ and the alternative hypothesis $H_A: \ p \neq 1/2$ (or one-sided versions of this). Thus under the null hypothesis H_0, the test statistic V has the following distribution:

$$V \sim \text{ Binomial}(n, 1/2) \ ,$$

and the sign test thus becomes a test of the parameter p of a binomial distribution.

The sign test test is always correct when the data are realizations of (4.4): i.e. the probability of a Type I error is controlled by α, whatever the distribution of the X_i. This is not true for the z- and t-tests. Due to the Central Limit Theorem, however, they do keep the probability of a Type I error under control by α for large n at least.

However, the power of the z- and t-tests generally becomes rapidly worse when the X_i in (4.4) no longer have a normal distribution. Thus the sign test often has a higher power than the z- or t-test when the data are non-normal (and not approximately normal, either). One drawback the sign test has, though, is that it does not use the information about how far from μ_0 the X_i are (see the above definition of the test statistic V).

Example (cont.): Aggregation of blood platelets
The null hypothesis is $H_0: \ \mu = \mu_0 = 0$. The realized value of the test statistic is then $v = 10$, and the p-value obtained by using the one-sided alternative $H_A: \ \mu > \mu_0 = 0$ is 0.005 (for the t-test, the p-value was 0.00082).

The Wilcoxon test

The Wilcoxon test is a compromise which does not assume a normal distribution (unlike the t-test), but which also makes better use of the data than the sign test.

The prerequisites for the Wilcoxon test are that the data be realizations of (4.4) and the distribution of the X_i continuous and symmetric (symmetric density around $\mu = \mathcal{E}(X_i)$). The p-value for a one- or two-sided alternative can be calculated using a computer.

The Wilcoxon test is preferable in most cases; it often has a better power than both the t-test and the sign test. Only when the data can be described extremely well by a normal distribution does the t-test remain "completely suitable" for good data analysis; this assumption or condition can be checked e.g. graphically, using a normal plot (cf. chapter 4.4.6).

Example (cont.): Aggregation of blood platelets
The null hypothesis is $H_0 : \mu = \mu_0 = 0$. The p-value when using the one-sided alternative $H_A : \mu > \mu_0 = 0$ is 0.002528.

4.7 Testing with two independent samples (Stahel, ch. 8.8)

One frequent aim in testing is to compare two methods (groups, experimental conditions, treatments) with respect to the location of the distribution.

4.7.1 Paired and unpaired samples

For all applications, it is not only good evaluation of the data which is important, but also good planning of the experiment. We need to ensure that any differences found are really due the methods compared, rather than some other disturbance. The main principles to ensure this are *blocking* and *randomization.*

Randomization means than the order of the experiments and allocation of the experimental subjects to experimental conditions are both random. Then our observations (realizations of random variables) are

$$x_1, x_2, \ldots, x_n \text{ under Condition 1 },$$
$$y_1, y_2, \ldots, y_m \text{ under Condition 2 }.$$

In general – but not always – we have $m \neq n$. Such **random allocation of experimental subjects to experimental conditions** leads to what is called an **unpaired sample**.

Example:
Random allocation of 100 test patients to a group of 60 patients treated using a real drug, and a group of 40 patients undergoing a treatment with placebos.

Example:
Data on latent heat of melting ice in chapter 4.1.

On the other hand, we have a **paired sample** if **both experimental conditions are applied to the same experimental subject**. In such a case, the data have the following structure:

$$x_1, \ldots, x_n \text{ under Condition 1 },$$
$$y_1, \ldots, y_n \text{ under Condition 2 }.$$

Of course this means that for paired samples, the sample size n must be the same for both groups.

Example:
Blood platelet aggregation data, cf. chapter 4.1.

4.7.2 Paired tests

To analyze paired comparisons, we use the differences insider each pair:

$$u_i = x_i - y_i \quad (i = 1, \ldots, n) \ ,$$

which we consider to be realizations of iid. random variables $U_1, \ldots, U_n$. Having no difference between the two experimental conditions then simply means that $E[U_i] = 0$ (or alternately: median$(U_i) = 0$). Tests of this hypothesis are described in chapter 4.6. Note that the symmetricity required of the distribution of U_i for the Wilcoxon test is always present under the null hypothesis that X_i and Y_i are identically distributed.

4.7.3 Unpaired tests

For unpaired samples our data $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ (cf. chapter 4.7.1) can be regarded as realizations of the following random variables:

$$\begin{aligned} &X_1, \ldots, X_n \ \text{ i.i.d.} \ , \\ &Y_1, \ldots, Y_m \ \text{ i.i.d.} \ , \end{aligned} \tag{4.5}$$

where all the X_i are independent of all the Y_j.

4.7.4 Two-sample t-test for equal variances

The simplest case can be solved using the following assumptions on (4.5):

$$\begin{aligned} &X_1, \ldots, X_n \text{ i.i.d.} \quad \sim \mathcal{N}(\mu_X, \sigma^2) \ , \\ &Y_1, \ldots, Y_m \text{ i.i.d.} \quad \sim \mathcal{N}(\mu_Y, \sigma^2) \ . \end{aligned} \tag{4.6}$$

Here the null hypothesis of interest is

$$H_0: \ \mu_X = \mu_Y \ .$$

The two-sample t-test (assuming equal variances) rejects the null hypothesis $H_0 : \mu_X = \mu_Y$ if

$$\begin{aligned} |T| &= \frac{|\overline{X}_n - \overline{Y}_m|}{S_{pool}\sqrt{1/n + 1/m}} > t_{n+m-2,1-\alpha/2} \text{ using the alternative } H_A : \mu_X \neq \mu_Y \ , \\ T &= \frac{\overline{X}_n - \overline{Y}_m}{S_{pool}\sqrt{1/n + 1/m}} > t_{n+m-2,1-\alpha} \text{ using the alternative } H_A : \mu_X > \mu_Y \ , \\ T &= \frac{\overline{X}_n - \overline{Y}_m}{S_{pool}\sqrt{1/n + 1/m}} < -t_{n+m-2,1-\alpha} \text{ using the alternative } H_A : \mu_X < \mu_Y \ . \end{aligned}$$

Here

$$S_{pool}^2 = \frac{1}{n+m-2}\left(\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 + \sum_{i=1}^{m}(Y_i - \overline{Y}_m)^2\right)$$

is the pooled estimator of the common variance σ^2. The choice of the denominator in the test statistic T comes from the identity

$$\mathrm{Var}(\overline{X}_n - \overline{Y}_m) = \sigma^2(\frac{1}{n} + \frac{1}{m}) \ . \tag{4.7}$$

Proof of (4.7):
1. $\overline{X}_n$ and $\overline{Y}_m$ are independent, as all the X_i are independent of all the Y_j.
2. Due to the independence of $\overline{X}_n$ and $\overline{Y}_m$, the variance can be decomposed as $\mathrm{Var}(\overline{X}_n - \overline{Y}_m) = \mathrm{Var}(\overline{X}_n) + \mathrm{Var}(-\overline{Y}_m) = \mathrm{Var}(\overline{X}_n) + \mathrm{Var}(\overline{Y}_m)$.
3. We have $\mathrm{Var}(\overline{X}_n) = \sigma^2/n$ and $\mathrm{Var}(\overline{Y}_m) = \sigma^2/m$.
Thus applying Step 2, we obtain $\mathrm{Var}(\overline{X}_n - \overline{Y}_m) = \sigma^2(1/n + 1/m)$. □

We can derive the two-sample t-test in the following way. We replace the unknown difference $\mu_X - \mu_Y$ by its estimate $\overline{X}_n - \overline{Y}_m$ and ascertain whether or not this estimate is "close to" 0 (if it were "far from" 0, this would be evidence for H_A). To quantify this, we divide by the square root of the variance estimate and use the quotient as our test statistic:

$$\begin{aligned} T &= \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\widehat{\mathrm{Var}}(\overline{X}_n - \overline{Y}_m)}} \\ &= \frac{\overline{X}_n - \overline{Y}_m}{S_{pool}\sqrt{1/n + 1/m}} \ . \end{aligned}$$

Under the assumption (4.6) and the null hypothesis $\mu_X = \mu_Y$ we then have

$$T \ \sim \ t_{n+m-2} \ .$$

In this way we obtain the decision rule mentioned above, in a manner analagous to the derivation of the one-sample t-test in chapter 4.6.2.

Example: Latent heat of melting ice, cf. chapter 4.1.
Let the null hypothesis be $H_0: \ \mu_X = \mu_Y$, and the alternative hypothesis $H_A: \ \mu_X \neq \mu_Y$. The characteristic numbers of this data set are: $\overline{x}_{13} = 80.021$, $\overline{y}_8 = 79.979$, $s^2_{pool} = 7.2 \ 10^{-4}$. Thus the test statistic is 3.47, which is clearly larger than the 97.5% quantile $t_{19,0.975} = 2.093$.

4.7.5 Further two-sample tests

Two-sample t-test for non-matching variances

Now replace the assumption in (4.6) by the following:

$$\begin{aligned} &X_1, \ldots, X_n \text{ i.i.d.} \quad \sim \mathcal{N}(\mu_X, \sigma_X^2), \\ &Y_1, \ldots, Y_m \text{ i.i.d.} \quad \sim \mathcal{N}(\mu_Y, \sigma_Y^2). \end{aligned}$$

This generalization of the two-sample t-test for unequal variances $\sigma_X^2 \neq \sigma_Y^2$ can be found in the literature. It has also been widely implemented in statistical software.

Two-sample Wilcoxon test (Mann-Whitney test)

The requirements of the two-sample Wilcoxon test, also known as the Mann-Whitney test, on (4.5) are:

$$\begin{aligned} X_1, \ldots, X_n \text{ i.i.d.} &\sim \text{ arbitrary cumulative distribution function } F(\cdot) \ , \\ Y_1, \ldots, Y_m \text{ i.i.d.} &\sim F(\cdot - \delta) \ . \end{aligned}$$

This means that the distribution of Y_j is the same as that of X_i up to a shift by δ, as: $P(Y_j \leq x + \delta) = F_Y(x + \delta) = F_X(x + \delta - \delta) = F_X(x) = P(X_i \leq x)$.

A p-value of a two-sample Wilcoxon test can be calculated using a computer. For the same reasons as in the one-sample case (cf. chapter 4.6.4), this Wilcoxon test is generally preferable to the t-test.

4.8* Design of experiments (Stahel, ch. 14.1 - 14.2)

Carefully planning how to obtain the data is just as important as their subsequent evaluation. So far we have mainly discussed comparisons between two treatments (paired or unpaired). When performing such comparisons, we should never compare a new treatment with results obtained for the standard treatment in previous studies. Instead, we should always use a **control group** in the same study, so that this group is as similar as possible to the group receiving the new treatment. Then the question arises as to how to create these groups. Similarly, a paired setup forces us to decide in which order to administer the new treatments. Systematic differences between the groups – or systematic effects caused by the choice of treatment order – can best be avoided by **randomization**, making the allocation or the order random. Here random does not mean arbitrary, but using random numbers.

Another important point is that where possible, the experiment should be **double blind**. This means that neither the person administering the treatment nor the person receiving it should know how the groups are allocated. This is necessary, as otherwise effects can appear that influence the outcome of the experiment (e.g. that it is not the treatment that is effective, but the effort put into it).

A randomized, double blind experiment is not always possible (out of ethical or practical considerations). In such circumstances this makes the evaluation and interpretation of the results vastly more difficult, as confounding effects cannot be ruled out in practice. One famous example is the connection between smoking and lung cancer – which was disputed for a long time, as the effects of genetic predisposition and lifestyle could not be ruled out.

Chapter 5

Regression

5.1 Correlation und empirical correlation

The joint distribution of dependent random variables X and Y is generally quite complicated; thus usually a **simplifying** characteristic number suffices for describing their dependence. We define the covariance and the correlation between X and Y in the following way:

$$\begin{aligned}
&\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] \quad \text{(covariance)}\\
&\mathrm{Corr}(X,Y) = \rho_{XY} = \mathrm{Cov}(X,Y)/(\sigma_X\sigma_Y) \quad \text{(correlation)}\ ,
\end{aligned}$$

where $\sigma_X = \sqrt{\mathrm{Var}(X)}$, and similarly for σ_Y.

The correlation ρ_{XY} of X and Y is a dimension-less, normalized number taking on values $\rho_{XY} \in [-1,1]$.

Correlation measures the strength and direction of the **linear dependence** between X und Y. Its extreme values ±1 are only attained under very special conditions:

$$\begin{aligned}
&\mathrm{Corr}(X,Y) = +1 \text{ if and only if } Y = a + bX \text{ for some } a \in \mathbb{R} \text{ and } b > 0,\\
&\mathrm{Corr}(X,Y) = -1 \text{ if and only if } Y = a + bX \text{ for some } a \in \mathbb{R} \text{ and } b < 0.
\end{aligned}$$

Furthermore the following implication holds true:

$$X \text{ and } Y \text{ independent} \quad \Longrightarrow \quad \mathrm{Corr}(X,Y) = 0. \tag{5.1}$$

The converse of this is generally not true.

5.1.1 Empirical correlation

In Chapter 4.2.2 and Figure 4.3 we saw an example of data $(x_1,y_1),\ldots,(x_n,y_n)$ that we could consider to be realizations of iid. random vectors $(X_1,Y_1),\ldots,(X_n,Y_n)$.

In this situation, the empirical correlation is

$$\widehat{\mathrm{Corr}}(X,Y) = \hat{\rho}_{XY} = \frac{\sum_{i=1}^n (x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^n (x_i-\overline{x})^2}\sqrt{\sum_{i=1}^n (y_i-\overline{y})^2}}\ .$$

This has the following properties analogous to the properties of correlation:

$$\hat{\rho}_{XY} \in [-1, 1] \ ,$$
$$\hat{\rho}_{XY} = +1 \ \Leftrightarrow \ y_i = a + bx_i \text{ for all } i = 1, \dots, n \text{ and for some } a \in \mathbb{R} \text{ and } b > 0 \ ,$$
$$\hat{\rho}_{XY} = -1 \ \Leftrightarrow \ y_i = a + bx_i \text{ for all } i = 1, \dots, n \text{ and for some } a \in \mathbb{R} \text{ and } b < 0 \ .$$

5.2 Simple linear regression

Consider the following example from chemistry. The dimerization of 1,3-butadiene proceeds according to a second-order reaction model, and can thus be characterized by the equation $\frac{d}{dt}C(t) = -\kappa C(t)^2$. Here C denotes the partial pressure of the input, and t is time. This equation admits solutions of the form

$$\frac{1}{C(t)} = \frac{1}{C(0)} + \kappa t \ .$$

Figure 5.1 shows measurements taken at different times over the course of the reaction. The above equation shows that the reciprocal of the partial pressure should exhibit a linear dependence on time. Due to random measurement errors and small systematic deviations from the simple model, the points do not lie perfectly on a line.

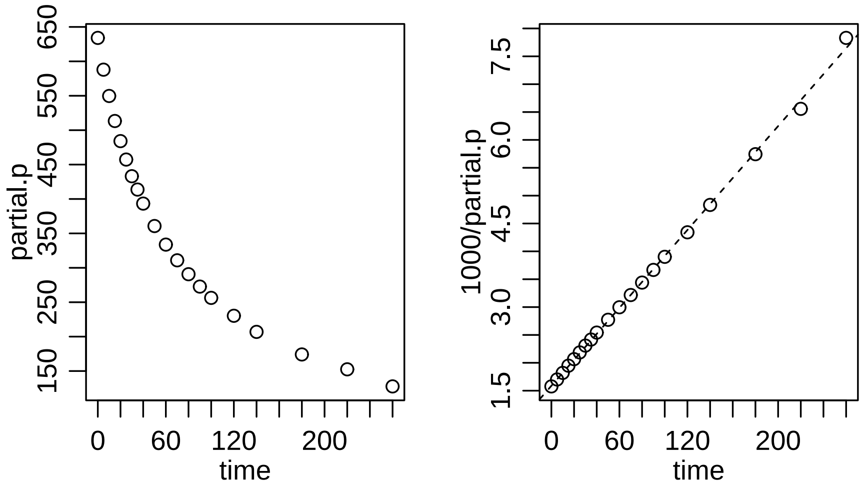

Figure 5.1: Partial pressure of butadiene (left), and its scaled reciprocal 1000/(partial pressure) (right), plotted against time

5.2.1 The simple linear regression model

In the example above, we used data

$$(x_1, y_1), \dots, (x_n, y_n) \ ,$$

where x_i denotes the time at which the i-th measurement was taken, and y_i the reciprocal of the partial pressure measured then. We can consider these data as realizations of the following model:

$$
\begin{aligned}
&Y_i = h(x_i) + E_i \ (i = 1, \ldots, n) \ , \\
&E_1, \ldots, E_n \text{ i.i.d. } , \ \mathcal{E}(E_i) = 0, \ \mathrm{Var}(\mathcal{E}_i)) = \sigma^2 \ .
\end{aligned}
$$

Here the variable Y is the **response variable**, and the variable x is the **explanatory variable**, **predictor variable** or **covariate**. The random variables E_i are often termed error terms or noise variables. They show that the connection between explanatory and response variables is not an exact one. The explanatory variables x_i $(i = 1, \ldots, n)$ are deterministic, while the response variables Y_i are actual random variables (due to the E_i).

Possible models for the function $h(\cdot)$ include:

$$
\begin{aligned}
&h(x) = \beta_0 + \beta_1 x && : \text{simple linear regression} \ , \\
&h(x) = \beta_1 x && : \text{simple linear regression through the origin} \ .
\end{aligned}
$$

We mostly consider the more general model, which includes an intercept β_0. This model is illustrated in Figure 5.2, where the distribution of the error terms is specified as $\mathcal{N}(0, 0.1^2)$.

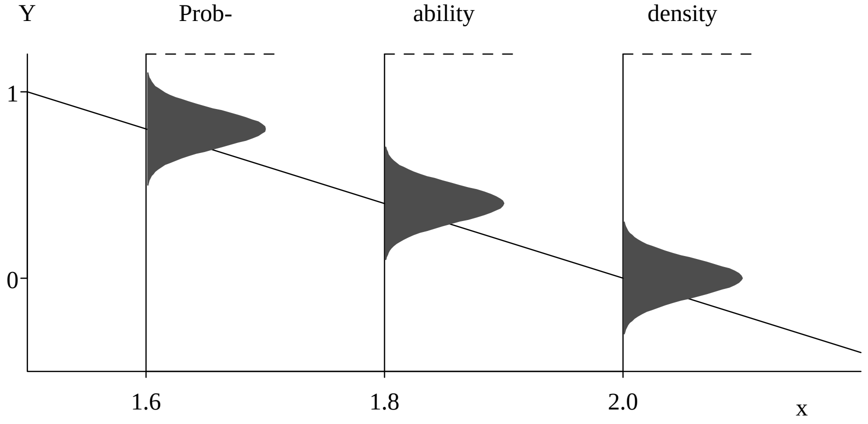

Figure 5.2: Illustration of the regression model $Y_i = 4 - 2x_i + E_i$, where $E_i \sim \mathcal{N}(0, 0.1^2)$, for three observations

5.2.2 Parameter estimation

In a simple linear regression, the unknown model parameters are β_0, β_1 and the error variance σ^2. The method of least squares gives us the following estimates:

$$
\hat{\beta}_0, \hat{\beta}_1 \text{ minimize } \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_i))^2 \ .
$$

This optimization problem admits a unique solution:

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (Y_i - \bar{Y}_n)(x_i - \bar{x}_n)}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2} \\
\hat{\beta}_0 &= \bar{y}_n - \hat{\beta}_1 \bar{x}_n \ .
\end{aligned}
$$

The least-squares principle gives us unbiased estimates of β_0 and β_1, i.e.:

$$\mathcal{E}(\hat{\beta}_0) = \beta_0 \ , \quad \mathcal{E}(\hat{\beta}_1) = \beta_1 \ .$$

This means the estimates exhibit no systematic error (for example, they do not systematically overestimate β_1, as otherwise we would have $\mathcal{E}(\hat{\beta}_1) > \beta_1$).

To estimate the error variance σ^2 we can use the concept of residuals. If we could observe realizations of the error terms E_i, we could apply the empirical estimate of variance to these to approximate σ. As we do not have realizations of the E_i, we approximate these using the **residuals**:

$$R_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \ (i = 1, \ldots, n) \ .$$

Since $E_i = Y_i - (\beta_0 + \beta_1 x_i)$, the approximation $R_i \approx E_i$ is a meaningful one. The estimate of variance is then:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} R_i^2 \ . \tag{5.2}$$

Note that (for simple linear regression with intercept β_0,) we have the identity $\sum_{i=1}^{n} R_i = 0$. Then the variance estimate in (5.2) is the empirical variance for a sample (cf. chapter 4.6.1), but with the factor $1/(n-1)$ replaced by a factor $1/(n-2)$. This factor follows the rule of thumb which says it should be $1/(n -$ number of parameters), where the variance parameter to be estimated is not counted towards the number of parameters (thus here the parameters that count are β_0 and β_1).

If the dataset contains realizations y_i $(i = 1, \ldots, n)$, the estimates are computed using the y_i instead of the Y_i. The realizations of the residuals, for example, then become $r_i = y_i - (\hat{\beta}_0 - \hat{\beta}_1 x_i)$.

5.2.3 Tests and confidence intervals

In this section we shall discuss the 2nd and 3rd key questions (cf. chapter 3.1) in the context of simple linear regression. In the course of this discussion we shall aim at more than just finding the best-fitting regression line.

Applying the t-Test in regression

As an example, consider the following dataset. The average daily temperature (x) and the average daily ozone concentration (Y) were measured on $n = 111$ days. The resulting data and the fitted regression line $\hat{\beta}_0 + \hat{\beta}_1 x$ are shown in Figure 5.2.3. The key practical question of interest is: does the temperature influence the concentration of ozone. We can translate this question into a testing problem:

$$\begin{aligned} H_0 &: \ \beta_1 = 0 \ , \\ H_A &: \ \beta_1 \neq 0 \ . \end{aligned}$$

"By default" here, we carry out a two-sided test: the t-test of the slope in a simple linear regression.

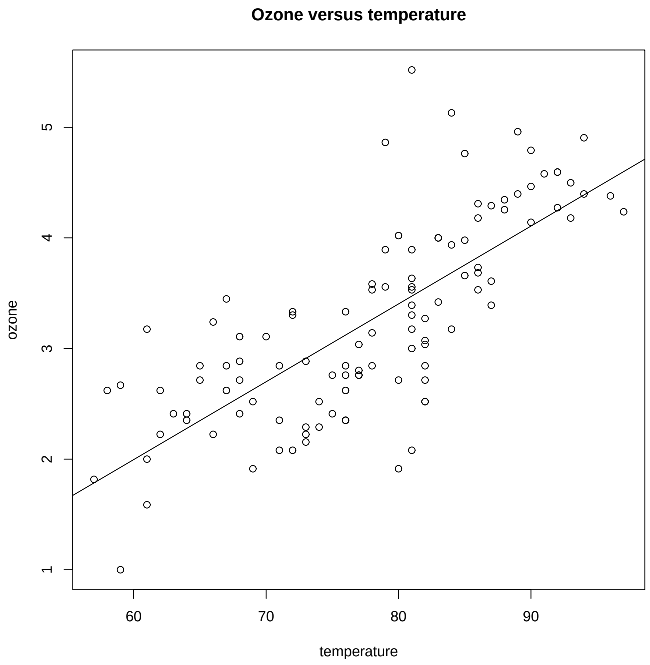

Figure 5.3: Scatterplot and fitted regression line for the dataset of ozone vs. temperature.

We assume here that

$$E_1, \ldots, E_n \text{ i.i.d. } \mathcal{N}(0, \sigma^2) \ . \tag{5.3}$$

The test statistic is

$$\frac{\hat{\beta}_1}{\widehat{s.e.}(\hat{\beta}_1)} \ ,$$
$$\widehat{s.e.}(\hat{\beta}_1) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_1)} = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}} \ .$$

Under the null hypothesis and the assumption of normally distributed errors (5.3), we have:

$$T \sim t_{n-2} \text{ under the null hypothesis } H_0 : \ \beta_1 = 0 \ ,$$

and we can compute the p-value of this two-sided t-test like in chapter 4.6.2 (using $n-2$ instead of $n-1$ degrees of freedom). This p-value is also given by standard statistical software.

In the same way we can obtain a test of $H_0 : \ \beta_0 = 0$ under the two-sided alternative $H_A : \ \beta_0 \neq 0$. The p-value here – under the normal assumption (5.3) – is also given by the standard statistical software packages.

When a simple linear regression is fitted for the ozone vs. temperature data using the software package R, the following output is produced:

```
Call:
  lm(formula = ozone ~ temperature)
```

```
Residuals:
       Min       1Q   Median       3Q      Max
-1.49016 -0.42579  0.02521  0.36362  2.04439

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.225984   0.461408  -4.824 4.59e-06 ***
temperature   0.070363   0.005888  11.951  < 2e-16 ***
---

Residual standard error: 0.5885 on 109 degrees of freedom
Multiple R-Squared: 0.5672,Adjusted R-squared: 0.5632
F-statistic: 142.8 on 1 and 109 DF,  p-value: < 2.2e-16
```

The second column in the table "Coefficients" describes the point estimate $\hat{\beta}_i$ $(i = 0, 1)$; the third column of this table gives the estimated standard error $\widehat{s.e.}(\hat{\beta}_i)$ $(i = 0, 1)$; and the fourth column lists the values of the test statistics $\hat{\beta}_i/\widehat{s.e.}(\hat{\beta}_i)$ $(i = 0, 1)$, which are the quotients of the second and the third column. The fifth column then gives us the p-values for $H_0 :\ \beta_i = 0$ and $H_A :\ \beta_i \neq 0$ $(i = 0, 1)$. We moreover find an estimate $\hat{\sigma}$ of the error standard deviation under "Residual standard error" – here the "degrees of freedom" are $n - 2$.

Confidence intervals

On the basis of the normality assumption, we obtain the following two-sided confidence intervals for β_i $(i = 0, 1)$ at level $1 - \alpha$:

$$\begin{aligned} &\hat{\beta}_0 \pm \widehat{s.e.}(\hat{\beta}_0) t_{n-2;1-\alpha/2} \ \text{ for } \beta_0 \ , \\ &\hat{\beta}_1 \pm \widehat{s.e.}(\hat{\beta}_1) t_{n-2;1-\alpha/2} \ \text{ for } \beta_1 \ . \end{aligned}$$

5.2.4 R^2, the coefficient of determination

The goodness of fit of a regression model can be quantified by the so-called coefficient of determination, which we denote R^2. To do this, we use an equation that describes the relationship between several sources of variation. Writing $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ for the fitted value at x_i, we have

$$\underbrace{\sum_{i=1}^{n} (y_i - \overline{y})^2}_{SS_Y} = \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{SS_E} + \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}_{SS_R} \ . \tag{5.4}$$

Here SS_Y describes the total variation of the response variables (without regarding the influence of the explanatory variables x), SS_E is the residual squared error, and SS_R denotes the variation explained by the regression (through the influence of the explanatory variables x). The coefficient of determination is then defined in the following way:

$$R^2 = \frac{SS_R}{SS_Y} \ , \tag{5.5}$$

and it quantifies the proportion of total variation explained by the regression. From Equation 5.4 we see that $0 \leq R^2 \leq 1$; if R^2 is close to 1, the regression model explains much of the variation and is thus good; if $R^2 \approx 0$, however, the model is not of much use. In standard computer output, the value of R^2 can generally be found under the description "Multiple R-squared".

For simple linear regression we have

$$R^2 = \hat{\rho}_{XY}^2$$

in general, i.e. R^2 ist just the square of empirical correlation.

5.2.5 General procedure for simple linear regression

Put simply, simple linear regression can be carried out as follows:

1. Fit the regression line, i.e. compute the point estimate $\hat{\beta}_0,\ \hat{\beta}_1$.
2. Test whether the explanatory variable x influences the response Y at all, using a t-test with $H_0:\ \beta_1 = 0$ and $H_a:\ \beta_1 \neq 0$. If the test finds nothing significant (i.e. if H_0 is not rejected, but kept), the problem is "not of interest in its current form".
3. Test whether the regression line goes through the origin, using a t-test with $H_0:\ \beta_0 = 0$ and $H_A:\ \beta_0 \neq 0$. If the test finds nothing significant (i.e. H_0 is kept), then use the smaller model containing no intercept.
4. Calculate confidence intervals for β_0 and β_1, where desired.
5. Calculate R^2, the coefficient of determination. In some ways this is a more informal (and extra) quantification than the test in point 2.
6. Check the model assumptions by analyzing the residuals. This key step is described at length in chapter 5.2.6.

5.2.6 Analysis of residuals

In this part we shall decribe graphical methods that can be used to check the model assumptions of simple linear regression on the basis of the known residuals $r_i (i = 1, \ldots, n)$. These model assumptions are as follows (in order of importance):

1. $\mathcal{E}(E_i) = 0$.
   Thus $\mathcal{E}(Y_i) = \beta_0 + \beta_1 x_i$, i.e. there is no bias in the model.
   A failure of this assumption to hold might mean e.g. that x and Y exhibit non-linear dependence.
2. $E_1, \ldots, E_n$ i.i.d.
   This assumption could be broken by differences in the error variances, i.e. $\mathrm{Var}(E_i) = \sigma_i^2$ with different σ_i^2 for different $i = 1, \ldots, n$. Alternatively, any problems here might also stem from the errors being correlated.
3. $E_1, \ldots, E_n$ i.i.d. $\mathcal{N}(0, \sigma^2)$.
   Deviations from this assumptions can be caused by heavy-tailed error distributions.

The Tukey-Anscombe plot

The most important plot in the analysis of residuals is the plot of residuals r_i against fitted values $\hat{y}_i$, known as the Tukey-Anscombe Plot.

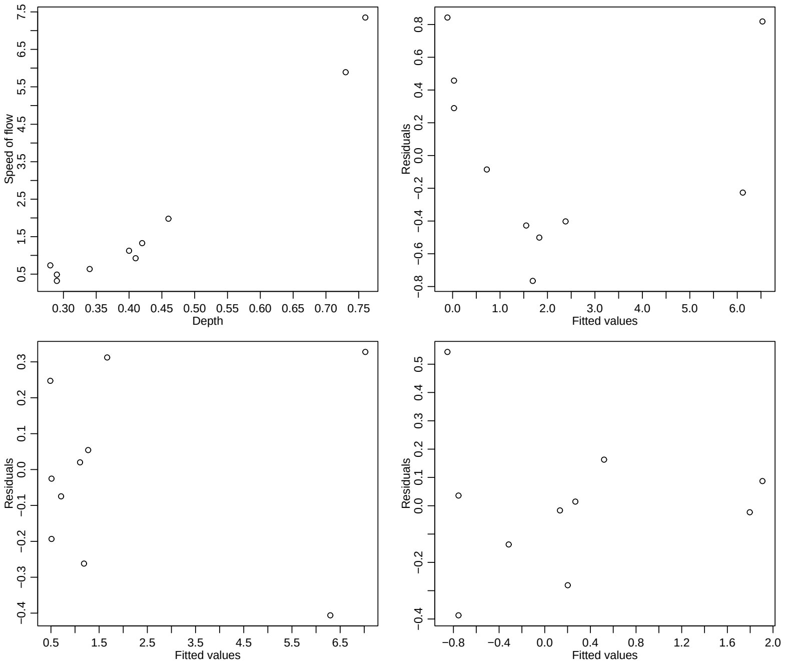

Figure 5.4: Scatterplot of depth and speed of flow (above left), Tukey-Anscombe plot for simple linear regression (above right), for square regression (cf. chapter 5.3.1) (below left) and for simple linear regression using the logarithms $\log(Y)$ and $\log(x)$ (below right).

Ideally the points in a Tukey-Anscombe plot are evenly scattered around zero.
Deviations from this:
- Conical increase of the scattered residuals as $\hat{y}_i$ increases
  Maybe transform the response variable by taking its logarithm (if the Y_i are positive), i.e. use the new model

$$\log(Y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i \ .$$

- Outliers
  Maybe use robust regression methods (see the literature for this)
- Irregular structure
  This indicates that there may be a non-linear connection between the variables
  Maybe transform the response and/or explanatory variables (also cf. the example given in Figure 5.1).

The Tukey-Anscombe plot for the ozone data is given in Figure 5.5.

Non-linear dependencies can of course occur in practice; they indicate a mistake in the specification of the regression function. This problem can be combatted by including further explanatory variables in the regression (e.g. quadratic terms as in chapter 5.3.1) or by transforming the response or explanatory variables as above. One simple example is shown in Figure 5.4, where the depth of streams is compared to the speed at which they flow. When doing a simple linear regression, the Tukey-Anscombe plot shows a clearly non-linear structure – which disappears after a quadratic term is added (cf. chapter 5.3.1) or if logarithms are taken of both variables (i.e. a x and Y are described by the power model

$$\log(Y_i) = \beta_0 + \beta_1 \log(x_i) + \varepsilon_i \ (i = 1, \ldots, n) \ .$$

Howver, this example contains insufficient data for distinguishing between these two models. The non-linear dependence here is naturally also visible upon a close look at the original scatterplot. In most cases, though, non-linear behaviour is easier to spot in the Tukey-Anscombe plot.

The serial correlation plot

To check the assumption that the errors $E_1, \ldots, E_n$ are independent, the following plot can be used: plot the residuals r_i against their indices i.

Ideally the points are evenly scattered around zero.
Deviations from this:
- Large areas in which the residuals are all positive or all negative
  The point estimates are still fine, but the tests and confidence intervals are no longer correct
  Maybe use regression methods that allow for correlation between errors (see the literature for this)

The serial correlation plot of residuals for the ozone dataset is shown in Figure 5.5.

The normal plot

The normal plot (cf. chapter 4.4.6) is useful for checking the assumption of normality (5.3).

Ideally the normal plot contains a more or less straight line
Deviations from this:
- Not a straight line
  Maybe use robust regression (see the literature for this)

The normal plot for the ozone dataset is shown in Figure 5.5.

Finding a good model

Often many models are looked at, and fitted, in a kind of “workflow-feedback” procedure. We start with an initial model, analyze its residuals and then use the results to modify the model. The modified model (still assumed to be linear, although perhaps with transformed variables) is then fitted once more using linear regression, and this new model is then evaluated itself using the analysis of residuals. This procedure is repeated until a “satisfactory” model has been found and fitted.

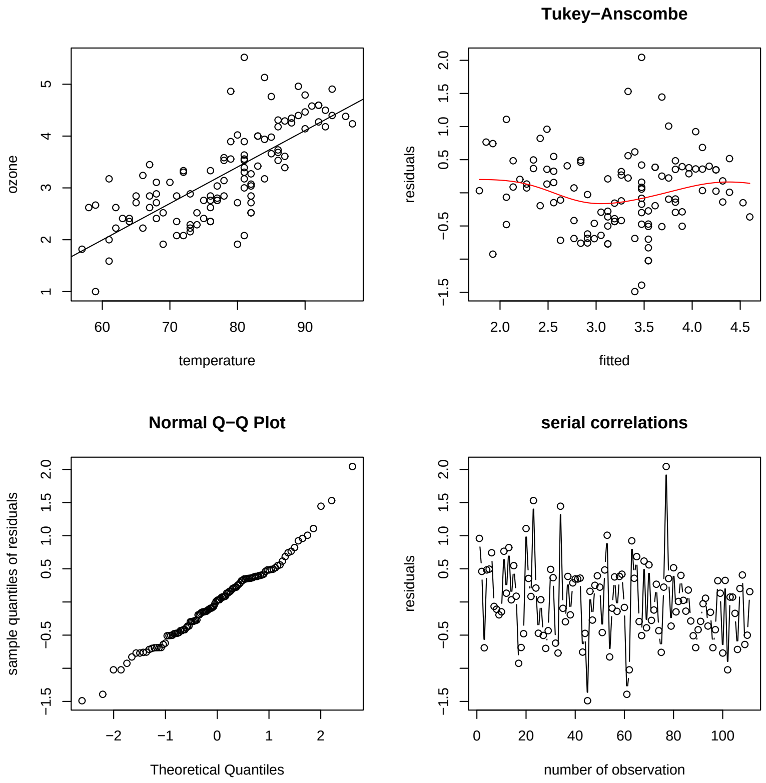

Figure 5.5: Ozone data: scatterplot with fitted regression line (above left); Tukey-Anscombe plot (above right); normal plot (below left); serial correlation plot (below right).

5.3 Multiple linear regression

Often there are several predictor variables $x_{i,1}, \ldots, x_{i,p-1}$ $(p > 2)$.

5.3.1 The multiple linear regression model

The multiple linear regression model is the following:

$$\begin{aligned} &Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{i,j} + E_i \ , \\ &E_1, \ldots, E_n \text{ i.i.d. } , \ \mathcal{E}(E_i) = 0, \ \mathrm{Var}(\mathcal{E}_i) = \sigma^2 \ . \end{aligned}$$

As with simple linear regression we assume that the predictor variables are deterministic. It often helps to have this model written with matrices:

$$\underset{n \times 1}{Y} \quad = \quad \underset{n \times p}{X} \quad \times \quad \underset{p \times 1}{\beta} \quad + \quad \underset{n \times 1}{E} \quad , \tag{5.6}$$

where X is an $(n \times p)$ matrix with columns $(1, 1, \ldots, 1)^T$, $(x_{1,1}, x_{2,1}, \ldots, x_{n,1})^T$ up to $(x_{1,p-1}, x_{2,p-1}, \ldots, x_{n,p-1})^T$.

Examples of multiple linear regression include:

Simple linear regression: $Y_i = \beta_0 + \beta_1 x_i + E_i$ $(i = 1, \ldots n)$.

$$p = 2 \qquad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} .$$

Quadratic regression: $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i$ $(i = 1, \ldots n)$.

$$p = 3, \qquad X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} .$$

Note that the regression function is quadratic in the predictors x_i, but *linear* in the coefficients β_j – and is thus a special case of the multiple linear regression model.

Regression with transformed predictor variables:
$Y_i = \beta_0 + \beta_1 \log(x_{i2}) + \beta_2 \sin(\pi x_{i3}) + E_i$ $(i = 1, \ldots n)$.

$$p = 3, \qquad X = \begin{pmatrix} 1 & \log(x_{12}) & \sin(\pi x_{13}) \\ 1 & \log(x_{22}) & \sin(\pi x_{23}) \\ \vdots & \vdots & \vdots \\ 1 & \log(x_{n2}) & \sin(\pi x_{n3}) \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} .$$

Once again the model is *linear* in the coefficients β_j and non-linear in the predictors x_{ij}.

5.3.2 Parameter estimation and the t-test

Similarly to simple linear regression, multiple linear regression mainly uses the method of least squares:

$$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{p-1} \text{ minimize } \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,p-1}))^2 .$$

If $p < n$, the unique solution of this optimization problem has an explicit form:

$$\hat{\beta} = (X^T X)^{-1} X^T Y ,$$

where $\hat{\beta}$ is the $p \times 1$ vector $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{p-1})^T$ and X, Y are as in (5.6).

The estimate of error variance is

$$\frac{1}{n-p} \sum_{i=1}^{n} R_i^2, \quad R_i = Y_i - (\hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_j x_{i,j}) .$$

Assuming normality of the errors as in (5.3), t-tests of the following hypotheses can be carried out like for simple linear regression:

$$H_{0,j}: \ \beta_j = 0; \quad H_{A,j}: \ \beta_j \neq 0 \ (j = 0, \ldots, p-1) .$$

The main distinction to be made lies in the interpretation of the parameters:

β_j measures the linear effect
of the jth predictor variable on the response variable Y
after the linear effects of all other variables on Y
have been eliminated. $(j = 1, \ldots, p-1)$

In particular, this implies that the coefficient β_j cannot be found by merely by carrying out a single simple linear regression of Y on the jth predictor variable.

Example: Take $p = 3$ and 2 predictor variables. We assume that the predictor variables have high empirical correlation. Then it may well happen that neither $H_{0,1}: \beta_1 = 0$ nor $H_{0,2}: \beta_2 = 0$ is rejected, even though at least one of the coefficients β_1 and β_2 is non-zero. To prevent the erroneous conclusion being drawn that none of the predictor variables has an effect on the response variable, we must use the so-called F-test.

5.3.3 The F-test

The (global) F-test gives a quantitative answer to the question whether or not at least one of the predictor variables has a relevant (regression) effect on the response variable. The (global) F-test looks at the following null hypothesis:

$$\begin{aligned} &H_0: \ \beta_1 = \ldots = \beta_{p-1} = 0 \\ &H_A: \ \text{at least one } \beta_j \neq 0 \ (j = 1, \ldots, p-1) \ . \end{aligned}$$

The p-value of the (global) F-test is given by the computer output as the “F statistic”.

5.3.4 R^2, the coefficient of determination

For multiple linear regression the coefficient of determination, R^2, is defined using the formula (5.5) (by means of the decomposition in (5.4). However the interpretation of this coefficient in terms of a squared sample covariance between the response and the predictor variables is longer possible here.

5.3.5 Analysis of residuals

The analysis of residuals proceeds in a completely analogous way to chapter 5.2.6. The general procedure for multiple linear regression is like in chapter 5.2.5, using the F-test after Step 1.

5.3.6 Strategies for data analysis: a closing example

We regard an example in which the quality of asphalt is analyzed using 6 explanatory variables.

```
  y = RUT  : "rate of rutting" = change of rut depth in inches per million
             wheel passes
 x1 = VISC : viscosity of asphalt
 x2 = ASPH : percentage of asphalt in surface course
```

```
  x3 = BASE : percentage of asphalt in  base   course
  x4 = RUN  : '0/1' indicator for two sets of runs.
  x5 = FINES: 10* percentage of fines in surface course
  x6 = VOIDS: percentage of voids in surface course
```

The data are visualized by pairwise scatterplots in Figure 5.6. The dependencies of these

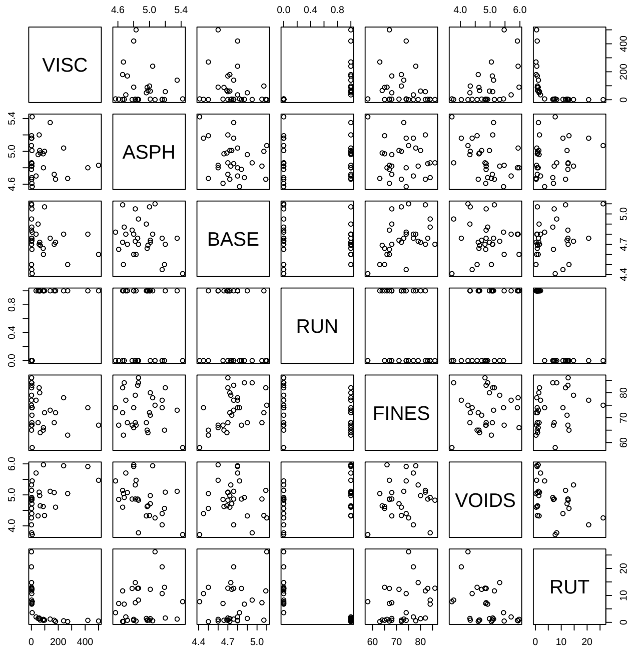

Figure 5.6: Pairwise scatterplots for the asphalt datset. The response variable is "RUT".

variables are more linear if the response variable "RUT" and the predictor variable "VISC" are both replaced by their respective logarithms.

```
   y = LOGRUT  : log("rate of rutting") = log(change of rut depth in inches
                 per million wheel passes)
  x1 = LOGVISC : log(viscosity of asphalt)
  x2 = ASPH : percentage of asphalt in surface course
  x3 = BASE : percentage of asphalt in  base   course
  x4 = RUN  : '0/1' indicator for two sets of runs.
  x5 = FINES: 10* percentage of fines in surface course
  x6 = VOIDS: percentage of voids in surface course
```

This transformed dataset is plotted in Figure 5.7.

Using the software package R, we fit a multiple linear model, and obtain the following output:

```
Call:
lm(formula = LOGRUT ~ ., data = asphalt1)

Residuals:
```

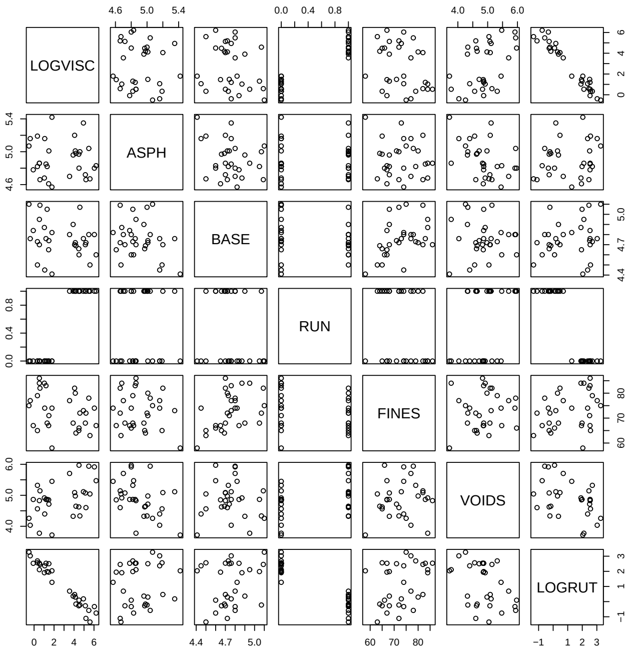

Figure 5.7: Pairwise scatterplots for the transformed asphalt dataset. The response variable is “LOGRUT”, which is the logarithm of the original response variable “RUT”. The predictor variable “LOGVISC” is the logarithm of the original predictor variable “VISC”.

```
     Min       1Q   Median       3Q      Max
-0.48348 -0.14374 -0.01198  0.15523  0.39652

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.781239   2.459179  -2.351 0.027280 *
LOGVISC     -0.513325   0.073056  -7.027 2.90e-07 ***
ASPH         1.146898   0.265572   4.319 0.000235 ***
BASE         0.232809   0.326528   0.713 0.482731
RUN         -0.618893   0.294384  -2.102 0.046199 *
FINES        0.004343   0.007881   0.551 0.586700
VOIDS        0.316648   0.110329   2.870 0.008433 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2604 on 24 degrees of freedom
Multiple R-Squared: 0.9722,     Adjusted R-squared: 0.9653
F-statistic: 140.1 on 6 and 24 DF,  p-value: < 2.2e-16
```

We can see that the predictor variables “LOGVISC”, “ASPH” and “VOID” are significant or even highly significant; furthermore the predictor variable “RUN” is merely weakly significant. The F-test is extremely significant, and R^2 is quite close to 1. Here we are using $n - p = 24$ degrees of freedom, where $p = 7$, and $n = 31$. In Figure 5.8 we see the analysis of residuals, summarized by the Tukey-Anscombe and normal plots: evidently the assumption of normality for the errors is a sensible one. The Tukey-Anscombe plot exhibits some systematic variation, which may be due to underlying non-linearity. As however R^2 already is quite close to 1, we can still conclude that the multiple linear regression model can explain very much of the total variation.

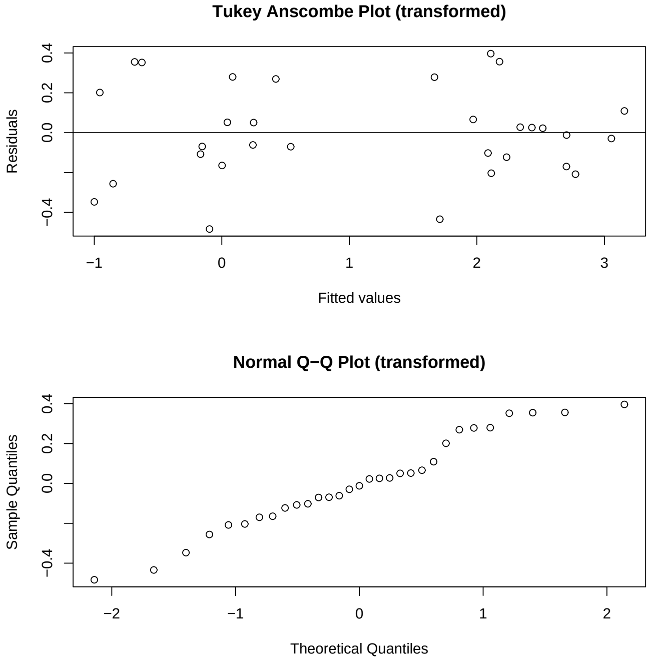

Figure 5.8: Tukey-Anscombe (above) and normal plots (below) for the asphalt dataset containing the transformed variables “LOGRUT” and “LOGVISC”.

Without logarithmic transformations, i.e. using the untransformed model from Figure 5.6, the coefficient of determination is $R^2 = 0.7278$, which is considerably worse than that of the transformed model.