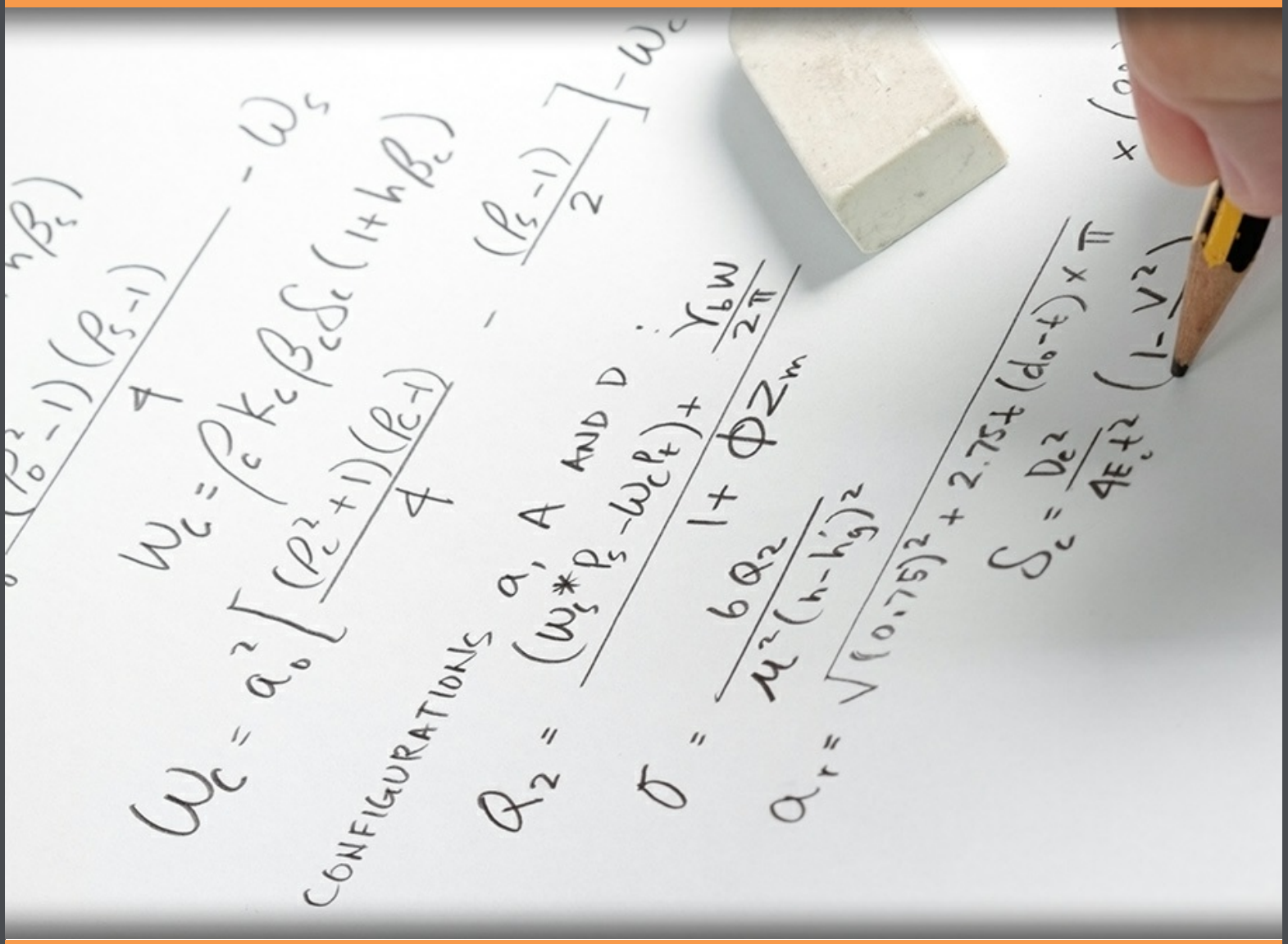
Exercises in Statistical Inference

with detailed solutions

Robert Jonsson



Robert Jonsson

Exercises in Statistical Inference with detailed solutions

Exercises in Statistical Inference with detailed solutions
1st edition

ISBN 978-87-403-XXXX-X

Contents

About the author

Robert Jonsson got his Ph.D. in Statistics from the Univ. of Gothenburg, Sweden, in 1983. He has been doing research as well as teaching undergraduate and graduate students at Dept. of Statistics (Gothenburg), Nordic School of Public Health (Gothenburg) and Swedish School of Economics (Helsinki, Finland). His researches cover theoretical statistics, medical statistics and econometrics that in turn have given rise to 14 articles in refereed international journals and some dozens of national papers. Teaching experience reaches from basic statistical courses for undergraduates to Ph.D. courses in Statistical Inference, Probability and Stochastic processes.

1 Introduction

1.1 Purpose of this book

The book is designed for students in statistics at the master level. It focuses on problem solving in the field of statistical inference and should be regarded as a complement to text books such as Wackerly *et al* 2007, Mat*hematical Statistics with Applications* or Casella & Berger 1990, *Statistical Inference*. The author has noticed that many students, although being well aware of the statistical ideas, fall short when being faced with the task of solving problems. This requires knowledge about statistical theory, but also about how to apply proper methodology and useful tricks. It is the aim of the book to bridge the gap between theoretical knowledge and problem solving.

Each of the following chapters contains a minimum of the theory needed to solve the problems in the *Exercises*. The latter are of two types. Some exercises with solutions are interspersed in the text while others, called *Supplementary Exercises*, follow at the end of the chapter. The solutions of the latter are found at the end of the book. The intention is that the reader shall try to solve these problems while having the solutions of the preceding exercises in mind. Towards the end of the following chapters there is a section called 'Final Words'. Here some important aspects are considered, some of which might have been overlooked by the reader.

1.2 Chapter content and plan of the book

Emphasis will be on the kernel areas of statistical inference: Point estimation – Confidence Intervals – Test of hypothesis. More specialized topics such as Prediction, Sample Survey, Experimental Design, Analysis of Variance and Multivariate Analysis will not be considered since they require too much space to be accommodated here. Results in the kernel areas are based on probability theory. Therefore we first consider some probabilistic results, together with useful mathematics. The set-up of the following chapters is as follows.

- *Ch. 2* Basic properties of discrete and continuous (random) variables are considered and examples of some common probability distributions are given. Elementary pieces of mathematics are presented, such as rules for derivation and integration. Students who feel that their prerequisites are insufficient in these topics are encouraged to practice hard, while others may skip much of the content of this chapter.
- *Ch. 3* The chapter is mainly devoted to sampling distributions, i.e. the distribution of quantities that are computed from a sample such as sums and variances. In more complicated cases methods are presented for obtaining asymptotic or approximate formulas. Results from this chapter are essential for the understanding of results that are derived in the subsequent chapters.

- *Ch. 4* Important concepts in point estimation are introduced, such as likelihood of a sample and sufficient statistics. Statistics used for point estimation of unknown quantities in the population are called estimators. (Numerical values of the latter are called estimates.) Some requirements on 'good' estimators are mentioned, such as being unbiased, consistent and having small variance. Four general methods for obtaining estimators are presented: Ordinary Least Squares (OLS), Moment, Best Linear Unbiased Estimator (BLUE) and Maximum Likelihood (ML). The performance of various estimators is compared. Due to limited space other estimation methods have to be omitted.
- *Ch. 5* The construction of confidence intervals (CIs) for unknown parameters in the population by means of so called pivotal statistics is explained. Guide lines are given for determining the sample size needed to get a CI of certain coverage probability and of certain length. It is also shown how CIs for functions of parameters, such as probabilities, can be constructed.
- *Ch. 6* Two alternative ways of testing hypotheses are described, the p-value approach and the rejection region (RR) approach. When a statistic is used for testing hypotheses it is called a test statistic. Two general principles for constructing test statistics are presented, the Chi-square principle and the Likelihood Ratio principle. Each of these gives raise to a large number of well-known tests. It's therefore a sign of statistical illiteracy when referring to a test as *the* Chi-Square test (probably supposed to mean the well-known test of independency between two qualitative variables). Furthermore, some miscellaneous methods are presented. A part of the chapter is devoted to nonparametric methods for testing goodness-of-fit, equality of two or more distributions and Fisher's exact test for independency.

  A general expression for the power (ability of a test to discriminate between the alternatives) is derived for (asymptotically) normally distributed test statistics and is applied to some special cases.

  When several hypotheses are tested simultaneously, we increase the probability of rejecting a hypothesis when it in fact is true. (This is one way to 'lie' when using statistical inference, more examples are given in the book.) One solution of this problem, called the Bonferroni-Holm correction is presented.

  We finally give some tests for linear models, although this topic perhaps should require their own book. Here we consider the classical Gauss-Markov model and simple cases of models with random coefficients.

From the above one might get the impression that statistical testing is a more 'important' in some sense than point and interval estimation. This is however not the case. It has been noticed that good point estimators also work well for constructing good CIs and good tests. (See e.g. Stuart *et al* 1999, p. 276.) A frequent question from students is: Which is best, to make a CI or to make a test? A nice answer to this somewhat controversial question can be found in an article by T. Wonnacott, 1987. He argues that in general a CI is to be preferred in front of a test because a CI is more informative. For the same reason he argues for a p-value approach in front of a RR approach. However, in practice there are situations where the construction of CIs becomes too complicated. Also the computation of p-values may be complicated. E.g. in nonparametric inference (Ch. 6.2.4) it is often much easier to make a test based on the RR approach than to use the p-value approach. The latter in turn being simpler than making a CI. An approach based on testing is also much easier to use when several parameters have to be estimated simultaneously.

1.3 Statistical tables and facilities

A great deal of the problem solving is devoted to computation of probabilities. For continuous variables this means that areas under frequency curves have to be computed. To this end various statistical tables are available. When using these there are two different quantities of interest.

- Given a value on the x-axis, what is the probability of a larger value, i.e. how large is the area under the curve above the value on the x-axis? This may be called computation of a p-value.
- Given a probability, i.e. an area under curve, what is the value on the x-axis that produced the probability? This may be called computation of an inverse p-value.

Statistical tables can show lower-tail areas or upper-tail areas. Lower-tail areas are areas below values on the x-axis and upper-tail areas are areas above. The reader should watch out carefully whether it is required to search for a p-value or an inverse p-value and whether the table show lower-or upper-tail areas. This seems to actually be a stumbling block for many students. It may therefore be helpful to remember some special cases for the normal-, Student's T-, Chi-square- and F-distributions. (These will be defined in Ch. 2.2.2 and Ch. 3.1.) The following will serve as hang-ups:

- In the *normal* distribution the area under curve above 1.96 is 0.025. The area under curve below 1.96 is thus 1-0.025=0.975.
- In *Student's T* distribution one needs to know the degrees of freedom (df) in order to determine the areas. With df = 1 the area under curve above 12.706 is 0.025.
- In the *Chi-square* distribution with df = 1 the area under curve above $3.84 \approx (1.96)^2$ is $2 \cdot 0.025 = 0.05$.
- In the *F* distribution one needs to know a pair of degrees of freedoms sometimes denoted $(\text{numerator}, \text{denominator}) = (f_1, f_2)$. With $f_1 = 1 = f_2$ the area under curve above $161.45 \approx (12.706)^2$ is 0.025.

Calculation of probabilities is facilitated by using either statistical program packages, so called 'calculators' or printed statistical tables.

- *Statistical program packages*. These are the most reliable ones to use and both p-values and inverse p-values can easily be computed by using programs such as SAS or SPSS, just to mention a few ones. E.g. in SAS the function *probt* can be used to find p-values for Student's T distribution and the function *tinv* to find inverse p-values. However, read manuals carefully.
- *'Calculators'*. These have quite recently appeared on the internet. They are easy to use (enter a value and click on 'calculate') and they are often free. Especially the calculation of areas in the F-distribution may be facilitated. An example is found under the address http://vassarstats.net/tabs.html.
- *Printed tables*. These are often found in statistical text books. Quality can be uneven, but an example of an excellent table is the table over the Chi-square distribution in Wackerly *et al*, 2007. This shows both small lower-tail areas and small upper-tail areas. Many tables can be downloaded from the internet. One example from the University of Glasgow is http://www.stats.gla.ac.uk.

Throughout this book we will compute exact probabilities obtained from functions in the program packet SAS. However, it is frequently enough to see whether a p-value is above or below 0.05 and in such cases it will suffice to use printed tables.

2 Basic probability and mathematics

2.1 Probability distributions of discrete and continuous random variables

A variable that is dependent on the outcome of an experiment (in a wide sense) is called a *random variable* (or just *variable*) and is denoted by an upper case letter, such as *Y*. A particular value taken by *Y* is denoted by a lower case letter *y*. For example, let *Y* = 'Number of boys in a randomly chosen family with 4 children', where *Y* may take any of the values *y* = 0,…,4. Before the 'experiment' of choosing such a family we do not know the value of *y*. But, as will be shown below, we can calculate the probability that the family has *y* boys. The probability of the outcome '*Y* = *y*' is denoted $P(Y = y)$ and since it is a function of *y* it is denoted $p(y)$. This is called *the probability function* (*pf*) of *the discrete variable Y*. A variable that can take any value in some interval, e.g. waiting time in a queue, is called continuous. The latter can be described by the *density* (*frequency function*) *of the continuous variable Y*, $f(y)$. The latter shows the relative frequency of values close to *y*.

Properties of *p*(*y*) (If not shown, summations are over all possible values of *y*.)

1) $0 \le p(y) \le 1, \sum p(y) = 1$
2) *Expected value, Population mean,* of *Y* : $\mu = E(Y) = \sum y \cdot p(y)$, center of gravity.
3) Expected value of a function of *Y*: $E(g(Y)) = \sum g(y) \cdot p(y)$.
4) (*Population*) *Variance* of *Y*: $\sigma^2 = V(Y) = \sum (y-\mu)^2 \cdot p(y) = E(Y^2) - \mu^2$, dispersion around population mean. The latter expression is often simpler for calculations. Notice that *(3)* is used with $g(y) = (y-\mu)^2$.
5) *Cumulative distribution function (cdf)* of *Y*. $F(y) = P(Y \le y) = p(y) + p(y-1) + ...$ and *Survival function* $S(y) = P(Y > y) = p(y+1) + p(y+2) + ... = 1 - F(y)$.

Properties of *f*(*y*) (If not shown, integration is over all possible values of *y*.)

1) $f(y) \ge 0, \int f(y)dy = 1, F(y) = \int_{y=-\infty}^{y} f(x)dx, S(y) = 1 - F(y).$
2) $\mu = E(Y) = \int y \cdot f(y)dy$, center of gravity.
3) Expected value of a function of *Y*, *g*(*Y*): $\mu = E(g(Y)) = \int g(y) \cdot f(y)dy$.
4) (*Population*) *Variance* of *Y*: $\sigma^2 = V(Y) = \int (y-\mu)^2 f(y)dy = E(Y^2) - \mu^2$.
5) *Cumulative distribution function* (*cdf*) of *Y*. $F(y) = P(Y \le y) = \int_{-\infty}^{y} f(x)dx$ and *Survival function* $= P(Y > y) = \int_{y}^{\infty} f(x)dx$.

6) The *Population median, M,* is obtained by solving the equation $F(M) = 1/2$ for M.
One may define a median also for a discrete variable, but this can cause problems when trying to obtain an unique solution. We illustrate these properties in two elementary examples. The mathematics needed to solve the problems is found in Section 2.2.3.

EX 1 You throw a symmetric six-sided dice and define the discrete Y = 'Number of dots that comes up'. The pf of Y is obviously $p(y) = 1/6,\ y = 1,...,6$.

1) $\sum p(y) = \sum_{y=1}^{6} \frac{1}{6} = 6 \cdot \frac{1}{6} = 1,$

2) $E(Y) = \sum y \cdot p(y) = \sum_{y=1}^{6} y \cdot \frac{1}{6} = \frac{1}{6} \cdot \frac{6 \cdot (6+1)}{2} = \frac{7}{2}$

3) $E(Y^2) = \sum y^2 \cdot p(y) = \sum_{y=1}^{6} y^2 \cdot \frac{1}{6} = \frac{1}{6} \cdot \frac{6(6+1)(2 \cdot 6+1)}{6} = \frac{91}{6}$

4) $V(Y) = E(Y^2) - \mu^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$

EX2 You arrive at a bus stop where buses run every ten minutes. Define the continuous variable *Y* = 'Waiting time for the next bus'. The density can be assumed to be $f(y) = 1/10, 0 \le y \le 10$.

1) $\sum p(y) = \sum_{y=1}^{6} \frac{1}{6} = 6 \cdot \frac{1}{6} = 1$

2) $E(Y) = \int y \cdot f(y)dy = \int_0^{10} y \cdot \frac{1}{10} dy = \frac{1}{10}\left[\frac{y^2}{2}\right]_0^{10} = \frac{1}{10} \cdot \left(\frac{100-0}{2}\right) = 5$

3) $E(Y^2) = \int y^2 \cdot f(y)dy = \int_0^{10} y^2 \cdot \frac{1}{10} dy = \frac{1}{10}\left[\frac{y^3}{3}\right]_0^{10} = \frac{1}{10}\left(\frac{1000-0}{3}\right) = \frac{100}{3}$

4) $V(Y) = E(Y^2) - \mu^2 = \frac{100}{3} - 5^2 = \frac{25}{3}$

5) $F(y) = \int_0^y \frac{1}{10} dx = \frac{y}{10}$. So, $F(M) = \frac{M}{10} = \frac{1}{2} \Rightarrow M = 5$.

Here the median equals the mean and this is always the case when the density is symmetric around the mean.

One may calculate probabilities such as the probability of having to wait more than 8 minutes,

$$P(Y > 8) = \int_8^{10} \frac{1}{10} dy = \frac{1}{10}(10-8) = \frac{1}{5}$$

More generally, $\alpha_r = E(Y^r)$ is the *r*th *moment* and $\mu_r = E\left((Y-\mu)^r\right)$ the *r*th *central moment*, *r* = 1,2,….

A *bivariate random variable* $\mathbf{Y}$ consists of a pair of variables (Y_1, Y_2). If the latter are discrete the pf of $\mathbf{Y}$ is $p(y_1, y_2) = P(Y_1 = y_1 \cap Y_2 = y_2)$, i.e. the probability of the simultaneous outcome. *Given that* $Y_2 = y_2$ *the conditional probability of* Y_1 is $p\left(y_1 | y_2\right) = P\left(Y_1 = y_1 | Y_2 = y_2\right)$.

Properties of $\mathbf{p(y_1, y_2)}$ (If not shown, summations are over all possible values of y_1 and y_2)

1) $0 \le p(y_1, y_2) \le 1, \sum\sum p(y_1, y_2) = 1$.

2) $\sum_{y_2} p(y_1, y_2) = p(y_1)$ $\sum_{y_1} p(y_1, y_2) = p(y_2)$, $p(y_1)$ and $p(y_2)$ are *marginal* pfs.

3) $p\left(y_1 | y_2\right) = \frac{p(y_1, y_2)}{p(y_2)}$, $p\left(y_2 | y_1\right) = \frac{p(y_1, y_2)}{p(y_1)}$.

4) $\sum_{y_1} p\left(y_1 | y_2\right) = \frac{1}{p(y_2)} \sum_{y_1} p(y_1, y_2) = \frac{1}{p(y_2)} \cdot p(y_2) = 1$.

5) Y_1 and Y_2 are *independent* if $p\left(y_1 | y_2\right) = p(y_1)$ or $p\left(y_2 | y_1\right) = p(y_2)$ or $p(y_1, y_2) = p(y_1) \cdot p(y_2)$.

6) $E\left(g(Y_1) \cdot h(Y_2)\right) = \sum\sum g(y_1) \cdot h(y_2) \cdot p(y_1, y_2)$.

7) *Covariance between* Y_1 and Y_2 :
$\sigma_{12} = Cov(Y_1, Y_2) = \sum\sum (y_1 - \mu_1)(y_2 - \mu_2) \cdot p(y_1, y_2) = E(Y_1 Y_2) - \mu_1 \mu_2.$

Notice that $\sigma_{11} = Cov(Y_1, Y_1)$ is simply the variance of Y_1.

8) *Correlation between* Y_1 and Y_2:

$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \cdot \sigma_2}$, where $\sigma_1^2 = V(Y_1)$ and $\sigma_2^2 = V(Y_2)$ Notice that $-1 \le \rho_{12} \le 1.$

9) The conditional expected value $\mu_{1|2} = E(Y_1 | Y_2 = y_2) = \sum_{y_1} y_1 \cdot p(y_1 | y_2)$ is termed the *regression function*. If this is a linear function of y_2 the regression is *linear*, $\alpha + \beta \cdot y_2$, where α is an *intercept* and β is the *slope* or *regression coefficient*.

10) The conditional variance $V(Y_1 | Y_2 = y_2) = \sum (y_1 - \mu_{1|2})^2 \cdot p(y_1 | y_2) =$

$E(Y_1^2 | Y_2 = y_2) - \mu_{1|2}^2$ is the *residual variance*.

More generally, *a n-dimensional random variable* $\mathbf{Y}$ has *n* components $(Y_1, ..., Y_n)$ and the pf is $p(y_1, ..., y_n) = P(Y_1 = y_1 \cap ... \cap Y_n = y_n)$. This can represent the outcomes in a sample of *n* observations. Assume for instance that we have chosen a sample of *n* families, each with 4 children. Define the variable Y_i = 'Number of boys in family *i*', $i = 1 \ldots n$. In this case it may be reasonable to assume that the number of boys in one chosen family is *independent* of the number of boys in another family. *The probability of the sample* is thus

$$p(y_1, ..., y_n) = p(y_1) \cdot ... \cdot p(y_n) = \prod_{i=1}^{n} p(y_i) \quad (1a)$$

If furthermore each Y_i has the same pf we say that the sequence $(Y_i)_{i=1}^n$ is *identically and independently distributed (iid)*.

Similar relations hold for *n*-dimensional continuous variables. For *n* independent variables the joint density is

$$f(y_1, ..., y_n) = \prod_{i=1}^{n} f(y_i) \quad (1b)$$

Linear form of random variables

Let $(Y_i)_{i=1}^n$ be variables with $E(Y_i)=\mu_i, V(Y_i)=\sigma_{ii}$ and $Cov(Y_i,Y_j)=\sigma_{ij}$. A *linear form* of the Y_j's is $L=\sum_{i=1}^n a_i Y_i$, where the a_i are constants. It is easy to show the following (Wackerly, Mendenhall &Scheaffer 2008, p. 271)

$$E(L)=\sum_{i=1}^n a_i\mu_i$$
$$V(L)=\sum_{i=1}^n a_i^2\sigma_{ii}+2\sum\sum_{1\le i<j\le n} a_i a_j\sigma_{ij} \qquad (2)$$

Consider e.g. the case $n = 3$ in which case $\sum\sum_{1\le i<j\le 3} a_i a_j \sigma_{ij} = a_1a_2\sigma_{12} + a_1a_3\sigma_{13} + a_2a_3\sigma_{23}$. We illustrate the use of eq. (2) below.

EX3 Variance of a sum and of a difference.

$V(Y_1 + Y_2) = [a_1 = 1 = a_2] = \sigma_{11} + \sigma_{22} + 2\sigma_{12}$, $V(Y_1 - Y_2) = [a_1 = 1, a_2 = -1] = \sigma_{11} + \sigma_{22} - 2\sigma_{12}$

Assume further that $\sigma_{11} = \sigma_{22} = \sigma^2$ say. Then $\rho_{12} = \frac{\sigma_{12}}{\sigma^2}$ and it follows that

$V(Y_1 + Y_2) = 2\sigma^2(1 + \rho_{12})$ and $V(Y_1 - Y_2) = 2\sigma^2(1 - \rho_{12})$.

This last equation is interesting because it shows that the variance in data with positively correlated observations can be reduced by forming differences. In fact $V(Y_1 - Y_2) \to 0$ as $\rho_{12} \to 1$. A typical example of positively correlated observations is in 'before-after' studies, e.g. when body weight is measured for each person before and after a slimming program.

2.2 Some distributions

Many discrete and continuous distributions have been found to be workable models for several important practical situations. Such distributions have been termed 'families of distributions' or 'distributional laws'. In this section we catalog some of these and give the basic assumptions on which they are based. We also give means and variances and indicate important properties and applications in following examples. When a certain variable Y follows a certain law L we use the notation $Y \sim L$.

2.2.1 Discrete distributions

1) $Y \sim Bernoulli(p)$. Y is a variable that takes the value 1 with probability p and 0 with probability (1-p). The outcome $Y = 1$ is often termed a 'success' and the outcome $Y = 0$ is termed a 'failure'. The pf is

$$p(y) = p^y(1-p)^{1-y},\ y = 0,1$$

with mean $\mu = p$ and variance $\sigma^2 = p(1-p)$.

2) $Y \sim Binomial(n, p)$. The pf can be derived under the following assumptions: n independent repetitions are made of the same experiment that each time can result in one of the outcomes 'success' with probability p and 'failure' with probability (1-p). Define the variable Y = 'Number of successes that occur in n trials'. The pf is

$$p(y) = \binom{n}{y} p^y(1-p)^{n-y},\ y = 0,1,\ldots,n$$

with $\mu = np$ and $\sigma^2 = np(1-p)$. Notice that $Y = \sum_{i=1}^{n} Y_i$, where $(Y_i)_{i=1}^n$ is a sequence of iid variables, each $\sim Bernoulli(p)$. For the meaning of $\binom{n}{y}$ see Ch.2.3.5 below.

3) $Y \sim Geometric(p)$. Assumptions: Independent repetitions are made of the same experiment that each time can result in one of the outcomes 'success' with probability *p* and 'failure' with probability (1-*p*). Define the variable *Y* = 'Number of trials when a 'success' occurs for the first time'. The pf is

$$p(y) = (1-p)^{y-1} p,\ y = 1,2,...\infty$$

with $\mu = 1/p$ and $\sigma^2 = (1-p)/p^2$. The survival function is $S(y) = P(Y > y) = (1-p)^y$. An interesting property of the Geometric distribution is *the lack of memory*, which means that the probability of a first 'success' in trial number (*y*+1), given that there has been no 'successes' in earlier trials, is the same as the probability of a 'success' in the first trial. Symbolically,

$$P\left(Y = y+1 \middle| Y > y\right) = \frac{P(Y = y+1 \cap Y > y)}{P(Y > y)} = \frac{P(Y = y+1)}{P(Y > y)} = \frac{(1-p)^y p}{(1-p)^y} = P(Y = 1)$$

4) $Y \sim Poisson(\lambda)$. The pf can be derived under a variety of different assumptions. One of the simplest way to obtain the pf is to start with a variable that is *Binomial(n,p)* and to let $n \to \infty$, while at the same time $p \to 0$ in such a way that $n \cdot p \to \lambda$. In practice this means that *n* is large and *p* is so small that the product $n \cdot p = \lambda$ is moderate, say within the interval (0.5, 20). The pf is

$$p(y) = \frac{\lambda^y}{y!} e^{-\lambda},\ y = 0,1,...\infty$$

with $\mu = \lambda$ and $\sigma^2 = \lambda$.

A more general random quantity is $Y(t)$. This is a counting function that describes the number of events that occurs during a time interval of length *t*. It is called a *stationary Poisson process of rate (intensity)* λ and the pf is

$$P\left(Y(t) = y\right) = \frac{(\lambda t)^y}{y!} e^{-\lambda t},\ y = 0,1,...\infty$$

with $E\left(Y(t)\right) = \lambda t = V\left(Y(t)\right)$. λ can be interpreted as the expected number of events per unit time since

$$E\left(\frac{Y(t)}{t}\right) = \frac{1}{t} \cdot E\left(Y(t)\right) = \lambda. \text{ Also, } V\left(\frac{Y(t)}{t}\right) = \frac{1}{t^2} \cdot V\left(Y(t)\right) = \frac{\lambda}{t}.$$

A Poisson process can be obtained under the assumption that the process is a superposition of a large number of independent general point processes, each of low intensity (Cox & Smith 1954, p. 91).

Let $X(s)$ and $Y(t)$ be two independent Poisson processes of rates λ_X and λ_Y, respectively, e.g. number of road accidents during s and t hours on roads with and without limited speed. We are interested in comparing the two intensities in order to draw conclusions about the effect of limited speed on road accidents. One elegant way to do this is to use the *Conditional Poisson Property* (cf. Cox & Lewis 1968, p 223)

$$\text{The conditional variable } (Y(t)\,|\,X(s)+Y(t)=n) \sim \textit{Binomial}(n,\ p=\frac{\lambda_Y \cdot t}{\lambda_X \cdot s+\lambda_Y \cdot t}) \qquad (3)$$

The problem of comparing two intensities can thus be reduced to the problem of drawing inference about one single parameter. Notice that if $\lambda_X = \lambda_Y$ then $p = t/(s+t)$.

The discrete variable $Y(t)$ that counts the number of events in intervals of length t is related to another continuous variable that expresses the length between successive events. (Cf. the theorem (4) in Section 2.2.2.)

5) $Y \sim$. *(Discrete) Uniform(N)*. The pf is

$$p(y) = \frac{1}{N},\ y = 1,2,...,n$$

with $\mu = (N+1)/2$ and $\sigma^2 = (N^2-1)/12$. The distribution put equal mass on each of the outcomes 1,2,…,*N*. A typical example with $N = 6$ is when you throw a symmetric six-sided dice and count the number of dots coming up.

6) $(Y_1,...,Y_k) \sim$ *Multinomial(n,* $p_1,...,p_k$*)*. This is the only example of a discrete many-dimensional variable that is considered in this book. The pf is derived under the same assumptions as for a Binomial variable. However, instead of two outcomes at each single trial, there are *k* mutually exclusive outcomes $A_1,...,A_k$ where the probability of A_i is p_i and $\sum_{i=1}^{k} p_i = 1$. The pf of the variables $Y_i = \text{'Number of times that } A_i \text{ occurs'}, i = 1,...,k$ is

$$p(y_1,...,y_k) = \frac{n!}{y_1!\cdots y_k!} p_1^{y_1}\cdots p_k^{y_k} \text{ with } \sum_{i=1}^{k} y_i = n$$

Verify that $k = 2$ gives the Binomial distribution. Here $\mu_i = E(Y_i) = n \cdot p_i$, $\sigma_{ii} = V(Y_i) = n \cdot p_i(1-p_i)$ and $\sigma_{ij} = Cov(Y_i, Y_j) = -n \cdot p_i p_j, i \neq j$.

EX 4 Let *Y* be the variable 'Number of boys in a randomly chosen family with 4 children'. This can be assumed to be Binomial(*n*, *p*) with $n = 4$ and $p = 53/103 \approx 0.516$, the latter figure being obtained from population statistics in the Scandinavian countries (106 born boys on 100 born girls). By using the pf in *(2)* above one gets

$$p(0) = \binom{4}{0}(53/103)^0(50/103)^4 = 0.056,\ p(1) = \binom{4}{1}(53/103)^1(50/103)^3 = 0.235,$$

$$p(2) = \binom{4}{2}(53/103)^2(50/103)^2 = 0.374,\ p(3) = \binom{4}{3}(53/103)^3(50/103)^1 = 0.265,$$

$$p(4) = \binom{4}{4}(53/103)^4(50/103)^0 = 0.070$$

These probabilities are very close to the actual relative frequencies. However, it should be kept in mind that calculations have been based on crude figures and the results may not be true in other populations. E.g. if both parents are smokers the proportion born boys is only 0.451 or 82 born boys on 100 born girls (Fukada *et al* 2002, p. 1407).

EX 5 In Russian roulette a revolver with place for 6 bullets is loaded with one bullet. You spin the revolver, direct it towards your head and then fire. Define the variable *Y* = 'Number of trials until the bullet hits your head for the first time (and probably the last).' The variable can be assumed to have a Geometric distribution with *p* = 1/6. In this case it is perhaps not that interesting to compute the probability that the revolver fires after exact *y* trials, but the probability to survive *y* trials. From the expression above in *(3)*, Ch. 2.2.1, we get the survival function

$$S(y) = P(Y > y) = (5/6)^y \text{, } y = 1,2,...\infty$$

A few values are:

y	1	2	3	4	5	6
S(y)	0.83	0.69	0.58	0.48	0.40	0.33

The median is somewhere between 3 and 4 trials which implies that after 3 successive trials most of the candidates will have been hit by the bullet. Russian roulette has been a motive in several films such as "The Deer Hunter", "The Way of the Gun" and "Leon", just to mention a few ones. The next time you are watching such a film you should have the table above in your mind.

EX 6 Let $X(s)$ be a Poisson process of rate λ_X representing the number of road accidents on a road segment. During 12 months it is noticed that there has been 18 accidents, so that λ_X may be put equal to 18/12 = 1.5. One can now calculate the probability of several outcomes such as

- At least one accident in *s* months, $P(X(s) \geq 1) = \sum_{x=1}^{\infty} p(x) = 1 - p(0) = 1 - e^{-1.5 \cdot s}$, which tends to 1with increasing values of *s*.
- At least one accident in 1 month, $P(X(1) \geq 1) = 1 - e^{-1.5} = 0.777$.
- At least two accidents in 1 month, $P(X(1) \geq 2) = 1 - p(0) - p(1) = 1 - e^{-1.5} - 1.5 \cdot e^{-1.5} = 0.442$.
- At least two accidents in one month given that at least one accident has occurred,

  $P(X(1) \geq 2 | X(1) \geq 1) = \frac{P(X(1) \geq 1 \cap X(1) \geq 2)}{P(X(1) \geq 1)}$ = [The intersection of the two events in the numerator is simply $X(1) \geq 2$]= $\frac{P(X(1) \geq 2)}{P(X(1) \geq 1)} = \frac{0.442}{0.777}$ = 0.569.

EX7 Assume that speed limits are introduced on the road segment in EX 6 and after this one observe 3 accidents in 3 months. The rate of accidents has thus decreased from 1.5 to 1.0 per month. Does this imply that restricted speed has had an effect on accidents, or is the decrease just temporary? We will later present some ways to tackle this question (Cf. Ch. 6), but for the moment we just show how the problem of comparing two rates can be reformulated.

Let $Y(t)$ be the Poisson process of accidents during time *t* after the introduction of speed limits and let the rate be λ_Y. According to formula (3) in this section the variable $(Y(3) | X(12) + Y(3) = 21)$ is Binomial (*n*,*p*) with *n* = 21 and $p = \lambda_Y \cdot 3/(\lambda_X \cdot 12 + \lambda_Y \cdot 3)$. If $\lambda_X = \lambda_Y$ then *p* = 1/5, to be compared with the observed proportion 3/21 = 1/7.

EX 8 (Y_1, Y_2, Y_3) is a Multinomial variable (n, p_1, p_2, p_3). The pf is $p(y_1, y_2, y_3 = \frac{n!}{y_1! y_2! y_3!} p_1^{y_1} p_2^{y_2} p_3^{y_3}$. The outcomes are often referred to as *cell frequencies*.

The mean and variance of $Y_1 - Y_2$ are

$E(Y_1 - Y_2) = \mu_1 - \mu_2 = np_1 - np_2 = n \cdot (p_1 - p_2)$

$V(Y_1 - Y_2) = \sigma_{11} + \sigma_{22} - 2\sigma_{12} = np_1(1 - p_1) + np_2(1 - p_2) - 2np_1p_2$ = [After some re-arrangements] = $n \cdot (p_1 + p_2)(1 - (p_1 + p_2))$

2.2.2 Continuous distributions

A convenient way to summarize the properties of a continuous distribution is to calculate the (symmetric) *variation limits* (c_1, c_2). These are the limits within which a certain percentage of all observations will fall. E.g. the 95% limits are obtained by solving the two equations $P(Y < c_1) = 0.025$ and $P(Y > c_2) = 0.025$ for c_1 and c_2. (Cf. EX 9-EX12.)

1. *Uniform distribution on the interval* $[a, b]$, $Y \sim Uniform[a, b]$.

   Density $f(y) = \begin{cases} \frac{1}{(b-a)}, & a \le y \le b, \\ 0, & \text{otherwise} \end{cases}$ cdf $F(y) = \begin{cases} 0, & y < a \\ \frac{(y-a)}{(b-a)}, & a \le y \le b \\ 1, & y > b \end{cases}$

   It is easy to show that $\mu = (b-a)/2$ and $\sigma^2 = (b-a)^2/12$.

2) *Gamma distribution,* $Y \sim Gamma(\lambda, k)$

   This is a class of distributions that is closely connected with the Gamma function $\Gamma(k)$ (Cf. Section 2.3.5.). The general form of the density is

   $$f(y) = \frac{\lambda^k}{\Gamma(k)} y^{k-1} e^{-\lambda \cdot y}, \; y \ge 0, \lambda > 0, k > 0.$$

   Notice that the integral of the density over all values of *y* is 1, a property that can be used in computations. Two important special cases are:

   - *Exponential distribution,* $k = 1$, $Y \sim Exponential(\lambda)$, with density $f(y) = \lambda e^{-\lambda \cdot y}$.
   - *Chi-square distribution with n degrees of freedom (df)* $\lambda = 1/2$ and $k = n/2$, $Y \sim \chi^2(n)$.

   The cdf can only be expressed explicitly if *k* is a positive integer, $F(y) = 1 - \sum_{i=0}^{k-1} \frac{(\lambda y)^i}{i!} e^{-\lambda \cdot y}$. In the exponential case we thus get $F(y) = 1 - e^{-\lambda \cdot y}$. An important theorem that links the Exponential distribution to the Poisson process in Section 2.2.1 is the following:

Let (X_i) be a sequence of times-between-successive events. Then we have the identity

$$Y(t) \text{ is a Poisson process of rate } \lambda \Leftrightarrow \begin{cases} 1)\ Each\ X_i \sim Exponential(\lambda) \\ 2)\ The\ X_i\ are\ independent \end{cases} \quad (4)$$

This gives us a simple clue to determine whether a given sequence of events follow the Poisson law or not: (1) Make a histogram of X_i and compare it with the Exponential density, (2) Make a plot of each interval length versus the length of the following interval (X_i versus X_{i+1}) and compute the correlation. For more refined methods the reader is referred a book by Cox & Lewis 1966, p. 152.

For $Y \sim Gamma(\lambda, k)$ we have $\mu = k/\lambda$ and $\sigma^2 = k/\lambda^2$. More generally $E(Y^r) = \frac{1}{\lambda^r}\frac{\Gamma(k+r)}{\Gamma(k)}$ which holds for r = K – 2, – 1, 0, 1, 2K. Special case: $k = 1 \Rightarrow \alpha_r = E(Y^r) = \frac{r!}{\lambda^r}$.

The following theorem makes it possible to calculate areas under the Gamma density by using tables for Chi-square variables that are found in most textbooks:

$$Y \sim Gamma(\lambda, k) \Rightarrow 2\lambda Y \sim \chi^2(2k) \quad (5)$$

An application of this is given in EX 11 below.

3) *Weibull distribution,* $Y \sim W(\alpha, \lambda)$. This has the density

$$f(y) = \alpha \cdot \lambda \cdot y^{\alpha-1} e^{-\lambda \cdot y^{\alpha}}, y \geq 0, \alpha > 0, \lambda > 0$$

Here $\mu = \frac{\Gamma(1+1/\alpha)}{\lambda^{1/\alpha}}$ and $\sigma^2 = \frac{\Gamma(1+2/\alpha) - \Gamma^2(1+1/\alpha)}{\lambda^{2/\alpha}}$. The cdf is $F(y) = 1 - e^{-\lambda \cdot y^{\alpha}}$. This distribution is obtained from the relation $Y = X^{1/\alpha}$, where $X \sim$ Exponential(λ).

Applications can be found in survival analysis and reliability engineering.

4) *Normal distribution,* $Y \sim N(\mu, \sigma^2)$ has the density

$$f(y) = \frac{1}{\sqrt{2\pi \cdot \sigma^2}} e^{\frac{(y-\mu)^2}{2\cdot\sigma^2}}, -\infty < y < \infty,$$

where μ and σ^2 is the mean and variance, respectively. A *standard normal variable* is obtained by putting $\mu = 0$ and $\sigma^2 = 1$. The latter is denoted $Z \sim N(0,1)$ and will be used to compute areas under the normal density in a way that is described in EX 12 below. Notice that $Z = (Y - \mu)/\sigma$, the transformation is called *standardization*.

The normal distribution can be obtained as a limiting distribution in several ways. Some of these are listed below in (a) to (c), where the one in (a) is formulated as a theorem due to its importance. A proof of (a) can be found in Casella & Berger 1990, p. 217. A proof of (c) can be found in Cramer 1957, p. 250.

a) Central Limit Theorem (*CLT*). Let $(Y_i)_{i=1}^n$ be a sequence of independent and identically distributed (iid) variables with mean μ and variance σ^2. Then the cdf of the standardized variable

$$Z_n = \frac{\sum_{i=1}^{n} Y_i - E\left(\sum_{i=1}^{nn} Y_i\right)}{\sqrt{V\left(\sum_{i=1}^{n} Y_i\right)}} = \frac{\sum_{i=1}^{n} Y_i - n \cdot \mu}{\sqrt{n \cdot \sigma^2}} = \frac{\bar{Y} - \mu}{\sqrt{\sigma^2 / n}}$$ tends to the cdf of $Z \sim N(0,1)$ as $n \to \infty$.

This is denoted $Z_n \xrightarrow{D} Z$, as $n \to \infty$. (6)

b) If $Y \sim Binomial(n, p)$ then $Z_n = \frac{Y - np}{\sqrt{np(1-p)}} \xrightarrow{D} Z \sim N(0,1)$, as $n \to \infty$

c) If $Y(t)$ is a Poisson process with rate λ then $Z(t) = \frac{Y(t) - \lambda t}{\sqrt{\lambda t}} \xrightarrow{D} Z \sim N(0,1)$ as $t \to \infty$, or alternatively, with $t = 1\, Z(1) \xrightarrow{D} Z \sim N(0,1)$ as $\lambda \to \infty$.

Comments

- The CLT was first formulated and proved by the French mathematician Laplace about 1778 (exact year is hard to establish). Notice that it is the *standardized* variable that has a normal distribution as a limit. In some textbooks you may find expressions like '$\overline{Y}$ has a limiting Normal distribution with mean μ and variance σ^2 / n'. But this is not true since the distribution of $\overline{Y}$ tends to a 'one-point' distribution at μ with variance zero.
- As you might suspect, the result in *(b)* is simply a result of the CLT since $Y \sim Binomial(n, p)$ can be expressed as $Y = \sum_{i=1}^{n} Y_i$ where the Y_i are iid with a Bernoulli distribution. However, this result was published earlier than that of the CLT, in November 12, 1733 by the French mathematician de Moivre and it seems to be the first time that the formula of the normal density appears.
- Further results were later obtained by the German mathematician K.F. Gauss (1809) and the Russians Markov (1900) and Liapuonov (1901). It has been found that the limiting Z -distribution exists under less restricted assumptions than mentioned in (a) above.
- Many distributions are related to $Z \sim N(0,1)$, e.g. $Z^2 \sim \chi^2(1)$.
- If $Y_i \sim N(\mu_i, \sigma_i^2)$ then $L = \sum a_i Y_i \sim N$ with mean and variance given in (2), Ch. 2.1.

5) *Laplace distribution, Laplace's first law or the double Exponential distribution,* $Y\sim L(\mu,b)$. The density and cdf are

$$f(y)=\frac{1}{2b}e^{-\frac{|y-\mu|}{b}} \text{ and } F(y)=\begin{cases}(1/2)\cdot e^{\frac{y-\mu}{b}}, y<b\\ 1-(1/2)\cdot e^{-\frac{(y-\mu)}{b}}, y\geq b\end{cases}$$

With mean μ and $\sigma^2=2b^2$.

This distribution and its generalizations to non-symmetric casas has important applications in engineering and finance.

EX 9 Assume that waiting times are distributed *U[0,b]*. Compute the mean and the median waiting time and also the 95% variation limits.

$\mu=b/2, F(M)=\frac{M}{b}=(\text{Put})=1/2\Rightarrow M=b/2$.

95 % variation limits are obtained from: $P(Y<c_1)=F(c_1)=\frac{c_1}{b}=(\text{Put})=\frac{0.05}{2}\Rightarrow c_1=0.025b$,

$P(Y>c_2)=1-F(c_2)=1-\frac{c_2}{b}=(\text{Put})=0.025\Rightarrow c_2=0.975b$. The 95 % variation limits are thus (0.025*b*, 0.975*b*). E.g. if a bus runs every 20 minutes from a bus stop, 95 % of the waiting times will range from 0.5 to 19.5 minutes.

EX 10 Intervals between arrivals to an intensive care are distributed $Exponential(\lambda)$. Compute the mean and median interval and give the 95% variation limits.

$\mu=1/\lambda, F(M)=1-e^{-\lambda\cdot M}=(\text{Put})=1/2\Rightarrow e^{-\lambda\cdot M}=1/2\Rightarrow\lambda\cdot M=\ln(2), \text{so } M=\frac{\ln(2)}{\lambda}\approx\frac{0.67}{\lambda}$

$P(Y<c_1)=1-e^{-\lambda\cdot c_1}=(\text{Put})=0.025\Rightarrow e^{-\lambda\cdot c_1}=0.975\Rightarrow c_1=-\ln(0{,}975)/\lambda\approx 0.025/\lambda.$

$P(Y>c_2)=1-P(Y<c_2)=e^{-\lambda\cdot c_2}=(\text{Put})=0.025\Rightarrow c_2=-\ln(0.025)/\lambda\approx 3.69/\lambda$

EX 11 Assume that service times (minutes) for a customer at a cash machine are distributed *Gamma*($\lambda = 2, k = 2$). Determine the mean and median service times and give the 95 % variation limits for the service times.

$\mu = k / \lambda = 2 / 2 = 1.$

$P(Y < M) = [\text{Notice the trick}] = P(2\lambda Y < 2\lambda M) = P(\chi^2(4) < 2\lambda M) = (\text{Put})$ =1/2. From a table of the Chi-square distribution we get $2\lambda M = 3.36 \Rightarrow M = 3.36/4 =$ 0.84.

$P(Y < c_1) = P(2\lambda Y < 2\lambda c_1) = P(\chi^2(4) < 2\lambda c_1) = (\text{Put}) = 0.025$. The same tables gives $2\lambda c_1 = 0.48$ so $c_1 = 0.12$. $P(Y > c_2) = P(2\lambda Y > 2\lambda c_2) = P(\chi^2(4) > 2\lambda c_2) = (\text{Put}) = 0.025$. From this $2\lambda c_2 = 11.14$, so $c_2 = 2.79$.

In this example we have used the theorem in (5)

EX 12 $Y \sim N(\mu, \sigma^2)$. Determine the 95 % variation limits for Y.

$P(Y < c_1) =$ [Notice the trick]$= P\left(\frac{Y-\mu}{\sigma} < \frac{c_1-\mu}{\sigma}\right) = P\left(Z < \frac{c_1-\mu}{\sigma}\right) = (\text{Put}) = 0.025 \Rightarrow$

$\frac{c_1-\mu}{\sigma} = -1.96 \Rightarrow c_1 = \mu - 1.96\sigma$. Similarly we get $c_2 = \mu + 1.96\sigma$

2.3 Mathematics

Some mathematics will be needed when solving problems in statistical inference. Here we consider a few results that will be needed.

2.3.1 Functions of a single variable

A function $y = f(x)$ maps one set of *x*- values on one set of *y*- values. The function is called one-to-one if only one *x*- value correspond to a *y*- value. In such a case one can obtain the reversed map, the *inverse function* $x = f^{-1}(y)$. Consider the function $y = x^2, -\infty < x < \infty$, which maps values along the whole *x*- values on the positive *y*- axis. It is not one-to-one since e.g. both *x* = -1 and *x* = 1 gives *y* =1. On the other hand, $y = x^2, 0 \le \mathrm{x} < \infty$ is one-to-one with the inverse function $x = \sqrt{y}$.

Some simple functions

- *Straight line,* $y = a + b \cdot x$, a is the *intercept* and b is the *slope*.
- *Exponential,* $y = ab^x$. With $a = 1$ and $b = e \approx 2.7182$, $y = e^x$ having the following properties: $e^{-x} = 1/e^x, e^{x_1} \cdot e^{x_2} = e^{x_1+x_2}, \left(e^{x_1}\right)^{x_2} = e^{x_1 x_2}$
- *Potense,* $y = ax^b$
- *Logarithmic (natural),* $y = \ln(x)$ having the following properties: $\ln(0) \to \infty$, $\ln(1) = 0$, $\ln(e) = 1$,

  $\ln(x_1 x_2) = \ln(x_1) + \ln(x_2)$, $\ln(x_1 / x_2) = \ln(x_1) - \ln(x_2)$, $\ln(x^b) = b\ln(x)$, $\ln(e^x) = x$. If $y = \ln(x)$ then $e^y = x$
- *Logistic (S-curve),* $y = e^l / (1 + e^l)$, where $l = a + b \cdot x$.

Linearization of non-linear functions

- $y = ab^x$. Taking logarithms on both sides gives $y' = \ln(y) = \ln(ab^x) = \ln(a) + x\ln(b) = a' + b'x$. So x plotted against $\ln(y)$ gives a straight line.
- $y = ax^b$. $y' = \ln(y) = \ln(ax^b) = \ln(a) + b\ln(x) = a' + bx'$. So, $\ln(x)$ plotted against ln(y) gives a straight line.
- $y = e^l/(1 + e^l)$, with $l = a + b \cdot x$. Now $y/(1 - y) = e^l$, so $y' = \ln\left(y/(1-y)\right) = l = a + b \cdot x$ and thus a plot of x against $\ln(y/(1-y))$ gives a straight line.

2.3.2 Sums and products

The *sum* of $x_1,...,x_n = x_1 + ... + x_n = \sum_{i=1}^{n} x_i$. The x_i are *terms* Sometimes we drop the lower or upper index in the summation sign if they are obvious. The *product* of $x_1,...,x_n = x_1 \cdots x_n = \prod_{i=1}^{n} x_i$. The x_i are now termed *factors.*

Some rules

- $(x_1 + x_2)^2 = x_1^2 + x_2^2 + 2x_1x_2$. More generally: $\left(\sum_{i=1}^{n} x_i\right)^2 = \sum_{i=1}^{n} x_i^2 + 2\sum_{1\le i<j\le n} x_i x_j$. Notice that the last sum contains $n^2 - n$ terms of the form $x_i x_j$.
- $\prod_{i=1}^{n} e^{x_i} = e^{x_1} \cdots e^{x_n} = e^{\sum_i x_i}$, $\ln\left(\prod e^{x_i}\right) = \sum x_i$
- $\sum_{i=1}^{n} a \cdot x_i = a\sum_{i=1}^{n} x_i$, $\prod_{i=1}^{n} a \cdot x_i = a^n \prod_{i=1}^{n} x_i$, where *a* is a constant.

EX 13 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is termed the *arithmetic mean*. Obviously $\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} =$

$\sum_{i=1}^{n} x_i - n \cdot \bar{x} = 0$.

Let *a* be an arbitrary constant. Then $\sum_{i=1}^{n}(x_i - a)^2$ is minimized if $a = \bar{x}$.

Proof: $\sum_{i=1}^{n}(x_i - a)^2 = [\text{Notice the trick}] = \sum_{i=1}^{n}\left((x_i - \bar{x}) + (\bar{x} - a)\right)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(\bar{x} - a)^2 +$

$2\sum_{i=1}^{n}(x_i - \bar{x})(\bar{x} - a) = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - a)^2 + 2(\bar{x} - a)\sum_{i=1}^{n}(x_i - \bar{x})$, where the last term is zero.

Notice that $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 + n \cdot \bar{x}^2 - 2\bar{x}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$. The latter expression is often simpler to use in calculations.

2.3.3 Derivatives

The *derivative* of $y = f(x)$ with respect to *x* is the limit $f'(x) = \lim\left(f(x+h) - f(x)\right)/h$ as $h \to 0$. Other notations for a derivative are $y', \frac{dy}{dx}, \frac{df}{dx}$ or $D_x f$. Rather than having to calculate the limit it is easier to use the following rules.

Derivation rules

1) *Special functions*

$f(x)$:	$a+bx$	x^b	e^x	$e^{g(x)}$	$\ln(x)$
$f'(x)$:	b	bx^{b-1}	e^x	$g'(x)e^{g(x)}$	$1/x$

2) $f(x)=g(x)\pm h(x)\Rightarrow f'(x)=g'(x)\pm h'(x)$
3) $f(x)=g(x)\cdot h(x)\Rightarrow f'(x)=g'(x)\cdot h(x)+g(x)\cdot h'(x)$
4) $f(x)=g(x)/h(x)\Rightarrow f'(x)=\big(g'(x)\cdot h(x)-g(x)\cdot h'(x)\big)/h^2(x)$
5) $f(x)=g\big(h(x)\big)\Rightarrow f'(x)=h'(x)g'(h)$. This is a very useful rule that is demonstrated in EX 14 below.

EX 14 $f(x)=\ln(3x+1)$. Put $h(x)=3x+1$ and $g(h)=\ln(h)$ in *(5)* above, with $h'(x)=3$, $g'(h)=1/h$. Then $f'(x)=3/(3x+1)$

$f(x)=\sqrt{2x}$. Put $h(x)=2x$ and $g(h)=\sqrt{h}=h^{1/2}$, with $h'(x)=2,\ g'(h)=1/2\cdot h^{-1/2}=\dfrac{1}{2\sqrt{h}}$. Then

$$f'(x)=\frac{1}{\sqrt{2x}}.$$

$y=(x-a)^2.\ \dfrac{dy}{dx}=2(x-a),\ \dfrac{dy}{da}=(-1)\cdot 2(x-a)=2(a-x)$. The function *y* can be considered as a function of either *x* or *a*.

$$y=\sum_{i=1}^{n}(x_i-a)^2.\ \frac{dy}{dx_i}=\big[\text{There is just one } x_i\big]=2(x_i-a),\ \frac{dy}{da}=\sum_{i=1}^{n}(-1)\cdot 2(x_i-a)=2\sum_{i=1}^{n}(a-x_i)$$

Two important theorems about extreme values

- If $f(x)$ has a local maximum (max) or minimum (min) at $x=x_0$ then this can be obtained by solving the equation $f'(x)=0$ for $x=x_0$. Furthermore, from the sign of the second derivative $f''(x)$, we draw the following conclusions:

$$f''(x_0)\begin{cases}>0\Rightarrow f(x)\text{ has a local min at } x=x_0\\<0\Rightarrow f(x)\text{ has a local max at } x=x_0\end{cases}$$

- If $f(x)>0$ then $f(x)$ has a local max or min at the same *x*- value as $\ln\big(f(x)\big)$

EX 14 Does the function $f(x)=e^{-(x-1)^2}$ have any max/min-values? Since $f(x)>0$ we prefer to study the simpler function $z(x)=\ln\big(f(x)\big)=-(x-1)^2$. Since $z'(x)=-2(x-1)=0\Rightarrow x_0=1$, this must be a value of interest. Now, $z''(x)=-2<0,$ from which we conclude that the function has a local maximum at $x=1.$

2.3.4 Integrals

The (Riemann) *integral* $\int_a^b f(x)dx$ is the area between a and b under the curve $f(x)$.

Integration rules

1) $\int_a^b f(x)dx = [F(x)]_{x=a}^{x=b} = F(b) - F(a)$ where F is a *primitive function* to f. Since $F'(x) = f(x)$ we can use the derivation rules above to find primitive functions.

2) $\int_a^b \big(g(x) \pm h(x)\big)dx = \int_a^b g(x)dx \pm \int_a^b h(x)dx$

3) $\int_a^b g(x) \cdot h(x)dx = [G(x)h(h)]_{x=a}^{x=b} - \int_a^b G(x)h'(x)dx$ (Partial integration)

EX 15 $\int_0^1 (1-x)dx = \left[x - \frac{x^2}{2} \right]_{x=0}^{x=1} = 1 - \frac{1}{2} - (0-0) = \frac{1}{2}.$

$\int_0^1 \frac{1}{\sqrt{x}} dx = \int_0^1 x^{-1/2} dx = \left[\frac{x^{1/2}}{1/2} \right]_{x=0}^{x=1} = 2 - 0 = 2.$

$\int_0^\infty e^{-x} dx = \left[-e^{-x} \right]_{x=0}^{x\to\infty} = -0 - (-e^{-0}) = 1.$ An area under an infinitely long interval can thus be finite. This is an example of a mathematical paradox since it would imply that we could paint an infinitely long fence, having an exponential shape, with a finite amount of paint.

2.3.5 Some special functions and relations

Let n be any of the integers 0,1,2,…. Then $n!$ ('n faculty') equals 1 for $n = 0$ and $1 \cdot 2 \cdots n$ for $n > 0$.

The *combination operator* $\binom{n}{x} = \frac{n!}{x!(n-x)!}$. E.g. $\binom{5}{2} = \frac{5!}{2!3!} = 10.$

Some series

- $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}, \ \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}.$
- *Geometric* $\sum_{i=0}^{n} x^i = \frac{1-x^{n+1}}{1-x}. \ \sum_{i=0}^{\infty} x^i = \frac{1}{1-x}$, provided that -1< x <1.
- *Binomial* $\sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i} = (a+b)^n$
- *Exponential* $\sum_{i=0}^{\infty} \frac{x^i}{i!} = e^x$
- *Taylor* Let $f^{(i)}(a)$ be the $i:th$ derivative of $f(x)$ computed at $x = a$ with $f^{(0)}(a) = f(a)$. Then $f(x) = \sum_{i=0}^{\infty} \frac{(x-a)^i}{i!} f^{(i)}(a)$. In practice this may be used to approximate $f(x)$ by a polynomial. E.g. $f(x) \approx f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2} f''(a)$. In this case $f(x)$ has been approximated by a Taylor polynomial of order 2 about a.

EX 16

$\sum_{i=1}^{\infty} 0.8^i = \sum_{i=0}^{\infty} 0.8^i - 0.8^0 = \frac{1}{1-0.8} - 1 = 4. \ \sum_{i=0}^{8} \binom{8}{i} = [\text{Put } a = 1 = b] = (1+1)^8 = 256.$

Let $0 < p < 1$ and consider $\sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} = (p+1-p)^n = 1$-

Gamma function

For any *p,* define the *Gamma function* $\Gamma(p) = \int_0^\infty x^{p-1}e^{-x}dx$. Tables of this function can be found in Standard Mathematical Tables. Tables can also be produced by using program packages such SAS, SPSS or Statistica. The behavior of the function is quite complicated but we will only need the following properties:

- $\Gamma(p+1) = p \cdot \Gamma(p)$
- $\Gamma(p+1) = p!$ if p =0, 1, 2,…

Cauchy-Schwarz inequality

Let x_i and y_i be real numbers. Then $\left(\sum x_i y_i\right)^2 \le \left(\sum x_i^2\right)\left(\sum y_i^2\right)$.

2.4 Final words

Notice the difference between a discrete and a continuous variable when calculating probabilities. For a continuous variable Y the probability $P(Y = y)$ is always 0. This implies that $P(Y \ge y) = P(Y > y)$. On the other hand, for a discrete variable, $P(Y \ge y = P(Y = y)) + P(Y > y)$.

The *population median M* is a value such that $F(M) = 1/2$ and nothing else. The *sample median m* is obtained by ranking the observations in a sample and to let *m* be the observation in the middle, or the average of the observations in the middle. *m* may be used as an estimate of *M*.

In Ch. 2 we only considered discrete bivariate distributions. Continuous bivariate distributions are treated analogously. The essential difference is that all summation symbols in properties *(1)-(10)* are replaced by integrals.

The reader is encouraged to use the summation symbol $\sum_{i=1}^{n} x_i$ rather than $x_1 + \ldots + x_n$ and the product symbol $\prod_{i=1}^{n} x_i$ rather than $x_1 \cdot \ldots \cdot x_n$. In the book we will use alternative symbols for division. To save space we write a/b instead of $\frac{a}{b}$. A typical example is $\frac{a/b}{c/d + e/f}$.

3 Sampling Distributions

Data consist of observations $y_{1,},...,y_n$ (numerical values) that have been drawn from a population. The latter may be called a *specific sample*. If we want to guess, or estimate, the value of a population characteristic such as the population mean μ one may take the sample mean $\bar{y} = \sum y_i / n$. Any new sample of n observations drawn from the population will give rise to a new set of y – values and thus also of $\bar{y}$. To understand this variation from sample to sample it is useful to introduce the concept of a *random sample of size n*, $Y_1,...,Y_n$. Throughout this book it will be assumed that the latter variables are independent so that the probability of the sample can be expressed as in (1a) and (1b).

The appropriateness of taking the sample mean as a guess for μ can be judged by studying the distribution of $\bar{Y}$ and calculate the dispersion around μ. However, $\bar{Y}$ is just one possible function of $Y_1,...,Y_n$, and there might be other functions that are better in some sense. Every function of the n-dimensional variable is termed a *statistic* with the general notation $T = g(Y_1,...,Y_n)$. The distribution of T is called a *sampling distribution*. If the purpose is to estimate a characteristic in the population, T is called an *estimator* and a numerical value of T is called an *estimate*, t. If the purpose is to find an interval (T_1, T_2) that covers the population characteristic with a certain probability it is called a *confidence interval* (*CI*). Finally, the statistic is called a *test-statistic* if the purpose is to use it for testing a statistical hypothesis. In this chapter we consider some exact and approximate results of sampling distributions..

3.1 Some exact sampling distributions

Sum of variables

1) $Y_i \sim Bernoulli(p) \Rightarrow \sum_{i=1}^{n} Y_i \sim Binomial(n, p)$

2) $Y_i \sim Binomial(n_i, p) \Rightarrow \sum_{i=1}^{k} Y_i \sim Binomial\left(\sum_{i=1}^{k} n_i, p\right)$

3) $Y_i \sim Poisson(\lambda_i) \Rightarrow \sum_{i=1}^{n} Y_i \sim Poisson\left(\sum_{i=1}^{n} \lambda_i\right)$

4) $Y_i \sim N(\mu_i, \sigma_i^2) \Rightarrow \sum_{i=1}^{n} a_i Y_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$

5) Special case with $\mu = \mu, \sigma_i^2 = \sigma^2$ and $a_i = 1/n$: $\overline{Y} \sim N(\mu, \sigma^2 / n)$

6) $Y_i \sim Gamma(\lambda, k_i) \Rightarrow \sum_{i=1}^{n} Y_i \sim Gamma\left(\lambda, \sum_{i=1}^{n} k_i\right), \frac{\sum_{i=1}^{n} Y_i}{n} \sim Gamma\left(n\lambda, \sum_{i=1}^{n} k_i\right)$

7) Special case with $k_i = 1$: $Y_i \sim Exponential(\lambda) \Rightarrow \sum_{i=1}^{n} Y_i \sim Gamma(\lambda, n)$

8) Special case with $\lambda = 1/2$ and $k_i = n_i / 2$: $Y_i \sim \chi^2(n_i) \Rightarrow \sum_{i=1}^{k} Y_i \sim \chi^2\left(\sum_{i=1}^{k} n_i\right)$

Sum of quadratic forms

9) $Y_i \sim N(\mu, \sigma^2) \Rightarrow \frac{\sum_{i=1}^{n} (Y_i - \mu)^2}{\sigma^2} \sim \chi^2(n)$, or $\sum_{i=1}^{n} (Y_i - \mu)^2 \sim \sigma^2 \cdot \chi^2(n)$. Notice that the sign '~' (distributed as) can be treated in the same way as the equality sign.

10) $Y_i \sim N(\mu, \sigma^2) \Rightarrow \frac{\left(\overline{Y} - \mu\right)^2}{\sigma^2 / n} \sim \chi^2(1)$, or $\left(\overline{Y} - \mu\right)^2 \sim \frac{\sigma^2}{n} \chi^2(1)$.

An important theorem on chi-square distributed quadratic forms is the following theorem (Cochran, 1934)

Cochran's Theorem: Let Q_1, Q_2 and Q_3 be quadratic forms such that $Q_1 = Q_2 + Q_3$ then

$$Q_1 \sim \chi^2(n_1) \text{ and } Q_2 \sim \chi^2(n_2) \Rightarrow \begin{cases} Q_3 \sim \chi^2(n_1 - n_2) \\ Q_2 \text{ and } Q_3 \text{ are independent} \end{cases} \quad (7)$$

EX 17 Prove the relations in *(9)* and *(10)* above.

$$Y_i \sim N(\mu,\sigma^2) \Rightarrow \frac{Y_i-\mu}{\sigma} \sim N(0,1) \Rightarrow \frac{(Y_i-\mu)^2}{\sigma^2} \sim \chi_i^2(1) \Rightarrow \frac{\sum_{i=1}^{n}(Y_i-\mu)^2}{\sigma^2} \sim \chi^2(n)\cdot$$

$$Y_i \sim N(\mu,\sigma^2) \Rightarrow \overline{Y} \sim N(\mu,\sigma^2/n) \Rightarrow \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \Rightarrow \frac{(\overline{Y}-\mu)^2}{\sigma^2/n} \sim \chi^2(1)\cdot$$

EX 18 Use Cochran's Theorem to show that $Y_i \sim N(\mu,\sigma^2) \Rightarrow \frac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}{\sigma^2} \sim \chi^2(n-1)$.

$$Y_i-\mu=(Y_i-\overline{Y})+(\overline{Y}-\mu) \Rightarrow \sum_{i=1}^{n}(Y_i-\mu)^2=\sum_{i=1}^{n}(Y_i-\overline{Y})^2+\sum_{i=1}^{n}(\overline{Y}-\mu)^2+$$

$\sum_{i=1}^{n}2(Y_i-\overline{Y})(\overline{Y}-\mu)$. Here the last term is $2(\overline{Y}-\mu)\sum_{i=1}^{n}(Y_i-\overline{Y})=0$ (cf. EX 13). So,

$$\frac{\sum_{i=1}^{n}(Y_i-\mu)^2}{\sigma^2}=\frac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}{\sigma^2}+\frac{(\overline{Y}-\mu)^2}{\sigma^2/n} \text{ or } Q_1=Q_2+Q_3$$

The result now follows from *(9)* and *(10)* above.

EX 18 (Continued) The *sample variance* is defined as $S^2=\frac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}{n-1} \sim \frac{\sigma^2}{(n-1)}\chi^2(n-1)$. Notice that Q_2 is a function of S^2 and Q_3 is a function of $\overline{Y}$. Since Q_2 and Q_3 are independent it follows that S^2 *and* $\overline{Y}$ *are independent random variables*. So, if we repeatedly compute S^2 and $\overline{Y}$ in samples from a normal distribution we will obtain a zero correlation. This may seem to be amazing since S^2 is functionally dependent of $\overline{Y}$, but it illustrates that statistical dependency and functional dependency are two different concepts.

Ratios

11) *Student's T with f degrees of freedom,* $T(f)$

$$Z \sim N(0,1) \text{ and } V \sim \chi^2(f) \text{ are independent} \Rightarrow \frac{Z}{\sqrt{V/f}} \sim T(f)$$

Tables showing areas under the density of T can be found in most elementary text books.

12) *Variance ratio F with* f_1 and f_2 *degrees of freedom,* $F(f_1,f_2)$

$$V_1 \sim \chi^2(f_1) \text{ and } V_2 \sim \chi^2(f_2) \text{ are independent} \Rightarrow \frac{V_1/f_1}{V_2/f_2} \sim F(f_1,f_2)$$

Tables showing areas under the density of F can also be found in elementary textbooks, but these are more comprehensive and seldom show areas for all values of f_1, f_2. Sometimes one can use the fact $F(f_1, f_2) = 1/F(f_2, f_1)$.

Order statistics

A random sample of n independent observations $(Y_i)_{i=1}^n$ can be arranged in increasing order, from the smallest to the largest $Y_{(1)} < Y_{(2)} < ... < Y_{(n)}$. Here only the distribution of the smallest and largest observations $Y_{(1)}$ and $Y_{(n)}$ are considered. We also restrict ourselves to the case with continuous variables. The distributional properties are summarized in the following theorem:

$Y_{(1)}$ has cdf $F_{Y_{(1)}}(y) = 1 - \prod_{i=1}^{n}\left(1 - F_{Y_i}(y)\right) = \left[\text{If all } Y_i \sim Y\right] = 1 - \left(1 - F_Y(y)\right)^n.$

In the latter case $f_{Y_{(1)}}(y) = nf_Y(y)\left(1 - F_Y(y)\right)^{n-1}.$

$Y_{(n)}$ has cdf $F_{Y_{(n)}}(y) = \prod_{i=1}^{n} F_{Y_i}(y) = \left[\text{If all } Y_i \sim Y\right] = \left(F_Y(y)\right)^n$ (8)

In the latter case $f_{Y_{(n)}}(y) = nf_Y(y)\left(F_Y(y)\right)^{n-1}.$

EX 19 Determine the cdf and density of $Y_{(1)}$ if all $Y_i \sim Y \sim Exponential(\lambda)$.

$F_Y(y) = 1 - e^{-\lambda y} \Rightarrow F_{Y_{(1)}}(y) = 1 - \left(e^{-\lambda y}\right)^n = 1 - e^{-n\lambda \cdot y}$, $f_{Y_{(1)}}(y) = n\lambda e^{-n\lambda \cdot y}$. Thus, the smallest of n observations is $Exponential(n\lambda)$, so the expected value of $Y_{(1)}$ is $\frac{1}{n\lambda}$

EX 20 Determine the cdf and density of $Y_{(n)}$ if all $Y_i \sim Y \sim Uniform[0, b]$.

$F_Y(y) = \frac{y}{b},\ 0 \le y \le b \Rightarrow F_{Y_{(n)}}(y) = \frac{y^n}{b^n},\ f_{Y_{(n)}} = \frac{ny^{n-1}}{b^n},\ 0 \le y \le b.$

$E(Y_{(n)}) = \int_0^b y \cdot \frac{ny^{n-1}}{b^n} dy = \frac{n}{b^n}\int_0^b y^n dy = \frac{n}{b^n} \cdot \frac{b^{n+1}}{(n+1)} = \frac{n}{(n+1)} \cdot b.$

3.2 Sample moments

In Ch. 2.1 we introduce the population moments $\alpha_r = E(Y^r)$ and the population central moments $\mu_r = E\left((Y - \mu)^2\right)$. By means of the Binomial series in Ch. 2.3.5 we can express μ_r in terms of α_r in the following way. $\mu_r = E\left((Y - \mu)^r\right) = E\left(\sum_{i=0}^{r}\binom{r}{i} Y^i (-\mu)^{r-i}\right) = \sum_{i=0}^{r}\binom{r}{i} E(Y^i)(-\mu)^{r-i} = \sum_{i=0}^{r}\binom{r}{i}\alpha_i (-\mu)^{r-i}$. From this we get e.g. $\mu_2 = \alpha_0\mu^2 - 2\alpha_1\mu + \alpha_2\mu^0 = \alpha_2 - \alpha_1^2$.

The corresponding sample moments are $a_r = \frac{1}{n}\sum_{i=1}^{n} Y_i^r$ with $a_1 = \bar{Y}$ and $m_r = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^r$. Instead of studying the properties of m_r in general we confine ourselves to $S^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

The following theorem gives some properties of sample moments.

If $(Y_i)_{i=1}^n$ are iid variables with mean μ and variance σ^2, then

$$E(a_r) = \alpha_r \ , V(a_r) = (\alpha_{2r} - \alpha_r^2)\frac{1}{n} \tag{9a}$$

$$E(S^2) = \sigma^2 \ , V(S^2) = \left(\mu_4 - \frac{(n-3)}{(n-1)}\mu_2^2\right)\frac{1}{n} \tag{9b}$$

The expressions for $V(S^2)$ above is proved in a book by C.R. Rao 1965, p.368. Proofs of the other relations are left as exercises for the reader in EX 21 below.

EX 21 $E(a_r)=\frac{1}{n}E\left(\sum Y_i^r\right)=\frac{1}{n}\sum E(Y_i^r)=\frac{1}{n}n\cdot\alpha_r=\alpha_r$ •

$E(a_r^2)=\frac{1}{n^2}E\left((\sum Y_i^r)^2\right)=[\text{Cf. Ch. 2.3.2}]=\frac{1}{n^2}E\left(\sum (Y_i^r)^2+2\sum\sum Y_i^r Y_j^r\right)=$

$\frac{1}{n^2}\left(\sum E(Y_i^{2r})+2\sum\sum E(Y_i^r)E(Y_j^r)\right)=\frac{1}{n^2}\left(n\cdot\alpha_{2r}+2\frac{(n^2-n)}{2}\cdot\alpha_r\alpha_r\right)=\frac{\alpha_{2r}}{n}+\alpha_r^2-\frac{\alpha_r^2}{n}$ •

So, $V(a_r)=E(a_r^2)-E^2(a_r)=(\alpha_{2r}-\alpha_r^2)\frac{1}{n}$.

$E\left(\sum(Y_i-\bar{Y})^2\right)=[\text{Cf. EX 13}]=E\left(\sum Y_i^2-(\sum Y_i)^2/n\right)=\sum E(Y_i^2)-\frac{1}{n}E\left((\sum Y_i)^2\right)=$

$[\text{Cf. expression above}]=n\cdot\alpha_2-\frac{1}{n}\left(n\cdot\alpha_2+(n^2-n)\cdot\alpha_1\alpha_1\right)=(n-1)\cdot(\alpha_2-\alpha_1^2)$.

EX 22 Let $(Y_i)_{i=1}^n$ be iid and distributed $Exponential(\lambda)$. Determine $V(S^2)$.

$\mu_2=V(Y)=\frac{1}{\lambda^2}$ and from Ch. 2.2.2 *(2)* $\alpha_r=\frac{r!}{\lambda^r}$ with $\alpha_1=\mu=\frac{1}{\lambda}$

$\mu_4=\sum_{i=0}^{4}\binom{4}{i}\alpha_i(-\mu)^{4-i}=\alpha_0\mu^4-4\alpha_1\mu^3+6\alpha_2\mu^2-4\alpha_3\mu+\alpha_4\cdot 1=$

$\frac{1}{\lambda^4}-4\frac{1}{\lambda}\frac{1}{\lambda^3}+6\frac{2}{\lambda^2}\frac{1}{\lambda^2}-4\frac{6}{\lambda^3}\frac{1}{\lambda}+\frac{24}{\lambda^4}=\frac{9}{\lambda^4}$.

Thus, from (9b) $V(S^2)=\left(\frac{9}{\lambda^4}-\frac{(n-3)}{(n-1)}\frac{1}{\lambda^4}\right)\frac{1}{n}=\frac{(8n-6)}{n(n-1)}\frac{1}{\lambda^4}$

3.3 Asymptotic and approximate results in sampling theory

Sometimes it is not possible, or very hard, to find the exact distribution of a statistic T_n based on n observations. In such a case one may try to find the asymptotic distribution when n is large. If also this is a stumbling block one can try to find at least approximate expressions for expectations and variances of T_n. In this section we present some ways to handle these problems.

3.3.1 Convergence in probability and in distribution

By *convergence of T_n in probability towards a constant c when* $n\to\infty$ we mean that the probability for the event that the distance between T_n and c is positive, tends to zero with increasing n. In symbols this is expressed by $T_n\xrightarrow{P}c$, as $n\to\infty$.In practice it is often cumbersome to verify if the latter probability tends to zero. Then one may use the following theorem.

$$E(T_n)=c \text{ and } V(T_n)\to 0\Rightarrow T_n\xrightarrow{P}c \tag{10}$$

By *convergence in distribution (or in law)* we mean that the cdf of T_n tends to the cdf of T, say.
In symbols we express this by $T_n\xrightarrow{D}T$. An example is the *CLT* given in (6).

Some important results

Let g be a continuous function, then the following relations hold. (For proofs the reader is referred to Ch 2c.4 and Ch 6a.2 in Rao 1965.)

$$T_n \xrightarrow{P} c \Rightarrow g(T_n) \xrightarrow{P} g(c)$$
$$T_n \xrightarrow{D} T \Rightarrow g(T_n) \xrightarrow{D} g(T) \tag{11}$$

$$T_n \xrightarrow{D} T \text{ and } U_n \xrightarrow{P} c \Rightarrow \begin{cases} T_n \pm U_n \xrightarrow{D} T \pm c \\ T_n \cdot U_n \xrightarrow{D} T \cdot c \\ T_n / U_n \xrightarrow{D} T / c \end{cases} \tag{12}$$

Let θ be a parameter and let the variance of T_n be $\sigma^2(\theta)$, a function of θ. Then

$$\sqrt{n}(T_n - \theta) \xrightarrow{D} Y \sim N(0, \sigma^2(\theta)) \Rightarrow \sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{D} X \sim N\left(0, [g'(\theta)]^2 \sigma^2(\theta)\right) \tag{13}$$

We now consider applications of (10)-(13)

EX 23 Let $(Y_i)_{i=1}^n$ be iid with $Y_i \sim Bernoulli(p)$. Put $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ =Relative frequency of 'success' after *n* trials.

a) Show that $\hat{p} \xrightarrow{P} p, \text{as } n \to \infty$.

This follows from (10) since [cf. 2.2.1 *(1)*] $E(\hat{p}) = p \text{ and } V(\hat{p}) = \frac{p(1-p)}{n} \to 0$ as $n \to \infty$.

The fact that $\hat{p} \xrightarrow{P} p$ has been termed *law of large numbers*. It can be empirically verified by throwing a thumbtack a number of times and noticing the relative frequency of the event 'tip of the tack is up'. The author, with his particular type of thumbtack found that the frequency stabilized around *p* = 0.6 after about 20 trials. The outcome is of course depending on the experimental conditions, but the reader is encouraged to repeat it, with a shoe or a coin. It is instructive to plot the relative frequency of the event on the Y-axis against the number of trials on the X-axis.

b) Show that $\frac{(\hat{p}-p)}{\sqrt{p(1-p)/n}} \xrightarrow{D} Z \sim N(0,1)$, as $n \to \infty$.

The left hand side is, after multiplication with *n* in both numerator and denominator,

$\frac{\sum_{i=1}^{n} Y_i - np}{\sqrt{np(1-p)}}$. The CLT in (6) now gives the result.

c) Show that $\frac{(\hat{p}-p)}{\sqrt{p(1-p)/n}} \xrightarrow{D} Z \sim N(0,1), \text{as } n \to \infty$.

The left hand side can be written

$$\frac{\frac{(\hat{p}_n - p)}{\sqrt{p(1-p)/n}} \xrightarrow{D} Z \sim N(0,1)}{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)/n}{p(1-p)/n}} \xrightarrow{P} 1}$$

The convergence in the numerator was shown in b). To prove the convergence in the denominator, notice that $\hat{p}_n \xrightarrow{P} p \Rightarrow g(\hat{p}_n) = \hat{p}_n(1-\hat{p}_n) \xrightarrow{P} g(p) = p(1-p)$ Finally, the result follows from (12).

Comment: The difference between the expressions in b) and c) is that in c) we have replaced *p* by an estimator $\hat{p}_n$ in the denominator. This will simplify calculations of confidence intervals (cf. Ch. 5). However, *n* in c) needs to be much larger than in b), for the approximation to normality to hold. If *p* is not too far from 0.5, then *n* about 50 is sufficient for normality in b), while *n* perhaps larger than 5000 may be required in c).

d) Show that $\frac{\sqrt{n}(\ln \hat{p} - \ln p)}{\sqrt{(1-p)/p}} \xrightarrow{D} Z \sim N(0,1)$

Multiplying the left hand side in b) by $\sqrt{p(1-p)}$ and using (10) gives $\sqrt{n}(\hat{p}_n - p) \xrightarrow{D}$

$\sqrt{p(1-p)} \cdot Z \sim N(0, p(1-p))$. Since $\ln x$ is continuous with derivative $1/x$ it follows from

(13) that $\sqrt{n}(\ln \hat{p}_n - \ln p) \xrightarrow{D} N\left(0, \left[\frac{d \ln p}{dp}\right]^2 \cdot p(1-p)\right)$ from which d) follows.

EX 24 Let $(Y_i)_{i=1}^n$ be iid variables with $E(Y_i) = \mu$ and $V(Y_i) = \sigma^2$. Show that

$$\frac{n(\overline{Y} - \mu)^2}{\sigma^2} \xrightarrow{D} \chi^2(1)$$

This follows because $\frac{\sqrt{n}(\overline{Y} - \mu))}{\sigma} \xrightarrow{D} Z \sim N(0,1)$ according to the *CLT* in (6). From (11)

$$\frac{n(\overline{Y} - \mu)^2}{\sigma^2} \xrightarrow{D} Z^2 = \chi^2(1)$$

EX 25 Let $(Y_i)_{i=1}^n$ be iid variables with $E(Y_i) = \mu$ and $V(Y_i) = \sigma^2$. Show that

$$\frac{(\overline{Y} - \mu)}{S/\sqrt{n}} \xrightarrow{D} Z \sim N(0,1) \text{ as } n \to \infty$$

Dividing numerator and denominator by $\sigma/\sqrt{n}$ yields

$$\frac{\frac{(\overline{Y} - \mu)}{\sigma/\sqrt{n}} \xrightarrow{D} Z \sim N(0,1)}{\frac{S/\sqrt{n}}{\sigma/\sqrt{n}} = \frac{S}{\sigma} \xrightarrow{P} 1}$$

, and the result follows from (12).

Here $\frac{S}{\sigma} \xrightarrow{P} 1$ for two reasons: (i) $V(S^2) \to 0 \left[\text{Cf. (9b)}\right] \Rightarrow S^2 \xrightarrow{P} \sigma^2$.

(ii) $g(S^2) = \sqrt{S^2/\sigma^2} \xrightarrow{P} \sqrt{\sigma^2/\sigma^2} = 1$ [Cf. (11)].

3.3.2 Approximations of moments

Let $Y_i, i = 1,2$ be two random variables with means μ_i and variances σ_i^2 and with covariance σ_{12}. From Taylor expansions of a function *g* of the variables, one can show the following. (Cf. Casella & Berger 1990, pp. 328–331.)

$$E(g(Y_i)) \approx g(\mu_i) + \frac{1}{2} g''(\mu_i) \cdot \sigma_i^2,\ V(g(Y_i)) \approx [g'(\mu_i)]^2 \cdot \sigma_i^2$$

$$Cov(g_1(Y_1), g_2(Y_2)) \approx g'_1(\mu_1) g'_2(\mu_2) \cdot \sigma_{12} \quad (14)$$

EX 26

a) Let $Y \sim Gamma(\lambda, k)$. Determine the approximate mean and variance of ln*Y*.

From 2.2.2 *(2)* we know that the mean and variance is $\mu = k/\lambda$ and $\sigma^2 = k/\lambda^2$, respectively. The derivatives of $g(y) = \ln y$ are $g'(y) = 1/y$ and $g''(y) = -1/y^2$. Thus (14) gives

$$E(\ln Y) \approx \ln \mu + \frac{1}{2} \cdot (-1/\mu^2) \cdot \sigma^2 = \ln(k/\lambda) - 1/2k, \; V(\ln Y) \approx (-1/\mu)^2 \cdot \sigma^2 = 1/k.$$

b) $(Y_i)_{i=1}^n$ is a sequence of iid variables, each being distributed $Gamma(\lambda, k)$. Determine the approximate mean and variance of $\ln \overline{Y}$, where as usual $\overline{Y} = \sum_{i=1}^{n} Y_i / n$.

From 3.1 *(6)* we know that $\sum_{i=1}^{n} Y_i \sim G = Gamma(\lambda, nk)$. Now, $\ln \overline{Y} = \ln(G/n) = \ln G - \ln n$. From a) above we get $E(\ln G) \approx \ln(nk/\lambda) \overset{i=1}{1}/2nk$ and $V(\ln G) \approx 1/nk$. Thus,

$$E(\ln \overline{Y}) \approx \ln(nk/\lambda) - 1/2nk - \ln n = \ln(k/\lambda) + \ln n - 1/2nk - \ln n = \ln(k/\lambda) - 1/2nk$$

$$V(\ln \overline{Y}) \approx V(\ln G) + V(\ln n) = V(\ln G) + 0 = 1/nk$$

Notice that, as $n \to \infty$, $E(\ln \overline{Y}) \to \ln E(Y_i)$ and $V(\ln \overline{Y}) \to 0$.

A generalization of (14) to a function of two variables is

$$E\big(g(Y_1,Y_2)\big)\approx g(\mu_1,\mu_2)+\frac{1}{2}\left(\frac{d^2g(\mu_1,\mu_2)}{dy_1^2}\cdot\sigma_1^2+\frac{d^2g(\mu_1,\mu_2)}{dy_2^2}\cdot\sigma_2^2+2\frac{d^2g(\mu_1,\mu_2)}{dy_1dy_2}\cdot\sigma_{12}\right)$$
$$V\big(g(Y_1,Y_2)\big)\approx\left(\frac{dg(\mu_1,\mu_2)}{dy_1}\right)^2\cdot\sigma_1^2+\left(\frac{dg(\mu_1,\mu_2)}{dy_2}\right)^2\cdot\sigma_2^2+2\left(\frac{dg(\mu_1,\mu_2)}{dy_1}\right)\left(\frac{dg(\mu_1,\mu_2)}{dy_2}\right)\cdot\sigma_{12} \quad (15)$$

EX 27

Let $Y_i, i=1,2$ be correlated variables with means μ_i and variances σ_i^2 and with covariance σ_{12}. Derive the approximate mean and variance of $R=Y_1/Y_2$.

We start by computing the derivatives of the function $g=y_1/y_2$ (cf. derivation rule *(4)* in Ch. 2.3.3).

$\frac{dg}{dy_1}=\frac{1}{y_2},\frac{d^2g}{dy_1^2}=0,\frac{dg}{dy_2}=-\frac{y_1}{y_2^2},\frac{d^2g}{dy_2^2}=\frac{2y_1}{y_2^3},\frac{d^2g}{dy_1dy_2}=-\frac{1}{y_2^2}$. Thus,

$$E\left(\frac{Y_1}{Y_2}\right)\approx\frac{\mu_1}{\mu_2}+\frac{1}{2}\left[0+\frac{2\mu_1\sigma_2^2}{\mu_2^3}+2\left(-\frac{1}{\mu_2^2}\right)\sigma_{12}\right]=\frac{\mu_1}{\mu_2}\left[1+\frac{\sigma_2^2}{\mu_2^2}-\frac{\sigma_{12}}{\mu_1\mu_2}\right]$$
$$V\left(\frac{Y_1}{Y_2}\right)\approx\frac{\sigma_1^2}{\mu_2^2}+\left(-\frac{\mu_1\sigma_2^2}{\mu_2^2}\right)+\frac{2}{\mu_2}\left(-\frac{\mu_1}{\mu_2}\right)\sigma_{12}=\left(\frac{\mu_1}{\mu_2}\right)^2\left[\frac{\sigma_1^2}{\mu_1^2}+\frac{\sigma_2^2}{\mu_2^2}-2\frac{\sigma_{12}}{\mu_1\mu_2}\right]$$

3.4 Final words

Uppercase letters or lowercase letters? Uppercase letters, such as S^2 for a sample variance, are used for statistics when we want to stress that the quantity has a distribution. Lowercase letters, such as s^2, are used for specific values of a statistic.

The distribution of a statistic is called a *sampling distribution*. This is a creation by statisticians for the purpose of drawing conclusions about parameters in the population and it has nothing to do with the real world. Distributions that are intended to reflect facts in nature or society are called *population distributions*.

Asymptotic results are obtained as a limit, e.g. when $n\to\infty$ and $p\to 0$ in the Poisson approximation of the Binomial distribution. *Approximate* results just mean that they are not exact.

Knowledge about sampling distributions is the key for understanding the content in the following chapters. It's therefore important that you are comfortable with the properties in *(1)-(10)*, and also of Cochran's theorem.

We have assumed that there is a given a random sample. This can be achieved in a verity of ways. In this book we don't bother how the sample has been collected. For readers interested in these matters there is a hugh amount of literature in the field. (See e.g. Scheaffer, *et al*, 2012).

Supplementary Exercises, Ch. 3

EX 28 Let $(Y_i)_{i=1}^n$ be iid variables.

Find cdf and density of the smallest observation $Y_{(1)}$ if $Y_i \sim Uniform[0,b]$.

EX 29 Let $(Y_i)_{i=1}^n$ be iid variables.

Find cdf and density of the largest observation $Y_{(n)}$ if $Y_i \sim Exponential(\lambda)$.

EX 30 Let $Y \sim Binomial(n,p)$ and put $\hat{p} = Y/n$ so that $E(\hat{p}) = p$ and $V(\hat{p}) = p(1-p)/n$. As an estimator of $V(\hat{p})$ one may use $\hat{V}(\hat{p}) = \hat{p}(1-\hat{p})/n$. Show that the exact mean $E\left(\hat{V}(\hat{p})\right)$ and the approximate mean obtained from (14) are identical.

EX 31 In medical statistics one often wants to study whether a factor F causes a disease. Data from two independent samples of sizes n_1 and n_2 can be summarized in the following frequency table:

	Diseased	Not-Diseased	Total
F is present	Y_1	$n_1 - Y_1$	n_1
F is absent	Y_2	$n_2 - Y_2$	n_2

Data are analyzed by comparing the *Relative Risk* $\hat{R} = \hat{p}_1 / \hat{p}_2$, where $\hat{p}_i = Y_i / n_i, i = 1,2$ with the hypothetical value of 1, being obtained if F does not cause the disease. The variance of $\hat{R}$ is estimated by

$$\hat{V}(\hat{R}) = \hat{R}^2 \left[\frac{n_1 - Y_1}{n_1 Y_1} + \frac{n_2 - Y_2}{n_2 Y_2} \right]$$

Justify this expression by using the result in EX 27.

[Hint: Use the fact that Y_1 and Y_2 can be treated as two independent variables that are $\sim Binomial(n_i, p_i), i = 1,2.$]

EX 32 The sample variance S^2 is in general unbiased for σ^2 (cf. (9b)). However, *S* is not in general unbiased for σ. Determine approximate expressions for $E(S)$ and $V(S)$ in the following cases:

a) $(Y_i)_{i=1}^n$ are iid with expectation μ and variance σ^2 with a general distribution for Y_i.
b) -"- with $Y_i \sim N(\mu, \sigma^2)$.

4 Point estimation

In this chapter we deal with the problem of how to estimate an unknown characteristic in the population based on a sample of n observations. Focus will be on the estimation of parameters, such as the variance σ^2 in a normal distribution or the upper point b in a Uniform distribution. We briefly also consider the estimation of functions of parameters and other quantities such as probability and cdf. First some concepts are introduced and then we discuss some requirements on good estimators. Finally some estimation methods are presented and evaluated.

4.1 Concepts

A *statistic* T is a function of the random variables $Y_1,...,Y_n$ in a sample. A *point estimator* is a statistic that is used to estimate the value of an unknown parameter in the population, in general denoted θ. A *point estimate* t is a numerical value of T, obtained in a specific sample.

In (1a) and (1b) we introduced the concept of *probability of a random sample* of independent observations. This is a function of the variable values $y_1,...,y_n$. If we instead consider it as a function of the parameter θ, it is termed *Likelihood* $L(\theta) = L(y_1,...,y_n,\theta)$. When we want to study the long-run behavior of the likelihood over all possible drawn samples, we use the notation $L(Y_1,...,Y_n,\theta)$. In the latter case L is a random variable.

Intuitively the observations in a sample contain information about θ in some sense. (The statistical concept of information will be defined formally below.) E.g. given the body weights in kg, 75, 50, 90, 72, 78 of five persons drawn from a certain population, we conclude that the population mean should be slightly larger than 70, but also that the dispersion is quite large. Sometimes all information about θ is contained in a single statistic T. In such a case T is termed a *sufficient statistic for* θ. If we have found a sufficient statistic T for θ we can, roughly speaking, skip the original observations and only use T for making inference about θ. The following factorization criterion can be used to find a sufficient statistic:

Assume that the likelihood L can be factorized into two parts such that

$$L(Y_1,...,Y_n,\theta) = L_1(T,\theta)\cdot L_2(Y_1,...,Y_n), \tag{16}$$

where L_1 only depends on T and θ and L_2 and does not depend on T and θ but possibly only on the observations, then T is sufficient for θ.

More generally, $T_1,...,T_p$ are simultaneous sufficient statistics for $\theta_1,...,\theta_p$ if (16) holds with T and θ being replaced by the corresponding vectors. The following results can be useful:

Let g be a continuous function. Then: T is sufficient for $\theta \Rightarrow g(T)$ is sufficient for $g(\theta)$ (17)

A sufficient statistic is unique. (There can't be several sufficient statistics for a parameter besides functions of T.) (18)

EX 33 $(Y_i)_{i=1}^n$ are independent variables in a random sample. Find sufficient statistics in the following cases: (See Ch. 2.2 to find the various distributions.)

a) $Y_i \sim Binomial(m_i, p)$

$L = \prod_{i=1}^n \binom{m_i}{y_i} p^{y_i}(1-p)^{m_i-y_i} = p^{\sum y_i}(1-p)^{\sum m_i - \sum y_i} \cdot \prod_{i=1}^n \binom{m_i}{y_i}$. From this we conclude that the statistic $\sum_{i=1}^n Y_i$ is sufficient for *p*.

b) $Y_i \sim Geometric(p)$

$L = \prod_{i=1}^n p^{y_i-1}(1-p) = p^{\sum_{i=1}^n y_i - n}(1-p)^n \cdot 1$. Thus, $\sum_{i=1}^n Y_i$ is sufficient for *p*.

c) $Y_i \sim Poisson(\lambda)$

$L = \prod_{i=1}^n \frac{\lambda^{y_i}}{y_i!} e^{-\lambda} = \lambda^{\sum_{i=1}^n y_i} e^{-n\lambda} \cdot \frac{1}{\prod_{i=1}^n y_i!}$. Thus, $\sum_{i=1}^n Y_i$ is sufficient for λ.

d) Y_i has the density $f(y) = 2\lambda y e^{-\lambda y^2}$. This is called the *Weibull* distribution and is used as a model for life lengths of materials.

$L = \prod_{i=1}^n 2\lambda y_i e^{-\lambda y_i^2} = \lambda^n e^{-\lambda y_i^2} \cdot 2^n \prod_{i=1}^n y_i$. Thus, $\sum_{i=1}^n Y_i^2$ is sufficient for λ.

e) $Y_i \sim Gamma(\lambda, k)$

$L = \prod_{i=1}^n \frac{\lambda^k}{\Gamma(k)} y_i^{k-1} e^{-\lambda y_i} = \frac{\lambda^{nk}}{\Gamma^n(k)} \left(\prod_{i=1}^n y_i\right)^{k-1} e^{-\lambda \sum_{i=1}^n y_i} \cdot 1$. Thus, $\left(\sum_{i=1}^n Y_i, \prod_{i=1}^n Y_i\right)$ are simultaneous sufficient for (λ, k).

f) $Y_i \sim N(\mu, \sigma^2)$

$L = \prod_{i=1}^n \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} e^{-\frac{\sum_{i=1}^n (y_i-\mu)^2}{2\sigma^2}}$. Here $\sum_{i=1}^n (y_i-\mu)^2 =$

$\sum_{i=1}^n (y_i-\mu)^2 = \sum_{i=1}^n \left((y_i-\bar{y})+(\bar{y}-\mu)\right)^2 = \sum_{i=1}^n (y_i-\bar{y})^2 + \sum_{i=1}^n (\bar{y}-\mu)^2 + \sum_{i=1}^n 2(y_i-\bar{y})(\bar{y}-\mu) =$

$\sum_{i=1}^n (y_i-\bar{y})^2 + n(\bar{y}-\mu)^2$, since the last term is zero (cf. EX 13). From this it follows that

$\left(\sum_{i=1}^n Y_i, \sum_{i=1}^n (Y_i-\bar{Y})^2\right)$ are simultaneous sufficient for (μ, σ^2).

4.2 Requirements on estimators

In order for an estimator T_n of θ, based on n observations, to be considered as good one usually requires the following:

- T_n is *consistent* for θ. This means that the estimator converges in probability towards the parameter, $T_n \xrightarrow{P} \theta$, as $n \to \infty$. Remember from (10), Ch. 3.3, that a sufficient condition for this is that $E(T_n) = \theta$ and $V(T_n) \to 0$. Estimators that are not consistent are useless in the sense that we do not necessarily get closer to θ by increasing n.
- T_n is *unbiased* for θ which means that $E(T_n) = \theta$. The difference $E(T_n) - \theta$ is the *bias* of the estimator, denoted $bias(\theta)$. The dispersion of T_n around θ can be measured by the *Mean Squared Error (MSE) of* T_n, $E\left[(T_n - \theta)^2\right] = V(T_n) + \left(bias(\theta)\right)^2$.
- T_n is a *minimum variance estimator (MVE)* which means that $V(T_n)$ is smaller than the variance of all other estimators. A MVE is unique, so there can only be one estimator with smallest variance.

The problem of finding a MVE is rather complicated. Before treating this we consider some results about derivatives of the log-likelihood function. The function $\frac{d \ln L}{d\theta}$ is called a *score function* and it plays an important role in statistical inference. From Eq. (1a) and Eq. (1b) it follows that

$$\frac{d \ln L}{d\theta} = \sum_{i=1}^{n} \frac{d \ln p(y_i, \theta)}{d\theta} \text{ (discrete case) and } = \sum_{i=1}^{n} \frac{d \ln f(y_i, \theta)}{d\theta} \text{ (continuous case)} \tag{19}$$

In order to obtain further results in the continuous case we set up the following conditions:

a) The range of *y*- values in $f(y, \theta)$ does not depend on θ. (20)

b) $\frac{d \ln f}{d\theta}$ and $\frac{d^2 \ln f}{d\theta^2}$ are continuous.

Notice that (20) does not hold for $Y \sim Uniform[0, b]$ with density $f(y, b) = 1/b, 0 \le y \le b$, but for all other densities we have considered so far.

If the conditions (a) and (b) in (20) holds then

a) $E\left(\frac{d \ln L}{d\theta}\right) = 0$ (21)

b) $I(\theta) = V\left(\frac{d \ln L}{d\theta}\right) = -E\left(\frac{d^2 \ln L}{d\theta^2}\right)$

The function $I(\theta)$ is the *information about θ that is contained in a sample of size n*. A solution to the problem of finding a MVE is given by the following theorem, called the *Information inequality* or the *Cramer-Rao inequality* after two of its discoverers who published the result in 1945.

$$V(T_n) \geq \frac{\left(1+\frac{dbias(\theta)}{d\theta}\right)^2}{I(\theta)} = (\text{If } T_n \text{ is unbiased}) = \frac{1}{I(\theta)} \tag{22}$$

The lower limit in (22) for the variance is called the *Cramer-Rao (C-R) limit*. The limit may not be attainable for a MVE, but no estimator can have smaller variance. Thus, if we have found an estimator with a variance that equals the C-R limit, then we have found a MVE. But, if the variance of an estimator is larger than the C-R limit, the estimator may still be a MVE. The search for a MVE in the latter case can be complicated. Some help may be obtained from a theorem of Rao and Blackwell (Casella & Berger 1990, p. 316), but this is beyond the level of this book.

EX 34 Let $(Y_i)_{i=1}^n$ be iid variables where $Y_i \sim Exponential(\lambda)$

a) Find an unbiased estimator of $E(Y_i) = 1/\lambda$ that is based on the smallest observation $Y_{(1)}$. Is this estimator consistent?

$$F_{Y_i}(y) = 1 - e^{-\lambda \cdot y} \Rightarrow [\text{See (8)}] \Rightarrow F_{Y_{(1)}}(y) = 1 - \left[1 - \left(1 - e^{-\lambda \cdot y}\right)\right]^n = 1 - e^{-n\lambda \cdot y} \Rightarrow$$

$Y_{(1)} \sim Exponential(n\lambda) \Rightarrow E(Y_{(1)}) = 1/n\lambda$. Thus, $T_n = nY_{(1)}$ is unbiased for $1/\lambda$.

b) The variance of this estimator is $V(T_n) = n^2 V(Y_{(1)}) = n^2 \cdot \dfrac{1}{(n\lambda)^2} = 1/\lambda^2$ which does not tend to 0 as $n \to \infty$, so the estimator is not consistent.

c) Find a sufficient statistic for λ.

$L = \prod_{i=1}^n \lambda e^{-\lambda \cdot y_i} = \lambda^n e^{-\lambda \sum_{i=1}^n y_i} \cdot 1 \Rightarrow \sum_{i=1}^n Y_i$ is sufficient for λ.

d) Determine the information about λ that is contained in a sample of *n* observations and also the Cramer-Rao (C-R) limit.

$$\ln L = \ln(\lambda^n) + \ln\left(e^{-\lambda \sum_{i=1}^n y_i}\right) = n \ln \lambda - \lambda \cdot \sum_{i=1}^n y_i \Rightarrow \frac{d \ln L}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^n y_i \Rightarrow \frac{d^2 \ln L}{d\lambda^2} = -\frac{n}{\lambda^2}$$. Thus,

considering the latter as a random unit we get $I(\lambda) = -\left(-\dfrac{n}{\lambda^2}\right) = \dfrac{n}{\lambda^2}$. The information increases with *n* and decreases with increasing λ.

The C-R limit for any unbiased estimator of λ is $\dfrac{1}{I(\lambda)} = \dfrac{\lambda^2}{n}$.

4.3 Estimation methods

In this section we present some general methods to obtain estimators. Focus will be on the case with a single parameter, but examples in the multi-parameter case are also given. It is required that the estimators are unbiased. In this case the precision of an estimator can be measured by its variance. When comparing estimators it is useful to use the concept of *relative efficiency* of an estimator T_1 relative to another T_2, $RE = V(T_1)/V(T_2)$. Examples of *RE*s for estimators produced by various methods are given in the Supplementary Exercises of this chapter.

4.3.1 Method of Ordinary Least Squares (OLS)

The method originates from the works by the mathematicians Legendre (1805) and Gauss (1809). Many students are familiar with this method as a way to fit a straight line to data points in the plane, but the method can be used in more general contexts.

Given the variables Y_i with expectations $E(Y_i) = g_i(\theta)$, $i = 1 \ldots n$, consider the sum of squares $SS = \sum_{i=1}^{n} [Y_i - g_i(\theta)]^2$ and determine the value of θ that minimizes *SS*, say $\hat{\theta}_{OLS}$. This can be obtained from the solution of $\frac{dSS}{d\theta} = 0$. By putting $\hat{\theta}_{OLS}$ into *SS* one gets an estimated sum of squares $SSE = \sum_{i=1}^{n} \left[Y_i - g_i(\hat{\theta}_{OLS})\right]^2$ which can be used for estimating dispersion.

EX 35 $(Y_i)_{i=1}^{n}$ are iid with $Y_i \sim Poisson(\lambda)$. Find the OLS estimator $\hat{\lambda}_{OLS}$.

$$SS = \sum_{i=1}^{n} [Y_i - \lambda]^2 \Rightarrow \frac{dSS}{d\lambda} = \sum_{i=1}^{n} (-1) \cdot 2[Y_i - \lambda] = -2\sum_{i=1}^{n} [Y_i - \lambda] = 0 \Rightarrow \sum_{i=1}^{n} Y_i - n\lambda = 0 \Rightarrow$$

$$\hat{\lambda}_{OLS} = \sum_{i=1}^{n} Y_i / n = \overline{Y} \cdot$$

Here we notice that $E(\hat{\lambda}_{OLS}) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n} n \cdot \lambda = \lambda$ (Unbiased.) and $V(\hat{\theta}_{OLS}) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) =$

$$\frac{1}{n^2} n \cdot \lambda = \frac{\lambda}{n}.$$

EX 36 $(Y_i)_{i=1}^{n}$ where Y_i are independent with $E(Y_i) = \beta x_i$ and $V(Y_i) = \sigma^2$. This model is often called 'Linear Regression trough the Origin with constant variance'. Here each x_i is fixed while Y_i is random. Find the OLS estimator of β.

$$SS = \sum_{i=1}^{n} [Y_i - \beta x_i]^2 \Rightarrow \frac{dSS}{d\beta} = 2\sum_{i=1}^{n} (-x_i)[Y_i - \beta x_i] = 0 \Rightarrow \beta \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i Y_i \Rightarrow \hat{\beta}_{OLS} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}$$

$$E\left(\hat{\beta}_{OLS}\right) = \frac{1}{\sum_{i=1}^{n} x_i^2} \cdot \sum_{i=1}^{n} x_i E(Y_i) = \frac{1}{\sum_{i=1}^{n} x_i^2} \cdot \sum_{i=1}^{n} x_i \cdot \beta x_i = \beta \text{ (Unbiased.)}$$

$$V\left(\hat{\beta}_{OLS}\right) = \frac{1}{\left(\sum_{i=1}^{n} x_i^2\right)^2} \cdot \sum_{i=1}^{n} x_i^2 V(Y_i) = \frac{1}{\left(\sum_{i=1}^{n} x_i^2\right)^2} \cdot \sum_{i=1}^{n} x_i^2 \cdot \sigma^2 = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$. Notice that the variance is small, i.e. the precision of the estimator is high, if the x_i-values are large. In practice this means that if we e.g. want to estimate the relation between Y_i = Fuel consumption and x_i = Speed, we should measure Fuel consumption when Speed is high.

When $E(Y_i) = g_i(\theta_1, \ldots, \theta_p)$, a function of several parameters, we put $SS = \sum_{i=1}^{n} \left[Y_i - g_i(\theta_1, \ldots, \theta_p)\right]^2$. By solving the equations $\frac{dSS}{d\theta_1} = 0, \ldots, \frac{dSS}{d\theta_p} = 0$ we get the OLS estimators of the parameters.

4.3.2 Method of Moments

This method was suggested by the statistician Karl Pearson in the late 1800s. The approach is to equal the sample moments $\overline{Y}, S^2, \ldots$ to the corresponding moments in the population $E(Y) = g_1(\theta_1, \theta_2, \ldots)$, $V(Y) = g_2(\theta_1, \theta_2, \ldots), \ldots$ and solve for the parameters. The method has several deficiencies, but moment estimates can be used when more ingenious methods require initial values in order to get iterative solutions. An example of this is given in EX 44.

EX 37 $(Y_i)_{i=1}^n$ are iid and $Y_i \sim Poisson(\lambda)$. Here $E(Y_i) = \lambda = V(Y_i)$.

Obviously $\hat{\lambda}_{Mom} = \overline{Y}$. We might have used $\hat{\lambda}_{Mom} = S^2$, but this is less appropriate since S^2 has larger variance than $\overline{Y}$.

EX 38 $(Y_i)_{i=1}^n$ are iid and $Y_i \sim Gamma(\lambda, k)$. Here $E(Y) = k/\lambda$ and $V(Y) = k/\lambda^2$.

Put $\begin{cases} k/\lambda = \overline{Y} & (1) \\ \mathrm{k}/\lambda^2 = S^2 & (2) \end{cases}$ from which we get $k = \lambda\overline{Y} = \lambda^2 S^2 \Rightarrow \hat{\lambda}_{Mom} = \overline{Y}/S^2$. The latter inserted into (1) yields $\hat{k}_{Mom} = \overline{Y}^2/S^2$.

4.3.3 Method of Best Linear Unbiased Estimator (BLUE)

The method has an unclear origin but seems to have been in use since the early 1900s. The approach is simply to put $T_n = \sum_{i=1}^{n} a_i Y_i$ and determine the constants a_i such that T_n is unbiased and has minimum variance. This problem belongs to the field 'minimization under restrictions'. In the examples below we show how solutions can be obtained in a simple way by using Lagrange's multiplier λ.

EX 39 $(Y_i)_{i=1}^{n}$ are independent with $E(Y_i) = \theta$ and $V(Y_i) = V_i$. Find the BLUE of θ.

$$T_n = \sum_{i=1}^{n} a_i Y_i \Rightarrow E(T_n) = \sum_{i=1}^{n} a_i E(Y_i) = \theta \sum_{i=1}^{n} a_i = (\text{Put}) = \theta \Rightarrow \sum_{i=1}^{n} a_i = 1 \text{ or } \sum_{i=1}^{n} a_i - 1 = 0 \quad \text{(i)}$$

We now minimize $Q = V(T_n) + \lambda\left[\sum_{i=1}^{n} a_i - 1\right] = \sum_{i=1}^{n} a_i^2 V_i + \lambda\left[\sum_{i=1}^{n} a_i - 1\right]$ with respect to a_i.

$$\frac{dQ}{da_i} = 2a_i V_i + \lambda = 0 \Rightarrow a_i = \frac{-\lambda}{2V_i} = \frac{\lambda'}{V_i}, \text{ say} \quad \text{(ii)}$$

Putting this into (i) gives $\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} \frac{\lambda'}{V_i} = 1 \Rightarrow \lambda' = \frac{1}{\sum_{i=1}^{n} \frac{1}{V_i}}$, which inserted into (ii) gives

$$a_i = \frac{1}{V_i \sum_{i=1}^{n} \frac{1}{V_i}} = \frac{1/V_i}{\sum_{i=1}^{n} 1/V_i}. \text{ So, } \hat{\theta}_{BLUE} = \frac{\sum_{i=1}^{n} Y_i / V_i}{\sum_{i=1}^{n} 1/V_i}.$$

The variance is $V(\hat{\theta}_{BLUE}) = \frac{1}{\left(\sum_{i=1}^{n} 1/V_i\right)^2} \cdot \sum_{i=1}^{n} (1/V_i)^2 \cdot V_i = \frac{1}{\sum_{i=1}^{n} 1/V_i}$.

Notice that if all variances are equal, $V_i = V$, then $\hat{\theta}_{BLUE} = \overline{Y}$. Otherwise BLUE estimates can't be computed in practice without further assumptions about the variances.

EX 40 $(Y_i)_{i=1}^n$ are independent with $E(Y_i) = \beta x_i$ and $V(Y_i) = \sigma^2 x_i^p$ for some *p*.

This is the same situation as in EX 36 with the exception that $V(Y_i)$ is no longer constant, but changes with x_i. Find the BLUE of β.

$$T_n = \sum_{i=1}^n a_i Y_i \Rightarrow E(T_n) = \sum_{i=1}^n a_i E(Y_i) = \beta \sum_{i=1}^n a_i x_i = (\text{Put}) = \beta \Rightarrow \sum_{i=1}^n a_i x_i - 1 = 0 \quad \text{(i)}$$

$$V(T_n) = \sum_{i=1}^n a_i^2 V(Y_i) = \sum_{i=1}^n a_i^2 \sigma^2 x_i^p = \sigma^2 \sum_{i=1}^n a_i^2 x_i^p .$$

$$Q = V(T_n) + \lambda\left[\sum_{i=1}^n a_i x_i - 1\right] = \sigma^2 \sum_{i=1}^n a_i^2 x_i^p + \lambda\left[\sum_{i=1}^n a_i x_i - 1\right], \ \frac{dQ}{d\beta} = 2\sigma^2 a_i x_i^p + \lambda x_i = 0 \Rightarrow$$

$$a_i = -\frac{\lambda x_i^{1-p}}{2\sigma^2} = \lambda' x_i^{1-p} \quad \text{(ii)}$$

(ii) into (i) gives $\sum_{i=1}^n \lambda' x_i^{1-p} \cdot x_i = \lambda' \sum_{i=1}^n x_i^{2-p} = 1 \Rightarrow \lambda' = \frac{1}{\sum_{i=1}^n x_i^{2-p}}$, which inserted into (ii) gives

$$a_i = \frac{x_i^{1-p}}{\sum_{i=1}^n x_i^{2-p}}. \text{ BLUE for } \beta \text{ is thus } \hat{\beta}_{BLUE} = \frac{\sum_{i=1}^n x_i^{1-p} Y_i}{\sum_{i=1}^n x_i^{2-p}}.$$

$$V(\hat{\beta}_{BLUE}) = \frac{1}{\left(\sum_{i=1}^n x_i^{2-p}\right)^2} \sum_{i=1}^n \left(x_i^{1-p}\right)^2 V(Y_i) = \frac{1}{\left(\sum_{i=1}^n x_i^{2-p}\right)^2} \sum_{i=1}^n \left(x_i^{1-p}\right)^2 \sigma^2 x_i^p = \frac{\sigma^2}{\sum_{i=1}^n x_i^{2-p}}.$$

Special cases

$$p = 0 \text{ so } V(Y_i) = \sigma^2 : \hat{\beta}_{BLUE} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2} \text{ with } V(\hat{\beta}_{BLUE}) = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}.$$

$$p = 1, \text{so } V(Y_i) = \sigma^2 x_i : \hat{\beta}_{BLUE} = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i} = \frac{\overline{Y}}{\overline{x}} \text{ with } V(\hat{\beta}_{BLUE}) = \frac{\sigma^2}{\sum_{i=1}^n x_i}.$$

$$p = 2, \text{so } V(Y_i) = \sigma^2 x_i^2 : \hat{\beta}_{BLUE} = \frac{\sum Y_i / x_i}{n} \text{ with } V(\hat{\beta}_{BLUE}) = \frac{\sigma^2}{n}.$$

This illustrates that estimators of the same parameter can differ very much, depending on which assumptions are made about the data structure. In practice it is therefore important that such structures are investigated before the estimation is done.

4.3.4 Method of Maximum Likelihood (ML)

The method was developed by the English statistician R.A. Fisher in a series of papers published during the period 1912–1922. The idea is to determine the value of θ that maximizes the likelihood $L(\theta)$, thereby finding 'the most likely value of the parameter given the outcomes $y_1,...,y_n$'. If $L(\theta) > 0$, the likelihood has maximum for the same value as $\ln L(\theta)$ (cf. Ch. 2.3.3). Since the latter function is more convenient to deal with, the ML estimator $\hat{\theta}_{ML}$ can be found by solving the *likelihood equations*

$$\frac{d \ln L}{d\theta} = 0 \text{ for one parameter, or } \begin{cases} \frac{d \ln L}{d\theta_1} = 0 \\ \vdots \\ \frac{d \ln L}{d\theta_p} = 0 \end{cases} \text{ for many parameters.}$$

Some properties of ML estimators:

- The likelihood equations give the ML estimators if the conditions in (20) Ch. 4.2 holds.
- Let $g(\theta)$ be a continuous function of θ. Then the ML estimator of $g(\theta)$ is $g(\hat{\theta}_{ML})$.
- ML estimators are seldom unbiased for finite n, but the bias can often easily be removed.
- If a sufficient statistic T_n for θ exists, then $\hat{\theta}_{ML}$ is a function of T_n.
- ML estimators are consistent.
- In large samples ($n \to \infty$) $V(\hat{\theta}_{ML}) = 1/I(\theta)$, so ML estimators are MVEs in large samples.
- As $n \to \infty$, $\dfrac{\hat{\theta}_{ML} - \theta}{\sqrt{1/I(\theta)}} \xrightarrow{D} Z \sim N(0,1)$.

ML estimators can be MVEs, also in small samples as will be demonstrated in the Supplementary Exercises of this Chapter.

EX 41 $(Y_i)_{i=1}^n$ are iid with $Y_i \sim Exponential(\lambda)$. Show that the ML estimator of λ is biased and correct it for bias. Determine the variance of the corrected estimator and compare it with the CR limit.

$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda \cdot y_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} y_i} \Rightarrow \ln L(\lambda) = n \ln(\lambda) - \lambda \sum_{i=1}^{n} y_i \Rightarrow \frac{d \ln L(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} y_i = 0 \Rightarrow$

$\hat{\lambda}_{ML} = \frac{n}{\sum_{i=1}^{n} y_i} = \frac{1}{\bar{y}}$. The corresponding estimator is $\hat{\lambda}_{ML} = \frac{n}{\sum_{i=1}^{n} Y_i} = \frac{1}{\bar{Y}}$. (1)

To compute expectation and variance, notice that $\sum_{i=1}^{n} Y_i \sim Gamma(\lambda, n)$ and use the expression in Ch.2.2.2*(2)* with $r = -1$ and $k = n$:

$E(\hat{\lambda}_{ML}) = n \cdot \lambda \frac{\Gamma(n-1)}{\Gamma(n)} = [\text{Cf. Ch. 2.3.5}] = n \cdot \lambda \frac{\Gamma(n-1)}{(n-1)\Gamma(n-1)} = \frac{n}{(n-1)} \lambda$ .. This is not unbiased, but the bias can easily be removed by considering $\hat{\lambda}'_{ML} = \frac{(n-1)}{n} \cdot \hat{\lambda}_{ML} = \frac{n-1}{\sum_{i=1}^{n} Y_i}$.

$V(\hat{\lambda}'_{ML}) = (n-1)^2 V\left[\left(\sum_{i=1}^{n} Y_i\right)^{-1}\right]$. Now, $V\left[\left(\sum_{i=1}^{n} Y_i\right)^{-1}\right] = E\left[\left(\sum_{i=1}^{n} Y_i\right)^{-2}\right] - E^2\left[\left(\sum_{i=1}^{n} Y_i\right)^{-1}\right] =$

$\lambda^2 \frac{\Gamma(n-2)}{\Gamma(n)} - \left(\lambda \frac{\Gamma(n-1)}{\Gamma(n)}\right)^2 = \lambda^2 \left(\frac{\Gamma(n-2)}{(n-1)(n-2)\Gamma(n-2)} - \left(\frac{\Gamma(n-1)}{(n-1)\Gamma(n-1)}\right)^2 \right) =$

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

Thus, $V(\hat{\lambda}'_{ML}) = \frac{\lambda^2}{(n-2)}$ which is larger than the C-R limit $\frac{\lambda^2}{n}$. It can be shown that the C-R limit can't be attained for any estimator. In fact, $\hat{\lambda}'_{ML}$ is an unbiased MVE.

EX 42 Let $Y(t)$ be a Poisson process of rate λ.

a) Find an unbiased estimator of λ and compute the variance of the estimator. Determine the C-R limit and compare this with the variance.

In this case the Likelihood is of a different form: $L(\lambda) = P\big(Y(t) = y\big) = \dfrac{(\lambda t)^y}{y!} e^{-\lambda \cdot t} \Rightarrow \ln\big(L(\lambda)\big) =$

$y\big(\ln(\lambda) + \ln(t)\big) - \ln(y!) - \lambda t \Rightarrow \dfrac{d\ln\big(L(\lambda)\big)}{d\lambda} = \dfrac{y}{\lambda} - t = 0 \Rightarrow \hat{\lambda}_{ML} = \dfrac{Y(t)}{t}$. (Notice here that we use the notation for an estimator rather than the estimate $\dfrac{y}{t}$.)

$E(\hat{\lambda}_{ML}) = \dfrac{1}{t} E\big(Y(t)\big) = \dfrac{1}{t}\lambda t = \lambda$ (Unbiased), $V(\hat{\lambda}_{ML}) = \dfrac{1}{t^2} V\big(Y(t)\big) = \dfrac{1}{t^2}\lambda t = \dfrac{\lambda}{t} \to 0 \text{ as } t \to \infty$.

$\dfrac{d\ln\big(L(\lambda)\big)}{d\lambda} = \dfrac{Y(t)}{\lambda} - t \Rightarrow \dfrac{d^2\ln\big(L(\lambda)\big)}{d\lambda^2} = -\dfrac{Y(t)}{\lambda^2} \Rightarrow I(\lambda) = \dfrac{1}{\lambda^2} E\big(Y(t)\big) = \dfrac{1}{\lambda^2}\lambda t = \dfrac{t}{\lambda}$. Thus, the variance equals the C-R limit $1/I(\lambda)$ and we conclude that $\hat{\lambda}_{ML}$ is an unbiased MVE.

b) Find the ML estimator of $P\big(Y(t) = 0\big) = e^{-\lambda \cdot t}$. Compute an estimate of the latter when *t* = 0.5 and we have observed that $Y(5) = 10$.

The ML estimator is $\left(e^{-\hat{\lambda}_{ML} \cdot t}\right)$ which gives the estimate $e^{-\frac{10}{5} \cdot 0.5} = e^{-1} \approx 0.37$. The latter estimator is in fact biased, but it can be shown that the bias tends to zero with increasing *t*.

EX 43 Let $(Y_1,...,Y_k) \sim Multinomial(n, p_1,..., p_k)$. Determine the ML estimators of $p_1,..., p_k$.

Here $L = C \cdot p_1^{y_1} \cdots p_k^{y_k} = C \cdot p_1^{y_1} \cdots p_{k-1}^{y_{k-1}} \left(1 - \sum_{i=1}^{k-1} p_i\right)^{y_k}$, where $C = \frac{n!}{y_1! \cdots y_k!}$.

We present two solutions, one without and one with the use of Lagrange's multiplier. *Without Lagrange's multiplier*:

$$\ln L = \ln C + \ln \prod_{i=1}^{k-1} p_i^{y_i} + \ln\left(1 - \sum_{i=1}^{k-1} p_i\right)^{y_k} = \ln C + \sum_{i=1}^{k-1} y_i \ln p_i + y_k \ln\left(1 - \sum_{i=1}^{k-1} p_i\right)$$

$$\frac{d \ln L}{dp_i} = 0 + \frac{y_i}{p_i} + y_k \left(\frac{-1}{1 - \sum_{i=1}^{k-1} p_i}\right) = \frac{y_i}{p_i} - \frac{y_k}{p_k} = 0 \Rightarrow p_i = y_i \frac{p_k}{y_k}, i = 1,...,k \quad \text{(i)}$$

Since $\sum_{i=1}^{k} p_i = 1$ we get $\sum_{i=1}^{k} y_i \frac{p_k}{y_k} = \frac{p_k}{y_k} \sum_{i=1}^{k} y_i = \frac{p_k}{y_k} n = 1 \Rightarrow \hat{p}_k = \frac{y_k}{n}$ and this inserted into (i) gives

$$\hat{p}_i = \frac{y_i}{n}.$$

With Lagrange's multiplier:

$\ln L = \ln C + \sum_{i=1}^{k} y_i \ln p_i$ is to be maximized subject to the condition $\sum_{i=1}^{k} p_i = 1$ (ii).

Put $Q = \ln C + \sum_{i=1}^{k} y_i \ln p_i + \lambda\left[\sum_{i=1}^{k} p_i - 1\right] \Rightarrow \frac{dQ}{dp_i} = 0 + \frac{y_i}{p_i} + \lambda = 0 \Rightarrow p_i = -\lambda y_i = \lambda' y_i$. (iii) Putting this into (ii) gives $\lambda' \sum_{i=1}^{k} y_i = \lambda' n = 1 \Rightarrow \lambda' = 1/n$, which inserted into (iii) gives $\hat{p}_i = y_i / n$.

The difficulty in this example arises from the fact that there are just *k*-1 genuine (linearly independent) parameters to estimate.

EX 44 $(Y_i)_{i=1}^n$ are iid where $Y_i \sim Gamma(\lambda, k)$

a) Determine the ML estimators of λ and p.

From EX 33 e) the likelihood is $L = \frac{\lambda^{nk}}{\Gamma^n(k)}\left(\prod_{i=1}^n y_i\right)^{k-1} e^{-\lambda\sum_{i=1}^n y_i} \Rightarrow$

$$\ln L = nk \cdot \ln\lambda - n\ln\Gamma(k) + (k-1)\sum_{i=1}^n \ln y_i - \lambda\sum_{i=1}^n y_i \Rightarrow \begin{cases} \frac{d\ln L}{d\lambda} = \frac{nk}{\lambda} - \sum_{i=1}^n y_i \\ \frac{d\ln L}{dk} = n\ln\lambda - n\frac{d\ln\Gamma(k)}{dk} + \sum_{i=1}^n \ln y_i \end{cases}$$

Putting these expressions equal to zero and solving for the two parameters yields:

$$\frac{nk}{\lambda} = \sum_{i=1}^n y_i \Rightarrow \hat{\lambda}_{ML} = \frac{\hat{k}_{ML}}{\bar{y}} \text{ (i) and } n\ln\left(\frac{k}{\bar{y}}\right) - n\frac{d\ln\Gamma(k)}{dk} + \sum_{i=1}^n \ln y_i = 0 \quad \text{(ii)}$$

Rearranging the terms in equation (ii) gives $\ln k - \frac{d\ln\Gamma(k)}{dk} = \ln\bar{y} - \frac{\sum_{i=1}^n \ln y_i}{n}$ (iii)

By first solving (iii) for $\hat{k}_{ML}$ and then putting this into (i) yields the solutions. However, (iii) has to be solved iteratively. How this can be done is illustrated in b) below.

b) From a sample of *n* = 100 observations the following quantities are calculated:

$$\sum y_i = 223.56,\ \sum y_i^2 = 619.0525,\ \sum \ln y_i = 66.3803$$

Compute the ML estimates of λ and k.

The right hand side of (iii) is ln(2.2356)-0.6638=0.1407. With $g(k) = \ln k - \frac{d\ln\Gamma(k)}{dk}$ we want to determine the value of *k* such that $g(k) = 0.1407$. The function $\frac{d\ln\Gamma(k)}{dk}$ is well known in mathematics and is called the *digamma function*. We can thus plot $g(k)$ against *k* to find a solution of *k*.

Some help in the search for solution is to use the estimate obtained by the Method of Moments. In EX 38 it was shown that $\hat{k}_{Mom} = \bar{y}^2 / s^2$. Now, $s^2 = \left(619.0525 - (223.56)^2/100\right)/(100-1)$ =1.2046 and $\bar{y} = 2.2356$, so $\hat{k}_{Mom} = 4.15$. It is felt that a search for *k* in the interval [3.00, 5.00] should suffice.

The following program code (written in SAS) can be used to find *k*.

```
data a;
do k=3 to 5 by 0.01; g=log(k)-digamma(k); output; end;
proc print; var k g; run;
```

The solution is $\hat{k}_{ML} = 3.71$ and putting this into (i) finally gives $\hat{\lambda}_{ML} = 1.69$.

4.4 Final words

In this chapter we have only considered estimation of parameters and function of parameters. E.g. we can estimate $p(y) = \lambda^y / y! e^{-\lambda}$ by plugging in an estimate of λ. It is also possible to estimate $p(y)$, $f(y)$, $F(y)$ etc. directly from data without model assumptions, but such procedures are beyond the scope of this book.

The information inequality in (22) seems to have been first discovered by Aitken and Silverstone in 1942 during the second World War. During the 1920s Fisher showed that $V(\hat{\theta}_{ML}) = 1/I(\theta)$ in large samples.

Consider the estimation of μ in the normal distribution. The estimator $\overline{Y}$ is unbiased for μ and has variance σ^2/n. An alternative estimator is the sample median, say m. This is also unbiased and has variance $\pi\sigma^2/2n$ in large samples (Rao (1965), p356). The relative efficiency is $V(\overline{Y})/V(m) = 2/\pi \approx 0.63$, and from this it seems obvious that the sample mean is to be preferred. However, there may be other aspects to take account of. In some cases the median is easier to use or can be computed more rapidly. As an example, consider estimation of the mean life length of rats that have been exposed to some drug. If we use the sample mean we have to wait until all rats have died (which may take years). By using the sample median we only have to wait until half of the rats have died.

In some text books one can find the concepts Best Asymptotic Normal (BAN) estimator and Consistent Asymptotic Normal (CAN) estimator. The ML estimator is both BAN and CAN.

Supplementary Exercises, Ch. 4

EX 45 $(Y_i)_{i=1}^n$ are iid with $Y_i \sim Uniform[0,b]$.

a) Find an unbiased estimator of *b* based on the largest observation $Y_{(n)}$ and determine the variance of the estimator.
b) Show that the methods of OLS and Moment give identical unbiased estimators and determine the variance of the estimators.
c) Compare the relative efficiency of the estimators in a) and b).
d) The waiting times at a red traffic light were recorded 10 times and gave the following values (in seconds): 8, 13, 16, 12, 46, 4, 22, 17, 34, 28. Use the data to estimate the time for each red light period. [Chose any estimator you want. Which one is most reliable?]

EX 46 $(Y_i)_{i=1}^n$ are independent with $Y_i \sim Binomial(n_i, p)$.

a) Find unbiased estimators of *p* by using the OLS- and ML methods. Compare the variances of the estimators.
b) Show that the ML estimator is BLUE, in contrast to the OLS estimator, and is in fact a MVE.
c) To estimate *p* = 'Proportion of students with back/neck pain', a sample of students in three class-rooms were taken with the following result:

Room	Total number of students	Number with back/neck pain
1	30	1
2	25	3
3	35	2

Compute the OLS- and ML estimates of *p*.

EX 47 $(Y_i)_{i=1}^n$ are iid with $Y_i \sim Geometric(p)$.

a) Find the ML estimator of *p*.
b) Sometimes it is practical to use *sequentially collected data*, rather than data with a fixed sample size *n*. Consider the following (fictive) data collected from a stream of students passing by:(0,0,0,1,1,0,0,1), where '1' indicate that the student visited a pub last night and '0' that the student did not visit a pub. Estimate the proportion of students who visited a pub last night.

EX 48 $(Y_i)_{i=1}^n$ are independent with $Y_i \sim N(\beta x_i, \sigma^2)$

a) Find the ML estimator of β. Show that it is BLUE and determine the distribution of the estimator.
b) Find the ML estimator of σ^2 and show that it is biased. Remove the bias and determine the distribution of the corrected estimator.

EX 49 In order to estimate mean μ and variance σ^2 in a normally distributed population, one takes independent samples from different regions. Determine the BLUE of the two parameters from the following data:

Region	1	⌜	*k*
Sample size	n_1	⌜	n_k
Sample mean	$\overline{Y}_1$	⌜	$\overline{Y}_k$
Sample variance	S_1^2	⌜	S_k^2

EX 50 Let $(Y_1, Y_2, Y_3) \sim Multinomial(n, p, 2p, (1-3p))$ Find the ML estimator of *p* and check whether it is MVU.

5 Interval estimation

In the previous chapter we considered the estimation of an unknown population parameter θ at a single point. In this chapter we will show how it is possible to construct intervals that enclose θ with a certain degree of confidence. This approach is more informative than that of Ch. 4 since it does not only tell us about the location of the parameter value, but also how confident we should be with the estimate.

5.1 Concepts

A *confidence interval (CI)* is a pair of statistics $(\hat{\theta}_L, \hat{\theta}_U), \hat{\theta}_L < \hat{\theta}_U$, that encloses θ with probability $1-\alpha$, the latter being termed *the confidence coefficient* or *confidence level*. The use of $1-\alpha$ is somewhat confusing, but its origin will be evident in Ch. 6.

Some properties of a CI:

- $\hat{\theta}_L$ and $\hat{\theta}_U$ are both functions of a random sample $Y_1 \ldots Y_n$ and therefore the location and length of the CI will vary randomly.
- There is no guarantee that a *specific CI*, which is a function of $y_1 \ldots y_n$, contains the true value of θ. All we know is that a sequence of specific CIs will contain θ in $100(1-\alpha)\%$ of all cases in the long run.
- It is desirable that the CI is short, in order to be informative, and also that $1-\alpha$ is high, so that the CI is reliable. However these two aspects are incompatible. By increasing $1-\alpha$ we are also increasing the length of the CI.
- There is no golden rule to solve the conflict between length and level of a CI. In practice it is up to the statistician to use common sense in this matter. If the sample size is small and if population variance $V(Y)$ is large, then one should be prepared to decrease the confidence level rather than stick to the conventional level of 0.95.

In Ch. 4 there were some requirements on an estimator, and especially that it should be an unbiased MVE. Here we define a *'best' interval estimator* by the requirement that $E(\hat{\theta}_U - \hat{\theta}_L)$ is minimum for given n and $1-\alpha$.

An important method for finding a CI for θ is to find a *pivotal statistic*, i.e. a statistic with the following properties:

1) It is a function of $Y_1 \ldots Y_n$ and θ.
2) Its probability distribution does not depend of θ.

An example of a pivotal statistic is $S^2(n-1)/\sigma^2 \sim \chi^2(n-1)$ (Cf. EX 18). The subsequent examples will show how the pivotal method works.

5.2 CIs in small samples by means of pivotal statistics

EX 51 $(Y_i)_{i=1}^n$ are iid where $Y_i \sim N(\mu,\sigma^2)$.

a) Construct 95% CIs for μ and σ^2 when *n* = 10.
b) Determine the specific CIs when $\bar{y} = 80$ and $S^2 = 81$.

a) CI for μ.

From Ch. 3.1*(11)*, $\dfrac{(\bar{Y}-\mu)}{S/\sqrt{n}} = \dfrac{\dfrac{(\bar{Y}-\mu)}{\sigma/\sqrt{n}} \sim N(0,1)}{\dfrac{S/\sqrt{n}}{\sigma/\sqrt{n}} \sim \sqrt{\dfrac{\chi^2(n-1)}{(n-1)}}} \sim T(n-1)$. This quantity is pivotal. Let *C* be some constant, then since the *T*- distribution is symmetric around zero

$$1-\alpha = P\left(-C < \frac{\bar{Y}-\mu}{S/\sqrt{n}} < C\right) = \begin{cases} P\left(\bar{Y} - C\dfrac{S}{\sqrt{n}} < \mu < \bar{Y} + C\dfrac{S}{\sqrt{n}}\right) & \text{(i)} \\ P\left(-C < T(n-1) < C\right) & \text{(ii)} \end{cases}$$

In (i) we have simply rearranged the inequality so that μ is centered. (ii) is used to determine *C* in which case we must know the confidence level $1-\alpha$ and *n*.

With $1-\alpha = 0.95$ we get: $P\left(-C < T(9) < C\right) = 0.95 \Rightarrow C = 2.262$, obtained from tables of the *T*- distribution. The 95% CI for μ is thus $\left(\bar{Y} - 2.262\dfrac{S}{\sqrt{10}}, \bar{Y} + 2.262\dfrac{S}{\sqrt{10}}\right)$.
Notice that the interval can be constructed before the sample is taken.

CI for σ^2

From EX 18, $S^2 \sim \dfrac{\sigma^2}{(n-1)}\chi^2(n-1)$ which is not pivotal, but $\dfrac{S^2(n-1)}{\sigma^2} \sim \chi^2(n-1)$ is. The chi-square distribution is not symmetric so we consider two constants *a* and *b* such that

$$1-\alpha = P\left(a < \frac{S^2(n-1)}{\sigma^2} < b\right) = \begin{cases} P\left(\dfrac{S^2(n-1)}{b} < \sigma^2 < \dfrac{S^2(n-1)}{a}\right) \\ P\left(a < \chi^2(n-1) < b\right) \end{cases}$$

With $1-\alpha = 0.95$ the area below *a* under the chi-square density is 0.025 and the area above *b* is 0.025. Thus, $P\left(a < \chi^2(9) < b\right) = 0.95 \Rightarrow a = 2.7004, b = 19.0228$. The 95% CI for σ^2 is $\left(\dfrac{9S^2}{19.0228}, \dfrac{9S^2}{2.7004}\right)$.

b)

The lower and upper limits in the specific CI for μ are $80 \pm 2.262\dfrac{9}{\sqrt{10}} = 80 \pm 6.4$, so the 95% CI is (73.6,86.4).

The specific CI for σ^2 is $\left(\dfrac{9\cdot 81}{19.0228}, \dfrac{9\cdot 81}{2.7004}\right)$ or $(38.3, 270.0)$. This interval is very wide, as it should be since variance is a squared quantity, such as dollar^2 or kg^2. It would be wise to use a confidence level lower than 95% in this case.

EX 52 $(Y_i)_{i=1}^n$ are iid with $Y_i \sim Uniform[0,b]$

a) Construct a 95% CI for *b* based on the largest observation in the sample.
b) Determine the specific 95% CI from the following data: 28, 19, 31, 12, 15.

a)

From (8) in Ch. 3.1the cdf of the largest observation $Y_{(n)}$ is $F_{Y_{(n)}}(y) = \left(\frac{y}{b}\right)^n$, $0 \le y \le b$. $Y_{(n)}$ is

not pivotal, but we may try to find a pivotal statistic by the following device. Consider $CY_{(n)}$ with cdf

$F_{CY_{(n)}}(y) = P\left(CY_{(n)} \le y\right) = P\left(Y_{(n)} \le y/C\right) = F_{Y_{(n)}}\left(y/C\right)) = \left(\frac{y}{Cb}\right)^n$. This is not dependent on *b* if $C = 1/b$.

Thus, $Y_{(n)}/b$ is pivotal with cdf $F_{Y_{(n)}/b}(y) = y^n, 0 \le y \le b$.

We now proceed as in EX 51.

$$0.95 = P\left(c_1 < Y_{(n)}/b < c_2\right) \Rightarrow \begin{cases} P\left(Y_{(n)}/b < c_1\right) = c_1^n = 0.025 \Rightarrow c_1 = 0.025^{\frac{1}{n}} \\ P\left(Y_{(n)}/b < c_2\right) = c_2^n = 0.975 \Rightarrow c_2 = 0.975^{\frac{1}{n}} \end{cases}$$

It remains to put the parameter *b* in the center,

$$P\left(c_1 < Y_{(n)}/b < c_2\right) = P\left(\frac{Y_{(n)}}{c_2} < b < \frac{Y_{(n)}}{c_1}\right) \Rightarrow \left(\frac{Y_{(n)}}{0.975^{\frac{1}{n}}}, \frac{Y_{(n)}}{0.025^{\frac{1}{n}}}\right)$$ is a 95% CI for *b*.

b)

Here $n = 5$ and $Y_{(n)} = 31 \Rightarrow \left(\frac{31}{0.975^{\frac{1}{5}}}, \frac{31}{0.025^{\frac{1}{5}}}\right) = (31.2, 64.8)$

EX 53 Given two independent sets of iid variables $(X_i)_{i=1}^{n_X}$ and $(Y_i)_{i=1}^{n_Y}$ where $X_i \sim N(\mu_X, \sigma_X^2)$ and $Y_i \sim N(\mu_Y, \sigma_Y^2)$.

a) Construct a CI for the ratio σ_X^2 / σ_Y^2.
b) Construct a CI for the difference $(\mu_X - \mu_Y)$ assuming that $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.

a)

Consider the two unbiased estimators $\hat{\sigma}_X^2 = S_X^2$ and $\hat{\sigma}_Y^2 = S_Y^2$. These are independent since they are based on independent sets of variables. From EX 18 in Ch. 3.1 we know that

$$S_X^2 \sim \frac{\sigma_X^2}{(n_X - 1)} \chi^2(n_X - 1) \text{ and } S_Y^2 \sim \frac{\sigma_Y^2}{(n_Y - 1)} \chi^2(n_Y - 1) \Rightarrow$$

$$\frac{S_X^2}{S_Y^2} \sim \frac{\sigma_X^2}{\sigma_Y^2} \cdot \frac{\chi^2(n_X - 1)(n_X - 1)}{\chi^2(n_Y - 1)(n_Y - 1)} \sim \frac{\sigma_X^2}{\sigma_Y^2} \cdot F(n_X - 1, n_Y - 1)$$ [Cf. the *F*- distribution in Ch. 3.1.]

The latter quantity is not pivotal, but $\frac{\sigma_Y^2}{\sigma_X^2} \cdot \frac{S_X^2}{S_Y^2} \sim F(n_X - 1, n_Y - 1)$ is.

EX 53 (Continued)

The *F*- distribution is not symmetric so we choose two constants c_1 and c_2 as limits in the following inequality

$$1-\alpha = P\left(c_1 < \frac{\sigma_Y^2}{\sigma_X^2} \cdot \frac{S_X^2}{S_Y^2} < c_2 \right) = \begin{cases} P\left(\frac{S_X^2}{c_2 S_Y^2} < \frac{\sigma_X^2}{\sigma_Y^2} < \frac{S_X^2}{c_1 S_Y^2} \right) & \text{(i)} \\ P\left(c_1 < F(n_X - 1, n_Y - 1) < c_2 \right) & \text{(ii)} \end{cases}$$

In (i) we have simply centered the variance ratio. The expression in (ii) is used to determine the two constants c_1 and c_2. However, this may be cumbersome, especially since *F*- tables often are incomplete. We devote a few lines to show how this can be done.

Let $n_1 = 25, n_2 = 10$ and $1-\alpha = 0.95$. We want to determine c_1 and c_2 so that

(ii) holds. Since $P(F(24,9) > c_2) = 0.025$ we obtain $c_2 = 3.61$ by using Table 7 in

Wackerly *et al*. It is harder to find the value of c_1 from the table. The value of c_1 giving

$P(F(24,9) > c_1) = 0.975$ or $P(F(24,9) < c_1) =$ 0.025 is not shown. Instead we use the fact that [See Ch. 3.1 *(12)*.]

$$P(F(24,9) < c_1) = P\left(\frac{1}{F(9,24)} < c_1 \right) = P\left(F(9,24) > \frac{1}{c_1} \right) = 0.025 \Rightarrow \frac{1}{c_1} = 2.70 \Rightarrow c_1 = 0.37 .$$

In this case the 95% CI for $\frac{\sigma_X^2}{\sigma_Y^2}$ is $\left(\frac{S_X^2}{3.61 \cdot S_Y^2}, \frac{S_X^2}{0.37 \cdot S_Y^2} \right)$.

b)

$\overline{X} \sim N(\mu_X, \sigma^2 / n_X), \overline{Y} \sim N(\mu_Y, \sigma^2 / n_Y) \Rightarrow (\overline{X} - \overline{Y}) \sim N(\mu_X - \mu_Y, \sigma^2(1/n_X + 1/n_Y))$, since a linear function of

normally distributed variables is itself normally distributed (Cf. Ch. 2.2.2). Thus,

$\frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma^2(1/n_X + 1/n_Y)}} \sim N(0,1)$. This is pivotal, but it can't be used since σ^2 is unknown. We need an estimator

of σ^2.

From EX 49 it follows that $\hat{\sigma}^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X - 1 + n_Y - 1}$ is BLUE for σ^2. Since $(n_X - 1)S_X^2 \sim \sigma^2 \chi^2(n_X - 1)$ and

$(n_Y - 1)S_Y^2 \sim \chi^2(n_y - 1)$ it follows that

$$\hat{\sigma}^2 \sim \frac{\sigma^2\left(\chi^2(n_X - 1) + \chi^2(n_Y - 1)\right)}{n_X - 1 + n_Y - 1} \sim \sigma^2 \frac{\chi^2(n_X + n_Y - 2)}{(n_X + n_Y - 2)}$$ [Cf. Ch. 3.1 *(8)*]

The following statistic is pivotal and useful

$$\frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\hat{\sigma}^2(1/n_X + 1/n_Y)}} = \frac{\frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma^2(1/n_X + 1/n_Y)}}}{\frac{\sqrt{\hat{\sigma}^2(1/n_X + 1/n_Y)}}{\sqrt{\sigma^2(1/n_X + 1/n_Y)}}} \sim \frac{N(0,1)}{\sqrt{\frac{\chi^2(n_X + n_Y - 2)}{(n_X + n_Y - 2)}}} \sim T(n_X + n_Y - 2)$$

Proceeding in the same way as in EX 51 we finally obtain the lower and upper CI limits as

$\overline{X} - \overline{Y} \pm C\sqrt{\hat{\sigma}^2(1/n_X + 1/n_Y)}$, where *C* is determined from the $T(n_X + n_Y - 2)$ -distribution.

Comment to EX 53 The results in this exercise are crucial for comparing two means. By making a CI for $(\mu_X - \mu_Y)$ the main interest is whether the CI encloses zero. If a 95% CI, or a CI of higher level, does not contain zero it is customary to conclude that the two means are *significally different*. This way of claiming statistical significance is different from another one based on statistical hypothesis testing that is considered in Ch. 6.

Notice that the first step was to make a CI for the variance ratio σ_X^2 / σ_Y^2. If the latter encloses 1, the customary conclusion is that there is no significant difference between the variances, which therefore could be set equal. The CI for $(\mu_X - \mu_Y)$ was then constructed under this assumptions. On the other hand, if the CI for the variance ratio does not enclose 1 we can't claim that variances are equal and a different approach is required. This is called the *Behrens-Fisher's problem* for which several approximate solutions have been suggested. The latter are however beyond the level of this book. Most statistical software present solutions both with and without the assumption of equal variances.

EX 54 $(Y_i)_{i=1}^n$ are independent with $Y_i \sim N(\beta x_i, \sigma^2)$. Construct CIs for β and σ^2.

CI for β

From EX 48 we know that $\hat{\beta}_{ML} = \hat{\beta}_{BLUE} = \frac{\sum x_i Y_i}{\sum x_i^2}$ and that $\frac{\hat{\beta}_{ML} - \beta}{\sqrt{\sigma^2 / \sum x_i^2}} \sim N(0,1)$. The latter statistic is pivotal but it can't be used since σ^2 is unknown. As an unbiased estimator of σ^2 we take $\hat{\sigma}^2 = \frac{\sum (Y_i - \hat{\beta}_{ML} x_i)^2}{(n-1)} \sim \frac{\sigma^2}{(n-1)} \chi^2(n-1)$ [Cf. EX 48], where $\hat{\beta}_{ML}$ and $\hat{\sigma}^2$ are independent.
A pivotal statistic that is useful is

$$\frac{\hat{\beta}_{ML} - \beta}{\sqrt{\hat{\sigma}^2 / \sum x_i^2}} = \frac{\frac{\hat{\beta}_{ML} - \beta}{\sqrt{\sigma^2 / \sum x_i^2}}}{\frac{\sqrt{\hat{\sigma}^2 / \sum x_i^2}}{\sqrt{\sigma^2 / \sum x_i^2}}} \sim \frac{N(0,1)}{\sqrt{\frac{\chi^2(n-1)}{(n-1)}}} \sim T(n-1),$$

the latter distribution being symmetric around zero. Thus

$$1 - \alpha = P\left(-C < \frac{\hat{\beta}_{ML} - \beta}{\sqrt{\hat{\sigma}^2 / \sum x_i^2}} < C \right) = \begin{cases} P\left(\hat{\beta}_{ML} - C\sqrt{\hat{\sigma}^2 / \sum x_i^2} < \beta < \hat{\beta}_{ML} + C\sqrt{\hat{\sigma}^2 / \sum x_i^2} \right) \\ P(-C < T(n-1) < C) \end{cases}$$

The CI for β is $\hat{\beta}_{ML} \pm C\sqrt{\hat{\sigma}^2 / \sum x_i^2}$ where *C* is determined from the $T(n-1)$ - distribution. To illustrate the computation of C, let *n* = 10 and assume that we want a 90% CI for β. Since the area under the *T* – density between –*C* and *C* is 0.90, the area above *C* is 0.05. (Most tables today show areas above *C*.) From the tables we get *C* = 1.833.

CI for σ^2

$\hat{\sigma}^2$ is not pivotal, but $\frac{\hat{\sigma}^2 (n-1)}{\sigma^2} \sim \chi^2(n-1)$ is. Therefore, and since the chi-square distribution is not symmetric, there are two constants *a* and *b* to be determined.

$$1 - \alpha = P\left(a < \frac{\hat{\sigma}^2 (n-1)}{\sigma^2} < b \right) = \begin{cases} P\left(\frac{\hat{\sigma}^2 (n-1)}{b} < \sigma^2 < \frac{\hat{\sigma}^2 (n-1)}{a} \right) \\ P(a < \chi^2(n-1) < b) \end{cases}$$

Here *a* and *b* are determined as in EX 51. The CI for σ^2 is thus $\left(\frac{\hat{\sigma}^2 (n-1)}{b}, \frac{\hat{\sigma}^2 (n-1)}{a} \right)$.

5.3 Approximate CIs in large samples based on Central Limit Theorems

In Ch. 2.2.2 (a) the Central Limit Theorem (CLT) was stated for a standardized sum of iid variables, denoted by Z_n, and for a Poisson process in Ch. 2.2.2 (c), denoted by $Z(t)$. In Ch. 4.3.4 it was stated that standardized ML estimators have asymptotic $N(0,1)$-distributions. These results can be used to find CIs that holds approximately in large samples. Since the CIs only hold approximately it is important to understand the meaning of 'a large sample'. Below some examples of CIs derived from CLTs are given.

EX 55 $Y \sim Binomial(n,p)$, or equivalently $(Y_i)_{i=1}^n$ are iid with $Y_i \sim Bernoulli(p)$. Determine a CI for *p* based on $Y = \sum Y_i$ and a large n. Put $\hat{p} = Y/n$.

a) Give a justification for the formula $\hat{p} \pm C\sqrt{\hat{p}(1-\hat{p})/n}$ that is often found in textbooks. (Sometimes division by *n* –1 is used instead of *n*.).

From EX 23c)

$$1-\alpha = P\left(-C < \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} < C\right) = \begin{cases} P\left(\hat{p} - C\sqrt{\hat{p}(1-\hat{p})/n} < p < \hat{p} + C\sqrt{\hat{p}(1-\hat{p})/n}\right) \\ P\left(-C < Z_1 < C\right) \end{cases}$$

where *C* is determined from tables of the Normal distribution

b) Show that a more accurate CI for *p* is given by the limits

$$\frac{2\hat{p} + C^2/n \pm \sqrt{\left(2\hat{p} + C^2/n\right)^2 - 4\hat{p}^2\left(1 + C^2/n\right)}}{2\left(1 + C^2/n\right)}$$

From EX 23b)

$$1-\alpha = P\left(-C < \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} < C\right) = \begin{cases} P\left((\hat{p}-p)^2 < C^2\,\frac{p(1-p)}{n}\right) & \text{(i)} \\ P\left(-C < Z_2 < C\right) & \text{(ii)} \end{cases}$$

By solving the inequality in (i) for *p* it can be shown that *p* is located between the limits stated in b). *C* in (ii) is obtained from the normal distribution. The CI in b) is more accurate since the approach to normality goes much faster for Z_2 than for Z_1. Notice that the approach to normality for Z_1 requires not only convergence in distribution, but also convergence in probability as was shown in EX 23 c).

Comment to EX 55 It is important to know the difference between the expressions given in EX 55 a) and b). The former is often stated in textbooks as being a result of the CLT and sometimes lowest sample sizes of 30–50 are advocated. This may hold for the validity of the expression in b), but definitely not for the expression in a).

In 1963 an interesting relation was found between the Binomial – and F distributions by G.H. Jowett. By using this it is possible to obtain a CI for *p* that holds for any sample size. The latter may be hard to find in text books at master level in statistics, but yet we present it here since the result is very useful. (Cf. Casella & Berger 1990, p. 499.)

Let $F_{.975}(f_1, f_2)$ be the 97.5% percentile of the F distribution, i.e. $P\big(F(f_1, f_2) > F_{.975}(f_1, f_2)\big) = 0.025$. Then a 95% CI for *p* is given by

$$\left(\frac{Y}{Y + (n - Y + 1)F_{.975}(2(n - Y + 1), 2n)}, \frac{(Y+1)F_{.975}(2(Y+1), 2(n-Y)}{n - Y + (Y+1)F_{.975}(2(Y+1), 2(n-Y)} \right) \quad (23)$$

If $Y = 0$ then the lower limit is 0 and if $Y = n$ then the upper limit is 1.

The expression in (23) gives CIs that are *conservative* in the sense that they give CIs with a confidence level of at least 95%. For simplicity a 95% CI was considered. If a 99% CI was required we would instead search for 99.5% percentiles in the F-distribution.

EX 56 Use the expressions for a CI in (21) and in EX 55 a) and b) to calculate 95% CIs for *p* in the two cases $(y = 2, n = 20)$ and $(y = 10, n = 100)$.

To use (23) we have to determine the 97.5% percentiles of the F-distribution. This can be a problem since F-tables are often incomplete and percentiles are only shown for a few degrees of freedom. The best one can do is to use statistical software, such as SAS or SPSS, to find the percentiles. In worst case one may be forced to use linear interpolation.

In the case $(y = 2, n = 20)$ we find $F_{.975}(38,4) = 8.4191$ and $F_{.975}(6,36) = 2.7846$. (These values were obtained by using the function $finv(0.975, f_1, f_2)$ in SAS.) Similarly, in the case $(y = 10, n = 100)$ we get $F_{.975}(182,20) = 2.1326$ and $F_{.975}(22,180) = 1.7503$.

In both cases we get the same point estimate of *p,* 2/20=10/100 = 10%, but the CIs are different:

	CI (%) for *p*		
Expression in:	(23)	55 b)	55 a)
$(y = 2, n = 20)$	(1.2, 31.7)	(2.8, 30.1)	(-3.1, 23.1)
$(y = 10, n = 100)$	(4.9, 17.6)	(5.5, 17.4)	(4.1, 15.7)

The CIs based on (23) are certainly wider, but they are more reliable since they are conservative as mentioned above. Notice that the expression in 55 a) can result in peculiar CIs in small samples.

EX 57 Determine a CI for the rate λ in a Poisson process.

In Ch. 2.2.2 *(4)(c)* it was seen that, if $Y(t)$ is a Poisson process of rate λ ,then

$Z(t)=\frac{Y(t)-\lambda t}{\sqrt{\lambda t}}=\frac{Y(t)/t-\lambda}{\sqrt{\lambda/t}}\xrightarrow{D}Z\sim N(0,1)\text{ as }t\to\infty$. This statistic is asymptotically pivotal, but there are some difficulties to obtain an inequality for λ by using this fact. Instead we notice that the statistic $\hat{\lambda}=Y(t)/t\xrightarrow{P}\lambda,\text{ as }t\to\infty.$ [This follows from (10) in Ch. 3.2.1 since $E(\hat{\lambda})=\frac{1}{t}E\big(Y(t)\big)=\frac{\lambda t}{t}=\lambda$ and $V(\hat{\lambda})=\frac{1}{t^2}V\big(Y(t)\big)=\frac{\lambda t}{t^2}=\frac{\lambda}{t}\to 0,\text{ as }t\to\infty.$]. Thus,

$\frac{\hat{\lambda}-\lambda}{\sqrt{\hat{\lambda}/t}}=\frac{\frac{\hat{\lambda}-\lambda}{\sqrt{\lambda/t}}\xrightarrow{D}Z}{\sqrt{\frac{\hat{\lambda}/t}{\lambda/t}}\xrightarrow{P}1}$, and from (11) in Ch. 3.2.1 we get $\frac{\hat{\lambda}-\lambda}{\sqrt{\hat{\lambda}/t}}\xrightarrow{D}Z$. Now,

$1-\alpha=P\left(-C<\frac{\hat{\lambda}-\lambda}{\sqrt{\hat{\lambda}/t}}<C\right)=\left\{P\left(\hat{\lambda}-C\sqrt{\hat{\lambda}/t}<\lambda<\hat{\lambda}+C\sqrt{\hat{\lambda}/t}\right)\right.$. So $\hat{\lambda}\pm C\sqrt{\hat{\lambda}/t}$ are the CI limits for λ ,

$\left.P(-C<Z<C)\right.$

where *C* is determined from tables over the normal distribution.

EX 58 $(Y_i)_{i=1}^n$ are iid variables from an unspecified distribution with mean μ and variance σ^2 . If *n* is large a 95% CI for μ is given by

$$\overline{Y}\pm 1.96\frac{S}{\sqrt{n}}$$

(This is perhaps the most cited expression in statistical inference and is found in most elementary text books. Sometimes 1.96 is replaced by the figure 2,)

Give a rigorous motivation for the expression!

From (9b) $V(S^2)\to 0,\text{ as }n\to\infty\Rightarrow S^2\xrightarrow{P}\sigma^2$ (Cf. (10) in Ch. 3.3.1) $\Rightarrow g(S^2)=\sqrt{\frac{S^2}{\sigma^2}}\xrightarrow{P}1$ (Cf. (11) in Ch. 3.3.1). Thus, $\frac{\overline{Y}-\mu}{S/\sqrt{n}}=\frac{\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}\xrightarrow{D}Z\sim N(0,1)}{\frac{S/\sqrt{n}}{\sigma/\sqrt{n}}\xrightarrow{P}1}\xrightarrow{D}Z\sim N(0,1)$.

For large *n*, $0.95=P\left(-1.96<\frac{\overline{Y}-\mu}{S/\sqrt{n}}<1.96\right)=P\left(\overline{Y}-1.96\frac{S}{\sqrt{n}}<\mu<\overline{Y}+1.96\frac{S}{\sqrt{n}}\right)$.

The simple expression above should be used with caution. Especially if the population distribution is heavily skewed or has multiple peaks, a very large *n* would be required.

5.4 Some further topics

5.4.1 Selecting the sample size

Looking back at the examples of this chapter it is seen that the bounds of a CI are functions of the sample size *n*. This opens the possibility to determine *n* in advance in such a way that the CI has a stipulated length. The problem is that the bounds of a CI are also dependent on the values of one or several statistics that not yet have been computed. There is no simple solution to this problem but some guide lines can be given when the CI has the structure $\hat{\theta} \pm C\sqrt{\hat{V}(\hat{p})}$. The term $C\sqrt{\hat{V}(\hat{\theta})}$ is called *Bound on the Error (*BE*)*. We consider two cases, a proportion and a mean.

- *Bernoulli proportion p in large samples*

The CI is $\hat{p} \pm C\sqrt{\hat{p}(1-\hat{p})/n}$, so $BE = C\sqrt{\hat{p}(1-\hat{p})/n}$. (Division by *n* – 1 instead of *n* is of minor importance.) Values of $\hat{p}$ can be obtained in several ways:

- *Worst case scenario.* Choose $\hat{p} = 1/2$. It is easily shown that this value maximizes $\hat{p}(1-\hat{p})$ for $0 \le \hat{p} \le 1$. The maximal BE now becomes $C/2\sqrt{n}$. This solution should only be used when there is no information whatsoever about $\hat{p}$.
- *Qualified guess.* Here one uses earlier experience to guess the value of $\hat{p}$. Notice that the function $\hat{p}(1-\hat{p})$ is symmetric around $\hat{p} = 1/2$ so e.g. $\hat{p} = 0.10$ gives the same BE as $\hat{p} = 0.90$
- *Pilot study.* The idea is to take a first small sample (pilot sample) to estimate *p*. Observations from the pilot sample could then be included into the final sample. The approach is appealing since it is free from more or less reliable assumptions. A problem is to decide how large the pilot sample shall be. One solution is to collect data *sequentially* and compute estimates $\hat{p}_n$ for increasing *n* until the estimates have stabilized. Usually this occurs for *n* less than 20-30.

After having determined an appropriate value of $\hat{p}$ it is instructive to plot BE on the Y-axes against *n* on the X-axis for various choices of *C*. (Remember that $C = 1.645, 1.960, 2.575$ corresponds to the confidence levels 90%, 95% and 99%, respectively.) Alternatively, in the expression for BE above one can solve for *n*, giving $n = \hat{p}(1-\hat{p})\left(C/B\right)^2$.

One should be aware that data collection in large samples can be costly. A simple expression for the total cost is $c_0 + c \cdot n$, where c_0 is a fixed cost and *c* is the cost for each sample unit.

EX 59 Determine the sample size needed to get a CI for *p* with a BE of 0.01 or alternatively 0.025. The CI levels shall be 90%, 95% or 99%.

a) Use the worst case scenario.
b) Use a pilot sample with the data $(Y_i)_{i=1}^{15}$ (0,0,1,0,0,0,0,1,0,0,0,0,0,0,1) where $P(Y_i = 1) = p$.

In a) we use $\hat{p} = 1/2$.

In b) we conclude from the table below that $\hat{p} = 0.20$ may be appropriate

n	10	11	12	13	14	15
$\sum_{i=1}^{n} y_i / n$	.20	.18.	.17	.15	.14	.20

The following table illustrates how the sample size *n* differs between the two approaches in a) and b):

a)	*p* = 0.5		**b)**	*p* = 0.2	
CI level	BE	*n*	CI level	BE	*n*
90%	0.01	6765	90%	0.01	4330
95%	"	9604	95%	"	6147
99%	"	16577	99%	"	10609
90%	0.025	1082	90%	0.025	693
95%	"	1537	95%	"	983
99%	"	2652	99%	"	1697

It is seen that the sample size increases with increasing CI level and decreases with increasing length of the CI. The approach in a) leads to unnecessary large samples compared with the approach in b).

- *Population mean μ in large samples*

In EX 58 it was shown that the limits $\overline{Y} \pm C \cdot S/\sqrt{n}$ gives a CI for μ in large samples provided that the observations are iid. The Bound on the Error is $BE = C \cdot S/\sqrt{n}$ from which $n = C^2 S^2 / BE^2$. In the latter expression *S* can be determined in at least two ways.

- '*Empirical rule*'. Replace S^2 by the true variance σ^2. Since 99% of the observations are found within the variation limits $\mu \pm 2.58\sigma$, the range of *y*–values is roughly $2 \cdot 2.58\sigma \approx 5.2\sigma$. From this we get $S \approx \sigma \approx range/5.2\sigma$. (Sometimes the figure 2.8 is replaced by $1.96 \approx 2$, corresponding to 95% variation limits, which gives $S \approx range/4\sigma$.)

  This approach has several drawbacks. There is a great amount of arbitrariness in the choice of coefficient, 2.58 or 2, and sometimes even 3. As a consequence there will be large differences in the choice of *n*. Furthermore, in many cases it can be hard to identify the range of possible y-values.

- *Pilot study*. As in the case with a Bernoulli proportion, we may take a first small sample to obtain a likely value of S^2. Data are collected sequentially until the value of S^2 has stabilized. The calculations can be performed in any of the following ways.

  For $n > 5$, say: $\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} y_i^2 \Rightarrow S_n^2 = \left(\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2 / n \right) / (n - 1)$ or from the recursive relations $\bar{y}_{n+1} = \frac{y_{n+1} + n \cdot \bar{y}_n}{n+1}, S_{n+1}^2 = \frac{(n-1)S_n^2}{n} + \frac{(y_{n+1} - \bar{y}_n)^2}{(n+1)}$. (The latter relations are found in Casella & Berger, p. 244.)

EX 60 In order to construct a 95% CI for the mean Area of Optic Disc (AOD) in a group of children, one wants to determine the sample size *n* that is needed to get a Bound on the Error (BE) of about 0.10 to 0.20 (mm^2). In a pilot study the following values were obtained sequentially.

n	1	2	3	4	5	6	7	8	9	10
AOD	2.39	3.33	2.12	1.90	2.66	2.53	2.30	2.98	2.59	2.70
n	11	12	13	14	15					
AOD	2.20	2.77	1.86	2.72	3.28					

Use the data to first calculate a value of S^2 and then suggest a proper sample size.

The sequentially calculated values of S^2 are, starting from $n = 7$.

n	7	8	9	10	11	12	13	14	15
S_n^2	0.21	0.21	0.19	0.17	0.16	0.15	0.18	0.17	0.20

An appropriate value seems to be $S^2 = 0.20$ that gives $n = 1.96^2 \cdot 0.20 / BE^2$. The desired sample sizes are thus $n = 77$ for $BE = 0.10$ and $n = 20$ for $BE = 0.20$.

5.4.2 CI for a function of a parameter

Given a CI for θ, say $\hat{\theta}_L < \theta < \hat{\theta}_U$, it is possible to make a CI for a function of θ, $g(\theta)$, provided that the latter is monotonous (decreasing or increasing). The approach is illustrated in the following examples.

EX 61 $(Y_i)_{i=1}^n$ are iid where $Y_i \sim Exponential(\lambda)$.

a) Determine a 95% CI for λ based on the statistic $\sum_{i=1}^n Y_i$. Compute the CI limits when *n*=50 and $\sum_{i=1}^n y_i = 65.0$.

Compute the corresponding CI limits for the survival function $P(Y > y) = e^{-\lambda \cdot y}$.

$Y_i \sim Gamma(\lambda,1) \Rightarrow [\text{Ch. } 3.1] \Rightarrow \sum_{i=1}^n Y_i \sim Gamma(\lambda,n) \Rightarrow [\text{Ch..2.2.2}] \Rightarrow$

$2\lambda \sum_{i=1}^n Y_i \sim \chi^2(2n)$. Thus, $0.95 = P\left(a < 2\lambda \sum_{i=1}^n Y_i < b\right) = \left\{ P\left(\frac{a}{2\sum Y_i} < \lambda < \frac{b}{2\sum Y_i}\right) \right.$
$\left. P(a < \chi^2(2n) < b) \right.$

From tables of the Chi square distribution (e.g. in Wackerly *et al* 2007, pp850-851) we get, with

$n = 50$, $\begin{cases} P\left(\chi^2(100) > b\right) = 0.025 \Rightarrow b = 129.56 \\ P\left(\chi^2(100) > a\right) = 0.975 \Rightarrow a = 74.22 \end{cases}$.

The CI for λ is thus $\left(\frac{74.22}{2 \cdot 65}, \frac{129.56}{2 \cdot 65}\right) = (0.57, 1.00)$.

b) $P(Y > y) = e^{-\lambda \cdot y}$ is monotonously decreasing with λ. Therefore

$\hat{\lambda}_L < \lambda < \hat{\lambda}_U \Rightarrow e^{-\hat{\lambda}_U \cdot y} < e^{-\lambda \cdot y} < e^{-\hat{\lambda}_L \cdot y}$. It follows that the 95% CI for the survival function is $\left(e^{-y}, e^{-0.57y}\right)$.

Notice that the latter CI will cover the true value in 95% of all cases at *one specific value of y*. It may be tempting to plot the lower and upper limits against *y*, thereby creating a so called confidence region. The latter will however not contain the true values in 95% of all the cases, since we are making several confidence statements simultaneously which in turn will reduce the confidence level. This *Multiple inference problem* is discussed further in Ch. 6.4.

In Ch. 2.2.1 (3) it was stated that if $X(s)$ and $Y(t)$ are Poisson processes of rates λ_X and λ_Y, respectively, then the conditional variable $\left(Y(t) \middle| X(s) + Y(t) = n\right) \sim Binomial(n, p = \frac{\lambda_Y t}{\lambda_X s + \lambda_Y t})$. This can be used to make a CI for the ratio $R = \lambda_Y / \lambda_X$. Due to its importance we formulate the solution of the problem as a theorem.

A CI for the ratio of two Poisson rates $R = \lambda_Y / \lambda_X$ can be constructed in the following way:

a) First, make a CI for the Binomial proportion *p* giving $\left(\hat{p}_L, \hat{p}_U\right)$.
b) A CI for *R* is then obtained as $\left(\hat{R}_L = \frac{\hat{p}_L}{(1-\hat{p}_L)} \frac{s}{t}, \hat{R}_U = \frac{\hat{p}_U}{(1-\hat{p}_U)} \frac{s}{t}\right)$. (24)

This follows easily from the fact that $p = \frac{Rt}{Rt + s} \Rightarrow R(p) = \frac{p}{(1-p)} \frac{s}{t}$ and this is a function that increases monotonously from $R(0) = 0$ to infinity as $p \to 1$.

EX 62 In the snow-free period April–November there were 85 road accidents on a certain stretch of road and during the winter period December–March there were 65 road accidents on the same stretch. Is the rate of road accidents significantly higher during the winter period?

Introduce the notations

$X(8) =$ Number of accidents in the snow - free period, of rate λ_X

$Y(4) =$ -"- winter period, of rate λ_Y

We will answer the question about significance by making a 95% CI for $R = \lambda_Y / \lambda_X$.

If the latter does not cover 1 we draw the conclusion that there is a significant difference.

The observed proportion of $Y(4)/\big(Y(4)+X(8)\big)$ is $\hat{p} = 65/(65+85) = 0.4333$. Since *n* is large we use the expression $\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$ for a 95% CI in EX 55 a). This yields the limits 0.4333 ± 0.0809 or the CI $(0.3524, 0.5142)$. The expression in (24) finally gives the CI limits for *R*:

$\hat{R}_L = \dfrac{0.3524}{(1-0.3524)}\dfrac{8}{4} = 1.09$, $\hat{R}_U = \dfrac{0.5142}{(1-0.5142)}\dfrac{8}{4} = 2.17$. Since the latter interval does not cover 1and is in fact located above 1, the conclusion is that the rate of road accidents is significantly higher in the winter period.

In Ch. 6 we will consider other ways to claim statistical significance.

5.5 Final words

Verify that you can find the points *a* and *b* in the χ^2- and $F-$ distributions such that 2.5% of the observations are smaller than *a* and 97.5% of the observations are larger than *b*. The intervals in the examples of Ch. 5 are 95% CIs. Change the confidence levels to 90% and 99% to study the effect on the lengths of the CIs.

Remember the interpretation of a CI. If you repeatedly construct 95% CIs, then in the long run there will be 1 interval of 20 that doesn't cover the true parameter value.

Notice that proportions around 1/2 require the largest sample size for a given confidence level and Bound on the Error. Many people do not agree about this. Therefor you should go through the arguments in Ch. 5.4.1, so you can persuade them.

Supplementary Exercises, Ch. 5

EX 63 The following data shows body weight (kg) of 10 males before (*X*) and after (*Y*) participating in a training program with the purpose to reduce weight.

Subject	1	2	3	4	5	6	7	8	9	10
X	88.3	94.6	88.4	102.5	94.3	79.3	86.3	96.9	88.5	101.8
Y	88.1	93.5	88.5	102.0	94.7	78.5	86.1	96.2	88.2	101.1

a) Does the training program have a significant effect on weight-loss? Answer the question by drawing conclusion from a 95% CI for the average weight-loss.
[Hint: Just look at the differences within subjects. Don't use the approach in EX 53. Why?]
b) When the same training program was used by a population of females it was found that the variance of the weight-loss was 0.7. Does the latter value differ significantly from the variance obtained for males?
c) Give a 95% CI for the proportion of males that loses weight. Compare the results that are obtained by using the expressions in EX 55 a) and in (23).
d) As expected, the CI in c) becomes very wide. Consider the sample above as a pilot sample and determine the sample size needed to get a 95% CI with a Bound on the Error that is 0.025.

EX 64 Data below summarizes measurements of Area of Optic Disk (AOD) in mm^2 from two samples of children called FAS and Control. Children in the FAS (Fetal Alcoholic Syndrome) group had mothers who were high-consumers of alcohol during pregnancy.

	FAS	Control
Sample size	22	30
Mean	2.01	2.55
Variance	0.3623	0.2305

Determine a 95% CI for the difference of mean AOD between the two groups.

EX 65 Let $(Y_i)_{i=1}^n$ be iid where $Y_i \sim Exponential(\lambda)$.

a) Determine a 95% CI for λ based on the fact that $\left(\overline{Y} - E(\overline{Y})\right)/\sqrt{V(\overline{Y})} \xrightarrow{D} Z \sim N(0,1)$.
b) Compute the expected length of the CI in a) when $n = 50$. Compare the latter with the expected length of the CI in X 61 a)

EX 66 During an epidemic a sample of five institutions at a university was randomly selected. These were asked how many of their employees who were on the sick-list. The result was

Institution	1	2	3	4	5
Sick-listed	4	10	8	2	6
Total staff	10	42	25	11	12

Give a 95% CI for the total proportion sick-listed at the university.

[Hint: Use the ML estimator in EX 46 together with the CLT.]

EX 67 The number of bacteria (per cm^3) in a certain type of food varies according to a Poisson distribution. In a sample of 4 units one obtained the following result

Unit	1	2	3	4
Number of bacteria	103	112	91	117

Determine a 95% CI for the mean number of bacteria.

[Hint: Use the asymptotic normality of the Poisson distribution.]

6 Hypothesis Testing

In Ch. 5 we considered one way to claim statistical significance, namely to construct a CI for an unknown parameter. In this chapter we will meet another way to claim significance, by setting up hypotheses about parameters and to see if these are in accordance with data. There are mainly two ways to do this, the *p-value approach* and the *rejection region approach*. Both of these are described below.

6.1 Concepts

6.1.1 p-value approach

In the p-value approach a basic hypothesis, called the *null hypothesis* H_0, is formulated about one or several parameters. In the next step a statistic $T = T(Y_1 \ldots Y_n)$, called a *test statistic,* is chosen and the value taken by T in a specific sample determines whether H_0 shall be rejected or not. The precise way in which this is done is illustrated in the following example.

EX 68 Let *p* be the proportion of born boys in a certain population. We want to test the hypothesis $H_0 : p = 1/2$. To this end we take sample of *n* born boys and calculate the value of the test statistic $\hat{p} = X/n$, where X is the number of born boys in the sample. If the value of $\hat{p}$ deviates 'very much' from the value specified by H_0 we should reject H_0. But what is the meaning of 'very much', is e.g. $X = 7$ out of $n = 10$ enough?

For assistance in this matter we calculate the *p-value* = $P(X \geq 7 | p = 1/2)$ where it can be assumed that $X \sim Binomial(10,1/2)$. Thus (Cf. Ch. 2.2.1 *(2)*)

$$p-value = \binom{10}{7}(1/2)^7(1/2)^3 + \binom{10}{8}(1/2)^8(1/2)^2 + \binom{10}{9}(1/2)^9(1/2)^1 + \binom{10}{10}(1/2)^{10}(1/2)^0 =$$

$$176 \cdot (1/2)^{10} = 0.1719.$$

The latter is called a *one-sided(one-tailed)* p-value. But there is nothing *a priori* that says that a deviation from H_0 only goes in one direction in this case. We should therefor also calculate $P(X \leq 3 | p = 1/2) = 0.1719$. (The Binomial pf is symmetric for $p = 1/2$.)

The *two-sided* p-value is thus 0.1719+0.1719 =0.34. The latter is the probability of getting observed extreme deviations from H_0 by mere chance, and it is quite large.

Assume now that we instead have observed $X = 70$ out of $n = 100$. Since $\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \xrightarrow{D} Z \sim N(0,1)$, we can calculate a two-sided p-value in the following way:

$$P(\hat{p} \geq 0.70 | p = 1/2) = P\left(\frac{\hat{p} - 1/2}{\sqrt{1/400}} \geq \frac{0.70 - 1/2}{\sqrt{1/400}}\right) \approx P(Z \geq 4.00) = 0.00003 =$$

$$P(\hat{p} \leq 0.30 | p = 1/2)$$

Thus, p-value $= 2 \cdot 0.00003 = 0.00006$, which is very small.

A p-value is thus the probability of obtaining a value on the test statistic as least as extreme as the one that is observed, provided that H_0 holds. Some comments on this.

- A p-value should normally be two-sided. Exceptions are when it is obvious that deviations only can go in one direction. If you present one-sided p-values in a paper that you send to a scientific journal, it is likely that it will be returned since referees often wants two-sided p-values.
- It is customary to reject H_0 when the p-value is less than 0.05. Here is some frequently used terminology for this:

$0.05<p<0.10$	'Weak significance'	
$0.01<p<0.05$	'Significance'	*
$0.001<p<0.01$	'Strong significance'	**
$p<0.001$	'Very strong significance'	***

The concept of 'Weak significance' can be found in areas such as Psychology and Sociology where sometimes sample sizes are small and the p-values become large only for this reason. The use of stars, similar to classification of brandy, has been popular in medical studies, but should be avoided. It has actually happened that it has been confused with foot notes.

- A p-value expresses the degree of evidence against H_0 that is found *in the present sample* and nothing else. Hypotheses such as $p = 1/2$ for the proportion of heads when tossing a coin, or mean = 0 for the difference in means between two groups, can strictly speaking be rejected without data. (1/2 is not the same as 0.5 or something with more decimals, it is exactly one divided by two.) These hypotheses can always be rejected by choosing n sufficiently large. Consider a study conducted som years ago of the effect of physical activity upon on the risk of getting heart disease. An 'active´ group consisting of 30 000 subjects and a 'control' group of 20 000 subjects were followed in time and the proportion of heart diseases were reported in each group. In the study a p-value just below 0.05 was obtained for the hypothesis 'no difference between the proportion of heart disease in the two groups', Newspapers reported that it is now proved that physical activity has a statistically significant positive effect on the risk of heart disease. The author's personal reaction to this as a statistician is that, if such large amount of data were needed to get a p-value below 5%, then the true difference must be marginal.
- The p-value concept seems to have been first used by Laplace in the 1770s when studying the excess of born boys compared to girls. It was later popularized by R Fisher in the 1920s and he invented the term *test of significance* for this approach. It was later displaced by the rejection region approach, to be described in the next section. During the last years the p-value approach has regained its leading position. This is probably due to the rapid development of computer programs by means of which the computation of p-values is easy, something that wasn't the case 30–50 years ago. Today most statistical soft-ware supply their users with a variety of p-values, obtained by using various test statistics and under various assumptions. This in turn has increased the need for a higher statistical level of knowledge.

6.1.2 Rejection region approach

As before there is a null hypothesis H_0 and a test statistic T. Now there is furthermore an *alternative hypothesis* H_a and a *rejection region, RR,* such that if T takes a value within RR then H_0 is rejected and H_a is accepted. (RR is sometimes called a critical region.) By using the symbol $\in$ (belongs to) this can be expressed as 'Reject H_0'$\Leftrightarrow$'$T \in RR$' . To types of errors can be made in reaching a decision. A *type I error* is made if H_0 is rejected when H_0 is true. The probability of this event is denoted α and it is customary to require that $\alpha < 0.05$. A *type II error* is made if H_0 is accepted when H_a is true. The probability of the latter event is denoted β.

An important concept is that of a *power function*, which is the probability of rejecting H_0. If θ is the parameter that is specified by H_0, then the power is $Pow(\theta) = P(T \in RR)$. Under H_0: $\theta = \theta_0$, $Pow(\theta_0) = \alpha$. The latter equality is seldom possible to achieve when the test statistic has a discrete distribution and in that case it is required that $Pow(\theta) < \alpha$. In general the power depends on: (i) θ, (ii) the sample size n, (iii) the choice of RR and (iv) the choice of test statistic T. The *best test statistic* is the one that maximizes the power for given θ,n and RR. This is often based on the best estimator. (Stuart *et al* 1999, Ch. 22.36.)

EX 69 Consider the test statistic *Y* = 'Number of born boys' $\sim Binomial(n=10,\ p)$ that is used for testing $H_0: p=1/2$ against $H_a: p\neq 1/2$.

Compute the power for each of the RRs: $(i)\ \{0,10\}, (ii)\ \{0,1,9,10\}, (iii)\ \{0,1,2,8,9,10\}$. Suggest a proper RR.

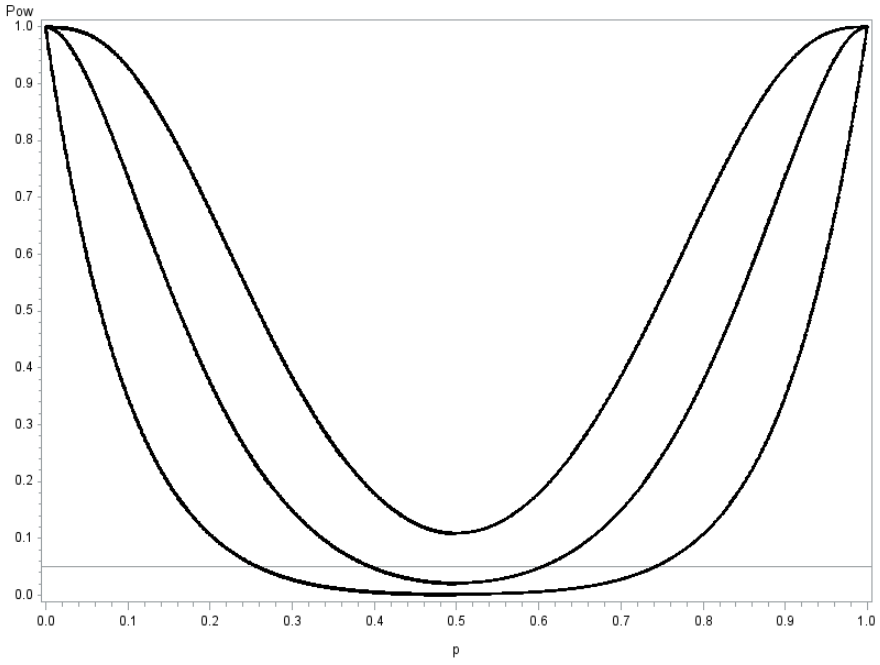
$$(i)\ Pow_1(p)=P(Y=0)+P(Y=10)=\binom{10}{0}p^0(1-p)^{10}+\binom{10}{10}p^{10}(1-p)^0=(1-p)^{10}+p^{10}.$$

$$(ii)\ Pow_2(p)=P(Y\le 1)+P(Y\ge 9)=Pow_1(p)+P(Y=1)+P(Y=9)=Pow_1(p)+$$

$$\binom{10}{1}p^1(1-p)^9+\binom{10}{9}p^9(1-p)^1=Pow_1(p)+10p(1-p)\left((1-p)^8+p^8\right).$$

$$(iii)\ Pow_3(p)=P(Y\le 2)+P(Y\ge 8)=Pow_2(p)+\binom{10}{2}p^2(1-p)^8+\binom{10}{8}p^8(1-p)^2=$$

$$Pow_2(p)+45p^2(1-p)^2\left((1-p)^6+p^6\right).$$



The three power curves are shown in Figure 1.

Of special interest is to compute the power under H_0, $\alpha_i = Pow_i(p=1/2)\ \ i=1,2,3.$

$\alpha_1=0.0062, \alpha_2=0.0480, \alpha_3=0.1796$. Since the latter value is larger than 0.05 we can't use the corresponding RR. The RR {0, 1, 9, 10} is to be preferred since it has a power that is less than 0.05 under the null hypothesis and it is constantly larger than the power in (*i*).

6.2 Methods of finding tests

In Ch. 4.3 some methods for finding point estimators were presented. We now consider some methods that can guide us when testing hypotheses. These are based on the Chi-square principle, the Likelihood-ratio (LR) principle and on miscellaneous methods.

6.2.1 The Chi-square principle

This requires that data are classified. For measurements on a continuous variable we thus have to create classes. How this can be done is illustrated in EX 109 below. We thus have the following data

Class	1	2	…	*k*	Total
Observed frequency	Y_1	Y_2	$\cdots$	Y_k	$\sum Y_i = n$
Hypothetical probability	p_1	p_2	$\cdots$	p_k	$\sum p_i = 1$

Here the hypothetical probability p_i is the probability of belonging to class i under H_0. Examples of such probabilities are given in the examples below. In general the null hypothesis can be formulated $H_0 : p_i = p_i(\theta_1, \theta_2, \ldots)$, where $\theta_1, \theta_2, \ldots$ are unknown parameters that need to be estimated (by the ML method) giving $\hat{\theta}_1, \hat{\theta}_2, \ldots$. The Chi-square statistic is

$$X^2 = \sum_{i=1}^{k} \frac{\left(Y_i - np_i(\hat{\theta}_1, \hat{\theta}_2, \ldots | H_0)\right)^2}{np_i(\hat{\theta}_1, \hat{\theta}_2, \ldots | H_0)} \quad (25)$$

As $n \to \infty$, $X^2 \xrightarrow{D} \chi^2(k-1-a)$ under H_0. a is the number of linearly independent parameters to estimate.

In practice (25) is used in the following way: Compute the value of X^2, giving X^2_{OBS}. Then calculate the p-value $P\left(\chi^2(k-1-a) > X^2_{OBS}\right)$ and reject H_0 if the latter is smaller than 0.05.

EX 70 Given the following data

Class	1	2	Total
Observed frequency	Y_1	Y_2	n
Probability	1-p	p	1

Test $H_0 : p = 1/2$ against $H_1 : p \neq 1/2$.

There are no parameters to estimate under H_0.

$X^2 = \frac{(Y_1 - n \cdot 1/2)^2}{n \cdot 1/2} + \frac{(Y_2 - n \cdot 1/2)^2}{n \cdot 1/2}$ gives X^2_{OBS}. p-value = $P(\chi^2(2-1-0) > X^2_{OBS})$. Notice that X^2 can also be written $\left(\frac{Y_2/n - 1/2}{\sqrt{1/4n}}\right)^2$.

Contingency tables (R × C cross tables).

The following frequency table, often called a 2 × 2 table, is a convenience way to summarize data.

		Factor II		
		1	2	Total
Factor I	1	Y_{11}	Y_{12}	Y_{1+}
	2	Y_{21}	Y_{22}	Y_{2+}
	Total	Y_{+1}	Y_{+2}	n

Here there are two 'Factors', each divided into two categories. Examples are when the factors are two doctors who classify the same n patients as either 'Healthy' (1) or 'Diseased' (2), or when the political opinion (left- or right wing) of n voters is measured at two times (Factor I and II). The table can be generalized to a R × C table with R rows and C columns. It can also be extended to more than two factors, e.g. when the political opinion of n voters about P >2 parties is measured at T > 2 times. In such a case the sample is called a *Panel*. Notice that all cell-frequencies are random, except for the fixed sample size n.

Corresponding to the 2 × 2 table of frequencies, there is a 2 × 2 table of probabilities

		Factor II		
		1	2	Total
Factor I	1	p_{11}	p_{12}	p_{1+}
	2	p_{21}	p_{22}	p_{2+}
	Total	p_{+1}	p_{+2}	1

The 'Total' probabilities are called *marginal proportions (probabilities)*. Notice that there are three genuine (or linearly independent) parameters p_{ij} and two genuine marginal proportions, since the sum equals 1. In the R × C table there are $R \cdot C - 1$ genuine parameters p_{ij} and $R - 1 + C - 1$ genuine marginal proportions.

The frequencies in a 2 x 2 table are distributed *Multinomid* $(n, p_{11}, p_{12}, p_{21}, p_{22})$ (Cf. Ch. 2.2.1 *(6)*), so the probability of the outcomes, or Likelihood if we are interested in the parameters, can be written $L = (const.) p_{11}^{y_{11}} p_{12}^{y_{12}} p_{21}^{y_{21}} p_{22}^{y_{22}} \cdot$ We will consider two types of hypotheses:

Equality of marginal proportions. $H_0 : p_{1+} = p_{+1}$. This is easily seen to be identical with $H_0 : p_{21} = p_{12} (= p)$. $(= p)$. Under H_0 there are 2 genuine parameters to estimate.

Independency between Factor I and Factor II. In this case $H_0 : p_{ij} = p_{i+} \cdot p_{+j}$. Under H_0 there are 2 genuine parameters to estimate.

EX 71 *Test of equal marginal proportions in the 2x2 table by means of McNemar's test.*

As just noticed we shall test $H_0 : p_{1+} = p_{+1} \Leftrightarrow H_0 : p_{21} = p_{12} (= p)$. Let the Likelihood under H_0 be denoted L_0. Then

$L_0 = (const.) p_{11}^{y_{11}} p^{y_{21}+y_{12}} (1 - p_{11} - 2p)^{y_{22}}$. (Notice that we only keep the genuine parameters.)

$$\ln L_0 = \ln(const.) + y_{11} \ln p_{11} + (y_{21} + y_{12}) \ln p + y_{22} \ln(1 - p_{11} - 2p)$$

$$\frac{d \ln L_0}{dp_{11}} = y_{11} \frac{1}{p_{11}} + y_{22} \left(\frac{-1}{(1 - p_{11} - 2p)} \right) = 0, \quad \frac{d \ln L_0}{dp} = (y_{21} + y_{12}) \frac{1}{p} + y_{22} \left(\frac{-2}{(1 - p_{11} - 2p)} \right) = 0$$

We thus have the equations

$$\begin{cases} y_{11} / p_{11} = y_{22} / (1 - p_{11} - 2p) & (1) \\ (y_{21} + y_{12}) / p = 2 y_{22} / (1 - p_{11} - 2p) & (2) \end{cases}$$

Since these may be somewhat tricky to solve we show an example of a solution.

Multiplying Eq. (1) by 2 and taking the difference between the left and right-hand sides in the two equations yields

$(y_{21} + y_{12}) / p - 2 y_{11} / p_{11} = 0 \Rightarrow$

$p_{11} = 2 p y_{11} / (y_{21} + y_{12})$ which inserted into Eq. (1) gives $\frac{y_{21} + y_{12}}{2p} = \frac{y_{22}}{1 - 2p y_{11} / (y_{21} + y_{12}) - 2p}$.

So, $\frac{(y_{21} + y_{12})}{2} = p(y_{11} + y_{21} + y_{12} + y_{22}) = p \cdot n \Rightarrow \hat{p} = \frac{(y_{21} + y_{12})}{2n}$ and this inserted into Eq. (1) gives

$\hat{p}_{11} = y_{11} / n$ and finally $\hat{p}_{22} = 1 - \hat{p}_{11} - 2\hat{p} = y_{22} / n$.

According to the Chi-square principle. $X^2 = \sum \frac{\left(Y_{ij} - n \cdot (\hat{p}_{ij} | H_0) \right)^2}{n \cdot (\hat{p}_{ij} | H_0)}$. Here

$$Y_{11} - n \cdot \left(\hat{p}_{11} | H_0 \right) = Y_{11} - n \cdot \frac{Y_{11}}{n} = 0, \qquad Y_{12} - n \cdot \left(\hat{p}_{12} | H_0 \right) = Y_{12} - n \cdot \frac{(Y_{21} + Y_{12})}{2n} = \frac{(Y_{12} - Y_{21})}{2},$$

$$Y_{21} - n \cdot \left(\hat{p}_{21} | H_0 \right) = Y_{21} - n \cdot \frac{(Y_{21} + Y_{12})}{2n} = \frac{(Y_{21} - Y_{12})}{2}, \qquad Y_{22} - n \cdot \left(\hat{p}_{22} | H_0 \right) = Y_{22} - n \cdot \frac{Y_{22}}{n} = 0.$$

Thus

$$X^2 = 0 + \frac{\left((Y_{12} - Y_{21}) / 2 \right)^2}{(Y_{21} + Y_{12}) / 2} + \frac{\left((Y_{21} - Y_{12}) / 2 \right)^2}{(Y_{21} + Y_{12}) / 2} + 0 = (!) = \frac{\left(Y_{12} - Y_{21} \right)^2}{\left(Y_{12} + Y_{21} \right)}.$$

This test statistic was derived by McNemar (McNemar 1947, p. 153) and has been termed *McNemar's Test*.

Under H_0 the statistic is distributed $\chi^2(4 - 1 - 2) = \chi^2(1)$ in large samples with p-value = $P\left(\chi^2(1) > X^2_{OBS} \right)$

EX 72 *Test of independency*

In the 2 x 2 table the hypothesis of independency is $H_0 : p_{ij} = p_{i+} \cdot p_{+j}$, for $i, j = 1,2$. Under H_0 the likelihood is $L_0 = (const.)(p_{1+} \cdot p_{+1})^{y_{11}}(p_{1+} \cdot p_{+2})^{y_{12}}(p_{2+} \cdot p_{+1})^{y_{21}}(p_{2+} \cdot p_{+2})^{y_{22}}$. Since we only want genuine parameters, put $p_{2+} = 1 - p_{1+}$ and $p_{+2} = 1 - p_{+1}$. Taking logarithms gives

$$\ln L_0 = \ln(const.) + y_{11}\left(\ln p_{1+} + \ln p_{+1}\right) + y_{12}\left(\ln p_{1+} + \ln(1 - p_{+1})\right) + y_{21}\left(\ln(1 - p_{1+}) + \ln p_{+1}\right) + y_{22}\left(\ln(1 - p_{1+}) + \ln(1 - p_{+1})\right).$$

$$\frac{d \ln L_0}{dp_{1+}} = \frac{y_{11}}{p_{1+}} + \frac{y_{12}}{p_{1+}} - \frac{y_{21}}{(1 - p_{1+})} - \frac{y_{22}}{(1 - p_{1+})} = \frac{y_{11} + y_{12}}{p_{1+}} - \frac{(y_{21} + y_{22})}{(1 - p_{1+})} = \frac{y_{1+}}{p_{1+}} - \frac{y_{2+}}{(1 - p_{1+})} = 0$$

$$\frac{d \ln L_0}{dp_{+1}} = \frac{y_{11}}{p_{+1}} - \frac{y_{12}}{(1 - p_{+1})} + \frac{y_{21}}{p_{+1}} - \frac{y_{22}}{(1 - p_{+1})} = \frac{y_{11} + y_{21}}{p_{+1}} - \frac{(y_{12} + y_{22})}{(1 - p_{+1})} = \frac{y_{+1}}{p_{+1}} - \frac{y_{+2}}{(1 - p_{+1})} = 0$$

From this we get

$y_{1+} - y_{1+}p_{1+} = y_{2+}p_{1+} \Rightarrow p_{1+} = \frac{y_{1+}}{(y_{1+} + y_{2+})} = \frac{y_{1+}}{n}$ and similarly $\hat{p}_{2+} = \frac{y_{2+}}{n}$. Also notice that

$\hat{p}_{+1} = 1 - \hat{p}_{1+} = 1 - \frac{y_{1+}}{n} = \frac{n - y_{1+}}{n} = \frac{y_{2+}}{n}$ and similarly $\hat{p}_{+2} = \frac{y_{+2}}{n}$

In the Chi-square statistic (25) $n \cdot \left(\hat{p}_{ij} \middle| H_0\right) = n \cdot \left(\hat{p}_{i+}\hat{p}_{+j}\right) = \frac{Y_{i+} \cdot Y_{+j}}{n}$. Thus we obtain the statistic

$$X^2 = \sum_{i,j=1}^{2} \frac{\left(Y_{ij} - Y_{i+} \cdot Y_{+j} / n\right)^2}{Y_{i+} \cdot Y_{+j} / n}.$$

In large samples this is distributed $\chi^2(4 - 1 - 2) = \chi^2(1)$ and the p-value is $P\left(\chi^2(1) > X^2_{OBS}\right)$.

In the table with *R* rows and *C* columns the Chi-square statistic remains the same, but now the degrees of freedom is changed to $R \cdot C - 1 - (R - 1) - (C - 1) = (R - 1)(C - 1)$. The p-value is now obtained as $P\left(\chi^2((R - 1)(C - 1)) > X^2_{OBS}\right)$.

When the hypothesis of independence between the two factors is rejected, one should go further in the analysis and determine which combination of levels from the factors that contributes to the dependency. This can be done by considering $D_{ij} = Y_{ij} - Y_{i+} \cdot Y_{+j} / n$. The latter is called *Deviation* and is supplied by many statistical soft-wares. If D_{ij} is greater than zero or below zero there is an over-or underrepresentation, respectively, of observations in cell (i, j) . Since deviations may be due merely to chance, one should study whether the deviation *is significantly different from zero* (a 'significant deviation'). A statistic for this purpose is the Cell Chi-square defined by

$$X^2_{ij} = \frac{\left(Y_{ij} - Y_{i+} \cdot Y_{+j} / n\right)^2}{Y_{1+} \cdot Y_{+j} / n}$$

In large samples this is distributed as a $\chi^2(1)$ variable (Cochran 1954, p. 417). This means that if the Cell Chi-square is larger than 3.85, then the deviation is significant at the 5% level. The single cell statistic is supplied by statistical soft-ware, e.g. in SAS where it is denoted 'Cell Chi-Square'.

When analyzing deviations in many cells one should be aware of the risk of making wrong decisions due to *the multiple inference* context. When several conclusions are to be drawn simultaneously with 5% significance one has to adjust the individual significance level so that the global level is maintained at 5%. This is explained further in Ch. 6.4.

We now turn to another application of the chi-square principle, the *test of fit*. In this case the null hypothesis specifies that data have a certain distribution. In (25) Y_i are the observed frequencies which are to be compared with the hypothetical ones under H_0.

EX 73 *A test of randomness for binary data.*

A sequence of digital numbers starts with 0,0,0,0,1,1,… and ends with …,0,0,1,0,0,1. We want to study whether these occur in a random order. There are several ways to do this, but one is the following: Define the variable *Y* = 'Number of digits until the first '1' occurs. The observations on *Y* are

Y	1	2	3	4	5	6	10
Frequency	26	13	9	2	1	1	1

From Ch. 2.2.1 *(3)* it follows that in this case H_0: 'Digits are in random order' is the same as $H_0 : Y \sim Geometric(p)$, where *p* is the probability of '1'.

In EX 47 it is seen that the ML estimate of *p* is $\hat{p} = 1/\bar{y}$. From the table above we get

$$\bar{y} = \frac{\sum y_i}{n} = \frac{26 \cdot 1 + 13 \cdot 2 + \ldots + 1 \cdot 10}{26 + 13 + \ldots + 1} = \frac{108}{53} \Rightarrow \hat{p} = \frac{53}{108} = 0.49 .$$

The estimated expected frequency under H_0 of the outcome $'Y = y'$ is $n \cdot (1 - \hat{p})^y \hat{p}, y = 1,2,...$ E.g. the expected frequency of the outcome $'Y = 2'$ is $53 \cdot (1 - 0.49)^{2-1} 0.49 = 13.2$. One obtains the following table

y	1	2	3	4	5
Expected frequency	26.0	13.2	6.8	3.4	1.8

Here the expected frequencies of the outcomes *Y* = 5 or larger are small so we throw them together in the following way: 53-(26.0+13.2+6.8+3.4) = 3.6. We now get the table

y	1	2	3	4	5-
Expected frequency	26.0	13.2	6.8	3.4	3.6
Observed frequency	26	13	9	2	3

$$X^2_{OBS} = \frac{(26 - 26.0)^2}{26.0} + \ldots + \frac{(3 - 3.6)^2}{3.6} = 1.39 \Rightarrow \text{p - value} = P\left(\chi^2(5 - 1 - 1) > 1.39\right) >> 0.10 .$$

There is thus no reason to reject the hypothesis of randomness.

Comment to EX 73 It has been recommended that expected frequencies under H0 shall be larger than 2 in the Chi-square test of fit (Stuart et al 1999, p. 409), earlier recommendations were larger than 5. In EX 73 all expected frequencies for y larger than 4 are definitely too small. At *y* = 3 there is an over-representation of observed frequencies with *Deviation* = 2.2, but this isn't serious since the cell-Chi-square statistic is only 0.71.

EX 74 *Challenge the computer in 'thinking randomly'.*

The original series in EX 73 was actually made by a random number generator. (The function ranbin(0,1,p) with p = 1/2 in SAS.)

It is a challenge to try to beat the computer in 'thinking randomly', as measured by the value of X^2_{OBS}.
Write down a sequence of slightly more than one hundred 0's and 1's. Try to place them in 'random' order, and repeat the analysis made in EX 73. You will probably not beat the computer and it is even likely that the sequence you have created will be rejected as random.

A tip! By the time you will learn from the table of observed and expected frequencies how to improve your skill. Then it is time to challenge your friends in tournaments.

EX 75 The following table shows the treatment times (minutes) for patients at a clinic.

Treatment time	0-10	10-20	20-30	30-40	40-	Total
Frequency	10	16	13	6	5	50

Mean = 20, Variance = 140, Max.value = 45

Test whether the treatment times have a $Uniform(b)$ - distribution.

The cdf is $F(y) = y/b,\ 0 \le y \le b$ and a ML estimate of *b* is $\hat{b} = \frac{(n+1)}{n} y_{(n)} = \frac{51}{50} \cdot 45 = 45.9$.

Thus, the estimated cdf is $\hat{F}(y) = y/45.9$. The expected frequencies are.

$50 \cdot \hat{P}(0 \le Y \le 10) = 50\left(\hat{F}(10) - \hat{F}(0)\right) = 10.9,\ \ 50 \cdot \hat{P}(10 \le Y \le 20) = 50\left(\hat{F}(20) - \hat{F}(10)\right) = 10.9$
$50 \cdot \hat{P}(20 \le Y \le 30) = 50\left(\hat{F}(30) - \hat{F}(20)\right) = 10.9, 50 \cdot \hat{P}(30 \le Y \le 40) = 50\left(\hat{F}(40) - \hat{F}(30)\right) = 10.9$
$50 \cdot \hat{P}(Y \ge 40) = 50\left(1 - \hat{F}(40)\right) = 6.4$. Thus,

Treatment time	0–10	10-20	20–30	30–40	40–
Expected frequency	10.9	10.9	10.9	10.9	6.4
Observed frequency	10	16	13	6	5

$X^2_{OBS} = \frac{(10-10.9)^2}{10.9} + \ldots + \frac{(5-6.4)^2}{6.4} = 5.37$, p-value = $= P\left(\chi^2(5-1-1) > 5.37\right) = 0.15.$

This p-value isn't small enough to reject the hypothesis of a $Uniform(b)$ -distribution. However, the sample size is small and there are some suspicious signs of a positive deviation for the cell 10–20 and a negative deviation for the cell 30–40 (although neither being significant). There seems to be reasons to search for a more realistic probability model.

EX 76 Check if a $Gamma(\lambda,k)$ - distribution gives a better fit to the data in EX 75.

ML estimates of this distribution are quite laborious to obtain (Cf. EX 44), therefore we confine ourselves with Moment estimates (Cf. EX 38) $\hat{\lambda} = \bar{y}/s^2 = 1/7$ and $\hat{k} = \bar{y}^2/s^2 = 2.9 \approx 3$.

These estimates inserted into the cdf (Cf. Ch.2.2.2 *(2)*) $F(y) = 1 - e^{-\lambda y}\sum_{i=0}^{k-1}(\lambda y)^i/i!$ gives

$\hat{F}(y) = 1 - e^{-y/7}\left(1 + x/7 + (x/7)^2/2\right)$ From this the expected frequencies are

$50 \cdot P(0 < Y < 10) = 8.7,\quad 50 \cdot P(10 < Y < 20) = 18.6,\quad 50 \cdot P(20 < Y < 30) = 12.9$

$50 \cdot P(30 < Y < 40) = 6.2,\quad 50 \cdot P(Y > 40) = 3.8$. Thus,

Treatment time	0–10	10–20	20–30	30–40	40–
Expected frequency	8.7	18.6	12.9	6.2	3.8
Observed frequency	10	16	13	6	5

$X^2_{OBS} = \frac{(10-8.7)^2}{8.7} + \ldots + \frac{(5-3.8)^2}{3.8} = 0.90$, p-value $= P\left(\chi^2(5-1-2) > 0.90\right) = 0.64.$

The gamma distribution seems to give a much better fit to data than the uniform distribution.

EX 77 The production of goods by a particular method has since a long time resulted in 25% god, 62% medium and 13% bad products. In a test with a new method 50 products were produced. Of these 20 were god, 19 were medium and 11 were bad.

Does the new method give products of a different quality or are the observed differences merely due to chance?

H_0: The new method gives products of the same quality as the older method.

Quality	God	Medium	Bad
Expected frequency	$0.25 \cdot 50 = 12.5$	$0.62 \cdot 50 = 31.0$	$0.13 \cdot 50 = 6.5$
Observed frequency	20	19	11

$X^2_{OBS} = \frac{(20-12.5)^2}{12.5} + \frac{(19-31.0)^2}{31.0} + \frac{(11-6.5)^2}{6.5} = 12.3$, p-value= $= P\left(\chi^2(3-1-0) > 12.3\right) = 0.002$.

(No parameters have been estimated.) There is thus a strong reason to reject H_0.

Let's look more closely at the differences.

Quality	God	Medium	Bad
Deviation	+7.5	-12.0	+4.5
Cell-Chi-square	4.50 p<5%	4.65 p<5%	3.12 (NS)

The total Chi-square of 12.3 above shows that *there is a significant difference*. The table of deviations *explains* in a way the nature of the difference. The new method involves an over-representation of god products and an under-representation of medium products (NS is an often used abbreviation for 'Not Significant'.)

6.2.2 The Likelihood Ratio principle

To test the hypothesis $H_0 : (\theta_1, \theta_2, \ \ldots) = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots)$ against $H_a : (\theta_1, \theta_2, \ldots) \neq (\theta_1^{(0)}, \theta_2^{(0)}, \ldots)$ we consider the Likelihood Ratio (LR) statistic $\Lambda = \hat{L}_0 / \hat{L}$. Here $\hat{L}_0$ is the likelihood under H_0 with ML estimators inserted for the parameters. $\hat{L}$ is the correspond likelihood under both H_0 and H_1, i.e. without any restrictions on the parameters. An obvious rejection region (RR) is $\Lambda < c$, or $-\ln \Lambda > c'$. (This follows because $0 < \Lambda < 1$.) The LR test is performed in the following way:

1. Compute the ML estimates of the parameters under H_0 and under H_1 (26)
2. Compute the value of the LR statistic, say Λ_{OBS}.
3. Compute the p-value for H_0 from the p-value $= P\left(\chi^2(r-s) > -2\ln\Lambda_{OBS}\right)$, where r = Number of parameters estimated without restrictions on the parameters and s = Number of parameters estimated under H_0.

Notice that this test is a large-sample test. (Strictly speaking the test should be termed estimated LR test (ELR), since estimates are plugged in for the parameters.)

EX 78

a) Make a LR test based on the fictive data in EX 70 to test H_0: p =1/2 against H_a: $p \neq 1/2$.
b) Compare the LR and the Chi-square tests when $n = 100$ and $Y_2 = 60$.

a) The likelihood is $L = c \cdot p^y (1-p)^{n-y}$, where $c = \binom{n}{y}$. Under H_0 we get

$L_0 = c \cdot (1/2)^y (1/2)^{n-y}$. (There are no parameters to estimate.) The unrestricted ML estimate of p is

$$\hat{p} = y/n \Rightarrow \hat{L} = c \cdot (y/n)^y (1-y/n)^{n-y} \Rightarrow \Lambda = \frac{L_0}{\hat{L}} = \frac{c \cdot (1/2)^y (1/2)^{n-y}}{c \cdot (y/n)^y (1-y/n)^{n-y}} =$$

$$\frac{1}{(2y/n)^y (2(1-y/n))^{n-y}} \Rightarrow -2\ln\Lambda = 2\left(y\ln(2y/n) + (n-y)\ln[2(1-y/n)]\right).$$

$$\text{p-value} = P\left(\chi^2(1-0) > -2\ln\Lambda_{OBS}\right).$$

b) *Chi-square test*

$$X^2 = \frac{(60-100\cdot 1/2)^2}{100\cdot 1/2} + \frac{(40-100\cdot 1/2)^2}{100\cdot 1/2} = 4.00 \Rightarrow \text{p - value} = P\left(\chi^2(1) > 4.00\right) = 0.0455.$$

LR test

$$-2\ln\Lambda = 2\left(60\ln(2\cdot 60/100) + (100-60)\ln[2(1-60/100)]\right) = 4.03 \Rightarrow \text{p - value} =$$

$$P\left(\chi^2(1) > 4.03\right) = 0.0447.$$

The two p-values are roughly the same. In practice it will suffice to just notice that the p-values are below 5%, so H_0 is rejected at the 5% level.

EX 79 *Test of equality between two proportions in independent Binomial samples*.

Often one is interested in comparing two proportions, e.g. the proportion of smokers among men and women. In such a case one takes a sample of men and a sample of women that is independent of the first sample. (A sample of couples is thus not appropriate.) Data can be summarized in the following way:

	Sample 1 (Men)		Sample 2 (Women)		
	Frequency	Probability	Frequency	Probability	Total
Smokers	Y_1	p_1	Y_2	p_2	$Y_1 + Y_2$
Non-smokers	$n_1 - Y_1$	$1 - p_1$	$n_2 - Y_2$	$1 - p_2$	$n_1 + n_2 - (Y_1 + Y_2)$
Total	n_1	1	n_2	1	$n_1 + n_2$

We want to test H_0: $p_1 = p_2 (= p)$ against H_a: $p_1 \neq p_2$

The unrestricted likelihood is $L = \binom{n_1}{y_1} p_1^{y_1} (1-p_1)^{n_1 - y_1} \binom{n_2}{y_2} p_2^{y_2} (1-p_2)^{n_2 - y_2}$ with two parameters to estimate.

The likelihood under H_0 is $L_0 = \binom{n_1}{y_1}\binom{n_2}{y_2} p^{y_1+y_2} (1-p)^{n_1+n_2-(y_1+y_2)}$ with one parameter to estimate.

$$\ln L = const. + y_1 \ln p_1 + (n_1 - y_1)\ln(1-p_1) + y_2 \ln p_2 + (n_2 - y_2)\ln(1-p_2) \Rightarrow$$

$$\frac{d \ln L}{dp_1} = \frac{y_1}{p_1} - \frac{(n_1 - y_1)}{(1-p_1)} = 0 \Rightarrow \hat{p}_1 = \frac{y_1}{n_1}, \qquad \frac{d \ln L}{dp_2} = \frac{y_2}{p_2} - \frac{(n_2 - y_2)}{(1-p_2)} = 0 \Rightarrow \hat{p}_2 = \frac{y_2}{n_2}$$

$$\ln L_0 = const. + (y_1 + y_2)\ln p + (n_1 + n_2 - (y_1 + y_2))\ln(1-p) \Rightarrow$$

$$\frac{d \ln L_0}{dp} = \frac{y_1 + y_2}{p} - \frac{\left(n_1 + n_2 - (y_1 + y_2)\right)}{(1-p)} = 0 \Rightarrow \hat{p} = \frac{y_1 + y_2}{n_1 + n_2}$$

$$\Lambda = \frac{\left(\frac{y_1 + y_2}{n_1 + n_2}\right)^{y_1+y_2} \left(1 - \frac{(y_1 + y_2)}{(n_1 + n_2)}\right)^{n_1+n_2-(y_1+y_2)}}{\left(\frac{y_1}{n_1}\right)^{y_1} \left(1 - \frac{y_1}{n_1}\right)^{n_1 - y_1} \left(\frac{y_2}{n_2}\right)^{y_2} \left(1 - \frac{y_2}{n_2}\right)^{n_2 - y_2}}$$

The p-value is finally obtained from $P\left(\chi^2(2-1) > -2\ln\Lambda_{OBS}\right)$

EX 80 *LR test of equality between marginal proportions.*

a) Show how the LR test can be used to test H_0: $p_{21} = p_{12}(= p)$:in the 2 x 2 table.
b) Perform the test in the following case where two methods A and B are used to classify the same products into two categories 1 and 2.

		Method B	
		1	2
Method A	1	32	19
	2	21	18

c) Test the same hypothesis by using a Chi-square test.

a) From EX 71 the estimated likelihood under H_0 is $L_0 = c \cdot \hat{p}_{11}^{y_{11}} \hat{p}^{y_{21}+y_{12}} \hat{p}_{22}^{y_{22}}$, where the estimates are $\hat{p}_{11} = \frac{y_{11}}{n}, \hat{p} = \frac{y_{21}+y_{12}}{2n}, \hat{p}_{22} = \frac{y_{22}}{n}$. The unrestricted estimated likelihood is $L = c \cdot \hat{p}_{11}^{y_{11}} \hat{p}_{21}^{y_{21}} \hat{p}_{12}^{y_{12}} \hat{p}_{22}^{y_{22}}$, where the estimates are $\hat{p}_{ij} = \frac{y_{ij}}{n},\ i,j = 1,2.$ The LR ratio is $\Lambda = \frac{\hat{p}^{y_{21}+y_{12}}}{\hat{p}_{21}^{y_{21}} \hat{p}_{12}^{y_{12}}} \Rightarrow -2\ln\Lambda = -2\{(y_{21}+y_{12})\ln\hat{p} - y_{21}\ln\hat{p}_{21} - y_{12}\ln\hat{p}_{12}\}$. p-value= $P\left(\chi^2(3-2) > -2\ln\Lambda_{OBS}\right)$

b) $-2\ln\Lambda_{OBS} = 3.9729 \Rightarrow \text{p - value} = P\left(\chi^2(1) > 3.9729\right) = 0.0462$. H_0 is thus rejected at the 5% level. This conclusion can of course also be reached by noticing that the RR consists of values larger than $3.8416 = (1.96)^2$.
c) The Chi-square test is based on the statistic $X^2 = \frac{(y_{21}-y_{12})^2}{y_{21}+y_{12}}$. In this case $X^2_{OBs} = 3.92$ giving the p-value $P\left(\chi^2(1) > 3.92\right) = 0.0472$, very close to that obtained by the LR test.

EX 81 *LR test of independency.*

a) Show how the LR test can be used to test H_0: $p_{ij} = p_{i+}p_{+j}$ (independence) in the 2 x 2 table.
b) Apply the test to the following data illustrating the relation between left/right-handedness and type of twin (identical = 1, fraternal = 2).

	Type of twin	
	1	2
Right-handed	207	228
Left-handed	41	18

c) Test the same hypothesis with the Chi-square test.

a) Under H_0 the likelihood is $L_0 = c \cdot (\hat{p}_{1+} \cdot \hat{p}_{+1})^{y_{11}} (\hat{p}_{1+} \cdot \hat{p}_{+2})^{y_{12}} (\hat{p}_{2+} \cdot \hat{p}_{+1})^{y_{21}} (\hat{p}_{2+} \cdot \hat{p}_{+2})^{y_{22}}$,

where $\hat{p}_{i+} = \frac{y_{i1} + y_{i2}}{n}$, i $= 1,2$ and $\hat{p}_{+j} = \frac{y_{1j} + y_{2j}}{n}$, $j = 1,2$.

The unrestricted likelihood is $L = c \cdot \hat{p}_{11}^{y_{11}} \hat{p}_{12}^{y_{12}} \hat{p}_{21}^{y_{21}} \hat{p}_{22}^{y_{22}}$, where $\hat{p}_{ij} = \frac{y_{ij}}{n}$.

The LR statistic is $\Lambda = \frac{L_0}{L}$ and the p-value is $P\left(\chi^2(3-2) > -2\ln\Lambda_{OBS}\right)$. In this case the likelihood ratio can't be simplified as in EX 78.

b) After laborious computations we obtain $-2\ln\Lambda_{OBS} = 10.21$ and p-value is $P\left(\chi^2(1) > 10.21\right) = 0.0014$.

3. $X^2_{OBS} = 9.97$ and p-value is $P\left(\chi^2(1) > 9.97\right) = 0.0016$. The following table is obtained for deviations: *Deviation/Cell Chi-square*

	Type of twin	
	1	2
Right-handed	-11.4/0.59	11.4/0.60
Left-handed	11.4/4.37	-11.4/4.41

From the table it is concluded that the rejection of the null hypothesis is mainly due to a significant over-representation of identical twins (1) that are left-handed.

Comment to EX 81 The LR test for independence is laborious compared with the Chi-square test. The latter is also more informative since the source of the total Chi-square can be explained in terms of separate deviations. Today most statistical software present results for both tests, so ease of calculation is not a problem. It may be tempting to report the p-value that is lowest and this seems often to be the one obtained from the LR test, but in that case you lose the informative aspect mentioned above.

The p-values in EX 81 have been reported with too many decimals just to illustrate differences.

EX 82 In EX 62 a CI for the ratio of two Poisson rates λ_X and λ_Y was used to claim a significant difference between the rates. We now show how the same problem can be solved by LR testing.

Consider the hypothesis H_0: $\lambda_X = \lambda_Y (= \lambda)$ against H_a: $\lambda_X \neq \lambda_Y$.

Since the two samples are independent (Cf. EX 62) the unrestricted likelihood is

$L = \frac{(\lambda_X \cdot s)^{x(s)}}{x(s)!} e^{-\lambda_X \cdot s} \cdot \frac{(\lambda_Y \cdot t)^{y(t)}}{y(t)!} e^{-\lambda_Y \cdot t}$ and the likelihood under H_0 is

$L_0 = \frac{\lambda^{x(s)+y(t)} s^{x(s)} t^{y(t)} e^{-\lambda(s+t)}}{x(s)!\, y(t)!}$. From this the following estimates are easily obtained:

$\hat{\lambda}_X = \frac{x(s)}{s}$, $\hat{\lambda}_Y = \frac{y(t)}{t}$, $\hat{\lambda} = \frac{x(s)+y(t)}{s+t}$.

The estimated LR reduces to $\Lambda = \frac{\hat{L}_0}{\hat{L}} = \frac{\hat{\lambda}^{x(s)+y(t)}}{\hat{\lambda}_X^{x(s)} \cdot \hat{\lambda}_Y^{y(t)}}$, since many factors cancel each other. The p-value is

$P\left(\chi'^2(2-1) > -2\ln \Lambda_{OBS}\right)$.

In EX 62, $s = 8, t = 4, x(8) = 85, y(t) = 65 \Rightarrow \hat{\lambda} = \frac{85+65}{8+4} = 12.5, \hat{\lambda}_X = \frac{85}{8} = 10.63, \hat{\lambda}_Y = \frac{65}{4} = 16.25$.

This gives $-2\ln\Lambda = 6.4791 \Rightarrow \text{p - value} = 0.0109$. The null hypothesis is rejected.

Ex 83 $\left(Y_i\right)_{i=1}^n$ are iid with $Y_i \sim N(\alpha+\beta x_i, \sigma^2)$. Show how we can test H_0: $\beta = 0$ against H_a: $\beta \neq 0$.

(This model is the same as in EX 54 with the exception that there is a further parameter $\alpha \neq 0$.)

The unrestricted likelihood is $L = \frac{1}{(2\pi\cdot\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum(y_i-\alpha-\beta x_i)^2} \Rightarrow \ln L = c - \frac{n}{2}\ln\sigma^2 - \frac{\sum(y_i-\alpha-\beta x_i)^2}{2\sigma^2} \Rightarrow$

$$\frac{d\ln L}{d\alpha} = \frac{-(-1)\cdot 2\sum(y_i-\alpha-\beta x_i)}{2\sigma^2} = 0 \Rightarrow \sum y_i = n\cdot\hat{\alpha} + \hat{\beta}\sum x_i,\ \bar{y} = \hat{\alpha} + \hat{\beta}\cdot\bar{x} \quad \text{(i)}$$

$$\frac{d\ln L}{d\beta} = -\frac{(-1)\cdot 2\sum x_i(y_i-\alpha-\beta x_i)}{2\sigma^2} = 0 \Rightarrow \sum x_i y_i = \hat{\alpha}\sum x_i + \hat{\beta}\sum x_i^2, \frac{\sum x_i y_i}{n} =$$

$$\hat{\alpha}\cdot\bar{x} + \hat{\beta}\frac{\sum x_i^2}{n} \quad \text{(ii)}$$

$$\frac{d\ln L}{d\sigma^2} = -\frac{n}{2\sigma^2} - \sum(y_i-\alpha-\beta x_i)^2\left(-\frac{1}{2}\right)\left(\frac{0-1}{(\sigma^2)^2}\right) = 0 \Rightarrow \hat{\sigma}^2 = \frac{\sum(y_i-\hat{\alpha}-\hat{\beta}\cdot x)^2}{n} \quad \text{(iii)}$$

(i), (ii) gives $\begin{cases} \hat{\beta} = \dfrac{\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)/n}{\sum x_i^2 - \left(\sum x_i\right)^2/n} = \dfrac{S_{XY}}{S_{XX}} \\ \hat{\alpha} = \bar{y} - \hat{\beta}\cdot\bar{x} \end{cases}$. (iv)

The likelihood under H_0 is $L_0 = \frac{1}{\left(2\pi\cdot\sigma^2\right)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum(y_i-\alpha)^2} \Rightarrow \ln L_0 = -\frac{n}{2}\ln\sigma^2 - \frac{\sum(y_i-\alpha)^2}{2\sigma^2}$

$$\Rightarrow \frac{d\ln L_0}{d\alpha} = \frac{-(-1)\cdot 2\sum(y_i-\alpha)}{2\sigma^2} = 0 \Rightarrow \hat{\alpha} = \bar{y}. \quad \text{(v)}$$

$$\frac{d\ln L_0}{d\sigma^2} = -\frac{n}{\sigma^2} - \sum(y_i-\alpha)^2\left(\frac{0-1}{(\sigma^2)^2}\right) = 0 \Rightarrow \hat{\sigma}^2 = \frac{\sum(y_i-\hat{\alpha})^2}{n} = (\text{From (v)}) = \frac{\sum(y_i-\bar{y})^2}{n}$$

The latter estimate, obtained under H_0 is denoted $\hat{\sigma}_0^2$.

EX 83 (Continued) Thus, $\Lambda = \frac{L_0}{L} = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{n/2}$ (Several factors cancel each other.)

$\Rightarrow -2\ln\Lambda = n\ln\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}$. The p-value is $P\left(\chi^2(2-1) > -2\ln\Lambda_{OBS}\right)$.

The factor S_{XY} in (iv) can be expressed in several ways, e.g. $S_{XY} = \sum(x_i - \bar{x})(y_i - \bar{y})$. Similarly $S_{XX} = \sum(x_i - \bar{x})^2$.

The Λ-test statistic can be expressed more simply. $\sum\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2 = \sum\left(y_i - (\bar{y} - \hat{\beta}\bar{x}) - \hat{\beta}x_i\right)^2 =$

$\sum\left((y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})\right)^2 = \sum(y_i - \bar{y})^2 + \hat{\beta}^2\sum(x_i - \bar{x})^2 - 2\hat{\beta}\sum(x_i - \bar{x})(y_i - \bar{y}) =$

$S_{YY} + \hat{\beta}^2 S_{XX} - 2\hat{\beta}S_{XY} = \left[\text{Notice that } \hat{\beta}S_{XY} = \hat{\beta}^2 S_{XX}\right] = S_{YY} - \hat{\beta}^2 S_{XX}$.

Thus, $\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} = \frac{S_{YY} - \hat{\beta}^2 S_{XX}}{S_{YY}} = 1 - \frac{S_{XY}^2}{S_{XX}S_{YY}} = 1 - r^2$, where *r* is the *sample correlation coefficient*.

It follows that $-2\ln\Lambda = -n\ln(1-r^2)$.

If (X, Y) has a bivariate normal distribution it can be shown that the conditional expectation is $E(Y|X = x) = \alpha + \beta x$, where $\beta = \rho \cdot \sigma_Y / \sigma_X$ and ρ is the *population correlation coefficient*. The hypothesis that $\beta = 0$ is thus equivalent with the hypothesis that $\rho = 0$ and can be tested in the same way.

6.2.3 Miscellaneous methods

In this section we consider tests that are of a 'common sense-nature'. This means that the test statistics are sensitive to changes in the parameters and are used together with a proper RR. For example, when there is a sample of observations on a variable $Y \sim N(\mu, \sigma^2)$ and we want to test $H_0: \mu = \mu_0$ against H_a: $\mu \neq \mu_0$, it is obvious that we shall use a test statistic based on $\left|\overline{Y} - \mu_0\right|$ and reject H_0 for large values of the latter quantity. There may be situations where it is less obvious how to perform a test. Then one may use the Neyman-Pearson Lemma which states that, when testing $H_0: \theta = \theta_0$ against $H_a: \theta \neq \theta_a$, the test with maximal power is obtained from the LR $L_0 / L_a < c$ (cf. Ch. 20.10-20.13 in Stuart *et al* 1999.) We will seldom need this Lemma since in the following applications the best RR agrees with the one obtained by common-sense reasoning.

EX 84 Consider again the situation in EX 70 and EX 78 where we test H_0: $p = 1/2$ against H_a: $p \neq 1/2$.

Put $\hat{p} = \frac{Y}{n}$, with $E(\hat{p}) = p$ and $V(\hat{p}) = \frac{p(1-p)}{n}$. Intuitively it seems reasonable to choose the test statistic $T = \frac{\hat{p} - E(\hat{p}|H_0)}{\sqrt{V(\hat{p}|H_0)}} = \frac{\hat{p} - 1/2}{\sqrt{1/4n}}$. H_0 is rejected for large values of $|T|$ or equivalently, for large values of $T^2 = \left(\frac{\hat{p} - 1/2}{\sqrt{1/4n}}\right)^2$. However, this is exactly the same test that was obtained by the Chi-square principle.

EX 85 Consider the situation in EX 79 where data were obtained from two independent Binomial samples with proportions p_1 and p_2 and one wanted to test H_0: $p_1 = p_2 (= p)$ against H_a: $p_1 \neq p_2$.

a) Construct a test of 'common sense-nature'.
b) In order to test a new vaccine 90 pupils from a school were vaccinated and 66 were not vaccinated. After six months it was noticed how many pupils who had got a flue, with the following result:

	Vaccinated (1)	Not vaccinated (2)
With flu	4	18
Without flu	86	48
	90	66

Test whether the vaccine has a significant preventive effect by using the test in a). Compare the result with that which is obtained by using the LR test in EX 79.

EX 85 (Continued)

a) Test statistic: $T = \dfrac{\hat{p}_1 - \hat{p}_2 - E\left(\hat{p}_1 - \hat{p}_2 \middle| H_0\right)}{\sqrt{V\left(\hat{p}_1 - \hat{p}_2 \middle| H_0\right)}}$. Here $E\left(\hat{p}_1 - \hat{p}_2 \middle| H_0\right) = p - p = 0$,

$V\left(\hat{p}_1 - \hat{p}_2 \middle| H_0\right) = \dfrac{p(1-p)}{n_1} + \dfrac{p(1-p)}{n_2} = p(1-p)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)$. p is unknown and has to be estimated. In EX 46 it was seen that if $(Y_i)_{i=n}^{n}$ are independent and $Y_i \sim Binomial(n_i, p)$ then $\hat{p} = \dfrac{\sum n_i \hat{p}_i}{\sum n_i} = \left[\hat{p}_i = \dfrac{Y_i}{n_i}\right] = \dfrac{\sum Y_i}{\sum n_i}$ is an ML estimator of p and also BLUE. In this case $\hat{p} = \dfrac{Y_1 + Y_2}{n_1 + n_2}$. Thus, the test statistic to be used is $T' = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$. What about the distribution of T'?

According to the CLT $T \xrightarrow{D} Z \sim N(0,1)$ as $n_1, n_2 \to \infty$. By similar arguments as in EX 23 c) it then follows that also T' has a limiting standard normal distribution. (Recall that $Y_1 + Y_2 \sim Binomial(n_1 + n_2, p)$.) The latter convergence is however slower.

H_0 is rejected for larges values of $|T'|$, so p-value is $2 \cdot P\left(Z > |T'_{OBS}|\right)$.

b) From the table we get $\hat{p}_1 = \dfrac{4}{90} = 0.0444$, $\hat{p}_2 = \dfrac{18}{66} = 0.2727$, $\hat{p} = \dfrac{4+18}{90+66} = 0.1410$, from which

$$T' = \frac{0.0444 - 0.2727}{\sqrt{0.141 \cdot 0.859(1/90 + 1/66)}} = -4.04 \Rightarrow \text{p - value} = 2 \cdot P(Z > 4.04) = 0.00006$$

The LR test in EX 77 gives

$$\Lambda = \frac{(22/156)^{22}(1-22/156)^{156-22}}{(4/90)^4(1-4/90)^{90-4}(18/66)^{18}(1-18/66)^{66-18}} \Rightarrow -2\ln\Lambda = 16.8547 \Rightarrow$$

p-value $= P\left(\chi^2(2-1) > 16.8547\right) = 0.00004$

[Don't calculate Λ directly, but instead $\ln\Lambda = 22\ln(22/156) + \ldots - (66-18)\ln(1-18/66)$.]

Both p-values are very small and are close to each other. The conclusion is that the vaccine has significant preventive effect (p-value<0.001). Avoid statements such as 'H_0 is rejected' or 'p-value = 0.00006' if results are to be reported in a scientific journal.

EX 86 In EX 62 two rates λ_X and λ_Y in a Poisson process were compared. A 95% CI for the ratio $R = \lambda_Y / \lambda_X$ was (1.09, 2.17) and it was concluded that there was a significant difference between the rates.

The same data were analyzed in EX 82 by performing a LR test, giving the p-value 0.0109. The hypothesis of equal rates was thus rejected.

Consider now a third way to analyze the data, by using a test based on the conditional Poisson property in (3).

EX 86 (Continued)

$\left(Y(t)\middle|X(s)+Y(t)=n\right)\sim Binomial\left(n, p=\frac{\lambda_Y t}{\lambda_X s+\lambda_Y t}\right)$ so $H_0:\lambda_X=\lambda_Y \Leftrightarrow H_0: p=\frac{t}{s+t}$. Here $X(8)=85, Y(4)=65, \hat{p}=\frac{65}{85+65}=0.4333$ and and $H_0: p=1/3$ against $H_a: p\neq 1/3$.

Test statistic $T=\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}=\frac{0.4333-1/3}{\sqrt{\frac{1}{3}\cdot\frac{2}{3}\cdot\frac{1}{150}}}=2.6025$. Since $T\xrightarrow{D}Z\sim N(0,1)$ the p-value is $2P(Z>2.6025)=0.0094$. The null hypothesis is thus strongly rejected (p-value < 0.01).

EX 87 In EX 51 it was shown how CIs for the parameters in the normal distribution can be constructed. Assume that $(Y_i)_{i=1}^n$ are iid where $Y_i\sim N(\mu,\sigma^2)$.

a) Show how to test $H_0:\mu=\mu_0$ against $H_a:\mu\neq\mu_0$.

b) Show how to test $H_0:\sigma^2=\sigma_0^2$ against $H_a:\sigma^2\neq\sigma_0^2$.

c) Apply the tests with $\mu_0=16.0$ and $\sigma_0^2=0.4$ when $n=10, \bar{y}=16.67, s^2=0.7312.$

Compare these results with the results that are obtained using an approach based on CIs.

a) $T=\frac{\overline{Y}-\mu_0}{S/\sqrt{n}}\sim T(n-1)$. (Cf. EX 51.) p-value $=2P(T(n-1)>|T_{OBS}|)$

b) $T=\frac{(n-1)S^2}{\sigma_0^2}\sim\chi^2(n-1)$. (Cf. EX 51.) p-value $=2P(\chi^2(n-1)>T_{OBS})$

c) $n=10,\ \bar{y}=16.67,\ s^2=0.7312$.

To test $H_0:\mu=16.0$ against $H_a:\mu\neq16.0$ consider the test statistic

$T=\frac{16.67-16.0}{\sqrt{0.7312/10}}=2.478\Rightarrow$ p - value $=2P(T(9)>2.478)=0.035$. Here it suffices to conclude from a T-table that p-value < 0.05.

A 95% CI for μ is given by $\overline{Y}\pm C\frac{S}{\sqrt{n}}$, where *C* is determined by $P(T(9)>C)=0.025\Rightarrow C=2.262$.

Thus, $16.67\pm2.262\sqrt{0.7312/10}, (16.06, 17.28)$. Both approaches suggest that H_0 is rejected at the 5% level.

To test $H_0:\sigma^2=0.4$ against $H_a:\sigma^2\neq0.4$ consider the test statistic $T=\frac{9\cdot0.7312}{0.4}=16.45\Rightarrow$ p-value $=2P(\chi^2(9)>16.45)=0.116$, so H_0 is not rejected.

A 95% CI for σ^2 is given by $\frac{(n-1)S^2}{b}<\sigma^2<\frac{(n-1)S^2}{a}$, where *a* = 2.7004 and *b* =19.0228 (See EX 51 for details.) This gives the interval (0.36, 2.44) and neither in this case is H_0 rejected.

EX 88 Given two independent sets of iid variables $(X_i)_{i=1}^{n_X}$ and $(Y_i)_{i=1}^{n_Y}$ where $X_i \sim N(\mu_X, \sigma_X^2)$ and $Y_i \sim N(\mu_Y, \sigma_Y^2)$.

a) Show how to test $H_0 : \sigma_X^2 = \sigma_Y^2$ against $\sigma_X^2 \neq \sigma_Y^2$.

b) Show how to test $H_0 : \mu_X = \mu_Y (= \mu)$ against $\mu_X \neq \mu_Y$.

c) In two independent data sets one obtains

$\left(n_X = 10,\ \sum x_i = 126,\ \sum x_i^2 = 1692\right)$ and $\left(n_Y = 8,\ \sum y_i = 127,\ \sum y_i^2 = 2122\right)$. Perform the tests in a) and b) above and compare the results with that which are obtained by making CIs.

a) Let the largest of the two sample variances be S_Y^2. Then (Cf. EX 53.) $\frac{\sigma_X^2}{\sigma_Y^2}\frac{S_Y^2}{S_X^2} \sim F(n_Y - 1, n_X - 1)$ which under H_0 becomes $\frac{S_Y^2}{S_X^2} \sim F(n_Y - 1, n_X - 1)$. The p-value (two-sided) is $2P\left(F(n_Y - 1, n_X - 1) > \frac{s_Y^2}{s_X^2}\right)$.
Notice that, if we for some reason, want to test $H_0 : \sigma_X^2 = c \cdot \sigma_Y^2$ then the test statistic is $c \cdot \frac{S_Y^2}{S_X^2}$.

b) A suitable test statistic is $T = \frac{\overline{X} - \overline{Y} - E\left(\overline{X} - \overline{Y} \middle| H_0\right)}{\sqrt{V\left(\overline{X} - \overline{Y} \middle| H_0\right)}} = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_X^2 / n_X + \sigma_Y^2 / n_Y}}$.

If σ_X^2 and σ_Y^2 were known then *T* would be distributed $N(0,1)$, but in practice the variances are unknown and has to be estimated. If the test in a) suggests that the variances could be assumed to be equal, $= \sigma^2$, then we estimate this by $\hat{\sigma}^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}$. (Cf. EX 53.) It follows that the test statistic to be used is $T' = \frac{\overline{X} - \overline{Y}}{\sqrt{\hat{\sigma}^2 (1/n_X + 1/n_Y)}}$. The p-value (two-sided) is $2P\left(T(n_X + n_Y - 2) > \left|T'_{OBS}\right|\right)$.

If the test in a) suggests that the variances are unequal then we are faced with the Behrens-Fisher problem mentioned in the Comments to EX 53.

When *both* n_X and n_Y *are large* things become simpler since we can use the fact that

$$T'' = \frac{\overline{X} - \overline{Y}}{\sqrt{S_X^2 / n_X + S_Y^2 / n_Y}} = \frac{\frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_X^2 / n_X + \sigma_Y^2 / n_Y}} \sim N(0,1)}{\frac{\sqrt{S_X^2 / n_X + S_Y^2 / n_Y}}{\sqrt{\sigma_X^2 / n_X + \sigma_Y^2 / n_Y}} \xrightarrow{P} 1} \xrightarrow{D} Z \sim N(0,1)$$

The convergence in probability in the denominator above can be motivated in the following way: From (9b) $E\left(S_X^2\right) = \sigma_X^2$ and $V\left(S_X^2\right) = const./n_X \Rightarrow S_X^2 \xrightarrow{P} \sigma_X^2$ (Cf. (10)). Similarly, $S_Y^2 \xrightarrow{P} \sigma_Y^2$. (11a) and (11b) then gives the result.

EX 88 (Continued)

c) $\bar{x} = 12.600, s_X^2 = \frac{1692 - (126)^2/10}{(10-1)} = 11.60$, $\bar{y} = 15.875, s_y^2 = \frac{2122 - (127)^2/8}{(8-1)} = 15.13$

Test of $H_0 : \sigma_X^2 = \sigma_Y^2$ against $H_a : \sigma_X^2 \neq \sigma_Y^2$

$T = \frac{s_y^2}{s_X^2} = \frac{15.13}{11.60} = 1.30 \Rightarrow \text{p - value} = P(F(7,9) > 1.30) = 0.34$. We can thus assume equal variances.

Test of $H_0 : \mu_X = \mu_Y$ against $H_a : \mu_X \neq \mu_Y$

$\hat{\sigma}^2 = \frac{(10-1)\cdot 11.60 + (8-1)\cdot 15.13}{10+8-2} = 13.14$, $T' = \frac{15.875 - 12.600}{\sqrt{13.14(1/10 + 1/8)}} = 1.91 \Rightarrow \text{p - value} =$

$2P(T(10+8-2) > 1.91) = 2 \cdot 0.0371 = 0.074$. We can't reject H_0 at the 5% level.

A 95% CI for the variance ratio is $\frac{S_Y^2}{S_X^2 \cdot c_2} < \frac{\sigma_Y^2}{\sigma_X^2} < \frac{S_Y^2}{S_X^2 \cdot c_1}$ (Cf. EX 53.). Here c_1 and c_2 are constants that are determined in the following way:

$P(F(7,9) < c_1) = 0.025$ can't be found in most tables, but $P\left(\frac{1}{F(7,9)} > \frac{1}{c_1}\right) = P(F(9,7) > 1/c_1)$

$= 0.025 \Rightarrow 1/c_1 = 4.82 \Rightarrow c_1 = 0.21$. $P(F(7,9) > c_2) = 0.025 \Rightarrow c_2 = 4.20$

Thus, the CI is $\left(\frac{15.13}{11.60 \cdot 4.20}, \frac{15.13}{11.60 \cdot 0.21}\right) = (0.31, 6.21)$, in accordance with the test result above.

A 95 % CI for $\mu_Y - \mu_X$ is $(\bar{Y} - \bar{X}) \pm C\sqrt{\hat{\sigma}^2(1/n_X + 1/n_Y)}$ (Cf. EX 53.). C is determined by

$P(T(16) > C) = 0.025 \Rightarrow C = 2.120$.

Thus, $(15.875\text{-}12.600 \pm 2.120\sqrt{3.14(1/10 + 1/8)} = 3.28 \pm 3.65$. Since the interval covers zero the difference between the means is not significant.

EX 89 $(Y_i)_{i=1}^n$ are iid variables with an arbitrary distribution and with $E(Y_i) = \mu$ and $V(Y_i) = \sigma^2$. Show how to test $H_0 : \mu = \mu_0$ when n is large.

In EX 58 it was shown that $\frac{\bar{Y} - \mu}{S/\sqrt{n}} \xrightarrow{D} Z \sim N(0,1)$ as $n \to \infty$. As a test statistic we thus chose $T = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}}$ and p - value is $2P(Z > |T_{OBS}|)$ for a two-sided alternative.

EX 90 $(X_i)_{i=1}^{n_X}$ and $(Y_i)_{i=1}^{n_Y}$ are independent sets of iid variables with arbitrary distributions and finite means and variances. Show how to test $H_0 : \mu_X = \mu_Y$ when both sample sizes are large.

The test statistic to use is $T = \dfrac{\overline{X} - \overline{Y}}{\sqrt{S_X^2 / n_X + S_Y^2 / n_Y}} \sim N(0,1)$.

An argument for the distribution follows from EX 88 b). In the latter case all variables were assumed to have normal distributions. But, looking back at the proof it is seen that the numerator tends in distribution to a $N(0,1)$ -variable also for iid variables with arbitrary distributions.

The statistic above can also be used for constructing a CI for the difference between means.

6.2.4 Nonparametric methods

In earlier chapters inference has been based on estimated parameters in probability models. Such problems are said to be *parametric*, and others are called *nonparametric*. The distinction between the two methods is not clear-cut. Test of independency or test of equal marginal proportions are sometimes referred to as nonparametric methods, although a lot of parameters are involved. A tentative position is that nonparametric methods are less affected by unrealistic assumptions. The latter are however also based on assumptions, something that is often overlooked, especially that the observations are assumed to be iid.

Goodness of Fit and comparison of distributions based on the sample-distribution function

We saw in Ch. 6.2.1 that the Chi-square test can be used to test the hypothesis that the observations in a sample come from a certain distribution. This *Goodness of Fit problem* was solved by first estimate the parameters in the hypothetical distribution and then compare observed and expected frequencies. In this section we consider an alternative way to test for Goodness of Fit which we call the *Kolmogorov test*, also called the Kolmogorov-Smirnov test, proposed by the Russian statistician A. Kolmogorov in 1933. An important difference between the two tests is that in the Chi-square test the parameters are estimated, but in the Kolmogorov test the parameters have to be specified (or known). A further limitation is that the Kolmogorov test (in its original form) can only be applied to continuous distributions. Some attempts have been made to use the Kolmogorov test when parameters are estimated, e.g. in the Exponential or the Normal distribution. In the latter case the adjusted test is called the Lilliefors test and is provided by many statistical soft-wares (e.g. in *proc univariate* in SAS).

The *sample cdf*, $S_n(y)$, is constructed in the following way. Rank all observations from the smallest to the largest $y_{(1)} \le y_{(2)} \le \ldots \le y_{(n)}$. From the sequence $\left(y_{(i)}, i/n\right)_{i=1}^{n}$ we then form the step function $S_n(y) = i/n,\ y_{(i)} \le y < y_{(i+1)}$. As an illustration consider the data (1,2,2,5). Then $S_4(y) = 0, y < 1$

$= 1/4,\ 1 \le y < 2,\ = 3/4,\ 2 \le y < 3,\ = 3/4,\ 3 \le y < 5,\ = 1,\ y \ge 5$.

$H_0 : F(y) = F_0(y)$, (the hypothetical cdf with known parameters.)

The Kolmogorov test statistic is $D_n = \max\left|S_n(y) - F_0(y)\right|$ and H_0 is rejected for large values of D_n. In this case it is too complicated to compute p-values and a RR approach is simpler. The smallest values for which H_0 is rejected (two-sided test) with the significance levels $\alpha = 0.05$ and 0.01 and for various sample sizes $n \ge 1$, can easily be downloaded from the internet. For large n (at least larger than 100) the RR consists of observed values of D_n larger than $1.36/\sqrt{n}$ with $\alpha = 0.05$ and larger than $1.63/\sqrt{n}$ with $\alpha = 0.01$.

EX 91 Test whether the following ranked numbers are generated by a variable that is ~ *Uniform* [0,1]
.09 .20 .23 .29 .32 .34 .34 .37 .41 .45 .53 .70 .83 .87 .94 .97
$H_0 : F(y) = F_0(y) = y$.

We get the following table: (Notice that in this case $F_0(y) = y = y_{(i)}$.)

$y_{(i)}$	.09	.20	.23	.29	.32	.34	.37
i/n	1/16= .0625	2/16= .1250	3/16= .1875	4/16= .2500	5/16= .3125	7/16= .4375	8/16= .5000
.41	.45	.53	.70	.83	.87	.94	.97
9/16= .5625	10/16= .6250	11/16= .6875	12/16= .7500	13/16= .8125	14/16= .8750	15/16= .9375	16/16= 1

The largest value of D_{16} is $|0.6250 - 0.45| = .175$ and this is far below the rejection limit .3273 $(\alpha = 0.05)$, so there is no reason to reject the null hypothesis

A different problem is to compare two distributions by means of the observations in two independent samples. This was done several times in Ch. 6.2.2-6.2.3 by assuming a specific form of the distributions, e.g. $\sim N(\mu_X, \sigma^2)$ and $\sim N(\mu_Y, \sigma^2)$. Now we will test the hypothesis $H_0 : F_1(x) = F_2(y)$ without specifying the form of the cdfs, which are assumed to be continuous. The latter hypothesis is also termed the *hypothesis of homogeneity*. The test that is used will be called the *Smirnov two-sample test*, also called the Kolmogorov-Smirnov two-sample test. N.V. Smirnov (1900-1966) was a great mathematician in the former Soviet Union who won prices in many areas. (He is said to have won "the bronze star in vodka distillation" in 1940, but this is may be a student jokes.)

The Smirnov test statistic is $D_{m,n} = \max|S_m(x) - S_n(y)|$, where $S_m(x)$ is the sample cdf from a sample of size *m* and $S_n(y)$ is the sample cdf from a sample of size *n*. H_0 is rejected for large values of $D_{m,n}$. Tables for the test can easily be downloaded from the internet. Critical values are given for each pair of *m,n* (often denoted n_1, n_2) and for $\alpha = 0.05$ For $\alpha = 0.01$. large sample sizes, say above 25, approximate critical values are given by $1.36\sqrt{1/m + 1/n}$ and $1.63\sqrt{1/m + 1/n}$ for $\alpha = 0.01$.

The Smirnov test shall in first place be used when very little is known about the distributional form, but also as a complement to parametric tests in situations where it is suspected that lack of significance in a test might be a result of choosing a bad probability model.

EX 92 Check whether the following two samples of observations on the variables X_i and Y_i are drawn from the same population.

X_i : 7.6 8.4 8.6 8.7 9.3 9.9 10.1 10.6 11.2 (m = 9)

Y_i : 5.2 5.7 5.9 6.5 6.8 8.2 9.1 9.8 10.8 11.3 11.5 12.3 12.5 13.4 14.6 (n = 15)

We make the following table of ranked observations:

$X_{(i)}$	$Y_{(i)}$	$S_X(x) - S_Y(y)$
	5.2	0-1/15 = -1/15
	5.7	0-2/15 = -2/15
	5.9	0-3/15 = -3/15
	6.5	0-4/15 = -4/15
	6.8	0-5/15 = -5/15
7.6		1/9-5/15 = -2/9
	8.2	1/9-6/15 = -13/45
8.4		2/9-6/15 = -8/45
8.6		3/9-6/15= -1/15
8.7		4/9-6/15= 2/45
	9.1	4/9-7/15= -1/45
9.3		5/9-7/15= 4/45
	9.8	5/9-8/15= 1/45
9.9		6/9-8/15= 2/15
10.1		7/9-8/15= 11/45
10.6		8/9-8/15= 16/45
	10.8	8/9-9/15= 13/45
11.2		1-9/15= 2/5= 0.400
	11.3	1-10/15= 1/3
	11.5	1-11/15= 4/15
	12.3	1-12/15= 1/5
	12.5	1-13/15= 2/15
	13.4	1-14/15= 1/15
	14.6	1-1= 0

We obtain $D_{9,12} = 0.400 = 18/45$. (Tables of critical values of this test often show fractions.) Since $P\left(D_{9,15} > 19/45 \approx 0.422\right) = 0.20$ it is concluded that the maximal difference isn't large enough to reject the hypothesis of equal distributions. In fact $P\left(D_{9,15} > 0.533\right) = 0.05$, so a much larger maximal difference would be required to reject the hypothesis.

The Smirnov test can be used to test the more general hypothesis $H_0 : F_1(y) = \cdots = F_k(y), k \geq 2$. If all the *k* sample sizes are large, the p-value for H_0 can be computed quite easily, although the computations may be heavy. It is wise to have a computer program that calculates the value of the test statistic. Due to the usability in situations where it is hard to know the population distribution, we outline the test procedure (following the work in Fisz 1963, p. 409).

Let the sample sizes be $n_1 \ldots n_i \ldots n_k$ and define the constants

$$K_2 = \sqrt{n_2 n_1 /(n_1 + n_2)}, K_3 = \sqrt{n_3(n_1 + n_2)/(n_1 + n_2 + n_3)} \ldots K_k = \sqrt{n_k(n_1 + \ldots + n_{k-1})/(n_1 + \ldots + n_k)}$$

Introduce the statistics

$$D_2 = \max\left|S_2(y) - S_1(y)\right|, D_3 = \max\left|S_3(y) - \frac{(n_1 S_1(y) + n_2 S_2(y)}{(n_1 + n_2)}\right| \quad \ldots$$

$$D_k = \max\left|S_k(y) - \sum_{i=1}^{k-1} n_i S_i(y) / \sum_{i=1}^{k-1} n_i\right|$$

Put $A_i = K_i D_i, i = 2 \ldots k$ and $A_{MAX} = \max\left(A_2 \ldots A_k\right)$. Then the p-value is $1 - \left(Q(A_{MAX})\right)^{k-1}$ where $Q(\lambda)$ is the Kolmogorov-Smirnov λ-distribution. (Cf. Table VIII in Fisz 1963.)

The Sign Test and The Wilcoxon Signed-Rank test for One Sample

Let *M* be the population median in a continuous distribution. Then the hypothesis $H_0 : F(y) = F_0(y)$ implies the hypothesis $H_0 : M = M_0$. E.g. if $Y \sim N(\mu, \sigma^2)$ then $M = \mu$. When data consist of matched pairs $\left(X_i, Y_i\right)_{i=1}^n$ one can reduce the problem of making inference from two dependent samples to a one-sample problem by considering the differences $D_i = X_i - Y_i, i = 1 \ldots n.$ In this case it is natural to test the hypothesis $H_0 : M = 0$ which is equivalent to $H_0 : P\left(X > Y\right) = P\left(X < Y\right) = 1/2$.

The Sign Test for $H_0 : M = M_0$ consists of computing the value of the test statistic Y = 'Number of observations below M_0 (if suspiciously few are below) or above M_0 (if suspiciously few are above)'. By suspiciously few we mean that they deviate much from the expectation *n*/2. Under H_0 the test statistic is distributed $Binomial(n, p = 1/2)$.

EX 93 Consider the following measurements of body temperature (in degrees Celsius):
37.1 37.0 37.3 37.2 36.9 37.4 36.8 37.1 37.3 37.3 36.9 37.0 37.5 37.2 37.1
Are these data in agreement with the hypothesis that the median body temperature in the population is 37.0?

Since there are just 3 values that are below 37.0 we compute the probability $P\left(Y \leq 3 \middle| p = 1/2\right) =$
$\sum_{y=0}^{3} \binom{15}{y} (1/2)^y (1/2)^{15-y} = (1/2)^{15} \sum_{y=0}^{3} \binom{15}{y} = \frac{576}{32768} = 0.0176$. So, the (two-sided) p-value is $2 \cdot 0.0176 = 0.035$ and the hypothesis is rejected.

The Sign Test for Matched Pairs is illustrated in the following example.

EX 94 Blood pressure measurements (in millimeters of mercury) were obtained before and after a training program with the following result:

Subject	1	2	3	4	5	6
Before	136.9	201.4	166.8	150.0	173.2	169.3
After	130.2	180.7	149.6	153.2	162.6	160.1
Difference	6.7	20.7	17.2	-3.2	10.6	9.2

Test the hypothesis that the median difference is zero.

Since there is just 1 difference that is negative we consider the variable *Y* = 'Number of differences that are negative' and calculate the probability $P(Y \le 1 | p = 1/2) = \binom{6}{0}(1/2)^0(1/2)^6 + \binom{6}{1}(1/2)^1(1/2)^5 = (1/2)^6(1+6) = 0.109$. The (two-sided) p-value is 0.22 so the hypothesis can't be rejected by the sign test.

As a comparison we use Student's T-test (Cf. EX 87.) for the hypothesis that the mean difference is zero.

$\sum d_i = 61.2, \sum d_i^2 = 976.46 \Rightarrow \bar{d} = 10.2, s_d^2 = \frac{1}{5}\left(976.46 - (61.2)^2/6\right) = 70.4440$.

$T = \frac{10.2 - 0}{\sqrt{70.4440/6}} = 2.977 \Rightarrow P(T(5) > 2.977) = 0.0155$. So, the (two-sided) p-value is 0.031 and the hypothesis is rejected.

Notice that the latter test is based on the assumption that the observed differences come from a normal distribution. It is to be expected that tests that make use of more information about the distribution are more efficient (provided that the distributional assumptions are valid). However, in the next example we introduce a nonparametric test that is more efficient than the sign test and is nearly as efficient as the T-test.

The Wilcoxon Signed-Rank Test for Matched-Pairs

We will test $H_0 : F_X(x) = F_Y(y)$ based on a sample of matched pairs $(X_i, Y_i)_{i=1}^n$. Proceed in the following steps:

- Form the differences $D_i = X_i - Y_i = \begin{cases} +, \text{if } D_i > 0 \\ -, \text{if } D_i < 0 \end{cases}$. Ties, i.e. cases with $D_i = 0$, are eliminated. The 'working' sample size after this elimination is denoted *n*'. It is assumed that the differences are continuous and have a symmetric distribution about 0.
- Rank the absolute differences from the smallest (1) to the largest (*n*') and put a + or a – sign above the absolute difference. If two or more absolute differences are tied for the same rank, then the average rank is assigned to each member of the tied group. E.g. the six observations 6<7=7=7=7<8 are given the ranks 1, 3.5, 3.5, 3.5, 3.5, 6 since (2+3+4+5)/4=3.5.

- Put T^- = 'Rank sum for negative differences' and T^+ = 'Rank sum for positive differences'. As a test statistic chose $T = \min(T^-, T^+)$ and reject H_0 if $T \le T_C$, where T_C is the critical value in the table. Tables are easily downloaded from the internet. A good table can be found in Wackerly *et al* 2008, Table 9 in Appendix 3. The latter shows critical values for working sample sizes up to 50 together with the p-values 0.10, 0.05, 0.02 and 0.01 (two-sided tests). p-values can be computed in an exact way but this is complicated. For $n > 50$ one can use the fact that $Z = \dfrac{T^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$ has approximately a $N(0,1)$-distribution. So, the p-value is $2P(Z > |Z_{OBS}|)$.

EX 95 Consider again the data in EX 94. The following table can be constructed:

Sign	-	+	+	+	+	+
$\lvert D_i \rvert$	3.2	6.7	9.2	10.6	17.2	20.7
Rank	1	2	3	4	5	6

From this we get $T^- = 1, T^+ = 20 \Rightarrow T = \min(1,20) = 1$. The critical value from the table in Wackerly *et al* mentioned above is $T_C = 1$ and this corresponds to a p-value less than 0.05. The exact p-value is $1/32 \approx 0.0313$, very close to that obtained by the T-test in EX 94.

The Mann-Whitney U Test for Two Independent Samples

Also now we test the hypothesis $H_0 : F_X(x) = F_Y(y)$, but there are two independent samples $(X_i)_{i=1}^{n_X}$ and $(Y_i)_{i=1}^{n_Y}$ each with iid observations. The distributions may be discrete or continuous. Assume that $n_X \leq n_Y$. The test proceeds in the following steps:

- Put all the $n_X + n_Y$ observations together and rank them according to their magnitude, from smallest to largest. Compute the rank sum for the observations that belong to the X – sample and call this W.
- The test statistic is $U = n_X n_Y + n_X(n_X + 1)/2 - W$. Under H_0 the distribution of U is symmetric about the expectation $E(U) = n_X n_Y / 2$. This in turn implies that $P(U \leq u_0) = P(U \geq n_X n_Y - u_0)$.
- H_0 is rejected for extremely large or small values of U with two-sided tests. Critical values are obtained from tables. We will show in the example below how p-values can be computed by using Table 8, Appendix 3 in Wackerly *et al.* The latter gives values of $P(U \leq u_0)$ for sample sizes up to 10 and $u_0 = 0,1,\ldots, n_X n_Y / 2$.
- For $n_X > 10$ and $n_Y > 10$ it can be shown that $Z = \dfrac{U - n_X n_Y / 2}{\sqrt{n_X n_Y (n_X + n_Y + 1)/12}}$ is close to a $N(0,1)$ -distribution and p-values (two-sided) are obtained from $2 \cdot P(Z > |Z_{OBS}|)$.

EX 96 Consider the following two independent series of independent observations:

x:	51	32	41	57	47	38	62	44	42	35
y:	61	34	60	59	63	45	49	53	46	58

Test whether the two series come from the same population:

a) By using the Mann-Whitney U Test.
b) -"- with Normal approximation.
c) By using the T-test for two independent samples.

a) Put all observations together and rank those observations that are coming from the x-series.

	32	34	36	38	41	42	44	45	46	47
Rank	1			4	5	6	7			10

	49	51	33	57	58	59	60	61	62	63
Rank		12		14					19	

The rank sum is $W = 1+4+5+6+7+10+12+14+19 = 78$ and the test statistic is $U = 10\cdot 10 + 10\cdot(10+1)/2 - 78 = 77$.

The latter value is larger than the expectation, $10\cdot 10/2 = 50$. The p-value is thus

$2\cdot P(U \geq 77) = 2\cdot P(U \leq 23)$. [Remember the property of the U-distribution listed above.]

From the table we get $P(U \leq 23) = 0.0216$. So, the p-value is about 0.04 and the hypothesis of similar populations can be rejected (p<0.05, Mann-Whitney U Test.).

b) The observed value of *Z* is $\dfrac{77 - 10\cdot 10/2}{\sqrt{10\cdot 10\cdot(10+1)/12}} = 2.041$ and p-value is

$2\cdot P(Z > 2.041) = 2\cdot 0.0207 = 0.0414$, very close to the p-value obtained in a).

c) $\sum x_i = 449, \sum x_i^2 = 20977, \bar{x} = 44.9, s_X^2 = \dfrac{1}{(10-1)}\left(20977 - (449)^2/10\right) = 90.77$

$\sum y_i = 528, \sum y_i^2 = 28642, \bar{y} = 52.8, s_Y^2 = \dfrac{1}{(10-1)}\left(28642 - (528)^2/10\right) = 84.84$

First we test $H_0 : \sigma_X^2 = \sigma_Y^2$ [Cf. EX 88 a).].

$\dfrac{s_X^2}{s_y^2} = 1.07 \Rightarrow \text{p - value} = 2\cdot P(F(9,9) > 1.07) = 2\cdot 0.46 = 0.92$. There is no reason to reject the null hypothesis and we can pool the two variances

$\hat{\sigma}^2 = \dfrac{(10-1)\cdot 90.77 + (10-1)\cdot 84.44}{(10+10-2)} = 87.81$.

EX 96 (Continued) We then test $H_0 : \mu_X = \mu_Y$ [Cf. EX 88 b)]

$$\frac{44.9 - 52.8}{\sqrt{87.81(1/10 + 1/10)}} = -2.041 \Rightarrow \text{p - value} = 2 \cdot P(T(10-1) > 2.041) = 2 \cdot 0.038 = 0.076 .$$

According to this test we can't reject H_0 at the 5% level.

One reason to the failure of the T-test to reject the null hypothesis in this case, may be that the normality assumption is violated. A histogram of the y-values gives the following pattern:

y	25–35	35–45	45–55	55–65
Frequency	1	1	3	5

The histogram suggests that the y-values are sampled from a population with a skew distribution rather than a normal distribution, although the sample size is too small to verify this.

This illustrates the strength of non-parametric methods since these are not, or to less extent, dependent on the form of the population distribution.

Fisher's Exact Test of Independency

In EX 70 and EX 79 it was shown how independency could be tested in contingency tables, by the chi-square principle and the LR principle, respectively. Both these tests require that the sample size *n* is large and p-values are computed by using the asymptotic distribution. In small samples one can use a test termed Fisher's exact test to test for independence. Using the same notation for the cell frequencies in the 2 x 2 table as in Ch. 6.2.1, we calculate the probability of a certain outcome in the four cells, *given that the marginal are fixed*, from the expression

$$P(y_{11}, y_{12}, y_{21}, y_{22}) = \frac{y_{1+}!\, y_{2+}!\, y_{+1}!\, y_{+2}!}{y_{11}!\, y_{12}!\, y_{21}!\, y_{22}!} \tag{27}$$

The p-value is obtained by calculating the sum of probabilities of all outcomes in the 2 × 2 table that are more extreme or equal to the observed outcome. This is illustrated in EX 95 below.

This test was first suggested by R.A. Fisher who discussed an experimental investigation of a lady's claim to be able to tell by taste whether the tea was added to the milk or the milk was added to the tea ('the tea drinking lady experiment'). In that case all margins were fixed since there were 4 cups of each type and the lady was informed about this fact. However, the test is used also in situations where just one margin is fixed, and even when only the total *n* is fixed. In the last case it is an example of a *conditional test* where we condition on the margins in the *present* data, although we are aware of that the margins will vary randomly from sample to sample.

EX 97 The following 2 x 2 table shows the frequencies of the two variables lower back pain (yes/no) and sex (males and females).

		Lower back pain: Yes	Lower back pain: No
Sex	Male	6	2
	Female	11	19

a) Use Fisher's exact test to investigate whether the two variables are independent.
b) Answer the same question by using the Chi-square principle and the LR principle.

a) Construct tables with actual and more extreme outcomes given fixed margins:

6	2	8
11	19	30
17	21	38

Table A

7	1	8
10	20	30
17	21	38

Table B

8	0	8
9	21	30
17	21	38

Table C

According to Eq. (27) the probability of the outcome in Table A is $\frac{8!30!17!21!}{6!2!11!19!38!}$. The sum of the probabilities of the outcomes in Table A to Table C is

$\frac{8!30!17!21!}{38!}\left(\frac{1}{6!2!11!19!}+\frac{1}{7!1!10!20!}+\frac{1}{8!0!9!21!}\right)=0.0620$. This is the p-value for a one-sided test. The p-value for a two-sided test is usually found by doubling the one-sided value (McCullagh & Nelder 1983, p99) and this is the simplest alternative, but there are also other more complicated possibilities (Rao 1965, p345). The two-sided p-value is thus 0.1240.

The calculation of exact probabilities is tedious, but tables can be downloaded from the internet and even calculators, e.g. 'Free Fisher's Exact Test Calculator' which can be found on www.danielsoper.com . In SAS p-values are obtained from *proc freq* by adding */chisq* (see SAS manuals for details). The p-values obtained may vary slightly depending on which alternative is used.

b) The Chi-square principle gives (cf. EX 70)

$X^2=1.6378+1.3258+0.4367+0.3536=3.7539$, p-value $=P\left(\chi^2(1)>3.7539\right)=0.05027$. The latter is a two-sided p-value that differs very much from the value 0.1240 obtained in a). It is obvious that the sample size n = 38 is too small for the Chi-square approximation to be valid. One simple way to deal with the problem is to use *Yates correction for continuity* (Yates 1934, p217)

$$X^2_{Yates}=\frac{\left(\left|y_{11}y_{22}-y_{12}y_{21}\right|-n/2\right)^2}{y_{1+}y_{2+}y_{+1}y_{+2}}=\frac{\left(\left|6\cdot19-2\cdot11\right|-38/2\right)^2}{8\cdot30\cdot17\cdot21}=2.3635,$$ p-value $=P\left(\chi^2(1)>2.3635\right)=0.1242$,

very close to the p-value in a).

The LR principle gives (cf. EX 81)

$\hat{p}_{1+}=8/38,\ \hat{p}_{2+}=30/38,\ \hat{p}_{+1}=17/38,\ \hat{p}_{+2}=21/38,\ \hat{p}_{11}=6/38,\ \hat{p}_{12}=2/38,\ \hat{p}_{21}=11/38,$
$\hat{p}_{22}=19/38 \Rightarrow -2\log\Lambda=-2\left(-3.10014+1.58646+2.18822-2.58980\right)=3.83052$
p-value $=P\left(\chi^2(1)>3.83052\right)=0.05033$. This is close to 0.05027 obtained with the Chi-square principle. It seems to be unknown how to obtain a corrected test statistic in this case.

Rank correlation

Pearson's correlation coefficient for the correlation between two variables (cf. the section about properties of $p(y_1, y_2)$ in Ch. 2.1) is estimated by the sample correlation coefficient defined as $r = s_{XY} / s_X s_Y$, where $(n-1)s_X^2 = \sum x_i^2 - (\sum x_i)^2 / n$, $(n-1)s_Y^2 = \sum y_i^2 - (\sum y_i)^2 / n$ and $(n-1)s_{XY} = \sum x_i y_i - (\sum x_i)(\sum y_i) / n$. However, the calculation of *r* requires that the scale of the variables is at least at an interval level, i.e. that the operations addition and subtraction make sense. For ordinal (rank order) data one may define other measures of correlation. The simplest is Spearman's coefficient r_S. Let $R(x_i)$ be the rank of x_i among $x_1 \ldots x_n$ and let $R(y_i)$ be the rank of y_i among $y_1 \ldots y_n$, then r_S is defined as *r* but with x_i and y_i replaced by $R(x_i)$ and $R(y_i)$. The ranks for tied observations are treated in the same way as was done for the Mann-Whitney U test. If there are no ties in both the *x* and the *y* observations the computation of r_S can be simplified, $r_S = 1 - \frac{6}{n(n^2-1)} \sum d_i^2$, where $d_i = R(x_i) - R(y_i)$.

r_S can be used to test *the hypothesis of no association* between two variables in situations where it isn't possible to obtain precise measurements, but only ranked values. In two-sided tests the null hypothesis shall be rejected for large or small values of r_S (remember that $-1 \le r_S \le 1$). Critical values are found in tables, e.g. Table 11, Appendix 3 in Wackerly *et al* 2007. Tables can be easily downloaded from the internet.

EX 98 A person was asked to make an assessment about the ability of 10 subjects and rank them. The ability of the subjects was then evaluated in a formal test. The result was

Rank	Test (*x*)	1	2	3	4	5	6	7	8	9	10
according to	Assessment (*y*)	3	4	1	5	6	8	2	10	7	9

We find $\sum d_i^2 = 52 \Rightarrow r_S = 1 - \frac{6}{10 \cdot (100-1)} \cdot 52 = 0.685$. This is quite large, but is it large enough in order to reject the hypothesis of no association between the ranked series? Referring to Table 11 mentioned earlier, one finds the critical value 0.648 for $\alpha = 0.025$ (one-sided test) and $\alpha = 0.05$ (two-sided test). The conclusion is that there is a significant association between the two series (p<0.05, Spearman's rank correlation).

6.3 The power of normally distributed statistics

Let T be a statistic with mean θ and variance $\sigma^2(\theta)/n$. The variance is thus allowed to be a function of the mean. An example of this is the sample proportion $\hat{p} = Y/n$, where $Y \sim Binomial(n, p)$, with mean p and variance $p(1-p)/n$. In this section it is assumed that T is normally distributed or at least that n is so large that $Z = \frac{T - \theta}{\sigma(\theta)/\sqrt{n}}$ can be assumed to be distributed $N(0,1)$. In the examples below we first show a general expression for the power and then we consider some special cases.

EX 99 *A general expression for the power.*

Find the RR for testing $H_0 : \theta = \theta_0$ against $H_a : \theta \neq \theta_0$ when the type I error is α, and determine the power.

The RR is obviously of the form $|T - \theta_0| > C_\alpha$, i.e. $(T - \theta_0) > C_\alpha$ or $(T - \theta_0) < -C_\alpha$, where C_α is a constant that depends on α.

$\alpha = P(\text{Reject } H_0 | H_0) = P(T - \theta_0 > C_\alpha | H_0) + P(T - \theta_0 < -C_\alpha | H_0) =$ [Do the same operations on both sides of the inequality sign.] $= P\left(\frac{T - \theta_0}{\sigma(\theta_0)/\sqrt{n}} > \frac{C_\alpha}{\sigma(\theta_0)/\sqrt{n}} \right) + P\left(\frac{T - \theta_0}{\sigma(\theta_0)/\sqrt{n}} < -\frac{C_\alpha}{\sigma(\theta_0)/\sqrt{n}} \right) =$

$P\left(Z > \frac{C_\alpha}{\sigma(\theta_0)/\sqrt{n}} \right) + P\left(Z < -\frac{C_\alpha}{\sigma(\theta_0)/\sqrt{n}} \right) =$ [Due to symmetry.] $= 2 \cdot P\left(Z > \frac{C_\alpha}{\sigma(\theta_0)/\sqrt{n}} \right)$.

In the sequel we choose $\alpha = 0.05$, so $\frac{C_{0.05}}{\sigma(\theta_0)/\sqrt{n}} = 1.96 \Rightarrow C_{0.05} = 1.96 \cdot \sigma(\theta_0)/\sqrt{n}$.

The RR, with $\alpha = 0.05$, is $|T - \theta_0| > 1.96 \cdot \sigma(\theta_0)/\sqrt{n}$ (28a)

The power is $Pow(\theta) = P(\text{Reject } H_0) = P(T > \theta_0 + 1.96 \cdot \sigma(\theta_0)/\sqrt{n}) + P(T < \theta_0 - 1.96 \cdot \sigma(\theta_0)/\sqrt{n}) =$
[Do the same operations on both sides of the inequality sign.]=

$$P\left(\frac{T - \theta}{\sigma(\theta)/\sqrt{n}} > \frac{\theta_0 + 1.96 \cdot \sigma(\theta_0)/\sqrt{n} - \theta}{\sigma(\theta)/\sqrt{n}} \right) + P\left(\frac{T - \theta}{\sigma(\theta)/\sqrt{n}} < \frac{\theta_0 - 1.96 \cdot \sigma(\theta_0)/\sqrt{n} - \theta}{\sigma(\theta)/\sqrt{n}} \right).$$

From this we get

$$Pow(\theta) = P\left(Z > 1.96\frac{\sigma(\theta_0)}{\sigma(\theta)} - \frac{(\theta - \theta_0)}{\sigma(\theta)}\sqrt{n} \right) + P\left(Z < -1.96\frac{\sigma(\theta_0)}{\sigma(\theta)} - \frac{(\theta - \theta_0)}{\sigma(\theta)}\sqrt{n} \right) \quad (28b)$$

In (28a) and (28b) $\alpha = 0.05$. If it is very important to not falsely reject the null hypothesis one should choose a lower type I error. E.g. with $\alpha = 0.01$ the figure 1.96 is replaced by 2.575.

EX 100 Find a RR for testing $H_0 : p = 1/2$ against $H_a : p \neq 1/2$, where *p* is a Binomial proportion, and study the power.

In EX 23 b) it was shown that $\frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$ can be assumed to be distributed $N(0,1)$ for large *n*. Here $E(\hat{p}) = p$ and $V(\hat{p}) = p(1-p)/n$, so in this case $\theta = p$ and $\sigma^2(\theta) = p(1-p)$. Putting this into Eq. (28a) gives the following RR: $|\hat{p} - 1/2| > 1.96\frac{\sqrt{1/2 \cdot 1/2}}{\sqrt{n}} \approx \frac{1}{\sqrt{n}}$.

The power is from (28b):

$$Pow(p) = P\left(Z > 1.96\frac{1/2}{\sqrt{p(1-p)}} - \frac{(p-1/2)}{\sqrt{p(1-p)}}\sqrt{n} \right) + P\left(Z < -1.96\frac{1/2}{\sqrt{p(1-p)}} - \frac{(p-1/2)}{\sqrt{p(1-p)}}\sqrt{n} \right).$$

The behavior of this power is studied when *n* = 50 (pow1) and when *n* =100 (pow2). The following program codes (in SAS) computes the power and depicts the shapes of the powers in Figure 2 below.

```
data ppow;
n1=50; n2=100;
do p=0.1 to 0.9 by 0.1;
A=1.96/2/sqrt(p*(1-p));
B1=(p-1/2)*sqrt(n1)/sqrt(p*(1-p));
B2=(p-1/2)*sqrt(n2)/sqrt(p*(1-p));
pow1=1-probnorm(A-B1)+probnorm(-A-B1);
pow2=1-probnorm(A-B2)+probnorm(-A-B2);
output; end;
proc print; var p pow1 pow2;
run;
```

Obs	p	pow1	pow2
1	0.1	1.00000	1.00000
2	0.2	0.99784	1.00000
3	0.3	0.82832	0.98699
4	0.4	0.28904	0.51631
5	0.5	0.05000	0.05000
6	0.6	0.28904	0.51631
7	0.7	0.82832	0.98699
8	0.8	0.99784	1.00000
9	0.9	1.00000	1.00000

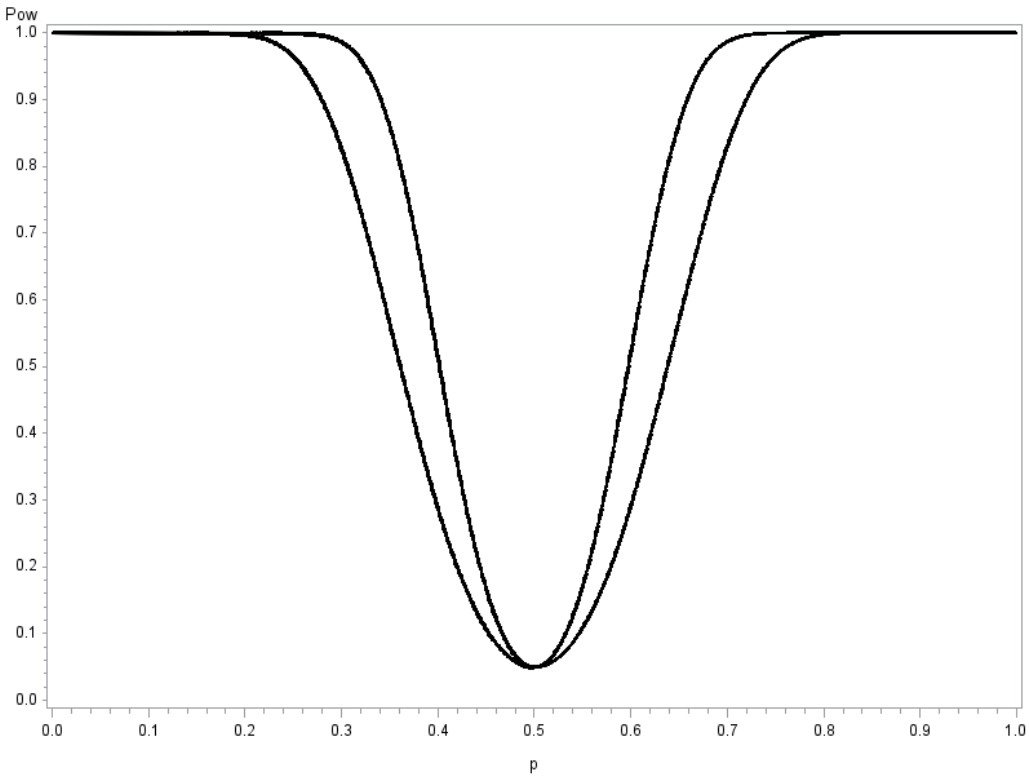


Figure 2

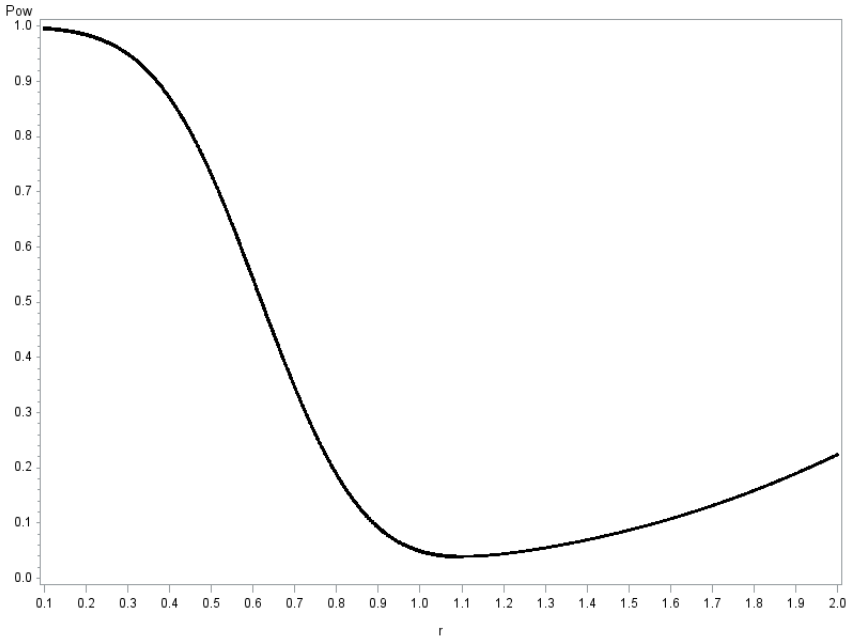


Figure 3

EX 101 Let $(Y_i)_{i=1}^n$ be iid observations where $Y_i \sim Exponential(\lambda)$. Construct a RR for testing $H_0; \lambda = \lambda_0$ against $H_a : \lambda \neq \lambda_0$ when *n* is so large that the CLT is applicable. Also, study the power and notice what happens if *n* is too small for the normality approximation to be valid, say *n* = 10.

$E(Y_i) = 1/\lambda, V(Y_i) = 1/\lambda^2 \Rightarrow E(\overline{Y}) = 1/\lambda, V(\overline{Y}) = \dfrac{1/\lambda^2}{n}$, so in this case $\theta = 1/\lambda$ and $\sigma^2(\theta) = 1/\lambda^2$. For large *n*, $\dfrac{\overline{Y} - 1/\lambda}{\sqrt{1/n\lambda^2}}$ is distributed $N(0,1)$.

(28a) gives the RR:

$$\left|\overline{Y} - 1/\lambda_0\right| > 1.96\frac{1/\lambda_0}{\sqrt{n}}.$$

(28b) gives the power, where we notice that $\dfrac{\sigma(\theta_0)}{\sigma(\theta)} = \dfrac{1/\lambda_0}{1/\lambda} = \dfrac{\lambda}{\lambda_0}$ and $\dfrac{\theta - \theta_0}{\sigma(\theta)} = \dfrac{1/\lambda - 1/\lambda_0}{1/\lambda} = 1 - \dfrac{\lambda}{\lambda_0}$. It turns out that the power is a function of λ/λ_0:

$$Pow(\lambda/\lambda_0) = P\left(Z > 1.96 \cdot (\lambda/\lambda_0) - (1 - (\lambda/\lambda_0))\sqrt{n}\right) + P\left(Z < -1.96 \cdot (\lambda/\lambda_0) - (1 - (\lambda/\lambda_0))\sqrt{n}\right).$$

The latter expression is obtained under the assumption that *n* is large. When *n* = 10 the power is illustrated in Figure 3 above. It is seen that the power is very weak for $r = \lambda/\lambda_0 > 1$ and perhaps more interesting is that the type I error can be smaller for *r* > 1 than for *r* = 1 (the value under H_0). Such a test is called *biased*.

6.4 Adjusted p-values for simultaneous inference

We have been told to reject the null hypothesis when the p-value is small (less than 0.05). In this case there is just one hypothesis to test. When we increase the number of hypothesis we increase the chance to reject at least one of the hypotheses when it's true. If α is the type I error for testing a single hypothesis, one has to make the p-values smaller so that the significance level of a whole family of hypotheses is (at most)α. When testing m hypotheses simultaneously the Italian statistician Bonferroni suggested that each hypothesis is tested at the level α/m. This advice had the drawback that extremely small individual p-values could be needed. An improved method was later suggested by Holm (1979, p. 65). The method can be described in the following way: If there are m simultaneous hypotheses to be tested, rank the p-values from the tests, from the smallest to the largest, $p_{(1)} < p_{(2)} < \ldots p_{(i)} < \ldots$. Then claim simultaneous significance for all p-values such that $p_{(i)} < \dfrac{\alpha}{m - i + 1}$. The method is illustrated in the following example.

EX 102 The following table shows the durations of a sick leave for various age groups.

		Duration (Weeks)			
		-1	1-4	4-	Total
Age Group (Years)	-30	48	32	12	92
	30-50	35	26	40	101
	50-	12	24	52	88
	Total	95	82	104	281

Is there an association between Age and Duration, and in such a case, which combinations can explain it?

The total chi-square is $X^2 = 47.35 \Rightarrow \text{p - value} = P\left(\chi^2(4) > 47.35\right) \approx 10^{-9}$, so the association is very strong. In order to search for an explanation to this we present a table with the measures *Deviation / Cell Chi-Square* (cf. Ch. 6.2.1).

		Duration (Weeks)		
		-1	1-4	4-
Age Group (Years)	-30	16.9 / 9.2	5.2 / 1.0	-22.1 / 14.3
	30-50	0.9 / 0.0	-3.5 / 0.4	2.6 / 0.2
	50-	-17.8 / 10.6	-1.7 / 0.1	19.4 /11.6

There are four Cell Chi-Square measures that are relatively large so we rank their corresponding p-values. The latter being obtained from a table showing X^2 and $p = P\left(\chi^2(1) > X^2\right)$.

i	1	2	3	4
Cell Chi-Square, X^2	14.3	11.6	10.6	9.2
p-value	0.0002	0.0007	0.0011	0.0024
$\frac{0.05}{3 \cdot 3 - i + 1}$	0.0056	0.0063	0.0071	0.0080

EX 102 (Continued) Here all p-values in the third row are smaller than the values in the fourth row. So, in the corresponding cells there are simultaneous significant deviations (at the 5% level). The conclusion is that there is an over-representation of members in the youngest age group with a short sick leave and also an over-representation of members in the oldest age group with a long sick leave.

Notice that if we want simultaneous significance at the 1% level, there are only three cells that meet the requirement. For the cell with $X^2 = 9.2$ one gets the p-value $= 0.0024 > \frac{0.01}{9 - 4 + 1} = 0.0016.$

There are other ways to adjust for multiple comparisons. E.g. when testing for pairwise equality of three or more means, one may apply the methods of Scheffe' or Tukey. These are used within the field of Analysis of Variance (ANOVA) and require that many assumptions are met. The so called Holm-Bonferroni method just described has the advantage that it can be used generally, although more specialized methods may be more efficient in certain situations.

There is no clear-cut answer to the question 'How many, and which hypotheses shall be considered in the simultaneous inference?'. When testing for significant individual cell deviations in an R × C contingency table it is quite natural to set up R × C hypotheses. In other cases it may be harder to reach a decision on this issue.

6.5 Randomized tests

In EX 69 it was noticed that with a discrete distribution, such as the Binomial, one can't expect that the type I error is exactly α. However, this can be achieved by introducing a further random component to the RR. The methodology is illustrated in the following frequently cited example. (Observe that a randomized test is not to be confused with a randomization test.)

EX 103 Let $(Y_i)_{i=1}^n$ be iid with $Y_i \sim Poisson(\lambda)$. We want to test $H_0 : \lambda = 0.1$ against the one-sided alternative $H_a : \lambda > 0.1$ with a type I error (α) of 0.05 and with n = 10

As a test statistic we take $\sum Y_i$ and reject H_0 if $\sum Y_i > c$. This choice of RR seems obvious, but can also be shown to follow from the Neyman-Pearson lemma. Since $\sum Y_i \sim Poisson(n\lambda)$ (cf. (3) in Ch. 3.1) we get, since $n\lambda = 10 \cdot 0.1 = 1$:

$\alpha = P\left(\sum Y_i > c \middle| H_0\right) = \sum_{y=c+1}^{\infty} \frac{1}{y!} e^{-1} = 1 - e^{-1} \sum_{y=0}^{c} \frac{1}{y!}$. From this we can construct the following table:

c	0	1	2	3
α	0.63	0.20	0.08	0.02

Since we can't find a value of c which gives $\alpha = 0.05$ we reformulate the RR in the following way:

$$\text{RR: } \begin{cases} \text{If } \sum Y_i > 3, \text{ reject } H_0 \text{ with probability 1} \\ \text{If } \sum Y_i = 3, \text{ reject } H_0 \text{ with probability P} \end{cases}$$

Now, $0.05 = P\left(\sum Y_i > 3 \middle| H_0\right) \cdot 1 + P\left(\sum Y_i = 3 \middle| H_0\right) \cdot \mathrm{P} = 0.02 \cdot 1 + 0.0613 \cdot \mathrm{P}$

Thus $\mathrm{P} = \frac{0.05 - 0.02}{0.0613} = 0.506 \approx 0.5$.

In practice this means that if $\sum y_i > 3$ then H_0 is rejected. But if $\sum y_i = 3$ it is not clear if H_0 should be rejected until you have tossed a coin where e.g. the outcome 'head' means rejection.

The above example with P = 0.5 has inspired a lot of jokers to make fun about theoretical statistical inference. An example: 'Patient: – Am I going to die in cancer? Statistician: – I just got the result from the lab but wait, first I have to toss a coin to decide about your future'. Randomized tests are not to be used in practice for several apparent reasons. But, there is one important application for randomized tests, and that is when the power functions of several discretely distributed test statistics are to be compared. In that case it is important that the all the power curves start at the same level.

6.6 Some tests for linear models

6.6.1 The Gauss-Markov model

Let (X,Y) be a bivariate random variable (cf. the properties (1)-(10) in Ch. 2.2.1). The conditional expectation $E(Y|X=x)$ is called the *regression function for the regression of Y on X* and the conditional variance $V(Y|X=x)$ is called the *residual variance*. We will use the following notations for the population parameters:

$$E(X)=\mu_X, E(Y)=\mu_Y, V(X)=\sigma_X^2, V(Y)=\sigma_Y^2, \sqrt{V(Y)}=\sigma_Y, Cov(X,Y)=\sigma_{XY}, \text{population correlation } \rho=\frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

$$E(Y|X=x)=\mu_{Y|x}=\alpha+\beta\cdot x \text{ if linear}, V(Y|X=x)=\sigma^2 \text{ if constant.}$$

An important special case is when (X,Y) has a bivariate Normal distribution. In that case

$$E(Y|X=x)=\alpha+\beta\cdot x, \text{ with } \beta=\rho\frac{\sigma_Y}{\sigma_X}=\frac{\sigma_{XY}}{\sigma_X^2},\ \alpha=\mu_Y-\beta\cdot\mu_X,\ V(Y|X=x)=\sigma^2=\sigma_Y^2(1-\rho^2) \quad (29)$$

In the Gauss-Markov model (Rao 1965, p179) the following assumptions are made:

$$\begin{aligned}&(Y_i|x_i)_{i=1}^n \text{ are independent and } \sim N(\alpha+\beta\cdot x_i,\sigma^2). \text{ This can alternatively be expressed}\\ &Y_i=\alpha+\beta\cdot x_i+E_i, \text{ where } (E_i)_{i=1}^n \text{ are iid and } \sim N(0,\sigma^2)\end{aligned} \quad (30)$$

The model in (30) is quite restrictive. It involves independency, linearity, constant variance and Normality. The model is not proper for follow-up, or panel data, where measurements are taken from several subjects that are followed in time. In the special case when $\alpha=0$ the model is called *regression through the origin* (cf. EX 54).

Corresponding to the population moments above there are sample moments.

$$S_{XX}=\sum(X_i-\bar{X})^2=\sum X_i^2-\frac{(\sum X_i)^2}{n}, S_{YY}=\sum Y_i^2-\frac{(\sum Y_i)^2}{n}, S_{XY}=\sum X_iY_i-\frac{(\sum X_i)(\sum Y_i)}{n},$$

sample correlation $r=\frac{S_{XY}}{S_XS_Y}$, where $S_X=\sqrt{S_{XX}}$, estimators $\hat{\beta}=\frac{S_{XY}}{S_{XX}}, \hat{\alpha}=\bar{Y}-\hat{\beta}\cdot\bar{x}, \hat{\sigma}^2=\frac{SSE}{(n-2)}$, where $SSE=\sum(Y_i-\hat{\alpha}-\hat{\beta}\cdot x_i)^2=S_{YY}-\hat{\beta}^2S_{XX}$ (Cf. EX 83) is the 'sum of square for errors'. From the assumptions in (30) it follows that $\hat{\beta}\sim N(\beta,\frac{\sigma^2}{S_{XX}}), \hat{\alpha}\sim N\left(\alpha,\sigma^2(\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}})\right), \hat{\sigma}^2\sim\sigma^2\frac{\chi^2(n-2)}{(n-2)}$. Furthermore, $\hat{\beta}$ and $\hat{\alpha}$ are independent of $\hat{\sigma}^2$ and $Cov(\hat{\alpha},\hat{\beta})=-\sigma^2\bar{x}/S_{XX}$.

EX 104

a) Show how to test $H_0 : \beta = \beta_0$ against $H_a : \beta \neq \beta_0$ and $H_0 : \alpha = \alpha_0$ against $H_a : \alpha \neq \alpha_0$.

b) The following data show the relation between Body weight in kg (x) and Body volume in liter (y) for twelve 4-year old boys:

x	17.1	10.5	13.8	15.7	11.9	10.4	15.0	16.0	17.8	15.8	15.1	12.1
y	16.1	10.4	13.5	15.9	11.6	10.2	14.1	15.8	17.6	15.5	14.8	11.9

Test the following hypotheses $H_0 : \beta = 1$ against $H_a : \beta \neq 1$ and $H_0 : \alpha = 0$ against $H_a : \alpha \neq 0$ by considering the regression of *Y* on *x*.

c) Repeat the analysis in b) but now by considering the regression of *X* on *y*.

a)

$$\frac{\hat{\beta} - \beta_0}{\sqrt{\sigma^2 / S_{XX}}} \sim N(0,1) \text{ and } \frac{\hat{\sigma}^2}{\sigma^2} \sim \frac{\chi^2(n-2)}{(n-2)} \Rightarrow \frac{\hat{\beta} - \beta_0}{\sqrt{\hat{\sigma}^2 / S_{XX}}} = \frac{\frac{\hat{\beta} - \beta_0}{\sqrt{\sigma^2 / S_{XX}}}}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}} \sim \frac{N(0,1)}{\sqrt{\frac{\chi^2(n-2)}{(n-2)}}} \sim T(n-2).$$

Similarly, $\frac{\hat{\alpha} - \alpha_0}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-2)S_{XX}}\right)}} \sim T(n-2)$.

b) $\sum x_i = 171.2,\ \sum x_i^2 = 2511.26, \sum y_i = 167.4, \sum y_i^2 = 2400.14, \sum x_i y_i = 2454.51$.

$$S_{XX} = 2511.26 - (171.2)^2/12 = 68.8067, S_{YY} = 2400.14 - (167.4)^2/12 = 64.9100,$$
$$S_{XY} = 2454.51 - (171.2)(167.4)/12 = 66.2700$$

$$\hat{\beta} = \frac{66.2700}{68.8067} = 0.96, \quad \hat{\alpha} = 13.95 - \hat{\beta} \cdot 14.27 = 0.21, \quad \hat{\sigma}^2 = \frac{64.9100 - \hat{\beta}^2 \cdot 68.8067}{(12-2)} = 0.1088$$

$$H_0 : \beta = 1, \quad T = \frac{0.96 - 1}{\sqrt{0.1088/68.8067}} = 1.01 \Rightarrow \text{p - value} = 2 \cdot P(T(10) > 1.01) = 2 \cdot 0.17 = 0.3$$

No reason to reject H_0.

$$H_0 : \alpha = 0, \quad T = \frac{0.21 - 0}{\sqrt{0.1088\left(\frac{1}{12} + \frac{(14.27)^2}{68.8067}\right)}} = 0.63 \Rightarrow \text{p - value} = 2 \cdot P(T(10) > 0.63) = 2 \cdot 0.27 = 0.54$$

No reason to reject H_0. We have an example of regression through the origin.

c) Now the regression function is $E(X|Y = y) = \alpha' + \beta' \cdot y$.

$$H_0 : \beta = 1, \quad T = \frac{1.02 - 1}{\sqrt{0.1148/(64.91)}} = 0.48 \Rightarrow \text{p - value} = 2 \cdot P(T(10) > 0.48) = 2 \cdot 0.32 = 0.64.$$

EX 104 (Continued)

$$H_0:\beta=1,\quad T=\frac{1.02-1}{\sqrt{0.1148/(11\cdot 64.91)}}=1.652\Rightarrow \text{p - value}=2\cdot P\left(\chi^2(10)>1.652\right)=2\cdot 0.065=0.13.$$

We can't reject H_0.

$$H_0:\alpha=0,\ T=\frac{0.024-0}{\sqrt{0.1148\left(\frac{1}{12}+\frac{(13.95)^2}{64.91}\right)}}=0.013\Rightarrow \text{p - value}=2\cdot P\left(T(10)>0.013\right)=2\cdot 0.495=0.99$$

No reason to reject j H_0. Very strong reason for regression through the origin.

Comment to EX 104 Sometimes it isn't crystal clear which of two variables that should be regarded as dependent. In such situations one may try to let both be dependent and check whether the two regression analyses give consistent results. Notice however that a regression relation is different from a mathematical relation. From the mathematical relation $y=\alpha+\beta\cdot x$ one can solve for the inverse relation $x=-\alpha/\beta+1/\beta\cdot y=\alpha'+\beta'\cdot y$. E.g. in EX 104 we don't get $\alpha'=-0.21/0.96=-0.22$, but $\alpha'=0.024$.

Several Y-observations at each x. Test of linearity.

Data is now $\left(\left(Y_{ij},x_i\right)_{j=1}^{n_i}\right)_{i=1}^{k}=\left(Y_{1j},x_1\right)_{j=1}^{n_1}\ldots\left(Y_{kj},x_k\right)_{j=1}^{n_k}$. The expressions for estimation and tests of parameters are the same as above. The data $\left(Y_{1j},x_1\right)_{j=1}^{2}$ is interpreted as $\left(Y_{11},x_1\right),\left(Y_{12},x_1\right)$. The difference is that we now can test whether there is one linear regression line through the data or not.

Introduce the notations $\overline{Y}_i=\frac{\sum_{j=1}^{n_i}Y_{ij}}{n_i}$, $\overline{Y}=\frac{\sum_{i=1}^{k}n_i\overline{Y}_i}{\sum_{i=1}^{k}n_i}$, $\overline{x}=\frac{\sum_{i=1}^{k}n_ix_i}{\sum_{i=1}^{k}n_i}$. Then the parameter estimates can be computed as $\hat{\beta}=\frac{\sum n_ix_i\overline{Y}_i-\left(\sum n_i\overline{Y}_i\right)\left(\sum n_ix_i\right)/\sum n_i}{\sum n_ix_i^2-\left(\sum n_ix_i\right)^2/\sum n_i}$, $\hat{\alpha}=\overline{Y}-\hat{\beta}\cdot\overline{x}$.

The hypothesis to test is $H_0:E\left(Y\middle|X=x\right)=\alpha+\beta\cdot x$. The test statistic for this is

$$F=\frac{\sum n_i\left(\overline{Y}_i-(\hat{\alpha}+\hat{\beta}x_i)\right)^2/(k-2)}{\sum\sum(Y_{ij}-\overline{Y}_i)^2/(\sum n_i-k)}.\ \text{The p-value for } H_0 \text{ is } P\left(F(k-2,\sum n_i-k)>F_{OBS}\right) \qquad (31)$$

The test in Eq. (31) is illustrated in the following example.

EX 105 Test the following (artificial) data for linearity:

x	1	2	3
Y	3, 4, 5	1, 3	3, 4, 5

We first make the table:

x_i	x_i^2	n_i	$n_i x_i$	$n_i x_i^2$	Y_{ij}	$\overline{Y}_i$	$n_i\overline{Y}_i$	$n_i x_i\overline{Y}_i$
1	1	3	3	3	3, 4, 5	4	12	12
2	4	2	4	8	1, 3	2	4	8
3	9	3	9	27	3, 4, 5	4	12	36
Total		8	16	38			28	56

From this we get $\hat{\beta} = \dfrac{56-(28)(16)/8}{38-(10)^2/8} = 0,\ \hat{\alpha} = \dfrac{28}{8} = 3.5$

The value of the test statistic can then be computed from the following table:

x_i	n_i	Y_{ij}	$\overline{Y}_i$	$n_i\left(\overline{Y}_i - (\hat{\alpha}+\hat{\beta}x_i)\right)^2$	$\sum\left(Y_{ij}-\overline{Y}_i\right)^2$
1	3	3, 4, 5	4	$3(4-3.5)^2 = 0.75$	1+0+1=2
2	2	1, 3	2	$2(2-3.5)^2 = 4.50$	1+1=2
3	3	3, 4, 5	4	$3(4-3.5)^2 = 0.75$	1+0+1=2
Total	8			6	6

The statistic is $F = \dfrac{6/(3-2)}{6/(8-3)} = 5.00 \Rightarrow \text{p - value} = P\left(F(1,5) > 5.00\right) = 0.076$. The hypothesis of linearity can't be rejected at the 5 % level. The p-value is however small and one should look for other alternatives than the straight line.

Regression towards the mean- how to 'lie' with regression analysis

When people with extreme values of the measurements, such as high blood pressure, are measured once more it is found that the mean of the extreme group is closer to the mean of the whole population. If people with extreme values are treated with some medicine the decrease may be interpreted as showing the effect of the treatment (a significant negative value of β). The problem is that *the mean level may go down (significantly) even if people are not treated*. This phenomenon, known as regression towards the mean, can be explained by measurement errors and natural biological variations (cf. Davis, p. 493). It is actually linked with the word "regression" used by F. Galton in a paper from 1885, who found that the height of children from very short or very tall parents move toward the average. This false pattern is more pronounced if we relate change with initial value. The following theoretical example is instructive if you want to make an experiment which proves that your 'hocus pocus drug' has a significant lowering effect on blood pressure, anxiety, cholesterol, body weight etc. The intention is of course that you shall use the knowledge to reveal others, not to use it for their own purposes.

EX 106 *Regression of 'Change' on 'Initial value'.*

Introduce the following notations and assumptions:

$Y_1 =$ Systolic Blood Pressure (SBP) at a point in time and $Y_2 =$ SBP at a time later. Put $D = Y_2 - Y_1$. For simplicity it is assumed that $V(Y_1) = V(Y_2) = \sigma^2$. The correlation between $\rho_{12} = \dfrac{Cov(Y_1, Y_2)}{\sigma^2}$

We are interested in the relation between Y_1 and D. Assume that (Y_1, D) has a bivariate Normal distribution and consider the regression function $E(D|Y_1 = y_1) = \alpha + \beta \cdot y_1$. From Eq. (29) we know that $\beta = \dfrac{Cov(Y_1, D)}{V(Y_1)}$ and from Ch. 2.1 we get

$Cov(Y_1, D) = E(Y_1 \cdot D) - E(Y_1)E(D) = E(Y_1(Y_2 - Y_1)) - E(Y_1)(E(Y_2) - E(Y_1)) = E(Y_1Y_2) - E(Y_1)E(Y_2) -$

$E(Y_1^2) + (E(Y_1))^2 = Cov(Y_1, Y_2) - V(Y_1) = \sigma^2\rho_{12} - \sigma^2 = -\sigma^2(1 - \rho_{12}) \Rightarrow \beta = -(1 - \rho_{12})$, which in practice is negative.

The true regression line $E(D|Y_1 = y_1)$ will have a negative slope and it is thus likely that the estimated line also has a negative slope.

Multiple regression

The regression function is now $E(Y|X_1 = x_1,\ldots,X_k = x_k) = \alpha + \beta_1 x_1 + \ldots + \beta_k x_k$. The assumptions about the variables $(Y_i|x_1,\ldots,x_k)_{i=1}^n$ are analogous to those in Eq. (30). Under the latter assumptions the best estimators of the parameters are given by the OLS method (cf. Ch. 4.3.1). The estimators $\hat{\alpha}, \hat{\beta}_1,\ldots,\hat{\beta}_k$ can easily be expressed in matrix form but this is beyond the scope of this book. The reader is advised to obtain the solutions by running some computer program, e.g. the procedures *proc glm* or *proc reg* in SAS.

Before considering the tests being of interest we make some comments about the variables $X_1 \ldots X_k$. The latter are called *independent variables* in contrast to Y which is the *dependent variable*. The independent variables may also be termed *explanatory variables* (often used by econometricians) or *predictors*. The interpretation of a single regression parameter β_i is that it shows how much the expectation of Y changes when x_i is increased by one unit and all other independent variables are fixed. But, if the x-variables are inter-related, or more or less *collinear*, this is impossible. Collinearity may lead to biased parameter estimates with great variance.

One special form of independent variables is the so called *dummy variable*. Consider the following examples:

- We want to study how Y = 'Amount of savings' depends on x = 'Salary' among men and women. Introduce the dummy $z = 1$ for men and $z = 0$ for women. The regression model can be written

$$E(Y|x,z) = \alpha + \beta_1 x + \beta_2 z + \beta_3 x \cdot z = \begin{cases} \alpha + \beta_2 + (\beta_1 + \beta_3)x, \text{ if } z = 1 \\ \alpha + \beta_1 x, \text{ if } z = 0 \end{cases}.$$

  This is a comparison of two lines, one for men and one for women. Hypotheses of interest are if $\beta_3 = 0$ (parallel lines), or if $\beta_3 = 0$ and $\beta_2 = 0$ (identical lines). Here β_1 is a separate salary-effect regardless of sex, β_2 is a separate sex- effect regardless of salary and β_3 is a salary-effect connected to sex. The latter parameter measures the *interaction effect*. When analyzing data with this model in a computer the input data consists of values in 3 columns, *Y, x, z*. Then you have to specify the model. E.g. in SAS you write the lines *proc glm; model y=x z x*z*;

- In the above example there was a comparison of two regression lines. When several regression lines are to be compared things are a bit more complicated. Assume that we want to study how Y = 'Household expenditure' depends on x = 'Salary' during the four seasons of the year. Since there are four seasons we introduce three dummies such that

Season	z_1	z_2	z_3
Spring	0	0	0
Summer	1	0	0
Autumn	0	1	0
Winter	0	0	1

The model can be written

$$E(Y|x,z_1,z_2,z_3)=\alpha+\beta x+\beta_1 z_1+\beta_2 z_2+\beta_3 z_3=\begin{cases}\alpha+\beta x,\text{Spring}\\ \alpha+\beta_1+\beta x,\text{Summer}\\ \alpha+\beta_2+\beta x,\text{Autumn}\\ \alpha+\beta_3+\beta x,\text{Winter}\end{cases}$$

This is a comparison of four parallel lines. If we allow the four lines to have different slopes we add $\beta_4 xz_1+\beta_5 xz_2+\beta_6 xz_3$ to the latter regression function.

Above we have defined dummies for two sexes and four seasons. In general, with c categories we need c-1 dummies taking the values 1 and 0. In computer programs for estimating linear models there may be other choices of dummies. The definition of those specific dummies that have been used is seen in the beginning of the print-out.

The first result of interest in a multiple regression study is the ANOVA (Analysis of Variance) table. This shows how the total variation of the Y –observations can be split up into two components. This is shown in the following table:

Variance source	Degrees of freedom	Sum of squares
Regression (Model)	k	$SSR=SST-SSE$
Error	$n-k-1$	$SSE=\sum_{i=1}^{n}\left(Y_i-(\hat{\alpha}+\sum_{j=1}^{k}\hat{\beta}_j x_j)\right)^2$
Corrected Total	$n-1$	$SST=\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2$

From the table we get an unbiased estimator of σ^2, $\hat{\sigma}^2=\frac{SSE}{n-k-1}\sim\frac{\sigma^2}{(n-k-1)}\chi^2(n-k-1)$.

We also get a measure of the fit of the linear model, the *Coefficient of Determination* $R^2_{Y|x_1\mathsf{K}\,x_p}=\frac{SSR}{SST}$, taking values in the interval [0, 1]. Values close to 1 indicates that the explanatory ability of the model is god. (R^2 is in fact the square of the *Multiple correlation coefficient* which is the correlation between Y and $\hat{\alpha}_0+\sum\hat{\beta}_j X_{ij}$.) R^2 can never become smaller when more x- variables are included in the regression model. As a measure of the gain in explanatory ability by including x_{p+1} beyond $x_1\ldots x_p$ one may use $\frac{R^2_{Y|x_1\mathsf{K}\,x_{p+1}}-R^2_{Y|x_1\mathsf{K}\,x_p}}{1-R^2_{Y|x_1\mathsf{K}\,x_p}}$, i.e. the actual increase in relation to the maximal possible increase.

Four classes of hypothesis to be tested

1) $H_0 : \text{All } \beta_j = 0,\ j = 1 \ldots p.$ Test statistic is $T = \dfrac{SSR/k}{SSE/(n-k-1)}$. p - value $= P\left(F(k, n-k-1) > T_{OBS}\right)$.

   This should be the first hypothesis to test and if it isn't rejected there is no reason to continue, but instead try to search for better explanatory variables.

2) $H_0 : \text{Some } \beta_j = 0.$ Test statistic is $T = \dfrac{\hat{\beta}_j}{\sqrt{\hat{V}(\hat{\beta}_j)}}$. p-value $= 2 \cdot P\left(T(n-k-1) > |T_{OBS}|\right)$.

   Here $\hat{\beta}_j$ and $\sqrt{\hat{V}(\hat{\beta}_j)}$ are found from the computer out-print under the names 'Estimate' and 'Std error of estimate'. Notice that since $T^2(f) = F(1, f)$ (cf. *(11)* and *(12)* in Ch. 3.1) the p-value can also be obtained from $P\left(F(1, n-k-1) > T_{OBS}^2\right)$.

   In this case it is perhaps more instructive to place a 95 % CI under each estimated $\hat{\beta}_j$. The latter is obtained from $\hat{\beta}_j \pm C\sqrt{\hat{V}(\hat{\beta}_j)}$ where C is determined from $P\left(T(n-k-1) > C\right) = 0.025$.

3. $H_0 : \text{All } \beta_j = 0 \text{ for } j = 1 \ldots k' < k.$ Test statistic is $T = \dfrac{(SSE' - SSE)/(k-k')}{SSE/(n-k-1)}$. Here SSE is the sum of square for 'Error' in the full model with k regression coefficients and SSE' is corresponding sum of squares in the reduced model with $k - k'$ regression coefficients. p - value $= P\left(F(k-k', n-k-1) > T_{OBS}\right)$.

   This test is perhaps the most useful one. It enables us to see whether the model with regression function $\alpha + \beta_1 x_1 + \ldots + \beta_{k'} x_{k'} + \ldots + \beta_k x_k$ can be replaced by the regression function $\alpha + \beta_{k'+1} x_{k'+1} + \ldots + \beta_k x_k$. The use of this test is illustrated in the following examples.

4. Tests about linear structures of the regression coefficients. Some examples are the following:

   $H_0 : E\left(Y \middle| X = x\right) = \alpha_0 + \beta_0 x$. Here, α_0, β_0 and x have fixed given values. E.g. α_0 and β_0 are the intercept and slope that has been observed during a long time for a production process and one wants to test whether a new process gives the same regression relation at *x*.

   Test statistic is $T = \dfrac{(\hat{\alpha} + \hat{\beta}x) - (\alpha_0 + \beta_0 x)}{\sqrt{\hat{\sigma}^2\left(\dfrac{1}{n} + \dfrac{(x - \bar{x})^2}{S_{XX}}\right)}}$ with p-value $= P\left(T(n-2) > |T_{OBS}|\right)$.

It is instructive to derive the above expression. Obviously, $\hat{\alpha} + \hat{\beta}x$ is unbiased for $\alpha + \beta x$.

$V(\hat{\alpha}+\hat{\beta}x)=[\text{Cf. Eq. (2) in Ch. 2.1}]=V(\hat{\alpha})+x^2V(\hat{\beta})+2xCov(\hat{\alpha},\hat{\beta})=\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}}\right)+\frac{x^2\sigma^2}{S_{XX}}-\frac{2x\bar{x}\sigma^2}{S_{XX}}=$

$=\sigma^2\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{XX}}\right)$. Since $\hat{\alpha}$ and $\hat{\beta}$ each has normal distributions it follows that

$\frac{(\hat{\alpha}+\hat{\beta}x)-(\alpha_0+\beta_0x)}{\sqrt{V(\hat{\alpha}+\hat{\beta}x)}}\sim N(0,1)$. Now, dividing numerator and denominator in the expression for T above by $\sqrt{V(\hat{\alpha}+\hat{\beta}x)}$ yields a statistic that is distributed as $\frac{N(0,1)}{\sqrt{\chi^2(n-2)/(n-2)}}\sim T(n-2)$.

An alternative to testing is to construct a CI for the true regression line at x:

$\hat{\alpha}+\hat{\beta}x\pm C\sqrt{\hat{\sigma}^2\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{XX}}\right)}$, where C is determined from $P(T(n-2)>C)=\alpha/2$ to get a $100(1-\alpha)\%$ CI. E.g. if we want a 90% CI when n =12, then Tables over the T-distribution shows that $P(T(10)>1.812)=0.10/2=0.05$, so C = 1.812. so C = 1.812.

For several x–variables the computations are heavy and will not be shown here. Results can be obtained from most computer programs. E.g. in SAS the codes *proc glm; model y=x1 x2 x3/clm p;* will give you 95% CIs for the expected means $E(Y|x_1,x_2,x_3)=\alpha+\beta_1x_1+\beta_2x_2+\beta_3x_3$, together with predicted (estimated) values.

Another type of linear structure among regression parameters is the following

$H_0 : \beta_2 = \beta_3 (= \beta)$ in the regression function $E\left(Y \middle| X = x\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$.

This has been termed a test of aggregation, in this case aggregation of the variables x_2 and x_3. A typical example is when Y is 'Prices of clothing', x_2 is 'Price of leather' and x_3 is 'Price of textile'. If H_0 is not rejected this means that the effects of prices of leather and can't be separated. The simplest way to perform the test is to run two regression analyses, one with x_1, x_2, x_3 as independent variables giving rise to SSE, and one with $x_1, (x_2 + x_3)$ as independent variables giving rise to SSE'. The test statistic is $T = \frac{(SSE' - SSE)/(3-2)}{SSE/(n-4)}$ with p-value $= P\left(F(1, n-4) > T_{OBS}\right)$.

EX 107 Data below shows a sample of 12 persons employed in a company where *Y* = 'Weekly earnings (in 1000 SEK)' of various ages (year) and sex (0 = Woman, 1 = Man). Assume that the *Y*–values are normally distributed.

Weekly earnings	5	6	8	8	6	10	8	11	9	11	13	13
Age (*x*)	20	30	35	35	40	40	45	45	50	55	55	60
Sex (*z*)	0	0	0	1	0	1	0	1	1	1	1	1

a) Test whether the mean salary differ between men and women. (Use an ordinary T-test.)

b) Study if mean salary increases with age by using the model $E(Y|x) = \alpha + \beta x$. From the out-print you get the following results:

ANOVA table

	df	SS	Parameter	Estimate	Std Error of Estimate
Regression	1	59.00	β	0.20	0.036
Error	10	19.00			
Corrected Total	11	78.00			

Test $H_0 : \beta = 0$ and compute the Coefficient of Determination. (Std Error of Estimate in the table above is simply $\sqrt{\hat{V}(\hat{\beta})}$).

c) Find a proper regression model that describes how mean salary depends on both age and sex.

Model: $E(Y|x,z) = \alpha + \beta_1 x + \beta_2 z$

ANOVA table

	df	SS	Parameter	Estimate	Std Error of Estimate
Regression	2	66.2396	β_1	0.14	0.039
Error	9	11.7604	β_2	2.07	0.879
Corrected Total	11	78.00			

Model: $E(Y|x,z) = \alpha + \beta_1 x + \beta_2 z + \beta_3 x \cdot z$

ANOVA table

	df	SS	Parameter	Estimate	Std Error of Estimate
Regression	3	67.1659	β_1	0.10	0.060
Error	8	10.8341	β_2	-0.61	3.358
Corrected Total	11	78.00	β_3	0.066	0.080

d) Comment on the following statement: 'It's true that men earn more than women, but this is due to the fact that men at the company tend to be older than women and salary increases with age'.

EX 107 (Continued)

a) Women ($z = 0$): $\bar{y}_W = \frac{33}{5} = 6.60, s_W^2 = \frac{225 - (33)^2/5}{(5-1)} = 1.80$

Men ($z = 1$): $\bar{y}_M = \frac{75}{7} = 10.714, s_M^2 = \frac{825 - (75)^2/7}{(7-1)} = 3.5714$

$H_0 : \sigma_W^2 = \sigma_M^2$, $F = \frac{3.5714}{1.80} = 1.984 \Rightarrow \text{p - value} = P(F(6,4) > 1.984) = 0.26$. No reason to reject H_0.

The pooled variance estimate is $\hat{\sigma}^2 = \frac{4 \cdot 1.80 + 6 \cdot 3.5714}{4+6} = 2.863$.

$H_0 : \mu_W = \mu_M$, $T = \frac{10.714 - 6.60}{\sqrt{2.863\left(\frac{1}{5} + \frac{1}{7}\right)}} = 4.153 \Rightarrow \text{p - value} = 2 \cdot P(T(10) > 4.153) = 2 \cdot 0.001 = 0.002$.

Reject H_0, women have significantly lower earnings.

b) $H_0 : \beta = 0$, $T = \frac{0.20}{0.036} = 5.56 \Rightarrow \text{p - value} = 2 \cdot P(T(10) > 5.56) = 2 \cdot 0.0001 = 0.0002$. Reject

Reject H_0, there is a strong linear relation between age and earnings.

The Coefficient of Determination is $R_{Y|x}^2 = \frac{59.00}{78.00} = 0.756\,(76\,\%)$.

c) Model: $E(Y|x,z) = \alpha + \beta_1 x + \beta_2 z$

$H_0 : \beta_1 = 0, T = \frac{0.14}{0.039} = 3.59 \Rightarrow \text{p - value} = 2 \cdot P(T(9) > 3.59) = 0.006$. Reject H_0.

$H_0 : \beta_2 = 0, T = \frac{2.07}{0.879} = 2.35 \Rightarrow \text{p - value} = 2 \cdot P(T(9) > 2.35) = 0.043$. Reject H_0.

Both *x* and *z* has a significant effect on salary. $R_{Y|x,z}^2 = \frac{66.2396}{78.00} = 0.849\,(85\,\%)$.

Model: $E(Y|x,z,xz) = \alpha + \beta_1 x + \beta_2 z + \beta_3 x \cdot z$

$H_0 : \beta_1 = 0, T = \frac{0.10}{0.060} = 1.70 \Rightarrow \text{p - value} = 2 \cdot P(T(8) > 1.70) = 0.13$. There is no significant effect of *x*.

$H_0 : \beta_2 = 0, T = \dfrac{-0.61}{3.36} = -0.18 \Rightarrow \text{p - value} = 2 \cdot P(T(8) > 0.18) = 0.86$. -"- z.

$H_0 : \beta_3 = 0, T = \dfrac{0.066}{0.080} = 0.83 \Rightarrow \text{p - value} = 2 \cdot P(T(8) > 0.83) = 0.43$. -"- $x \cdot z$.

In this case $R^2_{Y|x,z,xz} = \dfrac{67.1659}{78.00} = 0.861$. The relative increase of R^2 by introducing $x \cdot z$ is only

$\dfrac{0.861 - 0.849}{1 - 0.849} = 8\%$. There is no point in assuming that there are two lines with different slopes

d) It's true that earnings increase with age, but it's also true that there is a separate sex effect on the earnings that has nothing to do with the age effect.

6.6.2 Random Coefficient models

When repeated measurements are obtained in time from a sample of persons or companies, the Gauss-Markov model in the preceding chapter can be very poor. This type of data is called *panel data* by econometricians and *longitudinal data* or *follow-up data* by biometricians. The typical pattern in this data is that measurements from each chosen sampling unit have its own development. When the latter develop along straight lines, each line has its own intercept and possibly also its own slope. If the Gauss-Markov model is used and one single model is fitted to the data, the conclusions can be totally wrong. If each individual slope is positive, the line fitted by the Gauss-Markov model can be negative and vice versa. This is nicely illustrated in Diggle *et al*, p. 1. Models where slopes and intercepts are allowed to be random are called *Random Coefficient models*. A general exposition of these models is beyond the scope of this book. Here we only illustrate the inference when intercepts are random and slopes are fixed, *Error Components Regression (ECR) models*, and when both intercepts and slopes vary randomly, *Random Coefficient Regression (RCR) models*.

Assume that measurements are made on n persons at the same times $i = 1\ldots, t$. (The latter assumption will simplify the computations considerably.) The value obtained of the j:th person, $j = 1,\ldots,n$, at time i is denoted Y_{ij}. The two models are

$$ECR : Y_{ij} = A_j + \beta \cdot x_i + E_{ij}$$

$$RCR : Y_{ij} = A_j + B_j \cdot x_i + E_{ij}$$

The assumptions are: $E_{ij} \sim N(0,\sigma^2), A_j \sim N(\alpha,\sigma_A^2), B_j \sim N(\beta,\sigma_B^2), Cov(A_j,B_j) = \sigma_{AB}$. All other components are uncorrelated. It seems hard to motivate all those Normality assumptions, especially that slopes have Normal distributions, but the assumptions are needed to reach any results in the inference. In both models $E(Y_{ij}) = \alpha + \beta \cdot x_i$. In the *ECR* model $V(Y_{ij}) = \sigma_A^2 + \sigma^2$ and $Cov(Y_{ij},Y_{i'j}) =$ [cf. Ch. 2.1 Properties of $p(y_1,y_2)$ *(7)*] $E\big((A_j + \beta \cdot x_i + E_{ij})(A_j + \beta \cdot x_{i'} + E_{i'j})\big) - (\alpha + \beta \cdot x_i)(\alpha + \beta \cdot x_{i'}) =$

$$E\big(A_j^2 + A_j\beta x_{i'} + A_j E_{i'j} + A_j\beta x_i + \beta^2 x_i x_{i'} + \beta x_i E_{i'j} + A_j E_{ij} + \beta x_{i'} E_{ij} + E_{ij}E_{i'j}\big) -$$

$\alpha^2 - \alpha\beta x_{i'} - \alpha\beta x_i - \beta^2 x_i x_{i'}$ =[Many terms cancel each other out]= $E(A_j^2) - \alpha^2 = V(A_j) = \sigma_A^2$. Thus, the correlation between Y_{ij} and $Y_{i'j}$ is $Cov(Y_{ij}, Y_{i'j}) / \sqrt{V(Y_{ij})V(Y_{i'j})} = \sigma_A^2 / (\sigma_A^2 + \sigma^2)$. So, there is a constant correlation between measurements within each person. In the *RCR* model similar calculations yields $V(Y_{ij}) = \sigma_A^2 + 2x_i\sigma_{AB} + x_i^2\sigma_B^2$ and $Cov(Y_{ij}, Y_{i'j}) = \sigma_A^2 + (x_i + x_{i'})\sigma_{AB} + x_i x_{i''}\sigma_B^2$. From these expressions it is seen that a simple test to decide whether data follows an *ECR*- or a *RCR* model is to plot estimates of $V(Y_{ij})$ against x_i. If the latter relationship is constant, then we have an *ECR* model. On the other hand, if the relationship shows a quadratic pattern we have a *RCR* model. A formal statistical test that enables us to choose between the two models is presented below.

The following statistics will be needed. $\bar{x} = \frac{1}{t}\sum x_i, \bar{Y}_j = \frac{1}{t}\sum Y_{ij}, B_{YY} = \sum \bar{Y}_j^2 - \frac{1}{n}\left(\sum \bar{Y}_j\right)^2$

$$S_{xx} = \sum x_i^2 - \frac{1}{t}\left(\sum x_i\right)^2, S_{xY_j} = \sum x_i Y_j - \frac{1}{t}\left(\sum x_i\right)\left(\sum_i Y_{ij}\right), S_{Y_jY_j} = \sum_i Y_{ij}^2 - \frac{1}{t}\left(\sum_i Y_{ij}\right)^2$$

$$\hat{\beta}_j = \frac{S_{xY_j}}{S_{xx}}, \hat{\alpha}_j = \bar{Y}_j - \hat{\beta}_j\bar{x}, \hat{\beta} = \frac{1}{n}\sum \hat{\beta}_j, \hat{\alpha} = \frac{1}{n}\sum \hat{\alpha}_j, SSE_j = \sum_i \left(Y_{ij} - (\hat{\alpha}_j + \hat{\beta}_j x_i)\right)^2 = S_{Y_jY_j} - (\hat{\beta}_j)^2 S_{xx},$$

(This last relation is proved in EX 81.) $SSE = \sum SSE_j, S_{AA} = \sum \hat{\alpha}_j^2 - \frac{1}{n}\left(\sum \hat{\alpha}_j\right)^2,$

$$S_{BB} = \sum \hat{\beta}_j^2 - \frac{1}{n}\left(\sum \hat{\beta}_j\right)^2.$$

The hypotheses to test are $H_0 : ECR$ against $H_a : RCR \Rightarrow H_0 : \sigma_B^2 = 0$ against $H_a : \sigma_B^2 > 0$. The test statistic for this is $T = S_{xx}\dfrac{S_{BB}/(n-1)}{SSE/n(t-2)}$ with p-value $= P\left(F(n-1, n(t-2)) > T_{OBS}\right.$. It is shown in Petzold & Jonsson 2003, p6 that this test statistic is identical with the test statistic F_1 in Hsiao 2003, p15. For the more general model with *k* regression coefficients the reader is referred to the latter citations. Since computations can be quite heavy, the analysis with these types of models are facilitated by utilizing soft-ware, e.g. *proc mixed* in SAS.

Depending on the outcome of the latter test we go further.

ECR model: $\hat{\sigma}^2 = \dfrac{SSE + S_{xx}S_{BB}}{n(t-1)-1}, \hat{\sigma}_A^2 = \dfrac{B_{YY}}{n-1} - \dfrac{\hat{\sigma}^2}{t}, \hat{V}(\hat{\beta}) = \dfrac{\hat{\sigma}^2}{nS_{xx}}, \hat{V}(\hat{\alpha}) = \dfrac{1}{n}\left(\hat{\sigma}_A^2 + \hat{\sigma}^2(\dfrac{1}{t} + \dfrac{\bar{x}^2}{S_{xx}})\right)$

$H_0 : \beta = \beta_0$ against $H_a : \beta \neq \beta_0$ is tested by $T = \dfrac{\hat{\beta} - \beta_0}{\sqrt{\hat{V}(\hat{\beta})}}$ with p-value $= 2 \cdot P\left(T(n(t-1)-1) > \left|T_{OBS}\right|\right)$.

RCR model: $\hat{\sigma}^2 = \dfrac{SSE}{n(t-2)}, \hat{\sigma}_A^2 = \dfrac{S_{AA}}{n-1} - \hat{\sigma}^2\left(\dfrac{1}{t} + \dfrac{\bar{x}^2}{S_{xx}}\right), \hat{\sigma}_B^2 = \dfrac{S_{BB}}{n-1} - \dfrac{\hat{\sigma}^2}{S_{xx}}, \hat{V}(\hat{\beta}) = \dfrac{1}{n}\left(\hat{\sigma}_B^2 + \dfrac{\hat{\sigma}^2}{S_{xx}}\right)$,

$\hat{V}(\hat{\alpha})$ is the same as for the *ECR* model.

$H_0 : \beta = \beta_0$ against $H_a : \beta \neq \beta_0$ is tested by $T = \dfrac{\hat{\beta} - \beta_0}{\sqrt{\hat{V}(\hat{\beta})}}$ with p-value $= 2 \cdot P\left(T(n-1) > \left|T_{OBS}\right|\right)$.

EX 108

The table below shows the concentration of HbA1c (Glycosylated hemoglobin) measured at three points in time ($x_i = 1,2,3$ months) on 18 patients with diabetes. The purpose of the study was to see whether it is possible to reduce the HbA1c level among patients by dietary advice.

j	Y_{ij}	$\overline{Y}_j$	$\hat{\beta}_j$	$\hat{\alpha}_j$	SSE_j
1	6.4 6.3 7.6	6.77	0.60	5.57	0.327
2	9.1 8.5 8.2	8.60	-0.45	9.50	0.015
3	7.6 8.2 6.8	7.53	-0.40	8.33	0.667
4	7.3 7.3 7.0	7.20	-0.15	7.50	0.015
5	9.6 9.7 8.7	9.33	-0.45	10.23	0.202
6	9.3 8.8 8.5	8.87	-0.40	9.67	0.007
7	8.3 7.5 7.8	7.87	-0.25	8.37	0.202
8	8.1 7.9 7.3	7.77	-0.40	8.57	0.027
9	8.6 7.4 7.9	7.97	-0.35	8.67	0.482
10	8.2 8.1 7.5	7.93	-0.35	8.63	0.042
11	7.4 7.0 6.7	7.03	-0.35	7.73	0.002
12	6.8 6.7 6.5	6.67	-0.15	6.97	0.002
13	8.4 8.8 7.9	8.37	-0.25	8.87	0.282
14	9.2 8.9 8.8	8.97	-0.20	9.37	0.007
15	7.9 8.2 7.4	7.83	-0.25	8.33	0.202
16	7.2 6.8 6.4	6.80	-0.40	7.60	0.000
17	8.0 7.6 7.0	7.53	-0.50	8.53	0.007
18	10.2 11.2 8.9	10.10	-0.65	11.40	1.815
Total			-5.35	153.83	4.298

$$\bar{x} = 2, S_{xx} = 2, B_{YY} = 15.1805, \hat{\beta} = \frac{1}{18} \cdot (-5.35) = -0.297, \hat{\alpha} = \frac{1}{18} \cdot 153.83 = 8.546, S_{AA} = 28.1, S_{BB} = 1.127$$

$H_0 : ECR$ against $H_a : RCR$ is tested by

$T = 2 \cdot \frac{1.127/(18-1)}{4.298/18(3-2)} = 0.55 \Rightarrow \text{p - value} = P(F(17,18) > 0.55) = 0.88$. No reason to reject the *ECR* model.

For the *ECR* model, $\hat{\sigma}^2 = \frac{4.298 + 2 \cdot 1.127}{18(3-1)-1} = 0.187$, so $\hat{V}(\hat{\beta}) = \frac{0.187}{18 \cdot 2} = 0.0052$

$H_0 : \beta = 0$ against $H_a : \beta \neq 0$ is tested by

$T = \dfrac{-0.297}{\sqrt{0.0052}} = -4.12 \Rightarrow$ p - value $= 2 \cdot P\big(T(18(3-1)-1) > 4.12\big) = 0.0001$. This means that H_0 is strongly rejected.

The conclusion is that there is a significant reduction of the HbA1c level and this reduction is similar to all patients with a mean rate of about 0.3 units per month.

6.7 Final words

Statistics is not an exact science in the sense that there are clear-cut solutions to every problem. When analyzing linear models you find two opposite schools. The 'significance fundamentalists' argues that all non-significant parameters must be deleted from the model. The argument is that unnecessary parameters 'steal' degrees of freedom so that other parameters may not clear the 5% p-value threshold. On the other hand there are 'significance liberals' who retain all parameters in the model that they find interesting. The author's personal view is close to that of a 'significant fundamentalist'.

The square root of the estimated variance of a statistic is called *Standard Error of Estimate*. This is an old fashion name, but is now common in computer printouts.

Recall that there are two different ways to test for equality rates in a Poisson process. One is based on interval data, intervals between events (EX 80), and one based on counts, or frequency data (EX 84).

In this book you find several examples of *tests of linearity* in regression models. This is an area that has been overlooked. Many examples where significance can't be established may be due to the fact that a linear model is used where a non-linear model would be more adequate. Closeness to linearity is often said to be measured by R^2, the coefficient of determination. However, the F-test in (31) Ch. 6.6 is much more efficient in detecting deviation from linearity. In EX 146 where linearity was rejected by the F-test, one obtains $R^2 = 0.969$, which is large.

As you have noticed, tests of hypotheses in linear models require heavy computations. It is therefore desirable that you supply reliable statistical software to your computer. This is of special importance when dealing with random coefficient models (Ch. 6.2.2) where more or less sophisticated software are available under the name of 'mixed models'.

When communicating results from a statistical analysis you should avoid expressions like "H_0 against H_a". (This is for internal use among statisticians.) Instead use formulations like "The new method gives significantly lower values than the old method ($p<0.01$, two-sided Sign test), just to take an example.

Supplementary Exercises, Ch. 6

EX 109 Gregor Mendel is said to be the founder of the science of genetics. He performed a large number of experiments to test his theories and much of these data are still available. In one famous experiment he cross-pollinated smooth yellow pea plants with wrinkly green peas with the following result:

(Shape, Color)	(Round, Yellow)	(Wrinkly, Yellow)	(Round, Green)	(Wrinkly, Green)
Theoretical proportion	9/16	3/16	3/16	1/16
Observed frequency	315	108	101	32

a) Make a 2 x 2 table of the observed frequencies in terms of the factors Color and Shape.
b) Test whether the observed frequencies are in accordance with Mendel's theory.

EX 110 The number of white blood cells per cubic millimeters is known to vary according to a Poisson distribution. 10 blood samples from the same person showed the following number of white blood cells: 81 38 63 63 50 63 69 50 38 31.

a) Compare the sample mean and variance. Conclusion?
b) Use the Chi-square principle to test whether the observations come from the same Poisson distribution.

EX 111 The number of persons being on sick leave per day was recorded at a department, with the following result:

Number on sick leave	0	1	2	3	4	5-
Frequency	12	10	6	0	2	0

Determine whether the Poisson distribution is an adequate model for the outcome.

EX 112 The Normal distribution is often taken for granted without giving any support for this assumption. Show how the Chi-square principle can be used in order to test whether the following (ordered) data can be assumed to be Normally distributed:

23 23 24 27 29 31 32 33 35 36 36 37 40 42 43 43 44 45 48 48 54 54 56 57 57 58 58 58 58 59 61 61 62 63 64 65 66 68 68 70 73 74 75 77 81 87 89 93 97

[Hint: Use some classification, e.g. -39, 40-60, 61-80, 81-.]

EX 113 *Several independent Binomial samples.*

In EX 77 and EX 83 the proportion in two independent samples were compared. Consider now $\left(Y_i\right)_{i=1}^{k}$ where $Y_i \sim Binomial(n_i, p_i)$ and $H_0 : p_1 = \ldots = p_k (= p)$.

Under the null hypothesis $\hat{p} = \dfrac{\sum n_i \hat{p}_i}{\sum n_i} = \dfrac{\sum Y_i}{\sum n_i}$ is BLUE (cf. EX 46). The statistic for testing H_0 is

(Rao 1965, p. 333) $T = \dfrac{1}{\hat{p}(1-\hat{p})} \sum n_i (\hat{p}_i - \hat{p})^2 \sim \chi^2(k-1)$ under H_0.

Consider the following Norwegian data:

Season	Spring	Summer	Autumn	Winter
Number of born boys	9251	7967	7327	7662
Number of births	17866	15408	14251	14885

Test whether the proportion born boys is the same for all seasons.

EX 114 Test for independency between (A, Non-A) and (B, Non-B) in the following three contingency tables (fictive data).

Group 1

	B	Non-B	Total
A	10	40	50
Non-A	20	80	100
Total	30	120	150

Group 2

	B	Non-B	Total
A	60	40	100
Non-A	30	20	50
Total	90	60	150

Group 1+2

	B	Non-B	Total
A	70	80	150
Non-A	50	100	150
Total	120	180	300

Conclusions?

EX 115 In a sample of 100 couples, the husbands and wives were asked about their opinions about a politician. The result was:

		Husband	
		Positive	Negative
Wife	Positive	6	10
	Negative	24	60

a) Estimate the proportion positive husbands and wives, respectively. Determine whether the difference between the proportions is significant by using the Chi-square and the LR principles.
b) Are the opinions of husbands and wives independent?

EX 116 In a medical rehabilitation project patients with different degree of estimated working capacity (low, medium, high) received different types of training (physical, activation, education). The following frequencies were obtained:

		Working capacity			
		Low	Medium	High	Total
Type of training	Physical	119	80	21	220
	Activation	363	50	34	447
	Education	23	12	4	39
	Total	505	142	59	706

Is there an association between estimated working capacity and the type of training? If so, investigate which combinations are over/under-represented.

EX 117 During a severe epidemic 40 % of the population were on sick leave. A telephone survey to five randomly chosen institutions at a University gave the following result:

Institution no	1	2	3	4	5
Number on sick leave	4	10	8	2	6
Total number of employees	10	42	25	11	12

Test whether University employees are on sick leave to the same extent as the rest of the population.

EX 118 In a factory there were 10 accidents during 2 weeks. After this equipment were renewed and during the following 3 weeks there were 5 accidents. Did the measures have a significant effect on the rate of accidents?

[Hint: Use the conditional Poisson property and compute p-values from the Binomial distribution.]

EX 119 The conditional Poisson property can be generalized to *k* independent Poisson processes of rates $\lambda_1 \ldots \lambda_k$:

$P\left(Y_1(t_1)=y_1,\ldots,Y_k(t_k)=y_k \middle| \sum Y_i(t_i)=n\right)=\frac{n!}{y_1!\ldots y_k!}p_1^{y_1}\ldots p_k^{y_k}$, where $p_i=\frac{\lambda_i t_i}{\sum \lambda_i t_i}, i=1\ldots k$. The hypothesis $H_0:\lambda_1=\ldots=\lambda_k(=\lambda)$ is thus equivalent with $H_0: p_i=\frac{t_i}{\sum t_i}, i=1\ldots k$. It follows that the latter hypothesis can be tested by $X^2=\sum\frac{\left(\ Y_i-n(t_i/\sum t_i)\ \right)^2}{n(t_i/\sum t_i)}=\sum\frac{(Y_i-n/k)^2}{n/k}$ for a Poisson distribution (cf. Stuart *et al* 1999, p. 393).

Test whether the data 19 16 20 25 are observations on the same Poisson distributed variable.

EX 120 Let $(Y_i)_{i=1}^{n=10}$ be iid with $Y_i \sim Exponential(\lambda)$. Derive the RR for testing $H_0:\lambda=\lambda_0$ against $H_a:\lambda\neq\lambda_0$ at the 5% level. Consider the case *n* = 10.

[Hint: Use the Neyman-Pearson Lemma, mentioned in Ch. 6.2.3 and the property (5) in Ch. 2.2.2]

EX 121 Let $(X_i)_{i=1}^{n_X}$ and $(Y_i)_{i=1}^{n_Y}$ be two independent sets of iid variables with Exponential distributions with parameters λ_X and λ_Y , respectively. Show how the LR principle can be used to test $H_0:\lambda_X=\lambda_Y(=\lambda)$. Perform the test when $n_X=40, \sum x_i=20, n_Y=60, \sum y_i=40$.

EX 122 Let $(Y_i)_{i=1}^{n}$ be iid variables where $Y_i \sim Geometric(p)$.

a) Show how to test $H_0: p=1/2$ against $H_a: p\neq 1/2$ by means of the LR principle.
b) Perform the test when $n=50$ and $\sum y_i=80$.
c) Give examples where this test may be of interest.

EX 123 16 persons participated in a weight loss program. The body weight (in kg) of each person was measured initially (X) and after six months (Y). The following values were obtained of the difference D=X – Y (in increasing order):

-1.8, -1.7, -1.4, 0.3, 0.6, 1.6, 1.7, 1.9, 2.3, 2.4, 2.8, 3.7, 4.5, 5.8, 6.3, 6.8

Let $\mu_D=E(D)$ and test $H_0:\mu_D=0$ against $H_a:\mu_D\neq 0$

a) By assuming that differences are normally distributed.
b) By performing an exact Sign test based on the Binomial distribution.
c) By using a normality approximation of the test in b).

[Hint: In the last case the approximation can be improved by letting
$P(Y\leq y)\approx P\left(Z<\frac{y-E(Y)+1/2}{\sqrt{V(Y)}}\right)$ and $P(Y<y)\approx P\left(Z<\frac{y-E(Y)-1/2}{\sqrt{V(Y)}}\right)$]

EX 124 The same products were classified as Bad or God by Municipal- and State authorizes. The result was

		State: Bad	State: God
Municipal	Bad	20	10
	God	20	50

a) Test whether the classifications agree by testing equality between marginal frequencies. [Proportion Bad will suffice.]
b) Test whether the classifications from the two authorities are independent.

EX 125 Yeast cells were counted in a hemacytometer with the following result:

Number of yeast cells per square	0	1	2	3	4	5	6
Frequency	103	143	98	42	8	4	2

Check whether the frequencies are in accordance with the Poisson distribution

EX 126 At an industry men are working in three shifts: Morning, Day and Night. From each shift a random sample of 200 products were chosen and the number of defective products was recorded with the following result: 12 for Morning, 10 for Day and 23 for Night.

Use a Chi-square test to draw conclusions from the data.

[Hint: Construct a 2 x 3 table and test for independency.]

EX 127 Use the test for a difference between two Binomial proportions (Cf. EX 85.) to draw conclusions from the data in the preceding example. You may have to adjust the p-values for multiple comparisons (Cf. Ch. 6.4.).

EX 128 In a study it was found that 41 of 248 identical twins were left-handed and that 18 of 246 fraternal twins were left-handed. Is the difference significant?

EX 129 In the middle of 1950 the SALK vaccine against polio was tested in USA in several multi-center studies. In one such study 20 000 children were vaccinated and among these 1 case of polio was detected, compared with 114 cases of polio among 473 000 unvaccinated children. What conclusion can you draw about the effect of the vaccine?

EX 130 In a sample of 300 families the standard of the electronic equipment was classified as Cheap or Expensive. The families were also classified according to social class as Low-Middle-High. The result was

		Class: Low	Class: Middle	Class: High	Total
Standard	Cheap	38	88	31	160
	Expensive	62	42	39	140
	Total	100	130	70	300

Test whether Standard of equipment and Social class is independent and if not, try to find some significant patterns.

EX 131 In 2012 the yearly incidence of malignant melanoma in Sweden was 35 cases per 100 000 person. The same year 60 cases was observed in the city of Malmö in southern Sweden, with 110 000 inhabitants. Does this indicate that inhabitants in Malmö had a significantly higher risk for malignant melanoma than people in the rest of Sweden?

EX 132 A dealer takes a sample of 200 oranges from a large batch from his importer. He notices that 19 of these are of bad quality while the rest are acceptable. In a delivery from a new importer he finds that 10 oranges of 200 are bad while the rest are acceptable. Shall he prefer the new importer?

a) Discuss whether a one-sided or a two-sided test is preferably.
b) Test whether the proportion bad oranges is the same with the former and with the new importer.
c) Use the ordinary Chi-square test to test for independency between Quality of oranges and Importer. Compare the result with b).

EX 133 In a school with 156 pupils 90 were offered vaccination against a certain disease. After half a year the effect of the vaccination were studied, with the following result:

	Diseased	Not diseased	Total
Vaccinated	4	86	90
Not vaccinated	18	48	66
Total	22	134	156

Did the vaccination have a significant effect?

[Hint: Repeat the arguments that was given in the preceding example a)–c).

EX 134 A certain kind of surgical treatment may lead to complications. A comparison of two methods gave the following frequencies:

	Old method	New method	Total
Complications	15	1	16
No complications	83	46	129
Total	98	47	145

Is the new method better than the old method?

a) Use the Chi-square test for independence.
b) Use Fisher's exact test.

EX 135 A comparison of life lengths (hours) between two types of bulbs gave the following result:

	Sample size	Mean	Stand. Dev.
Type A	20	1128	62
Type B	20	1236	83

a) Test whether there is a difference in quality between the two types of bulbs under the assumption that life lengths are normally distributed.
b) Repeat the test in a) but now without assuming that life lengths are normally distributed. [Hint: Use the CLT.]

EX 136 Two varieties of wheat A and B were grown in 12 different areas. The yield is summarized in the following table:

Area	1	2	3	4	5	6	7	8	9	10	11	12
A	24	16	21	24	26	12	17	21	25	19	29	22
B	21	17	20	25	21	13	15	19	21	22	24	22

Assume that yield in an area is normally distributed and test whether the difference in yield is significant.

EX 137 We have seen how the Chi-square principle can be used for a variety of tests. Here is another example, called the *Median test*. (There are several versions of this test.)

The *population median* is defined in Ch. 2.1. As an estimator of this one may take the *sample median*, defined as the middle point of the ranked data in a sample, or the average of the two middle points in case of an even sample size.

Test whether the following two series of data come from populations with the same median.

A	44	40	46	22	51	41	48	38	58	60	28	40
B	24	54	80	35	36	23	15	21	43	18	12	29

[Hint: Rank the observations in each series, compute the sample median m for the combined series and count the number of observations that are above or below m in each of the two series. Then you summarize the result in a 2 x 2 table with frequencies of the two variables (above m /below m) and (Series A/ Series B). The classical Chi-square test will then give you the answer.]

EX 138 *A beer-tasting Binomial experiment*. 4 glasses of beer A and 4 glasses of beer B are served sequentially in random order to a person who has to decide which of the beers he is tasting.

a) How would you in practice arrange the experiment so that the variable *Y* = 'Number of correct answers' $\sim Binomial(n=8, p=1/2)$? Why assuming that $p=1/2$?
b) Assume that the person gives correct answers in *y* cases. What is the smallest value of *y* for which the hypothesis $H_0 : p=1/2$ is rejected by a one-sided test at the 5% level?
c) When people are appointed as professional tasters of beer, coffee, tea, etc. they have to undergo tests in very long series. Assume that the person has to compare 50 glasses of beer of each kind instead of 4. Which is the smallest number of correct answers required to reject the hypothesis in b)?

EX 139 In 160 families with four children the number of boys (*Y*) was:

Y	0	1	2	3	4
Frequency	6	38	58	47	11

Test whether the frequencies are in agreement with a variable that is distributed $Binomial(n=4, p=0.516)$.
[Hint: See EX 4.]

a) By using the Kolmogorov test.
b) By using a Chi-square test.

EX 140 Independent measurements of viscosity for a certain substance were measured during two days with the following result:

Day1: 37.0 31.4 34.4 33.3 34.9 36.2 31.0 33.5 33.7 33.4 34.8 30.8 32.9 34.3 33.3
Day2: 28.4 31.3 28.7 32.1 31.9 32.8 30.2 30.2 32.4 30.7

Has the population distribution changed from one day to the next?

a) Use the Smirnov two-sample test. [Hint: Rank the observations within each sample and estimates the two sample cdf's as described in CH. 6.2.4. Then search for the largest difference between the two cdf's. It could help to make a plot.]
b) Compare the result in a) with the result that is obtained by assuming that both series are normally distributed. Conclusions?

EX 141 Repeat the analysis of the two varieties of wheat in EX 136 by using the sign test.

EX 142 15 student were ranked according to their results in Mathematics and Statistics with the following result:

Math.	3	5	1	12	10	8	6	9	2	15	13	7	11	4	14
Stat.	2	1	3	15	12	5	9	4	6	13	14	10	7	8	11

Is there a significant association between the two series? [Hint: Compare with EX 98.]

EX 143 In a clinical trial one wants to study the effect of a drug on the concentration of a substance in blood. In a pilot study where the concentration was measured before and after the drug was added, the following result was obtained:

Before	1.10	0.98	0.95	0.99	1.05	1.20	0.96	1.07	0.96	1.06
After	1.05	0.95	0.90	0.98	1.01	1.08	0.94	1.08	0.98	1.04

The hypothesis of interest is $H_0 : \mu_D = 0$ against $H_a : \mu_D \neq 0$.

a) Test the hypotheses by means of a p-value argument under the assumption that the mean difference is normally distributed. Give reasons for the assumption.
b) Let the estimates from the pilot study represent the true population parameters and assume that the variance of the difference is the same under H_0 and H_a determine a rejection region (RR) as a function of the sample size *n* under normality assumptions. The type-I error is 0.05. [Hint: Cf. Ch. 6.3.]
c) Study the power function for the test in b). For which values of μ_D is the power larger than 0.90 when *n* = 100?

EX 144 In 2003 it was found that the proportion disabled (*p*) who were full-time workers in service profession was 8%. Ten years later it was decided to plan a study to see whether this proportion had changed. In a sequentially collected sample it was found that the proportion stabilized around 3/20 = 0.15.

The hypothesis to test is $H_0 : p = 0.08$ against $H_a : p \neq 0.08$. Determine the sample size, *n*, required to get a power of at least 0.90 when *p* = 0.15. Also, determine the rejection region (RR).

EX 145 $(Y_i)_{i=1}^n$ are iid with $Y_i \sim N(\mu,\sigma^2)$. One wants to test $H_0 : \sigma^2 = \sigma_0^2$ against $H_a : \sigma^2 \neq \sigma_0^2$ based on the test statistic $T = (n-1)S^2 = \sum (Y_i - \overline{Y})^2$ with a type-I error of 5%.

a) Specify a RR of the form $T < a$ or $T > b$ and derive the power function.
b) Let *n* = 10 and study the power as a function of $R = \sigma_0^2 / \sigma^2$.

EX 146 In EX 120, where iid observations were distributed $Exponential(\lambda)$, we determined the RR for testing $H_0 : \lambda = \lambda_0$ against $H_a : \lambda \neq \lambda_0$. Express the power of this test as a function of the ratio

$R = \lambda / \lambda_0$. For which values of *R* are the power larger than 0.90?

EX 147 A new rapid method to measure concentration of a certain substance was tested against an exact method with the following result:

Exact method (*X*)	1	2	3	4	5	6
New method (*Y*)	1.2	1.9	3.1	4.2	4.7	5.9

a) Apply the model $E(Y|x) = \alpha + \beta \cdot x$ and test the hypotheses $H_0 : \beta = 1$ against $H_a : \beta \neq 1$ and $H_0 : \alpha = 0$ against $H_a : \alpha \neq 0$. [Hint: Cf. EX 104.]
b) If 'non-significant parameters' appear in a) formulate an alternative model and perform the test.

EX 148 The concentration of a substance in blood (*Y*) was measured and compared with a known concentration (*X*). The following result was obtained:

X	*Y*
1	1.1 0.7 1.8 0.4
3	3.0 1.4 4.9 4.4 4.5
5	7.3 8.2 6.2
10	12.0 13.1 12.6 13.2
15	18.7 19.7 17.4 17.1

Test whether the model $E(Y|X = x) = \alpha + \beta \cdot x$ is adequate. [Hint: Cf. Ex105.]

EX 149 In an experiment one studied the relation between *x* = Temperature in minus degrees Celsius needed to reach the freezing point and *Y* = Concentration of an alcohol at which the freezing point was reached. The result was:

x	0.5	1	4	16
Y	1.2 1.5	1.9 2.1	3.9 4.1	7.8 8.2

Can the relation be described by a linear regression function?

[WARNING! The alcohol is not ethyl so don't put a bottle of Champagne (12–13%) in your freeze of about – $-16^o C$.]

EX 150 In the preceding example the linear model had to be abandoned. Plot the data and try to find a non-linear model that fits the data better.

[Hint: Try some of the non-linear models in Ch. 2.3.1 and test whether the linearized version can be accepted.]

EX 151 The body weight (*Y*) was recorded for 64 men and women after a diet period. Let *x* be the initial weight and let *z* be a variable taking the value 1 for men and 0 for women. By running the model $E(Y|x,z)=\alpha+\beta_1 x+\beta_2 z+\beta_3 x\cdot z$ one obtained the Sum of Squares for Error *SSE* = 18.4381. With the model $E(Y|x)=\alpha+\beta_1 x$ the Sum of Squares increased to *SSE* = 18.7454.

Formulate and test relevant hypotheses and draw conclusions.

[Hint: See EX 107.]

EX 152 A frequently used relation in econometrics is the production function $Q(I,P)=\alpha\cdot I^{\beta_1}P^{\beta_2}$, where *Q* = consumed quantity, *I* = income level of prospective consumers and *P* = price of the commodity. The parameters β_1 and β_2 are interpreted as Income elasticity and Price elasticity, respectively. Estimate the parameters in the linearized model from the following data:

(*I* =Total domestic private consumption (million SEK), *Q* = Yearly consumed quantity of strong beer (million liter), *P* = Total price of strong beer (million SEK). All prices in 1988 monetary value.)

Year	*I*	*P*	*Q*
-73	421 027	778.4	24.1
-74	437 067	754.5	25.0
-75	453 748	770.7	24.9
-76	472 681	756.3	24.5
-77	485 582	1321.1	44.5
-78	491 919	2193.8	76.6
-79	507 296	2494.6	89.7
-80	497 081	2640.8	91.8
-81	489 929	2669.9	90.4
-82	506 769	2840.3	99.5
-83	507 822	3070.2	101.1
-84	515 257	3286.3	104.9
-85	527 904	3645.2	105.8
-86	554 850	4196.2	120.1
-87	584 427	4746.3	131.5
-88	563 293	5018.0	147.9

[Hint: Use a computer!]

EX 153 Time to recovery after a certain disease vary according to the Weibull distribution with survival function $S(y) = e^{-\lambda y^{\alpha}}$. (Cf. Ch. 2.2.2.) Test $H_0 : \alpha = 1$ against $H_a : \alpha \neq 1$ based on the following data of the proportion patients that are recovered at *y* years which is used as an estimator of $S(y)$, $\hat{S}(y)$.

y	0.47	0.64	0.89	1.08	1.41
$\hat{S}(y)$	5/6	4/6	3/6	2/6	1/6

[Hint: Linearize the survival function and use the techniques for analyzing linear models in Ch. 6.6.1.]

Answers to Supplementary Exercises

Solutions to Supplementary Exercises Ch. 3

EX 28 $F_{Y_i}(y)=\frac{y}{b}, 0\le y\le b \Rightarrow F_{Y_{(1)}}(y)=1-\left[1-\frac{y}{b}\right]^n$. From derivation rule *(5)* in Ch. 2.3.3,

$$f_{Y_{(1)}}(y)=F_{Y_i}'(y)=\left(-\frac{1}{b}\right)\left(0-n\left[1-\frac{y}{b}\right]^{n-1}\right)=\frac{n}{b}\left[1-\frac{y}{b}\right]^{n-1}.$$

EX 29 $F_{Y_i}(y)=1-e^{-\lambda\cdot y}, y\ge 0 \Rightarrow F_{Y_{(n)}}(y)=\left[1-e^{-\lambda\cdot y}\right]^n$. From derivation rule *(5)* in Ch. 2.3.3,

$$f_{Y_{(n)}}(y)=F_{Y_{(n)}}'(y)=\left(-(-\lambda)e^{-\lambda\cdot y}\right)n\left[1-e^{-\lambda\cdot y}\right]^{n-1}=n\lambda e^{-\lambda\cdot y}\left[1-e^{-\lambda\cdot y}\right]^{n-1}.$$

EX 30 *Exact mean*: $E\left(\hat{p}(1-\hat{p})/n\right)=\frac{1}{n}E\left(\hat{p}-\hat{p}^2\right)=\frac{1}{n}\left(E(\hat{p})-E(\hat{p}^2)\right)=\frac{1}{n}\left(E(\hat{p})-\left[V(\hat{p})+E^2(\hat{p})\right]\right)=$

$$\frac{1}{n}\left(p-\frac{p(1-p)}{n}-p^2\right)=\frac{1}{n}p(1-p)\left(1-\frac{1}{n}\right)=\frac{(n-1)}{n^2}p(1-p)$$

Approximate mean:

First we notice that $g(p)=p-p^2 \Rightarrow g'(p)=1-2p, g''(p)=-2$. (13) with $\mu=p, \sigma^2=\frac{p(1-p)}{n}$ gives:

$$E\left(\hat{p}(1-\hat{p})/n\right)=\frac{1}{n}E\left(\hat{p}-\hat{p}^2\right)\approx\frac{1}{n}\left(p-p^2+\frac{1}{2}(-2)\frac{p(1-p)}{n}\right)=$$

$$\frac{(n-1)}{n^2}p(1-p).$$

It follows that $\hat{V}(\hat{p})=\frac{\hat{p}(1-\hat{p})}{(n-1)}$ is unbiased for $V(\hat{p})=\frac{p(1-p)}{n}$.

EX 31 $\hat{p}_i, i=1,2$ have $\mu_i=p_i$ and $\sigma_i^2=p_i(1-p_i)/n_i$ and $\sigma_{12}=0$ (due to independence). Thus we get from

EX 27: $V(\hat{R})\approx\left(\frac{p_1}{p_2}\right)^2\left[\frac{p_1(1-p_1)/n_1}{p_1^2}+\frac{p_2(1-p_2)/n_2}{p_2^2}+0\right]=\left(\frac{p_1}{p_2}\right)^2\left[\frac{1-p_1}{n_1p_1}+\frac{1-p_2}{n_2p_2}\right]$.

The expression for the estimated variance is finally obtained by replacing p_i by Y_i/n_i.

EX 32 From (9b) we know that $\mu = E(S^2) = \sigma^2$ and (don't use σ^2 in the following to avoid confusion)

$$V(S^2) = \left(\mu_4 - \frac{(n-3)}{(n-1)}\mu_2^2\right)\frac{1}{n} = \left[\text{If } Y_i \sim N(\mu,\sigma^2)\right] = \frac{2\sigma^4}{(n-1)}.$$ The latter relation follows since

$$S^2 \sim \frac{\sigma^2}{(n-1)}\chi^2(n-1) \Rightarrow V(S^2) = \frac{\sigma^4}{(n-1)^2}V(\chi^2(n-1)) = \frac{\sigma^4}{(n-1)^2}2(n-1).$$

Since $S = \sqrt{S^2} = (S^2)^{\frac{1}{2}}$ we consider the function $g(y) = y^{\frac{1}{2}}$ with $g'(y) = \frac{1}{2y^{1/2}}$ and $g''(y) = \frac{-1}{4y^{3/2}}$. Thus,

a) $$E(S) \approx (\sigma^2)^{\frac{1}{2}} + \frac{1}{2}\left(\frac{-1}{4(\sigma^2)^{3/2}}\right)V(S^2) = \sigma - \frac{V(S^2)}{8\sigma^3}$$

$$V(S) \approx \left(\frac{1}{2(\sigma^2)^{1/2}}\right)^2 V(S^2) = \frac{V(S^2)}{4\sigma^2}$$

b) $$E(S) \approx \sigma - \frac{1}{8\sigma^3}\frac{2\sigma^4}{(n-1)} = \sigma - \frac{\sigma}{4(n-1)}$$

$$V(S) \approx \frac{1}{4\sigma^2}\frac{2\sigma^4}{(n-1)} = \frac{\sigma^2}{2(n-1)}$$

Ch. 4

EX 45

a) $F(y)=\frac{y}{b}, 0\le y\le b$ and $F_{Y_{(n)}}(y)=\left(F(y)\right)^n \Rightarrow F_{Y_{(n)}}(y)=\frac{y^n}{b^n} \Rightarrow f_{Y_{(n)}}(y)=\frac{ny^{n-1}}{b^n}$

$$E(Y_{(n)})=\int_0^b y\cdot\frac{ny^{n-1}}{b^n}dy=\frac{n}{b^n}\int_0^b y^n dy=\frac{n}{b^n}\left[\frac{y^{n+1}}{n+1}\right]_{y=0}^{y=b}=\frac{n}{b^n}\cdot\frac{b^{n+1}}{(n+1)}=\frac{n}{(n+1)}b \Rightarrow$$

$\hat{b}=\frac{(n+1)}{n}Y_{(n)}$ is unbiased for *b*.

$V(\hat{b})=\frac{(n+1)^2}{n^2}V(Y_{(n)})$. Now, $E(Y_{(n)}^2)=\int y^2\cdot\frac{ny^{n-1}}{b^n}dy=\frac{n}{(n+2)}\cdot b^2$, so

$$V(\hat{b})=\frac{(n+1)^2}{n^2}\left[\frac{n}{(n+2)}\cdot b^2-\left(\frac{n}{(n+1)}b\right)^2\right]=\frac{(n+1)^2}{n}\cdot b^2\left[\frac{1}{(n+2)}-\frac{n}{(n+1)^2}\right]=\frac{b^2}{n(n+2)}$$

b) *OLS estimator*: $SS=\sum_{i=1}^n\left[Y_i-\frac{b}{2}\right]^2 \Rightarrow \frac{dSS}{db}=\sum_{i=1}^n(-\frac{1}{2})2\left[Y_i-\frac{b}{2}\right]=0 \Rightarrow \sum_{i=1}^n Y_i=\frac{nb}{2} \Rightarrow$

$\hat{b}_{OLS}=2\sum_{i=1}^n Y_i/n=2\overline{Y}$. This is easily seen to be unbiased for *b*. The variance is

$$V(\hat{b}_{OLS})=\left(\frac{2}{n}\right)^2\sum_{i=1}^n V(Y_i)=\left[\text{Cf. } 2.2.2\right]=\frac{4}{n^2}\cdot n\frac{b^2}{12}=\frac{b^2}{3n}.$$

Moment estimator: $\overline{Y}=b/2 \Rightarrow \hat{b}_{Mom}=2\overline{Y}=\hat{b}_{OLS}$.

a) Relative efficiency (RE) = $\frac{V(\hat{b})}{V(\hat{b}_{OLS})}=\frac{b^2/n(n+2)}{b^2/3n}=\frac{3}{n+2}\le 1$ with equality if $n = 1$. In the latter case $\hat{b}$ and $\hat{b}_{OLS}$ are identical.

b) *b* = Length of each red light period. From the data we obtain $y_{(n)}=46$ and $\overline{y}=20.0$, so the estimates are $\hat{b}=\frac{(10+1)}{10}\cdot 46=50.6$ and $\hat{b}_{OLS}=\hat{b}_{Mom}=2\cdot 20.0=40.0$. Of these two the first one should be more reliable since RE in this case is 1/4.

EX 46

a) *OLS estimator*: Remember that $E(Y_i) = n_i p$ and $V(Y_i) = n_i p(1-p)$.

$$SS = \sum_{i=1}^{n}[Y_i - n_i p]^2 \Rightarrow \frac{dSS}{dp} = \sum_{i=1}^{n}(-n_i)2[Y_i - n_i p] = 0 \Rightarrow \sum_{i=1}^{n} n_i^2 p - \sum_{i=1}^{n} n_i Y_i = 0 \Rightarrow \hat{p}_{OLS} = \frac{\sum_{i=1}^{n} n_i Y_i}{\sum_{i=1}^{n} n_i^2}$$

$$E(\hat{p}_{OLS}) = \frac{1}{\sum_{i=1}^{n} n_i^2}\sum_{i=1}^{n} n_i E(Y_i) = \frac{1}{\sum_{i=1}^{n} n_i^2}\sum_{i=1}^{n} n_i \cdot n_i p = p \text{ (Unbiased.)}$$

$$V(\hat{p}_{OLS}) = \frac{1}{\left(\sum_{i=1}^{n} n_i^2\right)^2}\sum_{i=1}^{n} n_i^2 V(Y_i) = \frac{1}{\left(\sum_{i=1}^{n} n_i^2\right)^2}\sum_{i=1}^{n} n_i^2 \cdot n_i p(1-p) = p(1-p)\frac{\sum_{i=1}^{n} n_i^3}{\left(\sum_{i=1}^{n} n_i^2\right)^2}$$

ML estimator:

$$L = \prod_{i=1}^{n}\binom{n_i}{y_i} p^{y_i}(1-p)^{n_i - y_i} = C \cdot p^{\sum_{i=1}^{n} y_i}(1-p)^{\sum_{i=1}^{n} n_i - \sum_{i=1}^{n} y_i} \Rightarrow$$ [*C* is a constant which doesn't contain *p*]

EX 46 (Continued)

$$\ln L = \ln C + \sum_{i=1}^{n} y_i \ln p + \left(\sum_{i=1}^{n} n_i - \sum_{i=1}^{n} y_i\right)\ln(1-p) \Rightarrow \frac{d\ln L}{dp} = 0 + \frac{\sum_{i=1}^{n} y_i}{p} + \frac{\left(\sum_{i=1}^{n} n_i - \sum_{i=1}^{n} y_i\right)(-1)}{(1-p)}$$

and putting this equal to zero yields $\hat{p}_{ML} = \sum_{i=1}^{n} y_i / \sum_{i=1}^{n} n_i$

$$E(\hat{p}_{ML}) = \frac{1}{\sum_{i=1}^{n} n_i}\sum_{i=1}^{n} E(Y_i) = \frac{1}{\sum_{i=1}^{n} n_i}\sum_{i=1}^{n} n_i p = p \text{ (Unbiased.)}$$

$$V(\hat{p}_{ML}) = \frac{1}{\left(\sum_{i=1}^{n} n_i\right)^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{\left(\sum_{i=1}^{n} n_i\right)^2}\sum_{i=1}^{n} n_i p(1-p) = \frac{p(1-p)}{\sum_{i=1}^{n} n_i}$$

Comparison: $RE = \dfrac{V(\hat{p}_{ML})}{V(\hat{p}_{OLS})} = \dfrac{\left(\sum_{i=1}^{n} n_i^2\right)^2}{\sum_{i=1}^{n} n_i \cdot \sum_{i=1}^{n} n_i^3} \leq 1$. This follows from Cauchy-Schwartz inequality (Cf. Ch. 2.3.5) by letting $x_i = n_i^{\frac{1}{2}}$ and $y_i = n_i^{\frac{3}{2}}$. (It may suffice if one demonstrates the inequality numerically by choosing a few values of n_i.)

(From now on we skip the upper and lower index in the summation signs.)

b) To show that the ML estimator is BLUE, put $T_n = \sum a_i Y_i \Rightarrow E(T) = \sum a_i E(Y_i) =$

$\sum a_i \cdot n_i p = p\sum a_i n_i = (\text{Put}) = p \Rightarrow \sum a_i n_i - 1 = 0$ (1)

$$Q = V(T_n) + \lambda\left[\sum a_i n_i - 1\right] = \sum a_i^2 V(Y_i) + \lambda\left[\sum a_i n_i - 1\right] = \sum a_i^2 \cdot n_i p(1-p) + \lambda\left[\sum a_i n_i - 1\right] \Rightarrow$$

$$\frac{dQ}{da_i} = 2a_i n_i p(1-p) + \lambda n_i = 0 \Rightarrow a_i = -\frac{\lambda}{2p(1-p)} = \lambda',$$ say. (2) Putting this into (1) gives $\lambda' = 1/\sum n_i$

which inserted into (2) gives $a_i = 1/\sum n_i$. Thus, $\hat{p}_{BLUE} = \sum Y_i / \sum n_i = \hat{p}_{ML}$.

To show that the ML estimator is a MVE we determine the C-R limit. From a) above,

$$\frac{d^2 \ln L}{dp^2} = \left[\text{Cf. derivation rules in Ch. 2.3.3}\right] = \sum y_i \left(\frac{0-1}{p^2}\right) - \left(\sum n_i - \sum y_i\right)\left(\frac{0-(-1)}{(1-p)^2}\right) =$$

$$-\frac{\sum y_i}{p^2} - \frac{\left(\sum n_i - \sum y_i\right)}{(1-p)^2} \Rightarrow \left[\text{Replacing } y_i \text{ by } Y_i\right] \Rightarrow -E\left(\frac{d^2 \ln L}{dp^2}\right) = \frac{\sum n_i p}{p^2} + \frac{\left(\sum n_i - \sum n_i p\right)}{(1-p)^2} =$$

$$\frac{\sum n_i}{p} + \frac{\sum n_i}{(1-p)} = \frac{\sum n_i}{p(1-p)} = I(p).$$ Thus, $V(\hat{p}_{ML}) = \frac{1}{I(p)}$ so the ML estimator is a MVE.

c) Let n_i = Total number of students in room *i*, Y_i = Number with back/neck pain in room *i*

From the data we get $\sum n_i = 90,\ \sum n_i^2 = 2750,\ \sum y_i = 6,\ \sum n_i y_i = 175$. This leads to the following estimates: $\hat{p}_{OLS} = 175/2750 = 0.064\ (6.4\%),\ \hat{p}_{ML} = 6/90 = 0.067\ (6.7\%)$.

EX 47

$$L = \prod (1-p)^{y_i - 1} p = (1-p)^{\sum y_i - n} p^n \Rightarrow \ln L = (\sum y_i - n)\ln(1-p) + n \ln p \Rightarrow$$

$$\frac{d \ln L}{dp} = (\sum y_i - n)\frac{(-1)}{(1-p)} + \frac{n}{p} = 0 \Rightarrow \hat{p}_{ML} = \frac{n}{\sum y_i} = \frac{1}{\bar{y}}$$

Define the variable Y_i = Number of trials until the first '1' appears. Then *n* = 3 and $y_1 = 4, y_2 = 1, y_3 = 3$.

Thus $\hat{p}_{ML} = \frac{3}{8} = 0.375.$

EX 48

a) $$L = \prod \left(2\pi\sigma^2\right)^{-\frac{1}{2}} e^{-\frac{(y_i - \beta x_i)^2}{2\sigma^2}} = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{\sum (y_i - \beta x_i)^2}{2\sigma^2}} \Rightarrow$$

$$\ln L = -\frac{n}{2}\ln\left(2\pi\sigma^2\right) - \frac{\sum \left(y_i - \beta x_i\right)^2}{2\sigma^2} \Rightarrow \frac{d \ln L}{d\beta} = 0 - \frac{\sum (-x_i) 2\left(y_i - \beta x_i\right)}{2\sigma^2} = 0 \Rightarrow \hat{\beta}_{ML} = \frac{\sum x_i Y_i}{\sum x_i^2}$$

EX 48 (continued)

$\hat{\beta}_{ML}$ is simply the BLUE estimator that was found in EX 30, without any distributional assumptions.

In Ch. 2.2.2 *(4)* it was stated in the comments to the normal distribution that a linear form of normally distributed variables is itself normally distributed. Applying this to $\hat{\beta}_{ML} = \sum l_i Y_i$, with $l_i = \frac{x_i}{\sum x_i^2}$, shows that $\hat{\beta}_{ML}$ is exactly normally distributed with $E(\hat{\beta}_{ML}) = \frac{\sum x_i \cdot \beta x_i}{\sum x_i^2} = \beta$ and $V(\hat{\beta}_{ML}) =$

$$\frac{\sum x_i^2 \cdot \sigma^2}{\left(\sum x_i^2\right)^2} = \frac{\sigma^2}{\sum x_i^2}.$$

b) From the expression for ln L in a) we get

$$\frac{d \ln L}{d\sigma^2} = \left[\text{The derivative is with repect to } \sigma^2, \text{ not } \sigma!\right] = \frac{-n}{2\sigma^2} - \frac{\sum (y_i - \beta x_i)^2}{2} \cdot \left(-\frac{1}{\sigma^4}\right) = 0 \Rightarrow$$

$\hat{\sigma}^2{}_{ML} = \frac{\sum (Y_i - \hat{\beta}_{ML} x_i)^2}{n}$. [Notice that the ML estimator has been inserted for β.] We now determine the distribution of this by using Cochran's theorem (7) in Ch. 3.1.

$$\sum (Y_i - \beta x_i)^2 = \sum \left((Y_i - \hat{\beta}_{ML} x_i) + (\hat{\beta}_{ML} x_i - \beta x_i)\right)^2 = \sum (Y_i - \hat{\beta}_{ML} x_i)^2 + \sum (\hat{\beta}_{ML} x_i - \beta x_i)^2$$

since the cross product vanishes. In fact

$$2\sum (Y_i - \hat{\beta}_{ML} x_i)(\hat{\beta} x_i - \beta x_i) = 2(\hat{\beta}_{ML} - \beta)\sum (Y_i - \hat{\beta}_{ML} x_i) x_i = 2(\hat{\beta}_{ML} - \beta)\left(\sum x_i Y_i - \hat{\beta}_{ML} \sum x_i^2\right) =$$

$2(\hat{\beta}_{ML} - \beta)\left(\sum x_i Y_i - \sum x_i Y_i\right) = 0$. Thus, $\sum \left(Y_i - \beta x_i\right)^2 = \sum \left(Y_i - \hat{\beta}_{ML} x_i\right)^2 + (\hat{\beta}_{ML} - \beta)^2 \sum x_i^2$.

Divide each term in the latter expression by σ^2 and write the identity as $Q_1 = Q_2 + Q_3$.

We now find the distributions of Q_1 and Q_3.

$$Y_i \sim N(\beta x_i, \sigma^2) \Rightarrow \frac{(Y_i - \beta x_i)}{\sigma} \sim N(0,1) \Rightarrow \frac{(Y_i - \beta x_i)^2}{\sigma^2} \sim \chi_i^2(1) \Rightarrow Q_1 = \frac{\sum (Y_i - \beta x_i)^2}{\sigma^2} \sim \chi^2(n)$$

$$\hat{\beta}_{ML} \sim N\left(\beta, \frac{\sigma^2}{\sum x_i^2}\right) \Rightarrow \frac{(\hat{\beta}_{ML} - \beta)}{\sqrt{\sigma^2 / \sum x_i^2}} \sim N(0,1) \Rightarrow Q_3 = \frac{(\hat{\beta}_{ML} - \beta)^2}{\sigma^2 / \sum x_i^2} \sim \chi^2(1)$$

From Cochran's theorem it follows that $Q_2 = \frac{\sum \left(Y_i - \hat{\beta}_{ML} x_i\right)^2}{\sigma^2} \sim \chi^2(n-1) \Rightarrow$

$$S^2 = \frac{\sum \left(Y_i - \hat{\beta}_{ML} x_i\right)^2}{(n-1)} \sim \frac{\sigma^2}{(n-1)} \cdot \chi^2(n-1) \text{ with } E(S^2) = \frac{\sigma^2}{(n-1)} E\left(\chi^2(n-1)\right) = \frac{\sigma^2 (n-1)}{(n-1)} = \sigma^2.$$

So, $\hat{\sigma}^2{}_{ML}$ is not unbiased, but the corrected estimator S^2 is unbiased.

EX 49

BLUE of μ. $T_n = \sum a_i \overline{Y}_i$ has $E(T_n) = \sum a_i E(\overline{Y}_i) = \sum a_i \mu = \mu \sum a_i = (\text{Put}) = \mu \Rightarrow$

$\sum a_i - 1 = 0$ (1)

$V(T_n) = \sum a_i^2 V(\overline{Y}_i) = \sum a_i^2 \sigma^2 / n_i = \sigma^2 \sum a_i^2 / n_i$. Thus, $Q = V(T_n) + \lambda\left[\sum a_i - 1\right] =$

$$\sigma^2 \sum a_i^2 / n_i + \lambda\left[\sum a_i - 1\right] \Rightarrow \frac{dQ}{da_i} = 2\sigma^2 a_i / n_i + \lambda = 0 \Rightarrow a_i = -\frac{\lambda n_i}{2\sigma^2} = \lambda' n_i \quad (2)$$

(2) inserted into (1) gives $\sum \lambda' n_i = \lambda' \sum n_i = 1 \Rightarrow \lambda' = 1 / \sum n_i$, which inserted into (2) gives

$a_i = n_i / \sum n_i$. So, $\hat{\mu}_{BLUE} = \sum n_i \overline{Y}_i / \sum n_i$.

$$V\left(\hat{\mu}_{BLUE}\right) = \frac{1}{\left(\sum n_i\right)^2} \sum n_i^2 V(\overline{Y}_i) = \frac{1}{\left(\sum n_i\right)^2} \sum n_i^2 \frac{\sigma^2}{n_i} = \frac{\sigma^2}{\sum n_i}$$

BLUE of σ^2. Notice first that $S_i^2 \sim \frac{\sigma^2}{(n_i - 1)} \chi^2(n_i - 1)$ (Cf. EX 16) $\Rightarrow E\left(S_i^2\right) =$

$$\frac{\sigma^2}{(n_i - 1)} E\left(\chi^2(n_i - 1)\right) = \frac{\sigma^2}{(n_i - 1)}(n_i - 1) = \sigma^2 \text{ and } V\left(S_i^2\right) = \frac{\sigma^4}{(n_i - 1)^2} V\left(\chi^2(n_i - 1)\right) =$$

$$\frac{\sigma^4}{(n_i - 1)^2} 2(n_i - 1) = \frac{2\sigma^4}{(n_i - 1)}.$$

Put $T_n = \sum a_i S_i^2 \Rightarrow E(T_n) = \sum a_i E\left(S_i^2\right) = \sigma^2 \sum a_i = (\text{Put}) = \sigma^2 \Rightarrow \sum a_i - 1 = 0$ (3)

$V(T_n) = \sum a_i^2 V\left(S_i^2\right) = 2\sigma^4 \sum a_i^2 \ (n_i - 1)$. Thus,

$$Q = V(T_n) + \lambda\left[\sum a_i - 1\right] = 2\sigma^4 \sum \frac{a_i^2}{(n_i - 1)} + \lambda\left[\sum a_i - 1\right] \Rightarrow \frac{dQ}{da_i} = \frac{4\sigma^4 a_i}{(n_i - 1)} + \lambda = 0 \Rightarrow$$

$a_i = -\frac{\lambda(n_i - 1)}{4\sigma^4} = \lambda'(n_i - 1)$ (4), which inserted into (3) yields

$\sum \lambda'(n_i - 1) = 1 \Rightarrow \lambda' = \frac{1}{\sum (n_i - 1)}$ and this inserted into (4) gives $a_i = \frac{(n_i - 1)}{\sum (n_i - 1)}$. So,

$$\hat{\sigma}^2_{BLUE} = \frac{\sum (n_i - 1) S_i^2}{\sum (n_i - 1)}$$

$$V\left(\hat{\sigma}^2_{BLUE}\right) = \frac{\sum (n_i - 1)^2 V(S_i^2)}{\left(\sum (n_i - 1)\right)^2} = \frac{\sum (n_i - 1)^2 \frac{2\sigma^4}{(n_i - 1)}}{\left(\sum (n_i - 1)\right)^2} = \frac{2\sigma^4}{\sum (n_i - 1)}$$

EX 50 In this example the cell probabilities are specified as hypothetical proportions. In Ch.6, where tests of hypothesis are considered, there will be more examples of this.

$L = C \cdot p^{y_1} (2p)^{y_2} (1-3p)^{y_3} \Rightarrow \ln L = \ln C + y_1 \ln p + y_2 (\ln(2) + \ln(p)) + y_3 \ln(1-3p) \Rightarrow$

$\frac{d \ln L}{dp} = 0 + \frac{y_1}{p} + \frac{y_2}{p} + y_3 \frac{(-3)}{(1-3p)} = 0 \Rightarrow (y_1 + y_2)(1-3p) = 3y_3 p \Rightarrow \hat{p}_{ML} = \frac{y_1 + y_2}{3(y_1 + y_2 + y_3)} =$

$\frac{y_1 + y_2}{3n}$. The corresponding ML estimator is $\hat{p}_{ML} = \frac{Y_1 + Y_2}{n}$.

$E(\hat{p}_{ML}) = \frac{1}{3n}\left(E(Y_1) + E(Y_2)\right) = \left[\text{Cf. Ch. 2.2.1 (6)}\right] = \frac{1}{3n}\left(np + n \cdot 2p\right) = p$ (Unbiased.).

$V(\hat{p}_{ML}) = \frac{1}{(3n)^2}\left(V(Y_1) + V(Y_2) + 2Cov(Y_1, Y_2)\right) = \frac{1}{9n^2}\left(np(1-p) + n \cdot 2p(1-2p) - 2n \cdot p \cdot 2p\right) =$

$\frac{np}{9n^2}(1 - p + 2(1-2p) - 4p) = \frac{p(1-3p)}{n}$.

We now find the C-R limit.

$$\frac{d^2 \ln L}{dp^2} = -\frac{y_1}{p^2} - \frac{y_2}{p^2} - 3y_3\left(\frac{0-(-3)}{(1-3p)^2}\right) = -\left(\frac{y_1 + y_2}{p^2} + \frac{9y_3}{(1-3p)^2}\right) \Rightarrow -E\left(\frac{d^2 \ln L}{dp^2}\right) =$$

$$E\left(\frac{Y_1 + Y_2}{p^2} + \frac{9Y_3}{(1-3p)^2}\right) = \frac{np + n \cdot 2p}{p^2} + \frac{9n(1-3p)}{(1-3p)^2} = \frac{3n}{p} + \frac{9n}{(1-p)} = \frac{3n}{p(1-3p)} = I(p)$$

Since $V(\hat{p}_{ML}) = 1/I(p)$ we conclude that the ML estimator is MVE.

Ch. 5

EX 63 Data consist of naturally paired observations that are strongly dependent. Therefore the approach in EX 53 can't be used. Instead we consider the weight-loss $D = X - Y$ for each subject which gives the series

subject	1	2	3	4	5	6	7	8	9	10
D	0.2	1.1	-0.1	0.3	-0.4	0.8	0.2	0.7	0.3	0.7

Here we get $n = 10, \hat{\mu}_D = 0.40, \hat{\sigma}_D^2 = 0.2022\ (\hat{\sigma}_D = 0.4497)$

a) The 95% CI for μ_D is $0.40 \pm C\frac{0.4497}{\sqrt{10}}$, where C is obtained from $P(T(9) > C) = 0.025 \Rightarrow C = 2{,}262$.

Thus, the 95% CI is (0.08, 0.72), so the conclusion is that the training program has a significant positive effect on weight-loss.

b) The 95% CI for σ^2 is (Cf. EX 43) $\left(\frac{0.2022 \cdot 9}{b}, \frac{0.2022 \cdot 9}{a}\right)$, where *a* and *b* are determined from $P(a < \chi^2(9) < b) = 0.95$. $\begin{cases} P(\chi^2(9) > b) = 0.025 \Rightarrow b = 19.0226 \\ P(\chi^2(9) > a) = 0.975 \Rightarrow a = 2.7004 \end{cases}$

This gives the CI (0.096, 0.674) and since the latter is far below 0.7 (the value for females) we conclude that the variation in weight-loss is significantly smaller among males.

c) A weight-loss was observed for *Y* = 8 subjects of *n* = 10, giving $\hat{p} = 0.80$. From the large-sample expression in EX 55 a) we get $0.80 \pm 1.96\sqrt{0.80 \cdot 0.20/10}$, or (-0.05, 1.05), which is unreasonable.

In order to use the conservative expression in (23) we need the percentiles $F_{.975}(6,20) = 3.13$ and $F_{.975}(18,4)$. The latter is hard to obtain from tables in textbooks, but one solution may be to use linear interpolation between $F_{.975}(20,4) = 8.56$ and $F_{.975}(15,4) = 8.66$, yielding $F_{.975}(18,4) \approx 8.60$. The latter is close to the true value 8.59 obtained by using the function *finv* in SAS (Cf. EX 56).

The 95% CI is $\left(\frac{8}{8 + (10-8+1) \cdot 3.13}, \frac{(8+1) \cdot 8.59}{10-8+(8+1) \cdot 8.59}\right) = (0.107, 0.976)$

d) From Ch. 5.4.1 $n = \hat{p}(1-\hat{p})(C/B)^2$, where *C* = 1.645 since we want a CI of 90%. Thus, $n = 0.8 \cdot 0.2(1.645/0.025)^2 = 693$.

The reason for choosing a 90% CI in this case is that we need to have *Bound of Error* small and this in turn is due to the fact that $\hat{p}$ is quite close to 1.

The balance between the choice of confidence level and *Bound of Error* can sometimes be a delicate problem.

EX 64 Introduce the notations μ_F and μ_C for the population means in the two groups and σ_F^2 and σ_C^2 for the population variances. We first make a 95% CI for the ratio σ_C^2 / σ_F^2.

In accordance with EX 53 a) we get

$\frac{S_C^2 / S_F^2}{c_2} < \frac{\sigma_C^2}{\sigma_F^2} < \frac{S_C^2 / S_F^2}{c_1}$, where $P(c_1 < F(n_C - 1, n_F - 1) < c_2) = 0.95$.

$S_C^2 / S_F^2 = 0.6362$ and $\begin{cases} P(F(29,21) > c_2) = 0.025 \Rightarrow c_2 = 2.317 \\ P(F(29,21) > c_1) = 0.975 \Rightarrow c_1 = 0.455 \end{cases}$ gives the CI (0.27, 1.40).

Since this interval well covers 1we can assume that the two population variances are equal.
The BLUE of the common variance is

$$\hat{\sigma}^2 = \frac{(n_C - 1)S_C^2 + (n_F - 1)S_F^2}{n_C - 1 + n_F - 1} = \frac{29 \cdot 0.2305 + 21 \cdot 0.3623}{50} = 0.2859$$

From EX 53 b) the 95% CI for $\mu_C - \mu_F$ is $\overline{Y}_C - \overline{Y}_F \pm C\sqrt{\hat{\sigma}^2(1/n_C + 1/n_F)}$, where C is determined from

$P(-C < T(50) < C) = 0.95 \Rightarrow P(T(50) > C) = 0.025 \Rightarrow C = 2.009$. This gives the CI (0.24, 0.85).

Since the CI is far above 0 the conclusion is that the mean AOD in the C group is significantly larger than in the FAS group.

EX 65 $Y \sim Exponential$ has $E(Y) = 1/\lambda$ and $V(Y) = 1/\lambda^2 \Rightarrow E(\overline{Y}) = 1/\lambda$ and $V(\overline{Y}) = \frac{1/\lambda^2}{n}$.

According to the CLT $\frac{\overline{Y} - 1/\lambda}{\sqrt{\frac{1/\lambda^2}{n}}} = \sqrt{n}\left(\lambda\overline{Y} - 1\right) \xrightarrow{D} Z \sim N(0,1)$.

From this we get $0.95 = P\left(-1.96 < \sqrt{n}(\lambda\overline{Y} - 1) < 1.96\right)$ and centering λ finally gives the CI $\left(\frac{1}{\overline{Y}}(1 - 1.96/\sqrt{n}), \frac{1}{\overline{Y}}(1 + 1.96/\sqrt{n})\right)$.

In order to calculate the expected length of the CI we need to know that $\sum_{i=1}^{n} Y_i \sim Gamma(\lambda, n)$ so (Cf. Ch. 2.2.2 (2)) $E\left(1/\sum_{i=1}^{n} Y_i\right) = \frac{1}{\lambda^{-1}}\frac{\Gamma(n-1)}{\Gamma(n)} = \lambda\frac{\Gamma(n-1)}{(n-1)\Gamma(n-1)} = \frac{\lambda}{(n-1)}$. The expected length of the CI above is thus $\frac{\lambda}{(n-1)}\frac{n \cdot 2 \cdot 1.96}{\sqrt{n}} = [n = 50] = 0.566\lambda$. This is to be compared with the CI in EX 61 a) when *n* = 50, $\frac{\lambda}{(n-1)}\frac{(129.56 - 74.22)}{2} = 0.565\lambda$.

EX 66 From EX 46 $\hat{p}_{ML} = \sum Y_i / \sum n_i$ has $E(\hat{p}_{ML}) = p$ and $V(\hat{p}_{ML}) = p(1-p)/\sum n_i$. Here $\sum Y_i$ has a Binomial distribution, so according to the CLT $\frac{\hat{p}_{ML} - p}{\sqrt{\hat{p}_{ML}(1-\hat{p}_{ML})/\sum n_i}} \xrightarrow{D} Z \sim N(0,1)$. (Cf. EX 23 c).) This gives the 95% CI limits $\hat{p}_{ML} \pm 1.96\sqrt{\hat{p}_{ML}(1-\hat{p}_{ML})/\sum n_i}$.

From the table we get $\hat{p}_{ML} = 30/100 = 0.30$, so the CI is (0.21, 0.39).

EX 67 $\hat{\lambda} = \frac{103 + 112 + 91 + 117}{4} = 105.76$. (This is the OLS-, Moment-, BLUE- and ML estimate.)

$E(\hat{\lambda}) = \lambda$ and $V(\hat{\lambda}) = \lambda/n$, with $\hat{V}(\hat{\lambda}) = \hat{\lambda}/n$.

Now, following the same lines as in EX 23 c) it can be shown that

$$\frac{(\hat{\lambda} - \lambda)}{\sqrt{\hat{\lambda}/n}} = \frac{\frac{(\hat{\lambda} - \lambda)}{\sqrt{\lambda/n}} \xrightarrow{D} Z \sim N(0,1)}{\frac{\sqrt{\hat{\lambda}/n}}{\sqrt{\lambda/n}} \xrightarrow{P} 1} = Z \sim N(0,1) \Rightarrow \hat{\lambda} \pm 1.96\sqrt{\hat{\lambda}/n}$$

gives a 95% CI for λ in large samples. The CI is in this case (96, 116).

Ch. 6

EX 109

a) A 2 x 2 table may look as follows:

		Shape		
		Round	Wrinkly	Total
Color	Yellow	315	101	416
	Green	108	32	140
	Total	423	133	556

b) We obtain the following table:

Characteristic	Obs. Freq.	Exp. Freq.	Deviation	Cell Chi.square
Yellow, Round	315	312.75	2.25	0.016
Green, Round	108	104.25	3.75	0.135
Yellow, Wrinkly	102	104.25	-3.25	0.101
Green, Wrinkly	32	34.75	-2.75	0.218
Total	556	556.00	0	0.470

$\text{p-value} = P\left(\chi^2(1) > 0.47\right) = 0.49$. No reason to reject Mendel's theory. (It's interesting that the famous statistician R. A. Fisher concluded that Mendel's results were far too perfect, indicating that adjustments had been made to the data to make observations fit the hypothesis.)

EX 110

a) In the Poisson distribution $E(Y) = \lambda = V(Y)$. The corresponding sample estimates are $\bar{y} = 54.6$ and $s^2 = 251.8$. The variance is more than four times larger than the mean, which seems strange.

b) $H_0 : Y \sim Poisson(\lambda)$

The test statistic is $X^2 = \sum \frac{(Y_i - \bar{Y})^2}{\bar{Y}} = \frac{(81-54.6)^2}{54.6} + \ldots + \frac{(31-54.6)^2}{54.6} = 41.5$

p-value $= P\left(\left(\chi^2(10-1) > 41.5\right) < 0.005 \Rightarrow \text{Reject } H_0\right.$.

EX 111 Let *Y* = 'Number on sick leave per day'. $H_0 : Y \sim Poisson(\lambda)$

The pf is $P(Y=y)=\frac{\lambda^y}{y!}e^{-\lambda}$. $\hat{\lambda}=\bar{y}=\frac{12\cdot 0+10\cdot 1+\ldots+0\cdot 5}{12+10+\ldots+0}=\frac{30}{30}=1.0$

Expected frequencies:

$$30\hat{P}(Y=0)=30\frac{(1.0)^0}{0!}e^{-1.0}=11.04,\quad 30\hat{P}(Y=1)=30\frac{(1.0)^1}{1!}e^{-1.0}=11.04,$$

$$30\hat{P}(Y=2)=30\frac{(1.0)^2}{2!}e^{-1.0}=5.52,\quad 30\hat{P}(Y\geq 3)=30\{1-\hat{P}(Y\leq 2)\}=30-11.04-11.04-5.52=2.40$$

The latter is computed to avoid small expected frequencies (cf. comment to EX 71).

$$X^2=\frac{(12-11.04)^2}{11.04}+\frac{(10-11.04)^2}{11.04}+\frac{(6-5.52)^2}{5.52}+\frac{(2-2.40)^2}{2.40}=0.29$$

p-value $=P(\chi^2(4-1-1)>0.29)>>0.10$. There is no reason to reject the Poisson model.

EX 112 $H_0 : Y \sim N(\mu, \sigma^2)$

From the *n* = 48 observations we obtain the estimates $\hat{\mu} = 55.125$ and $\hat{\sigma} = 18.96$.

Approximate cell probabilities in the four cells are (*Z* denote a $N(0,1) - \text{variable}$):

$$\hat{P}(Y < 40) = \hat{P}\left(Z < \frac{40 - 55.125}{18.96}\right) = 0.2125$$

$$\hat{P}(40 < Y < 60) = \hat{P}\left(Z < \frac{60 - 55.125}{18.96}\right) - \hat{P}\left(Z < \frac{40 - 55.125}{18.96}\right) = 0.3889$$

$$\hat{P}(60 < Y < 80) = \hat{P}\left(Z < \frac{80 - 55.125}{18.96}\right) - \hat{P}\left(Z < \frac{60 - 55.125}{18.96}\right) = 0.3038$$

$$\hat{P}(Y > 80) = 1 - \hat{P}(Y < 80) = 1 - \hat{P}\left(Z < \frac{80 - 55.125}{18.96}\right) = 0.0948$$

Multiplying these probabilities by *n* = 48 gives the expected cell frequencies.

$$X^2 = \frac{(11 - 10.2)^2}{10.2} + \frac{(18 - 18.7)^2}{18.7} + \frac{(14 - 14.6)^2}{14.6} + \frac{(5 - 4.5)^2}{4.5} = 0.17$$

p-value= $= P(\chi^2(4 - 2 - 1) > 0.17) = 0.68$, No reason to reject H_0.

EX 113 $H_0 : p_1 = p_2 = p_3 = p_4 (= p)$

We get the following proportions of born boys

Season	Spring	Summer	Autumn	Winter	Over all ($\hat{p}$)
Proportion	0.51780	0.51707	0.51414	0.51475	0.516055

$$\sum_{i=1}^{4} n_i (\hat{p}_i - \hat{p})^2 = 0.054351 + 0.015844 + 0.052297 + 0.025490 = 0.147982$$

$$T = \frac{0.147982}{0.516055(1 - 0.516055)} = 0.59 \Rightarrow \text{p - value} = P(\chi^2(4 - 1) > 0.59) >> 0.10$$. There is no reason to reject H_0, even though the sample is very large.

EX 114 For Group1 and Group2 $X^2 = 0$, whereas for Group (1+2) $(1+2)\ X^2 = 5.01 \Rightarrow \text{p - value} < 0.05$. In the latter case the combination of data from tables with unequal proportions and marginal frequencies has created an impression of association which in fact does not exist.

This example illustrates that it is possible to 'create non-significance' by searching for sub-groups where no association is found. On the other hand, you may find significant associations in sub groups while no significant association is found in the total group.

The problem can be settled by a clear definition of the population to be studied.

EX 115

a) The proportion positive among husbands is roughly twice as large as among wives, 0.30 and 0.16, respectively. But is the difference significant?

Using the Chi-square principle in the form of McNemar's test gives $X^2 = \frac{(24-10)^2}{24+10} = 5.76 \Rightarrow$

p-value < 0.05 (p-value = 0.016). There is a significant difference.

To apply the LR test (cf. EX 78) we need the estimates $\hat{p} = \frac{24+10}{2 \cdot 100} = 0.17,\ p_{21} = 0.24,\ p_{12} = 0.10$

$-2\ln\Lambda = -2\{(24+10)\ln(0.17) - 24\ln(0.24) - 10\ln(0.10)\} = 5.9340 \Rightarrow$ p-value < 0.05
(p-value =0.015)

b) The ordinary Chi Square test of independency yields $X^2 = 0.48 \Rightarrow \text{p - value} = P(\chi^2(1) > 0.48) >> 0.10$. The opinions of husbands and wives are independent.

EX 116

H_0 : No association between Working capacity and Type of training.

The total Chi-square measure is $X^2 = 65.71 \Rightarrow \text{p - value} << 0.001$. There is thus a strong association between the two factors.

The next step is to search for the combination of factors that are 'most responsible' for the high Chi-square measure. This can be done by considering the table of deviations and cell Chi-square measures and then apply the Bonferroni-Holm principle described in Ch. 6.4.

Table showing Deviation / Cell Chi-square

	Low	Medium	High
Physical	-38.4 / 9.35	35.75 / 28.88	2.62 / 0.37
Activation	43.26 / 5.85	-39.91 / 17.71	-3.36 / 0.30
Education	-4.90 / 0.86	4.16 / 2.20	0.74 / 0.17

From this we get the table of ranked Cell Chi-square measures

i	1	2	3	4
Cell Chi-square	28.88	17.71	9.35	5.85
p-value	<0.001	<0.001	0.0022	0.015
$0.05/(3 \cdot 3 - i + 1)$	0.0056	0.0063	0.0071	0.0083

Here the first three p-values are smaller than the value in the bottom row. The corresponding Cell Chi-square measures are thus significant after *adjustment for multiple significance*. There is an over-representation for the combination 'Medium working capacity' x 'Physical training' and also an under-representation for the combinations 'Medium working capacity' x 'Activation' and 'Low working capacity' x 'Physical training'.

EX 117 The hypothesis to test is $H_0 : p = 0.4$.

An estimate of *p* is $\hat{p} = \frac{\sum n_i \hat{p}_i}{\sum n_i} = \frac{\sum y_i}{\sum n_i} = \frac{4 + \ldots + 6}{10 + \ldots + 12} = 0.30$ (cf. EX 46b)) and an estimate of the variance of the latter is $\hat{V}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{\sum n_i} = \frac{0.30(1-0.30)}{100} = 0.0021$.

As a test statistic we take $T = \frac{\hat{p} - 0.4}{\sqrt{\hat{V}(\hat{p})}} = \frac{0.30 - 0.40}{\sqrt{0.0021}} = -2.182$. From EX 23 c) it follows that we can use the normal distribution to compute the (two-sided) p-value $2 \cdot P(Z > 2.182) = 2 \cdot 0.015 = 0.03$. The conclusion is that University employees are one sick leave to a less extent than the rest of the population.

EX 118 Introduce the notations:

$X(2) =$ Number of accidents during 2 weeks before the renewal, with intensity λ_X.

$Y(3) =$ -"- 3 " after -"- λ_Y.

Then $\left(Y(3) \middle| X(2) + Y(3) = n = 15\right) \sim Binomial(n = 15, p = \frac{3\lambda_Y}{2\lambda_X + 3\lambda_Y})$.

We want to test $H_0 : \lambda_X = \lambda_Y \Leftrightarrow p = \frac{3}{5}$. The observed number of accidents is 5 and therefore a one-sided p-value is obtained by by $P(Y(3) \le 5) = \binom{15}{0}(3/5)^0(2/5)^{15} + \ldots + \binom{15}{5}(3/5)^5(2/5)^{10} = 0.0338$.

The computation of the last expression is simplified by using tables over the Binomial distribution, e.g. Table 1 p839 in Wackerly *et al* 2007.

For a two-tailed test we also have to compute the probability of extremely large values. Since the expected number of accidents is $n \cdot p = 15 \cdot 3/5 = 9$, we compute $P(Y(3) \ge 13) =$

$\binom{15}{13}(3/5)^{13}(2/5)^2 + \binom{15}{14}(3/5)^{14}(2/5)^1 + \binom{15}{15}(3/5)^{15}(2/5)^0 = 0.0271$.

The p-value for a two-tailed test is 0.0338+0.0271 > 0.05 and the conclusion is that the effect of renewed equipment is not significant.

EX 119 $H_0 : p_i = 1/4$. The sum of observations is 80, so $n/k = 80/4 = 20$. The test statistic is

$$X^2 = \frac{(19-20)^2}{20} + \frac{(16-20)^2}{20} + \frac{(20-20)^2}{20} + \frac{(25-20)^2}{20} = 2.10 \Rightarrow \text{p - value} = P(\chi^2(4-1) > 2.10) > 0.10$$

H_0 is not rejected, the observations may be generated by the same Poisson variable.

EX 120 The likelihood is $L = \lambda^n e^{-\lambda \sum y_i} \Rightarrow \hat{\lambda}_{ML} = \frac{n}{\sum y_i} = \frac{1}{\bar{y}}$ (cf. EX 41)

The LR is $\Lambda = \frac{L_0}{\hat{L}} = \frac{\lambda_0{}^n e^{-\lambda_0 \sum y_i}}{(1/\bar{y})^n e^{-n}} = (\lambda_0 \bar{y})^n e^{-\lambda_0 n\bar{y}+n} = (\lambda_0 \bar{y})^n e^{n(1-\lambda_0 \bar{y})}$. H_0 is rejected for small values of Λ. Which values of $\lambda_0 \bar{y}$ will make Λ small? To answer this consider the function $g(x) = x^n e^{n(1-x)} \Rightarrow$

$\ln(g(x)) = n\ln(x) + n(1-x)$ which has max/min for the same values as $g(x)$ (cf. Ch. 2.3.3).

$\frac{d\ln(g(x))}{dx} = \frac{n}{x} - n = 0 \Rightarrow x = 1, \frac{d^2 \ln(g(x))}{dx^2} = -\frac{n}{x^2} < 0 \Rightarrow$ Local max for $x = \lambda_0 \bar{y} = 1$. Thus H_0 rejected for extremely large or small values of $\lambda_0 \bar{y}$, or equivalently for large values of $\lambda_0 \sum y_i$, but how large or small?

To this end we use property (5) in Ch. 2.2.2.

$Y_i \sim Exponential(\lambda) \Rightarrow \sum_{i=1}^{n} Y_i \sim Gamma(n, \lambda) \Rightarrow 2\lambda_0 \sum_{i=1}^{n} Y_i \sim \chi^2(2n)$. By using the latter relation probabilities can easily be computed from the Chi-square distribution. The RR is thus of the form $2\lambda_0 \sum Y_i < C_1$ or $2\lambda_0 \sum Y_i > C_2$, where C_1 and C_2 are constants that are determined in the following way:

The test is two-sided at the 5% level and *n* = 10 which yields

$P(\chi^2(2 \cdot 10) < C_1) = 0.025 \Rightarrow C_1 = 9.5908$ and $P(\chi^2(2 \cdot 10) > C_2 = 0.025) \Rightarrow C_2 = 34.1696$.

The RR is $\sum Y_i < \frac{9.5908}{2\lambda_0}$ or $\sum Y_i > \frac{34.1696}{2\lambda_0}$.

Notice that the sample has not yet been collected and nor has the value of λ_0 been specified. Once this has been the case the test can be performed.

EX 121 The Likelihood without restrictions on the parameters is

$L=\prod_{i=1}^{n_X}\lambda_X e^{-\lambda_X x_i}\cdot\prod_{i=1}^{n_Y}\lambda_Y e^{-\lambda_Y y_i}=\lambda_X^{n_X}\lambda_Y^{n_Y}e^{-\lambda_X\sum x_i}e^{-\lambda_Y\sum y_i}$ and the Likelihood under H_0 is

$L_0=\lambda^{(n_X+n_Y)}e^{-\lambda(\sum x_i+\sum y_i)}$.

$\ln L=n_X\ln\lambda_X+n_Y\ln\lambda_Y-\lambda_X\sum x_i-\lambda_Y\sum y_i\Rightarrow\frac{d\ln L}{d\lambda_X}=\frac{n_X}{\lambda_X}-\sum x_i=0\Rightarrow\hat{\lambda}_X=\frac{n_X}{\sum x_i}$.

Similarly, $\hat{\lambda}_Y=\frac{n_Y}{\sum y_i}$.

$\ln L_0=(n_X+n_Y)\ln\lambda-\lambda\left(\sum x_i+\sum y_i\right)\Rightarrow\frac{d\ln L_0}{d\lambda}=\frac{(n_X+n_Y)}{\lambda}-\left(\sum x_i+\sum y_i\right)\Rightarrow\hat{\lambda}=\frac{n_X+n_Y}{\sum x_i+\sum y_i}$

$\Lambda=\frac{\hat{L}_0}{\hat{L}}=\frac{\hat{\lambda}^{(n_X+n_Y)}}{\hat{\lambda}_X{}^{n_X}\hat{\lambda}_Y{}^{n_Y}}\cdot\frac{e^{-(n_X+n_Y)}}{e^{-n_X}e^{-n_Y}}=\frac{\hat{\lambda}^{(n_X+n_Y)}}{\hat{\lambda}_X{}^{n_X}\hat{\lambda}_Y{}^{n_Y}}\Rightarrow$

$-2\ln\Lambda=-2\left\{(n_X+n_Y)\ln\hat{\lambda}-n_X\ln\hat{\lambda}_X-n_Y\ln\hat{\lambda}_Y\right\}$ which is distributed $\chi^2(2-1)$ under H_0.

The estimates are $\hat{\lambda}_X=\frac{40}{20}=2.00,\hat{\lambda}_Y=\frac{60}{40}=1.50,\hat{\lambda}=\frac{40+60}{20+40}=1.67\Rightarrow$

$-2\ln\Lambda=-2(-0.9712)=1.9425\Rightarrow\text{p - value}=P\left(\chi^2(1)>1.9425\right)=0.16>0.05.$

There is no reason to reject H_0

The link between the Poisson process and exponentially distributed intervals was stated in property (4)

Ch. 2.2.2. In EX 86 it was shown how to test the equality between two Poisson rates based on count data (frequency of occurrences). In the present example we have shown how to perform the same test based on interval data.

EX 122

a) The unrestricted Likelihood is $L=\prod_{i=1}^{n}(1-p)^{y_i-1}p=(1-p)^{\sum y_i-n}p^n \Rightarrow$

$$\ln L=\left(\sum y_i-n\right)\ln(1-p)+n\ln p \Rightarrow \frac{d\ln L}{dp}=\frac{\left(\sum y_i-n\right)}{(1-p)}(-1)+\frac{n}{p}=0\Rightarrow \hat{p}=\frac{n}{\sum y_i}=\frac{1}{\bar{y}}.$$

The Likelihood under H_0 is $L_0=(1/2)^{\sum y_i-n}(1/2)^n=(1/2)^{\sum y_i}$. (No parameters need to be estimated.)

$$\Lambda=\frac{L_0}{\hat{L}}=\frac{(1/2)^{\sum y_i}}{\left(1-\frac{1}{\bar{y}}\right)^{\sum y_i-n}\left(\frac{1}{\bar{y}}\right)^n}=\text{[After some simplification, but not needed.]}=\frac{(\bar{y}/2)^{\sum y_i}}{(\bar{y}-1)^{\sum y_i-n}}.$$

$-2\ln\Lambda=-2\left\{\sum y_i\cdot\ln(\bar{y}/2)-\left(\sum y_i-n\right)\ln(\bar{y}-1)\right\}$ which is distributed $\chi^2(1-0)$ under H_0.

b) $-2\ln\Lambda=-2\{80\ln(0.8)-(80-50)\ln(8/5-1)\}=5.0534$. p-value $=P\left(\chi^2(1)>5.0534\right)=0.02$ <0.05.

H_0 is rejected and since $\hat{p}=50/80=0.625>0.5$ we draw the conclusion that *p* is significantly larger than 1/2.

c) There are many situations where we can collect data on the variable Y_i = 'Number of trials until an (0, 1) -event occurs for the first time' in order to test the hypothesis that *p* =1/2. An example is a sequence of ups and downs on the stock market.

EX 123

a) From the data we get $n=16, \bar{y}_D=2.2375, s_D^2=7.3265$. Thus,
$T=\frac{2.2375-0}{\sqrt{7.3265/16}}=3.307\Rightarrow \text{p - value}=2\cdot P(T(16-1)>3.307)=2\cdot 0.0024=0.004$. The conclusion is that the weight loss program had a significant positive effect.

b) Define *T* = 'Number of positive signs'. Under $H_0: P(+)=P(-)=1/2$ the variable *T* is distributed $Binomial(n=16, p=1/2)$. In the data we observe $T=13$ and a one-sided p-value is

$$P(T\geq 13)=\sum_{y=13}^{16}\binom{16}{y}(1/2)^y(1/2)^{16-y}=(1/2)^{16}\left\{\binom{16}{13}+\binom{16}{14}+\binom{16}{15}+\binom{16}{16}\right\}=0.0106.$$

However, to compute a two-sided p-value we should also consider the possibility that the outcome may be in the 'opposite direction'. Since the expected value of *T* is 16/2 = 8, the 'opposite direction' consists of the outcomes $T\leq 3$.

$$P(T\leq 3)=\sum_{y=0}^{3}\binom{16}{y}(1/2)^y(1/2)^{16-y}=(1/2)^{16}\left\{\binom{16}{0}+\binom{16}{1}+\binom{16}{2}+\binom{16}{3}\right\}=0.0106.$$ (This result is to be expected since the Binomial distribution is symmetric for *p* = 1/2.)

The two-sided p-value is thus 0.02, which is much larger than 0.004 in a), but still less than 0.05.

c) $T\sim Binomial(n=16, p=1/2)\Rightarrow P(T\geq 13)=1-P(T<13)\approx 1-P\left(Z<\frac{13-16/2-1/2}{\sqrt{16/4}}\right)=$

EX 123 (Continued)

$= 1 - P(Z < 2.25) = P(Z > 2.25) = 0.0122$.

Similarly $P(T \le 3) \approx P\left(Z < \dfrac{3 - 16/2 + 0.5}{\sqrt{16/4}}\right) = P(Z < -2.25) = 0.0122$ because of symmetry.

The two-sided p-value is $2 \cdot 0.0122 = 0.02$.

EX 124

a) Introduce the notations p_S and p_M for the proportion of products that are classified 'Bad' by State authorities and Municipal authorities, respectively. The observed difference is $\hat{p}_S - \hat{p}_M =$

$\frac{(20+20)}{100} - \frac{(20+10)}{100} = 0.10$, but is the difference significant?

$H_0 : p_S = p_M$

McNemar's test (Cf. EX 71) yields: $X^2 = \frac{(20-10)^2}{20+10} = 3.33 \Rightarrow \text{p - value} = P\left(\chi^2(1) > 3.33\right) = 0.068.$

We can't reject the null hypothesis at the 5% level.

b) H_0 : Independency between the two types of classifications. (It's close to an insult to set up this hypothesis.)

The ordinary Chi-square test of independency yields:

$X^2 = \frac{(20-12)^2}{12} + \frac{(10-18)^2}{18} + \frac{(20-28)^2}{28} + \frac{(50-42)^2}{42} = 12.7 \Rightarrow \text{p - value} = P\left(\chi^2(1) > 12.7\right) << 0.005$

There is strong reason to reject H_0.

EX 125 The pf. is $p(y) = \frac{\lambda^y}{y!} e^{-\lambda}$. $\hat{\lambda} = \frac{0 \cdot 103 + 1 \cdot 143 + \ldots + 6 \cdot 2}{103 + 143 + \ldots + 2} = 1.3225.$ (This is simply $\hat{\lambda} = \bar{y}$.)

Now, $\hat{p}(0) = e^{-\hat{\lambda}} = 0.2667$, $\hat{p}(1) = \hat{\lambda} \cdot e^{-\hat{\lambda}} = 0.3527$, and so on. In this way we get the following table, where *Y* = Observed frequency, $n\hat{p}(y)$ = Expected frequency and *n* = 400.

y	0	1	2	3	4	5	6	Total
$\hat{p}(y)$	.2667	.3527	.2332	.1028	.0337	.0090	.0019	1
$n\hat{p}(y)$	106.7	141.1	93.3	41.1	13.5	3.6	0.7	400
Y	103	143	98	42	8	4	2	400
$\frac{\left(Y - n\hat{p}(y)\right)^2}{n\hat{p}(y)}$	0.127	0.026	0.237	0.020	2.241	0.044	2.414	5.109

p-value $= P\left(\chi^2(7-1-1) > 5.109\right) = 0.40$, so there is no reason to reject the Poisson distribution. The degrees of freedom in the Chi-square distribution is due to the fact that there are 7 cells and 1 parameter has been estimated.

EX 126 A 2 x 3 frequency table is

	Morning	Day	Night	Total
Defective	12	10	23	45
Not defective	188	190	177	555
Total	200	200	200	600

The table with *Deviation / Cell Chi-Square* (Cf. Ch. 6.2.1) is

	Morning	Day	Night
Defective	-3 / 0.6	-5 / 1.6667	8 / 4.2667
Not defective	3 /0.0486	5 /0.1351	-8 / 0.3459

The total sum of *Cell Chi-Square* is
$X^2 = 0.6 + \ldots + 0.3459 = 7.0631 \Rightarrow \text{p - value} = P\left(\chi^2((3-1)(2-1)) > 7.0631\right) = 0.0293$. We reject the hypothesis of no association.

In the last table there is a large excess of observations (+8) in the cell Defective x Night. The p-value for this *Cell Chi-Square* is $P\left(\chi^2(1) > 4.2667\right) = 0.0389 < 0.05$. However, there are deviations in 6 cells to take account of. Since 0.0389 > 0.05/6 (Cf. Ch, 6.4.) we can't claim that there is a significant over-representation of observations in the cell Defective x Night after having adjusted p-values for multiple comparisons.

EX 127 There are three differences between proportions to consider. The largest difference is for the proportion defective in Night and Day, $23/200 - 10/200 = 0.065$. The corresponding test statistic for testing that the true difference is zero is (Cf. EX 83 a).)

$$T = \frac{0.065}{\sqrt{\hat{p}(1-\hat{p})(1/200 + 1/200)}} = \left[\hat{p} = \frac{23+10}{200+200}\right] = 2.36 \Rightarrow \text{p - value} = 2P(Z > 2.36) = 0.018 < 0.05.$$

However, since there are three comparisons to make we should require that 0.018 is less than $0.05/3 \approx 0.017$ (Cf. Ch. 6.4.). The difference between defectives in Night and Day is approximately significant after adjustment for multiple comparisons. The other two differences are not.

EX 128 Let p_I and p_F be the proportion left-handed among identical- and fraternal twins, respectively. We want to test $H_0 : p_I = p_F (= p)$. The test statistic for this is

$T = \dfrac{\hat{p}_I - \hat{p}_F - 0}{\sqrt{\hat{p}(1-\hat{p})(1/n_I + 1/n_F)}}$ which is distributed $N(0,1)$ in large samples. $\hat{p}_I = \dfrac{41}{248}$, $\hat{p}_F = \dfrac{18}{246}$,

$\hat{p} = \dfrac{41+18}{248+246} \Rightarrow T = 3.153 \Rightarrow \text{p - value} = 2P(Z > 3.153) \approx 0.002$. There is thus a strongly significant difference between the proportions.

EX 129 We give two types of solutions, one 'Straight' as in EX 85 a) and one in a 'Bio-statistical style'.

Straight solution: The two estimated proportions to be compared are $\hat{p}_1 = \frac{1}{20000}$ and $\hat{p}_2 = \frac{114}{473000}$.

$\hat{p} = \frac{1+114}{20000+473000}$. $T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{20000}+\frac{1}{473000}\right)}}$ takes the value -1.733

$\Rightarrow \text{p - value} = P(Z > 1.733) = 0.042 < 0.05$. (In this case it seems reasonable to use a one-sided test.)

Hint: If you have problems to perform calculations with very small numbers in both numerator and denominator in *T*, just multiply *T* by e.g. $\frac{1000}{1000}$.

Solution in a Bio-statistical style: Let *Y* ='Number of polio cases among 20000 vaccinated children'.

Under the hypothesis that the SALK vaccine has no effect $Y \sim Binomial(n = 20000, p = \frac{114}{473000})$

and approximately $Y \sim Poisson(\lambda = np = 4.8)$. The latter variable has expected value 4.8 and therefore the hypothesis of no effect is rejected for small values of *Y*.

p-value = $P(Y \le 1) = \sum_{y=0}^{1} \frac{4.8^y}{y!} e^{-4.8} = (1+4.8)e^{-4.8} = 0.046 < 0.05$.

In this case the group under study (vaccinated children) has been exposed to a standard value of a parameter (*p* = 114/473000) and the outcome of this is evaluated. Such an approach is very common in bio-statistical studies, especially in epidemiology.

EX 130 Table of *Deviation / Cell Chi-Square*:

	Low	Middle	High
Cheap	-14.3/ 3.93	20.0/ 5.86	-5.6/ 0.9
Expensive	14.3/ 4.31	-20/ 6.44	5.6/ 1.0

The total Chi-square measure is $X^2 = 3.93 + \ldots + 1.0 = 22.4 \Rightarrow \text{p - value} = 0.00001 << 0.05$. There is thus strong evidence against independency.

The largest *Cell Chi-Square* is 6.44 and p-value $= P(\chi^2(1) > 6.44) = 0.011 < 0.05$. For a multiple comparison in 6 cells it is required that the latter is less than 0.05/6 = 0.008, which is nearly true. No further significant patterns can be seen.

The conclusion is that there is a significant under-representation of Middle- class families with Expensive electronic equipment.

EX 131 Let Y = 'Yearly number of cases in Malmö' $\sim Binomial(n = 110000, p)$. We want to test $H_0 : p = 35/100000$. Since p is small, $Y \sim Poisson(\lambda = 110000 \cdot 35/100000 = 38.5)$ under H_0.

The expected value of Y is 38.5 and the observed value is 60. It is therefore natural to compute the p-value as $P(Y \geq 60) = \sum_{y=60}^{\infty} \frac{\lambda^y}{y!} e^{-\lambda}$, but this is a heavy task. Instead we use the fact that for large λ, a Poisson variable can be approximated by a $N(\lambda, \lambda)$ -variable (Cf. Ch. 2.2.2.).

$$\text{p-value} = P(Y \geq 60) \approx P\left(Z > \frac{60 - 38.5}{\sqrt{38.5}}\right) = P(Z > 3.47) = 0.0002$$

The conclusion is that inhabitants in Malmö have a significant higher risk for malignant melanoma than the rest of the Swedish population.

EX 132

a) As the question is formulated, a one-sided test seems to be appropriate. On the other hand, if the problem was to find out which of the two importers is best, a two-sided test is preferably.

b) The two estimated proportions are $\hat{p}_1 = \frac{19}{200} = 0.095$ and $\hat{p}_2 = \frac{10}{200} = 0.050$. To test $H_0 : p_1 = p_2 (= p)$ we use the test statistic $T = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$ where $\hat{p} = \frac{19+10}{200+200}$.

This gives $T = 1.735 \Rightarrow \text{p - value} = P(Z > 1.735) = 0.0413 < 0.05$. The new importer is better!

c) **c)** We construct the following 2 x 2 table:

		Quality		
		Bad	Acceptable	Total
Importer	Former	19	181	200
	New	10	190	200
	Total	29	371	400

$X^2 = 1.3966 + 0.1092 + 1.3966 + 0.1092 = 3.0114 \Rightarrow \text{p - value} = P(\chi^2(1) > 3.0114) = 0.0826 \cdot$

Notice that the p-value in c) is twice that of b). *The Chi-square test is by its nature a two-sided test.*

It has been demonstrated that the test of equality between Binomial proportions in EX 83 a) is equivalent to the Chi-square test of independence if the test is two-sided. The Chi-square test can also be used as a one-sided test if the p-value is divided by 2. However, both tests are approximate and require that sample sizes are large.

EX 133 We first consider one-sided vs. two-sided tests.

Most people would probably agree that vaccination could not have a negative effect on the state of health. The test should thus be one-sided.

(However, there may be other opinions on this issue. Some might e.g. argue that vaccine can be contaminated.)

The test for equality of two Binomial proportions yields

$$T = \frac{4/90 - 18/66}{\sqrt{\frac{22}{156}\left(1-\frac{22}{156}\right)\left(\frac{1}{90}+\frac{1}{66}\right)}} = -4.0476 \Rightarrow \text{p - value} = P(Z > 4.0476) \approx 0.000025 .$$

From the 2 x 2 table the following table over *Deviation / Cell Chi-Square* is constructed:

	Diseased	Not diseased
Vaccinated	-8.7 / 5.9529	8.7 /0.98
Not vaccinated	8.7 / 8.1176	-8.7 / 1.33

From this, $X^2 = 16.38 \Rightarrow \text{p - value} = P(\chi^2(1) > 16.38) = 0.000050$. The latter is the result of a two-sided test. To get a one-sided p-value we simply divide it by 2.

In the last table there are two significant *Cell Chi-Square*. $P\left(\chi^2(1) > 8.1176\right) = 0.0044 < \frac{0.05}{4} = .00125$ and

$P\left(\chi^2(1) > 5.9529\right) = 0.0147 < \frac{0.05}{3} = 0.0166$ (Cf. Ch. 6.4.).

EX 134

a) $X^2 = 6.6197 \Rightarrow \text{p - value} = P\left(\chi^2(1) > 6.6197\right) = 0.0178$ (two-sided test). For a one-sided test the p-value is halved, 0.0139.

b) p-value $= \frac{98!\cdot 47!\cdot 129!\cdot 16!}{145!}\left(\frac{1}{15!\cdot 1!\cdot 83!\cdot 46!} + \frac{1}{16!\cdot 0!\cdot 82!\cdot 47!}\right) = 0.0123$ (one-sided test)

EX 135

a) We first test whether the population variances of the two bulbs are equal (Cf. EX 88.).

$T = \frac{(83)^2}{(62)^2} = 1.79 \Rightarrow \text{p - value} = 2 \cdot P\left(F(19,19) > 1.79\right) = 0.22$. The variances can be considered equal.

The pooled sample variance is $S^2 = \frac{(20-1)(62)^2 + (20-1)(83)^2}{(20-1)+(20-1)} = 5367$.

The test statistic for testing that the two population means are equal is

$T = \frac{1128 - 1236}{\sqrt{5367(1/20 + 1/20)}} = -4.66 \Rightarrow \text{p - value} = 2 \cdot P\left(T(20-1+20-1) > 4.66\right) = 0.00004$. The difference is clearly significant.

b) Referring to EX 88, we get $T = \frac{1128 - 1236}{\sqrt{(62)^2/20 + (83)^2/20}} = -4.66 \Rightarrow \text{p - value} = 2 \cdot P\left(Z > 4.66\right) < 0.00001$.

EX 136 Let Y_A and Y_B be the yield from variety A and B, respectively. Form the difference $D = Y_A - Y_B$ and compute the estimates $\overline{d} = 1.33$ and $s_d^2 = 6.7882$. The hypothesis that the difference between the means is zero is the same as that $E(D) = 0$. This is tested by $T = \dfrac{1.33 - 0}{\sqrt{6.7882/12}} = 1.77$

$\Rightarrow$ p - value $= 2 \cdot P\big(T(12-1) > 1.77\big) =$ 0.10. We can't claim that the difference is significant.

EX 137 The ranked series are:

A	22	28	38	40	40	41	44	46	48	51	58	60
B	12	15	18	21	23	24	29	35	36	43	54	80

The combined ranked series is easily shown to have the sample median $m = \dfrac{38+40}{2} = 39$. This leads to the 2 x 2 frequency table

	A	B	Total
>39	9	3	12
<39	3	9	12
Total	12	12	24

The hypothesis of interest is H_0 :No association between Type of series and Distribution around sample median.

$X^2 = 6.0 \Rightarrow$ p - value $= P\big(\chi^2(1) > 6.0\big) = 0.0143$. The two series have significantly different medians.

EX 138

a) Remember the conditions for the Binomial distribution in Ch. 2.2.1,' *independent* repetitions of the *same* experiment'. The taster has thus to spit out the beer after each tasting. He also has to reset his memory, so that he can't compare the taste of a new glass with the taste of the preceding one. Therefore, a so called wash-out period is needed between tastings. The arrangement of the glasses is easily done by tossing a coin.
The hypothesis $H_0 : p = 1/2$ means that we assume that the taster is merely guessing.

b) From the Binomial distribution we get $P(Y=8) = \binom{8}{8}\frac{1}{2^8}\frac{1}{2^0} = \frac{1}{2^8} = 0.004 < 0.05$,
$P(Y \geq 7) = P(Y=7) + P(Y=8) = \binom{8}{7}\frac{1}{2^7}\frac{1}{2} + \frac{1}{2^8} = \frac{1}{2^8}(8+1) = 0.035 < 0.05$. For *y* smaller than 7 we get probabilities that are larger than 0.05. The smallest acceptable value of *y* is thus 7.

c) We have to solve *x* in the relation $P(Y > x) \approx P\left(Z > \frac{x - 100 \cdot (1/2)}{\sqrt{100(1/2)(1/2)}}\right) = 0.05 \Rightarrow \frac{x-50}{5} = 1.645 \Rightarrow$ $x = 59$ will be enough.

EX 139

a) From EX 4 we get

y	0	1	2	3	4
Sample cdf	6/160	44/160	102/160	149/160	160/160
$p(y)$	0.056	0.235	0.374	0.265	0.070
$F_0(y)$	0.056	0.291	0.665	0.930	1.000

Here, $F_0(y) = p(0) + \ldots + p(y)$.

The largest absolute difference between the sample cdf and $F_0(y)$ is $D_{160} = |102/160 - 0.665| = 0.028$.

The critical value for a two-sided test at the 5 % level is $1.36/\sqrt{160} = 0.11$ (Cf. Ch. 6.2.4) > 0.028 so there is no reason to reject the Binomial distribution.

b)

y	0	1	2	3	4
Expected frequency	9.0	37.6	59.8	42.4	11.2
Observed frequency	6	38	58	47	11

Here, Expected frequency is $160 \cdot p(y)$ $y = 0,1,2,3,4$.

$$X^2 = \frac{(6-9.0)^2}{9.0} + \frac{(38-37.6)^2}{37.6} + \frac{(58-59.8)^2}{59.8} + \frac{(47-42.4)^2}{42.4} + \frac{(11-11.2)^2}{11.2} = 1.56 \Rightarrow$$

p-value $= P(\chi^2(5-1-1) > 1.56) = 0.67$ and there is again no reason to reject the Binomial distribution.

EX 140

a) The ranked series and sample cdf's are:

Day1 (15 observations):

y	30.8	31.0	31.4	32.9	33.3	33.4	33.5
$S_{15}(y)$	0.067	0.133	0.200	0.267	0.400	0.467	0.533

y	33.7	34.3	34.4	34.8	34.9	36.2	37.0
$S_{15}(y)$	0.600	0.667	0.733	0.800	0.867	0.933	1.000

Here, 0.067 = 1/15, 0.133 = 2/15, and so on. Notice that there are two observations at $y = 33.3$.

Day2 (10 observations):

y	28.4	28.7	30.2	30.7	31.3	31.9	32.1	32.4	32.8
$S_{10}(y)$	0.100	0.200	0.400	0.500	0.600	0.700	0.800	0.900	1.000

For $31.4 \le y < 32.9, S_{15}(y) = 0.20$ and for $y \ge 32.8, S_{10}(y) = 1.000$. The largest absolute difference between the two cdf's is $D_{15,10} = |0.20 - 1.00| = 0.80$.

From tables over the Smirnov two-sample distribution it is seen that the critical values for rejecting the hypothesis of equal cdf's are 15/30 (5% level) and 19/30 (1% level). The observed absolute difference of 0.80 is larger than both of these, so the hypothesis of equal population distributions can be rejected at least at the 1% level.

b) From the data we get Day 1: Mean = 33.66, Variance = 3.05
Day2: Mean =30.87, Variance = 2.28

As in EX 86 we first test whether the population variances of the two series are equal.

$F = \dfrac{3.05}{2.28} = 1.34 \Rightarrow \text{p - value} = P\big(F(15-1,10-1) > 1.34\big) = 0.34 \Rightarrow$ Population variances can be assumed to be equal. The pooled estimate of the common variance is $s^2 = \dfrac{(15-1)\cdot 3.05 + (10-1)\cdot 2.28}{(15-1)+(10-1)} = 2.75$.

To test whether the population means are equal,

$T = \dfrac{33.66 - 30.87}{\sqrt{2.75(1/15 + 1/10)}} = 4.18 \Rightarrow \text{p - value} = 2 \cdot P\big(T(15-1+10-1) > 4.18\big) = 0.0004 \Rightarrow$ Population means differ significantly.

According to both tests in a) and b) the population distribution of viscosity has changed from one day to another. The test in b) gave a somewhat stronger rejection of equal means. On the other hand, the test in a) is free from assumptions about the population distribution. The latter is to prefer in the absence of evidence for that viscosity is normally distributed.

EX 141 Introduce the notations $\begin{cases} +\text{ if yield for A is higher than B} \\ -\text{ if yield for B is higher than A} \\ 0\text{ if yield for A equals yield for B} \end{cases}$. We now get the following pattern:

Area	1	2	3	4	5	6	7	8	9	10	11	12
Sign	+	-	+	-	+	-	+	+	+	-	+	0

Let Y = 'Number of minus signs' of $n = 11$'. (The observation with a 0 is deleted.) Under the hypothesis of no difference in yield, $Y \sim Binomial(n = 11, p = 1/2)$.

p-value = $P(Y \le 4) = \sum_{y=0}^{4} \binom{11}{y} (1/2)^y (1/2)^{11-y} = (1/2)^{11} \left[\binom{11}{0} + \binom{11}{1} + \binom{11}{2} + \binom{11}{3} + \binom{11}{4} \right] =$

$\frac{1}{2048}(1 + 11 + 55 + 165 + 330) = 0.27$. A two-sided p-value is obtained by doubling the latter value. In neither case the difference is not significant.

EX 142 We want to test the hypothesis of no association between the two series. Let d_i be the difference between the ranks for the *i*:th student.

$$\sum d_i^2 = 1+16+4+9+2+9+2+25+16+14+1+9+16+16+9 = 146$$

$$r_S = 1 - \frac{6}{15\cdot(15^2-1)}\cdot 146 = 0.7393 .$$

From Table 11, Appendix 3 in Wackerly *et al* one gets the critical value 0.525
(two-sided test, $\alpha = 0.05$). Since this is smaller than 0.7393 we reject the hypothesis of no association.

EX 143 Let $X_i = \text{Value before}, Y_i = \text{Value after}$ for the *i*:th sample unit and put $D_i = X_i - Y_i$.

From the data we get $n = 10, \overline{D} = 0.03, S_D^2 = 0.001533\ (S_D = 0.0392)$.

a) $Z = \dfrac{0.03-0}{0.0392/\sqrt{10}} = 2.42 \Rightarrow \text{p - value} = 2\cdot P(Z > 2.42) = 0.0156$. There has been a significant decrease of the concentration due to the effect of the drug.
There are good reasons for assuming normality in this case. Notice that $\overline{D}$ involves a sum of 20 variables.

b) RR is $\left|\overline{D} - 0\right| > 1.96\cdot 0.0392/\sqrt{n}$

c) $$Pow(\mu_D) = P\left(Z > 1.96\cdot 1 - \frac{(\mu_D - 0)\sqrt{n}}{0.0392}\right) + P\left(Z < -1.96\cdot 1 - \frac{(\mu_D - 0)\sqrt{n}}{0.0392}\right).$$

For $n = 100$ this will be larger than 0.90 if $\left|\mu_D\right| > 0.013$.

EX 144 From EX 97–98 we obtain

$$Pow(p) = P\left(Z > 1.96\cdot\frac{0.08(1-0.08)}{\sqrt{p(1-p)}} - \frac{(p-0.08)\sqrt{n}}{\sqrt{p(1-p)}}\right) + P\left(Z < -1.96\cdot\frac{0.08(1-0.08)}{\sqrt{p(1-p)}} - \frac{(p-0.08)\sqrt{n}}{\sqrt{p(1-p)}}\right)$$

By inserting *p* = 0.15 one gets the requirement $Pow(0.15) = 0.90$, an equation in *n* that has to be solved by 'trial and error'. It is found that for *n* = 200, $Pow(0.15) = 0.90031$.

RR is $\left|\hat{p} - 0.08\right| > 1.96\cdot\dfrac{\sqrt{0.08(1-0.08)}}{\sqrt{n}} = [n = 200] = 0.0376$.

Now the study can begin and the sample is collected. In practice it would be wise to include somewhat more than 200 persons in the sample due to non-responses or drop-outs.

EX 145 $T=(n-1)S^2 \sim \sigma^2\chi^2(n-1)$. The RR is $T<a$ or $T>b$ and $\alpha=0.05$.

a) $\frac{\alpha}{2}=P(T<a|H_0)=P(\sigma_0^2\chi^2(n-1)<a)=P\left(\chi^2(n-1)<\frac{a}{\sigma_0^2}\right)\Rightarrow \frac{a}{\sigma_0^2}=\chi^2_{\alpha/2},\ a=\sigma_0^2\chi^2_{\alpha/2}$

$\frac{\alpha}{2}=P(T>b|H_0)=P(\sigma_0^2\chi^2(n-1)>b)=P\left(\chi^2(n-1)>\frac{b}{\sigma_0^2}\right)=1-P\left(\chi^2(n-1)<\frac{b}{\sigma_0^2}\right)\Rightarrow$

$P\left(\chi^2(n-1)<\frac{b}{\sigma_0^2}\right)=1-\frac{\alpha}{2}\Rightarrow \frac{b}{\sigma_0^2}=\chi^2_{1-\alpha/2},\ b=\sigma_0^2\chi^2_{1-\alpha/2}$.

Here we have used the notation $P(\chi^2<\chi^2_p)=p$.

The RR is thus $T<\sigma_0^2\chi^2_{\alpha/2}$ or $T>\sigma_0^2\chi^2_{1-\alpha/2}$.

$Pow(\sigma^2)=P(T<\sigma_0^2\chi^2_{\alpha/2})+P(T>\sigma_0^2\chi^2_{1-\alpha/2})=P(\sigma^2\chi^2(n-1)<\sigma_0^2\chi^2_{\alpha/2})+$

$P(\sigma^2\chi^2(n-1)>\sigma_0^2\chi^2_{1-\alpha/2})=P\left(\chi^2(n-1)<\frac{\sigma_0^2}{\sigma^2}\chi^2_{\alpha/2}\right)+P\left(\chi^2(n-1)>\frac{\sigma_0^2}{\sigma^2}\chi^2_{1-\alpha/2}\right)$, which turns out to be a function of $R=\sigma_0^2/\sigma^2$.

b) With *n* = 10 and $\alpha/2=0.025$, $\chi^2_{0.025}=2.70039$ and $\chi^2_{0.975}=19.0229$. The RR is

$T<\sigma_0^2\cdot 2.70039$ or $T>\sigma_0^2\cdot 19.0229$.

$Pow(R)=P(\chi^2(9)<R\cdot 2.70039)+P(\chi^2(9)>R\cdot 19.0229)$. This is a non-symmetric function. Some values are:

R	0.2	0.5	1.0	1.5	2.0	5.0
$Pow(R)$	0.92	0.39	0.05	0.09	0.20	0.94

EX 146 With *n* = 10 the RR is $\sum Y_i<\frac{9.5908}{2\lambda_0}$ or $\sum Y_i>\frac{34.1696}{2\lambda_0}$.

$Pow(\lambda)=P\left(\sum Y_i<\frac{9.5908}{2\lambda_0}\right)+P\left(\sum Y_i>\frac{34.1696}{2\lambda_0}\right)=$ [Multiply each factor within braces with 2λ and use the result that $2\lambda\sum_{i=1}^{n}Y_i\sim\chi^2(2n)$] $=P\left(\chi^2(2n)<\frac{\lambda}{\lambda_0}\cdot 9.5908\right)+P\left(\chi^2(2n)>\frac{\lambda}{\lambda_0}\cdot 34.1696\right)$.

Put $R=\frac{\lambda}{\lambda_0}$ and plot the power as a function of *R*.It is seen that the power is larger than 0.90 for

$R<0.35$ and $R>3.0$.

EX 147 From the data we get $n=6, \sum x=21, \sum x^2=91, \sum y=21, \sum y^2=89.2, \sum xy=90$,

$S_{XX}=91-(21)^2/6=17.5, S_{YY}=89.2-(21)^2/6=15.7, S_{XY}=90-(21)(21)/6=16.5$.

a) $\hat{\beta}=\frac{16.5}{17.5}=0.9429,\ \hat{\alpha}=\frac{21}{6}-\hat{\beta}\frac{21}{6}=0.2,\ SSE=89.2-\hat{\beta}^2\cdot 91=8.3029,\ \hat{\sigma}^2=\frac{SSE}{n-2}=2.0757$

$H_0:\beta=1$

$T=\frac{0.9429-1}{\sqrt{2.0757/(6-2)17.5}}=-0.33,\ \text{p - value}=2\cdot P\left(T(6-2)>0.33\right)=0.66$. Don't reject H_0.

$H_0:\alpha=0$

$T=\frac{0.2-0}{\sqrt{2.0757\left(\frac{1}{6}+\frac{(21/6)^2}{(6-2)17.5}\right)}}=0.237,\ \text{p - value}=2\cdot P\left(T(6-2)>0.237\right)=0.82$.

Don't reject H_0.

b) Since $H_0:\alpha=0$ can't be rejected we apply the model $E(Y|x)=\beta\cdot x$.

From EX 54, $\hat{\beta}=\frac{90}{91}=0.9890,\ SSE=89.2-(90/91)^2\cdot 91=0.1890,\ \hat{\sigma}^2=\frac{SSE}{(n-1)}=0.0378$ (The *SSE* in the latter expression is different from the one in a).)

$H_0:\beta=1$

$T=\frac{0.9890-1}{\sqrt{0.0378/91}}=-0.54,\ \text{p - value}=2\cdot P\left(T(6-1)>0.54\right)=0.62$. Don't reject H_0.

The conclusion is that the model in b) is to prefer.

EX 148 First we construct the following table:

x	x^2	n	nx	nx^2	$\bar{y}$	$n\bar{y}$	$nx\bar{y}$
1	1	4	4	4	1.00	4.0	4.0
3	9	5	15	45	3.64	18.2	54.6
5	25	3	15	75	7.23	21.7	108.5
10	100	4	40	400	12.725	50.9	509.0
15	225	4	60	900	18.225	72.9	1093.5
Total		20	134	1424		167.7	1769.6

$$\hat{\beta}=\frac{1769.6-(167.7)(134)/20}{1424-(134)^2}=1.2277,\ \hat{\alpha}=\frac{167.7}{20}-\hat{\beta}\frac{134}{20}=0.1544$$

x	n	$\bar{y}$	$n(\bar{y}-\hat{\alpha}-\hat{\beta}\cdot x)^2$	$\sum y_{ij}^2-(\sum y_{ij})^2/n_i$
1	4	1.00	0.584	$5.1-(4.0)^2/4=1.1$
3	5	3.64	0.195	$74.58-(18.2)^2/5=8.332$
5	3	7.23	2.653	$158.97-(21.7)^2/3=2.0067$
10	4	12.725	0.345	$648.61-(50.9)^2/4=0.9075$
15	4	18.225	0.476	$1332.95-(72.9)^2/4=4.3475$
Total			4.253	15.6612

$$H_0: E\left(Y\middle|X=x\right)=\alpha+\beta\cdot x$$

$F=\frac{4.253/(5-2)}{15.6612/(20-5)}=1.358\Rightarrow \text{p - value}=P\left(F(3,15)>1.358\right)=0.29$. No reason to reject the hypothesis.

EX 149

x	y	n	nx	nx^2	$\bar{y}$	$n\bar{y}$	$nx\bar{y}$
0.5	1.3 1.5	2	1	0.5	1.4	2.8	1.4
1	1.9 2.1	2	2	2	2.0	4.0	4.0
4	3.9 4.1	2	8	32	4.0	8.0	32.0
16	7.8 8.2	2	32	512	8.0	16.0	256.0
Total	30.8	8	43	546.5		30.8	293.4

$$\hat{\beta}=\frac{293.4-(30.8)(43)/8}{546.5-(43)^2/8}=0.4054,\ \hat{\alpha}=\frac{30.8}{8}-\hat{\beta}\frac{43}{8}=1.67$$

EX 149 (Continued)

x	n	y	$\bar{y}$	$n(\bar{y}-\hat{\alpha}-\hat{\beta}\cdot x)^2$	$\sum(y_{ij}-\bar{y}_j)^2$
0.5	2	1.3 1.5	1.4	0.4469	0.01+0.01 = 0.02
1	2	1.9 2.1	2.0	0.0114	0.01+0.01 = 0.02
4	2	3.9 4.1	4.0	1.0037	0.01+0.01 = 0.02
16	2	7.8 8.2	8.0	0.0488	0.04+0.04 = 0.08
Total				1.5108	0.14

$F=\dfrac{1.5108/(4-2)}{0.14/(8-4)}=21.6\Rightarrow \text{p - value}=P(F(2,4)>21.6)=0.007$. The linear model has to be rejected.

EX 150 The model $y = a \cdot x^b$ may be a candidate. Here, $\ln(y) = \ln(a) + b\ln(x)$ or $y' = a' + b \cdot x'$.

We now test whether $E(Y'|x') = a' + b \cdot x'$ is a proper model.

x'	y'	n	nx'	$n(x')^2$	$\bar{y}''$	$nx'\bar{y}''$	$n\bar{y}'$
-.069315	0.2624 0.4046	2	-1.3863	0.9609	0.3339	-0.4629	0.6678
0	0.6419 0.7419	2	0	0	0.6919	0	1.3838
1.38629	1.3610 1.4110	2	2.7726	3.8436	1.3860	3.8428	2.7720
2.77259	2.0541 2.1041	2	5.5452	15.3745	2.0791	11.5291	4.1583
3.4657	8.9818	8	6.9315	20.1790	4.4909	14.9090	8.9818

$$\hat{b}' = \frac{14.9090 - (8.9818)(6.9315)/8}{20.1790 - (6.9315)^2/8} = 0.5028,\ \hat{a}' = \frac{8.9818}{8} - \hat{b}'\frac{6.9315}{8} = 0.687$$

From this we get a new table:

n	$n(\bar{y}' - \hat{a}' - \hat{b}' \cdot x')^2$	$\sum (y'_{ij} - \bar{y}'_i)^2$
2	.000044	0.0102
2	.000046	0.0050
2	.000007	0.0013
2	.000008	0.0013
Total	0.000105	0.0078

$F = \frac{0.000105/(4-2)}{0.0078/(8-4)} = 0.027 \Rightarrow \text{p - value} = P(F(2,4) > 0.027) = 0.97$. There is definitely no reason to reject the new model.

EX 151 The hypothesis to test is $H_0 : \beta_2 = 0, \beta_3 = 0$ against $H_a : \beta_2 \neq 0, \beta_3 \neq 0$, i.e. there is no effect of the sex on body weight beyond the initial weight.

$T = \frac{(18.7454 - 18.4381)/(3-1)}{18.4381/(64-3-1)} = 0.50 \Rightarrow \text{p - value} = P(F(2,60) > 0.50) = 0.61$. No reason to reject H_0

EX 152 The model is linearized by the transformation $\ln(Q) = \ln(\alpha) + \beta_1 \ln(I) + \beta_2 \ln(P) =$

$\alpha' + \beta_1 I' + \beta_2 P'$.

From the print out we obtain:

$R^2 = 0.990$ (a good agreement).

Parameter	Estimate	*T-value*	$P(T > \lvert T - value \rvert)$
α'	9.0795	1.43	0.1765
β_1	-0.9994	-1.92	0.0766
β_2	1.0779	16.23	< 0.0001

Here the *T-value* and the corresponding p-value are computed under the hypothesis that the parameter is zero. Since $\hat{\alpha} = e^{9.0795} = 8773$ the estimated model is $Q = 8773 \cdot I^{-1} P^{1.08}$. A 'significance-fundamentalist' (a person who argues that all non-significant parameters should be deleted in a model) would object against including *I* in the model.

Comment to the solution of EX 152. To solve non-linear problems by linearization and using least-squares techniques, as in the present example, is very frequent (perhaps too frequent) among econometricians. One should be aware of that if $\hat{\beta}$ is unbiased for β, it does not follow that $P^{\hat{\beta}}$ is unbiased for P^{β}. Let's look at this a bit closer.

Put $g(\hat{\beta}) = P^{\hat{\beta}}$, where $E(\hat{\beta}) = \beta$. According to (14) in Ch. 3.3.2, $E\left(g(\hat{\beta})\right) \approx g(\beta) + \frac{1}{2} g''(\beta) \cdot V(\hat{\beta})$.
Now we have to find $g''(\beta)$. For simplicity, put $g = P^y \Rightarrow \ln(g) = y \ln(P) \Rightarrow \frac{d \ln(g)}{dy} = \frac{g'}{g} = \ln(P) \Rightarrow$
$g' = \ln(P) \cdot g \Rightarrow g'' = \ln(P) \cdot g' = \left(\ln(P)\right)^2 \cdot g = \left(\ln(P)\right)^2 P^y$. From this we finally obtain
$E\left(P^{\hat{\beta}}\right) \approx P^{\beta} + \frac{1}{2}\left(\ln(P)\right)^2 P^{\beta} \cdot V(\hat{\beta}) = P^{\beta}\left(1 + \frac{\left(\ln(P)\right)^2 V(\hat{\beta})}{2}\right) > P^{\beta}$, so there will be a positive bias. However, since $V(\hat{\beta}) = \frac{const.}{n}$, the bias can be ignored in large samples.

EX 153 $S(y) = e^{-\lambda y^{\alpha}} \Rightarrow \ln\left(S(y)\right) = -\lambda y^{\alpha} \Rightarrow y' = \ln\left(-\ln\left(S(y)\right)\right) = \ln(\lambda) + \alpha \ln(y) = \lambda' + \alpha \cdot x$, say.

We thus run a regression of $\hat{y}' = \ln\left(-\ln\left(\hat{S}(y)\right)\right)$ on x = ln(y). This yields:

Parameter	Estimate	Standard Error of Estimate
λ'	-0.0914	0.0378
α	2.0470	0.0888

Here, Standard Error of Estimate $= \sqrt{\hat{V}(Estimator)}$

$H_0 : \alpha = 1$ against $H_a : \alpha \neq 1$ is tested by $T = \frac{2.0470 - 1}{0.0888} = 11.79 \Rightarrow$ p - value $= 2 \cdot P\left(T(5-2) > 11.79\right) =$ 0.001.
Reject H_0!

References

Casella, G. & Berger, R.L. 1990, *Statistical Inference*, Duxbury Press, Belmont, California.

Cochran, W.G. 1934,'The distribution of quadratic forms in a normal system, with applications to the analysis of covariance', *Proc. Camb. Phil. Soc*, 30, pp. 178–191.

Cochran, W.G. 1950, 'The Comparison of Percentages in Matched Samples', *Biometrika*, 37, pp. 256–266.

Cochran, W.G. 1954, 'Some methods for strengthening the common χ^2 test', *Biometrics*, 10, pp. 417–451.

Cox, D.R. & Smith, W.L. 1954, 'On the superposition of renewal processes', *Biometrika*, 41, pp. 91–99.

Cramer, H. 1957, *Mathematical Methods of Statistics*, 7th edn, Princeton University Press, Princeton.

Davis, C.E. 1976, 'The effect of regression to the mean in epidemiologic and clinical studies', *Am. J. Epidemiol.*, 104, pp. 493–498.

Diggle, P.J., Liang, K-Y, & Zeeger, S.L. 1994, *Analysis of Longitudinal Data*, Oxford University Press, New York.

Fisz, M. 1963, *Probability Theory and Mathematical Statistics*, 3rd edn, Wiley, New York.

Fukuda, M., Fukuda, K., Shimizu, T., Andersen, C.Y. & Byskov, A.G. 2002, 'Parental periconceptional smoking and male:female ratio of new born infants', *The Lancet*, 359, pp. 1407–1408.

Holm, S. 1979, 'A Simple Sequentially Rejective Multiple Test Procedure', *Scandinavian J. of Statistics*, 6, No 2, pp. 65–70.

Hsiao, C. 2002, Analysis of Panel Data, Cambridge University Press, Cambridge.

McNemar, Q. 1947, 'Note on the sampling error of the difference between correlated proportions or percentages', *Psychometrika*, 12(2), pp. 153–157.

Petzold, M. & Jonsson, R. 2003, 'Maximum Likelihood Ratio-based small sample tests for random coefficients in linear regression', *Working Paper in Economics,* No 110, 2003.

Rao, C.R. 1965, *Linear Statistical Inference and Its Applications*, Wiley, New York.

Scheaffer, R.L., Mendenhall, W. Ott, R.L. & Gerow, K. 2012, *Survey Sampling*, 7th edn, Brooks/Cale CENGAGE Learning.

Shuster, J.J. 1992, 'Exact unconditional tables for significance in the 2 × 2 multinomial trial', *Statistics in Medicine*, 11, pp. 913–922.

Stuart, A., Ord, J.K. & Arnold, S. 1999, *Kendall's Advanced Theory of Statistics*, Vol 2A, Arnold, London.

Wackerly, D., Mendenhall, W. & Scheaffer, R.L. 2007, *Mathematical Statistics with Applications*, 7th edn, Thomson, Toronto.

Wonnacott, T. 1987, 'Confidence intervals or hypothesis tests?', *J. of Applied Statistics*, 14, No 3, pp. 195–201.

Yates, F. 1934, 'Contingency table involving small numbers and the χ^2 test', *Supplement to the J. of the Royal Statistical Society*, 1(2), 217–235.