A First Course in Mathematical Statistics

MATH 472 Handout, Spring 04

Michael Nussbaum*

May 3, 2004

*Department of Mathematics,
Malott Hall, Cornell University,
Ithaca NY 14853-4201,
e-mail nussbaum@math.cornell.edu,
http://www.math.cornell.edu/~nussbaum

CONTENTS

0.2 Preface

"Spring. 4 credits. Prerequisite: MATH 471 and knowledge of linear algebra such as taught in MATH 221. Some knowledge of multivariate calculus helpful but not necessary.
Classical and recently developed statistical procedures are discussed in a framework that emphasizes the basic principles of statistical inference and the rationale underlying the choice of these procedures in various settings. These settings include problems of estimation, hypothesis testing, large sample theory." (The Cornell Courses of Study 2000-2001).
This course is a sequel to the introductory probability course MATH471. These notes will be used as a basis for the course **in combination with a textbook** (to be found among the references given below).

0.3 References

[BD] Bickel, P., Doksum, K., *Mathematical Statistics, Basic Ideas and Selected Topics, Vol. 1,* (2d Edition), Prentice Hall, 2001

[CB] Casella, G. and R. Berger, R. *Statistical Inference*, Duxbury Press, 1990.

[D] Durrett, R., *The Essentials of Probability,* Duxbury Press, 1994.

[DE] Devore, J., *Probability and Statistics for Engineering and the Sciences,* Duxbury - Brooks/Cole, 2000

[FPP] Freedman, D., Pisani, R., and Purves, R: *Statistics* (3rd Edition) 1997.

[HC] Hogg, R. V. and Craig, A. T., *Introduction to Mathematical Statistics* (5 Edition), Prentice-Hall, 1995

[HT] Hogg, R. V. and Tanis, E. A.., *Probability and Statistical Inference* (6 Edition), Prentice-Hall, 2001

[LM] Larsen, R. and Marx, M., *An Introduction to Mathematical Statistics and its Applications,* Prentice Hall 2001

[M] Moore, D. *The Basic Practice of Statistics*, (2d Edition), W. H. Freeman and Co, 2000

[R] Rice, J., *Mathematical Statistics and Data Analysis,* Duxbury Press, 1995

[ROU] Roussas, G., *A Course in Mathematical Statistics*, (2d Edition), Academic Press, 1997

[RS] Rohatgi, V and Ehsanes Saleh, A. K., *An Introduction to Probability and Statistics,* John Wiley 2001

[SH] Shao, Jun, *Mathematical Statistics,* Springer Verlag, 1998

[TD] Tamhane, A and Dunlop, D., *Statistics and Data Analysis: from Elementary to Intermediate* Prentice Hall 2000

Chapter 1
INTRODUCTION

1.1 Hypothesis testing

This course is a sequel to the introductory probability course Math471, the basis of which has been the book "The Essentials of Probability" by R. Durrett (quoted as [D] henceforth). Some statistical topics are already introduced there. We start by discussing some sections and examples (essentially reproducing, sometimes extending the text of [D]).

Testing biasedness of a roulette wheel ([D] chap. 5.4 p. 244)**.** Suppose we run a casino and we wonder if our roulette wheel is biased. A roulette wheel has 18 outcomes that are red, 18 that are black and 2 that are green. If we bet \$1 on red then we win \$1 with probability $p = 18/38 = 0.4736$ and lose \$1 with probability 20/38 (see [D] p. 81 on gambling theory for roulette). To phrase the biasedness question in statistical terms, let p be the probability that red comes up and introduce two hypotheses:

$$\begin{aligned} H_0 &: p = 18/38 \text{ null hypothesis} \\ H_1 &: p \neq 18/38 \text{ alternative hypothesis} \end{aligned}$$

To test to see if the null hypothesis is true, we spin the roulette wheel n times and let $X_i = 1$ if red comes up on the ith trial and 0 otherwise, so that $\bar{X}_n$ is the fraction of times red comes up in the first n trials. The test is specified by giving a critical region $\mathcal{C}_n$ so that we reject H_0 (that is, decide H_0 is incorrect) when $\bar{X}_n \in \mathcal{C}_n$. One possible choice in this case is

$$\mathcal{C}_n = \left\{ x : \left| x - \frac{18}{38} \right| > 2\sqrt{\frac{18}{38} \cdot \frac{20}{38}}/\sqrt{n} \right\}.$$

This choice is motivated by the fact that if H_0 is true then using the central limit theorem (ξ is a standard normal variable)

$$P\left(\bar{X}_n \in \mathcal{C}_n\right) = P\left(\left| \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \right| \geq 2 \right) \approx P\left(|\xi| \geq 2\right) = 0.05. \tag{1.1}$$

Rejecting H_0 when it is true is called a **type I error.** . In this test we have set the type I error to be 5%.

The basis for the approximation "≈" is the **central limit theorem**. Indeed the results X_i, $i = 1, \ldots, n$ of the n trials are independent random variables all having the same distribution (or probability law). This distribution is a Bernoulli law, where X_i takes only values 0 and 1:

$$P\left(X_i = 0\right) = 1 - p, \; P\left(X_i = 1\right) = p.$$

If μ is the expectation of this law then

$$\mu = EX_1 = p.$$

If σ^2 is the variance then

$$\sigma^2 = EX_1^2 - (EX_1)^2 = p - p^2 = p(1-p).$$

The central limit theorem (CLT) gives

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \stackrel{L}{\Longrightarrow} N(0,1) \tag{1.2}$$

where $N(0,1)$ is the standard normal distribution and $\stackrel{L}{\Longrightarrow}$ signifies convergence in law (in distribution). Thus as $n \to \infty$

$$P\left(\left|\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right| \geq 2\right) \to P\left(|\xi| \geq 2\right).$$

So in fact our reasoning is based on a *large sample approximation* for $n \to \infty$. The relation $P\left(|\xi| \geq 2\right) = 0.05$ is then taken from a tabulation of the standard normal law $N(0,1)$.
Now $\sqrt{\frac{18}{38} \cdot \frac{20}{38}} = 0.4993 \approx 1/2$ so to simplify the arithmetic the test can be formulated as

$$\text{reject } H_0 \text{ if } \left|\bar{X}_n - \frac{18}{38}\right| > \frac{1}{\sqrt{n}}$$

or in terms of the total number of reds $S_n = X_1 + \ldots + X_n$

$$\text{reject } H_0 \text{ if } \left|S_n - \frac{18n}{38}\right| > \sqrt{n}.$$

Suppose now that we spin the wheel $n = 3800$ times and get red 1868 times. Is the wheel biased? We expect $18n/38 = 1800$ reds, so the excess number of reds is $|S_n - 1800| = 68$. Given the large number of trials, this might not seem like a large excess. However $\sqrt{3800} = 61.6$ and $68 > 61.6$ so we reject H_0 and think "if H_0 were correct then we would see an observation this far from the mean less than 5% of the time.
Testing academic performance ([D] chap. 5.4 p. 244). Do married college students with children do less well because they have less time to study or do they do better because they are more serious ? The average GPA at the university is 2.48, so we might set up the following hypothesis test concerning μ, the mean grade point average of married students with children:

$$\begin{aligned} H_0 &: \mu = 2.48 \text{ null hypothesis} \\ H_1 &: \mu \neq 2.48 \text{ alternative hypothesis.} \end{aligned}$$

Suppose that to test this hypothesis we have records $X_1, \ldots, X_n$ of 25 married college students with children. Their average GPA is $\bar{X}_n = 2.35$ and sample standard deviation $\hat{\sigma}_n = 0.5$. Recall that the standard deviation of a sample $X_1, \ldots, X_n$ with sample mean $\bar{X}_n$ is

$$\hat{\sigma}_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2}.$$

Using (1.1) from the last example we see that to have a test with a type I error of 5% we should reject H_0 if

$$\left|\bar{X}_n - 2.48\right| > \frac{2\sigma}{\sqrt{n}}.$$

The basis is again the central limit theorem: no particular assumption is made about the distribution of the X_i, they are just *independent and identically distributed random variables* (i.i.d. r.v.'s) with a finite variance σ^2 (standard deviation σ). We again have the CLT (1.2) and we can take $\mu = 2.48$ to test our hypothesis. But contrary to the previous example, the value of σ is then still undetermined (in the previous example both μ and σ are given by p). Thus σ is unknown, but we can estimate it by the sample standard deviation $\hat{\sigma}_n$. Taking $n = 25$ we see that

$$\frac{2(0.5)}{\sqrt{25}} = 0.2 > 0.13 = \left|\bar{X}_n - 2.48\right|$$

so we are not 95% certain that $\mu \neq 2.48$. Note the inconclusiveness in the outcome: the result of the test is the negative statement "we are not 95% certain that H_0 is not true.. ", but not that there is particularly strong evidence for H_0. That nonsymmetric role of the two hypotheses is specific to the setup of statistical testing; this will be discussed in detail later.

Testing the difference of two means.([D] p. 245) Suppose we have independent random samples of size n_1 and n_2 from two populations with unknown means μ_1, μ_2 and variances σ_1^2, σ_2^2 and we want to test

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \text{ null hypothesis} \\ H_1 &: \mu_1 \neq \mu_2 \text{ alternative hypothesis.} \end{aligned}$$

Now the CLT implies that

$$\bar{X}_1 \approx N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \; -\bar{X}_2 \approx N\left(-\mu_1, \frac{\sigma_2^2}{n_2}\right).$$

Indeed these are just reformulations of (1.2) using properties of the normal law:

$$\mathcal{L}(\xi) = N(0,1) \text{ if and only if } \mathcal{L}\left(\frac{\sigma}{\sqrt{n}}\xi + \mu\right) = N\left(\mu, \frac{\sigma^2}{n}\right),$$

and $\mathcal{L}(-\xi) = \mathcal{L}(\xi) = N(0,1)$. Here we are using the standard notation $\mathcal{L}(\xi)$ for "probability law of ξ" (law of ξ, distribution of ξ). We have assumed that the two samples are independent, so if H_0 is correct,

$$\bar{X}_1 - \bar{X}_2 \approx N\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

Based on the last result, if we want a test with a type I error of 5% then we should

$$\text{reject } H_0 \text{ if } \left|\bar{X}_1 - \bar{X}_2\right| > 2\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

For a concrete example we consider a study of passive smoking reported in the *New England Journal of Medicine.* (cf. [D] p.246). A measurement of the size S of lung airways called "FEF 25-75%" was taken for 200 female nonsmokers who were in a smoky environment and for 200 who were not.

In the first group the average value of S was 2.72 with a standard deviation of 0.71 while in the second group the average was 3.17 with a standard deviation of 0.74 (Larger values are better.). To see that there is a significant difference between the averages we note that

$$2\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} = 2\sqrt{\frac{(0.71)^2}{200}+\frac{(0.74)^2}{200}} = 0.14503$$

while $\left|\bar{X}_1 - \bar{X}_2\right| = 0.45$. With these data, H_0 is rejected, based on the reasoning, similar to the first example, "if H_0 were true, then what we are seeing would be very improbable, i.e. would have probability not higher than 0.05." But again the reasoning is based on a normal approximation, i.e. an belief that sample size 200 is a large enough.

1.2 What is statistics ?

The Merriam-Webster Dictionary says:
"Main Entry: **sta·tis·tics**
Pronunciation: st&-'tis-tiks
Function: noun plural but singular or plural in construction
Etymology: German *Statistik* study of political facts and figures, from New Latin *statisticus* of politics, from Latin *status* state
Date: 1770
1 : a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data
2 : a collection of quantitative data"
In ENCYCLOPÆDIA BRITANNICA we find
" **Statistics:** the science of collecting, analyzing, presenting, and interpreting data. Governmental needs for census data as well as information about a variety of economic activities provided much of the early impetus for the field of statistics. Currently the need to turn the large amounts of data available in many applied fields into useful information has stimulated both theoretical and practical developments in statistics. Data are the facts and figures that are collected, analyzed, and summarized for presentation and interpretation. Data may be classified as either quantitative or qualitative. Quantitative data measure either how much or how many of something, and qualitative data provide labels, or names, for categories of like items". ..... "Sample survey methods are used to collect data from observational studies, and experimental design methods are used to collect data from experimental studies. The area of descriptive statistics is concerned primarily with methods of presenting and interpreting data using graphs, tables, and numerical summaries. Whenever statisticians use data from a sample–i.e., a subset of the population–to make statements about a population, they are performing statistical inference. Estimation and hypothesis testing are procedures used to make statistical inferences. Fields such as health care, biology, chemistry, physics, education, engineering, business, and economics make extensive use of statistical inference. Methods of probability were developed initially for the analysis of gambling games. Probability plays a key role in statistical inference; it is used to provide measures of the quality and precision of the inferences. Many of the methods of statistical inference are described in this article. Some of these methods are used primarily for single-variable studies, while others, such as regression and correlation analysis, are used to make inferences about relationships among two or more variables."
The subject of this course is **statistical inference** Let us again quote Merriam-Webster:
"Main Entry: **in·fer·ence**
Pronunciation: 'in-f(&-)r&n(t)s, -f&rn(t)s

Function: *noun*
Date: 1594
1 : the act or process of inferring: as a : the act of passing from one proposition, statement, or judgment considered as true to anotherwhose truth is believed to follow from that of the former b : the act of passing from statistical sample data to generalizations (as of the value of population parameters) usually with calculated degrees of certainty."

1.3 Confidence intervals

Suppose a country votes for president, and there are two candidates, A and B. An opinion poll institute wants to predict the outcome, by sampling a limited number of voters. We assume that 10 days ahead of the election, all voters have formed an opinion, no one intends to abstain, and all voters are willing answer the opinion poll if asked (these assumptions are not realistic, but are made here in order to explain the principle). The proportion intending ot vote for A is p, where $0 < p < 1$, so if a voter is picked at random and asked, the probability that he favors A is p. The proportion p is unknown; if $p > 1/2$ then A wins the election.
The institute samples n voters, and assigns value 1 to a variable X_i if the vote intention is A, 0 otherwise ($i = 1, \ldots, n.$). The institute selects the sample in a random fashion, throughout the voter population, so that X_i can be assumed to be independent Bernoulli $B(1, p)$ random variables. (Again in practice the choice is not entirely random, but follows some elaborate scheme in order to capture different parts of the population, but we disregard this aspect). Recall that p is unknown; an estimate of p is required to form the basis of a prediction ($< 1/2$ or $> 1/2$?). In the theory of statistical inference a common notation for estimates based on a sample of size n is $\hat{p}_n$. Suppose that the institute decides to use the sample mean

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

as an estimate of p, so $\hat{p}_n = \bar{X}_n$.

1.3.1 The Law of Large Numbers

We have for $\bar{X}_n = \hat{p}_n$ and any $\varepsilon > 0$

$$P(|\hat{p}_n - p| > \varepsilon) \to 0 \text{ as } n \to \infty. \tag{1.3}$$

In words, for any small fixed number ε the probability that $\hat{p}_n$ is outside the interval $(p-\varepsilon, p+\varepsilon)$ can be made arbitrarily small by selecting n sufficiently large. If the institute samples enough voters, it can believe that its estimate $\hat{p}_n$ is close enough to the true value. In statistics, estimates which converge to the true value in the above probability sense are called **consistent estimates** (recall that the convergence type (1.3) is called *convergence in probability*). As a basic requirement the institute needs a good estimate $\hat{p}_n$ to base its prediction upon. The LLN tells the institute that it actually pays to get more opinions.
Suppose the institute has sampled a large number of voters, and the estimate $\hat{p}_n$ turns out $> 1/2$ but near. It is natural to proceed with caution in this case, as the reputation of the institute depends on the reliability of its published results. Results which are deemed unreliable will not be published. A controversy might arise within the institute:
Researcher *a*: *'We spent a large amount of money on this poll on we have a really large n. So let us go ahead and publish the result that A will win.'*

Researcher *b*: '*This result is too close to the critical value. I do not claim that B will win, but I favor not to publish a prediction'.*
Clearly a sound and rational criterion is needed for a decision whether to publish or not. A method for this should be fixed in advance.

1.3.2 Confidence statements with the Chebyshev inequality

Recall Chebyshev's inequality ([D] chap. 5.1, p. 222). If Y is a random variable with finite variance $\mathrm{Var}(Y)$ and $Y > 0$ then

$$P\left(|Y - EY| \geq y\right) \leq \mathrm{Var}(Y)/y^2.$$

Applying this for $Y = \bar{X}_n = \hat{p}_n$, we obtain

$$P(|\hat{p}_n - p| > \varepsilon) \leq \frac{\mathrm{Var}(X_1)}{n\varepsilon^2} = \frac{p(1-p)}{n\varepsilon^2} \tag{1.4}$$

for any $\varepsilon > 0$. Suppose we wish to guarantee that

$$P(|\hat{p}_n - p| > \varepsilon) \leq \alpha \tag{1.5}$$

for an α given in advance (e.g. $\alpha = 0.05$ or $\alpha = 0.01$). Now $p(1-p) \leq \frac{1}{4}$ so

$$P(|\hat{p}_n - p| > \varepsilon) \leq \frac{1}{4n\varepsilon^2} \leq \alpha$$

provided we select $\varepsilon = 1/\sqrt{4n\alpha}$.
The Chebyshev inequality thus allows to to **quantitatively assess** the accuracy when the sampe size is given (or alternatively, to determine the neccessary sample size to attain to a given desired level of accuracy ε). The convergence in probability (or LLN) (1.3) is just a **qualitative statement** on $\hat{p}_n$; actually it also derived from the Chebyshev inequality (cf. the proof of the LLN).
Another way of phrasing (1.5) would be: '*the probability that the interval* $(\hat{p}_n - \varepsilon, \hat{p}_n + \varepsilon)$ *covers* p *is more than* 95%', or

$$P((\hat{p}_n - \varepsilon, \hat{p}_n + \varepsilon) \ni p) \geq 1 - \alpha. \tag{1.6}$$

Such statements are called **confidence statements**, and the interval $(\hat{p}_n - \varepsilon, \hat{p}_n + \varepsilon)$ is a **confidence interval** Note that $\hat{p}_n$ is a random variable, so the interval is in fact a random interval. Therefore the element sign $\in$ is written in reverse form $\ni$ to stress the fact that (1.6) the interval is random, not p (p is merely unknown).
To be even more precise, we note that the probability law depends on p, so we should properly write P_p (as is done usually in statistics, where the probabilities depend on an unknown parameter). So we have

$$P_p((\hat{p}_n - \varepsilon, \hat{p}_n + \varepsilon) \ni p) \geq 1 - \alpha. \tag{1.7}$$

When α is a preset value, and $(\hat{p}_n - \varepsilon, \hat{p}_n + \varepsilon)$ is known to fulfill (1.7), the common practical point of view is: '*we believe that our true unknown* p *is within distance* ε *of* $\hat{p}_n$.' When $\hat{p}_n$ happens to be more than ε away from 1/2 (and e.g larger) then the opinion poll institute has enough evidence; this immediately implies '*we believe that our true unknown* p *is greater than* 1/2', and they can go ahead and publish the result. They know that if the true p is actually less than 1/2, then the

outcome they see ($\hat{p}_n \geq 1/2 + \varepsilon$) has less 0.05 probability:

$$\begin{aligned} P_p(\hat{p}_n &\geq 1/2 + \varepsilon) = 1 - P_p(\hat{p}_n - \varepsilon < 1/2) \\ &\leq 1 - P_p(\hat{p}_n - \varepsilon < p) \\ &\leq 1 - P_p(\hat{p}_n - \varepsilon < p < \hat{p}_n + \varepsilon) \\ &= 1 - P_p((\hat{p}_n - \varepsilon, \hat{p}_n + \varepsilon) \ni p) \\ &\leq \alpha. \end{aligned}$$

Note that the α is to some extent arbitrary, but common values are $\alpha = 0.05$ (95% confidence) and $\alpha = 0.01$ (99% confidence).
The reasoning " *When I observe a fact (an outcome of a random experiment) and I know that under a certain hypothesis, this fact would have less than* 5% *probability then I reject this hypothesis'* is very common; it is the basis of statistical testing theory. In our case of confidence intervals, the 'fact' (event) would be "1/2 is not within $(\hat{p}_n - \varepsilon, \hat{p}_n + \varepsilon)$" and the hypothesis would be $p = 1/2$. When we reject $p = 1/2$ because of $\hat{p}_n \geq 1/2 + \varepsilon$ then we can also reject all values $p < 1/2$.
But this type of decision rule (rational decision making, testing) cannot give reasonable certainty in all cases. When 1/2 is in the 95% confidence interval the institute would be well advised to be cautious, and not publish the result. It just means '*unfortunately, I did not observe a fact which would be very improbable under the hypothesis*', so there there is not enough evidence against the hypothesis; nothing can really be ruled out. .
In summary: the prediction is suggested by $\hat{p}_n$; the confidence interval is a rational way of deciding whether to publish or not.
Note that, contrary to the above testing examples, the confidence interval did not involve any large sample approximation. However such arguments (normal approximation, estimate $\mathrm{Var}(X_1)$ by $\hat{p}_n(1 - \hat{p}_n)$) can alternatively be used here.

Chapter 2
ESTIMATION IN PARAMETRIC MODELS

2.1 Basic concepts

After hypothesis testing and confidence intervals, let us introduce the third major branch of statistical inference: *parameter estimation.*

Recall that *a random variable* is a number depending on the outcome of an experiment. When the space of outcomes is Ω, then a random variable X is a function on Ω with values in a space $\mathcal{X}$, written $X(\omega)$. The ω is often omitted.

We also need the concept of a *realization* of a random variable. This is a particular value $x \in \mathcal{X}$ which the random variable has taken, i.e. when ω has taken a specific value. Conceptually, by "random variable" we mean the whole function $\omega \mapsto X(\omega)$, whereas the realization is just a value $x \in \mathcal{X}$ (the "data" in a statistical context).

Population and sample.

Suppose a large shipment of transistors is to be inspected for defective ones. One would like to know the proportion of defective transistors in the shipment; assume it is p where $0 \leq p \leq 1$ (p is unknown). A sample of n transistors is taken from the shipment and the proportion of defective ones is calculated. This is called the sample proportion:

$$\hat{p} = \frac{\#\{\text{defectives in sample}\}}{n}. \tag{2.1}$$

Here the shipment is called the **population** in the statistical problem and p is called a **parameter** of the population. When we randomly select one individual (transistor) from the population, this transistor is defective with probability p. We may define a random variable X_1 in the following way:

$$\begin{aligned} X_1 &= 1 \text{ if defective} \\ X_1 &= 0 \text{ otherwise.} \end{aligned}$$

That is, X_1 takes value 1 with probability p and value 0 with probability $1-p$. Such a random variable is called a **Bernoulli random variable** and the corresponding probability distribution is the Bernoulli distribution, or Bernoulli law , written $B(1,p)$. The sample space of X_1 is the set $\{0,1\}$.

When we take a sample of n transistors, this should be a **simple random sample,** which means the following:

a) each individual is equally likely to be included in the sample

b) results for different individuals are independent one from another.

In mathematical language, a simple random sample of size n yields a set $X_1, \ldots, X_n$ of independent, identically distributed random variables (i.i.d. random variables).. They are identically distributed,

in the above example, because they all follow the Bernoulli law $B(1,p)$.(they are a random selection from the population which has population proportion p). The $X_1,\ldots,X_n$ are independent as random variables because of property b) of a simple random sample.
Denote $X=(X_1,\ldots,X_n)$ the totality of observations, or the vector observations. This is now a random variable with values in the n-dimensional Euclidean space $\mathbb{R}^n$. (We also call this a random vector, or also a random variable with values in $\mathbb{R}^n$. Some probability texts assume random variables to take values in $\mathbb{R}$ only; the higher dimensional objects are called random elements or random vectors.) The sample space of X is now the set of all sequences of length n which cosnsist of 0's and 1′s, written also $\{0,1\}^n$. In general, we denote $\mathcal{X}$ the sample space of an observed random variable X..

Notation *Let X be a random variable with values in a space $\mathcal{X}$. We write $\mathcal{L}(X)$ for the probability distribution (or the law) of X.*

Recall that the probability distribution (or the law) of X is given by the the totality of the values X can take, togehter with the associated probabilities. That definition is true for discrete random variables (the totality of values is finite or countable); for continuous random variables the probability density function defines the distribution. When X is real valued, either discrte or continuous, the law of X can also be described by the distribution function

$$F(x)=P\left(X\leq x\right).)$$

In the above exampe, each individial X_i is Bernoulli: $\mathcal{L}(X_i)=B(1,p)$ but the law of $X=(X_1,\ldots,X_n)$ is not Bernoulli: it is the law of n i.i.d. random variables having Bernoulli law $B(1,p)$. (In probability theory, such a law is called the product law, written $B(1,p)^{\otimes n}$). Note that in our statistical problem above, p is not known so we have a whole set of laws for X: all the laws $B(1,p)^{\otimes n}$ where $p\in[0,1]$.
The parametric estimation problem. Let X be an observed random variable with values in $\mathcal{X}$ and $\mathcal{L}(X)$ be the probability distribution (or the law) of X. Assume that $\mathcal{L}(X)$ is unknown, but known to be from a certain set of laws:

$$\mathcal{L}(X)\in\left\{P_\vartheta;\vartheta\in\Theta\right\}.$$

Here ϑ is an index (a parameter) of the law and Θ is called the parameter space (the set of admitted ϑ). The problem is to estimate a function ϑ based on a realization of X. The set $\{P_\vartheta;\vartheta\in\Theta\}$ is also called a *parametric family of laws.*
In the sequel we assume Θ is a subset of the real line $\mathbb{R}$ and $X=(X_1,\ldots,X_n)$ where $X_1,\ldots,X_n$ are independent random variables. Here n is called *sample size.* In most of this section we confine ourselves to the case that $\mathcal{X}$ is a finite set (with the exception of some examples). In the above example, the ϑ above takes the role of the population proportion p; since the population proportion is known to be in $[0,1]$, the parameter space Θ would be $[0,1]$.

Definition 2.1.1 *(i) A **statistic** T is an arbitrary function of the observed random variable X. (ii) As an **estimator** T of ϑ we admit any mapping with values in Θ.*

$$T:\mathcal{X}\mapsto\Theta$$

In this case, for any realization x the statistic $T(x)$ gives the estimated value of ϑ.

Note that $T = T(X)$ is also a random variable. Statistical terminology is such that an "estimate" is a realized value of that random variable, i.e. $T(x)$ above (the estimated value of ϑ), whereas "estimator" denotes the whole function T. (also called "estimating function") . Sometimes the words "estimate" and "estimator" are used synonymously.
Thus an estimator is a special kind of statistic. Other instances of statistics are those used for building tests or confidence intervals.

Notation Since the distribution of X depends on an unknown parameter, we stress this dependence and write

$$P_\vartheta(X \in B),\ E_\vartheta h(X) \text{ etc.}$$

for probabilities, expectations etc. which are computed under the assumption that ϑ is the true parameter of X.

For later reference we write the model of i.i.d. Bernoulli random variables in the following form.

Model M_1 A random vector $X = (X_1, \ldots, X_n)$ is observed, with values in $\mathcal{X} = \{0,1\}^n$; the distribution of X is the joint law $B(1,p)^{\otimes n}$ of n independent and identically distributed Bernoulli random variables X_i each having law $B(1,p)$, where $p \in [0,1]$.

This fits into the above parametric estimation problems as follows: we have to set $\vartheta = p$, $\Theta = [0,1]$.and $P_\vartheta = B(1,p)^{\otimes n}$.

Remark 2.1.2 We set $p = \vartheta$ and write $P_p(\cdot)$ for probabilities depending on the unknown p. Thus for a particular value $x = (x_1, \ldots, x_n) \in \{0,1\}^n$ we have

$$P_p(X = x) = \prod_{i=1}^n P_p(X_i = x_i) = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i}.$$

Note that the above probability can be written

$$p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} = p^z(1-p)^{n-z},$$

where z denotes the value $z = \sum_{i=1}^n x_i$. Thus the probability $P_p(X = x)$ depends only on the number of realized 1's among the $x_1, \ldots, x_n$, or the number of "successes" in n Bernoulli trials (the number of times the original event has occurred). □

The problem is to estimate the parameter p from the data $X_1, \ldots, X_n$. (Note that now we used the word "data" for the random variables, not for the realizations; but this also corresponds to usage in statistics). As estimators all mappings from $\mathcal{X}$ into $[0,1]$ are admitted, for instance the relative frequency of observed 1's:

$$T(X) = n^{-1}\sum_{j=1}^n X_j = \bar{X}_n = \hat{p}_n$$

i.e. the sample mean $\bar{X}_n$ which coincides with the sample proportion $\hat{p}_n$ from (2.1). In the sequel we shall identify certain requirements which "good" estimators have to fulfill, and we shall develop criteria to quantify the performance of estimators.
Let us give sime other examples of parametric estimation problems. In each case, we observe a random vector $X = (X_1, \ldots, X_n)$ consisting of independent and identically distributed random

variables X_i. For the X_i we shall specify a parametric family of laws $\{Q_\vartheta, \vartheta \in \Theta\}$; then this defines the parametric family of laws $\{P_\vartheta, \vartheta \in \Theta\}$ for X, i.e. the specification $\mathcal{L}(X) \in \{P_\vartheta, \vartheta \in \Theta\}$. It suffices to give the family of laws of X_1; then $\mathcal{L}(X_1) \in \{Q_\vartheta, \vartheta \in \Theta\}$ determines the family $\{P_\vartheta, \vartheta \in \Theta\}$.
(i) **Poisson family:** $\{\mathrm{Po}(\lambda),\ \lambda > 0\}$. Here $\Theta = (0, \infty)$.
(ii) **Normal location family:** $\left\{N(\mu, \sigma^2),\ \mu \in \mathbb{R}\right\}$ where σ^2 is fixed (known). Here $\vartheta = \mu$, $\Theta = \mathbb{R}$ and the expectation parameter μ of the normal law describes its location on the real line.
(iii) **Uniform family:** $\{U(0, \vartheta),\ \vartheta > 0\}$ where $U(0, \vartheta)$ is the uniform law with endpoints 0 and ϑ, having density

$$p_\vartheta(x) = \begin{cases} \vartheta^{-1} \text{ for } 0 \le x \le \vartheta \\ 0 \text{ otherwise.} \end{cases}$$

Choosing the best estimator. In the above example involving transistors the sample proportion $\hat{p}$ suggested itself as a reasonable estimator of the population proportion p. However its not a priori clear that this is the best; we may also consider other function of the observations $X = (X_1, \ldots, X_n)$. First of all, we have to define what it means that an estimator is good. A quantitative comparison of estimators is made possible by the approach of *statistical decision theory* . We choose a **loss function** $L(t, \vartheta)$ which measures the loss (inaccuracy) if the unknown parameter ϑ is estimated by a value t. We stress that t must be be chosen as an estimate depending on the data, so the criterion becomes more complicated (randomness intervenes) and the choice of the loss function is just a first step. The loss is assumed to be nonnegative, i.e. the minimal possible loss is zero.
Natural choices, in case $\vartheta \in \Theta \subseteq \mathbb{R}$, are the distance of t and ϑ (estimation error)

$$L(t, \vartheta) = |t - \vartheta|$$

or the quadratic estimation error

$$L(t, \vartheta) = (t - \vartheta)^2. \tag{2.2}$$

Another possible choice is

$$L(t, \vartheta) = \mathbf{1}_{[\delta, \infty)}(|t - \vartheta|) \tag{2.3}$$

for some $\delta > 0$, which means that emphasis is put on the distance being less than δ, not on its actual value.
As has been said above, the value of the loss becomes random when we insert an estimate of t based on the data. The loss then is a random variable $L(T(X), \vartheta)$ where $T(X)$ is the estimator. To judge the performance of an estimator, it is natural to proceed to the *average* or *expected loss* for given ϑ.

Definition 2.1.3 *The risk of an estimator T at parameter value ϑ is*

$$R(T, \vartheta) = E_\vartheta L(T(X), \vartheta).$$

The risk $R(T, \cdot)$ as a function of ϑ is called the ***risk function*** *of the estimator T.*

Note that in our present model ($\vartheta = p$) we do not have to worry about existence of the expected value (since the law $B^n(1, p)$ has finite support). (For later generalizations, note that since L is nonnegative, if its expectation is not finite then it must be $+\infty$ which may then also be regarded as the value of the risk.)

Thus the random nature of the loss is sucessfully dealt with by taking the expectation, but for judging an estimator $T(X)$, the fact remains that ϑ is unknown. Thus we still have a whole risk function $\vartheta \mapsto R(T, \vartheta)$ as a criterion for performance. But it is desirable to express the quality of T by *just one number* and then try to minimize it. There is a further choice involved for the method to obtain such a number; in the sequel we shall discuss several approaches.
Suppose that, rather than reducing the problem in this fashion, we try to minimize the whole risk function simultaneously, i.e. try to find an estimator T^* such that

$$R(T^*, \vartheta) = \min_T R(T, \vartheta) \text{ for all } \vartheta \in \Theta.$$

Such an estimator would be called *a uniformly best estimator* . In general such an estimator will not exist: for each $\vartheta_0 \in \Theta$ consider an estimator

$$T_{\vartheta_0}(x) = \vartheta_0.$$

This estimator ignores the data and always selects ϑ_0; then $R(T_{\vartheta_0}, \vartheta_0) = 0$, i.e. this estimator is very good if ϑ_0 is true. Thus if T^* were uniformly best it would have to compete with T_{ϑ_0}, i.e. fulfill

$$R(T^*, \vartheta) = 0 \text{ for all } \vartheta \in \Theta$$

i.e. T^* always achieves 0 risk. If the risk adequately expresses a distance to the parameter ϑ this means that a *sure decision* is possible, which is not realistic for statistical problems, and possible only if the problem itself is degenerate or trivial. Thus (we argued informally) uniformly best estimators do not exist in general.
There are two ways out of this dilemma:

- reduce the problem to minimizing one characteristic, as argued above (i.e. the maximal risk or an average risk over Θ)
- restrict the class of estimators such as to rule out "unreasonable" competitors like T_{ϑ_0} which are likely to be very bad for most $\vartheta \in \Theta$. Within a restricted class of estimators, a uniformly best one may very well exist.

Consistency of estimators. At least an estimator should converge towards the true (unknown) parameter to be estimated when the sample size increases. To emphasize the dependence of the data vector on the sample size n we write the statistical model

$$\mathcal{L}(X^{(n)}) \in \{P_{\vartheta,n}; \vartheta \in \Theta\}.$$

Recall that *convergence in probability* for a sequence of random variables is denoted by the symbol $\xrightarrow[P]{}$.

Definition 2.1.4 *A sequence $T_n = T_n(X^{(n)})$ of estimators (each based on a sample of size n) for the parameter ϑ is called* ***consistent*** *if for all $\vartheta \in \Theta$*

$$P_{\vartheta,n}\left(|T_n(X^{(n)}) - \vartheta| > \varepsilon\right) \to 0 \text{ as } n \to \infty, \text{ for all } \varepsilon > 0,$$

or in other notation

$$T_n = T_n(X^{(n)}) \xrightarrow[P]{} \vartheta \text{ as } n \to \infty, \text{ if } \mathcal{L}(X^{(n)}) = P_{\vartheta,n}.$$

Proposition 2.1.5 *In Model $\boldsymbol{M}_1$ the estimator*

$$T_n(X_1, \ldots, X_n) = \bar{X}_n$$

is consistent for the probability of success p.

Proof. Since $X_1, \ldots, X_n$ are i.i.d. Bernoulli $B(1, p)$ random variables, we have $EX_1 = p$ and the result immediately follows from the law of large numbers. ■
The consistency requirement restricts class of estimators to "reasonable" ones; consistency can be seen as a minimal requirement. But there are still many mappings of the observation space into the parameter space which define consistent sequences of estimators.

2.2 Bayes estimators

In the Bayesian approach, one assumes that an a priori distribution is given on the parameter space, reflecting one's prior belief about the parameter $\vartheta \in \Theta$. In the case of Model $\mathbf{M}_1$, assume that this prior distribution is given in the form of a density g on the interval $[0, 1]$. For an estimator T, this allows to reduce the risk function $R(T, p)$, $p \in [0, 1]$ to just one number, by integration:

$$B(T) = \int_0^1 R(T, p) g(p) dp \tag{2.4}$$

$$= \int_0^1 E_p (T - p)^2 g(p) dp \tag{2.5}$$

(the latter equality being true in the case of quadratic loss). This quantity $B(T)$ is called *integrated risk* or *mixed risk* of the estimator T. A Bayes estimator T_B then is defined by the property of minimizing the integrated risk:

$$B(T_B) = \inf_T B(T) \tag{2.6}$$

and the **Bayes risk** is minimal integrated risk, i.e. (2.6).
The name "Bayesian" derives from the analogy with the Bayes formula: if B_i, $i = 1, \ldots, k$ is a partition of the sample space then

$$P(A) = \sum_{i=1}^{k} P(A|B_i) \cdot P(B_i).$$

In our example involving the parameter p above, p takes continuous values in the interval $(0, 1)$, but to make the connection to the Bayes formula above, consider a model where the parameter ϑ takes only finitely many values ($\Theta = (\vartheta_1, \ldots, \vartheta_k)$). Consider the case that P is the joint distribution of (X, U) where X is the data and U is a random variable which takes the k possible values of the parameter ϑ ($U \in \Theta$). For $A =$ "$X \in A$" and $B_i =$ "$U = \vartheta_i$" we get

$$P(X \in A) = \sum_{i=1}^{k} P(X \in A | U = \vartheta_i) P(U = \vartheta_i)$$

Here $g_i = P(U = \vartheta_i)$ can be construed as a prior probability that the parameter ϑ takes value for the parameter ϑ_i, and the conditional probabilities $P(X \in A | U = \vartheta_i)$ can be construed as a family of probability measures depending on $\vartheta \in \Theta$. In other words, in the Bayesian approach

a family of probability measures $\{P_\vartheta; \vartheta \in \Theta\}$ is understood as a conditional distribution given $U = \vartheta$. The marginal distribution of ϑ "awaits specification" as a prior distribution based on belief or experience. In our example (2.5), the probabilities $g_i = P(U = \vartheta_i)$, $i = 1, \ldots, k$ are replaced by a probability density $g(p)$, $p \in (0, 1)$. It is also possible of course to admit only finitely many values of p: p_i, $i = 1, \ldots, k$ with prior probabilities g_i; then (2.5) would read

$$B(T) = \sum_{i=1}^{k} E_{p_i} (T - p_i)^2 g_i$$

The expectations $E_{p_i} (T - p_i)^2$ for a given p_i can then be interpreted as conditional expectation given $U = \vartheta_i$, and the integrated risk $B(T)$ above then is an unconditional expecation, namely $B(T) = E (T - p)^2$, the expected loss squared loss $(T - p)^2$ with respect to the joint distribution of observations X and random parameter p.

In the philosophical foundations of statistics, or in theories of how statistical decisions should be made in the real world, the Bayesian approach has developed into an important separate school of thought; those who believe that prior distributions on ϑ should always be applied (and are always available) are sometimes called "Bayesians", and Bayesian statistics is the corresponding part of Mathematical Statistics.

In Model $\mathbf{M}_1$, consider the following family of prior densities for the Bernoulli parameter p: for $\alpha, \beta > 0$

$$g_{\alpha,\beta}(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)},\ p \in [0, 1].$$

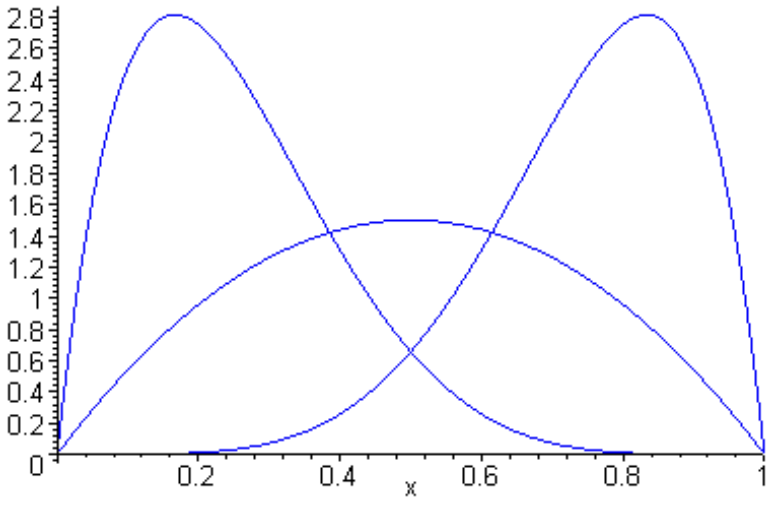
where

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}. \tag{2.7}$$

The corresponding distributions are called the **Beta distributions** ; here $B(\alpha, \beta)$ stands for the beta function defined by (2.7). Recall that

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp(-x) dx.$$

Thus we consider a whole family ($\alpha, \beta > 0$) of possible prior distributions for p, allowing a wide range of choices for "prior belief". The plot below shows three different densities from this family; let us mention that the uniform density on $[0, 1]$ is also in this class (for $\alpha = \beta = 1$). We will discuss the Beta family in more detail later, establishing also that $B(\alpha, \beta)$ is the correct normalization factor. It will also become clear that Bayesian methods are very useful to prove non-Bayesian optimality properties of estimators.



Beta densities for $(\alpha, \beta) = (2,6), (2,2), (6,2)$

Proposition 2.2.1 *In Model $\boldsymbol{M}_1$, let g be an arbitrary prior density for the parameter $p \in [0,1]$. The corresponding Bayes estimator of p is*

$$T_B(x) = \frac{\int_0^1 p^{z(x)+1}(1-p)^{n-z(x)}g(p)dp}{\int_0^1 p^{z(x)}(1-p)^{n-z(x)}g(p)dp} \tag{2.8}$$

for $x = (x_1, \ldots, x_n) \in \{0,1\}^n$ and

$$z(x) = \sum_{i=1}^{n} x_i.$$

Remark. (i) Note that the Bayes estimator depends on the sample x only via the statistic $z(x)$, or equivalently, via the the sample mean $\bar{X}_n(x) = n^{-1}z(x)$.
(ii) The function

$$g_x(p) = \frac{p^{z(x)}(1-p)^{n-z(x)}g(p)}{\int_0^1 p^{z(x)}(1-p)^{n-z(x)}g(p)dp}$$

is a probability density 0n $[0,1]$, depending on the observed x. We see from (2.8) that $T_B(x)$ is the expectation for that density:

$$T_B(x) = \int_0^1 p g_x(p)dp.$$

Proof. For any estimator T and $\mathcal{X} = \{0,1\}^n$

$$B(T) = \sum_{x\in\mathcal{X}} \int_0^1 (T(x)-p)^2 p^{z(x)}(1-p)^{n-z(x)}g(p)dp.$$

To minimize $B(T)$, the value $T(x)$ should be chosen optimally; one could try to do this for every term in the sum, given x. Thus we look for the minimum of

$$b_x(t) = \int_0^1 (t-p)^2 p^{z(x)}(1-p)^{n-z(x)}g(p)dp.$$

This can be written as a polynomial of degree 2 in t:

$$b_x(t) = c_0 - 2c_1 t + c_2 t^2$$

where

$$\begin{aligned} c_2 &= \int_0^1 p^{z(x)}(1-p)^{n-z(x)} g(p)dp, \\ c_1 &= \int_0^1 p^{z(x)+1}(1-p)^{n-z(x)} g(p)dp, \\ c_0 &= \int_0^1 p^{z(x)+2}(1-p)^{n-z(x)} g(p)dp. \end{aligned}$$

Since

$$0 \leq \left| p^{z(x)}(1-p)^{n-z(x)} \right| \leq 1,$$

all these integrals are finite, and $c_2 > 0$. It follows that the unique minimum t_0 of $b_x(t)$ can be obtained by setting the derivative 0. The solution of

$$b'_x(t) = 2tc_2 - 2c_1 = 0$$

is $t_0 = c_1/c_2$; we have $0 \leq c_1 \leq c_2$ since $p \leq 1$. This implies that $0 \leq t_0 \leq 1$, and

$$T_B(x) = t_0(x)$$

is of the form (2.8). ■

2.3 Admissible estimators

Bayes estimators also enjoy an optimality property in terms of the original risk function $p \mapsto R(T_B, p)$ even before a weighted average over the parameter p is taken.

Definition 2.3.1 *An estimator T of a parameter ϑ is called* ***admissible ,*** *if for every estimator S of ϑ, the relation*

$$R(S, \vartheta) \leq R(T, \vartheta) \text{ for all } \vartheta \in \Theta \tag{2.9}$$

implies

$$R(S, \vartheta) = R(T, \vartheta) \text{ for all } \vartheta \in \Theta. \tag{2.10}$$

Thus admissibility means that there can be no estimator S which is uniformly at least as good (2.9 holds), and strictly better for one $\vartheta_0 \in \Theta$: then $R(S, \vartheta_0) < R(T, \vartheta_0)$ contradicts (2.10). Non-admissibility of T means that T can be improved by another estimator S.

Proposition 2.3.2 *Suppose that in Model $\boldsymbol{M}_1$, the prior density g is such that $g(p) > 0$ for all $p \in [0,1]$ with the exception of a finite number of points p. Then the Bayes estimator T_B for this prior density is admissible, for quadratic loss.*

Proof. Suppose that T_B is not admissible. Thus there is an estimator S and a $p_0 \in [0,1]$ with

$$\begin{aligned} R(S,p) &\leq R(T_B,p) \text{ for all } p \in [0,1], \\ \text{and } R(S,p_0) &< R(T_B,p_0). \end{aligned} \tag{2.11}$$

Note that $R(S,p)$ is continuous in p (continuous on $p \in (0,1)$, and right resp. left continuous at the endpoints 0 and 1):

$$R(S,p) = \sum_{x \in \mathcal{X}} (S(x)-p)^2 \, p^{z(x)} (1-p)^{n-z(x)}$$

for $z(x) = \sum_{i=1}^n x_i$, $x \in \mathcal{X}$. Thus (2.11) implies that there must be whole neighborhood of p_0 within which S is better: for some $\varepsilon > 0$

$$R(S,p) < R(T_B,p)\,,\ p \in [p_0-\varepsilon, p_0+\varepsilon] \cap [0,1].$$

It follows that

$$B(S) = \int_0^1 R(S,p) g(p) dp < \int_0^1 R(T_B,p) g(p) dp = B(T_B).$$

This contradicts the fact that T_B is a Bayes estimator; thus T_B must be admissible. ■

2.4 Bayes estimators for Beta densities

Let us specify (2.8) to the case of Beta densities $g_{\alpha,\beta}(p)$ introduced above. We have

$$T_B(x) = \frac{\int_0^1 p^{z(x)+1}(1-p)^{n-z(x)} g_{\alpha,\beta}(p) dp}{\int_0^1 p^{z(x)}(1-p)^{n-z(x)} g_{\alpha,\beta}(p) dp} \tag{2.12}$$

$$= \frac{\int_0^1 p^{\alpha+z(x)}(1-p)^{\beta+n-z(x)-1} dp}{\int_0^1 p^{\alpha+z(x)-1}(1-p)^{\beta+n-z(x)-1} dp}. \tag{2.13}$$

With a partial integration we obtain for $\gamma, \delta > 0$

$$\int_0^1 p^\gamma (1-p)^{\delta-1} dp = \left[-\frac{1}{\delta} p^\gamma (1-p)^\delta\right]_0^1 + \frac{\gamma}{\delta} \int_0^1 p^{\gamma-1}(1-p)^\delta dp \tag{2.14}$$

$$= \frac{\gamma}{\delta} \int_0^1 p^{\gamma-1}(1-p)^\delta dp. \tag{2.15}$$

(*Technical remark:* for $\gamma < 1$, the function $p^{\gamma-1}$ tends to ∞ as $p \to 0$, so the second integral in (2.14) is improper [a limit of $\int_t^1$ for $t \searrow 0$], similarly in case of $\delta < 1$. Hence, strictly speaking, one should argue in terms of $\int_t^1$ first and then take a limit. Note that these limits exist since the function $p^{\gamma-1}$ is integrable on $(0,1)$ for $\gamma > 0$:

$$\int_t^1 p^{\gamma-1}(1-p)^\delta dp \leq \int_t^1 p^{\gamma-1} dp = \left[\frac{1}{\gamma} p^\gamma\right]_t^1 = \frac{1}{\gamma}(1-t^\gamma) \leq \frac{1}{\gamma}. \ )$$

Relation (2.15) implies

$$\begin{aligned}\int_0^1 p^{\gamma-1}(1-p)^{\delta-1}dp &= \int_0^1 (p+1-p)p^{\gamma-1}(1-p)^{\delta-1}dp \\ &= \int_0^1 p^{\gamma}(1-p)^{\delta-1}dp + \int_0^1 p^{\gamma-1}(1-p)^{\delta}dp \\ &= \left(1+\frac{\delta}{\gamma}\right)\int_0^1 p^{\gamma}(1-p)^{\delta-1}dp\end{aligned}$$

(for the last equality, we reversed the roles of δ and γ in (2.15)). Setting now $\gamma = \alpha + z(x)$, $\delta = \beta + n - z(x)$, we obtain from (2.13)

$$\begin{aligned}T_B(x) &= \frac{\int_0^1 p^{\gamma}(1-p)^{\delta-1}dp}{\int_0^1 p^{\gamma-1}(1-p)^{\delta-1}dp} = \left(1+\frac{\delta}{\gamma}\right)^{-1} \\ &= \frac{\gamma}{\gamma+\delta} = \frac{\alpha+z(x)}{\alpha+\beta+n},\end{aligned}$$

thus (:= means defining equality)

$$T_{\alpha,\beta}(X) := T_B(X) = \frac{\bar{X}_n + \alpha/n}{1+\alpha/n+\beta/n}.$$

We already know that the Bayes estimator is a function of the sample mean; we specified formula (2.8).

Let us discuss some limiting cases.

Limiting case A) Sample size n is fixed, $\alpha \to 0$, $\beta \to 0$. In the limit $\bar{X}_n$ is obtained. However, the family of densities $g_{\alpha,\beta}$ does not converge to a density for $\alpha \to 0$, $\beta \to 0$, since the function

$$g^*(p) = p^{-1}(1-p)^{-1}$$

is not integrable. Indeed

$$\int_t^{1/2} p^{-1}(1-p)^{-1}dp \geq \int_t^{1/2} p^{-1}dp = \log(1/2) - \log t \to \infty \text{ for } t \to \infty.$$

This means that $\bar{X}_n$ is not a Bayes estimator for one of the densities $g_{\alpha,\beta}$.

Limiting case B) α and β are fixed, sample size $n \to \infty$. In this case we have for $p \in (0,1)$

$$T_{\alpha,\beta}(X) = \bar{X}_n + o_p(n^{-\gamma}) \text{ for all } 0 \leq \gamma < 1. \tag{2.16}$$

Definition 2.4.1 *A sequence of r.v. Z_n is called $o_p(n^{-\gamma})$ if*

$$n^{\gamma} Z_n \to_p 0,\ n \to \infty.$$

Relation (2.16) thus means that

$$n^{\gamma}\left|T_{\alpha,\beta}(X) - \bar{X}_n\right| \to_p 0.$$

To see (2.16), note that

$$\frac{\bar{X}_n + \alpha/n}{1+\alpha/n+\beta/n} - \bar{X}_n = \frac{\alpha/n - \bar{X}_n(\beta/n+\alpha/n)}{1+\alpha/n+\beta/n}$$

and since $\bar{X}_n \to_P p$, it is obvious that the above quantity is $o_p(n^{-\gamma})$. At the same time, $\bar{X}_n$ converges in probability to p, but slower: we have

$$\bar{X}_n = p + o_p(n^{-\gamma}) \text{ for all } 0 \le \gamma < 1/2. \tag{2.17}$$

Indeed, $Var_p(\bar{X}_n) = p(1-p)/n$, hence

$$\mathrm{Var}_p(n^\gamma(\bar{X}_n - p)) = n^{2\gamma}\mathrm{Var}_p(\bar{X}_n - p) = n^{2\gamma}\mathrm{Var}_p\bar{X}_n = n^{2\gamma-1}p(1-p)$$

which tends to 0 for $0 \le \gamma < 1/2$ but not for $\gamma = 1/2$. Rather, for $\gamma = 1/2$ we have by the central limit theorem

$$n^{1/2}(\bar{X}_n - p) \to_d N(0, p(1-p))$$

so that $(\bar{X}_n - p)$ is not $o_p(n^{-1/2})$.
The interpretation of (2.16), (2.17) is that $n \to \infty$, the influence of the a priori information diminishes. The Bayes estimators become all close to the sample mean (and to each other) at rate $n^{-\gamma}$, $\gamma < 1$, whereas they converge to the true p only at rate $n^{-\gamma}$, $\gamma < 1/2$.

2.5 Minimax estimators

Let us consider the case $\alpha = \beta = n^{1/2}/2$. For the risk we obtain

$$\begin{aligned} R\left(T_{\alpha,\beta}, p\right) &= E_p\left(T_{\alpha,\beta} - p\right)^2 = \frac{1}{(n+\alpha+\beta)^2} E_p\left(n\bar{X}_n + \alpha - (n+\alpha+\beta)p\right)^2 & (2.18)\\ &= \frac{1}{(n+n^{1/2})^2}\left\{E_p\left(n\bar{X}_n - np\right)^2 + (\alpha(1-p) - \beta p)^2\right\} & (2.19)\end{aligned}$$

(for the last equality, note that $E(Y - EY + a)^2 = E(Y - EY)^2 + a^2$ for nonrandom a). Set $q := 1 - p$ and note that $n\bar{X}_n$ has the binomial distribution $B(n, p)$, so that

$$E_p\left(n\bar{X}_n - np\right)^2 = \mathrm{Var}(n\bar{X}_n) = npq.$$

Hence (2.19) equals

$$\begin{aligned} R\left(T_{\alpha,\beta}, p\right) &= \frac{n}{4(n+n^{1/2})^2}\left(4pq + (p-q)^2\right) = \frac{n}{4(n+n^{1/2})^2}(p+q)^2 \\ &= \frac{1}{4n\left(1+n^{-1/2}\right)^2} =: m_n. \end{aligned}$$

The above reasoning is valid for all $p \in [0,1]$; it means that the risk $R\left(T_{\alpha,\beta}, p\right)$ for this special choice of α, β does not depend upon p. In addition we know that $T_{\alpha,\beta}$ is admissible. This implies that $T_{\alpha,\beta}$ is a *minimax estimator ;* let us define that important notion.

Definition 2.5.1 *In model $\boldsymbol{M}_1$, for any estimator T set*

$$M(T) = \max_{0 \le p \le 1} R(T, p). \tag{2.20}$$

*An estimator T_M is called **minimax** if*

$$M(T_M) = \min_T M(T) = \min_T \max_{0 \le p \le 1} R(T, p).$$

Note that the maximum in (2.20) exists since $R(T,p)$ is continuous in p for every estimator. The minimax approach is similar to the Bayes approach, in that *one characteristic* of the risk function $R(T,p)$, $p \in [0,1]$ is selected as performance criterion for an estimator. In that case the characteristic is the worst case risk .

Theorem 2.5.2 *In model $\boldsymbol{M}_1$, the Bayes estimator $T_{\alpha,\beta}$ for $\alpha = \beta = n^{1/2}/2$ is a minimax estimator for quadratic loss.*

Proof. Let T be an arbitrary estimator of p. Since $T_{\alpha,\beta}$ is admissible according to Proposition 2.3.2, there must be a $p_0 \in [0,1]$ such that

$$R(T,p_0) \geq R\left(T_{\alpha,\beta},p_0\right) = m_n. \tag{2.21}$$

If there were no such p_0, then we would have

$$R(T,p) < R\left(T_{\alpha,\beta},p\right)$$

for all p which contradicts admissibility of $T_{\alpha,\beta}$. Now (2.21) implies

$$M(T) \geq m_n = M(T_{\alpha,\beta}).$$

■

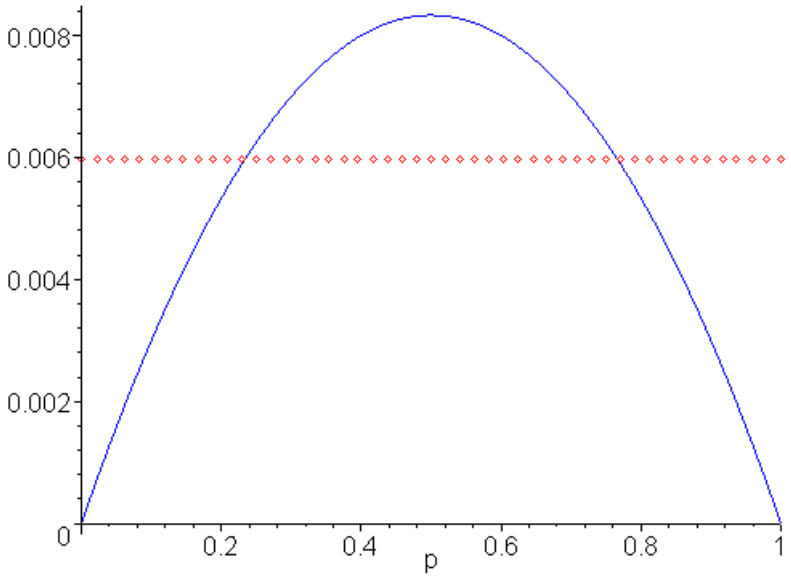
The estimator $T_{\alpha,\beta} = T_M$ has the form

$$T_M = \frac{\bar{X}_n + \alpha/n}{1+\alpha/n+\beta/n} = \frac{\bar{X}_n + n^{-1/2}/2}{1+n^{-1/2}}. \tag{2.22}$$

Let us compare the risk functions of the minimax estimator T_M and of the sample mean. We have

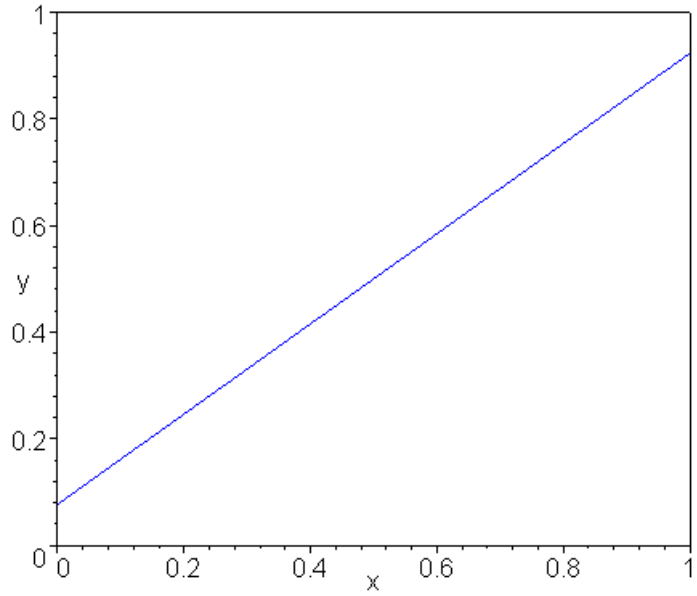
$$\begin{aligned} R\left(\bar{X}_n,p\right) &= \mathrm{Var}_p(\bar{X}_n) = \frac{p(1-p)}{n}, \\ R\left(T_M,p\right) &= m_n = \frac{1}{4n\left(1+n^{-1/2}\right)^2}. \end{aligned}$$

For $n = 30$ the two risk function are plotted below.



Risk of sample mean (line) and minimax risk (dots), n=30

It is clearly visible that the "problem region" for the sample mean is the area around $p = 1/2$; the minimax estimator is better in the center, at the expense of the tail regions, and thus achieves a smaller overall risk.
It is instructive to look at the form of the minimax estimator T_M itself. The function $f_M(x) = \frac{x+n^{-1/2}/2}{1+n^{-1/2}}$ is plotted below, also for $n = 30$.



The minimax estimator as a function of the sample mean

This function has values $f_M(1/2) = 1/2$ and $f_M(0) = \frac{n^{-1/2}/2}{1+n^{-1/2}}$ and it is linear; by symmetry we obtain $f_M(1) = 1 - f_M(0)$. This means that the values of the sample mean are "moved" towards the value $1/2$, which is the one where the risk would be maximal. This can be understood as a kind of prudent behaviour of the minimax estimator, which tends to be closer to the value $1/2$ since it is more damaging to make an error if this value were true. We can also write

$$\begin{aligned} \bar{X}_n &= \frac{1}{2} + (\bar{X}_n - 1/2), \\ T_M &= \frac{1}{2} + \frac{(\bar{X}_n - 1/2)}{1 + n^{-1/2}} \end{aligned}$$

where it is seen that the minimax estimator takes the distance of $\bar{X}_n$ to $1/2$ and "shrinks" it by a factor $(1 + n^{-1/2})^{-1}$ which is < 1.

Chapter 3
MAXIMUM LIKELIHOOD ESTIMATORS

In the general statistical model where X is an observed random variable with values in $\mathcal{X}$, $\mathcal{X}$ is a countable set (i.e. X is a discrete r.v.) and $\{P_\vartheta; \vartheta \in \Theta\}$ is a family of probability laws on $\mathcal{X}$:

$$\mathcal{L}(X) \in \{P_\vartheta; \vartheta \in \Theta\},$$

consider the probability function for given $\vartheta \in \Theta$, i.e. $P_\vartheta(X = x)$. For each $x \in \mathcal{X}$, the function

$$L_x(\vartheta) = P_\vartheta(X = x)$$

is called the **likelihood function** of ϑ given x. The name reflects the heuristic principle that when observations are realized, i.e. X took the value x, the most "likely" parameter values of ϑ are those where $P_\vartheta(X = x)$ is maximal. This does not mean that $L_x(\vartheta)$ gives a probability distribution on the parameter space; the **likelihood principle** has its own independent rationale, on a purely heuristic basis.
Under special conditions however the likelihood function can be interpreted as a probability function. Consider the case that $\Theta = \{\vartheta_1, \ldots, \vartheta_k\}$ is a finite set, and consider a prior distribution on Θ which is uniform:

$$P(U = \vartheta_i) = k^{-1},\ i = 1, \ldots, k.$$

Understanding $P_\vartheta(X = x)$ as a conditional distribution given ϑ, i.e. setting (as always in the Bayesian approach)

$$P(X = x|U = \vartheta) = P_\vartheta(X = x),$$

we immediately obtain a **posterior distribution** of ϑ, i.e. the conditional distribution of U given $X = x$:

$$\begin{aligned} P(U &= \vartheta|X = x) = \frac{P(U = \vartheta, X = x)}{P(X = x)} && (3.1) \\ &= \frac{P_\vartheta(X = x)P(U = \vartheta)}{P(X = x)} = \frac{P_\vartheta(X = x)}{k \cdot P(X = x)} \\ &= L_x(\vartheta) \cdot (k \cdot P(X = x))^{-1}. && (3.2) \end{aligned}$$

Here the factor $(k \cdot P(X = x))^{-1}$ does not depend on ϑ, so that in this case, the posterior probability function of ϑ is proportional to the likelihood function $L_x(\vartheta)$. Recall that the marginal probability function of X is

$$P(X = x) = \sum_{i=1}^{k} P_{\vartheta_i}(X = x)P(U = \vartheta_i)$$

Thus in this special case, the likelihood principle can be derived from the Bayesian approach, for a "noninformative" prior distribution (i.e. the uniform distribution on Θ). However in cases where Θ there is no natural uniform distribution on Θ, such as for $\Theta = \mathbb{R}$ or $\Theta = \mathbb{Z}_+$ (the nonnegative integers), such a reasoning is not straightforward. (A limiting argument for a sequence of prior distributions is often possible).
A **maximum likelihood estimator** (MLE) of ϑ is an estimator $T(x) = T_{ML}(x)$ such that

$$L_x(T_{ML}(x)) = \max_{\vartheta \in \Theta} L_x(\vartheta),$$

i.e. for each given x, the estimator is a value of ϑ which maximizes the likelihood.
In the case of model $\mathbf{M}_1$, the probability function for given $p = \vartheta$ is

$$P_p(X = x) = p^{z(x)}(1-p)^{n-z(x)} = L_x(p).$$

for $x \in \mathcal{X}$. In this case the parameter space is $\Theta = [0,1]$, on which there is a natural uniform distribution (the beta density for $\alpha = \beta = 1$). It can be shown that also in this case, something analogous to (3.2) holds, i.e. the likelihood function is proportional to the density of the posterior distribution. Without proving this statement, let us directly compute the maximum likelihood estimate.
Assume first that x is such that $z(x) = \sum_{i=1}^n x_i \in (0,n)$. Then the likelihood function has $L_x(0) = L_x(1) = 0$ and is positive and continuously differentiable on the open interval $p \in (0,1)$. Thus also the **logarithmic likelihood function**

$$l_x(p) = \log L_x(p)$$

is continuously differentiable, and since log is a monotone function, local extrema of the likelihood function in $(0,1)$ coincide with local extrema of the log-likelihood in $p \in (0,1)$. We have

$$\begin{aligned} l_x(p) &= z(x)\log p + (n - z(x))\log(1-p), \\ l'_x(p) &= z(x)p^{-1} - (n - z(x))(1-p)^{-1}. \end{aligned}$$

We look for zeros of this function in $p \in (0,1)$; the local extrema are among these. We obtain

$$\begin{aligned} z(x)p^{-1} &= (n - z(x))(1-p)^{-1}, \\ z(x)(1-p) &= (n - z(x))p, \\ np &= z(x). \end{aligned}$$

Thus if $z(x) = \sum_{i=1}^n x_i \in (0,n)$ then $\hat{p} = n^{-1}z(x) = \bar{x}_n$ gives the unique local extremum of $l_x(p)$ on $(0,1)$, which must be a maximum of $L_x(p)$.
If either $z(x) = 0$ or $z(x) = n$ then

$$L_x(p) = (1-p)^n \text{ or } L_x(p) = p$$

such that $\hat{p} = 0$ or $\hat{p} = 1$ resp. are maximum likelihood estimates. We have established

Proposition 3.0.3 *In model $\boldsymbol{M}_1$, the sample mean $\bar{X}_n$ is the unique maximum likelihood estimator (MLE) of the parameter $p \in [0,1]$:*

$$T_{ML}(X) = \bar{X}_n.$$

We now turn to the **continuous likelihood principle**. Let X be a random variable with values in $\mathbb{R}^k$ such that $\mathcal{L}(X) \in \{P_\vartheta; \vartheta \in \Theta\}$, and each law P_ϑ has a density $p_\vartheta(x)$ on $\mathbb{R}^k$. For each $x \in \mathbb{R}^k$, the function

$$L_x(\vartheta) = p_\vartheta(x)$$

is called the likelihood function of ϑ given x. A maximum likelihood estimator (MLE) of ϑ is an estimator $T(x) = T_{ML}(x)$ such that

$$L_x(T_{ML}(x)) = \max_{\vartheta \in \Theta} L_x(\vartheta),$$

i.e. for each given x, the estimator is a value of ϑ which maximizes the likelihood.
Let us consider an example in which all densities are Gaussian.

Model $\mathbf{M}_2$ Observed are n independent and identically random variables $X_1, \ldots, X_n$, each having law $N(\mu, \sigma^2)$, where $\sigma^2 > 0$ is given (known) and $\mu \in \mathbb{R}$ is unknown.

This is also called the **Gaussian location model** (or Gaussian shift model). Consider the case of sample size $n = 1$. We can represent X_i as

$$X_i = \mu + \xi_i$$

where ξ_i are i.i.d. centered normal: $\mathcal{L}(\xi_i) = N(0, \sigma^2)$. The parameter ϑ is μ and parameter space Θ is $\mathbb{R}$.

Proposition 3.0.4 *In the Gaussian location model $\mathbf{M}_2$, the sample mean $\bar{X}_n$ is the unique maximum likelihood estimator of the expectation parameter $\mu \in \mathbb{R}$:*

$$T_{ML}(X) = \bar{X}_n.$$

Proof. We have

$$\begin{aligned} L_x(\mu) &= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \end{aligned}$$

thus the logarithmic likelihood function is

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

Note that

$$\begin{aligned} \frac{d}{d\mu} l_x(\mu) &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(-2(x_i-\mu)) \\ &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right) = 0 \end{aligned}$$

if and only if

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

Thus $l_x(\mu)$ which is continuously differentiable on $\mathbb{R}$ has a unique local extremum at $\mu = \bar{x}_n$. Since $l_x(\mu) \to -\infty$ for $\mu \to \pm\infty$, this must be a maximum. ■

We can now ask what happens if the variance σ^2 is also unknown. In that connection we introduce the **Gaussian location-scale model**

Model $\mathbf{M}_3$ Observed are n independent and identically random variables $X_1, \ldots, X_n$, each having law $N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are both unknown.

In addition to the sample mean $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$, consider the statistic

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

This statistic is called the **sample variance** . The *empirical second (central) moment (e.s.m.)* is

$$\tilde{S}_n^2 = \frac{n-1}{n} S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

At first glance, it would seem more natural to call the expression $\tilde{S}_n^2$ the sample variance, since it is the sample analog of the variance of the random variable X_1 :

$$\sigma^2 = \mathrm{Var}(X_1) = E\left(X_1 - EX_1\right)^2.$$

However it is customary in statistics to call S_n^2 the sample variance; the reason is unbiasedness for σ^2 , which will be discussed later. Note that we need a sample size $n \geq 2$ for S_n^2 and $\tilde{S}_n^2$ to be nonzero.

Proposition 3.0.5 *In the Gaussian location-scale model $\boldsymbol{M}_3$, for a sample size $n \geq 2$, if $\tilde{S}_n^2 > 0$ then sample mean and e.s.m. $(\bar{X}_n, \tilde{S}_n^2)$ are the unique maximum likelihood estimators of the parameter $\vartheta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$:*

$$T_{ML}(X) = (\bar{X}_n, \tilde{S}_n^2)$$

The event $\tilde{S}_n^2 > 0$ has probability one for all $\vartheta \in \Theta$.

Proof. We write $\bar{x}_n$, $\tilde{s}_n^2$ for sample mean and e.s.m. when $x_1, \ldots, x_n$ are realized data. Note that

$$\tilde{s}_n^2 = n^{-1} \sum_{i=1}^{n} x_i^2 - (\bar{x}_n)^2.$$

Hence

$$\sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} x_i^2 - 2n\bar{x}_n\mu + n\mu^2$$

$$= \sum_{i=1}^{n} x_i^2 - n\bar{x}_n^2 + n\left(\bar{x}_n - \mu\right)^2 = n\tilde{s}_n^2 + n\left(\bar{x}_n - \mu\right)^2.$$

Thus the likelihood function $L_x(\vartheta)$ is

$$\begin{aligned} L_x(\vartheta) &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\tilde{s}_n^2+(\bar{x}_n-\mu)^2}{2\sigma^2 n^{-1}}\right) \qquad (3.3) \\ &= \frac{1}{\sigma n^{-1/2}}\varphi\left(\frac{\bar{x}_n-\mu}{\sigma n^{-1/2}}\right)\cdot\frac{1}{n^{1/2}(2\pi\sigma^2)^{(n-1)/2}}\exp\left(-\frac{\tilde{s}_n^2}{2\sigma^2 n^{-1}}\right). \qquad (3.4) \end{aligned}$$

where

$$\varphi(x) = \frac{1}{(2\pi)^{1/2}}\exp\left(-\frac{x^2}{2}\right)$$

is the standard normal density. We see that the first factor is a normal density in $\bar{x}_n$ and the second factor does not depend on μ. To find MLE's of μ and σ^2 we first maximize for fixed σ^2 over all possible $\mu \in \mathbb{R}$. The first factor is the likelihood function in model $\mathbf{M}_2$ for a sample size $n = 1$, variance $n^{-1}\sigma^2$ and an observed value $x_1 = \bar{x}_n$. This gives an MLE $\hat{\mu} = \bar{x}_n$. We can insert this value into (3.3); we now have to maximize

$$L_x\left(\bar{x}_n, \sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{\tilde{s}_n^2}{2\sigma^2 n^{-1}}\right)$$

over $\sigma^2 > 0$. For notational convenience, we set $\gamma = \sigma^2$; equivalently, one may minimize

$$\tilde{l}_x(\gamma) = -\log L_x(\bar{x}_n, \gamma) = \frac{n}{2}\log\gamma + \frac{n\tilde{s}_n^2}{2\gamma}.$$

Note that if $\tilde{s}_n^2 > 0$, for $\gamma \to 0$ we have $\tilde{l}_x(\gamma) \to \infty$ and for $\gamma \to \infty$ also $\tilde{l}_x(\gamma) \to \infty$, so that a minimum exists and is a zero of the derivative of $\tilde{l}_x$. The event $\tilde{S}_n^2 > 0$ has probability 1 since otherwise $x_i = \bar{x}_n$, $i = 1, \ldots, n$. i.e. all x_i are equal, which clearly has probability 0 for independent continuous random variables X_i. We obtain

$$\begin{aligned} \tilde{l}_x'(\gamma) &= \frac{n}{2\gamma} - \frac{n\tilde{s}_n^2}{2\gamma^2} = 0, \\ \gamma &= \tilde{s}_n^2 \end{aligned}$$

as the unique zero, so $\hat{\sigma}^2 = \tilde{s}_n^2$ is the MLE of σ^2. ■

For our next example, let us introduce the *double exponential distribution* $DE(\mu, \lambda)$. It has density

$$f(x) = (2\lambda)^{-1}\exp\left(-|x-\mu|\lambda^{-1}\right)$$

for given $\mu \in \mathbb{R}$, $\lambda > 0$. We shall assume $\lambda = 1$ here; clearly the above is a density for any μ since

$$\int_0^\infty \exp(-x)dx = 1.$$

Clearly $DE(\mu, \lambda)$ has finite moments of any order, and by symmetry reasons μ is the expectation. We introduce the **double exponential location model .**

Model $\mathbf{M}_4$ Observed are n independent and identically random variables $X_1, \ldots, X_n$, each having law $DE(\mu, 1)$, where $\mu \in \mathbb{R}$ is unknown.

We will show that the MLE in this case is the sample median. For any vector $(x_1, \ldots, x_n)$, let $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ be the vector of ordered values; this is always uniquely defined. For n i.i.d. random variables $X = (X_1, \ldots, X_n)$, define the **order statistics** to be components of the vector $X_{(1)}, \ldots, X_{(n)}$. Recall that a statistic was any function of the data; thus for given i, the i-th order statistic is a well defined data function. In particular $X_{(n)} = \max_{i=1,\ldots,n} X_i$ etc, $X_{(1)} = \min_{i=1,\ldots,n} X_i$. Define the **sample median** as

$$\mathrm{med}(X) = \begin{cases} X_{((n+1)/2)} \text{ if } n \text{ is uneven} \\ \frac{1}{2}(X_{(n/2)} + X_{(n/2+1)}) \text{ if } n \text{ is even} \end{cases}$$

In other words, the sample median is the central order statistic if n is uneven, and the center of the two central order statistics if n is even.

Proposition 3.0.6 *In the double exponential location model $\boldsymbol{M}_4$, the sample median* $\mathrm{med}(X)$ *is a maximum likelihood estimator of the location parameter $\mu \in \Theta = \mathbb{R}$: $T_{ML}(X) = \mathrm{med}(X)$.*

Note that uniqueness is not claimed here.
Proof. The likelihood function is

$$\begin{aligned} L_x(\mu) &= \prod_{i=1}^{n} (2)^{-1} \exp(-|x_i - \mu|) \\ &= (2)^{-n} \exp\left(-\sum_{i=1}^{n} |x_i - \mu|\right) \end{aligned}$$

and maximizing it is equivalent to minimizing $-\log L_x(\mu)$, or minimizing

$$\tilde{l}_x(\mu) = \sum_{i=1}^{n} |x_i - \mu| .$$

For given x this is a piecewise linear continuous function in μ; for $\mu > x_{(n)}$ it is a linear function in μ tending to ∞ for $\mu \to \infty$, and for $\mu > x_{(1)}$ it is also a linear function in μ tending to ∞ for $\mu \to -\infty$. Thus the minimum must be attained in the range of the sample, i. e. in the interval $[x_{(1)}, x_{(n)}]$. Assume that $x_{(1)} < x_{(1)} < \ldots x_{(n)}$, i.e. no two values of the sample are equal (there are no ties). That event has probability one since the X_i are continuous random variables. Inside an interval $(x_{(i)}, x_{(i+1)})$, $1 \leq i \leq n-1$ the derivative of $\tilde{l}_x(\mu)$ is

$$\tilde{l}'_x(\mu) = \sum_{j=1}^{i} 1 + \sum_{j=i+1}^{n} (-1) = i - (n - i) = 2i - n.$$

Consider first the case of even n. Then $\tilde{l}'_x(\mu)$ is negative for $i < n/2$, positive for $i > n/2$ and 0 for $i = n/2$. Thus the minimum in μ is attained by any value $\mu \in [x_{(n/2)}, x_{(n/2+1)}]$, in particular by the center of that interval which is the sample median. If n is uneven then $\tilde{l}'_x(\mu)$ is negative for $i < n/2$ and positive for $i > n/2$. Since the function $\tilde{l}_x(\mu)$ is continuous, the minimum is attained in the beginning point of the first interval where $\tilde{l}'_x(\mu)$ is positive, which is $X_{((n+1)/2)}$. ■

Since the sample median minimizes $\sum_{i=1}^{n} |X_i - \mu|$ it may be called a **least absolute deviation estimator** . In contrast the sample minimizes the sum of squares $\sum_{i=1}^{n} (X_i - \mu)^2$, i.e. it is a **least squares estimator** . Note that the sample mean remains the same when $X_{(1)} \to \infty$ whereas $\bar{X}_n$ also tends to infinity in that case. By that reason, the sample median is often applied independently of a maximum likelihood justification, which holds when the data are double exponential.
In analogy to the sample median, the population median is defined as as a "half point" of the distribution of a random variable X representing the population. A value m is a median of a r.v. X if simultaneously $P(X \geq m) \geq 1/2$ and $P(X \leq m) \geq 1/2$ hold. For continuous distributions, we always have $P(X = m) = 0$, and therefore m is a solution of $P(X > m) = 1 - P(X < m) = 1/2$ which may not be unique.

Chapter 4
UNBIASED ESTIMATORS

Consider again the general parametric estimation problem**:** let X be a random variable with values in $\mathcal{X}$ and $\mathcal{L}(X)$ be the distribution (or the law) of X. Assume that $\mathcal{L}(X)$ is known up to a ϑ from a parameter space $\Theta \subseteq \mathbb{R}^k$:

$$\mathcal{L}(X) \in \{P_\vartheta; \vartheta \in \Theta\}.$$

The problem is to estimate a real valued function $g(\vartheta)$ based on a realization of X. Up to now we primarily considered the case where Θ is an interval in $\mathbb{R}$ and $g(\vartheta) = \vartheta$ (i.e. we are interested in estimation of ϑ itself), but the Gaussian location-scale model $\mathbf{M}_3$ was an instance of a model with a two-dimensional parameter $\vartheta = (\mu, \sigma^2)$. Here we might set $g(\vartheta) = \mu$.
Consider an estimator T of $g(\vartheta)$ such that $E_\vartheta T$ exists. In our i.i.d. Bernoulli model $\mathbf{M}_1$, that is true for every estimator.

Definition 4.0.7 *(i) The quantity*

$$E_\vartheta T - g(\vartheta)$$

is called the ***bias*** *of the estimator T.*
(ii) If

$$E_\vartheta T = g(\vartheta) \text{ for all } \vartheta \in \Theta$$

then the estimator T is called ***unbiased*** *for $g(\vartheta)$.*

In model $\mathbf{M}_1$ we have

$$E_p \bar{X}_n = p \text{ for all } p \in [0,1],$$

i.e. the sample mean is an unbiased estimator for the parameter p. Similarly, in the Gaussian location and location-scale models, the sample mean is unbiased for $\mu = X_1$.
Unbiasedness is sometimes considered as a value in itself, i.e. a desirable property for an estimator, independently of the risk optimality. Recall that the risk function (for quadratic loss) for estimation of a real valued $\vartheta \in \Theta$ was defined as

$$R(T, \vartheta) = E_\vartheta \left(T(X) - \vartheta\right)^2.$$

Suppose T is an estimator with $R(T, \vartheta) < \infty$. Then

$$\begin{aligned} R(T,\vartheta) &= E_\vartheta \left(T(X) - E_\vartheta T + E_\vartheta T - \vartheta)\right)^2 \qquad (4.1) \\ &= E_\vartheta \left(T(X) - E_\vartheta T\right)^2 + \left(E_\vartheta T - \vartheta\right)^2 \\ &= \mathrm{Var}_\vartheta T(X) + \left(E_\vartheta T - \vartheta\right)^2. \qquad (4.2) \end{aligned}$$

The last line is called the **bias-variance decomposition** of the quadratic risk of T; it holds for any T with finite risk at ϑ (or equivalently with $\mathrm{Var}_\vartheta T(X) < \infty$); T need not be unbiased. If T

is unbiased then the second term in (4.2), i.e. the squared bias vanishes; thus for unbiased T the quadratic risk is the variance.
We saw that in model $\mathbf{M}_1$ the estimator

$$T_M = \frac{\bar{X}_n + n^{-1/2}/2}{1 + n^{-1/2}}$$

is minimax with respect to quadratic risk $R(T, \vartheta)$, and it is clearly biased. The unbiased estimator $\bar{X}_n$ performs less well in the minimax sense. It is thus a matter of choice how to judge the performance: one may strictly adhere to the quadratic risk as a criterion, leaving the bias problem aside, or impose unbiasedness as an a priori requirement for "good" estimators.
If the latter point of view is taken, within the restrict the class of unbiased estimators it is often possible to find *uniformly best elements* (optimal for all values of ϑ), as we shall see now.

4.1 The Cramer-Rao information bound

An *information bound* , relative to a statistical decision problem, is a bound for the best possible performance of a decision procedure, either within all possible procedures or within a certain restricted class of methods. We shall discuss a bound which relates to the class of unbiased estimators.
Let us formalize a set of assumptions which was already made several times.

Model $\mathbf{M}_f$ The sample space $\mathcal{X}$ for the observed random variable X is finite, and $\mathcal{L}(X) \in \{P_\vartheta, \vartheta \in \Theta\}$, where Θ is an open (possibly infinite) interval in $\mathbb{R}$.

The case $\Theta = \mathbb{R}$ is included. Note that model $\mathbf{M}_1$ is a special case, if the open interval $\Theta = (0, 1)$ is taken as parameter space. With this model we associate a family of probability functions $p_\vartheta(x)$, $x \in \mathcal{X}, \vartheta \in \Theta$. Let T be an unbiased estimator of the parameter:

$$E_\vartheta T(X) = \vartheta, \tag{4.3}$$

Since $\mathcal{X}$ is finite, the expectation always exists for all $\vartheta \in \Theta$. Let us add a differentiability assumption on the dependence of $p_\vartheta(x)$ on ϑ.

Assumption D_1 For every $x \in \mathcal{X}$, the probability function $p_\vartheta(x)$ is differentiable in ϑ, in every point $\vartheta \in \Theta$.

Now (4.3) can be written

$$\sum_{x \in \mathcal{X}} T(x) p_\vartheta(x) = \vartheta$$

and both sides are differentiable in ϑ. Taking the derivative we obtain

$$\sum_{x \in \mathcal{X}} T(x) p'_\vartheta(x) = 1. \tag{4.4}$$

where p'_ϑ is the derivative wrt to ϑ. The first derivative of a probability function has an interesting property. We can also differentiate the equality

$$\sum_{x \in \mathcal{X}} p_\vartheta(x) = 1$$

valid for any probability function $p_\vartheta(x)$. Differentiation gives

$$\sum_{x\in\mathcal{X}} p'_\vartheta(x) = 0. \tag{4.5}$$

This means that in (4.4) we can replace $T(x)$ by $T(x)+c$ where c is any constant; indeed (4.5) implies that the right side of (4.4) is still 1. Choosing $c=-\vartheta$, we obtain

$$\sum_{x\in\mathcal{X}} (T(x)-\vartheta)\, p'_\vartheta(x) = 1. \tag{4.6}$$

Define now the **score function**

$$l_\vartheta(x) = \begin{cases} p'_\vartheta(x)/p_\vartheta(x) \text{ if } p_\vartheta(x) > 0 \\ 0 \text{ if } p_\vartheta(x) = 0. \end{cases}$$

Note that if $p_\vartheta(x) = 0$ at $\vartheta = \vartheta_0 \in \Theta$ then $p'_{\vartheta_0}(x)$ must also be 0: since $p_\vartheta(x)$ is nonnegative and differentiable, it has a local minimum at ϑ_0 and hence $p'_{\vartheta_0}(x) = 0$. This fact and (4.6) imply

$$1 = \sum_{x\in\mathcal{X}} (T(x)-\vartheta)\, l_\vartheta(x) p_\vartheta(x) = E_\vartheta\,(T(X)-\vartheta)\, l_\vartheta(X) \tag{4.7}$$

Let us apply now the Cauchy-Schwarz inequality : for any two random variables Z_1, Z_2 on a common probability space

$$|EZ_1Z_2|^2 \le \left(EZ_1^2\right)\left(Z_2^2\right)$$

provided the left side exists, i.e. $EZ_i^2 < \infty$, $i = 1, 2$. Squaring both sides of (4.7), we obtain

$$1 \le E_\vartheta\,(T(X)-\vartheta)^2 \cdot E_\vartheta l_\vartheta^2(X).$$

The quantity

$$I_F(\vartheta) = E_\vartheta l_\vartheta^2(X) \tag{4.8}$$

is called the **Fisher information** of the parametric family $p_\vartheta, \vartheta \in \Theta$ of probability functions, at point $\vartheta \in \Theta$.

Theorem 4.1.1 ***(Cramer-Rao bound)*** *In model* $\mathbf{M}_f$*, under smoothness assumption* $\mathbf{D}_1$*, assume that* $I_F(\vartheta) > 0$*,* $\vartheta \in \Theta$*. Then for every unbiased estimator* T *of* ϑ

$$\mathrm{Var}_\vartheta T(X) \ge (I_F(\vartheta))^{-1},\ \vartheta \in \Theta \tag{4.9}$$

where $I_F(\vartheta)$ *is the Fisher information (4.8).*

Note that the score function $l_\vartheta(x)$ can also be written

$$l_\vartheta(x) = \begin{cases} \frac{d}{d\vartheta}\log p_\vartheta(x) \text{ if } p_\vartheta(x) > 0 \\ 0 \text{ if } p_\vartheta(x) = 0 \end{cases}$$

so that the Fisher information is

$$\begin{aligned} I_F(\vartheta) &= E_\vartheta l_\vartheta^2(X) = E_\vartheta \left(\frac{d}{d\vartheta}\log p_\vartheta(X)\right)^2 \\ &= \sum_{x:p_\vartheta(x)>0} \left(p'_\vartheta(x)\right)^2 / p_\vartheta(x). \end{aligned} \tag{4.10}$$

The form (4.10) involving the logarithm of p_ϑ is convenient in many cases for computation of the Fisher information.
Note also that for the score function we have as a consequence of (4.5)

$$E_\vartheta l_\vartheta(X) = 0. \tag{4.11}$$

The Cramer-Rao bound (4.9) gives a benchmark against which to measure the performance of any unbiased estimator. An unbiased estimator attaining the Cramer-Rao bound at ϑ is called a **best unbiased estimator** (or *uniformly best* if that is true for all $\vartheta \in \Theta$; "uniformly" is usually omitted). Another terminology is **uniformly minimum variance unbiased estimator (UMVUE).**

Example 4.1.2 Consider the case of Model $\mathbf{M}_1$ for sample size $n = 1$, i.e. we observe one Bernoulli r.v. with law $B(1, p)$, $p \in (0, 1)$. The probability function is, for $x \in \{0, 1\}$

$$q_p(x) = (1-p)^{1-x}p^x.$$

For the score function we obtain

$$\begin{aligned} l_p(x) &= \frac{d}{dp}\log(1-p)^{1-x}p^x = (1-x)\frac{d}{dp}\log(1-p) + x\frac{d}{dp}\log p \\ &= -(1-x)\frac{1}{(1-p)} + x\frac{1}{p} \\ &= \frac{x(1-p)-(1-x)p}{p(1-p)} = \frac{x-p}{p(1-p)}. \end{aligned}$$

We check again (4.11) in this case from $E_p(X - p) = 0$. The Fisher information is thus

$$I_F(p) = E_p l_p^2(X) = (p(1-p))^{-2}\mathrm{Var}_p X = (p(1-p))^{-1}.$$

Thus $p(1-p)$ is the Cramer-Rao bound. It follows that X is a best unbiased estimator for all $p \in (0, 1)$, *i.e. uniform minimum variance unbiased.* □

The next theorem specializes the situation to the case of independent and identically distributed data.

Theorem 4.1.3 *Suppose that $X = (X_1, \ldots, X_n)$ are independent and identically distributed where the statistical model for X_1 is a model of type $\mathbf{M}_f$, with probability function q_ϑ, satisfying assumption $\mathbf{D}_1$. Assume that the Fisher information for q_ϑ satisfies $I_F(\vartheta) > 0$, $\vartheta \in \Theta$. Then for every unbiased estimator T of ϑ*

$$\mathrm{Var}_\vartheta T(X) \geq n^{-1}(I_F(\vartheta))^{-1},\ \vartheta \in \Theta. \tag{4.12}$$

where $I_F(\vartheta)$ is the Fisher information for q_ϑ.

Proof. The model for the vector $X = (X_1, \ldots, X_n)$ is also of type $\mathbf{M}_f$, and the probability function is $p_\vartheta(x) = \prod_{i=1}^n q_\vartheta(x_i)$. It suffices to find the Fisher information of p_ϑ, call it $I_{F,n}(\vartheta)$. We have

$$I_{F,n}(\vartheta) = E_\vartheta\left(\frac{d}{d\vartheta}\log p_\vartheta(X)\right)^2 = E_\vartheta\left(\sum_{i=1}^n \frac{d}{d\vartheta}\log q_\vartheta(X_i)\right)^2.$$

Expanding the square, we note for the mixed terms ($i \neq j$), when l_ϑ is the score function for q_ϑ

$$E_\vartheta l_\vartheta(X_i) l_\vartheta(X_j) = E_\vartheta l_\vartheta(X_i) \cdot E_\vartheta l_\vartheta(X_j) = 0$$

in view of independence of X_i and X_j and (4.11). Hence

$$I_{F,n}(\vartheta) = \sum_{i=1}^{n} E_\vartheta l_\vartheta^2(X_i) = n \cdot I_F(\vartheta). \tag{4.13}$$

Now Theorem 4.1.1 is valid for the whole model for $X = (X_1, \ldots, X_n)$ and implies (4.12). ■

Interpreting the Fisher information. To draw an analogy to concepts from physics or mechanics, let t be a time parameter and and suppose $x(t) = (x_i(t))_{i=1,2,3}$ describes a movement in three dimensional space $\mathbb{R}^3$ (a path in space, as a function of time). The velocity $v(t_0)$ at time t_0 is the length of the gradient, i.e.

$$v(t_0) = \left\|\mathbf{x}'(t)\right\| = \left(\sum_{i=1}^{3} (x_i'(t))^2\right)^{1/2}. \tag{4.14}$$

If the sample space $\mathcal{X}$ has k elements and all $p_\vartheta(x_i) > 0$

$$I_F(\vartheta) = \sum_{i=1}^{k} \frac{\left(\frac{d}{d\vartheta} p_\vartheta(x_i)\right)^2}{p_\vartheta(x_i)} \tag{4.15}$$

$$= 4 \sum_{i=1}^{k} \left(\frac{d}{d\vartheta} p_\vartheta^{1/2}(x_i)\right)^2. \tag{4.16}$$

Thus $I_F(\vartheta)$ is similar to a squared velocity if we identify ϑ with time, i.e. we consider the "curve" described in $\mathbb{R}^k$ by the vector $\left(p_\vartheta^{1/2}(x_1), \ldots, p_\vartheta^{1/2}(x_k)\right)$ when ϑ varies in an interval Θ. Note that the vector $\mathbf{q}_\vartheta := \left(p_\vartheta^{1/2}(x_1), \ldots, p_\vartheta^{1/2}(x_k)\right)$ has length one ($\|\mathbf{q}_\vartheta\|^2 = 1$) since $p_\vartheta(x_i)$ sum up to one. Thus the family $(\mathbf{q}_\vartheta, \vartheta \in \Theta)$ describes a parametric curve on the surface of the unit sphere in $\mathbb{R}^k$. The Fisher information can thus be interpreted as a "sensitivity" of the location $\mathbf{q}_\vartheta \in \mathbb{R}^k$ with respect to small changes in the parameter, at position ϑ.
Note that in Theorem (4.1.1) we assumed that $I_F(\vartheta) > 0$. Consider the case where $p_\vartheta(x)$ does not depend on ϑ; formally then $p_\vartheta(\cdot)$ is differentiable, and $l_\vartheta^2 = 0$, $I_F(\vartheta) = 0$. Then Theorem (4.1.1) has to be interpreted as meaning that unbiased estimators T do not exist (indeed there is insufficient information in the family to enable $E_\vartheta T = \vartheta$; the left side does not depend on ϑ). The relation $I_F(\vartheta) = 0$ may also occur in other families or at particular points ϑ; then unbiased estimators do not exist.

4.2 Countably infinite sample space

Let us now consider the case where the sample space $\mathcal{X}$ is still discrete but only countable (not necessarily finite), e.g. the case where it consists of the nonnegative integers $\mathcal{X} = \mathbb{Z}_+$.

Model $\mathbf{M}_d$ The sample space $\mathcal{X}$ for the observed random variable X is countable, and $\mathcal{L}(X) \in \{P_\vartheta, \vartheta \in \Theta\}$, where Θ is an open (possibly infinite) interval in $\mathbb{R}$.

The proof of the Cramer-Rao bound is very similar here, only we need to be sure that diffentiation is possible under the infinite sum sign. If $\mathcal{X} = \{x_1, x_2, \ldots\}$ then we need to differentiate both sides of the unbiasedness equation

$$\sum_{i=1}^{\infty} T(x_i) p_\vartheta(x_i) = \vartheta \tag{4.17}$$

If ϑ and $\vartheta' = \vartheta + \Delta$ are two close values then we would take the ratio of differences on both sides

$$\sum_{i=1}^{\infty} T(x_i) \left(p_\vartheta(x_i) - p_{\vartheta+\Delta}(x_i)\right) \Delta^{-1} = 1 \tag{4.18}$$

and compute the derivative by letting $\Delta \to 0$. Both sides are differentiable in ϑ. Here for every x_i

$$\left(p_\vartheta(x_i) - p_{\vartheta+\Delta}(x_i)\right) \Delta^{-1} \to p'_\vartheta(x_i).$$

hence if $p_\vartheta(x_i) > 0$ then

$$\frac{p_\vartheta(x_i) - p_{\vartheta+\Delta}(x_i)}{\Delta p_\vartheta(x_i)} \to \frac{p'_\vartheta(x_i)}{p_\vartheta(x_i)} = l_\vartheta(x_i) \text{ as } \Delta \to 0.$$

Condition $\mathbf{D}_2$ (i) *The probability function $p_\vartheta(x)$ is positive and differentiable in ϑ, at every $x \in \mathcal{X}$ and every $\vartheta \in \Theta$*

(ii) *For every $\vartheta \in \Theta$, there exist an $\varepsilon = \varepsilon_\vartheta > 0$ and a function $b_\vartheta(x)$ satisfying $E_\vartheta b_\vartheta^2(X) < \infty$ and*

$$\left|\frac{p_\vartheta(x) - p_{\vartheta+\Delta}(x)}{\Delta p_\vartheta(x)}\right| \le b_\vartheta(x) \text{ for all } |\Delta| \le \varepsilon \text{ and all } x \in \mathcal{X}.$$

Using that condition, we find for the right hand side of (4.18)

$$\begin{aligned}\sum_{i=1}^{\infty} T(x_i) \left(p_\vartheta(x_i) - p_{\vartheta+\Delta}(x_i)\right) \Delta^{-1} &= \sum_{i=1}^{\infty} \left(T(x_i) \frac{p_\vartheta(x_i) - p_{\vartheta+\Delta}(x_i)}{\Delta p_\vartheta(x_i)}\right) p_\vartheta(x_i) \\ &= \sum_{i=1}^{\infty} r_\Delta(x_i) p_\vartheta(x_i).\end{aligned}$$

The sequence of functions for $\Delta \to 0$

$$r_\Delta(x) = T(x_i) \frac{p_\vartheta(x) - p_{\vartheta+\Delta}(x)}{\Delta p_\vartheta(x)}$$

converges to a limit function $r_0(x) = T(x) l_\vartheta(x)$ pointwise (i.e. for every $x \in \mathcal{X}$). We would like to show

$$\sum_{i=1}^{\infty} r_\Delta(x_i) p_\vartheta(x_i) \to \sum_{i=1}^{\infty} r_0(x_i) p_\vartheta(x_i) = E_\vartheta T(X) l_\vartheta(X) \tag{4.19}$$

since in that case we could infer that $E_\vartheta T(X) l_\vartheta(X) = 1$ (differentiate both sides of (4.17) under the integral sign). By a result from real analysis, the Lebesgue dominated convergence theorem (see Appendix) it suffices to show that there is a function $\tilde{r}(x) \ge 0$ and a $\delta > 0$ such that

$$|r_\Delta(x)| \le \tilde{r}(x) \text{ for all } |\Delta| \le \delta, \text{ and } E_\vartheta \tilde{r}(X) < \infty.$$

To establish that, we use

$$|r_\Delta(x)| = \left| T(x_i) \frac{p_\vartheta(x) - p_{\vartheta+\Delta}(x)}{\Delta p_\vartheta(x)} \right| \le |T(x_i) b_\vartheta(x)| =: \tilde{r}(x)$$

and the Cauchy-Schwarz inequality

$$E_\vartheta \tilde{r}(X) \le \left(E_\vartheta T^2(X)\right)^{1/2} \left(E_\vartheta b_\vartheta^2(X)\right)^{1/2} < \infty$$

according to condition $\mathbf{D}_2$ and the finiteness of $E_\vartheta T^2(X)$, which is natural to assume for the estimator. $T(x)$. Thus we have established

$$E_\vartheta T(X) l_\vartheta(X) = 1.$$

. The same technique allows us to differentiate the relation

$$\sum_{i=1}^{\infty} p_\vartheta(x_i) = 1$$

under the series sign, i.e. to obtain

$$\sum_{i=1}^{\infty} p'_\vartheta(x_i) = E_\vartheta l_\vartheta(X) = 0.$$

Theorem 4.2.1 ***(Cramer-Rao bound, discrete case)*** *In model* $\mathbf{M}_d$*, assume that the Fisher information exists and is positive:*

$$0 < I_F(\vartheta) = E_\vartheta \left(\frac{p'_\vartheta(X)}{p_\vartheta(X)} \right)^2 < \infty.$$

for all $\vartheta \in \Theta$*, and also that condition* $\mathbf{D}_2$ *holds. Then for for every unbiased estimator* T *of* ϑ *with finite variance*

$$\mathrm{Var}_\vartheta T(X) \ge (I_F(\vartheta))^{-1},\ \vartheta \in \Theta. \tag{4.20}$$

Remark 4.2.2 Clearly an analog of Theorem 4.1.3 holds, we do not state it explicitly. It is evident that (4.13) still holds. One would have to verify that it suffices to impose Condition $\mathbf{D}_2$ only on the law of X_1, but we omit this argument.

Example. Let X have Poisson law $\mathrm{Po}(\vartheta)$, where $\vartheta > 0$ is unknown. Let us check condition $\mathbf{D}_2$ (here $\Theta = (0, \infty)$). We have for $x_k = k$, $k = 0, 1, \ldots$

$$p_\vartheta(x_k) = \frac{\vartheta^k}{k!} \exp(-\vartheta)$$

This is continuously differentiable in ϑ, for every $k \ge 0$, and

$$p'_\vartheta(x_k) = \frac{1}{k!} \left(k\vartheta^{k-1} - \vartheta^k \right) \exp(-\vartheta) = \left(k\vartheta^{-1} - 1 \right) \frac{\vartheta^k}{k!} \exp(-\vartheta).$$

By the mean value theorem, for every Δ and every $k \geq 0$ there exists $\Delta^*(k)$ (lying in the interval between 0 and Δ) such that

$$p_{\vartheta+\Delta}(k) - p_\vartheta(k) = \Delta \cdot p'_{\vartheta+\Delta^*(k)}(k).$$

so that for $|\Delta| \leq \varepsilon$, $\varepsilon > 0$ sufficiently small

$$\begin{aligned}
\left|\frac{p_\vartheta(k) - p_{\vartheta+\Delta}(k)}{\Delta p_\vartheta(k)}\right| &= \left|\frac{p'_{\vartheta+\Delta^*(k)}(k)}{p_\vartheta(k)}\right| \\
&= \left|\frac{k}{\vartheta+\Delta^*(k)} - 1\right| \left|\frac{\vartheta+\Delta^*(k)}{\vartheta}\right|^k \exp(-\Delta^*(k)) \\
&\leq \left(\frac{k}{\vartheta-\varepsilon} + 1\right)\left(1 + \frac{\varepsilon}{\vartheta}\right)^k \exp\varepsilon =: b_\vartheta(k)
\end{aligned}$$

Let us denote by C_0, C_1, C_2 etc. constants which do not depend on k (but may depend on ε and ϑ). We have

$$\frac{k}{\vartheta-\varepsilon} + 1 \leq C_0\,(k+1) \leq C_1 k,$$

$$b^2_\vartheta(k) \leq C_1 k^2 C_2^{2k} \leq C_1 \exp(2k) \cdot C_2^{2k} = C_3 \exp(C_4 k).$$

Thus for $E_\vartheta b^2_\vartheta(X) < \infty$ it suffices to show that the Poisson law has *finite exponential moments of all order*, i.e. for all $c > 0$ and all $\vartheta > 0$

$$E_\vartheta \exp(cX) < \infty.$$

To see this, note

$$\begin{aligned}
E_\vartheta \exp(cX) &= \sum_{k=1}^{\infty} \exp(ck) \frac{\vartheta^k}{k!} \exp(-\vartheta) = \sum_{k=1}^{\infty} \frac{(\vartheta \exp c)^k}{k!} \exp(-\vartheta) \\
&= \exp(\vartheta \exp c - \vartheta).
\end{aligned}$$

□

4.3 The continuous case

Model $\mathbf{M}_c$ The observed random variable $X = (X_1, \ldots, X_k)$ is continuous with values in $\mathbb{R}^k$ and $\mathcal{L}(X) \in \{P_\vartheta, \vartheta \in \Theta\}$. Each law P_ϑ is described by a joint density $p_\vartheta(x) = p_\vartheta(x_1, \ldots, x_k)$, and Θ is an open subset of $\mathbb{R}^d$.

The essential work for the Cramer-Rao bound was already done in the previous subsection; indeed this time we need *differentiation under the integral sign* which is analogous to the infinite series case. The reasoning is very similar. As estimator T is called unbiased if $E_\vartheta |T| < \infty$ and $E_\vartheta T = \vartheta$ for all $\vartheta \in \Theta$. We again start with the unbiasedness relation

$$\int T(x) p_\vartheta(x) dx = \vartheta$$

and we need to differentiate the left side under the integral sign. Let us assume that our parameter ϑ is one dimensional, as in the previous subsections; multivariate versions of the Cramer-Rao bound can be derived but need not interest us here.

Condition $\mathbf{D}_3$ (i) *The parameter space Θ is an open (possibly infinite) interval in $\mathbb{R}$. There is a subset $\mathcal{S} \subseteq \mathbb{R}^k$ such that for all $\vartheta \in \Theta$, we have $p_\vartheta(x) > 0$ for $x \in \mathcal{S}$, $p_\vartheta(x) = 0$ for $x \notin \mathcal{S}$.*
(ii) *The density $p_\vartheta(x)$ is positive and differentiable in ϑ, at every $x \in \mathcal{S}$ and every $\vartheta \in \Theta$*
(iii) *For every $\vartheta \in \Theta$, there exist an $\varepsilon = \varepsilon_\vartheta > 0$ and a function $b_\vartheta(x)$ satisfying $E_\vartheta b_\vartheta^2(X) < \infty$ and*

$$\left| \frac{p_\vartheta(x) - p_{\vartheta+\Delta}(x)}{\Delta p_\vartheta(x)} \right| \leq b_\vartheta(x) \text{ for all } |\Delta| \leq \varepsilon \text{ and all } x \in \mathcal{X}.$$

We can define a score function $l_\vartheta(x)$

$$l_\vartheta(x) = \begin{cases} \frac{d}{d\vartheta} \log p_\vartheta(x) \text{ if } p_\vartheta(x) > 0 \\ 0 \text{ if } p_\vartheta(x) = 0 \end{cases}$$

and the Fisher information

$$I_F(\vartheta) = E_\vartheta l_\vartheta^2(X) = \int_S \left(p'_\vartheta(x)\right)^2 / p_\vartheta(x) dx.$$

Theorem 4.3.1 ***(Cramer-Rao bound, continuous case)*** *In Model $\mathbf{M}_c$, assume that that the smoothness condition $\mathbf{D}_3$ holds, and that the Fisher information exists and is positive: $0 < I_F(\vartheta) < \infty$ for all $\vartheta \in \Theta$. Then for for every unbiased estimator T of ϑ with finite variance*

$$\mathrm{Var}_\vartheta T(X) \geq (I_F(\vartheta))^{-1}, \vartheta \in \Theta. \tag{4.21}$$

The proof is exactly as in the preceding countably infinite (discrete case), but the infinite series $\sum_{i=1}^{\infty} T(x_j) p_\vartheta(x_j)$ has to be substituted by an integral $\int T(x) p_\vartheta(x) dx$. Remark 4.2.2 on the i.i.d. case can be made here analogously.

Example 4.3.2 Consider the Gaussian location model $\mathbf{M}_2$ for sample size $n = 1$, i.e. we observe a Gaussian r.v. with law $N(\vartheta, \sigma^2)$ where $\sigma^2 > 0$ is known, $\Theta = \mathbb{R}$. Let us verify condition $\mathbf{D}_3$. For ease of notation assume $\sigma^2 = 1$. If φ is the standard Gaussian density then $p_\vartheta(x) = \varphi(x - \vartheta)$, and

by a transformation $y = x - \vartheta$ it suffices to show the condition at parameter point $\vartheta = 0$. Now the density

$$p_\vartheta(x) = \varphi(x - \vartheta)$$

is continuously differentiable in ϑ *with derivative*

$$p'_\vartheta(x) = -\varphi'(x - \vartheta)$$

Since

$$\varphi(x) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{x^2}{2}\right)$$

we have

$$\varphi'(x) = -x\varphi(x).$$

Now for $|\Delta| \le \varepsilon$

$$|\varphi(x - \Delta) - \varphi(x)| = \left| \int_{-\Delta}^{0} \varphi(x + u) dy \right| \le \Delta \sup_{|u| \le \varepsilon} \left|\varphi'(x - u)\right|.$$

Also note that

$$\begin{aligned} \sup_{|u| \le \varepsilon} \left|\varphi'(x - u)\right| &\le (|x| + \varepsilon) \sup_{|u| \le \varepsilon} \varphi(x - u), \\ \varphi(x - u) &= (2\pi)^{-1/2} \exp\left(-\frac{(x - u)^2}{2}\right) \\ &= \varphi(x) \exp\left(xu - u^2/2\right) \end{aligned}$$

so that for $|\Delta| \le \varepsilon$, $\varepsilon > 0$ *sufficiently small and* $\vartheta = 0$

$$\begin{aligned} \left| \frac{p_\vartheta(x) - p_{\vartheta+\Delta}(x)}{\Delta p_\vartheta(x)} \right| &\le \left| \frac{\sup_{|u| \le \varepsilon} |\varphi'(x + u)|}{\varphi(x)} \right| \\ &\le (|x| + \varepsilon) \left| \sup_{|u| \le \varepsilon} \exp\left(xu - u^2/2\right) \right| \\ &\le (|x| + \varepsilon) \exp\left(|x|\varepsilon\right) \end{aligned}$$

Since $|x| \le C_1 \exp(|x|\varepsilon)$ *for an appropriate constant* C_1 *(depending on* ε*), we have*

$$\left| \frac{p_\vartheta(x) - p_{\vartheta+\Delta}(x)}{\Delta p_\vartheta(x)} \right| \le C_2 =: b_\vartheta(x).$$

Now we need to show that

$$E_\vartheta b_\vartheta^2(x) = C_2^2 E_0 \exp\left(|X|2\varepsilon\right) < \infty,$$

The right side clearly has a finite expectation under $\mathcal{L}(X) = N(0, 1)$. Indeed for all $t > 0$ *we have*

$$\begin{aligned} & \frac{1}{(2\pi)^{1/2}} \int \exp(t|x|) \exp\left(-\frac{x^2}{2}\right) dx \\ \le\ & 2\frac{1}{(2\pi)^{1/2}} \int \exp(tx) \exp\left(-\frac{x^2}{2}\right) dx \\ =\ & 2\frac{1}{(2\pi)^{1/2}} \int \exp\left(-\frac{(x - t)^2}{2}\right) dx \cdot \exp(\frac{t^2}{2}) = 2\exp(\frac{t^2}{2}). \end{aligned}$$

. □

Proposition 4.3.3 *In the Gaussian location model* $\mathbf{M}_2$*, the Fisher information w.r.t. the expectation parameter* $\mu \in \mathbb{R}$ *is* n/σ^2*, and the sample mean* $\bar{X}_n$ *is is a UMVUE of* μ*.*

Proof. We have

$$Var_\mu \bar{X}_n = \sigma^2/n$$

so we need only find the Fisher information. Condition $\mathbf{D}_3$ is easily verified, analogously to the case $n = 1$: in (3.4) we found an expression for the joint density

$$\prod_{i=1}^{n} p_\mu(x_i) = \frac{1}{\sigma n^{-1/2}} \varphi\left(\frac{\bar{x}_n - \mu}{\sigma n^{-1/2}}\right) \cdot \frac{1}{n^{1/2}(2\pi\sigma^2)^{(n-1)/2}} \exp\left(-\frac{s_n^2}{2\sigma^2 n^{-1}}\right). \tag{4.22}$$

where $s_n^2 = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$ is the sample variance. The second factor does not depend on μ (σ^2 is fixed), and therefore cancels in the term $(p_\mu(x) - p_{\mu+\Delta}(x))/p_\mu(x)$ in condition $\mathbf{D}_3$. The first factor is the density (as a function of $\bar{x}_n$) of the law $N(\mu, n^{-1}\sigma^2)$; thus reasoning as in the above example (4.3.2), we finally have to show that

$$E_0 \exp\left(2|\bar{x}_n| n\sigma^{-2}\varepsilon\right) < \infty$$

which follows as above. To find the Fisher information, we can use the factorization (4.22) again. Indeed

$$\begin{aligned} I_{F,n}(\mu) &= E_\mu \left(\frac{d}{d\mu} \log \prod_{i=1}^{n} p_\mu(x_i)\right)^2 \\ &= E_\mu \left(\frac{d}{d\mu} \log \frac{1}{\sigma n^{-1/2}} \varphi\left(\frac{\bar{x}_n - \mu}{\sigma n^{-1/2}}\right)\right)^2 \end{aligned}$$

so here the Fisher information is the same as if we observed only $\bar{x}_n$ with law $N(\mu, n^{-1}\sigma^2)$, i.e. in a Gaussian location model with variance $\gamma^2 = n^{-1}\sigma^2$. The score function in such a model is

$$\begin{aligned} l_\mu(x) &= \frac{d}{d\mu} \log \varphi\left(\frac{x-\mu}{\sigma n^{-1/2}}\right) \\ &= \frac{d}{d\mu}\left(-\frac{(x-\mu)^2}{2\gamma^2}\right) = \gamma^{-2}(x-\mu), \end{aligned}$$

and

$$I_F(\mu) = \gamma^{-4} E(x-\mu)^2 = \gamma^{-2} = n\sigma^{-2}.$$

■

We will now discuss an example where the Cramer-Rao bound does not hold; cf. Casella and Berger [CB], p. 312, Example 7.3.5. Suppose that $X_1, \ldots, X_n$ are i.i.d with uniform density on $[0, \vartheta]$. That means the density is

$$p_\vartheta(x) = \frac{1}{\vartheta},\ 0 \le x \le \vartheta.$$

We might try to formally apply the Cramer-Rao bound, neglecting the differentiability condition for a moment. For one observation, the score function is

$$l_\vartheta(x) = \frac{\partial}{\partial\vartheta} \log p_\vartheta(x) = \begin{cases} -\frac{1}{\vartheta},\ 0 < x < \vartheta \\ 0,\ x > \vartheta \end{cases},$$

hence (formally)

$$I_F(\vartheta) = E_\vartheta l_\vartheta^2(x) = \frac{1}{\vartheta^2}$$

which suggests that for any unbiased estimator based on n observations

$$\mathrm{Var}_\vartheta T(X) \geq \frac{\vartheta^2}{n}. \tag{4.23}$$

However an estimator can be found which is better. Consider the maximum, i.e the largest order statistic

$$X_{[n]} = \max_{i=1,\ldots,n} X_i$$

. The density of $X_{[n]}$ (q_ϑ, say) can be found as follows: for $t \leq \vartheta$

$$\begin{aligned} P_\vartheta\left(X_{[n]} \leq t\right) &= P_\vartheta\left(X_1 \leq t, \ldots, X_n \leq t\right) \\ &= \prod_{i=1}^{n} P_\vartheta(X_i \leq t) = (t/\vartheta)^n, \end{aligned}$$

and $P_\vartheta\left(X_{[n]} \leq t\right) = 1$ for $t > \vartheta$, hence

$$q_\vartheta(t) = \frac{d}{dt}(t/\vartheta)^n = nt^{n-1}/\vartheta^n,\ 0 \leq t \leq \vartheta$$

and $q_\vartheta(t) = 0$ for $t > \vartheta$. For the expectation of $X_{[n]}$ we get

$$\begin{aligned} E_\vartheta X_{[n]} &= \vartheta^{-n} n \int_0^\vartheta y \cdot y^{n-1} dy = \vartheta^{-n} n \left[\frac{1}{n+1} y^{n+1}\right]_0^\vartheta \\ &= \vartheta \frac{n}{n+1}. \end{aligned}$$

Remark 4.3.4 *Assume that $\vartheta = 1$, i.e. we have the uniform distribution on $[0,1]$. Then*

$$\vartheta - E_\vartheta \max_{i=1,\ldots,n} X_i = \frac{1}{n+1}$$

which can be interpreted as follows: when n points are randomly thrown into $[0,1]$ (independently, with uniform law) then the largest of these value tends to be $\frac{1}{n+1}$ away from the right boundary. The same is true by symmetry for the smallest value and the left boundary. □

The estimator $X_{[n]}$ is thus biased, but the bias can easily be corrected: the estimator

$$T_c(X) = \frac{n+1}{n} X_{[n]}$$

(which moves $X_{[n]}$ up towards the interval boundary) is unbiased. To find its variance we note

$$\begin{aligned} E_\vartheta X_{[n]}^2 &= \vartheta^{-n} n \int_0^\vartheta y^2 \cdot y^{n-1} dy = \vartheta^{-n} n \left[\frac{1}{n+2} y^{n+2}\right]_0^\vartheta \\ &= \vartheta^2 \frac{n}{n+2}, \end{aligned}$$

$$\begin{aligned}
\mathrm{Var}_\vartheta\left(T_c(X)\right) &= \left(\frac{n+1}{n}\right)^2 \left(\mathrm{Var}_\vartheta X_{[n]}^2\right) \\
&= \vartheta^2 \left(\frac{n+1}{n}\right)^2 \left(\frac{n}{n+2} - \left(\frac{n}{n+1}\right)^2\right) \\
&= \vartheta^2 \left(\frac{(n+1)^2}{n(n+2)} - 1\right) \\
&= \frac{\vartheta^2}{n(n+2)} \qquad (4.24)
\end{aligned}$$

which is smaller than the bound ϑ^2/n *suggested by* (4.23). In fact the variance (4.24) decreases much faster with n than the bound (4.23), by an additional factor $1/(n+2)$.
Theorem 4.3.1 is in fact not applicable since the conditions are violated (the support depends on ϑ, and the density is not differentiable everywhere). That suggests that the statistical model where ϑ describes the support of the density $[0, \vartheta]$ is more informative than the "regular" cases where the Cramer-Rao bound applies. Indeed if $X_{[n]} = y$ then all values $\vartheta < y$ can be excluded with certainty, and certainty is a lot of information in a statistical sense.

Chapter 5
CONDITIONAL AND POSTERIOR DISTRIBUTIONS

5.1 The mixed discrete / continuous model

In section 2.2 we introduced the integrated risk with respect to a prior density g; it was mentioned already that in the Bayesian approach, the parametric family P_ϑ, $\vartheta \in \Theta$ for the observations X is understood as a conditional distribution given ϑ. In model $\mathbf{M}_1$, the data X were discrete and the assumed prior distribution for p was continuous (the Beta densities.) It is clear that this should give rise to a **joint distribution of observation and parameter**. Most basic probability courses treat joint distributions of two r.v.'s (X, U) only for the case that the joint law is either discrete or continuous (hence X and U are of the same type). Let us fill in some technical details here for this mixed case.

When X and U are both discrete or both continuous then a joint distribution is given by a joint probability function $q(x, u)$, or a joint density $q(x, u)$ respectively. In our mixed case where Θ is an interval in $\mathbb{R}$ and our sample space $\mathcal{X}$ for X is finite we can expect that a joint distribution is given by $q(x, u)$ which is of mixed type: $q(x, u) \geq 0$ for all $x \in \mathcal{X}$, $u \in \Theta$ and

$$\sum_{x \in \mathcal{X}} \int_\Theta q(x, u) du = 1. \tag{5.1}$$

Such a function defines a joint distribution Q of X and U: for any subset A of $\mathcal{X}$ and any interval B in Θ

$$Q(X \in A, U \in B) = Q(A \times B) = \sum_{x \in A} \int_B q(x, u) du.$$

It is then possible to extend the class of sets where the probabilities Q are defined: consider sets $C \in \mathcal{S}$ of form

$$C = \bigcup_{x \in A} \{x\} \times B_x$$

where A is an arbitrary subset of $\mathcal{X}$ and for each $x \in A$, B_x is a finite union of intervals in Θ. Define

$$Q(C) = \sum_{x \in A} \int_{B_x} q(x, u) du \tag{5.2}$$

Let $\mathfrak{S}_0$ be the collection of all such sets C as above; the function Q on sets $C \in \mathfrak{S}_0$ fulfills all the axioms of a probability.

Suppose now that P_ϑ, $\vartheta \in \Theta$ is a parametric family of distributions on $\mathcal{X}$ and $g(\vartheta)$ is a density on Θ. Setting

$$q(x, \vartheta) = P_\vartheta(x) g(\vartheta)$$

we have an object q as described above: it is nonnegative and (5.1) is fulifilled:

$$\begin{aligned}\sum_{x\in\mathcal{X}}\int_\Theta P_\vartheta(x)g(\vartheta)d\vartheta &= \int_\Theta\left(\sum_{x\in\mathcal{X}}P_\vartheta(x)\right)g(\vartheta)d\vartheta\\ &= \int_\Theta g(\vartheta)d\vartheta = 1\end{aligned}$$

since g is a density. The joint distribution of X and U thus defined gives rise to marginal and conditional probabilities, for $x\in\mathcal{X}$

$$P(X = x) = P(X = x, U\in\Theta) = \int_\Theta P_\vartheta(x)g(\vartheta)d\vartheta, \tag{5.3}$$

$$\begin{aligned}P(U \in B|X = x) &= \frac{P(U\in B, X = x)}{P(X = x)} \qquad (5.4)\\ &= \frac{\int_B P_\vartheta(x)g(\vartheta)d\vartheta}{\int_\Theta P_\vartheta(x)g(\vartheta)d\vartheta}.\end{aligned}$$

If $P_\vartheta(x) > 0$ for all x and ϑ (which is the case in M_1 if $\Theta = (0,1)$) then also $P(X = x) > 0$, and the conditional probability (5.4) is well defined. Then it is immediate (and shown in probability courses) that the function

$$Q_x(B) = P(U\in B|X = x)$$

fulfills all the axioms of a probability on Θ; it defines the conditional law of U given $X = x$. In the statistical context this is called the **posterior distribution** of ϑ given $X = x$.
It is obvious that this distribution has a density on $\Theta = (a,b)$: for $B = (a,t]$ we have

$$Q_x((a,t]) = \int_a^t P_\vartheta(x)g(\vartheta)\left(\int_\Theta P_u(x)g(u)du\right)^{-1}d\vartheta$$

which identifies the density of Q_x as

$$q_x(\vartheta) = P_\vartheta(x)g(\vartheta)\left(\int_\Theta P_u(x)g(u)du\right)^{-1} \tag{5.5}$$

This density is called the **conditional density** , or in the context of Bayesian statistics, the **posterior density** of ϑ given $X = x$. The formula (5.5) is very simple: for given prior density $g(\vartheta)$, adjoin the probability function $P_\vartheta(x)$ (when $X = x$ is observed) and renormalize $P_\vartheta(x)g(\vartheta)$ such that it integrates to one w.r.t. ϑ (i.e. becomes a density).

Remark 5.1.1 *The formula (5.5) suggests an analog for the case that X is a continuous r.v. as well, with density $p_\vartheta(x)$ say. In this case the formula (5.4) cannnot be used to define a conditional distribution since all events $\{X = x\}$ have probability 0 for the laws $P_\vartheta(x)$ and thus also for the marginal law of X. For continuous r.v.'s X the conditional (posterior) density $q_x(\vartheta)$ is directly defined by replacing the probability $P_\vartheta(x)$ in (5.5) by the density $p_\vartheta(x)$; cf. [D], sect. 3.8 and our discussion to follow in later sections.*

Exercise. To prepare for the purely continuous case, let us see what happens when we reverse the roles of ϑ and X in (5.5), i.e. we take the marginal probability function for X given by (5.3) and

combine it with the conditional density for ϑ given by (5.4). Consider the expression $q_x(\vartheta)P(X=x)$ and, analogously to (5.5), divide it by its sum over all possible values of x ($x \in \mathcal{X}$). Call the result $q_\vartheta(x)$. Show that for any ϑ with $g(\vartheta) > 0$, the relation

$$q_\vartheta(x) = P_\vartheta(x),\ x \in \mathcal{X}$$

holds. This result justifies to call $P_\vartheta(x)$ a conditional probability function under $U = \vartheta$:

$$P_\vartheta(x) = P\left(X = x | U = \vartheta\right),$$

even though U is continuous and the event $U = \vartheta$ has probability 0.

Remark 5.1.2 Consider the uniform prior density on Θ: $g(\vartheta) = (b-a)^{-1}$. Then

$$q_x(\vartheta) = L_x(\vartheta)\left(\int_\Theta L_x(u)du\right)^{-1}$$

i.e. the posterior density is proportional as a function of ϑ to the likelihood function $L_x(\vartheta)$ (the normalizing factor does not depend on ϑ).

5.2 Bayesian inference

Computation of a posterior distribution and of a Bayes estimator (which is associated, as we shall see) can be subsumed under the term *Bayesian inference* . Let us compute the posterior density in our case of model $\mathbf{M}_1$ and prior densities of the Beta family. We have

$$\begin{aligned} P_p(x)g(p) &= p^{z(x)}(1-p)^{n-z(x)} \cdot \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)} \\ &= \frac{p^{z(x)+\alpha-1}(1-p)^{n-z(x)+\beta-1}}{B(\alpha,\beta)}. \end{aligned}$$

The posterior density is proportional to the Beta density $g_{\alpha+z(x),\beta+n-z(x)}$, and hence must coincide with this density:

$$q_x(p) = g_{\alpha+z(x),\beta+n-z(x)}(p).$$

We see that if the prior density is in the Beta class, then the posterior density is also in this class, for any observed data x. Such a family is called a **conjugate family of prior distributions** The next subsection presents a more technical discussion of the Beta family. Using the formula given below for the expectation, we find the expected value of the posterior distribution as

$$\begin{aligned} E(U|X &= x) = \frac{\alpha + z(x)}{\alpha + z(x) + \beta + n - z(x)} \\ &= \frac{\alpha + z(x)}{\alpha + \beta + n} = \frac{\bar{X}_n + \alpha/n}{1 + \alpha/n + \beta/n} = T_{\alpha,\beta}(X), \end{aligned} \tag{5.6}$$

i.e. it coincides with the Bayes estimator for the prior density $g_{\alpha,\beta}$ already found in section 2.4. This is no coincidence, as the next proposition shows. Notationally, the separate symbol U for ϑ as a random variable is only needed for expressions like $P\left(X = x | U = \vartheta\right)$; we suppress U and set $U = \vartheta$ wherever possible. Recall that $E(U|X = x)$ is a general notation for conditional expectation ([D] sec. 4.6), i.e the expectation of a conditional law. In our statistical context $E(\vartheta|X = x)$ is called the **posterior expectation .**

Proposition 5.2.1 *In the statistical model* $\mathbf{M}_f$ *(the sample space* $\mathcal{X}$ *for* X *is finite and* Θ *is an open interval in* $\mathbb{R}$*), let* Θ *be a finite interval and* g *be a prior density on* Θ*. Then, for a quadratic loss function a Bayes estimator* T_B *of* ϑ *is given by the posterior expectation*

$$T_B(x) = E(\vartheta|X = x),\ x \in \mathcal{X}.$$

Proof. Note that for a r.v. $U = \vartheta$ taking values in a finite interval, the expectation and all higher moments always exists, so both for prior and posterior distributions the expectation exists. The integrated risk for any estimator is (it was defined previously for the special case of Model $\mathbf{M}_1$)

$$\begin{aligned} B(T) &= \int_\Theta R(T,\vartheta)g(\vartheta)d\vartheta \\ &= \int_\Theta E_\vartheta \left(T-\vartheta\right)^2 g(\vartheta)d\vartheta. \end{aligned}$$

Obviously this can be written

$$\begin{aligned} B(T) &= \int_\Theta \sum_x \left(T(x)-\vartheta\right)^2 P_\vartheta(x)g(\vartheta)d\vartheta \\ &= \sum_x \int_\Theta \left(T(x)-\vartheta\right)^2 P_\vartheta(x)g(\vartheta)d\vartheta. \end{aligned}$$

From (5.5) we have

$$P_\vartheta(x)g(\vartheta) = q_x(\vartheta)P(X=x),$$

where $P(X=x)$ is the marginal probability function of X. Hence

$$B(T) = \sum_x \left(\int_\Theta \left(T(x)-\vartheta\right)^2 q_x(\vartheta)d\vartheta \right) P(X=x). \tag{5.7}$$

Let $q(\vartheta)$ be an arbitrary density in ϑ, $E_q(\cdot)$, be expectation under q and a be a constant (a does not depend on ϑ). Then we claim

$$E_q \left(a-\vartheta\right)^2 \geq E_q \left(E_q\vartheta - \vartheta\right)^2 = \mathrm{Var}_q\vartheta \tag{5.8}$$

with equality if and only if $a = \vartheta$. (Note that $E_q \left(a-\vartheta\right)^2$ is always finite in our model). Indeed

$$\begin{aligned} E_q \left(a-\vartheta\right)^2 &= E_q \left(a - E_q\vartheta - (\vartheta - E_q\vartheta)\right)^2 \\ &= \left(a-E_q\vartheta\right)^2 + \mathrm{Var}_q\vartheta \end{aligned}$$

in view of $E_q(\vartheta - E_q\vartheta) = 0$, which proves (5.8) and our claim. Now apply this result to the expression in round brackets under the sum sign in (5.7) and obtain that for any given x, $T(x) = E_{q_x}\vartheta = E(\vartheta|X=x) = T_B(x)$ is the unique minimizer. Hence

$$\begin{aligned} B(T) &\geq \int_\Theta E_\vartheta \left(T_B(X)-\vartheta\right)^2 g(\vartheta)d\vartheta \\ &= E\left(T_B(X)-\vartheta\right)^2. \end{aligned}$$

where the last expectation refers to the joint distribution X and ϑ. ■

In the last display we wrote $T_B(x) = E(\vartheta|X=x)$ as a random variable when the conditioning x is seen as random. The common notation for this random variable is $E(\vartheta|X)$.

5.3 The Beta densities

Let us discuss more carefully the Beta class of densities which were used as prior distributions in model $M_{d,1}$. Consider the following family of prior densities for the Bernoulli parameter p: for $\alpha, \beta > 0$

$$g_{\alpha,\beta}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)},\ x \in [0,1].$$

where $B(\alpha,\beta)$ is the beta function

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{5.9}$$

Recall that the Gamma function is defined (for $\alpha > 0$) as

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

and satisfies

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha). \tag{5.10}$$

It was already argued that for $\alpha, \beta > 0$, the function $x \mapsto x^{\alpha-1}(1-x)^{\beta-1}$ is integrable on $[0,1]$ (cf. relation (2.15). Relation (5.9) is proved below. The moments are (k integer)

$$\begin{aligned} m_k &:= \int_0^1 x^k g_{\alpha,\beta}(x)dx \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 x^{k+\alpha-1}(1-x)^{\beta-1}dx \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\cdot\frac{\Gamma(k+\alpha)\Gamma(\beta)}{\Gamma(k+\alpha+\beta)}. \end{aligned}$$

Invoking (5.10), we find

$$m_k = \frac{(k-1+\alpha)(k-2+\alpha)\dots(\alpha)}{(k-1+\alpha+\beta)(k-2+\alpha+\beta)\dots(\alpha+\beta)}.$$

especially (for a r.v. U with beta density $g_{\alpha,\beta}$)

$$\begin{aligned} EU &= m_1 = \frac{\alpha}{\alpha+\beta}, \\ EU^2 &= m_2 = \frac{(\alpha+1)\alpha}{(1+\alpha+\beta)(\alpha+\beta)}, \\ \mathrm{Var}(U) &= EU^2 - (EU)^2 = \frac{(\alpha+1)(\alpha+\beta)\alpha - (1+\alpha+\beta)\alpha^2}{(1+\alpha+\beta)(\alpha+\beta)^2} \\ &= \frac{\alpha\beta}{(1+\alpha+\beta)(\alpha+\beta)^2}. \end{aligned} \tag{5.11}$$

Thus $\alpha = \beta$ implies $EU = 1/2$; in particular the prior distribution for which the Bayes estimator is minimax ($\alpha = \beta = n^{1/2}/2$, cf. Theorem 2.5.2) has expected value $1/2$.

We now show that

$$B(\alpha,\beta)=\int_0^1 u^{\alpha-1}\,(1-u)^{\beta-1}\,du=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{5.12}$$

Consider the **gamma distribution** for parameter α: it has density for $x>0$

$$f_\alpha(x)=\frac{1}{\Gamma(\alpha)}x^{\alpha-1}\exp(-x).$$

Consider independent r.v.s X, Y with gamma laws for parameters α, β, and their joint density

$$\tilde{g}_{\alpha,\beta}(x,y)=f_\alpha(x)f_\beta(y)=\frac{1}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}y^{\beta-1}\exp(-(x+y)).$$

Consider the change of variables $x=rt$, $y=r(1-t)$ for $0<r<\infty$, $0\le t\le 1$. (Indeed every (x,y) with $x>0$, $y>0$ can be uniquely represented in this form: $r=x+y$, $t=x/(x+y)$). The Jacobian matrix is

$$\begin{pmatrix} \frac{\partial}{\partial t}rt & \frac{\partial}{\partial r}rt \\ \frac{\partial}{\partial t}r(1-t) & \frac{\partial}{\partial r}r(1-t)\end{pmatrix}=\begin{pmatrix} r & t \\ -r & (1-t)\end{pmatrix}$$

with determinant $(1-t)r+rt=r$. The new density in variables t,r is

$$\begin{aligned} g^*_{\alpha,\beta}(t,r) &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)}(rt)^{\alpha-1}(r(1-t))^{\beta-1}r\exp(-r)\\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)}r^{\alpha+\beta-1}\exp(-r)t^{\alpha-1}(1-t)^{\beta-1}.\end{aligned}$$

When we integrate over r, the result is the marginal density of t (call this $\tilde{f}_{\alpha,\beta}(t)$) and hence must integrate to one. We find from the definition of the Gamma function

$$\tilde{f}_{\alpha,\beta}(t)=\int_0^\infty g^*_{\alpha,\beta}(t,r)dr=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}t^{\alpha-1}(1-t)^{\beta-1}.$$

This is the density of the Beta law; since this density integrates to one, we proved (5.12).

5.4 Conditional densities in continuous models

Consider a two dimensional random variable (X,Y) with joint density $p(x,y)$. Consider an interval $B_h=(x-h/2,x+h/2)$ and assume $P(X\in B_h)>0$. The conditional probability $P\,(Y\in A|X\in B_h)$ is well defined:

$$P\,(Y\in A|X\in B_h)=\frac{P(Y\in A,X\in B_h)}{P(X\in B_h)} \tag{5.13}$$

It is clear that $P\,(A|X\in B_h)$ considered as a function of A, is a probability distribution, i.e. the conditional distribution of Y given $X\in B_h$. This distribution has a density: for $A=\{Y\le y\}$ we obtain

$$P\,(Y\le y|X\in B_h)=\int_{-\infty}^{y}\int_{B_h}p(x,y)dxdy\,(P(X\in B_h))^{-1} \tag{5.14}$$

which shows that the density is

$$p\,(y|X\in B_h)=\int_{B_h}p(x,y)dx\,(P(X\in B_h))^{-1}\,.$$

When X is continuous, it is desirable to define conditional distributions under an event $\{X = x\}$. Indeed, when X is realized, conditioning on an interval B_h which contains x does not correspond to the actual information we have. We might be tempted to condition on smaller and smaller intervals, all containing the realized value x, and each of these conditions would reflect more accurately the information we have, yet none of them would correspond to the actual information $X = x$.
It is natural to try to pass to a limit here as $h \to 0$. Let $p_X(x)$ be the marginal density of X:

$$p_X(x) = \int p(x, y)dy.$$

Without trying to be rigorous in this limit argument, for $B_h = (x - h/2, x + h/2)$, $h \to 0$

$$\begin{aligned} \int_{(x-h/2,x+h/2)} p(x, y)dx &\approx hp(x, y), \\ P(X \in B_h) = \int_{(x-h/2,x+h/2)} p_X(x)dx &\approx hp_X(x) \end{aligned}$$

Based on this heuristic, we introduce the **conditional density given** $X = x$ by definition: if $p_X(x) > 0$ then

$$p\left(y|x\right) = \frac{p(x, y)}{p_X(x)}. \tag{5.15}$$

The figure is meant to illustrate the idea of the of $p\left(y|x\right)$ as giving the "relative weight " given to different y's by the joint density, once the other argument x is fixed.
The expression $p\left(y|x\right)$ is defined only in case $p_X(x) > 0$. In that case however it clearly is a density, since it is nonnegative and, as a function of y, it integrates to one. Then, for such x for which $p_X(x) > 0$, we define the **conditional distribution of** Y **given** $X = x$: for any A

$$P(Y \in A|X = x) = \int_A p\left(y|x\right) dy.$$

Note that formula (5.15) is analogous to the conditional probability function in the discrete case: if X and Y can take only finitely many values x, y and

$$p(x, y) = P(Y = y, X = x),\ p_X(x) = P(X = x)$$

are the joint and marginal probability functions then if $p_X(x) > 0$, by the classical definition of conditional probability

$$p\left(Y = y|X = x\right) = \frac{P(Y = y, X = x)}{P(X = x)} = \frac{p(x, y)}{p_X(x)}.$$

which looks exactly like (5.15), but all the terms involved are probability functions, not densities. To state a connection between conditional densities and independence, we need a slightly more precise definition. First note that densities are not unique; they can be modified in certain points or subsets without affecting the corresponding probability measure.

Definition 5.4.1 *Let Z be a random variable with values in $\mathbb{R}^k$ and density q on $\mathbb{R}^k$. A **version** of q is a density $\tilde{q}$ on $\mathbb{R}^k$ such that for all sets $A \subset \mathbb{R}^k$ for which the k-dimensional volume (measure) is defined,*

$$\int_A q(z)dz = \int_A \tilde{q}(z)dz$$

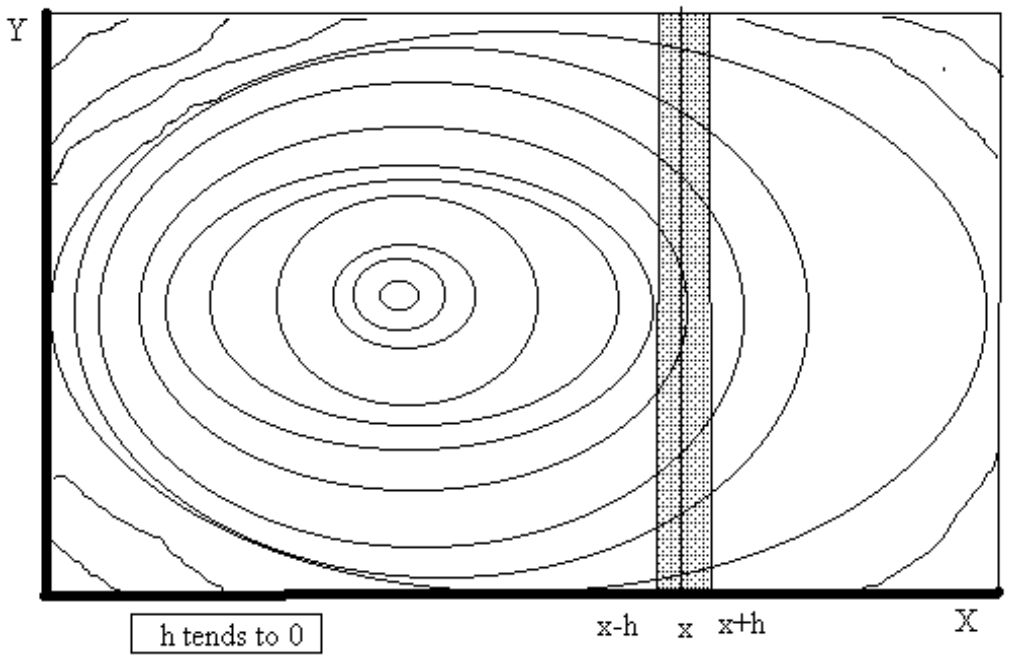


Figure 1 Conditional density of Y given $X = x$. The altitude lines symbolize the joint density of X and Y. On the dark strip, in the limit when h tends to 0, this gives the conditional density of Y given $X = x$

In particular, if $A = A_1 \cup A_2$ and A_2 has volume 0 then the density q can be arbitrarily modified on A_2. On the real line, A_2 might consists of a countable number of points; on $\mathbb{R}^k$, A_2 might consists of hyperplanes, smooth surfaces etc.

Definition 5.4.2 *Suppose that $Z = (Z_1, \ldots, Z_k)$ is a continuous random variable with values in $\mathbb{R}^k$ and with joint density $p(z) = p(z_1, \ldots, z_k)$. Set $X = Z_1$, $Y = (Z_2, \ldots, Z_k)$ and $p(x, y) = p(z)$. A version of the conditional density of Y given $X = x$ is any function $p(y|x)$ with properties:*
(i) For all x, $p(y|x)$ is a density in y, i.e.

$$p(y|x) \geq 0, \int p(y|x)\, dy = 1$$

(ii) There is a version of $p(x, y)$ such that for $p_X(x) = \int p(x, y)dy$ we have

$$p(x, y) = p(y|x)\, p_X(x) \tag{5.16}$$

Lemma 5.4.3 *A conditional density as defined above exists.*

Proof. If x is such that $p_X(x) > 0$ then we can set

$$p(y|x) = \frac{p(x, y)}{p_X(x)}. \tag{5.17}$$

Clearly this is a density in y since

$$\int \frac{p(x, y)}{p_X(x)} dy = \int \frac{p_X(x)}{p_X(x)} dy = 1.$$

Let $A = \{x \in \mathbb{R} : p_X(x) = 0\}$. Clearly $P(X \in A) = 0$, and hence for

$$A_0 = \{(x, y) : x \in A\}$$

we obtain

$$P((X, Y) \in A_0) = P(X \in A) = 0.$$

Now modify $p(x, y)$ on A_0, to obtain another version $\tilde{p}$, namely set

$$\tilde{p}(x, y) = 0 \text{ for } (x, y) \in A_0.$$

Clearly $\tilde{p}$ is a version of p: for any event B

$$\begin{aligned} \int_B p(z)dz &= P(Z \in B) = P(Z \in B \cap A_0) \\ &= \int_{B \cap A_0} p(z)dz = \int_{B \cap A_0} \tilde{p}(z)dz = \int_B \tilde{p}(z)dz. \end{aligned}$$

For this version $\tilde{p}(x, y)$ we have: $\tilde{p}_X(x) = 0$ implies $\tilde{p}(x, y) = 0$ and hence for such x, $p(y|x)$ can be chosen as an arbitrary density to fulfill (5.16). ■

Proposition 5.4.4 *(X, Y) with joint density $p(x, y)$ are independent if and only if there is a version of $p(y|x)$ which does not depend on x.*

Proof. If X and Y are independent then $p(x, y) = p_X(x)p_Y(y)$. Thus $p_Y(y) = p(y|x)$ is such a version. Conversely, if $p(y|x)$ is such a version then

$$\begin{aligned} p_Y(y) &= \int p(x, y)dx = \int p(y|x)\, p_X(x)dx \\ &= p(y|\cdot) \int p_X(x)dx = p(y|\cdot) \end{aligned}$$

hence

$$p(x, y) = p_Y(y)p_X(x)$$

which implies that (X, Y) are independent. ■

In case that all occurring joint and marginal densities are positive, there is really no need to consider modifications; the conditional densities can just be taken as (5.17).

Let now $f(Y)$ be any function of the random variable Y. The **conditional expectation** of $f(Y)$ given $X = x$ is

$$E(f(Y)|X = x) = \int f(y)p(y|x)dy,$$

i.e. the expectation with respect to the conditional distribution of Y given $X = x$, given by the conditional density $p(y|x)$. Note that this conditional expectation depends on x, i.e. is a function of x- the realization of the random variable X. Sometimes it is useful to "keep in mind" the original random nature of this realization, i.e. consider the conditional expectation to be a function of the random variable X. The common notation for this random variable is

$$E(f(Y)|X)$$

it is a random variable which is a function of X. For the expectation we then have

$$\begin{aligned} EE\left(f(Y)|X\right) &= E\int f(y)p(y|X)dy = \int\left(\int f(y)p(y|x)dy\right)p_X(x)dx \\ &= \int f(y)p(x,y)d(x,y) = \int f(y)p_Y(y)dy = Ef(Y), \end{aligned}$$

i.e. expectation of conditional expectation yields the expectation. These notions are the same in the case of discrete random variables; only in the continuous case we had to pay attention to the fact that $P(X = x) = 0$.

5.5 Bayesian inference in the Gaussian location model

Suppose that we observe a continuous random variable $X = (X_1, \ldots, X_n)$ with density unknown density $p_\vartheta(x)$, $x \in \mathbb{R}^n$, $\vartheta \in \Theta$ where Θ is an open subset of $\mathbb{R}^d$ (this was called model $\mathbf{M}_c$ before). In a Bayesian statistical approach, assume that ϑ "becomes random" as well. Suppose we have a prior density $\pi(\vartheta)$ on the parameter space Θ and wish to build a joint density of X and ϑ from this.
We need some regularity condition on functions and sets. A set $A \subseteq \mathbb{R}^k$ is called *measurable* if its k-dimensional volume is defined (volume may be 0 or ∞). One also defines *measurable functions* on $\mathbb{R}^k$; we do not give the definition here but remark only that measurability is necessary for integrals $\int f(x)dx$ to be defined (thus densities must be measurable). All continuous functions are measurable, and also functions which are continuous except on a set of volume 0. In this course, we need not dwell on these technicalities (measure theory); we assume that all occurring sets and functions are are measurable.

Proposition 5.5.1 *Suppose that in model $\mathbf{M}_c$ a density $\pi(\vartheta)$ on Θ is given such that $\pi(\vartheta) > 0$, $\vartheta \in \Theta$, (and that $p_\vartheta(x)$ is jointly measurable as a function of (x, ϑ)). Then the function*

$$p(x, \vartheta) = p_\vartheta(x)\pi(\vartheta) \tag{5.18}$$

is a density on $\mathbb{R}^k \times \Theta$. When this is construed as a joint density of (X, ϑ), then $p_\vartheta(x)$ is a version of the conditional density $p(x|\vartheta)$.

Proof. The function $p(x, \vartheta)$ is nonnegative; we have

$$\int_\Theta \int_{\mathbb{R}^n} p(x, \vartheta)dxd\vartheta = \int_\Theta \int_{\mathbb{R}^n} p_\vartheta(x)dx\pi(\vartheta)d\vartheta = \int_\Theta \pi(\vartheta)d\vartheta = 1,$$

thus $p(x, \vartheta)$ is a density. Then $\pi(\vartheta)$ is the marginal density of ϑ, derived from the joint density: indeed

$$\int p(x, \vartheta)dx = \int p_\vartheta(x)\pi(\vartheta)dx = \pi(\vartheta).$$

Then, we immediately see from (5.18) that $p_\vartheta(x)$ is a version of the conditional density $p(x|\vartheta)$. ■
This justifies the notation that for a parametric family of densities, one writes interchangeably $p_\vartheta(x)$ or $p(x|\vartheta)$. Then $p(\vartheta|x)$ is again called the *posterior density*. If Θ is an interval in $\mathbb{R}$ then the conditional expectation

$$E(\vartheta|X = x) = \int \vartheta p(\vartheta|x)d\vartheta$$

if it exists, is again called *posterior expectation.* Let us discuss the case of the Gaussian location model, first in the case of sample size $n = 1$. We can represent $X = X_1$ as

$$X = \mu + \xi$$

where ξ is centered normal: $\mathcal{L}(\xi_i) = N(0, \sigma^2)$. The parameter ϑ is μ and parameter space Θ is $\mathbb{R}$. Assume that μ becomes random as well: $\mathcal{L}(\mu) = N(m, \tau^2)$ where m and τ^2 are known (these are sometimes called *hyperparameters*). For Bayesian statistical inference, we wish to compute the conditional distribution of μ given X, i.e. the posterior distribution of the parameter.
The joint density of X and μ is

$$\begin{aligned} p(x,\mu) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\cdot\frac{1}{\sqrt{2\pi}\tau}\exp\left(-\frac{(\mu-m)^2}{2\tau^2}\right) \\ &= \frac{1}{2\pi\sigma\tau}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}-\frac{(\mu-m)^2}{2\tau^2}\right). \end{aligned}$$

To find the marginal density of X, we define $\xi = X - \mu$ and compute the joint distribution of ξ and μ. First find the conditional law $\mathcal{L}(\xi|\mu)$ which is the law of $X - \mu$ when μ is fixed. This is $N(0, \sigma^2)$, and since it does not depend on μ, ξ and μ are independent in their joint law (Proposition 5.4.4). In other words, we have

$$X = \mu + \xi$$

where μ, ξ are independent $N(m, \tau^2)$, $N(0, \sigma^2)$ respectively. By the properties of the normal distribution, we conclude that X has marginal law $\mathcal{L}(X) = N(m, \sigma^2 + \tau^2)$, with density ($\varphi$ is the standard Gaussian density)

$$\begin{aligned} p_X(x) &= \frac{1}{(\sigma^2+\tau^2)^{1/2}}\varphi\left(\frac{x-m}{(\sigma^2+\tau^2)^{1/2}}\right) \\ &= \frac{1}{(2\pi)^{1/2}(\sigma^2+\tau^2)^{1/2}}\exp\left(-\frac{(x-m)^2}{2(\sigma^2+\tau^2)}\right) \end{aligned}$$

and the posterior density is

$$\begin{aligned} p(\mu|x) &= p(x,\mu)/p_X(x) \\ &= \frac{(\sigma^2+\tau^2)^{1/2}}{(2\pi)^{1/2}\sigma\tau}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}-\frac{(\mu-m)^2}{2\tau^2}+\frac{(x-m)^2}{2(\sigma^2+\tau^2)}\right). \end{aligned}$$

We conjecture that this is a Gaussian density (if X and μ were independent then $p(\mu|x) = p_\mu(\mu)$ certainly would be Gaussian). We shall establish that $p(\mu|x)$ is indeed Gaussian with variance

$$\frac{\sigma^2\tau^2}{\sigma^2+\tau^2} := \rho^2. \tag{5.19}$$

Note that for $\mu^* = \mu - m$, $x^* = x - m$

$$-\frac{(x-\mu)^2}{2\sigma^2}-\frac{(\mu-m)^2}{2\tau^2}+\frac{(x-m)^2}{2(\sigma^2+\tau^2)} = -\frac{1}{2\rho^2}\left(\mu^*-\frac{\tau^2}{\sigma^2+\tau^2}x^*\right)^2$$

and setting

$$\beta = \frac{\tau^2}{\sigma^2+\tau^2}, \tag{5.20}$$

we obtain

$$p(\mu|x) = \frac{1}{\rho}\varphi\left(\frac{\mu - m - \beta(x-m)}{\rho}\right).$$

Proposition 5.5.2 *In the Gaussian location model* $\mathbf{M}_2$ *for sample size* $n = 1$ *and a normal prior distribution* $\mathcal{L}(\mu) = N(m, \tau^2)$*, the posterior distribution is*

$$\mathcal{L}(\mu|X = x) = N(m + \beta(x-m), \rho^2)$$

where ρ *and* β *are defined by (5.19), (5.20). The normal family* $\left\{N(m, \tau^2),\ m \in \mathbb{R},\ \tau^2 > 0\right\}$ *is a conjugate family of prior distributions.*

Let us interpret this result. The posterior distribution of μ has expected value $m + \beta(x-m)$; note that $0 < \beta < 1$. The prior distribution of μ had expectation m, so the posterior expectation of μ intuitively represents a "compromise" between the prior belief and the empirical evidence about μ, i.e. $X = (\mu + \xi) = x$. Indeed

$$E(\mu|X = x) = m + \beta(x-m) = m(1-\beta) + \beta x$$

so $E(\mu|X = x)$ is a convex (linear) combination of data x and prior mean m (is always between these two points). In other words, the data x are "shrunken" towards m when $x - m$ is multiplied by β. A similar shrinkage effect was observed for the Bayes estimator in the Bernoulli model $\mathbf{M}_1$ when the prior mean was $1/2$ (recall that the minimax estimator there was Bayes for a Beta prior with $\alpha = \beta$, which has mean $1/2$).
Moreover

$$\rho^2 = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2} < \min(\tau^2, \sigma^2).$$

The posterior variance ρ^2 is thus smaller than both the prior variance and the variance of the data given ϑ (i.e. σ^2). It is seen that the information in the data and in the prior distribution is "added up" to give a smaller variability (a posteriori) than is contained in either of the two sources. In fact

$$\frac{1}{\rho^2} = \frac{1}{\sigma^2} + \frac{1}{\tau^2}.$$

The inverse variance of a distribution can be seen as a measure of concentration, sometimes called **precision** . Then the precision of the posterior is the sum of precisions of the data and of the prior distribution.
The posterior expectation again can be shown to be a Bayes estimator for quadratic loss. Before establishing this, let us discuss two limiting cases.

Case 1 Variance τ^2 of the prior density is large: $\tau^2 \to \infty$. In this case

$$\beta = \frac{\tau^2}{\sigma^2 + \tau^2} \to 1,\ \rho^2 \to \sigma^2. \tag{5.21}$$

A large prior variance means that the prior density is more and more spread out, i.e more "diffuse" or noninformative . $\beta \to 1$ then means that the prior information counts less and less in comparison to the data, and the posterior expecation of μ tends to x. This means, that in the limit the only evidence we have about μ is the realized value $X = x$.

Case 2 Variance τ^2 of the prior density is small: $\tau^2 \to 0$. In this case

$$\beta = \frac{\tau^2}{\sigma^2 + \tau^2} \to 0,\ \rho^2 \to 0. \tag{5.22}$$

A small prior variance means that the prior density is more and more concentrated around m. Then (5.22) means that the "belief" that μ is near m becomes overwhelmingly strong, and forces the posterior distribution to concentrate around m as well.

Case 3 Variance σ^2 of the data (given μ) is large: $\sigma^2 \to \infty$. In this case

$$\beta = \frac{\tau^2}{\sigma^2 + \tau^2} \to 0,\ \rho^2 \to \tau^2$$

The posterior density tends to the prior density, since the quality of the information $X = x$ becomes inferior (large variance σ^2)

Case 4 Variance σ^2 of the data (given μ) is small: $\sigma^2 \to 0$. We expect the prior distribution to matter less and less, since the data are more and more reliable. Indeed

$$\beta = \frac{\tau^2}{\sigma^2 + \tau^2} \to 1,\ \rho^2 \to 0$$

which is similar to (5.21) for case 1.

In the case of prior mean $m = 0$, the quantity

$$r = \frac{\tau^2}{\sigma^2}$$

is often called the **signal-to-noise ratio** . Recall that in the Gaussian location model (Model $\mathbf{M}_{d,1}$) for $n = 1$ the data are

$$X = \mu + \xi$$

where $\mathcal{L}(\mu) = N(0, \tau^2)$ and $\mathcal{L}(\xi) = N(0, \sigma^2)$. The parameter μ can be seen as a "signal" which is observed with noise ξ. For $m = 0$ we have

$$\tau^2 = E\mu^2,\ \sigma^2 = E\xi^2$$

so that τ^2, σ^2 represent the average absolute value the (squared) signal and noise. The parameters of the posterior density can be expressed in terms of the signal-to-noise ratio r and σ^2:

$$\beta = \frac{\tau^2}{\sigma^2 + \tau^2} = \frac{r}{1+r},\ \rho^2 = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2} = \sigma^2\frac{r}{1+r}$$

and the discussion of the limiting cases 1-4 above could have been in terms of r.

Remark 5.5.3 *The source of randomness of the parameter ϑ may not only be "prior belief", as argued until now, but may be entirely natural and part of the model, along with the randomness of the noise ξ. This randomness assumption even dominates in the statistical literature related to communication and signal transmission (assumption of a random signal). The problem of estimation of ϑ then becomes the problem of* ***prediction .*** *Predicting ϑ still means to find an estimator*

$T(X)$ depending on the data, but for assessing the quality of a predictor, the randomness of ϑ is always taken into account. The risk of a predictor is the expected loss

$$EL(T(X), \vartheta).$$

w.r. to the joint distribution of X and ϑ. This coincides with the mixed risk of an estimator

$$B(T) = E_\pi \left(E_\vartheta L(T(X), \vartheta) \right)$$

*in Bayesian statistics when a prior distribution π on ϑ is used to build the joint distribution of (X, ϑ). Thus an **optimal predictor** is the same as a Bayes estimator, when the loss functions L coincide.*

In statistical communication theory (or **information theory)**, a family of probability distributions $\{P_\vartheta, \vartheta \in \Theta\}$ on some sample space X is often called a **noisy (communication) channel,** the parameter ϑ is the **signal at the input** and observations X are called the **signal at the output**. It is assumed that the signal ϑ is intrinsically random, e.g. the finite set $\Theta = \{\vartheta_1, \ldots, \vartheta_k\}$ may represent letters of an alphabet which occur at different frequencies (or probabilities) $\pi(\vartheta)$. Thus $\pi(\vartheta)$ may be given naturally, by the frequency of certain letters in a given language. Typical examples of noisy channels are

- **the binary channel :** $\Theta = \{0, 1\}$, $P_\vartheta = B(1, p_\vartheta)$ where $p_\vartheta \in (0, 1)$, $\vartheta = 0, 1$. The signals are either 0 and 1, and given $\vartheta = 1$ the channel then produces the correct signal with probability p_1 and the distorted signal 0 with probability $1 - p_1$; analogously for $\vartheta = 0$. It is naturally to assume here $p_1 > 1/2$, $p_0 < 1/2$ lest the channel would be entirely distorted (gives the wrong signal with higher probability than the correct one). A natural prior π, would be the uniform here, $(\pi(0) = \pi(1) = 1/2)$ since in most data streams is probably not sensible to assume that 0 is more frequent than 1.

- **the n-fold product of the binary channel:** $\Theta = \{0, 1\}^n$, $P_\vartheta = \bigotimes_{i=1}^n B(1, p_{\vartheta(i)})$ where $\vartheta(i)$ is the i-th component of a sequence of 0's and 1's of length n and $\bigotimes_{i=1}^n$ signifies the n-fold product of laws. The numbers $p_\vartheta \in (0, 1)$, $\vartheta = 0, 1$ are as above in the simple binary channel. Here a signal is a sequence of length n, and the channel transmits this sequence such that each component is sent independently in a simple binary channel. The difference to the previous case is that the signal is a sequence of length n, not just 0 or 1; thus $n = 8$ gives $card(\Theta) = 2^8 = 256$ and this suffices to encode and send all the ASCII signs. Here it is natural to assume a non-uniform distribution π on the alphabet Θ since ASCII signs (e.g. letters) do not all occur with equal probability.

- **the Gaussian channel .** Let $\vartheta \in \mathbb{R}$ and $P_\vartheta = N(\vartheta, \sigma^2)$ where σ^2 is fixed. Here the signal at the output has the form

$$X = \vartheta + \xi \tag{5.23}$$

  where $\mathcal{L}(\xi) = N(0, \sigma^2)$, i.e. the channel transmits the signal in additive Gaussian noise. Here the real numbers ϑ are "codes" agreed upon for any other signal, such as sequences of of 0's and 1's as above. Thus the channel (5.23) itself coincides with the Gaussian location model. However since there are usually only a finite number of signals possible, in information theory one considers prior distributions π for the signal ϑ which are concentrated on finite sets $\Theta \subset \mathbb{R}$.

After this digression, our next task is Bayesian inference in the Gaussian location model for general sample size n. Our prior distribution for ϑ will again be $N(m, \tau^2)$. We have

$$X_i = \vartheta + \xi_i$$

where ξ_i are $N(0, \sigma^2)$ and independent. Consider the density of X given $\vartheta = \mu$:

$$\begin{aligned} p_\mu(x) &= p(x|\mu) = p(x_1, \ldots, x_n|\mu) \\ &= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sigma n^{-1/2}} \varphi\left(\frac{\bar{x}_n - \mu}{\sigma n^{-1/2}}\right) \cdot \frac{1}{n^{1/2}(2\pi\sigma^2)^{(n-1)/2}} \exp\left(-\frac{s_n^2}{2\sigma^2 n^{-1}}\right) \end{aligned}$$

where the last line is the representation found in (3.4) in the proof of Proposition 3.0.5, with φ the standard normal density and s_n^2 the sample variance. The first factor is a normal density in $\bar{x}_n$ and the second factor (call it $\tilde{p}(s_n^2)$ for now) does not depend on μ. If we denote $\varphi_{\mu,\sigma}(x)$ the density of the normal law $N(\mu, \sigma^2)$ then

$$p_\mu(x) = \varphi_{\mu, n^{-1/2}\sigma}(\bar{x}_n) \cdot \tilde{p}(s_n^2).$$

If $\pi(\mu)$ is the prior density then

$$p(\mu|x) = \frac{p_\mu(x)\pi(\mu)}{\int p_\mu(x)\pi(\mu)d\mu} = \frac{\varphi_{\mu, n^{-1/2}\sigma}(\bar{x}_n)\pi(\mu)}{\int \varphi_{\mu, n^{-1/2}\sigma}(\bar{x}_n)\pi(\mu)d\mu}$$

This coincides with the posterior distribution for one observation ($n = 1$) at value $X_1 = \bar{x}_n$ and variance $n^{-1}\sigma^2$. In other words, the posterior distribution given $X = x$ may be computed as if we observed only the sample mean $\bar{X}_n$, taking into account that its law is $N(\mu, n^{-1}\sigma^2)$. Thus the posterior density depends on the vector x only via the function $\bar{x}_n$ of x.

Proposition 5.5.4 *In the Gaussian location model* $\mathbf{M}_2$ *for general sample size* n *and a normal prior distribution* $\mathcal{L}(\mu) = N(0, \tau^2)$*, the posterior distribution is*

$$\mathcal{L}(\mu|X = x) = N(m + \beta(\bar{x}_n - m), \rho^2) \tag{5.24}$$

where ρ *and* β *are defined by*

$$\beta = \frac{\tau^2}{n^{-1}\sigma^2 + \tau^2},\ \rho^2 = \frac{n^{-1}\sigma^2\tau^2}{n^{-1}\sigma^2 + \tau^2} \tag{5.25}$$

The normal family $\left\{N(m, \tau^2),\ m \in \mathbb{R},\ \tau^2 > 0\right\}$ *is a conjugate family of prior distributions.*

We again emphasize the special role of the sample mean.

Corollary 5.5.5 *In Model* $\mathbf{M}_2$ *for general sample size* n*, consider the statistic* $\bar{X}_n$ *and regard these as data in a derived model:*

$$\mathcal{L}(\bar{X}_n|\mu) = N(\mu, n^{-1}\sigma^2)$$

(Gaussian location for sample size $n = 1$ *and variance* $n^{-1}\sigma^2$*). In this model, the normal prior* $\mathcal{L}(\mu) = N(m, \tau^2)$ *leads to the same posterior distribution (5.24) for* μ*.*

Remark 5.5.6 ***Sufficient statistics*** *Properties of statistics (data functions) $T(X)$ like these suggest that $T(X)$ may contain all the relevant information in the sample, i.e. that is suffices to take the statistic $T(X)$ and perform all inference about the parameter ϑ in the parametric model for $T(X)$ derived from the original family $\{P_\vartheta, \vartheta \in \Theta\}$ for X:*

$$\{Q_\vartheta, \vartheta \in \Theta\} = \{\mathcal{L}(T(X)|\vartheta), \vartheta \in \Theta\}$$

which is a family of distributions in the space $\mathcal{T}$ where $T(X)$ takes its values. This idea is called the ***sufficiency principle*** *and T would be a sufficient statistic. At this point we do not rigorously define this concept; noting only that it is of fundamental importance in the theory of statistical inference.*

The expectation of the posterior distribution in Proposition 5.5.4 is $m + \beta(\bar{x}_n - m)$. It is natural to call this the *conditional expectation* of μ given $X = x$, or the posterior expecation, written $E(\mu|X = x)$. (We have not shown so far that is is unique; more accurately we should call it a version of the conditional expectation.) Clearly for $n \to \infty$ we have $\beta \to 1$ and $E(\mu|X = x)$ will be close to the sample mean. Moreover, the above corollary shows that in a discussion of limiting cases as in Case 1 – Case 4 above, all statements carry over when X (the one observation for $n = 1$) is replaced by the sample mean. In addition, the noise variance in this discussion now is replaced by $n^{-1}\sigma^2$. Thus e.g. case 4 can be taken to mean that as sample size n increases, the prior distribution matters less and less (indeed we have more and more information in the sample). The analog of the signal-to-noise ratio would be

$$r = \frac{n\tau^2}{\sigma^2} \to \infty \text{ as } n \to \infty.$$

Again we see that intuitively large sample size n amounts to "small noise".

5.6 Bayes estimators (continuous case)

The following discussion of Bayes estimators as posterior expectations is analogous to the case of finite sample space (Proposition 5.2.1). Consider the again the general continuous statistical model $\mathbf{M}_c$ where a density $\pi(\vartheta)$ on Θ is given such that $\pi(\vartheta) > 0$, $\vartheta \in \Theta$. Consider the mixed quadratic risk of an estimator T

$$\begin{aligned} B(T) &= \int_\Theta R(T, \vartheta)\pi(\vartheta)d\vartheta \\ &= \int_\Theta E_\vartheta \left(T(X) - \vartheta\right)^2 \pi(\vartheta)d\vartheta = E\left(T(X) - \vartheta\right)^2 \end{aligned}$$

where the last expectation is taken w.r. to the joint distribution of X and ϑ. A **Bayes estimator** is an estimator T_B which minimizes $B(T)$:

$$B(T_B) = \inf_T B(T).$$

It is also called a **best predictor**, depending on the context (cf. remark 5.5.3).
In the statistical model $\mathbf{M}_f$ (the sample space $\mathcal{X}$ for X is finite and Θ is an finite interval in $\mathbb{R}$), let Θ be a finite interval and g be a prior density on Θ. Then, for a quadratic loss function a Bayes estimator T_B of ϑ is given by the posterior expectation

$$T_B(x) = E(\vartheta|X = x),\ x \in \mathcal{X}.$$

Proposition 5.6.1 *In model* $\mathbf{M}_c$ *assume a prior density* $\pi(\vartheta)$ *on* Θ *such that* $\pi(\vartheta) > 0$, $\vartheta \in \Theta$. *Suppose that for all realizations* x *of* X, *there is a version of the posterior density with finite second moment* ($E(\vartheta^2|X = x) < \infty$). *Then the posterior expectation* $T(x) = E(\vartheta|X = x)$ *is a Bayes estimator for quadratic loss.*

Proof. In Proposition 5.2.1 the finitene second moment was automatically ensured by the condition that Θ was a finite interval; otherwise the proof is entirely analogous. Under the condition $E(\vartheta|X = x)$ exists. We have

$$\begin{aligned} B(T) &= \int_{\mathbb{R}^k \times \Theta} (T(x) - \vartheta)^2 \, p(x|\vartheta)\pi(\vartheta) dx d\vartheta \\ &= \int_{\mathbb{R}^k} \left(\int_\Theta (T(x) - \vartheta)^2 \, p(\vartheta|x) d\vartheta \right) p_X(x) dx \\ &= \int_{\mathbb{R}^k} E\left((T(X) - \vartheta)^2 \, | X = x \right) p_X(x) dx. \end{aligned}$$

Invoking relation (5.8) again, we find for any x and $T_B(x) = E(\vartheta|X = x)$

$$\int_\Theta (T(x) - \vartheta)^2 \, p(\vartheta|x) d\vartheta \geq \int_\Theta (T_B(x) - \vartheta)^2 \, p(\vartheta|x) d\vartheta.$$

This holds true even if the left side is infinite. The right side is the variance of $p(\vartheta|x)$ and is finite under the assumptions. Hence

$$B(T) \geq E\left(T_B(X) - \vartheta\right)^2.$$

■

Corollary 5.6.2 *In Model* $\mathbf{M}_2$ *for general sample size* n *and a normal prior distribution* $\mathcal{L}(\mu) = N(m, \tau^2)$, *the Bayes estimator for quadratic loss is*

$$T_B(X) = m + \beta(\bar{X}_n - m)$$

where β *is given by (5.25). The Bayes risk is*

$$B(T_B(X)) = E\left(T_B(X) - \mu\right)^2 = \ \rho^2 = \frac{n^{-1}\sigma^2\tau^2}{n^{-1}\sigma^2 + \tau^2}. \tag{5.26}$$

Proof. The second statement follows from the fact that if $\mathrm{Var}(\mu|X)$ denotes the variance of the posterior density $p(\vartheta|X)$ then, from the proof of Proposition 5.6.1,

$$E\left(T_B(X) - \mu\right)^2 = E(\mathrm{Var}(\mu|X))$$

and Proposition (5.5.4), which gives

$$\mathrm{Var}(\mu|X = x) = \rho^2$$

Thus $\mathrm{Var}(\mu|X = x)$ does not depend on x and $E(\mathrm{Var}(\mu|X)) = \rho^2$. ■

Note that speaking of "the" Bayes estimator is justified here: indeed when we start with the normal conditional density for our data $X_1, \ldots, X_n$. and normal prior, then there is always a version of the posterior density which is normal. As long as we do not arbitrarily modify these normal densites in certain points (which is theoretically allowed) we obtain a unique posterior expectation for all data points x.

5.7 Minimax estimation of Gaussian location

The Bayes estimators have also interesting properties with regard to the original risk function $R(T,\vartheta)$, without integration over Θ. One such statement (admissibility) was proved in Proposition 2.3.2. It is easy to find the risk $R(T,\vartheta)$ of the Bayes estimator: the usual bias-variance decomposition gives

$$E_\vartheta\left(T_B(X)-\vartheta\right)^2 = E_\vartheta\left(T_B(X)-E_\vartheta T_B(X)\right)^2 + E_\vartheta\left(E_\vartheta T_B(X)-\vartheta\right)^2.$$

Consider the Gaussian location model with a mean zero Gaussian prior $\mathcal{L}(\mu)=N(0,\tau^2)$. According to Corollary 5.6.2 we have $T_B(X)=\beta\bar{X}_n$ and hence

$$\begin{aligned} E_\mu T_B(X)-\mu &= \frac{\tau^2}{n^{-1}\sigma^2+\tau^2}E_\mu\bar{X}_n-\mu = \frac{\tau^2\mu}{n^{-1}\sigma^2+\tau^2}-\mu \\ &= -\frac{n^{-1}\sigma^2}{n^{-1}\sigma^2+\tau^2}\mu, \\ \mathrm{Var}_\mu\left(T_B(X)\right) &= \beta^2\mathrm{Var}_\mu\left(\bar{X}_n\right) = \left(\frac{\tau^2}{n^{-1}\sigma^2+\tau^2}\right)^2 n^{-1}\sigma^2, \end{aligned}$$

hence

$$\begin{aligned} R(T_B,\mu) &= E_\mu\left(T_B(X)-\mu\right)^2 = \left(\frac{1}{n^{-1}\sigma^2+\tau^2}\right)^2\left(\tau^4 n^{-1}\sigma^2+(n^{-1}\sigma^2)^2\mu^2\right) \\ &= n^{-1}\sigma^2\left(\frac{\tau^2}{n^{-1}\sigma^2+\tau^2}\right)^2\left(1+\frac{n^{-1}\sigma^2}{\tau^4}\mu^2\right). \end{aligned}$$

The sample mean is an unbiased estimator, hence

$$R(\bar{X}_n,\mu) = E_\mu\left(\bar{X}_n-\mu\right)^2 = \mathrm{Var}_\mu\left(\bar{X}_n\right) = n^{-1}\sigma^2. \tag{5.27}$$

Comparing the two risks, we note the following facts.

A) Since

$$\frac{\tau^2}{n^{-1}\sigma^2+\tau^2} < 1$$

for small values of μ we have

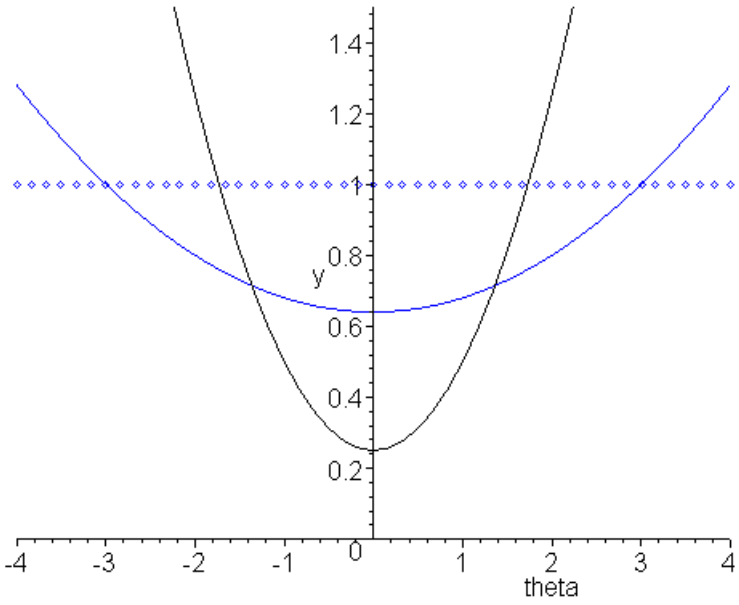
$$R(T_B,\mu) < R(\bar{X}_n,\mu)$$

i.e. for small values of μ the estimator T_B is better.

B) For $|\mu|\to\infty$, $R(T_B,\mu)\to\infty$ whereas $R(\bar{X}_n,\mu)$ remains constant. For large values of μ the estimator $\bar{X}_n$ is better.

C) As $\tau\to\infty$, the risk of $T_B(X)$ approaches the risk of $\bar{X}_n$, at any given value of μ.

The sample mean recommends itself as particularly prudent, in the sense that it takes into account possibly large values of μ, whereas the Bayes estimator is more "optimistic" in the sense that it is geared towards smaller values of μ.



Risks of the Bayes estimators for $\tau = 1$ and $\tau = 2$, and risk of sample mean (dotted), for $n^{-1}\sigma^2 = 1$

Recall Definition 2.5.1 of a minimax estimator; that concept is the same in the general case: for any estimator T set

$$M(T) = \sup_{\vartheta \in \Theta} R(T, \vartheta). \tag{5.28}$$

Here more generally sup is written in place of the maximum, since it is not claimed that the supremum is attained. Conceptually, the criterion is again the *worst case risk.* An estimator T_M is called *minimax* if

$$M(T_M) = \min_T M(T).$$

Theorem 5.7.1 *In the Gaussian location model for $\mu \in \Theta = \mathbb{R}$ (Model $\mathbf{M}_2$), the sample mean $\bar{X}_n$ is a minimax estimator for quadratic loss.*

Proof. Suppose that for an estimator T_0 we have

$$M(T_0) < M(\bar{X}_n). \tag{5.29}$$

Above in (5.27) it was shown that $R(\bar{X}_n, \mu) = n^{-1}\sigma^2$ does not depend on μ, hence $M(\bar{X}_n) = n^{-1}\sigma^2$. It follows that there is $\varepsilon > 0$ such that

$$R(T_0, \mu) \le n^{-1}\sigma^2 - \varepsilon \text{ for all } \mu \in \mathbb{R}.$$

Take a prior distribution $N(0, \tau)$ for μ; then for the mixed (integrated) risk $B(T_0)$ of the estimator T_0

$$B(T_0) = E_\tau R(T_0, \mu) \le n^{-1}\sigma^2 - \varepsilon$$

If $T_{B,\tau}$ is the Bayes estimator for $N(0, \tau)$ given by Corollary (5.6.2) then

$$B(T_{B,\tau}) = \rho^2 = \frac{n^{-1}\sigma^2\tau^2}{n^{-1}\sigma^2 + \tau^2} \to n^{-1}\sigma^2 \text{ as } \tau \to \infty.$$

Hence for τ large enough we have

$$B(T_0) < B(T_{B,\tau})$$

which contradicts the fact that $T_{B,\tau}$ is the Bayes estimator. Hence there can be no T_0 with (5.29).
■

It essential for this argument that the parameter space Θ is the whole real line. It turns out (see exercise below) that for e.g. $\Theta = [-K, K]$ we can find an estimator $T_{B,\tau}$ which is uniformly strictly better than $\bar{X}_n$, so that $\bar{X}_n$ is no longer minimax.

Let us consider the case $\Theta = [-K, K]$ in more detail; here we have an a priori restriction $|\mu| \leq K$. It is a more complicated problem to find a minimax estimator here. A common approach is to simplify the problem again, by looking for minimax estimators within a *restricted class of estimators*. For simplicity consider the case of the Gaussian location for $n = 1$, $\sigma^2 = 1$. A **linear estimator** is any estimator

$$T(X) = aX + b$$

where a, b are real numbers. The appropriate worst case risk is again (5.28), for the present parameter space Θ. A **minimax linear estimator** is a linear estimator T_{LM} such that

$$M(T_{LM}) = \sup_{\mu \in \Theta} R(T_{LM}, \mu) \leq \min_{T \text{ linear}} M(T).$$

Exercise 5.7.2 *Consider the Gaussian location model with restricted parameter space $\Theta = [-K, K]$, where $K > 0$, sample size $n = 1$ and $\sigma^2 = 1$. **(i)** Find the minimax linear estimator T_{LM}. **(ii)** Show that T_{LM} is strictly better than the sample mean $\bar{X}_n = X$, everywhere on $\Theta = [-K, K]$ (this implies that X is not admissible). **(iii)** Show that T_{LM} is Bayesian in the unrestricted model $\Theta = R$ for a certain prior distribution $N(0, \tau^2)$, and find the τ^2.*

Chapter 6
THE MULTIVARIATE NORMAL DISTRIBUTION

Recall the Gaussian location model (Model $\mathbf{M}_2$), for sample size $n = 1$. We can represent X as

$$X = \mu + \xi$$

where ξ is centered normal: $\mathcal{L}(\xi) = N(0, \sigma^2)$. In a Bayesian statistical approach, assume that μ becomes random as well: $\mathcal{L}(\mu) = N(0, \tau^2)$, in such a way that it is independent of the "noise" ξ. Thus ξ and μ are independent normal r.v.'s. For Bayesian statistical inference, we computed the joint distribution of X and μ; this is well defined as the distribution of the r.v. $(\mu + \xi, \mu)$. The joint density was

$$p(x, \mu) = \frac{1}{2\pi\sigma\tau} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} - \frac{\mu^2}{2\tau^2} \right). \tag{6.1}$$

Here we started with a pair of independent Gaussian r.v.'s (ξ, μ) and obtained a pair (X, μ). We saw that the marginal of X is $N(0, \sigma^2 + \tau^2)$; the marginal of μ is $N(0, \tau^2)$. We have a joint distribution of (X, μ) in which both marginals are Gaussian, but (X, μ) are not independent: indeed (6.1) is not the product of its marginals. Note that we took a linear transformation of (ξ, μ):

$$\begin{aligned} X &= 1 \cdot \xi + 1 \cdot \mu, \\ \mu &= 0 \cdot \xi + 1 \cdot \mu \end{aligned}$$

We could have written $\xi = \sigma x_1$ and $\mu = \tau x_2$ for independent standard normals x_1,x_2; then the linear transformation is

$$\begin{aligned} X &= \sigma \cdot x_1 + \tau \cdot x_2, \\ \mu &= 0 \cdot x_1 + \tau \cdot x_2. \end{aligned} \tag{6.2}$$

Let us consider a more general situation. We have independent standard normals x_1,x_2; and consider

$$\begin{aligned} y_1 &= m_{11} x_1 + m_{12} x_2, \\ y_2 &= m_{21} x_1 + m_{22} x_2. \end{aligned}$$

where the linear transformation is nonsingular: $m_{11} m_{22} - m_{21} m_{12} \neq 0$. Nonsingularity is true for (6.2): it means just $\sigma\tau > 0$. Define a matrix

$$M = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix}$$

and vectors

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}.$$

Then nonsingularity means $|M| \neq 0$ (where $|M|$ is the determinant of M). Let us find the joint distribution of y_1, y_2, i.e. the law of the vector $\mathbf{y}$. For this we immediately proceed to the k-dimensional case, i.e. consider i.i..d standard normal $x_1, \ldots, x_k$, and let

$$\mathbf{x} = (x_1, \ldots, x_k)^\top,$$

M be a nonsingular $k \times k$ matrix and $\mathbf{y} = M\mathbf{x}$.
Let A be a measurable set in $\mathbb{R}^k$, i.e. a set which has a well defined volume (finite or infinite). Then (φ is the standard normal density))

$$\begin{aligned} P(\mathbf{y} \in A) = P(M\mathbf{x} \in A) &= \int_{M\mathbf{x}\in A} \left(\prod_{i=1}^{k} \varphi(x_i) \right) dx_1 \ldots dx_k \\ &= \frac{1}{(2\pi)^{k/2}} \int_{M\mathbf{x}\in A} \exp\left(-\frac{1}{2} \sum_{i=1}^{k} x_i^2 \right) dx_1 \ldots dx_k \\ &= \frac{1}{(2\pi)^{k/2}} \int_{M\mathbf{x}\in A} \exp\left(-\frac{1}{2} \mathbf{x}^\top \mathbf{x} \right) d\mathbf{x}. \end{aligned}$$

where $\mathbf{x}^\top$is the transpose of $\mathbf{x}$, $\mathbf{x}^\top\mathbf{x}$ is the inner product (scalar product) and $\mathbf{x}^\top\mathbf{x} = \sum_{i=1}^{k} x_i^2$. Now let M^{-1} be the inverse of M and write

$$\mathbf{x} = M^{-1}M\mathbf{x},\ \mathbf{x}^\top\mathbf{x} = (M\mathbf{x})^\top (M^{-1})^\top M^{-1} M\mathbf{x}.$$

Set $\Sigma = MM^\top$; this is a nonsingular symmetric matrix. Recall the following matrix rules:

$$\begin{aligned} (M^{-1})^\top M^{-1} &= \left(MM^\top \right)^{-1} = \Sigma^{-1},\ |\Sigma| = |M| \left| M^\top \right| = |M|^2, \\ \left| M^{-1} \right| &= |M|^{-1}. \end{aligned}$$

We thus obtain

$$\mathbf{x}^\top\mathbf{x} = (M\mathbf{x})^\top \Sigma^{-1} M\mathbf{x},$$

$$P(\mathbf{y} \in A) = \frac{1}{(2\pi)^{k/2}} \int_{M\mathbf{x}\in A} \exp\left(-\frac{1}{2} (M\mathbf{x})^\top \Sigma^{-1} M\mathbf{x} \right) d\mathbf{x}$$

This suggests a multivariate change of variables: setting $\mathbf{y} = M\mathbf{x}$, we have to set $\mathbf{x} = M^{-1}\mathbf{y}$ and (formally)

$$d\mathbf{x} = \left| M^{-1} \right| d\mathbf{y} = |M|^{-1} d\mathbf{y}.$$

Writing $|M| = |\Sigma|^{1/2}$, we obtain

$$P(\mathbf{y} \in A) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \int_{\mathbf{y}\in A} \exp\left(-\frac{1}{2} \mathbf{y}^\top \Sigma^{-1} \mathbf{y} \right) d\mathbf{y}.$$

Definition 6.1.3 *The distribution in $\mathbb{R}^k$ given by the joint density of $y_1, \ldots, y_k$*

$$p(\mathbf{y}) = \frac{1}{(2\pi)^{k/2}\,|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\mathbf{y}^\top \Sigma^{-1}\mathbf{y}\right) \tag{6.3}$$

*for some $\Sigma = MM^\top$, $|\Sigma| \neq 0$ ($\mathbf{y} = (y_1, \ldots, y_k)^\top$), is called the **k-variate normal distribution** $N_k(\mathbf{0}, \Sigma)$.*

Theorem 6.1.4 *Let $x_1, \ldots, x_k$ be i.i..d standard normal random variables, M be a nonsingular $k \times k$ matrix and*

$$\mathbf{x} = (x_1, \ldots, x_k)^\top,\ \mathbf{y} = M\mathbf{x}.$$

Then $\mathbf{y}$ has a distribution $N_k(\mathbf{0}, \Sigma)$, where $\Sigma = MM^\top$.

Clearly all matrices $\Sigma = MM^\top$ with $|M| \neq 0$ are possible here. Let us describe the class of possible matrices Σ differently.

Lemma 6.1.5 *Any matrix $\Sigma = MM^\top$ with $|M| \neq 0$ is* positive definite: *for any $\mathbf{x} \in \mathbb{R}^k$ which is not identically zero.*

$$\mathbf{x}^\top \Sigma \mathbf{x} > 0.$$

Proof. For any $\mathbf{x} \neq \mathbf{0}$ ($\mathbf{0}$ is the null vector) and $\mathbf{z} = M^\top \mathbf{x}$

$$\mathbf{x}^\top \Sigma \mathbf{x} = \mathbf{x}^\top MM^\top \mathbf{x} = (M^\top \mathbf{x})^\top (M^\top \mathbf{x}) = \mathbf{z}^\top \mathbf{z} > 0,$$

since $\mathbf{z}^\top \mathbf{z} = 0$ would mean $\mathbf{z} = \mathbf{0}$ and thus $\mathbf{x} = M^{-1}\mathbf{z} = \mathbf{0}$, which was excluded by assumption. ■

Recall the following basic fact from linear algebra.

Proposition 6.1.6 *(**Spectral decomposition**) Any positive definite $k \times k$-matrix Σ can be written*

$$\Sigma = C^\top \Lambda C$$

where Λ is a diagonal $k \times k$-matrix

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdot & \cdot & 0 \\ 0 & \lambda_2 & 0 & \cdot & \cdot \\ \cdot & 0 & \cdot & 0 & \cdot \\ \cdot & \cdot & 0 & \cdot & 0 \\ 0 & \cdot & \cdot & 0 & \lambda_k \end{pmatrix}$$

*with positive diagonal elements $\lambda_i > 0$, $i = 1, \ldots, k$ (called **eigenvalues** or **spectral values**) and C is an orthogonal $k \times k$-matrix :*

$$C^\top C = CC^\top = I_k$$

(I_k is the unit $k \times k$-matrix). If all the λ_i, $i = 1, \ldots, k$ are different, Λ is unique and C is unique up to sign changes of its row vectors.

Lemma 6.1.7 *Every positive definite $k \times k$-matrix Σ can be written as $MM^\top$ where M is a nonsingular $k \times k$-matrix.*

Proof. It is easy to take a square root $\Lambda^{1/2}$ of a diagonal matrix Λ: let $\Lambda^{1/2}$ be the diagonal matrix with diagonal elements $\lambda_i^{1/2}$; then $\Lambda^{1/2}\Lambda^{1/2} = \Lambda$. Now take $M = C^\top \Lambda^{1/2}$ where C, Λ are from the spectral decomposition:

$$MM^\top = C^\top \Lambda^{1/2}\Lambda^{1/2} C = C^\top \Lambda C = \Sigma$$

and M is nonsingular:

$$|M| = \left|C^\top\right| \left|\Lambda^{1/2}\right| = \left|C^\top\right| \prod_{i=1}^{k} \lambda_i^{1/2},$$

and for any orthogonal matrix C one has $\left|C^\top\right| = \pm 1$ since

$$\left|C^\top\right|^2 = \left|C^\top C\right| = |I_k| = 1.$$

■

As a result, the k-variate normal distribution $N_k(\mathbf{0}, \Sigma)$ is defined for any positive definite matrix Σ.

Recall that for any random vector $\mathbf{y}$ in $\mathbb{R}^k$, the covariance matrix is defined by

$$(\mathrm{Cov}(\mathbf{y}))_{i,j} = \mathrm{Cov}(y_i, y_j) = E(y_i - Ey_i)(y_j - Ey_j).$$

if the expectations exist. (Here $(A)_{i,j}$ is the $(i.j)$ entry of a matrix A). This existence is guaranteed by a condition $\sum_{i=1}^k Ey_i^2 < \infty$ (via the Cauchy-Schwarz inequality).

Lemma 6.1.8 *The law $N_k(\mathbf{0}, \Sigma)$ has expectation $\mathbf{0}$ and covariance matrix Σ.*

Proof. In the present case, we have $\mathbf{y} = M\mathbf{x}$, hence

$$y_i = \sum_{r=1}^{k} m_{ir} x_r,$$

which immediately shows $Ey_i = 0$. Furthermore,

$$\begin{aligned} \mathrm{Cov}(y_i, y_j) &= Ey_i y_j = E\left(\sum_{r=1}^{k} m_{ir} x_r\right)\left(\sum_{s=1}^{k} m_{js} x_s\right) \\ &= E\left(\sum_{r,s=1}^{k} m_{ir} m_{js} x_r x_s\right). \end{aligned}$$

Since

$$Ex_r x_s = \begin{cases} 1 \text{ if } r = s \\ 0 \text{ otherwise} \end{cases}$$

we obtain

$$\mathrm{Cov}(y_i, y_j) = \sum_{r=1}^{k} m_{ir} m_{jr} = \left(MM^\top\right)_{i,j} = (\Sigma)_{i,j}.$$

■

For the matrix Σ one commonly writes

$$(\Sigma)_{i,j} = \sigma_{ij}, (\Sigma)_{i,i} = \sigma_{ii} = \sigma_i^2.$$

Definition 6.1.9 *Let* $\mathbf{x}$ *be a random vector with distribution* $N_k(\mathbf{0}, \Sigma)$ *where* Σ *is a positive definite matrix, and let* $\boldsymbol{\mu}$ *be a (nonrandom) vector from* $\mathbb{R}^k$*. The distribution of the random vector*

$$\mathbf{y} = \mathbf{x} + \boldsymbol{\mu}$$

is called the k-variate normal distribution $N_k(\boldsymbol{\mu}, \Sigma)$*.*

The following result is obvious.

Lemma 6.1.10 *(i) The law* $N_k(\boldsymbol{\mu}, \Sigma)$ *has expectation* $\boldsymbol{\mu}$ *and covariance matrix* Σ*.*
(ii) The density is

$$\varphi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{y}) := \frac{1}{(2\pi)^{k/2}\,|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})\right).$$

Lemma 6.1.11 $\mathcal{L}(\mathbf{x}) = N_k(\mathbf{0}, I_k)$ *if and only if* $\mathbf{x}$ *is a vector of i.i.d. standard normals.*

Proof. In this case the joint density is

$$\varphi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x}) = \varphi_{\mathbf{0},I_k}(\mathbf{x}) = \prod_{i=1}^{k} \varphi(x_i)$$

which proves that the components x_i are i.i.d. standard normal. ▲

Lemma 6.1.12 *Let* $\mathcal{L}(\mathbf{x}) = N_k(\boldsymbol{\mu}, \Sigma)$ *where* Σ *is positive definite, and let* A *be a (nonrandom)* $l \times k$ *matrix with rank* l *(this implies* $l \le k$*). Then*

$$\mathcal{L}(A\mathbf{x}) = N_l(A\boldsymbol{\mu}, A\Sigma A^\top)$$

Proof. It suffices to show

$$\mathcal{L}(A\mathbf{x} - \mathbf{A}\boldsymbol{\mu}) = \mathcal{L}(\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})) = N_l(\mathbf{0}, A\Sigma A^\top),$$

so we can assume $\boldsymbol{\mu} = \mathbf{0}$. Let also $\mathbf{x} = M\boldsymbol{\xi}$ where $\mathcal{L}(\boldsymbol{\xi}) = N_k(\mathbf{0}, I_k)$ and $\Sigma = MM^\top$. Then $A\mathbf{x} = AM\boldsymbol{\xi}$ so for $l = k$ the claim is immediate. In general, for $l \le k$, consider also the $l \times l$-matrix $A\Sigma A^\top$; note that it is positive definite:

$$\mathbf{a}^\top A\Sigma A^\top \mathbf{a} > 0$$

for every nonzero l-vector $\mathbf{a}$ since $A^\top\mathbf{a}$ is then a nonzero k-vector (A has rank l). Let

$$A\Sigma A^\top = C^\top \Lambda C$$

be a spectral decomposition. Define $D = \Lambda^{-1/2}C$; then

$$\begin{aligned} A\mathbf{x} &= AM\boldsymbol{\xi}, \\ DA\mathbf{x} &= DAM\boldsymbol{\xi}. \end{aligned}$$

Suppose first that $l = k$. Then $DA\mathbf{x}$ is multivariate normal with expectation 0 and covariance matrix

$$\begin{aligned} DAM\,(DAM)^\top &= DAMM^\top A^\top D^\top \\ &= DA\Sigma A^\top D^\top = DC^\top \Lambda C D^\top \\ &= \Lambda^{-1/2} C \left(C^\top \Lambda C\right) C^\top \Lambda^{-1/2} \\ &= \Lambda^{-1/2}\Lambda\Lambda^{-1/2} = I_l. \end{aligned}$$

Hence

$$\mathcal{L}(DA\mathbf{x}) = N_l(\mathbf{0}, I_l),$$

i.e. $DA\mathbf{x}$ is a vector of standard normals.
If $l < k$ then we find a $(k-l) \times k$ matrix F such that

$$DAMF^\top = \mathbf{0},\ FF^\top = I_{k-l}.$$

For this it suffices to select the rows of F as $k-l$ orthonormal vectors which are a basis of the subspace of $\mathbb{R}^k$ which is orthogonal to the subspace spanned by the rows of DAM. Then for the $k \times k$ matrix

$$F_0 = \begin{pmatrix} DAM \\ F \end{pmatrix}$$

we have

$$F_0 F_0^\top = \begin{pmatrix} I_l & \mathbf{0} \\ \mathbf{0} & I_{k-l} \end{pmatrix} = I_k,$$

i.e. F_0 is orthogonal. Hence $F_0\boldsymbol{\xi}$ is multivariate normal $N_k(\mathbf{0}, I_k)$, i.e. a vector of independent standard normals. Since $DAM\boldsymbol{\xi}$ consists of the first l elements of $F_0\boldsymbol{\xi}$, it is a vector of l standard normals. We have shown again that $DA\mathbf{x}$ is a vector of standard normals.
Note that $D^{-1} = C^\top \Lambda^{1/2}$, since

$$DC^\top \Lambda^{1/2} = \Lambda^{-1/2} C C^\top \Lambda^{1/2} = \Lambda^{-1/2}\Lambda^{1/2} = I_l.$$

Hence for $A\mathbf{x} = D^{-1}DA\mathbf{x}$

$$\begin{aligned} \mathcal{L}(A\mathbf{x}) &= N_l(\mathbf{0}, D^{-1}(D^{-1})^\top) = \\ &= N_l(\mathbf{0}, C^\top \Lambda^{1/2}\Lambda^{1/2} C) = N_l(\mathbf{0}, A\Sigma A^\top). \end{aligned}$$

■

Two random vectors $\mathbf{x}$, $\mathbf{y}$ with dimensions k, l respectively, are said to have a *joint normal distribution* if the vector

$$\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}$$

has a $k+l$-multivariate normal distribution. The *covariance matrix* $\mathrm{Cov}(\mathbf{x},\mathbf{y})$ of $\mathbf{x}$, $\mathbf{y}$ is defined by

$$(\mathrm{Cov}(\mathbf{x},\mathbf{y}))_{i,j} = (\mathrm{Cov}(x_i y_j))\,,\ i = 1,\ldots,k,\ j = 1,\ldots,l;$$

it is a $k \times l$-matrix. We then have a block structure for the joint covariance matrix $\mathrm{Cov}(\mathbf{z})$:

$$\mathrm{Cov}(\mathbf{z}) = \begin{pmatrix} \mathrm{Cov}(\mathbf{x}) & \mathrm{Cov}(\mathbf{x},\mathbf{y}) \\ \mathrm{Cov}(\mathbf{y},\mathbf{x}) & \mathrm{Cov}(\mathbf{y}) \end{pmatrix}.$$

Theorem 6.1.13 *Two random vectors* $\mathbf{x}$, $\mathbf{y}$ *with joint normal distribution are independent if and only if they are uncorrelated, i.e. if*

$$\mathrm{Cov}(\mathbf{x},\mathbf{y}) = \mathbf{0} \tag{6.4}$$

(where $\mathbf{0}$ *stands for the null matrix).*

Proof. Independence here means that the joint density is the product of its marginals. Assume that both $\mathbf{x}$,$\mathbf{y}$, are centered, i.e. have zero expectation (otherwise expectation can be subtracted, with (6.4) still true). Write $\Sigma = \mathrm{Cov}(\mathbf{z})$, $\Sigma_{11} = \mathrm{Cov}(\mathbf{x})$, $\Sigma_{22} = \mathrm{Cov}(\mathbf{y})$, $\Sigma_{12} = \mathrm{Cov}(\mathbf{x},\mathbf{y})$, $\Sigma_{21} = \Sigma_{12}^\top$. Then (6.4) means that

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \mathbf{0} \\ \mathbf{0} & \Sigma_{22} \end{pmatrix}.$$

where $\mathbf{0}$ represents null matrices of appropriate dimensions. A matrix of such a structure is called **block diagonal.** It is easy to see that

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \Sigma_{22}^{-1} \end{pmatrix}.$$

Indeed, both inverses Σ_{11}^{-1}, Σ_{22}^{-1} exist, since for the determinants

$$|\Sigma| = |\Sigma_{11}|\,|\Sigma_{22}| \tag{6.5}$$

and the matrix multiplication rules show

$$\Sigma^{-1}\Sigma = \begin{pmatrix} \Sigma_{11}^{-1}\Sigma_{11} & \mathbf{0} \\ \mathbf{0} & \Sigma_{22}^{-1}\Sigma_{22} \end{pmatrix} = I_{k+l}.$$

For the joint density of $\mathbf{x},\mathbf{y}$ (density of $\mathbf{z}^\top = (\mathbf{x}^\top,\mathbf{y}^\top)$) we obtain

$$\begin{aligned} p(\mathbf{x},\mathbf{y}) &= \varphi_{0,\mathbf{\Sigma}}(\mathbf{z}) = \\ &= \frac{1}{(2\pi)^{(k+l)/2}\,|\Sigma_{11}|^{1/2}\,|\Sigma_{22}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}^\top\Sigma_{11}^{-1}\mathbf{x} + \mathbf{y}^\top\Sigma_{22}^{-1}\mathbf{y})\right) \\ &= \varphi_{0,\mathbf{\Sigma}_{11}}(\mathbf{x}) \cdot \varphi_{0,\mathbf{\Sigma}_{22}}(\mathbf{y}). \end{aligned}$$

This immediately implies that the marginal of $\mathbf{x}$ is

$$p(\mathbf{x}) = \int p(\mathbf{x},\mathbf{y})d\mathbf{y} = \varphi_{0,\mathbf{\Sigma}_{11}}(\mathbf{x})$$

and the marginal of $\mathbf{y}$ is $\varphi_{0,\mathbf{\Sigma}_{22}}(\mathbf{y})$. Thus $\varphi_{0,\mathbf{\Sigma}}(\mathbf{z})$ is the product of its marginals.
Conversely, assume that $\varphi_{0,\mathbf{\Sigma}}(\mathbf{z})$ is the product of its marginals, i.e. $\mathbf{x},\mathbf{y}$ are independent. This implies that for any real valued functions $f(\mathbf{x})$, $g(\mathbf{y})$ of the two vectors we have $Ef(\mathbf{x})g(\mathbf{y}) = Ef(\mathbf{x})Eg(\mathbf{y})$. Let $e_{i(k)}$ be the i-th unit vector in $\mathbb{R}^k$, so $e_{i(k)}^\top\mathbf{x} = x_i$; then for the covariance matrix we obtain

$$\begin{aligned} \mathrm{Cov}(\mathbf{x},\mathbf{y})_{i,j} &= Ex_iy_j = E\left(e_{i(k)}^\top\mathbf{x}\cdot e_{j(l)}^\top\mathbf{y}\right) \\ &= Ee_{i(k)}^\top\mathbf{x}\cdot Ee_{j(l)}^\top\mathbf{y} = 0 \end{aligned}$$

since both $\mathbf{x},\mathbf{y}$ are centered. ■

Exercise. Prove (6.5) for a block diagonal Σ, using the spectral decomposition for each of the blocks Σ_{11}, Σ_{22} and the rule $|AB| = |A||B|$.

Chapter 7
THE GAUSSIAN LOCATION-SCALE MODEL

Recall the Gaussian location-scale model which was already introduced (see page 32). :

Model $\mathbf{M}_3$ Observed are n independent and identically random variables $X_1, \ldots, X_n$, each having law $N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are both unknown..

The parameter is this model is two dimensional: $\vartheta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$.

7.1 Confidence intervals

To illustrate the consequences of an unknown variance σ^2, let us look at the problem of constructing a **confidence interval,** beginning with the location model, model M_2. Suppose the confidence level $1 - \alpha$ is given (e.g. $\alpha = 0.05$ or $\alpha = 0.01$), and try to construct a random interval $(\hat{\mu}_-, \hat{\mu}_+)$:

$$P_\mu([\hat{\mu}_-, \hat{\mu}_+] \ni \mu) \geq 1 - \alpha. \tag{7.1}$$

meaning that *the probability that the interval* $[\hat{\mu}_-, \hat{\mu}_+]$ *covers* μ *is more than* 95%', Note both $\hat{\mu}_-, \hat{\mu}_+$ are random variables (functions of the data), so the interval is in fact a random interval. Therefore the element sign $\in$ is written in reverse form $\ni$ to stress the fact that (7.1) the interval is random, not μ (μ is merely unknown)
When σ^2 is known it is easy to build a confidence interval based on the sample mean: since

$$\mathcal{L}(\bar{X}_n) = N(\mu, n^{-1}\sigma^2),$$

it follows that

$$\mathcal{L}(\left(\bar{X}_n - \mu\right) n^{1/2}/\sigma) = N(0, 1). \tag{7.2}$$

Definition 7.1.1 *Suppose that P is a continuous law on $\mathbb{R}$ with density p such that $p(x) > 0$ for $x > 0$ and $P((0, \infty)) \geq 1/2$. For every $\alpha \in (0, 1/2)$, the uniquely defined number $z_a > 0$ fulfilling*

$$\int_{z_\alpha}^{\infty} p(x)dx = \alpha$$

is called the ***(upper) α-quantile*** *of the distribution P.*

The word "upper" is usually omitted for the standard normal distribution, since it is symmetric around 0. In our case it follows immediately from (7.2) that

$$P_\mu\left(\left|\left(\bar{X}_n - \mu\right) n^{1/2}/\sigma\right| \leq z_{\alpha/2}\right) = 1 - \alpha,$$

hence

$$[\hat{\mu}_-, \hat{\mu}_+] = [\bar{X}_n - \sigma z_{\alpha/2}/\sqrt{n}, \bar{X}_n + \sigma z_{\alpha/2}/\sqrt{n}] \tag{7.3}$$

is a $1-\alpha$-confidence interval for μ.

Obviously we have to know the variance σ^2 for this confidence interval, so the procedure breaks down for the location-scale model. In Proposition 3.0.5, page 32 we already encountered the sample variance:

$$S_n^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = n^{-1}\sum_{i=1}^n X_i^2 - (\bar{X}_n)^2.$$

This appears as a reasonable estimate to substitute for the unknown σ^2, for various reasons. First, S_n^2 is the variance of the **empirical distribution** $\hat{P}_n$**:** when $x_1, \ldots, x_n$ are observed, the empirical distribution is a discrete law which assigns probability n^{-1} to each point x_i. From the point of view that $x_1, \ldots, x_n$ should be identified with the random variables $X_1, \ldots, X_n$, this is a *random probability distribution*, with distribution function

$$\hat{F}_n(t) = n^{-1}\sum_{i=1}^n 1_{(-\infty,t]}(X_i)$$

This is the **empirical distribution function, (**e.d.f.**).** Obviously if Z is a random variable with law $\hat{P}_n$ then (assuming $x_1, \ldots, x_n$ fixed)

$$\begin{aligned} \bar{x}_n &= \sum_{i=1}^n n^{-1} x_i = EZ \\ s_n^2 &= \sum_{i=1}^n n^{-1} x_i^2 - (\bar{x}_n)^2 = \mathrm{Var}(Z). \end{aligned}$$

Analogously to the sample mean, we write S_n^2 for s_n^2 when this is construed as a random variable. For the expectation of S_n^2 we obtain (when ξ_i are ind. standard normals)

$$\begin{aligned} ES_n^2 &= En^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = En^{-1}\sum_{i=1}^n (X_i - \mu - (\bar{X}_n - \mu))^2 \\ &= En^{-1}\sum_{i=1}^n (\sigma\xi_i - (\sigma\bar{\xi}_n))^2 = \sigma^2 E\left(n^{-1}\sum_{i=1}^n \xi_i^2 - (\bar{\xi}_n)^2\right) \end{aligned}$$

Since $E\xi_i^2 = 1$ and $\mathcal{L}(\bar{\xi}_n) = N(0, n^{-1})$, we have $E(\bar{\xi}_n)^2 = n^{-1}$, and thus

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

Thus S_n^2 is not unbiased, but an unbiased estimator is

$$\hat{S}_n^2 = \frac{n}{n-1}S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Suppose we plug in $\hat{S}_n^2$ for σ^2 in the confidence interval (7.1). We cannot expect this to be an α-confidence interval, since it is based on the standard normal law for $\left(\bar{X}_n - \mu\right) n^{1/2}/\sigma$ via (7.2), and the quantity

$$T_\mu = T_\mu(X) = \frac{\left(\bar{X}_n - \mu\right) n^{1/2}}{\hat{S}_n}$$

cannot be expected to have a normal law. (Here $\hat{S}_n = \sqrt{\hat{S}_n^2}$). However we can hope to identify this distribution, and then base a confidence interval upon it (by taking a quantile).
Note that T_μ is not a statistic, since it depends on the unknown parameter μ. However, reverting again to the representation $X_i - \mu = \sigma\xi_i$, we obtain

$$T_\mu(X) = \frac{\sigma\bar{\xi}_n n^{1/2}}{\left(\frac{1}{n-1}\sum_{i=1}^n \sigma^2(\xi_i - \bar{\xi}_n)^2\right)^{1/2}} = \frac{\bar{\xi}_n n^{1/2}}{\left(\frac{1}{n-1}\sum_{i=1}^n (\xi_i - \bar{\xi}_n)^2\right)^{1/2}}.$$

We see that the distribution of T_μ does not depend on the parameters μ and σ^2; it depends only on the sample size n (i..e the number of i.i.d standard normals ξ_i involved).

7.2 Chi-square and t-distributions

Definition 7.2.1 *Let $X_1, \ldots, X_n$ be independent $N(0,1)$. The distribution of the statistic*

$$T = T(X) = \frac{\bar{X}_n n^{1/2}}{\hat{S}_n}$$

is called the ***t-distribution*** *with $n-1$ degrees of freedom (denoted t_{n-1}). The statistic T is called the* ***t-statistic.***

Lemma 7.2.2 *The t-distribution is symmetric around 0, i.e. for $x \geq 0$*

$$P(T \geq x) = P(T \leq -x)$$

Proof. This follows immediately form the fact that $\mathcal{L}(-\xi_i) = N(0,1) = \mathcal{L}(\xi_i)$. ■
This suggest that a confidence interval based on the t-distribution can be built in the same fashion as for the standard normal, i.e. taking an upper quantile.
It remains to find the actual form of the t-distribution, in order to compute its quantiles. The following result prepares this derivation.

Theorem 7.2.3 *Let $X_1, \ldots, X_n$ be n i.i.d. r.v.'s each having law $N(\mu, \sigma^2)$.*
(i) *The sample mean $\bar{X}_n$ and the sample variance S_n^2 are independent random variables.*
(ii) *The bias corrected sample variance $\hat{S}_n^2 = nS_n^2/(n-1)$ can be represented as*

$$\hat{S}_n^2 = \frac{\sigma^2}{n-1}\sum_{i=1}^{n-1} \xi_i^2$$

where $\xi_1, \ldots, \xi_{n-1}$ are i.i.d standard normals, independent of $\bar{X}_n$.

Proof. Let X be the vector $X = (X_1, \ldots, X_n)^\top$; then X has a multivariate normal distribution with covariance matrix $\sigma^2 I_n$. To describe the expectation vector, let $\mathbf{1}_n = (1, \ldots, 1)^\top$ be the vector in $\mathbb{R}^n$ consisting of 1's. Then

$$\mathcal{L}(X) = N_n(\mu\mathbf{1}_n, \sigma^2 I_n).$$

Consider the linear subspace of $\mathbb{R}^n$ of dimension $n-1$ which is orthogonal to $\mathbf{1}_n$. Let $b_1, \ldots, b_{n-1}$ be an orthonormal basis of this subspace and $b_n = n^{-1/2}\mathbf{1}_n$. Then b_n has also norm 1 ($\|b_n\| = 1$), the whole set $b_1, \ldots, b_n$ is an orthonormal basis of $\mathbb{R}^n$ and the $n \times n$-matrix

$$B = \begin{pmatrix} b_1^\top \\ b_2^\top \\ \ldots \\ b_n^\top \end{pmatrix}$$

is orthogonal. Define $Y = BX$; then

$$\mathcal{L}(Y) = N_n(\mu B\mathbf{1}_n, \sigma^2 I_n)$$

and the components $Y_1, \ldots, Y_n$ are independent. (Indeed when the covariance matrix of a multivariate normal is of form $\sigma^2 I_n$, or more generally a diagonal matrix, then the joint density of y_i is the product of its marginals). Moreover

$$\begin{aligned} Y_n &= b_n^\top X = n^{-1/2}\mathbf{1}_n^\top X = n^{1/2}\bar{X}_n, \\ EY_j &= Eb_j^\top X = \mu b_j^\top \mathbf{1}_n = 0,\ j = 1, \ldots, n-1. \end{aligned} \tag{7.4}$$

It follows that $Y_1, \ldots, Y_{n-1}$ are independent $N(0, \sigma^2)$ random variables and are independent of $\bar{X}_n$. Now

$$\sum_{i=1}^n Y_i^2 = \|Y\|^2 = X^\top B^\top BX = X^\top X = \sum_{i=1}^n X_i^2.$$

On the other hand, in view of (7.4)

$$\sum_{i=1}^n Y_i^2 = \sum_{i=1}^{n-1} Y_i^2 + n\bar{X}_n^2,$$

so that

$$\sum_{i=1}^{n-1} Y_i^2 = \sum_{i=1}^n X_i^2 - n\bar{X}_n^2 = nS_n^2.$$

This shows that S_n^2 is a function of $Y_1, \ldots, Y_{n-1}$ and hence independent of $\bar{X}_n$, which establishes (i). Dividing both sides in the last display by $n-1$ and setting $\xi_i = Y_i/\sigma$ establishes (ii). ■

Definition 7.2.4 *Let $X_1, \ldots, X_n$ be independent $N(0,1)$. The distribution of the statistic*

$$\chi^2 = \chi^2(X) = \sum_{i=1}^n X_i^2$$

*is called the **chi-square distribution** with n degrees of freedom, denoted χ_n^2.*

To find the form of the t-distribution, we need to find the density of the ratio of a normal and the square root of an independent χ^2-variable. We begin with deriving the density of the χ^2-distribution. The following lemma immediately follows from the above definition.

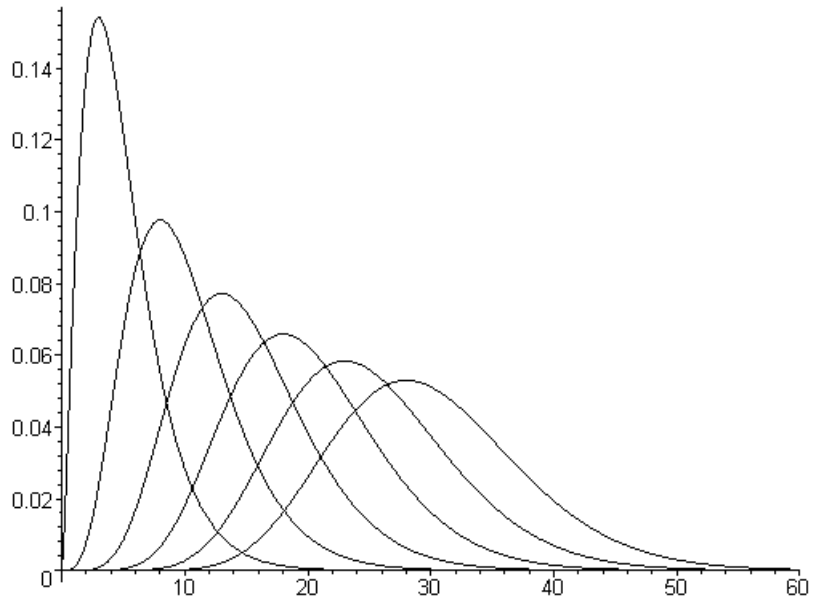


Figure 1 The densities of χ^2_n for $n = 5, 10, 15, \ldots, 30$

Lemma 7.2.5 *Let Y_1, Y_2 be independent r.v's with laws*

$$\mathcal{L}(Y_1) = \chi^2_k,\ \mathcal{L}(Y_2) = \chi^2_l,\ k, l \geq 1.$$

Then

$$\mathcal{L}(Y_1 + Y_2) = \chi^2_{k+l}.$$

Proposition 7.2.6 *The density of the law χ^2_n is*

$$f_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} \exp(-x/2),\ x \geq 0.$$

Comment. In Section 5.3 a family of Gamma densities was introduced as

$$f_\alpha(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} \exp(-x)$$

for $\alpha > 0$. Define more generally, for some $\gamma > 0$

$$f_{\alpha,\gamma}(x) = \frac{1}{\Gamma(\alpha)\gamma^\alpha} x^{\alpha-1} \exp(-x\gamma^{-1}).$$

The corresponding law is called the $\Gamma(\alpha, \gamma)$-distribution . The chi-square distribution χ^2_n thus coincides with the distribution $\Gamma(n/2, 2)$. The figure shows the χ^2_n-densities for 6 values of n ($n = 5, 10, \ldots, , 30$). From the definition it is clear that χ^2_n has expectation n.

Proof of the Proposition. Recall that for $\alpha > 0$

$$\begin{aligned} \Gamma(\alpha) &= \int_0^\infty x^{\alpha-1} \exp(-x) dx \\ \Gamma(\alpha+1) &= \alpha\Gamma(\alpha). \end{aligned}$$

Our proof will proceed by induction. Start with χ^2_1: let X_1 be $N(0,1)$; then

$$P\left(X_1^2 \leq t\right) = P\left(-t^{1/2} \leq X_1 \leq t^{1/2}\right) = 2\int_0^{t^{1/2}} \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{z^2}{2}\right) dz$$

A change of variable $x = z^2$, $dz = (1/2x^{1/2})dx$ gives

$$P\left(X_1^2 \leq t\right) = \int_0^t \frac{1}{(2\pi x)^{1/2}} \exp\left(-\frac{x}{2}\right) dx$$

and we obtain

$$f_1(x) = \frac{1}{2^{1/2}\pi^{1/2}} x^{1/2-1} \exp(-x/2),\ x \geq 0.$$

Now

$$\Gamma(1/2) = \pi^{1/2} \tag{7.5}$$

follows from the fact that f_1 integrates to one and the definition of the gamma-function. We obtained the density of χ^2_1 as claimed.
For the induction step, we assume that $\mathcal{L}(Y_1) = \chi^2_n$ and $\mathcal{L}(Y_2) = \chi^2_1$. By the previous lemma, f_{n+1} is the *convolution* of f_n and f_1: assuming the densities zero for negative argument, we obtain

$$f_{n+1}(x) = \int_0^\infty f_n(y) f_1(x-y) dy$$

(Indeed the convolution of densities is the operation applied to densities of two independent r.v.'s for obtaining the density of the sum). Hence

$$\begin{aligned} f_{n+1}(x) &= \int_0^x \frac{1}{2^{n/2}\Gamma(n/2)} y^{n/2-1} \exp(-y/2) \frac{1}{2^{1/2}\Gamma(1/2)} (x-y)^{-1/2} \exp(-(x-y)/2) dy \\ &= \frac{1}{2^{(n+1)/2}\Gamma(n/2)\Gamma(1/2)} \exp(-x/2) \int_0^x y^{n/2-1} (x-y)^{-1/2} dy \end{aligned}$$

To compute the integral, we note (with a change of variables $u = y/x$)

$$\begin{aligned} \int_0^x y^{n/2-1}(x-y)^{-1/2} dy &= x \cdot x^{n/2-1} \cdot x^{-1/2} \int_0^x \left(\frac{y}{x}\right)^{n/2-1} \left(1-\frac{y}{x}\right)^{-1/2} \frac{1}{x} dy \\ &= x^{(n+1)/2-1} \int_0^1 u^{n/2-1} (1-u)^{-1/2} du \\ &= x^{(n+1)/2-1} B(n/2, 1/2) \end{aligned}$$

where $B(n/2, 1/2)$ is the Beta integral:

$$B(\alpha,\beta) = \int_0^1 u^{\alpha-1} (1-u)^{\beta-1} du = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

(cf. Section 5.3 on the Beta densities). Thus, collecting results, we get

$$\begin{aligned} f_{n+1}(x) &= \frac{1}{2^{(n+1)/2}\Gamma(n/2)\Gamma(1/2)} \exp(-x/2) x^{(n+1)/2-1} \frac{\Gamma(n/2)\Gamma(1/2)}{\Gamma((n+1)/2)} \\ &= \frac{1}{2^{(n+1)/2}\Gamma((n+1)/2)} x^{(n+1)/2-1} \exp(-x/2) \end{aligned}$$

which is the form of $f_{n+1}(x)$ claimed. ■

Proposition 7.2.7 ***(i)*** *The law t_n is the distribution of*

$$\frac{n^{1/2}Z_1}{\sqrt{Z_2}}$$

where Z_1, Z_2 are independent r.v.'s with standard normal and χ^2_n-distribution, respectively. ***(ii)*** *The density of the law t_n is*

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

Proof. (i) follows immediately from Definition 7.2.1 and Theorem 7.2.3: for a r..v T with t_{n-1}-distribution we obtain, if ξ_i are defined as in the proof of Theorem 7.2.3

$$\begin{aligned} T &= \frac{\bar{X}_n n^{1/2}}{\hat{S}_n} = \frac{\xi_n}{\left(\frac{1}{n-1}\sum_{i=1}^{n-1}\xi_i^2\right)^{1/2}} \\ &= \frac{(n-1)^{1/2}\xi_n}{\left(\sum_{i=1}^{n-1}\xi_i^2\right)^{1/2}} = \frac{(n-1)^{1/2}Z_1}{Z_2^{1/2}}. \end{aligned}$$

To prove (ii), we note that the law t_n is symmetric, so we can proceed via the law of the squared variable nZ_1^2/Z_1. Here $\mathcal{L}(Z_1^2) = \chi^2_1$, and the joint density of Z_1^2, Z_2 is

$$g(t,u) = \frac{1}{2^{1/2}\Gamma(1/2)}t^{1/2-1}\exp(-t/2)\frac{1}{2^{n/2}\Gamma(n/2)}u^{n/2-1}\exp(-u/2),\ t,u \geq 0.$$

Consequently

$$P\left(\frac{Z_1^2}{Z_2} \leq x\right) = \int_{t/u\leq x, u\geq 0, t\geq 0}\frac{1}{2^{(n+1)/2}\Gamma(1/2)\Gamma(n/2)}u^{n/2-1}t^{-1/2}\exp(-(t+u)/2)\,dt\,du$$

Substitute t by $v = t/u$, then $dt = udv$ and the above integral is

$$\begin{aligned} &\int_{0\leq v\leq x, u\geq 0}\frac{1}{2^{(n+1)/2}\Gamma(1/2)\Gamma(n/2)}u^{n/2-1}(vu)^{-1/2}\exp(-(vu+u)/2)\,udv\,du \\ = &\int_{0\leq v\leq x, u\geq 0}\frac{1}{2^{(n+1)/2}\Gamma(1/2)\Gamma(n/2)}u^{(n+1)/2-1}v^{-1/2}\exp(-((v+1)u)/2)\,dv\,du \end{aligned}$$

Another substitution of u by $z = (v+1)u$ makes this

$$\int_{0\leq v\leq x}\int_{z\geq 0}\frac{1}{2^{(n+1)/2}\Gamma(1/2)\Gamma(n/2)}\frac{1}{(v+1)^{(n+1)/2-1}}z^{(n+1)/2-1}v^{-1/2}\exp(-z/2)\,\frac{1}{v+1}dz\,dv.$$

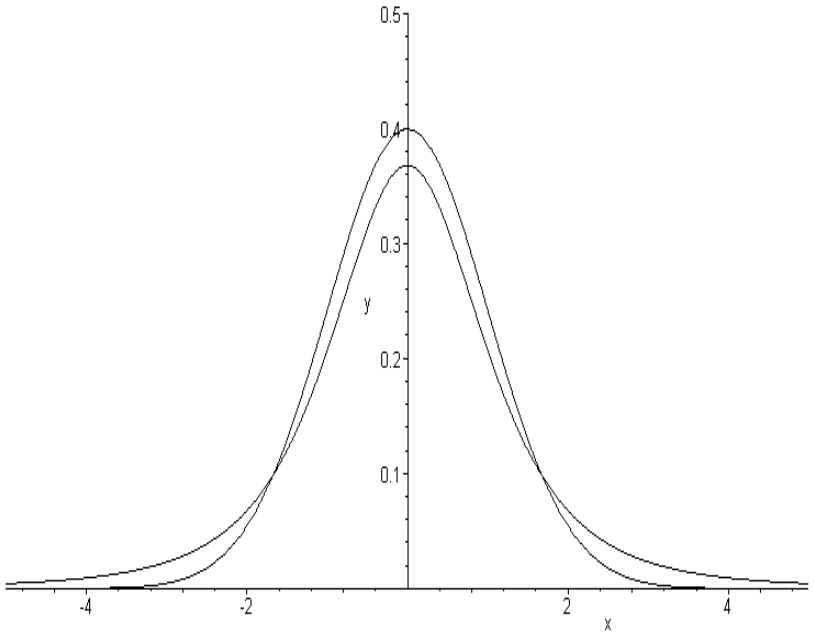
Here the terms depending on z together with appropriate constants (when we also divide and multiply by $\Gamma((n+1)/2)$) form the density of χ^2_{n+1}, so that the expression becomes, after integrating out z,

$$\int_{0\leq v\leq x}\frac{\Gamma((n+1)/2)}{\Gamma(1/2)\Gamma(n/2)}\frac{v^{-1/2}}{(v+1)^{(n+1)/2}}\,dv = P\left(\frac{Z_1^2}{Z_2}\leq x\right).$$

A final change of variables $s = (nv)^{1/2}$, $dv = n^{-1}2sds$ gives

$$\begin{aligned} P\left(0 \leq \frac{n^{1/2}|Z_1|}{Z_2^{1/2}} \leq a\right) &= 2\int_{0\leq s\leq a} \frac{\Gamma((n+1)/2)}{(n\pi)^{1/2}\Gamma(n/2)} \left(1+\frac{s^2}{n}\right)^{-(n+1)/2} ds \\ &= 2P\left(0 \leq \frac{n^{1/2}Z_1}{Z_2^{1/2}} \leq a\right). \end{aligned}$$

Now differentiating w.r. to a gives the form of the density (ii). ■
Let us visualize the forms of the normal and the t-distribution.



The t-density with 3 degrees of freedom against the standard normal (dotted)

Clearly the t-distribution has heavier tails, which means that the quantiles (now called $t_{\alpha/2}$ in place of $z_{\alpha/2}$) are farther out and the confidence interval is wider. A narrower confidence interval, for the same level α, is preferable. Thus the absence of knowledge of the sample variance σ^2 is reflected in less sharp confidence statements.

Theorem 7.2.8 *Let $t_{\alpha/2}$ be the upper $\alpha/2$-quantile of the t-distribution with $n-1$ degrees of freedom. Then in Model $\mathbf{M}_3$, the interval*

$$[\hat{\mu}_-, \hat{\mu}_+] = [\bar{X}_n - \hat{S}_n t_{\alpha/2}/\sqrt{n}, \bar{X}_n + \hat{S}_n t_{\alpha/2}/\sqrt{n}]$$

is a $1-\alpha$-confidence interval for the unknown expected value μ.

The t-distribution has been found by Gosset ("The probable error of a mean", Biometrika 6, 1908), who wrote under the pseudonym "Student". The distribution is frequently called "Student's t". A notion of "**studentization**" has been derived from it: in the Gaussian location model $\mathbf{M}_2$, the statistic

$$Z = \frac{\bar{X}_n n^{1/2}}{\sigma}$$

is sometimes called the Z-statistic. It is standard normal when $\mu = 0$ and can therefore be used to test the hypothesis $\mu = 0$ (for the theory if testing hypotheses cf. later sections). In the location-scale model, σ is not known, and by substituting $\hat{S}_n$ one forms the t-statistic

$$T = \frac{\bar{X}_n n^{1/2}}{\hat{S}_n}.$$

The procedure of substituting the unknown variance σ^2 by its estimate $\hat{S}_n^2$ is called studentization.

Remark 7.2.9 *Consider the absolute moment of order r (integer) of the t_n-distribution:*

$$E\,|T|^r = m_r = C_n \int_0^\infty x^r \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2} dx$$

where C_n is the appropriate normalizing constant. For $x \to \infty$, the integrand is of order $x^r x^{-(n+1)} = x^{r-n-1}$, so this integral is finite only for $r < n$. For $r = n$, since $\int_0^\infty x^{-1} dx = \infty$, the r-th moment does not exist. This illustrates the fact that the t-distribution has "heavy tails" compared to the normal distribution; indeed $E\,|Z|^r < \infty$ for all r if Z is standard normal.

7.3 Some asymptotics

We saw in the plot of the t-distribution that already for 3 degrees of freedom it appears close to the normal. Write

$$\frac{n^{1/2} Z_1}{\sqrt{Z_2}} = \frac{Z_1}{\sqrt{n^{-1} Z_2}}$$

where Z_1, Z_2 are independent r.v.'s with standard normal and χ^2_n-distribution, respectively. Since

$$n^{-1} Z_2 = n^{-1} \sum_{i=1}^n \xi_i^2$$

for some i.i.d. standard normals ξ_i, we have by the law of large numbers

$$n^{-1} \sum_{i=1}^n \xi_i^2 \to_P E\xi_1^2 = 1.$$

This suggests that the law of $n^{1/2} Z_1/\sqrt{Z_2}$, i.e. the law t_n, should become close to the law of Z_1 as $n \to \infty$. Let us formally prove that statement. We begin with a recap of some probability notions.

Definition 7.3.1 *Let F_n, $n = 1, 2, \ldots$ be a sequence of distribution functions. The F_n are said to* ***converge in distribution*** *to a limit F (a distribution function) if*

$$F_n(t) \to F(t) \text{ as } n \to \infty$$

for every point of continuity t of F.

A *point of continuity* is obviously a point where F is continuous. Any distribution function F has left and right side limits at every point t, so it means that these limits coincide in t. When F is

continuous then in the above statement t must run through all $t \in \mathbb{R}$. For instance, the distribution function of every $N(\mu, \sigma^2)$ is continuous.
It is also said that a sequence of r.v.'s Y_n converges in distribution (or in law), written

$$Y_n \Longrightarrow_d F$$

when the d.f. of Y_n converge in d. to F. One also writes $Y_n \stackrel{\mathcal{L}}{\Rightarrow} Y$ for a r.v. Y having that distribution function F (or also $F_n \stackrel{\mathcal{L}}{\Rightarrow} F$, $F_n \stackrel{\mathcal{D}}{\Rightarrow} F$)

Example 7.3.2 *Convergence in probability* : if F is the d.f. of the random variable $Y = 0$ (which is always 0 !) then

$$F(t) = \mathbf{1}_{[1,\infty)}(t)$$

i.e. $F(t)$ jumps only in 0. Now $Y_n \Longrightarrow_d Y$ is equivalent to $Y_n \to_P 0$ *(Exercise)*

Example 7.3.3 Let $\mathcal{L}(Y_n) = B(n, \lambda n^{-1})$ and $\mathcal{L}(Y) = \mathrm{Po}(\lambda)$ then for all events A

$$\sup_A |P(Y_n \in A) - P(Y \in A)| \to 0$$

which implies $Y_n \Longrightarrow_d Y$ (take $A = (-\infty, t]$).

Example 7.3.4 ***(Central Limit Theorem)***. *Let $Y_1, \ldots, Y_n$ be independent identically distributed r.v.'s with distribution function F and finite second moment $EY_1^2 < \infty$. Let*

$$\bar{Y}_n = n^{-1} \sum_{i=1}^{n} Y_i$$

be the average (or sample mean) and $\sigma^2 = \mathrm{Var}(Y_1)$. Then for fixed F and $n \to \infty$

$$n^{1/2} \left(\bar{Y}_n - EY_1 \right) \Longrightarrow_d N(0, \sigma^2). \tag{7.6}$$

The normal law $N(0, \sigma^2)$ is continuous, so in the CLT the relation $\Longrightarrow_d$ means convergence of the d.f. of the standardized sum $n^{1/2} \left(\bar{Y}_n - EY_1 \right)$ to the normal d.f. at every $t \in \mathbb{R}$.

Lemma 7.3.5 *Suppose X_n is a sequence of r.v. which converges in distribution to a continuous limit law P_0:*

$$\mathcal{L}(X_n) \Longrightarrow_d P_0$$

and let Y_n be a sequence of r.v. which converges in probability to 0:

$$Y_n \to_P 0.$$

Then

$$\mathcal{L}(X_n + Y_n) \Longrightarrow_d P_0.$$

Note that no independence assumptions were made.

Proof. Let F_n be the distribution function of X_n and F_0 be the respective d.f. of the law P_0. Convergence in distribution means that

$$P(X_n \leq t) = F_n(t) \to F_0(t)$$

for every continuity point of the limit d.f. F_0. We assumed that P_0 is a continuous law, so it means convergence for every t. Now for $\varepsilon > 0$

$$\begin{aligned} P(X_n + Y_n &\leq t) = P(\{X_n + Y_n \leq t\} \cap (\{|Y_n| \leq \varepsilon\} \cup \{|Y_n| > \varepsilon\})) \\ &= P(\{X_n + Y_n \leq t\} \cap \{|Y_n| \leq \varepsilon\}) + P(\{X_n + Y_n \leq t\} \cap \{|Y_n| > \varepsilon\}). \end{aligned}$$

The first term on the right is

$$\begin{aligned} P(\{X_n \leq t - Y_n\} \cap \{|Y_n| \leq \varepsilon\}) &\leq P(\{X_n \leq t + \varepsilon\} \cap \{|Y_n| \leq \varepsilon\}) \\ &\leq P(X_n \leq t + \varepsilon). \end{aligned}$$

The other term is

$$P(\{X_n + Y_n \leq t\} \cap \{|Y_n| > \varepsilon\}) \leq P(|Y_n| > \varepsilon) \to 0 \text{ as } n \to \infty.$$

On the other hand we have

$$P(X_n \leq t + \varepsilon) \to F_0(t + \varepsilon) \text{ as } n \to \infty.$$

Hence for every $\delta > 0$ we can find m_1 such that for all $n \geq m_1$

$$P(X_n + Y_n \leq t) \leq F_0(t + \varepsilon) + 2\delta. \tag{7.7}$$

Now take the same $\varepsilon > 0$; we have

$$P(X_n \leq t - \varepsilon) =$$

$$\begin{aligned} & P(\{X_n \leq t - \varepsilon\} \cap \{|Y_n| \leq \varepsilon\}) + P(\{X_n \leq t - \varepsilon\} \cap \{|Y_n| > \varepsilon\}) \\ \leq\ & P(\{X_n + Y_n \leq t\} \cap \{|Y_n| \leq \varepsilon\}) + P(|Y_n| > \varepsilon) \\ \leq\ & P(X_n + Y_n \leq t) + P(|Y_n| > \varepsilon). \end{aligned}$$

Consequently

$$P(X_n + Y_n \leq t) \geq P(X_n \leq t - \varepsilon) - P(|Y_n| > \varepsilon).$$

Using again the two limits for the probabilities on the right, for every $\delta > 0$ we can find m_2 such that for all $n \geq m_2$

$$P(X_n + Y_n \leq t) \geq F_0(t - \varepsilon) - 2\delta. \tag{7.8}$$

Taking $m = \max(m_1, m_2)$ and collecting (7.7), (7.8), we obtain for $n \geq m$

$$F_0(t - \varepsilon) - 2\delta \leq P(X_n + Y_n \leq t) \leq F_0(t + \varepsilon) + 2\delta.$$

Since F_0 is continuous at t, and ε was arbitrary, we can select ε such that

$$\begin{aligned} F_0(t + \varepsilon) &\leq F_0(t) + \delta, \\ F_0(t - \varepsilon) &\geq F_0(t) - \delta \end{aligned}$$

so that for n large enough

$$F_0(t) - 3\delta \leq P(X_n + Y_n \leq t) \leq F_0(t) + 3\delta$$

and since δ was also arbitrary, the result follows. ■

Lemma 7.3.6 *Under the assumptions of Lemma 7.3.6, we have*

$$X_n Y_n \to_P 0.$$

Proof. Let $\varepsilon > 0$; and $\delta > 0$ be arbitrary and given. Suppose $|X_n Y_n| \geq \varepsilon$. Then, for every $T > 0$, either $\{|X_n| > T\}$, or if that is not the case, then $|Y_n| \geq \varepsilon/T$. Hence .

$$P\left(|X_n Y_n| \geq \varepsilon\right) \leq P\left(|X_n| \geq T\right) + P\left(|Y_n| \geq \varepsilon/T\right). \tag{7.9}$$

Now for every $T > 0$

$$\begin{aligned} P\left(|X_n| > T\right) &= 1 - F_n\left(X_n \leq T\right) + F_n\left(X_n < -T\right) \\ &\leq 1 - F_n\left(X_n \leq T\right) + F_n\left(X_n \leq -T\right). \end{aligned}$$

Since F_n converges to F_0 at both points T, $-T$, we find $m_1 = m_1(T)$ (depending on T) such that for all $n \geq m_1$

$$P\left(|X_n| \geq T\right) \leq 1 - F_0(T) + F_0(-T) + \delta$$

Select now T large enough such that

$$1 - F_0(T) \leq \delta,\ F_0(-T) \leq \delta.$$

Then for all $n \geq m_1(T)$

$$P\left(|X_n| \geq T\right) \leq 3\delta.$$

On the other hand, once T is fixed, in view of convergence in probability to 0 of $|Y_n|$, one can find m_2 such that for all $n \geq m_2$

$$P\left(|Y_n| \geq \varepsilon/T\right) \leq \delta.$$

In view of (7.9) we have for all $n \geq m = \max(m_1, m_2)$

$$P\left(|X_n Y_n| \geq \varepsilon\right) \leq 4\delta.$$

Since $\delta > 0$ was arbitrary, the result is proved. ■

We need an auxiliary result which despite its simplicity is still frequently cited with a name attached to it.

Lemma 7.3.7 ***(Slutsky's theorem)***. *Suppose a sequence of random variables X_n converges in probability to a number x_0 ($X_n \to_p x_0$ as $n \to \infty$). Suppose f is a real valued function defined in a neighborhood of x_0 and continuous there. Then*

$$f(X_n) \to_p f(x_0),\ n \to \infty$$

Proof. Consider an arbitrary $\varepsilon > 0$. Select $\delta > 0$ small enough such that $(x_0 - \delta, x_0 + \delta)$ is in the abovementioned neighborhood of x_0 and also fulfilling the condition that $|z - x_0| \leq \delta$ implies $|f(z) - f(x_0)| \leq \varepsilon$ (by continuity of f such a δ can be found). Then the event $|f(X_n) - f(x_0)| > \varepsilon$ implies $|X_n - x_0| > \delta$ and hence

$$P\left(|f(X_n) - f(x_0)| > \varepsilon\right) \leq P\left(|X_n - x_0| > \delta\right).$$

Since the latter probability tends to 0 as $n \to \infty$, we also have

$$P\left(|f(X_n) - f(x_0)| > \varepsilon\right) \to 0 \text{ as } n \to \infty$$

and since ε was arbitrary, the result is proved. ■

Theorem 7.3.8 *The t-distribution with n degrees of freeedom converges (in distribution) to the standard normal law as $n \to \infty$.*

Proof. Let

$$X_n = \frac{Z_1}{\sqrt{n^{-1}Z_2}}$$

where Z_1, Z_2 are independent r.v.'s with standard normal and χ^2_n-distribution, respectively. We know already

$$n^{-1}Z_2 \to_P 1$$

By Slutsky's theorem, for the r.v. $Y_{1,n}$

$$Y_{1,n} := \frac{1}{\sqrt{n^{-1}Z_2}} \to_P 1$$

since the function $g(x) = x^{-1/2}$ is continuous at 1 and defined for $x > 0$. (Indeed we need consider only those x, since $n^{-1}Z_2 > 0$ with probability one). Now

$$X_n = Z_1 Y_{1,n} = Z_1 + Z_1 (Y_{1,n} - 1).$$

Now $Y_{1,n} - 1 \to_P 0$, hence by Lemma 7.3.6 $Z_1 (Y_{1,n} - 1) \to_P 0$. Now Z_1 is a constant sequence with law $N(0,1)$, which certainly converges in law to $N(0,1)$. Then by Lemma 7.3.5 $X_n \Longrightarrow_d N(0,1)$. ■

We can translate this limiting statement about the t-distribution into a confidence statement.

Theorem 7.3.9 *Let $z^*_{\alpha/2}$ be the upper $\alpha/2$-quantile of $N(0,1)$. Then in the Gaussian location-scale model $\mathbf{M}_{c,2}$, the interval*

$$[\hat{\mu}^*_-, \hat{\mu}^*_+] = [\bar{X}_n - \hat{S}_n z^*_{\alpha/2}/\sqrt{n}, \bar{X}_n + \hat{S}_n z^*_{\alpha/2}/\sqrt{n}]$$

*is an **asymptotic α-confidence interval** for the unkown expected value μ:*

$$\lim_{n\to\infty} P_{\mu,\sigma^2}([\hat{\mu}^*_-, \hat{\mu}^*_+] \ni \mu) \geq 1 - \alpha.$$

Here the same quantiles as in the exact interval (7.3) are used, but $\hat{S}_n$ replaces the unknown σ. In summary: if σ^2 is unknown, one has the choice between an exact confidence interval (which keeps level $1-\alpha$) based on the t-distribution, or an asymptotic interval (which keeps the confidence level only approximately) based on the normal law. The normal interval would be shorter in general: consider e.g. degrees of freedom 10 and $\alpha = 0.05$; then for the t-distribution we have $z_{\alpha/2} = 2.228$, whereas the normal quantile is $z^*_{\alpha/2} = 1.96$ (cf. the tables of the normal and t-distribtions on pp. 608/609 of [CB]).

Note that in subsection (1.3.2) we discussed an nonasymptotic confidence interval for the Bernoulli parameter p based on the Chebyshev inequality, and mentioned that alternatively, a large sample approximation based on the CLT could also have been used. In this section we developed the tools for this.

Chapter 8
TESTING STATISTICAL HYPOTHESES

8.1 Introduction

Consider again the basic statistical model where X is an observed random variable with values in $\mathcal{X}$ and the law $\mathcal{L}(X)$ is known up to a parameter ϑ from a parameter space Θ: $\mathcal{L}(X) \in \{P_\vartheta; \vartheta \in \Theta\}$. This time we do not restrict the nature of Θ; this may be a general set, possibly even a set of laws (in this case ϑ is identified with $\mathcal{L}(X)$). Suppose the parameter set Θ is divided into two subsets: $\Theta = \Theta_0 \cup \Theta_1$ where $\Theta_0 \cap \Theta_1 = \emptyset$. The problem is to decide on the basis of an observation of X whether the unknown ϑ belongs to Θ_0 or to Θ_1. Thus two *hypotheses* are formulated:

H $\vartheta \in \Theta_0$, the **hypothesis**

K $\vartheta \in \Theta_1$, the **alternative**.

Of course both of these are "hypotheses", but in testing theory they are treated in a nonsymmetric way (to be explained). In view of this nonsymmetry, one of them is called the hypothesis and the other the alternative. It is traditional to write them as above, with letters H for the first (*the* hypothesis) and K for the second (the alternative).

Example 8.1.1 *Assume that a new drug has been developed, which is supposed to have a higher probability p of success when applied to an average patient. The new drug will be introduced only if a high degree of certainty can be obtained that it is better. Suppose p_0 is the known probability of success of the old drug. Clinical trials are performed to test the hypothesis that $p > p_0$. For the new drug, n patients are tested independently, and succes of the drug is measured (we assume that only success or failure of the treatment can be seen in each case). Let the j-th experiment (patient) for the new drug be X_j; assume that the X_j are independent $B(1,p)$. Thus observations are $X = (X_1, \ldots, X_n)$, X_j are i.i.d. Bernoulli r.v.'s., where $\vartheta = p$ and $\Theta = (0,1)$. The hypotheses are $\Theta_0 = (0, p_0]$ and $\Theta_1 = (p_0, 1)$.*

The motivation for a nonsymmetric treatment of the hypotheses is evident in this example: if the statistical evidence is inconclusive, one would always stay with the old drug. There can be no question of treating H and K the same way. Thus in section 5.5 we briefly discussed the problem of estimating a signal $\vartheta \in \{0,1\}$ (binary channel, Gaussian channel), where basically both values 0 and 1 are treated the same way, e.g. we use a Bayesian decision rule for prior probabilities 1/2. In contrast, one will decide 1, i.e decide in favor of the new drug only if there is "reasonable statistical certainty".

Definition 8.1.2 *A **test** is a decision rule characterized by an **acceptance region** $S \subset \mathcal{X}$, i.e. a (measurable) subset of the sample space, such that*

$X \in S$ means that $\vartheta \in \Theta_0$ is accepted
$X \in S^c$ means that $\vartheta \in \Theta_0$ is rejected (thus $\vartheta \in \Theta_1$ is accepted.)
*The complement S^c is called **the critical region** (rejection region).*

Formally, a test ϕ is usually defined as a statistic which is the indicator function of a set S^c, such that

$$\phi(X) = \mathbf{1}_{S^c}(X)$$

where a value $\phi(X) = 1$ is understood as a decision that $\vartheta \in \Theta_0$ is rejected (and 0 that it is accepted).
In the above example, a reasonable test would be given by a rejection region

$$S^c = \left\{ x : n^{-1} \sum_{i=1}^{n} x_j > c \right\}$$

for realizations $x = (x_1, \ldots, x_n)$, where c is some number fixed in advance. One would decide "$\vartheta \in \Theta_1$" if the number of successes in the sample is large enough.
Tests and estimators have in common that they are statistics which are also decision rules. But the nature of the decisions is different, which is reflected in the loss functions. The natural common loss function for testing problems is "loss is 0 if the decision is correct, and is 1 otherwise". Thus if $d \in \{0, 1\}$ is one of the two possible decisions then the **loss function** is

$$L(d, \vartheta) = \{ \begin{array}{l} d \text{ if } \vartheta \in \Theta_0 \\ 1 - d \text{ if } \vartheta \in \Theta_1 \end{array} .$$

As for estimation, the **risk** of a decision rule ϕ at parameter value ϑ is the expected loss when ϑ is the true parameter. The decision rule is $\phi(X) = \mathbf{1}_{S^c}(X)$; its risk is

$$R(\phi, \vartheta) = E_\vartheta L(\phi(X), \vartheta) = \{ \begin{array}{l} E_\vartheta \phi(X) = P_\vartheta(S^c) \text{ if } \vartheta \in \Theta_0 \\ 1 - E_\vartheta \phi(X) = P_\vartheta(S) \text{ if } \vartheta \in \Theta_1. \end{array}$$

Thus the risk coincides with the **probability of error** in each case: for $\vartheta \in \Theta_0$, an erroneous decision is made when $X \in S^c$; when $\vartheta \in \Theta_1$ is true, then the error in the decision occurs when $X \in S$.
Since both probability of errors are functions of $E_\vartheta(1 - \phi(X)) = P_\vartheta(S)$, one can equivalently work with the **operation characteristic (OC):**

$$OC(\phi, \vartheta) = P_\vartheta(S) = \{ \begin{array}{l} 1 - R(\phi, \vartheta) \text{ if } \vartheta \in \Theta_0 \\ R(\phi, \vartheta) \text{ if } \vartheta \in \Theta_1. \end{array}$$

A test with zero risk would require a set S such that $P_\vartheta(S) = 1$ if $\vartheta \in \Theta_0$, $P_\vartheta(S) = 0$ if $\vartheta \in \Theta_1$. This is possible only in degenerate cases: if such an $S \subset \mathcal{X}$ exists then the families $\{P_\vartheta; \vartheta \in \Theta_0\}$ and $\{P_\vartheta; \vartheta \in \Theta_1\}$ are said to be **orthogonal**. In this case "sure" decisions are possible according to whether $X \in S$ or not, and one is led outside the realm of statistics.

Example 8.1.3 *Let $U(a, b)$ be the uniform law on the interval (a, b), $\Theta = \{0, 1\}$ and $P_0 = U(0, 1)$, $P_1 = U(1, 2)$. Here $S = (0, 1)$ gives zero error probabilities.*

In typical statistical situations, the probabilities $P_\vartheta(S)$ depends continuously on the parameter, and the sets Θ_0, Θ_1 are bordering each other. In this case the risk $R(\phi, \vartheta)$ cannot be near 0 on the common border.

Example 8.1.4 *Let $\Theta = \mathbb{R}$, $P_\vartheta = N(\vartheta, \sigma^2)$, ($\sigma^2$ fixed), $\Theta_0 = (-\infty, 0]$, $\Theta_1 = (0, \infty)$. (Gaussian location model, Model $\mathbf{M}_{c,1}$, for sample size $n = 1$). Reasonable acceptance regions are*

$$S_a = \{x : x \leq a\}$$

for some a. The OC of any such test $\phi_a = \mathbf{1}_{S_a^c}$ is (if ξ is a standard normal r.v. such that $X = \vartheta + \sigma\xi$)

$$\begin{aligned} OC(\phi, \vartheta) &= P_\vartheta(S) = P_\vartheta(X \leq a) = P(\xi \leq (a - \vartheta)/\sigma) \\ &= \Phi\left((a - \vartheta)/\sigma\right) \end{aligned}$$

where Φ is the standard normal distribution function.

Example 8.1.1 (continued). Suppose $p_0 = 0.6$ and use a critical region $\left\{\bar{X}_n > c\right\}$ for $c = 0.8$. We assume n is not small and use the De Moivre-Laplace central limit theorem for an approximation to the OC. We have

$$\begin{aligned} OC(\phi, p) &= P_p(\bar{X}_n \leq c) = P_p\left(\frac{n^{1/2}(\bar{X}_n - p)}{(p(1-p))^{1/2}} \leq \frac{n^{1/2}(c-p)}{(p(1-p))^{1/2}}\right) \\ &\approx \Phi\left(\frac{n^{1/2}(c-p)}{(p(1-p))^{1/2}}\right) \end{aligned}$$

where $\approx$ means approximate equality. For a visualization of this approximation to the OC, see the plot below. □

These examples show that it is not generally possible to keep error probabilities uniformly small under both hypotheses.

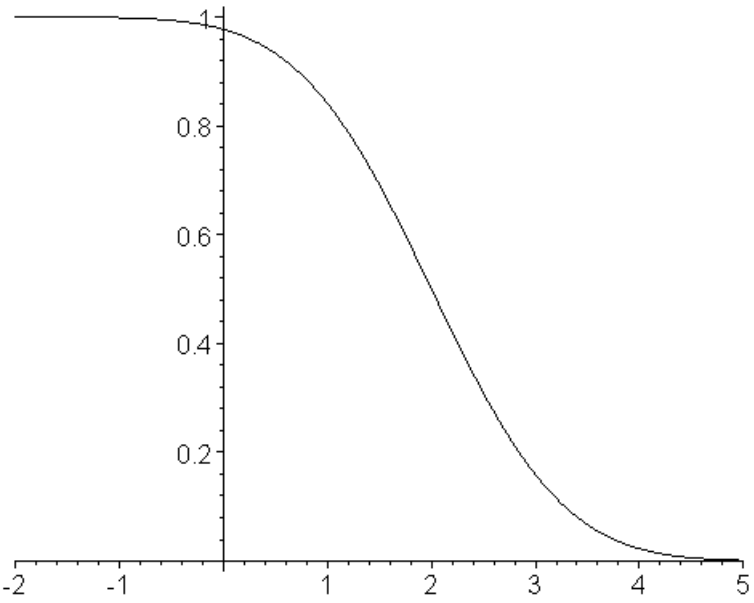
Definition 8.1.5 *Suppose that $\mathcal{L}(X) \in \{P_\vartheta; \vartheta \in \Theta\}$ and ϕ is test for $H : \vartheta \in \Theta_0$ vs. $K : \vartheta \in \Theta_1$ with acceptance region S.*
***(i)** An **error of the first kind** is: ϕ takes value 1 when $\vartheta \in \Theta_0$ is true;*
*An **error of the second kind** is: ϕ takes value 0 when $\vartheta \in \Theta_1$ is true;*
***(ii)** ϕ has **significance level** (or level) α if*

$$P_\vartheta(\phi(X) = 1) \leq \alpha \text{ for all } \vartheta \in \Theta_0$$

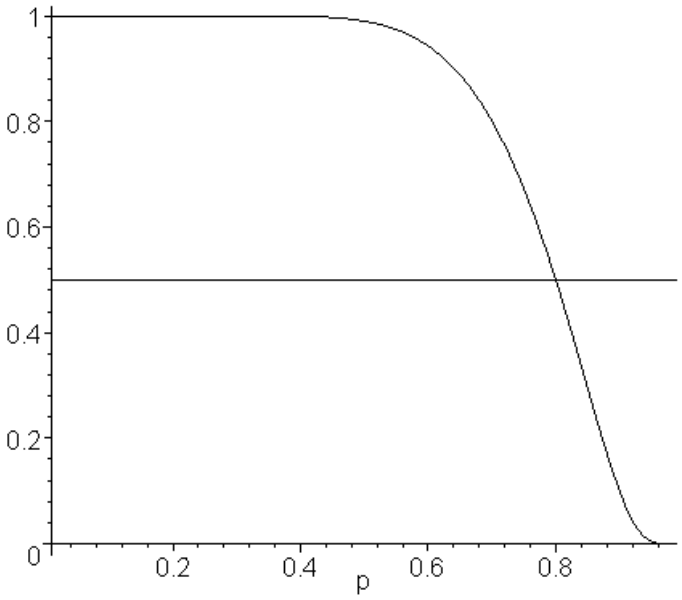
(iii) For $\vartheta \in \Theta_1$, the probability

$$\beta(\vartheta) = P_\vartheta(\phi(X) = 1)$$

*is called the **power** of the test ϕ.*



OC of the critical region $\{x > 2\}$ for testing $H : \vartheta \leq 0$ vs. $K : \vartheta > 0$ for the family $N(\vartheta, 1)$, $\vartheta \in \mathbb{R}$



Approximation to OC for critical region $\{\bar{X}_n > c\}$ for testing $H : p \leq p_0$ vs. $K : p > p_0$ for i.i.d. Bernoulli model (Model $\mathbf{M}_{d,1}$), with $p_0 = 0.6$, $c = 0.8$, $n = 15$

Thus significance level α means that probability of an error of the first kind is uniformly less than α. The power is the probability of not making an error of the second kind for a given ϑ in the alternative.
In terms of the risk, ϕ has level α if $R(\phi, \vartheta) \leq \alpha$ for all $\vartheta \in \Theta_0$, and the power is

$$\beta(\vartheta) = 1 - R(\phi, \vartheta),\ \vartheta \in \Theta_1.$$

In terms of the OC, ϕ has level α if $OC(\phi, \vartheta) \geq 1 - \alpha$ and the power is

$$\beta(\vartheta) = 1 - OC(\phi, \vartheta),\ \vartheta \in \Theta_1$$

In example 8.1.1, it is particularly apparent why the error of the first kind is very sensitive, so that its probability should be kept under a given small α. When actually the old drug is better ($\vartheta \in \Theta_0$), but we decide erroneously that the new is better, it is a very painful error indeed, with potentially grave consequences. We wish a decision procedure which limits the probability of such a catastrophic misjudgment. But given this restriction, opportunities for switching to a better drug should not be missed, i.e. when the new drug is actually better, then the decision should be able to detect it with as high a probability as possible.
For drug testing, procedures like this (significance level α should be kept for some small α) are required by law in every developed country. In general statistical practice, common values are $\alpha = 0.05$ and $\alpha = 0.01$.
If one of the hypothesis sets Θ_0, Θ_1 consists of only one element, the respective hypothesis (or alternative) called **simple**, otherwise it is called **composite** . An important special case of testing theory is the one where both hypothesis and alternative are simple, i.e. $\Theta_0 = \{\vartheta_0\}$, $\Theta_1 = \{\vartheta_1\}$ (which means the whole parameter space Θ consists of the two elements ϑ_0, ϑ_1; in this case the *Neyman-Person fundamental lemma* applies (see below).
A test where Θ_0 is simple is called a **significance test**. The question to be decided there is only whether the hypothesis $H : \vartheta_0 = \vartheta$ can be rejected with significance level α; the alternatives are usually not specified.

8.2 Tests and confidence sets

A **confidence set** is a random subset $A(X)$ of the parameter space which covers the true parameter with probability at least $1 - \alpha$:

$$P_\vartheta(A(X) \ni \vartheta) \geq 1 - \alpha,\ \vartheta \in \Theta.$$

Confidence intervals $A(X) = [\hat{\mu}_-(X), \hat{\mu}_+(X)]$ were treated in detail in section 7.1 for the Gaussian location-scale model and in the introductory section 1.3. Confidence sets are also called **domain estimators** of the parameter ϑ (the estimators which pick a value of ϑ rather than a covering set are called **point estimators**).
There is a close connection between confidence sets and significance tests.

1 Suppose a confidence set $A(X)$ for level $1 - \alpha$ is given. Let $\vartheta_0 \in \Theta$ be arbitrary and consider a simple hypothesis $H : \vartheta = \vartheta_0$ (vs. alternative $K : \vartheta \neq \vartheta_0$). Construct a test ϕ_{ϑ_0} by

$$\phi_{\vartheta_0}(X) = 1 - \mathbf{1}_{A(X)}(\vartheta_0)$$

where $\mathbf{1}_{A(X)}$ is the indicator of the confidence set, as a function of ϑ_0. In other words, $H : \vartheta = \vartheta_0$ is rejected if ϑ_0 is outside the confidence set. Then

$$P_{\vartheta_0}(\phi_{\vartheta_0} = 1) = 1 - P_{\vartheta_0}(A(X) \ni \vartheta_0) \leq \alpha$$

hence ϕ_{ϑ_0} is an α-significance test for $H : \vartheta = \vartheta_0$.

2 We saw that a confidence set generates a family of significance tests, one for each $\vartheta_0 \in \Theta$. Assume now conversely that such a family ϕ_ϑ, $\vartheta \in \Theta$ is given, and they all observe level α. Define

$$A(X) = \{\vartheta \in \Theta : \phi_\vartheta(X) = 0\}.$$

Then

$$\begin{aligned} P_\vartheta(A(X) \ni \vartheta) &= P_\vartheta(\phi_\vartheta(X) = 0) = 1 - P_\vartheta(\phi_\vartheta(X) = 1) \\ &\geq 1 - \alpha. \end{aligned}$$

i.e. we found a $1-\alpha$-confidence set $A(X)$.

For a more general setting, let $\gamma(\vartheta)$ be a function of the parameter (with values in an arbitrary set Γ). A confidence set for $\gamma(\vartheta)$ is defined by

$$P_\vartheta(A(X) \ni \gamma(\vartheta)) \geq 1 - \alpha,\ \vartheta \in \Theta.$$

For instance ϑ might have two components: $\vartheta = (\vartheta_1, \vartheta_2)$. and $\gamma(\vartheta) = \vartheta_1$. Then the above family of tests should be indexed by $\gamma \in \Gamma$, and ϕ_{γ_0} has level α for a hypothesis H: $\gamma(\vartheta) = \gamma_0$. This hypothesis is composite if γ is not one-to-one (then ϕ_{γ_0} cannot be called a significance test).
As an example, consider the Gaussian location-scale model for unknown σ^2 (Model $M_{c,2}$). Here $\vartheta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$. Consider the quantity

$$T_\mu = T_\mu(X) = \frac{\left(\bar{X}_n - \mu\right) n^{1/2}}{\hat{S}_n}$$

used to build a confidence interval

$$[\hat{\mu}_-, \hat{\mu}_+] = [\bar{X}_n - \hat{S}_n z_{\alpha/2}/\sqrt{n}, \bar{X}_n + \hat{S}_n z_{\alpha/2}/\sqrt{n}]$$

for the unkown expected value μ. The level $1-\alpha$ was kept for all unknown $\sigma^2 > 0$, i.e. we have a confidence interval for the parameter function $\gamma(\vartheta) = \mu$.
As we remarked, $T_\mu(X)$ depends on the parameter μ and therefore is not a statistic. Such a function of both the observations and the parameter used to build a confidence interval is called a **pivotal quantity.** The knowledge of the law of the pivotal quantity under the respective parameter is the basis for a confidence interval. When looking at the significance test derived from $T_\mu(X)$, for a hypothesis $H : \mu = \mu_0$, we find that the test is

$$\phi_{\mu_0}(X) = 1 \text{ if } \left|T_{\mu_0}(X)\right| > z_{\alpha/2} \tag{8.1}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$-quantile of the t-distribution for $n-1$ degrees of freedom. From this point of view, when $\mu = \mu_0$ is a "known" hypothesis, $T_{\mu_0}(X)$ does not depend on an unknown parameter, and is thus a statistic. In the example, it is the **t-statistic** for testing $H : \mu = \mu_0$, and the test (8.1) is the **two sided t-test**. The basic result implied by Theorem 7.2.8 about this test is the following.

Theorem 8.2.1 *In the Gaussian location-scale model Model $M_{c,2}$, for sample size n, consider the hypothesis $H : \mu = \mu_0$. Let $z_{\alpha/2}$ be the upper $\alpha/2$-quantile of the t-distribution with $n-1$ degrees of freedom. Then the two sided t-test (8.1) has level α for any unknown $\sigma^2 > 0$.*

Analogously , when σ^2 is known, the test

$$\phi_{\mu_0}(X) = 1 \text{ if } \left|Z_{\mu_0}(X)\right| > z^*_{\alpha/2} \tag{8.2}$$

where $z^*_{\alpha/2}$ is the upper $\alpha/2$-quantile of $N(0,1)$ and

$$Z_\mu = Z_\mu(X) = \frac{\left(\bar{X}_n - \mu\right) n^{1/2}}{\sigma}$$

is the Z-statistic, is called the **two sided Gauss test** for $H : \mu = \mu_0$. Then the t-statistic $T_{\mu_0}(X)$ can be construed as a studentized Z-statistic $Z_{\mu_0}(X)$.
Let us investigate the relationship between the power of ϕ_ϑ and the confidence interval. Suppose that we have two $1-\alpha$-confidence sets $A^*(X)$, $A(X)$ for ϑ itself, where

$$A^*(X) \subseteq A(X)$$

i.e. $A^*(X)$ is contained $A(X)$ (in the case of intervals, $A^*(X)$ would be shorter or of equal length). For the respective families ϕ_{ϑ_0}, $\phi^*_{\vartheta_0}$, $\vartheta_0 \in \Theta$ of α-significance tests this means

$$\phi_{\vartheta_0}(X) = 1 \text{ implies } \phi^*_{\vartheta_0}(X) = 1$$

hence

$$P_\vartheta\left(\phi_{\vartheta_0} = 1\right) \leq P_\vartheta\left(\phi^*_{\vartheta_0} = 1\right) \text{ for all } \vartheta \in \Theta.$$

At $\vartheta \neq \vartheta_0$ these are precisely the respective powers of the two tests ϕ_{ϑ_0}, $\phi^*_{\vartheta_0}$. (At $\vartheta = \vartheta_0$ the relation implies that for $A(X)$ to keep level $1-\alpha$, it is sufficient that $A^*(X)$ keeps this level). Thus $\phi^*_{\vartheta_0}$ has uniformly better (or at least as good) power for all $\vartheta \neq \vartheta_0$.
It was mentioned that shorter confidence intervals are desirable (given a confidence level), since they enable "sharper" decision making. Translating this into a power relation for tests, we have made the idea more transparent. The assumed inclusion $A^*(X) \subseteq A(X)$ implies a larger critical region for $\phi^*_{\vartheta_0}$: $\left\{\phi_{\vartheta_0} = 1\right\} \subseteq \left\{\phi^*_{\vartheta_0} = 1\right\}$. This does not describe all situations in which $\phi^*_{\vartheta_0}$ might have better power (and thus $A^*(X)$ is better in some sense); we shall not further investigate the "power" of confidence intervals here but will concentrate on tests instead.
However *asymptotic confidence intervals* should be discussed briefly. The statement of Theorem 7.3.9 can be translated immediately into the language of test theory. When the law of the observed r.v. X depends on n, we write X_n and $\mathcal{L}(X_n) \in \{P_{\vartheta,n}; \vartheta \in \Theta\}$ for the family of laws. Here the observation space $\mathcal{X}_n$ might also depend on n, as is the case of n i.i.d. observed variables.

Definition 8.2.2 ***(i)*** *A sequence of tests $\phi_n = \phi_n(X_n)$ for testing $H : \vartheta \in \Theta_0$ vs. $K : \vartheta \in \Theta_1$ has* ***asymptotic level*** *α if*

$$limsup_{n\to\infty} P_{\vartheta,n}(\phi_n(X) = 1) \leq \alpha \text{ for all } \vartheta \in \Theta_0$$

(ii) *The sequence ϕ_n is* ***consistent*** *if it has asymptotic power one, i.e.*

$$\lim_n P_{\vartheta,n}(\phi_n(X) = 1) = 1 \text{ for all } \vartheta \in \Theta_1.$$

We saw in Theorem 7.3.9 that in Model $\mathbf{M}_{c,2}$, the interval

$$[\hat{\mu}_-^*, \hat{\mu}_+^*] = [\bar{X}_n - \hat{S}_n z_{\alpha/2}^*/\sqrt{n}, \bar{X}_n + \hat{S}_n z_{\alpha/2}^*/\sqrt{n}]$$

is an asymptotic α-confidence interval for the unkown expected value μ:

$$\text{liminf}_{\,n\to\infty} P_{\mu,\sigma^2}([\hat{\mu}_-, \hat{\mu}_+] \ni \mu) \geq 1 - \alpha$$

where $z_{\alpha/2}^*$ is a normal quantile. Thus if $\phi_{\mu_0}^*$ is the derived test for $H : \mu = \mu_0$ then

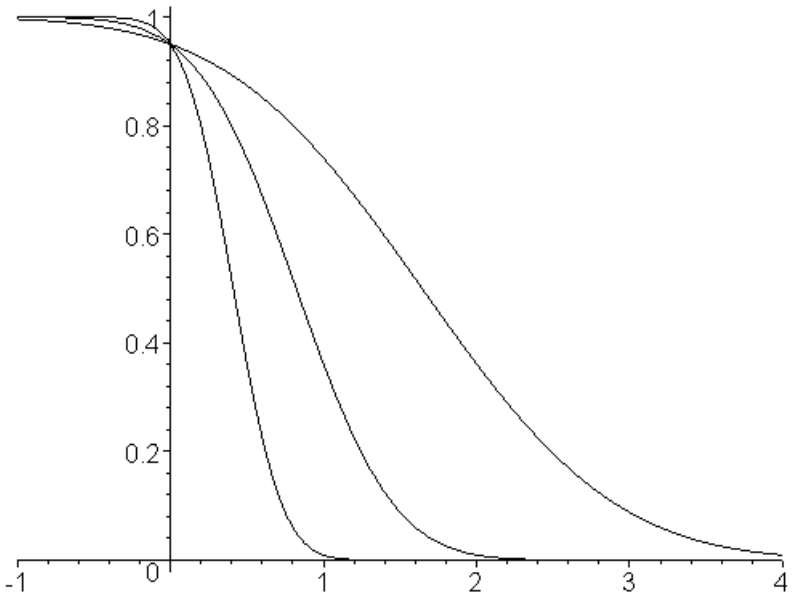
$$\begin{aligned} \text{limsup}_{\,n\to\infty} P_{\mu,\sigma^2}\left(\phi_{\mu_0} = 1\right) &= 1 - \text{liminf}_{\,n\to\infty} P_{\mu,\sigma^2}([\hat{\mu}_-, \hat{\mu}_+] \ni \mu_0) \\ &\leq \alpha \end{aligned}$$

so that every $\phi_{\mu_0}^*$ is an asymptotic α-test.

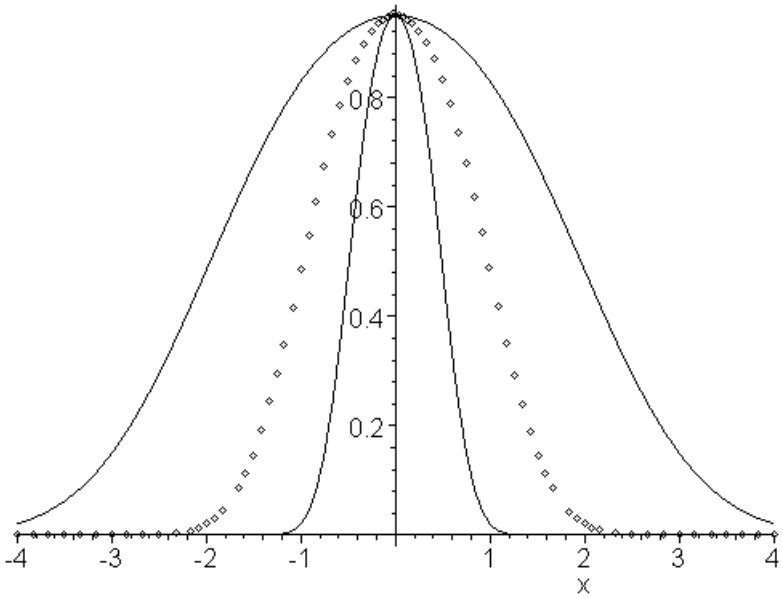
Exercise. Consider the test $\phi_{\mu_0}^*$ (i.e. the test based on the t-statistic as in (8.1), where the quantile $z_{\alpha/2}^*$ of the standard normal is used in place of $z_{\alpha/2}$). Show that in the Gaussian location-scale model $\mathbf{M}_{c,2}$, for sample size n, this test is consistent as $n \to \infty$ on the pertaining alternative

$$\Theta_1(\mu_0) = \left\{(\mu, \sigma^2) : \mu \neq \mu_0\right\}.$$

.



Consistency of one sided Gauss tests (OC-plot)



Consistency of two sided Gauss tests (OC-plot)

Above, the first plot, in the Gaussian location model with $\sigma^2 = 1$ gives the OC of a test of $H{:}\mu \leq \mu_0$ vs $K : \mu > \mu_0$ of type

$$\phi(X) = \{ \begin{array}{l} 1 \text{ if } \bar{X}_n > c_n \\ 0 \text{ otherwise} \end{array}$$

where c_n is selected such that ϕ is an α-test, for $\alpha = 0.05$ and sample sizes $n = 1, 2, 4$ respectively. This is the same situation as in the first plot on p. 96, only α is selected as one of the common values (in the other figure we just took $c_1 = 2$ and did not care about the resulting α), and three sample sizes are plotted. This is a **one sided Gauss test**; we have not yet discussed the respective theory, but it can be observed that these test keep level α on the whole composite hypothesis $H{:}\mu \leq \mu_0$. Moreover, the behaviour of consistency is visible (for larger n, the power increases).
The second plot concerns the simple hypothesis $H{:}\mu = \mu_0$ in the same model ($\sigma^2 = 1$, $\alpha = 0.05$, $n = 1, 2, 4$) and the two sided Gauss test (8.2) derived from the confidence interval (7.3). The middle OC-line for $n = 2$ is dotted.

8.3 The Neyman-Pearson Fundamental Lemma

We saw that in testing theory, the a basic goal is to maximize the power of tests, while keeping significance level α.

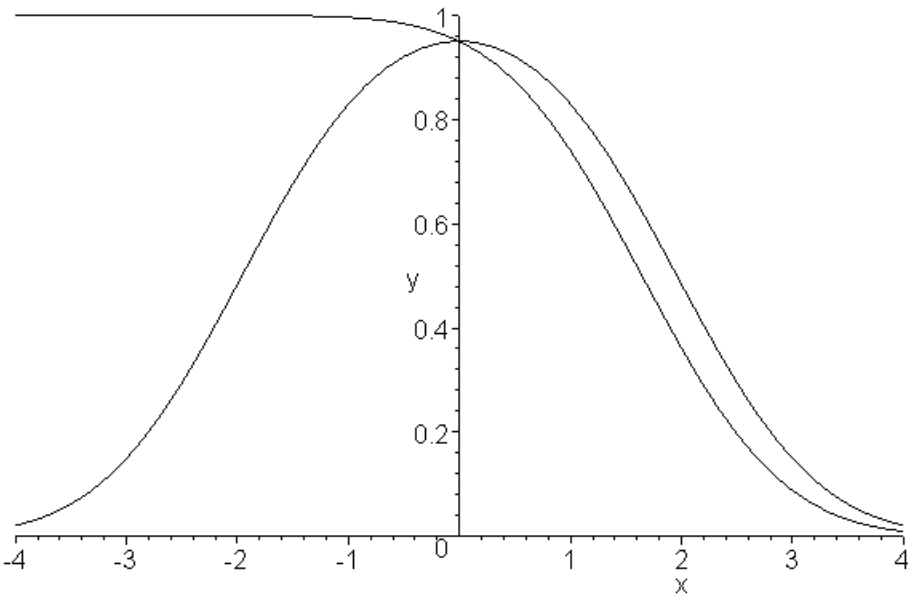
Definition 8.3.1 *Suppose that $\mathcal{L}(X) \in \{P_\vartheta; \vartheta \in \Theta\}$. An α-test $\phi = \phi(X)$ for testing $H : \vartheta \in \Theta_0$ vs. $K : \vartheta \in \Theta_1$ is **uniformly most powerful (UMP)** if for every other test ψ which is an α-test for the problem:*

$$\sup_{\vartheta \in \Theta_0} E_\vartheta \psi \leq \alpha$$

we have

$$E_\vartheta \psi \leq E_\vartheta \phi, \text{ for all } \vartheta \in \Theta_1.$$

Typically is not possible to find such a UMP test; some tests do better at particular points in the alternative, at the the expense of the power at other points. An example is given in the following plot.



Power of one sided and two sided Gauss test for $H : \mu = \mu_0$ vs. $K : \mu \neq 0$

In the Gausian location model with $\sigma^2 = 1$ and $n = 1$, for hypotheses H: $\mu = 0$ vs $K : \mu \neq 0$ we look at the OC of two tests:

$$\text{one sided Gauss test: } \phi_1(X) = \begin{cases} 1 \text{ if } \bar{X}_n > c_1 \\ 0 \text{ otherwise} \end{cases}$$

$$\text{two sided Gauss test: } \phi_2(X) = \begin{cases} 1 \text{ if } \left|\bar{X}_n\right| > c_2 \\ 0 \text{ otherwise} \end{cases}$$

where c_1, c_2 are selected such both ϕ_1, ϕ_2 are α-tests for $\alpha = 0.05$. This is the same situation as in the two plots on p. 8.2 and the following, where we put the two tests in one picture and omitted the larger sample sizes $n = 2, 4$. (Here of course $\bar{X}_n = X_1$ if $n = 1$, but the same curves appear if only $\mathrm{Var}(\bar{X}_n) = 1$, i.e. $\sigma^2/n = 1$).
We see that ϕ_1 is better than ϕ_2 for alternatives $\mu > 0$, but it is much worse for alternatives $\mu < 0$. If we really are interested also in alternatives $\mu < 0$ (i.e. we wish to detect these, and not just be content with a statement $\mu \leq 0$) we should apply the two sided test; the one sided ϕ_1 is totally implausible for the two sided alternative $\mu \neq 0$. However, even though it is implausible, ϕ_1 has better power for $\mu > 0$.
In one special situation it is possible to find a UMP test, namely when both hypothesis and alternative are simple. In this case, we have only one point in the alternative, and maximizing the power turns out to be possible. Recall the continuous statistical model, first defined in section 4.3:

Model $\mathbf{M}_c$ The observed random variable $X = (X_1, \ldots, X_k)$ is continuous with values in $\mathbb{R}^k$ and $\mathcal{L}(X) \in \{P_\vartheta, \vartheta \in \Theta\}$. Each law P_ϑ is described by a joint density $p_\vartheta(x) = p_\vartheta(x_1, \ldots, x_k)$, and $\Theta \subseteq \mathbb{R}^d$.

(Earlier we required that Θ be an open set, but this is omitted now).

Definition 8.3.2 *Assume Model* $\mathbf{M}_c$*, and that* $\Theta = \{\vartheta_0, \vartheta_1\}$ *consists of only two elements. A test* ϕ *for the hypotheses*
$H : \vartheta = \vartheta_0$
$K : \vartheta = \vartheta_1$
is called a ***Neyman-Pearson test*** *of level* α *if*

$$\phi(X) = \begin{cases} 1 \text{ if } p_{\vartheta_1}(x) > c\, p_{\vartheta_0}(x) \\ 0 \text{ otherwise} \end{cases}$$

where the value c *is chosen such that*

$$P_{\vartheta_0}(\phi(X) = 1) = \alpha. \tag{8.3}$$

We should first show that a Neyman-Person test exists. Of course we can take any c and build a test according to the above rule. This rule seems plausible: given x, each of the two densities can be regarded as a likelihood. We might say that ϑ_1 is more "likely" if the ratio of likelihoods $p_{\vartheta_1}(x)/p_{\vartheta_0}(x)$ is sufficiently large. We recognize an application of the *likelihood principle* (recall that this consists in regarding the density as a function of $\vartheta \in \Theta$ when x is already observed, and assigning corresponding likelihoods to each ϑ).
The question is only whether c can be chosen such that (8.3) holds. Define a random variable

$$L = L(X) = \begin{cases} p_{\vartheta_1}(X)/p_{\vartheta_0}(X) \text{ if } p_{\vartheta_0}(X) > 0 \\ 0 \text{ otherwise} \end{cases}$$

where X has distribution P_{ϑ_0}, and let F_L be its distribution function

$$F_L(t) = P_{\vartheta_0}(L(X) \le t).$$

Recall that distribution functions F are monotone increasing, right continuous and 0 at $-\infty$, 1 at ∞. To ensure (8.3), we make
Assumption L. *The distribution function* $F_L(t)$ *is continuous.*
In this case

$$P_{\vartheta_0}(L(X) = t) = 0$$

for all t**,** and for $c > 0$ consider the probability $P_{\vartheta_0}(\phi(X) = 1)$. Since

$$P_{\vartheta_0}(p_{\vartheta_0}(X) = 0) = 0,$$

we have

$$\begin{aligned} P_{\vartheta_0}(\phi(X) = 1) &= P_{\vartheta_0}(\{\phi(X) = 1\} \cap \{p_{\vartheta_0}(X) > 0\}) \\ &= P_{\vartheta_0}(\{L(X) > c\} \cap \{p_{\vartheta_0}(X) > 0\}) \\ &= P_{\vartheta_0}(L(X) > c) = 1 - P_{\vartheta_0}(L(X) \le c) \\ &= 1 - P_{\vartheta_0}(L(X) < c) = 1 - F_L(c) \end{aligned}$$

which is continuous and monotone a with limit 0 for $c \to \infty$. Furthermore, $F_L(t) = 0$ for all $t < 0$ since $L(X)$ is nonnegative, and in view of continuity of F_L (Assumption L) we have

$$1 - F_L(c) \to 0 \text{ as } c \to 0.$$

Thus a value c with (8.3) exists for every $\alpha \in (0, 1)$.

Example 8.3.3 *Let $P_\vartheta = N(\vartheta, 1)$, $\vartheta \in \Theta$ (Gaussian location). Here*

$$\begin{aligned} L(x) &= \frac{p_{\vartheta_1}(x)}{p_{\vartheta_0}(x)} = \exp\left(-\frac{(x-\vartheta_1)^2 - (x-\vartheta_0)^2}{2}\right) \\ &= \exp\left(x(\vartheta_1 - \vartheta_0)\right) \exp\left(-\frac{\vartheta_1^2 - \vartheta_0^2}{2}\right) \end{aligned}$$

Now for any $t > 0$

$$\begin{aligned} P_{\vartheta_0}\left(L(X) = t\right) &= P_{\vartheta_0}\left(X(\vartheta_1 - \vartheta_0) - \frac{\vartheta_1^2 - \vartheta_0^2}{2} = \log t\right) \\ &= P_{\vartheta_0}\left(X = t_0(\vartheta_1, \vartheta_0, t)\right) \end{aligned}$$

*where t_0 is a well defined number if $\vartheta_1 \neq \vartheta_0$. However for a normal X this probability is 0 for any t_0. Moreover $L(X) = 0$ cannot happen since $\exp(z) > 0$ for all z, so that assumption **L** is fulfilled if $\vartheta_1 \neq \vartheta_0$.*

Example 8.3.4 *Let $U(a, b)$ be the uniform law on the interval $[a, b]$ and $P_{\vartheta_0} = U(0, 1)$, $P_{\vartheta_1} = U(0, a)$ where $0 < a < 1$. Then $p_{\vartheta_1}(x) = a^{-1}$ for $x \in [0, a]$, 0 otherwise, and*

$$L(x) = p_{\vartheta_1}(x) = \begin{cases} a^{-1} \text{ if } x \in [0, a] \\ 0 \text{ otherwise.} \end{cases}.$$

Thus if X has law P_{ϑ_0} then L takes only values a^{-1} and 0, and

$$\begin{aligned} P(L(X) &= a^{-1}) = a \\ P(L(X) &= 0) = 1 - a \end{aligned}$$

Thus F_L is not continuous; it jumps at 0 and a^{-1} is constant at other points.

In the latter example we cannot guarantee the existence of a Neyman-Pearson test. We will remedy this situation later; let us first prove the optimality of Neyman-Pearson tests (abbreviated N-P tests henceforth). We do not claim that there is only one N-P test for a given level α (the c may not be unique, and also one can use different versions of the densities, e.g. modifying them in some points etc.).

Theorem 8.3.5 ***(Neyman-Pearson fundamental lemma)**. In model Model $\mathbf{M}_c$, for $\Theta = \{\vartheta_0, \vartheta_1\}$, under assumption **L**, any N-P test of level α ($0 < \alpha < 1$) is a most powerful α-test for the hypotheses $H : \vartheta = \vartheta_0$ vs. $K : \vartheta = \vartheta_1$.*

Proof. Let ϕ be a N-P test and S_{NP} its acceptance region

$$S_{NP} = \{x : p_{\vartheta_1}(x) \leq c\, p_{\vartheta_0}(x)\}.$$

We can assume $c > 0$ here since

$$1 - \alpha = P_{\vartheta_0}\left(L(X) \leq c\right)$$

and for $c = 0$, $\alpha < 1$ this would mean that $P_{\vartheta_0}\left(L(X) = 0\right) > 0$, which we excluded by assumption **L**. Let ψ be any α-test and S its acceptance region. We have to show

$$P_{\vartheta_1}(S) \geq P_{\vartheta_1}(S_{NP}).$$

Define $A = S \setminus S_{NP}$, $A' = S_{NP} \setminus S$; then

$$\begin{aligned} S &= (S \cap S_{NP}) \cup A, \\ S_{NP} &= (S \cap S_{NP}) \cup A'. \end{aligned}$$

and it suffices to show

$$P_{\vartheta_1}(A) \geq P_{\vartheta_1}\left(A'\right). \tag{8.4}$$

Now since ψ is an α-test, we have

$$P_{\vartheta_0}(S) \geq 1 - \alpha = P_{\vartheta_0}(S_{NP})$$

which implies

$$P_{\vartheta_0}(A) \geq P_{\vartheta_0}\left(A'\right). \tag{8.5}$$

Since $A \subseteq S_{NP}^c$, we have for any $x \in A$ that $p_{\vartheta_1}(x) > c\, p_{\vartheta_0}(x)$, hence

$$P_{\vartheta_0}(A) = \int_A p_{\vartheta_0}(x)dx \leq c^{-1} \int_A p_{\vartheta_1}(x)dx = c^{-1} P_{\vartheta_1}(A), \tag{8.6}$$

and since $A' \subseteq S_{NP}$, we have

$$P_{\vartheta_0}\left(A'\right) \geq c^{-1} \int_{A'} p_{\vartheta_1}(x)dx = c^{-1} P_{\vartheta_1}\left(A'\right). \tag{8.7}$$

Relations (8.7), (8.6), (8.5) imply (8.4). ■

Let us consider the case where Assumption L is not fulfilled. For any test, the quantity $P_{\vartheta_0}(\phi(X) = 1)$ is called the **size** of the test. We saw in the above proof that the assumption that the Neyman-Pearson test ϕ has size exactly α was essential. Recall that

$$P_{\vartheta_0}(\phi(X) = 1) = 1 - F_L(c)$$

and that $F_L(c)$ is continuous from the right. When Assumption L is not fulfilled, the following situation may occur: there is a c_0 such that for all $c < c_0$, $F_L(c) < 1 - \alpha$, and at c_0 we have $F_L(c_0) > 1 - \alpha$. In other words, the function $1 - F_L(c)$ jumps in such a way that α is not attained. In order to deal with this situation, let us generalize the notion of a test function.

Definition 8.3.6 *A **randomized test** ϕ (based on the data X) is any statistic such that $0 \leq \phi(X) \leq 1$.*

When the value of ϕ is between 0 and 1, the interpretation is that the decision between hypothesis H and alternative K is taken randomly, such that ϕ is the probability of deciding K. Thus, given the data $X = x$, a Bernoulli random variable Z is generated with law (conditional on x) $\mathcal{L}(Z) = B(1, \phi(x))$, and the decision is Z. The former nonrandomized test functions are special cases: when $\phi(x) = 1$ or $\phi(x) = 0$, the Bernoulli r.v. Z is degenerate and takes the corresponding value 1 or 0 with probability one. For a randomized test, we have for $\vartheta = \vartheta_0$ or $\vartheta = \vartheta_1$, and writing $P_\vartheta(Z = \cdot)$ for the unconditional probability in Z (when X is random)

$$\begin{aligned} P_\vartheta(Z = 1) &= E_\vartheta P(Z = 1 | X = x) = E_\vartheta \phi(X), \\ P_\vartheta(Z = 0) &= 1 - E_\vartheta \phi(X) \end{aligned} \tag{8.8}$$

so that both the errors of first and second kind are a function of the expected value of $\phi(X)$ under the respective hypothesis.
This method of introducing artificial randomness into the decision process should be regarded with **common sense reservations** from a practical point of view. However randomized tests provide a completion of theory and therefore a better understanding of the basic problems of statistics. For instance, inclusion of randomized tests allows to state that there is always a level α test: take $\phi(X) = \alpha$, independently of the data. But the power of that trivial test is also α, i. e. not very good.
In the above situation, when there is no c such that $F_L(c) = 1 - \alpha$, consider the left limit of F_L at c_0:

$$F_{L,-}(c_0) := \lim_{c \nearrow c_0} F_L(c) = P_{\vartheta_0}\left(L(X) < c_0\right)$$

(this always exists for monotone functions), then the height of the jump of F_L at c_0 is the probability that the r.v. $L(X)$ takes the value c_0:

$$\begin{aligned} P_{\vartheta_0}\left(L(X) = c_0\right) &= F_L(c_0) - F_{L,-}(c_0) \\ &= P_{\vartheta_0}\left(L(X) \leq c_0\right) - P_{\vartheta_0}\left(L(X) < c_0\right) > 0. \end{aligned}$$

Define

$$\gamma_\alpha = \frac{\alpha - P_{\vartheta_0}\left(L(X) > c_0\right)}{P_{\vartheta_0}\left(L(X) = c_0\right)},$$

then

$$\alpha = P_{\vartheta_0}\left(L(X) > c_0\right) + \gamma_\alpha P_{\vartheta_0}\left(L(X) = c_0\right). \tag{8.9}$$

Moreover, since $F_{L,-}(c_0)$ is a limit of values which are all $< 1 - \alpha$,

$$F_{L,-}(c_0) \leq 1 - \alpha$$

and by assumption $F_L(c_0) > 1 - \alpha$, hence

$$0 < \gamma_\alpha = \frac{F_L(c_0) - (1 - \alpha)}{F_L(c_0) - F_{L,-}(c_0)} \leq 1.$$

That allows us to construe γ_α as the value of a randomized test ϕ, which is taken if the event $L(X) = c_0$ occurs. We can then define Neyman-Pearson tests *for any statistical model*, provided the likelihood ratio $L(X)$ is defined as a random variable. Then the distribution function F_L is defined, and we may construct a level α test as above.

Definition 8.3.7 *Assume Model* $\mathbf{M}_d$ *(discrete) or Model* $\mathbf{M}_c$ *(continuous), and that* $\Theta = \{\vartheta_0, \vartheta_1\}$. *Let* p_ϑ *be either probability functions or densities and define the likelihood ratio as a function of the data* x

$$L = L(x) = \begin{cases} p_{\vartheta_1}(x)/p_{\vartheta_0}(x) & \text{if } p_{\vartheta_0}(x) > 0 \\ +\infty & \text{otherwise} \end{cases}$$

A test ϕ *for the hypotheses*
$H : \vartheta = \vartheta_0$
$K : \vartheta = \vartheta_1$

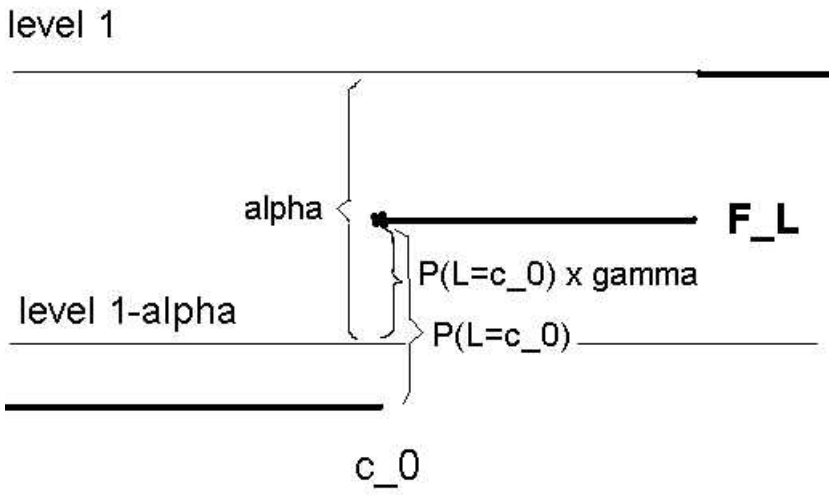


Figure 1 Construction of the randomized Neyman-Pearson test

is called a ***randomized Neyman-Pearson test*** *of level* α *if there exist* $c \in [0, \infty)$ *and* $\gamma \in [0, 1]$ *such that*

$$\phi(x) = \begin{cases} 1 \text{ if } L(x) > c \\ \gamma \text{ if } L(x) = c \\ 0 \text{ if } L(x) < c \end{cases}$$

and such that

$$P_{\vartheta_0}(L(X) > c) + \gamma P_{\vartheta_0}(L(X) = c) = \alpha. \tag{8.10}$$

Note we modified the definition of $L(x)$: formerly we took $L(x) = 0$ if $p_{\vartheta_0}(x) = 0$. But under H such values of x do not occur anyway (or with probability 0), so that F_L as considered before remains the same and a level α is attained. The modification ensures that if we *decide* on the basis of L (rejct if L is large enough), then $p_{\vartheta_0}(x) = 0$ implies that the decision is always 1.

Theorem 8.3.8 ***(Neyman-Pearson fundamental lemma, general case).*** *In Model* $\mathbf{M}_d$ *(discrete) or Model* $\mathbf{M}_c$ *(continuous), for* $\Theta = \{\vartheta_0, \vartheta_1\}$*, and any* α *(*$0 < \alpha < 1$*) a N-P test of level* α *exists and is a most powerful* α*-test (among all randomized) for the hypotheses* $H : \vartheta = \vartheta_0$ *vs.*$K : \vartheta = \vartheta_1$*.*

Proof. We have shown existence above: $L(X)$ is a well defined r.v. if X has law P_{ϑ_0} (it takes value ∞ only with probability 0); it has distribution function F_L. If c_0 exists such that $F_L(c_0) = 1 - \alpha$ then take $c = c_0$ and $\gamma = 0$. Otherwise, find c and γ as above, fulfilling (8.9). Only the properties of distribution functions were used for establishing (8.9), i.e. (8.10). Let Z be the "randomizing" random variable, which has conditional law $\mathcal{L}(Z) = B(1, \phi(x))$ given $X = x$. Then the probability under H that H is rejected is

$$\begin{aligned} P_{\vartheta_0}(Z = 1) &= E_{\vartheta_0}\phi(X) \\ &= 1 \cdot P_{\vartheta_0}(L(X) > c) + \gamma \cdot P_{\vartheta_0}(L(X) = c) = \alpha, \end{aligned} \tag{8.11}$$

i.e. the test has indeed size α. The optimality proof is analogous to Theorem 8.3.5. We assume that $p_\vartheta(x)$ are densities; the case of probability functions requires only changes in notation. According to (8.11), let ϕ be a N-P test with given c, γ and

$$\begin{aligned} S_> &:= \{x\colon\ L(x) > c\}, \\ S_= &:= \{x\colon\ L(x) = c\}, \\ S_< &:= \{x\colon\ L(x) < c\}. \end{aligned}$$

Then for any randomized α-test ψ

$$\begin{aligned} E_{\vartheta_1}\psi(X) &= \int_{S_>} \psi(x)p_{\vartheta_1}(x)dx + \int_{S_=\cup S_<} \psi(x)p_{\vartheta_1}(x)dx \\ &\leq \int_{S_>} \psi(x)p_{\vartheta_1}(x)dx + c\int_{S_=\cup S_<} \psi(x)p_{\vartheta_0}(x)dx \\ &= \int_{S_>} \psi(x)\left(p_{\vartheta_1}(x) - cp_{\vartheta_0}(x)\right)dx + c\int \psi(x)p_{\vartheta_0}(x)dx. \end{aligned}$$

The second term on the right is bounded from above by $c\alpha$, since ψ is a level α test. For the first term, since $p_{\vartheta_1}(x) - cp_{\vartheta_0}(x) > 0$ on $S_>$ and $\psi(x) \leq 1$, we obtain an upper bound by substituting 1 for ψ. Hence

$$\begin{aligned} E_{\vartheta_1}\psi(X) &\leq \int_{S_>} \left(p_{\vartheta_1}(x) - cp_{\vartheta_0}(x)\right)dx + c\alpha \\ &= \int_{S_>} \phi(x)\left(p_{\vartheta_1}(x) - cp_{\vartheta_0}(x)\right)dx + c\int \phi(x)p_{\vartheta_0}(x)dx \\ &= \int_{S_>} \phi(x)p_{\vartheta_1}(x)dx + c\int_{S_=} \phi(x)p_{\vartheta_0}(x)dx \\ &= \int_{S_>\cup S_=} \phi(x)p_{\vartheta_1}(x)dx \leq E_{\vartheta_1}\phi(X). \end{aligned}$$

■

In some cases the Neyman-Pearson lemma allows the construction of UMP tests for composite hypotheses. Consider the Gaussian location model (Model $\mathbf{M}_{c,1}$), for sample size n, and the hypotheses $H : \mu \leq \mu_0$. $K : \mu > \mu_0$. Consider the **one sided Gauss test** ϕ_{μ_0}: for the test statistic

$$Z(X) = \frac{\left(\bar{X}_n - \mu_0\right) n^{1/2}}{\sigma}$$

the test is defined by

$$\phi_{\mu_0}(X) = 1 \text{ if } Z_{\mu_0}(X) > z_\alpha \tag{8.12}$$

(0 otherwise), where z_α is the upper α -quantile of $N(0,1)$. As was argued in Example 8.3.3, condition L is fulfilled here; hence for Neyman-Pearson tests of two simple hypotheses within this model randomization is not needed. We have composite hypotheses now, but the following can be shown.

Proposition 8.3.9 *In the Gaussian location model (Model $\mathbf{M}_{c,1}$), for sample size n, for the test problem $H : \mu \leq \mu_0$ vs. $K : \mu > \mu_0$, for any $0 < \alpha < 1$ the one sided Gauss test (8.12) is a UMP α-test.*

Proof. Note first that ϕ_{μ_0} is an α-test for $H : \mu \leq \mu_0$. Indeed for any $\mu \leq \mu_0$

$$\begin{aligned} P_\mu\left(Z(X) > z_\alpha\right) &= P_\mu\left(\frac{\left(\bar{X}_n - \mu\right) n^{1/2}}{\sigma} > z_\alpha + \frac{\left(\mu_0 - \mu\right) n^{1/2}}{\sigma}\right) \\ &\leq P_\mu\left(\frac{\left(\bar{X}_n - \mu\right) n^{1/2}}{\sigma} > z_\alpha\right) = \alpha. \end{aligned}$$

Consider now any point $\mu_1 > \mu_0$. We claim that for simple hypotheses $H^* : \mu = \mu_0$, $K^* : \mu = \mu_1$ the test ϕ_{μ_0} is a Neyman-Pearson test of level α. Indeed, when p_{μ_0}, p_{μ_1} are the densities then for $x = (x_1, \ldots, x_n)$ (φ is the standard normal density)

$$\begin{aligned} L(x) &= \prod_{i=1}^{n} \frac{\varphi\left((x_i - \mu_1)/\sigma\right)}{\varphi\left((x_i - \mu_0)/\sigma\right)} = \exp\left(-\frac{\sum_{i=1}^n (x_i - \mu_1)^2 - \sum_{i=1}^n (x_i - \mu_0)^2}{2\sigma^2}\right) \\ &= \exp\left(\frac{n\bar{x}_n(\mu_1 - \mu_0)}{\sigma^2}\right) \exp\left(-\frac{n\mu_1^2 - n\mu_0^2}{2\sigma^2}\right) \end{aligned}$$

Since $\mu_1 > \mu_0$, $L(x)$ is a monotone function of $\bar{x}_n$, and $L(x) > c$ is equivalent to $\bar{x}_n > c^*$ for some c^*. In turn, $\bar{x}_n$ is a monotone function of $Z(x) = (\bar{x}_n - \mu_0)\, n^{1/2}/\sigma$, thus $L(x) > c$ is equivalent to $Z(x) > c^{**}$. We find c^{**} from the level α condition:

$$P_{\mu_0}\left(Z(X) > c^{**}\right) = \alpha,$$

and since $Z(X)$ is standard normal under μ_0, we find $c^{**} = z_\alpha$. Thus ϕ_{μ_0} is a Neyman-Pearson α-test for $H^* : \mu = \mu_0$ vs. $K^* : \mu = \mu_1$. Any test ψ of level α for the original composite hypotheses is also a test of level α for $H^* : \mu = \mu_0$,vs. $K^* : \mu = \mu_1$, so that the fundamental lemma (8.3.5) implies

$$E_{\mu_1}\psi \leq E_{\mu_1}\phi_{\mu_0}.$$

Since $\mu_1 > \mu_0$ was arbitrary, the proposition follows. ■

8.4 Likelihood ratio tests

We saw that the Neyman-Pearson lemma is closely connected to the likelihood principle: the N-P test rejects if $L(x) = p_{\vartheta_1}(x)/p_{\vartheta_0}(x)$ is too large at the observed x. The densities (or probability functions) $p_\vartheta(x)$ is a *likelihood* of ϑ when x is fixed (observed) and ϑ varies, and the decision is taken as a function of the two likelihoods. That idea can be carried over to general composite hypotheses.

Definition 8.4.1 *Assume Model* $\mathbf{M}_d$ *(discrete) or Model* $\mathbf{M}_c$ *(continuous), and that* $\Theta = \Theta_0 \cup \Theta_1$ *where* $\Theta_0 \cap \Theta_1 = \emptyset$. *Let* p_ϑ *be either probability functions or densities and define the likelihood ratio as a function of the data* x

$$L = L(x) = \begin{cases} \frac{\sup_{\vartheta\in\Theta_1} p_\vartheta(x)}{\sup_{\vartheta\in\Theta_0} p_\vartheta(x)} & \text{if } \sup_{\vartheta\in\Theta_0} p_\vartheta(x) > 0 \\ +\infty & \text{otherwise} \end{cases}$$

A (possibly randomized) test ϕ *for the hypotheses*
$H : \vartheta \in \Theta_0$

$K : \vartheta \in \Theta_1$
is called a ***likelihood ratio test*** *(LR test) if there exist* $c \in [0, \infty)$ *such that*

$$\phi(x) = \{ \begin{array}{l} 1 \text{ if } L(x) > c \\ 0 \text{ if } L(x) < c \end{array} .$$

Note that for $L(x) = c$ we made no requirement; any value $\phi(x)$ in $[0, 1]$ is possible, so that the test is possibly a randomized one. Neyman-Pearson tests are a special case for simple hypotheses. One interpretation is the following. Suppose that the suprema over both hypotheses are attained, so that for certain $\hat{\vartheta}_i(x) \in \Theta_i$, $i = 0, 1$ we have

$$\sup_{\vartheta \in \Theta_i} p_\vartheta(x) = \max_{\vartheta \in \Theta_i} p_\vartheta(x) = p_{\hat{\vartheta}_i(x)}(x),\ i = 0, 1.$$

Then $\hat{\vartheta}_i(x)$ are *maximum likelihood estimators (MLE)* of ϑ under assumptions $\vartheta \in \Theta_i$, and the LR test can be interpreted as a Neyman-Pearson test for simple hypotheses $H : \vartheta = \hat{\vartheta}_0(x)$ vs. $K : \vartheta = \hat{\vartheta}_1(x)$. Of course this is pure heuristics and none of the Neyman-Pearson optimality theory applies, since the "hypotheses" have been formed on the basis of the data.
Consider the Gaussian location-scale model and recall the form of the t-statistic for given μ_0

$$T_{\mu_0}(X) = \frac{\left(\bar{X}_n - \mu_0\right) n^{1/2}}{\hat{S}_n}.$$

The two sided t-test was already defined (cp. (8.1); it rejects when $\left|T_{\mu_0}(X)\right|$ is too large. The **one sided t-test** is the test which rejects when $T_{\mu_0}(X)$ is too large (in analogy to the one sided Gauss test for known σ^2).

Proposition 8.4.2 *Consider the Gaussian location-scale model (Model* $\mathbf{M}_{c,2}$*), for sample size* n*.*
(i) *For hypotheses* $H : \mu \leq \mu_0$ *vs.* $K : \mu > \mu_0$*, the one sided t-test is a LR test.*
(ii) *For hypotheses* $H : \mu = \mu_0$ *vs.* $K : \mu \neq \mu_0$*, the two sided t-test is a LR test.*

Proof. In relation (3.4) in the proof of Proposition 3.0.5, we obtained the following form of the density p_{μ,σ^2} of the data $x = (x_1, \ldots, x_n)$:

$$p_{\mu,\sigma^2}(x) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{S_n^2 + (\bar{x}_n - \mu)^2}{2\sigma^2 n^{-1}}\right), \tag{8.13}$$

$$S_n^2 = n^{-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2. \tag{8.14}$$

Consider first the two sided case (ii). To find MLE's of μ and σ^2 under $\mu \neq \mu_0$, we first maximize for fixed σ^2 over all possible $\mu \in \mathbb{R}$. This gives an unrestricted MLE $\hat{\mu} = \bar{x}_n$, and since $\bar{x}_n = \mu_0$ with probability 0, we obtain that $\hat{\mu}_1 = \bar{x}_n$ is the MLE of μ with probability 1 under K. We now have to maximize

$$l_x(\sigma^2) = \frac{1}{(\sigma^2)^{n/2}} \exp\left(-\frac{S_n^2}{2\sigma^2 n^{-1}}\right)$$

over $\sigma^2 > 0$. For notational convenience, we set $\gamma = \sigma^2$; equivalently, one may minimize

$$\tilde{l}_x(\gamma) = -\log l_x(\gamma) = \frac{n}{2} \log \gamma + \frac{n S_n^2}{2\gamma}.$$

Note that if $S_n^2 > 0$, for $\gamma \to 0$ we have $\tilde{l}_x(\gamma) \to \infty$ and for $\gamma \to \infty$ also $\tilde{l}_x(\gamma) \to \infty$, so that a minimum exists and is a zero of the derivative of $\tilde{l}_x$. The event $s_n^2 > 0$ has probability 1 since otherwise $x_i = \bar{x}_n$, $i = 1, \ldots, n$, i.e. all x_i are equal, which clearly has probability 0 for independent continuous x_i. We obtain

$$\begin{aligned} \tilde{l}'_x(\gamma) &= \frac{n}{2\gamma} - \frac{nS_n^2}{2\gamma^2} = 0 \\ \gamma &= S_n^2 \end{aligned}$$

as the unique zero, so $\hat{\sigma}_1^2 = s_n^2$ is the MLE of σ^2 under K. Thus

$$\begin{aligned} \max_{\mu \neq \mu_0, \sigma^2 > 0} p_{\mu,\sigma^2}(x) &= \frac{1}{(2\pi\hat{\sigma}_1^2)^{n/2}} \exp\left(-\frac{S_n^2 + (\bar{x}_n - \hat{\mu}_1)^2}{2\hat{\sigma}_1^2 n^{-1}}\right) \\ &= \frac{1}{(\hat{\sigma}_1^2)^{n/2}} \cdot \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{n}{2}\right) \end{aligned}$$

Now under H the MLE of μ is $\hat{\mu}_0 = \mu_0$. Defining

$$S_{0,n}^2 := S_n^2 + (\bar{x}_n - \mu_0)^2 = n^{-1} \sum_{i=1}^{n} (x_i - \mu_0)^2,$$

we see that the MLE of σ^2 under H is $\hat{\sigma}_0^2 = S_{0,n}^2$. Hence

$$\max_{\mu = \mu_0, \sigma^2 > 0} p_{\mu,\sigma^2}(x) = \frac{1}{(\hat{\sigma}_0^2)^{n/2}} \cdot \frac{1}{(2\pi)^{n/2}} \exp\left(-n/2\right)$$

and the likelihood ratio L is

$$\begin{aligned} L(x) &= \frac{\max_{\mu \neq \mu_0, \sigma^2 > 0} p_{\mu,\sigma^2}(x)}{\max_{\mu = \mu_0, \sigma^2 > 0} p_{\mu,\sigma^2}(x)} = \frac{(\hat{\sigma}_0^2)^{n/2}}{(\hat{\sigma}_1^2)^{n/2}} \\ &= \left(\frac{S_n^2 + (\bar{x}_n - \mu_0)^2}{s_n^2}\right)^{n/2} \end{aligned}$$

Note that

$$T_{\mu_0}(X) = \frac{n^{1/2}\left(\bar{X}_n - \mu_0\right)}{\hat{S}_n} = \frac{(n-1)^{1/2}(\bar{X}_n - \mu_0)}{S_n}$$

hence

$$L(X) = \left(1 + \frac{1}{n-1} T_{\mu_0}^2\right)^{n/2}.$$

Thus $L(x)$ is a strictly monotone increasing function of $\left|T_{\mu_0}\right|$, which proves (ii).
Consider now claim (i). For hypotheses $H : \mu \leq \mu_0$ vs. $K : \mu > \mu_0$, the one sided t-test which rejects when the t-statistic

$$T_{\mu_0}(X) = \frac{\left(\bar{X}_n - \mu_0\right) n^{1/2}}{\hat{S}_n}.$$

is too large (with a proper choice of critical value, such that an α-test results). It is easy to see that the rejection region $T_{\mu_0}(X) > z_\alpha$ where z_α is the upper α-quantile of the t_{n-1}-distribution

leads to an α-test for H (exercise). to shoe equivalence to the LR test, note that when maximizing $p_{\mu,\sigma^2}(x)$ over the alternative, the supremum is not attained ($\mu > \mu_0$ is an open interval). However the supremum is the same as the maximum over $\mu \geq \mu_0$ which is attained by certain maximum likelihod estimators $\hat{\mu}_1, \hat{\sigma}_1^2$. (we will find these, and also MLE's $\hat{\mu}_0, \hat{\sigma}_0^2$ under H).
The density p_{μ,σ^2} of the data $x = (x_1, \ldots, x_n)$ is again (8.13), (8.14). To find MLE's of μ and σ^2 under $\mu > \mu_0$, we first maximize for fixed σ^2 over all possible μ. When $\bar{x}_n > \mu_0$ the solution is $\hat{\mu} = \bar{x}_n$. When $\bar{x}_n \leq \mu_0$, the problem is to minimize $(\bar{x}_n - \mu)^2$ under a condition $\mu > \mu_0$. This minimum is not attained (μ can be selected arbitrarily close to μ_0, such that still $\mu > \mu_0$, which makes $(\bar{x}_n - \mu)^2$ arbitrarily close to $(\bar{x}_n - \mu_0)^2$, never attaining this value). However

$$\inf_{\mu > \mu_0} (\bar{x}_n - \mu)^2 = \min_{\mu \geq \mu_0} (\bar{x}_n - \mu)^2 = (\bar{x}_n - \mu_0)^2 .$$

Thus the MLE of μ under $\mu \geq \mu_0$ is $\hat{\mu}_1 = \max(\bar{x}_n, \mu_0)$. This is not the MLE under K, but gives the supremal value of the likelihood under K for given σ^2. To continue, we have to maximize in σ^2. Now

$$(\bar{x}_n - \hat{\mu}_1)^2 = (\min(0, \bar{x}_n - \mu_0))^2$$

and defining

$$S_{n,1}^2 := S_n^2 + (\bar{x}_n - \hat{\mu}_1)^2$$

we obtain

$$\sup_{\mu > \mu_0, \sigma^2 > 0} p_{\mu,\sigma^2}(x) = \sup_{\sigma^2 > 0} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{S_{n,1}^2}{2\sigma^2 n^{-1}}\right).$$

The maximization in σ^2 is now analogous to the argument above given for part (ii). The maximizing value is $\hat{\sigma}_1^2 = S_{n,1}^2$ and the maximized likelihood (which is also the supremal likelihood under K) is

$$\sup_{\mu > \mu_0, \sigma^2 > 0} p_{\mu,\sigma^2}(x) = \frac{1}{(\hat{\sigma}_1^2)^{n/2}} \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2n^{-1}}\right).$$

Now under the hypothesis, since $\mu \leq \mu_0$ is a closed interval, the MLE's can straightforwardly be found. An analogous argument to the one above gives

$$\begin{aligned}
\hat{\mu}_0 &= \min(\bar{x}_n, \mu_0), \\
\min_{\mu \leq \mu_0} (\bar{x}_n - \mu)^2 &= (\bar{x}_n - \hat{\mu}_0)^2 = (\max(0, \bar{x}_n - \mu_0))^2 , \\
\hat{\sigma}_0^2 &= S_{n,0}^2 \text{ where } S_{n,0}^2 := S_n^2 + (\bar{x}_n - \hat{\mu}_0)^2 , \\
\sup_{\mu \leq \mu_0, \sigma^2 > 0} p_{\mu,\sigma^2}(x) &= \frac{1}{(\hat{\sigma}_0^2)^{n/2}} \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2n^{-1}}\right).
\end{aligned}$$

Thus the likelihood ratio is

$$L(x) = \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2}\right)^{n/2} = \left(\frac{S_n^2 + (\bar{x}_n - \hat{\mu}_0)^2}{S_n^2 + (\bar{x}_n - \hat{\mu}_1)^2}\right)^{n/2}$$

Suppose first that the t-statistic $T_{\mu_0}(X)$ has values ≤ 0; this is equivalent to $\bar{x}_n \leq \mu_0$. In this case $\hat{\mu}_0 = \bar{x}_n$, $\hat{\mu}_1 = \mu_0$, hence

$$\begin{aligned} L(x) &= \left(\frac{S_n^2}{S_n^2 + (\bar{x}_n - \mu_0)^2}\right)^{n/2} = \left(\frac{1}{1 + (\bar{x}_n - \mu_0)^2 / S_n^2}\right)^{n/2} \\ &= \left(\frac{1}{1 + (T_{\mu_0}(X))^2/(n-1)}\right)^{n/2}. \end{aligned}$$

Thus for nonpositive values of $T_{\mu_0}(X)$, the likelihood ratio $L(x)$ is a monotone decreasing function of the absolute value of $T_{\mu_0}(X)$, which means it is *monotone increasing in* $T_{\mu_0}(X)$, on values $T_{\mu_0}(X) \leq 0$.
Consider now nonnegative values of $T_{\mu_0}(X)$: $T_{\mu_0}(X) \geq 0$. Then $\bar{x}_n \geq \mu_0$, hence $\hat{\mu}_0 = \mu_0$, $\hat{\mu}_1 = \bar{x}_n$ and

$$\begin{aligned} L(x) &= \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2}\right)^{n/2} = \left(\frac{S_n^2 + (\bar{x}_n - \mu_0)^2}{S_n^2}\right)^{n/2} \\ &= \left(1 + (T_{\mu_0}(X))^2/(n-1)\right)^{n/2}. \end{aligned}$$

Thus for values $T_{\mu_0}(X) \geq 0$, the likelihood ratio $L(x)$ is monotone increasing in $T_{\mu_0}(X)$.
The two areas of values of $T_{\mu_0}(X)$ we considered do overlap (in $T_{\mu_0}(X) = 0$); and we showed that $L(x)$ is a monotone increasing function of $T_{\mu_0}(X)$ on both of these. Hence $L(x)$ is a monotone increasing function of $T_{\mu_0}(X)$. ■

Chapter 9
CHI-SQUARE TESTS

9.1 Introduction

Consider the following problem related to Mendelian heredity. Two characteristica of pea plants (phenotypes) are observed: *form*, which may be smooth or wrinkled, and *color*, which may be yellow or green. Thus there are 4 combinations of form and color (4 combined phenotypes). Mendelian theory predicts certain frequencies of these in the total population of pea plants; call them $M_1, \ldots, M_4$ (here M_1 is the frequency of smooth/yellow etc); assume these are normed as $\sum_{j=1}^4 M_j = 1$. We observe a sample of n pea plants; let the observed frequency be Z_j for each phenotype ($j = 1, \ldots, 4$), then $\sum_{j=1}^4 Z_k = n$. We wish to find out whether these observations support the *Mendelian hypothesis* that $(M_1, \ldots, M_4)$ are the frequencies of phenotypes in the total population.

Model $\mathbf{M}_{d,2}$ The observed random vector $Z = (Z_1, \ldots, Z_k)$ has a multinomial distribution $\mathfrak{M}_k(n, \mathbf{p})$ with unknown probability vector $\mathbf{p} = (p_1, \ldots, p_k)$ ($\sum_{j=1}^k p_j = 1$, $p_j \geq 0$, $j = 1, \ldots, k$).

Recall the basic facts about the multinomial law. Consider a random k-vector Y of form $(0, \ldots, 0, 1, 0, \ldots, 0)$ where exactly one component is 1 and the others are 0. The probability that the 1 is at position j with is p_j; thus Y can describe an individual falling into one of k categories (phenotypes in the above example). This Y is said to have law $\mathfrak{M}_k(1, \mathbf{p})$. If $Y_1, \ldots, Y_n$ are i.i.d. with law $\mathfrak{M}_k(1, \mathbf{p})$ then $Z = \sum_{i=1}^n Y_i$ has the law $\mathfrak{M}_k(n, \mathbf{p})$. (The Y_i may be called *counting vectors).* The probability function is

$$P\left(Z = (z_1, \ldots, z_k)\right) = \frac{n!}{\prod_{j=1}^k z_j!} \prod_{j=1}^k p_j^{z_j} \tag{9.1}$$

where $\sum_{j=1}^k z_j = n$, $z_j \geq 0$ integer. Since the j-th component of Y_1 is Bernoulli $B(1, p_j)$, the j-th component Z_j of Z has binomial law $B(n, p_j)$. The Z_j are not independent; in fact $\sum_{j=1}^k Z_j = n$. For $k = 2$ all the information is in Z_1 since $Z_2 = n - Z_1$; thus for $k = 2$ observing a multinomial $\mathfrak{M}_2(n, (p_1, p_2))$ is equivalent to observing a binomial $B(n, p_1)$.
In Model $\mathbf{M}_{d,2}$ consider the hypotheses
$H : \mathbf{p} = \mathbf{p}_0$
$K : \mathbf{p} \neq \mathbf{p}_0$.
The test we wish to find thus is a significance test. Recall the basic rationale of hypothesis testing: what we wish to statistically ascertain at level α is K; if K is accepted then it can be claimed that "the deviation from the null hypothesis* H is statistically significant"; and there can be reasonable confidence in the truth of K. On the contrary, when H is accepted, no statistical significance

*The hypothesis H is often called the "null hypothesis", even if it is not of the form "$\vartheta = 0$"

claim can be attached to this result. When formulating a test problem, the statement for which "reasonable statistical certainty" is desired is taken as K.
Let us find the likelihood ratio test for this problem. Setting $\vartheta = \mathbf{p}$, $\vartheta_0 = \mathbf{p}_0 = (p_{0,1}, \ldots, p_{0,k})$, denoting $p_\vartheta(z)$ the probability function (9.1) of $\mathfrak{M}_k(n, \mathbf{p})$ and

$$\Theta = \left\{ \mathbf{p} : p_j \geq 0, \sum_{j=1}^{k} p_j = 1 \right\}, \ \Theta_1 = \Theta \setminus \{\vartheta_0\}$$

we obtain the likelihood ratio statistic

$$L(z) = \frac{\sup_{\vartheta \in \Theta_1} p_\vartheta(z)}{p_{\vartheta_0}(z)} = \frac{\sup_{\mathbf{p} \in \Theta_1} \prod_{j=1}^{k} p_j^{z_j}}{\prod_{j=1}^{k} p_{0,j}^{z_j}}.$$

Consider the numerator; let us first maximize over $\mathbf{p} \in \Theta$ (this will be justified below). If some of the z_j are 0, we can set the corresponding $p_j = 0$ (making the other p_j larger; we set $0^0 = 1$). We now maximimze over p_j such that $z_j > 0$. Taking a logarithm, we have to maximize

$$\sum_{j : z_j > 0} z_j \log p_j$$

over all $\mathbf{p} \in \Theta$. Since $\log x \leq x - 1$, we have

$$\begin{aligned} \sum_{j : z_j > 0} z_j \log \frac{p_j}{n^{-1} z_j} &\leq \sum_{j : z_j > 0} z_j \left(\frac{p_j}{n^{-1} z_j} - 1 \right) = n \sum_{j : z_j > 0} p_j - \sum_{j : z_j > 0} z_j \\ &= 0 \end{aligned}$$

so that

$$\sum_{j : z_j > 0} z_j \log p_j \leq \sum_{j : z_j > 0} z_j \log n^{-1} z_j$$

and for $p_j = n^{-1} z_j$ equality is attained. This is the unique maximizer since $\log x = x - 1$ only for $x = 1$. We proved

Proposition 9.1.1 *In the multinomial Model* $\mathbf{M}_{d,2}$, *with* $\mathcal{L}(Z) = \mathfrak{M}_k(n, \mathbf{p})$ *with no restriction on the parameter* $\mathbf{p}$ *the maximum likelihood estimator* $\hat{\mathbf{p}}$ *is*

$$\hat{\mathbf{p}}(Z) = n^{-1} Z.$$

The interpretation is that $\hat{\mathbf{p}}$ is a vector valued sample mean of the counting vectors $Y_1, \ldots, Y_n$. In this sense, we have a generalization of the result for binomial observations (Proposition 3.0.3).
Recall that for the LR statistic we have to find the supremum over $\Theta_1 = \Theta \setminus \{\mathbf{p}_0\}$, i.e. only one point $\mathbf{p}_0$ is taken out. Since the target function $\mathbf{p} \mapsto \prod_{j=1}^{k} p_j^{z_j}$ is continuous (with $0^0 = 1$) on Θ, we have

$$\begin{aligned} L(z) &= \frac{\sup_{\vartheta \in \Theta_1} p_\vartheta(z)}{p_{\vartheta_0}(z)} = \frac{\sup_{\vartheta \in \Theta} p_\vartheta(z)}{p_{\vartheta_0}(z)} \\ &= \frac{\max_{\vartheta \in \Theta} p_\vartheta(z)}{p_{\vartheta_0}(z)} = \prod_{j=1}^{k} \left(\frac{n^{-1} z_j}{p_{0,j}} \right)^{z_j}. \end{aligned}$$

Since the logarithm is a monotone function, the acceptance region S (complement of critical / rejection region) can also be written

$$S = \left\{ z : \log\left(L(z)\right)^{-1} = \sum_{j=1}^{k} z_j \log \frac{np_{0,j}}{z_j} \geq c \right\}.$$

Even the logarithm is a relatively involved function of the data, so it is difficult to find its distribution under H and to determine the critical value c from that. We will use a Taylor approximation of the logarithm to simplify it. The basis is the observation that the estimator $\hat{\mathbf{p}}(Z)$ is *consistent*, i.e. converges in probability to the true probability vector $\mathbf{p}$

$$\hat{\mathbf{p}}(Z) = n^{-1}Z \to_p \mathbf{p}.$$

Under the hypothesis, this true vector is $\mathbf{p}_0$, so all values $n^{-1}z_j/p_{0,j}$ converge to one. Note the Taylor expansion

$$\log(1+x) = x - \frac{x^2}{2} + o(x^2) \text{ as } x \to 0$$

where $o(x^2)$ is a term which is of smaller order than x^2 (such that $o(x^2)/x^2 \to 0$). Thus, assuming that each term $p_{0,j}/n^{-1}z_j - 1$ is small, we obtain

$$\begin{aligned} \log\left(L_1(z)\right)^{-1} &= \sum_{j=1}^{k} z_j \log\left(1 + \frac{p_{0,j}}{n^{-1}z_j} - 1\right) \\ &\approx \sum_{j=1}^{k} z_j \left(\frac{p_{0,j}}{n^{-1}z_j} - 1\right) - \frac{1}{2}\sum_{j=1}^{k} z_j \left(\frac{p_{0,j}}{n^{-1}z_j} - 1\right)^2. \end{aligned}$$

Here the first term on the right vanishes, since the $p_{0,j}$ sum to one and the z_j sum to n. We obtain

$$\log\left(L_1(z)\right)^{-1} \approx -\sum_{j=1}^{k} \frac{1}{2} z_j \left(\frac{p_{0,j}}{n^{-1}z_j} - 1\right)^2.$$

We need not make the approximation "≈" more rigorous, if we do not insist on using the likelihood ratio test. In fact we will use the LR principle only to find a reasonable test (which should be shown to have asymptotic level α). In this spirit, we proceed with another approximation $n^{-1}z_j \approx p_{0,j}$ to obtain

$$\begin{aligned} -\frac{1}{2}\sum_{j=1}^{k} z_j \left(\frac{p_{0,j}}{n^{-1}z_j} - 1\right)^2 &= -\frac{1}{2}\sum_{j=1}^{k} z_j \left(\frac{p_{0,j} - n^{-1}z_j}{n^{-1}z_j}\right)^2 \\ &\approx -\frac{1}{2}\sum_{j=1}^{k} \frac{n\left(p_{0,j} - n^{-1}z_j\right)^2}{p_{0,j}}. \end{aligned}$$

Definition 9.1.2 *In the multinomial Model* $\mathbf{M}_{d,2}$*, with* $\mathcal{L}(Z) = \mathfrak{M}_k(n, \mathbf{p})$*, the* χ^2***-statistic*** *relative to a given parameter vector* $\mathbf{p}_0$ *is*

$$\chi^2(Z) = \sum_{j=1}^{k} \frac{\left(n^{1/2}\left(p_{0,j} - n^{-1}Z_j\right)\right)^2}{p_{0,j}}.$$

The name is derived from the asymptotic distribution of this statistic, which we will establish below (the statistic does not have a χ^2-distribution). The hypothesis $H : \mathbf{p} = \mathbf{p}_0$ will be rejected if $\chi^2(Z)$ is too large; as shown above, that idea was obtained from the likelihood ratio principle.
But the $\chi^2(Z)$ has an interpretation of its own, as a measure of deviation from the hypothesis. Indeed $n^{-1}Z_j$ are consistent estimators of the true parameter $\mathbf{p}$, so the sum of squares $\sum_{j=1}^{k}(p_{0,j}-p_j)^2$ can be seen as a measure of departure from H. In the chi-square statistic, we have a weighted sum of squares with weights $p_{0,j}^{-1}$.
We know that since each Z_j has a marginal binomial law, for each j we have a convergence in distribution

$$n^{1/2}\left(n^{-1}Z_j - p_{0,j}\right) \overset{\mathcal{L}}{\Rightarrow} N(0, p_{0,j}(1-p_{0,j})) \tag{9.2}$$

i.e. has a limiting normal law by the CLT under H. The χ^2 distribution is a sum of squares of independent normals. However the Z_j are not independent in the multinomial law; so we need more than the CLT for each Z_j: in fact a **multivariate CLT** for the joint law of $(Z_1, \ldots, Z_k)$ is required.

9.2 The multivariate central limit theorem

Recall the Central Limit Theorem (CLT) for i.i.d. real valued random variables (Example 7.3.4. Recall also that since there are no further conditions on the law of Y_1, the CLT is valid for both continuous and discrete Y_i. Here the symbol $\overset{\mathcal{L}}{\Rightarrow}$ refers to *convergence in distribution (or in law),* see Definition 7.3.1.
Suppose now that $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are i.i.d. random vectors of dimension k; we also write $\mathbf{Y}$ for a random vector which has the distribution of $\mathbf{Y}_1$. Assume that

$$E\left\|\mathbf{Y}\right\|^2 < \infty. \tag{9.3}$$

Here $\|\mathbf{Y}\|^2 = \sum_{j=1}^{k} Y_j^2$ is the Euclidean norm of the random vector $\mathbf{Y}$. Recall that for any random vector $\mathbf{Y}$ in $\mathbb{R}^k$, the covariance matrix is defined by

$$\left(\mathrm{Cov}(\mathbf{Y})\right)_{j,l} = \mathrm{Cov}(Y_j, Y_l) = E\left(Y_j - EY_j\right)\left(Y_l - EY_l\right),\ 1 \le j, l \le k.$$

if the expectations exist. (Here $(A)_{i,j}$ is the $(i.j)$ entry of a matrix A). This existence is guaranteed by condition (9.3), as a consequence of the Cauchy-Schwarz inequality:

$$\begin{aligned} \left|\mathrm{Cov}(Y_j, Y_l)\right|^2 &\le \mathrm{Var}(Y_j)\mathrm{Var}(Y_l) \le \left(EY_j^2\right)\left(EY_l^2\right) \\ &\le \left(E\left\|\mathbf{Y}\right\|^2\right)^2. \end{aligned}$$

Note that for expectations of vectors and matrices, the following convention holds: the expectation of a vector (matrix) is the vector (matrix) of expectations. This means

$$E\mathbf{Y} = \left(EY_j\right)_{j=1,\ldots,k}$$

and since the $k \times k$-matrix whose components are Y_jY_l can be written

$$\mathbf{Y}\mathbf{Y}^\top = \left(Y_jY_l\right)_{j=1,\ldots,k}^{l=1,\ldots,k},$$

the covariance matrix can be written

$$\mathrm{Cov}(\mathbf{Y}) = E\left(\mathbf{Y}-E\mathbf{Y}\right)\left(\mathbf{Y}-\mathbf{E}\mathbf{Y}\right)^{\top}.$$

Our starting point for the multivariate CLT is the observation that if $\mathbf{t}$ is a nonrandom k-vector, then the r.v.'s $\mathbf{t}^{\top}\mathbf{Y}_i$, $i = 1, \ldots, n$ are real-valued i.i.d. r.v.'s with finite second moment.Indeed, since for any vector $\mathbf{t}, \mathbf{x}$ we have

$$\left(\mathbf{t}^{\top}\mathbf{x}\right)^2 = \mathbf{t}^{\top}\mathbf{x}\mathbf{x}^{\top}\mathbf{t},$$

we obtain

$$\begin{aligned}\mathrm{Var}(\mathbf{t}^{\top}\mathbf{Y}) &= E\left(\mathbf{t}^{\top}\mathbf{Y}-E\mathbf{t}^{\top}\mathbf{Y}\right)^2 = E\left(\mathbf{t}^{\top}(\mathbf{Y}-E\mathbf{Y})\right)^2 = \\ &= E\mathbf{t}^{\top}(\mathbf{Y}-E\mathbf{Y})(\mathbf{Y}-E\mathbf{Y})^{\top}\mathbf{t} \\ &= \mathbf{t}^{\top}E(\mathbf{Y}-E\mathbf{Y})(\mathbf{Y}-E\mathbf{Y})^{\top}\mathbf{t} = \mathbf{t}^{\top}\mathrm{Cov}(\mathbf{Y})\mathbf{t} < \infty\end{aligned}$$

since $\mathrm{Cov}(\mathbf{Y})$ is a finite matrix. Thus according to the (univariate) CLT, if

$$\sigma_{\mathbf{t}}^2 = \mathbf{t}^{\top}\mathrm{Cov}(\mathbf{Y})\mathbf{t}$$

is not 0, then

$$n^{1/2}\left(n^{-1}\sum_{i=1}^{n}\mathbf{t}^{\top}\mathbf{Y}_i - E\mathbf{t}^{\top}\mathbf{Y}\right) \stackrel{\mathcal{L}}{\Rightarrow} N(0, \sigma_{\mathbf{t}}^2). \tag{9.4}$$

Define the sample mean of the random vectors $\mathbf{Y}_i$ by

$$\bar{\mathbf{Y}}_n = n^{-1}\sum_{i=1}^{n}\mathbf{Y}_i;$$

then (9.4) can be written

$$n^{1/2}\mathbf{t}^{\top}\left(\bar{\mathbf{Y}}_n - E\bar{\mathbf{Y}}_n\right) \stackrel{\mathcal{L}}{\Rightarrow} N(0, \mathbf{t}^{\top}\mathrm{Cov}(\mathbf{Y})\mathbf{t}).$$

This suggests a multivariate normal distribution $N_k(\mathbf{0}, \Sigma)$ with $\Sigma = \mathrm{Cov}(\mathbf{Y})$ as the limit law for the vector $n^{1/2}\left(\bar{\mathbf{Y}}_n - E\bar{\mathbf{Y}}_n\right)$. Indeed recall that for a nonsingular (positive definite) matrix Σ

$$\mathcal{L}(\mathbf{Z}) = N_k(\mathbf{0}, \Sigma) \text{ implies that } \mathcal{L}(\mathbf{t}^{\top}\mathbf{Z}) = N(0, \mathbf{t}^{\top}\Sigma\mathbf{t}) \text{ for every } \mathbf{t} \neq \mathbf{0}.$$

(Lemma 6.1.12. Actually the converse is also true; cf. Proposition 9.2.3 below). In the sequel, for the multivariate CLT we will impose the condition that $\mathrm{Cov}(\mathbf{Y})$ is nonsingular. This means that $\mathbf{t}^{\top}\mathrm{Cov}(\mathbf{Y})\mathbf{t} > \mathbf{0}$ for every $\mathbf{t} \neq \mathbf{0}$.
For an interpretation of this condition, note if it is violated then there exists a $\mathbf{t} \neq \mathbf{0}$ such that $\mathbf{t}^{\top}\mathrm{Cov}(\mathbf{Y})\mathbf{t} = \mathbf{0}$ (this number is the variance of a random variable and thus it cannot be negative). Then the r.v. $\mathbf{t}^{\top}\mathbf{Y}$ is 0 with probability 1. Define the hyperplane (linear subspace) in $\mathbb{R}^k$

$$\mathcal{H} = \left\{\mathbf{x} \in \mathbb{R}^k : \mathbf{t}^{\top}\mathbf{x} = \mathbf{0}\right\};$$

then $\mathbf{Y} \in \mathcal{H}$ with probability one. The condition that $\mathrm{Cov}(\mathbf{Y})$ is nonsingular thus excludes this case of a "degenerate" random vector which is actually concentrated on a linear subspace of $\mathbb{R}^k$. However a multivariate CLT is still possible if the vector $\mathbf{Y}$ is linearly transformed (to a space of lower dimension).

Definition 9.2.1 *Let Q_n, $n = 1, 2, \ldots$ be a sequence of distributions in $\mathbb{R}^k$. The Q_n are said to* ***converge in distribution*** *to a limit Q_0 if for random vectors $\mathbf{Y}_n$ such that $\mathcal{L}(\mathbf{Y}_n) = Q_n$, $\mathcal{L}(\mathbf{Y}_0) = Q_0$*

$$\begin{aligned} \mathbf{t}^\top \mathbf{Y}_n \stackrel{\mathcal{L}}{\Rightarrow} Q_{0,\mathbf{t}} \text{ as } n &\to \infty, \\ Q_{0,\mathbf{t}} &= \mathcal{L}(\mathbf{t}^\top \mathbf{Y}_0), \end{aligned}$$

for every $\mathbf{t} \in \mathbb{R}^k$, $\mathbf{t} \neq \mathbf{0}$. In this case one writes

$$\mathbf{Y}_n \stackrel{\mathcal{L}}{\Rightarrow} Q_0.$$

Note that it is not excluded here that the limit law has a singular covariance matrix ($\mathbf{t}^\top \mathbf{Y}_0$ might be 0 with probability one for certain $\mathbf{t}$, or even for all $\mathbf{t}$). However for the multivariate CLT we will exclude this case by assumption, since we did not systematically treat the multivariate normal $N_k(\mathbf{0}, \Sigma)$ with singular Σ.
With this definition, it is not immediately clear that the limit law Q_0 is unique. It is desirable to have this uniqueness; otherwise there could be two different limit laws Q_0, Q_0^* such that (for $\mathcal{L}(\mathbf{Y}_0^*) = Q_0^*$)

$$\mathcal{L}(\mathbf{t}^\top \mathbf{Y}_0) = \mathcal{L}(\mathbf{t}^\top \mathbf{Y}_0^*) \text{ for all } \mathbf{t} \neq \mathbf{0}.$$

That this is not possible will follow from the Proposition 9.2.3 below.

Theorem 9.2.2 ***(Multivariate CLT)*** *Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be i.i.d. random vectors of dimension k, each with distribution Q, fulfilling condition*

$$E \|\mathbf{Y}_1\|^2 < \infty. \tag{9.5}$$

Let

$$\bar{\mathbf{Y}}_n = n^{-1} \sum_{i=1}^{n} \mathbf{Y}_i$$

and assume that the covariance matrix $\Sigma = \mathrm{Cov}(\mathbf{Y}_1)$ is nonsingular. Then for fixed Q and $n \to \infty$

$$n^{1/2} \left(\bar{\mathbf{Y}}_n - E\mathbf{Y}_1 \right) \stackrel{\mathcal{L}}{\Rightarrow} N_k(\mathbf{0}, \Sigma).$$

Proof. With our definition of convergence in distribution, it is an immediate consequence of the univariate (one dimensional) CLT and the properties of the multivariate normal distribution (following the pattern outlined above, that everything is reduced to the one dimensional case). ■
The uniqueness of the limiting law follows from the next statement. It means that we can in fact use all properties of the the multivariate normal when it is a limiting law (e.g. independence of components when they are uncorrelated).

Proposition 9.2.3 *Let Q, Q^* be the distributions of two random vectors $\mathbf{Y}$, $\mathbf{Y}^*$ with values in $\mathbb{R}^k$. Then*

$$\mathcal{L}(\mathbf{Y}) = \mathcal{L}(\mathbf{Y}^*) \text{ if and only if } \mathcal{L}(\mathbf{t}^\top \mathbf{Y}) = \mathcal{L}(\mathbf{t}^\top \mathbf{Y}^*) \text{ for all } \mathbf{t} \neq \mathbf{0}.$$

Proof. The complete argument is beyond the scope of this course; let us discuss some elements (cp. also the arguments for the proof of the univariate CLT in [D]). Suppose that for all $\mathbf{t} \in \mathbb{R}^k$, the expression

$$M_{\mathbf{Y}}(\mathbf{t}) = E \exp(\mathbf{t}^\top \mathbf{Y})$$

is finite (i. e. the expectation is finite). In that case, $M_{\mathbf{Y}}$ is called the *moment generating function* (m.g.f.) of $\mathcal{L}(\mathbf{Y})$. Analogously to the one dimensional case, it can be shown that the m. g.f. determines $\mathcal{L}(\mathbf{Y})$ uniquely (that is the key argument). Thus if $\mathcal{L}(\mathbf{t}^\top \mathbf{Y}) = \mathcal{L}(\mathbf{t}^\top \mathbf{Y}^*)$ then their univariate m. g.f. coincide:

$$E \exp(u\mathbf{t}^\top \mathbf{Y}) = E \exp(u\mathbf{t}^\top \mathbf{Y}^*)$$

and conversely, if that is the case for all u and $\mathbf{t}$ then $M_{\mathbf{Y}} = M_{\mathbf{Y}^*}$, hence $\mathcal{L}(\mathbf{Y}) = \mathcal{L}(\mathbf{Y}^*)$.
Existence of the m.g.f. is a strong additional assumption on a distribution. The proof in the general case (without any conditions on the laws $\mathcal{L}(\mathbf{Y}), \mathcal{L}(\mathbf{Y}^*)$) is based on the so called *characteristic function* of a random vector

$$\phi_{\mathbf{Y}}(\mathbf{t}) = E \exp(i\mathbf{t}^\top \mathbf{Y})$$

where the complex-valued expression

$$\exp(iz) = \cos(z) + i \sin(z)$$

occurs (for $z = \mathbf{t}^\top \mathbf{Y}$). Since

$$|\exp(iz)| = 1$$

(absolute value for complex numbers), no special strong assumptions have to be made for the existence of the characteristic function: it exists for any random vector and can be used in the proof in much the same way as above. The essential part is again that $\phi_{\mathbf{Y}}$ uniquely determines $\mathcal{L}(\mathbf{Y})$. ■
Let A be a subset of $\mathbb{R}^k$. A set is called *regular* if it has a volume and a boundary of zero volume (the boundary is the intersection of $\bar{A}$ and $\overline{A^c}$, where A^c is the complement and $\bar{A}$ is the closure of A). Rectangles and balls are regular. The following statement is similar to Proposition 9.2.3, in the sense that advanced mathematical tools are neeeded for its proof, and we only quote it here.

Proposition 9.2.4 *Let $\mathbf{Y}_n$ be a sequence of random vectors in $\mathbb{R}^k$ such that*

$$\mathbf{Y}_n \stackrel{\mathcal{L}}{\Rightarrow} Q_0 \text{ as } n \to \infty$$

where Q_0.is a continuous law in $\mathbb{R}^k$ (Q_0 has a density). Then

$$P(\mathbf{Y}_n \in A) \to Q_0(A)$$

for all regular sets $A \subset \mathbb{R}^k$.

9.3 Application to multinomials

Let us apply the multivariate CLT to the multinomial random vector $\mathbf{Z}$. Since the components are linearly dependent (they sum to n), we cannot expect a nonsingular covariance matrix. Recall that if $\mathcal{L}(\mathbf{Z}) = \mathfrak{M}_k(n, \mathbf{p})$ then $\mathbf{Z} = \sum_{i=1}^n \mathbf{Y}_i$ where $\mathbf{Y}_i$ are independent $\mathfrak{M}_k(1, \mathbf{p})$. If $\mathcal{L}(\mathbf{Y}) = \mathfrak{M}_k(1, \mathbf{p})$ then for $(Y_1, \ldots, Y_k)^\top = \mathbf{Y}$

$$EY_j = p_j,$$

and for $j = l$, since Y_j is binomial,

$$EY_jY_l = EY_j^2 = EY_j = p_j,$$

while for $j \neq l$

$$EY_jY_l = P\left(Y_j = 1, Y_l = 1\right) = 0$$

(the random vector $\mathbf{Y}$ has a 1 in exactly one position). We can now write down the covariance matrix:

$$\mathrm{Cov}(Y_jY_l) = EY_jY_l - EY_jEY_l = \left\{ \begin{array}{l} p_j - p_jp_l,\, j = l \\ -p_jp_l,\, j \neq l. \end{array} \right.$$

Introduce transformed variables $\tilde{Y}_j = (Y_j - p_j)/p_j^{1/2}$; then

$$\mathrm{Cov}(\tilde{Y}_j\tilde{Y}_l) = p_j^{-1/2}p_l^{-1/2}\mathrm{Cov}(Y_jY_l) = \left\{ \begin{array}{l} 1 - p_j^{1/2}p_l^{1/2},\, j = l \\ -p_j^{1/2}p_l^{1/2},\, j \neq l. \end{array} \right.$$

Let Λ be the diagonal matrix with diagonal elements $p_j^{-1/2}$. Then, for vectors

$$\tilde{\mathbf{Y}} = \Lambda(\mathbf{Y} - \mathbf{p}),\, \tilde{\mathbf{p}} = \Lambda\mathbf{p} = (p_1^{1/2}, \ldots, p_k^{1/2})^\top$$

we have $(\tilde{Y}_1, \ldots, \tilde{Y}_k)^\top = \tilde{\mathbf{Y}}$ and

$$\mathrm{Cov}(\tilde{\mathbf{Y}}) = I_k - \tilde{\mathbf{p}}\tilde{\mathbf{p}}^\top. \tag{9.6}$$

Let $e_1, \ldots, e_k$ be an orthonormal basis of vectors in $\mathbb{R}^k$ such that $e_k = \tilde{\mathbf{p}}$. That is possible since $\tilde{\mathbf{p}}$ has length 1 :

$$\|\tilde{\mathbf{p}}\|^2 = \sum_{j=1}^{k} \left(p_k^{1/2}\right)^2 = 1.$$

Then $e_1, \ldots, e_{k-1}$ are all orthogonal to $\tilde{\mathbf{p}}$. Let G, G_0 be matrices of dimension $k \times k$, $(k-1) \times k$

$$G = \left(\begin{array}{c} e_1^\top \\ \ldots \\ e_k^\top \end{array} \right), G_0 = \left(\begin{array}{c} e_1^\top \\ \ldots \\ e_{k-1}^\top \end{array} \right). \tag{9.7}$$

These are orthogonal matrices. Moreover we have

$$e_k^\top\tilde{\mathbf{Y}} = \tilde{\mathbf{p}}^\top\tilde{\mathbf{Y}} = \sum_{j=1}^{k} p_j^{1/2}p_j^{-1/2}(Y_j - p_j) \tag{9.8}$$

$$= \sum_{j=1}^{k}(Y_j - p_j) = 1 - 1 = 0. \tag{9.9}$$

Lemma 9.3.1 *Let $\mathcal{L}(\mathbf{Z}) = \mathfrak{M}_k(n, \mathbf{p})$, let Λ be the diagonal matrix with diagonal elements $p_j^{-1/2}$ and let the $(k-1) \times k$-matrix F be defined by*

$$F = G_0\Lambda$$

where G_0 is defined by (9.7). Then

$$\|F(\mathbf{Z}-n\mathbf{p})\|^2=\sum_{j=1}^{k}p_j^{-1}(Z_j-np_j)^2 \tag{9.10}$$

and the random vector $F\mathbf{Z}$ has covariance matrix:

$$\mathrm{Cov}(F\mathbf{Z})=nI_{k-1}. \tag{9.11}$$

Proof. Note that

$$\Lambda(\mathbf{Z}-n\mathbf{p})=\sum_{i=1}^{n}\tilde{\mathbf{Y}}_i$$

where $\mathbf{Y}_i$ are i.i..d $\mathfrak{M}_k(1,\mathbf{p})$ and

$$\tilde{\mathbf{Y}}_i=\Lambda(\mathbf{Y}_i-\mathbf{p}).$$

Above it was shown (9.9) that

$$e_k^\top\tilde{\mathbf{Y}}_i=\tilde{\mathbf{p}}^\top\tilde{\mathbf{Y}}_i=0,\ i=1,\dots,n.$$

Hence

$$e_k^\top\Lambda(\mathbf{Z}-n\mathbf{p})=0.$$

This implies

$$\begin{aligned}\sum_{j=1}^{k}p_j^{-1}(Z_j-np_j)^2 &= \|\Lambda(\mathbf{Z}-n\mathbf{p})\|^2=\|G\Lambda(\mathbf{Z}-n\mathbf{p})\|^2\\ &= \sum_{j=1}^{k}\left(e_j^\top\Lambda(\mathbf{Z}-n\mathbf{p})\right)^2=\sum_{j=1}^{k-1}\left(e_j^\top\Lambda(\mathbf{Z}-n\mathbf{p})\right)^2\\ &= \|G_0\Lambda(\mathbf{Z}-n\mathbf{p})\|^2=\|F(\mathbf{Z}-n\mathbf{p})\|^2\end{aligned}$$

thus the first claim (9.10) is proved. For the second claim, we note that in view of (9.6) and the additivity of covariance matrices for independent vectors

$$\begin{aligned}\mathrm{Cov}(F\mathbf{Z}) &= n\mathrm{Cov}(F\mathbf{Y}_1)=n\mathrm{Cov}(G_0\tilde{\mathbf{Y}}_1)\\ &= nG_0\left(I_k-\tilde{\mathbf{p}}\tilde{\mathbf{p}}^\top\right)G_0^\top\\ &= nG_0G_0^\top=nI_{k-1}.\end{aligned}$$

(computation rules for covariance matrices can be obtained from the rules for the multivariate normal, cp. Lemma 6.1.12:

$$\mathrm{Cov}(A\mathbf{X})=A\mathrm{Cov}(\mathbf{X})A^\top.$$

■

The following is the Central Limit Theorem for a multinomial random variable, which generalizes the de Moivre-Laplace CLT for binomials (sums of i.i..d Bernoulli r.v's, cp. (1.2) and (9.2)). Since the components of the multinomial are dependent, we need to multiply with a $(k-1)\times k$-matrix F first, otherwise we would get a multivariate normal limiting distribution with singular covariance matrix.

Proposition 9.3.2 *Let $\mathcal{L}(\mathbf{Z}) = \mathfrak{M}_k(n, \mathbf{p})$. Then for the $(k-1) \times k$-matrix F defined above we have*

$$n^{1/2} F\left(n^{-1}\mathbf{Z} - \mathbf{p}\right) \stackrel{\mathcal{L}}{\Rightarrow} N_{k-1}(\mathbf{0}, I_k) \text{ as } n \to \infty.$$

Proof. We can represent $\mathbf{Z}$ as a sum of i.i.d. vectors

$$\mathbf{Z} = \sum_{i=1}^{n} \mathbf{Y}_i$$

where each $\mathbf{Y}_i$ is $\mathfrak{M}_k(1, \mathbf{p})$. Hence

$$F\mathbf{Z} = \sum_{i=1}^{n} F\mathbf{Y}_i$$

the $F\mathbf{Y}_i$ are again i.i.d. vectors expectation $F\mathbf{p}$ and with unit covariance matrix, according to (9.11) for $n = 1$. For the multivariate CLT, the second moment condition is fulfilled trivially since the vector $\mathbf{Y}_1$ takes only k possible values. The multivariate CLT (Theorem 9.2.2) yields the result. ■

The next result justifies the name of the χ^2-statistic, by establishing an asymptotic distribution.

Theorem 9.3.3 *Let $\mathcal{L}(\mathbf{Z}) = \mathfrak{M}_k(n, \mathbf{p})$. Then for the χ^2-statistic*

$$\chi^2(\mathbf{Z}) = \sum_{j=1}^{k} \frac{\left(n^{1/2}\left(n^{-1}Z_j - p_j\right)\right)^2}{p_j}$$

we have

$$\chi^2(\mathbf{Z}) \stackrel{\mathcal{L}}{\Rightarrow} \chi^2_{k-1} \text{ as } n \to \infty.$$

Proof. We have according to (9.10)

$$\begin{aligned} \chi^2(\mathbf{Z}) &= n^{-1} \sum_{j=1}^{k} p_j^{-1} (Z_j - np_j)^2 \\ &= n^{-1} \left\| F(\mathbf{Z} - n\mathbf{p}) \right\|^2 = \left\| n^{1/2} F(n^{-1}\mathbf{Z} - \mathbf{p}) \right\|^2. \end{aligned}$$

The above Proposition 9.3.2 implies that the expression inside $\|\cdot\|^2$ is asymptotically multivariate $k-1$-standard normal. Denote this expression

$$\mathbf{V}_n = n^{1/2} F(n^{-1}\mathbf{Z} - \mathbf{p}).$$

For convergence in law to χ^2_{k-1} we have to show that

$$P\left(\|\mathbf{V}_n\|^2 \leq t\right) \to F(t)$$

where F is the distribution function of the law χ^2_{k-1}, at every continuity point t of F. Since this law has a density, F is continuous, so it has to be shown for every t (it suffices for $t \geq 0$). The set

$\left\{\mathbf{x} : \|\mathbf{x}\|^2 \leq t\right\}$ is a ball in $\mathbb{R}^{k-1}$ and hence regular in the sense of Proposition 9.2.4. Thus if ξ is a random vector with law $N_{k-1}(\mathbf{0}, I_k)$ then

$$P\left(\|\mathbf{V}_n\|^2 \leq t\right) \to P\left(\|\xi\|^2 \leq t\right).$$

By definition of the χ^2_{k-1} distribution, we have

$$P\left(\|\xi\|^2 \leq t\right) = F(t)$$

so that the theorem follows from Proposition 9.2.4. ■

9.4 Chi-square tests for goodness of fit

In a **goodness-of-fit test** we test whether the distribution of the data is different from a given specified distribution Q_0. More generally, this term is used also when we test whether the distribution is outside a specified family (such as the normal family). In the first case, the hypothesis H is simple (consists of Q_0) and the test may also be called a significance test, but the terminology "goodness-of-fit" emphasizes that we specifically focus on the actual shape of the distribution. Goodness of fit to whole families of distributions (taken as H) will be discussed in the next section.

Theorem 9.4.1 *Consider Model* $\mathbf{M}_{d,2}$*: the observed random k-vector* $\mathbf{Z}$ *has law* $\mathcal{L}(\mathbf{Z}) = \mathfrak{M}_k(n, \mathbf{p})$ *where* $\mathbf{p}$ *is unknown. Consider the hypotheses*
$H : \mathbf{p} = \mathbf{p}_0$
$K : \mathbf{p} \neq \mathbf{p}_0$.
Let z_α *be the upper* α*-quantile of the distribution* χ^2_{k-1}*. The test* $\varphi(\mathbf{Z})$ *defined by*

$$\varphi(\mathbf{Z}) = \begin{cases} 1 \text{ if } \chi^2(\mathbf{Z}) > z_\alpha \\ 0 \text{ otherwise} \end{cases} \tag{9.12}$$

where

$$\chi^2(\mathbf{Z}) = \sum_{j=1}^{k} \frac{(np_{0,j} - Z_j)^2}{np_{0,j}}. \tag{9.13}$$

is the χ^2*-statistic relative to* H*, is an asymptotic* α*-test, i. e. under* $\mathbf{p} = \mathbf{p}_0$

$$\text{limsup}_{n\to\infty} P\left(\varphi(\mathbf{Z}) = 1\right) \leq \alpha.$$

The form (9.13) of the χ^2-statistic is easy to memorize: take observed frequency minus expected frequency, square it, and divide by expected frequency, and sum over components (components are also called *cells*).

Remark 9.4.2 ***On quantiles***. In the literature (especially in tables) it is more common to use **lower quantiles**; i.e. values q_γ such that for a given random variable X

$$P\left(X \leq q_\gamma\right) \geq \gamma,\ P\left(X \geq q_\gamma\right) \geq 1 - \gamma.$$

For $\gamma = 1/2$ one obtains a **median** (i.e. a theoretical median of the random variable X; note that quantiles q_γ need not be unique). When X has a continuous strictly monotone distribution function F (at least in a neighborhood of q_γ) then q_γ is the unique value

$$F(q_\gamma) = \gamma.$$

Lower quantiles are often written in the form $\chi^2_{k;\gamma}$ if F corresponds to χ^2_k. Thus for the upper quantile z_α used above we have

$$z_\alpha = \chi^2_{k-1;1-\alpha}.$$

Example 9.4.3 Consider again the heredity example (beginning of Section 9.1). Suppose the values of $(M_1, \ldots, M_4) = (p_{0,1}, \ldots, p_{0,4})$ predicted by the theory are

$$\begin{aligned}(p_{0,1}, \ldots, p_{0,4}) &= \frac{1}{100^2}\left(91^2, 9\cdot 91, 9\cdot 91, 9^2\right)\\ &= (0.828\,1, 0.0819, 0.0819, 0.00\,81)\end{aligned}$$

(we do not claim that these are the correct values corresponding to Mendelian theory). Suppose we have 1000 observations with observed frequency vector

$$\mathbf{Z} = (822, 96, 75, 7)$$

(these are hypothetical, freely invented data). We have

$$\chi^2(\mathbf{Z}) = \sum_{j=1}^{4} \frac{(np_{0,j} - Z_j)^2}{np_{0,j}}$$

$$\begin{aligned}&= \frac{(828\,.1 - 822)^2}{828\,.1} + \frac{(81.9 - 96)^2}{81.9} + \frac{(81.9 - 75)^2}{81.9} + \frac{(8.1 - 7)^2}{8.1}\\ &= 3.\,2031\end{aligned}$$

At significance level $\alpha = 0.05$, we find $z_\alpha = \chi^2_{3;0.95} = 7.82$. The hypothesis is not rejected.

Example 9.4.4 Suppose we have a die and want to test whether it is fair, i.e. all six outcomes are equally probable. For n independent trials, the frequency vector for outcomes $(1, \ldots, 6)$ is

$$\mathbf{Z} = (Z_1, \ldots, Z_6)$$

and the expected frequency vector would be

$$n\mathbf{p}_0 = \left(\frac{n}{6}, \ldots, \frac{n}{6}\right).$$

This can easily be simulated on a computer; in fact one would then test whether the computer die (i.e. a uniform random number taking values on integers $\{1, \ldots, 6\}$) actually has the uniform distribution. The random number generator of QBasic (an Ms-Dos Basic version) was tested in this way, with $n = 10000$, and a result

$$\mathbf{z} = (1686, 1707, 1739, 1583, 1643, 1642).$$

We have $n/6 = 1666.667$ and a value for the χ^2-statistic

$$\chi^2(\mathbf{z}) = \sum_{j=1}^{6} \frac{(n/6 - z_j)^2}{n/6} = 9.2408.$$

The quantile at $\alpha = 0.05$ is $z_\alpha = \chi^2_{5;0.95} = 11.07$, so the hypothesis of a uniform distribution cannot be rejected.

Exercise. Suppose you are intent on proving that the number generator is bad, and run the above simulation program 20 times. You claim that the random number generator is bad when the test rejects at least once. Are you still doing an α-test ? (assuming that n above is large enough so that the level of *one* test is practically α)

The χ^2-test can also be used to test hypotheses that the data follow a specific distribution, not necessarily multinomial. Suppose observations are i.i.d.real valued $X_1, \ldots, X_n$, with distribution Q. Suppose Q_0 is a specific distribution, and consider hypotheses

$H : Q = Q_0$

$K : Q \neq Q_0$.

This is transformed into multinomial hypotheses by selecting a partition of the real line into subsets or cells $A_1, \ldots, A_k$

$$\bigcup_{j=1}^{k} A_k = \mathbb{R},\ A_i \cap A_j = \emptyset,\ j \neq i\ .$$

The A_j are often called **cells** or **bins**; they are usually intervals. For a real r.v. X, define an indicator vector

$$\mathbf{Y}(X) = (Y_1, \ldots, Y_k),\ Y_j = \mathbf{1}_{A_j}(X),\ j = 1, \ldots, k \tag{9.14}$$

i.e. $\mathbf{Y}(X)$ indicates into which of the k cells the r.v. X falls. Then obviously $\mathbf{Y}(X)$ has a multinomial distribution:

$$\mathcal{L}(\mathbf{Y}(X)) = \mathfrak{M}_k(1, \mathbf{p}),\ \mathbf{p} = \mathbf{p}(Q) = (Q(A_1), \ldots, Q(A_k))$$

This $\mathbf{p}(Q)$ is the vector of **cell probabilities**, corresponding to the given partition. Thus, in the above problem, the vectors $\mathbf{Y}_i = \mathbf{Y}(X_i)$ are multinomial; they are sometimes called **binned data**. Then

$$\mathbf{Z} := \sum_{i=1}^{n} \mathbf{Y}(X_i) \tag{9.15}$$

is multinomial $\mathfrak{M}_k(n, \mathbf{p})$ with the above value of $\mathbf{p}$. When Q takes the value Q_0 then also the vector of cells probabilities takes the value

$$\mathbf{p}_0 = \mathbf{p}(Q_0) = (Q_0(A_1), \ldots, Q_0(A_k)).$$

From initial hypotheses H, K one obtains derived hypotheses

$H' : \mathbf{p}(Q) = \mathbf{p}_0(Q)$

$K' : \mathbf{p}(Q) \neq \mathbf{p}_0(Q)$.

From Theorem 9.4.1 it is clear that as $n \to \infty$, the χ^2-test based $\mathbf{Z}$ for the multinomial hypothesis H' is again an asymptotic α-test.

Corollary 9.4.5 *Suppose observations are i.i.d. real valued $X_1, \ldots, X_n$, with distribution Q. Consider hypotheses*

$H : Q = Q_0$

$K : Q \neq Q_0$.

For a partition of the real line into nonintersecting cells $A_1, \ldots, A_k$, define the vector of cell frequencies $\mathbf{Z}$ by (9.15). Then the χ^2-test $\varphi(\mathbf{Z})$ defined by (9.12) based $\mathbf{Z}$ is an asymptotic α-test as $n \to \infty$.

This χ^2-test has very wide range of applicability; it is not specified whether Q_0 is discrete or continuous. *Every* distribution Q_0 gives rise to a specific multinomial distribution $\mathfrak{M}_k(n, \mathbf{p}(Q_0))$ which is then tested. For instance, a random number generator for standard normal variables can be tested in this way. On the real line, at least one of the cells A_j contains an unbounded interval. However there is a certain arbitrariness involved in the choice of the cells $A_1, \ldots, A_k$. In fact partitioning the data into groups amounts to a "coarsening" of the hypothesis: there are certainly distributions $Q \neq Q_0$ which have the same cell probabilities, i.e. $\mathbf{p}(Q_0) = \mathbf{p}(Q)$. These cannot be told apart from Q_0 by this method. If one choses a large number of groups k, the number of observations in each cell may be small, so that the approximation based on the CLT appears less credible.

Remark 9.4.6 *A family of probability distributions $\mathcal{P} = \{P_\vartheta, \vartheta \in \Theta\}$, indexed by ϑ, is called* ***parametric*** *if all $\vartheta \in \Theta$ are finite dimensional vectors ($\Theta \subseteq \mathbb{R}^k$ for some k), otherwise $\mathcal{P}$ is called* ***nonparametric****. In hypothesis testing, any hypothesis corresponds to some $\mathcal{P}$, thus the terminology is extended to hypotheses. Any simple hypothesis (consisting of* one *probability distribution $\mathcal{P} = \{Q_0\}$) is parametric. In Corollary 9.4.5, the alternative $K : Q \neq Q_0$ is nonparametric: the set of all distributions $Q = \mathcal{L}(X_1)$ cannot be parametrized by a finite dimensional vector (take e.g. only all discrete distributions $\neq Q_0$ characterized by probabilitities $q_1, \ldots, q_r$, r arbitrarily large).*

Thus we encountered the first nonparametric hypothesis, in the form of the alternative $Q \neq Q_0$. In this sense, the χ^2-test for goodness of fit in Corollary 9.4.5 is a *nonparametric test*; in a narrower sense this term is used for tests which have level α on a nonparametric hypothesis. (However this χ^2-test actually tests the hypothesis on the cell probabilities $\mathbf{p}(Q) = \mathbf{p}(Q_0)$, with asymptotic level α, and the set of all Q fulfilling this hypothesis is also nonparametric).

9.5 Tests with estimated parameters

Back in the multinomial model, consider now the situation that the hypothesis is not $H : \mathbf{p} = \mathbf{p}_0$ but also composite: let $\mathcal{H}$ be a $d+1$-dimensional linear subspace of $\mathbb{R}^k$ $(0 \leq d \leq k-1)$and assume the hypothesis is $H : \mathbf{p} \in \mathcal{H}$. An example would be the hypothesis $p_1 = p_2$ (when $k > 2$). Now $\mathbf{p}$ is already in a $k-1$ dimensional affine manifold:

$$\mathcal{S}_P = \left\{ \mathbf{x} : \mathbf{1}^\top \mathbf{x} = 1,\, x_j \geq 0,\, j = 1, \ldots, k \right\}, \tag{9.16}$$

which is called the **probability simplex** in $\mathbb{R}^k$. It is the set of all k-dimensional probability vectors. (Here $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^k$). Instead of fixing $\mathbf{p}_0$ as before, we now have only linear restrictions on $\mathbf{p}$: if $\mathcal{H}^\top$ is the orthogonal complement of $\mathcal{H}$ and $\mathbf{h}_1, \ldots, \mathbf{h}_{k-d-1}$ an orthonormal basis then

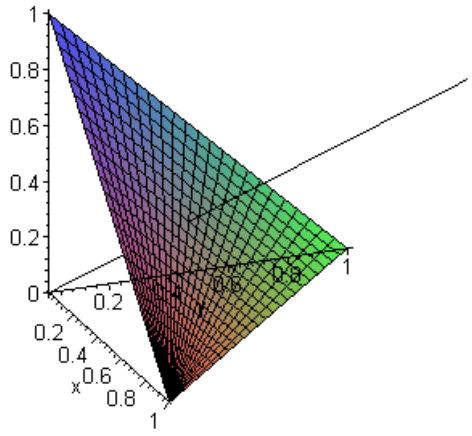
$$\mathbf{h}_j^\top \mathbf{p} = \mathbf{0},\, j = 1, \ldots, k-d-1. \tag{9.17}$$

If there is a probability vector in $\mathcal{H}$, then $\mathbf{h}_1, \ldots, \mathbf{h}_{k-d-1}, \mathbf{1}$ must be linearly independent (otherwise $\mathbf{1}$ would be a linear combination of $\mathbf{h}_j$, and we cannot have $\mathbf{1}^\top \mathbf{p} = 1$), and it follows that

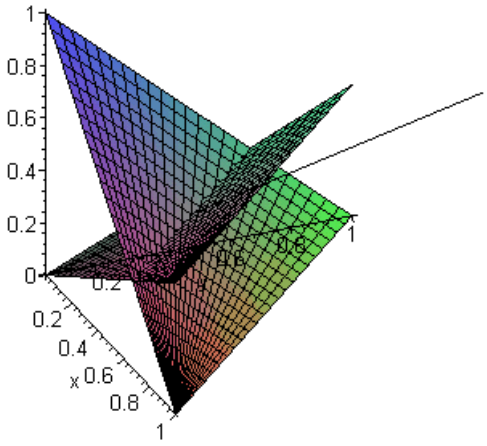
$$\mathcal{H}_0 = \mathcal{S}_P \cap \mathcal{H} \tag{9.18}$$

has dimension $k - (k - d - 1 + 1) = d$. (The dimension of $\mathcal{H}_0$ should be understood in the sense that there are $d+1$ points $\mathbf{x}_0, \ldots, \mathbf{x}_d$ in $\mathcal{H}_0$ such that $\mathbf{x}_j - \mathbf{x}_0$, $j = 1, \ldots, k$ are linearly independent, and no more such points. Simlarly, $\mathcal{S}_P$ has dimension $k-1$).When $d = 0$ then $\mathcal{H}_0$ consists of only

one point $\mathbf{p}_0$. The setting can be can be visualized as follows.



The probability simplex for $k = 3$ intersected with the linear space spanned by $\mathbf{1}$ (dimension 1). The intersection is $n^{-1}\mathbf{1}$ (dimension $d = 0$).



The probability simplex intersected with a linear subspace $\mathcal{H}$ (dimension 2). The intersection is $\mathcal{H}_0$ (dimension $d = 1$).

The multinomial data vector $n^{-1}\mathbf{Z}$ also takes values in $\mathcal{S}_P$, which means intuitively that there are $k - 1$ "degrees of freedom". Our parameter vector $\mathbf{p}$ varies in $\mathcal{H}_0$ with dimension d, which means that there are d "free" parameters under the hypothesis which must be estimated. We now claim

that the corresponding χ^2-statistic has a limiting χ^2-distribution with degrees of freedom

$$\dim(\mathcal{S}_P) - \dim(\mathcal{H}_0) = (k-1) - d = k - d - 1.$$

Let us discuss what we mean by estimated parameters. A guiding principle is still the likelihood ratio principle: consider the LR statistic

$$L(z) = \frac{\sup_{\vartheta \in \Theta_1} p_\vartheta(z)}{\sup_{\vartheta \in \Theta_0} p_\vartheta(z)} \tag{9.19}$$

where Θ_1, Θ_0 are the parameter spaces under H, K respectively. In the case $\Theta_0 = \{\mathbf{p}_0\}$ this led us to the χ^2-statistic relative to $\mathbf{p}_0$

$$\chi^2(Z) = \sum_{j=1}^{k} \frac{\left(n^{1/2}\left(p_{0,j} - n^{-1}Z_j\right)\right)^2}{p_{0,j}}.$$

Since now in (9.19) we also have to maximize over the hypothesis, we should expect that in place of $p_{0,j}$ we now obtain estimated values under the hypothesis: $\hat{\mathbf{p}} = \hat{\mathbf{p}}(Z)$, which are the maximum likelihood estimators under H.
Write $o_p(n^{-r})$ for a random vector such that $n^r \left\| o_p(n^{-r}) \right\| \to_p 0.$, $r \geq 0$.

Lemma 9.5.1 *The MLE $\hat{\mathbf{p}}$ in a multinomial model $\{\mathfrak{M}_k(n, \mathbf{p}),\ \mathbf{p} \in \mathcal{H}_0\}$ for a parameter space $\mathcal{H}_0$ given by (9.18) fulfills*

$$F\left(\hat{\mathbf{p}} - \mathbf{p}\right) = \Pi F\left(n^{-1}\mathbf{Z} - \mathbf{p}\right) + o_p(n^{-1/2}) \tag{9.20}$$

where F is the $(k-1) \times k$-matrix defined in Lemma (9.3.1), $\mathbf{p}$ is the true parameter vector, and Π is a $(k-1) \times (k-1)$ projection matrix of rank d.

Comment. The result means that the $k-1$-vector $F\left(\hat{\mathbf{p}} - \mathbf{p}\right)$ "almost" lies in the d-dimensional linear subspace of $\mathbb{R}^{k-1}$ associated to the projection Π. This is related to the fact that both $\hat{\mathbf{p}}$, $\mathbf{p}$ are in the d-dimensional manifold $\mathcal{H}_0$.
Proof. We present only a sketch, suppressing some technical arguments. Let $\mathbf{p}$ denote the true value and $\hat{\mathbf{p}}$ the one over which one maximizes (and ultimately the maximizing value). Consider the log-likelihood; up to an additive term which does not depend on $\hat{\mathbf{p}}$ it is

$$\sum_{j=1}^{k} Z_j \log \hat{p}_j$$

Maximizing this is the same as minimizing

$$-2\sum_{j=1}^{k} Z_j \log\left(\frac{\hat{p}_j}{n^{-1}Z_j}\right).$$

A preliminary argument (which we cannot give here) shows that $\hat{\mathbf{p}}$ is consistent, i.e. $\hat{\mathbf{p}} \to_p \mathbf{p}$. Also $n^{-1}\mathbf{Z} \to_p \mathbf{p}$, so that $\hat{p}_j / n^{-1}Z_j \to_p 1.$. A Taylor expansion of the logarithm yields

$$\begin{aligned} & -2\sum_{j=1}^{k} Z_j \log\left(1 + \frac{\hat{p}_j}{n^{-1}Z_j} - 1\right) \\ = \quad & -2\sum_{j=1}^{k} Z_j \left(\frac{\hat{p}_j}{n^{-1}Z_j} - 1\right) + \sum_{j=1}^{k} Z_j \left(\frac{\hat{p}_j}{n^{-1}Z_j} - 1\right)^2 + o_p(1). \end{aligned}$$

Here the first term vanishes and the remainder is

$$\sum_{j=1}^{k} \frac{n\left(n^{-1}Z_j - \hat{p}_j\right)^2}{n^{-1}Z_j} + o_p(1).$$

In another approximation step, $n^{-1}Z_j$ is replaced by its limit p_j to yield

$$\sum_{j=1}^{k} \frac{n\left(n^{-1}Z_j - \hat{p}_j\right)^2}{p_j} + o_p(1).$$

The above expression is one which $\hat{\mathbf{p}}$ minimizes. Similarly to (9.10) we now have

$$\sum_{j=1}^{k} p_j^{-1}(n^{-1}Z_j - \hat{p}_j)^2 = \left\|F(n^{-1}\mathbf{Z} - \hat{\mathbf{p}})\right\|^2$$

since with the choice of the vector e_k and Λ as in Lemma (9.3.1) we have $\Lambda e_k = \Lambda\tilde{\mathbf{p}} = \Lambda^2\mathbf{p} = \mathbf{1}$ and

$$e_k^\top \Lambda(n^{-1}\mathbf{Z} - \hat{\mathbf{p}}) = 1 - 1 = 0.$$

Furthermore, denoting

$$\hat{\mathbf{q}} = F(\hat{\mathbf{p}} - \mathbf{p}),\ \mathbf{Z}^* = F(n^{-1}\mathbf{Z} - \mathbf{p}) \tag{9.21}$$

we obtain

$$\left\|F(n^{-1}\mathbf{Z} - \hat{\mathbf{p}})\right\|^2 = \|\mathbf{Z}^* - \hat{\mathbf{q}}\|^2 \tag{9.22}$$

and $\hat{\mathbf{q}}$ minimizes

$$n\|\mathbf{Z}^* - \hat{\mathbf{q}}\|^2 + o_p(1). \tag{9.23}$$

Let us disregard the requirement that all components of $\hat{\mathbf{p}}$ must be nonnegative; it can be shown that since $n^{-1}\mathbf{Z} \in \mathcal{S}_P$, this requirement is fulfilled automatically for the minimizer $\hat{\mathbf{p}}$ if n is large enough. With this agreement, the vector $\hat{\mathbf{q}}$ varies in the set

$$\mathcal{H}_1 = \{F(\mathbf{x} - \mathbf{p}), \mathbf{x} \in \mathcal{S} \cap \mathcal{H}\}. \tag{9.24}$$

where

$$\mathcal{S} = \left\{\mathbf{x} : \mathbf{1}^\top\mathbf{x} = 1\right\}.$$

This set $\mathcal{H}_1$ can be described as follows. The set $\mathcal{S}$ is an affine subspace of $\mathbb{R}^k$; for the given $\mathbf{p} \in \mathcal{S}_P$ it can be represented

$$\mathcal{S} = \left\{\mathbf{p} + \mathbf{z} : \mathbf{1}^\top\mathbf{z} = 0\right\}.$$

Then since $\mathbf{p} \in \mathcal{H}$, we have

$$\begin{aligned}
\mathcal{H}_1^0 \quad &: \quad = \{(\mathbf{x} - \mathbf{p}), \mathbf{x} \in \mathcal{S} \cap \mathcal{H}\} \\
&= \quad \left\{(\mathbf{z}, \mathbf{1}^\top\mathbf{z} = \mathbf{0},\ \mathbf{z} \in \mathcal{H}\right\} \\
&= \quad \left\{z,\ \mathbf{h_j}^\top\mathbf{z} = 0, j = 1, \ldots, k - d - 1,\ \mathbf{1}^\top\mathbf{z} = 0\right\}
\end{aligned}$$

and since $\mathbf{1}$ and the $\mathbf{h_j}^\top$ are linearly independent, $\mathcal{H}_1^0$ is a d-dimensional linear subspace of $\mathbb{R}^k$. Then $\mathcal{H}_1 = \left\{F\mathbf{y},\ \mathbf{y}{\in}\mathcal{H}_1^0\right\}$ is a linear subspace of $\mathbb{R}^{k-1}$, and by the construction of the matrix F the space $\mathcal{H}_1$ has not smaller dimension than $\mathcal{H}_1^0$, i.e. also dimension d. (Indeed, if the dimension were less then there would be an nonzero element $\mathbf{z}$ of $\mathcal{H}_1^0$ orthogonal to all rows of F. Since these are orthogonal to $\mathbf{p}$ by construction, cp. (9.7), $\mathbf{z}$ must be a multiple of $\mathbf{p}$, which implies $\mathbf{1}^\top\mathbf{p} = 0$, in contradiction to $\mathbf{1}^\top\mathbf{p} = 1$). Let Π be the projection matrix onto $\mathcal{H}_1$ in $\mathbb{R}^{k-1}$. Since $\hat{\mathbf{q}}$ minimizes the distance to $\mathbf{Z}^*$ within the space $\mathcal{H}_1$, we must have (approximately)

$$\hat{\mathbf{q}} = \Pi\mathbf{Z}^*$$

i.e. $\hat{\mathbf{q}}$ is the projection of $\mathbf{Z}^*$ onto $\mathcal{H}_1$. This is already (9.20) up to the size $o_p(n^{-1/2})$ of the error term, for which a more detailed argument based on (9.23) is necessary. ■

Consider now the χ^2 statistic relative to H, with (maximum likelihood) estimated parameter $\hat{\mathbf{p}} = \hat{\mathbf{p}}(Z)$. We obtain

$$\chi^2(Z) = \sum_{j=1}^{k} \frac{n\left(\hat{p}_j(\mathbf{Z}) - n^{-1}Z_j\right)^2}{\hat{p}_j(\mathbf{Z})}.$$

To find the asymptotic distribution, we substitute the denominator by its probability limit p_j (the true parameter):

$$\begin{aligned}
\chi^2(\mathbf{Z}) &= \sum_{j=1}^{k} \frac{n\left(\hat{p}_j(\mathbf{Z}) - n^{-1}Z_j\right)^2}{p_j} + o_p(1) \\
&\approx n\left\|F(n^{-1}\mathbf{Z} - \hat{\mathbf{p}})\right\|^2 \\
&= n\left\|\mathbf{Z}^* - \hat{\mathbf{q}}\right\|^2 = n\left\|(I_{k-1} - \Pi)\mathbf{Z}^* + o_p(n^{-1/2})\right\|^2
\end{aligned}$$

according to the approximation of the Lemma above. Now the matrix $\Pi_* = I_{k-1} - \Pi$ is also a projection matrix, namely onto the orthogonal complement of $\mathcal{H}_1$, of rank $k-1-d$. It can be represented

$$\Pi_* = CC^\top$$

where C is a $(k-1-d)\times(k-1)$ orthogonal matrix (such that $C^\top C = I_{k-1-d}$). Thus

$$\chi^2(\mathbf{Z}) = \left\|n^{1/2}C^\top\mathbf{Z}^* + o_p(1)\right\|^2 + o_p(1).$$

Since

$$n^{1/2}\mathbf{Z}^* = n^{1/2}F(n^{-1}\mathbf{Z} - \mathbf{p}) \overset{\mathcal{L}}{\Rightarrow} N_{k-1}\left(\mathbf{0}, I_{k-1}\right)$$

it follows

$$n^{1/2}C^\top\mathbf{Z}^* \overset{\mathcal{L}}{\Rightarrow} N_{k-1-d}\left(\mathbf{0}, I_{k-1-d}\right).$$

This implies

$$\chi^2(\mathbf{Z}) \overset{\mathcal{L}}{\Rightarrow} \chi^2_{k-d-1}.$$

We see that *if d parameters must be estimated under H, then the degrees of freedom in the limiting χ^2 law are $k-d-1$.* We argued for a hypothesis $H : \mathbf{p} \in \mathcal{H}_0$ where $\mathcal{H}_0$ is a d-dimensional affine manifold in $\mathbb{R}^k$ described in (9.18) ($0 \le d < k-1$).

Theorem 9.5.2 *Consider Model* $\mathbf{M}_{d,2}$*: the observed random k-vector* $\mathbf{Z}$ *has law* $\mathcal{L}(\mathbf{Z}) = \mathfrak{M}_k(n, \mathbf{p})$ *where* $\mathbf{p}$ *is unknown. Let* $\mathcal{H}_0$ *be a d-dimensional set of probability vectors of form (9.18). Consider the hypotheses*
$H : \mathbf{p} \in \mathcal{H}_0$
$K : \mathbf{p} \notin \mathcal{H}_0$
Let $\chi^2_{k-d-1;1-\alpha}$ *be the lower* $1-\alpha$*-quantile of the distribution* χ^2_{k-d-1}*. The test* $\varphi(\mathbf{Z})$ *defined by*

$$\varphi(\mathbf{Z}) = \begin{cases} 1 \text{ if } \chi^2(\mathbf{Z}) > \chi^2_{k-d-1;1-\alpha} \\ 0 \text{ otherwise} \end{cases} \tag{9.25}$$

where

$$\chi^2(\mathbf{Z}) = \sum_{j=1}^{k} \frac{(n\hat{p}_j - Z_j)^2}{n\hat{p}_j}. \tag{9.26}$$

is the χ^2*-statistic and* $\hat{\mathbf{p}} = (\hat{p}_1, \ldots, \hat{p}_k)$ *is the MLE of* $\mathbf{p}$ *relative to* $\mathcal{H}_0$*, is an asymptotic* α*-test.*

Consider now a hypothesis of form $H : \mathbf{p} \in \mathcal{P}$ where $\mathcal{P}$ is a parametric family of probability vectors

$$\mathcal{P} = \{\mathbf{p}_\vartheta,\, \vartheta \in \Theta\}$$

and $\Theta \subseteq \mathbb{R}^d$. Under some smoothness conditions, and assuming that the mapping $\vartheta \mapsto \mathbf{p}_\vartheta$ is one-to-one, the set $\mathcal{P}$ can be regarded as a "smooth" subset of the probability simplex $\mathcal{S}_P$. We can assume that in every point $\mathbf{p} \in \mathcal{P}$ there is a tangent set of form $\mathcal{H}_0$ (or tangent affine subspace $\mathcal{H}_1$) which has the same dimension d. In this sense, "locally" we are back in the previous case of Theorem 9.5.2. Here "locally" means that if the MLE $\hat{\vartheta}$ of ϑ is consistent, it will point us to a vicinity of the true parameter ϑ, i.e. the true underlying probability vector $\mathbf{p}_\vartheta$, and in this vicinity we can substitute $\mathcal{P}$ by its tangent space $\mathcal{H}_0$ at $\mathbf{p}_\vartheta$. This is the outline of the proof that the χ^2-statistic with estimated parameters

$$\chi^2(\mathbf{Z}) = \sum_{j=1}^{k} \frac{\left(np_j(\hat{\vartheta}) - Z_j\right)^2}{np_j(\hat{\vartheta})} \tag{9.27}$$

where $p = (p_1(\hat{\vartheta}), \ldots, p_k(\hat{\vartheta}))$ has still a limiting χ^2-distribution with $k-d-1$ degrees of freedom. The essential condition is that d parameters are to be estimated, and $\vartheta \mapsto \mathbf{p}_\vartheta$ is smooth and one-to-one.
In conjunction with binned data, this procedure can be used to test that the data are in a specific class of distributions, not necessarily multinomial. Suppose observations are i.i.d.real valued $X_1, \ldots, X_n$, with distribution Q. Suppose $\mathcal{Q} = \{Q_\vartheta,\, \vartheta \in \Theta\}$ is a specific class of distributions, and consider hypotheses
$H : Q \in \mathcal{Q}$
$K : Q \notin \mathcal{Q}$.
This is transformed into multinomial hypotheses by selecting a partition of the real line into subsets or cells $A_1, \ldots, A_k$, as discussed above. We obtain a vector of cell probabilities

$$\mathbf{p}(Q) = (Q(A_1), \ldots, Q(A_k)).$$

The observed vector of cell frequencies $\mathbf{Z}$ is multinomial $\mathfrak{M}_k(n, \mathbf{p})$ with the above value of $\mathbf{p}$. When Q takes values inside the family $\mathcal{Q}$, then $\mathbf{p}(Q)$ also takes values inside a family $\mathcal{P}$ defined by

$$\mathcal{P} := \{\mathbf{p}(Q_\vartheta),\, \vartheta \in \Theta\}\,.$$

From initial hypotheses H, K one obtains derived hypotheses
$H' : \mathbf{p}(Q) \in \mathcal{P}$
$K' : \mathbf{p}(Q) \notin \mathcal{P}$.
Under smoothness conditions on $\mathcal{Q}$ it is clear that as $n \to \infty$, the χ^2-test based on $\mathbf{Z}$, with estimated parameters relative to the hypothesis H' is again an asymptotic α-test. Here the degrees of freedom in the limiting χ^2-distribution is $k - d - 1$ if d the family $\mathcal{Q}$ has d parameters.
It should be stressed however that the estimator $\hat{\vartheta}$ should be the MLE based on the binned multinomial data $\mathbf{Z}$, not on the original data $X_1, \ldots, X_n$. Thus, for a **test of normality,** strictly speaking, one cannot use sample mean and sample variance as estimators, but one has to get the binned data first and then estimate mean and variance from these multinomial data by maximum likelihood.

9.6 Chi-square tests for independence

Consider a bivariate random variable $\mathbf{X} = (X_1, X_2)$ taking values in the finite set of pairs (j, l), $1 \le j \le r$, $1 \le l \le s$, where $r, s \ge 2$, with probabilities

$$P\left((X_1, X_2) = (j, l)\right) = p_{jl}.$$

These probabilities give the joint distribution of (X_1, X_2),with marginal distributions

$$P\left(X_1 = j\right) = \sum_{l=1}^{s} p_{jl} =: q_{1,j},\ P\left(X_2 = l\right) = \sum_{j=1}^{r} p_{jl} =: q_{2,l}.$$

We are interested in the problem whether X_1, X_2 are independent, i.e. whether the joint distribution is the product of its marginals:

$$p_{jl} = q_{1,j} q_{2,l},\ j = 1, \ldots, r,\ l = 1, \ldots, s. \tag{9.28}$$

Suppose that there are n i.i.d observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$, all having the distribution of $\mathbf{X}$.
This can easily be transformed into a hypothesis about a multinomial distribution. Call the pairs (j, l) **cells**; it is not important that they are pairs of natural numbers- these can just be symbols for certain categories. Thus there are rs cells; they can be written as an $r \times s$-matrix. Define a counting variable $\mathbf{Y}_i$ associated to observation $\mathbf{X}_i = (X_{1i}, X_{2i})$: $\mathbf{Y}_i$ is a $r \times s$-matrix such that

$$\begin{aligned} \mathbf{Y}_i &= (Y_{i,jl})_{j=1\ldots,r}^{l=1,\ldots,s}, \\ Y_{i,jl} &= \begin{cases} 1 \text{ if } (X_{1i}, X_{2i}) = (j, l) \\ 0 \text{ otherwise.} \end{cases} \end{aligned}$$

These $\mathbf{Y}_i$ can be identified with vectors of dimension $k = rs$; when they are looked upon as vectors, they have a multinomial distribution

$$\mathcal{L}(\mathbf{Y}_i) = \mathfrak{M}_k(1, \mathbf{p})$$

where $\mathbf{p}$ is the $r \times s$-matrix of cell probabilities p_{jl}. We can also define counting vectors for each variable X_1, X_2 separately:

$$\begin{aligned} \mathbf{Y}_{1,i} &= (Y_{1,i,j})_{j=1\ldots,r}, \mathbf{Y}_{2,i} = (Y_{2,i,l})_{l=1\ldots,s}, \\ Y_{1,i,j} &= \begin{cases} 1 \text{ if } X_{1i} = j, \\ 0 \text{ otherwise.} \end{cases}, \; Y_{2,i,l} = \begin{cases} 1 \text{ if } X_{2i} = l, \\ 0 \text{ otherwise.} \end{cases}, \end{aligned}$$

Then the counting matrix $\mathbf{Y}_i$ is obtained as

$$\mathbf{Y}_i = \mathbf{Y}_{1,i}\mathbf{Y}_{2,i}^{\top}.$$

Again we have n of these observed counting vectors, and the matrix (or vector) of observed cell frequencies is

$$\mathbf{Z} = \sum_{i=1}^{n} \mathbf{Y}_i.$$

Let us stress again that we identify a $r \times s$ matrix with a rs-vector here. We write $\mathfrak{M}_{r\times s}(n, \mathbf{p})$, for the multinomial distribution of a $r \times s$-matrix with corresponding matrix of probabilities $\mathbf{p}$. (This can be identified with $\mathfrak{M}_{rs}(n, \mathbf{p})$ when $\mathbf{p}$ is construed as a vector). We can also define cell frequencies for the variables X_1, X_2 separately:

$$\mathbf{Z}_1 = \sum_{i=1}^{n} \mathbf{Y}_{1,i}, \; \mathbf{Z}_2 = \sum_{i=1}^{n} \mathbf{Y}_{2,i}. \tag{9.29}$$

The hypothesis of independence of X_1, X_2 translates into a hypothesis about the shape of the probability matrix $\mathbf{p}$: according to (9.28), for

$$\mathbf{q}_1 = (q_{1,j})_{j=1,\ldots,r}, \; , \; \mathbf{q}_2 = (q_{2,l})_{l=1,\ldots,s},$$

we have
$H : \mathbf{p} = \mathbf{q}_1\mathbf{q}_2^{\top}$,
K : "$\mathbf{p}$ is not of this form".
The hypotheses H can be written in the form $H : \mathbf{p} \in \mathcal{P}_I$ where $\mathcal{P}_I$ is a parametric family of probability vectors (the lower index I in $\mathcal{P}_I$ stands for "independence"):

$$\mathcal{P}_I = \left\{ \mathbf{q}_1\mathbf{q}_2^{\top}, \; \mathbf{q}_1 \in \mathcal{S}_{P,r}, \; \mathbf{q}_2 \in \mathcal{S}_{P,s} \right\}$$

where $\mathcal{S}_{P,r}$ is the probability simplex in $\mathbb{R}^r$. Indeed, in the case of independence we have just the two marginals, which are two probability vectors in $\mathbb{R}^r$, $\mathbb{R}^s$ respectively. These marginal probability vectors have $r-1$, and $s-1$ independent parameters respectively (the respective $r-1$, $s-1$ first components). Thus $\mathcal{P}_I$ can be smoothly parametrized by a $r+s-2$-dimensional parameter $\vartheta \in \Theta$, where Θ is a subset of $\mathbb{R}^{r+s-2}$ (but we do not make this explicit). Thus the hypotheses are
$H : \mathbf{p} \in \mathcal{P}_I$
$K : \mathbf{p} \notin \mathcal{P}_I$.
Define "marginal" cell frequencies

$$Z_{j\cdot} = \sum_{l=1}^{s} Z_{jl}, \; Z_{\cdot l} = \sum_{j=1}^{r} Z_{jl}.$$

These coincide with the components of the vectors $\mathbf{Z}_1$, $\mathbf{Z}_2$ defined in (**??**):

$$Z_{j\cdot} = \#\left\{i : X_{1i} = j\right\}, \; Z_{\cdot l} = \#\left\{i : X_{2i} = l\right\}.$$

Proposition 9.6.1 *In the multinomial Model* $\mathbf{M}_{d,2}$*, when* $\mathbf{Z}$ *is a multinomial* $r \times s$ *matrix with law* $\mathcal{L}(\mathbf{Z}) = \mathfrak{M}_{r\times s}(n, \mathbf{p})$*,* $r, s \geq 2$*, the maximum likelihood estimator* $\hat{\mathbf{p}}$ *under the hypothesis of independence* $\mathbf{p} \in \mathcal{P}_I$ *is*

$$\hat{\mathbf{p}} = \hat{\mathbf{q}}_1 \hat{\mathbf{q}}_2^\top = (\hat{q}_{1,j}\hat{q}_{2,l})_{j=1,\ldots,r}^{l=1,\ldots,s}$$

where

$$\hat{\mathbf{q}}_1 = n^{-1}\mathbf{Z}_1,\ \hat{\mathbf{q}}_2 = n^{-1}\mathbf{Z}_2$$

and $\mathbf{Z}_1$*,* $\mathbf{Z}_2$ *are the vectors of marginal cell frequencies*

$$\mathbf{Z}_1 = (Z_{j\cdot})_{j=1,\ldots,r},\ \mathbf{Z}_2 = (Z_{\cdot l})_{l=1,\ldots,s}.$$

Proof. The probability function for $\mathbf{Z}$ is (for a matrix $\mathbf{z} = (z_{jl})_{j=1,\ldots,r}^{l=1,\ldots,s}$)

$$P(\mathbf{Z} = \mathbf{z}) = C(n, \mathbf{z}) \prod_{j=1}^{r}\prod_{l=1}^{s} p_{jl}^{z_{jl}}$$

where $C(n, \mathbf{z})$ is a factor which does not depend on the parameters. Independence means $p_{jl} = q_{1,j}q_{2,l}$, so the likelihood function is

$$\begin{aligned} l(\mathbf{q}_1, \mathbf{q}_2) &= C(n, \mathbf{z}) \prod_{j=1}^{r}\prod_{l=1}^{s} q_{1,j}^{z_{jl}} q_{2,l}^{z_{jl}} \\ &= C(n, \mathbf{z}) \left(\prod_{j=1}^{r}\prod_{l=1}^{s} q_{1,j}^{z_{jl}}\right)\left(\prod_{j=1}^{r}\prod_{l=1}^{s} q_{2,l}^{z_{jl}}\right) \\ &= C(n, \mathbf{z}) \prod_{j=1}^{r} q_{1,j}^{z_{j\cdot}} \prod_{l=1}^{s} q_{2,l}^{z_{\cdot l}}. \end{aligned}$$

Now the factor $\prod_{l=1}^{s} q_{2,l}^{z_{\cdot l}}$ is the likelihood (up to a factor) for the multinomial vector $\mathbf{Z}_2$ defined in (9.29) with parameter $\mathbf{q}_2$, and $\prod_{j=1}^{r} q_{1,j}^{z_{j\cdot}}$ is proportional to the the likelihood for $\mathbf{Z}_1$. Thus maximizing over $\mathbf{q}_1, \mathbf{q}_2$ amounts to maximizing the product of two multinomial likelihoods, each in its own parameter $\mathbf{q}_1$,$\mathbf{q}_2$. The maximizer of each likelihood is the unrestricted MLE in a multinomial model for $\mathbf{Z}_1$ or $\mathbf{Z}_2$, thus according to Proposition 9.1.1

$$\hat{\mathbf{q}}_1 = n^{-1}\mathbf{Z}_1,\ \hat{\mathbf{q}}_2 = n^{-1}\mathbf{Z}_2.$$

This proves the result. ■

We can now write down the χ^2-statistic with estimated parameters (estimated under the independence hypothesis $\mathbf{p} \in \mathcal{P}_I$), according to (9.27):

$$\begin{aligned} \chi^2(\mathbf{Z}) &= \sum_{j=1}^{r}\sum_{l=1}^{s} \frac{\left(Z_{jl} - n(n^{-1}Z_{j\cdot})(n^{-1}Z_{\cdot l})\right)^2}{n(n^{-1}Z_{j\cdot})(n^{-1}Z_{\cdot l})} \\ &= \sum_{j=1}^{r}\sum_{l=1}^{s} \frac{(nZ_{jl} - Z_{j\cdot}Z_{\cdot l})^2}{nZ_{j\cdot}Z_{\cdot l}}. \end{aligned} \tag{9.30}$$

The dimension of $\mathbf{Z}$ is $k = rs$, while the number of estimated parameters is $d = r - 1 + s - 1$, so according to the result in the previous section, as $n \to \infty$

$$\begin{aligned} \chi^2(\mathbf{Z}) \overset{\mathcal{L}}{\Rightarrow} \chi^2_{k-d-1} &= \chi^2_{rs-1-(r-1)-(s-1)} \\ &= \chi^2_{(r-1)(s-1)}. \end{aligned}$$

We thus obtain an asymptotic α-test for independence.
We assumed initially that two real random variables (X_1, X_2) take discrete values (j, l), $1 \le j \le r$, $1 \le l \le s$. But obviously, when these both take real values, one can use two partitions $A_{1,j}$, $j = 1, \ldots, r$ and $A_{2,l}$, $l = 1, \ldots, s$ and define cells

$$B_{j,l} = A_{1,j} \times A_{2,l}.$$

For n i.i..d data (X_{1i}, X_{2i}), this gives rise to observed cell frequencies

$$Z_{jl} = \#\{i : (X_{1i}, X_{2i}) \in B_{j,l}\}$$

which is then a multinomial matrix $\mathbf{Z}$ as above. The hypotheses of independence of X_1 and X_2 gives rise to a derived hypotheses $\mathbf{p} \in \mathcal{P}_I$ *as* above. Thus the χ^2-statistic can again be used to test independence.
A **contingency table** is a matrix of form

	$l=1$	$\ldots$			$l=s$		
$j=1$	$\overline{\overline{Z_{11}}}$	$\overline{\overline{\quad}}$	$\overline{\overline{\quad}}$	$\overline{\overline{\quad}}$	$\overline{\overline{Z_{1s}}}$	$\vert$	$Z_{1\cdot}$
		$\ldots$		$\ldots$		$\vert$	$\ldots$
$\ldots$	Z_{j1}		Z_{jl}		Z_{js}	$\vert$	$Z_{j\cdot}$
		$\ldots$		$\ldots$		$\vert$	$\ldots$
$j=r$	Z_{r1}				Z_{rs}	$\vert$	$Z_{r\cdot}$
	$\overline{\overline{Z_{\cdot 1}}}$	$\overline{\overline{\ldots}}$	$\overline{\overline{Z_{\cdot l}}}$	$\overline{\overline{\ldots}}$	$\overline{\overline{Z_{\cdot s}}}$		

It serves as a symbolic aid in computing the χ^2-statistic (9.30). The χ^2-test for independence is also called **χ^2-test in a contingency table.**
Exercise. Test your random number generator for independence in consecutive pairs. If $N_1, N_2, \ldots$is the sequence generated then take pairs $\mathbf{X}_1 = (N_1, N_2)$, $\mathbf{X}_2 = (N_3, N_4), \ldots$, and test independence of the first from the second component. Note: if they are not independent then presumably the pairs $\mathbf{X}_i$ are also not independent, so the alternatives which one might formulate are different from the above. Still the contingency table. provides an asymptotic α-test.

Chapter 10
REGRESSION

10.1 Regression towards the mean

To introduce the term, we start with a quotation from the Merriam-Webster dictionary* for that entry.

Regression: a: *a trend or shift toward a lower or less perfect state: as a : progressive decline of a manifestation of disease* ***b*** *:... **c** : reversion to an earlier mental or behavioral level* ***d*** *: a functional relationship between two or more correlated variables that is often empirically determined from data and is used especially to predict values of one variable when given values of the others* <the regression of y on x is linear>*; specifically : a function that yields the mean value of a random variable under the condition that one or more independent variables have specified values.*

Let us explain the origin of the usage **d** in mathematical statistics. Around 1886 the biometrist Francis Galton observed the size of pea plants and their offspring. He observed an effect which can equivalently be described by observing body height of fathers and sons in the human species; since the latter is a popular example, we shall phrase his results like this† Predictably, he noticed that tall fathers tended to have tall sons and short fathers tended to have short sons. At the same time, he noticed that the sons of tall fathers tended to be shorter then their fathers, while the sons of short fathers tended to be less short. Thus the respective height of sons tended to be closer to the overall average body height of the total population. He called this effect "regression towards the mean".

This phenomenon can be confirmed in a probabilistic model, when we assume a joint normal distribution for the height of fathers X and sons Y. Let (X, Y) have a bivariate normal distribution with mean vector (a, a), where $a > 0$. This a is the average total population height; for simplicity assume it is the same for X and Y, i.e. height does not increase from one one generation to the next. It should also be assumed that X and Y have the same variance:

$$\mathrm{Var}(X) =: \sigma_X^2 = \mathrm{Var}(Y) =: \sigma_Y^2$$

and that they are positively correlated:

$$\mathrm{Cov}(X, Y) =: \sigma_{XY} > 0$$

(recall the correlation between X and Y is

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}).$$

*http://www.m-w.com

†Sources differ as to what he actually observed; the textbook has fathers and sons (p. 555), while the "Encyclopedia of Statististal Sciences" quotes Galton about seeds.

Thus

$$\mathcal{L}\begin{pmatrix} X \\ Y \end{pmatrix} = N_2(a\mathbf{1}, \mathbf{\Sigma}),\ \mathbf{\Sigma} = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}.$$

The average body height of sons, given the height of the father, is described by the *conditional expectation* $E(Y|X = x)$. To find it, we state a basic result on the conditional distribution $\mathcal{L}(Y|X = x)$ in a bivariate normal. Some special cases appeared already (Proposition 5.5.2).

Proposition 10.1.1 *Suppose that X and Y have a joint bivariate normal distribution with expectation vector $\boldsymbol{\mu} = (\mu_X, \mu_Y)^\top$ and positive definite covariance matrix $\mathbf{\Sigma}$:*

$$\mathcal{L}\begin{pmatrix} X \\ Y \end{pmatrix} = N_2(\boldsymbol{\mu}, \mathbf{\Sigma}),\ \mathbf{\Sigma} = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}. \tag{10.1}$$

Then

$$\mathcal{L}(Y|X = x) = N\left(\mu_y + \beta(x - \mu_x), \sigma_{Y|X}^2\right) \tag{10.2}$$

where

$$\begin{aligned} \beta &= \frac{\sigma_{XY}}{\sigma_X^2}, \\ \sigma_{Y|X}^2 &= \sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2}. \end{aligned}$$

Proof. Recall the form of the joint density p of X and Y: (Lemma 6.1.10: for $\mathbf{z} = (x, y)^\top$

$$p(\mathbf{z}) = p(x, y) = \frac{1}{(2\pi)\left|\Sigma\right|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{z} - \boldsymbol{\mu})\right).$$

The marginal density of x is

$$p_1(x) = \frac{1}{(2\pi)^{1/2}\sigma_X} \exp\left(-\frac{(x - \mu_x)^2}{2\sigma_X^2}\right).$$

The density of $\mathcal{L}(Y|X = x)$ (conditional density) is given by

$$p(y|x) = \frac{p(x, y)}{p_1(x)},$$

and the conditional expectation $E(Y|X = x)$ can be read off that density. Note that

$$\begin{aligned} |\mathbf{\Sigma}| &= \sigma_X^2\sigma_Y^2 - \sigma_{XY}^2, \\ \mathbf{\Sigma}^{-1} &= \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}^{-1} = \frac{1}{|\mathbf{\Sigma}|}\begin{pmatrix} \sigma_Y^2 & -\sigma_{XY} \\ -\sigma_{XY} & \sigma_X^2 \end{pmatrix}. \end{aligned}$$

For ease of notation write $\tilde{x} = x - \mu_x$, $\tilde{y} = y - \mu_y$. This gives

$$\begin{aligned} p(x, y) &= \frac{1}{(2\pi)\left|\Sigma\right|^{1/2}} \exp\left(-\frac{1}{2\left|\mathbf{\Sigma}\right|}\left(\tilde{x}^2\sigma_Y^2 - 2\tilde{x}\tilde{y}\sigma_{XY} + \tilde{y}^2\sigma_X^2\right)\right), \\ p(y|x) &= \frac{\sigma_X}{(2\pi)^{1/2}\left|\Sigma\right|^{1/2}} \exp\left(-\frac{1}{2\left|\mathbf{\Sigma}\right|}\left(\tilde{x}^2\sigma_Y^2 - 2\tilde{x}\tilde{y}\sigma_{XY} + \tilde{y}^2\sigma_X^2\right) + \frac{\left|\mathbf{\Sigma}\right|\tilde{x}^2}{2\left|\mathbf{\Sigma}\right|\sigma_X^2}\right) \end{aligned}$$

Note that

$$\begin{aligned}\tilde{x}^2\sigma_Y^2 - 2\tilde{x}\tilde{y}\sigma_{XY} + \tilde{y}^2\sigma_X^2 &= \tilde{x}^2\sigma_Y^2 + \sigma_X^2\left(\tilde{y} - \tilde{x}\sigma_{XY}\sigma_X^{-2}\right)^2 - \tilde{x}^2\sigma_{XY}^2\sigma_X^{-2} \\ &= \sigma_X^2\left(\tilde{y} - \beta\tilde{x}\right)^2 + \tilde{x}^2\sigma_{Y|X}^2, \\ |\boldsymbol{\Sigma}|\,\sigma_X^{-2} &= \sigma_Y^2 - \sigma_{XY}^2\sigma_X^{-2} = \sigma_{Y|X}^2.\end{aligned}$$

The terms involving $\tilde{x}^2$ now cancel out, and we obtain

$$p(y|x) = \frac{1}{(2\pi)^{1/2}\sigma_{Y|X}} \exp\left(-\frac{1}{2\sigma_{Y|X}^2}\left(\tilde{y} - \beta\tilde{x}\right)^2\right).$$

In view of

$$\tilde{y} - \beta\tilde{x} = y - \mu_y - \beta(x - \mu_x)$$

this proves (10.2). ■

Corollary 10.1.2 *If X, Y have a bivariate normal distribution, then*
(i) $E(Y|X = x)$ is a linear function of x
(ii) The variance of $\mathcal{L}(Y|X = x)$ (conditional variance) does not depend on x.

Definition 10.1.3 *Let X and Y have a joint bivariate normal distribution (10.1).*
(i) The quantity

$$\beta = \frac{\sigma_{XY}}{\sigma_X^2}$$

*is called the **regression coefficient** for the regression of Y on X.*
(ii) The linear function

$$y = E(Y|X = x) = \mu_y + \beta(x - \mu_x) \tag{10.3}$$

*is called the **regression function** or **regression line** (for Y on X).*

Thus β is the slope of the regression line and $\mu_y - \beta\mu_x$ is its intercept.
Note that the regression line always goes through the point (μ_x, μ_y) given by the two means. Indeed for $x = \mu_x$ in (10.3) we obtain $y = \mu_y$.
Furthermore, the absolute value of β is bounded by

$$|\beta| \leq \frac{\sigma_X\sigma_Y}{\sigma_X^2} = \frac{\sigma_Y}{\sigma_X}$$

as a consequence of the Cauchy-Schwarz inequality. For the conditional expectation we obtain

$$E(Y|X = x) = \mu_y + \beta(x - \mu_x).$$

For father / son height model of Galton, we assumed that $\mu_y = \mu_x = a$, furthermore $\sigma_Y = \sigma_X$, and positive correlation: $\sigma_{XY} > 0$. This implies

$$0 < \beta \leq 1.$$

Here equality is not possible in : since Σ is positive definite, it cannot be singular, hence $|\boldsymbol{\Sigma}| = \sigma_X^2\sigma_Y^2 - \sigma_{XY}^2 > 0$, so that

$$|\sigma_{XY}| < \sigma_X\sigma_Y$$

which means that

$$0 < \beta < 1.$$

This gives the desired mathematical explanation of Galton's "regression toward the mean". We have

$$\begin{aligned} E(Y|X &= x) = a + \beta(x - a) \\ &= (1 - \beta)a + \beta x. \end{aligned}$$

which means that $E(Y|X = x)$ is a convex combination of x and a; the average height of sons, given the height x of fathers, is always "pulled toward the mean height" a. It is less than x if $x > a$ and is greater than x if $x < a$.
It is interesting to note that an analogous phenomenon is observed for the relationship of sons with given height to their fathers. Indeed, reversing the roles of X and Y, we obtain the **second regression line**

$$\begin{aligned} x &= E(X|Y = y) = \mu_x + \beta'(y - \mu_y), \qquad (10.4) \\ \beta' &= \frac{\sigma_{XY}}{\sigma_Y^2} \end{aligned}$$

Under the assumption $\sigma_Y^2 = \sigma_X^2$ we have $\beta' = \beta$, hence $0 < \beta' < 1$, so that the fathers of tall sons tend to be shorter etc. In (10.4) x is given as a function of y; when we put in the same form as the first regression line, with y a function of x, we obtain

$$y = \mu_y + \frac{1}{\beta'}(x - \mu_x)$$

and turns out that the other regression line also goes through the point (μ_x, μ_y), but has different slope $1/\beta'$. This slope is higher $(1/\beta' > 1)$ if $\sigma_Y^2 = \sigma_X^2$. The linear function (10.4) is said to pertain to the regression of X on Y.
Back in the general bivariate normal $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, consider the random variable

$$\eta = Y - E(Y|X = x) = Y - \mu_y - \beta(x - \mu_x);$$

we know from (10.2) that it has conditional law $N(0, \sigma_{Y|X}^2)$, given x. It is often called the **residual** (random variable). With the conditional law of η, we can form the joint law of η and X; since the conditional law does not depend on x (Corollary 10.1.2 (ii)), it turns out that η and X are independent.

Corollary 10.1.4 *If X, Y have a bivariate normal distribution, then the residual*

$$\eta = Y - E(Y|X) = Y - \mu_y - \beta(X - \mu_x)$$

and X are independent.

Recall that $E(Y|X)$ denotes the conditional expectation as a random variable, i.e. $E(Y|X = x)$ when X is understood as random.
As a consequence, we can write any bivariate normal distribution as

$$\begin{aligned} Y &= \mu_y + \beta\xi + \eta \qquad (10.5) \\ X &= \mu_x + \xi \qquad (10.6) \end{aligned}$$

where ξ, η are independent normal with laws $N(0, \sigma_X^2)$, $N(0, \sigma_{Y|X}^2)$ respectively. If $\mu_x = 0$ we can write

$$Y = \mu_y + \beta X + \eta.$$

Assume that we wish to obtain a representation (10.5), (10.5) in terms of standard normals. Define $\eta_0 = \sigma_{Y|X}^{-1}\eta$, $\xi_0 = \sigma_X^{-1}\xi$; then for a matrix

$$M = \begin{pmatrix} \sigma_X & 0 \\ \beta\sigma_X & \sigma_{Y|X} \end{pmatrix} = \begin{pmatrix} \sigma_X & 0 \\ \sigma_{XY}\sigma_X^{-1} & \left(\sigma_Y^2 - \sigma_{XY}^2\sigma_X^{-2}\right)^{1/2} \end{pmatrix}$$

we have for $\mathbf{Z} = (X, Y)^\top$, $\mathbf{U} = (\xi_0, \eta_0)^\top$

$$\mathbf{Z} = \mu + M\mathbf{U},\ \mathcal{L}(\mathbf{U}) = N_2(\mathbf{0}, I_2),\ . \tag{10.7}$$

Since $\mathbf{\Sigma}$ is the covariance matrix of $\mathbf{Z}$ we should have

$$\mathbf{\Sigma} = MM^\top$$

according to the rules for the multivariate normal. The above relation can indeed be verified, and represents a decomposition of $\mathbf{\Sigma}$. into a product of a lower triangular with its transpose.
Next we define a regression function for possibly nonnormal distributions

Definition 10.1.1 *Let X, Y have a continuous or discrete joint distribution, where the first moment of Y exists: $E|Y| < \infty$. The* ***regression function*** *(for Y on X) is defined as*

$$r(x) := E(Y|X = x).$$

In the normal case, we have seen that r is a linear function, determined by the regression coefficient and the two means. In the general case, one should verify that $E(Y|X = x)$ exists, and clarify the problem of uniqueness. Let us do that in the continuous case.

Proposition 10.1.1 *Let X, Y have a continuous joint distribution, where $E|Y| < \infty$.*
(i) There is a version of the conditional density $p(y|x)$ such that for this density, $E(Y|X = x)$ exists for all x.
(ii) For all versions of $p(y|x)$, the law of the random variable $E(Y|X) = r(X)$ is the same, thus $\mathcal{L}(E(Y|X))$ is unique.

Proof. Recall definition 5.4.2: any version of the conditional density $p(y|x)$ fulfills

$$p(x, y) = p(y|x)p_X(x). \tag{10.8}$$

Now $E|Y| < \infty$ means that

$$\begin{aligned} \infty &> \int\int |y|p(x, y)dydx = \int\int |y|p(y|x)p_X(x)dydx \\ &= \int p_X(x)\left\{\int |y|p(y|x)dy\right\}dx. \end{aligned}$$

Let A be the set of x such that $g(x) = \int |y|p(y|x)dy$ is infinite. For this A we must have $\int_A p_X(x)dx = 0$, or else the whole expression above would be infinite. (This needs a few lines

of reasoning with integrals.) Hence we must have $P(X \in A) = 0$. For these x we can modify our version of $p(y|x)$, e.g. take it as the standard normal density (as in the proof of Lemma 5.4.3), so that $\int |y|p(y|x)dy$ is also finite for these x. Thus we found a version $p(y|x)$ such that

$$E(Y|X = x) = \int yp(y|x)dy \tag{10.9}$$

exists for all x. This proves (i).
For (ii), integrate (10.8) over y to obtain

$$\int yp(x, y)dy = r(x)p_X(x). \tag{10.10}$$

Note that two densities for a probability law must coincide except on a set of probability 0. This holds for $p(x, y)$, and it implies that two versions of the function $\int yp(x, y)dy$ must coincide for all x. (in terms of X). But the versions of densities $p_X(x)$ coincide for a set of probability 0 in terms of X; and then (10.10) implies such a property for $r(x)$. This mean that $\mathcal{L}(r(X))$ is uniquely determined. ■
Note that strictly speaking, we are not entitled to speak of "the" regression function $r(x)$ above, as it is not unique. However the law of the random variable $E(Y|X) = r(X)$ is unique.
The next statement recalls a "best prediction" property of the conditional expectation. In the framework of discrete distributions, this was already discussed in Remark 2.4, in connection with the properties of the conditional expectation $E(Y|X)$ as a random variable. For the normal case, this was exercise H5.2.

Proposition 10.1.2 *Let X, Y have a continuous or discrete joint distribution, where the second moment of Y exists: $E|Y|^2 < \infty$. Then the regression function has the property that*

$$E\left(Y - r(X)\right)^2 = \min_{f \in \mathfrak{M}_X} E\left(Y - f(X)\right)^2$$

where $\mathfrak{M}_X$ is the set of all (measurable) functions of X such that $E\left(f(X)\right)^2 < \infty$.

Proof. This resembles other calculations with conditional expectations (cf. Remark 2.4, p. 9). A little more work is needed now to ensure that all expectations exist. We concentrate on the continuous case. First note that under the conditions, $E\left(Y - f(X)\right)^2$ is finite for every $f \in \mathfrak{M}_X$. We claim that also $r \in \mathfrak{M}_X$. Indeed the Cauchy-Schwarz inequality gives

$$\begin{aligned} |r(x)|^2 &= \left|\int yp(y|x)dy\right|^2 \leq \int p(y|x)dy \cdot \int y^2 p(y|x)dy \\ &= \int y^2 p(y|x)dy. \end{aligned}$$

The last integral can be shown to be finite for all x when $E|Y|^2 < \infty$, similarly to Lemma 10.1.1 above (possibly with a modification of $p(y|x)$). Thus

$$\begin{aligned} E\left(r(X)\right)^2 &\leq \int p_X(x) \left(\int y^2 p(y|x)dy\right) dx \\ &= \int\int y^2 p(x, y)dxdy = EY^2 < \infty \end{aligned}$$

which proves $r \in \mathfrak{M}_X$. This implies that all expectations in the following reasoning are finite. We have for any $f \in \mathfrak{M}_X$

$$\begin{aligned} E\left(Y - f(X)\right)^2 &= E\left(Y - r(X) - f(X) + r(X)\right)^2 \\ &= E\left(Y - r(X)\right)^2 + 2E\left(Y - r(X)\right)\left(f(X) - r(X)\right) + E\left(f(X) - r(X)\right)^2, \end{aligned}$$

and the middle term vanishes, when we use the formulae.

$$E(\cdot) = E\left(E(\cdot|X)\right),\ E(Yh(X)|X) = h(X)E(Y|X).$$

Hence

$$E\left(Y - f(X)\right)^2 \geq E\left(Y - r(X)\right)^2.$$

■

The regression function is thus a characteristic of the joint distribution of X and Y. In general $r(x) := E(Y|X = x)$ is a **nonlinear regression function**; it is linear if the joint distribution is normal.

In the general case, it is no longer true that the residual $\eta = Y - r(X)$ and X are independent; it can only be shown that they are uncorrelated (Exercise). However one can build a bivariate distribution of X, Y from independent η (with zero mean) and X:

$$Y = r(X) + \eta. \tag{10.11}$$

Assume that $E\left|r(X)\right| < \infty$, so that Y has finite expectation and hence $E\left(Y|X\right)$ exists. It then follows that

$$E\left(Y|X\right) = E\left(r(X)|X\right) + E\left(\eta|X\right) = r(X) + E\eta = r(X)$$

so r is the regression function for (X, Y).

10.2 Bivariate regression models

For the joint distribution of X,Y the relation (10.11) describes approximately the dependence of Y on X, or a "noisy" causal relationship between X and Y. Suppose one has i.i.d. observations $\mathbf{Z}_i = (X_i, Y_i)^\top$, and one wishes to draw inferences on this dependence, i.e. on the regression function $r(x)$:

$$Y_i = r(X_i) + \eta_i$$

where η_i are the corresponding residuals. Since the regression function $r(x) = E\left(Y|X\right)$ depends only on the conditional distribution of Y given X, and X is observed, it makes sense to take a conditional point of view, and assume that $X_i = x_i$ where x_i are nonrandom values. These can be taken to be the realized X_i Thus

$$Y_i = r(x_i) + \eta_i$$

where the η_i are still independent. (Exercise: show that the joint conditional densities of all $\eta_i = Y_i - E(Y_i|X_i)$ under all X_i is the product of the individual conditional densities of η_i).

Definition 10.2.1 *Suppose that x_i, $i = 1, \ldots, n$ are nonrandom values, and η_i, $i = 1, \ldots, n$ are i.i.d. random variables with zero expectation, $\mathcal{L}(\eta) = Q$. Let $\mathcal{R}$ be a set of functions on $\mathbb{R}$ and $\mathcal{Q}$ be a set of probability laws on $\mathbb{R}$. A **bivariate regression model** is given by observed data*

$$Y_i = r(x_i) + \eta_i,$$

where it is assumed that $r \in \mathcal{R}$ and $Q \in \mathcal{Q}$.
***(i)** A **linear regression model** is obtained when $\mathcal{R}$ is assumed to be a set of linear functions*

$$r(x) = \alpha + \beta x$$

(where linear restrictions on α, β may be present)
***(ii)** A **normal linear regression model** is obtained when in (i) $\mathcal{Q}$ is assumed as a set of normal laws $N(0, \sigma^2)$ (σ^2 fixed or unknown)*
***(iii)** A **nonlinear regression model** (wide sense) is obtained when for some parameter set $\Theta \subseteq \mathbb{R}^k$, $k \geq 1$*

$$\mathcal{R} = \{r_\vartheta, \vartheta \in \Theta\}$$

and all functions $r_\vartheta(x)$ are nonlinear in x.
***(iv)** A **nonparametric regression model** is obtained when $\mathcal{R}$ is a class of functions of x which cannot be smoothly parametrized by some $\vartheta \in \Theta \subseteq \mathbb{R}^k$, e.g. a set of differentiable functions*

$$\mathcal{R} = \left\{ r : r' \text{ exists, } \left|r'(x)\right| \leq C, \text{ all } x \right\}.$$

In a nonparametric regression model, the functions are also nonlinear in x, but the term is reserved for families $r_\vartheta, \vartheta \in \Theta$ indexed by a finite dimensional ϑ. A typical example for nonlinear regression is **polynomial regression**

$$r(x) = \sum_{j=0}^{k} \vartheta_j x^j$$

or more generally

$$r_\vartheta(x) = \sum_{j=0}^{k} \vartheta_j \varphi_j(x)$$

where $\{\varphi_j\}$ is a system of functions (e.g. **trigonometric regression**). However in these examples, the function values $r_\vartheta(x)$ depend linearly on the parameter $\vartheta = (\vartheta_j)_{j=1,\ldots,n}$. We shall see below that these can be treated similarly to the linear case. Therefore the name of *nonlinear regression model in a narrow sense* is reserved for models where $r_\vartheta(x)$ depends on ϑ nonlinearly, e.g.

$$r_\vartheta(x) = \sin(\vartheta x).$$

In the linear case, we have

$$Y_i = \alpha + \beta x_i + \eta_i,\ i = 1, \ldots, n \tag{10.12}$$

and commonly it is assumed that the variance exists: $EY_1^2 < \infty$. The linear case (without normality assumption) is simply called **a linear model.**
Note that the normal location and location-scale models are special cases of the normal linear model, when a restriction $\alpha = 0$ is assumed and $x_i = 1$, $i = 1, \ldots, n$. Here the x_i do not resemble realizations of i.i.d. random variables; this possibility enlarges considerably the scope of the linear model. Indeed the x_i can be chosen or designed, e. g. taken all as 1 as above or as as an equidistant grid (or mesh):

$$x_i = i/n,\ i = 1, \ldots, n.$$

The set $\{x_i,\ i = 1, \ldots, n\}$ is also called **a regression design**. Especially the case of nonparametric regression with equidistant grid resembles the problems of function interpolation from noisy data (also called smoothing) in numerical analysis.

10.3 The general linear model

Our starting point above was a bivariate normal distribution of variables (X, Y), and the description of $E(Y|X)$, as the best predictor of Y given X. The same reasoning is possible when we have several variables $X_1, \ldots, X_k$, i.e. a whole vector $\mathbf{X} = (X_1, \ldots, X_k)^\top$ and we are interested in an approximate causal relationship between $\mathbf{X}$ and Y. This of course allows modelling of much more complex relationships. The $X_1, \ldots, X_k$ are called **regressor variables** and Y the **regressand**. Again $E(Y|\mathbf{X})$ can be shown to be a linear function of $\mathbf{X}$, but we forgo this and proceed directly to setting up a linear regression model in several nonrandom regressors.
Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a set of nonrandom k-vectors. These might be independent realizations of a random vector $\mathbf{X}$, but they might also be designed values. We assume that for some vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^\top$, observations are

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i,\ i = 1, \ldots, n$$

where ε_i are i.i.d. with $E\varepsilon_1 = 0$, $E\varepsilon_1^2 < \infty$. This is a direct generalization of (10.12).

Definition 10.3.1 *Let X be a nonrandom $n \times k$-matrix of rank k, and $\boldsymbol{\varepsilon}$ be a random n-vector such that*

$$E\boldsymbol{\varepsilon} = \mathbf{0},\ \mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 I_n.$$

***(i)** A (general) **linear model** is given by an observed n-vector*

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where X is fixed (known), $\boldsymbol{\varepsilon}$ is unobserved and $\boldsymbol{\beta} \in \mathbb{R}^k$ is unknown, and σ^2 is either known or unknown ($\sigma^2 > 0$).
*(ii) A **normal linear model** is obtained when $\boldsymbol{\varepsilon}$ is assumed to have a normal distribution.*

Notation and terminology. *Now X is an $n \times k$-matrix,whereas in the previous section X was a random variable (the regressor variable), and generalizing this, in the first paragraph of this section $X = (X_1, \ldots, X_k)$ was a random vector of regressor variables. This reflects the situation that the matrix X may arise from independent realizations $\mathbf{x}_i$ of the random vector X, in conjunction with a conditional point of view (the $\mathbf{x}_i$ are considered nonrandom, and form the rows of the matrix X) Therefore we keep the symbol X for the matrix above;X may be called* **regression matrix**. *The columns of X ($\xi_1, \ldots, \xi_k$, say) may be called* **nonrandom regressor variables**; *they correspond to the random regressors $X_1, \ldots, X_k$.*
In the normal linear model, we have $\mathcal{L}(\boldsymbol{\varepsilon}) = N_n(\mathbf{0}, I_n)$, and the components ε_i are i.i.d. standard normal. Hence $\mathcal{L}(\mathbf{Y}) = N_n(X\boldsymbol{\beta}, I_n)$. In the general case, ε_i are only uncorrelated; however in most cases, when modelling real world phenomena by a by a linear model, r.v.'s which are uncorrelated but not independent will not often occur.

Example 10.3.2 Let U have the uniform distribution on $[0, 1]$ and

$$Z = \cos(2\pi U),\ Y = \sin(2\pi U).$$

Then $Z^2 + Y^2 = 1$ and the pair (Z, Y) takes values on the unit circle, which implies that Z, Y are not independent (they do not have a joint density on $\mathbb{R}^2$ which is the product of its marginals.)

But Z, Y are uncorrelated:

$$\begin{aligned} EZ &= \int_0^1 \cos(2\pi u)du = 0,\ EY = 0, \\ \mathrm{Cov}(Z,Y) &= EZY = \int_0^1 \cos(2\pi u)\sin(2\pi u)du = 0. \end{aligned}$$

□

Identifiability.

Let us explain the assumption $\mathrm{rank}(X) = k$. Let $\mathbf{x}_i^\top, i = 1, \ldots, n$ be the rows of the $n \times k$-matrix X and $\boldsymbol{\xi}_j$, $j = 1, \ldots, k$ the colums:

$$X = \begin{pmatrix} \mathbf{x}_1^\top \\ \ldots \\ \mathbf{x}_n^\top \end{pmatrix} = \begin{pmatrix} \boldsymbol{\xi}_1, & \ldots, & \boldsymbol{\xi}_k \end{pmatrix}.$$

Recall that $\mathrm{rank}(X) = k$ is equivalent to each of the two:

$$\mathrm{Lin}\left(\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}\right) = \mathbb{R}^k, \tag{10.13}$$

$$\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k \text{ are linearly independent } n\text{-vectors.} \tag{10.14}$$

(where Lin() denotes the linear space spanned by a set of vectors, also called linear span, linear hull).

Definition 10.3.3 *Let $\mathcal{P} = \{P_\vartheta; \vartheta \in \Theta\}$ be a family of probability laws on a general space $\mathcal{Z}$. The parameter ϑ is called **identifiable** in $\mathcal{P}$ if for all pairs $\vartheta_1, \vartheta_2 \in \Theta$, $\vartheta_1 \neq \vartheta_2$ implies $P_{\vartheta_1} \neq P_{\vartheta_2}$.*

If $\mathcal{P}$ is a statistical model, i.e. is the set of possible (assumed) distributions of an observed random variable Z with values in $\mathcal{Z}$, then nonidentifiablity means that there exist two parameters $\vartheta_1 \neq \vartheta_2$ which lead to the same law of Z. These cannot be distinguished in any statistical sense: for hypotheses $\vartheta_1 = \vartheta_2$ vs. $\vartheta_1 \neq \vartheta_2$ the trivial randomized test $\varphi(Z) = \alpha$ is a most powerful α-test. A parameter ϑ is just an index or name for a probability law; identifiability means that no law in $\mathcal{P}$ has two names.
Identifiablity thus is a basic condition for a statistical model, if inference on the parameter ϑ is desired. If ϑ is nonindentifiable in $\mathcal{P}$ then it is advisable to **reparametrize**, i.e. give the laws other names which are identifiable.
Assume for a moment that the assumption $\mathrm{rank}(X) = k$ is not part of the definition of the linear model (Definition 10.3.1(i)).

Lemma 10.3.4 *In the normal linear model, $\boldsymbol{\beta}$ is identifiable if and only if* $\mathrm{rank}(X) = k$.

Proof. Assume $\mathrm{rank}(X) = k$ and $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$. Since

$$E\mathbf{Y} = X\boldsymbol{\beta},$$

it suffices to show that $X\boldsymbol{\beta}_1 \neq X\boldsymbol{\beta}_2$. If $X\boldsymbol{\beta}_1 = X\boldsymbol{\beta}_2$ the we would have $X(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) = 0$ and $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \neq \mathbf{0}$, which contradicts (10.14).

Conversely, assume identifiability; recall $\mathcal{L}(\mathbf{Y}) = N_n(\mathbf{X}\boldsymbol{\beta}, I_n)$. If $\text{rank}(X) < k$ then (10.14) is violated, i.e. $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k$ are linearly dependent. Hence there is $\boldsymbol{\beta} \neq \mathbf{0}$ such that $X\boldsymbol{\beta} = \mathbf{0}$. Since also $\mathbf{X0} = \mathbf{0}$, the parameters $\boldsymbol{\beta}$ and $\mathbf{0}$ lead to the same distribution $\mathcal{L}(\mathbf{Y}) = N_n(\mathbf{0}, I_n)$, hence $\boldsymbol{\beta}$ is not identifiable. This contradicts the assumption, hence $\text{rank}(X) = k$. ■

In the linear model without the normality assumption, another parameter is present, namely the distribution of the random noise vector $\boldsymbol{\varepsilon}$. When this law $\mathcal{L}(\boldsymbol{\varepsilon})$ is assumed unknown, except for the assumptions $E\boldsymbol{\varepsilon} = \mathbf{0}$, $\text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$, then the parameter is $\vartheta = (\boldsymbol{\beta}, \mathcal{L}(\boldsymbol{\varepsilon}))$ and $\mathcal{L}(\mathbf{Y}) = P_\vartheta$ is indexed by ϑ. In this situation, we call $\boldsymbol{\beta} = \boldsymbol{\beta}(\vartheta)$ identifiable if $P_{\vartheta_1} = P_{\vartheta_2}$ implies $\boldsymbol{\beta}(\vartheta_1) = \boldsymbol{\beta}(\vartheta_2)$. It is easy to see (exercise) that also in this model, the condition $\text{rank}(X) = k$ is necessary and sufficient for identifiability of $\boldsymbol{\beta}$.

10.3.1 Special cases of the linear model

1. Bivariate linear regression. We have

$$Y_i = \alpha + \beta x_i + \varepsilon_i,\ i = 1, \ldots, n$$

Here $k = 2$, the rows of X are $\mathbf{x}_i^\top = (1, x_i)$, $\boldsymbol{\beta} = (\alpha, \beta)^\top$, and identifiability holds as soon as not all x_i are equal.

2. Normal location-scale model. Here

$$Y_i = \beta + \varepsilon_i,\ i = 1, \ldots, n,$$

$k = 1$, $X = (1, \ldots, 1)^\top$, ε_i are i.i.d. $N(0, \sigma^2)$, $\sigma^2 > 0$.

3. **Polynomial regression.** Here for some design points x_i, $i = 1, \ldots, n$

$$Y_i = \sum_{j=1}^{k} \beta_j \varphi_j(x_i) + \varepsilon_i,\ i = 1, \ldots, n$$

where $\varphi_j(x) = x^{j-1}$. The matrix X is made up of the columns

$$\boldsymbol{\xi}_j = (\varphi_j(x_1), \ldots, \varphi_j(x_n))^\top, j = 1, \ldots, k. \tag{10.15}$$

Note that we obtained as a special case of the linear model one which was earlier classified as a nonlinear regression model (in a wide sense), since the functions

$$r(x) = \sum_{j=1}^{k} \beta_j \varphi_j(x)$$

are nonlinear in x (polynomials). However they are linear in the parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^\top$, and for purposes of estimating $\boldsymbol{\beta}$ this can be treated as a linear model.

Lemma 10.3.5 *In the linear model arising from polynomial regression, the parameter $\boldsymbol{\beta}$ is identifiable if and only if among the design points x_i, $i = 1, \ldots, n$, there are at least k different points.*

Proof. Identifiability means linear independence of the vectors $\boldsymbol{\xi}_j$ in (10.15). This in turn means that for any coefficients $\beta_1, \ldots, \beta_k$, the relation

$$r(x_i) = \sum_{j=1}^{k} \beta_j \varphi_j(x_i) = 0,\ i = 1, \ldots, n \tag{10.16}$$

implies $\beta_j = 0$, $j = 1, \ldots, k$. Now $r(x)$ is a polynomial of degree $k-1$; if not all β_j vanish, then r can have at most $k-1$ different zeros. Thus if among the design points x_i, $i = 1, \ldots, n$, there are at least k different points then X has rank k, thus $\boldsymbol{\beta}$ is identifiable. Conversely, assume that only $x_1, \ldots, x_{k-1}$ are different. Consider the polynomial r_0 of degree $k-1$ having these points as zeros:

$$r_0(t) = \prod_{j=1}^{k-1}(t - x_i).$$

Let $\beta_1, \ldots, \beta_k$ be the coefficients of r_0; then .(10.16) holds for $r = r_0$, hence the vectors $\boldsymbol{\xi}_j$ are linearly dependent. ■

The result that some of the design points x_i may be the same opens up the possibility of a **design with replication**: for a fixed number $m \geq k$, take m different points x_j^*, $j = 1, \ldots, m$ and repeated measurements at these points, l times say, so that $n = ml$. The entire design may then be written with a double index:

$$x_{jk} = x_j^*,\ j = 1, \ldots, m,\ k = 1, \ldots, l$$

and each of the x_j^* appears l times. As a result, using double index notation again, we obtain observations

$$Y_{jk} = r(x_j^*) + \varepsilon_{jk},\ j = 1, \ldots, m,\ k = 1, \ldots, l \tag{10.17}$$

which suggests taking averages $\bar{Y}_{j\cdot} = l^{-1}\sum_{k=1}^{l} Y_{jk}$ to obtain a simplified model, with more accurate "data" $\bar{Y}_{j\cdot}$. The choice of such a replicated design may be advantageous.

Comment on notation: we now use k also as a running index, even though above k denoted the number of functions φ_j involved, i.e. the dimension of the regression parameter $\boldsymbol{\beta}$. In the sequel, we will use d for the dimension of $\boldsymbol{\beta}$. The reason is that use of k in expressions such as Y_{jk} is traditional, in connection with regression and replicated designs.

4. Analysis of variance. Consider a model

$$Y_{jk} = \mu_j + \varepsilon_{jk},\ j = 1, \ldots, m,\ k = 1, \ldots, l \tag{10.18}$$

where ε_{jk} are independent noise variables. Here again a replication structure is present, similary to (10.17), *but no particular form is assumed for the function* r. Thus, if r is an arbitrary function in (10.17), we might as well write $\mu_j = r(x_j^*)$ and assume μ_j unrestricted. The case $m = 2$ (for normal ε_{jk}) was already encountered in the *two sample problem*. Suppose there are two "treatments", $j = 1, 2$, respective observations Y_{jk}, $j = 1, 2$, and one wishes to test whether the treatments have an effect: H:$\mu_1 = \mu_2$ vs. $K : \mu_1 \neq \mu_2$. To explain the name "analysis of variance", let us find the likelihood ratio test. In HW 7.1 we found a certain t-test for this problem; cf. also HW 6.1; this will turn out to be the LR test.

Define the two sample means and variances

$$\bar{y}_{i\cdot} = l^{-1}\sum_{k=1}^{l} y_{ik},\ S_{il}^2 = l^{-1}\sum_{k=1}^{l}(y_{ik} - \bar{y}_{i\cdot})^2,\ i = 1, 2.$$

and also **pooled estimates (where** $n = 2l$)

$$\bar{y}_{\cdot\cdot} = n^{-1}\left(l\bar{y}_{1\cdot} + l\bar{y}_{2\cdot}\right) = \frac{1}{2}\left(\bar{y}_{1\cdot} + \bar{y}_{2\cdot}\right), \tag{10.19}$$

$$S_n^2 = \frac{1}{n}\left(\sum_{k=1}^{l}\left(y_{1k} - \bar{y}_{\cdot\cdot}\right)^2 + \sum_{k=1}^{l}\left(y_{2k} - \bar{y}_{\cdot\cdot}\right)^2\right). \tag{10.20}$$

Note that the pooled sample variance can be decomposed: since

$$\sum_{k=1}^{l}(y_{1k}-\bar{y}_{\cdot\cdot})^2=\sum_{k=1}^{l}(y_{1k}-\bar{y}_{1\cdot})^2+l\,(\bar{y}_{1\cdot}-\bar{y}_{\cdot\cdot})^2\,,$$

we obtain

$$S_n^2=\frac{1}{2}\left(S_{1l}^2+S_{2l}^2\right)+\frac{1}{2}\sum_{j=1,2}(\bar{y}_{j\cdot}-\bar{y}_{\cdot\cdot})^2 \tag{10.21}$$

Note that the second term has the form of a sample variance, for a "sample" of size 2 with "observations" $\bar{y}_{1\cdot}$, $\bar{y}_{2\cdot}$. The first term can be seen as the" **variability within groups"** and the second term as the "**variability between groups"**.

Theorem 10.3.6 *In the Gaussian two sample problem (10.18),* $\mathcal{L}(\varepsilon_{jk})=N(0,\sigma^2)$, $\sigma^2>0$ *unknown, for hypotheses*
$H:\mu_1=\mu_2$
$K:\mu_1\neq\mu_2$
the LR statistic is

$$L(y_1,y_2)=\left(1+\frac{\frac{1}{2}\sum_{j=1,2}(\bar{y}_{j\cdot}-\bar{y}_{\cdot\cdot})^2}{\frac{1}{2}(S_{1l}^2+S_{2l}^2)}\right)^{n/2}.$$

The LR test is equivalent to a t-test which rejects when $|T|$ *is too large, where*

$$\begin{aligned}T &= \frac{l^{1/2}\,(\bar{y}_{1\cdot}-\bar{y}_{2\cdot})}{\left(\hat{S}_{1l}^2+\hat{S}_{2l}^2\right)^{1/2}},\\ \hat{S}_{il}^2 &= \frac{l}{l-1}S_{il}^2, i=1,2,\end{aligned}$$

and $\mathcal{L}(T)=t_{2l-2}$ *under* H.

Comment. *This form of the likelihood ratio statistic explains the name "analysis of variance".* The pooled sample variance is decomposed according to (10.21), and the LR test rejects if the variability between groups is too large, compared to the variability within groups.

Proof. The argument is very similar to the proof of Proposition 8.4.2 about the LR tests for the one sample Gaussian location-scale model; it can be considered the "two sample analog". Since we have two independent samples with different expectations μ_i and the same variance, the joint density p_{μ_1,μ_2,σ^2} of all the data $y_1=(y_{11},\ldots,x_{1l})$, $y_2=(y_{21},\ldots,y_{2l})$ is, under the alternative,

$$p_{\mu_1,\mu_2,\sigma^2}(y_1,y_2)=\prod_{i=1,2}\frac{1}{(2\pi\sigma^2)^{l/2}}\exp\left(-\frac{S_{il}^2+(\bar{y}_{i\cdot}-\mu_i)^2}{2\sigma^2 l^{-1}}\right). \tag{10.22}$$

Maximizing the likelihood under the alternative we obtain estimates $\hat{\mu}_i=\bar{y}_{i\cdot}$ and for $n=2l$

$$\begin{aligned}\max_{\mu_1\neq\mu_2,\sigma^2>0}p_{\mu_1,\mu_2,\sigma^2}(y_1,y_2) &= \frac{1}{(2\pi\hat{\sigma}^2)^{l}}\exp\left(-\frac{S_{1l}^2+S_{2l}^2}{2\hat{\sigma}^2 l^{-1}}\right)\\ &= \frac{1}{(2\pi\hat{\sigma}^2)^{n/2}}\exp\left(-\frac{(S_{1l}^2+S_{2l}^2)/2}{2\hat{\sigma}^2 n^{-1}}\right)\\ &= \frac{1}{(\hat{\sigma}^2)^{n/2}}\cdot\frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2n^{-1}}\right)\end{aligned}$$

where

$$\hat{\sigma}^2 = \frac{1}{2}\left(S_{1l}^2 + S_{2l}^2\right).$$

Under the hypothesis $\mu_1 = \mu_2 = \mu$ the data y_{1k}, y_{2k} are identically distributed, as $N(\mu, \sigma^2)$. Thus we can treat the two samples as one pooled sample of size n, and form the pooled mean and variance estimates (10.19), (10.20). To maximize the likelihood in μ, σ^2, we just refer to the results in the one sample case (Proposition 8.4.2) to obtain

$$\max_{\mu,\sigma^2>0} p_{\mu,\mu,\sigma^2}(y_1, y_2) = \frac{1}{(\hat{\sigma}_0^2)^{n/2}} \cdot \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2n^{-1}}\right).$$

Thus the likelihood ratio is

$$\begin{aligned} L(y_1, y_2) &= \frac{\max_{\mu_1\neq\mu_2,\sigma^2>0} p_{\mu_1,\mu_2,\sigma^2}(y_1, y_2)}{\max_{\mu,\sigma^2>0} p_{\mu,\mu,\sigma^2}(y_1, y_2)} \\ &= \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\right)^{n/2} = \left(\frac{S_n^2}{\left(S_{1l}^2 + S_{2l}^2\right)/2}\right)^{n/2}. \end{aligned}$$

The LR test compares the sample variance under the hypothesis S_n^2 with the sample variance under the alternative, which can taken to be $\left(S_{1l}^2 + S_{2l}^2\right)/2$. Using (10.21), we obtain

$$L(y_1, y_2) = \left(1 + \frac{\sum_{j=1,2}\left(\bar{y}_{j\cdot} - \bar{y}_{\cdot\cdot}\right)^2}{S_{1l}^2 + S_{2l}^2}\right)^{n/2}.$$

As a consequence of (10.19), we have

$$\bar{y}_{1\cdot} - \bar{y}_{\cdot\cdot} = (\bar{y}_{1\cdot} - \bar{y}_{2\cdot})/2,\ \bar{y}_{2\cdot} - \bar{y}_{\cdot\cdot} = (\bar{y}_{2\cdot} - \bar{y}_{1\cdot})/2$$

and hence

$$L(y_1, y_2) = \left(1 + \frac{1}{2(l-1)}T^2\right)^{n/2}.$$

The statistic T has a t-distribution with $2l-2$ degrees of freedom, since $\bar{y}_{1\cdot}$, $\bar{y}_{2\cdot}$, S_{1l}^2, S_{2l}^2 are all independent,

$$T = \frac{(l/2)^{1/2}\left(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}\right)}{\left((\hat{S}_{1l}^2 + \hat{S}_{2l}^2)/2\right)^{1/2}},$$

$$\begin{aligned} \mathcal{L}\left((l/2)^{1/2}(\bar{y}_{1\cdot} - \bar{y}_{2\cdot})\right) &= N(0, \sigma^2), \\ \mathcal{L}\left(\frac{2(l-1)}{\sigma}(\hat{S}_{1l}^2 + \hat{S}_{2l}^2)/2\right) &= \chi_{2l-2}^2. \end{aligned}$$

■

The ANOVA (analysis of variance) model (10.18) is a special case of the linear model. Write

$$\boldsymbol{\beta} = (\mu_1, \ldots, \mu_m)^\top, n = ml, \tag{10.23}$$

$$X = \begin{pmatrix} \mathbf{1}_l & & \\ & \ldots & \\ & & \mathbf{1}_l \end{pmatrix} \tag{10.24}$$

where $\mathbf{1}_l$ is the l-vector consisting of 1's, X is the $ml \times m$ matrix where the column vectors $\mathbf{1}_l$ are arranged in a diagonal fashion (with 0's elsewhere),

$$\mathbf{Y} = (Y_{11}, \ldots, Y_{1l}, Y_{21}, \ldots, Y_{ml})^\top,$$

and the n-vector $\boldsymbol{\varepsilon}$ is formed analogously. Then

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$\text{rank}(X) = m$, and the hypothesis $H : \mu_1 = \ldots = \mu_m$ can be expressed as

$$H : \boldsymbol{\beta} \in \text{Lin}(\mathbf{1}_m)$$

where $\text{Lin}(\mathbf{1}_m)$ is the linear subspace of $\mathbb{R}^m$ spanned by the column vector $\mathbf{1}_m$. That is a special case of a **linear hypothesis** about $\boldsymbol{\beta}$ (i.e. $\boldsymbol{\beta}$ is in some specified subspace of $\mathbb{R}^m$).

10.4 Least squares and maximum likelihood estimation

Consider again the general linear model

$$\begin{aligned} \mathbf{Y} &= X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, & (10.25)\\ E\boldsymbol{\varepsilon} &= \mathbf{0},\ \text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 I_n. & (10.26) \end{aligned}$$

and the problem of estimating $\boldsymbol{\beta}$, with known or unknown σ^2. If X is a $n \times k$-matrix and $\text{rank}(X) = k$ (identifiability), then the expectation of the random vector $\mathbf{Y}$ is in the linear subspace of $\mathbb{R}^n$ spanned by the matrix X (or by the k columns $\boldsymbol{\xi}_j$ of X):

$$\begin{aligned} \text{Lin}(\mathbf{X}) &= \text{Lin}(\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k\}) = \left\{ X\boldsymbol{\beta},\ \boldsymbol{\beta} \in \mathbb{R}^k \right\} \\ &= \left\{ \mathbf{z} = \sum_{j=1}^{k} \boldsymbol{\xi}_j \beta_j,\ \beta_j \text{ arbitrary} \right\}. \end{aligned}$$

An equivalent way of writing the linear model is

$$E\mathbf{Y} \in \text{Lin}(X),\ \text{Cov}(\mathbf{Y}) = \sigma^2 I_n.$$

Indeed no distributional assumptions about $\boldsymbol{\varepsilon}$ are made in the general linear model, so the above is all the information one has. For obtaining an estimator of $\boldsymbol{\beta}$, one could try to apply the principle of maximum likelihood. However since the distribution of $\boldsymbol{\varepsilon}$ (Q, say) is unspecified, it would have to be considered a parameter, along with β. Then Q and $\boldsymbol{\beta}$ specify the distribution of $\mathbf{Y}$, and hence a likelihood for any realization $\mathbf{Y} = \mathbf{y}$. But maximizing it in Q does not lead to satisfactory results, since the class for Q (arbitrary distributions on $\mathbb{R}^n$) is too large. Thus one has to look for a different principle to get an estimator of $\boldsymbol{\beta}$. The distribution Q of the noise is not of primary interest in most cases; it can be considered a nuisance parameter . The regression parameter $\boldsymbol{\beta}$ is the parameter of interest, since it describes the dependence of $\mathbf{Y}$ on X, and possibly also σ^2.
Of course one could assume normality, and then find maximum likelihood estimators. However another principle can be invoked, without a normality assumption.

Definition 10.4.1 *In the general linear model (10.25), (10.26), with observed vector* $\mathbf{Y} \in \mathbb{R}^n$, $\boldsymbol{\beta} \in \mathbb{R}^k$ *and* $\operatorname{rank}(X) = k$, *a* ***least squares estimator (LSE)*** $\hat{\boldsymbol{\beta}}$ *of* $\boldsymbol{\beta}$ *is a function* $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{Y})$ *of the observations such that*

$$\left\| \mathbf{Y} - X\hat{\boldsymbol{\beta}} \right\|^2 = \min_{\boldsymbol{\beta} \in \mathbb{R}^k} \|\mathbf{Y} - X\boldsymbol{\beta}\|^2$$

where $\|\cdot\|^2$ *is the squared Euclidean norm in* $\mathbb{R}^n$.

The name "least squares" derives from the fact that the squared Euclidean norm of any vector $\mathbf{z} \in \mathbb{R}^n$ is the sum of squares of the components z_i:

$$\|\mathbf{z}\|^2 = \sum_{i=1}^{n} z_i^2.$$

Recall that another from of writing the linear model was

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i,\ i = 1, \ldots, n$$

where $\mathbf{x}_i^\top$ are the rows of X ($\mathbf{x}_i$ are k-vectors). Thus another way of describing the least squares estimator $\hat{\boldsymbol{\beta}}$ is to say that for given $\mathbf{Y}$ it is a minimizer of

$$\operatorname{Ls}_{\mathbf{Y}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta} \right)^2.$$

This expression, depending on the observations $\mathbf{Y}$, to be minimized in $\boldsymbol{\beta}$ is also called the **least squares criterion.**

Exercise. *Show that if* β *is the true parameter in (10.25), (10.26), then* β *provides a best approximation to the data* Y *in an average sense:*

$$E\,\|\mathbf{Y} - X\boldsymbol{\beta}\|^2 = \min_{\boldsymbol{\gamma} \in \mathbb{R}^k} E\,\|\mathbf{Y} - X\boldsymbol{\gamma}\|^2$$

This minimization property can serve as a justification for the least squares criterion. If we could compute $E\,\|\mathbf{Y} - X\boldsymbol{\gamma}\|^2$ for any $\boldsymbol{\gamma}$, we would take the minimizer and obtain the true $\boldsymbol{\beta}$. However to find the expectation involved we already have to know β, so we take just $\|\mathbf{Y} - X\boldsymbol{\gamma}\|^2$ and minimize it. In this sense, $\hat{\boldsymbol{\beta}}$ can be considered an empirical analog of $\boldsymbol{\beta}$.

Theorem 10.4.2 *Consider the general linear model (10.25), (10.26), in the case* $k < n$, *with observed vector* $\mathbf{Y} \in \mathbb{R}^n$, $\boldsymbol{\beta} \in \mathbb{R}^k$ *and* $\operatorname{rank}(X) = k$, *with* σ^2 *either known or unknown* ($\sigma^2 > 0$)
(i) The LSE $\hat{\boldsymbol{\beta}}$ *of* $\boldsymbol{\beta}$ *is uniquely determined and given by*

$$\hat{\boldsymbol{\beta}} = \left(X^\top X \right)^{-1} X^\top \mathbf{Y}. \tag{10.27}$$

(ii) Under a normality assumption $\mathcal{L}(\boldsymbol{\varepsilon}) = N_n(\mathbf{0}, \sigma^2 I_n)$, *with probability one the LSE* $\hat{\boldsymbol{\beta}}$ *coincides with the maximum likelihood estimator (MLE) for* $\boldsymbol{\beta}$, *both in cases of known* σ^2 *and unknown* σ^2 *(when* $\sigma^2 > 0$*).*

Proof. Note that $\text{rank}(X) = k$ implies that $\left(X^\top X\right)^{-1}$ exists. The key argument is that the matrix

$$\Pi_X = X\left(X^\top X\right)^{-1} X^\top$$

represents the linear projection operator in $\mathbb{R}^n$ onto the linear subspace $\text{Lin}(\mathbf{X})$. Indeed, note that Π_X is a projection matrix, i.e. idempotent ($\Pi_X\Pi_X = \Pi_X$) and symmetric: $\Pi_X^\top = \Pi_X$. To see these two properties, note

$$\Pi_X\Pi_X = X\left(X^\top X\right)^{-1} X^\top X\left(X^\top X\right)^{-1} X^\top = X\left(X^\top X\right)^{-1} X^\top = \Pi_X,$$

and using the matrix rule $(AB)^\top = B^\top A^\top$, (which implies $(H^{-1})^\top = (H^\top)^{-1}$ for symmetric nonsingular H)

$$\begin{aligned}\Pi_X^\top &= \left(\left(X\left(X^\top X\right)^{-1}\right)X^\top\right)^\top = X\left(X\left(X^\top X\right)^{-1}\right)^\top = X\left(\left(X^\top X\right)^{-1}\right)^\top X^\top \\ &= X\left(\left(X^\top X\right)^\top\right)^{-1} X^\top = \Pi_X.\end{aligned}$$

Hence Π_X is a projection matrix. It has rank k, and it leaves the space $\text{Lin}(X)$ invariant: if $\mathbf{z} \in \text{Lin}(X)$ then $\mathbf{z} = X\mathbf{a}$ for some $\mathbf{a} \in \mathbb{R}^k$, and

$$\Pi_X\mathbf{z} = \Pi_X X\mathbf{a} = X\left(X^\top X\right)^{-1} X^\top X\mathbf{a} = X\mathbf{a} = \mathbf{z}.$$

Also for any $\mathbf{y}$ the vector $\Pi_X\mathbf{y}$ is in $\text{Lin}(X)$:

$$\Pi_X\mathbf{y} = X\left(X^\top X\right)^{-1} X^\top\mathbf{y} = X\tilde{\mathbf{y}}, \text{ where } \tilde{\mathbf{y}} = \left(X^\top X\right)^{-1} X^\top\mathbf{y}.$$

Moreover, consider the orthogonal complement of $\text{Lin}(X)$, i.e. $\text{Lin}(X)^\perp$. Any $\mathbf{z} \in \text{Lin}(X)^\perp$ is mapped into $\mathbf{0}$ by the linear map Π_X : since $X^\top\mathbf{z} = \mathbf{0}$, we have

$$\Pi_X\mathbf{z} = X\left(X^\top X\right)^{-1} X^\top\mathbf{z} = \mathbf{0}.$$

These facts establish that indeed Π_X is a matrix which represents the projection onto $\text{Lin}(X)$. (As a consequence, $I_n - \Pi_X$ is the projection operator onto $\text{Lin}(X)^\perp$).
It is well known that the projection operator has a minimizing property: for any $\mathbf{y} \in \mathbb{R}^n$, $\Pi_X\mathbf{y}$ gives the element of $\text{Lin}(X)$ closest to $\mathbf{y}$ (in the sense of $\|\cdot\|$). Indeed for any $\mathbf{z} \in \text{Lin}(X)$

$$\|\mathbf{y} - \mathbf{z}\|^2 = \|\mathbf{y} - \Pi_X\mathbf{y} + \Pi_X\mathbf{y} - \mathbf{z}\|^2 \tag{10.28}$$

Note that $\Pi_X\mathbf{y} - \mathbf{z} \in \text{Lin}(X)$, and

$$\mathbf{y} - \Pi_X\mathbf{y} = (I_n - \Pi_X)\mathbf{y} \in \text{Lin}(X)^\perp$$

since $I_n - \Pi_X$ projects onto $\text{Lin}(X)^\perp$. It follows that, when we compute (10.28) via $\|\mathbf{z}\|^2 = \mathbf{z}^\top\mathbf{z}$, since the two vectors are orthogonal, we get a sum of squared norms:

$$\|\mathbf{y} - \mathbf{z}\|^2 = \|\mathbf{y} - \Pi_X\mathbf{y}\|^2 + \|\Pi_X\mathbf{y} - \mathbf{z}\|^2.$$

The right side is minimized for $\mathbf{z} = \Pi_X \mathbf{y}$.
Apply this minimizing property of Π_X to obtain

$$\|\mathbf{Y} - X\boldsymbol{\beta}\|^2 \geq \|\mathbf{Y} - \Pi_X \mathbf{Y}\|^2$$

where equality is achieved for

$$X\boldsymbol{\beta} = \Pi_X \mathbf{y} = X\left(X^\top X\right)^{-1} X^\top \mathbf{Y}.$$

Pre-multiply by $X^\top$ to obtain

$$X^\top X\boldsymbol{\beta} = X^\top \mathbf{Y}$$

which gives a unique solution

$$\hat{\boldsymbol{\beta}} = \left(X^\top X\right)^{-1} X^\top \mathbf{Y}. \tag{10.29}$$

Part (i) is proved. For (ii), write down the likelihood function: since $\mathcal{L}(\mathbf{Y}) = N_n(X\boldsymbol{\beta}, \sigma^2 I_n)$, it is

$$L_{\mathbf{Y}}(\boldsymbol{\beta},\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{Y} - X\boldsymbol{\beta}\|^2\right).$$

For known σ^2, it is obvious that maximizing $L_{\mathbf{Y}}$ in $\boldsymbol{\beta}$ is equivalent to minimizing the least squares criterion

$$\mathrm{Ls}_{\mathbf{Y}}(\boldsymbol{\beta}) = \|\mathbf{Y} - X\boldsymbol{\beta}\|^2.$$

For unknown σ^2, restricted only by $\sigma^2 > 0$, minimize $L_{\mathbf{Y}}(\boldsymbol{\beta},\sigma^2)$ first in $\boldsymbol{\beta}$, for given σ^2. The solution is again $\hat{\boldsymbol{\beta}}$ given by (10.29). We now have to maximize

$$L_{\mathbf{Y}}(\hat{\boldsymbol{\beta}},\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{n^{-1}\mathrm{Ls}_{\mathbf{Y}}(\hat{\boldsymbol{\beta}})}{2\sigma^2 n^{-1}}\right)$$

in σ^2. This procedure was already carried out in the proof of Proposition 3.0.5 (insert now $n^{-1}\mathrm{Ls}_{\mathbf{Y}}(\hat{\boldsymbol{\beta}})$ for S_n^2). Provided that $\mathrm{Ls}_{\mathbf{Y}}(\hat{\boldsymbol{\beta}}) > 0$, the solution is

$$\hat{\sigma}^2 = n^{-1}\mathrm{Ls}_{\mathbf{Y}}(\hat{\boldsymbol{\beta}}).$$

Thus (ii) is proved if we show that $\mathrm{Ls}_{\mathbf{Y}}(\hat{\boldsymbol{\beta}}) > 0$ happens with probability one. Indeed $\mathrm{Ls}_{\mathbf{Y}}(\hat{\boldsymbol{\beta}}) = 0$ means that $\mathbf{Y} \in \mathrm{Lin}(X)$. Under normality, with covariance matrix $\sigma^2 I_n$ it is obvious that $\mathbf{Y} \in \mathrm{Lin}(X)$ happens with probability 0, if $k < n$ is fulfilled, since any proper linear subspace of $\mathbb{R}^n$ has probability 0 (indeed for any nonrandom vector $\mathbf{z} \in \mathbb{R}^n$, $\mathbf{z} \neq \mathbf{0}$ the event $\mathbf{z}^\top \mathbf{Y} = 0$ has probability 0, since $\mathbf{z}^\top \mathbf{Y}$ is normally distributed with variance $\sigma^2 \|\mathbf{z}\| > 0$). ■
It is easy to see that if $k = n$ then the LSE of β is still given by (10.27) and coincides with the MLE under normality if σ^2 is known, but if σ^2 is unknown then $\mathrm{Ls}_{\mathbf{Y}}(\hat{\boldsymbol{\beta}}) = 0$ and the MLE of σ^2 under normality does not exist (or should be taken as 0, with the likelihood function taking value ∞). The assumption $k < n$ is realistic, since k represents the number of independent regressor variables in most cases, and can be expected to be less than n.
Exercise. Consider the special cases of the linear model discussed in the previous subsection.
1. Bivariate linear regression:

$$Y_i = \alpha + \beta x_i + \varepsilon_i,\ i = 1, \ldots, n$$

Here $k = 2$, the rows of X are $\mathbf{x}_i^\top = (1, x_i)$, $\boldsymbol{\beta} = (\alpha, \beta)^\top$, and identifiability holds as soon as not all x_i are equal). Show that the LSE of α, β are

$$\hat{\alpha}_n = \bar{Y}_n - \hat{\beta}_n \bar{x}_n,\ \hat{\beta}_n = \frac{\sum_{i=1}^n (Y_i - \bar{Y}_n)(x_i - \bar{x}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}. \tag{10.30}$$

where $\bar{x}_n$ is the mean of the nonrandom x_i.

Remark 10.4.3 *Note that the formula for $\hat{\beta}_n$ is analogous to the regression coefficient in a bivariate normal distribution for (X, Y) according to Definition 10.1.3:*

$$\beta = \frac{\sigma_{XY}}{\sigma_X^2}$$

Therefore $\hat{\beta}_n$ is also called the ***empirical regression coefficient*** *(and β the theoretical coefficient). The regression function for (X, Y) was found as*

$$y = E(Y|X = x) = \mu_y + \beta(x - \mu_x) = \alpha + \beta x$$

which shows that $\hat{\alpha}$ is also the analog of $\alpha = \mu_y - \beta\mu_x$.

Since the bivariate regression is very important for applications (it is programmed in scientific pocket calculators), we summarize again what was done there.

Fitting a straight line to data (bivariate linear regression). *Given pairs (Y_i, x_i), $i = 1, \ldots, n$, find the straight line $y = \alpha + \beta x$ which best fits the data in the sense that the least squares criterion*

$$\sum_{i=1}^n (Y_i - \alpha + \beta x_i)^2 .$$

is minimal. The solutions $\hat{\alpha}_n$, $\hat{\beta}_n$ are given by (10.30). The fitted straight line $y = \hat{\alpha}_n + \hat{\beta}_n x$ passes through the point $(\bar{x}_n, \bar{Y}_n)$ with slope $\hat{\beta}_n$.

2. Location-scale model. (normality not assumed). Here

$$Y_i = \beta + \varepsilon_i,\ i = 1, \ldots, n,$$

This can be obtained from 1. above when $\alpha = 0$ is assumed, and $x_i = 1$. Show that the LSE of β is the sample mean:

$$\hat{\beta}_n = \bar{Y}_n. \tag{10.31}$$

3. ANOVA: Here

$$Y_{jk} = \mu_j + \varepsilon_{jk},\ j = 1, \ldots, m,\ k = 1, \ldots, l$$

Show that the LSE of μ_j is the group mean:

$$\hat{\mu}_j = \bar{Y}_{j\cdot} = l^{-1} \sum_{k=1}^l Y_{jk},\ j = 1, \ldots, m.$$

10.5 The Gauss-Markov Theorem

Consider again the general linear model:

Model (LM) (General linear model). Observations are

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{10.32}$$
$$E\boldsymbol{\varepsilon} = \mathbf{0},\ \mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 I_n. \tag{10.33}$$

where $\mathbf{Y}$ is a random n-vector, X is a $n \times k$-matrix and $\mathrm{rank}(X) = k$, $\boldsymbol{\beta} \in \mathbb{R}^k$ is unknown, $\boldsymbol{\varepsilon}$ is an unobservable random n-vector, and $\mathcal{L}(\boldsymbol{\varepsilon})$ satisfies (10.33) where σ^2 is known **(Model ($\mathbf{LM}_1$)** or unknown, restricted by $\sigma^2 > 0$ **(Model ($\mathbf{LM}_2$)**.

Recall that the *normal linear model* is obtained when we add the assumption $\mathcal{L}(\boldsymbol{\varepsilon}) = N(\mathbf{0},\sigma^2 I_n)$. This might be called **NLM** (possibly NLM_1 or NLM_2).
In (LM) we are now interested in optimality properties of the least squares estimator of $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = \left(X^\top X\right)^{-1} X^\top \mathbf{Y}.$$

Above in (10.31) it was remarked that the sample mean $\bar{Y}_n$ is a special case of $\hat{\boldsymbol{\beta}}$, for $X = \mathbf{1}$ (the n-vector consisting of 1's). In Section 5.7 it was shown that $\bar{Y}_n$ is a *minimax estimator* under normality, for known σ^2 (i.e. in the Gaussian location model). In Proposition 4.3.3 it was shown that in the Gaussian location model, the sample mean is also a *uniformly minimum variance unbiased estimator (UMVUE)*. Both optimality properties depend on the normality assumption: e.g. for the second optimality, within the class of unbiased estimators, it was shown that the quadratic risk of $\bar{Y}_n$ attains the Cramer-Rao bound, which is the inverse Fisher information. The Fisher information I_F is a function of the density (or probability function) of the data: recall the general form of I_F for a density p_ϑ depending on ϑ

$$I_F(\vartheta) = E_\vartheta \left(\frac{\partial}{\partial \vartheta} \log p_\vartheta(Y) \right)^2.$$

In the linear model, the distribution of $\mathbf{Y}$ is left unspecified. We might ask whether the sample mean, or more generally $\hat{\boldsymbol{\beta}}$, still has any optimality properties. A natural choice for the risk is $E\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|^2$, i.e. expected loss for the squared Euclidean distance.
Note that $\hat{\boldsymbol{\beta}}$ is **unbiased**:

$$\begin{aligned} E\hat{\boldsymbol{\beta}} &= E\left(X^\top X\right)^{-1} X^\top \mathbf{Y} = \left(X^\top X\right)^{-1} X^\top E\mathbf{Y} \\ &= \left(X^\top X\right)^{-1} X^\top X\boldsymbol{\beta} = \boldsymbol{\beta}. \end{aligned}$$

Without normality, it cannot be shown that $\hat{\boldsymbol{\beta}}$ is *best unbiased*. But when we further restrict the class of estimators admitted for comparison, to linear estimators, $\hat{\boldsymbol{\beta}}$ turns out to be optimal. Recall that a similar device was already employed for the sample mean and minimaxity: in Exercise 5.7.2 a minimax linear estimator was proposed, i.e. an estimator which is minimax within the class of linear ones.

Definition 10.5.1 *Consider the general linear model (LM).*
(i) *A* ***linear estimator*** *of* $\boldsymbol{\beta}$ *is any estimator* $\hat{\mathbf{b}}$ *of form*

$$\hat{\mathbf{b}} = A\mathbf{Y}$$

where A *is a (nonrandom)* $k \times n$*-matrix.*
(ii) *A linear estimator* $\tilde{\mathbf{b}}$ *is called* ***best linear unbiased estimator (BLUE)*** *if* $\tilde{\mathbf{b}}$ *is unbiased:*

$$E\tilde{\mathbf{b}} = \boldsymbol{\beta} \tag{10.34}$$

and if

$$E\left\|\tilde{\mathbf{b}} - \boldsymbol{\beta}\right\|^2 \leq E\left\|\hat{\mathbf{b}} - \boldsymbol{\beta}\right\|^2 \tag{10.35}$$

for all linear unbiased $\hat{\mathbf{b}}$*. Relations (10.34), (10.35) are assumed to hold for all values of the unknown parameters (*$\boldsymbol{\beta} \in \mathbb{R}^k$*, and* $\mathcal{L}(\boldsymbol{\varepsilon})$ *as specified).*

Theorem 10.5.2 ***(Gauss, Markov)*** *In the linear model (LM), the least squares estimator is the unique BLUE.*

Proof. Consider a linear unbiased estimator $\hat{\mathbf{b}} = A\mathbf{Y}$. The unbiasedness condition implies

$$\boldsymbol{\beta} = E\hat{\mathbf{b}} = EA\mathbf{Y} = AX\boldsymbol{\beta}$$

for all $\boldsymbol{\beta} \in \mathbb{R}^k$, hence

$$AX = I_k. \tag{10.36}$$

The loss is

$$\begin{aligned} \left\|\hat{\mathbf{b}} - \boldsymbol{\beta}\right\|^2 &= \|A\mathbf{Y} - \boldsymbol{\beta}\|^2 = \|A\mathbf{X}\boldsymbol{\beta} + A\boldsymbol{\varepsilon} - \boldsymbol{\beta}\|^2 && (10.37) \\ &= \|A\boldsymbol{\varepsilon}\|^2 = (A\boldsymbol{\varepsilon})^\top \mathbf{A}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^\top A^\top A\boldsymbol{\varepsilon}. && (10.38) \end{aligned}$$

For the risk we have to compute the expected loss.
For any $n \times n$-matrix $B = (B_{ij})_{i=1,\ldots,n}^{j=1,\ldots,n}$, define the **trace** as

$$\operatorname{tr}[B] = \sum_{i=1}^{n} B_{ii}.$$

In other words, the trace is the sum of diagonal elements. Note that for $n \times k$-matrices B, C and for n-vectors $\mathbf{z}, \mathbf{y}$ we have

$$\begin{aligned} \operatorname{tr}\left[BC^\top\right] &= \sum_{i=1}^{n}\sum_{j=1}^{k} B_{ij}C_{ji}^\top = \operatorname{tr}\left[C^\top B\right], && (10.39) \\ \operatorname{tr}\left[B^\top B\right] &= \sum_{i=1}^{n}\sum_{j=1}^{n} B_{ij}^2 = \operatorname{tr}\left[BB^\top\right], && (10.40) \\ \operatorname{tr}\left[\mathbf{z}\mathbf{y}^\top\right] &= \sum_{i=1}^{n} z_i y_i = \mathbf{y}^\top\mathbf{z},\ \operatorname{tr}\left[\mathbf{z}\mathbf{z}^\top\right] = \mathbf{z}^\top\mathbf{z} = \|\mathbf{z}\|^2. \end{aligned}$$

From (10.39) we see that $C \mapsto \mathrm{tr}\,[BC]$ is a linear operation on matrices, so that when C is a random matrix, then for the expectation we have

$$E\mathrm{tr}\,[BC] = \mathrm{tr}\,[BEC]\,.$$

Applying this to the quadratic loss (10.37), (10.38), we have

$$\begin{aligned} E\left\|\hat{\mathbf{b}} - \boldsymbol{\beta}\right\|^2 &= E\boldsymbol{\varepsilon}^\top A^\top A\boldsymbol{\varepsilon} = E\mathrm{tr}\left[A\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top A^\top\right] \\ &= E\mathrm{tr}\left[A^\top A\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right] = \mathrm{tr}\left[A^\top AE(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)\right] \\ &= \mathrm{tr}\left[A^\top A\sigma^2 I_n\right] = \sigma^2\mathrm{tr}\left[A^\top A\right]. \end{aligned}$$

The problem thus is to minimize $\mathrm{tr}\left[A^\top A\right]$ under the unbiasedness restriction $AX = I_k$ from (10.36). From (10.40) we see that if $\mathrm{vec}(A)$ is the kn-vector formed of all elements of A, then

$$\mathrm{tr}\left[A^\top A\right] = \sum_{i=1}^{k}\sum_{j=1}^{n} A_{ij}^2 = \|\mathrm{vec}(A)\|^2$$

i. e. problem is to minimize the length of the vector $\mathrm{vec}(A)$ under a set of affine linear restrictions $AX = I_k$. Consider first the special case $k = 1$; then $a = A^\top$ and X are vectors of dimension k and the set $\{a : a^\top X = 1\}$ is an affine hyperplane in $\mathbb{R}^k$. To minimize the norm of a within this set means taking a perpendicular to the hyperplane, i.e. having the same direction as X. This gives $a = X(X^\top X)^{-1}$ as a minimizer

This argument is generalized to $k \geq 1$ as follows. Let $\Pi_X = X\left(X^\top X\right)^{-1}X^\top$ be the projection operator onto $\mathrm{Lin}(X)$ in the space $\mathbb{R}^n$. We have

$$\begin{aligned} \mathrm{tr}\left[A^\top A\right] &= \mathrm{tr}\left[AA^\top\right] = \mathrm{tr}\left[A(I_n - \Pi_X + \Pi_X)A^\top\right] \\ &= \mathrm{tr}\left[A(I_n - \Pi_X)A^\top\right] + \mathrm{tr}\left[A\Pi_X A^\top\right] \\ &= \mathrm{tr}\left[A(I_n - \Pi_X)A^\top\right] + \mathrm{tr}\left[\left(X^\top X\right)^{-1}\right]. \end{aligned}$$

in view of $AX = I_k$. Here the term $\mathrm{tr}\left[A(I_n - \Pi_X)A^\top\right]$ is nonnegative, since for $C = A(I_n - \Pi_X)$ we have (recall that $I_n - \Pi_X$ is idempotent, since it is a projection)

$$\mathrm{tr}\left[A(I_n - \Pi_X)A^\top\right] = \mathrm{tr}\left[CC^\top\right] \geq 0$$

since $\mathrm{tr}\left[CC^\top\right]$ is a sum of squares (10.40).Thus

$$E\left\|\hat{\mathbf{b}} - \boldsymbol{\beta}\right\|^2 \geq \sigma^2\mathrm{tr}\left[\left(X^\top X\right)^{-1}\right].$$

This lower bound is attained for $A = \left(X^\top X\right)^{-1}X^\top$:

$$(I_n - \Pi_X)A^\top = (I_n - \Pi_X)X\left(X^\top X\right)^{-1} = \mathbf{0} \tag{10.41}$$

since $I_n - \Pi_X$ projects onto $\text{Lin}(X)^\perp$. Thus $\hat{\boldsymbol{\beta}}$ is a BLUE.
It remains to show uniqueness. This follows from the fact that (10.41) is necessary for attainment of the lower bound, and this implies

$$A\Pi_X = A$$

The left side is

$$A\Pi_X = AX\left(X^\top X\right)^{-1} X^\top = \left(X^\top X\right)^{-1} X^\top,$$

and the result is proved. ■
The BLUE property of Gauss-Markov Theorem is not a very strong optimality, since the class of estimators is quite restricted. On the other hand, it says that that $\hat{\boldsymbol{\beta}}$ is *uniformly best* within that class. Recall that the sample mean $\bar{Y}_n$ is a special case, so we obtained another optimality property of the sample mean. This also suggests more optimality properties of $\hat{\boldsymbol{\beta}}$ under normality (minimaxity, best unbiased estimator); these indeed can be established, but we do not discuss this here.

Chapter 11
LINEAR HYPOTHESES AND THE ANALYSIS OF VARIANCE

11.1 Testing linear hypotheses

In the normal linear model (NLM),

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

recall that the parameter $\boldsymbol{\beta}$ varies in the whole space $\mathbb{R}^k$. Consider a linear subspace of $\mathbb{R}^k$, $\mathcal{S}$ say, of dimension $s < k$, and a hypothesis $H : \boldsymbol{\beta} \in \mathcal{S}$. Such a problem, i.e. **testing a linear hypothesis**, arises naturally in a number of situations.

1. Bivariate linear regression. We have

$$Y_i = \alpha + \beta x_i + \varepsilon_i,\ i = 1, \ldots, n.$$

Here $k = 2$, the rows of X are $\mathbf{x}_i^\top = (1, x_i)$, $\boldsymbol{\beta} = (\alpha, \beta)^\top$. A linear hypotheses could be
$H : \beta = 0$,
meaning that the nonrandom regressor variable x_i has **no influence** upon Y_i. If the x_i are i.i.d. realizations of a normal random variable X, $\mathcal{L}(X) = N(\mu_x, \sigma_X^2)$, then the regression coefficient β is (see Definition (10.1.3)

$$\beta = \frac{\sigma_{XY}}{\sigma_X^2}.$$

The hypothesis then means that $\sigma_{XY} = 0$, i.e. X and Y are independent random variables, or equivalenty X and Y are uncorrelated. Indeed the correlation coefficient is

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

and it is related to β by

$$\beta = \rho \frac{\sigma_X}{\sigma_Y}$$

(in "Galton's case" of fathers and sons where $\sigma_X^2 = \sigma_Y^2$ they are actually equal).
Using the linear algebra formalism of (NLM), the linear subspace $\mathcal{S}$ would be the subspace of $\mathbb{R}^2$ spanned by the vector $(1, 0)^\top$, i. e.

$$\mathcal{S} = \mathrm{Lin}((1, 0)^\top).$$

2. Normal location-scale model. Here

$$Y_i = \beta + \varepsilon_i,\ i = 1, \ldots, n,$$

$k = 1$, $X = (1, \ldots, 1)^\top$, ε_i are i.i.d. $N(0, \sigma^2)$, $\sigma^2 > 0$. The only possible linear hypothesis is $H : \beta = 0$.

3. **Polynomial regression.** Here for some design points x_i, $i = 1, \ldots, n$

$$Y_i = \sum_{j=1}^{k} \beta_j \varphi_j(x_i) + \varepsilon_i,\ i = 1, \ldots, n$$

the functions φ_j are $\varphi_j(x) = x^{j-1}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^\top$. A linear hypotheses could be
$H : \beta_k = 0$,
meaning that the regression function

$$r(x) = \sum_{j=1}^{k} \beta_j \varphi_j(x_i)$$

is actually a polynomial of degree $k-2$, and not of degree $k-1$ as postulated in the model (LM). A special case is **1.** above (the bivariate linear regression) for $k = 2$. The hypothesis means that the mean function $r(x)$ is a polynomial of lesser degree, or that the model is *less complex*. Of course, as always with testing problems, the statistically significant result would be a rejection; for e.g. $k = 3$ this means that it is not possible to describe the mean function of the Y_i by a straight line, and a higher degree polynomial is needed.

4. Analysis of variance (ANOVA). Here

$$Y_{jk} = \mu_j + \varepsilon_{jk},\ j = 1, \ldots, m,\ k = 1, \ldots, l \tag{11.1}$$

where ε_{jk} are independent noise variables. The index j is associated with m "treatments" or groups, and one might wish to test whether the treatments have any effect:
H:$\mu_1 = \ldots = \mu_m$.
or in other words, whether the groups differ in the characteristic Y_{jk} measured. The linear subspace $\mathcal{S}$ of $\mathbb{R}^m$ would be

$$\mathcal{S} = \mathrm{Lin}(\mathbf{1})$$

where $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^m$, and $\dim(\mathcal{S}) = 1$ (note that the m here corresponds to k in (10.32), (10.33)).

5. General case. Recall that $\boldsymbol{\xi}_j$ were the colums of the matrix X, so that model (LM) can be written

$$\mathbf{Y} = \sum_{j=1}^{k} \beta_j \boldsymbol{\xi}_j + \boldsymbol{\varepsilon}$$

where the j-th column may have arisen from n independent realizations of the j-th component of a random vector $\mathbf{X}$, or as designed nonrandom values. In either case, $\boldsymbol{\xi}_j$ may be construed as one of k regressor variables which influences the regressand $\mathbf{Y}$*). Such a variable $\boldsymbol{\xi}_j$ is also called an **explanatory variable**, or one of j independent variables, when $\mathbf{Y}$ is the "dependent" variable. Thus a hypothesis $H : \beta_j = 0$ postulates that $\boldsymbol{\xi}_j$ *is without influence on* $\mathbf{Y}$, and can be dispensed with. Clearly the polynomial regression above is a special case for $\boldsymbol{\xi}_j = (\varphi_j(x_1), \ldots, \varphi_j(x_n))^\top$ (designed nonrandom values). Here again, what is sought is the "statistical certainty" in the case of rejection, when it can be claimed that $\boldsymbol{\xi}_j$ *actually does influence* the measured quantity $\mathbf{Y}$.
In the normal linear model (NLM), to find a test statistic for the problem
$H : \boldsymbol{\beta} \in \mathcal{S}$.

*Note that throughout math and statistics software, a vector of values is frequently called a "variable".

$K : \boldsymbol{\beta} \notin \mathcal{S}$
we could again apply the likelihood ratio principle.That was already done in a special case of ANOVA, namely the two sample problem (cf. Theorem 10.3.6). We repeat some of that reasoning, with general notation, assuming first σ^2 unknown. The density of $\mathbf{Y}$ is (as a function of $\mathbf{y} \in \mathbb{R}^n$)

$$p_{\boldsymbol{\beta},\sigma^2}(\mathbf{y}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{y}-X\boldsymbol{\beta}\right\|^2\right).$$

The LR statistic is

$$L(\mathbf{y}) = \frac{\max_{\boldsymbol{\beta}\notin\mathcal{S},\sigma^2>0} p_{\boldsymbol{\beta},\sigma^2}(\mathbf{y})}{\max_{\boldsymbol{\beta}\in\mathcal{S},\sigma^2>0} p_{\boldsymbol{\beta},\sigma^2}(\mathbf{y})}.$$

To find the numerator, we first maximize over $\boldsymbol{\beta} \notin \mathcal{S}$ and then over $\sigma^2 > 0$. Maximizing first over $\boldsymbol{\beta} \in \mathbb{R}^k$**,** we obtain the LSE of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = \left(X^\top X\right)^{-1} X^\top \mathbf{Y}.$$

We now claim that with probability one, $\hat{\boldsymbol{\beta}} \notin \mathcal{S}$, so that $\hat{\boldsymbol{\beta}}$ is also the maximizer over $\boldsymbol{\beta} \notin \mathcal{S}$. Note that

$$\mathcal{L}(\hat{\boldsymbol{\beta}}) = N_k(\boldsymbol{\beta},\sigma^2 \left(X^\top X\right)^{-1})$$

and that the matrix $\left(X^\top X\right)^{-1}$ is nonsingular. It was already argued that a multivariate normal vector, with nonsingular covariance matrix, takes values in a given lower dimensional subspace with probability 0 (conclusion of proof of Theorem 10.4.2). Thus $P(\hat{\boldsymbol{\beta}} \notin \mathcal{S}) = 1$ and almost surely

$$\begin{aligned}\max_{\boldsymbol{\beta}\notin\mathcal{S},\sigma^2>0} p_{\boldsymbol{\beta},\sigma^2}(\mathbf{y}) &= \max_{\sigma^2>0} p_{\hat{\boldsymbol{\beta}},\sigma^2}(\mathbf{y}) \\ &= \max_{\sigma^2>0} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\left\|(I_n - \boldsymbol{\Pi}_X)\mathbf{y}\right\|^2\right).\end{aligned}$$

This maximization problem has been solved already in the proof of Proposition 3.0.5 (insert now $\|(I_n - \boldsymbol{\Pi}_X)\mathbf{y}\|^2$ for S_n^2): the result is

$$\begin{aligned}\max_{\boldsymbol{\beta}\notin\mathcal{S},\sigma^2>0} p_{\boldsymbol{\beta},\sigma^2}(\mathbf{y}) &= \frac{1}{(\hat{\sigma}^2)^{n/2}} \cdot \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{n}{2}\right), \\ \hat{\sigma}^2 &= n^{-1}\left\|(I_n - \boldsymbol{\Pi}_X)\mathbf{y}\right\|^2. \end{aligned} \tag{11.2}$$

Key idea for linear hypotheses in linear models. *To find the maximized likelihood under the hypothesis, we note that $\boldsymbol{\beta} \in \mathcal{S}$ implies that $X\,\boldsymbol{\beta}$ varies in a linear subspace of* $\mathrm{Lin}(X)$*. Indeed if S is a $k \times s$-matrix such that $\mathcal{S} = \mathrm{Lin}(S)$ then every $\boldsymbol{\beta} \in \mathcal{S}$ can be represented as $\boldsymbol{\beta} = S\mathbf{b}$ for some $\mathbf{b} \in \mathbb{R}^s$, thus*

$$X\boldsymbol{\beta} = XS\mathbf{b}$$

where $\mathrm{rank}(XS) = s$ *and we see that*

$$\{X\boldsymbol{\beta},\, \boldsymbol{\beta} \in \mathcal{S}\} = \mathrm{Lin}(XS).$$

We can now apply the results for least squares (or ML) estimation of $\boldsymbol{\beta} \in \mathbb{R}^k$ to estimation of $\mathbf{b} \in \mathbb{R}^s$. We can immediately write down a LSE for $\mathbf{b}$ and a derived one for $\boldsymbol{\beta} \in \mathcal{S}$, but we skip this and proceed directly to the MLE of σ^2 under the hypothesis, analogously to (11.2):

$$\hat{\sigma}_0^2 = n^{-1}\left\|(I_n - \boldsymbol{\Pi}_{XS})\mathbf{y}\right\|^2$$

where the projection $I_n - \Pi_X$ is substituted by $I_n - \Pi_{XS}$. Write the respective linear subspaces of $\mathbb{R}^n$ as

$$\mathcal{X} := \mathrm{Lin}(X),\ \mathcal{S}_0 := \mathrm{Lin}(XS)$$

$\Pi_{\mathcal{X}} = \Pi_X$, $\Pi_{\mathcal{S}_0} = \Pi_{XS}$ for the associated projection operators in $\mathbb{R}^n$. It follows that the maximized likelihood under H is

$$\begin{aligned} \max_{\boldsymbol{\beta}\notin\mathcal{S},\sigma^2>0} p_{\boldsymbol{\beta},\sigma^2}(\mathbf{y}) &= \frac{1}{(\hat\sigma_0^2)^{n/2}}\cdot\frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{n}{2}\right), \\ \hat\sigma_0^2 &= n^{-1}\left\|(I_n - \mathbf{\Pi}_{\mathcal{S}_0})\mathbf{y}\right\|^2. \end{aligned} \tag{11.3}$$

Thus the LR statistic is

$$L(\mathbf{y}) = \left(\frac{\hat\sigma_0^2}{\hat\sigma^2}\right)^{n/2} = \left(\frac{\left\|(I_n - \mathbf{\Pi}_{\mathcal{S}_0})\mathbf{y}\right\|^2}{\left\|(I_n - \mathbf{\Pi}_{\mathcal{X}})\mathbf{y}\right\|^2}\right)^{n/2}. \tag{11.4}$$

For a further transformation, note that the matrix $\mathbf{\Pi}_{\mathcal{X}} - \mathbf{\Pi}_{\mathcal{S}_0}$ is again a projection matrix, namely onto the orthogonal complement of $\mathcal{S}_0$ in $\mathcal{X}$. Denote this orthogonal complement as $\mathcal{X}\ominus\mathcal{S}_0$ (the space of all $\mathbf{x}\in\mathcal{X}$ which are orthogonal to $\mathcal{S}_0$). We then have a decomposition

$$\mathbb{R}^n = \mathcal{X}^\perp \oplus (\mathcal{X}\ominus\mathcal{S}_0)\oplus\mathcal{S}_0 \tag{11.5}$$

where $\oplus$ denotes the orthogonal sum, $\mathcal{X}^\perp = \mathbb{R}^n\ominus\mathcal{X}$ and all three spaces are orthogonal. We have a corresponding decomposition of I_n (which is the projection onto $\mathbb{R}^n$)

$$I_n = (I_n - \mathbf{\Pi}_{\mathcal{X}}) + (\mathbf{\Pi}_{\mathcal{X}} - \mathbf{\Pi}_{\mathcal{S}_0}) + \mathbf{\Pi}_{\mathcal{S}_0}$$

and all three matrices on the right are projection matrices, orthogonal to one another. (Exercise: show that $\mathbf{\Pi}_{\mathcal{X}} - \mathbf{\Pi}_{\mathcal{S}_0}$ is the projection operator onto $\mathcal{X}\ominus\mathcal{S}_0$). As a consequence

$$\begin{aligned} \left\|(I_n - \mathbf{\Pi}_{\mathcal{S}_0})\mathbf{y}\right\|^2 &= \left\|(I_n - \mathbf{\Pi}_{\mathcal{X}} + \mathbf{\Pi}_{\mathcal{X}} - \mathbf{\Pi}_{\mathcal{S}_0})\mathbf{y}\right\|^2 \\ &= \left\|(I_n - \mathbf{\Pi}_{\mathcal{X}})\mathbf{y}\right\|^2 + \left\|(\mathbf{\Pi}_{\mathcal{X}} - \mathbf{\Pi}_{\mathcal{S}_0})\mathbf{y}\right\|^2, \end{aligned}$$

and we obtain from (11.4)

$$L(\mathbf{y}) = \left(1 + \frac{\left\|(\mathbf{\Pi}_{\mathcal{X}} - \mathbf{\Pi}_{\mathcal{S}_0})\mathbf{y}\right\|^2}{\left\|(I_n - \mathbf{\Pi}_{\mathcal{X}})\mathbf{y}\right\|^2}\right)^{n/2}.$$

Note that the form obtained for the two sample problem in Theorem 10.3.6 is a special case; there we could further argue that the LR test is equivalent to a certain t-test.

Definition 11.1.1 *Let Z_1, Z_2 be independent r.v.'s having χ^2-distributions of k_1 and k_2 degrees of freedom, respectively. The* ***F-distribution with k_1, k_2 degrees of freedom*** *(denoted F_{k_1,k_2}) is the distribution of*

$$Y = \frac{k_1^{-1}Z_1}{k_2^{-1}Z_2}.$$

It is possible to write down the density of the F-distribution explicitly, with methods similar those for the t-distribution (Proposition 7.2.7). Note that for $k_1 = 1$, Z_1 is the square of a standard normal, hence Y is the square of a t-distributed r. v. with k_2 degrees of freedom.

Definition 11.1.2 *In the normal linear model (NLM), consider a linear hypothesis given by a linear subspace* $\mathcal{S} \subset \mathbb{R}^k$, $\dim(\mathcal{S}) = s < k$
$H : \boldsymbol{\beta} \in \mathcal{S}$.
$K : \boldsymbol{\beta} \notin \mathcal{S}$.
The F***-statistic*** *for this problem is*

$$F(\mathbf{Y}) = \frac{\|(\mathbf{\Pi}_{\mathcal{X}} - \mathbf{\Pi}_{\mathcal{S}_0})\mathbf{Y}\|^2/(k-s)}{\|(I_n - \mathbf{\Pi}_{\mathcal{X}})\mathbf{Y}\|^2/(n-k)} \tag{11.6}$$

where $\mathcal{X} = \mathrm{Lin}(X)$, $\mathcal{S}_0 = \mathrm{Lin}(XS)$ *and* $\mathcal{S} = \mathrm{Lin}(S)$.

Proposition 11.1.3 *In the normal linear model (NLM), under a hypothesis* $H : \boldsymbol{\beta} \in \mathcal{S}$, *the pertaining* F*-statistic has an* F*-distribution with* $k - s$, $n - k$ *degrees of freedom.*

Proof. Note that under the hypothesis,

$$E\mathbf{Y} = X\boldsymbol{\beta} \in \mathcal{S}_0,$$

hence

$$(\mathbf{\Pi}_{\mathcal{X}} - \mathbf{\Pi}_{\mathcal{S}_0})\mathbf{Y} = (\mathbf{\Pi}_{\mathcal{X}} - \mathbf{\Pi}_{\mathcal{S}_0})(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = (\mathbf{\Pi}_{\mathcal{X}} - \mathbf{\Pi}_{\mathcal{S}_0})\boldsymbol{\varepsilon}.$$

But

$$E\mathbf{Y} = X\boldsymbol{\beta} \in \mathcal{X}$$

is already in the model assumption (LM); it holds under H in particular. Thus

$$(I_n - \mathbf{\Pi}_{\mathcal{X}})\mathbf{Y} = (I_n - \mathbf{\Pi}_{\mathcal{X}})\boldsymbol{\varepsilon}.$$

Let now $\mathbf{e}_1, \ldots, \mathbf{e}_n$ be an orthonormal basis of $\mathbb{R}^n$ such that

$$\begin{aligned} \mathrm{Lin}(\mathbf{e}_1, \ldots, \mathbf{e}_s) &= \mathcal{S}_0,\ \mathrm{Lin}(\mathbf{e}_{s+1}, \ldots, \mathbf{e}_k) = \mathcal{X} \ominus \mathcal{S}_0, \\ \mathrm{Lin}(\mathbf{e}_{k+1}, \ldots, \mathbf{e}_n) &= \mathcal{X}^{\perp} \end{aligned}$$

i.e. the basis conforms to the decomposition (11.5). Then the three projection operators are given by

$$\mathbf{\Pi}_{\mathcal{S}_0}\mathbf{y} = \sum_{i=1}^{s} (\mathbf{e}_i^{\top}\mathbf{y})\mathbf{e}_i, \text{ etc.},$$

hence

$$\|(I_n - \mathbf{\Pi}_{\mathcal{X}})\boldsymbol{\varepsilon}\|^2 = \sum_{i=k+1}^{n} (\mathbf{e}_i^{\top}\boldsymbol{\varepsilon})^2, \|(\mathbf{\Pi}_{\mathcal{X}} - \mathbf{\Pi}_{\mathcal{S}_0})\boldsymbol{\varepsilon}\|^2 = \sum_{i=s+1}^{k} (\mathbf{e}_i^{\top}\boldsymbol{\varepsilon})^2.$$

Note that $z_i = \sigma^{-1}\mathbf{e}_i^{\top}\boldsymbol{\varepsilon}$, $i = 1, \ldots, n$ are i.i.d. standard normals. For the F-statistic we obtain

$$F(\mathbf{Y}) = \frac{\sum_{i=s+1}^{k} z_i^2/(k-s)}{\sum_{i=k+1}^{n} z_i^2/(n-k)}.$$

This proves the result. ■
An immediate consequence is the following.

Theorem 11.1.4 *In the normal linear model (NLM), consider a linear hypothesis given by a linear subspace* $\mathcal{S} \subset \mathbb{R}^k$, $\dim(\mathcal{S}) = s < k$
$H : \boldsymbol{\beta} \in \mathcal{S}$.
$K : \boldsymbol{\beta} \notin \mathcal{S}$.
Let $F(\mathbf{Y})$ *be the pertaining* F*-statistic and* $F_{k-s,n-k;1-\alpha}$ *the lower* $1-\alpha$*-quantile of the distribution* $F_{k-s,n-k}$. *The* ***F-test*** *which rejects when*

$$F(\mathbf{Y}) > F_{k-s,n-k;1-\alpha}$$

is an α*-test.*

11.2 One-way layout ANOVA

Consider a model

$$Y_{jk} = \mu_j + \varepsilon_{jk},\ k = 1, \ldots, l_j,\ j = 1, \ldots, m, \tag{11.7}$$

where ε_{jk} are i.i.d. normal noise variables and μ_j are unknown parameters. The model is slightly more general than (10.18) since we admit different observations numbers l_j for each group j. The total number of observations then is $n = \sum_{j=1}^m l_j$. Clearly this is again a special case of the normal linear model (NLM). Assume that $l_j > 1$ for at least one j.
The groups $j = 1, \ldots, m$ are also called **factors** or **treatments**. The question of interest is "*do the factors have an influence upon the measured quantity* Y_{jk}?" This corresponds to a hypothesis $H : \mu_1 = \ldots = \mu_m$. At first sight, the question appears similar to the one in bivariate regression, where one asks whether a measured regressor quantity x_i has an influence upon the regressand Y_i (the contrary is the linear hypothesis $\beta = 0$). This in turn is related to a question "are the r.v.'s Y and X correlated ?". However the difference is that in ANOVA no numerical value x_i is attached to the groups; the groups or factors are just different categories (they are qualitative in nature). Thus, even if one is willing to assume that individuals are randomly selected from one of the groups, to compute a correlation or regression does not make sense- it is not clear which values X_i should be associated to the groups. An example is the drug testing problem, where one has two samples, one for old and new drug ($m = 2$). For ANOVA with m groups, an example would be that j represents different social groups and Y_{jk} the cholesterol level, or j might represent different regions of space and Y_{jk} the size of cosmic background radiation.
Recall Theorem 10.3.6 where it was shown that the two sample problem (with $H : \mu_1 = \mu_2$) can be treated by a t-test. and that it is a special case of ANOVA. The general ANOVA case for $m \geq 2$ can be treated by an F-test; it suffices to formulate the test problem as a linear hypothesis in a (normal) linear model and apply Theorem 11.1.4.
To write down the F-statistic,

$$F(\mathbf{Y}) = \frac{\|(\mathbf{\Pi}_{\mathcal{X}} - \mathbf{\Pi}_{\mathcal{S}_0})\mathbf{Y}\|^2 / (k-s)}{\|(I_n - \mathbf{\Pi}_{\mathcal{X}})\mathbf{Y}\|^2 / (n-k)} \tag{11.8}$$

it suffices to identify the linear spaces and projection operators involved. We have

$$\boldsymbol{\beta} = (\mu_1, \ldots, \mu_m)^\top, n = \sum_{j=1}^m l_j, \tag{11.9}$$

$$X = \begin{pmatrix} \mathbf{1}_{l_1} & & \\ & \ldots & \\ & & \mathbf{1}_{l_m} \end{pmatrix} \tag{11.10}$$

where $\mathbf{1}_l$ is the l-vector consisting of 1's, X is the $n \times m$ matrix where the vectors $\mathbf{1}_l$ are arranged in a diagonal fashion (with 0's elsewhere),

$$\mathbf{Y} = (Y_{11}, \ldots, Y_{1l_1}, Y_{21_2}, \ldots, Y_{ml_m})^\top, \tag{11.11}$$

and the n-vector $\boldsymbol{\varepsilon}$ is formed analogously. Then

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$\mathrm{rank}(X) = m$, and the hypothesis $H : \mu_1 = \ldots = \mu_m$ can be expressed as

$$H : \boldsymbol{\beta} \in \mathcal{S} = \mathrm{Lin}(\mathbf{1}_m)$$

where $\mathrm{Lin}(\mathbf{1}_m)$ is a one dimensional linear space ($\dim(\mathcal{S}) = s = 1$). Thus

$$\mathcal{S}_0 = \{X\boldsymbol{\beta},\, \boldsymbol{\beta} \in \mathcal{S}\} = \mathrm{Lin}(\mathbf{1}_n)$$

which can also be seen by noting that if all μ_j are equal to one value μ, then then $Y_{jk} = \mu + \varepsilon_{jk}$. To find the value $\|(\boldsymbol{\Pi}_{\mathcal{X}} - \boldsymbol{\Pi}_{\mathcal{S}_0})\mathbf{Y}\|^2$, note that

$$\begin{aligned} \boldsymbol{\Pi}_{\mathcal{S}_0} &= \mathbf{1}_n \left(\mathbf{1}_n^\top \mathbf{1}_n\right)^{-1} \mathbf{1}_n^\top = n^{-1}\mathbf{1}_n\mathbf{1}_n^\top, \\ \boldsymbol{\Pi}_{\mathcal{S}_0}\mathbf{Y} &= \mathbf{1}_n \left(n^{-1}\mathbf{1}_n^\top \mathbf{Y}\right) = \mathbf{1}_n \bar{Y}_{\cdot\cdot} \end{aligned}$$

where $\bar{Y}_{\cdot\cdot}$ is the *overall sample mean*:

$$\bar{Y}_{\cdot\cdot} = n^{-1} \sum_{j=1}^{m} \sum_{k=1}^{l_j} Y_{jk}.$$

Furthermore let

$$\bar{Y}_{j\cdot} = l_j^{-1} \sum_{k=1}^{l_j} Y_{jk}$$

be the *mean within the j-th group*. We claim that

$$\boldsymbol{\Pi}_{\mathcal{X}}\mathbf{Y} = X \left(X^\top X\right)^{-1} X^\top \mathbf{Y} = X \begin{pmatrix} \bar{Y}_{1\cdot} \\ \ldots \\ \bar{Y}_{m\cdot} \end{pmatrix}. \tag{11.12}$$

Indeed, $X^\top\mathbf{Y}$ gives the m-vector of the sums $\sum_{k=1}^{l_j} Y_{jk}$ for each group, and $X^\top X$ is a $m \times m$-diagonal matrix with l_j as diagonal elements. But we can also write

$$\boldsymbol{\Pi}_{\mathcal{S}_0}\mathbf{Y} = \mathbf{1}_n \bar{Y}_{\cdot\cdot} = X \begin{pmatrix} \bar{Y}_{\cdot\cdot} \\ \ldots \\ \bar{Y}_{\cdot\cdot} \end{pmatrix}$$

hence

$$\begin{aligned} \boldsymbol{\Pi}_{\mathcal{X}}\mathbf{Y} - \boldsymbol{\Pi}_{\mathcal{S}_0}\mathbf{Y} &= X \begin{pmatrix} \bar{Y}_{1\cdot} - \bar{Y}_{\cdot\cdot} \\ \ldots \\ \bar{Y}_{m\cdot} - \bar{Y}_{\cdot\cdot} \end{pmatrix}, \\ \|(\boldsymbol{\Pi}_{\mathcal{X}} - \boldsymbol{\Pi}_{\mathcal{S}_0})\mathbf{Y}\|^2 &= \sum_{j=1}^{m} l_j \left(\bar{Y}_{j\cdot} - \bar{Y}_{\cdot\cdot}\right)^2. \end{aligned}$$

Similarly we obtain from (11.12) and (11.11)

$$\begin{aligned}(I_n - \mathbf{\Pi}_{\mathcal{X}})\mathbf{Y} &= (Y_{11} - \bar{Y}_{1\cdot}, \ldots, Y_{1l_1} - \bar{Y}_{1\cdot}, Y_{21} - \bar{Y}_{2\cdot}, \ldots, Y_{ml_m} - \bar{Y}_{l\cdot})^\top, \\ \|(I_n - \mathbf{\Pi}_{\mathcal{X}})\mathbf{Y}\|^2 &= \sum_{j=1}^{m}\sum_{k=1}^{l_j} \left(Y_{jk} - \bar{Y}_{j\cdot}\right)^2.\end{aligned}$$

Thus the F-statistic (11.8) takes the form

$$F(\mathbf{Y}) = \frac{(m-1)^{-1}\sum_{j=1}^{m} l_j \left(\bar{Y}_{j\cdot} - \bar{Y}_{\cdot\cdot}\right)^2}{(n-m)^{-1}\sum_{j=1}^{m}\sum_{k=1}^{l_j}\left(Y_{jk} - \bar{Y}_{j\cdot}\right)^2}. \tag{11.13}$$

The terms

$$\hat{D}_w = \sum_{j=1}^{m}\sum_{k=1}^{l_j}\left(Y_{jk} - \bar{Y}_{j\cdot}\right)^2,\ \hat{D}_b = \sum_{j=1}^{m} l_j \left(\bar{Y}_{j\cdot} - \bar{Y}_{\cdot\cdot}\right)^2$$

can be called the **sums of squares within groups** and and **sums of squares between groups** respectively**.** Recall that we introduced this terminology essentially already in the two sample problem (just before Theorem 10.3.6; there we divided by n and called this "variability"). Consider also the quantity

$$\hat{D}_0 = \sum_{j=1}^{m}\sum_{k=1}^{l_j}\left(Y_{jk} - \bar{Y}_{\cdot\cdot}\right)^2 = \|(I_n - \mathbf{\Pi}_{\mathcal{S}_0})\mathbf{Y}\|^2.$$

This is the total **sums of squares**; then $n^{-1}\hat{D}_0 = S_n^2$ is the total sample variance . We have a decomposition

$$\hat{D}_0 = \hat{D}_w + \hat{D}_b \tag{11.14}$$

as an immediate consequence of the identity

$$\|(I_n - \mathbf{\Pi}_{\mathcal{S}_0})\mathbf{Y}\|^2 = \|(I_n - \mathbf{\Pi}_{\mathcal{X}})\mathbf{Y}\|^2 + \|(\mathbf{\Pi}_{\mathcal{X}} - \mathbf{\Pi}_{\mathcal{S}_0})\mathbf{Y}\|^2.$$

The decomposition (11.14) of the total sample variance S^2 generalizes (10.21); *it gives the name to the test procedure "analysis of variance".* The hypothesis H of equality of means is rejected when the between groups sum of squares $\hat{D}_b$ is too large, compared to the within groups sum of squares $\hat{D}_w$.
Note that the quantities

$$\hat{d}_w = (n-m)^{-1}\hat{D}_w,\ \hat{d}_b = (m-1)^{-1}\hat{D}_b$$

are both unbiased estimates of σ^2 under the hypothesis (cf. the proof of Proposition 11.1.3); they can be called the respective "mean sum of squares". The F-statistic (11.13) involves these quantities $\hat{d}_w$, $\hat{d}_b$ which differ only by a factor from the sums of squares $\hat{D}_w$, $\hat{D}_b$.
A common way of visualizing all the quantities involved is the ANOVA table:

	sum of squares	degrees of freedom	mean s. of squ.	F-value
between groups	$\hat{D}_b$	$m-1$	$(m-1)^{-1}\hat{D}_b = \hat{d}_b$	$F(\mathbf{Y}) = \hat{d}_b/\hat{d}_w$
within groups	$\hat{D}_w$	$n-m$	$(n-m)^{-1}\hat{D}_w = \hat{d}_w$	
total	$\hat{D}_0$	$n-1$		

11.3 Two-way layout ANOVA

Return to the cholesterol /social group example and suppose we have data from various countries, where it can be assumed that the general level of cholesterol varies from country to country (e.g. via different nutritional habits). Nevertheless we are still interested in the same question, namely whether the cholesterol level differs across social groups or not. This can be modeled by a "two way layout", where we allow for additional effects (the countries). It will become clear why the simpler ANOVA in the previous subsection is called "one way layout".
In the previous subsection we could have written the nonrandom group means μ_j as

$$\mu_j \quad = \quad \mu + \delta_j,\, j = 1, \ldots, m,\, \sum_{j=1}^{m} \delta_j = 0$$

.

Any vector $\boldsymbol{\beta} = (\mu_1, \ldots, \mu_m)^\top$ can be written uniquely in this form: set

$$\mu = m^{-1} \sum_{j=1}^{m} \mu_j,\ \delta_j = \mu_j - \mu.$$

The one way layout can be written

$$Y_{jk} = \mu + \delta_j + \varepsilon_{jk},\, k = 1, \ldots, l_j,\, j = 1, \ldots, m,$$

and the hypothesis is

$$\delta_1 = \ldots = \delta_m = 0.$$

The δ_j are called **factor effects** and μ is called the **main effect** . Then the Y_{jk} can be decomposed

$$\begin{aligned} Y_{jk} &= \bar{Y}_{\cdot\cdot} + \hat{\delta}_j + \ \hat{\varepsilon}_{jk},\, k = 1, \ldots, l_j,\, j = 1, \ldots, m, \\ \hat{\delta}_j &= \bar{Y}_{j\cdot} - \bar{Y}_{\cdot\cdot},\, \hat{\varepsilon}_{jk} = Y_{jk} - \bar{Y}_{j\cdot}, \end{aligned} \tag{11.15}$$

where $\hat{\varepsilon}_{jk}$ are called **residuals** and $\hat{\delta}_j$ can be interpreted as estimates of the factor effects δ_j. The relation (11.15) can be written as a decomposition of the data vector $\mathbf{Y}$

$$\mathbf{Y} = \mathbf{\Pi}_{\mathcal{S}_0}\mathbf{Y} + (\mathbf{\Pi}_{\mathcal{X}} - \mathbf{\Pi}_{\mathcal{S}_0})\mathbf{Y} + (I_n - \mathbf{\Pi}_{\mathcal{X}})\mathbf{Y}.$$

The $F-$statistic (11.13) then takes the form

$$F(\mathbf{Y}) = \frac{(m-1)^{-1} \sum_{j=1}^{m} l_j \hat{\delta}_j^2}{(n-m)^{-1} \sum_{j=1}^{m} \sum_{k=1}^{l_j} \hat{\varepsilon}_{jk}^2}.$$

This indicates how the two way layout can be treated. The model is

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk},\, k = 1, \ldots, l,\, i = 1, \ldots, q, \quad j = 1, \ldots, m,$$

where ε_{ijk} are i.i.d normal variables with variance σ^2. For simplicity we assume that all groups (i, j) have an equal number of observations l. The index i is called the **first factor** and j is called the **second factor**. Again this is a normal linear model, but the matrix X has an involved form. It is somewhat laborious to work out all the projections and derived sums of squares; we therefore

forgo the projection approach and use the more elementary multiple "dot" notation. Note that the projection approach is still needed for a rigourous proof that the tests statistics below have an F-distribution. The array of nonrandom mean values μ_{ij} can be written

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \tag{11.16}$$

$$\sum_{i=1}^{q} \alpha_i = \sum_{j=1}^{m} \beta_j = 0 \tag{11.17}$$

where

$$\begin{aligned}
\mu &= \bar{\mu}_{..} = q^{-1}\sum_{i=1}^{q}\bar{\mu}_{i.} = m^{-1}\sum_{j=1}^{m}\bar{\mu}_{.j} \text{ is the } \textbf{global effect} \\
\alpha_i &= \bar{\mu}_{i.} - \bar{\mu}_{..} \text{ is the } \textbf{main effect} \text{ of value } i \text{ of the first factor} \\
\beta_j &= \bar{\mu}_{.j} - \bar{\mu}_{..} \text{ is the } \textbf{main effect} \text{ of value } j \text{ of the second factor} \\
\gamma_{ij} &= \mu_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \bar{\mu}_{..} \text{ is the } \textbf{interaction} \\
&\quad \text{of value } i \text{ of the first factor and value } j \text{ of the second factor.}
\end{aligned}$$

(note that relation (11.16, 11.17) immediately follow from the definitions above of the quantities involved). It follows also that

$$\begin{aligned}
\sum_{i=1}^{q}\gamma_{ij} &= q\bar{\mu}_{.j} - q\bar{\mu}_{..} - q(\bar{\mu}_{.j} - \bar{\mu}_{..}) = 0 \text{ for all } j = 1, \ldots, m, \\
\sum_{j=1}^{m}\gamma_{ij} &= 0 \text{ for all } i = 1, \ldots, q.
\end{aligned}$$

Assume now that there is no interaction between the factors 1 **and** 2, i.e. $\gamma_{ij} = 0$. In this case from (11.16) we obtain

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk},\ k = 1, \ldots, l,\ i = 1, \ldots, q,\ \ j = 1, \ldots, m.$$

Set $n = mql$. The two hypotheses of interest are
$H_1 : \alpha_1 = \ldots = \alpha_q = 0$
$H_2 : \beta_1 = \ldots = \beta_m = 0$
In our example, if factor 1 (indexed by i) is social group and factor 2 (indexed by j) is country, then H_1 would be "cholesterol level does not depend on social group, even though it may vary across countries" and H_2 would be "cholesterol level does not depend on country, even though it may vary across social groups".
The analog of (11.14) is

$$\hat{D}_0 = \hat{D}_w + \hat{D}_{b1} + \hat{D}_{b2}, \tag{11.18}$$

$$\hat{D}_{b1} = lm\sum_{i=1}^{q}\left(\bar{Y}_{i..} - \bar{Y}_{...}\right)^2,\ \hat{D}_{b2} = lq\sum_{j=1}^{m}\left(\bar{Y}_{.j.} - \bar{Y}_{...}\right)^2,$$

$$\hat{D}_w = \sum_{i,j,k}\hat{\varepsilon}_{ijk}^2,\ \hat{D}_0 = \sum_{i,j,k}(Y_{ijk} - \bar{Y}_{...})^2.$$

where

$$\hat{\varepsilon}_{ijk} = Y_{ijk} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}$$

are the residuals. The two-way ANOVA table is

	sum of squares	degrees of freedom	mean s. of squ.	F-value
1. factor, between gr.	$\hat{D}_{b1}$	$q-1$	$\hat{D}_{b1}/(q-1) = \hat{d}_{b1}$	$\hat{d}_{b1}/\hat{d}_w$
2. factor, between gr.	$\hat{D}_{b2}$	$m-1$	$\hat{D}_{b2}/(m-1) = \hat{d}_{b2}$	$\hat{d}_{b2}/\hat{d}_w$
residuals (within gr)	$\hat{D}_w$	$n-m-q+1$	$\hat{D}_w/(n-m-q+1) = \hat{d}_w$	
total	$\hat{D}_0$	$n-1$		

Here the F-value in the row for the first factor is for testing H_1.

Remark 11.3.1 *We briefly outline the associated projection arguments needed for the proof of the F-distributions. Let Y be the n-vector of the data Y_{ijk}, e.g. arranged in a lexicographic fashion:*

$$\mathbf{Y} = (Y_{111}, Y_{112}, \ldots, Y_{11l}, Y_{121}, \ldots, Y_{qml})^\top$$

let $\mathcal{S}_{00}$ be the subspace of $\mathbb{R}^n$ spanned by the vector $\mathbf{1}_n$,

$$\begin{aligned} \mathcal{S}_{10} &= \left\{ \mathbf{Z} : Z_{ijk} = \alpha_i, \text{ for some } \alpha_i \text{ where } \sum_{i=1}^{q} \alpha_i = 0 \right\}, \\ \mathcal{S}_{01} &= \left\{ \mathbf{Z} : Z_{ijk} = \beta_j, \text{ for some } \beta_j \text{ where } \sum_{i=1}^{q} \beta_j = 0 \right\}. \end{aligned}$$

Note that $\mathcal{S}_{10}, \mathcal{S}_{01}$ and $\mathcal{S}_{00}$ are mutually orthogonal in $\mathbb{R}^n$: e.g. for any $\mathbf{Z}_1 \in \mathcal{S}_{10}$ and $\mathbf{Z}_2 \in \mathcal{S}_{01}$ we have $\mathbf{Z}_1^\top \mathbf{Z}_2 = 0$ etc. Consider the linear space $\mathcal{X}$ spanned by these three subspaces, i.e. the set of all linear combinations $\sum_{r=1}^{3} \lambda_r \mathbf{Z}_r$ where $\mathbf{Z}_1 \in \mathcal{S}_{10}$, $\mathbf{Z}_2 \in \mathcal{S}_{01}$, $\mathbf{Z}_3 \in \mathcal{S}_{00}$. This can be represented

$$\mathcal{X} = \left\{ \begin{array}{c} \mathbf{Z} : Z_{ijk} = \mu + \alpha_i + \beta_j, \text{ for some } \mu, \alpha_i, \beta_j, \\ \text{where } \sum_{i=1}^{q} \alpha_i = 0, \ \sum_{i=1}^{q} \beta_j = 0 \end{array} \right\}.$$

or in short notation (using the orthogonal sum operation $\oplus$)

$$\mathcal{X} = \mathcal{S}_{10} \oplus \mathcal{S}_{00} \oplus \mathcal{S}_{01}.$$

To this corresponds a representation of projection operators:

$$\Pi_{\mathcal{X}} = \Pi_{\mathcal{S}_{00}} + \Pi_{\mathcal{S}_{10}} + \Pi_{\mathcal{S}_{01}}$$

where $\Pi_{\mathcal{S}_{00}} = \Pi_{\mathbf{1}_n} = n^{-1} 1_n 1_n^\top$. The basic assumption of no interaction $\gamma_{ij} = 0$ in (11.16) means that

$$E\mathbf{Y} \in \mathcal{X}.$$

(If the assumption of no interaction is not made then we were only able to claim $E\mathbf{Y} \in \mathcal{X}_0$ where

$$\mathcal{X}_0 = \left\{ \mathbf{Z} : Z_{ijk} = \mu_{ij}, \text{ for some } \mu_{ij} \right\}).$$

The basic (normal) linear model (under the assumption of no interaction) can be expressed as

$$\begin{aligned} \mathcal{L}(\mathbf{Y}) &= N_n(E\mathbf{Y}, \sigma^2 I_n) \\ \text{where } E\mathbf{Y} &\in \mathcal{X}. \end{aligned}$$

To write in the form $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ *we need a matrix* X *which spans the space* $\mathcal{X}$*; we can avoid this here. The two way ANOVA decomposition (11.18) can be written as*

$$\begin{aligned} \hat{D}_0 &= \|(I_n - \Pi_{\mathcal{S}_{00}})\mathbf{Y}\|^2 = \|(I_n - \Pi_{\mathcal{X}})\mathbf{Y}\|^2 + \|\Pi_{\mathcal{S}_{10}}\mathbf{Y}\|^2 + \|\Pi_{\mathcal{S}_{01}}\mathbf{Y}\|^2 \\ &= \hat{D}_w + \hat{D}_{b1} + \hat{D}_{b2}. \end{aligned}$$

In this linear model, the expression

$$\hat{d}_w = (n - m - q + 1)^{-1}\hat{D}_w = (n - m - q + 1)^{-1}\|(I_n - \Pi_{\mathcal{X}})\mathbf{Y}\|^2$$

is an unbiased estimator of σ^2 *and* $\hat{D}_w/\sigma^2$ *has a law* $\chi^2_{n-m-q+1}$*. The two linear hypotheses are*
$H_1 : \alpha_1 = \ldots = \alpha_q = 0$ *or equivalently* $E\mathbf{Y} \in \mathcal{S}_{01} \oplus \mathcal{S}_{00}$
$H_2 : \beta_1 = \ldots = \beta_m = 0$ *or equivalently* $E\mathbf{Y} \in \mathcal{S}_{10} \oplus \mathcal{S}_{00}$.
Under H_1*, we have* $\Pi_{\mathcal{S}_{10}} E\mathbf{Y} = \mathbf{0}$ *and hence the expression*

$$\hat{d}_{b1} = (q - 1)^{-1}\hat{D}_{b1} = (q - 1)^{-1}\|\Pi_{\mathcal{S}_{10}}\mathbf{Y}\|^2$$

is independent of $\hat{d}_w$ *and such that* $\hat{D}_{b1}/\sigma^2$ *has a law* χ^2_{q-1}*. Thus* $\hat{d}_{b1}/\hat{d}_w$ *has law* $F_{q-1,n-m-q+1}$.

The textbook in chap. 11.3 treats the case where $l = 1$ and β_j are random variables (RCB, randomized complete block design). This model is very similar to the two way layout treated here. The theory of ANOVA with its associated design problems has many further ramifications.

Chapter 12
SOME NONPARAMETRIC TESTS

Recall Remark 9.4.6: a family of probability distributions $\mathcal{P} = \{P_\vartheta, \vartheta \in \Theta\}$, indexed by ϑ, is called parametric if all $\vartheta \in \Theta$ are finite dimensional vectors ($\Theta \subseteq \mathbb{R}^k$ for some k), otherwise $\mathcal{P}$ is called nonparametric. In hypothesis testing, any hypothesis corresponds to some $\mathcal{P}$, thus the terminology is extended to hypotheses. Any simple hypothesis (consisting of *one* probability distribution $\mathcal{P} = \{Q_0\}$) is parametric. The understanding is that the family $\mathcal{P}$ is *too large* to be indexed by some $\vartheta \subseteq \mathbb{R}^k$; it should be indexed by some set of functions or other objects (e.g. the associated distribution functions, or the densities if they exist, or the infinite series of their Fourier ceofficients, or the probability laws themselves). Typically some restrictions are then placed on the functions involved, such that: the density is symmetric around 0, or is differentiable with derivative bounded by a constant etc.
In the χ^2-goodness-of-fit test treated in Corollary 9.4.5, we already encountered a nonparametric alternative $K : Q \neq Q_0$. In a more narrow sense, a nonparametric test is one in which the hypothesis H is a nonparametric set. It was also noted that the χ^2-test for goodness of fit actually tests the hypothesis on the cell probabilities $\mathbf{p}(Q) = \mathbf{p}(Q_0)$, with asymptotic level α, and the set of all Q fulfilling this hypothesis is also nonparametric.
A genuinely nonparametric test was the χ^2-test for independence (χ^2-test in a contingency table) treated in Section 9.6: the hypothesis consists of all joint distributions of two r.v.s's X_1, X_2 which are the product of their marginals.

12.1 The sign test

Suppose that for a pair of (possibly dependent) random variables X, Y, one is interested in the difference $Z = X - Y$, more precisely in the median of Z, i.e. med(Z). If X and Y represent some measurements or "treatments", then med$(Z) > 0$ would mean that X is "better" in some sense than Y. Assume that Z has a continuous distribution, so that $P(Z = 0) = 0$. If one observes n independent pairs of r.v.'s (X_i, Y_i), $i = 1, \ldots, n$ all having the distribution of (X, Y) then corresponding hypotheses would be
H :med$(Z) = 0$
K :med$(Z) > 0$.
A possible test statistic is

$$S = \sum_{i=1}^{n} \text{sgn}(Z_i)$$

where $Z_i = X_i - Y_i$ and

$$\text{sgn}(x) = \begin{cases} 1 \text{ if } x > 0 \\ 0 \text{ if } x = 0 \\ -1 \text{ if } x < 0. \end{cases}$$

It is plausible to reject H if the number of positive Z_i is too large compared to the number of negative Z_i. This is the (one sided) **sign test** for H. Define

$$S_0 = \sum_{i=1}^{n} \mathbf{1}_{(0,\infty)}(Z_i);$$

then $S = 2S_0 - n$. Moreover if $p = P(Z > 0)$ then the statistic S_0 has a binomial distribution $B(n, p)$. Under H we have

$$P(Z < 0) = P(Z > 0)$$

so that $p = 1/2$ and the distribution of S_0 under H is $B(n, 1/2)$. Since S is a strictly monotone function of S_0, the sign test can also be based on S and the binomial distribution under H (rejection when S is too large). Note that since $B(n, 1/2)$ is a discrete distribution, in order to achieve exact size α under the hypothesis, for any given α one has to use a randomized test in general (alternatively, for a nonrandomized α-test the size may be less than α).
This is a prototypical nonparametric test; the hypothesis H contains all distributions Q for Z which have median 0, and this is a large nonparametric class of laws. The statistic S is **distribution-free** under H, since its law is $B(n, 1/2)$ which does not depend on Q in H.

12.2 The Wilcoxon signed rank test

The hypothesis $H : \text{med}(Z) = 0$ in the sign test has the disadvantage that it includes distributions for Z which are very skewed, in the sense e.g. that most of the positive values may be concentrated around a large value ($\mu_+ > 0$ say) whereas the probability mass on the negative half-axis may be concentrated around a moderate negative value ($\mu_- < 0$ say). In that case one would tend to say that X is "better" than Y in some sense, even though the median of $Z = X - Y$ is still 0. Thus one formulates the **hypothesis of symmetry**
$H : P(X - Y > c) = P(Y - X > c)$ for all $c > 0$
or equivalently
$H : P(Z > c) = P(Z < -c)$ for all $c > 0$.
Assume that Z_i, $i = 1, \ldots, n$ are i.i.d. having the distribution of $Z = X - Y$ (a continuous distribution). The rank of Z_i among $Z_1, \ldots, Z_n$ denoted by R_i is defined to be the number of Z_j's satisfying $Z_j \leq Z_i$, $j = 1, \ldots, n$. The rank of $|Z_i|$ among $|Z_1|, \ldots, |Z_n|$ is similarly defined and denoted by $\tilde{R}_i$. Because of the continuity assumption, we assume that no two Z_i are equal and that no Z_i is zero. Define **signed ranks**

$$\tilde{S}_i = \text{sgn}(Z_i)\tilde{R}_i$$

and the **Wilcoxon signed rank statistic** W

$$W_n = \sum_{i=1}^{n} \tilde{S}_i.$$

Under the hypothesis of symmetry, about one half of the $\tilde{S}_i$ would be negative; thus $\tilde{S}_i$ would be close to 0. Thus it seems plausible to reject H if W is too large (one sided test) or if $|W|$ is too large (two sided test).

Lemma 12.2.1 *Under the hypothesis H of symmetry, W has the same distribution as*

$$V_n = \sum_{i=1}^{n} M_i i \tag{12.1}$$

where M_i are independent Rademacher random variables, i.e.

$$P(M_i = -1) = P(M_i = -1) = 1/2$$

(or $M_i = 2B_i - 1$ where B_i are Bernoulli $B(1, 1/2)$).

Proof. If Z has a symmetric distribution then $\mathrm{sgn}(Z)$ and $|Z|$ are independent:

$$\begin{aligned} P(|Z| &> c|\mathrm{sgn}(Z) = 1) = P(|Z| > c|Z > 0) = 2P(Z > c) \\ &= 2P(Z < -c) = P(|Z| > c|Z < 0) = P(|Z| > c|\mathrm{sgn}(Z) = -1) \end{aligned}$$

thus the conditional distribution of $|Z|$ given $\mathrm{sgn}(Z)$ does not depend on $\mathrm{sgn}(Z)$, which means independence. Thus W has the same law as

$$V_n^* = \sum_{i=1}^{n} M_i \tilde{R}_i$$

where M_i are independent of the original sample of Z_i. Moreover, a random permutation of the $(M_1, \ldots, M_n)$ (independent of the M_i) does not change the law of the vector $(M_1, \ldots, M_n)$, so that $\mathcal{L}(V^*) = \mathcal{L}(V)$. ■

Note that $EM_i = 0$ so that $EW = 0$ under H. The variance of M_i is

$$\mathrm{Var}(M_1) = EM_1^2 = \frac{1}{2}(-1)^2 + \frac{1}{2}(1)^2 = 1$$

so that

$$\mathrm{Var}(W_n) = \mathrm{Var}(V_n) = \sum_{i=1}^{n} i^2 = \frac{1}{6}\left(n(n+1)(2n+1)\right) =: v_n.$$

(the last formula can easily be proved by induction).

To obtain a test, we can now use a normal approximation for the law of W_n or find the quantiles of its exact law. The normal approximation for the law of V_n seems plausible since M_i are i.i.d. zero mean and the weights i in (12.1) are deterministic. An appropriate central limit theorem (Lyapunov CLT) gives

$$\frac{W_n}{v_n^{1/2}} \Longrightarrow_d N(0, 1).$$

The corresponding asymptotic α-test of H then rejects if $W_n > z_\alpha v_n^{1/2}$ where z_α is the upper α-quantile of $N(0, 1)$.

For the exact distribution, define

$$W_n^+ = \sum_{i:\mathrm{sgn}(Z_i)=1} \tilde{R}_i, W_n^- = \sum_{i:\mathrm{sgn}(Z_i)=-11} \tilde{R}_i.$$

The test can also be based on W_n^+ since

$$\begin{aligned} W_n^+ + W_n^- &= \sum_{i=1}^{n} i = \frac{n(n+1)}{2} \\ W &= W_n^+ - W_n^- = 2W_n^+ - n(n+1)/2 \end{aligned}$$

By Lemma (12.2.1), the law of W_n^+ coincides with that of

$$V_n^+ = \sum_{i=1}^{n} B_i i$$

where B_i are i.i.d. Bernoulli $B(1, 1/2)$. This distribution has been tabulated in the past (one sided critical values, without randomization, for selected values of α and $n = 1, \ldots, 20$ *) and this can easily be included in statistical software today. Note that the two sided critical values for W can be obtained from the fact that W has a symmetric distribution around 0.

Further justification. Recall that when Z_i are i.i.d. normal $N(\mu, \sigma^2)$ with unknown σ^2 then the hypothesis of symmetry around 0 reduces to $H : \mu = 0$. For this the t-test is available, based on $T = n^{1/2}\bar{Z}_n/\hat{S}_n$ or when σ^2 is known we could even use a Z-test based on the statistic $Z^0 = n^{1/2}\bar{Z}_n/\sigma$ and its normal distribution. The one sided Z-test was UMP test against alternatives $H : \mu > 0$, and the two sided Z- and t-tests can also be shown to have an optimality property (UMP unbiased tests). When the assumption of normality of $Q = \mathcal{L}(Z_i)$ is not justified, for testing symmetry we could try still to use the t-test: we have

$$n^{1/2}\bar{Z}_n/\hat{S}_n \Longrightarrow_d N(0, 1)$$

if the second moment of the Q exists. In that case we would obtain an asymptotic α-test for the hypothesis of symmetry, but this breaks down if the class of admitted distributions is too large, i.e. the second moment of Q may not exist. For instance, we may have $Q(\cdot) = Q_0(\cdot - \mu)$ where Q_0 is a Cauchy distribution (which is symmetric about 0). Symmetry of Q means $\mu = 0$, and T is not an appropriate test statistic (does not provide an asymptotic α-test). In contrast, the Wilcoxon signed statistic W is distribution free under the hypothesis of symmetry: according to Lemma (12.2.1) i.e. its law does not depend on Q as long as Q is symmetric.

Nonparametric alternatives. A random variable Z_1 with distribution function F_1 is **stochastically larger** than Z_2 with distribution function F_2 if

$$\begin{aligned} P(Z_1 &> x) \geq P(Z_1 > x) \text{ for all } x \in \mathbb{R} \\ \text{and there exists } x_0 &: P(Z_1 > x_0) > P(Z_1 > x_0). \end{aligned}$$

If, for distribution functions we define the symbol $F_1 \precsim F_2$ to mean "$F_1 \leq F_2$ and $F_1(x_0) < F_2(x_0)$ for at least one x_0" then is equivalent to

$$F_1 \precsim F_2.$$

An

Invariance considerations. Let z be a point in $\mathbb{R}^n$ (thought of as representing a realization of $\mathbf{Z} = (Z_1, \ldots, Z_n)$) and let π be a transformation $\pi(z) = (\tau(z_1), \ldots, \tau(z_n))$ where τ is a continuous,

*cf. e.g. the table given in Rohatgi and Saleh, Introduction to Prob. and Statistics, 2nd Ed.

strictly monotone increasing and *uneven* real valued function on $\mathbb{R}$ (uneven means $\tau(x) = -\tau(-x)$). If $Q = \mathcal{L}(Z)$ is a symmetric law then the law of $\tau(Z)$ is also symmetric:

$$\begin{aligned} P(\tau(Z) &> c) = P(Z > \tau^{-1}(c)) = P(Z < -\tau^{-1}(c)) \\ &= P(Z < \tau^{-1}(-c)) = P(\tau(Z) < -c) \text{ for all } c > 0. \end{aligned}$$

Note that the statistic

$$L(\mathbf{Z}) = (\operatorname{sgn}(Z_i), \tilde{R}_i)_{i=1,\ldots,n}$$

is invariant under any transformation π applied to $\mathbf{Z}$. It can be shown that L is *maximal invariant* under the group of transformation of $\mathbb{R}^n$ given by all π, i.e. any other invariant map is a function of $L(\mathbf{Z})$. Thus the set of all signs and of all ranks $\tilde{R}_i$ of $|Z_i|$ is a maximal invariant. This provides a justification for use of the Wilcoxon signed rank statistic W.

Chapter 13
EXERCISES

13.1 Problem set H1

Exercise H1.1. Show that in model $\mathbf{M}_1$, with parameter $p \in \Theta = [0,1]$ and risk function $R(T,p)$ defined by

$$R(T,p) = E_p\left(T(X) - p\right)^2 \qquad (*)$$

there is no estimator $T(X)$ such that

$$R(T,p) = 0 \text{ for all } p \in \Theta.$$

Exercise H1.2. Let p_i, $i = 1,\ldots,k$ be a finite subset of $(0,1)$ (where $k \geq 2$), and consider a statistical model $\mathbf{M}_1'$ with the same data X and parameter p as $\mathbf{M}_1$ but where p is now restricted to $\Theta = \{p_1,\ldots,p_k\}$. Assume that each parameter value $p_i \in \Theta$ is assigned a prior probability $q_i > 0$ where $\sum_{i=1}^k q_i = 1$. For any estimator T, define the mixed risk

$$B(T) = \sum_{i=1}^{k} R(T,p_i)q_i$$

where $R(T,p)$ is again the quadratic risk (*).
a) Find the form of the Bayes estimator T_B (i.e the minimizer of $B(T)$) and show that it is unique.
b) Show that T_B is admissible in the model $\mathbf{M}_1'$.

13.2 Problem set H2

Exercise H2.1. Let $X_1,...,X_n$ be independent and identically distributed with Poisson law $\mathrm{Po}\,(\lambda)$, where $\lambda \geq 0$ is unknown. (The Poisson law with parameter $\lambda = 0$ is defined as the one point distribution at 0, where $X_1 = 0$ with probability one). Find the maximum likelihood estimator (MLE) of λ (proof).
Exercise H2.2 Let $X_1,\ldots,X_n$ be independent and identically distributed such that X_1 has the uniform law on the set $\{1,\ldots,r\}$ for some integer $r \geq 1$ (i.e. $P_r(X_1 = k) = 1/r$, $k = 1,\ldots,r$). In the statistical model where $r \geq 1$ is unknown, find the MLE of r (proof).
Exercise H2.3. Let $X_1,\ldots,X_n$ be independent and identically distributed such that X_1 has the geometric law $\mathrm{Geom}(p)$, i.e.

$$P_p(X_1 = k) = (1-p)^{k-1}\,p,\ k = 1,2,\ldots$$

i) In the statistical model where $p \in (0,1]$ is unknown, find the MLE of p (proof).
ii) Compute (or find in a textbook) the expectation of X_1 under p.

Exercise H2.4 Let $X_1, \ldots, X_n$ be independent and identically distributed such that X_1 has the uniform law on the interval $[0, \vartheta]$ for some $\vartheta > 0$ (i. e. X_1 has the uniform density $p_\vartheta(x) = \vartheta^{-1}\mathbf{1}_{[0,\vartheta]}(x)$. Here $\mathbf{1}_A(x)$ is the indicator function of a set A: $\mathbf{1}_A(x) = 1$ if $x \in A$, $\mathbf{1}_A(x) = 0$ otherwise). In the statistical model where $\vartheta > 0$ is unknown, find the MLE of ϑ (proof).
Hint: it can be assumed that not all X_i are 0, since this event has probability 0 under any $\vartheta > 0$.
Exercise H2.5 Let $X_1, \ldots, X_n$ be independent and identically distributed such that X_1 has the exponential law $\mathrm{Exp}(\lambda)$ on $[0, \infty)$ with parameter λ (i. e. X_1 has density $p_\vartheta(x) = \lambda^{-1}\exp(-x\lambda^{-1})\mathbf{1}_{[0,\infty)}(x)$).
i).In the statistical model where $\lambda > 0$ is unknown, find the MLE of λ (proof)
ii) Recall the expectation of X_1 under λ.
.

13.3 Problem set H3

Exercise H3.1. Let $X_1, ..., X_n$ be independent and identically distributed with Poisson law $\mathrm{Po}(\vartheta)$, where $\vartheta \in \Theta = (0, \infty)$ is unknown.
(i) Compute the Fisher information at each $\vartheta \in \Theta$.
(ii) Assume that condition $\mathbf{D}_2$ for the validity of the Cramer-Rao bound is fulfilled (for $n = 1$ it is shown in the handout). Find a minimum variance unbiased estimator (UMVUE) of ϑ for this model.

Exercise H3.2. Let X be an observed (integer-valued) random variable with binomial law $B(n, \vartheta)$ where $\vartheta \in \Theta = (0, 1)$ is unknown.
(i) Compute the Fisher information at each $\vartheta \in \Theta$.
(ii) Clearly condition $\mathbf{D}_1$ for the validity of the Cramer-Rao bound is fulfilled (sample space is finite, prob. function differentiable). Find a minimum variance unbiased estimator (UMVUE) of ϑ for this model.

Exercise H3.3. Let $X_1, ..., X_n$ be independent and identically distributed with Geometric law $\mathrm{Geom}(\vartheta^{-1})$, i. e.

$$P_\vartheta(X_1 = k) = (1 - \vartheta^{-1})^{k-1}\vartheta^{-1}$$

where $\vartheta \in \Theta = (1, \infty)$ is unknown. (Note that in Exercise H2.3 the family $\mathrm{Geom}(p)$, $p \in (0, 1]$ was considered. Here we just took $\vartheta = p^{-1}$ as parameter and also excluded the value $p = 1$)
(i) Compute the Fisher information at each $\vartheta \in \Theta$.
(ii) Assume that condition $\mathbf{D}_2$ for the validity of the Cramer-Rao bound is fulfilled . Find a minimum variance unbiased estimator (UMVUE) of ϑ for this model.
Hint: Here the variance of X_1 is important; compute or look up in a book.

13.4 Problem set H4

Exercise H4.1. Let $X_1, \ldots, X_n$ be independent and identically distributed such that X_1 has the exponential law $\mathrm{Exp}(\lambda)$ on $[0, \infty)$ with parameter λ (i. e. X_1 has density $p_\lambda(x) = \lambda^{-1}\exp(-x\vartheta\lambda^{-1})\mathbf{1}_{[0,\infty)}(x)$), where λ is unknown and varies in the set $\Theta = (0, \infty)$.
(i) Compute the Fisher information at each $\lambda \in \Theta$.
(ii) Assume that condition $\mathbf{D}_3$ for the validity of the Cramer-Rao bound is fulfilled. Find a minimum variance unbiased estimator (UMVUE) of λ for this model.

Exercise H4.2. Consider the **Gaussian scale model:** observations are $X_1, \ldots, X_n$, independent and identically distributed such that X_1 has the normal law $N(0, \sigma^2)$ with variance σ^2, where σ^2 is unknown and varies in the set $\Theta = (0, \infty)$.
(i) Compute the Fisher information $I_F(\sigma^2)$ for one observation X_1.
(ii) Assume that condition $\mathbf{D}_3$ for the validity of the Cramer-Rao bound is fulfilled. Show that for n observations in the Gaussian scale model, the sample variance

$$S^2 = n^{-1} \sum_{i=1}^{n} X_i^2$$

is a uniformly best unbiased estimator.
Hints: (a) Note that σ^2 is treated as parameter, not σ; so it may be convenient to write ϑ for σ^2 when taking derivatives.
(b) Note that

$$\mathrm{Var}_{\sigma^2} X_1^2 = 2\sigma^4.$$

A short proof runs as follows. We have $X_1 = \sigma Y$ for standard normal Z, so it suffices to prove $\mathrm{Var} Z^2 = 2$. Now $\mathrm{Var} Z^2 = EZ^4 - (EZ^2)^2$, so it suffices to prove $EZ^2 = 3$. For the standard normal density φ we have by partial integration, using $\varphi'(x) = -x\varphi(x)$

$$\int x^4 \varphi(x) dx = -\int x^3 \varphi'(x) dx = 3 \int x^2 \varphi(x) dx = 3.$$

Exercise H4.3. In the handout, sec. 8.3 a family of Gamma densities was introduced as

$$f_\alpha(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha - 1} \exp(-x)$$

for $\alpha > 0$. Define more generally, for some $\gamma > 0$

$$f_{\alpha,\gamma}(x) = \frac{1}{\Gamma(\alpha)\gamma^\alpha} x^{\alpha - 1} \exp(-x\gamma^{-1}).$$

The corresponding law is called the $\Gamma(\alpha, \gamma)$-distribution.
(i) Let $X_1, ..., X_n$ be independent and identically distributed with Poisson law $\mathrm{Po}\,(\vartheta)$, where $\vartheta \in \Theta = (0, \infty)$ is unknown. Assume that the results on Bayesian inference in section 8 carry over from finite to countable sample space. Show that the family $\{\Gamma(\alpha, \gamma),\ \alpha > 0, \gamma > 0\}$ is a conjugate family of prior distributions.
(ii) Show that for a r.v. U with law $\Gamma(\alpha, \gamma)$

$$EU = \alpha\gamma.$$

(iii) In the above Poisson model, with prior $\Gamma(\alpha, \gamma)$, find the posterior expectation $E(\vartheta|X)$ of ϑ and discuss its relation to the sample mean $\bar{X}_n$ for large n (α, γ fixed).
Remark: Note that if Proposition 8.1 carries over to the case of countable sample space then $E(\vartheta|X)$ is a Bayes estimator for quadratic risk.

13.5 Problem set H5

Exercise H5.1. Consider the Gaussian location model with restricted parameter space $\mu \in \Theta = [-K, K]$, where $K > 0$, sample size $n = 1$ and $\sigma^2 = 1$. A linear estimator is of form $T(X) = aX + b$ where a, b are nonrandom (fixed) real numbers.
(i) Find the minimax linear estimator T_{LM} (note that all a, b are allowed)
(ii) Show that T_{LM} is strictly better than the sample mean $\bar{X}_n = X$, everywhere on $\Theta = [-K, K]$ (this implies that X is not admissible).
(iii) Show that T_{LM} is Bayesian in the unrestricted model $\Theta = \mathbb{R}$ for a certain prior distribution $N(0, \tau^2)$, and find the τ^2.
Exercise H5.2. Consider the binary channel for information transmission, as a statistical model**:** $\Theta = \{0, 1\}$, $P_\vartheta = B(1, p_\vartheta)$ where $p_\vartheta \in (0, 1)$, $\vartheta = 0, 1$. Assume also symmetry: $p_1 = 1 - p_0$ and $1/2 < p_1$. Note that for estimating ϑ, the quadratic loss coincides with the 0-1-loss: for $t, \vartheta \in \Theta$ we have

$$(t - \vartheta)^2 = \begin{cases} 0 \text{ if } t = \vartheta \\ 1 \text{ if } t \neq \vartheta. \end{cases}$$

Consider estimators with values in Θ (note there exist only four possible estimators here (maps $\{0, 1\} \mapsto \{0, 1\}$: $T_1(x) = x$, $T_2(x) = 1 - x$, $T_3(x) = 0$, $T_4(x) = 1$).
(i) Find the maximum likelihood estimator of $\vartheta \in \Theta$
(ii) For a prior distribution Q on Θ with $q_0 = Q(\{0\})$, $q_1 = Q(\{1\})$, and the above 0-1-loss, find the Bayes estimator with values in Θ. Note: the posterior expectation cannot be used since it will be between 0 and 1 in general.
(iii) Assume that q_1 tends to 1. Find the value such that if $q_1 > z$ then the Bayes estimator is T_4 (disregards the data and takes always 1 as estimated value).
Exercise H5.3. Consider the binary Gaussian channel: $\Theta = \{0, 1\}$, $P_\vartheta = N(\mu_\vartheta, 1)$ where $\mu_0 < \mu_1$ are some fixed values (a restricted Gaussian location model for sample size $n = 1$). Note that estimators with values Θ are described by indicators of sets $A \subset \mathbb{R}$ such that $T(x) = \mathbf{1}_A(x)$ (i.e. $A = \{x : T(x) = 1\}$).
(i) as in H5.2.
(ii) as in H5.2
(iii) Show that the Bayes estimator for a uniform prior ($q_1 = 1/2$) is minimax.

13.6 Problem set H6

Exercise H6.1. Let $X_1, \ldots, X_{n_1}$, be independent $N(\mu_1, \sigma^2)$ and $Y_1, \ldots, Y_{n_2}$ be independent $N(\mu_2, \sigma^2)$, also independent of $X_1, \ldots, X_{n_1}$ (n_1,$n_2 > 1$) For each of the two samples, form the sample means $\bar{X}$, $\bar{Y}$ and the bias corrected sample variances

$$S^2_{(1)} = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2, \; S^2_{(2)} = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$$

Consider the statistic

$$Z = \frac{\bar{X} - \bar{Y}}{\sigma \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{1/2}}$$

which is standard normal if $\mu_1 = \mu_2$. In a model where μ_1, μ_2 are unknown but σ^2 is known, it obviously can be used to build a confidence interval for the difference $\mu_1 - \mu_2$.

For the case that in addition σ^2 is unknown, find a statistic which has a t-distribution if $\mu_1 = \mu_2$ (this would then be called a "studentized" statistic) , and find the degrees of freedom.

13.7 Problem set H7

Exercise H7.1. Let $X_1, \ldots, X_{n_1}$, be independent $N(\mu_1, \sigma^2)$ and $Y_1, \ldots, Y_{n_2}$ be independent $N(\mu_2, \sigma^2)$, also independent of $X_1, \ldots, X_{n_1}$ (n_1, $n_2 > 1$). In a model where these r.v.'s are observed, and μ_1, μ_2 and σ^2 are all unknown, find an α-test for the hypothesis

$$H : \mu_1 = \mu_{2.}$$

Exercise H7.2. Let $z_{\alpha/2,n}$ be the upper $\alpha/2$-quantile of the t-distribution with n degrees of freedom and $z^*_{\alpha/2}$ the respective quantile for the standard normal distribution. Use the tables at the end of the textbook or a computer program to find
a) $z_{\alpha/2,n}$ for $n = 5$, $n = 20$ and $z^*_{\alpha/2}$ for a value $\alpha = 0.05$
b) the same for $\alpha = 0.01$.

Exercise H7.3. In the introductory subsection 1.3.2 "Confidence statements with the Chebyshev inequality" (early in the handout, p. 10) we constructed a confidence interval for the parameter p for n i.i.d. Bernoulli observations $X_1, \ldots, X_n$ (model $\mathbf{M}_{d,1}$) of form $[\bar{X}_n - \varepsilon, \bar{X}_n + \varepsilon]$ where $\bar{X}_n$ is the sample mean and $\varepsilon = \varepsilon_{n,\alpha} = 2(n\alpha)^{-1/2}$. Using the Chebyshev inequality, it was shown that this has coverage probability at least $1 - \alpha$.
(i) Use the central limit theorem

$$n^{1/2}(\bar{X}_n - p) \Longrightarrow_d N(0, p(1-p)) \text{ as } n \to \infty$$

and the upper bound $p(1-p) \leq 1/4$ to construct an asymptotic $1-\alpha-$ confidence interval of form $[\bar{X}_n - \varepsilon^*_{n,\alpha}, \bar{X}_n + \varepsilon^*_{n,\alpha}]$ for p, involving a quantile of the standard normal distribution $N(0,1)$.
(ii) Show that the ratio of the lengths of the two confidence intervals, i.e. $\varepsilon^*_{n,\alpha}/\varepsilon_{n,\alpha}$, does not depend on n and find its numerical values for $\alpha = 0.05$ and $\alpha = 0.01$, using the table of $N(0,1)$ on p. 608 textbook.
(iii) Use the property of the standard normal distribution function $\Phi(x)$

$$1 - \Phi(x) \leq x^{-1} \exp\left(-x^2/2\right)$$

([D] p. 108) to prove that $\varepsilon^*_{n,\alpha}/\varepsilon_{n,\alpha} \to 0$ for $\alpha \to 0$ (n fixed).
Comment: the interval based on the normal approximation turns out to be shorter, and this effect becomes more pronounced for smaller α.

Exercise H7.4. Consider the test $\phi^*_{\mu_0}$ (i.e. the test based on the t-statistic as in (8.1) handout, where the quantile $z^*_{\alpha/2}$ of the standard normal is used in place of $z_{\alpha/2}$). On p. 94 handout it is argued that this is an asymptotic α-test. Show that in the Gaussian location-scale model $\mathbf{M}_{c,2}$, for sample size n, this test is consistent as $n \to \infty$ on the pertaining alternative

$$\Theta_1(\mu_0) = \left\{(\mu, \sigma^2) : \mu \neq \mu_0\right\}.$$

Comment: The proof of consistency of the two-sided Gauss test, which is illustrated in the figure p. 95, is similar but more direct since σ^2 is known there.

13.8 Problem set E1

Exercise E1.1. (10%). Let $X_1, \ldots, X_n$ be independent identically distributed with unknown distribution Q, where it is known only that

$$\mathrm{Var}(X_1) \leq K$$

for some known positive K. Then also $\mu = EX_1$ exists. Consider hypotheses
$H : \mu = \mu_0$
$K : \mu \neq \mu_0$.
Find an α-test (exact α-test; i.e. level is observed for every n, not just asymptotic as $n \to \infty$). (**Hint:** Chebyshev inequality, p. 10 or [D], p. 222).

Exercise E1.2. Let $X_1, \ldots, X_n$, be independent Poisson Po(λ). Consider some λ_0, λ_1 such that $0 < \lambda_0 < \lambda_1$.

(i) (5%) Consider simple hypotheses
$H : \lambda = \lambda_0$
$K : \lambda = \lambda_1$.
Find a most powerful α-test.

Note: the distribution of any proposed test statistic can be expected to be discrete, so that a *randomized* test might be most powerful. For the solution, this aspect can be ignored; just indicate the statistic, its distribution under H and the type of rejection region (such as "reject when T is too large").

(ii) (10%) Consider composite hypotheses
$H : \lambda = \lambda_0$
$K : \lambda > \lambda_0$.
Find a uniformly most powerful (UMP) α-test.

Hint: take a solution of (i) which does not depend on λ_1.

(iii) (10%) Consider composite hypotheses
$H : \lambda \leq \lambda_0$
$K : \lambda > \lambda_0$.
Find a uniformly most powerful (UMP) α-test.

Hint: take a solution of (ii) and show that it preserves level α on $H : \lambda \leq \lambda_0$. Properties of the Poisson distribution are useful.

Exercise E1.3. (20%**)** Consider the Gaussian location-scale model (Model $\mathbf{M}_{c,2}$), for sample size n, i. e. observations are i.i.d. $X_1, \ldots, X_n$ with distribution $N(\mu, \sigma^2)$ where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown. For a certain $\sigma_0^2 > 0$, consider hypotheses $H : \sigma^2 \leq \sigma_0^2$ vs. $K : \sigma^2 > \sigma_0^2$.

Find an α-test with rejection region of form (c, ∞) (i.e. a one-sided test) where c is a quantile of a χ^2-distribution. (Note: it is not asked to find the LR test; but the test should have level α. This includes an argument that the level is observed on *all* parameters in the hypothesis $H : \sigma^2 \leq \sigma_0^2$.)

Hint: A good estimator of σ^2 might be a starting point.

.

Exercise E1.4 (Two sample problem, F-test for variances). Let $X_1, \ldots, X_n$ be independent $N(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_n$ be independent $N(\mu_2, \sigma_2^2)$, also independent of $X_1, \ldots, X_n$ ($n > 1$) where

μ_1, σ_1^2 and μ_2, σ_2^2 are all unknown. Define the statistics

$$F = F(X,Y) = \frac{S_X^2}{S_Y^2}, \tag{13.1}$$
$$S_X^2 = n^{-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2,\ S_Y^2 = n^{-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}_n\right)^2$$

(here (X,Y) symbolizes the total sample).
Define the **F-distribution with k_1, k_2 degrees of freedom** (denoted F_{k_1,k_2}) as the distribution of $k_1^{-1}Z_1/k_2^{-2}Z_2$ where Z_i are independent r.v.'s having χ^2-distributions of k_1 and k_2 degrees of freedom, respectively.
i) (15%) Show that $F(X,Y)$ has an F-distribution if $\sigma_1^2 = \sigma_2^2$, and find the degrees of freedom.
ii) (20%) For hypotheses $H : \sigma_1^2 \leq \sigma_2^2$ vs. $K : \sigma_1^2 > \sigma_2^2$, find an α-test with rejection region of form (c, ∞) (i.e. a one-sided test) where c is a quantile of an F-distribution. (Note: it is not asked to find the LR test; but the test should have level α. This includes an argument that the level is observed on *all* parameters in the hypothesis $H : \sigma_1^2 \leq \sigma_2^2$).

Exercise E1.5 (F-test for equality of variances). Consider the two sample problem of exercise E1.4, but hypotheses $H : \sigma_1^2 = \sigma_2^2$ vs. $K : \sigma_1^2 \neq \sigma_2^2$.
i) (5%) Find the likelihood ratio test and show that it is equivalent to a test which rejects if the F-statistic (13.1) is outside a certain interval of form $[c^{-1}, c]$.
ii) (5%) Show that the c of i) can be chosen as the upper $\alpha/2$ quantile of the distribution $F_{r,r}$ for a a certain $r > 0$.

13.9 Problem set H8

Exercise H8.1 *(Exercise 8.59 e, p. 399 textbook).* A famous medical experiment was conducted by Joseph Lister in the late 1800s. Mortality associated with surgery was quite high and Lister conjectured that the use of a disinfectant, carbolic acid, would help. Over a period of several years Lister performed 75 amputations with an without using carbolic acid. The data are

		Carbolic acid	*used ?*
		Yes	No
Patient	Yes	34	19
lived ?	No	6	16

Use these data to test whether the use of carbolic acid is associated with patient mortality.
Exercise H8.2. Let $X_1, \ldots, X_n$ be independent $N(\mu_1, 1)$ and $Y_1, \ldots, Y_n$ be independent $N(\mu_2, 1)$, also independent of $X_1, \ldots, X_n$ and consider hypotheses
$H : \mu_1{=}\mu_2 = 0$
$K : (\mu_1, \mu_2) \neq (0,0)$.
Find the likelihood ratio test and show that it is equivalent to a test based on a statistic which has a certain χ^2-distribution under H (thus the critical value can be taken as a quantile of this χ^2-distribution).
Exercise H8.3. Let $X_1, \ldots, X_n$ be independent Poisson $\mathrm{Po}(\lambda_1)$ and $Y_1, \ldots, Y_n$ be independent $\mathrm{Po}(\lambda_2)$, also independent of $X_1, \ldots, X_n$. Let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ be the parameter vector, $\boldsymbol{\lambda}_0$ a particular value for this vector (with positive components) and consider hypotheses

$H : \boldsymbol{\lambda} = \boldsymbol{\lambda}_0$
$K : .\boldsymbol{\lambda} \neq \boldsymbol{\lambda}_0$.
Find an asymptotic α-test. **Hint:** find a statistic similar to the χ^2-statistic in the multinomial case (Definition 9.1.2, p. 111 handout) and its asymptotic distribution under H.

Exercise H8.4 *(Adapted from exercise 8.60, p. 399 textbook)* Let $\mathbf{Z} = (Z_1, \ldots, Z_k)$ have a multinomial law $\mathfrak{M}_k(n, \mathbf{p})$ with unknown $\mathbf{p} = (p_1, p_2, \ldots, p_k)$, where $k > 2$. Consider hypotheses on the first two components
$H : p_1 = p_2$
$K : p_1 \neq p_2$
A test that if often used, called *McNemar's Test*, rejects H if

$$\frac{(X_1 - X_2)^2}{X_1 + X_2} > \chi^2_{1;1-\alpha} \tag{13.2}$$

where $\chi^2_{1;1-\alpha}$ is the lower $1-\alpha$ quantile of the distribution $\chi^2_{1;1-\alpha}$*.
(i) Find the maximum likelihood estimator $\hat{\mathbf{p}}$ of the parameter $\mathbf{p}$ under the hypothesis.
(ii) Show that the appropriate χ^2-statistic with estimated parameter $\hat{\mathbf{p}}$ (maximum likelihood estimator under H as above), as defined in relation (9.26) on p.127 handout, coincides with McNemar's statistic (13.2) (exact equality, not approximate with an error term).
Comment: It follows that McNemar's test is the χ^2-test for this problem and is an asymptotic α-test, cf. Theorem 9.5.2, p.127 handout.

13.10 Problem set H9

Exercise H9.1. Consider the general linear model

$$\begin{aligned} \mathbf{Y} &= X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \\ E\boldsymbol{\varepsilon} &= \mathbf{0},\ \mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 I_n. \end{aligned}$$

where X is a $n \times k$-matrix, $\mathrm{rank}(X) = k$. Show that $\boldsymbol{\beta}$ provides a best approximation to the data Y in an average sense:

$$E\,\|\mathbf{Y} - X\boldsymbol{\beta}\|^2 = \min_{\boldsymbol{\gamma} \in \mathbb{R}^k} E\,\|\mathbf{Y} - X\boldsymbol{\gamma}\|^2$$

Exercise H9.2. Consider the bivariate linear regression model:

$$Y_i = \alpha + \beta x_i + \varepsilon_i,\ i = 1, \ldots, n$$

where not all x_i are equal, α, β are real valued and ε_i are uncorrelated with variance σ^2.
(i) Show that the LSE of α, β are

$$\hat{\alpha}_n = \bar{Y}_n - \hat{\beta}_n \bar{x}_n,\ \hat{\beta}_n = \frac{\sum_{i=1}^n (Y_i - \bar{Y}_n)(x_i - \bar{x}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}. \tag{13.3}$$

where $\bar{x}_n$ is the mean of the nonrandom x_i.

*The textbook writes an upper α-quantile, called $\chi^2_{1,\alpha}$ there; it coincides the lower quantile $\chi^2_{1;1-\alpha}$.

(ii) Find the distribution of $\hat{\beta}_n$ when ε_i are independent normal: $\mathcal{L}(\varepsilon_i) = N(0, \sigma^2)$.
Hint: for (ii), a possibility is to find the projection matrix $\Pi_{\mathbf{1}}$ projecting onto the space Lin($\mathbf{1}$), where $\mathbf{1}$ is the n-vector consisting of 1's, and use

$$(Y_1 - \bar{Y}_n, \ldots, Y_n - \bar{Y}_n)^\top = (I_n - \Pi_{\mathbf{1}})(Y_1, \ldots, Y_n)^\top.$$

More elementary arguments are also possible.
Exercise H9.3. Consider the general linear model as in H9.1, but with an assumption $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \Sigma$ where Σ is a known positive definite (symmetric) $n \times n$-matrix. Find the BLUE (best linear unbiased estimator) of β, as defined in Def. 10.5.1 .
Hint: Recall Lemma 6.1.7, p. 70 and the fact that $\mathrm{Cov}(A\boldsymbol{\varepsilon}) = A\Sigma A^\top$ for any $n \times n$-matrix A.
Exercise H9.4. Suppose that a r.v. Y and the random k-vector $\mathbf{X}$ have a joint normal distribution $N_{k+1}(\mathbf{0}, \Sigma)$, where Σ is a positive definite $(k+1) \times (k+1)$-matrix. Write Σ in partitioned form

$$\Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \sigma_Y^2 \end{pmatrix}$$

where

$$\Sigma_{XY} = \Sigma_{YX}^\top = E\mathbf{X}Y$$

is a k-vector, $\sigma_Y^2 = EY^2$ and $\Sigma_{XX} = \mathrm{Cov}(\mathbf{X})$. Show that

$$\begin{aligned} E(Y|\mathbf{X}) &= \mathbf{X}^\top \boldsymbol{\beta}, \\ \text{where } \boldsymbol{\beta} &= \Sigma_{XX}^{-1} \Sigma_{XY}. \end{aligned} \tag{13.4}$$

Comment. *Compare this with the form of β in the case $k = 1$ (i.e $\beta = \sigma_X^{-2}\sigma_{XY}$ as in Def. 10.1.3, p. 135 handout), and with the form of the LSE in the linear model (Theorem 10.42, p. 149): $\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{Y}$.*
Hint. Consider independent $\mathbf{X}^*$, ε where $\mathbf{X}^*$ has the same law as $\mathbf{X}$ and $\mathcal{L}(\varepsilon) = N(0, \sigma_\varepsilon^2)$, for some $\sigma_\varepsilon^2 > 0$ ($\mathbf{X}^*$ is a random k-vector, ε a random variable) and define

$$Y^* = \mathbf{X}^{*\top} \boldsymbol{\beta} + \varepsilon.$$

with $\boldsymbol{\beta}$ as above. Find a value of the variance σ_ε^2 such that ($\mathbf{X}^*$,Y^*) have the same joint distribution as ($\mathbf{X}$,Y). This solves the problem, since then

$$E(Y|\mathbf{X}) = E(Y^*|\mathbf{X}^*).$$

13.11 Problem set H10

A PRACTICE EXAM (2 1/2 hours time)

Exercise H10.1. *(15 %)* A company registered 100 cases within a year where some employee was missing exactly one day at work. These were distributed among the days of the week as follows:

Day	M	T	W	Th	F
No.	22	19	16	18	25

Test the hypothesis that these one day absences are uniformly distributed among the days of the week, at level $\alpha = 0.05$.

(Solution consists of: value of the test statistic, critical value for the test [quantile of the pertaining distribution], resulting decision.)

Exercise H10.2 *(15 %)* The personnel manager of a bank wants to find out whether the chance to successfully pass a job interview depends on the sex of the applicant. For 35 randomly selected applicants, 21 of which were male, the results of the interview were evaluated. It turned out that exactly 16 applicants passed the interview, 5 of which were female. Use a χ^2-test in a contingency table to test whether interview result and sex are independent, at level $\alpha = 0.05$.

(Solution consists of: value of the test statistic, critical value for the test [quantile of the pertaining distribution], resulting decision.)

Comment*: most sources would recommend Fisher's exact test here, but this was not treated and the χ^2-test is also applicable.*

Exercise H10.3. *(20 %)* Suppose 18 wheat fields have been divided into $m = 3$ groups, with $l_1 = 5$, $l_2 = 7$ and $l_3 = 6$ members. There are three kinds of fertilizer Φ_j, and the group j of fields is treated with fertilizer Φ_j, $j = 1, 2, 3$. The yield results for all fields are given in the following table.

Φ_1	781	655	611	789	596		
Φ_2	545	786	976	663	790	568	720
Φ_3	696	660	639	467	650	380	

Assume that these values are realizations of independent random variables Y_{jk} with distribution $N(\mu_j, \sigma^2)$, $k = 1, \ldots, l_j$, $j = 1, 2, 3$, where the mean values μ_j correspond to fertilizer Φ_j. Test the hypothesis that the three group means coincide, at level $\alpha = 0.05$.

(Solution consists of: value of the test statistic, critical value for the test [quantile of the pertaining distribution], resulting decision.)

Exercise H10.4. *(25 %)* Consider a normal linear model of type NLM_2 (cf. section 10.5, p. 152 handout), for dimension $k = 1$, i.e. observations are

$$\mathbf{Y} = X\beta + \boldsymbol{\varepsilon},$$

where X is a nonrandom $n \times 1$-matrix (i.e. an n-vector in this case), β is an unknown real valued parameter and

$$\mathcal{L}(\boldsymbol{\varepsilon}) = N_n(\mathbf{0}, \sigma^2 I_n)$$

where $\sigma^2 > 0$ is unknown. Assume also that $\mathrm{rank}(X) = 1$ (identifiability condition; i. e. $X \neq \mathbf{0}$). Consider some value β_0 and hypotheses

$H : \beta \leq \beta_0$

$K : \beta > \beta_0$.

Let $\hat{\beta}$ be the LSE of β and define the statistic

$$T_{\beta_0}(\mathbf{Y}) = \sqrt{n-1}\frac{(\hat{\beta} - \beta_0)(X^\top X)^{1/2}}{(\mathbf{Y}^\top (I_n - \Pi_X)\mathbf{Y})^{1/2}}. \tag{13.5}$$

where Π_X is the projection matrix onto the linear space $\mathrm{Lin}(X)$. Show that $T_{\beta_0}(\mathbf{Y})$ can be used as a test statistic to construct an α-test, and indicate the distribution and the quantile used to find the rejection region.

Comment: *Note that this is not a linear hypothesis on β.*

Hint: *in the case that all elements of X are 1 we obtain the Gaussian location-scale model (mean value β), for which the present hypothesis testing problem was discussed extensively.*

Exercise H10.5. *(25 %)* Suppose that the random vector $Z = (X, Y)$ has a bivariate normal distribution with $EX = EY = 0$ and covariance matrix

$$\mathrm{Var}(X) = \sigma_X^2,\ \mathrm{Var}(Y) = \sigma_Y^2,\ \mathrm{Cov}(X, Y) = \rho\sigma_X\sigma_Y,$$

where $\rho = EXY/\sigma_X\sigma_Y$ is the correlation coefficient between X and Y. Suppose that $Z_i = (X_i, Y_i)$, $i = 1, \ldots, n$ are i.i.d. observed random 2-vectors each having the distribution of Z. Define the empirical covariance matrix

$$\begin{aligned} S_X^2 &= n^{-1}\sum_{i=1}^n X_i^2,\ S_Y^2 = n^{-1}\sum_{i=1}^n Y_i^2, \\ S_{XY} &= n^{-1}\sum_{i=1}^n X_iY_i \end{aligned}$$

(note that we do not use centered data $X_i - \bar{X}_n$ etc. here for empirical variances / covariances since $EX = EY = 0$ is known) and the empirical correlation coefficient

$$\hat{\rho} = \frac{S_{XY}}{S_X S_Y}$$

(where $S_X = \sqrt{S_X^2}$ etc.).
Consider hypotheses
$H : \rho = 0$
$K : \rho \neq 0.$
Define the statistic

$$T_0(\mathbf{Z}) = \sqrt{n-1}\frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}} \tag{13.6}$$

(where $\mathbf{Z}$ represents all the data Z_i, $i = 1, \ldots, n$). Show that $T_0(\mathbf{Z})$ can be used as a test statistic to construct an α-test, and indicate the distribution and the quantile used to find the rejection region.
Hint: *Note that this is closely related to the previous exercise H10.4. In H*10.4, *set* $\beta_0 = 0$, *consider the two sided problem* $H : \beta = \beta_0$ vs. $K : \beta \neq \beta_0$ *(then H is a linear hypothesis) and look for similarities of the statistics (13.5) and (13.6). The distribution of (13.5) was found in H10.4, but now the* X_i *are random. How does this affect the distribution of the test statistic ?*
Further comment: *when it is not assumed that* $EX = EY = 0$, *the definition of* $\hat{\rho}$ *has to be modified in an obvious way, by using centered data* $X_i - \bar{X}_n$, $Y_i - \bar{Y}_n$, *and* $\sqrt{n-2}$ *appears in place of* $\sqrt{n-1}$. *In this form the test is found in the literature.*

13.12 Problem set E2

ANOTHER PRACTICE EXAM (2 1/2 hours time)

Exercise E2.1. *(25 %)* A course in economics was taught to two groups of students, one in a classrooom situation and the other on TV. There were 24 students in each group. These students were first paired according to cumulative grade-point averages and background in economics, and then assigned to the courses by a flip of a coin (this was repeated 24 times). At the end of the course

each class was given the same final examination. Use the Wilcoxon signed rank test (level $\alpha = 0.05$, normal approximation to the test statistic) to test that the two methods of teaching are equally effective against a two-sided alternative. The differences in final scores for each pair of students, the TV student's score having been subtracted from the corresponding classroom student's score were as follows:

14	−4	−6	−2	−1	18
6	12	8	−4	13	7
2	6	21	7	−2	11
−3	−14	−2	17	−4	−5

Hint: treatment of ties. Let $Z_1, \ldots, Z_n$ be the data. If some $|Z_i|$ have the same absolute values (i.e. ties occur) then they are assigned values $\tilde{R}_i$ which are the averages of the ranks. *Example:* 4 values $|Z_{i_1}|, \ldots, |Z_{i_4}|$ have the same absolute value c and only two of the other $|Z_i|$ are smaller. The $|Z_{i_1}|, \ldots, |Z_{i_4}|$ would then occupy ranks $3, 4, 5, 6$. Since they are tied, they are all assigned the average rank $(3+4+5+6)/4 = 4.5$. The next rank assigned is then 7 (or higher if there is another tie).

Remark: For the two-sided version of the Wilcoxon signed rank test, as described in section 12.2 handout, the last sentence on p. 175 should read as *"The corresponding asymptotic α-test of H then rejects if $|W_n| > z_{\alpha/2} v_n^{1/2}$ where $z_{\alpha/2}$ is the upper $\alpha/2$-quantile of $N(0,1)$)"*.

As a starting point, to limit the bookkeeping effort in the solution, the following table gives the ordered absolute values of the data $|Z_i|$, starring those that were originally negative. Below each entry (every other row of the table) is the rank $\tilde{R}_i$ of $|Z_i|$ (in the notation of the handout), where ties are treated as indicated:

1*	2*	2*	2*	2	3*
1	*3.5*	*3.5*	*3.5*	*3.5*	*6*
4*	4*	4*	5*	6*	6
8	*8*	*8*	*10*	*12*	*12*
6	7	7	8	11	12
12	*14.5*	*14.5*	*16*	*17*	*18*
13	14*	14	17	18	21
19	*20.5*	*20.5*	*22*	*23*	*24*

Solution.

Exercise E2.2. Consider the one-way layout ANOVA treated in handout sec. 11.2, relation (11.7):

$$Y_{jk} = \mu_j + \varepsilon_{jk},\ k = 1, \ldots, l,\ j = 1, \ldots, m, \tag{13.7}$$

where ε_{jk} are i.i.d. normal $N(0, \sigma^2)$ noise variables and μ_j are unknown parameters, and there is an equal number of observation $l > 1$ for each factor j. The total number of observations is $n = ml$. In this case the F test given by the statistic (11.13) (handout) can be considered an average t test.

(i) (25 %) Let i and j be two different factor indices from $\{1, \ldots, m\}$. Show that a t test of
$H : \mu_i = \mu_j$
$K : \mu_i \neq \mu_j$
can be based on the statistic

$$T_{ij}(\mathbf{Y}) = \frac{\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}}{\left(2\hat{d}_w/l\right)^{1/2}}$$

where

$$\hat{d}_w = (n-m)^{-1} \sum_{j=1}^{m} \sum_{k=1}^{l} \left(Y_{jk} - \bar{Y}_{j\cdot}\right)^2$$

is the mean sum of squares within groups. (More precisely, show that T_{ij} has a t-distribution under H and find the degrees of freedom).

(ii) (25 %) Show that

$$\frac{1}{m(m-1)} \sum_{j=1}^{m} \sum_{i=1}^{m} T_{ij}^2(\mathbf{Y}) = F(\mathbf{Y})$$

where $F(\mathbf{Y})$ is the F statistic (11.13) (handout). This relation shows that $F(\mathbf{Y})$ is an average of the (nonzero) $T_{ij}^2(\mathbf{Y})$.

Exercise E2.3. (25 %) Consider again the model (13.7) of exercise F2, with the same assumptions, but in an asymptotic framework where the number l of observation in each group tends to infinity (m stays fixed). Consider the F-statistic

$$F(\mathbf{Y}) = \frac{(m-1)^{-1} \sum_{j=1}^{m} l \left(\bar{Y}_{j\cdot} - \bar{Y}_{\cdot\cdot}\right)^2}{(n-m)^{-1} \sum_{j=1}^{m} \sum_{k=1}^{l} \left(Y_{jk} - \bar{Y}_{j\cdot}\right)^2}$$

Find the limiting distribution, as $l \to \infty$, of the statistic $(m-1)F(\mathbf{Y})$ under the hypothesis of equality of means $H : \mu_1 = \ldots = \mu_m$. (It follows that this distribution can be used to obtain an asymptotic α-test of H).

Hint: Limiting distributions of other test statistics in the handout have been obtained e.g. in Theorem 7.3.8 and Theorem 9.3.3.

Chapter 14
APPENDIX: TOOLS FROM PROBABILITY, REAL ANALYSIS AND LINEAR ALGEBRA

14.1 The Cauchy-Schwartz inequality

Proposition 14.1.1 ***(Cauchy-Schwartz-inequality)***. *Suppose that for the random variables Y_i, $i = 1, 2$, the second moments EY_i^2 exist, for $i = 1, 2$. Then the expectation of $Y_1 \cdot Y_2$ exists, and*

$$|EY_1Y_2| \leq \left(EY_1^2\right)^{1/2} \left(EY_2^2\right)^{1/2}.$$

Proof. For any $\lambda > 0$

$$0 \leq \left(\lambda^{1/2}Y_1 \pm \lambda^{-1/2}Y_2\right)^2 = \lambda Y_1^2 \pm 2Y_1Y_1 + \lambda^{-1}Y_2^2,$$

thus

$$\mp Y_1Y_2 \leq \frac{1}{2}\left(\lambda Y_1^2 + \lambda^{-1}Y_2^2\right).$$

This proves that EY_1Y_2 exists and

$$|EY_1Y_2| \leq \frac{1}{2}\left(\lambda EY_1^2 + \lambda^{-1}EY_2^2\right).$$

If both EY_2^2, $EY_1^2 > 0$ then for $\lambda = \left(EY_2^2\right)^{1/2} / \left(EY_1^2\right)^{1/2}$ we obtain the assertion. If one of them is 0 ($EY_2^2 = 0$, say), then by taking $\lambda > 0$ arbitrarily small, we obtain $|EY_1Y_2| = 0$. ■

14.2 The Lebesgue Dominated Convergence Theorem

This is a result from real analysis for measure theoretic integrals, which contain both sums and integrals as a special cases. We formulate here a special case relating to expectation of random variables. For a statement and proof in full generality see Durrett [D], Appendix, (5.6), p. 468.

Theorem 14.2.1 *(Lebesgue) Let X be a random variable taking values in a sample space $\mathcal{X}$ and let $r_n(x)$, $n = 1, 2, \ldots$ be a sequence of functions on $\mathcal{X}$ such that $r_n(x) \to r_0(x)$ for all $x \in \mathcal{X}$. Assume furthermore that there exists a function $\tilde{r}(x) \geq 0$ such that $|r_n(x)| \leq \tilde{r}(x)$ for all $x \in \mathcal{X}$ and all n (domination property), and $E\,\tilde{r}(X) < \infty$. Then*

$$E\,r_n(X) \to E\,r_0(X), \text{ as } n \to \infty.$$