Lecture Notes
Mathematical Statistics

Winter semester 2020/21

Friedrich-Schiller-Universität Jena

Michael H. Neumann

Important information

- Lecture period: November 02 – February 12
  Prior to a possible start of classroom lectures you will find your weekly program as a corresponding part of lecture notes. You will find the material under the following link: `https://users.fmi.uni-jena.de/~jschum/lehre/lectures.php?name=Neumann`
- Examination period: February 15 – March 12
  There will be oral examinations, hopefully within the official examination period. Dates of these examinations will be fixed in good time. Possible retakes will preferably take place before the lecture period of the Summer semester.
- Contact
  - Office: room 3516
  - Phone: 03641/946260
  - Email: michael.neumann@uni-jena.de

Literature

[1] Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics*. Holden-Day. San Francisco.

[2] Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*. 2nd Edition. Wiley, New York.

[3] Montgomery, D. C. and Peck, E. A. (1992). *Introduction to Linear Regression Analysis*. Wiley, New York.

[4] Shao, J. (2003). *Mathematical Statistics*. 2nd edition. Springer, New York.

[5] Weisberg, S. (2005). *Applied Linear Regression*. 3rd Edition. Wiley, Hoboken.

Contents

1 Linear regression **4**
1.1 Introduction and data examples 4
1.2 Least squares estimation in a linear model 10
1.3 Choice of a good model: hypothesis testing 14

2 Statistical estimation of parameters **21**
2.1 A model for a statistical experiment 21
2.2 Some methods of estimation 23
2.3 Consistency of estimators 28
2.4 Comparison of estimators – optimality theory 36
2.5 The information inequality 39
2.6 Bayes and minimax estimators 52

3 Testing statistical hypotheses **62**
3.1 The elements of hypothesis testing 62
3.2 Optimal tests 65
3.3 Likelihood ratio tests 80

1 Linear regression

1.1 Introduction and data examples

A common problem in statistics is that of estimating the relationship that exists (if any) between two random variables X and Y; for instance, height and weight, income and intelligence quotient (IQ), ages of husband and wife at marriage, length and breadth of leaves, temperature and pressure of a certain volume of gas, or the length of a metal rod and its temperature. Regression analysis is a statistical technique for investigating, modeling and representing the relationship between variables. But why do we use the word "regression"? It appears that the British anthropologist and meteorologist Sir Francis Galton (1822-1911) was responsible for the introduction of the word "regression". Originally he used the term "reversion" in an unpublished address "Typical laws of heredity in man" to the Royal Institution on February 9, 1877. The later term "regression" appears in his Presidential address made before Section H of the British Association in Aberdeen, 1885, printed in *Nature*, September 1885, 507-510, and also in a paper "Regression Towards Mediocrity in Hereditary Stature," *The Journal of the Anthropological Institute of Great Britain and Ireland*, **15**, 1886, 246-263. In the latter, Galton reports on his initial discovery that the offspring of seeds are taller than the mean if the parents are taller than the mean and vice versa; however, the size of the offsprings are usually less extreme than the size of the parents. In other words, extreme characteristics are not completely passed on to the next generation. Galton also reports that the same effect was observed in the records of adult children and their parents. Today the term "regression" is used for statistical methods which are designed to detect and quantify dependence between (usually random) aspects. Here is a reprint from Galton's paper (Figure 1) which shows the original data (heights of females are adjusted by a factor of 1.08 and heights of "mid-parents" are computed accordingly).

TABLE I.

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.
(All Female heights have been multiplied by 1·08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.													Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

Figure 1: Original data from Galton's paper

The following "scatter plot" (Figure 2) shows pairs of measurements (height of adult child vs. height of mid-parent).

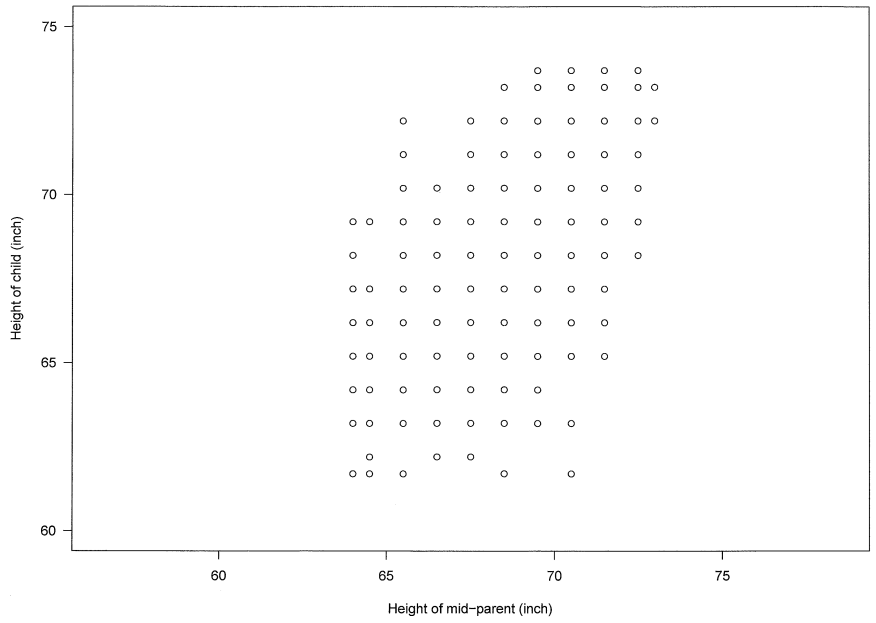

Figure 2: Scatter plot of the data from Galton's paper

Only a very careful inspection of this plot would reveal two features:

- There is a tendency that parents which are taller than the mean produce children that are also taller than the mean and, vice versa, children from shorter parents are shorter than the mean.
- Extreme characteristics are not completely passed on to children, that is, the heights of children are less extreme than the heights of their parents.

In order to obtain a more clear picture of this relationship, we can try to fit a smooth curve through the points in such a way that the points are as "close" to the curve as possible. Of course, we would not expect an exact fit because both variables in the above example are subject to chance fluctuations owing to factors outside our control. Although the heights of children depend on the parent's heights, factors such as diseases, nutrition etc. influence the growth of children. The relationship between the heights of the parents and their children can be conveniently described by the following statistical model.

$$Y_i = \theta_1 + \theta_2 x_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where

- x_i height of ith "mid-parent" (...),
- Y_i height of ith (adult) child,
- ε_i a random variable ("error"), accounting for uncontrolled influences,
- n sample size.

Here and in the following, the explanatory variables (**"regressors"**) are assumed to be nonrandom while the dependent variables (**"regressands"**) are modeled as random variables. The "errors" are introduced in order to transform the causal relationship into an equality. Such error terms cannot be avoided whenever there are certain factors that influence the cause-effect relationship in an uncontrollable way. Even if there is an exact relationship between variables like temperature and pressure, fluctuations would still show up in a scatter plot because of errors of measurement. For theoretical considerations, it is usually assumed that the errors have zero mean.

Model (1.1) contains unknown parameters, θ_1 and θ_2. The most popular method of approximating (or "estimating") these parameters is the method of least squares.[1] Suppose we have **realizations** $(x_1, y_1), \ldots, (x_n, y_n)$ of the random pairs $(x_1, Y_1), \ldots, (x_n, Y_n)$ at our disposal. Then we obtain an estimate $\widehat{\theta} = \left(\widehat{\theta}_1, \widehat{\theta}_2\right)^T$ of the vector $\theta = (\theta_1, \theta_2)^T$ (The superscript T stands for transposition.) as a solution to

$$\sum_{i=1}^{n} \left(y_i - \widehat{\theta}_1 - \widehat{\theta}_2 x_i\right)^2 = \inf_{(\theta_1,\theta_2)^T \in \mathbb{R}^2} \sum_{i=1}^{n} \left(y_i - \theta_1 - \theta_2 x_i\right)^2. \tag{1.2}$$

It will be shown below that the infimum on the right-hand side of (1.2) is actually attained and that the minimizer $\widehat{\theta}$ is uniquely defined unless all $x_1, \ldots, x_n$ are equal; see Theorem 1.1 below for such results in a general context. It turns out that the estimate has the form $\widehat{\theta} = \sum_{i=1}^{n} w_i y_i$, where $w_1, \ldots, w_n$ are certain weights depending on the regressors $x_1, \ldots, x_n$ alone. The corresponding *random variable* is $\sum_{i=1}^{n} w_i Y_i$, the so-called **least squares estimator** of θ, which we also denote by $\widehat{\theta}$. We will also show below that this estimator has certain optimality properties. The least squares estimates of θ_1 and θ_2 can be represented as

$$\widehat{\theta}_1 = \bar{y}_n - \widehat{\theta}_2 \bar{x}_n$$

and

$$\widehat{\theta}_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)(x_i - \bar{x}_n)}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2},$$

where $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$. The function

$$\widehat{g}(x) = \widehat{\theta}_1 + \widehat{\theta}_2 x$$

is drawn in Figure 3 as a bold solid line. The "1-1 curve", $h(x) = x$, is displayed as a dashed line.

A comparison of these two curves reveals the claimed principles: Adult children of tall parents tend to be taller than the average and, vice versa, children of short parents tend to be shorter than the average. This is also corroborated by $\widehat{\theta}_2 \approx 2/3 > 0$. On the other hand, their deviation in height from the average is less extreme than for their parents; this is in line with $\widehat{\theta}_2 < 1$.

[1] In Draper and Smith [2, p. 11] you will find some historical details: There has been a dispute about who first discovered the methods of least squares. It appears that it was proposed *independently* by Carl Friedrich Gauß (1777-1855) and the French mathematician Adrien-Marie Legendre (1752-1833), that Gauß started using it before 1803 (he claimed in about 1795, but there is no corroboration of this earlier date), and that the first account was published by Legendre in 1805. When Gauß wrote in 1809 that he had used the method earlier than the date of Legendre's publication, controversy concerning the priority began. Today, the term "least squares method" is used as a direct translation from the French "méthode des moindres carrés".

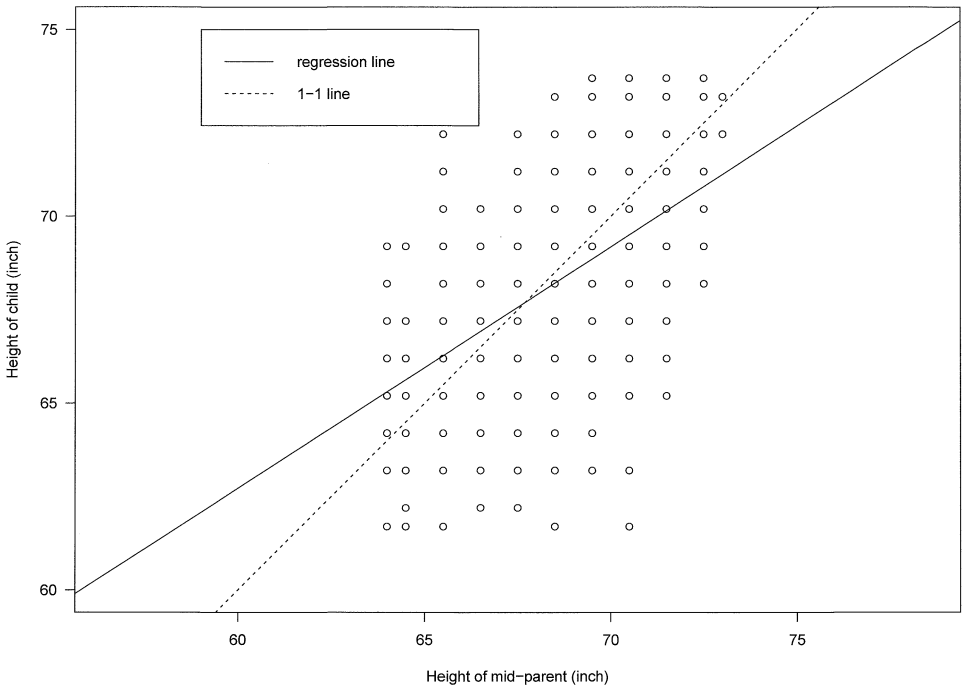

Figure 3: Function fitted by least squares (bold solid line)

Before we systematically investigate the method of least squares we consider one more example. In order to find out what would be safe speed limits to permit on different streets one has to know in what distance an automobile could be stopped when traveling at different speeds. By comparing this distance with the length of view at intersections we could judge how fast cars might be able to travel without risk of collisions at street intersections. One way to determine what is the relation between speed and stopping distance would be to make a number of tests taking different types of machines and different drivers. The following (hypothetical) data can be found in Chapter 3 of the book "Methods of Correlation and Regression Analysis. Linear and curvilinear" by M. Ezekiel and K A. Fox (1959).

RELATION BETWEEN SPEED OF AUTOMOBILE AND DISTANCE REQUIRED FOR STOPPING AFTER SIGNAL, AS SHOWN BY 63 OBSERVATIONS

Speed When Signal is Given	Distance Traveled After Signal Before Stopping	Speed	Distance	Speed	Distance
Miles per hour	*Feet*	*Miles per hour*	*Feet*	*Miles per hour*	*Feet*
5	2	21	39	28	84
10	8	26	39	27	57
10	17	25	33	30	67
10	14	24	56	16	34
8	9	18	29	18	34
16	19	25	59	8	8
17	29	27	78	5	8
12	11	25	48	5	4
9	5	21	42	13	15
7	6	25	56	14	14
7	7	30	60	8	13
9	13	29	68	9	5
4	4	17	22	14	16
5	8	16	14	8	11
13	18	13	27	35	85
15	16	12	21	40	110
18	47	12	19	39	138
19	30	26	41	31	77
20	48	28	64	35	107
21	55	29	54	22	35
36	79	30	101	40	134

A slightly more clear picture is provided by the following scatter plot (Figure 4).

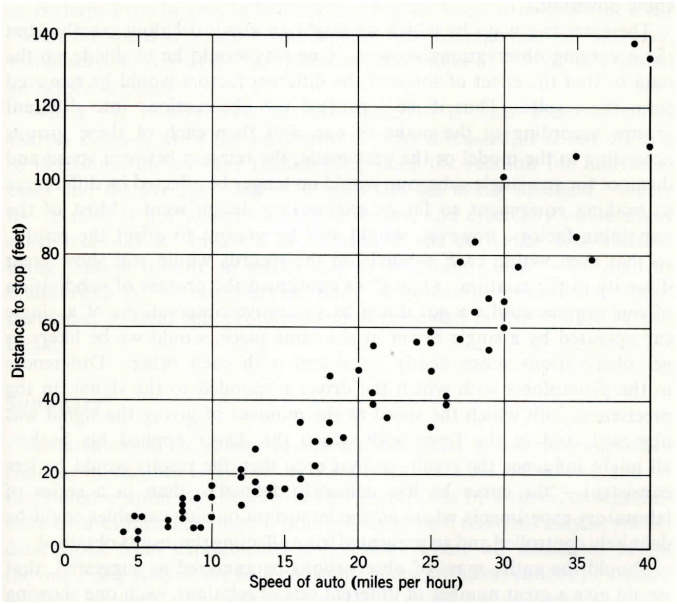

Figure 4: Scatter plot of the automobile data

It is visible with the naked eye that there is a tendency that higher speeds require longer stopping distances. Of course, there are variations in the distances which different cars or different drivers required to stop, even when traveling at the same speed. It should not be surprising that this relation is not more definite since there are many factors outside control that influence the stopping distance, e.g. cars with strong or worn brakes, experienced or inexperienced drivers, drivers with almost instantaneous reaction to signal to stop and others with lagging response and so on. So, we can at best hope to uncover the relation between speed and **average** stopping distance. Figure 5 shows the relation between speed and the respective arithmetic mean of the measured stopping distances.

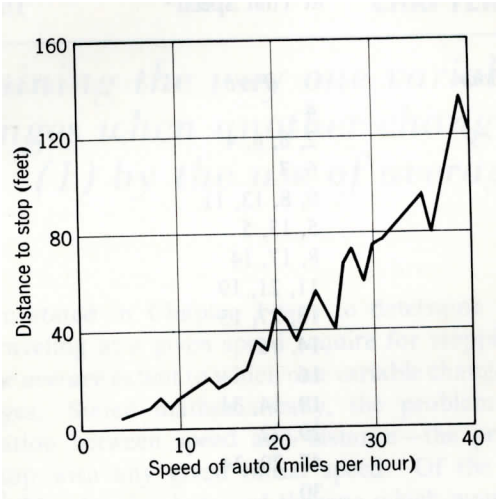

Figure 5: Averaged stopping distances

This curve is again not that we expect since the searched-for function should be at least monotonically increasing. At this point it makes sense to use logical arguments to find an

appropriate model. We denote by $(v_1, y_1), \ldots, (v_n, y_n)$ the pairs of measurements, where v_i is the speed of the ith vehicle and y_i the measured stopping distance. We regard $y_1, \ldots, y_n$ as **realizations of random variables** $Y_1, \ldots, Y_n$, which leads to the following regression model.

$$Y_i \;=\; f(v_i) \;+\; \varepsilon_i, \qquad i = 1, \ldots, n.$$

The random variables $\varepsilon_1, \ldots, \varepsilon_n$ are used to model variations that are outside control. We aim at approximating the unknown function $f\colon\ [0, \infty) \to [0, \infty)$. Since the interpolation of the data point as in the above picture does not lead to the desired result, we include some additional considerations that lead to a reasonable model. The mean stopping distance for a given speed v can be split up into a distance caused by the reaction time of the driver (**thinking distance**) and the distance covered by the car after applying the brake (**braking distance**), that is

$$f(v) \;=\; f_R(v) \;+\; f_B(v) \qquad \forall v \geq 0. \tag{1.3}$$

It is clear that the thinking distance is proportional to the velocity of the car,

$$f_R(v) \;=\; v \cdot \theta_1, \tag{1.4}$$

where $\theta_1 = t_R$ is the average reaction time.

To simplify matters we impose the assumption that the drivers apply a **constant** braking force. Newton's second law of motion tells us that

$$\underbrace{F}_{\text{force}} \;=\; \underbrace{m}_{\text{mass}} \cdot \underbrace{a}_{\text{acceleration}}.$$

Therefore, we have a constant deceleration ("negative acceleration"). Let, w.l.o.g., $t_0 = 0$ be the time when the brake is applied and t_1 be the time when the car comes to a halt. Then the velocity $v(t)$ at time $t \in [0, t_1]$ is given by

$$v(t) \;=\; \underbrace{v(t_0)}_{=v} + \int_{t_0}^{t_1} a\, du \;=\; v \;+\; a\, t.$$

Since $v(t_1) = 0$ we obtain that

$$t_1 \;=\; -v/a.$$

Therefore, the braking distance for an initial velocity v is equal to

$$\begin{aligned} f_B(v) \;&=\; \int_{t_0}^{t_1} v(u)\, du \;=\; \int_0^{-v/a} \big(v + au\big)\, du \\ &=\; \left[\frac{(v+au)^2}{2a}\right]_0^{-v/a} \;=\; -\frac{v^2}{2a}. \end{aligned} \tag{1.5}$$

(Note that $a < 0$ which means that $f_B(v)$ is actually positive.)
We conclude from (1.3) to (1.5) that

$$f(v) \;=\; v \cdot \theta_1 \;+\; v^2 \cdot \theta_2,$$

where $\theta_1 = t_R$ and $\theta_2 = -1/(2a)$ are unknown constants. This leads us to the following model:

$$Y_i \;=\; v_i\theta_1 \;+\; v_i^2\theta_2 \;+\; \varepsilon_i, \qquad i = 1, \ldots, n.$$

(This model is said to be a **linear regression model** since the unknown parameters θ_1 and θ_2 enter linearly.) The solid line in Figure 6 shows the graph of the function $\widehat{f}(v) = v\widehat{\theta}_1 + v^2\widehat{\theta}_2$, where $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are determined by the least squares method.

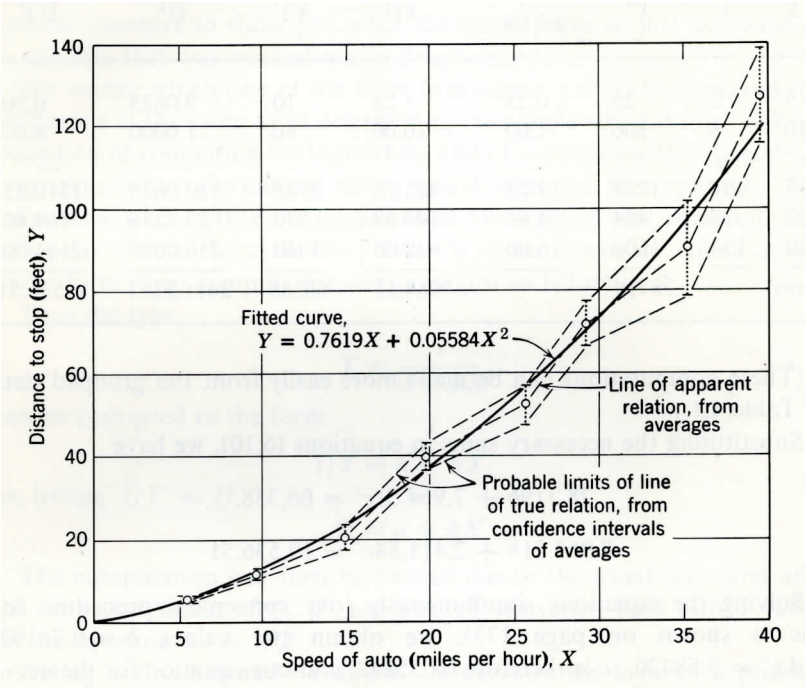

Figure 6: Curve obtained by a least squares fit (bold solid line)

1.2 Least squares estimation in a linear model

In this subsection we consider a general linear regression model,

$$Y_i = x_i^T\theta + \varepsilon_i, \qquad i = 1, \dots, n,$$

$x_i = \big(x_{i1}, \dots, x_{ik}\big)^T$ is the ith vector of explanatory variables (independent variables or regressors), Y_i the ith dependent variable, ε_i the ith error variable, and $\theta = \big(\theta_1, \dots, \theta_k\big)^T$ the vector of unknown coefficients. It will be convenient for a further analysis to rewrite the n model equations in matrix/vector form:

$$\underbrace{\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}}_{=:Y} = \underbrace{\begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}}_{=:X} \theta + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{=:\varepsilon}.$$

Let

$$S(\theta) = \sum_{i=1}^{n} \big(Y_i - x_i^T\theta\big)^2 = \big\|Y - X\theta\big\|^2$$

be the sum of squared deviations, where $\|\cdots\|$ denotes the Euclidean norm in $\mathbb{R}^n$. In what follows we examine the least squares estimator of θ which is defined as the minimizer of the functional $S(\theta)$. For now it is neither clear that a minimizer exists (i.e. that $\inf\big\{\|Y - X\theta\|^2 \colon\ \theta \in \mathbb{R}^k\big\}$ is attained) nor that such a minimizer is unique. The following theorem gives an answer to these questions.

Theorem 1.1. *(i) There exists some $\widehat{\theta} \in \mathbb{R}^k$ such that*

$$\big\|Y - X\widehat{\theta}\big\|^2 = \inf\Big\{\, \|Y - X\theta\|^2 :\ \ \theta \in \mathbb{R}^k\Big\}.$$

(ii) *Any least squares estimator $\widehat{\theta}$ is a solution to the normal equation, i.e.*

$$X^T X\widehat{\theta} = X^T Y.$$

(iii) *If the matrix X has rank k, then $X^T X$ is regular, $\widehat{\theta}$ is uniquely defined, and*

$$\widehat{\theta} = \left(X^T X\right)^{-1} X^T Y.$$

Proof. (i) Let $d := \inf\left\{\|Y - X\theta\|^2 \colon\ \theta \in \mathbb{R}^k\right\} = \inf\left\{\|Y - v\|^2 \colon\ v \in M\right\}$, where $M := \left\{Xb \colon\ b \in \mathbb{R}^k\right\}$ is a linear subspace of $\mathbb{R}^n$. There exists a sequence $(v_n)_{n\in\mathbb{N}}$, $v_n \in M\ \forall n \in \mathbb{N}$, such that

$$\left\|Y - v_n\right\|^2 \underset{n\to\infty}{\longrightarrow} d.$$

Next we show that $(v_n)_{n\in\mathbb{N}}$ is a Cauchy sequence. Indeed, using $\|a+b\|^2+\|a-b\|^2 = 2\|a\|^2 + 2\|b\|^2\ \forall a, b \in \mathbb{R}^n$ we obtain that

$$\begin{aligned}\left\|v_n - v_m\right\|^2 &= \left\|\left(Y - v_m\right) - \left(Y - v_n\right)\right\|^2 \\ &= \underbrace{2\left\|Y - v_m\right\|^2}_{\underset{n\to\infty}{\longrightarrow} 2d} + \underbrace{2\left\|Y - v_n\right\|^2}_{\underset{n\to\infty}{\longrightarrow} 2d} - \underbrace{\left\|\left(Y - v_m\right) + \left(Y - v_n\right)\right\|^2}_{=4\|Y - \frac{v_m+v_n}{2}\|^2 \geq 4d}.\end{aligned}$$

This yields

$$\left\|v_n - v_m\right\|^2 \underset{m,n\to\infty}{\longrightarrow} 0.$$

(This statement means that for all $\epsilon > 0$ there exists some $N_\varepsilon \in \mathbb{N}$ such that $\|v_n - v_m\|^2 \leq \epsilon$ if $m, n \geq N_\epsilon$.) Hence, $(v_n)_{n\in\mathbb{N}}$ is a Cauchy sequence. Since the finite-dimensional linear space M is complete there exists some $v \in M$ such that $v_n \underset{n\to\infty}{\longrightarrow} v$. Since $\|Y - v\| \leq \underbrace{\|Y - v_n\|}_{\underset{n\to\infty}{\longrightarrow}\sqrt{d}} + \underbrace{\|v_n - v\|}_{\underset{n\to\infty}{\longrightarrow} 0}$ we obtain that $\|Y - v\|^2 = d$. We can find some $\widehat{\theta} \in \mathbb{R}^k$ such that $v = X\widehat{\theta}$ and we conclude that

$$\left\|Y - X\widehat{\theta}\right\|^2 = \inf\left\{\ \|Y - X\theta\|^2 :\ \ \theta \in \mathbb{R}^k\right\}.$$

(ii) Suppose that $\widehat{\theta}$ is a least squares estimator which does not satisfy the normal equation, that is $X^T(Y - X\widehat{\theta}) \neq 0_k$. ($0_k$ denotes the vector consisting of k zeroes.) Then there exists some $\alpha \in \mathbb{R}^k$ such that $(Y - X\widehat{\theta})^T X\alpha > 0$. Let $\bar{\alpha} = \varepsilon\alpha$, where $\varepsilon > 0$. Then

$$\begin{aligned}\left\|Y - X(\widehat{\theta} + \bar{\alpha})\right\|^2 &= \left\|Y - X\widehat{\theta}\right\|^2 + \varepsilon^2\alpha^T X^T X\alpha - 2\varepsilon\underbrace{(Y - X\widehat{\theta})^T X\alpha}_{>0} \\ &< \left\|Y - X\widehat{\theta}\right\|^2,\end{aligned}$$

for sufficiently small $\varepsilon > 0$. This is a contradiction to our assumption that $\widehat{\theta}$ is a least squares estimator. Hence, $\widehat{\theta}$ is a solution to the normal equation.

(iii) Suppose that the matrix X has rank k. This means that the columns of X are linearly independent and for all $\alpha \in \mathbb{R}^k$ such that $\alpha \neq 0_k$ we obtain $X\alpha \neq 0_n$. Therefore we have that $\alpha^T X^T X \alpha \neq 0$, which implies $X^T X \alpha \neq 0_k$. Hence, $X^T X$ is regular. Since any least squares estimator satisfies the normal equation we obtain that $\widehat{\theta} = \left(X^T X\right)^{-1} X^T Y$.

□

The following theorem provides a certain optimality property of the least squares estimator. In the class of all linear and unbiased estimators, $\widehat{\theta}$ has the smallest covariance matrix.

Theorem 1.2. *Suppose that*

$$Y \;=\; X\theta \,+\, \varepsilon,$$

where X is an $(n \times k)$-matrix with rank$(X) = k$, $\theta \in \mathbb{R}^k$, $E\varepsilon = 0_n$, and $\mathrm{Cov}(\varepsilon) = \sigma^2 I_n$, *$\sigma^2 > 0$. ($I_n$ denotes the n-dimensional unit matrix.) Then*

(i) *$\widehat{\theta}$ is an unbiased estimator of θ, that is*

$$E_\theta \widehat{\theta} \;=\; \theta \qquad \forall \theta \in \mathbb{R}^k.$$

(The notation E_θ means that the expectation refers to θ as the true parameter.)

(ii) $E_\theta\Big[\big(\widehat{\theta} - \theta\big)\big(\widehat{\theta} - \theta\big)^T\Big] \;=\; \sigma^2\big(X^T X\big)^{-1} \qquad \forall \theta \in \mathbb{R}^k.$

Let $\widetilde{\theta} = LY$ be any linear and unbiased estimator of θ. Then, for all $\theta \in \mathbb{R}^k$,

$$E_\theta\Big[\big(\widetilde{\theta} - \theta\big)\big(\widetilde{\theta} - \theta\big)^T\Big] \;-\; E_\theta\Big[\big(\widehat{\theta} - \theta\big)\big(\widehat{\theta} - \theta\big)^T\Big] \;=\; \sigma^2\Big(LL^T \;-\; \big(X^T X\big)^{-1}\Big)$$

is non-negative definite. Furthermore

$$E_\theta\Big[\big(\widetilde{\theta}_i - \theta_i\big)^2\Big] \;\geq\; E_\theta\Big[\big(\widehat{\theta}_i - \theta_i\big)^2\Big] \qquad \forall i = 1, \ldots, k.$$

As announced, the least squares estimator $\widehat{\theta}$ has the smallest "matrix risk" (which is equal to its covariance matrix since $\widehat{\theta}$ is unbiased) among all linear and unbiased estimators. In this sense, it is the **best linear unbiased estimator**, abbreviated as **BLUE**.

Proof of Theorem 1.2. (i) According to Theorem 1.1(iii), the least squares estimator has the form

$$\widehat{\theta} \;=\; \big(X^T X\big)^{-1} X^T Y.$$

This implies that

$$E_\theta \widehat{\theta} \;=\; E_\theta\Big[\big(X^T X\big)^{-1} X^T (X\theta + \varepsilon)\Big] \;=\; \big(X^T X\big)^{-1} X^T \underbrace{E_\theta\big[X\theta + \varepsilon\big]}_{=X\theta} \;=\; \theta \qquad \forall \theta \in \mathbb{R}^k.$$

(ii) For the least squares estimator $\widehat{\theta}$, we obtain that

$$\begin{aligned} & E_\theta\Big[\big(\widehat{\theta}-\theta\big)\big(\widehat{\theta}-\theta\big)^T\Big] \\ &= E_\theta\Big[\big(X^TX\big)^{-1}X^T\varepsilon\,\varepsilon^TX\big(X^TX\big)^{-1}\Big] \\ &= \big(X^TX\big)^{-1}X^T\underbrace{E_\theta\big[\varepsilon\,\varepsilon^T\big]}_{=\sigma^2I_n}X\big(X^TX\big)^{-1} \\ &= \sigma^2\big(X^TX\big)^{-1}. \end{aligned}$$

Let $\widetilde{\theta}=LY$ be any linear and unbiased estimator of θ. The matrix risk of this estimator is equal to

$$\begin{aligned} E_\theta\Big[\big(\widetilde{\theta}-\theta\big)\big(\widetilde{\theta}-\theta\big)^T\Big] &= E_\theta\big[L\varepsilon\varepsilon^TL^T\big] \\ &= L\underbrace{E_\theta\big[\varepsilon\varepsilon^T\big]}_{=\sigma^2I_n}L^T \;=\; \sigma^2LL^T. \end{aligned}$$

From the unbiasedness of $\widetilde{\theta}$ we conclude

$$E_\theta\big[L\big(X\theta+\varepsilon\big)\big] \;=\; LX\theta \;=\; \theta \qquad \forall\theta\in\mathbb{R}^k,$$

which implies that $LX=I_k$.

Using the fact that any matrix of the form MM^T is non-negative definite we obtain that

$$\begin{aligned} 0_{k\times k} &\preceq \Big(L\;-\;\big(X^TX\big)^{-1}X^T\Big)\Big(L\;-\;\big(X^TX\big)^{-1}X^T\Big)^T \\ &= LL^T\;-\;\underbrace{LX}_{=I_k}\big(X^TX\big)^{-1}\;-\;\big(X^TX\big)^{-1}\underbrace{X^TL^T}_{=I_k}+\big(X^TX\big)^{-1}X^TX\big(X^TX\big)^{-1} \\ &= LL^T\;-\;\big(X^TX\big)^{-1}. \end{aligned}$$

This implies that

$$E_\theta\Big[\big(\widetilde{\theta}-\theta\big)\big(\widetilde{\theta}-\theta\big)^T\Big]\;-\;E_\theta\Big[\big(\widehat{\theta}-\theta\big)\big(\widehat{\theta}-\theta\big)^T\Big] \;=\; \sigma^2\Big(LL^T\;-\;\big(X^TX\big)^{-1}\Big)$$

is actually non-negative definite.

Since the diagonal elements of any non-negative definite matrix are non-negative we obtain, for all $i=1,\ldots,k$,

$$\begin{aligned} E_\theta\Big[\big(\widetilde{\theta}_i-\theta_i\big)^2\Big] &= \sigma^2\big(LL^T\big)_{i,i} \\ &\geq \sigma^2\Big(\big(X^TX\big)^{-1}\Big)_{i,i} \;=\; E_\theta\Big[\big(\widehat{\theta}_i-\theta_i\big)^2\Big]. \end{aligned}$$

which completes the proof.

□

1.3 Choice of a good model: hypothesis testing

As an illustrating example, we consider again the stopping distance problem. Recall that we had the following experimental design:

- n vehicles were driven with speeds $v_1, \ldots, v_n$
- the respective stopping distances $y_1, \ldots, y_n$ were recorded, these measurements are modeled as realizations of random variables $Y_1, \ldots, Y_n$

Logical considerations led to the following linear regression model:

$$Y_i \;=\; v_i\theta_1 \;+\; v_i^2\theta_2 \;+\; \varepsilon_i, \qquad i = 1, \ldots, n. \tag{1.6}$$

This model can be rewritten in a more compact vector/matrix form,

$$Y \;=\; X\theta \;+\; \varepsilon,$$

where

$$Y \;=\; \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad X \;=\; \begin{pmatrix} v_1 & v_1^2 \\ \vdots & \vdots \\ v_n & v_n^2 \end{pmatrix}, \qquad \theta \;=\; \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \qquad \varepsilon \;=\; \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

If not all the initial velocities v_i are equal, then the matrix X has rank 2 (see Exercise 2, first problem sheet), the least squares estimator is uniquely defined and has the following form:

$$\widehat{\theta} \;=\; \big(X^TX\big)^{-1}X^TY.$$

Now it could well happen that a statistician is not sure whether or not a model of this complexity is needed. For example, in case of cars equipped with advanced driver assistance systems, the time needed to react to an obstacle should be very small. This means that the unknown parameter θ_1 is close to zero. If so, then the simpler model

$$Y_i \;=\; v_i^2\theta_2 \;+\; \varepsilon_i, \qquad i = 1, \ldots, n. \tag{1.7}$$

could be used as well. This raises the following question: If θ_1 is close to zero (or, ideally, equal to zero), is there any gain by fitting model (1.7) rather than (1.6)? Apart from the lower complexity of model (1.7), can we estimate the remaining parameter θ_2 with a higher precision? To see what typically happens in such a situation, we compare the quadratic risk of the respective least squares estimators of θ_2 in both models. We assume that $E\varepsilon = 0_n$, $\mathrm{Cov}(\varepsilon) = \sigma^2 I_n$ ($\sigma^2 > 0$), and that not all the v_i are equal. Furthermore, we assume that model (1.7) is adequate which means that model (1.6) is adequate as well, with $\theta_1 = 0$. Under these assumptions, it follows from Theorem 1.2 that the matrix risk of the least squares estimator in model (1.6) is equal to

$$E\Big[\big(\widehat{\theta} - \theta\big)\big(\widehat{\theta} - \theta\big)^T\Big] \;=\; \sigma^2\big(X^TX\big)^{-1}.$$

Since

$$X^TX \;=\; \begin{pmatrix} \sum_{i=1}^n v_i^2 & \sum_{i=1}^n v_i^3 \\ \sum_{i=1}^n v_i^3 & \sum_{i=1}^n v_i^4 \end{pmatrix} \;=\; \begin{pmatrix} V_2 & V_3 \\ V_3 & V_4 \end{pmatrix},$$

where $V_k = \sum_{i=1}^n v_i^k$ $(k = 2, 3, 4)$ we obtain that

$$\left(X^T X\right)^{-1} \;=\; \frac{1}{V_2 V_4 \,-\, V_3^2} \begin{pmatrix} V_4 & -V_3 \\ -V_3 & V_2 \end{pmatrix}.$$

This yields for the squared error risk of the second component $\widehat{\theta}_2$ of the least squares estimator that

$$E\Big[\big(\widehat{\theta}_2 - \theta_2\big)^2\Big] \;=\; \sigma^2 \,\Big[\left(X^T X\right)^{-1}\Big]_{2,2} \;=\; \sigma^2 \,\frac{1}{V_4 \,-\, V_3^2/V_2}. \tag{1.8}$$

Now we consider the least squares estimator of θ_2 in the alternative model (1.7). To distinguish it from the corresponding estimator in the full model, we denote it by $\widetilde{\theta}_2$. It follows again from Theorem 1.1 that

$$\widetilde{\theta}_2 \;=\; \frac{\sum_{i=1}^n v_i^2 Y_i}{\sum_{i=1}^n v_i^4},$$

and from Theorem 1.2 that

$$E\Big[\big(\widetilde{\theta}_2 - \theta_2\big)^2\Big] \;=\; \sigma^2 \,\frac{1}{\sum_{i=1}^n v_i^4} \;=\; \sigma^2 \,\frac{1}{V_4}. \tag{1.9}$$

Under the above assumption that not all the velocities v_i are equal we obtain that $V_i > 0$ for $i = 2, 3, 4$, which implies that

$$E\Big[\big(\widetilde{\theta}_2 - \theta_2\big)^2\Big] \;=\; \sigma^2 \,\frac{1}{V_4} \;<\; \sigma^2 \,\frac{1}{V_4 \,-\, V_3^2/V_2} \;=\; E\Big[\big(\widehat{\theta}_2 - \theta_2\big)^2\Big]. \tag{1.10}$$

This shows that, given we know in advance that model (1.7) is adequate, we should better estimate the parameter θ_2 using this reduced model.

This result can be generalized. Suppose that we have the regression model

$$Y \;=\; \underbrace{\begin{pmatrix} X_0 & Z \end{pmatrix}}_{=:X} \theta \,+\, \varepsilon, \tag{1.11}$$

where X_0 is an $(n \times k)$-matrix and Z an $(n \times l)$-matrix such that $\operatorname{rank}(X) = k + l$. Accordingly, the parameter vector θ consists of a subvector θ_0 of length k and a subvector γ of length l $(\theta = (\theta_0^T, \gamma^T)^T)$. If we knew that the last l components of the explanatory variables do not contribute to an explanation of Y, that is if $\gamma = 0_l$, then we could also use the reduced model

$$Y \;=\; X_0 \theta_0 \,+\, \varepsilon \tag{1.12}$$

and estimate the remaining components of θ by least squares. Since the matrix X has rank $k + l$ it follows that its columns are linearly independent, which means that the submatrix X_0 has full column rank l. We conclude from Theorem 1.1 that the least squares estimator of θ_0 in model (1.12) is uniquely defined and is given by

$$\widetilde{\theta}_0 \;=\; \left(X_0^T X_0\right)^{-1} X_0^T Y.$$

According to Theorem 1.2, the matrix risk of this estimator is equal to

$$E_{\theta_0}\Big[\big(\widetilde{\theta}_0 - \theta_0\big)\big(\widetilde{\theta}_0 - \theta_0\big)^T\Big] \;=\; \sigma^2 \big(X_0^T X_0\big)^{-1}. \tag{1.13}$$

In the full model (1.11), the least squares estimator is given by

$$\widehat{\theta} \;=\; \big(X^TX\big)^{-1}X^TY$$

and its matrix risk is equal to

$$E_\theta\Big[\big(\widehat{\theta}-\theta\big)\big(\widehat{\theta}-\theta\big)^T\Big] \;=\; \sigma^2\big(X^TX\big)^{-1}. \tag{1.14}$$

The corresponding estimator $\widehat{\theta}_0$ of the subvector θ_0 is given by the first k components of $\widehat{\theta}$ and, irrespectively of the value of γ, its matrix risk is given by the matrix spanned by the first k columns and the first k rows of $\sigma^2\big(X^TX\big)^{-1}$. Since the matrix

$$X^TX \;=\; \begin{pmatrix} X_0^T \\ Z^T \end{pmatrix}\Big(X_0\;Z\Big) \;=\; \begin{pmatrix} X_0^TX_0 & X_0^TZ \\ Z^TX_0 & Z^TZ \end{pmatrix}$$

is positive definite we have that $\det\big(Z^TZ\big)$ is a regular (positive definite) matrix and we obtain

$$0 \;<\; \det\big(X^TX\big) \;=\; \det\big(Z^TZ\big)\,\det\Big(X_0^TX_0 \;-\; X_0^TZ\big(Z^TZ\big)^{-1}Z^TX_0\Big);$$

see fact 9.11.2(4b) on page 147 in Lütkepohl, H. "Handbook of Matrices" (1996). Therefore, the matrix $\Big(X_0^TX_0 - X_0^TZ\big(Z^TZ\big)^{-1}Z^TX_0\Big)$ is also regular and it follows from statement 9.11.3(2b) on page 148 in the same book that

$$\begin{pmatrix} X_0^TX_0 & X_0^TZ \\ Z^TX_0 & Z^TZ \end{pmatrix}^{-1} \;=\; \begin{pmatrix} \Big(X_0^TX_0 \;-\; X_0^TZ\big(Z^TZ\big)^{-1}Z^TX_0\Big)^{-1} & E \\ E^T & F \end{pmatrix},$$

for some matrices E and F. Therefore we obtain from (1.14) that

$$E_\theta\Big[\big(\widehat{\theta}_0-\theta_0\big)\big(\widehat{\theta}_0-\theta_0\big)^T\Big] \;=\; \sigma^2\Big(X_0^TX_0 \;-\; X_0^TZ\big(Z^TZ\big)^{-1}Z^TX_0\Big)^{-1}. \tag{1.15}$$

Now we can distinguish between two relevant cases:

1) If the columns of X_0 are orthogonal to those of Z, i.e. $X_0^TZ = 0_{k\times l}$, then

$$E_\theta\Big[\big(\widehat{\theta}_0-\theta_0\big)\big(\widehat{\theta}_0-\theta_0\big)^T\Big] \;=\; E_\theta\Big[\big(\widetilde{\theta}_0-\theta_0\big)\big(\widetilde{\theta}_0-\theta_0\big)^T\Big] \;=\; \sigma^2\big(X_0^TX_0\big)^{-1}.$$

In this case, there will be no gain by using the reduced model since the matrix risks of $\widehat{\theta}_0$ and $\widetilde{\theta}_0$ are equal.

2) If $X_0^TZ \neq 0_{k\times l}$, then $\Big(X_0^TX_0 \;-\; \big(X_0^TX_0 - X_0^TZ\big(Z^TZ\big)^{-1}Z^TX_0\big)\Big) \;=\; X_0^TZ\big(Z^TZ\big)^{-1}Z^TX_0$ is a **non-zero** and non-negative definite (positive semidefinite) matrix. This implies that

$$\begin{aligned} &E_\theta\Big[\big(\widehat{\theta}_0-\theta_0\big)\big(\widehat{\theta}_0-\theta_0\big)^T\Big] \;-\; E_\theta\Big[\big(\widetilde{\theta}_0-\theta_0\big)\big(\widetilde{\theta}_0-\theta_0\big)^T\Big] \\ &\quad=\; \sigma^2\Big(X_0^TX_0 \;-\; X_0^TZ\big(Z^TZ\big)^{-1}Z^TX_0\Big)^{-1} \;-\; \sigma^2\Big(X_0^TX_0\Big)^{-1} \end{aligned}$$

is a non-negative definite and **nonzero** matrix. In other words, the risk matrix of $\widehat{\theta}_0$ is "greater" than that of $\widetilde{\theta}_0$. In this case it is clearly advisable to use the reduced model for estimating θ_0.

But what happens if the model we use is **inappropriate**? Let us assume that we employ a linear regression model,

$$Y \;=\; X\theta \,+\, \varepsilon,$$

but that, in contrast to our assumption above, there is no $\theta \in \mathbb{R}^k$ such that $EY = X\theta$. We assume again that the $(n \times k)$-matrix X has rank k and that $\mathrm{Cov}(Y) = \sigma^2 I_n$, for some $\sigma^2 > 0$. Let $\widehat{\theta} = \big(X^T X\big)^{-1} X^T Y$ be the ordinary least squares estimator. As in the case of an adequate model, we have that

$$\mathrm{Cov}\,\big(\widehat{\theta}\big) \;=\; \mathrm{Cov}\,\Big(\big(X^T X\big)^{-1} X^T Y\Big) \;=\; \big(X^T X\big)^{-1} X^T \underbrace{\mathrm{Cov}(Y)}_{=\sigma^2 I_n} X \big(X^T X\big)^{-1} \;=\; \sigma^2 \,\big(X^T X\big)^{-1}.$$

But what does the least squares estimator approximate? In fact, since $EY \neq X\theta \;\forall \theta \in \mathbb{R}^k$, there will be no "true parameter" θ. Recall that $X\widehat{\theta} = X\big(X^T X\big)^{-1} X^T Y$ is the orthogonal projection of Y onto the linear subspace $\big\{Xb\colon\; b \in \mathbb{R}^k\big\}$ which is spanned by the columns of X, that is

$$\big\|Y \,-\, X\widehat{\theta}\big\|^2 \;=\; \inf\big\{\|Y - Xb\|^2\colon\; b \in \mathbb{R}^k\big\}.$$

Let $\bar{\theta} := E\widehat{\theta} = \big(X^T X\big)^{-1} X^T EY$. Then $X\bar{\theta}$ is just the orthogonal projection of EY onto $\big\{Xb\colon\; b \in \mathbb{R}^k\big\}$, that is

$$\big\|EY \,-\, X\bar{\theta}\big\|^2 \;=\; \inf\big\{\|EY - Xb\|^2\colon\; b \in \mathbb{R}^k\big\}. \tag{1.16}$$

In this sense, $\widehat{\theta}$ is still a meaningful quantity. However, it is also clear that

$$E\big[X\widehat{\theta}\big] \;\neq\; EY,$$

that is, this model does not primarily provide an approximation to the conditional expectation of the dependent variable given the explanatory variables. Rather, as (1.16) shows, the least squares method aims at delivering a **"best approximation"** of the true regression function. Hence, it is important to guard against using an inadequate model. To conclude, the above considerations show that it pays off to strive for a **minimal** model which describes the relation between explanatory variables and the independent variable in an **adequate** way.

Now we come back to the problem of deciding whether we should use model (1.6) or the alternative model (1.7). We are convinced by our logical considerations that the former model provides an adequate description of the relation between speed of a car and the corresponding average stopping distance. However, if the reduced model (1.7) is adequate, we should use this to estimate the remaining parameter θ_2; (1.10) shows that this improves the accuracy of the estimator. On the other hand, if (1.7) is not adequate, we face the risk of an insufficient approximation to the true relation between speed and stopping distance. What we need is an objective rule which leads to a decision between these two alternatives. Suppose we are inclined to believe in the reduced model (1.7), which corresponds to $\theta_1 = 0$ in model (1.6). In this context, we want to "test" the hypothesis of $\theta_1 = 0$ versus the alternative $\theta_1 \neq 0$. This pair of hypotheses can be reformulated as

$$H_0\colon\; \theta_1 \,=\, 0 \qquad \text{vs.} \qquad H_1\colon\; \theta_1 \neq 0.$$

In this context, H_0 is called **null hypothesis** while H_1 is the **alternative**. A decision for or against H_0 has to be based on the information given by the observed realization $y =$

$(y_1, \ldots, y_n)^T$ of the random vector $Y = (Y_1, \ldots, Y_n)^T$. (Recall that $v_1, \ldots, v_n$ were predetermined, non-random speeds.) A decision rule (which is called **test**) can be described by a function $\varphi\colon\ \mathbb{R}^n \to \{0,1\}$, where $\varphi(y) = 1$ means that we **reject** H_0 in case of $Y = y$ whereas $\varphi(y) = 0$ describes the fact that we **accept** H_0 if $Y = y$. (An advanced concept where φ may also attain values between 0 and 1 will be described in Section 3.) The performance of a test is reasonably measured by the frequency with which we make correct judgments when we use it. There are two types of error we can commit (not simultaneously):

- H_0 is true, but we reject H_0 ("**type one error**"),
- H_0 is false, but we accept H_0 ("**type two error**").

When we are looking for a "good" test we are seeking a test such that the probabilities of such wrong decisions are as small as possible. In our case, it would be desirable to keep both $P_{(0,\theta_2)}\big(\varphi(Y) = 1\big)$ and $P_{(\theta_1,\theta_2)}\big(\varphi(Y) = 0\big)$ small, the latter probability for all $\theta_1 \neq 0$. (P_θ denotes the probability measure corresponding to a given value for the parameter θ.)

To simplify matters, we make the assumption that

$$Y \ \sim\ N\big(\gamma, \sigma^2 I_n\big),$$

where $\gamma \in \big\{Xb\colon\ b \in \mathbb{R}^2\big\}$ and $\sigma^2 > 0$. Here, σ^2 is also an unknown parameter. The above test problem can be formulated as follows:

$$H_0\colon\ \begin{pmatrix} \gamma \\ \sigma^2 \end{pmatrix} \in \Theta_0 \qquad \text{vs.} \qquad H_1\colon\ \begin{pmatrix} \gamma \\ \sigma^2 \end{pmatrix} \in \Theta_1 \setminus \Theta_0,$$

where

$$\Theta_0 \ =\ \big\{X_0 b\colon\ b \in \mathbb{R}\big\} \times (0,\infty), \qquad \Theta_1 \ =\ \big\{Xb\colon\ b \in \mathbb{R}^2\big\} \times (0,\infty)$$

and

$$X_0 \ =\ \begin{pmatrix} v_1^2 \\ \vdots \\ v_n^2 \end{pmatrix}, \qquad X \ =\ \begin{pmatrix} v_1 & v_1^2 \\ \vdots & \vdots \\ v_n & v_n^2 \end{pmatrix}.$$

In Section 3 we will show that an "ideal" test such that both error probabilities are zero does not exist. Typically, one uses the following strategy: One assigns a small bound α to the probability of a type one error, and then one attempts to minimize the probability of a type two error, under the side condition that the probability of a type one error does not exceed the chosen bound α. We will also see in Section 3 that we can only "approximately" reach this goal in our case. A "reasonably good" test is the so-called **F test** (named after the British statistician Sir Ronald A. Fisher) which is described by a function $\varphi\colon\ \mathbb{R}^n \to \{0,1\}$ such that

$$\varphi(y) \ =\ \begin{cases} 1, & \text{if } T_n(y) \geq c, \\ 0, & \text{if } T_n(y) < c, \end{cases}$$

where

$$T_n(y) \ =\ \frac{\|P_1 y \,-\, P_0 y\|^2}{\frac{1}{n-2}\|y \,-\, P_1 y\}^2}.$$

$P_1 = X\big(X^T X\big)^{-1} X^T$ and $P_0 = X_0\big(X_0^T X_0\big)^{-1} X_0^T$ are the orthogonal projection matrices onto the linear subspace spanned by X and X_0, respectively. The numerator of T_n,

$\|P_1Y - P_0Y\|^2$, characterizes a possible inadequacy of the smaller model while its denominator, $\|Y - P_1Y\|/(n-2)$ is an unbiased estimator of the variance σ^2. It turns out that under the null hypothesis $T_n(Y)$ has a so-called F distribution with 1 and $n-2$ degrees of freedom, irrespectively of the actual value of σ^2. Therefore, if we choose the **"critical value"** c equal to the $(1-\alpha)$-quantile of such an F distribution, then we obtain a statistical test where the probability of a type one error is equal to α; for details see Section 3.

If we are even not sure that the larger model (1.6) is correct, then we could also test its adequacy. To do this, we need again a model which is guaranteed to be adequate. Recall how the above test statistic $T_n(Y)$ is built: Its numerator, $\|P_1Y - P_0Y\|^2$, compares the goodness of fit of the smaller model with that of the larger one while its denominator, $\|Y - P_1Y\|/(n-2)$ is an unbiased estimator of the variance σ^2. Now we can reliably estimate σ^2 if we have **"replications"**, that is, if $v_i = v_j$, for some pair(s) (i,j), $i \neq j$. Suppose that the n cars were driven with only m different initial speeds, $2 < m < n$. Let us assume that the ith velocity v_i appears n_i times and that the corresponding measured stopping distances are $y_{i1}, \ldots, y_{in_i}$, $n_1 + \cdots + n_m = n$. We avoid any assumption on the particular form of the function f and assume only that $EY_{i1} = \ldots = EY_{in_i}$, for $i = 1, \ldots, m$. This leads to the following linear regression model:

$$\underbrace{\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{m1} \\ \vdots \\ Y_{mn_m} \end{pmatrix}}_{=:Y} = \underbrace{\begin{pmatrix} \beta_1 \mathbb{1}_{n_1} \\ \vdots \\ \beta_m \mathbb{1}_{n_m} \end{pmatrix}}_{=:X_1} + \underbrace{\begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \vdots \\ \varepsilon_{m1} \\ \vdots \\ \varepsilon_{mn_m} \end{pmatrix}}_{=:\varepsilon},$$

where $\beta_1, \ldots, \beta_m$ are unknown parameters. Here and in the following $\mathbb{1}_n$ denotes the vector consisting of n ones. The model we want to test has the form

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{m1} \\ \vdots \\ Y_{mn_m} \end{pmatrix} = \underbrace{\begin{pmatrix} v_1 \mathbb{1}_{n_1} & v_1^2 \mathbb{1}_{n_1} \\ \vdots & \vdots \\ v_m \mathbb{1}_{n_m} & v_m^2 \mathbb{1}_{n_m} \end{pmatrix}}_{=:X} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \vdots \\ \varepsilon_{m1} \\ \vdots \\ \varepsilon_{mn_m} \end{pmatrix}.$$

If we assume again that

$$Y \sim N(\gamma, \sigma^2 I_n),$$

then we obtain in analogy to the previous considerations the following test problem:

$$H_0\colon \begin{pmatrix} \gamma \\ \sigma^2 \end{pmatrix} \in \Theta_0 \qquad \text{vs.} \qquad H_1\colon \begin{pmatrix} \gamma \\ \sigma^2 \end{pmatrix} \in \Theta_1 \setminus \Theta_0,$$

where

$$\Theta_0 = \{Xb\colon\ b \in \mathbb{R}^2\} \times (0,\infty), \qquad \Theta_1 = \{X_1 b\colon\ b \in \mathbb{R}^m\} \times (0,\infty).$$

In this case, the corresponding F test is given by

$$\varphi(y) \;=\; \begin{cases} 1, & \text{if } T_n(y) \geq c, \\ 0, & \text{if } T_n(y) < c, \end{cases}$$

where

$$T_n(y) \;=\; \frac{\frac{1}{m-2}\|P_1 y \,-\, P_0 y\|^2}{\frac{1}{n-m}\|y \,-\, P_1 y\|^2}.$$

$P_1 = X_1\big(X_1^T X_1\big)^{-1} X^T$ and $P_0 = X\big(X^T X\big)^{-1} X^T$ are the orthogonal projection matrices onto the linear subspace spanned by X_1 and X, respectively. It will be shown in Section 3 of this course that the statistic $T_n(Y)$ has an F distribution with $m-1$ and $n-m$ degrees of freedom under the null hypothesis. The probability of a type 1 error will be equal to $\alpha \in (0,1)$ if the critical value c is chosen equal to the $(1-\alpha)$-quantile of an F distribution with $m-1$ and $n-m$ degrees of freedom.

2 Statistical estimation of parameters

2.1 A model for a statistical experiment

Statistical studies and experiments produce data whose analysis is the ultimate goal of the venture. Mathematical statistics deals with situations in which the data can be thought of as the outcome of a random experiment. In order to arrive at a suitable formulation of a model for such experiments, we consider two simple, but nevertheless typical, examples:

Example 2.1. Suppose we are faced with a population of N elements, for example, a shipment of N manufactured items. An unknown number θ of these elements are defective. It might be desirable to know this number of defective items, for example, because a too large number of defective items gives the recipient the right to reject the shipment. Suppose that it is not possible (e.g. too expensive) to examine all of the items. To get informationot worse about θ, a sample of n items is drawn without replacement and inspected. The data gathered are the number of defective items found in the sample.

Example 2.2. An experimenter makes n independent measurements of the distance required to stop a car traveling with respective speeds $v_1, \ldots, v_n$. These measurements are subject to random fluctuations which might be caused by several factors, e.g. changing environmental conditions or even by an imprecise reading from a measurement device. Therefore, the data can be thought of as stopping distance according to the initial velocity plus some random errors.

We use these two simple examples to develop an abstract framework for such experiments. Let us begin with Example 2.1. The possible outcomes of this experiments are described by the numbers in the set $\Omega = \{0, 1, \ldots, n\}$. On this space we can define the random variable X by $X(\omega) = \omega$, $\omega = 0, 1, \ldots, n$. If θ is the **(unknown)** number of defective items in the shipment, then the random variable X has a hypergeometric distribution with parameters θ, N and n, that is

$$P\big(X = k\big) \;=\; \frac{\binom{\theta}{k}\binom{N-\theta}{n-k}}{\binom{N}{n}} \qquad \text{for } k = 0, 1, \ldots, n,$$

where

$$\binom{n}{k} \;=\; \begin{cases} \frac{n!}{k!\,(n-k)!} & \text{if } k \in \{0, 1, \ldots, n\}, \\ 0 & \text{if } k \notin \{0, 1, \ldots, n\}. \end{cases}$$

For a given value of the parameter θ, we encounter the well known probability model: We have a probability space $\big(\Omega, 2^\Omega, P\big)$ and, on this probability, space a random variable X. (We can use without hesitation the power set 2^Ω in this context.) The probability measure P is a equal to $\mathcal{H}(\theta, N; n)$, a hypergeometric distribution with parameters θ, N and n.

At this point, however, a striking difference to typical models in probability theory becomes apparent: Since we do not know in advance which of the possible values for the parameter θ is the right one, we have to take into account that in principle all possible values $\big\{0, 1, \ldots, N\big\}$ are candidates for being this one. This aspect will become vitally important when we compare the performance of competing methods to estimate an unknown parameter. It will be shown that, apart from trivial and therefore meaningless cases, a uniformly best method does not exist. Different methods of estimation have usually their pros and cons in respective parts of the parameter space. Hence it is important to specify for which set of parameters a given method has some desirable property.

In view of this, we have to replace the single distribution $\mathcal{H}(\theta, N; n)$ by an appropriate **family of distributions**, $\mathcal{P} = \big\{\mathcal{H}(\theta, N; n)\colon\ \theta \in \{0, 1, \ldots, N\}\big\}$, any one of which could have generated the data actually observed. The above **statistical experiment** can be described by the triple $\big(\Omega, 2^\Omega, \mathcal{P}\big)$ or, more completely, by the quadruple $\big(X, \Omega, 2^\Omega, \mathcal{P}\big)$. In most cases it will not be necessary to use this cumbersome notation. To simplify matters we can choose the basic space Ω equal to the set of the possible values of the random variable (or random vector) which generates the data as its realization. In this case it is also not necessary to specify the random variable in such a model. Moreover, the σ-field (here 2^Ω) is usually of minor interest in this context. Therefore, to simplify notation, we would consider the family of distribution $\big\{\mathcal{H}(\theta, N; n)\colon\ \theta \in \{0, 1, \ldots, N\}\big\}$ as an appropriate description of our statistical experiment.

We turn now to Example 2.2 of determining the distance required for stopping a car traveling with an initial speed v. Here, the choice of an appropriate model is less clear than in case of the first example. First of all, we would probably choose as a set containing all possible values of our measurements $\Omega = \mathbb{R}^n$ or, since we know in this case that all measurements will be non-negative, $\Omega = [0, \infty)^n$. As a suitable σ-field we could take the corresponding Borel σ-field, $\mathcal{B}^n$, or a suitable trace σ-field. But what about a suitable family of possible distributions? We can describe the setting by a simple regression model:

$$X_i \,=\, f(v_i) \,+\, \varepsilon_i, \qquad i = 1, 2, \ldots, n, \tag{2.1}$$

where $f(v)$ describes the average stopping distance of a car traveling at speed v. We still have to specify conditions on the vector of errors $\varepsilon = \big(\varepsilon_1, \ldots, \varepsilon_n\big)^T$. Of course, an appropriate specification depends on how the experiment is carried out. The following minimal assumptions are usually made:

(i) The errors ε_i have expectation 0.

(ii) The value of the error committed on one measurement does not affect the value of the error at other times. That is, $\varepsilon_1, \ldots, \varepsilon_n$ are independent.

According to our logical considerations described in Subsection 1.1 we could specify the function f as

(iii) $f(v) \,=\, v\theta_1 \,+\, v^2\theta_2$, where θ_1 and θ_2 are unknown parameters.

This leads to the linear regression model that we know already from our considerations in the previous section:

$$X_i \,=\, v_i\theta_1 \,+\, v_i^2\theta_2 \,+\, \varepsilon_i, \qquad i = 1, 2, \ldots, n. \tag{2.2}$$

And finally, we could also make a more specific assumption on the errors, for example:

(iv) $\varepsilon \,\sim\, \text{Uniform}[-\sigma v_i, \sigma v_i]$.

(If $\sigma \,\geq\, \theta_1$, then the corresponding random variables X_i are guaranteed to be non-negative.)

If we assume (i) to (iv), then the distribution of the random vector depends only on the parameters θ_1, θ_2 and σ, that is, $X = (X_1, \ldots, X_n)^T \,\sim\, P_{(\theta_1, \theta_2, \sigma)}$ for some appropriate choice of $P_{(\theta_1, \theta_2, \sigma)}$, where $\theta_1, \theta_2, \sigma \,\geq\, 0$. In this case the statistical experiment is conveniently described by $\mathcal{P} = \big\{P_{(\theta_1, \theta_2, \sigma)}\colon\ \theta_1, \theta_2, \sigma \geq 0\big\}$. This is a so-called **parametric model** since the family of possible distributions is parametrized by a finite-dimensional

parameter $(\theta_1, \theta_2, \sigma)^T \in \Theta := [0, \infty)^3$. One part of the parameter vector, $(\theta_1, \theta_2)^T$, parametrizes the quantity of interest (the function f), and the other subparameter σ is a so-called **nuisance parameter**. At the other end of the scale, when we assume only (i) and (ii), both the the function f we are primarily interested in and the distribution of the errors cannot be described by any finite-dimensional parameter. Such a model is called **nonparametric**. If we assume (i), (ii), and (iii), then we obtain a somewhat mixed situation. The function of interest f is completely described by the finite-dimensional parameter $(\theta_1, \theta_2)^T$ but the distribution of the errors may vary more or less freely. Such a model is said to be a **semiparametric model**.

How do we settle on a set of assumptions? In the first example the assumption of a hypergeometric distribution seems to be well motivated and the parametric model $\mathcal{P} = \big\{\mathcal{H}(\theta, N; n)\colon\ \theta \in \{0, 1, \ldots, N\}\big\}$ seems to be appropriate. In the second example, however, the decision about a model is more difficult. The choice of the assumptions can be based on experience, physical considerations (these probably lead to (iii)), and wishful thinking. Maybe (i), (ii), and (iii) are a good compromise between too weak and too rigorous assumptions. The advantage of assuming (i) to (iv) is that, if they are true, we know how to combine our measurements to estimate f in a highly efficient way. The danger is that, if they are false, our analysis, though correct for the model written down, may be quite irrelevant to the experiment that was actually performed.

2.2 Some methods of estimation

Once we have constructed a statistical model, we usually want to estimate the parameter of the unknown distribution generating the data. In some cases we are not primarily interested in the parameter itself, but in some subparameter or in some quantity related to the parameter used for parametrizing the model. Let us assume that the model is given by the family $\mathcal{P} = \big\{P_\theta\colon\ \theta \in \Theta\big\}$ of possible distributions of $X = \big(X_1, \ldots, X_n\big)^T$, where Θ is the parameter space. For instance, in the hypergeometric Example 2.1, we have $\mathcal{P} = \Big\{\mathcal{H}(\theta, N; n)\colon\ \theta \in \{0, 1, \ldots, n\}\Big\}$ and the quantity we want to estimate is the parameter θ itself. In the stopping distance Example 2.2, we might be only interested in estimating the function f which describes the relation between speed and the distance required for stopping. If we assume (i) to (iv), then the statistical model can be written in the form $\big\{P_\theta\colon\ \theta = (\theta_1, \theta_2, \sigma)^T \in [0, \infty)^3\big\}$. Since, according to assumption (iii), $f(v) = v\theta_1 + v^2\theta_2$, we only need to estimate $q(\theta) := (\theta_1, \theta_2)^T$. To estimate $q(\theta)$ we select a statistic $T = T(X_1, \ldots, X_n)$ and evaluate it at the outcome $(x_1, \ldots, x_n)$ of the experiment. Thus, if the true value of θ is θ_0, we observe $X_1 = x_1, \ldots, X_n = x_n$ and we approximate the unknown quantity $q(\theta_0)$ by the known value $T(x_1, \ldots, x_n)$ of the statistic $T(X_1, \ldots, X_n)$.

At this point it seems to be appropriate to introduce a few notions that will be used throughout this course. In classical probability, we fix a probability space $\big(\Omega, \mathcal{A}, P\big)$ as a starting point, where $\mathcal{A}$ is a σ-field in the non-empty set Ω and $P\colon\ \mathcal{A} \to [0, 1]$ is a probability measure on $\mathcal{A}$. On Ω, we may define random variables $X_1, X_2, \ldots$, where $X_i\colon\ \Omega \to \Omega_i$ is $(\mathcal{A} - \mathcal{A}_i)$-measurable, for some σ-field in Ω_i. As already explained, in mathematical statistics, we have to replace the single probability measure P by a **family** $\mathcal{P} = \big\{P_\theta\colon\ \theta \in \Theta\big\}$ of probability measures. In this course, we will exclusively deal with real-valued or $\mathbb{R}^d$-valued random variables, i.e. $\Omega_i = \mathbb{R}$ or $\Omega_i = \mathbb{R}^d$ and $\mathcal{A}_i = \mathcal{B}$ or $\mathcal{A}_i = \mathcal{B}^d$ are the respective Borel σ-fields. In most cases we can choose Ω as a set

containing all possible values of a random variable X and it will not be necessary to make our choice of Ω and $\mathcal{A}$ explicit. Our statistical analysis will be based on data x which will be thought of a realization of this random variable. Formally, the random variable X is a function $X\colon \Omega \to \Omega$ such that $X(\omega) = \omega\ \forall \omega \in \Omega$. A function $T\colon\ \Omega \to \Omega_T$ such that $T(\omega) = g(X(\omega))$ for some function g which is constructed for a certain purpose is called a **statistic**. In other words, a statistic is a random variable which is a function of the originally observed random variable X. As we will see later, we will require that T is $(\mathcal{A} - \mathcal{A}_T)$-measurable, where $\mathcal{A}_T$ is a suitable σ-field in Ω_T. (This is fulfilled if the above function g is $(\mathcal{A} - \mathcal{A}_T)$-measurable since the identity $X\colon \Omega \to \Omega$ is obviously $(\mathcal{A} - \mathcal{A})$-measurable.) A statistic T which is constructed with the aim to approximate the parameter θ is called **estimator** of θ and will be denoted by $\widehat{\theta}$. If we are merely interested in some related quantity, $q(\theta)$, a statistics T which is constructed with the purpose of approximating $q(\theta)$ will be denoted by $\widehat{q}$ or $\widehat{q}(\theta)$. A **realization** $T(\omega)$ of an estimator T of θ (or $q(\theta)$) is called **estimate** of θ (or $q(\theta)$). If there is no chance of confusion, we will identify an estimator with its value at a point. Thus $\widehat{\theta}$ can stand for both the statistic $\widehat{\theta}(X_1, \ldots, X_n)$ and the realization $\widehat{\theta}(x_1, \ldots, x_n)$.

In what follows we introduce two popular methods of parameter estimation. We assume that the observable random quantity X has a distribution P_θ, where $\theta \in \Theta$, for some parameter space Θ. Typically the sample will consists of more than one observation, that is $X = (X_1, \ldots, X_n)^T$ and n is the so-called **sample size**.

Method of Moments

Suppose that the parameter space Θ is a subset of $\mathbb{R}^d$. Suppose further that $X_1, \ldots, X_n$ are identically distributed under P_θ, and that the first d theoretical moments

$$\mu_k(\theta) \ = \ E_\theta\big[X_1^k\big], \qquad k = 1, \ldots, d,$$

are finite. We define the corresponding sample moments $\widehat{\mu}_k$ by

$$\widehat{\mu}_k \ = \ \frac{1}{n} \sum_{i=1}^{n} X_i^k.$$

Then the method of moments estimator $\widehat{\theta}_{MM}$ of θ is obtained by equating the first d theoretical moments with the corresponding sample moments, that is $\widehat{\theta}_{MM}$ is defined as any parameter from Θ such that

$$\mu_k\big(\widehat{\theta}_{MM}\big) \ = \ \widehat{\mu}_k \qquad \forall k = 1, \ldots, d.$$

We have to admit that a method of moments estimator need neither exist nor be unique. These properties have to checked case by case. For typical "textbook examples", the method of moments estimators are easy to compute. Moreover, if the sample size n is large, these estimators are likely to be close to the true value of the parameter. If we are not interested in θ but in a related quantity $q(\theta)$, then the corresponding method of moments estimator is given by $\widehat{q}_{MM} = q(\widehat{\theta}_{MM})$.

Examples

1) Suppose that $X_1, \ldots, X_n \sim \mathrm{Bin}(1, \theta)$, $\theta \in \Theta := [0, 1]$. Then $\mu_1(\theta) = E_\theta[X_1] = \theta$. Since $\widehat{\mu}_1 = \frac{1}{n} \sum_{i=1}^{n} X_i$ we obtain that

$$\widehat{\theta}_{MM} \ = \ \frac{1}{n} \sum_{i=1}^{n} X_i.$$

2) Suppose that $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, $\theta = \binom{\mu}{\sigma^2} \in \Theta := \mathbb{R} \times (0, \infty)$. Then the first two theoretical moments are given by

$$\begin{aligned} \mu_1(\theta) &= \mu, \\ \mu_2(\theta) &= \sigma^2 + \mu^2 \end{aligned}$$

while the corresponding sample moments are

$$\begin{aligned} \widehat{\mu}_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i, \\ \widehat{\mu}_2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2. \end{aligned}$$

Since the method of moments estimator $\widehat{\theta}_{MM} = \binom{\widehat{\mu}_{MM}}{\widehat{\sigma}^2_{MM}}$ has to satisfy $\mu_i(\widehat{\theta}_{MM}) = \widehat{\mu}_i$ for $i = 1, 2$ we obtain that

$$\begin{aligned} \widehat{\mu}_{MM} &= \bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i, \\ \widehat{\sigma}^2_{MM} &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \big(\bar{X}_n\big)^2 = \ldots = \frac{1}{n}\sum_{i=1}^{n} \big(X_i - \bar{X}_n\big)^2. \end{aligned}$$

Maximum likelihood method

The maximum likelihood method has a long history. It seems that it was first proposed by Carl Friedrich Gauß in 1821, and rediscovered and popularized by the British statistician Sir Ronald A. Fisher in about 1921.

Suppose that realizations $x_1, \ldots, x_n$ of random variables $X_1, \ldots, X_n$ are observed. To simplify notation, we set $X = \big(X_1, \ldots, X_n\big)^T$ and $x = (x_1, \ldots, x_n)^T$ and we assume that $X \sim P_\theta$, for some $\theta \in \Theta$. We consider first the special case that the random variable X takes its values in a finite or countably infinite set Ω_X, that is X has a discrete distribution P_θ. To obtain an estimate of the parameter θ we consider for all $\theta \in \Theta$ the respective probabilities of the event $X = x$ and choose the **most plausible** among these parameters. Hence, the so-called **maximum likelihood estimate** $\widehat{\theta}_{ML}(x)$ is given as a solution to

$$P_{\widehat{\theta}_{ML}(x)}\big(X = x\big) = \sup_{\theta \in \Theta} P_\theta\big(X = x\big),$$

that is, $\widehat{\theta}_{ML}(x)$ is chosen as such a value of the parameter θ which maximizes the probability that the event $X = x$ occurs. We denote by $\widehat{\theta}_{ML} = \widehat{\theta}_{ML}(X)$ the corresponding **maximum likelihood estimator**. (As for a methods of moments estimator, existence and uniqueness of such an estimator is not guaranteed in general. However, for typical "textbook examples" a unique solution to the above optimization problem exists.)

The method of maximum likelihood can also be used in the case of continuous distributions P_θ which have a densities p_θ w.r.t. some σ-finite measure μ. For the definition of a maximum likelihood estimator, the probabilities $P_\theta(X = x)$ have to be replaced by the values $p_\theta(x)$ of the densities at the point x. From now on we assume that either:

(i) All of the P_θ are continuous with densities p_θ w.r.t. one and the same σ-finite dominating measure μ;

(ii) All of the P_θ are discrete, and there exists a set $\{x_1, \dots, x_N\}$ or $\{x_1, x_2, \dots\}$ which is **independent of** θ such that $\sum_i P_\theta(X = x_i) = 1$ for all $\theta \in \Theta$.

(Note that the discrete case can also be cast in the form (i); then the counting measure plays the role of the dominating measure μ.) To unify our notation, we define a so-called **likelihood function** L by

$$L\big(\theta; x\big) \;=\; \begin{cases} p_\theta(x) & \text{in case (i),} \\ P_\theta(X = x) & \text{in case (ii).} \end{cases}$$

The maximum likelihood estimate $\widehat{\theta}_{ML}(x)$ of θ is defined by

$$L\big(\widehat{\theta}_{ML}(x); x\big) \;=\; \sup\big\{L(\theta; x)\colon\ \theta \in \Theta\big\}.$$

If the parameter space Θ contains finitely many points, then a maximum likelihood estimate can always be obtained by comparing finitely many values $L(\theta; x)$, $\theta \in \Theta$. If $L(\theta; x)$ is differentiable on the interior Θ° of Θ, then possible candidates for maximum likelihood estimates are the values $\theta \in \Theta^\circ$ satisfying

$$\frac{\partial}{\partial \theta_i} L\big(\theta; x\big) \;=\; 0 \qquad i = 1, \dots, d.$$

Note that θ's satisfying these so-called **likelihood equations** may be local or global minima, local or global maxima, or simply stationary points. Also, extrema may occur at the boundary of Θ or when $\|\theta\| \to \infty$. Hence, it is important to analyze the entire likelihood function to find its maxima.

Examples

1) (Normal distribution with known variance)
Suppose that $X_1, \dots, X_n$ are independent, $X_i \sim N(\theta, \sigma^2)$, where $\sigma^2 > 0$ is assumed to be known and $\theta \in \Theta := \mathbb{R}$. Then $X = (X_1, \dots, X_n)^T$ has a multivariate normal distribution with a density p_θ w.r.t. the n-dimensional Lebesgue measure λ^n given by

$$p_\theta(x) \;=\; \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\theta)^2}{2\sigma^2}}.$$

In this case, the maximum likelihood estimate $\widehat{\theta}_{ML}(x)$ of θ can be easily obtained by inspection. Indeed, since

$$L\big(\theta; x\big) \;=\; p_\theta(x) \;=\; \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \big((x_i - \bar{x}_n)^2 \;+\; (\bar{x}_n - \theta)^2\big) \Big\}$$

we obtain that $\widehat{\theta}_{ML}(x) = \bar{x}_n$. The corresponding maximum likelihood estimator is given by $\widehat{\theta}_{ML} = \widehat{\theta}_{ML}(X) = \bar{X}_n$.

2) (Normal distribution with unknown mean and variance)
Suppose that $X_1, \dots, X_n$ are independent, $X_i \sim N(\mu, \sigma^2)$, where this time both the location parameter $\mu \in \mathbb{R}$ and the variance $\sigma^2 > 0$ are unknown. Now $\theta = \binom{\mu}{\sigma^2}$ is the parameter of interest, $\theta \in \Theta := \mathbb{R} \times (0, \infty)$. The density p_θ of $X = (X_1, \dots, X_n)^T$ w.r.t. λ^n is given by

$$p_\theta(x) \;=\; \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \big((x_i - \mu)^2\big) \Big\}.$$

The likelihood can be written in the following form:

$$\begin{aligned} & L\big(\theta; x\big) \\ &= \; p_\theta(x) \;=\; \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \big((x_i - \bar{x}_n)^2 \,+\, (\bar{x}_n - \mu)^2\big)\Big\} \\ &= \; \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \big(x_i - \bar{x}_n\big)^2\Big\} \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \big(\bar{x}_n - \mu\big)^2\Big\}. \qquad (2.3) \end{aligned}$$

The first component μ of the parameter θ appears on the right-hand side of (2.1) only in the third factor. Thus, irrespectively of the choice of σ^2, the value $\widehat{\mu}_{ML}$ which maximizes the likelihood function is given by

$$\widehat{\mu}_{ML} \;=\; \bar{x}_n.$$

To find the maximum likelihood estimate $\widehat{\sigma}^2_{ML}$ of the second component of θ, we have to identify the maximizer of the function $\sigma^2 \mapsto L\big(\binom{\widehat{\mu}_{ML}}{\sigma^2}; x\big)$. Let

$$\begin{aligned} g\big(\sigma^2\big) \;&:=\; \ln L\left(\binom{\widehat{\mu}_{ML}}{\sigma^2}; x\right) \\ &=\; -\frac{n}{2}\ln(2\pi) \,-\, \frac{n}{2}\ln\big(\sigma^2\big) \,-\, \frac{1}{2\sigma^2}\sum_{i=1}^{n}\big(x_i - \bar{x}_n\big)^2. \end{aligned}$$

The function $g\colon\ (0,\infty) \to \mathbb{R}$ is infinitely often differentiable and we obtain that

$$\frac{d}{d\sigma^2} g\big(\sigma^2\big) \;=\; -\frac{n}{2}\frac{1}{\sigma^2} \,+\, \frac{1}{2\sigma^4}\sum_{i=1}^{n}\big(x_i - \bar{x}_n\big)^2.$$

To find a candidate for a maximizer of g, we seek zeroes of $\frac{d}{d\sigma^2}g$. It is easy to see that $\frac{d}{d\sigma^2}g\big(\sigma^2\big) = 0$ if and only if $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\big(x_i - \bar{x}_n\big)^2$. Since

$$\begin{aligned} \frac{d^2}{d(\sigma^2)^2} g\big(\sigma^2\big) \;&=\; \frac{n}{2}\frac{1}{\sigma^4} \,-\, \frac{1}{\sigma^6}\sum_{i=1}^{n}\big(x_i - \bar{x}_n\big)^2 \\ &=\; \frac{n}{2\sigma^6}\underbrace{\left(\sigma^2 \,-\, \frac{1}{n}\sum_{i=1}^{n}\big(x_i - \bar{x}_n\big)^2\right)}_{=0 \text{ if } \sigma^2 = \frac{1}{n}\sum_i (x_i - \bar{x}_n)^2} \,-\, \underbrace{\frac{1}{2\sigma^6}\sum_{i=1}^{n}\big(x_i - \bar{x}_n\big)^2}_{<0} \;<\; 0 \end{aligned}$$

if $\sum_i (x_i - \bar{x}_n)^2 > 0$, we see that g attains its **global maximum** at $\sigma^2 = \frac{1}{n}\sum_i (x_i - \bar{x}_n)^2$. Hence the maximum likelihood estimate of σ^2 is given by

$$\widehat{\sigma}^2_{ML} \;=\; \frac{1}{n}\sum_{i=1}^{n}\big(x_i \,-\, \bar{x}_n\big)^2.$$

The value of the maximum likelihood estimator $\widehat{\theta}$ for a realization $x = (x_1, \ldots, x_n)^T$ of the random variable $X = \big(X_1, \ldots, X_n\big)^T$ is therefore

$$\widehat{\theta}_{ML}(x) \;=\; \begin{pmatrix} \bar{x}_n \\ \frac{1}{n}\sum_{i=1}^{n}\big(x_i \,-\, \bar{x}_n\big)^2 \end{pmatrix}.$$

The maximum likelihood estimator of θ (a random variable!) is then

$$\widehat{\theta}_{ML}(X) \;=\; \begin{pmatrix} \bar{X}_n \\ \frac{1}{n}\sum_{i=1}^{n}\big(X_i \,-\, \bar{X}_n\big)^2 \end{pmatrix}.$$

The maximum likelihood method is based on the maximization of the value $p_\theta(x)$ of the densities of the observed random variable X w.r.t. a dominating measure μ. The following lemma states the choice of this measure μ does not influence the result of this method.

Lemma 2.1. *The maximum likelihood estimator $\widehat{\theta}_{ML}$ does* **not** *depend on the choice of the dominating measure.*

Proof. Let μ_1 and μ_2 be σ-finite dominating measures for the family $\{P_\theta : \theta \in \Theta\}$ and let $p_{\theta,1}$ and $p_{\theta,2}$ be the respective densities of P_θ. Then

$$P_\theta \ll \mu_i \ll \mu_1 + \mu_2 \qquad \forall \theta \in \Theta, i = 1, 2.$$

($\mu \ll \nu$ for two measures μ and ν means that μ is absolutely continuous w.r.t. ν.)
Then P_θ has a density $p_{\theta,12}$ w.r.t. $\mu_1 + \mu_2$ and we obtain that

$$p_{\theta,12}(x) = \frac{dP_\theta}{d(\mu_1+\mu_2)}(x) = \underbrace{\frac{dP_\theta}{d\mu_i}(x)}_{=p_{\theta,i}(x)} \frac{d\mu_i}{d(\mu_1+\mu_2)}(x) \qquad i = 1, 2,$$

which implies that

$$p_{\theta,1}(x)\frac{d\mu_1}{d(\mu_1+\mu_2)}(x) = p_{\theta,2}(x)\frac{d\mu_2}{d(\mu_1+\mu_2)}(x) \qquad (\mu_1+\mu_2)-\text{almost everywhere.}$$

But this means in other words that, with probability 1, the maximization of $p_{\theta,1}(x)$ is equivalent to that of $p_{\theta,2}(x)$. □

2.3 Consistency of estimators

The method of moments and the maximum likelihood method are general approaches to obtaining "reasonable" estimators. They seem to be intuitive but they are not constructed with the ambitious goal to obtain an "optimal" procedure. On the other hand, if there is a great deal of data, the sample size is "large," then we should expect that an estimator $\widehat{\theta}$ of a parameter is with high probability close to its target θ. This is some sort of minimal property for a method of estimation and we can formulate this in a rigorous manner by asymptotic considerations. Thus we consider, for a general method of estimation, **sequences of estimators** $\left(\widehat{\theta}_n\right)_{n\in\mathbb{N}}$ where $\widehat{\theta}_n = \widehat{\theta}_n(X_1, \ldots, X_n)$ and we show that this sequence converges to θ as the sample size n tends to infinity. In line with this, we can no longer stick to a fixed statistical experiment but rather we have to consider appropriate sequences of statistical experiments. We think, however, that such a rigorous formalization is not necessary in this subsection and, in order not to overburden ourselves with a too cumbersome formalism, we formulate our results in a more loose way. The following definition of concepts of consistency is based on the commonly used modes of convergence in probability theory.

Definition 2.1. Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of random variables, where $X_i \sim P_\theta^{X_i}$, for some $\theta \in \Theta$, $\Theta \subseteq \mathbb{R}^d$.

(i) A sequence of estimators $\widehat{q}_n = \widehat{q}_n(X_1, \ldots, X_n)$ of an $\mathbb{R}^d$-valued parameter $q(\theta)$ is said to be **consistent** (weakly consistent) for $q(\theta)$ if

$$\widehat{q}_n \xrightarrow{P_\theta} q(\theta) \qquad \forall \theta \in \Theta,$$

that is, for all $\epsilon > 0$,

$$P_\theta\Big(\big\|\widehat{q}_n - q(\theta)\big\| > \epsilon\Big) \underset{n\to\infty}{\longrightarrow} 0.$$

(ii) $\widehat{q}_n$ is called **strongly consistent** for $q(\theta)$ if

$$\widehat{q}_n \overset{P_\theta - a.s.}{\longrightarrow} q(\theta) \qquad \forall \theta \in \Theta.$$

Although consistency is a concept relating to a **sequence** of estimators $\big(\widehat{q}_n\big)_{n\in\mathbb{N}}$, we often say "consistency of $\widehat{q}_n$" for simplicity.

Example 2.3. Suppose that $(X_i)_{i\in\mathbb{N}}$ is a sequence of independent and identically distributed random variables, where $X_i \sim N(\mu, \sigma^2)$, $\theta = \binom{\mu}{\sigma^2} \in \Theta := \mathbb{R} \times (0, \infty)$. Then, for $\widehat{\theta}_n$ being both the method of moments and the maximum likelihood estimator,

$$\widehat{\theta}_n \overset{P_\theta - a.s.}{\longrightarrow} \theta \qquad \forall \theta \in \Theta.$$

Indeed, it follows from the strong law of large numbers that $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i \overset{P_\theta - a.s.}{\longrightarrow} E_\theta X_1 = \mu$ holds for all $\theta \in \Theta$. Likewise, we have that $\frac{1}{n}\sum_{i=1}^n X_i^2 \overset{P_\theta - a.s.}{\longrightarrow} E_\theta X_1^2$ which implies that

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \big(X_i - \bar{X}_n\big)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2 \overset{P_\theta - a.s.}{\longrightarrow} E_\theta\big[X_1^2\big] - \big(E_\theta X_1\big)^2 = \sigma^2.$$

Moreover, since almost sure convergence implies convergence in probability, we also obtain

$$\widehat{\theta}_n \xrightarrow{P_\theta} \theta \qquad \forall \theta \in \Theta.$$

Hence, $\widehat{\theta}_n$ is both weakly and strongly consistent.

The above result for the special case of normally distributed random variables can be generalized. In what follows we provide sufficient conditions for the consistency of method of moments and maximum likelihood estimators.

Theorem 2.2. *Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables, where $X_i \sim P_\theta^{X_1}$, $\theta \in \Theta \subseteq \mathbb{R}^d$. Furthermore, we assume that $E_\theta\big[|X_1|^d\big] < \infty$ $\forall \theta \in \Theta$, which ensures that $\mu_k(\theta) := E_\theta\big[X_1^k\big]$ $(k = 1, \ldots, d)$ are well-defined and finite. We assume that the mapping $g\colon \big(\mu_1(\theta), \ldots, \mu_d(\theta)\big)^T \mapsto \theta$ is continuous and that $\big\{\big(\mu_1(\theta), \ldots, \mu_d(\theta)\big)^T\colon \theta \in \Theta\big\}$ is an open subset of $\mathbb{R}^d$.*

Then the system of equations

$$\mu_k(\theta) \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i^k, \qquad k = 1, \ldots, d \tag{2.4}$$

has with probability tending to one a solution $\widehat{\theta}_{MM,n}$ (otherwise we define $\widehat{\theta}_{MM,n}$ arbitrarily) and

$$\widehat{\theta}_{MM,n} \;\overset{P_{\theta_0}-a.s.}{\longrightarrow}\; \theta_0 \qquad \forall \theta_0 \in \Theta.$$

(θ_0 plays the role of the true parameter.)

Proof. Let $\theta_0 \in \Theta$ be arbitrary. It follows from $E_{\theta_0}\big[|X_1|^d\big] < \infty$ that $E_{\theta_0}\big[|X_1|^k\big] < \infty$ holds for all $k < d$. Therefore, we obtain from the strong law of large numbers that

$$\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n X_i^d \end{pmatrix} \;\overset{P_{\theta_0}-a.s.}{\longrightarrow}\; \begin{pmatrix} \mu_1(\theta_0) \\ \vdots \\ \mu_d(\theta_0) \end{pmatrix}. \tag{2.5}$$

Since $\big(\mu_1(\theta_0), \ldots, \mu_d(\theta_0)\big)^T$ is an inner point of the set $\mathcal{M} = \big\{\big(\mu_1(\theta), \ldots, \mu_d(\theta)\big)^T \colon\, \theta \in \Theta\big\}$ we see that

$$P_{\theta_0}\big((2.4) \text{ has a solution }\big) \;=\; P_{\theta_0}\left(\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n X_i^d \end{pmatrix} \in \mathcal{M}\right) \underset{n\to\infty}{\longrightarrow} 1.$$

Since the mapping g is continuous we conclude from (2.5) that

$$\widehat{\theta}_{MM,n} \;=\; g\left(\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n X_i^d \end{pmatrix}\right) \;\overset{P_{\theta_0}-a.s.}{\longrightarrow}\; g\left(\begin{pmatrix} \mu_1(\theta_0) \\ \vdots \\ \mu_d(\theta_0) \end{pmatrix}\right) \;=\; \theta_0.$$

□

In the previous proof we made use of the explicit representation

$$\widehat{\theta}_{MM,n} \;=\; g\left(\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n X_i^d \end{pmatrix}\right)$$

of the method of moments estimator which led to a short proof of the property of consistency. Likewise, the maximum likelihood estimators based on normally distributed random variables considered in the previous Subsection 2.2 also have an explicit representation. In other cases, however, no closed-form solutions to the maximization problem are known or available, and maximum likelihood estimates can only be found via numerical optimization. Having primarily such cases in mind, we establish a consistency result without making use of a closed-form representation of maximum likelihood estimates. To describe the basic idea of our approach in a transparent way, we consider first a simple special case.

Lemma 2.3. *Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables, where $X_i \sim P_\theta^{X_1}$, for some $\theta \in \Theta := \{\theta_1, \ldots, \theta_K\}$. Furthermore, suppose that $P_{\theta_j}^{X_1} \neq P_{\theta_k}^{X_1}$ if $j \neq k$. Then*

$$P_{\theta_0}\big(\widehat{\theta}_{ML,n} = \theta_0\big) \underset{n\to\infty}{\longrightarrow} 1 \qquad \forall \theta_0 \in \Theta.$$

Proof. Let $p_\theta = \frac{dP_\theta^{X_1}}{d\mu}$ be the density w.r.t. a σ-finite dominating measure μ. (For example, $\mu = P_{\theta_1}^{X_1} + \cdots + P_{\theta_K}^{X_1}$ is a possible choice.) Recall that $\widehat{\theta}_{ML,n}$ maximizes

$$L\big(\theta; X_1, \ldots, X_n\big) = \prod_{i=1}^n p_\theta\big(X_i\big).$$

Let $\theta_0 \in \{\theta_1, \ldots, \theta_K\}$ be arbitrary. We define

$$\Lambda_n(\theta) := \begin{cases} \prod_{i=1}^n \sqrt{p_\theta(X_i)/p_{\theta_0}(X_i)}, & \text{if } \prod_{i=1}^n p_{\theta_0}(X_i) > 0, \\ 1, & \text{if } \prod_{i=1}^n p_{\theta_0}(X_i) = 0. \end{cases}$$

We have that $\Lambda_n\big(\widehat{\theta}_{ML,n}\big) \geq \Lambda_n\big(\theta_0\big) = 1$ which implies that

$$P_{\theta_0}\Big(\widehat{\theta}_{ML,n} = \theta\Big) \leq P_{\theta_0}\Big(\Lambda_n(\theta) \geq 1\Big) \leq E_{\theta_0}\big[\Lambda_n(\theta)\big] \qquad \forall \theta \in \Theta.$$

We show below that

$$E_{\theta_0}\big[\Lambda_n(\theta)\big] \underset{n\to\infty}{\longrightarrow} 0 \qquad \forall \theta \neq \theta_0. \tag{2.6}$$

It follows that

$$P_{\theta_0}\Big(\widehat{\theta}_{ML,n} \neq \theta_0\Big) = \sum_{i:\ \theta_i \neq \theta_0} P_{\theta_0}\Big(\widehat{\theta}_{ML,n} = \theta_i\Big) \leq \sum_{i:\ \theta_i \neq \theta_0} E_{\theta_0}\big[\Lambda_n(\theta_i)\big] \underset{n\to\infty}{\longrightarrow} 0,$$

which completes the proof.

It remains to prove (2.6). Let $\theta \neq \theta_0$. Since $P_{\theta_0}\big(\prod_{i=1}^n p_{\theta_0}(X_i) = 0\big) = 0$ and $X_1, \ldots, X_n$ are independent we obtain that

$$E_{\theta_0}\big[\Lambda_n(\theta)\big] = \Big(E_{\theta_0}\big[\sqrt{p_\theta(X_1)}/\sqrt{p_{\theta_0}(X_1)}\big]\Big)^n = \Big(\int \sqrt{p_\theta(x)}\sqrt{p_{\theta_0}(x)}\, d\mu(x)\Big)^n.$$

Since $P_\theta^{X_1} \neq P_{\theta_0}^{X_1}$ we have that

$$\int \big(\sqrt{p_\theta(x)} - \sqrt{p_{\theta_0}(x)}\big)^2\, d\mu(x) > 0,$$

which implies

$$\begin{aligned} \rho\big(\theta, \theta_0\big) &:= \int \sqrt{p_\theta(x)}\sqrt{p_{\theta_0}(x)}\, d\mu(x) \\ &= \frac{1}{2}\Big\{ \underbrace{\int p_\theta(x)\, d\mu(x)}_{=1} + \underbrace{\int p_{\theta_0}(x)\, d\mu(x)}_{=1} - \int \big(\sqrt{p_\theta(x)} - \sqrt{p_{\theta_0}(x)}\big)^2\, d\mu(x)\Big\} < 1. \end{aligned}$$

Therefore, for $\theta \neq \theta_0$,

$$E_{\theta_0}\big[\Lambda_n(\theta)\big] = \rho(\theta, \theta_0)^n \underset{n\to\infty}{\longrightarrow} 0,$$

i.e. (2.6) is fulfilled. □

Some of the quantities that appeared in the previous proof characterize the closeness of two probability distributions. Here is a formal definition of these notions:

Definition 2.2. Let P_1 and P_2 be probability distributions on a measurable space $\big(\Omega, \mathcal{A}\big)$ which are absolutely continuous w.r.t. a σ-finite measure μ. Denote by $p_i := dP_i/d\mu$ the respective densities $(i = 1, 2)$. Then

$$H\big(P_1, P_2\big) \;:=\; \sqrt{\frac{1}{2}\int_\Omega \big(\sqrt{p_1} - \sqrt{p_2}\big)^2\, d\mu}$$

is the **Hellinger distance** between P_1 and P_2.

$$\rho\big(P_1, P_2\big) \;:=\; \int_\Omega \sqrt{p_1}\sqrt{p_2}\, d\mu$$

is called **Hellinger affinity** between P_1 and P_2.

Remark 2.1. $H\big(P_1, P_2\big)$ and $\rho\big(P_1, P_2\big)$ do not depend on the choice of the dominating measure μ and it holds that

$$\rho\big(P_1, P_2\big) \;=\; 1 \;-\; H^2\big(P_1, P_2\big).$$

The following theorem generalizes the result of Lemma 2.3 and states weak consistency of maximum likelihood estimators.

Theorem 2.4. *Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables, $X_i \sim P_\theta^{X_1}$, where $\theta \in \Theta$ and Θ being a* **compact** *subset of $\mathbb{R}^d$. Suppose that the distributions $P_\theta^{X_1}$ have respective densities p_θ w.r.t. a σ-finite measure μ, and that $\theta \mapsto p_\theta(x)$ is a continuous mapping for all x.*

Then the maximum likelihood estimator $\widehat{\theta}_{ML,n} = \widehat{\theta}_{ML,n}(X_1, \ldots, X_n)$ exists, that is, the likelihood function attains its supremum.

If additionally

(i) $\quad P_\theta^{X_1} \neq P_{\theta'}^{X_1} \quad$ *if $\theta \neq \theta'$,*

(ii) $\quad \omega_\theta(\delta) := \int \sup_{\bar\theta:\ \bar\theta\in\Theta, \|\bar\theta-\theta\|<\delta} \big(\sqrt{p_{\bar\theta}(x)} \,-\, \sqrt{p_\theta(x)}\big)^2\, d\mu(x) \,<\, \infty,$
for some $\delta = \delta(\theta) > 0$,

then $\big(\widehat{\theta}_{ML,n}\big)_{n\in\mathbb{N}}$ is weakly consistent, i.e.

$$\widehat{\theta}_{ML,n} \xrightarrow{P_{\theta_0}} \theta_0 \qquad \forall \theta_0 \in \Theta.$$

Proof. The existence of a maximizer $\widehat{\theta}_{ML,n}$ of the likelihood function follows from continuity of $\theta \mapsto p_\theta(x)$ and compactness of Θ.

Let $\theta_0 \in \Theta$ be arbitrary. We have to show that

$$P_{\theta_0}\Big(\big\|\widehat{\theta}_{ML,n} - \theta_0\big\| \geq \epsilon\Big) \underset{n\to\infty}{\longrightarrow} 0 \qquad \forall \epsilon > 0. \tag{2.7}$$

Let $\epsilon > 0$ be arbitrary. As in the proof of Lemma 2.3, we define

$$\Lambda_n(\theta) := \begin{cases} \prod_{i=1}^n \sqrt{p_\theta(X_i)/p_{\theta_0}(X_i)}, & \text{if } \prod_{i=1}^n p_{\theta_0}(X_i) > 0, \\ 1, & \text{if } \prod_{i=1}^n p_{\theta_0}(X_i) = 0. \end{cases}$$

We show that

$$\sup_{\theta\in\Theta:\ \|\theta-\theta_0\|\geq\epsilon} \Lambda_n(\theta) \xrightarrow{P_{\theta_0}} 0. \tag{2.8}$$

Since $\Lambda_n\big(\widehat{\theta}_{ML,n}\big) \geq \Lambda_n\big(\theta_0\big) \geq 1$, (2.8) implies that

$$P_{\theta_0}\Big(\big\|\widehat{\theta}_{ML,n} - \theta_0\big\| \geq \epsilon\Big) \leq P_{\theta_0}\Big(\sup_{\theta\in\Theta:\ \|\theta-\theta_0\|\geq\epsilon} \Lambda_n(\theta) \geq 1\Big) \underset{n\to\infty}{\longrightarrow} 0,$$

i.e. (2.7) and, therefore, the statement of the theorem is proved.

It remains to show (2.8). Let $\Theta_\epsilon := \Theta \cap \big\{\theta\colon\ \|\theta - \theta_0\| \geq \epsilon\big\}$. In order to show (2.8) we cover Θ_ϵ by a suitable set of open balls $\mathcal{U}_{\delta_1}(\theta_1), \ldots, \mathcal{U}_{\delta_N}(\theta_N)$ ($\mathcal{U}_{\delta_k}(\theta_k) = \big\{\theta \in \mathbb{R}^d\colon\ \|\theta - \theta_k\| < \delta_k\big\}$) such that

$$\sup_{\theta\in\mathcal{U}_{\delta_k}(\theta_k)\cap\Theta} \Lambda_n(\theta) \xrightarrow{P_{\theta_0}} 0, \qquad k = 1, \ldots, N. \tag{2.9}$$

Since

$$\sup_{\theta\in\Theta:\ \|\theta-\theta_0\|\geq\epsilon} \Lambda_n(\theta) \leq \sum_{k=1}^N \sup_{\theta\in\mathcal{U}_{\delta_k}(\theta_k)\cap\Theta} \Lambda_n(\theta)$$

we see that (2.8) follows from (2.9). We still have to prove that there exist open balls $\mathcal{U}_{\delta_1}(\theta_1), \ldots, \mathcal{U}_{\delta_N}(\theta_N)$ such that (2.9) is satisfied.

Construction of the balls $\mathcal{U}_\delta(\theta)$
Let $\theta \in \Theta_\epsilon$ be arbitrary. Then

$$\mathcal{U}_\delta(\theta) := \big\{\theta' \in \mathbb{R}^d\colon\ \|\theta' - \theta\| < \delta\big\}$$

is an open subset in $\mathbb{R}^d$. Let $\bar{\mathcal{U}}_\delta(\theta) = \mathcal{U}_\delta(\theta) \cap \Theta$. Then, with probability 1,

$$\begin{aligned} \sup_{\theta'\in\bar{\mathcal{U}}_\delta(\theta)} \Lambda_n(\theta') &= \sup_{\theta'\in\bar{\mathcal{U}}_\delta(\theta)} \prod_{i=1}^n \sqrt{p_{\theta'}(X_i)}/\sqrt{p_{\theta_0}(X_i)} \\ &\leq \prod_{i=1}^n p_{\theta_0}^{-1/2}(X_i)\Big\{p_\theta^{1/2}(X_i) \quad + \quad \sup_{\theta'\in\bar{\mathcal{U}}_\delta(\theta)} \big|\sqrt{p_{\theta'}(X_i)} - \sqrt{p_\theta(X_i)}\big|\Big\}. \end{aligned}$$

Since $X_1, \ldots, X_n$ are independent we obtain that

$$\begin{aligned} &E_{\theta_0}\Big[\sup_{\theta'\in\bar{\mathcal{U}}_\delta(\theta)} \Lambda_n(\theta')\Big] \\ &\leq \left(\int \sqrt{p_\theta(x)}\sqrt{p_{\theta_0}(x)}\,d\mu(x) \quad + \quad \int \sup_{\theta'\in\bar{\mathcal{U}}_\delta(\theta)} |\sqrt{p_{\theta'}(x)} - \sqrt{p_\theta(x)}|\sqrt{p_{\theta_0}(x)}\,d\mu(x)\right)^n. \end{aligned}$$

We have that

$$\int \sqrt{p_\theta(x)}\sqrt{p_{\theta_0}(x)}\,d\mu(x) \;=\; 1 \;-\; \frac{1}{2}\underbrace{\int \big(\sqrt{p_\theta(x)} - \sqrt{p_{\theta_0}(x)}\big)^2\,d\mu(x)}_{>0 \text{ since } P_\theta^{X_1} \neq P_{\theta_0}^{X_1}} \;=:\; K_{\theta,\theta_0} \;<\; 1.$$

Furthermore, we obtain by the Cauchy-Schwarz inequality

$$\begin{aligned}
&\int \sup_{\theta' \in \bar{\mathcal{U}}_\delta(\theta)} \big|\sqrt{p_{\theta'}(x)} - \sqrt{p_\theta(x)}\big|\sqrt{p_{\theta_0}(x)}\,d\mu(x) \\
&\leq \sqrt{\underbrace{\int \sup_{\theta' \in \bar{\mathcal{U}}_\delta(\theta)} \big(\sqrt{p_{\theta'}(x)} - \sqrt{p_\theta(x)}\big)^2\,d\mu(x)}_{\omega_\delta(\theta) \to_{\delta\to 0} 0 \text{ by dominated convergence}}}\sqrt{\underbrace{\int p_{\theta_0}(x)\,d\mu(x)}_{=1}}.
\end{aligned}$$

For sufficiently small $\delta = \delta(\theta) > 0$ we obtain

$$\underbrace{\int \sqrt{p_\theta(x)}\sqrt{p_{\theta_0}(x)}\,d\mu(x)}_{<1} \quad + \quad \underbrace{\int \sup_{\theta' \in \bar{\mathcal{U}}_\delta(\theta)} |\sqrt{p_{\theta'}(x)} - \sqrt{p_\theta(x)}|\sqrt{p_{\theta_0}(x)}\,d\mu(x)}_{\to 0 \text{ as } \delta \to 0} \;<\; 1$$

and, therefore,

$$\begin{aligned}
&E_{\theta_0}\big[\sup_{\theta' \in \bar{\mathcal{U}}_\delta(\theta)} \Lambda_n(\theta)\big] \qquad (2.10)\\
&\leq \left(\int \sqrt{p_\theta(x)}\sqrt{p_{\theta_0}(x)}\,d\mu(x) \;+\; \int \sup_{\theta' \in \bar{\mathcal{U}}_\delta(\theta)} |\sqrt{p_{\theta'}(x)} - \sqrt{p_\theta(x)}|\sqrt{p_{\theta_0}(x)}\,d\mu(x)\right)^n \underset{n\to\infty}{\longrightarrow} 0.
\end{aligned}$$

(2.10) implies that the neighborhood $\bar{\mathcal{U}}_\delta(\theta)$ about θ is small enough such that

$$\sup_{\theta \in \bar{\mathcal{U}}_\delta(\theta)} \Lambda_n(\theta) \xrightarrow{P_{\theta_0}} 0$$

is satisfied.

The rest of the proof consists of a compactness argument. The open balls $\big(\mathcal{U}_\delta(\theta)\big)_{\theta\in\Theta_\epsilon}$ cover Θ_ϵ. Since Θ is compact and $\big\{\theta \in \mathbb{R}^d\colon\ \|\theta - \theta_0\| \geq \epsilon\big\}$ is a closed subset of $\mathbb{R}^d$, $\Theta_\epsilon = \Theta \cap \big\{\theta\colon\ \|\theta - \theta_0\| \geq \epsilon\big\}$ is also a compact subset of $\mathbb{R}^d$. Therefore, we can choose a **finite** subcover of Θ_ϵ:

$$\mathcal{U}_{\delta_1}(\theta_1), \ldots, \mathcal{U}_{\delta_N}(\theta_N).$$

Since $\Theta_\epsilon \subseteq \Theta$, $\bar{\mathcal{U}}_{\delta_1}(\theta_1), \ldots, \bar{\mathcal{U}}_{\delta_N}(\theta_N)$ cover Θ_ϵ and are chosen such that (2.9) is fulfilled. This completes the proof. □

At the end of this subsection we consider an example where we do not obtain a simple closed-form expression for the maximum likelihood estimator. Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables having a Lebesgue density f_θ, where $f_\theta(x) = \theta g_1(x) + (1-\theta)g_2(x)$ and g_1 and g_2 are two known Lebesgue densities such that $\lambda\big(\{x\colon\ g_1(x) = g_2(x)\}\big) = 0$. The parameter $\theta \in [0,1]$ is unknown and should be estimated by the maximum likelihood method. The following calculations show that the

maximum likelihood estimator $\widehat{\theta}_{ML}$ of θ exists and is unique but a closed-form representation is not available.

We consider the log-likelihood function,

$$\ln L\big(\theta; X_1, \ldots, X_n\big) \;=\; \sum_{i=1}^{n} \ln\big(\theta g_1(X_i) \;+\; (1-\theta) g_2(X_i)\big)$$

and prove that this function has a unique minimizer. To this end, we also consider the derivative of the log-likelihood function,

$$s(\theta) \;:=\; \frac{d}{d\theta} \ln L\big(\theta; X_1, \ldots, X_n\big) \;=\; \sum_{i=1}^{n} \frac{g_1(X_i) - g_2(X_i)}{\theta g_1(X_i) + (1-\theta) g_2(X_i)},$$

and its second derivative,

$$s'(\theta) \;:=\; \frac{d^2}{d\theta^2} \ln L\big(\theta; X_1, \ldots, X_n\big) \;=\; -\sum_{i=1}^{n} \frac{\big(g_1(X_i) - g_2(X_i)\big)^2}{\big(\theta g_1(X_i) + (1-\theta) g_2(X_i)\big)^2}.$$

It turns out that $s'(\theta) < 0$ holds for all $\theta \in (0,1)$ with probability 1, that is, s is strictly monotonically decreasing and has at most one zero. If $s(\theta_0) = 0$ for some $\theta_0 \in (0,1)$, then $\widehat{\theta}_{ML} = \theta_0$. The necessary and sufficient condition that this happens is that $\lim_{\theta \to 0} s(\theta) > 0$ and $\lim_{\theta \to 1} s(\theta) < 0$. If $\lim_{\theta \to 0} s(\theta) \leq 0$, then the log-likelihood function is decreasing and $\widehat{\theta}_{ML} = 0$. And finally, if $\lim_{\theta \to 1} s(\theta) \geq 0$, then the log-likelihood function is increasing and $\widehat{\theta}_{ML} = 1$.

Although a closed-form representation of $\widehat{\theta}_{ML}$ seems to be out of reach, Theorem 2.4 yields that $\widehat{\theta}_{ML}$ is consistent.

2.4 Comparison of estimators – optimality theory

In the previous section we considered two widely applicable approaches to estimating an unknown parameter, the method of moments and the maximum likelihood method. Both methods seem to be motivated on a heuristic level but they are not devised with the primary aim of producing estimators that are in some well-defined sense optimal. In this subsection, we introduce criteria for evaluating and comparing the performance of estimators and discuss how we can find estimators that are optimal w.r.t. these measures.

Let us suppose that we observe realizations $x_1, \ldots, x_n$ of random variables $X_1, \ldots, X_n$, where the distribution of $X = \big(X_1, \ldots, X_n\big)^T$ depends on some unknown parameter $\theta_0 \in \Theta$, i.e. $X \sim P_{\theta_0}$. Suppose further that we are interested in a real-valued quantity $q(\theta_0)$ and that $T = T(X)$ is an estimator of $q(\theta_0)$. A possible measure of error is given by the absolute value of the deviation of $T(X)$ from its target, $\big|T(X) - q(\theta_0)\big|$. However, such a measure is unsatisfactory for two reasons:

(i) It depends on the unknown true value θ_0 of the parameter.

(ii) It is random and therefore cannot be computed even as a function of θ_0 before the experiment is carried out.

A way out of the second difficulty is to consider average measures of error, for example the **mean absolute error** $E_\theta\big|T(X) - q(\theta)\big|$ or the **mean squared error** (MSE) $R(T, \theta)$, given by

$$R\big(T, \theta\big) \;=\; E_\theta\big[(T(X) - q(\theta))^2\big].$$

The mean squared error is usually easier to compute than the mean average error and is therefore the preferred criterion for theoretical investigations. The MSE is determined by the mean and variance of T,

$$R\big(T, \theta\big) \;=\; \mathrm{var}\,\big(T(X)\big) \;+\; b^2(T, \theta),$$

where $b(T, \theta) = E_\theta[T(X)] - q(\theta)$ is the **bias** of T as an estimator of $q(\theta)$.

In order to highlight typical difficulties with the choice among different estimators we consider one of our toy examples. Suppose that we observe realizations of independent random variables $X_1, \ldots, X_n$, where $X_i \sim \mathrm{Bin}(1, \theta)$, and that we are interested in estimating the parameter θ. Without additional information about the true value of θ, the natural choice for the parameter space is $\Theta = [0, 1]$. As shown in Subsection 2.2, $T_1(X) = \bar{X}_n$ is the method of moments estimator of θ. It is at the same time the maximum likelihood estimator; see Exercise 5, Problem sheet 2. We have

$$E_\theta T_1 \;=\; \theta \qquad \forall \theta \in \Theta,$$

that is, T_1 is an **unbiased estimator** of θ. Moreover, its mean squared error is given by

$$R\big(T_1, \theta\big) \;=\; \mathrm{var}_\theta\,\big(T_1\big) \;=\; \frac{\theta(1-\theta)}{n} \qquad \forall \theta \in \Theta.$$

On the other hand, if one conjectures that the true parameter does not exceed some $\bar{\theta} \in (0, 1)$, then a natural candidate for an estimator of θ is given by

$$T_2(X) \;:=\; \begin{cases} \bar{X}_n, & \text{if } \bar{X}_n \le \bar{\theta}, \\ \bar{\theta}, & \text{if } \bar{X}_n > \bar{\theta}. \end{cases}$$

It is easy to see that

$$P_\theta\Big(\big|T_2-\theta\big| \le \big|T_1-\theta\big|\Big) \;=\; 1 \qquad \forall\theta \le \bar\theta$$

and

$$P_\theta\Big(\big|T_2-\theta\big| < \big|T_1-\theta\big|\Big) \;>\; 0 \qquad \forall\theta \in (0,\bar\theta].$$

This implies

$$R\big(T_2,\theta\big) \;\le\; R\big(T_1,\theta\big) \qquad \forall\theta \in [0,\bar\theta] \tag{2.11a}$$

and

$$R\big(T_2,\theta\big) \;<\; R\big(T_1,\theta\big) \qquad \forall\theta \in (0,\bar\theta]. \tag{2.11b}$$

On the other hand, if $\theta = 1$, then $P_1\big(T_1=1\big) = P_1\big(T_2=\bar\theta\big) = 1$, which implies

$$R\big(T_2,1\big) \;>\; R\big(T_1,1\big). \tag{2.12}$$

(2.11b) and (2.12) show the typical picture that one estimator is better than the other in one part of the parameter space whereas the other one is the winner in another part of the space. Comparing the MSE's of T_1 and T_2 means comparing two functions of θ and there is no obvious way to decide which of the two estimators should be used.

On the other hand, if we know for some reason that $\theta \le \bar\theta$, then the space of potentially true parameters is $\bar\Theta = [0,\bar\theta]$ but no longer Θ. On this reduced parameter space, it follows from (2.11a) and (2.11b) that the estimator T_2 is clearly preferable to T_1. These considerations lead to the following definition.

Definition 2.3. Suppose that realizations of $X_1,\ldots,X_n$ are observed, $X = \big(X_1,\ldots,X_n\big)^T \sim P_\theta$, where $\theta \in \Theta$.

(i) An estimator T_1 for $q(\theta)$ is **better** than T_2 if

$$R\big(T_1,\theta\big) \;\le\; R(T_2,\theta\big) \qquad \forall\theta \in \Theta$$

and

$$R\big(T_1,\theta\big) \;<\; R(T_2,\theta\big) \qquad \text{for some } \theta \in \Theta.$$

(ii) An estimator T for $q(\theta)$ is **admissible** if there does not exist a better estimator. Otherwise, T is **inadmissible**.

According to this definition, estimator $T_1 = \bar X_n$ in the binomial example above is inadmissible for $q(\theta) = \theta$ if the corresponding parameter space is equal to $\bar\Theta = [0,\bar\theta]$. If the parameter space is chosen to be $[0,1]$, then neither one of these estimators is better than the other. This does not necessarily mean that these estimators are admissible. It can still be the case that one of them or both can be improved by a third estimator. A possible approach to proving admissibility for a given estimator will be presented below.

A logical consequence of these considerations is the question whether there exists an estimator which improves all others. The answer will be no, except in trivial cases. To see why there does not exist a uniformly best estimator, consider once more the binomial example. Fix any parameter $\theta \in \Theta$ and consider the estimator T such that $T(x) = \theta$ for all possible realizations x of X. Then

$$R\big(T,\theta\big) \;=\; 0.$$

Suppose now that $T^* = T^*(X_1, \ldots, X_n)$ is a uniformly best estimator. We choose $\theta_1, \theta_2 \in (0,1)$, $\theta_1 \neq \theta_2$. Then

$$\begin{aligned}
0 &= R\big(T^*, \theta_1\big) + R\big(T^*, \theta_2\big) \\
&= \sum_{x \in \{0,1\}^n} \big(T^*(x) - \theta_1\big)^2 \theta_1^{\sum_{i=1}^n x_i} (1-\theta_1)^{n - \sum_{i=1}^n x_i} \\
&\quad + \sum_{x \in \{0,1\}^n} \big(T^*(x) - \theta_2\big)^2 \theta_2^{\sum_{i=1}^n x_i} (1-\theta_2)^{n - \sum_{i=1}^n x_i} \\
&\geq \sum_{x \in \{0,1\}^n} \underbrace{\Big\{\big(T^*(x) - \theta_1\big)^2 + \big(T^*(x) - \theta_2\big)^2\Big\}}_{>0} \\
&\qquad\qquad \times \underbrace{\min\Big\{\theta_1^{\sum_{i=1}^n x_i} (1-\theta_1)^{n - \sum_{i=1}^n x_i}, \theta_2^{\sum_{i=1}^n x_i} (1-\theta_2)^{n - \sum_{i=1}^n x_i}\Big\}}_{>0}.
\end{aligned}$$

Since all terms of the sum on the right-hand side of this equation are strictly positive we obtain a contradiction. This shows that a uniformly best estimator does not exist. Taking a look at this short proof reveals that such an ideal estimator will also not exist in other estimation problems, except trivial ones. Whenever two competing distributions P_{θ_1} and P_{θ_2} do not have a disjoint support (This is expressed in our example above by $\sum_{x \in \{0,1\}^n} \min\big\{\theta_1^{\sum_{i=1}^n x_i} (1-\theta_1)^{n - \sum_{i=1}^n x_i}, \theta_2^{\sum_{i=1}^n x_i} (1-\theta_2)^{n - \sum_{i=1}^n x_i}\big\} > 0$.), then we could use the above pattern of proof to show the non-existence of a uniformly best estimator.

A way out is to consider a class of procedures which does not contain foolish estimators, such as $T(X) \equiv \theta$ above, and to look for an estimator that improves all others in this class. A typical choice for such a class is given by that of unbiased estimators. In our binomial example above, this requires that such an estimator $T = T(X)$ has to satisfy

$$E_\theta\big[T(X)\big] \;=\; \sum_{x \in \{0,1\}^n} T(x) \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i} \;=\; \theta \qquad \forall \theta \in [0,1].$$

Using advanced tools we show in Subsection 2.5 that the estimator $\bar{X}_n$ is the (uniformly) **best unbiased estimator** of θ. For the related case that we observe a realization of only one random variable $X \sim \mathrm{Bin}(n, \theta)$, there is a quick proof that the natural estimator $T(X) = X/n$ is the best unbiased estimator for θ. Indeed, let $S = S(X)$ be any arbitrary unbiased estimator. Then

$$\sum_{k=0}^n \Big[\frac{k}{n} - S(k)\Big] \binom{n}{k} \theta^k (1-\theta)^{n-k} \;=\; 0 \qquad \forall \theta \in [0,1],$$

which implies

$$(1-\theta)^n \sum_{k=0}^n \Big[\frac{k}{n} - S(k)\Big] \binom{n}{k} \left(\frac{\theta}{1-\theta}\right)^k \;=\; 0 \qquad \forall \theta \in [0,1).$$

This is equivalent to

$$\sum_{k=0}^n \Big[\frac{k}{n} - S(k)\Big] \binom{n}{k} \rho^k \;=\; 0 \qquad \forall \rho \in [0,\infty),$$

which implies that $S(k) = k/n$ for all $k = 0, \ldots, n$. Therefore, $\bar{X}_n$ is the unique unbiased estimator for θ.

Sometimes an additional restriction is imposed on the class of estimators. In the case of linear regression models, we directed our focus on estimators for the parameter θ which are both unbiased and linear in the observations. Under these restrictions, we were able to identify an estimator which deserved the qualification "uniformly best" (best linear unbiased estimator).

While the restriction to unbiased estimators sometimes paves the way finding an optimal procedure, this approach has also its drawbacks:

(a) Unbiased estimators may not exist; see e.g. Exercise 7, Problem sheet 3.

(b) Even when best unbiased estimators exist, they may be inadmissible.

(c) The property of unbiasedness is not invariant under functional transformations; that is, $\widehat{\theta}$ can be unbiased for θ, but $q(\widehat{\theta})$ biased for $q(\theta)$.

There are ways other than unbiasedness out of the "no best procedure" dilemma. One popular possibility is to reduce the difficult comparison between the **functions** $\theta \mapsto R\big(S,\theta\big)$ and $\theta \mapsto R\big(T,\theta\big)$ to one between **numbers** based on these functions.

(i) We can use a probability measure π and average over θ. That is, if θ is real, we measure the performance of and estimator T for θ by

$$\int_\Theta R\big(T,\theta\big)d\pi(\theta).$$

This approach enables us to compute a best procedure. It corresponds to putting a prior probability measure π on θ. The practical calculation and properties of such **Bayes** procedures are discussed below in more detail.

(ii) We measure the performance of T by the worst that can happen, namely,

$$\sup_{\theta\in\Theta} R\big(T,\theta\big).$$

Best procedures in this sense are called **minimax** and are also discussed below.

2.5 The information inequality

In this section we derive a **universal** lower bound for the variances of all unbiased estimators of a real-valued parameter θ. This lower bound can be used to prove optimality of an unbiased estimator whose variance is always the same as the lower bound. An extended version which also includes biased estimator will be used later to show admissibility of certain estimators.

We begin with a simple special case. Suppose that we observe a discrete random variable X, where $X \sim P_\theta$, $\theta \in \Theta$. We assume that

- $\Omega_X := \big\{X(\omega)\colon\ \omega \in \Omega\big\}$ is finite,
- Θ is an open subset of $\mathbb{R}$,
- $p_\theta(x) := P_\theta\big(\{x\}\big)$ is differentiable in θ for all x. (p_θ is the density of P_θ w.r.t. the counting measure.)

Now we assume that $T = T(X)$ is an arbitrary unbiased estimator of θ, i.e.,

$$E_\theta\big[T(X)\big] \;=\; \sum_{x\in\Omega_X} T(x)p_\theta(x) \;=\; \theta \qquad \forall\theta\in\Theta.$$

Then

$$\begin{aligned}
1 \;&=\; \frac{d}{d\theta}E_\theta\big[T(X)\big] \;=\; \frac{d}{d\theta}\sum_{x\in\Omega_X} T(x)p_\theta(x)\\
&=\; \sum_{x\in\Omega_X} T(x)p'_\theta(x) && \text{(since } \Omega_X \text{ is finite)}\\
&=\; \sum_{x\in\Omega_X} \big(T(x)-\theta\big)p'_\theta(x) && \Big(\text{since } \sum_{x\in\Omega_X} p'_\theta(x) = \underbrace{\frac{d}{d\theta}\sum_{x\in\Omega_X} p_\theta(x)}_{=1\;\forall\theta} = 0\Big)\\
&=\; \sum_{x\in\Omega_X:\;p_\theta(x)\neq 0} \big(T(x)-\theta\big)p'_\theta(x) && \text{(since } p_\theta(x)=0 \text{ implies } p'_\theta=0)\\
&=\; \sum_{x\in\Omega_X} \big(T(x)-\theta\big)l_\theta(x)p_\theta(x),
\end{aligned}$$

where

$$l_\theta(x) \;:=\; \begin{cases} p'_\theta(x)/p_\theta(x), & \text{if } p_\theta(x)\neq 0,\\ 0, & \text{if } p_\theta(x)=0.\end{cases}$$

Therefore it follows from the Cauchy-Schwarz inequality that

$$1 \;=\; E_\theta\big[\big(T(X)-\theta\big)l_\theta(X)\big] \;\le\; \sqrt{E_\theta\big[(T-\theta)^2\big]}\sqrt{E_\theta\big[(l_\theta(X))^2\big]}. \tag{2.13}$$

Since $\sum_x T(x)p'_\theta(x) = 1$ we see that $p'_\theta(x_0)\neq 0$ for some x_0, which also implies $p_\theta(x_0) > 0$. Therefore,

$$I(\theta) \;:=\; E_\theta\big[(l_\theta(X))^2\big] \;=\; \sum_{x\in\Omega_X:\;p_\theta(x)\neq 0} \Big(\frac{p'_\theta(x)}{p_\theta(x)}\Big)^2 p_\theta(x) \;>\; 0.$$

Hence, it follows from (2.13)

$$E_\theta\big[(T-\theta)^2\big] \;\ge\; \frac{1}{I(\theta)} \qquad \forall\theta\in\Theta. \tag{2.14}$$

$I(\theta)$ measures the amount of information that the random variable X carries about the unknown parameter θ and is called **Fisher information**. The right-hand side of (2.14) is a universal lower bound for the mean squared error of unbiased estimators of θ. The number $1/I(\theta)$ is often referred to as the **Cramér-Rao lower bound**. Since priority of discovery is now given to the French mathematician M. Frêchet it is sometimes also called **information inequality**.

Now we come back to our binomial example already considered in the previous Subsection 2.4. Suppose that $X_1,\ldots,X_n$ are independent, where $X_i \sim \mathrm{Bin}(1,\theta)$, $\theta\in\Theta := [0,1]$. Armed with inequality (2.14), we show that $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ is the best unbiased estimator of θ.

We compute first the Fisher information for $X = (X_1, \ldots, X_n)^T$. Let $\theta \in (0,1)$. For $x = (x_1, \ldots, x_n)^T \in \{0,1\}^n$, we have

$$p_\theta(x) \;=\; P_\theta(X = x) \;=\; \theta^k (1-\theta)^{n-k}, \qquad \text{where } k = \sum_{i=1}^n x_i.$$

Since

$$\begin{aligned} \frac{d}{d\theta} p_\theta(x) &= k\theta^{k-1}(1-\theta)^{n-k} \;-\; (n-k)\theta^k (1-\theta)^{n-k} \\ &= \Big(\frac{k}{n} \;-\; \frac{n-k}{1-\theta}\Big)\theta^k (1-\theta)^k \end{aligned}$$

and $p_\theta(x) \neq 0 \;\forall x \in \{0,1\}^n$ we obtain that

$$l_\theta(x) \;=\; \frac{k}{n} \;-\; \frac{n-k}{1-\theta} \;=\; \frac{k - n\theta}{\theta(1-\theta)}$$

and, therefore,

$$I(\theta) \;=\; E_\theta\big[(l_\theta(X))^2\big] \;=\; E_\theta\left[\left(\frac{\sum_{i=1}^n X_i \;-\; n\theta}{\theta(1-\theta)}\right)^2\right] \;=\; \frac{n}{\theta(1-\theta)}.$$

On the other hand, $\bar{X}_n$ is an unbiased estimator of θ and it holds that

$$E_\theta\big[(\bar{X}_n - \theta)^2\big] \;=\; \frac{\theta(1-\theta)}{n} \;=\; \frac{1}{I(\theta)} \qquad \forall \theta \in (0,1).$$

Therefore, $\bar{X}_n$ is the best unbiased estimator of θ for $\theta \in (0,1)$. Since $E_0\big[(\bar{X}_n - 0)^2\big] = E_1\big[(\bar{X}_n - 1)^2\big] = 0$ we conclude that $\bar{X}_n$ has this optimality property also on the complete parameter space $\Theta = [0,1]$.

In what follows we want to generalize the above results to general families of distributions. Here is a first set of regularity conditions which guarantee that the Fisher information of a family of distributions can be defined:

(C1) Suppose that $\{P_\theta\colon\ \theta \in \Theta\}$ is a family of distributions on a measurable space $(\Omega, \mathcal{A})$, where Θ is an open subset of $\mathbb{R}$. Assume that there exists a σ-finite measure μ on $(\Omega, \mathcal{A})$ such that $P_\theta \ll \mu$ holds for all $\theta \in \Theta$. Denote by $p_\theta = dP_\theta/d\mu$ the density (Radon-Nikodym derivative) of P_θ w.r.t. μ. Assume that the function $x \mapsto \frac{d}{d\theta} p_\theta(x)$ is $(\mathcal{A} - \mathcal{B})$-measurable.

Then the **score function** l_θ is defined by

$$l_\theta(x) \;=\; \begin{cases} \frac{\frac{d}{d\theta} p_\theta(x)}{p_\theta(x)}, & \text{if } p_\theta(x) > 0, \\ 0, & \text{if } p_\theta(x) = 0. \end{cases}$$

The number

$$\begin{aligned} I(\theta) &= E_\theta\big[(l_\theta(X))^2\big] \;=\; \int_\Omega (l_\theta(x))^2 \, d\mu(x) \\ &= \int_{\{x\colon\ p_\theta(x) > 0\}} \frac{(\frac{d}{d\theta} p_\theta(x))^2}{p_\theta(x)} \, d\mu(x) \end{aligned}$$

is the **Fisher information** of the family $\{P_\theta\colon\ \theta \in \Theta\}$ at θ.

The next lemma shows that the Fisher information number does not depend on the choice of a dominating measure.

Lemma 2.5. *Suppose that* **(C1)** *is satisfied for two σ-finite measures μ_1 and μ_2, and let $I_1(\theta)$ and $I_2(\theta)$ be the corresponding Fisher information numbers. Then*

$$I_1(\theta) \;=\; I_2(\theta) \qquad \forall \theta \in \Theta.$$

Proof. We consider the Fisher information number $I_{12}(\theta)$ based on the dominating measure $\mu_1 + \mu_2$ and show that $I_i(\theta) = I_{12}(\theta)$, $i = 1, 2$, which proves the statement.

Let $p_{\theta,1}(x) := dP_\theta/d\mu_1$ and $p_{\theta,2}(x) := dP_\theta/d\mu_2$. Since $P_\theta \ll \mu_i \ll \mu_1 + \mu_2$ we obtain that

$$p_{\theta,12}(x) \;:=\; \frac{dP_\theta}{d(\mu_1+\mu_2)}(x) \;=\; \frac{dP_\theta}{d\mu_i}(x)\frac{d\mu_i}{d(\mu_1+\mu_2)}(x) \;=\; p_{\theta,i}(x)\frac{d\mu_i}{d(\mu_1+\mu_2)}(x) \qquad (\mu_1+\mu_2)\text{−a.e.}$$

Therefore,

$$\begin{aligned}
l_{\theta,12}(x) \quad &= \quad \begin{cases} \frac{\frac{d}{d\theta}p_{\theta,12}(x)}{p_{\theta,12}(x)}, & \text{if } p_{\theta,12}(x) > 0, \\ 0, & \text{if } p_{\theta,12}(x) = 0 \end{cases} \\
&= \quad \begin{cases} \frac{\frac{d}{d\theta}p_{\theta,i}(x)}{p_{\theta,i}(x)}, & \text{if } p_{\theta,i}(x)\frac{d\mu_i}{d(\mu_1+\mu_2)}(x) > 0, \\ 0, & \text{if } p_{\theta,i}(x)\frac{d\mu_i}{d(\mu_1+\mu_2)}(x) = 0 \end{cases} \\
&\underbrace{=}_{*} \quad \begin{cases} \frac{\frac{d}{d\theta}p_{\theta,i}(x)}{p_{\theta,i}(x)}, & \text{if } p_{\theta,i}(x) > 0, \\ 0, & \text{if } p_{\theta,i}(x) = 0 \end{cases} \\
&= \quad l_{\theta,i}(x) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \mu_i - \text{a.e.}
\end{aligned}$$

Equality (*) is actually true since

$$\mu_i\Big(\underbrace{\big\{x \in \Omega\colon\; \frac{d\mu_i}{d(\mu_1+\mu_2)}(x) = 0\big\}}_{=:\Omega_{i,0}}\Big) \;=\; \int_{\Omega_{i,0}} \frac{d\mu_i}{d(\mu_1+\mu_2)}(x)\, d(\mu_1+\mu_2)(x) \;=\; 0.$$

Therefore, we obtain

$$\begin{aligned}
I_{12}(\theta) \;&=\; \int \big(l_{\theta,12}(x)\big)^2 p_{\theta,12}(x)\, d(\mu_1+\mu_2)(x) \\
&=\; \int \big(l_{\theta,12}(x)\big)^2 p_{\theta,i}(x)\frac{d\mu_i}{d(\mu_1+\mu_2)}(x)\, d(\mu_1+\mu_2)(x) \\
&=\; \int \big(l_{\theta,12}(x)\big)^2 p_{\theta,i}(x)\, d\mu_i(x) \\
&=\; \int \big(l_{\theta,i}(x)\big)^2 p_{\theta,i}(x)\, d\mu_i(x) \\
&=\; I_i(\theta)
\end{aligned}$$

holds for $i = 1, 2$. □

Before we prove the information inequality in a general context, we collect a few additional conditions which allow in particular that the operations of differentiation w.r.t. θ and taking expectation can be interchanged.

(C2)
- The set $\Omega^0 := \{x \in \Omega\colon\ p_\theta(x) = 0\}$ does **not** depend on θ.
- $I(\theta) < \infty$ for all $\theta \in \Theta$.
- The derivatives $\frac{d}{d\theta} p_\theta(x)$ are continuous in θ for all x.
- The function $q_{\theta,\epsilon}$ defined by

$$q_{\theta,\epsilon}(x) := \begin{cases} \sup_{\bar\theta\colon\ |\bar\theta-\theta|\le\epsilon} \left|\frac{d}{d\theta} p_{\bar\theta}(x) - \frac{d}{d\theta} p_\theta(x)\right| / p_\theta(x), & \text{if } x \in \Omega \setminus \Omega^0, \\ 0 & \text{if } x \in \Omega^0 \end{cases}$$

is $(\mathcal{A} - \mathcal{B})$-measurable and

$$E_\theta\big[(q_{\theta,\epsilon}(X))^2\big] = \int (q_{\theta,\epsilon}(x))^2 p_\theta(x)\, d\mu(x) < \infty, \qquad \text{for some } \epsilon = \epsilon(\theta) > 0.$$

Lemma 2.6. *Suppose that* **(C1)** *and* **(C2)** *are fulfilled. Let $T = T(X)$ be a statistic such that $E_\theta\big[T^2\big] < \infty\ \forall \theta \in \Theta$.*

Then $\theta \mapsto E_\theta T$ is differentiable and

$$\frac{d}{d\theta} E_\theta T = \int_{\Omega\setminus\Omega^0} T(x) \frac{d}{d\theta} p_\theta(x)\, d\mu(x) = E_\theta\big[T(X) l_\theta(X)\big].$$

Proof. First of all, we convince ourselves that the expectation on the right-hand side exists. We have

$$\begin{aligned} E_\theta\big|T(X) l_\theta(X)\big| &= \int \big|T(x)\sqrt{p_\theta(x)}\ \ l_\theta(x)\sqrt{p_\theta(x)}\big|\, d\mu(x) \\ &\le \sqrt{\int T^2(x) p_\theta(x)\, d\mu(x)} \sqrt{\int (l_\theta(x))^2 p_\theta(x)\, d\mu(x)} \\ &= \sqrt{E_\theta\big[T^2\big]}\sqrt{I(\theta)} < \infty. \end{aligned}$$

Therefore, $E_\theta\big[T(X) l_\theta(X)\big]$ exists and is finite.

For $\bar\theta \ne \theta$, we obtain that

$$\begin{aligned} &\frac{E_{\bar\theta} T - E_\theta T}{\bar\theta - \theta} - \int_{\Omega\setminus\Omega^0} T(x) \frac{d}{d\theta} p_\theta(x)\, d\mu(x) \\ &= \int_{\Omega\setminus\Omega^0} T(x) \Big[\frac{p_{\bar\theta}(x) - p_\theta(x)}{\bar\theta - \theta} - \frac{d}{d\theta} p_\theta(x)\Big]\, \mu(x). \end{aligned}$$

Since, by assumption, $\frac{d}{d\theta}p_\theta(x)$ is continuous in θ, the integrand on the right-hand side converges to 0 as $\bar\theta \to \theta$. For $\epsilon = \epsilon(\theta)$, we obtain that

$$\begin{aligned}
&\int_{\Omega\setminus\Omega^0} \sup_{\bar\theta:\ |\bar\theta-\theta|\le\epsilon} \left\{ |T(x)| \Big| \frac{p_{\bar\theta}(x) - p_\theta(x)}{\bar\theta - \theta} \ - \ \frac{d}{d\theta}p_\theta(x) \Big| \right\} \mu(x) \\
&\quad \le \ \int_{\Omega\setminus\Omega^0} |T(x)| \sqrt{p_\theta(x)} \sup_{\bar\theta:\ |\bar\theta-\theta|\le\epsilon} \left\{ \big| \frac{d}{d\theta}p_{\bar\theta}(x) - \frac{d}{d\theta}p_\theta(x) \big| / p_\theta(x) \right\} \sqrt{p_\theta(x)}\, d\mu(x) \\
&\quad \le \ \sqrt{E_\theta\big[T^2\big]} \sqrt{\int_{\Omega\setminus\Omega^0} \big(q_{\theta,\epsilon}(x)\big)^2 p_\theta(x)\, d\mu(x)} \\
&\quad < \ \infty.
\end{aligned}$$

Therefore, it follows from Lebesgue's dominated convergence theorem that

$$\frac{E_{\bar\theta}T - E_\theta T}{\bar\theta - \theta} \ \underset{\bar\theta\to\theta}{\longrightarrow} \int_{\Omega\setminus\Omega^0} T(x) \frac{d}{d\theta}p_\theta(x)\, d\mu(x).$$

□

An immediate implication of Lemma 2.6 is that, under **(C1)** and **(C2)**,

$$E_\theta\big[l_\theta(X)\big] \ = \ 0. \tag{2.15}$$

Indeed, if we choose $T(x) = 1$ for all x, then $E_\theta\big[l_\theta(X)\big] = E_\theta\big[1\, l_\theta(X)\big] = \frac{d}{d\theta}1 = 0$.

The following result of often helpful in calculating the Fisher information in connection with independent random variables.

Proposition 2.7. *Let X_1 and X_2 be random variables which are independent under P_θ, for $\theta \in \Theta$. Suppose that Θ is an open subset of $\mathbb{R}$ and that the families of distributions $\big\{P_\theta^{X_i}\colon\ \theta \in \Theta\big\}$, $i = 1, 2$, satisfy the conditions* **(C1)** *and* **(C2)**. *Denote by $I_1(\theta)$ and $I_2(\theta)$ the respective Fisher information numbers about θ.*

Then the family of distributions $\big\{P_\theta^{(X_1,X_2)}\colon\ \theta \in \Theta\big\}$ satisfies **(C1)** *and* **(C2)** *and the Fisher information about θ contained in $X_1 + X_2$ is equal to $I_1(\theta) + I_2(\theta)$.*

Proof. Suppose that the distributions $P_\theta^{X_1}$ and $P_\theta^{X_2}$ have densities $p_\theta^{(1)}$ and $p_\theta^{(2)}$ w.r.t. σ-finite measures μ_1 and μ_2, respectively, and that these densities satisfy conditions **(C1)** and **(C2)**. Then $P_\theta^{(X_1,X_2)}$ has a density $p_\theta^{(1,2)}$ w.r.t. the product measure $\mu_1 \otimes \mu_2$ such that

$$p_\theta^{(1,2)}(x_1, x_2) \ = \ p_\theta^{(1)}(x_1) p_\theta^{(2)}(x_2) \qquad \forall x_1, x_2.$$

Since

$$\frac{d}{d\theta}p_\theta^{(1,2)}(x_1, x_2) \ = \ \Big(\frac{d}{d\theta}p_\theta^{(1)}(x_1)\Big) p_\theta^{(2)}(x_2) \ + \ p_\theta^{(1)}(x_1) \Big(\frac{d}{d\theta}p_\theta^{(2)}(x_2)\Big),$$

we obtain

$$l_\theta^{(1,2)}(x_1, x_2) \ = \ \begin{cases} \frac{\frac{d}{d\theta}p_\theta^{(1)}(x_1)}{p_\theta^{(1)}(x_1)} \ + \ \frac{\frac{d}{d\theta}p_\theta^{(2)}(x_2)}{p_\theta^{(2)}(x_2)} & \text{if } p_\theta^{(1)}(x_1) p_\theta^{(2)}(x_2) > 0, \\ 0 & \text{if } p_\theta^{(1)}(x_1) p_\theta^{(2)}(x_2) = 0. \end{cases}$$

It follows that

$$l_\theta^{(1)}(x_1) + l_\theta^{(2)}(x_2) = \begin{cases} l_\theta^{(1,2)}(x_1,x_2) & \text{if } p_\theta^{(1)}(x_1)p_\theta^{(2)}(x_2) > 0, \\ l_\theta^{(1)}(x_1) + l_\theta^{(2)}(x_2) & \text{if } p_\theta^{(1)}(x_1)p_\theta^{(2)}(x_2) = 0. \end{cases}$$

Since $P_\theta^{X_i}\big(\{x\colon\ p_\theta^{(i)}(x) = 0\}\big) = 0$ we therefore obtain that

$$l_\theta^{(1,2)}(X_1,X_2) = l_\theta^{(1)}(X_1) + l_\theta^{(2)}(X_2) \qquad P_\theta - a.s.$$

Hence, the Fisher information contained in (X_1, X_2) is given by

$$\begin{aligned} I_{1,2}(\theta) &= E_\theta\big[\big(l_\theta^{(1,2)}(X_1,X_2)\big)^2\big] \\ &= E_\theta\big[\big(l_\theta^{(1)}(X_1) + l_\theta^{(2)}(X_2)\big)^2\big] \\ &= I_1(\theta) + I_2(\theta) + 2E_\theta\big[l_\theta^{(1)}(X_1)l_\theta^{(2)}(X_2)\big]. \end{aligned}$$

Since the families $\big\{P_\theta^{X_i}\colon\ \theta\in\Theta\big\}$, $i = 1, 2$, satisfy conditions **(C1)** and **(C2)** we obtain by (2.15) that

$$E_\theta\big[l_\theta^{(i)}(X_i)\big] = 0, \qquad i = 1, 2,$$

which implies by independence of X_1 and X_2 that

$$E_\theta\big[l_\theta^{(1)}(X_1)l_\theta^{(2)}(X_2)\big] = 0.$$

This completes the proof of the proposition. □

Theorem 2.8. *Suppose that* **(C1)** *and* **(C2)** *are fulfilled and let $T = T(X)$ be an arbitrary estimator of θ such that $E_\theta\big[T^2\big] < \infty$ for all $\theta \in \Theta$.*

(i) If T is unbiased for θ, then

$$E_\theta\big[(T-\theta)^2\big] \geq \frac{1}{I(\theta)} \qquad \forall\theta\in\Theta.$$

(ii) The bias at θ, $b(\theta) = E_\theta T - \theta$, is differentiable in θ and

$$E_\theta\big[(T-\theta)^2\big] \geq \frac{\big(1+\frac{d}{d\theta}b(\theta)\big)^2}{I(\theta)} + \big(b(\theta)\big)^2 \qquad \forall\theta\in\Theta.$$

The inequality in (i) is the classical Cramér-Rao inequality which provides a universal lower bound for the mean squared error of an unbiased estimator of θ. The inequality in (ii) a **not** universal lower risk bound since the right-hand side depends on the estimator $T(X)$ through $b(\theta)$. It can be occasionally used for proving admissibility of a given estimator; see Proposition 2.10 below.

Proof of Theorem 2.8. (i) We have that

$$\begin{aligned} E_\theta\big[(T-\theta)l_\theta(X)\big] &= E_\theta\big[Tl_\theta(X)\big] - \theta \underbrace{E_\theta\big[l_\theta(X)\big]}_{=0} \\ &= \frac{d}{d\theta}E_\theta T = \frac{d}{d\theta}\theta = 1. \end{aligned}$$

Hence, we obtain by the Cauchy-Schwarz inequality

$$1 = E_\theta\big[(T-\theta)l_\theta(X)\big] \le \sqrt{E_\theta\big[(T-\theta)^2\big]}\sqrt{E_\theta\big[\big(l_\theta(X)\big)^2\big]},$$

which implies

$$E_\theta\big[(T-\theta)^2\big] \ge \frac{1}{I(\theta)}.$$

(ii) The mean squared error of T can be decomposed as

$$E_\theta\big[(T-\theta)^2\big] = E_\theta\big[(T-E_\theta T)^2\big] + \big(b(\theta)\big)^2.$$

Since

$$\begin{aligned} E_\theta\big[(T-E_\theta T)l_\theta(X)\big] &= E_\theta\big[Tl_\theta(X)\big] - E_\theta T \underbrace{E_\theta l_\theta(X)}_{=0} \\ &= \frac{d}{d\theta}E_\theta T = \frac{d}{d\theta}\big(b(\theta)+\theta\big) = \frac{d}{d\theta}b(\theta)+1 \end{aligned}$$

we obtain, again by Cauchy-Schwarz, that

$$\big(\frac{d}{d\theta}b(\theta)+1\big)^2 \le E_\theta\big[(T-E_\theta T)^2\big]E_\theta\big[\big(l_\theta(X)\big)^2\big],$$

which completes the proof.

□

Sufficiency

Before we explain what is meant by the word "sufficiency" we consider once more our binomial example: Suppose that realizations $x_1,\ldots,x_n$ of i.i.d. random variables $X_1,\ldots,X_n$ are observed, where $X_i \sim \mathrm{Bin}(1,\theta)$ and $\theta \in \Theta := (0,1)$. Let $X = \big(X_1,\ldots,X_n\big)^T$ and $x = \big(x_1,\ldots,x_n\big)^T$. Then

$$P_\theta\Big(X=x\Big) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \qquad \forall x \in \{0,1\}^n.$$

Now we consider the statistic $T = T(X) := \sum_{i=1}^n X_i$. We have that $T \sim \mathrm{Bin}(n,\theta)$, that is

$$P_\theta\big(T=k\big) = \binom{n}{k}\theta^k(1-\theta)^{n-k} \qquad \forall k = \{0,1,\ldots,n\}.$$

The conditional distribution of X given $T=k$ is given by

$$\begin{aligned} P_\theta\big(X=x \mid T=k\big) &= \frac{P_\theta\big(X=x,\ T=k\big)}{P_\theta\big(T=k\big)} \\ &= \begin{cases} 1/\binom{n}{k} & \text{if } \sum_{i=1}^n x_i = k, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{2.16}$$

Since $P_0(T=0)=P_1(T=n)=1$ we see that the conditional distribution can be chosen such that (2.16) holds in case of $\theta\in\{0,1\}$ as well. Hence, we see that the conditional distribution of X given $T=k$ does **not** depend on the unknown value of the parameter θ. The following considerations indicate that no information about θ is lost by recording only T rather than the original random variable X.

Suppose that $U\sim\text{Uniform}\big([0,1]\big)$ is independent of T under P_θ, for all possible values $\theta\in\Theta$. In what follows we define a random variable $\widetilde{X}=\widetilde{X}(T,U)$ such that

$$P_\theta^{\widetilde{X}} \;=\; P_\theta^X \qquad \forall\theta\in\Theta. \tag{2.17}$$

Recall that

$$P_\theta\big(X=x\mid T=k\big) \;=\; \begin{cases} 1/\binom{n}{k} & \text{if } x\in\Omega_k,\\ 0 & \text{if } x\notin\Omega_k,\end{cases}$$

where $\Omega_k := \big\{x\in\{0,1\}^n\colon\ \sum_{i=1}^n x_i = k\big\}$. We number the elements of Ω_k as $x_{k,1},\dots,x_{k,\binom{n}{k}}$ and choose

$$\widetilde{X} \;=\; \widetilde{X}(T,U) \;:=\; x_{k,l} \qquad \text{if } T=k \text{ and } U\in\left((l-1)/\binom{n}{k}, l/\binom{n}{k}\right].$$

Then

$$P_\theta\big(\widetilde{X}=x\mid T=k\big) \;=\; \begin{cases} 1/\binom{n}{k} & \text{if } x\in\Omega_k,\\ 0 & \text{if } x\notin\Omega_k,\end{cases}$$

and, therefore, (2.17) is fulfilled. These considerations show that, even if only T rather than X is recorded, we can always generate a random variable $\widetilde{X}$ with the same distribution as X. Since this does not require prior knowledge of θ, this shows that no essential information about θ is lost by recording only T. This suggests the following definition.

Definition 2.4. Let $X\sim P_\theta$, where $\theta\in\Theta$. A statistic $T=T(X)$ is **sufficient** for the parameter θ if there exists a version of the conditional distribution $P_\theta\big(X\in\cdot\mid T\in\cdot\big)$ that does not depend on θ.

In the following we show a second consequence of sufficiency: Whenever we have an estimator based on the original random variable X, we can find an at least equally good one which is based on a sufficient statistic $T=T(X)$.
Suppose as above that $X_1,\dots,X_n$ are i.i.d. such that $X_i\sim\text{Bin}(1,\theta)$, where $\theta\in\Theta\subseteq[0,1]$, and that $T=\sum_{i=1}^n X_i$. Let $\widehat{\theta}=\widehat{\theta}(X)$ be an arbitrary estimator of θ such that its mean squared error is finite. We define a new estimator $\check{\theta}=\check{\theta}(T)$ by

$$\check{\theta}(k) \;:=\; E_\theta\big(\widehat{\theta}(X)\mid T=k\big) \;=\; \sum_{x\in\Omega_k}\widehat{\theta}(x)\frac{1}{\binom{n}{k}}.$$

(This is indeed a feasible estimator of θ since the conditional distribution of X given $T=k$, and therefore the definition of $\check{\theta}(k)$ as well, does not depend on θ.)
Now we obtain by Jensen's inequality

$$\begin{aligned} E_\theta\big[(\widehat{\theta}-\theta)^2\big] \;&=\; \sum_{k=0}^n E_\theta\big[\big(\widehat{\theta}(X)-\theta\big)^2\mid T=k\big]P_\theta\big(T=k\big)\\ &\geq\; \sum_{k=0}^n \big(E_\theta\big(\widehat{\theta}-\theta\mid T=k\big)\big)^2 P_\theta\big(T=k\big)\\ &=\; E_\theta\big[(\check{\theta}-\theta)^2\big], \end{aligned}$$

that is, $\check{\theta}$ is not worse than $\widehat{\theta}$.

As a simple, maybe somewhat naive, special case, we consider the estimator $\widehat{\theta} = X_1$. Since

$$\begin{aligned} E_\theta\big(X_1 \mid T = k\big) &= P_\theta\big(X_1 = 1 \mid T = k\big) \\ &= \frac{\#\big\{x \in \Omega_k\colon\ x_1 = 1\big\}}{\#\Omega_k} = \frac{\binom{n-1}{k-1}}{\binom{n}{k}} = \frac{k}{n} \end{aligned}$$

we obtain that $\check{\theta} = \bar{X}_n$. Needless to say that $E_\theta\big[(\check{\theta} - \theta)^2\big] \le E_\theta\big[(\widehat{\theta} - \theta)^2\big]$.

Before we derive a general criterion for a statistic to be sufficient, we consider a second example where we can directly identify a sufficient statistic.

Let $X = \big(X_1, \ldots, X_n\big)^T$, where $X_1, \ldots, X_n$ are i.i.d. real-valued random variables such that $X_i \sim P \in \mathcal{P} := \big\{Q\colon\ Q \text{ is a probability measure on } (\mathbb{R}, \mathcal{B})\big\}$. (Here, the family of possible distributions is a nonparametric one.) Let $X_{n:k}$ denote the kth smallest value among $X_1, \ldots, X_n$. ($X_{n:k}$ is the kth **order statistic**.) Then $X_\uparrow = \big(X_{n:1}, \ldots, X_{n:n}\big)^T$ is a sufficient statistic for $P \in \mathcal{P}$.

To see why this holds true, we will first guess how $P\big(X \in \cdot \mid X_\uparrow = x\big)$ looks like. It seems to be natural to conjecture that the conditional distribution is such that $X_1, \ldots, X_n$ can only attain values from $\big\{X_{n:1}, \ldots, X_{n:n}\big\}$ and that, for reasons of symmetry, every order can appear with the same probability. For formalize this, let $\mathcal{P}_n := \big\{\pi\colon\ \pi \text{ is a permutation of } 1, \ldots, n\big\}$ and, for $x \in \mathbb{R}^n$, $\pi \in \mathcal{P}_n$, $x_\pi := \big(x_{\pi(1)}, \ldots, x_{\pi(n)}\big)^T$. In line with the above discussion, it seems natural to conjecture that

$$P\big(X \in C \mid X_\uparrow = x\big) = \frac{1}{n!} \sum_{\pi \in \mathcal{P}_n} \mathbb{1}_C(x_\pi) \qquad \forall C \in \mathcal{B}^n. \tag{2.18}$$

In order to justify this conjecture, we choose arbitrary Borel sets $C, D \in \mathcal{B}^n$ and prove that

$$P\big(X \in C, X_\uparrow \in D\big) = \int_D \frac{1}{n!} \sum_{\pi \in \mathcal{P}_n} \mathbb{1}_C(x_\pi)\, dP^{X_\uparrow}(x). \tag{2.19}$$

We have that

$$\begin{aligned} P\big(X \in C, X_\uparrow \in D\big) &= \frac{1}{n!} \sum_{\pi \in \mathcal{P}_n} P\big(X_\pi \in C, X_\uparrow \in D\big) \\ &= \frac{1}{n!} \sum_{\pi \in \mathcal{P}_n} E_P\big[\mathbb{1}_C(X_\pi)\ \mathbb{1}_D(X_\uparrow)\big] \\ &= \frac{1}{n!} \sum_{\pi \in \mathcal{P}_n} E_P\big[\mathbb{1}_C((X_\uparrow)_\pi)\ \mathbb{1}_D(X_\uparrow)\big] \\ &= \int_{\mathbb{R}^n} \frac{1}{n!} \sum_{\pi \in \mathcal{P}_n} \mathbb{1}_C(x_\pi)\ \mathbb{1}_D(x)\, dP^{X_\uparrow}(x) \\ &= \int_D \underbrace{\frac{1}{n!} \sum_{\pi \in \mathcal{P}_n} \mathbb{1}_C(x_\pi)}_{= P(X \in C \mid X_\uparrow = x)}\ dP^{X_\uparrow}(x), \end{aligned}$$

which implies (2.18).

In general, finding a sufficient statistic by means of the definition of sufficiency is not convenient since it involves guessing a statistic T that might be sufficient and computing the conditional distribution. Fortunately, a simple, necessary and sufficient criterion for a statistic to be sufficient is available. The results below is often referred to as the **factorization theorem** for sufficient statistics.

Theorem 2.9. *Suppose that $X \sim P_\theta$, where $\theta \in \Theta$. Suppose further that there exists a σ-finite measure μ such that $P_\theta \ll \mu \ \forall \theta \in \Theta$ and denote by p_θ the density of P_θ w.r.t. μ.*

Then a statistic $T = T(X)$ is sufficient for θ if and only if there exist non-negative functions g_θ and h such that

$$p_\theta(x) \ = \ g_\theta(T(x))h(x) \qquad \forall x \in \Omega_X.$$

Proof. We give the proof only in the discrete case. The proof in the general case can be found e.g. in Shao [4, pp. 105-106].

We suppose that X is a discrete random variable with possible values $x_1, \dots, x_N$ or $x_1, x_2, \dots$ ($x_j \neq x_k$ if $j \neq k$) and that p_θ is the density of P_θ w.r.t. the counting measure, that is $p_\theta(x) = P_\theta(X = x)$.

($\Longrightarrow$) Suppose that T is sufficient. Then $P_\theta\big(X = x \mid T = T(x)\big)$ does not depend on θ and we obtain that

$$\begin{aligned} p_\theta(x) \ &= \ P_\theta\big(X = x\big) \ = \ P_\theta\big(X = x, \ T = T(x)\big) \\ &= \ \underbrace{P_\theta\big(T = T(x)\big)}_{=: g_\theta(T(x))} \ \underbrace{P_\theta\big(X = x \mid T = T(x)\big)}_{=: h(x)}. \end{aligned}$$

($\Longleftarrow$) Suppose that

$$p_\theta(x) \ = \ P_\theta\big(X = x\big) \ = \ g_\theta(T(x))h(x) \qquad \forall x \in \Omega_X.$$

Then

$$P_\theta\big(T = t\big) \ = \ \sum_{k:\ T(x_k)=t} p_\theta(x_k) \ = \ \sum_{k:\ T(x_k)=t} g_\theta(T(x_k))h(x_k).$$

If $P_\theta\big(T = t\big) > 0$, then

$$\begin{aligned} P_\theta\big(X = x_j \mid T = t\big) \ &= \ \frac{P_\theta\big(X = x_j, \ T(X) = t\big)}{P_\theta\big(T = t\big)} \\ &= \ \begin{cases} \frac{P_\theta(X=x_j)}{P_\theta(T=t)} & \text{if } T(x_j) = t, \\ 0 & \text{if } T(x_j) \neq t \end{cases} \\ &= \ \begin{cases} \frac{h(x_j)}{\sum_{k:\ T(x_k)=t} h(x_k)} & \text{if } T(x_j) = t, \\ 0 & \text{if } T(x_j) \neq t \end{cases} \end{aligned}$$

If $P_\theta\big(T = t\big) = 0$, then we can define $P_\theta(X = x_j \mid T = t)$ in an arbitrary way, e.g.

$$P_\theta\big(X = x_j \mid T = t\big) \ = \ \begin{cases} \frac{h(x_j)}{\sum_{k:\ T(x_k)=t} h(x_k)} & \text{if } T(x_j) = t, \sum_{k:\ T(x_k)=t} h(x_k) > 0, \\ 0 & \text{otherwise} \end{cases}$$

The latter formula for the conditional distribution is correct in both cases and does not depend on θ. Hence, $T = T(X)$ is a sufficient statistic for θ.

□

Occasionally the (generalized) Cramér-Rao lower bound (see (ii) of Theorem 2.8) can be used for proving admissibility of a given estimator.

Proposition 2.10. *Suppose that* **(C1)** *and* **(C2)** *are fulfilled,* $\Theta = (a,b)$*, where* $-\infty \le a < b \le \infty$*, and* $I(\theta) > 0\ \forall \theta \in \Theta$*. Furthermore, suppose that*

$$\int_a^\theta I(u)\,du \;=\; \int_\theta^b I(u)\,du \;=\; \infty \qquad \forall \theta \in (a,b).$$

If T *is an unbiased estimator of* θ *such that*

$$E_\theta\big[(T-\theta)^2\big] \;=\; \frac{1}{I(\theta)} \qquad \forall \theta \in \Theta, \tag{2.20}$$

then T *is admissible in the class of* **all** *estimators.*

Proof. Suppose that some estimator T^* is better than T, that is,

$$E_\theta\big[\big(T^*-\theta\big)^2\big] \;\le\; E_\theta\big[\big(T-\theta\big)^2\big] \qquad \forall \theta \in (a,b)$$

and

$$E_{\theta_0}\big[\big(T^*-\theta_0\big)^2\big] \;<\; E_{\theta_0}\big[\big(T-\theta_0\big)^2\big] \qquad \text{for some } \theta_0 \in (a,b).$$

Let $b(\theta) = E_\theta T^* - \theta$ be the bias of T^* at θ. The estimator T^* fulfills the conditions of Theorem 2.8 and it follows from (ii) of this theorem that

$$\frac{(1+b'(\theta))^2}{I(\theta)} \;+\; b^2(\theta) \;\le\; E_\theta\big[\big(T^*-\theta\big)^2\big] \;\le\; E_\theta\big[\big(T-\theta\big)^2\big] \;=\; \frac{1}{I(\theta)} \qquad \forall \theta \in (a,b).$$

This implies

$$\frac{2b'(\theta) + \big(b'(\theta)\big)^2}{I(\theta)} \;\le\; -b^2(\theta) \qquad \forall \theta \in (a,b)$$

and therefore

$$2b'(\theta) \;\le\; -b^2(\theta)\; I(\theta) \qquad \forall \theta \in (a,b).$$

Hence, $\theta \mapsto b(\theta)$ is monotonically non-increasing. If $b(\theta) \neq 0$, we obtain in particular that

$$\left(\frac{1}{b(\theta)}\right)' \;=\; -\frac{b'(\theta)}{b^2(\theta)} \;\ge\; \frac{1}{2} I(\theta). \tag{2.21}$$

It follows from (2.20) that T is a best unbiased estimator of θ. Since T^* is assumed to be better there exists some $\theta^* \in (a,b)$ such that $b(\theta^*) \neq 0$. In what follows we show that there exist θ_1, θ_2 such that $\theta_1 < \theta^* < \theta_2$ and $|b(\theta_i)| < |b(\theta^*)|$, $i = 1,2$, which contradicts the monotonicity of $b(\cdot)$.

(i) If $b(\theta) \neq 0\ \forall \theta > \theta^*$, then it follows from

$$\frac{1}{b(\theta)} \;-\; \frac{1}{b(\theta^*)} \;=\; \int_{\theta^*}^\theta \left(\frac{1}{b(u)}\right)' du \;\ge\; \frac{1}{2}\int_{\theta^*}^\theta I(u)\,du \;\underset{\theta\to b}{\longrightarrow}\; \infty,$$

which implies that $b(\theta) \longrightarrow_{\theta\to b} 0$. Hence, there exists some $\theta_2 > \theta$ such that $|b(\theta_2)| < |b(\theta^*)|$.

(ii) If $b(\theta) \neq 0 \ \forall \theta < \theta^*$, then it follows from

$$\frac{1}{b(\theta^*)} - \frac{1}{b(\theta)} = \int_\theta^{\theta^*} \left(\frac{1}{b(u)}\right)' du \geq \frac{1}{2} \int_\theta^{\theta^*} I(u)\, du \underset{\theta \to a}{\longrightarrow} \infty,$$

which implies that $b(\theta) \longrightarrow_{\theta \to a} 0$. Hence, there exists some $\theta_1 < \theta$ such that $|b(\theta_1)| < |b(\theta^*)|$.

This leads, however, to a contradiction:

- If $b(\theta^*) > 0$, then $b(\theta) < b(\theta^*)$ for some $\theta < \theta^*$.
- If $b(\theta^*) < 0$, then there exists some $\theta > \theta^*$ such that $|b(\theta)| < |b(\theta^*)|$, which implies that $b(\theta) > b(\theta^*)$.

In both cases, we obtain a contradiction to the monotonicity of $b(\cdot)$. Hence, our assumption that T is inadmissible is wrong. □

Here is an example of an application of Proposition 2.10:
Suppose that $X_1, \ldots, X_n$ are i.i.d., $X_i \sim N(\theta, \sigma^2)$, where $\sigma^2 > 0$ is fixed and $\theta \in \Theta := \mathbb{R}$. The sample mean $\bar{X}_n$ is an unbiased estimator of θ and it holds that

$$E_\theta\left[\left(\bar{X}_n - \theta\right)^2\right] = \frac{\sigma^2}{n} = \frac{1}{I(\theta)} \qquad \forall \theta \in \Theta.$$

Since

$$\int_{-\infty}^{\theta} I(u)\, du = \int_\theta^{\infty} I(u)\, du = \infty$$

we conclude that $\bar{X}_n$ is admissible in the class of **all** estimators.

2.6 Bayes and minimax estimators

As indicated by the binomial example, apart from trivial estimation problems, a uniformly best estimator of an unknown parameter θ does not exist in general. This is basically because there is only a partial order but not a total order between the risk functions of candidate estimators. A common way out of this dilemma is to consider some "average value" of the risk functions, which leads to the following concept. To simplify matters, we use the mean squared error $R(T,\theta) = E_\theta\big[(T - q(\theta))^2\big]$ of a candidate estimator T for $q(\theta)$ as a measure of performance.

Definition 2.5. Suppose that a realization of $X \sim P_\theta$ is observed, where $\theta \in \Theta$. Let π be a probability measure on $\big(\Theta, \mathcal{A}_\Theta\big)$, where $\mathcal{A}_\Theta$ is a σ-field on the parameter space Θ.

Then, for an estimator $T = T(X)$ of $q(\theta)$,

$$\begin{aligned} r(T,\pi) &= \int_\Theta R(T,\theta)\, d\pi(\theta) \\ &= \int_\Theta \Big[\int \big(T(x) - q(\theta)\big)^2\, dP_\theta(x) \Big]\, d\pi(\theta) \end{aligned}$$

is called the **Bayes risk** of T w.r.t. a **prior distribution** π.

T^* is called **Bayes estimator** of $q(\theta)$, if

$$r\big(T^*,\pi\big) \;=\; \inf\big\{r(T,\pi)\colon\ T \text{ estimator of } q(\theta)\big\}.$$

Remark 2.2. (i) The existence of the integral $\int_\Theta R(T,\theta)\, d\pi(\theta)$ requires measurability of the function $\theta \mapsto R(T,\theta)$ and is therefore not ensured in general. This measurability will be obvious in simple cases, e.g. if Θ is a finite or countably infinite set. Sufficient conditions which ensure measurability will be given by Lemma 2.11 below.

(ii) So far T^* is only implicitly defined, as a minimizer of $r(\,\cdot\,,\pi)$. A simple method of computing a Bayes estimator is applicable in "textbook cases" and will be described below.

One possible interpretation of this approach is the following one. We can think of θ as a **random variable** following a distribution π. The goal is to estimate the **unobserved** realization θ_0 of this random variable, with the help of an observed realization of $X \sim P_{\theta_0}$. The probability measure P_{θ_0} may then be interpreted as the **conditional** distribution of X given $\theta = \theta_0$. The Bayes risk $r(T,\pi)$ is the expected value of $\big(T(X) - q(\theta)\big)^2$ w.r.t. the joint distribution $P^{X,\theta}$ of X and θ, which is defined by

$$P^{X,\theta}(B \times C) \;=\; \int_C P_\theta(B)\, \pi(d\theta).$$

In this sense, $r(T,\pi)$ characterizes the overall performance of T w.r.t. $\theta \sim \pi$.

We will show below that Bayes estimators are admissible under certain conditions. Therefore, this approach makes also sense without the interpretation of θ as a random variable. The Bayes approach is also a method of constructing admissible estimators.

Recall that the Bayes risk of an estimator T for $q(\theta)$ w.r.t. a prior distribution π is defined by $r(T,\pi) = \int_\Theta R(T,\theta)\,d\pi(\theta)$. The following lemma provides sufficient conditions for the measurability of the function $\theta \mapsto R(T,\theta)$ and, therefore, for the existence of the Bayes risk.

Lemma 2.11. *Suppose that $X \sim P_\theta$ is observed, where X takes values in $\mathbb{R}^n$ and $\theta \in \Theta \subseteq \mathbb{R}^d$. Assume that there exists a σ-finite measure μ on $(\mathbb{R}^n, \mathcal{B}^n)$ such that $P_\theta \ll \mu$ for all $\theta \in \Theta$. Assume further that, for $p_\theta := dP_\theta/d\mu$, the function $\theta \mapsto p_\theta(x)$ is continuous for all x. Let $q\colon \mathbb{R}^d \to \mathbb{R}$ be a continuous function and let $T = T(X)$ be an* **arbitrary** *estimator of $q(\theta)$ which takes values in $q(\Theta) = \big\{q(\theta)\colon\ \theta \in \Theta\big\}$.*

(i) If $q(\Theta)$ is bounded, then the risk function $\theta \mapsto R(T,\theta)$ is bounded and continuous, therefore $(\mathcal{B}^d - \mathcal{B})$-measurable.

(ii) If $q(\Theta)$ is unbounded, then the risk function $\theta \mapsto R(T,\theta)$ takes values in $[0,\infty)\cup\{\infty\}$ and is $(\mathcal{B}^d - \bar{\mathcal{B}})$-measurable, where $\bar{\mathcal{B}} = \sigma(\mathcal{B} \cup \{\infty\})$.

Proof. (i) Let $T = T(X)$ be an arbitrary estimator of $q(\theta)$ and let $\theta \in \Theta$ be arbitrary. We have to show that, for any sequence $(\theta_n)_{n\in\mathbb{N}}$ such that $\theta_n \underset{n\to\infty}{\longrightarrow} \theta$,

$$R\big(T,\theta_n\big) \underset{n\to\infty}{\longrightarrow} R\big(T,\theta\big).$$

Suppose that $(\theta_n)_{n\in\mathbb{N}}$ is such a sequence. We have that

$$\Big|R\big(T,\theta_n\big) - R\big(T,\theta\big)\Big| = \Big|\int \big\{\big(T(x)-q(\theta_n)\big)^2 p_{\theta_n}(x) - \big(T(x)-q(\theta)\big)^2 p_\theta(x)\big\}\,d\mu(x)\Big|.$$

The term in curly brackets tends to zero, as $n \to \infty$. This, however, does **not** automatically imply that $\big|\int\{\ldots\}\,d\mu(x)\big| \underset{n\to\infty}{\longrightarrow} 0$. On the other hand, we know from Lebesgues's dominated convergence theorem that the latter result would follow if there were a function $h\colon \Omega_X \to [0,\infty)$ such that $|\{\ldots\}| \le h(x)$ for all x and $\int h(x)\,d\mu(x) < \infty$. Since such a function does not exist in general we have to modify our approach and split up

$$\begin{aligned}\Big|R\big(T,\theta_n\big) - R\big(T,\theta\big)\Big| &\le \int \Big|\big(T(x)-q(\theta_n)\big)^2 - \big(T(x)-q(\theta)\big)^2\Big| p_\theta(x)\,d\mu(x)\\ &\quad + \int \big(T(x)-q(\theta_n)\big)^2 \big|p_{\theta_n}(x) - p_\theta(x)\big|\,d\mu(x)\\ &= I_{n,1} + I_{n,2}. \qquad (2.22)\end{aligned}$$

Since $q(\Theta)$ is bounded a dominating integrable function for the integrand in $I_{n,1}$ is given by $Mp_\theta(x)$, where $M := \sup\{(s-t)^2\colon\ s,t \in q(\Theta)\}$, and we obtain by Lebesgues's dominated convergence theorem that

$$I_{n,1} \underset{n\to\infty}{\longrightarrow} 0. \qquad (2.23)$$

The second term on the right-hand side of (2.22) can be estimated by

$$I_{n,2} \le M \int \big|p_{\theta_n}(x) - p_\theta(x)\big|\,d\mu(x). \qquad (2.24)$$

To complete the proof of (i), it remains to show that

$$\int \big|p_{\theta_n}(x) \,-\, p_\theta(x)\big|\, d\mu(x) \underset{n\to\infty}{\longrightarrow} 0. \tag{2.25}$$

As above, there does not exist a dominating integrable function for the integrand in (2.25). Since $|p_{\theta_n}(x) - p_\theta(x)| = (p_\theta(x) - p_{\theta_n}(x))^+ + (p_{\theta_n}(x) - p_\theta(x))^+$ we obtain that

$$\int \big|p_{\theta_n}(x) - p_\theta(x)\big|\, d\mu(x) \;=\; \int \big(p_\theta(x) - p_{\theta_n}(x)\big)^+ d\mu(x) + \int \big(p_{\theta_n}(x) - p_\theta(x)\big)^+ d\mu(x).$$

Since p_θ is obviously a dominating integrable function for the integrand of the first term on the right-hand side of this formula, we obtain from Lebesgues's dominated convergence theorem

$$\int \big(p_\theta(x) - p_{\theta_n}(x)\big)^+ d\mu(x) \underset{n\to\infty}{\longrightarrow} 0.$$

The second integral has to be treated in a different way. Since

$$\begin{aligned} &\int \big(p_\theta(x) - p_{\theta_n}(x)\big)^+ d\mu(x) \;-\; \int \underbrace{\big(p_{\theta_n}(x) - p_\theta(x)\big)^+}_{=(p_\theta(x) - p_{\theta_n}(x))^-} d\mu(x) \\ &= \int \big(p_\theta(x) - p_{\theta_n}(x)\big)\, d\mu(x) \;=\; 0 \end{aligned}$$

we can conclude, **without resorting to the dominated convergence theorem**, that $\int \big(p_{\theta_n}(x) - p_\theta(x)\big)^+ d\mu(x) \underset{n\to\infty}{\longrightarrow} 0$, which completes the proof of (2.25). (i) follows now from (2.22) to (2.25).

As a continuous real-valued function, $\theta \mapsto R(T,\theta)$ is also $(\mathcal{B}^d - \mathcal{B})$-measurable.

(ii) Now we allow $q(\Theta)$ to be unbounded. It follows from the same considerations as in the proof of (i) that, for any $M < \infty$,

$$\theta \;\mapsto\; R_M\big(T,\theta\big) \;:=\; E_\theta\big[\big(T(X) - \theta\big)^2 \wedge M\big]$$

is continuous and therefore $(\mathcal{B}^d - \mathcal{B})$-measurable, which also yields $(\mathcal{B}^d - \bar{\mathcal{B}})$-measurability. By monotone convergence,

$$R\big(T,\theta\big) \;=\; \lim_{M\to\infty} R_M\big(T,\theta\big),$$

i.e. $R(T,\cdot)$ is the pointwise limit of $(\mathcal{B}^d - \bar{\mathcal{B}})$-measurable functions. Hence, $\theta \mapsto R(T,\theta)$ is also $(\mathcal{B}^d - \bar{\mathcal{B}})$-measurable.

□

As an illustration, we consider once more our binomial example. Suppose that realizations $x_1, \ldots, x_n$ of i.i.d. random variables $X_1, \ldots, X_n$ are observed, where $X_i \sim \mathrm{Bin}(1,\theta)$, $\theta \in \Theta := [0,1]$. As prior distribution, we choose a so-called Beta distribution with parameters $\alpha, \beta > 0$. This distribution has a density $p_{\alpha,\beta}$ w.r.t. the Lebesgue measure, where

$$p_{\alpha,\beta}(t) \;=\; c_{\alpha,\beta}\, t^{\alpha-1}(1-t)^{\beta-1}\, \mathbb{1}_{(0,1)}(t) \qquad \forall t \in \mathbb{R},$$

and the constant $c_{\alpha,\beta}$ is chosen such that $\int_0^1 p_{\alpha,\beta}(t)\,dt = 1$. (The normalization constant $c_{\alpha,\beta}$ may also be expressed by the gamma function, but knowledge of this fact is not necessary here.)
We seek a Bayes estimator T^* of θ w.r.t. the mean squared error.

Solution:
Let $X = \big(X_1, \ldots, X_n\big)^T$. Then $p_\theta(x) := P_\theta\big(X = x\big) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$ and the risk function of an arbitrary estimator $T = T(X)$ is given by

$$R\big(T,\theta\big) \;=\; E_\theta\big[(T(X)-\theta)^2\big] \;=\; \sum_{x\in\{0,1\}^n} \big(T(x)-\theta\big)^2 p_\theta(x).$$

The Bayes risk of T is equal to

$$\begin{aligned} r(T,\theta) \;&=\; \int_0^1 \Big[\sum_{x\in\{0,1\}^n} \big(T(x)-\theta\big)^2 p_\theta(x)\Big] p_{\alpha,\beta}(\theta)\,d\theta \\ &=\; \sum_{x\in\{0,1\}^n} \int_0^1 \big(T(x)-\theta\big)^2 \underbrace{p_\theta(x)\; p_{\alpha,\beta}(\theta)}_{=c_{\alpha,\beta}\theta^{\sum x_i+\alpha-1}(1-\theta)^{n-\sum x_i+\beta-1}} d\theta \qquad (2.26) \\ &=\; c_{\alpha,\beta} \sum_{x\in\{0,1\}^n} \frac{1}{c_{\sum x_i+\alpha,\, n-\sum x_i+\beta}} \underbrace{\int_0^1 \big(T(x)-\theta\big)^2 p_{\sum x_i+\alpha,\, n-\sum x_i+\beta}(\theta)\,d\theta}_{=:h(T(x))}. \end{aligned}$$

To minimize $r(T,\pi)$, we can minimize each of the terms $h(T(x))$ separately. For a random variable Y such that $E\big[Y^2\big] < \infty$, it is well-known that $c \mapsto E\big[(c-Y)^2\big]$ attains its minimum at $c = EY$. With $h(T(x)) = \int_0^1 \big(T(x)-\theta\big)^2 p_{\sum x_i+\alpha,\, n-\sum x_i+\beta}(\theta)\,d\theta$, θ takes the role of the random variable Y, having a Beta distribution with parameters $\sum_i x_i + \alpha$ and $n - \sum_i x_i + \beta$. Therefore, $h(T(x))$ is minimized by the choice

$$T^*(x) \;=\; \int_0^1 \theta p_{\sum x_i+\alpha,\, n-\sum x_i+\beta}(\theta)\,d\theta. \qquad (2.27)$$

Before we compute the integral on the right-hand side of (2.27), we stop for a moment and comment on the calculations in (2.26): The quantities θ and X can be thought of as random variables of a two-stage experiment. First, "nature" chooses a value θ_0 of the random variable $\theta \sim \mathrm{Beta}(\alpha,\beta)$, and afterwards the random variable X is generated with a distribution P_{θ_0} having a probability mass function (i.e. a density w.r.t. the counting measure) p_{θ_0}. With this view, P_{θ_0} is the conditional distribution of X given $\theta = \theta_0$. In the third line of display (2.26), the roles of θ and X are interchanged: $p_{\sum x_i+\alpha,\, n-\sum x_i+\beta}$ in the integral on the right-hand side can be thought of as the density of the conditional distribution of θ given $X = x$. This distribution is called **posterior distribution** since it is an update of the prior distribution after the event $X = x$ has occurred. While π expresses our prior belief about the parameter θ, this posterior distribution reflects our updated belief about θ after a realization x of X was revealed.

Now we compute the integral on the right-hand side of (2.27). Let $a = \sum_{i=1}^n x_i + \alpha$ and $b = n - \sum_{i=1}^n x_i + \beta$. Then

$$\int_0^1 \theta\, p_{a,b}(\theta)\,d\theta \;=\; \int_0^1 \theta\; c_{a,b}\theta^{a-1}(1-\theta)^{b-1}\,d\theta \;=\; \frac{\int_0^1 \theta^a(1-\theta)^{b-1}\,d\theta}{\int_0^1 \theta^{a-1}(1-\theta)^{b-1}\,d\theta}. \qquad (2.28)$$

The integrals on the right-hand side of (2.28) can be represented in terms of the gamma function, however, there does not exist a simple closed-form expression for these integrals. Fortunately, we do not need to compute each of these integrals; rather it suffices to compute their ratio. We take the numerator and apply integration by parts:

$$\begin{aligned}
&\int_0^1 \underbrace{\theta^a}_{f(\theta)} \underbrace{(1-\theta)^{b-1}}_{g'(\theta)}\, d\theta \\
&= \Big[\theta^a \underbrace{\big(-\frac{1}{b}(1-\theta)^b\big)}_{g(\theta)}\Big]_0^1 - \int_0^1 \underbrace{a\theta^{a-1}}_{f'(\theta)} \underbrace{\big(-\frac{1}{b}(1-\theta)^b\big)}_{g(\theta)}\, d\theta \\
&= 0 + \frac{a}{b}\int_0^1 \theta^{a-1}(1-\theta)^b\, d\theta \\
&= \frac{a}{b}\int_0^1 \theta^{a-1}(1-\theta)^{b-1}\, d\theta - \frac{a}{b}\int_0^1 \theta^a(1-\theta)^{b-1}\, d\theta.
\end{aligned}$$

Rearranging terms we obtain that

$$(a+b)\int_0^1 \theta^a(1-\theta)^{b-1}\, d\theta = a\int_0^1 \theta^{a-1}(1-\theta)^{b-1}\, d\theta,$$

which yields

$$\int_0^1 \theta p_{a,b}(\theta)\, d\theta = \frac{a}{a+b}.$$

Therefore,

$$T^*(x) = \frac{\sum_{i=1}^n x_i + \alpha}{n + \alpha + \beta}$$

and the Bayes estimator is given by

$$T^* = \frac{\sum_{i=1}^n X_i + \alpha}{n + \alpha + \beta}.$$

Note that the case of "no prior information about θ" corresponds to $\pi = \text{Uniform}\big([0,1]\big)$, which is achieved with $\alpha = \beta = 1$. In this case, the Bayes estimator is given by $T^* = \frac{\sum_{i=1}^n X_i + 1}{n+2}$.

Here is again a brief sketch of the algorithm to compute the Bayes estimator of a parameter $q(\theta)$: Denote by $P^{X,\theta}$ the joint distribution of X and θ, which is given by $P^{X,\theta}(B \times C) = \int_C P_\theta(B)\, d\pi(\theta)$. For an arbitrary estimator $T = T(X)$, where X takes values in Ω_X, it follows from Fubini's theorem that

$$\begin{aligned}
r\big(T,\theta\big) &= \int_\Theta \Big[\int_{\Omega_X} \big(T(x) - q(\theta)\big)^2\, dP_\theta(x)\Big] d\pi(\theta) \\
&= \int_{\Omega_X \times \Theta} \big(T(x) - q(\theta)\big)^2\, dP^{X,\theta}(x,\theta) \\
&= \int_{\Omega_X} \Big[\underbrace{\int_\Theta \big(T(x) - q(\theta)\big)^2\, dP^{\theta|X=x}(\theta)}_{h(T(x))}\Big]\, dP^X(x). \qquad (2.29)
\end{aligned}$$

To compute $T^*(x)$, only the inner integral on the right-hand of (2.29) has to be minimized and it is not necessary to determine the unconditional distribution P^X of X. In case of the squared error loss $(T - q(\theta))^2$, we obtain that

$$T^*(x) \;=\; \int_\Theta q(\theta)\, dP^{\theta|X=x}(\theta).$$

If we replace the squared error loss by the absolute value loss $|T - q(\theta)|$, then the Bayes estimator $T^* = T^*(X)$ is given by

$$T^*(x) \;=\; \text{median}\big(P^{q(\theta)|X=x}\big).$$

We consider a second example. Suppose that realizations of i.i.d. random variables $X_1, \ldots, X_n$ are observed, where $X_i \sim N(\theta, \sigma^2)$. To simplify matters, we assume that $\sigma^2 > 0$ is fixed. As a measure for the performance of an estimator T of θ, we choose again the mean squared error, $R(T, \theta) = E_\theta\big[(T - \theta)^2\big]$. Let $\pi = N(0, \tau^2)$ be the chosen prior distribution for the parameter θ. We seek a Bayes estimator T^* of θ.

Solution
Let $T = T(X)$ be an arbitrary estimator of θ, where $X = (X_1, \ldots, X_n)^T$. Then its Bayes risk is equal to

$$\begin{aligned} r(T, \pi) &= \int_{\mathbb{R}} \left[\int_{\mathbb{R}^n} \big(T(x) - \theta\big)^2 \, dP_\theta^X(x) \right] d\pi(\theta) \\ &= \int_{\mathbb{R}} \left[\int_{\mathbb{R}^n} \big(T(x) - \theta\big)^2 \, \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\sum_{i=1}^n (x_i - \theta)^2}{2\sigma^2}} \, d\lambda^n(x) \right] \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{\theta^2}{2\tau^2}} \, d\lambda(\theta). \end{aligned}$$

Before we proceed, we combine the exponents of the exponential functions and rearrange the terms:

$$\begin{aligned} -\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta)^2 \;-\; \frac{1}{2\tau^2}\theta^2 &= -\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \bar{x}_n)^2 \;-\; \frac{1}{2\sigma^2} n(\bar{x}_n - \theta)^2 \;-\; \frac{1}{2\tau^2}\theta^2 \\ &= -\frac{1}{2}\left\{ \Big(\frac{n}{\sigma^2} \;+\; \frac{1}{\tau^2}\Big)\theta^2 \;-\; \frac{2n}{\sigma^2}\theta\bar{x}_n \;+\; g(x) \right\} \\ &= -\frac{1}{2}\left\{ \frac{n\tau^2 + \sigma^2}{\sigma^2\tau^2}\theta^2 \;-\; \frac{2n\tau^2}{\sigma^2\tau^2}\theta\bar{x}_n \;+\; g(x) \right\} \\ &= -\frac{1}{2}\left\{ \frac{n\tau^2 + \sigma^2}{\sigma^2\tau^2}\Big(\theta \;-\; \frac{n\tau^2}{n\tau^2 + \sigma^2}\bar{x}_n\Big)^2 \;+\; \widetilde{g}(x) \right\}, \end{aligned}$$

where $g(x)$ ad $\widetilde{g}(x)$ are terms which only depend on x but not on θ. We obtain from Fubini's theorem that

$$\begin{aligned} & r(T, \pi) \\ &= \int_{\mathbb{R}^n} \left[\int_{\mathbb{R}} \big(T(x) - \theta\big)^2 \exp\Big\{ -\frac{n\tau^2 + \sigma^2}{2\sigma^2\tau^2}\Big(\theta \;-\; \frac{n\tau^2}{n\tau^2 + \sigma^2}\bar{x}_n\Big)^2 \Big\} \, d\lambda(\theta) \right] c(x)\, d\lambda^n(x) \qquad (2.30) \\ &= \int_{\mathbb{R}^n} \left[\int_{\mathbb{R}} \big(T(x) - \theta\big)^2 \frac{1}{\sqrt{2\pi\, \frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}}} \exp\Big\{ -\frac{n\tau^2 + \sigma^2}{2\sigma^2\tau^2}\Big(\theta \;-\; \frac{n\tau^2}{n\tau^2 + \sigma^2}\bar{x}_n\Big)^2 \Big\} \, d\lambda(\theta) \right] dP^X(x), \end{aligned}$$

where $c(x)$ is some constant only depending on x. It can be seem from this formula that the posterior distribution of θ given $X = x$ is a normal distribution with location parameter $\frac{n\tau^2}{n\tau^2+\sigma^2}\bar{x}_n$ and variance $\frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}$. Therefore, the Bayes estimate given $X = x$ is

$$T^*(x) \;=\; \frac{n\tau^2}{n\tau^2+\sigma^2}\bar{x}_n.$$

In line with this, the Bayes estimator is given by

$$T^*(X) \;=\; \frac{n\tau^2}{n\tau^2+\sigma^2}\bar{X}_n.$$

Since the value of the inner integral on the right-hand side of (2.30) is equal to $\operatorname{var}\big(\theta \mid X = x\big) = \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}$ we see that the Bayes risk of T^* is equal to

$$r\big(T^*, \pi\big) \;=\; \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}. \tag{2.31}$$

The following theorem shows that the Bayes approach can be used for identifying admissible estimators. As before, we confine ourselves to the case where the performance of an estimator is measured by the mean squared error.

Theorem 2.12. *Suppose that a realization of $X \sim P_\theta$ is observed, where $\theta \in \Theta \subseteq \mathbb{R}^d$. Let $T = T(X)$ be a Bayes estimator for a parameter $q(\theta) \in \mathbb{R}$ w.r.t. a prior distribution π.*

(i) If $\Theta = \big\{\theta_1, \dots, \theta_N\big\}$ or $\Theta = \big\{\theta_1, \theta_2, \dots\big\}$ (i.e. Θ is finite or countably infinite), $\pi(\{\theta\}) > 0\ \forall \theta \in \Theta$, and $r(T, \pi) < \infty$, then T is admissible.

(ii) Assume that there exists a σ-finite measure μ such that $P_\theta \ll \mu$ for all $\theta \in \Theta$. Assume further that, for $p_\theta := dP_\theta/d\mu$, the function $\theta \mapsto p_\theta(x)$ is continuous for all x. Suppose that $\pi(\mathcal{U}_\delta(\theta)) > 0$ for all $\theta \in \Theta$, $\delta > 0$, where $\mathcal{U}_\delta(\theta) = \big\{\theta' \in \Theta\colon \|\theta' - \theta\| < \delta\big\}$. If $q(\Theta) = \{q(\theta)\colon \theta \in \Theta\}$ is a bounded subset of $\mathbb{R}$, $q\colon \Theta \to \mathbb{R}$ is a continuous mapping, and $r(T, \pi) < \infty$, then T is admissible.

Proof. (i) We prove this assertion by contradiction. Suppose that T is inadmissible. Then there exists some estimator T^* which is better than T, that is

$$\begin{aligned} R\big(T^*, \theta_i\big) &\le R\big(T, \theta_i\big) \qquad \text{for all } i \ge 1, \\ R\big(T^*, \theta_{i_0}\big) &< R\big(T, \theta_{i_0}\big) \qquad \text{for some } i_0. \end{aligned}$$

This implies that

$$\begin{aligned} r\big(T, \pi\big) &= \sum_{i\ge1} R\big(T, \theta_i\big)\pi\big(\{\theta_i\}\big) \\ &> \sum_{i\ge1} R\big(T^*, \theta_i\big)\pi\big(\{\theta_i\}\big) \;=\; r\big(T^*, \pi\big). \end{aligned}$$

(The case of "$\infty = \infty$" is excluded since $r(T, \pi) < \infty$.) Hence, we have a contradiction and T is therefore admissible.

(ii) Here we have to take into account that it could be the case that $\pi\big(\{\theta\}\big) = 0$ holds for some $\theta \in \Theta$. We adapt the above method of proof and assume again that T is inadmissible. Then there exists some estimator T^* which is better than T, that is

$$\begin{aligned} R\big(T^*,\theta\big) &\leq R\big(T,\theta\big) \qquad \text{for all } \theta \in \Theta, \\ R\big(T^*,\theta_0\big) &< R\big(T,\theta_0\big) \qquad \text{for some } \theta_0 \in \Theta. \end{aligned}$$

It follows from (i) of Lemma 2.11 that the risk functions $R\big(T,\cdot\big)$ and $R\big(T^*,\cdot\big)$ are continuous, which implies that there exists some sufficiently small $\delta > 0$ such that

$$R\big(T^*,\theta\big) < R\big(T,\theta\big) \qquad \text{for all } \theta \in \mathcal{U}_\delta(\theta_0).$$

Since $\pi\big(\mathcal{U}_\delta(\theta_0)\big) > 0$ we obtain

$$\begin{aligned} r\big(T,\pi\big) &= \int_{\mathcal{U}_\delta(\theta_0)} \underbrace{R\big(T,\theta\big)}_{>R(T^*,\theta)}\, d\pi(\theta) + \int_{\Theta\setminus\mathcal{U}_\delta(\theta_0)} \underbrace{R\big(T,\theta\big)}_{\geq R(T^*,\theta)}\, d\pi(\theta) \\ &> \int_{\mathcal{U}_\delta(\theta_0)} R\big(T^*,\theta\big)\, d\pi(\theta) + \int_{\Theta\setminus\mathcal{U}_\delta(\theta_0)} R\big(T^*,\theta\big)\, d\pi(\theta) \\ &= r\big(T^*,\theta\big). \end{aligned}$$

Hence, we obtain a contradiction to our assumption and T is therefore admissible. □

Instead of averaging the risk function as done with the Bayes approach we can look at the worst possible risk. That is, we prefer an estimator T to T', if and only if,

$$\sup_{\theta\in\Theta} R\big(T,\theta\big) < \sup_{\theta\in\Theta} R\big(T',\theta\big).$$

This leads to the following definition.

Definition 2.6. Suppose that $X \sim P_\theta$ is observed, where $\theta \in \Theta$, and let $q(\theta)$ be the parameter of interest. An estimator T^* of $q(\theta)$ is called **minimax estimator** (it **mini**mizes the **max**imum risk) if

$$\bar{r}\big(T^*,\Theta\big) := \sup_{\theta\in\Theta} R\big(T^*,\theta\big) \leq \inf_T \bar{r}\big(T,\Theta\big) = \inf_T \sup_{\theta\in\Theta} R\big(T,\theta\big).$$

According to this definition, minimax estimators are given in an implicit manner. Unfortunately, there does not exist a general algorithm for the computation of minimax estimators. A proof of the minimax property can sometimes be accomplished with the help of Bayes estimators. The idea is quite simple: If π is any prior distribution for the parameter θ, then

$$r\big(T,\pi\big) = \int_\Theta R\big(T,\theta\big)\, d\pi(\theta) \leq \sup_{\theta\in\Theta} R\big(T,\theta\big) = \bar{r}\big(T,\theta\big),$$

that is, a Bayes risk is always less than or equal to the minimax risk. In case of the opposite relation we can conclude that the corresponding estimator is minimax.

Theorem 2.13. *Suppose that the assumptions of Lemma 2.11 are fulfilled. (This ensures that the risk function of any arbitrary estimator is measurable.) Let $(T_k)_{k\in\mathbb{N}}$ be a sequence of Bayes estimators of $q(\theta)$ w.r.t. respective prior distributions $(\pi_k)_{k\in\mathbb{N}}$. If an estimator T satisfies*

$$\bar{r}\big(T,\Theta\big) \;=\; \sup_{\theta\in\Theta} R\big(T,\theta\big) \;\le\; \limsup_{k\to\infty} \int_\Theta R\big(T_k,\theta\big)\, d\pi_k(\theta),$$

then T is a minimax estimator of $q(\theta)$ in Θ.

Proof. Let T^* be an arbitrary estimator of $q(\theta)$. Then

$$\bar{r}\big(T^*,\Theta\big) \;\ge\; \int_\Theta R\big(T^*,\theta\big)\, d\pi_k(\theta) \;\ge\; \int_\Theta R\big(T_k,\theta\big)\, d\pi_k(\theta).$$

Taking the upper limit on both sides we obtain

$$\bar{r}\big(T^*,\Theta\big) \;\ge\; \limsup_{k\to\infty} \int_\Theta R\big(T_k,\theta\big)\, d\pi_k(\theta) \;\ge\; \bar{r}\big(T,\Theta\big),$$

that is, T minimizes the maximum risk. □

Example
Suppose that realizations of i.i.d. random variables $X_1,\dots,X_n$ are observed, where $X_i \sim N(\theta,\sigma^2)$. $\sigma^2 > 0$ is fixed and $\theta \in \Theta$ is the parameter of interest. As a measure of performance for any estimator we take the mean squared error.

(i) If $\Theta = \mathbb{R}$, then $\bar{X}_n$ is a minimax estimator.

(ii) If $\Theta = [a,b]$, $-\infty < a < b < \infty$, then $\bar{X}_n$ is **not** a minimax estimator.

Proof. (i) First of all, we can actually apply Lemma 2.11 since the conditions imposed there are satisfied. (For example, $X = \big(X_1,\dots,X_n\big)^T$ follows a multivariate normal distribution with a density p_θ w.r.t. λ^n and $\theta \mapsto p_\theta(x)$ is a continuous function for all x.)

To prove (i), we consider a sequence of Bayes estimators $(T_k)_{k\in\mathbb{N}}$ w.r.t. $(\pi_k)_{k\in\mathbb{N}}$, $\pi_k = N(0,k)$. According to (2.31), the corresponding Bayes risks are given by

$$r\big(T_k,\pi_k\big) \;=\; \frac{\sigma^2 k}{nk+\sigma^2}.$$

Since

$$r\big(T_k,\pi_k\big) \underset{k\to\infty}{\longrightarrow} \frac{\sigma^2}{n} \;=\; \sup_{\theta\in\mathbb{R}} R\big(\bar{X}_n,\theta\big)$$

it follows from Theorem 2.13 that $\bar{X}_n$ is minimax in $\Theta = \mathbb{R}$.

(ii) If $\Theta = [a,b]$, then $\bar{X}_n$ is inadmissible and can be improved by the estimator T_n given by

$$T_n \;=\; \begin{cases} \bar{X}_n & \text{if } \bar{X}_n \in [a,b], \\ a & \text{if } \bar{X}_n < a, \\ b & \text{if } \bar{X}_n > b. \end{cases}$$

For $\theta \in \Theta$, we obtain

$$P_\theta\big((T_n - \theta)^2 \leq (\bar{X}_n - \theta)^2\big) \;=\; 1$$

and

$$P_\theta\big((T_n - \theta)^2 < (\bar{X}_n - \theta)^2\big) \;>\; 0,$$

which implies that

$$R\big(T_n, \theta\big) \;<\; R\big(\bar{X}_n, \theta\big) \qquad \forall \theta \in \Theta. \tag{2.32}$$

Furthermore, it follows from (i) of Lemma 2.11 that $\theta \mapsto R(T_n, \theta)$ is continuous. Therefore, there exists some $\theta_0 \in \Theta$ such that

$$\sup_{\theta \in \Theta} R\big(T_n, \theta\big) \;=\; R\big(T_n, \theta_0\big).$$

It follows from (2.32) that

$$\bar{r}\big(T_n, \Theta\big) \;=\; R\big(T_n, \theta_0\big) \;<\; R\big(\bar{X}_n, \theta_0\big) \;=\; \frac{\sigma^2}{n} \;=\; \bar{r}\big(\bar{X}_n, \Theta\big),$$

that is, $\bar{X}_n$ is **not** minimax in Θ.

□

3 Testing statistical hypotheses

There are many questions in the sciences, in industry, and in life generally that require a definite answer. For example, does a new (pharmaceutical) drug help? Or, is one type of car safer than another? Does a lot of manufactured items contain an excessive number of defectives? We begin with a simple but common example. Suppose that a pharmaceutical company has developed a new drug and involved scientists believe that this drug increases the rate of recovery from some disease over the recovery rate when an established treatment (a well-tested drug or even no treatment at all) is applied. Suppose that it is known from past experience that a fixed proportion $\theta_0 = 0.2$ recover from the disease with the established treatment. In view of possible side effects of the new drug, but also in order to avoid any sort of costs after the drug is introduced into the market, people involved in the decision about the introduction into the market want to make sure that the new drug increases the chance of recovery. To this end, a random experiment has to be performed. Most simply, one would select n patients, administer the new drug, and then base the decision on the observed rate of recovery.

3.1 The elements of hypothesis testing

We use the simple drug example to develop the framework of the classical hypothesis testing theory. Let θ denote the rate of recovery when the new drug is given. When the n patients that are given the drug are selected from a large pool of patients, then the random number X of recoveries follows (approximately) a binomial distribution with parameters n and θ. For reasons explained below we choose as our hypothesis **(null hypothesis)** H_0 that the new drug has no effect or even that the new drug has either no or a negative effect on the recovery rate. This corresponds to $\theta \in \Theta_0$, where $\Theta_0 = \{\theta_0\}$ (the drug never harms) or $\Theta = [0, \theta_0]$, respectively. The **alternative** H_1 ("positive effect") is described by $\theta \in \Theta_1$, where $\Theta_1 = (\theta_0, 1]$. On the basis of our observed number x of recoveries among the n randomly selected patients who have been administered the drug we are to decide whether to **accept** H_0 and state that the true value of θ is in Θ_0 or to **reject** H_0 (i.e. accept H_1) and state that the true value of θ is in Θ_1. We can distinguish between two structural possibilities for Θ_i, $i = 1, 2$. If Θ_0 consists of one point only we call the null hypothesis H_0 **simple**. Otherwise, if Θ_0 consists of more than one point we call H_0 **composite**. The same convention applies to Θ_1.

To end up with a decision between H_0 and H_1 we need a rule for action. If Ω_X denotes the set of possible values of the random variable X, then we have to determine how we decide if the event $X = x$ occurs, for all $x \in \Omega_X$. Such a rule is conveniently described by a function $\varphi\colon\ \Omega_X \to \{0, 1\}$, where $\varphi(x) = 1$ means that we reject H_0 if the event $X = x$ occurs. $\varphi(x) = 0$ describes that we do not reject H_0 in case of $X = x$. Such a function $\varphi\colon\ \Omega_X \to \{0, 1\}$ is called **test** of H_0 versus H_1. In our drug example, it seems reasonable to reject H_0 if the observed value of X exceeds or equals some natural number k, and accept H_0 otherwise, which leads to a test function (or test) φ_k given by

$$\varphi_k(x) = \begin{cases} 1 & \text{if } x \geq k, \\ 0 & \text{if } x < k. \end{cases}$$

In this simple case, X is called **test statistic** since it is constructed for the purpose of testing whether or not H_0 is true. The value k which completes the specification of our test is referred to as the **critical value** of the test.

In more involved cases we cannot describe our decision rule by reference to test statistics or critical values. This can then be done by describing all sample points x for which we reject H_0. Suppose that a test φ is given by

$$\varphi(x) \;=\; \begin{cases} 1 & \text{if } x \in \Omega_{X,1}, \\ 0 & \text{if } x \in \Omega_{X,0}. \end{cases}$$

$\Omega_{X,1}$ is then called the **critical** or **rejection region** of the test φ whereas $\Omega_{X,0}$ is called the **acceptance region**. In the above example, we have that $\Omega_{X,1} = \{k, k+1, \ldots, n\}$ and $\Omega_{X,0} = \{0, 1, \ldots, k-1\}$.

The only reasonable measure of performance of a test is given by the probabilities that we make correct judgments when we use it. There are two types of error we can commit: we can reject the null hypothesis, when we should have accepted; or we can accept the null hypothesis, when we should have rejected. The first of these errors is a **type I error** and the second one a **type II error**. In the drug example, a type I error is committed if H_0 is rejected, when in fact $\theta = \theta_0$ (or $\theta \leq \theta_0$, respectively), and a type II error is committed if H_0 is accepted, when in fact $\theta > \theta_0$. We consider the case of a simple null hypothesis, that is $\Theta_0 = \{\theta_0\}$. Then the probability of a type I error is given by

$$P_{\theta_0}\big(\varphi_k(X) = 1\big) \;=\; P_{\theta_0}\big(X \in \Omega_{X,1}\big) \;=\; P_{\theta_0}\big(X \geq k\big) \;=\; \sum_{j=k}^{n} \binom{n}{j} \theta_0^j (1-\theta_0)^{n-j}.$$

The probability of a type II error is harder to deal with since it depends not only on the test but also on the particular alternative being considered. For the test φ_k and an alternative $\theta > \theta_0$, this probability is given by

$$P_{\theta}\big(\varphi_k(X) = 0\big) \;=\; P_{\theta}\big(X \in \Omega_{X,0}\big) \;=\; P_{\theta}\big(X \geq k\big) \;=\; \sum_{j=0}^{k-1} \binom{n}{j} \theta^j (1-\theta)^{n-j}.$$

This error probability is closely connected to the power of a test, which is defined as follows.

Definition 3.1. The **power** of a test against the alternative θ is the probability of rejecting H_0 when θ is true.

Thus, the power is one minus the probability of a type II error. It can be thought of as the probability that the test will "detect" that the alternative θ holds. The power is a function of θ on Θ_1. If Θ_0 is composite as well, then the probability of a type I error is also a function of θ. Both the power and the probability of a type I error are contained in the **power function** which is defined **for all** $\theta \in \Theta := \Theta_0 \cup \Theta_1$ by

$$\beta(\theta, \varphi) \;=\; P_\theta\big(\text{ "Rejection" }\big) \;=\; P_\theta\big(\varphi(X) = 1\big).$$

In the drug example above, we have

$$\beta(\theta, \varphi_k) \;=\; \sum_{j=k}^{n} \binom{n}{j} \theta^j (1-\theta)^{n-j}.$$

For a given test problem, say $H_0\colon\ \theta \in \Theta_0$ vs. $H_1\colon\ \theta \in \Theta_1$, there are typically many tests that seem to be reasonable. In fact, any function $\varphi\colon\ \Omega_X \to \{0, 1\}$ can serve as a

test. It is natural to choose a test on the basis of its performance which is characterized by the corresponding probabilities of type I and type II errors. We have seen in Section 2.4 of these Lecture Notes that a uniformly best estimator does not exist, except in trivial cases. Using analogous arguments as in the case of parameter estimation we can easily see that an ideal test also does not exist. We consider again the drug example, that is we test

$$H_0\colon \quad \theta = \theta_0 \qquad \text{vs.} \qquad H_1\colon \quad \theta > \theta_0,$$

where $\theta_0 = 0.2$. Information about θ is provided by a random variable $X \sim \mathrm{Bin}(n, \theta)$. A test $\varphi\colon \{0, 1, \ldots, n\} \to \{0, 1\}$ has a probability of a type I error 0 if and only if $\varphi(x) = 0$ for all $x \in \{0, 1, \ldots, n\}$. (This follows from the fact that $P_{\theta_0}\big(X = x\big) > 0$ $\forall x \in \{0, 1, \ldots, n\}$.) On the other hand, the power of this test is 0, for all $\theta > \theta_0$, whereas a test $\bar{\varphi}$ such that $\bar{\varphi}(x) = 1$ $\forall x \in \{0, 1, \ldots, n\}$ has maximum power 1. This explains why the search for a "uniformly best" test is pointless.

The Russian born American mathematician Jerzy Neyman (1894-1981) and the British statistician Egon Sharpe Pearson (1895-1980) proposed a possible way out of this dilemma. They remarked there there is often an asymmetry between the two types of error which grows out of a corresponding asymmetry between hypothesis and alternative. In the drug example with a simple hypothesis H_0 (i.e. $\Theta_0 = \{\theta_0\}$), the alternative H_1 is composite and there are alternatives in Θ_1 that are arbitrarily close to θ_0. Rejection of H_0 has the clear meaning that the drug works, but acceptance could well occur because an alternative practically indistinguishable from θ_0 holds. More importantly, committing a type I error means that a useless new treatment, possibly with unknown side effects, will be introduced in the market, while a reliable and well-tested treatment will be abandoned. In this case it becomes clear that it is more important to avoid the commitment of a type I error, which leads to the Neyman-Pearson proposal.

We begin by specifying a (usually small) number $\alpha > 0$ and restrict our attention to tests which in fact have the probability of rejection less than or equal to α for all $\theta \in \Theta_0$. Such tests are said to have **level (of significance)** α and we speak of rejecting H_0 at level α. The values $\alpha = 0.01$ and 0.05 are commonly used in practice. Since a test of level α is also of level $\alpha' > \alpha$ it is customary to give a name to the smallest level of significance of a test. This quantity is called the **size** of a test, and is evidently the maximum probability of a type I error.

In case of the drug example, one would choose as null hypothesis H_0 that the drug does **not** work, that is $\Theta_0 = \{\theta_0\}$ or even $\Theta_0 = [0, \theta_0]$. Of course, persons involved in the development of this drug hope that the hypothesis is rejected, which offers the opportunity to announce that the drug leads to **significant** improvements. By using a small level α one controls the probability of wrongly rejecting the hypothesis. Thus, if $\alpha = 0.05$, one can be 95% sure about not issuing a false claim, when the improvement is announced. If H_0 is not rejected, this does not necessarily mean that the new drug is useless. In such a case, usually nothing is announced and one goes on to an additional experiment.

Having restricted attention to tests of level α, Neyman and Pearson then propose that we select within this class on the basis of the power against the alternatives we are interested in. If, as is sometimes possible, there is a test which has maximal power among all α tests against all alternatives, it is chosen. Otherwise subsidiary criteria are brought in. We pursue this matter in the next subsection.

3.2 Optimal tests

In the previous subsection we introduced the framework of testing hypotheses. In particular, we introduced a test statistic on an ad hoc basis in order to construct a "reasonable" level α test, for a prescribed value of $\alpha > 0$. Now we want to derive tests which deserve the term "optimal". We assume that a realization x of a random variable $X \sim P_\theta$ whith possible values in a set Ω_X is observed, where $\theta \in \Theta$ is the unknown parameter, and that we want to decide which of the following statements is true:

$$H_0\colon\ \theta \in \Theta_0 \qquad \text{or} \qquad H_1\colon\ \theta \in \Theta_1,$$

where $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$. Any rule of action can be conveniently described by a function $\varphi\colon\ \Omega_X \to \{0,1\}$, where

$$\varphi(x) \;=\; \begin{cases} 1, & \text{if } X = x \text{ implies rejection of } H_0, \\ 0, & \text{if } X = x \text{ implies acceptance of } H_0. \end{cases}$$

Recall that the power function $\beta(\cdot,\varphi)$ of a test φ is defined by

$$\beta\big(\theta,\varphi\big) \;=\; P_\theta\big(\varphi(X) = 1\big).$$

If $\theta \in \Theta_0$, then $\beta\big(\theta,\varphi\big)$ describes the probability of a type I error (under P_θ). Alternatively, if $\theta \in \Theta_1$, then $\beta\big(\theta,\varphi\big)$ describes the power of φ against the alternative θ, i.e., one minus the probability of a type II error (under P_θ). As indicated in the previous subsection, we follow the proposal by Neyman and Pearson, fix a level of significance α and try to find a best test within the class of level α tests.

Definition 3.2. A test φ^* is a **uniformly most powerful** (UMP) level α test for $H_0\colon\ \theta \in \Theta_0$ versus $H_1\colon\ \theta \in \Theta_1$ if

$$\beta\big(\theta,\varphi^*\big) \;\le\; \alpha \qquad \forall \theta \in \Theta_0 \tag{3.1a}$$

and

$$\beta\big(\theta,\varphi^*\big) \;\ge\; \beta\big(\theta,\varphi\big) \qquad \forall \theta \in \Theta_1 \tag{3.1b}$$

holds for all level α tests φ.

To simplify matters, we confine ourselves first to the case of testing a simple hypothesis ($\Theta_0 = \{\theta_0\}$) versus a simple alternative ($\Theta_1 = \{\theta_1\}$). If X is a discrete random variable with values in a finite or countably infinite set Ω_X, then the probability of a type I error for a test φ is given by

$$P_{\theta_0}\big(\varphi(X) = 1\big) \;=\; \beta\big(\theta_0,\varphi\big) \;=\; \sum_{x \in \Omega_X\colon\, \varphi(x)=1} P_{\theta_0}\big(X = x\big).$$

On the other hand, the probability of a type II error is equal to

$$P_{\theta_1}\big(\varphi(X) = 0\big) \;=\; 1 \,-\, \beta\big(\theta_1,\varphi\big) \;=\; \sum_{x \in \Omega_X\colon\, \varphi(x)=0} P_{\theta_1}\big(X = x\big).$$

In view of this, it seems advisable to reject H_0 in case of $X = x$ if $P_{\theta_0}\big(X = x\big)$ is small while $P_{\theta_1}\big(X = x\big)$ is large. The following lemma shows that an optimal test can be obtained on the basis of the ratio of the densities under P_{θ_1} and P_{θ_0}.

Lemma 3.1. *(Neyman-Pearson lemma)*
Suppose that $X \sim P_\theta$, where $\theta \in \{\theta_0, \theta_1\}$, and that P_{θ_0} and P_{θ_1} have respective densities p_{θ_0} and p_{θ_1} w.r.t. some σ-finite measure μ. (For example, $\mu = P_{\theta_0} + P_{\theta_1}$ does the job.) For the problem of testing

$$H_0\colon\ \theta = \theta_0 \qquad \textit{vs.} \qquad H_1\colon\ \theta = \theta_1$$

a test φ may have the form

$$\varphi(x) \;=\; \begin{cases} 1, & \textit{if } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} > c, \\ 0, & \textit{if } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} < c, \end{cases} \tag{3.2}$$

where $c \in [0, \infty)$ is a constant. (If $p_{\theta_0}(x) = 0$ then the ratio $p_{\theta_1}(x)/p_{\theta_0}(x)$ takes on the value ∞ when $p_{\theta_1}(x) > 0$; and, by convention, equals 0 when $p_{\theta_1}(x) = 0$.)

If φ^* is another test such that

$$P_{\theta_0}\big(\varphi^*(x) = 1\big) \;\le\; P_{\theta_0}\big(\varphi(x) = 1\big),$$

then

$$P_{\theta_1}\big(\varphi^*(x) = 0\big) \;\ge\; P_{\theta_1}\big(\varphi(x) = 0\big),$$

that is, φ is a most powerful test for the significance level given by the size of φ, $\bar{\alpha} := P_{\theta_0}\big(\varphi(x) = 1\big)$.

Proof. Let φ^* be an arbitrary test such that

$$P_{\theta_0}\big(\varphi^*(x) = 1\big) \;\le\; P_{\theta_0}\big(\varphi(x) = 1\big).$$

We show first that

$$\big(\varphi(x) - \varphi^*(x)\big)\, c\, p_{\theta_0}(x) \;\le\; \big(\varphi(x) - \varphi^*(x)\big)\, p_{\theta_1}(x) \qquad \forall x \in \Omega_X. \tag{3.3}$$

Indeed, if $p_{\theta_0}(x) > 0$, we have that

a) $p_{\theta_1}(x)/p_{\theta_0}(x) > c$ implies $p_{\theta_1}(x) > c\, p_{\theta_0}(x)$ and $\big(\underbrace{\varphi(x)}_{=1} - \varphi^*(x)\big) \ge 0$,

b) $p_{\theta_1}(x)/p_{\theta_0}(x) < c$ implies $p_{\theta_1}(x) < c\, p_{\theta_0}(x)$ and $\big(\underbrace{\varphi(x)}_{=0} - \varphi^*(x)\big) \le 0$,

c) $p_{\theta_1}(x)/p_{\theta_0}(x) = c$ implies $p_{\theta_1}(x) = c\, p_{\theta_0}(x)$.

In these three cases (3.3) is satisfied.
If $p_{\theta_0}(x) = 0$, then (3.3) follows immediately if $p_{\theta_1}(x) = 0$. Otherwise, if $p_{\theta_1}(x) > 0$, then $p_{\theta_1}(x)/p_{\theta_0}(x) = \infty$, which implies that $\big(\underbrace{\varphi(x)}_{=1} - \varphi^*(x)\big) \ge 0$. Hence, (3.3) is again satisfied.

It follows from $P_{\theta_0}\big(\varphi^*(x) = 1\big) \le P_{\theta_0}\big(\varphi(x) = 1\big)$ that

$$\begin{aligned} 0 \;&\le\; c \cdot \big\{P_{\theta_0}\big(\varphi(x) = 1\big) \;-\; P_{\theta_0}\big(\varphi^*(x) = 1\big)\big\} \\ &=\; \int_{\Omega_X} \big(\varphi(x) - \varphi^*(x)\big)\, c\, p_{\theta_0}(x)\, d\mu(x) \\ &\le\; \int_{\Omega_X} \big(\varphi(x) - \varphi^*(x)\big)\, p_{\theta_1}(x)\, d\mu(x) \qquad \text{(by (3.3))} \\ &=\; P_{\theta_1}\big(\varphi(x) = 1\big) \;-\; P_{\theta_1}\big(\varphi^*(x) = 1\big), \end{aligned}$$

which completes the proof. □

In certain cases, Lemma 3.1 can be used to derive a most powerful level α test. We consider two examples.

Example 1
Suppose that we observe realizations of i.i.d. random variables $X_1, \ldots, X_n$, where $X_i \sim \mathcal{N}(\theta, \sigma^2)$. To simplify matters, we assume that $\theta \in \{\theta_0, \theta_1\}$ and that $\sigma^2 > 0$ is known. We want to test

$$H_0\colon \quad \theta = \theta_0 \qquad \text{vs.} \qquad H_1\colon \quad \theta = \theta_1,$$

where $\alpha \in (0,1)$ is the chosen level of significance. The following calculations show how Lemma 3.1 can be used for determining a most powerful test.

Solution
The random vector $X := \big(X_1, \ldots, X_n\big)^T$ has possible densities p_{θ_j} w.r.t. the Lebesgue measure λ^n $(j = 0, 1)$, where

$$p_{\theta_j}(x) \;=\; \Big(\frac{1}{\sqrt{2\pi}\,\sigma}\Big)^n \;\exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n \big(x_i - \theta_j\big)^2 \Big\}.$$

Let, w.l.o.g., $\theta_0 < \theta_1$. In line with Lemma 3.1, we confine our attention to tests of the form

$$\varphi(x) \;=\; \begin{cases} 1, & \text{if } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} > c, \\ 0, & \text{if } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} < c, \end{cases} \tag{3.4a}$$

and try to find some $c \in [0, \infty)$ such that

$$P_{\theta_0}\big(\varphi(X) \;=\; 1\big) \;=\; \alpha. \tag{3.4b}$$

If we achieve this goal, then it follows from Lemma 3.1 that φ is a most powerful level α test.

Before we proceed, we perform an auxiliary calculation. It holds that

$$\begin{aligned} \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} \;&=\; \exp\Big\{ \frac{1}{2\sigma^2} \sum_{i=1}^n \Big[\underbrace{\big(x_i - \theta_1 + \theta_1 - \theta_0\big)^2 \;-\; \big(x_i - \theta_1\big)^2}_{= 2(x_i - \theta_1)(\theta_1 - \theta_0) + (\theta_1 - \theta_0)^2} \Big] \Big\} \\ &=\; \exp\Big\{ \frac{1}{2\sigma^2} \Big[\sum_{i=1}^n 2\big(x_i - \theta_1\big)\big(\theta_1 - \theta_0\big) \;+\; n\big(\theta_1 - \theta_0\big)^2 \Big] \Big\} \\ &=\; \exp\Big\{ \frac{1}{2\sigma^2} \Big[2n\big(\bar{x}_n - \theta_1\big)\big(\theta_1 - \theta_0\big) \;+\; n\big(\theta_1 - \theta_0\big)^2 \Big] \Big\}. \end{aligned}$$

Hence, the function $x \mapsto p_{\theta_1}(x)/p_{\theta_0}(x)$ is strictly monotonically increasing in $\bar{x}_n$. Therefore, (3.4a) can be rewritten as

$$\varphi(x) \;=\; \begin{cases} 1, & \text{if } \bar{x}_n > \widetilde{c}, \\ 0, & \text{if } \bar{x}_n < \widetilde{c}, \end{cases}$$

for some $\widetilde{c}$. Since, under the null hypothesis H_0, $\bar{X}_n \sim \mathcal{N}\big(\theta_0, \sigma^2/n\big)$ we represent the test φ in the equivalent form

$$\varphi(x) \;=\; \begin{cases} 1, & \text{if } \frac{\bar{x}_n - \theta_0}{\sigma/\sqrt{n}} > \bar{c}, \\ 0, & \text{if } \frac{\bar{x}_n - \theta_0}{\sigma/\sqrt{n}} < \bar{c}. \end{cases} \tag{3.5}$$

(3.5) allows us to determine the critical value of the test appropriately. We have

$$P_{\theta_0}\big(\varphi(X) = 1\big) = P_{\theta_0}\Big(\frac{\bar{X}_n - \theta_0}{\sigma/\sqrt{n}} \geq \bar{c}\Big) = 1 - \Phi(\bar{c}) = \alpha$$

if and only if $\bar{c} = \Phi^{-1}(1-\alpha)$. ($\Phi$ denotes the distribution function of a $\mathcal{N}(0,1)$ distribution and its inverse, Φ^{-1}, is the corresponding **quantile function**.) Therefore,

$$\varphi(x) = \begin{cases} 1, & \text{if } \frac{\bar{x}_n - \theta_0}{\sigma/\sqrt{n}} \geq \Phi^{-1}(1-\alpha), \\ 0, & \text{if } \frac{\bar{x}_n - \theta_0}{\sigma/\sqrt{n}} < \Phi^{-1}(1-\alpha). \end{cases}$$

is a most powerful level α test. Unfortunately, there does not exist a closed-form expression for the values of Φ^{-1}. However, tables of these functions can be found in many textbooks. Examples for typical values of this function are $\Phi^{-1}(0.95) \approx 1.64$, $\Phi^{-1}(0.975) \approx 1.96$, and $\Phi^{-1}(0.99) \approx 2.33$.

The next example reveals limitations of our approach to obtain most powerful tests.

Example 2
Suppose that a realization of a random variable $X \sim \mathrm{Bin}(n, \theta)$ is observed. For simplicity we constrain ourselves again to the case that the parameter θ can only take on two possible values and we want to test

$$H_0\colon \quad \theta = \theta_0 \qquad \text{vs.} \qquad H_1\colon \quad \theta = \theta_1.$$

We assume that $0 < \theta_0 < \theta_1 < 1$ and we want to find a most powerful level α test, where $\alpha \in (0,1)$. We denote by p_{θ_0} and p_{θ_1} the respective densities of X w.r.t. the counting measure, where $p_{\theta_i}(k) = P_{\theta_i}(X = k) = \binom{n}{k}\theta_i^k(1-\theta_i)^{n-k}$, for $k = 0, 1, \ldots, n$, $i = 1, 2$. Guided by Lemma 3.1, we consider tests $\varphi\colon \{0, 1, \ldots, n\} \to \{0, 1\}$ of the form

$$\varphi(k) = \begin{cases} 1, & \text{if } \frac{p_{\theta_1}(k)}{p_{\theta_0}(k)} > c, \\ 0, & \text{if } \frac{p_{\theta_1}(k)}{p_{\theta_0}(k)} < c, \end{cases} \tag{3.6a}$$

and try to find some $c \in [0, \infty)$ such that

$$P_{\theta_0}\big(\varphi(X) = 1\big) = \alpha. \tag{3.6b}$$

Since

$$\frac{p_{\theta_1}(k)}{p_{\theta_0}(k)} = \frac{\binom{n}{k}\theta_1^k(1-\theta_1)^{n-k}}{\binom{n}{k}\theta_0^k(1-\theta_0)^{n-k}} = \Big(\underbrace{\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}}_{>1}\Big)^k \Big(\frac{1-\theta_1}{1-\theta_0}\Big)^n$$

we see that the mapping $k \mapsto p_{\theta_1}(k)/p_{\theta_0}(k)$ is strictly monotonically increasing. Therefore, (3.6a) can be equivalently rewritten as

$$\varphi_{\bar{c}}(k) = \begin{cases} 1, & \text{if } k \geq \bar{c}, \\ 0, & \text{if } k < \bar{c}. \end{cases}$$

To obtain a level α test, the critical value $\bar{c}$ must be chosen such that

$$P_{\theta_0}\big(\varphi_{\bar{c}}(X) = 1\big) = P_{\theta_0}\big(X \geq \bar{c}\big) = \sum_{k=\bar{c}}^{n} \binom{n}{k} \theta_0^k (1-\theta_0)^{n-k} \leq \alpha,$$

that is, $\bar{c}$ must be large enough. If we succeed to find some $\bar{c}$ such that the size of the corresponding test $\varphi_{\bar{c}}$ is equal to α, then it follows from Lemma 3.1 that this test is a most powerful level α test. Otherwise, taking into account that the probability of a type II error is equal to

$$P_{\theta_1}\big(\varphi_{\bar{c}}(X)=0\big) \;=\; P_{\theta_1}\big(X<\bar{c}\big) \;=\; \sum_{k=0}^{\bar{c}-1}\binom{n}{k}\theta_1^k(1-\theta_1)^{n-k},$$

we could choose $\bar{c}$ as the smallest integer such that the size of $\varphi_{\bar{c}}$ does not exceed α, i.e.

$$\bar{c} \;=\; \min\Big\{c\colon\; \sum_{k=c}^{n}\binom{n}{k}\theta_0^k(1-\theta_0)^{n-k}\Big\}.$$

This test is the best level α test within the class of tests of the form (3.6a), however, it is not clear if it is also the most powerful level α test among all tests. In fact, any function $\varphi\colon\; \{0,1,\ldots,n\}\to\{0,1\}$ can serve as a test for H_0 versus H_1. In order to find the most powerful level α test we could consider all possible tests, sort those out which have a size greater than α, and then choose that one among the remaining tests which has maximum power against θ_1. In fact, we have to consider 2^{n+1} possible tests, which is clearly not feasible unless n is very small. A way out of this dilemma is described in what follows.

Recall that in Example 1, the equality in (3.4b) can always be achieved by a suitable choice of the critical value c in (3.4a). This is, however, not true in general. In Example 2, for instance, there are only $n+2$ different tests of the form (3.6a). Hence, it could well happen that we want to set a level of significance α such that

$$P_{\theta_0}\big(X\geq k\big) \;\neq\; \alpha \qquad \forall k\in\{0,1,\ldots,n+1\}.$$

In this case, the Neyman-Pearson lemma (Lemma 3.1) does not help us to find a most powerful level α test. In such cases, we may consider so-called **randomized tests**, which are introduced next.

Definition 3.3. Suppose that we observe a realization of a random variable $X\sim P_\theta$ with values in Ω_X and that we want to test

$$H_0\colon\; \theta\in\Theta_0 \qquad \text{vs.} \qquad H_1\colon\; \theta\in\Theta_1.$$

A **randomized test** for the test problem H_0 vs. H_1 is a function $\varphi\colon\; \Omega_X\to[0,1]$, where $\varphi(x)$ is the (conditional) probability of rejecting H_0 if $X=x$. If $\varphi(x)\in\{0,1\}$ for all $x\in\Omega_X$, then the test φ is called **nonrandomized**.

To carry out a randomized test, we need an additional random experiment (to flip a coin, ...). In case of simple hypotheses, the probability of a type I error is given by

$$P_{\theta_0}\big(H_0 \text{ is rejected}\big) \;=\; \int_{\Omega_X} \underbrace{P_{\theta_0}\big(H_0 \text{ is rejected} \mid X=x\big)}_{=\varphi(x)}\, dP_{\theta_0}(x) \;=\; E_{\theta_0}\big[\varphi(X)\big]$$

and that of a type II error by

$$P_{\theta_1}\big(H_0 \text{ is accepted}\big) \;=\; \int_{\Omega_X} \underbrace{P_{\theta_1}\big(H_0 \text{ is accepted} \mid X=x\big)}_{=1-\varphi(x)}\, dP_{\theta_0}(x) \;=\; E_{\theta_1}\big[1-\varphi(X)\big].$$

Accordingly, the power function $\beta(\cdot, \varphi)$ is given by

$$\beta\big(\theta, \varphi\big) \;=\; E_\theta\big[\varphi(X)\big] \qquad \forall \theta \in \Theta_0 \cup \Theta_1.$$

The next theorem shows that, for testing a simple hypothesis versus a simple alternative, there always exists a most powerful test for a given level of significance.

Theorem 3.2. *Suppose that $X \sim P_\theta$, $\theta \in \{\theta_0, \theta_1\}$, and that P_{θ_0} and P_{θ_1} have respective densities p_{θ_0} and p_{θ_1} w.r.t. some σ-finite measure μ. Let the test problem be given by*

$$H_0\colon\ \theta = \theta_0 \qquad \textit{vs.} \qquad H_1\colon\ \theta = \theta_1.$$

(i) For each $\alpha > 0$, there exists a most powerful level α test φ_α, which is given by

$$\varphi_\alpha(x) \;=\; \begin{cases} 1, & \textit{if } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} > c_\alpha, \\ \gamma_\alpha, & \textit{if } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} = c_\alpha, \\ 0, & \textit{if } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} < c_\alpha, \end{cases} \tag{3.7a}$$

where $c_\alpha \in [0, \infty)$ and $\gamma_\alpha \in [0, 1]$ are chosen such that

$$E_{\theta_0}\big[\varphi_\alpha(X)\big] \;=\; \alpha. \tag{3.7b}$$

(ii) If φ^* is a most powerful level α test, then

$$\varphi^*(x) \;=\; \begin{cases} 1, & \textit{if } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} > c_\alpha, \\ 0, & \textit{if } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} < c_\alpha, \end{cases}$$

with a possible exception on a set of zero μ-measure, i.e.

$$\mu\Big(\big\{x \in \Omega_X\colon\ \varphi^*(x) \neq \varphi_\alpha(x) \ \ \textit{and} \ \ p_{\theta_1}(x)/p_{\theta_0}(x) \neq c_\alpha\big\}\Big) \;=\; 0.$$

Proof. (i) Let $\alpha > 0$ be arbitrary. For any test of the form (3.7a) we have

$$E_{\theta_0}\big[\varphi_\alpha(X)\big] \;=\; P_{\theta_0}\big(\big\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x) > c_\alpha\big\}\big) \;+\; \gamma_\alpha\, P_{\theta_0}\big(\big\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x) = c_\alpha\big\}\big).$$

We have to find $c_\alpha \in [0, \infty)$ and $\gamma_\alpha \in [0, 1]$ such that the right-hand side of this equation is equal to α. To this end, we consider a function $g\colon [0, \infty) \to [0, 1]$ defined by

$$g(c) \;:=\; P_{\theta_0}\big(\big\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x) \geq c\big\}\big).$$

This function has the following properties:

- g is monotonically non-increasing,
- $g(0) = P_{\theta_0}\big(\big\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x) \geq 0\big\}\big) = 1$,

– by continuity from above,

$$\begin{aligned}\lim_{c\to\infty} g(c) &= \lim_{c\to\infty} P_{\theta_0}\big(\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x)\geq c\}\big)\\ &= P_{\theta_0}\Big(\underbrace{\bigcap_{c>0}\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x)\geq c\}}_{=\{x\colon\ p_{\theta_0}(x)=0\}}\Big) = 0.\end{aligned}$$

(Note that probability densities should take on values in $[0,\infty)$.)

– g is left-continuous. It follows again from continuity from above that

$$\begin{aligned}g\big(c_0-0\big) &= \lim_{c\nearrow c_0,\, c<c_0} P_{\theta_0}\big(\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x)\geq c\}\big)\\ &= P_{\theta_0}\Big(\underbrace{\bigcap_{c\colon\ c<c_o}\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x)\geq c\}}_{=\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x)\geq c_0\}}\Big) = g\big(c_0\big).\end{aligned}$$

Let $c_\alpha := \sup\big\{c\colon\ g(c)\geq\alpha\big\}$. Then, by left-continuity of g,

$$P_{\theta_0}\big(\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x)\geq c_\alpha\}\big) = g\big(c_\alpha\big) = \lim_{c\nearrow c_\alpha,\, c<c_\alpha}\underbrace{g(c)}_{\geq\alpha}\geq\alpha$$

and

$$g\big(c_\alpha+0\big) = \lim_{c\searrow c_\alpha,\, c>c_\alpha}\underbrace{g(c)}_{<\alpha}\leq\alpha.$$

On the other hand, by continuity from below,

$$\begin{aligned}g\big(c_\alpha+0\big) &= \lim_{c\searrow c_\alpha,\, c>c_\alpha} P_{\theta_0}\big(\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x)\geq c\}\big)\\ &= P_{\theta_0}\Big(\underbrace{\bigcup_{c\colon\ c>c_\alpha}\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x)\geq c\}}_{=\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x)>c_\alpha\}}\Big)\\ &= P_{\theta_0}\big(\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x)>c_\alpha\}\big).\end{aligned}$$

If $P_{\theta_0}\big(\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x)>c_\alpha\}\big)=\alpha$, then choose $\gamma_\alpha=0$.
If $P_{\theta_0}\big(\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x)>c_\alpha\}\big)<\alpha$, then

$$P_{\theta_0}\big(\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x)=c_\alpha\}\big) > 0$$

and

$$\begin{aligned}\gamma_\alpha &:= \frac{\alpha - P_{\theta_0}\big(\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x)>c_\alpha\}\big)}{P_{\theta_0}\big(\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x)=c_\alpha\}\big)}\\ &= \frac{\alpha - P_{\theta_0}\big(\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x)>c_\alpha\}\big)}{P_{\theta_0}\big(\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x)\geq c_\alpha\}\big) - P_{\theta_0}\big(\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x)>c_\alpha\}\big)} \in [0,1].\end{aligned}$$

In both cases, we obtain that

$$E_{\theta_0}\big[\varphi_\alpha(X)\big] = P_{\theta_0}\big(\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x)>c_\alpha\}\big) + \gamma_\alpha\, P_{\theta_0}\big(\{x\colon\ p_{\theta_1}(x)/p_{\theta_0}(x)=c_\alpha\}\big) = \alpha,$$

as required.

The proof that φ_α is a most powerful level α test follows the same lines as the proof of Lemma 3.1 above.

(ii) Let φ^* be an arbitrary test such that

$$P_{\theta_0}\big(\varphi^*(x)=1\big) \;\le\; P_{\theta_0}\big(\varphi_\alpha(x)=1\big).$$

We can show analogously to (3.3) that

$$\big(\varphi_\alpha(x)-\varphi^*(x)\big)\,c_\alpha\,p_{\theta_0}(x) \;\le\; \big(\varphi_\alpha(x)-\varphi^*(x)\big)\,p_{\theta_1}(x) \qquad \forall x\in\Omega_X. \tag{3.8}$$

If

$$\mu\Big(\big\{x\in\Omega_X\colon\ \varphi^*(x)\neq\varphi_\alpha(x)\ \text{ and }\ p_{\theta_1}(x)/p_{\theta_0}(x)\neq c_\alpha\big\}\Big) \;>\; 0,$$

then

$$\mu\Big(\big\{x\colon\ \big(\varphi_\alpha(x)-\varphi^*(x)\big)\,c_\alpha\,p_{\theta_0}(x) \;<\; \big(\varphi_\alpha(x)-\varphi^*(x)\big)\,p_{\theta_1}(x)\big\}\Big) \;>\; 0,$$

which yields in conjunction with (3.8) that

$$\begin{aligned}
0 \;&\le\; c_\alpha\cdot\big\{P_{\theta_0}\big(\varphi_\alpha(x)=1\big) \;-\; P_{\theta_0}\big(\varphi^*(x)=1\big)\big\}\\
&=\; \int_{\Omega_X}\big(\varphi_\alpha(x)-\varphi^*(x)\big)\,c_\alpha\,p_{\theta_0}(x)\,d\mu(x)\\
&<\; \int_{\Omega_X}\big(\varphi_\alpha(x)-\varphi^*(x)\big)\,p_{\theta_1}(x)\,d\mu(x)\\
&=\; P_{\theta_1}\big(\varphi_\alpha(x)=1\big) \;-\; P_{\theta_1}\big(\varphi^*(x)=1\big),
\end{aligned}$$

i.e. φ^* has less power against θ_1 than φ_α.

□

Example 2 (continued)
To obtain a most powerful level α test, we choose c_α as the maximum value from $\{0,1,\ldots,n+1\}$ such that

$$P_{\theta_0}\big(X\ge c_\alpha\big) \;\ge\; \alpha.$$

Then

$$P_{\theta_0}\big(X>c_\alpha\big) \;=\; P_{\theta_0}\big(X\ge c_\alpha+1\big) \;<\; \alpha.$$

With $\gamma_\alpha := \big(\alpha-P_{\theta_0}\big(X>c_\alpha\big)\big)/P_{\theta_0}\big(X=c_\alpha\big)$ we obtain that

$$P_{\theta_0}\big(X>c_\alpha\big) \;+\; \gamma_\alpha\,P_{\theta_0}\big(X=c_\alpha\big) \;=\; \alpha.$$

Hence, φ_α defined by

$$\varphi_\alpha(k) \;=\; \begin{cases} 1, & \text{if } k>c_\alpha,\\ \gamma_\alpha, & \text{if } k=c_\alpha,\\ 0, & \text{if } k<c_\alpha \end{cases}$$

satisfies $E_{\theta_0}\varphi_\alpha(X) \;=\; \alpha$, i.e. the size of this test equals α. Since φ_α has Neyman-Pearson structure it is a most powerful level α test.

Tests of composite hypotheses

Our restriction to simple hypotheses guaranteed that we were able to find best (most powerful) tests for any given level of significance $\alpha > 0$; see Theorem 3.2 above. On the other hand, cases where we know in advance that an unknown parameter can only attain two possible values are rather rare and it is certainly desirable to extend our results to the more realistic case of composite hypotheses. Recall that a test φ^* is a **uniformly most powerful** (UMP) level α test for $H_0\colon\ \theta \in \Theta_0$ versus $H_1\colon\ \theta \in \Theta_1$ if

$$\beta\big(\theta, \varphi^*\big) \ \leq \ \alpha \qquad \forall \theta \in \Theta_0$$

and

$$\beta\big(\theta, \varphi^*\big) \ \geq \ \beta\big(\theta, \varphi\big) \qquad \forall \theta \in \Theta_1$$

holds for all level α tests φ. In the following we consider again the problem of testing hypotheses about the location parameter of normally distributed random variables and we investigate under which circumstances a UMP test can be obtained.

Suppose that realizations $x_1, \ldots, x_n$ of i.i.d. random variables $X_1, \ldots, X_n$ are observed, where $X_i \sim \mathcal{N}(\theta, \sigma^2)$ and σ^2 is known. Recall that we found for the problem of testing a simple hypothesis versus a simple alternative,

$$H_0\colon\ \theta = \theta_0 \qquad \text{vs.} \qquad H_1\colon\ \theta = \theta_1$$

with $\theta_0 < \theta_1$, a most powerful level α test φ_α, where

$$\varphi_\alpha(x) \ = \ \begin{cases} 1, & \text{if } \ \frac{\bar{x}_n - \theta_0}{\sigma/\sqrt{n}} \geq \Phi^{-1}(1-\alpha), \\ 0, & \text{if } \ \frac{\bar{x}_n - \theta_0}{\sigma/\sqrt{n}} < \Phi^{-1}(1-\alpha). \end{cases}$$

Now we consider the test problem

$$H_0'\colon\ \theta \leq \theta_0 \qquad \text{vs.} \qquad H_1'\colon\ \theta > \theta_0.$$

Since the corresponding sets of parameters, $\Theta_0 = (-\infty, \theta_0]$ and $\Theta_1 = (\theta_0, \infty)$ contain both more than one element, we are faced with a composite hypothesis and a composite alternative. In this particular case it turns out that the test φ_α, which was originally derived as a most powerful test for H_0 versus H_1, is actually an optimal test for H_0' versus H_1'. To see this, we proceed in two steps.

1) The test φ_α depends on θ_0 and the fact that $\theta_1 > \theta_0$, however, it does **not** depend on the particular value of θ_1. Therefore, we obtain that

$$E_{\theta_1}\big[\varphi_\alpha(X)\big] \ = \ \sup\big\{E_{\theta_1}\big[\bar{\varphi}(X)\big]\colon\ E_{\theta_0}\big[\bar{\varphi}(X)\big] \leq \alpha\big\} \qquad \forall \theta_1 > \theta_0, \tag{3.9}$$

i.e. φ_α is a **uniformly** most powerful level α test for H_0 versus H_1'.

2) We prove first that the power function $\theta \mapsto \beta(\theta, \varphi_\alpha) = E_\theta\big[\varphi_\alpha(X)\big]$ is monotonically increasing. Indeed, we obtain, for $\theta < \theta'$,

$$\begin{aligned} E_\theta\big[\varphi_\alpha(X)\big] \ &= \ P_\theta\Big(\sqrt{n}\,\frac{\bar{X} - \theta_0}{\sigma} \ \geq \ \Phi^{-1}(1-\alpha)\Big) \\ &= \ P_\theta\Big(\underbrace{\sqrt{n}\,\frac{\bar{X} - \theta}{\sigma}}_{\sim\mathcal{N}(0,1)} \ \geq \ \Phi^{-1}(1-\alpha) \ + \ \sqrt{n}\,\frac{\theta_0 - \theta}{\sigma}\Big) \\ &< \ P_{\theta'}\Big(\underbrace{\sqrt{n}\,\frac{\bar{X} - \theta'}{\sigma}}_{\sim\mathcal{N}(0,1)} \ \geq \ \Phi^{-1}(1-\alpha) \ + \ \sqrt{n}\,\frac{\theta_0 - \theta'}{\sigma}\Big) \\ &= \ E_{\theta'}\big[\varphi_\alpha(X)\big]. \end{aligned}$$

Therefore we obtain that

$$E_\theta\big[\varphi_\alpha(X)\big] \;\le\; E_{\theta_0}\big[\varphi_\alpha(X)\big] \;=\; \alpha \qquad \forall \theta \le \theta_0,$$

i.e. φ_α is a level α test for H_0'.

Since the family of level α tests for H_0' is contained in the family of level α tests for H_0 we conclude from (3.9) that φ_α is a uniformly most powerful level α tests for H_0' versus H_1'.

Now we consider a test problem which reveals the limitation of our approach. Suppose again that realizations $x_1, \ldots, x_n$ of i.i.d. random variables $X_1, \ldots, X_n$ are observed, where $X_i \sim \mathcal{N}(\theta, \sigma^2)$ and σ^2 is known. We consider the problem of testing a simple hypothesis versus a **two-sided** alternative,

$$H_0\colon\ \theta = \theta_0 \qquad \text{vs.} \qquad H_1''\colon\ \theta \neq \theta_0.$$

The two-sided alternative hypothesis H_1'' claims that the parameter θ is simply not equal to the value θ_0 given by the null hypothesis – the direction does not matter. It follows from Theorem 3.2 that a uniformly most powerful level α for H_0 versus H_1'' does **not** exist if $\alpha \in (0,1)$. This can be seen as follows. Suppose that such a test φ does exist. Then φ is in particular a most powerful level α for $H_0\colon\ \theta = \theta_0$ versus $H_1^{>}\colon\ \theta = \theta_0 + 1$. According to Theorem 3.2, this implies that

$$\varphi(x) \;=\; \begin{cases} 1, & \text{if } \sqrt{n}\,\frac{\bar{x}_n - \theta_0}{\sigma} > \Phi^{-1}(1-\alpha), \\ 0, & \text{if } \sqrt{n}\,\frac{\bar{x}_n - \theta_0}{\sigma} < \Phi^{-1}(1-\alpha) \end{cases} \qquad \lambda^n - \text{almost everywhere.} \tag{3.10a}$$

On the other hand, φ must also be a most powerful level α for $H_0\colon\ \theta = \theta_0$ versus $H_1^{<}\colon\ \theta = \theta_0 - 1$. According by Theorem 3.2, this requires that

$$\varphi(x) \;=\; \begin{cases} 1, & \text{if } \sqrt{n}\,\frac{\bar{x}_n - \theta_0}{\sigma} < \Phi^{-1}(\alpha), \\ 0, & \text{if } \sqrt{n}\,\frac{\bar{x}_n - \theta_0}{\sigma} > \Phi^{-1}(\alpha) \end{cases} \qquad \lambda^n - \text{a.e.,} \tag{3.10b}$$

which contradicts (3.10a). Hence, for any non-trivial level of significance $\alpha \in (0,1)$, there does not exist a uniformly most powerful test for $H_0\colon\ \theta = \theta_0$ versus $H_1''\colon\ \theta \neq \theta_0$.

Next we are going to prove a general result regarding the existence and uniqueness of a UMP test for composite hypotheses. Before we proceed, we look back to the problems of testing hypotheses about the location parameter of normally distributed random variables. In this case, a uniformly most powerful level α test does not exist when we have to deal with a two-sided alternative. On the other hand, it exists when a one-sided hypothesis versus a one-sided alternative has to be tested. In this case, we could start with a most powerful level α test φ_α for a simple null hypothesis $H_0\colon\ \theta = \theta_0$ versus a **simple** alternative $H_1\colon\ \theta = \theta_1$. It turned out that this test does not depend on the particular value of θ_1; it was only essential that $\theta_1 > \theta_0$. This led to the fact that φ_α is also uniformly most powerful against all values of θ contained in the composite alternative hypothesis $H_1'\colon\ \theta > \theta_0$. The reason why most powerful tests of H_0 versus alternatives $H_1\colon\ \theta = \theta_1$ have the same form for all $\theta_1 > \theta_0$ is that the corresponding likelihood ratios $p_{\theta_1}(x)/p_{\theta_0}(x)$ have some sort of similar structure. The next definition makes this point clear.

Definition 3.4. Suppose that the distribution of a random variable X taking values in Ω_X is in $\mathcal{P} = \{P_\theta\colon\ \theta \in \Theta\}$, a parametric family indexed by a real-valued parameter θ, and that $\mathcal{P}$ is dominated by a σ-finite measure μ. Let $p_\theta = dP_\theta/d\mu\ \forall \theta \in \Theta$.

The family $\mathcal{P}$ is said to have a **monotone likelihood ratio** in the real-valued statistic $t(X)$ if and only if, for any $\theta_1 < \theta_2$ $(\theta_1, \theta_2 \in \Theta)$, there exists a monotonically non-decreasing function $g_{\theta_1,\theta_2}\colon\ \mathbb{R} \to (0,\infty)$ such that

$$\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} \ = \ g_{\theta_1,\theta_2}\big(t(x)\big) \qquad \forall x \in \Omega_X.$$

Example
Suppose that $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2)$ are independent, $\theta := \Theta := \mathbb{R}$ and $\sigma^2 > 0$ fixed. Let p_θ be the density of $X = \big(X_1, \ldots, X_n\big)^T$ if θ is the corresponding location parameter.

Then, for $\theta_1 < \theta_2$,

$$\begin{aligned}
\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} \ &= \ \frac{\exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n \big(x_i - \theta_2\big)^2 \Big\}}{\exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n \big(x_i - \theta_1\big)^2 \Big\}} \\
&= \ \exp\Big\{ \frac{1}{2\sigma^2} \sum_{i=1}^n \big(x_i - \theta_2 + \theta_2 - \theta_1\big)^2 \ - \ \big(x_i - \theta_2\big)^2 \Big\} \\
&= \ \exp\Big\{ \frac{n}{2\sigma^2} \big(\theta_2 - \theta_1\big)^2 \quad + \quad \frac{n}{\sigma^2} \big(\bar{x}_n - \theta_2\big)\big(\theta_2 - \theta_1\big) \Big\}.
\end{aligned}$$

Hence, $x \mapsto p_{\theta_2}(x)/p_{\theta_1}(x)$ is monotonically increasing in $t(x) = \bar{x}_n$.

Theorem 3.3. *Let $X \sim P_\theta$, where $\theta \in \Theta \subseteq \mathbb{R}$. Suppose that there exists some σ-finite measure μ such that $P_\theta \ll \mu$ for all $\theta \in \Theta$, let $p_\theta = dP_\theta/d\mu$ be the corresponding densities, and suppose that the family $\{P_\theta\colon\ \theta \in \Theta\}$ has a monotone likelihood ratio in $t(X)$. We consider the test problem*

$$H_0\colon\ \theta \le \theta_0 \qquad \textit{vs.} \qquad H_1\colon\ \theta > \theta_0,$$

where $\theta_0 \in \Theta$ and $\{\theta \in \Theta\colon\ \theta \le \theta_0\}$, $\{\theta \in \Theta\colon\ \theta > \theta_0\}$ are non-empty sets.

(i) For each $\alpha \in (0,1)$, there exists a uniformly most powerful level α test φ_α, which is equal to

$$\varphi_\alpha(x) \ = \ \begin{cases} 1, & \text{if } \ t(x) > c_\alpha, \\ \gamma_\alpha, & \text{if } \ t(x) = c_\alpha, \\ 0, & \text{if } \ t(x) < c_\alpha, \end{cases} \tag{3.11a}$$

where $c_\alpha \in \mathbb{R}$ and $\gamma_\alpha \in [0,1]$ are chosen such that

$$E_{\theta_0}\varphi_\alpha(X) \ = \ \alpha. \tag{3.11b}$$

(ii) Assume in addition that the function g_{θ_0,θ_1} is **strictly** *monotonically increasing for some $\theta_1 > \theta_0$. If $\bar{\varphi}$ is a uniformly most powerful level α test, $\alpha \in (0,1)$, then*

$$\mu\Big(\big\{x\colon \quad \bar{\varphi}(x) \neq \varphi_\alpha(x) \quad \textit{and} \quad t(x) \neq c_\alpha\big\}\Big) \ = \ 0.$$

Proof.

(i) a) (Existence of c_α and γ_α)
We have to show that there exist $c_\alpha \in \mathbb{R}$ and $\gamma_\alpha \in [0,1]$ such that

$$P_{\theta_0}\big(t(X) > c_\alpha\big) \,+\, \gamma_\alpha \, P_{\theta_0}\big(t(X) = c_\alpha\big) \,=\, \alpha. \tag{3.12}$$

Let $h\colon\, \mathbb{R} \to [0,1]$ be defined by

$$h(c) \,:=\, P_{\theta_0}\big(t(X) \geq c\big).$$

Then h is monotonically non-increasing, left-continuous and

$$\begin{aligned} h(c) \quad &\underset{c\to\infty}{\longrightarrow} \quad 0, \\ h(c) \quad &\underset{c\to-\infty}{\longrightarrow} \quad 1. \end{aligned}$$

Let $c_\alpha := \sup\big\{c\colon\, h(c) \geq \alpha\big\}$. Then $c_\alpha \in \mathbb{R}$,

$$P_{\theta_0}\big(t(X) \geq c_\alpha\big) \,\geq\, \alpha \qquad (\text{since } \ h \ \text{ is left-continuous})$$

and

$$P_{\theta_0}\big(t(X) > c_\alpha\big) \,=\, P_{\theta_0}\Big(\bigcup_{c\colon\, c>c_\alpha} \big\{x\colon\, t(x) \geq c\big\}\Big) \,=\, \lim_{c \searrow c_\alpha,\, c>c_\alpha} P_{\theta_0}\big(t(X) \geq c\big) \,\leq\, \alpha.$$

Now we distinguish between two cases:
If $P_{\theta_0}\big(t(X) > c_\alpha\big) = \alpha$, then we set $\gamma_\alpha := 0$.
If $P_{\theta_0}\big(t(X) > c_\alpha\big) < \alpha$, then $P_{\theta_0}\big(t(X) = c_\alpha\big) = P_{\theta_0}\big(t(X) \geq c_\alpha\big) - P_{\theta_0}\big(t(X) > c_\alpha\big) > 0$ and we set

$$\gamma_\alpha \,:=\, \frac{\alpha \,-\, P_{\theta_0}\big(t(X) > c_\alpha\big)}{P_{\theta_0}\big(t(X) = c_\alpha\big)} \in [0,1].$$

In both cases, c_α and γ_α are such that (3.12) is satisfied.

b) (Optimality of φ_α)
Let $\bar\varphi$ be an arbitrary test such that $E_{\theta_0}\big[\bar\varphi(X)\big] \,\leq\, \alpha$ and let $\theta_1 \,>\, \theta_0$ be arbitrary. We have that

$$\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} \,=\, g_{\theta_0,\theta_1}\big(t(x)\big).$$

Let $d_\alpha := g_{\theta_0,\theta_1}\big(c_\alpha\big)$.
Since $g\colon\, \mathbb{R} \,\to\, (0,\infty)$ is monotonically non-decreasing, we obtain that $p_{\theta_1}(x) \geq d_\alpha \, p_{\theta_0}(x)$ when $t(x) \,\geq\, c_\alpha$ and $p_{\theta_1}(x) \,\leq\, d_\alpha \, p_{\theta_0}(x)$ when $t(x) \,\leq\, c_\alpha$. Therefore,

$$\begin{aligned} p_{\theta_1}(x) < d_\alpha \, p_{\theta_0}(x) \qquad &\text{implies} \qquad t(x) < c_\alpha, \\ p_{\theta_1}(x) > d_\alpha \, p_{\theta_0}(x) \qquad &\text{implies} \qquad t(x) > c_\alpha. \end{aligned}$$

Hence,

$$\big(\varphi_\alpha(x) - \bar\varphi(x)\big)\big(p_{\theta_1}(x) - d_\alpha \, p_{\theta_0}(x)\big) \,\geq\, 0 \qquad \forall x \in \Omega_X.$$

Now we obtain optimality of the test φ_α:

$$\begin{aligned} 0 &\leq d_\alpha \, E_{\theta_0}\big[\varphi_\alpha(X) - \bar{\varphi}(X)\big] \\ &= \int_{\Omega_X} \big(\varphi_\alpha(x) - \bar{\varphi}(x)\big)\, d_\alpha \, p_{\theta_0}(x)\, d\mu(x) \\ &\leq \int_{\Omega_X} \big(\varphi_\alpha(x) - \bar{\varphi}(x)\big)\, p_{\theta_1}(x)\, d\mu(x) \\ &= E_{\theta_1}\varphi_\alpha(X) \; - \; E_{\theta_1}\bar{\varphi}(X). \end{aligned}$$

This means that the power of $\bar{\varphi}$ against the alternative θ_1 is not greater than that of φ_α. Since this relation holds true for all $\theta_1 > \theta_0$ we conclude that φ_α is a uniformly most powerful test of $H_0^=\colon\ \theta = \theta_0$ versus $H_1\colon\ \theta > \theta_0$.

c) (Admissibility w.r.t. H_0)
It remains to show that

$$E_\theta\varphi_\alpha(X) \;\leq\; \alpha \qquad \forall\theta < \theta_0. \tag{3.13}$$

We can see by calculations analogous to those in part b) that $(1 - \varphi_\alpha)$ is a UMP test of size $1-\alpha$ for $H_0^=\colon\ \theta = \theta_0$ versus $H_1^<\colon\ \theta < \theta_0$. Since φ^* such that $\varphi^*(x) = 1-\alpha\ \forall x \in \Omega_X$ is also a level $1-\alpha$ test for $H_0^=$ vs. $H_1^<$ we obtain that

$$E_\theta\big[1 - \varphi_\alpha(X)\big] \;\geq\; E_\theta\big[\varphi^*(X)\big] \;=\; 1 \;-\; \alpha \qquad \forall\theta < \theta_0,$$

which implies that (3.13) is satisfied. Hence, φ_α is a level α test for $H_0\colon\ \theta \leq \theta_0$ versus H_1. Since the set of level α tests for $H_0\colon\ \theta \leq 0$ is contained in the set of level α tests for $H_0^=\colon\ \theta = \theta_0$ we conclude in conjunction with part b) that φ_α is a UMP level α test of H_0 versus H_1.

(ii) (Uniqueness)
Let $\bar{\varphi}$ be an arbitrary level α test for H_0 vs. H_1, that is $\sup\big\{E_\theta\bar{\varphi}(X) \leq \alpha\colon\ \theta \leq \theta_0\big\}$. By assumption, there exists some $\theta_1 > \theta_0$ such that the function g_{θ_0,θ_1} is **strictly** monotonically increasing. Let, as above, $d_\alpha := g_{\theta_0,\theta_1}(c_\alpha)$. Then

$$\begin{aligned} p_{\theta_1}(x) < d_\alpha\, p_{\theta_0}(x) \qquad &\text{if and only if} \qquad t(x) < c_\alpha, \\ p_{\theta_1}(x) > d_\alpha\, p_{\theta_0}(x) \qquad &\text{if and only if} \qquad t(x) > c_\alpha. \end{aligned}$$

Hence,

$$\begin{aligned} &\big\{x \in \Omega_X\colon\ \bar{\varphi}(x) \neq \varphi_\alpha(x) \quad \text{and} \quad t(x) \neq c_\alpha\big\} \\ &\quad = \ \big\{x \in \Omega_X\colon\ \bar{\varphi}(x) \neq \varphi_\alpha(x) \quad \text{and} \quad p_{\theta_1}(x) \neq d_\alpha\, p_{\theta_0}(x)\big\}. \end{aligned}$$

Assume that

$$\mu\Big(\big\{x \in \Omega_X\colon\ \bar{\varphi}(x) \neq \varphi_\alpha(x) \quad \text{and} \quad t(x) \neq c_\alpha\big\}\Big) \;>\; 0.$$

Since

$$\big(\varphi_\alpha(x) - \bar{\varphi}(x)\big)\big(p_{\theta_1}(x) - d_\alpha\, p_{\theta_0}(x)\big) \;\geq\; 0 \qquad \forall x \in \Omega_X$$

we obtain that

$$\begin{aligned} 0 &\leq d_\alpha \Big(E_{\theta_0}\varphi_\alpha(X) - E_{\theta_0}\bar{\varphi}(X) \Big) \\ &= \int_{\Omega_X} \big(\varphi_\alpha(x) - \bar{\varphi}(x)\big)\, d_\alpha\, p_{\theta_0}(x)\, d\mu(x) \\ &< \int_{\Omega_X} \big(\varphi_\alpha(x) - \bar{\varphi}(x)\big)\, p_{\theta_1}(x)\, d\mu(x) \\ &= E_{\theta_1}\varphi_\alpha(X) - E_{\theta_1}\bar{\varphi}(X). \end{aligned}$$

Hence, $\bar{\varphi}$ has less power against the alternative θ_1 than φ_α. Therefore, $\bar{\varphi}$ is **not** a UMP level α test of $H_0\colon\ \theta \leq \theta_0$ versus $H_1\colon\ \theta > \theta_0$.

□

The p-value

Our approach to testing hypotheses contains a subjective choice of a level of significance. It could happen that experimenter I may be satisfied to reject the hypothesis H_0 using a test with size $\alpha = 0.05$, while experimenter II insists on using $\alpha = 0.01$. Even if both use the same test statistic, it is then possible that experimenter I rejects the hypothesis H_0 while experimenter II accepts H_0 on the basis of the same outcome x of the experiment. Moreover, in some cases an ultimate decision for or against a hypothesis H_0 is not necessary. For example, econometricians often use regression models with a large number of parameters. Of course, it would be desirable to reduce such a model by dropping all superfluous parameters which do not contribute to explain a cause-effect relation. In such a case, it is natural to test hypotheses about the significance of these parameters. However, an ultimate decision about inclusion or exclusion of such a parameters is not absolutely necessary since one can maintain the maximal model. The difficulty of selecting a level of significance may be overcome by reporting the outcome of the experiment in terms of the **observed size** or **p-value** of the test. Recall that the size of a nonrandomized test based on a statistic T and a critical value t is given by

$$\alpha(t) = \sup\big\{P_\theta(T(X) \geq t)\colon\ \theta \in \Theta_0\big\}.$$

(This should not be mixed up with the level of significance; a test with size α is a level $\bar{\alpha}$ test for all $\bar{\alpha} \geq \alpha$.) The p-value is a statistic which is defined as the smallest level of significance at which an experimenter using T would reject on the basis **of the observed outcome** x.

Example

Suppose that realizations $x_1, \ldots, x_n$ of i.i.d. random variables $X_1, \ldots, X_n$ are observed, $X_i \sim N(\theta, \sigma^2)$ $(i = 1, \ldots, n)$, where $\sigma^2 > 0$ is known. A most powerful level α test φ_α for

$$H_0\colon\ \theta = \theta_0 \qquad \text{vs.} \qquad H_1\colon\ \theta = \theta_1,$$

$\theta_0 < \theta_1$ is given by

$$\varphi_\alpha(x) = \begin{cases} 1, & \text{if } \sqrt{n}\frac{\bar{x}-\theta_0}{\sigma} \geq \Phi^{-1}(1-\alpha), \\ 0, & \text{if } \sqrt{n}\frac{\bar{x}-\theta_0}{\sigma} < \Phi^{-1}(1-\alpha \end{cases}$$

If $X = x$, then the actual p-value $\widehat{\alpha}(x)$ is such that

$$\Phi^{-1}\big(1 - \widehat{\alpha}(x)\big) \;=\; \sqrt{n}\frac{\bar{x} - \theta_0}{\sigma},$$

which is equivalent to

$$\widehat{\alpha}(x) \;=\; 1 \;-\; \Phi\Big(\sqrt{n}\frac{\bar{x} - \theta_0}{\sigma}\Big).$$

Note that $\widehat{\alpha}(x)$ depends on the realization x of X and is therefore random; $\widehat{\alpha}(X)$ is the corresponding statistic (random variable).

While a small p-value provides a strong evidence against H_0, a medium value does not mean much. This is indicated by the following lemma.

Lemma 3.4. *Suppose that we observe a realization x of some random varieble X and that $\big(\varphi_\alpha\big)_{\alpha\in[0,1]}$ is a family of nonrandomized tests for $H_0\colon\ \theta = \theta_0$ versus any alternative, where $P_{\theta_0}\big(\varphi_\alpha(X) = 1\big) = \alpha$. Moreover, we assume that, for $0 \le \alpha \le \beta \le 1$,*

$$\varphi_\alpha(x) \;\le\; \varphi_\beta(x) \qquad \forall x \in \Omega_X.$$

Then

$$P_{\theta_0}^{\widehat{\alpha}(X)} \;=\; \mathit{Uniform}[0,1].$$

Proof. We have to show that

$$P_{\theta_0}\big(\widehat{\alpha}(X) \le u\big) \;=\; u \qquad \forall u \in (0,1).$$

First of all, we have

$$\big\{x\colon\ \widehat{\alpha}(x) < u\big\} \;=\; \big\{x\colon\ \inf\{\alpha\colon\ \varphi_\alpha(x) = 1\} < u\big\} \;\subseteq\; \big\{x\colon\ \varphi_u(x) = 1\big\}$$

and

$$\big\{x\colon\ \varphi_u(x) = 1\big\} \;\subseteq\; \big\{x\colon\ \inf\{\alpha\colon\ \varphi_\alpha(x) = 1\} \le u\big\} \;=\; \big\{x\colon\ \widehat{\alpha}(x) \le u\big\}.$$

Therefore,

$$P_{\theta_0}\big(\widehat{\alpha}(X) < u\big) \;\le\; P_{\theta_0}\big(\varphi_u(X) = 1\big) \;=\; u \;\le\; P_{\theta_0}\big(\widehat{\alpha}(X) \le u\big),$$

which implies, for $0 \le u < u + \epsilon \le 1$,

$$u \;\le\; P_{\theta_0}\big(\widehat{\alpha}(X) \le u\big) \;\le\; P_{\theta_0}\big(\widehat{\alpha}(X) < u + \epsilon\big) \;\le\; u + \epsilon.$$

□

3.3 Likelihood ratio tests

We have seen in the previous subsection that a uniformly most powerful level α test may not exist. Indeed, when realizations of i.i.d. random variables $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2)$ are observed, there does not exist a UMP test for $H_0\colon\ \theta = \theta_0$ versus the **two-sided alternative** $H_1\colon\ \theta \neq \theta_0$. In view of this, there is no clear guideline how a good test should be constructed. In this subsection, we introduce a generalization of the Neyman-Pearson statistics which yields good procedures for a great number of hypothesis testing problems. Suppose that we observe a realization $x = (x_1, \ldots, x_n)^T$ of a random variable $X = (X_1, \ldots, X_n)^T \sim P_\theta$ and that we wish to test $H_0\colon\ \theta \in \Theta_0$ versus $H_1\colon\ \theta \in \Theta_1$. Suppose further that there exists a σ-finite measure μ such that $P_\theta \ll \mu$ for all $\theta \in \Theta = \Theta_0 \cup \Theta_1$ and let $p_\theta := dP_\theta/d\mu$ denote the corresponding densities. The test statistic we want to consider is the **likelihood ratio** given by

$$L(x) \;=\; \frac{\sup\big\{p_\theta(x)\colon\ \theta \in \Theta_1\big\}}{\sup\big\{p_\theta(x)\colon\ \theta \in \Theta_0\big\}}. \tag{3.14}$$

Tests that reject H_0 for large values of $L(x)$ are called **likelihood ratio tests**. The statistic $L(x)$ coincides with the optimal test statistic when $\Theta_0 = \big\{\theta_0\big\}$ and $\Theta_1 = \big\{\theta_1\big\}$. In some cases we shall consider, $\theta \mapsto p_\theta(x)$ is a continuous function of θ for all x and the set Θ_1 is dense in $\Theta = \Theta_0 \cup \Theta_1$. Therefore, $\sup\big\{p_\theta(x)\colon\ \theta \in \Theta_1\big\} = \sup\big\{p_\theta(x)\colon\ \theta \in \Theta\big\}$, and the test statistic $L(x)$ is equal to

$$\widetilde{L}(x) \;=\; \frac{\sup\big\{p_\theta(x)\colon\ \theta \in \Theta\big\}}{\sup\big\{p_\theta(x)\colon\ \theta \in \Theta_0\big\}}. \tag{3.15}$$

For typical "textbook examples" maximum likelihood estimators exist and we may proceed as follows:

1) Calculate the maximum likelihood estimate $\widehat{\theta}$ of θ where θ may vary in $\Theta = \Theta_0 \cup \Theta_1$.

2) Calculate the maximum likelihood estimate $\widehat{\theta}_0$ of θ in the restricted model where θ may vary only in Θ_0.

3) Form $\widetilde{L}(x) = p_{\widehat{\theta}}(x)/p_{\widehat{\theta}_0}(x)$.

4) Find a function h which is strictly increasing on the range of $\widetilde{L}$ such that $h\big(\widetilde{L}(X)\big)$ has a simple form and a tabulated distribution under H_0. Since $h\big(\widetilde{L}(X)\big)$ is equivalent to $\widetilde{L}(X)$ we specify the size α likelihood ratio test through the test statistic $h\big(\widetilde{L}(X)\big)$ and its $(1-\alpha)$-quantile obtained from the table.

In what follows we consider a few examples.

Example 1: Two-sided z test
Suppose that we observe realizations $x_1, \ldots, x_n$ of i.i.d. random variables $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2)$, where $\theta \in \Theta := \mathbb{R}$ is the parameter of interest and $\sigma^2 > 0$ is assumed to be known. We consider the test problem

$$H_0\colon\ \theta = \theta_0 \qquad \text{vs.} \qquad H_1\colon\ \theta \neq \theta_0$$

and we intend to derive a size α likelihood ratio test.

Solution:
Let $X = \big(X_1, \ldots, X_n\big)^T$ and $x = \big(x_1, \ldots, x_n\big)^T$ and let p_θ be the density of X under P_θ. Since $\theta \mapsto p_\theta(x)$ is a continuous function and $\Theta_1 = \big\{\theta \in \Theta\colon\ \theta \neq \theta_0\big\}$ is dense in Θ, the statistic $L(x)$ is equal to $\widetilde{L}(x)$ and we only have to find the respective maximum likelihood estimates of θ in the unrestricted model and in the model given by $\Theta_0 = \{\theta_0\}$. The former is given by $\widehat{\theta} = \bar{X}_n$ and, since H_0 is a simple hypothesis, the latter by $\widehat{\theta_0} = \theta_0$. The statistic $\widetilde{L}(x)$ is equal to

$$\begin{aligned}
\widetilde{L}(x) &= \frac{p_{\widehat{\theta}}(x)}{p_{\widehat{\theta_0}}(x)} \\
&= \frac{\exp\big\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \bar{x}_n)^2\big\}}{\exp\big\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta_0)^2\big\}} \\
&= \exp\Big\{\frac{1}{2\sigma^2}\Big[\sum_{i=1}^n (x_i - \bar{x}_n + \bar{x}_n - \theta_0)^2 \ - \ \sum_{i=1}^n (x_i - \bar{x}_n)^2\Big]\Big\} \\
&= \exp\Big\{\frac{n}{2\sigma^2}(\bar{x}_n - \theta_0)^2\Big\}.
\end{aligned}$$

We see that $\widetilde{L}(x)$ is strictly increasing in $\big|\bar{x}_n - \theta_0\big|$. Therefore, the searched-for likelihood ratio test has the following form:

$$\begin{aligned}
\varphi(x) &= \begin{cases} 1, & \text{if } \ \widetilde{L}(x) > c_\alpha, \\ \gamma_\alpha, & \text{if } \ \widetilde{L}(x) = c_\alpha, \\ 0, & \text{if } \ \widetilde{L}(x) < c_\alpha \end{cases} \\
&= \begin{cases} 1, & \text{if } \ \big|\bar{x}_n - \theta_0\big| > c'_\alpha, \\ \gamma_\alpha, & \text{if } \ \big|\bar{x}_n - \theta_0\big| = c'_\alpha, \\ 0, & \text{if } \ \big|\bar{x}_n - \theta_0\big| < c'_\alpha \end{cases}
\end{aligned}$$

It remains to determine an appropriate critical value c'_α and the randomization constant γ_α. Under the null hypothesis H_0 we have that $\bar{X}_n \sim \mathcal{N}(\theta_0, \sigma^2/n)$. Therefore, it is most convenient if φ is represented in the following equivalent form:

$$\varphi(x) = \begin{cases} 1, & \text{if } \ \frac{\sqrt{n}\big|\bar{x}_n - \theta_0\big|}{\sigma} > c''_\alpha, \\ \gamma_\alpha, & \text{if } \ \frac{\sqrt{n}\big|\bar{x}_n - \theta_0\big|}{\sigma} = c''_\alpha, \\ 0, & \text{if } \ \frac{\sqrt{n}\big|\bar{x}_n - \theta_0\big|}{\sigma} < c''_\alpha \end{cases}$$

This test has size $\alpha \in (0,1)$ if $c''_\alpha \in \mathbb{R}$ and $\gamma_\alpha \in [0,1]$ are chosen such that

$$\begin{aligned}
\alpha &= \sup_{\theta \in \Theta_0} E_\theta \varphi(X) \ = \ E_{\theta_0}\varphi(X) \\
&= P_{\theta_0}\Big(\Big|\underbrace{\frac{\sqrt{n}(\bar{X}_n - \theta_0)}{\sigma}}_{\sim \mathcal{N}(0,1)}\Big| > c''_\alpha\Big) \ + \ \gamma_\alpha \ \underbrace{P_{\theta_0}\Big(\Big|\frac{\sqrt{n}(\bar{X}_n - \theta_0)}{\sigma}\Big| = c''_\alpha\Big)}_{=0},
\end{aligned}$$

which is accomplished by the choice $c''_\alpha = \Phi^{-1}(1 - \alpha/2)$ and $\gamma_\alpha = 1$.

Example 2: Two-sided t test
Suppose that we observe realizations $x_1, \ldots, x_n$ of i.i.d. random variables $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Now we assume that both μ and σ^2 are unknown and the underlying distributions are therefore parametrized by $\theta = \binom{\mu}{\sigma^2}$, where $\theta \in \Theta := \mathbb{R} \times (0, \infty)$. We intend to test again whether the location parameter is equal to some particular value, say μ_0. However, we have to take into account that σ^2 is also unknown and the test problem is therefore correctly stated as follows.

$$H_0\colon\ \theta \in \Theta_0 := \{\mu_0\} \times (0, \infty) \qquad \text{vs.} \qquad H_1\colon\ \theta = \Theta_1 = \Theta \setminus \Theta_0 = (\mathbb{R} \setminus \{\mu_0\}) \times (0, \infty).$$

We seek again a size α likelihood ratio test.

Solution:
Let p_θ be the density of $X = \big(X_1, \ldots, X_n\big)^T$ under P_θ. Since $\theta \mapsto p_\theta(x)$ is continuous for all $x \in \mathbb{R}^n$ and Θ_1 is dense in Θ we obtain that $L(x) = \widetilde{L}(x) = p_{\widehat{\theta}}(x)/p_{\widehat{\theta}_0}(x)$, where $\widehat{\theta}$ and $\widehat{\theta}_0$ are the maximum likelihood estimators when θ varies in Θ and Θ_0, respectively.

In the unrestricted case, the corresponding maximum likelihood estimator $\widehat{\theta}$ of θ is already known. It is given by $\widehat{\theta} = \binom{\widehat{\mu}}{\widehat{\sigma^2}} = \binom{\bar{X}_n}{n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2}$. If θ varies in Θ_0, then the maximum likelihood estimator $\widehat{\mu}_0$ of the location parameter is given by μ_0 and the maximum likelihood estimator $\widehat{\sigma_0^2}$ of the variance parameter can be shown to be equal to $n^{-1}\sum_{i=1}^n (X_i - \mu_0)^2$. Hence, $\widehat{\theta}_0 = \binom{\mu_0}{n^{-1}\sum_{i=1}^n (X_i - \mu_0)^2}$. The test statistic $\widetilde{L}(X)$ is given by

$$\begin{aligned}
\widetilde{L}(x) &= \frac{\frac{1}{(2\pi\,\widehat{\sigma^2})^{n/2}} \exp\Big\{ -\overbrace{\frac{1}{2\widehat{\sigma^2}} \sum_{i=1}^n (x_i - \bar{x}_n)^2}^{=n/2} \Big\}}{\frac{1}{(2\pi\,\widehat{\sigma_0^2})^{n/2}} \exp\Big\{ -\underbrace{\frac{1}{2\widehat{\sigma_0^2}} \sum_{i=1}^n (x_i - \mu_0)^2}_{=n/2} \Big\}} \\
&= \left(\frac{\widehat{\sigma_0^2}}{\widehat{\sigma^2}}\right)^{n/2} = \left(\frac{n^{-1}\sum_{i=1}^n (x_i - \bar{x}_n + \bar{x}_n - \mu_0)^2}{n^{-1}\sum_{i=1}^n (x_i - \bar{x}_n)^2}\right)^{n/2} \\
&= \left(1 + \frac{(\bar{x}_n - \mu_0)^2}{\widehat{\sigma^2}}\right)^{n/2}.
\end{aligned}$$

Therefore, the searched-for likelihood ratio test can be represented as

$$\varphi(x) = \begin{cases} 1, & \text{if } \frac{(\bar{x}-\mu_0)^2}{n^{-1}\sum_{i=1}^n (x_i - \bar{x}_n)^2} > c_\alpha, \\ \gamma_\alpha, & \text{if } \frac{(\bar{x}-\mu_0)^2}{n^{-1}\sum_{i=1}^n (x_i - \bar{x}_n)^2} = c_\alpha, \\ 0, & \text{if } \frac{(\bar{x}-\mu_0)^2}{n^{-1}\sum_{i=1}^n (x_i - \bar{x}_n)^2} < c_\alpha. \end{cases}$$

To obtain a size α test, we have to find $c_\alpha \in \mathbb{R}$ and $\gamma_\alpha \in [0, 1]$ such that

$$\sup_{\theta \in \Theta_0} E_\theta \varphi(X) = \alpha.$$

It will be shown below that, in case of $\mu = \mu_0$, the distribution of

$$T_n(X) = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sqrt{(n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2}}$$

does **not** depend on the particular value of σ^2. In fact, we will see that $T_n(X)$ has under the null hypothesis a so-called t distribution with $n-1$ degrees of freedom. Therefore, it is most convenient to represent the likelihood ratio test in the following form:

$$\varphi(x) \;=\; \begin{cases} 1, & \text{if } \big|T_n(x)\big| > c'_\alpha, \\ \gamma_\alpha, & \text{if } \big|T_n(x)\big| = c'_\alpha, \\ 0, & \text{if } \big|T_n(x)\big| < c'_\alpha. \end{cases}$$

The constants $c'_\alpha \in \mathbb{R}$ and $\gamma_\alpha \in [0,1]$ have to be chosen such that the size of the test equals a prescribed value $\alpha > 0$. Since a t distribution has a density which is symmetric about 0, we choose c'_α as the $(1-\alpha/2)$-quantile of a t distribution with $n-1$ degrees of freedom. Since $P_{\binom{\mu_0}{\sigma^2}}\big(|T_n| = c'_\alpha\big) = 0$ we can choose γ_α as an arbitrary number from $[0,1]$. In particular, there is no need for a randomization and we may choose $\gamma_\alpha = 1$.

In the following we give a detailed derivation of the distribution of the statistic $T_n(X)$ under the null hypothesis. We begin with a constructive definition of the t distribution.

Definition 3.5.

(i) Let $X_1, \ldots, X_k$ be independent and identically distributed, $X_i \sim \mathcal{N}(0,1)$ $(i = 1, \ldots, k)$. Then

$$Y \;:=\; X_1^2 + \cdots + X_k^2$$

has a χ^2 **distribution with** k **degrees of freedom**. $(Y \sim \chi^2_k)$

(ii) Let $X \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2_k$ be independent. Then

$$Z \;:=\; \frac{X}{\sqrt{Y/k}}$$

has a t **distribution with** k **degrees of freedom**. $(Z \sim t_k)$

The following theorem states that the above statistic $T_n(X)$ has a t_{n-1} distribution under the null hypothesis.

Theorem 3.5. *Let* $X_1, \ldots, X_n \sim \mathcal{N}\big(\mu, \sigma^2\big)$ *be independent. Then*

$$T_n(X) \;=\; \frac{\sqrt{n}(\bar{X}_n - \mu)}{\widehat{\sigma}_n} \;\sim\; t_{n-1},$$

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

Before we prove this theorem we derive two auxiliary results.

Lemma 3.6. *Let* $X \sim \mathcal{N}\big(\mu, \sigma^2 I_n\big)$, $\mu \in \mathbb{R}^n$, $\sigma^2 > 0$, *and let* A *and* B *be* $(k \times n)$- *and* $(l \times n)$*-matrices, respectively.*

If $AB^T = 0_{k\times l}$, *then* AX *and* BX *are independent.*

($0_{k\times l}$ *denotes the matrix with* k *rows and* l *columns where all entries are zero.)*

Proof. We derive first the characteristic function of a multivariate normal distribution with parameters $\mu \in \mathbb{R}^n$ and Σ. It is well-known that the characteristic function φ_X of a standard normal variable $X \sim \mathcal{N}(0,1)$ is given by

$$\varphi_X(t) \;=\; E\big[e^{itX}\big] \;=\; e^{-t^2/2} \qquad \forall t \in \mathbb{R}.$$

If now $X = (X_1, \ldots, X_n)^T \sim \mathcal{N}\big(0_n, I_n\big)$, then $X_1, \ldots, X_n$ are independent and follow a standard normal distribution. Therefore, the characteristic function φ_X of X is given by

$$\begin{aligned}\varphi_X(t) \;&=\; E\big[e^{it^TX}\big] \;=\; E\Big[\prod_{j=1}^n e^{it_jX_j}\Big] \;=\; \prod_{j=1}^n E\big[e^{it_jX_j}\big] \\ &=\; \prod_{j=1}^n e^{-t_j^2/2} \;=\; e^{-t^Tt/2} \qquad \forall t = (t_1, \ldots, t_n)^T \in \mathbb{R}^n.\end{aligned}$$

If $Y = \Sigma^{1/2}X + \mu$, then $Y \sim \mathcal{N}(\mu, \Sigma)$ and the corresponding characteristic function φ_Y is given by

$$\begin{aligned}\varphi_Y(t) \;&=\; E\big[e^{it^T(\Sigma^{1/2}X+\mu)}\big] \;=\; e^{it^T\mu}\, Ee^{it^T(\Sigma^{1/2}X)} \;=\; e^{it^T\mu}\, Ee^{i(\Sigma^{1/2}t)^TX} \\ &=\; e^{it^T\mu}\,\varphi_X(\Sigma^{1/2}t) \;=\; e^{it^T\mu - t^T\Sigma t/2} \qquad \forall t = (t_1, \ldots, t_n)^T \in \mathbb{R}^n.\end{aligned}$$

Let now $X \sim \mathcal{N}\big(\mu, \sigma^2 I_n\big)$. Then $\varphi_X(t) = e^{it^T\mu - t^T\sigma^2 I_n t/2}\ \forall t \in \mathbb{R}^n$. For arbitrary $t_1 \in \mathbb{R}^k$, $t_2 \in \mathbb{R}^l$, we obtain that

$$\begin{aligned}\varphi_{\binom{AX}{BX}}\Big(\begin{pmatrix} t_1 \\ t_2 \end{pmatrix}\Big) \;&=\; E\big[e^{i\binom{t_1}{t_2}^T\binom{A}{B}X}\big] \;=\; E\big[e^{i(t_1^TA+t_2^TB)X}\big] \\ &=\; e^{i(t_1^TA+t_2^TB)\mu - (t_1^TA+t_2^TB)\sigma^2 I_n(A^Tt_1+B^Tt_2)} \\ &=\; e^{it_1^TA\mu - \sigma^2t_1^TAA^Tt_1/2}\, e^{it_2^TB\mu - \sigma^2t_2^TBB^Tt_2/2}.\end{aligned}$$

($e^{-\sigma^2t_1^TAB^Tt_2} = 1$ since, by assumption, $AB^T = 0_{k\times l}$.)
Moreover, we see that

$$\begin{aligned}\varphi_{AX}(t_1) \;&=\; Ee^{it_1^T(AX)} \;=\; Ee^{i(t_1^TA)X} \;=\; \varphi_X(A^Tt_1) \\ &=\; e^{it_1^TA\mu - \sigma^2t_1^TAA^Tt_1/2}\end{aligned}$$

and, analogously,

$$\varphi_{BX}(t_2) \;=\; e^{it_2^TB\mu - \sigma^2t_2^TBB^Tt_2/2}. \tag{3.16}$$

Therefore, we have

$$\varphi_{\binom{AX}{BX}}\Big(\begin{pmatrix} t_1 \\ t_2 \end{pmatrix}\Big) \;=\; \varphi_{AX}(t_1)\,\varphi_{BX}(t_2) \qquad \forall t_1 \in \mathbb{R}^k,\ \forall t_2 \in \mathbb{R}^l,$$

which implies that AX and BX are independent. □

Lemma 3.7. *Let $X \sim \mathcal{N}\big(0_n, I_n\big)$ and let M be a symmetric $(n \times n)$-matrix such that $M^2 = M$. (M is an orthogonal projection matrix.)*

Then

$$X^TMX \;\sim\; \chi^2_m,$$

where $m = rank(M)$.

Proof. We use the spectral decomposition of M:

$$M \;=\; D^T \mathrm{Diag}\big(\underbrace{1,\ldots,1}_{m \text{ times}},\underbrace{0,\ldots,0}_{n-m \text{ times}}\big) D,$$

where D is an orthogonal matrix, i.e. $DD^T = D^T D = I_n$. Then

$$Y \;=\; \big(Y_1,\ldots,Y_n\big)^T \;:=\; DX \;\sim\; \mathcal{N}\big(D0_n, DI_nD^T\big) \;=\; \mathcal{N}\big(0_n, I_n\big)$$

and, therefore,

$$\begin{aligned} X^TMX \;&=\; X^TD^T\mathrm{Diag}\big(1,\ldots,1,0,\ldots,0\big)DX \\ &=\; \sum_{i=1}^{m} Y_i^2 \;\sim\; \chi^2_m. \end{aligned}$$

□

Proof of Theorem 3.5. We have that

$$T_n(X) \;=\; \frac{\sqrt{n}\big(\bar{X}_n-\mu\big)}{\widehat{\sigma}_n} \;=\; \frac{\sqrt{n}\big(\bar{X}_n-\mu\big)/\sigma}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i-\bar{X}_n)^2/\sigma^2}} \;=:\; \frac{V_1}{\sqrt{V_2/(n-1)}}.$$

We show that

a) $V_1 \sim \mathcal{N}(0,1)$,

b) $V_2 \sim \chi^2_{n-1}$,

c) V_1 and V_2 are independent.

Then we obtain, according to our definition of a t distribution, that

$$T_n(X) \;=\; \frac{V_1}{\sqrt{V_2/(n-1)}} \;\sim\; t_{n-1}.$$

While a) is obvious, it takes a few lines to prove b). Let $Y_i := (X_i-\mu)/\sigma$. Then $Y_1,\ldots,Y_n$ are independent and follow a standard normal distribution. We have

$$\begin{aligned} V_2 \;&=\; \sum_{i=1}^n \Big(\frac{X_i-\mu}{\sigma} - \frac{\bar{X}_n-\mu}{\sigma}\Big)^2 \;=\; \sum_{i=1}^n \big(Y_i - \bar{Y}_n\big)^2 \\ &=\; \sum_{i=1}^n Y_i^2 - n\bar{Y}_n^2 \;=\; Y^TMY, \end{aligned}$$

where

$$M \;=\; I_n - \begin{pmatrix} 1/n & \ldots & 1/n \\ \vdots & \ddots & \vdots \\ 1/n & \ldots & 1/n \end{pmatrix}, \qquad Y \;=\; \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}.$$

Since M is an orthogonal projection matrix of rank $n-1$ it follows from Lemma 3.7 that $V_2 \sim \chi^2_{n-1}$.

Note that $V_1 = \big(1/\sqrt{n},\cdots,1/\sqrt{n}\big)Y$ and $V_2 = Y^TM^TMY$. Since $\big(1/\sqrt{n},\cdots,1/\sqrt{n}\big)M^T = 0_{1\times n}$ it follows from Lemma 3.6 that the random variables V_1 and MY, and therefore V_1 and $V_2 = Y^TM^TMY$ as well, are independent. □

After studying the distribution of our test statistic under H_0 we summarize our findings. On the basis of realizations $x_1, \dots, x_n$ of i.i.d. random variables $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ we may test

$$H_0\colon\ \theta \in \Theta_0 \ := \ \{\mu_0\} \times (0, \infty) \qquad \text{vs.} \qquad H_1\colon\ \theta = (\mathbb{R} \setminus \{\mu_0\}) \times (0, \infty).$$

Our likelihood ratio approach led us to the so-called two-sided t test φ, which for a given level of significance $\alpha \in (0,1)$ has the form

$$\varphi(x) \ = \ \begin{cases} 1, & \text{if } \ \big|T_n(x)\big| \geq t_{n-1,1-\alpha/2}, \\ 0, & \text{if } \ \big|T_n(x)\big| < t_{n-1,1-\alpha/2}, \end{cases}$$

where $t_{n-1,1-\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of a t distribution with $n-1$ degrees of freedom. Indeed, if $\theta \in \Theta_0$, then $P_\theta^{T_n(X)} = t_{n-1}$ and we obtain

$$P_\theta\big(\big|T_n(X)\big| \geq t_{n-1,1-\alpha/2}\big) \ = \ \underbrace{P_\theta\big(T_n(X) \geq t_{n-1,1-\alpha/2}\big)}_{=\alpha/2} + \underbrace{P_\theta\big(T_n(X) \leq -t_{n-1,1-\alpha/2}\big)}_{=\alpha/2} \ = \ \alpha.$$

Example 3: One-sided t test
Suppose again that we observe realizations $x_1, \dots, x_n$ of i.i.d. random variables $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Now we intend to test whether the location parameter is less than or equal to some particular value μ_0. For $\theta = \binom{\mu}{\sigma^2}$, the corresponding pair of hypotheses is given by

$$H_0\colon\ \theta \in \Theta_0 \ := \ (-\infty, \mu_0] \times (0, \infty) \qquad \text{vs.} \qquad H_1\colon\ \theta = \Theta_1 = (\mu_0, \infty) \times (0, \infty).$$

We derive again a size α likelihood ratio test.

Solution:
The density $p_{\binom{\mu}{\sigma^2}}$ of $X = \big(X_1, \dots, X_n\big)^T$ under $P_{\binom{\mu}{\sigma^2}}$ is given by

$$p_{\binom{\mu}{\sigma^2}}(x) \ = \ \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big\{ -\frac{1}{2\sigma^2}\Big[\sum_{i=1}^n (x_i - \bar{x}_n)^2 \ + \ n(\bar{x}_n - \mu)^2\Big]\Big\}$$

It can be shown that

$$p_{\widehat{\theta}_0}(x) \ = \ \sup\big\{p_\theta(x)\colon\ \theta \in \Theta_0\big\},$$

for $\widehat{\theta}_0 = \binom{\widehat{\mu}_0}{\widehat{\sigma}_0^2}$ such that

$$\widehat{\mu}_0 \ = \ \begin{cases} \bar{x}_n, & \text{if } \ \bar{x}_n \leq \mu_0, \\ \mu_0, & \text{if } \ \bar{x}_n > \mu_0 \end{cases}$$

and

$$\widehat{\sigma}_0^2 \ = \ \frac{1}{n}\sum_{i=1}^n \big(x_i - \widehat{\mu}_0\big)^2.$$

Analogously we can see that

$$p_{\widehat{\theta}_1}(x) \ = \ \sup\big\{p_\theta(x)\colon\ \theta \in \Theta_1\big\},$$

for $\widehat{\theta}_1 = \binom{\widehat{\mu}_1}{\widehat{\sigma}_1^2}$ such that

$$\widehat{\mu}_1 \ = \ \begin{cases} \bar{x}_n, & \text{if } \ \bar{x}_n \geq \mu_0, \\ \mu_0, & \text{if } \ \bar{x}_n < \mu_0 \end{cases}$$

and

$$\widehat{\sigma}_1^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}\big(x_i-\widehat{\mu}_1\big)^2.$$

Since

$$\begin{aligned} L(x) \;&=\; \frac{\sup\big\{p_\theta(x)\colon\ \theta\in\Theta_1\big\}}{\sup\big\{p_\theta(x)\colon\ \theta\in\Theta_0\big\}} \\ &=\; \frac{p_{\widehat{\theta}_1}(x)}{p_{\widehat{\theta}_0}(x)} \\ &=\; \frac{\frac{1}{(2\pi\widehat{\sigma}_1^2)^{n/2}}\,\exp\Big\{-\overbrace{\frac{1}{2\widehat{\sigma}_1^2}\sum_{i=1}^{n}(x_i-\widehat{\mu}_1)^2}^{=n/2}\Big\}}{\frac{1}{(2\pi\widehat{\sigma}_0^2)^{n/2}}\,\exp\Big\{-\underbrace{\frac{1}{2\widehat{\sigma}_0^2}\sum_{i=1}^{n}(x_i-\widehat{\mu}_0)^2}_{=n/2}\Big\}} \end{aligned}$$

we see that the test statistic is strictly increasing in $\widehat{\sigma}_0^2/\widehat{\sigma}_1^2$. We obtain

$$\widehat{\sigma}_0^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}\big(x_i-\widehat{\mu}_0\big)^2 \;=\; \underbrace{\frac{1}{n}\sum_{i=1}^{n}\big(x_i-\bar{x}_n\big)^2}_{=:\,\widetilde{\sigma}_n^2} \;+\; \big(\bar{x}_n-\widehat{\mu}_0\big)^2$$

and, analogously,

$$\widehat{\sigma}_1^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}\big(x_i-\bar{x}_n\big)^2 \;+\; \big(\bar{x}_n-\widehat{\mu}_1\big)^2.$$

This leads to

$$\begin{aligned} \frac{\widehat{\sigma}_0^2}{\widehat{\sigma}_1^2} \;&=\; \frac{\widetilde{\sigma}_n^2+(\bar{x}_n-\widehat{\mu}_0)^2}{\widetilde{\sigma}_n^2+(\bar{x}_n-\widehat{\mu}_1)^2} \\ &=\; \begin{cases} \frac{\widetilde{\sigma}_n^2+(\bar{x}_n-\mu_0)^2}{\widetilde{\sigma}_n^2}, & \text{if } \bar{x}_n\ge\mu_0, \\ \frac{\widetilde{\sigma}_n^2}{\widetilde{\sigma}_n^2+(\bar{x}_n-\mu_0)^2}, & \text{if } \bar{x}_n\le\mu_0. \end{cases} \end{aligned}$$

See see from this representation that $\widehat{\sigma}_0^2/\widehat{\sigma}_1^2$ is strictly increasing in $\frac{\bar{x}_n-\mu_0}{\widetilde{\sigma}_n}$ and therefore also in $T_n(x) := \frac{\bar{x}_n-\mu_0}{\widehat{\sigma}_n}$. Hence, the searched-for likelihood ratio test for $H_0'\colon\ \theta\in\Theta_0'$ versus $H_1'\colon\ \theta\in\Theta_1'$ is given by

$$\varphi(x) \;=\; \begin{cases} 1, & \text{if } T_n(x)\ge c, \\ 0, & \text{if } T_n(x)<c. \end{cases}$$

This test has size α if $c=t_{n-1,1-\alpha}$.

While most powerful tests derived from the Neyman-Pearson lemma provide an optimal tradeoff between size and power against all alternatives, we do not have such a property for likelihood ratio tests in general. In fact, when constructing a likelihood ratio test, we have a prescribed test statistic and the focus is only on the choice of the critical value such that the size of the test equals some given value. On the other hand, it seems to be natural when the probability of rejection under the null hypothesis is smaller than that under the possible alternatives. Such a property can actually be shown to hold for some tests derived by the likelihood ratio approach. We consider the example of a two-sided z test.

Example
Suppose that we observe realizations $x_1, \ldots, x_n$ of i.i.d. random variables $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2)$, where $\theta \in \Theta := \mathbb{R}$ is the parameter of interest, $\sigma^2 > 0$ is assumed to be known, and we wish to test

$$H_0\colon\ \theta = \theta_0 \qquad \text{vs.} \qquad H_1\colon\ \theta \neq \theta_0.$$

As already shown, a size α likelihood ratio test φ_α ($\alpha \in (0,1)$) is given by

$$\varphi_\alpha(x) \;=\; \begin{cases} 1, & \text{if } \left|\frac{\sqrt{n}(\bar{x}_n - \theta_0)}{\sigma}\right| \geq \Phi^{-1}(1-\alpha/2), \\ 0, & \text{if } \left|\frac{\sqrt{n}(\bar{x}_n - \theta_0)}{\sigma}\right| < \Phi^{-1}(1-\alpha/2). \end{cases}$$

While we have that

$$E_{\theta_0}\big[\varphi_\alpha(X)\big] \;=\; \alpha,$$

it follows that the power against any alternative is strictly greater than α. Since the density of a $\mathcal{N}(\theta, 1)$ distribution is strictly increasing on $(-\infty, \theta]$, strictly decreasing on $[\theta, \infty)$, and symmetric about θ we obtain, for all $\theta \neq \theta_0$,

$$P_\theta\big(-c < \bar{X}_n - \theta_0 < c\big) \;<\; P_\theta\big(-c < \bar{X}_n - \theta < c\big) \qquad \forall c > 0, \tag{3.17}$$

which implies that

$$\begin{aligned} E_\theta\big[\varphi_\alpha(X)\big] \;&=\; 1 - \underbrace{P_\theta\Big(-\Phi^{-1}(1-\alpha/2) < \frac{\sqrt{n}(\bar{X}_n - \theta_0)}{\sigma} < \Phi^{-1}(1-\alpha/2)\Big)}_{< P_\theta\left(-\Phi^{-1}(1-\alpha/2) < \frac{\sqrt{n}(\bar{X}_n - \theta)}{\sigma} < \Phi^{-1}(1-\alpha/2)\right) = 1-\alpha} \\ &>\; \alpha. \end{aligned}$$

We formalize this property by the following definition.

Definition 3.6. Suppose that φ is a test based on a realization x of a random variable $X \sim P_\theta$.

φ is said to be an **unbiased test** for $H_0\colon\ \theta \in \Theta_0$ versus $H_1\colon\ \theta \in \Theta_1$ if

$$\sup\big\{E_\theta\varphi(X)\colon\ \theta \in \Theta_0\big\} \;\leq\; \inf\big\{E_\theta\varphi(X)\colon\ \theta \in \Theta_1\big\}.$$

The following examples show that some commonly used test share the property of unbiasedness.

Examples

1) If φ based on X is a **uniformly most powerful** level α test for $H_0\colon\ \theta \in \Theta_0$ versus $H_1\colon\ \theta \in \Theta_1$, then φ is an unbiased test. Indeed, we have that

$$\sup\big\{E_\theta\varphi(X)\colon\ \theta\in\Theta_0\big\}\ \le\ \alpha.$$

Since $\bar\varphi$ given by $\bar\varphi(x)=\alpha$ for all x is also a level α test we obtain that

$$E_\theta\varphi(X)\ \ge\ E_\theta\bar\varphi(X)\ =\ \alpha \qquad \forall\theta\in\Theta_1,$$

that is, φ is unbiased.

2) **(Two-sided t test)**
Recall that, based on i.i.d. random variables $X_1,\dots,X_n\sim\mathcal{N}(\mu,\sigma^2)$, a size α likelihood ratio test φ for

$$H_0\colon\ \theta:=\begin{pmatrix}\mu\\ \sigma^2\end{pmatrix}\in\Theta_0=\{\mu_0\}\times(0,\infty) \quad \text{vs.} \quad H_1\colon\ \theta\in\Theta_1=(\mathbb{R}\setminus\{\mu_0\})\times(0,\infty)$$

is given by

$$\varphi(x)\ =\ \begin{cases}1, & \text{if } \left|\frac{\sqrt{n}(\bar x_n-\mu_0)}{\widehat\sigma}\right|\ \ge\ t_{n-1,1-\alpha/2},\\ 0, & \text{if } \left|\frac{\sqrt{n}(\bar x_n-\mu_0)}{\widehat\sigma}\right|\ <\ t_{n-1,1-\alpha/2}.\end{cases}$$

As already shown, the critical value $t_{n-1,1-\alpha/2}$ is chosen such that

$$E_\theta\varphi(X)\ =\ \alpha \qquad \forall\theta\in\Theta_0.$$

Let now $\theta\in\Theta_1$, i.e. $\mu\neq\mu_0$. Using the independence of $\bar X_n$ and $\widehat\sigma=\sqrt{\frac{1}{n-1}\sum_{i=1}^n(X_i-\bar X_n)^2}$ we obtain from (3.17) that

$$\begin{aligned}
E_\theta\varphi(X) &= P_\theta\Big(\Big|\frac{\sqrt{n}(\bar X_n-\mu_0)}{\widehat\sigma}\Big|\ \ge\ t_{n-1,1-\alpha/2}\Big)\\
&= \int_{(0,\infty)} P_\theta\big(\sqrt{n}|\bar X_n-\mu_0|\ \ge\ \widehat\sigma\, t_{n-1,1-\alpha/2}\big|\widehat\sigma=u\big)\,dP_\theta^{\widehat\sigma}(u)\\
&= \int_{(0,\infty)} \underbrace{P_\theta\big(\sqrt{n}|\bar X_n-\mu_0|\ \ge\ u\,t_{n-1,1-\alpha/2}\big)}_{>P_{\binom{\mu_0}{\sigma^2}}\left(\sqrt{n}|\bar X_n-\mu_0|\ge u\,t_{n-1,1-\alpha/2}\right)}\ d\ \underbrace{P_\theta^{\widehat\sigma}}_{=P_{\binom{\mu_0}{\sigma^2}}^{\widehat\sigma}}(u)\\
&> \int_{(0,\infty)} P_{\binom{\mu_0}{\sigma^2}}\big(\sqrt{n}|\bar X_n-\mu_0|\ \ge\ u\,t_{n-1,1-\alpha/2}\big)\,dP_{\binom{\mu_0}{\sigma^2}}^{\widehat\sigma}(u)\\
&= \ \dots\ =\ E_{\binom{\mu_0}{\sigma^2}}\varphi(X)\ =\ \alpha.
\end{aligned}$$

Therefore, φ is an unbiased test.

We would like to add that we can show by analogous arguments that one-sided z and t tests are also unbiased for the corresponding test problems with one-sided hypotheses and alternatives.

In what follows, we introduce one of the most important class of likelihood ratio tests, the so-called F tests. These tests are derived in the context of linear regression models, where the errors are assumed to be independent and identically distributed, following a normal distribution. Recall that such a linear regression model can be represented in vector/matrix form as

$$\underbrace{\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}}_{=:Y} = \underbrace{\begin{pmatrix} x_{11} & \dots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nk} \end{pmatrix}}_{=:X} \underbrace{\begin{pmatrix} \theta_1 \\ \vdots \\ \theta_k \end{pmatrix}}_{=:\theta} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{=:\varepsilon}.$$

$Y_1, \dots, Y_n$ are the dependent variables. The matrix X is called design matrix and contains the explanatory variables x_{ij}. We suppose throughout that the x_{ij} are nonrandom. The term design refers to the fact that, in case of a planned experiment, it it is related to the actual experimental design (the specific setting of the explanatory variables). The vector θ contains unknown parameters which specify the linear relationship between the explanatory variables and the corresponding dependent ones. The vector ε of errors is assumed to follow a multivariate normal distribution, $\varepsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$.

Typical hypotheses to be tested are e.g. $H_{0,i}\colon\ \theta_i = 0$ which means that the variables $x_{1i}, \dots, x_{ni}$ do not contribute to an explanation (prediction) of the respective dependent variables $Y_1, \dots, Y_n$. If such a hypothesis is actually true, then we could simplify the model by deleting the corresponding column of the design matrix and dropping the corresponding component of the vector θ. Another field of application of F tests is the comparison of different treatments. Suppose that an experiment is performed where k different treatments (drug A, drug B, placebo,...) are applied to respective groups of $n_1, \dots, n_k$ patients and that in each case a certain variable is measured which characterizes the success of the corresponding treatment. In this case, we could employ the regression model

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{k1} \\ \vdots \\ Y_{kn_k} \end{pmatrix} = \begin{pmatrix} \mathbb{1}_{n_1} & 0_{n_1} & \dots & \dots & 0_{n_1} \\ 0_{n_2} & \mathbb{1}_{n_2} & 0_{n_2} & \dots & 0_{n_2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0_{n_{k-1}} & \dots & 0_{n_{k-1}} & \mathbb{1}_{n_{k-1}} & 0_{n_{k-1}} \\ 0_{n_k} & \dots & \dots & 0_{n_k} & \mathbb{1}_{n_k} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \vdots \\ \varepsilon_{k1} \\ \vdots \\ \varepsilon_{kn_k} \end{pmatrix}.$$

Here $\mathbb{1}_l$ and 0_l denote the vectors of length l consisting of ones and zeroes, respectively. A common hypothesis that we may wish to test is that $\theta_1 = \ldots = \theta_k$ which means that there is no specific effect due to the different treatments.

A general framework which covers both cases mentioned above is given by the following formulation of the test problem. It is assumed that a realization x of a random vector X is available, where

$$X \sim \mathcal{N}\big(\gamma, \sigma^2 I_n\big),$$

and $\gamma \in \Gamma \subseteq \mathbb{R}^n$ and $\sigma^2 > 0$ are both unknown parameters. We intend to test

$$H_0\colon\ \begin{pmatrix} \gamma \\ \sigma^2 \end{pmatrix} \in \Gamma_0 \times (0, \infty) \quad \text{vs.} \quad H_1\colon\ \begin{pmatrix} \gamma \\ \sigma^2 \end{pmatrix} \in (\Gamma \setminus \Gamma_0) \times (0, \infty),$$

where

- Γ_0 is an l-dimensional subspace of $\mathbb{R}^n$,
- Γ is a k-dimensional subspace of $\mathbb{R}^n$,
- $\Gamma_0 \subset \Gamma$,
- $0 \leq l < k < n$. (If $l = 0$, then $\Gamma_0 := \{0_n\}$.)

The test statistic $L(X)$ for a likelihood ratio test is given by

$$L(x) \;=\; \frac{\sup\{p_{\gamma,\sigma^2}(x)\colon\ \gamma \in \Gamma \setminus \Gamma_0,\, \sigma^2 > 0\}}{\sup\{p_{\gamma,\sigma^2}(x)\colon\ \gamma \in \Gamma_0,\, \sigma^2 > 0\}},$$

where p_{γ,σ^2} is the density of a $\mathcal{N}(\gamma, \sigma^2 I_n)$ distribution, i.e.

$$p_{\gamma,\sigma^2}(x) \;=\; \frac{1}{(2\pi\sigma^2)^{n/2}}\, \exp\Big\{ -\frac{1}{2\sigma^2}\|x-\gamma\|^2\Big\},$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^n$. Since $(\Gamma\setminus\Gamma_0)\times(0,\infty)$ is dense in $\Gamma\times(0,\infty)$ and $\binom{\gamma}{\sigma^2} \mapsto p_{\gamma,\sigma^2}(x)$ is continuous for all $x \in \mathbb{R}^n$ we obtain that $L(x) = \widetilde{L}(x)$, where

$$\widetilde{L}(x) \;=\; \frac{\sup\{p_{\gamma,\sigma^2}(x)\colon\ \gamma \in \Gamma,\, \sigma^2 > 0\}}{\sup\{p_{\gamma,\sigma^2}(x)\colon\ \gamma \in \Gamma_0,\, \sigma^2 > 0\}}.$$

Next we determine the maximum likelihood estimators for $\binom{\gamma}{\sigma^2}$ in the cases where this parameter varies in $\Gamma \times (0,\infty)$ and $\Gamma_0 \times (0,\infty)$, respectively.

a) MLE in $\Gamma \times (0,\infty)$
For arbitrary $\sigma^2 > 0$, the function $\gamma \mapsto p_{\gamma,\sigma^2}(x)$ is maximized by

$$\widehat{\gamma} \;=\; Px,$$

where Px is the orthogonal projection of x onto Γ. (P is the $(n \times n)$-orthogonal projection matrix onto Γ.)
It follows by simple calculations that $\sigma^2 \mapsto p_{Px,\sigma^2}(x)$ is maximized by

$$\widehat{\sigma}^2 \;=\; \frac{1}{n}\big\|x - Px\big\|^2.$$

b) MLE in $\Gamma_0 \times (0,\infty)$
We obtain in complete analogy to a) that the maximum likelihood estimators in the smaller parameter space $\Gamma_0 \times (0,\infty)$ are given by

$$\widehat{\gamma}_0 \;=\; P_0 x,$$

where $P_0 x$ is the orthogonal projection of x onto Γ_0, and

$$\widehat{\sigma}_0^2 \;=\; \frac{1}{n}\big\|x - P_0 x\big\|^2.$$

Now we consider the test statistic $\widetilde{L}(X)$, and represent it in such a way that we recognize an equivalent statistic which has a textbook distribution under the null hypothesis.

$$\begin{aligned}
\widetilde{L}(x) &= \frac{p_{\widehat{\gamma},\widehat{\sigma}^2}(x)}{p_{\widehat{\gamma}_0,\widehat{\sigma}_0^2}(x)} \\
&= \frac{(2\pi\widehat{\sigma}^2)^{-n/2}\,\exp\big\{-\overbrace{\frac{1}{2\widehat{\sigma}^2}\,\|x-Px\|^2}^{=n/2}\big\}}{(2\pi\widehat{\sigma}_0^2)^{-n/2}\,\exp\big\{-\underbrace{\frac{1}{2\widehat{\sigma}_0^2}\,\|x-P_0x\|^2}_{=n/2}\big\}} \\
&= \left(\frac{\|x-P_0x\|^2}{\|x-Px\|^2}\right)^{n/2} \\
&= \left(\frac{\|x-P_0x\|^2\,-\,\|x-Px\|^2}{\|x-Px\|^2}\,+\,1\right)^{n/2} \\
&= \left(\frac{\|(P-P_0)x\|^2}{\|(I_n-P)x\|^2}\,+\,1\right)^{n/2}. \qquad (3.18)
\end{aligned}$$

The last equation holds since

$$\begin{aligned}
\|x-P_0x\|^2\,-\,\|x-Px\|^2 &= x^T(I_n-P_0)^T(I_n-P_0)x\,-\,x^T(I_n-P)^T(I_n-P)x \\
&= x^T(I_n-P_0)x\,-\,x^T(I_n-P)x \\
&= x^T(P-P_0)x\,=\,\|(P-P_0)x\|^2.
\end{aligned}$$

A slight modification of the right-hand side leads to the following representation of the searched-for likelihood ratio test.

$$\varphi(x)\,=\,\begin{cases} 1, & \text{if } \frac{\frac{1}{k-l}\|(P-P_0)x\|^2}{\frac{1}{n-k}\|(I_n-P)x\|^2}\,>\,c_\alpha, \\ \gamma_\alpha, & \text{if } \frac{\frac{1}{k-l}\|(P-P_0)x\|^2}{\frac{1}{n-k}\|(I_n-P)x\|^2}\,=\,c_\alpha, \\ 0, & \text{if } \frac{\frac{1}{k-l}\|(P-P_0)x\|^2}{\frac{1}{n-k}\|(I_n-P)x\|^2}\,<\,c_\alpha. \end{cases} \qquad (3.19a)$$

In order to obtain a size α test, we still have to find constants $c_\alpha \in \mathbb{R}$ and $\gamma_\alpha \in [0,1]$ such that

$$\sup\big\{E_{\binom{\gamma}{\sigma^2}}\varphi(X)\colon\ \gamma\in\Gamma_0,\,\sigma^2>0\big\}\,=\,\alpha. \qquad (3.19b)$$

To this end, we need the following definition.

Definition 3.7. Let X_1 and X_2 be independent and have χ^2_r and χ^2_s distributions, respectively. Then the distribution of

$$Y\,=\,\frac{X_1/r}{X_2/s}$$

is called F **distribution with** r **and** s **degrees of freedom**. We write $Y \sim F_{r,s}$.

In the following we identify the distribution of the statistic $\frac{\frac{1}{k-l}\|(P-P_0)X\|^2}{\frac{1}{n-k}\|(I_n-P)X\|^2}$ under $\binom{\gamma}{\sigma^2} \in \Gamma_0 \times (0,\infty)$ as such an F distribution. Since quantiles of these distributions are tabulated we can find appropriate constants such that (3.19b) is satisfied. Equipped with these constants, the test φ given by (3.19a) has size α.

Theorem 3.8. *Let $X \sim \mathcal{N}(\gamma, \sigma^2 I_n)$ and let Γ_0 and Γ be l- and k-dimensional subspaces of $\mathbb{R}^n$, respectively, where $\Gamma_0 \subset \Gamma$, $0 \leq l < k < n$. Furthermore, let P_0 and P be the orthogonal projection matrices onto Γ_0 and Γ, respectively. If $\gamma \in \Gamma_0$ and $\sigma^2 > 0$, then*

$$T(X) := \frac{\frac{1}{k-l}\|(P-P_0)X\|^2}{\frac{1}{n-k}\|(I_n-P)X\|^2} \sim F_{k-l,n-k}.$$

Before we proceed with the proof of this theorem, we recall a few basic facts about orthogonal projection matrices. Let P be a real orthogonal projection matrix of dimension $n \times n$. Then $P^2 = P = P^T$, i.e. the matrix P is idempotent and symmetric. We consider the spectral decomposition of P,

$$P = D^T \mathrm{Diag}(\lambda_1, \ldots, \lambda_n) D.$$

Here, $\{\lambda_1, \ldots, \lambda_n\}$ are the eigenvalues of P (according to their multiplicity) and $D = (e_1, \ldots, e_n)^T$, where $\{e_1, \ldots, e_n\}$ is an orthonormal system of eigenvectors corresponding to $\lambda_1, \ldots, \lambda_n$. By idempotence of P we obtain that

$$P^2 = D^T \mathrm{Diag}(\lambda_1^2, \ldots, \lambda_n^2) D = P = D^T \mathrm{Diag}(\lambda_1, \ldots, \lambda_n) D,$$

which implies $\lambda_i^2 = \lambda_i$ and therefore $\lambda_i \in \{0, 1\}$ for all $i = 1, \ldots, n$. Moreover, the number of nonzero eigenvalues is equal to the rank of P.

To summarize, if P is an $(n \times n)$ orthogonal projection matrix of rank m, then we can represent it as

$$P = D^T \mathrm{Diag}(\underbrace{1, \ldots, 1}_{m \text{ times}}, \underbrace{0, \ldots, 0}_{n-m \text{ times}}) D = \sum_{i=1}^{m} e_i e_i^T,$$

where $D = (e_1, \ldots, e_n)^T$, $\{e_1, \ldots, e_m\}$ is an arbitrary orthonormal system of eigenvectors to the eigenvalue 1, and $\{e_{m+1}, \ldots, e_n\}$ is an arbitrary orthonormal system of eigenvectors to the eigenvalue 0.

If in particular P_0 and P are orthogonal projection matrices on the respective subspaces Γ_0 and Γ_n of $\mathbb{R}^n$ such that $\Gamma_0 \subset \Gamma$, where the dimension of these subspaces is l and k, respectively, $(0 \leq l < k \leq n)$, then we can obtain the following representation of the corresponding projection matrices P_0 and P. First, we choose an orthonormal basis $\{e_1, \ldots, e_l\}$ of Γ_0. $e_1, \ldots, e_l$ are eigenvectors of P_0 to the eigenvalue 1 and we obtain that

$$P_0 = \sum_{i=1}^{l} e_i e_i^T.$$

We can augment $\{e_1, \ldots, e_l\}$ by $e_{l+1}, \ldots, e_k$ such that $\{e_1, \ldots, e_k\}$ forms an orthonormal basis of Γ. The matrix P can be represented as

$$P = \sum_{i=1}^{k} e_i e_i^T.$$

Using these particular representations of P_0 and P we see that

$$P \, - \, P_0 \, = \, \sum_{i=l+1}^{k} e_i e_i^T$$

is also an orthogonal projection matrix which has rank $k - l$.

Proof of Theorem 3.8. Since, in case of $\gamma \in \Gamma_0$, $P_0\gamma = P\gamma = I_n\gamma = \gamma$ we obtain

$$T(X) \, = \, \frac{\frac{1}{k-l}\|(P - P_0)X\|^2}{\frac{1}{n-k}\|(I_n - P)X\|^2} \, = \, \frac{\frac{1}{k-l}\|(P - P_0)(X - \gamma)/\sigma\|^2}{\frac{1}{n-k}\|(I_n - P)\underbrace{(X - \gamma)/\sigma}_{=: Y}\|^2}. \tag{3.20}$$

The random vector Y has a $\mathcal{N}(0_n, I_n)$ distribution. Now we collect the following results.

a) Since $P - P_0$ is an orthogonal projection matrix of rank $k - l$ we obtain $(P - P_0)^T(P - P_0) = P - P_0$ and therefore, by Lemma 3.7,

$$\big\|(P - P_0)(X - \gamma)/\sigma\|^2 \, = \, Y^T(P - P_0)Y \, \sim \, \chi^2_{k-l}.$$

b) $I_n - P$ is an orthogonal projection matrix of rank $n - k$ which implies that

$$\big\|(I_n - P)(X - \gamma)/\sigma\|^2 \, = \, Y^T(I_n - P)Y \, \sim \, \chi^2_{n-k}.$$

c) Since $(I_n - P)(P - P_0)^T \, = \, P - P_0 - P^2 + \underbrace{PP_0}_{= P_0} = 0_{n\times n}$ it follows from Lemma 3.6 that the numerator and the denominator on the right-hand side of (3.20) are independent.

We obtain from a) to c) that the term on right-hand side of (3.20) has a structure as in our definition of an F distribution above, and we obtain that

$$T(X) \, \sim \, F_{k-l,n-k}.$$

□

Applications of F tests

After deriving the F test in a general framework we consider important applications. In all cases, we consider a linear regression model,

$$Y_i = \sum_{j=1}^{k} x_{ij}\beta_j + \varepsilon_i, \qquad \forall i = 1, \ldots, n,$$

and we impose the condition that the errors $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. and follow a normal distribution with mean zero and a common variance $\sigma^2 > 0$. The experimental design varies between the different fields of application.

(i) **Nested regression models**
We consider the general linear regression model

$$Y = X\beta + \varepsilon,$$

where

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{11} & \ldots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{nk} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

We impose the condition that $\varepsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$. We suppose that the design matrix X has full column rank k, where $1 < k < n$. A frequently encountered problem in practice is to test if one or more columns of the design matrix can be dropped which means that the corresponding components of the parameter β are zero. Suppose, for definiteness, that we wish to test if $\beta_i = 0$. The framework for a corresponding F test is given by

$$Y \sim \mathcal{N}(\gamma, \sigma^2 I_n),$$

$$H_0\colon \quad \begin{pmatrix} \gamma \\ \sigma^2 \end{pmatrix} \in \Gamma_0 \times (0, \infty) \qquad \text{vs.} \qquad H_1\colon \quad \begin{pmatrix} \gamma \\ \sigma^2 \end{pmatrix} \in (\Gamma \setminus \Gamma_0) \times (0, \infty),$$

where

$$\begin{aligned} \Theta_0 &= \{X\alpha\colon\ \alpha_i = 0,\ \alpha_1, \ldots, \alpha_{i-1}, \alpha_{i+1}, \ldots, \alpha_k \in \mathbb{R}\} \\ &= \{X_0\alpha\colon\ \alpha \in \mathbb{R}^{k-1}\}, \qquad X_0 = \begin{pmatrix} x_{11} & \ldots & x_{1,i-1} & x_{1,i+1} & \ldots & x_{1k} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{n,i-1} & x_{n,i+1} & \ldots & x_{nk} \end{pmatrix}, \\ \Theta &= \{X\alpha\colon\ \alpha \in \mathbb{R}^k\}. \end{aligned}$$

The orthogonal projections of Y onto Θ and Θ_0 are given by PY and P_0Y, respectively, where $P = X(X^TX)^{-1}X^T$ and $P_0 = X_0(X_0^TX_0)^{-1}X_0^T$. Since X has full column rank k and $1 < k < n$ a size α test φ_α of H_0 versus H_1 is given by

$$\varphi_\alpha(y) = \begin{cases} 1, & \text{if } \frac{\|(P-P_0)y\|^2}{\|(I_n-P)y\|^2/(n-k)} \geq F_{1,n-k;1-\alpha}, \\ 0, & \text{if } \frac{\|(P-P_0)y\|^2}{\|(I_n-P)y\|^2/(n-k)} < F_{1,n-k;1-\alpha}, \end{cases}$$

where $F_{1,n-k;1-\alpha}$ denotes the $(1-\alpha)$-quantile of an F distribution with 1 and $n-k$ degrees of freedom.

(ii) **Lack-of-fit test**

It is not uncommon that one is not sure about the adequacy of a linear regression model. The underlying idea of an appropriate test is the same as in the previous example, we compare the goodness-of-fit of the proposed model with that of a model which is known to be adequate. Suppose e.g. that we wish to test a polynomial regression model, where

$$Y_i = \sum_{j=1}^{k} x_i^{j-1}\beta_j + \varepsilon_i, \qquad i = 1, \ldots, n.$$

In this case and without additional information about the relationship between the explanatory variables and the response, a linear model which guaranteed to be adequate is not available in general. On the other hand, such a model exists if we have multiple experimental runs with the same settings for the explanatory variables, that is, $x_i = x_j$ for some pair(s) (i, j), $i \neq j$. For each setting x_i, we assume to have respectively n_i runs which leads to the regression model

$$Y_{ij} = \gamma_i + \varepsilon_{ij}, \qquad j = 1, \ldots, n_i,\ i = 1, \ldots, p.$$

In matrix/vector notation we obtain the model

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{p1} \\ \vdots \\ Y_{pn_p} \end{pmatrix} = \underbrace{\begin{pmatrix} \mathbb{1}_{n_1} & 0_{n_1} & \ldots & \ldots & 0_{n_1} \\ 0_{n_2} & \mathbb{1}_{n_2} & 0_{n_2} & \ldots & 0_{n_2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0_{n_{p-1}} & \ldots & 0_{n_{p-1}} & \mathbb{1}_{n_{p-1}} & 0_{n_{p-1}} \\ 0_{n_p} & \ldots & \ldots & 0_{n_p} & \mathbb{1}_{n_p} \end{pmatrix}}_{=X} \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_p \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \vdots \\ \varepsilon_{p1} \\ \vdots \\ \varepsilon_{pn_p} \end{pmatrix}.$$

On the other hand, if we believe that the relationship between an explanatory variable x_i and the corresponding response(s) Y_{ij} can be described by a polynomial model of order $k-1$, we could also use the model ($x_i \neq x_j$ for $i \neq j$)

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{p1} \\ \vdots \\ Y_{pn_p} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & x_1 & \ldots & x_1^{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1 & \ldots & x_1^{k-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_p & \ldots & x_p^{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_p & \ldots & x_p^{k-1} \end{pmatrix}}_{=X_0} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \vdots \\ \varepsilon_{p1} \\ \vdots \\ \varepsilon_{pn_p} \end{pmatrix}.$$

It follows from Exercise 2 (Problem sheet 1) that the matrix X_0 has full column rank k if there are at least k different x_i. However, if $k = p$, then the matrices X and X_0 have both rank k and the two matrices have the same image. In order to employ an F test for testing our polynomial model, we have to take care that $p > k$. Let Γ_0 and Γ be the subspaces of $\mathbb{R}^n$ which are spanned by the columns of the matrices X_0 and X, respectively, and let $P_0 = X_0(X_0^T X_0)^{-1} X_0^T$ and $P = X(X^T X)^{-1} X^T$ be the respective $(n \times n)$ projection matrices, where $n = n_1 + \cdots + n_p$.

It is possible to derive a simple explicit formula for the matrix P. Since $\{e_1, \ldots, e_p\}$ with $e_i = (\underbrace{0, \ldots, 0}_{n_1+\cdots+n_{i-1}}, \underbrace{1/\sqrt{n_i}, \ldots, 1/\sqrt{n_i}}_{n_i \text{ times}}, \underbrace{0, \ldots, 0}_{n_{i+1}+\cdots+n_p})^T$ is an orthonormal basis of Γ it follows that

$$P \;=\; \sum_{i=1}^{p} e_i e_i^T \;=\; \mathrm{DIAG}\big(D_1, \ldots, D_p\big) \tag{3.21}$$

is a block diagonal matrix, i.e. a block matrix that is a square matrix such that the main-diagonal blocks are square matrices and all off-diagonal blocks are zero matrices. The ith block D_i is an $(n_i \times n_i)$-matrix with all entries equal to $1/n_i$. For the projection matrix P_0 we may use the above representation $P_0 = X_0(X_0^T X_0)^{-1} X_0^T$.

The test problem "polynomial model is adequate" versus "polynomial model is not adequate" can be described in the usual way: Given $Y \sim \mathcal{N}(\gamma, \sigma^2 I_n)$, we wish to test

$$H_0\colon \quad \theta = \begin{pmatrix} \gamma \\ \sigma^2 \end{pmatrix} \in \Gamma_0 \times (0, \infty) \qquad \text{vs.} \qquad H_1\colon \quad \theta = \begin{pmatrix} \gamma \\ \sigma^2 \end{pmatrix} \in (\Gamma \setminus \Gamma_0) \times (0, \infty),$$

where

$$\Gamma_0 \;=\; \big\{X_0\alpha\colon\ \alpha \in \mathbb{R}^k\big\} \qquad \text{and} \qquad \Gamma \;=\; \big\{X\alpha\colon\ \alpha \in \mathbb{R}^p\big\}.$$

We have that $\Gamma_0 \subset \Gamma$, $\dim(\Gamma_0) = k$ and $\dim(\Gamma) = p$, which leads to a size α test φ_α such that

$$\varphi_\alpha(x) \;=\; \begin{cases} 1, & \text{if } \frac{\|(P-P_0)x\|^2/(p-k)}{\|(I_n-P)x\|^2/(n-p)} \geq F_{p-k,n-p;1-\alpha}, \\ 0, & \text{if } \frac{\|(P-P_0)x\|^2/(p-k)}{\|(I_n-P)x\|^2/(n-p)} < F_{p-k,n-p;1-\alpha} \end{cases}$$

Note that the denominator of the test statistic can be represented in an alternative form. It follows from (3.21) that

$$PY \;=\; \begin{pmatrix} \bar{Y}_{1\cdot}\mathbb{1}_{n_1} \\ \vdots \\ \bar{Y}_{p\cdot}\mathbb{1}_{n_p} \end{pmatrix},$$

where $\bar{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$. We have in particular

$$\big\|(I_n - P)Y\big\|^2 \;=\; \sum_{i=1}^{p}\sum_{j=1}^{n_i}\big(Y_{ij} - \bar{Y}_{i\cdot}\big)^2,$$

which implies that

$$E_\theta\big[\big\|(I_n-P)Y\big\|^2/(n-p)\big] \;=\; \frac{1}{n-p}\sum_{i=1}^{p}\underbrace{\sum_{j=1}^{n_i} E_\theta\big[\big(Y_{ij} - \bar{Y}_{i\cdot}\big)^2\big]}_{=(n_i-1)\sigma^2} \;=\; \sigma^2 \qquad \forall \theta \in \Gamma\times(0,\infty),$$

i.e., the denominator of the test statistic is an unbiased estimator of σ^2.

(iii) **Analysis of variance (ANOVA)**

In examples (i) and (ii) we have seen applications of F tests in the context of linear regression models in which the explanatory variables are usually quantitative. Now we consider so-called p-sample problems in which the explanatory variables are qualitative.

To fix ideas, suppose we are interested in comparing the performance of $p \geq 2$ treatments on a population (e.g. different medical drugs administered to groups of patients, different fertilizers given to plants,...). We suppose that we administer only one treatment to each subject and n_i subjects get treatment i, $1 \leq i \leq p$, $n_1 + \cdots + n_p = n$. The effect of the treatment given to the jth individual from the ith group is measured by some quantitative feature $y_{i,j}$ which is modeled as a realization of a random variable Y_{ij}. This leads to a regression model

$$Y_{ij} \,=\, \theta_i \,+\, \varepsilon_{ij}, \qquad j = 1, \ldots, n_i, \;\; i = 1, \ldots, p.$$

We assume that the errors ε_{ij} have zero mean. The parameters $\theta_1, \ldots, \theta_p$ characterize the respective effects of the different treatments and a common hypothesis that we may wish to test is that $\theta_1 = \ldots = \theta_p$ which means that there is no specific effect due to the different treatments. To derive an appropriate test, we impose the additional condition that the errors ε_{ij} are independent $\mathcal{N}(0, \sigma^2)$ variables. This might be justified (to some extent) on the basis of central limit behavior, experience, and hope. To find an appropriate test, we rewrite the above regression model in vector/matrix form:

$$\underbrace{\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{p1} \\ \vdots \\ Y_{pn_p} \end{pmatrix}}_{=Y} = \underbrace{\begin{pmatrix} \mathbb{1}_{n_1} & 0_{n_1} & \ldots & \ldots & 0_{n_1} \\ 0_{n_2} & \mathbb{1}_{n_2} & 0_{n_2} & \ldots & 0_{n_2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0_{n_{p-1}} & \ldots & 0_{n_{p-1}} & \mathbb{1}_{n_{p-1}} & 0_{n_{p-1}} \\ 0_{n_p} & \ldots & \ldots & 0_{n_p} & \mathbb{1}_{n_p} \end{pmatrix}}_{=X} \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \vdots \\ \varepsilon_{p1} \\ \vdots \\ \varepsilon_{pn_p} \end{pmatrix}.$$

Here $\mathbb{1}_l$ and 0_l denote the vectors of length l consisting of ones and zeroes, respectively. The random vector Y has a $\mathcal{N}(\gamma, \sigma^2 I_n)$ distribution. The test problem is given by

$$H_0\colon \quad \begin{pmatrix} \gamma \\ \sigma^2 \end{pmatrix} \in \Gamma_0 \times (0, \infty) \qquad \text{vs.} \qquad H_1\colon \quad \begin{pmatrix} \gamma \\ \sigma^2 \end{pmatrix} \in \big(\Gamma \setminus \Gamma_0\big) \times (0, \infty),$$

where $\Gamma_0 = \{\alpha \mathbb{1}_n\colon \, \alpha \in \mathbb{R}\}$ and Γ is spanned by the columns of the matrix X. The test statistic for an appropriate F test is given by

$$T(Y) \,=\, \frac{\big\|(P - P_0)Y\big\|^2/(p-1)}{\big\|(I_n - P)Y\big\|^2/(n-p)},$$

where P and P_0 are the orthogonal projection matrices onto Γ and Γ_0, respectively. It is not difficult to find simple explicit formulas for these projections. As above, it

follows from (3.21) that

$$PY \;=\; \begin{pmatrix} \bar{Y}_{1\cdot}\mathbb{1}_{n_1} \\ \vdots \\ \bar{Y}_{p\cdot}\mathbb{1}_{n_p}, \end{pmatrix}.$$

where $\bar{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$. Analogously, we can see that

$$P_0 Y \;=\; \bar{Y}_{\cdot\cdot}\mathbb{1}_n,$$

where $\bar{Y}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n_i} Y_{ij}$. Therefore we can represent the test statistic in the form

$$T(Y) \;=\; \frac{\frac{1}{p-1}\sum_{i=1}^{p} n_i\big(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\big)^2}{\frac{1}{n-p}\sum_{i=1}^{p}\sum_{j=1}^{n_i}\big(Y_{ij} - \bar{Y}_{i\cdot}\big)^2},$$

and a size α test φ_α for H_0 versus H_1 is given by

$$\varphi(y) \;=\; \begin{cases} 1, & \text{if } T(y) \geq F_{p-1,n-p;1-\alpha}, \\ 0, & \text{if } T(y) < F_{p-1,n-p;1-\alpha}. \end{cases}$$

But why are methods such as the above test called "analysis of variance"? We can break up the sum of squares as follows:

$$\begin{aligned} &\underbrace{\sum_{i=1}^{p}\sum_{j=1}^{n_i}\big(Y_{ij} - \bar{Y}_{\cdot\cdot}\big)^2}_{=SST} \\ &\quad= \;\big\|(I_n - P + P - P_0)Y\big\|^2 \\ &\quad= \;\big\|(I_n - P)Y\big\|^2 + \big\|(P - P_0)Y\big\|^2 \\ &\quad= \;\underbrace{\sum_{i=1}^{p} n_i\big(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\big)^2}_{=SSA} + \underbrace{\sum_{i=1}^{p}\sum_{j=1}^{n_i}\big(Y_{ij} - \bar{Y}_{i\cdot}\big)^2}_{=SSR}, \end{aligned}$$

i.e., the total sum of squares SST is broken up into two sources of variation, the variation among the p groups (measured by SSA) and the variation within each group of observations (measured by SSR). The above test is based on a comparison of these two sources of variation which also explains the name analysis of variance (ANOVA).

So far, we imposed the condition that the dependent variables $Y_{11}, \ldots, Y_{1n_1}, \ldots, Y_{p1}, \ldots, Y_{pn_p}$ are independent and normally distributed with a common variance σ^2. This allowed us to show that the test statistic has a known distribution under the null hypothesis, and the critical value could be chosen such that the test has the desired size α. Of course, the assumption that the errors are normally distributed is at best approximately adequate when such a test is used in practice. We show in the following that an F test has also an **asymptotic** justification when the sample size n is large. To set up an appropriate framework for our asymptotic considerations we assume that we have p independent sequences $(Y_{i,j})_{j\in\mathbb{N}}$, $i = 1, \ldots, p$, of i.i.d. random variables, where $EY_{ij} = \theta_i$, $\operatorname{var}(Y_{ij}) = \sigma^2$. In order to apply a central limit theorem, we suppose that all subsample sizes

$n_1, \ldots, n_p$ tend to infinity. To fix this idea, we suppose that $n_i = n_i(n)$ such that $n_i(n) \underset{n\to\infty}{\longrightarrow} \infty$ for all $i = 1, \ldots, p$. (We do not assume that $n_i(n)/n \underset{n\to\infty}{\longrightarrow} c_i$ for some $c_i \in (0,1)$ and all i, i.e., we do not exclude different rates of growth of the subsample sizes.) To describe our asymptotic consideration in a transparent way, we equip the relevant quantities with the additional subindex n, i.e.,

$$T_n(Y) \;=\; \frac{\big\|(P_n - P_{n0})Y\big\|^2/(p-1)}{\big\|(I_n - P_n)Y\big\|^2/(n-p)}.$$

The following theorem states the asymptotic correctness of the F test.

Theorem 3.9. *Suppose that $(Y_{i,j})_{j\in\mathbb{N}}$, $i = 1, \ldots, p$, are independent sequences of i.i.d. random variables, where $EY_{ij} = \theta_i$, $\mathrm{var}(Y_{ij}) = \sigma^2$. If $\theta_1 = \ldots = \theta_p =: \theta$, then*

(i) $T_n(Y) \overset{d}{\longrightarrow} T_\infty \sim \frac{1}{p-1}\chi^2_{p-1}$,

(ii) $P\big(T_n \geq F_{p-1,n-p;1-\alpha}\big) \underset{n\to\infty}{\longrightarrow} \alpha \qquad \forall \alpha \in (0,1)$.

Proof.

(i) We have that

$$T_n(Y) \;=\; \frac{\frac{1}{p-1}\sum_{i=1}^p n_i\big(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\big)^2}{\frac{1}{n-p}\sum_{i=1}^p \sum_{j=1}^{n_i}\big(Y_{ij} - \bar{Y}_{i\cdot}\big)^2} \;=:\; \frac{T_{n1}}{T_{n2}}.$$

We consider the numerator T_{n1} and the denominator T_{n2} separately. First we prove that

$$S_n \;:=\; \begin{pmatrix} \sqrt{n_1(n)}(\bar{Y}_{1\cdot} - \theta)/\sigma \\ \vdots \\ \sqrt{n_p(n)}(\bar{Y}_{p\cdot} - \theta)/\sigma \end{pmatrix} \overset{d}{\longrightarrow} Z \sim \mathcal{N}(0_p, I_p). \tag{3.22}$$

However, to prove (3.22), we have to be careful. We cannot use a classical multivariate central limit theorem since the subsample sizes $n_1(n), \ldots, n_p(n)$ can vary independently from each other and the left-hand side S_n of (3.22) cannot be rewritten in the form $\frac{1}{\sqrt{m}}\sum_{j=1}^m Y_i$, for some i.i.d. vectors $Y_1, \ldots, Y_m$. However, we can prove asymptotic normality for the components of S_n and then use their independence to deduce that (3.22) holds true. In fact, it follows from the Lindeberg-Lévy central limit theorem that

$$S_{ni} \;:=\; \sqrt{n_i(n)}\frac{\bar{Y}_{i\cdot} - \theta}{\sigma} \;=\; \frac{1}{\sqrt{n_i(n)}}\sum_{j=1}^{n_i(n)} \frac{Y_{ij} - \theta}{\sigma} \;\overset{d}{\longrightarrow}\; Z_i \sim \mathcal{N}(0,1) \qquad \forall i = 1, \ldots, p.$$

Using the independence of the sequences $(Y_{1,j})_{j\in N}, \ldots, (Y_{p,j})_{j\in N}$ we obtain that

$$\varphi_{S_n}(t) \;=\; E\big[e^{it^T S_n}\big] \;=\; E\big[\prod_{j=1}^p e^{it_j S_{nj}}\big] \;=\; \prod_{j=1}^p E\big[e^{it_j S_{nj}}\big] \underset{n\to\infty}{\longrightarrow} \prod_{j=1}^p e^{-t_j^2/2} \;=\; e^{-\|t\|^2/2}$$

holds for all $t = (t_1, \ldots, t_p)^T \in \mathbb{R}^p$. Since pointwise convergence of characteristic functions is equivalent to convergence in distribution of corresponding random variables we obtain (3.22). Moreover, it follows from (3.22) that

$$D_n S_n \xrightarrow{d} Z \sim \mathcal{N}(0_p, I_p), \tag{3.23}$$

where $(D_n)_{n\in\mathbb{N}}$ is an arbitrary sequence of orthogonal matrices. Indeed, since (3.22) implies uniform convergence of the corresponding functions on bounded sets, i.e.

$$\sup_{t:\, \|t\|\le K} \left| E e^{it^T S_n} - e^{-\|t\|^2/2} \right| \underset{n\to\infty}{\longrightarrow} 0,$$

for all $K < \infty$, we obtain that

$$E e^{it^T D_n S_n} = E e^{i(D_n^T t)^T S_n} \underset{n\to\infty}{\longrightarrow} e^{-\|D_n^T t\|^2/2} = e^{-\|t\|^2/2} \qquad \forall t \in \mathbb{R}^p,$$

which proves (3.23).

Next we have

$$\begin{pmatrix} \sqrt{n_1(n)}(\bar{Y}_{1\cdot} - \bar{Y}_{\cdot\cdot})/\sigma \\ \vdots \\ \sqrt{n_p(n)}(\bar{Y}_{p\cdot} - \bar{Y}_{\cdot\cdot})/\sigma \end{pmatrix}$$
$$= \underbrace{\left(I_p - \begin{pmatrix} \sqrt{n_1}/\sqrt{n} \\ \vdots \\ \sqrt{n_p}/\sqrt{n} \end{pmatrix} \left(\sqrt{n_1}/\sqrt{n} \cdots \sqrt{n_1}/\sqrt{n} \right) \right)}_{=: M_n} \begin{pmatrix} \sqrt{n_1(n)}(\bar{Y}_{1\cdot} - \theta)/\sigma \\ \vdots \\ \sqrt{n_p(n)}(\bar{Y}_{p\cdot} - \theta)/\sigma \end{pmatrix}.$$

Note that the matrix M_n is a $(p \times p)$-projection matrix of rank $p-1$, i.e., we have in particular $M_n^T M_n = M_n$ and $M_n = D_n^T \mathrm{Diag}(1, \ldots, 1, 0) D_n$, for some orthogonal matrix D_n. Since $y \mapsto \frac{1}{p-1} y^T \mathrm{Diag}(1, \ldots, 1, 0) y$ is a continuous function we obtain from (3.23) and the continuous mapping theorem

$$\begin{aligned} T_{n1} &= \frac{\sigma^2}{p-1} \left\| \begin{pmatrix} \sqrt{n_1(n)}(\bar{Y}_{1\cdot} - \bar{Y}_{\cdot\cdot})/\sigma \\ \vdots \\ \sqrt{n_p(n)}(\bar{Y}_{p\cdot} - \bar{Y}_{\cdot\cdot})/\sigma \end{pmatrix} \right\|^2 \\ &= \frac{\sigma^2}{p-1} \begin{pmatrix} \sqrt{n_1(n)}(\bar{Y}_{1\cdot} - \theta)/\sigma \\ \vdots \\ \sqrt{n_p(n)}(\bar{Y}_{p\cdot} - \theta)/\sigma \end{pmatrix}^T M_n \begin{pmatrix} \sqrt{n_1(n)}(\bar{Y}_{1\cdot} - \theta)/\sigma \\ \vdots \\ \sqrt{n_p(n)}(\bar{Y}_{p\cdot} - \theta)/\sigma \end{pmatrix} \\ &= \frac{\sigma^2}{p-1} (D_n S_n)^T \,\mathrm{Diag}(1, \ldots, 1, 0)\, D_n S_n \\ &\xrightarrow{d} \frac{\sigma^2}{p-1} Z^T \mathrm{Diag}(1, \ldots, 1, 0) Z. \end{aligned}$$

We obtain from Lemma 3.7 that

$$Z^T \mathrm{Diag}(1, \ldots, 1, 0) Z \sim \chi^2_{p-1},$$

which implies that

$$T_{n1} \xrightarrow{d} \frac{\sigma^2}{p-1} \chi^2_{p-1}. \tag{3.24}$$

It follows from the strong law of large numbers that

$$T_{n2} \;=\; \frac{1}{n-p}\sum_{i=1}^{p}(n_i-1)\;\underbrace{\frac{1}{n_i-1}\sum_{j=1}^{n_i}\big(Y_{ij}-\bar{Y}_{i\cdot}\big)^2}_{\xrightarrow{a.s.}\,\sigma^2}\;\xrightarrow{a.s.}\;\sigma^2,$$

which implies that

$$T_n(Y) \;=\; \frac{T_{n1}}{T_{n2}} \;\xrightarrow{d}\; T_\infty \sim \frac{1}{p-1}\chi^2_{p-1}.$$

(ii) Let $\alpha \in (0,1)$ be arbitrary. Since T_∞ has a continuous distribution function we obtain from (i)

$$\sup_x \big|P\big(T_n(Y)\geq x\big) \;-\; P\big(T_\infty \geq x\big)\big| \underset{n\to\infty}{\longrightarrow} 0.$$

Note that (i) holds in particular when $Y_{ij} \sim \mathcal{N}(\theta,\sigma^2)$. Let T_n^0 be the test statistic with such normally distributed Y_{ij}. Then

$$\sup_x \big|\big(T_n^0\geq x\big) \;-\; P\big(T_\infty \geq x\big)\big| \underset{n\to\infty}{\longrightarrow} 0.$$

Furthermore, we have that $T_n^0 \sim F_{p-1,n-p}$, which implies that $P\big(T_n^0 \geq F_{p-1,n-p;1-\alpha}\big) = \alpha$. Therefore, we obtain

$$\begin{aligned}
&\big|P\big(T_n(Y)\geq F_{p-1,n-p;1-\alpha}\big) \;-\; \alpha\big| \\
&\quad\leq\quad \big|P\big(T_n(Y)\geq F_{p-1,n-p;1-\alpha}\big) \;-\; P\big(T_\infty \geq F_{p-1,n-p;1-\alpha}\big)\big| \\
&\qquad\quad + \big|P\big(T_\infty \geq F_{p-1,n-p;1-\alpha}\big) \;-\; \alpha\big| \\
&\underset{n\to\infty}{\longrightarrow}\quad 0.
\end{aligned}$$

□

We would like to mention that there exist extensions of the so-called **one-way ANOVA** model described above. The **two-way analysis of variance (ANOVA)** examines the influence of two different categorical variables on one continuous dependent variable. The corresponding regression model is

$$Y_{ijk} \;=\; \mu \;+\; \alpha_i \;+\; \beta_j \;+\; \varepsilon_{ijk}, \qquad i=1,\ldots,p,\; j=1,\ldots,q,\; k=1,\ldots,n_{ij},$$

where the side conditions $\sum_i \alpha_i = \sum_j \beta_j = 0$ ensure that the parameters are clearly determined. Common hypotheses that may be tested are $\alpha_1 = \ldots = \alpha_p = 0$ and $\beta_1 = \ldots = \beta_q = 0$. It is also possible to take interaction effects between the independent variables into account which leads to the model

$$Y_{ijk} \;=\; \mu \;+\; \alpha_i \;+\; \beta_j \;+\; \gamma_{ij} \;+\; \varepsilon_{ijk}, \qquad i=1,\ldots,p,\; j=1,\ldots,q,\; k=1,\ldots,n_{ij}.$$

Under the side conditions $\sum_i \alpha_i = \sum_j \beta_j = \sum_{i,j} \gamma_{ij} = 0$, μ is the grand mean, α_i is the additive main effect of level i from the first factor, β_j is the additive main effect of level j from the second factor, and γ_{ij} is the non-additive interaction effect of treatment (i,j) from both factors. In all of these cases, likelihood ratio tests can be derived and it turns out that these tests are again F tests.