NMSA331 Mathematical statistics 1

Lecture notes

Last updated October 12, 2021.

Department of Probability and Mathematical Statistics
Faculty of Mathematics and Physics, Charles University

Contents

The list of symbols

$\boldsymbol{a}^\mathsf{T}$	the vector $\boldsymbol{a}$ transposed
$\boldsymbol{a}^{\otimes 2}$	$\boldsymbol{a}\boldsymbol{a}^\mathsf{T}$
$\|\boldsymbol{a}\|$	the Euclidean norm of the vector $\boldsymbol{a}$
$\xrightarrow{\mathsf{P}}$	convergence in probability
$\xrightarrow{\mathsf{a.s.}}$	convergence almost surely
$\xrightarrow{\mathsf{d}}$	convergence in distribution
$X \sim \mathcal{L}$	X has the exact distribution $\mathcal{L}$
$X \overset{\mathsf{as.}}{\sim} \mathcal{L}$	X has an asymptotic distribution $\mathcal{L}$
α	level of the test
$\beta_n(F), \beta_n(\theta)$	power of the test, powerfunction
γ_3	skewness random variable
γ_4	kurtosis random variable
$\widehat{\gamma}_4$	empirical kurtosis
Θ	parametric space
Θ_0	null hypothesis
Θ_1	alternative hypothesis
λ	Lebesgue measure on $\mathbb{R}$
μ_S	counting measure on a countable S
μ_k	k-th central moment of the random random variable
$\widehat{\mu}_k$	empirical odhad of the k-th central moment
μ'_k	k-tý moment random variable
$\widehat{\mu}'_k$	empirical odhad k-tého momentu
σ_X^2	the variance of the random variable X
$\widehat{\sigma}_n^2$	empirical estimator of variance
$\widehat{\Sigma}_n$	sample variance matrix
φ	the density of the standard normal distribution
Φ	the cumulative distribution function of the standard normal distribution

$\chi^2_f(\alpha)$	α-quantile of χ^2-distribution with f degress of freedom
Ω	the probability space
$\mathbb{1}_B$	the indicator of the set B
$\mathbf{1}_n$	the column vector of ones of the length n
$\mathcal{A}$	σ-algebra náhodných jevů na Ω
$\mathcal{B}_0$	Borel σ-algebra on $\mathbb{R}$
$\mathcal{B}_0^n$	Borel σ-algebra on $\mathbb{R}^n$
$C,\ C(\alpha)$	critical region of the test
$c_L(\alpha),\ c_U(\alpha)$	critical values
$\mathsf{cov}\,(X_1, X_2)$	the covariance of the random variables X_1 and X_2
$\mathsf{cov}\,(\boldsymbol{X}_1, \boldsymbol{X}_2)$	the covariance matrix of the random vectors $\boldsymbol{X}_1$ a $\boldsymbol{X}_2$
$\mathsf{diag}\,(\boldsymbol{a})$	diagonal matrix with the components of the vector $\boldsymbol{a}$ on the diagonal
$\mathsf{E}\,X$	expected value of the random variable (vector) X
$\mathcal{F}$	the model for the observed data
$\mathcal{F}_0$	distribution under the null hypothesis
$\mathcal{F}_1$	distribution under the alternative hypothesis
f_X	density of the random variable (vector) X
F_X	cumulative distribution function of the random variable (vector) X
F_X^{-1}	quantile function of the random variable X
$\widehat{F}_n$	empirical cumulative distribuiton function
$F_{m,n}(\alpha)$	α-quantile distribution $F_{m,n}$
H_0	null hypothesis
H_1	alternative hypothesis
$\mathbb{I}_n$	$n \times n$ matrix of identity
$\mathcal{L}^p$	the set of random varialbes on $(\Omega, \mathcal{A}, \mathsf{P})$ with the finite pth absolute moment
$\mathcal{L}^2_+$	the set of random varialbes on $(\Omega, \mathcal{A}, \mathsf{P})$ with finite and nonzero variance
$\mathcal{L}(X)$	distribution random variable (vector) X
m_X	median of the random variable X
$\widehat{m}_n$	sample median
MSE	mean squared error

P	probability
P_X	distribution random of the random variable X, i.e. the measure induced by thi random variabl
P_θ	distribution when the true value of the parameter is θ
$h(\mathbb{A})$	rank of matrix $\mathbb{A}$
$\mathbb{R}$	set of real numbers
R_i	the rank of the i-th observation
SE	standard error
S_n^2	sample variance
S_{jm}	sample covariance of the jth and the mth component of the random vector
S_X	support of distribution of the random variable X
$t_f(\alpha)$	α-quantile of the distribution t_f
$\mathsf{tr}\,(\mathbb{A})$	trace of the matrix $\mathbb{A}$
$u_X(\alpha)$	α-quantile of the random variable X
u_α	α-quantile of the distribution $\mathsf{N}(0,1)$
$\widehat{u}_n(\alpha)$	sample α-quantile
$\mathsf{var}\,X$	variance of the random variable X
$\mathsf{var}\,\boldsymbol{X}$	variance matrix of the random vector $\boldsymbol{X}$
$\mathcal{X}$	sample space
$X_{(k)}$	the k-th order statistics
$\overline{X}_n$	sample mean of $X_1, \ldots, X_n$

1. Clippings from the asymptotic theory

1.1. The convergence of random vectors

Let $\boldsymbol{X}$ be a k-dimensional random vector (with the cumulative distribution function $F_{\boldsymbol{X}}$) and $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$ be a sequence of k-dimensional random vectors (with the cumulative distribution functions $F_{\boldsymbol{X}_n}$).

Definition 1.1 We say that $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\mathsf{d}} \boldsymbol{X}$ (i.e. $\boldsymbol{X}_n$ converges *in distribution* to $\boldsymbol{X}$), if

$$\lim_{n\to\infty} F_{\boldsymbol{X}_n}(\boldsymbol{x}) = F_{\boldsymbol{X}}(\boldsymbol{x})$$

for each point $\boldsymbol{x}$ of the continuity of $F_{\boldsymbol{X}}$.

Let d be a metric in $\mathbb{R}^k$, e.g. the Euclidean metric $d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$.

Definition 1.2 We say that

- $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\mathsf{P}} \boldsymbol{X}$ (i.e. $\boldsymbol{X}_n$ converges *in probability* to $\boldsymbol{X}$), if

$$\forall \varepsilon > 0 \ \lim_{n\to\infty} \mathsf{P}\Big[\omega : d\big(\boldsymbol{X}_n(\omega), \boldsymbol{X}(\omega)\big) > \varepsilon\Big] = 0;$$

- $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{a.s.}} \boldsymbol{X}$ (i.e. $\boldsymbol{X}_n$ converges *almost surely* to $\boldsymbol{X}$), if

$$\mathsf{P}\Big[\omega : \lim_{n\to\infty} d(\boldsymbol{X}_n(\omega), \boldsymbol{X}(\omega)) = 0\Big] = 1.$$

Remark. For random vectors the convergence in probability and almost surely can be defined also component-wise. That is let $\boldsymbol{X}_n = (X_{n1}, \ldots, X_{nk})^{\mathsf{T}}$ and $\boldsymbol{X} = (X_1, \ldots, X_k)^{\mathsf{T}}$. Then

$$\boldsymbol{X}_n \xrightarrow[n\to\infty]{\mathsf{P}} \boldsymbol{X} \ (\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{a.s.}} \boldsymbol{X}) \quad \text{if} \quad X_{nj} \xrightarrow[n\to\infty]{\mathsf{P}} X_j \ (X_{nj} \xrightarrow[n\to\infty]{\text{a.s.}} X_j), \quad \forall j = 1, \ldots, k.$$

But this is not true for the convergence in distribution for which we have the Cramér-Wold tvrz that states

$$\boldsymbol{X}_n \xrightarrow[n\to\infty]{\mathsf{d}} \boldsymbol{X} \iff \boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{X}_n \xrightarrow[n\to\infty]{\mathsf{d}} \boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{X}, \quad \forall \boldsymbol{\lambda} \in \mathbb{R}^k.$$

Proposition 1.1 (**Continuous Mapping Theorem, CMT**) Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be continuous in each point of an open set $C \subset \mathbb{R}^k$ such that $\mathsf{P}(\boldsymbol{X} \in C) = 1$. Then

1. $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{a.s.}} \boldsymbol{X} \Rightarrow \boldsymbol{g}(\boldsymbol{X}_n) \xrightarrow[n\to\infty]{\text{a.s.}} \boldsymbol{g}(\boldsymbol{X})$;
2. $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\mathsf{P}} \boldsymbol{X} \Rightarrow \boldsymbol{g}(\boldsymbol{X}_n) \xrightarrow[n\to\infty]{\mathsf{P}} \boldsymbol{g}(\boldsymbol{X})$;
3. $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\mathsf{d}} \boldsymbol{X} \Rightarrow \boldsymbol{g}(\boldsymbol{X}_n) \xrightarrow[n\to\infty]{\mathsf{d}} \boldsymbol{g}(\boldsymbol{X})$.

Proposition 1.2 (**Cramér-Slutsky, CS**) Let $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\mathsf{d}} \boldsymbol{X}$, $\boldsymbol{Y}_n \xrightarrow[n\to\infty]{\mathsf{P}} \boldsymbol{c}$, then

1. $\boldsymbol{X}_n + \boldsymbol{Y}_n \xrightarrow[n\to\infty]{\mathsf{d}} \boldsymbol{X} + \boldsymbol{c}$;
2. $\boldsymbol{Y}_n \boldsymbol{X}_n \xrightarrow[n\to\infty]{\mathsf{d}} \boldsymbol{c}\,\boldsymbol{X}$,

where $\boldsymbol{Y}_n$ can be a sequence of random variables or vectors or matrices of appropriate dimensions ($\mathbb{R}$ or $\mathbb{R}^k$ or $\mathbb{R}^{m\times k}$) and analogously $\boldsymbol{c}$ can be either a number or a vector or a matrix of an appropriate dimension.

1.2. Basic asymptotic results

Proposition 1.3 (SLLN for i.id.) Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ be independent and identically distributed random vectors with a finite expectation $\mathsf{E}\,\boldsymbol{X}_i = \boldsymbol{\mu}$. Then

$$\overline{\boldsymbol{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \xrightarrow[n\to\infty]{\text{a.s.}} \boldsymbol{\mu}.$$

Proposition 1.4 (CLT for i.id.) Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ be independent and identically distributed random with the expectation $\mathsf{E}\,\boldsymbol{X}_i = \boldsymbol{\mu}$ and a finite variance matrix $\mathsf{var}\,\boldsymbol{X}_i = \boldsymbol{\Sigma}$. Then

$$\sqrt{n}\left(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}\right) \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}_k(\mathbf{0}_k, \boldsymbol{\Sigma}).$$

1.3. Δ-method

Let $\boldsymbol{T}_n = (T_{n1}, \ldots, T_{nk})^\mathsf{T}$ be an estimator of a k-dimensional parameter $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)^\mathsf{T}$ and $\boldsymbol{g} = (g_1, \ldots, g_m)^\mathsf{T}$ be a function from $\mathbb{R}^k \to \mathbb{R}^m$. Denote the Jacobi matrix of the function $\boldsymbol{g}$ at the point $\boldsymbol{x}$ as $\mathbb{D}_{\boldsymbol{g}}(\boldsymbol{x})$, i.e.

$$\mathbb{D}_{\boldsymbol{g}}(\boldsymbol{x}) = \begin{pmatrix} \nabla g_1(\boldsymbol{x}) \\ \vdots \\ \nabla g_m(\boldsymbol{x}) \end{pmatrix} = \begin{pmatrix} \frac{\partial g_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial g_1(\boldsymbol{x})}{\partial x_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_m(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial g_m(\boldsymbol{x})}{\partial x_k} \end{pmatrix}.$$

Proposition 1.5 (Δ-method) Let

$$\sqrt{n}\,(\boldsymbol{T}_n - \boldsymbol{\mu}) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}_k(\mathbf{0}_k, \boldsymbol{\Sigma}),$$

Further let $\boldsymbol{g} : A \to \mathbb{R}^m$, where $A \subset \mathbb{R}^k$, $\boldsymbol{\mu}$ is an interior point of A and the first-order partial derivatives of $\boldsymbol{g}$ are continuous in a neighbourhood of $\boldsymbol{\mu}$. Then

$$\sqrt{n}\,\big(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\mu})\big) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}_m\big(\mathbf{0}_m, \mathbb{D}_{\boldsymbol{g}}(\boldsymbol{\mu})\,\boldsymbol{\Sigma}\,\mathbb{D}_{\boldsymbol{g}}^{\mathsf{T}}(\boldsymbol{\mu})\big).$$

Theorem 1.5 is most often applied for $k = m = 1$ and $T_n = \overline{X}_n$, where $X_1, \ldots, X_n$ are i.i.d. random variables. Then by the central limit theorem

$$\sqrt{n}\,\big(\overline{X}_n - \mathsf{E}\,X_i\big) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}\big(0, \mathsf{var}\,(X_i)\big).$$

So if the function $g : \mathbb{R} \to \mathbb{R}$ has a continuous derivative in a neighbourhood of $\mu = \mathsf{E}\,X_i$, then

$$\sqrt{n}\big(g(\overline{X}_n) - g(\mu)\big) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}\big(0, [g'(\mu)]^2\,\mathsf{var}\,(X_i)\big). \tag{1.1}$$

Sometimes instead of (1.1) we write shortly $g(\overline{X}_n) \overset{\text{as}}{\approx} \mathsf{N}\big(g(\mu), \frac{[g'(\mu)]^2\,\mathsf{var}\,(X_i)}{n}\big)$. The quantity $\frac{[g'(\mu)]^2\,\mathsf{var}\,(X_i)}{n}$ is then called the **asymptotic variance** of $g(\overline{X}_n)$ and it is denoted as $\mathsf{avar}\,\big(g(\overline{X}_n)\big)$. Note that the asymptotic variance has to be understood as the **variance of the asymptotic distribution**, but not as some kind of a limiting variance.

As the following examples show for a sequence of random variables $\{Y_n\}$ the asymptotic variance $\mathsf{avar}\,(Y_n)$ may exist even if $\mathsf{var}\,(Y_n)$ does not exist for any $n \in \mathbb{N}$. Further even if $\mathsf{var}\,(Y_n)$ exists, then it **does not hold that** $\mathsf{var}\,(Y_n)/\mathsf{avar}\,(Y_n) \to 1$ as $n \to \infty$.

Example. Let $X \sim \mathsf{N}(0,1)$ and $\{\varepsilon_n\}$ be a sequence of random variables independent with X such that

$$\mathsf{P}(\varepsilon_n = -\sqrt{n}) = \tfrac{1}{2n}, \qquad \mathsf{P}(\varepsilon_n = 0) = 1 - \tfrac{1}{n}, \qquad \mathsf{P}(\varepsilon_n = \sqrt{n}) = \tfrac{1}{2n}.$$

Define $Y_n = X + \varepsilon_n$ and show that $Y_n \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}(0,1)$. Thus $\mathsf{avar}\,(Y_n) = 1$. On the other hand $\mathsf{var}\,(Y_n) = 2$ for each $n \in \mathbb{N}$.

Example. Suppose you have a random sample $X_1, \ldots, X_n$ from a Bernoulli distribution with parameter p_X and you are interested in estimating the logarithm of the odd, i.e. $\theta_X = \log\big(\frac{p_X}{1-p_X}\big)$. Compare the variance and the asymptotic variance of $\widehat{\theta}_X = \log\big(\frac{\overline{X}_n}{1-\overline{X}_n}\big)$.

2. Random sample

2.1. Definition of a random sample

Let the probability space $(\Omega, \mathcal{A}, P)$ be given.

Definition 2.1 *The random sample* from distribution F_X is defined as the sequence of $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ independent identically distributed random vectors defined on $(\Omega, \mathcal{A}, P)$ such that each random vector has a cumulative distribution F_X. The constant n is called *the sample size*.

The elements of random sample can be either real random variables or random vectors (matrices and so on). We can call them "observations" or "data". The whole random sample will be denoted as $\boldsymbol{X}$.

Remark. The true cumulative distribution function F_X from which our observations $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ comes are not known. We aim to use observations in order to learn something about F_X. We assume that the cumulative distribution F_X belongs to a set of distributions množiny distributions $\mathcal{F}$, which we call *the model*.

Definition 2.2 *The model* for the random sample $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ is a given set distributions $\mathcal{F}$ such that we assume that $F_X \in \mathcal{F}$.

Remark. The distribution F_X is unknown. Our goal is to use the observed data $\boldsymbol{X}$ in order to determine some characteristics of F_X that we call *parameters*. Formally the parameter is a constant (or a vector of constants) $\boldsymbol{\theta}_X \in \mathbb{R}^k$ that could be calculated if the distribution F_X was known. The parameter of interest thus can be written in the form $\boldsymbol{\theta}_X \equiv t(F_X)$, where t is a given functional.

Examples (Types of models for real random variables)**.**

1. The model $\mathcal{F}$ can be for instance the set of all distributions on $\mathbb{R}$ with a finite expectation (or a finite variance). The parameters of interest can be for instance $\mathsf{E}\, X_i$, $\mathsf{var}\, X_i$, $\mathsf{P}[X \leq x] \equiv F_X(x)$ or the quantile $F_X^{-1}(\alpha)$. Such a model is called non-parametric, as we cannot describe all the distributions in $\mathcal{F}$ with a finite number of parameters. By Θ we denote the set of possible values of $\boldsymbol{\theta} \equiv t(F)$ when $F \in \mathcal{F}$.
2. The model $\mathcal{F}$ can be the set of all distributions with densities (with respect to σ-finite measure) of the form $f(\boldsymbol{x}; \boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$, where $f(\cdot;\cdot)$ is a known function and $\boldsymbol{\theta}$ is an unknown constant (e.g. exponential distributions, normal distributions, geometric distributions). These models are called *parametric*. In

parametric models each parameter of interest $\theta_X = t(F_X)$ can be expressed as a function of the finite-dimensional parameter $\boldsymbol{\theta}$.

Examples (Parametric models)**.**

- $\mathcal{F} = \left\{\mathsf{N}(\mu, \sigma_0^2),\ \mu \in \mathbb{R},\ \sigma_0^2 \text{ be given}\right\}$; $\theta = \mu$, $\Theta = \mathbb{R}$.
- $\mathcal{F} = \left\{\mathsf{N}(\mu, \sigma^2),\ \mu \in \mathbb{R},\ \sigma^2 \in \mathbb{R}^+\right\}$; $\boldsymbol{\theta} = (\mu, \sigma^2)^\mathsf{T}$, $\Theta = \mathbb{R} \times \mathbb{R}^+$.
- $\mathcal{F} = \left\{\mathsf{Exp}(\lambda),\ \lambda \in \mathbb{R}^+\right\}$; $\theta = \lambda$, $\Theta = \mathbb{R}^+$.
- $\mathcal{F} = \left\{\mathsf{Alt}(p),\ p \in (0,1)\right\}$; $\theta = p$, $\Theta = (0,1)$.

Remark. We choose the model $\mathcal{F}$ and the parameter of interest $\boldsymbol{\theta}$. The model represents our apriori knowledge (not affected by the observed data) about the distributions of the random variables. The choice of the parameter depends on the question that we are trying answer by the statistical analysis. The choice of the model and parameter affects the choice of the method for the data analysis (as well as the obtained results).

2.2. Statistics

During statistical analysis we that from the random sample we calculate variables, that contain (summarize) information about the parameters of interests. These variables are called statistics. Consider the random sample $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$.

Definition 2.3 We call a *statistic* an arbitrary measurable function $\boldsymbol{S}(\boldsymbol{X})$ of observations calculated from the random sample $\boldsymbol{X}$. Statistic is a random variable (or a random vector).

A statistic cannot depend on the values that we do not know or that we do not observe. A statistic is a function of observed data (and known constants). The most commonly used statistics are the sample mean and the sample variance. To define them denote $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^\mathsf{T}$.

Definition 2.4

(i) A random variable $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is called a sample mean of the random sample $\boldsymbol{X}$.

(ii) Pro $n \geq 2$ the random variable $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ is called a sample variance of the random sample $\boldsymbol{X}$.

2.2.1. Properties of the sample mean

Consider the model $\mathcal{F} = \mathcal{L}^2$. I.e. we work with the random sample $\boldsymbol{X}$ whose components X_i are independent random variables with an arbitrary distribution with a finite second moment. Denote $\mu \equiv \mathsf{E}\, X_i$ a $\sigma^2 = \mathsf{var}\, X_i$.

Lemma 2.1

$$\overline{X}_n = \arg\min_{c\in\mathbb{R}} \sum_{i=1}^{n} (X_i - c)^2.$$

Proof. Introduce the function $f(c) = \sum_{i=1}^{n}(X_i - c)^2$. The statement of the lemma follows from the fact that $f'(\overline{X}_n) = 0$ and that $f''(c) > 0$ for each $c \in \mathbb{R}$. □

Theorem 2.2 (Properties of the sample mean)

(i) $\mathsf{E}\,\overline{X}_n = \mu$, $\mathsf{var}\,\overline{X}_n = \frac{\sigma^2}{n}$;

(ii) $\overline{X}_n \xrightarrow{\mathsf{P}} \mu$ as $n \to \infty$;

(iii) $\sqrt{n}\,(\overline{X}_n - \mu) \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}(0, \sigma^2)$, i.e. $\overline{X}_n \overset{\text{as.}}{\sim} \mathsf{N}\left(\mu, \frac{\sigma^2}{n}\right)$

Proof. (i) follows by the straightforward calculation. (ii) follow from the law of large numbers (Proposition 1.3 pro $k = 1$) and (iii) from the central limit theorem (Proposition 1.4 for $k = 1$). □

Remark. Suppose that the random variables in our sample are normally distributed, i.e. $\mathcal{F} = \left\{\mathsf{N}(\mu, \sigma^2),\ \mu \in \mathbb{R},\ \sigma^2 \in \mathbb{R}^+\right\}$. Then the statemetns (i) a (iii) of the previous proposition can be strengthened to

$$\sqrt{n}\,(\overline{X}_n - \mu) \sim \mathsf{N}(0, \sigma^2) \quad \text{i.e.} \quad \overline{X}_n \sim \mathsf{N}\left(\mu, \tfrac{\sigma^2}{n}\right).$$

Proof. From the assumptions it follows that the random vector $\boldsymbol{Z} = (X_1 - \mu, \ldots, X_n - \mu)^\mathsf{T}$ has independent components each of them having $\mathsf{N}(0, \sigma^2)$ distribution. By the defition of the multivariate normal distribution it follows that $\boldsymbol{Z} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2 \mathbb{I}_n)$. Denote $\boldsymbol{c} = \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)^\mathsf{T} \in \mathbb{R}^n$. Now from the properties of the multivariate normal distribution it follows that

$$\boldsymbol{c}^\mathsf{T}\boldsymbol{Z} = \sqrt{n}\,(\overline{X}_n - \mu) \sim \mathsf{N}(0, \sigma^2).$$

□

2.2.2. Relative (empirical) frequency

In applications often the random variable X_i takes only two values usually denoted as 0 and 1. The number one then means that in the ith trial an event B has occurred and the number zero otherwise. Denote $p = \mathsf{P}(X_i = 1)$. Then random variables $X_1, \ldots, X_n$ represent a random sample from the Bernoulli distribution $\mathsf{Be}(p)$.

The sample mean $\overline{X}_n$ is now empirical (or relative) frequency of the event B. Thus Theorem 2.2 immediately implies.

Theorem 2.3 (Properties of empirical frequency)

(i) $\mathsf{E}\,\overline{X}_n = p$, $\mathsf{var}\,\overline{X}_n = \frac{p(1-p)}{n}$;

(ii) $\overline{X}_n \xrightarrow[n\to\infty]{\mathsf{P}} p$;

(iii) $\sqrt{n}\,(\overline{X}_n - p) \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}\left(0, p(1-p)\right)$

(iv) $n\overline{X}_n \sim \mathsf{Bi}(n,p)$, where $\mathsf{Bi}(n,p)$ stands for the binomial distribution with n trials and p being the parameter of success.

Proof. (i), (ii) and (iii) follows directly from Theorem 2.2 together with $\mathsf{E}\,X_i = p$ and $\mathsf{var}\,X_i = p(1-p)$. (iv) follows from the fact that $n\overline{X}_n = \sum_{i=1}^n X_i$ and from the definition of the binomial distribution. □

Statement (ii) says that provided we have enough observations then we can find the value p with an arbitrary precision.

The end of lectures for week 1 (3.10.-8.10.).

2.2.3. Properties of the sample variance

First consider the model $\mathcal{F} = \mathcal{L}^2$. Denote $\mu = \mathsf{E}\,X_i$ and $\sigma^2 = \mathsf{var}\,X_i$. Sample variance can be rewritten in several useful ways.

Theorem 2.4 (i)

$$S_n^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2\right). \tag{2.1}$$

(ii) Let $\mathbf{1}_n$ be a column vector of n ones. Denote $\mathbb{A} = \mathbb{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ (matrix $n\times n$). Then

$$S_n^2 = \frac{1}{n-1}\boldsymbol{X}^\top\mathbb{A}\boldsymbol{X} = \frac{1}{n-1}\boldsymbol{Y}^\top\mathbb{A}\boldsymbol{Y}, \tag{2.2}$$

where $\boldsymbol{Y} = \boldsymbol{X} - c\mathbf{1}_n$ for some $c \in \mathbb{R}$.

Proof. Part (i):

$$\begin{aligned}\tfrac{n-1}{n}S_n^2 &= \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2 = \frac{1}{n}\sum_{i=1}^n\left(X_i^2 - 2X_i\overline{X}_n + \overline{X}_n^2\right) = \frac{1}{n}\sum_{i=1}^n X_i^2 - \frac{2}{n}\sum_{i=1}^n X_i\overline{X}_n + \overline{X}_n^2\\ &= \frac{1}{n}\sum_{i=1}^n X_i^2 - 2\overline{X}_n^2 + \overline{X}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2.\end{aligned}$$

Part (ii):

$$\begin{aligned}\boldsymbol{X}^\top\mathbb{A}\boldsymbol{X} &= \boldsymbol{X}^\top\left(\mathbb{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\right)\boldsymbol{X} = \boldsymbol{X}^\top\boldsymbol{X} - \frac{1}{n}\boldsymbol{X}^\top\mathbf{1}_n\mathbf{1}_n^\top\boldsymbol{X}\\ &= \sum_{i=1}^n X_i^2 - \frac{1}{n}\left(\sum_{i=1}^n X_i\right)^2 = \sum_{i=1}^n X_i^2 - n\overline{X}_n^2 = (n-1)S_n^2.\end{aligned}$$

The last part of the proposition follows from the fact that

$$\mathbf{1}_n^\top\mathbb{A} = \mathbf{0} = \mathbb{A}\mathbf{1}_n.$$

□

Remark. Both formulas (2.1) and (2.2) are useful in particular in theoretical derivations. Formula (2.2) shows that S_n^2 can be expressed in a quadratic form and shows that S_n^2 is location invariant.

Note that the matrix $\mathbb{A}$ is idempotentní, i.e. $\mathbb{A}\mathbb{A} = \mathbb{A}$. This will be used later on when deriving the distribution of S_n^2 (see Theorem 2.8 below).

We have a useful formula for calculating the expectations of the quadratic forms.

Lemma 2.5 Let $\boldsymbol{Z}$ be a random vector of length n with the mean value $\boldsymbol{\mu}$ and a finite variance matrix Σ. Let $\mathbb{B}$ be an arbitrary matrix $n \times n$. Then it holds that

$$\mathsf{E}\, \boldsymbol{Z}^\mathsf{T}\mathbb{B}\boldsymbol{Z} = \boldsymbol{\mu}^\mathsf{T}\mathbb{B}\boldsymbol{\mu} + \mathsf{tr}\left(\mathbb{B}\Sigma\right).$$

Proof.

$$\begin{aligned}\mathsf{E}\, \boldsymbol{Z}^\mathsf{T}\mathbb{B}\boldsymbol{Z} &= \mathsf{E}\,\mathsf{tr}\left(\boldsymbol{Z}^\mathsf{T}\mathbb{B}\boldsymbol{Z}\right) = \mathsf{E}\,\mathsf{tr}\left(\mathbb{B}\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T}\right) = \mathsf{tr}\left(\mathbb{B}\mathsf{E}\,\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T}\right) = \mathsf{tr}\left(\mathbb{B}(\boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T} + \Sigma)\right) \\ &= \mathsf{tr}\left(\mathbb{B}\boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T}\right) + \mathsf{tr}\left(\mathbb{B}\Sigma\right) = \boldsymbol{\mu}^\mathsf{T}\mathbb{B}\boldsymbol{\mu} + \mathsf{tr}\left(\mathbb{B}\Sigma\right),\end{aligned}$$

where we make use of the fact that

$$\Sigma = \mathsf{E}\,(\boldsymbol{Z} - \boldsymbol{\mu})(\boldsymbol{Z} - \boldsymbol{\mu})^\mathsf{T} = \mathsf{E}\,\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} - \boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T}.$$

□

Theorem 2.6 (Properties sample variance)

(i) $S_n^2 \xrightarrow[n\to\infty]{\mathsf{P}} \sigma^2$.

(ii) $\mathsf{E}\, S_n^2 = \sigma^2$.

(iii) If $\mathcal{F} = \mathcal{L}^4$ (i.e. if the fourth moment of X_i is finite), then

$$\sqrt{n}\left(S_n^2 - \sigma^2\right) \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}\left(0, \sigma^4(\gamma_4 - 1)\right),$$

where $\gamma_4 = \frac{\mathsf{E}\,(X_i-\mu)^4}{\sigma^4}$ is the kurtosis of X_i.

Proof. Part (i): With the help of Theorem 2.4(i) one can write

$$S_n^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2\right).$$

As $\frac{n}{n-1} \xrightarrow[n\to\infty]{} 1$, it is sufficient to show that

$$\frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2 \xrightarrow[n\to\infty]{\mathsf{P}} \sigma^2.$$

By the law of large numbers (Proposition 1.3) it holds that

$$\left(\overline{X}_n, \frac{1}{n}\sum_{i=1}^{n} X_i^2\right)^\mathsf{T} \xrightarrow[n\to\infty]{\mathsf{P}} \left(\mathsf{E}\, X_i, \mathsf{E}\, X_i^2\right)^\mathsf{T}.$$

Now the function $g(y_1, y_2) = y_2 - y_1^2$ is continuous on $\mathbb{R}^2$, i.e. it is continuous in (the unknown point) $\left(\mathsf{E}\, X_i, \mathsf{E}\, X_i^2\right)$, which is the support of the limit distribution. Now we can use the Continuous Mapping Theorem (Proposition 1.1(ii)) a dostáváme

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2 \xrightarrow[n\to\infty]{\mathsf{P}} \mathsf{E}\, X_i^2 - \left(\mathsf{E}\, X_i\right)^2 = \mathsf{var}\, X_i = \sigma^2,$$

which was to be proved.

Part (ii): Put $\boldsymbol{Y} = \boldsymbol{X} - \mu\mathbf{1}_n$ and note that $\mathsf{E}\,\boldsymbol{Y} = \mathbf{0}$. Then according to Theorem 2.4(ii) and Lemma 2.5 one can calculate

$$(n-1)\mathsf{E}\, S_n^2 = \mathsf{E}\,\boldsymbol{Y}^\mathsf{T}\mathbb{A}\boldsymbol{Y} = \mathsf{E}\,\boldsymbol{Y}^\mathsf{T}\mathbb{A}\mathsf{E}\,\boldsymbol{Y} + \mathsf{tr}\left(\mathbb{A}\sigma^2\mathbb{I}_n\right) = 0 + (n-1)\sigma^2,$$

as

$$\mathsf{tr}\left(\mathbb{A}\sigma^2\mathbb{I}_n\right) = \sigma^2\left(\mathsf{tr}\left(\mathbb{I}_n\right) - \tfrac{1}{n}\mathsf{tr}\left(\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\right)\right) = \sigma^2(n-1).$$

Part (iii): First we rewrite the sample variance as

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2 - \tfrac{n}{n-1}\left(\overline{X}_n - \mu\right)^2.$$

And thus

$$\sqrt{n}\left(S_n^2 - \sigma^2\right) = \frac{\sqrt{n}}{n-1}\sum_{i=1}^{n}\left[(X_i-\mu)^2 - \sigma^2\right] + \tfrac{\sqrt{n}}{n-1}\,\sigma^2 - \tfrac{n}{n-1}\sqrt{n}\left(\overline{X}_n - \mu\right)^2 \overset{ozn.}{=} A_n + B_n + C_n,$$

where A_n, B_n and C_n denotes the corresponding terms on the right-hand side of the above equation. Obviously

$$B_n = \tfrac{\sqrt{n}}{n-1}\,\sigma^2 \xrightarrow[n\to\infty]{} 0.$$

Further

$$C_n = \tfrac{n}{n-1}\sqrt{n}\left(\overline{X}_n - \mu\right)^2 = \tfrac{n}{n-1}\sqrt{n}\left(\overline{X}_n - \mu\right)\left(\overline{X}_n - \mu\right) \xrightarrow[n\to\infty]{\mathsf{P}} 0,$$

where we make use of the fact that

$$\tfrac{n}{n-1} \xrightarrow[n\to\infty]{} 1, \quad \sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}(0, \sigma^2), \quad \overline{X}_n - \mu \xrightarrow[n\to\infty]{\mathsf{P}} 0$$

and Cramér-Slucky theorem (Proposition 1.2).

Thus it is sufficient to deal with the term A_n. For $i \in \{1, \dots, n\}$ denote $Y_i = (X_i - \mu)^2$. Then with the help of the central limit theorem for the random variables Y_i (Proposition 1.4) a Cramér-Slucky theorem (Proposition 1.2)

$$A_n = \frac{\sqrt{n}}{n-1} \sum_{i=1}^{n} \left[(X_i - \mu)^2 - \sigma^2\right] = \frac{n}{n-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[(X_i - \mu)^2 - \sigma^2\right]$$
$$= \frac{n}{n-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[Y_i - \mathsf{E}\, Y_i\right] \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}\big(0, \mathsf{var}\,(Y_i)\big).$$

Now it remains to calculate

$$\mathsf{var}\,(Y_i) = \mathsf{var}\left((X_i - \mu)^2\right) = \mathsf{E}\,(X_i - \mu)^4 - \left(\sigma^2\right)^2 = \sigma^4 \left[\mathsf{E}\left(\tfrac{X_i - \mu}{\sigma}\right)^4 - 1\right] = \sigma^4 \left[\gamma_4 - 1\right].$$

□

Remark.

- Theorem 2.6(iii) says, that the asymptotic variance of the sample variance depends on the kurtosis.

Remark. Alternativaly one can prove Theorem 2.6(ii) (i.e. unbiasedness of the sample variance) by the following straightforward calculation

$$\mathsf{E}\, S_n^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} \mathsf{E}\, X_i^2 - n\, \mathsf{E}\, \overline{X}_n^2\right) = \frac{1}{n-1}\left(n\, \mathsf{E}\, X_1^2 - n\, \mathsf{var}\,(\overline{X}_n) - n\left(\mathsf{E}\, \overline{X}_n\right)^2\right)$$
$$= \frac{1}{n-1}\left(n(\sigma^2 + \mu^2) - n\, \tfrac{\sigma^2}{n} - n\, \mu^2\right) = \frac{1}{n-1}\left(n\sigma^2 - \sigma^2\right) = \sigma^2,$$

where we make us of the fact $\mathsf{E}\, X_1^2 = \mathsf{var}\,(X_1) + \left(\mathsf{E}\, X_1\right)^2$ and analogously also of $\mathsf{E}\left(\overline{X}_n\right)^2 = \mathsf{var}\left(\overline{X}_n\right) + \left(\mathsf{E}\, \overline{X}_n\right)^2$.

Exercise. Prove that, when X_i are zero-one variables then $S_n^2 = \frac{n}{n-1} \overline{X}_n (1 - \overline{X}_n)$. *Hint: Use the fact that* $X_i^2 = X_i$.

Now we add **the assumption of the normal distribution**, e.g. we are going to work in the smaller model $\mathcal{F} = \left\{\mathsf{N}(\mu, \sigma^2),\ \mu \in \mathbb{R},\ \sigma^2 \in \mathbb{R}^+\right\}$. Thus we have a random sample $\boldsymbol{X} = (X_1, X_2, \dots, X_n)^\mathsf{T}$, with X_i being independent with the distribution $\mathsf{N}(\mu, \sigma^2)$. Thanks to the independence it holds that $\boldsymbol{X} \sim \mathsf{N}_n(\mu \mathbf{1}_n, \sigma^2 \mathbb{I}_n)$.

First we give two results that hold for random vectors with (arbitrary) normal distributions.

Lemma 2.7 Let $\boldsymbol{X} \sim \mathsf{N}_n(\boldsymbol{\mu}, \Sigma)$ a $\mathbb{A}$ be a positive semidefinit matrix of the dimension $n \times n$.

(i) Let $\mathbb{B}$ be a matrix of dimension $m \times n$ such that $\mathbb{B}\Sigma\mathbb{A} = \mathbb{0}_{m\times n}$. Then the random variable $\boldsymbol{X}^\mathsf{T}\mathbb{A}\boldsymbol{X}$ and the random vector $\mathbb{B}\boldsymbol{X}$ are independent.

(ii) Let $\mathbb{B}$ be a positive semidefinite matrix of dimension $n \times n$ which satisfies $\mathbb{B}\Sigma\mathbb{A} = \mathbb{0}_{n\times n}$. Then the random variables $\boldsymbol{X}^\mathsf{T}\mathbb{A}\boldsymbol{X}$ and $\boldsymbol{X}^\mathsf{T}\mathbb{B}\boldsymbol{X}$ are independent.

Proof. Part (i). As the matrix $\mathbb{A}$ is positive semidefinite there exists an ortonormal matrix $\mathbb{U}$ such that

$$\mathbb{A} = \mathbb{U}\,\mathbb{D}\,\mathbb{U}^\mathsf{T}$$

where $\mathbb{D} = \mathsf{diag}\,(\lambda_1, \ldots, \lambda_n)$ is a diagonal with eigenvalues of the matrix $\mathbb{A}$ on the diagonal. Note that these eigenvalues are non-negative.

Further from the assumptions of lemma we have

$$\mathbb{0}_{m\times n} = \mathbb{B}\Sigma\mathbb{A} = \mathbb{B}\Sigma\mathbb{U}\,\mathbb{D}\,\mathbb{U}^\mathsf{T}.$$

Denote by $\mathbb{D}^{-1/2}$ the diagonal matrix with the ith diagonal element i given by $\frac{1}{\sqrt{\lambda_i}}$ if λ_i is positive and zero otherwise. Multiplying the above equation with the matrix $\mathbb{U}\mathbb{D}^{-1/2}$ from the right we get

$$\mathbb{0}_{m\times n} = \mathbb{B}\,\Sigma\,\mathbb{U}\,\mathbb{D}^{1/2}.$$

Thus random vectors $\mathbb{B}\boldsymbol{X}$ and $\mathbb{D}^{1/2}\mathbb{U}^\mathsf{T}\boldsymbol{X}$ are not correlated as

$$\mathsf{cov}\left(\mathbb{B}\boldsymbol{X}, \mathbb{D}^{1/2}\mathbb{U}^\mathsf{T}\boldsymbol{X}\right) = \mathbb{B}\,\Sigma\,\mathbb{U}\,\mathbb{D}^{1/2} = \mathbb{0}_{m\times n}.$$

Now from the definition multivariate normal distribution it follows that random vectors has the joint normal distribution as we can write

$$\begin{pmatrix} \mathbb{B}\boldsymbol{X} \\ \mathbb{D}^{1/2}\mathbb{U}^\mathsf{T}\boldsymbol{X} \end{pmatrix} = \begin{pmatrix} \mathbb{B} \\ \mathbb{D}^{1/2}\mathbb{U}^\mathsf{T} \end{pmatrix} \boldsymbol{X}.$$

Now the joint normality and the fact the random vectors are not correlated imply the independence of the random vectors $\mathbb{B}\boldsymbol{X}$ and $\mathbb{D}^{1/2}\mathbb{U}^\mathsf{T}\boldsymbol{X}$ (P.6.2(ii)). Thus also $\mathbb{B}\boldsymbol{X}$ and $\boldsymbol{X}^\mathsf{T}\mathbb{U}\mathbb{D}^{1/2}\mathbb{D}^{1/2}\mathbb{U}^\mathsf{T}\boldsymbol{X} = \boldsymbol{X}^\mathsf{T}\mathbb{A}\boldsymbol{X}$ are independent.

Part (ii). Analogously as above using the spectral decompositions one gets

$$\mathbb{A} = \mathbb{U}_A\,\mathbb{D}_A\,\mathbb{U}_A^\mathsf{T} \quad \text{and} \quad \mathbb{B} = \mathbb{U}_B\,\mathbb{D}_B\,\mathbb{U}_B^\mathsf{T},$$

where $\mathbb{U}_A$, $\mathbb{U}_B$ is ortonormal matrix and $\mathbb{D}_A$, $\mathbb{D}_B$ is diagonal matrix with non-negative elements on diagonals.

Further from the assumption of the lemmat

$$\mathbb{0}_{n\times n} = \mathbb{B}\Sigma\mathbb{A} = \mathbb{U}_B\mathbb{D}_B\mathbb{U}_B^\mathsf{T}\,\Sigma\,\mathbb{U}_A\mathbb{D}_A\mathbb{U}_A^\mathsf{T}$$

Let $\mathbb{D}_A^{-1/2}$ and $\mathbb{D}_B^{-1/2}$ are as the matrix $\mathbb{D}^{-1/2}$ above. Then multiplying the above equation with the matrix $\mathbb{U}_A\,\mathbb{D}_A^{-1/2}$ from the rate and with the matrix $\mathbb{D}_B^{-1/2}\mathbb{U}_B^\mathsf{T}$ from the left we get

$$\mathbb{0}_{n\times n} = \mathbb{D}_B^{1/2}\mathbb{U}_B^\mathsf{T}\,\Sigma\,\mathbb{U}_A\mathbb{D}_A^{1/2}.$$

Thus similarly as in part (i) we get that the random vectors $\mathbb{D}_B^{1/2}\mathbb{U}_B^\mathsf{T}\boldsymbol{X}$ a $\mathbb{D}_A^{1/2}\mathbb{U}_A^\mathsf{T}\boldsymbol{X}$ are independent. Thus also

$$\boldsymbol{X}^\mathsf{T}\mathbb{U}_B\mathbb{D}_B^{1/2}\,\mathbb{D}_B^{1/2}\mathbb{U}_B^\mathsf{T}\boldsymbol{X} = \boldsymbol{X}^\mathsf{T}\mathbb{B}\boldsymbol{X}$$

and

$$\boldsymbol{X}^\mathsf{T}\mathbb{U}_A\mathbb{D}_A^{1/2}\,\mathbb{D}_A^{1/2}\mathbb{U}_A^\mathsf{T}\boldsymbol{X} = \boldsymbol{X}^\mathsf{T}\mathbb{A}\boldsymbol{X}.$$

are independent.

□

Theorem 2.8 (Properties sample variance za normality) Let $X_i \sim \mathsf{N}(\mu,\sigma^2)$, $i = 1,\dots,n$ be independent. Then it holds

(i)
$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}.^{*} \tag{2.3}$$

(ii) $\overline{X}_n$ and S_n^2 are independent random variables.

Proof. Part (i). Using Theorem 2.4 one can rewrite

$$\frac{(n-1)S_n^2}{\sigma^2} = \boldsymbol{Y}^\mathsf{T}\mathbb{A}\boldsymbol{Y},$$

where

$$\boldsymbol{Y} = \left(\tfrac{X_1-\mu}{\sigma},\dots,\tfrac{X_n-\mu}{\sigma}\right)^\mathsf{T} \sim \mathsf{N}_n(\mathbf{0},\mathbb{I}_n)$$

and $\mathbb{A} = \mathbb{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$. As matrix $\mathbb{A}$ is idempotentní with the rank $n-1$, then the statement of the proposition follows from lemma A.1 (where $\Sigma = \mathbb{I}_n$).

Part (ii) Note that one can write

$$\overline{X}_n = \frac{1}{n}\mathbb{B}\boldsymbol{X}, \qquad S_n^2 = \frac{1}{n-1}\boldsymbol{X}^\mathsf{T}\mathbb{A}\boldsymbol{X},$$

where $\mathbb{B} = \mathbf{1}_n^\mathsf{T}$ a $\mathbb{A} = \mathbb{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$. Further $\boldsymbol{X} \sim \mathsf{N}_n(\mu\mathbf{1}_n,\sigma^2\mathbb{I}_n)$ and thus proposition follows from lemma 2.7(i) as

$$\mathbb{B}\Sigma\mathbb{A} = \mathbf{1}_n^\mathsf{T}\sigma^2\mathbb{I}_n\left(\mathbb{I}_n - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\right) = \sigma^2\left(\mathbf{1}_n^\mathsf{T} - \tfrac{1}{n}\,n\mathbf{1}_n^\mathsf{T}\right) = \mathbf{0}_n^\mathsf{T}.$$

□

Remark. From the definition of χ^2 distributions we know that random variable with χ^2_{n-1} distribution can be represented as $\sum_{i=1}^{n-1} Y_i^2$, where $Y_1,\dots,Y_{n-1}$ are independent and identically distributed random variables with $\mathsf{N}(0,1)$ distribution. From the central limit theorem and (2.3) it follows that

$$\frac{\frac{(n-1)S_n^2}{\sigma^2} - (n-1)}{\sqrt{n-1}} \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}(0,2)$$

* Viz definice A.1.

and thus

$$\sqrt{\tfrac{n-1}{n}}\sqrt{n}\left(S_n^2-\sigma^2\right) \overset{\text{as.}}{\sim} \mathsf{N}(0,2\sigma^4).$$

Taking into consideration that the skewness of normal distribution is 3, we see that statement (i) of Theorem 2.8 is in agreement with the asymptoic result of Theorem 2.6(iii). Theorem 2.8(i) now gives the exact distribution of S_n^2 for random sample from the normal distribution, while Theorem 2.6(iii) gives the asymptotic distribution S_n^2 for random sample from an aribtrary distribution that has the finite fourth moment.

Remark. One can remember the statement (i) of Theorem 2.8(i) as follows. Note that

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

If one uses the true expectation μ instead of $\overline{X}_n$ in the above formula, then $\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi_n^2$. By replacing the unknown expectation μ with its estimator $\overline{X}_n$ we loose one degrees of freedom (as we estimate one parametr).

Remark. Theorem 2.8(ii) says, that when the random sample comes from the normal distribution, then $\overline{X}_n$ and S_n^2 are independent for each finite $n>1$.

Theorem 2.9 (limitní Theorem o T_n) Let $X_1,\ldots,X_n$ be a random sample from an arbitrary distribution with the expectation μ and with the finite and non-zero variance σ^2. Then

$$T_n=\frac{\sqrt{n}\left(\overline{X}_n-\mu\right)}{S_n}\xrightarrow[n\to\infty]{\mathsf{d}}\mathsf{N}(0,1).$$

Proof. The random variable T_n can be now rewritten in the form

$$T_n=\frac{\sqrt{n}\left(\overline{X}_n-\mu\right)}{\sigma}\,\frac{\sigma}{S_n}.$$

By the central limit theorem (Proposition 1.4, pro $k=1$) one has that

$$\frac{\sqrt{n}\left(\overline{X}_n-\mu\right)}{\sigma}\xrightarrow[n\to\infty]{\mathsf{d}}\mathsf{N}(0,1).$$

Further z $S_n^2\xrightarrow[n\to\infty]{\mathsf{P}}\sigma^2$ (Theorem 2.6(i)) and byt the continuous mapping theorem (Proposition 1.1(ii)) for $g(y)=\sigma/\sqrt{y}$ one gets

$$\frac{\sigma}{S_n}\xrightarrow[n\to\infty]{\mathsf{P}}1.$$

The statement now follows from Cramér-Slucky věty (Proposition 1.2). □

Now we again add the assumption of **normal distribution**.

Theorem 2.10 Let $X_1, \ldots, X_n$ be a random sample from the distribution $\mathsf{N}(\mu, \sigma^2)$. Then

$$T_n = \frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{S_n} \sim t_{n-1}.^{*}$$

Proof. The random variable T_n can now be rewritten as

$$T_n = \frac{\frac{\sqrt{n}\,(\overline{X}_n - \mu)}{\sigma}}{\sqrt{\frac{(n-1)S_n^2}{\sigma^2}/(n-1)}}. \tag{2.4}$$

From the remark below Theorem 2.2 we know that $\sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma} \sim \mathsf{N}(0,1)$. Further $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$ (Theorem 2.8(i)), and at the same time the numerator and the denominator in fraction (2.4) are independent (Theorem 2.8(ii)). The statement now follows from the definition of the t-distributions (see Definition A.2). □

Remark. Theorem 2.10 gives the exact distribution of T_n for normally distributed data while Theorem 2.9 gives the asymptotic distribution of T_n for random sample from an arbitrary distribution with the finite and non-zero variance. Note that for $n \to \infty$ the distribution t_{n-1} converges in distribution to $\mathsf{N}(0,1)$.

* Viz definice A.2.

3. Parameter Estimation

We are given a random sample $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$, a model $\mathcal{F}$ and a parameter $\boldsymbol{\theta} = t(F) \in \mathbb{R}^p$ for $F \in \mathcal{F}$, which we need to estimate. Let $F_X \in \mathcal{F}$ be the true distribution of the random vector $\boldsymbol{X}_i$ and let $\boldsymbol{\theta}_X \equiv t(F_X)$ be the true value of $\boldsymbol{\theta}$.

3.1. Point estimation

Definition 3.1 An *estimator* of $\boldsymbol{\theta}_X \equiv t(F_X) \in \mathbb{R}^p$ is a p-dimensional random vector $\widehat{\boldsymbol{\theta}}_n$ which is given as $\widehat{\boldsymbol{\theta}}_n = T_n(\boldsymbol{X}) \equiv T_n(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, where T_n is some Borel measurable function of data.

Remark. An estimator is a statistic in sense of definition 2.3. It cannot depend on unknown parameters.

Definition 3.2 (Unbiasedness and consistency) Let us suppose that we are given a random sample $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$ from distribution $F_X \in \mathcal{F}$ and an estimator $\widehat{\boldsymbol{\theta}}_n \equiv T_n(\boldsymbol{X})$ of a parameter $\boldsymbol{\theta}_X \equiv t(F_X)$.

(i) $\widehat{\boldsymbol{\theta}}_n$ is said to be an *unbiased estimator* of the parameter $\boldsymbol{\theta}_X$ in the model $\mathcal{F}$ if and only if $\mathsf{E}\,\widehat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_X$ for every n (for which the estimator is well-defined) and for every distribution $F_X \in \mathcal{F}$.

(ii) $\widehat{\boldsymbol{\theta}}_n$ is said to be a *consistent estimator* of the parameter $\boldsymbol{\theta}_X$ in the model $\mathcal{F}$ if and only if $\widehat{\boldsymbol{\theta}}_n \xrightarrow{\mathsf{P}} \boldsymbol{\theta}_X$ as $n \to \infty$ for every distribution $F_X \in \mathcal{F}$.

Remark.

- Properties of a given estimator must be studied in context of the given model. It can easily happen that an estimator $\widehat{\boldsymbol{\theta}}_n$ is unbiased and consistent in some model $\mathcal{F}$, while in a different model $\mathcal{F}'$ it does not retain these properties.
- Unbiasedness is supposed to hold for each number of observations n for which the estimator is defined (e.g. in case of the sample variance for $n \geq 2$). Unbiasedness, however, does not guarantee that the estimator will approach the true value of the parameter being estimated as the sample size n increases. For some models there are no reasonable (or even none at all) unbiased estimators.
- Consistency is an asymptotic property, which does not say anything about behaviour of an estimator for finite n. (e.g. $\widehat{\theta}_n = 21$ for $n \leq 10^{10}$, $\widehat{\theta}_n = \overline{X}_n$ for $n > 10^{10}$ is a consistent estimator of $\theta_X = \mathsf{E}\,X_i$.)

- The aforementioned notion of consistency is sometimes called *weak consistency*. In addition, an estimator is said to be *strongly consistent* if and only if $\widehat{\theta}_n \xrightarrow[n\to\infty]{\text{a.s.}} \theta_X$.
- In statistics, estimators which are consistent, albeit not unbiased, are commonly used. On the other hand, estimators which are not consistent are typically unused because they either estimate "something different" or they do not get more accurate as the sample size increases.

Examples.

1. *Estimation of parameter* $\theta_X = \mathsf{E}\, X_i$ *in model* $\mathcal{F} = \mathcal{L}^1$*:*
   - The sample mean $\overline{X}_n$ is an unbiased and consistent estimator of θ_X [follows from theorem 2.2, (i) a (ii)].
   - The estimator $\widehat{\theta}_n = X_1$ is an unbiased estimator of θ_X, but it is not consistent.

2. *Estimation of parameter* $\theta_X = \mathsf{var}\, X_i$ *in model* $\mathcal{F} = \mathcal{L}^2$*:*
   - The sample variance S_n^2 is an unbiased and consistent estimator of θ_X [follows from theorem 2.6, (i) a (ii)].
   - The estimator $\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2$ is a consistent estimator of θ_X, but it is not unbiased.

3. *Estimation of parameter* $\theta_X = \mathsf{P}[X_i = 0]$ *in model* $\mathcal{F} = \left\{\mathsf{Po}(\lambda), \lambda > 0\right\}$*:*
   - The estimator $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{0\}}(X_i)$ is an unbiased and also consistent estimator of θ_X (unbiasedness and consistency of $\widehat{\theta}_n$ are preserved even in the model of all discrete distributions).
   - The estimator $\widetilde{\theta}_n = \left(\frac{n-1}{n}\right)^{\sum_{i=1}^n X_i}$ is also an unbiased and consistent estimator of θ_X (in model $\mathcal{F}$ but not in the model of all discrete distributions).

4. *Estimation of parameter* $\theta_X = \mathrm{e}^{-2\lambda_X}$ *in model* $\mathcal{F} = \left\{\mathsf{Po}(\lambda), \lambda > 0\right\}$ *for* $n = 1$*:*
   The only unbiased estimator is $\widehat{\theta} = (-1)^{X_1}$ and the only 2 values which this estimator attains are -1 and 1. However, $\mathrm{e}^{-2\lambda_X}$ only attains values from the interval $(0, 1)$.

Definition 3.3 (Bias) Let us suppose that the estimator $\widehat{\boldsymbol{\theta}}_n \equiv T_n(\boldsymbol{X})$ of a parameter $\boldsymbol{\theta}_X$ has finite expectation. Then the difference $\mathsf{E}\,(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X)$ is called *bias* of the estimator $\widehat{\boldsymbol{\theta}}_n$.

Definition 3.4 Let us suppose that the estimator $\widehat{\theta}_n \equiv T_n(\boldsymbol{X})$ of a parameter $\theta_X \in \mathbb{R}$ has finite variance.

(i) Expression

$$\mathrm{MSE}(\widehat{\theta}_n) = \mathsf{E}\left(\widehat{\theta}_n - \theta_X\right)^2$$

is called *mean squared error* of the estimator $\widehat{\theta}_n$.

(ii) Expression

$$\mathrm{SE}(\widehat{\theta}_n) = \sqrt{\mathsf{var}\,(\widehat{\theta}_n)}$$

is called *standard error* of the estimator $\widehat{\theta}_n$.

Remark.

- Beware of subtle differences in terminology. The term *standard deviation* (SD) usually refers to the square root of the variance of one random observation i.e. $\sqrt{\mathsf{var}\,X_i}$. The term *standard error* (SE) usually refers to the square root of the variance of some estimator calculated from the whole random sample. Some authors, however, use the term *standard error* when they want to refer to

$$\mathrm{SE}(\widehat{\theta}_n) = \sqrt{\widehat{\mathsf{var}}\,(\widehat{\theta}_n)},$$

where $\widehat{\mathsf{var}}\,(\widehat{\theta}_n)$ is an estimator of $\mathsf{var}\,(\widehat{\theta}_n)$
- Both the mean squared error and the standard error are measures of estimation accuracy. Furthermore, while the standard error disregards the bias, the mean squared error does not.
- It holds that the mean squared error can be decomposed as a sum of variance and bias squared:

$$\mathrm{MSE}(\widehat{\theta}_n) = \mathsf{var}\,(\widehat{\theta}_n) + [\mathsf{E}\,(\widehat{\theta}_n - \theta_X)]^2.$$

Proof of the aforementioned assertion is a direct calculation:

$$\begin{aligned}\mathrm{MSE}(\widehat{\theta}_n) &= \mathsf{E}\,(\widehat{\theta}_n - \mathsf{E}\,\widehat{\theta}_n + \mathsf{E}\,\widehat{\theta}_n - \theta_X)^2 \\ &= \mathsf{E}\,(\widehat{\theta}_n - \mathsf{E}\,\widehat{\theta}_n)^2 + 2\,\mathsf{E}\,(\widehat{\theta}_n - \mathsf{E}\,\widehat{\theta}_n)\mathsf{E}\,(\widehat{\theta}_n - \theta_X) + [\mathsf{E}\,(\widehat{\theta}_n - \theta_X)]^2 \\ &= \mathsf{var}\,(\widehat{\theta}_n) + 0 + [\mathsf{E}\,(\widehat{\theta}_n - \theta_X)]^2.\end{aligned}$$

- The mean squared error is one of the most appropriate criteria for comparison of estimators. If we have several different estimators of the same parameter in the same model, we try to find the one with the smallest MSE. Thus, in the case of unbiased estimators, we select the one with the smallest variance.
- MSE often cannot be calculated analytically. In many cases, however, one can decide on the basis of asymptotic variances of estimators. Assume that we have 2 estimators $\widehat{\theta}_n$ and $\widetilde{\theta}_n$, which satisfy

$$\sqrt{n}\,(\widehat{\theta}_n - \theta_X) \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}(0, \sigma_1^2), \qquad \sqrt{n}\,(\widetilde{\theta}_n - \theta_X) \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}(0, \sigma_2^2).$$

Then (for large sample sizes) estimator $\widehat{\theta}_n$ is preferred if $\sigma_1^2 < \sigma_2^2$. Conversely, if $\sigma_1^2 > \sigma_2^2$, then estimator $\widetilde{\theta}_n$ is preferred.

Example. *Estimation of parameter* $\sigma_X^2 = \mathsf{var}\,X_i$ *in model* $\mathcal{F} = \{\mathsf{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$. Show that $\mathrm{MSE}(S_n^2) > \mathrm{MSE}(\widehat{\sigma}_n^2)$.

Theorem 3.1 Let $\widehat{\theta}_n$ be an estimator of a parameter $\theta_X \in \mathbb{R}$ for which it holds that $\mathsf{E}\,\widehat{\theta}_n \xrightarrow[n\to\infty]{} \theta_X$ (bias converges to zero) and $\mathsf{var}\left(\widehat{\theta}_n\right) \xrightarrow[n\to\infty]{} 0$, for each $F_X \in \mathcal{F}$. Then $\widehat{\theta}_n$ is a consistent estimator of θ_X.

Proof. Let $\varepsilon > 0$. Then from Markov's inequality (theorem P.2.6) it follows that:

$$\mathsf{P}\left(\left|\widehat{\theta}_n - \theta_X\right| > \varepsilon\right) \leq \frac{\mathrm{MSE}(\widehat{\theta}_n)}{\varepsilon^2} = \frac{\mathsf{var}\left(\widehat{\theta}_n\right)}{\varepsilon^2} + \frac{\left(\mathsf{E}\,\widehat{\theta}_n - \theta_X\right)^2}{\varepsilon^2}.$$

Now, both terms on the right-hand side converge to zero because thanks to the assumptions of the theorem $\mathsf{var}\left(\widehat{\theta}_n\right) \to 0$ and $\mathsf{E}\,\widehat{\theta}_n \to \theta_X$ as $n \to \infty$. □

Remark.

- The opposite implication is not true. There exist consistent estimators which satisfy that $\mathsf{E}\left|\widehat{\theta}_n\right| = \infty$ for every finite n.
- Theorem 3.1 is useful in situations when the bias and the variance of the estimator $\widehat{\theta}_n$ are available (or can be easily calculated). If, however, it is possible to express $\widehat{\theta}_n$ as $\widehat{\theta}_n = g\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i\right)$ (i.e. as a transformation of the sample mean), then it is easier to study consistency of $\widehat{\theta}_n$ using the law of large numbers (theorem 1.3) in combination with the continuous mapping theorem (theorem 1.1).

Example. Let $X_1, \ldots, X_n$ be a random sample from the alternative distribution $\mathsf{Alt}(p_X)$. Consider $\widehat{\theta}_n = \frac{1}{\overline{X}_n}$ as an estimator of $\theta_X = \frac{1}{p_X}$. Show that although it holds that $\mathsf{E}\,\widehat{\theta}_n = \infty$, it also holds that $\widehat{\theta}_n \xrightarrow[n\to\infty]{\mathsf{P}} \theta_X$.

3.2. Choice of the parameter of interest

The parameter $\theta = t(F)$ which we are trying to estimate can be in principle anything. Not all parameters, however, make sense in context of the practical problem we are solving. Therefore, we must distinguish for which parameters it is reasonable to estimate them and for which it is not. This depends on the meaning of the values of the measured quantities, on the procedure by which they were obtained, processed, etc. The statistical methods that will be introduced, will be divided according to the type of measurements for which they are intended. We will consider the following data types or *measurement scales*.

3.2.1. Quantitative data

A random variable X will be called *quantitative* if its values have some specific numerical meaning (e.g. number, percentage, length, volume, weight, interest rate, concentration, temperature, duration, angle, latitude, calendar year). For quantitative data there exists a meaningful ordering of their values (temperature 10 °C is higher than −11,4 °C). Furthermore, differences of these values are interpretable. Quantitative random variables can be both discrete and continuous.

Quantitative variables can be further subdivided into two subgroups: *interval* and *ratio*. **Ratio variables** are typically non-negative with a clearly defined zero value and interpretable ratios. For example, the weight 0 kg has a clear interpretation and an object whose weight is 20 kg is 4 times heavier than 5 kg. Examples of ratio variables are number, length, volume, weight, interest rate, concentration, time duration, temperature measured in kelvins. **Interval variables** are quantitative variables which do not follow properties of ratio variables, i.e. they do not have a fixed zero value or ratios of their values are not interpretable. For instance, direction given by azimuth is an interval quantity because azimuth 360° is not six times greater than 60°. Similarly, temperature measured in °C is an interval quantity because the temperature of 16 °C is not four times higher than the temperature of 4 °C. Calendar year is also an interval quantity, because it does not make sense to calculate the ratio of this year and the year of your birth.

3.2.2. Categorical data

A random variable X is called *categorical* if its values encode affiliation (or *classification*) of an object with a certain category, or with one of several disjoint sets. Categorical variables are always discrete and have a finite number K of possible values, usually $1, \ldots, K$ or $0, \ldots, K-1$. Values of categorical variables do not have a direct numerical interpretation. Their sole purpose is to distinguish possible states. Individual states are called *levels* or *categories*.

We further subdivide categorical variables into *nominal* and *ordinal*. For **nominal variables** there is no ordering of their categories - it cannot be said that some category j precedes the category $j+1$. An example of a nominal variable is, for instance, residence categorised in terms of regions (1 = Prague, 2 = Central Bohemian, ..., 14 = Moravian-Silesian) or social status (1 = underage; 2 = student; 3 = employee; 4 = self-employed ; 5 = unemployed ; 6 = pensioner). Categories of **ordinal variables** are in some sense ordered. Thus, it is possible to claim that category j precedes category $j+1$ or that it is smaller, worse, etc. An example of an ordinal variable may be an answer to a question with options 1 = strongly disagree, 2 = rather disagree, 3 = do not know, 4 = rather agree, 5 = totally agree. A different example is a variable encoding the highest attained level of education as 1 = primary education; 2 = lower secondary education; 3 = upper secondary education; 4 = post-secondary non-tertiary education; 5 = short-cycle tertiary education; 6 = bachelor's or equivalent; 7 = master's or equivalent; 8 = doctorate or equivalent.

3.2.3. Binary data

Binary variables are a special case of categorical variables when $K = 2$. Hence, they classify observations into one of two possible states. Their values are typically chosen as 0 vs. 1 or, alternatively, 1 vs. 2. An example of a binary variable is the truth value of some statement (0 = false, 1 = true), realisation of a random phenomenon (0 = did not occur/failure, 1 = occurred/success) or sex (1 = male, 2 = female).

3.2.4. Choice of the parameter according to the type of data

In general, for nominal quantities it does not make sense to consider parameters such as $\mathsf{E}\,X$, $\mathsf{var}\,X$, cumulative distribution function, quantiles, covariance and correlation, in short, no characteristics that depend on encoding and ordering of individual categories. Although these parameters are properly defined, they have no practical interpretation. The only parameters which in case of nominal variables do have an interpretation are probabilities of individual categories, i.e. $p_j = \mathsf{P}[X = j]$ for all admissible values of j.

One exception are binary variables. If value 0 encodes failure and value 1 encodes success, then $\mathsf{E}\,X = \mathsf{P}[X = 1]$, i.e. expectation and probability of success are equal. For ordinal variables, thanks to natural ordering of their categories, it makes sense to consider their cumulative distribution functions. It is often possible to attach to them the interval interpretation (doctoral education is two levels higher than bachelor), however, it is not usually feasible to afford them ratio interpretation (we cannot say that bachelor's education is 2 times higher than upper secondary education). Ordinal variables are sometimes assigned non-integer values, so-called *scores*. For example we can create an ordinal variable in a way that we take some quantitative variable Z and categorise it according to some chosen partition, e.g. $X = 1$ if $Z \in \langle 0, 5)$, $X = 2$ if $Z \in \langle 5, 20)$, $X = 3$ if $Z \in \langle 20, 100)$ and $X = 4$ if $Z \geq 100$. Such quantities usually arise in questionnaires, when respondents are supposed to choose one of four options instead of writing down the exact number. The resulting variable X is obviously ordinal. Perhaps, instead of the values $1, \ldots, 4$ we could choose, as the values of X, midpoints of the intervals which were used to define X, i.e. 2,5; 12,5 a 60 for the first three intervals. There is clearly a problem with the last one since it does not have the right endpoint - thus, we would somehow need to add the last score (for example take 150). Variables encoded in this way are not only ordinal, but they also retain some properties of quantitative variables.

Ordinal variables can always be analysed as if they were nominal but it is often possible to also apply methods originally devised for quantitative variables, estimate their expectation or calculate their differences. Moreover, there exist special methods designed specifically for the ordinal data, but we will not encounter them for a while.

Our explanation of statistical methods (starting with chapter 4) will distinguish between methods for quantitative data, where we will work with characteristics such as expectation, variance, median, cumulative distribution function, covariance, etc., and methods for nominal data, where we will work with probabilities of individual categories.

3.3. Method of moments

The method of moments belongs, together with the method of maximum likelihood, to basic methods of parameter estimation.

Let us consider a parametric model: we are given a random sample $X_1, \ldots, X_n$ from

a distribution with a probability density function $f(x;\boldsymbol{\theta}_X)$ with respect to some σ-finite measure μ, where the form of the function $f(\cdot;\cdot)$ is known and $\boldsymbol{\theta}_X$ is an unknown (vector-valued) parameter, which belongs to some space of parameters $\Theta \subseteq \mathbb{R}^d$, $d \geq 1$. Thus, we are working with the following model:

$$\mathcal{F} = \left\{\text{distributions with density } f(x;\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d\right\}$$

The goal is to estimate the parameter $\boldsymbol{\theta}_X$. We will take advantage of the fact that we have at our disposal consistent estimators of moments and that we can usually express moments of X_i as functions of unknown parameters. We will assume that $\mathsf{E}\,|X_i|^d < \infty$.

Consider first $d = 1$. Let us assume that $\mathsf{E}\,X_i = \tau(\theta_X)$, where $\tau : \Theta \to \mathbb{R}$. Since $\overline{X}_n$ is a consistent estimator, it is reasonable to try to find the *moment estimator* $\widehat{\theta}_n$ as a solution of the *estimating equation*:

$$\overline{X}_n = \tau\left(\widehat{\theta}_n\right). \tag{3.1}$$

If the function τ is strictly monotone, it is possible to express the estimator as $\widehat{\theta}_n = \tau^{-1}(\overline{X}_n)$ and the estimated parameter as $\theta_X = \tau^{-1}(\mathsf{E}\,X_i)$.

Properties of the estimator $\widehat{\theta}_n$:

- If τ^{-1} is continuous at $\mathsf{E}\,X_i$, then $\widehat{\theta}_n \xrightarrow[n\to\infty]{\mathsf{P}} \theta_X$ (theorem 1.1).
- If τ^{-1} has a continuous derivative on some neighbourhood of $\mathsf{E}\,X_i$, then thanks to the Δ-method (theorem **??**)

  $$\sqrt{n}\left(\widehat{\theta}_n - \theta_X\right) \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}\left(0, V(\theta_X)\right),$$

  where

  $$V(\theta_X) = \left\{\left[\tau^{-1}(\mathsf{E}\,X_i)\right]'\right\}^2 \mathsf{var}\,X_i = \frac{\mathsf{var}\,X_i}{\left[\tau'\left(\tau^{-1}(\mathsf{E}\,X_i)\right)\right]^2} = \frac{\mathsf{var}\,X_i}{\left[\tau'(\theta_X)\right]^2}. \tag{3.2}$$

  Note that in the expression of the asymptotic variance (last equality) we do not need to know the explicit formula for τ^{-1}. This formula is therefore useful if τ^{-1} is given only implicitly and the estimate $\widehat{\theta}_n$ is being searched for using numerical methods as a solution of the estimating equation (3.1).
  In applications, the asymptotic variance $V(\theta_X)$ is estimated by

  $$\widehat{V}_n = \left\{\left[\tau^{-1}(\overline{X}_n)\right]'\right\}^2 S_n^2 = \frac{S_n^2}{\left[\tau'\left(\widehat{\theta}_n\right)\right]^2}.$$

  The last expression is again suitable especially when we do not have the explicit formula for τ^{-1}.

Examples.

1. $X_1, \ldots, X_n$ is a random sample from $\mathsf{Po}(\lambda_X)$ distribution, $\mathsf{E}\, X_i = \lambda_X$. The moment estimator of λ_X is $\widehat{\theta}_n = \overline{X}_n$.

2. $X_1, \ldots, X_n$ is a random sample from $\mathsf{Geo}(p_X)$ distribution, $\mathsf{E}\, X_i = \frac{1-p_X}{p_X}$ and $\mathsf{var}\, X_i = \frac{1-p_X}{p_X^2}$. Thus, $\tau(x) = \frac{1-x}{x}$ and $\tau^{-1}(x) = \frac{1}{1+x}$. The moment estimator of p_X is $\widehat{p}_n = \frac{1}{1+\overline{X}_n}$. Further,
$$\sqrt{n}\,(\widehat{p}_n - p_X) \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}\big(0, p_X^2(1-p_X)\big),$$
where the asymptotic variance $p_X^2(1-p_X)$ follows either from the first equality in (3.2)
$$V(p_X) = \left\{\frac{-1}{(1+\mathsf{E}\, X_i)^2}\right\}^2 \mathsf{var}\, X_i = p_X^4\,\frac{1-p_X}{p_X^2}$$
or, alternatively, also from the third equality in (3.2)
$$V(p_X) = \frac{\mathsf{var}\, X_i}{\left\{-\frac{1}{p_X^2}\right\}^2} = \frac{\frac{1-p_X}{p_X^2}}{\frac{1}{p_X^4}}.$$

3. $X_1, \ldots, X_n$ is a random sample from $\mathsf{R}(0, \theta_X)$ distribution, $\mathsf{E}\, X_i = \theta_X/2$. The moment estimator of θ_X is $\widehat{\theta}_n = 2\overline{X}_n$. It holds that $\sqrt{n}\,(\widehat{\theta}_n - \theta_X) \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}\big(0, \theta_X^2/3\big)$.

$d = 1$, but a different moment than $\mathsf{E}\, X_i$

Sometimes it can happen that $\mathsf{E}\, X_i = 0$ for every $\theta_X \in \Theta$. For example, this is true for distributions with finite expectations which are symmetric around zero. Then we can consider the second moment, i.e. $\mathsf{E}\, X_i^2 = \tau(\theta_X)$ and the estimator $\widehat{\theta}_n$ will be acquired as a solution of the equation
$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \tau(\widehat{\theta}_n).$$
Generally, we can consider some suitable (measurable) function t such that $\mathsf{E}\,|t(X_i)| < \infty$ and $\mathsf{E}\, t(X_i) = \tau(\theta_X)$. The estimator $\widehat{\theta}_n$ will be obtained as a solution of the equation
$$\frac{1}{n}\sum_{i=1}^{n} t(X_i) = \tau(\widehat{\theta}_n).$$

Now we will generalise the method for $d > 1$.

The most straightforward method is to consider the first d-moments, i.e. we will calculate
$$\mathsf{E}\, X_i = \tau_1(\boldsymbol{\theta}_X), \mathsf{E}\, X_i^2 = \tau_2(\boldsymbol{\theta}_X), \ldots, \mathsf{E}\, X_i^d = \tau_d(\boldsymbol{\theta}_X),$$

and thus, we will obtain mappings $\tau_1, \dots, \tau_d : \Theta \to \mathbb{R}$. The estimator of the parameter $\widehat{\boldsymbol{\theta}}_n$ is then obtained as a solution of the following system of d equations with d unknowns:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \tau_1(\widehat{\boldsymbol{\theta}}_n), \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \tau_2(\widehat{\boldsymbol{\theta}}_n), \dots, \frac{1}{n}\sum_{i=1}^{n} X_i^d = \tau_d(\widehat{\boldsymbol{\theta}}_n).$$

Once we define mapping $\boldsymbol{\tau} = (\tau_1, \dots, \tau_d)^\mathsf{T} : \Theta \to \mathbb{R}^d$, then under the assumption of existence of $\boldsymbol{\tau}^{-1}$ we can write

$$\widehat{\boldsymbol{\theta}}_n = \boldsymbol{\tau}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Z}_i\right), \quad \text{where } \boldsymbol{Z}_i = \left(X_i, X_i^2, \dots, X_i^d\right)^\mathsf{T}.$$

From this expression, similarly as in the case of $d = 1$, we can derive consistency and the asymptotic normality of the estimator $\widehat{\boldsymbol{\theta}}_n$.

Special case $d = 2$

Suppose that $(\mathsf{E}\, X_i, \mathsf{var}\, X_i)^\mathsf{T} = \boldsymbol{\tau}(\boldsymbol{\theta}_X)$, where $\boldsymbol{\tau} : \Theta \to \mathbb{R}^2$. Then it is reasonable to try to find the estimator of $\boldsymbol{\theta}_X$ as a solution of the system of estimating equations (more precisely 2 equations with 2 unknowns)

$$\left(\overline{X}_n, S_n^2\right)^\mathsf{T} = \boldsymbol{\tau}\left(\widehat{\boldsymbol{\theta}}_X\right).$$

If the function $\boldsymbol{\tau}$ is injective, then we can express the estimator as $\widehat{\boldsymbol{\theta}}_X = \boldsymbol{\tau}^{-1}\left(\overline{X}_n, S_n^2\right)$ and the estimated parameter as $\boldsymbol{\theta}_X = \tau^{-1}(\mathsf{E}\, X_i, \mathsf{var}\, X_i)$.

Properties of the estimator $\widehat{\theta}_n$:

- We know that $\overline{X}_n$ and S_n^2 are consistent estimators of $\mathsf{E}\, X_i$ and $\mathsf{var}\, X_i$. Hence, if the function $\boldsymbol{\tau}^{-1}$ is continuous at $(\mathsf{E}\, X_i, \mathsf{var}\, X_i)$, then $\widehat{\boldsymbol{\theta}}_n \xrightarrow[n\to\infty]{\mathsf{P}} \boldsymbol{\theta}_X$.
- From theorem 2.6, part (iv) we know that if $\mathsf{E}\, X_i^4 < \infty$, then $\overline{X}_n$ and S_n^2 are jointly asymptotically normal. If $\boldsymbol{\tau}^{-1}$ has a continuous derivative, then according to the Δ-method also $\widehat{\boldsymbol{\theta}}_n$ has jointly asymptotically normal distribution with variance matrix which can be calculated using theorem 2.6 and the Δ-method.

Examples.

4. $X_1, \dots, X_n$ is a random sample from gamma distribution with density $f(x; a, p) = \frac{a^p}{\Gamma(p)} x^{p-1} \mathrm{e}^{-ax} \mathbb{1}\{x > 0\}$. Then $\mathsf{E}\, X_i = \frac{p}{a}$ and $\mathsf{var}\, X_i = \frac{p}{a^2}$ (see chapter 8.2.6 of Kulich, 2018). The moment method yields consistent and asymptotically normal estimators

$$\widehat{a}_n = \frac{\overline{X}_n}{S_n^2} \quad \text{and} \quad \widehat{p}_n = \frac{\overline{X}_n^2}{S_n^2}.$$

5. $X_1, \dots, X_n$ is a random sample from $\mathsf{R}(\theta_1, \theta_2)$ distribution. We know that

$$\mathsf{E}\, X_i = \frac{\theta_1 + \theta_2}{2} \qquad \text{and} \qquad \mathsf{var}\, X_i = \frac{(\theta_2 - \theta_1)^2}{12}.$$

In this case, the system of estimating equations is of the form

$$\overline{X}_n = \frac{\widehat{\theta}_{1n} + \widehat{\theta}_{2n}}{2}, \qquad \operatorname{var} X_i = \frac{(\widehat{\theta}_{2n} - \widehat{\theta}_{1n})^2}{12}.$$

By solving this system we get

$$\widehat{\theta}_{1n} = \overline{X}_n - \sqrt{3S_n^2} \quad \text{and} \quad \widehat{\theta}_{2n} = \overline{X}_n + \sqrt{3S_n^2}.$$

Since from theorem 2.6 we know that

$$\sqrt{n}\left[\begin{pmatrix}\overline{X}_n \\ S_n^2\end{pmatrix} - \begin{pmatrix}\mu \\ \sigma^2\end{pmatrix}\right] \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}_2(\mathbf{0}, \Sigma),$$

where $\Sigma = \begin{pmatrix}\sigma^2 & \sigma^3\gamma_3 \\ \sigma^3\gamma_3 & \sigma^4(\gamma_4 - 1)\end{pmatrix}$ and $\gamma_3 = \frac{\mathsf{E}(X_i-\mu)^3}{\sigma^3}$, then using the Δ-method it is possible to show that

$$\sqrt{n}\left[\begin{pmatrix}\widehat{\theta}_{1n} \\ \widehat{\theta}_{2n}\end{pmatrix} - \begin{pmatrix}\theta_1 \\ \theta_2\end{pmatrix}\right] \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}_2(\mathbf{0}, \mathbb{D}\Sigma\mathbb{D}^\mathsf{T}),$$

where $\mathbb{D}$ denotes the Jacobian matrix of the mapping $\boldsymbol{\tau}^{-1}(x_1, x_2) = (x_1 - \sqrt{3x_2}, x_1 + \sqrt{3x_2})$ at point $(\mathsf{E}\, X_i, \operatorname{var} X_i)$. Therefore, the estimator $\widehat{\boldsymbol{\theta}}_n = (\widehat{\theta}_{1n}, \widehat{\theta}_{2n})$ is asymptotically normal (and according to theorem **??** also consistent).

6. $X_1, \ldots, X_n$ is a random sample from $\mathsf{B}(\alpha, \beta)$ distribution (see for instance chapter 8.2.7 Kulich, 2018), i.e. $\mathsf{E}\, X_i = \frac{\alpha}{\alpha+\beta}$ and $\operatorname{var} X_i = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. By the moment method we get consistent and asymptotically normal estimators

$$\widehat{\alpha}_n = \overline{X}_n\left(\frac{\overline{X}_n(1-\overline{X}_n)}{S_n^2} - 1\right) \quad \text{and} \quad \widehat{\beta}_n = (1-\overline{X}_n)\left(\frac{\overline{X}_n(1-\overline{X}_n)}{S_n^2} - 1\right)$$

(estimators are meaningful only if $S_n^2 < \overline{X}_n(1-\overline{X}_n)$).

Remark.

- Estimators obtained by the method of moments tend to have larger asymptotic variance compared to the estimators obtained by the method of maximum likelihood. Maximum likelihood theory will be discussed in detail in Mathematical Statistics 2.
- Using the implicit function theorem it can be proved that it is sufficient that $\boldsymbol{\tau}$ has continuous derivative on some neighbourhood of $(\mathsf{E}\, X_i, \operatorname{var} X_i)$.

3.4. Interval estimation

We are given a random sample $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$, a model $\mathcal{F}$ and a parameter $\theta = t(F) \in \mathbb{R}$ for $F \in \mathcal{F}$, which we need to estimate. Let $F_X \in \mathcal{F}$ be the true distribution of some random vector $\boldsymbol{X}_i$ and $\theta_X \equiv t(F_X)$ be the true value of the estimated parameter.

3.4.1. Definitions

Definition 3.5 An interval $B_n = B_n(\boldsymbol{X}) \subset \mathbb{R}$ is called a *confidence interval* for parameter $\theta_X \in \mathbb{R}$ with *confidence level* $1-\alpha$ in model $\mathcal{F}$ if and only if

$$\mathsf{P}[\omega \in \Omega : B_n(\omega) \ni \theta_X] = 1-\alpha, \quad \text{for every distribution } F_X \in \mathcal{F}.$$

An interval B_n is called an *asymptotic confidence interval* for parameter $\theta_X \in \mathbb{R}$ with *(asymptotic) confidence level* $1-\alpha$ in model $\mathcal{F}$ if and only if

$$\mathsf{P}[\omega \in \Omega : B_n(\omega) \ni \theta_X] \xrightarrow[n\to\infty]{} 1-\alpha \quad \text{for every distribution } F_X \in \mathcal{F}.$$

Remark.

- Interval B_n is random (calculated from the data) while the parameter θ_X is not. Expression $B_n \ni \theta_X$ is read as "interval B_n covers (the true value of) θ_X".
- Number $\alpha \in (0,1)$ is preselected; usually $\alpha = 0{,}05$ is chosen, which leads to confidence intervals with confidence levels of 0,95. However, we can also encounter intervals whose confidence levels are 0,90 or 0,99.
- It is not always possible or appropriate to calculate confidence intervals with exact prescribed coverage. We are often satisfied with asymptotic confidence intervals whose coverage converges to the prescribed level as the sample size increases.
- We defined confidence intervals only for real parameters. Nevertheless, similar concept can also be introduced for vector parameters: we need to find some random set B_n which covers the true value of the parameter with specified probability. This set is then called the *confidence set*. The shape of the set B_n, however, can be chosen in many different ways.

Remark. We distinguish between two-sided and one-sided confidence intervals (lower and upper).

- An interval of the form $\big(\eta_L(\boldsymbol{X}), \eta_U(\boldsymbol{X})\big)$, where $\eta_L(\boldsymbol{X})$ and $\eta_U(\boldsymbol{X})$ are two random variables satisfying $\mathsf{P}\big[\eta_L(\boldsymbol{X}) < \eta_U(\boldsymbol{X})\big] = 1$, $\eta_L(\boldsymbol{X}) > -\infty$ and $\eta_U(\boldsymbol{X}) < \infty$ a.s., is called *two-sided confidence interval*. Usually we construct it so that it holds (at least asymptotically) that

$$\mathsf{P}\big[\theta_X \le \eta_L(\boldsymbol{X})\big] = \frac{\alpha}{2}, \quad \mathsf{P}\big[\theta_X \ge \eta_U(\boldsymbol{X})\big] = \frac{\alpha}{2}.$$

- An interval of the form $\big(\eta_L(\boldsymbol{X}), \infty\big)$ is called *lower one-sided confidence interval*. We have that $\mathsf{P}\big[\eta_L(\boldsymbol{X}) < \theta_X\big] = 1-\alpha$.
- An interval of the form $\big(-\infty, \eta_U(\boldsymbol{X})\big)$ is called *upper one-sided confidence interval*. We have that $\mathsf{P}\big[\theta_X < \eta_U(\boldsymbol{X})\big] = 1-\alpha$.

Example (expectation in normal model with known variance)**.** Consider the problem of interval estimation of the expected value for normally distributed data with known variance.

Data: $X_1, \ldots, X_n \sim F_X$
Model: $F_X \in \mathcal{F} = \{\mathsf{N}(\mu, \sigma_X^2), \mu \in \mathbb{R}, \sigma_X^2 \text{ known}\}$
Estimated parameter: $\theta_X = \mathsf{E}\, X_i \equiv \mu_X$
Procedure:

1. We have an unbiased and consistent estimator of the parameter μ_X - the sample mean $\overline{X}_n$. We know that $\overline{X}_n \sim \mathsf{N}(\mu_X, \sigma_X^2/n)$. Thus
$$\frac{\sqrt{n}\,(\overline{X}_n - \mu_X)}{\sigma_X} \sim \mathsf{N}(0,1).$$
2. We will use the equality
$$\mathsf{P}\left[u_{\frac{\alpha}{2}} < \frac{\sqrt{n}\,(\overline{X}_n - \mu_X)}{\sigma_X} < u_{1-\alpha/2}\right] = 1-\alpha,$$
where $u_\alpha = \Phi^{-1}(\alpha)$ is α-quantile of the standard normal distribution and after several manipulations of the expression (using symmetry of the density of $\mathsf{N}(0,1)$ distribution around 0) we will arrive at
$$\mathsf{P}\left[\overline{X}_n - u_{1-\alpha/2}\frac{\sigma_X}{\sqrt{n}} < \mu_X < \overline{X}_n + u_{1-\alpha/2}\frac{\sigma_X}{\sqrt{n}}\right] = 1-\alpha.$$
3. We obtained a two-sided confidence interval (η_L, η_U). Its endpoints are
$$\eta_L(\boldsymbol{X}) = \overline{X}_n - u_{1-\alpha/2}\frac{\sigma_X}{\sqrt{n}}, \qquad \eta_U(\boldsymbol{X}) = \overline{X}_n + u_{1-\alpha/2}\frac{\sigma_X}{\sqrt{n}}.$$
Quantiles of the standard normal distribution which are needed for the construction of the confidence interval are listed in Table 3.1.
For $\alpha = 0{,}05$ we take quantile $u_{0{,}975} \doteq 1{,}96$ and obtain 95% two-sided confidence interval. This means that the interval covers the true value μ_X with probability 0,95.
4. One-sided interval would be obtained by a small modification of step 2. *Lower one-sided confidence interval* will be given as
$$\big(\eta_L(\boldsymbol{X}), \infty\big), \quad \text{where} \quad \eta_L(\boldsymbol{X}) = \overline{X}_n - u_{1-\alpha}\frac{\sigma_X}{\sqrt{n}}.$$
Upper one-sided confidence interval will be of the form
$$\big(-\infty, \eta_U(\boldsymbol{X})\big), \quad \text{where} \quad \eta_U(\boldsymbol{X}) = \overline{X}_n + u_{1-\alpha}\frac{\sigma_X}{\sqrt{n}}.$$

Table 3.1.: Some values of quantiles of the standard normal distribution.

κ	0,9	0,95	0,975	0,99	0,995
$u_\kappa = \Phi^{-1}(\kappa)$	1,282	1,645	1,960	2,326	2,576

One-sided confidence intervals differ from two-sided by the value of the normal quantile ($u_{1-\alpha}$ quantile is used instead of $u_{1-\alpha/2}$). For a 95% one-sided confidence interval we would take $u_{0,95} \doteq 1{,}645$.

Remark. Length of the confidence interval:
- decreases with increasing number of observations n,
- increases with increasing data variance σ_X^2,
- increases with increasing confidence level $1-\alpha$.

Example. Let $X_1, \ldots, X_n$ be a random sample from $\mathsf{N}(\mu_X, \sigma_X^2)$ distribution, the variance σ_X^2 is known. How many observations do we need so that the length of the two-sided confidence interval for the expected value μ_X does not exceed the specified limit $d > 0$?

We have that $2u_{1-\alpha/2}\,\sigma_X/\sqrt{n} \leq d$. Therefore we need at least $4u_{1-\alpha/2}^2\,\sigma_X^2/d^2$ observations. It is worth noting that if we want to halve the confidence interval, then we need to increase the sample size 4 times.

Lemma 3.2 (confidence interval after parameter transformation) If (η_L, η_U) is a(n) (asymptotic) confidence interval for parameter θ_X with the confidence level of $1-\alpha$ and if ψ is an increasing continuous real-valued function on the space of parameters $\Theta = \{t(F), F \in \mathcal{F}\} \subseteq \mathbb{R}$, then $\big(\psi(\eta_L), \psi(\eta_U)\big)$ is a(n) (asymptotic) confidence interval for parameter $\psi(\theta_X)$ with the confidence level of $1-\alpha$.

Proof. From the assumptions of the lemma we have that for a confidence interval with exact coverage it holds that

$$1-\alpha = \mathsf{P}\big[\eta_L(\boldsymbol{X}) < \theta_X < \eta_U(\boldsymbol{X})\big] = \mathsf{P}\big[\psi\big(\eta_L(\boldsymbol{X})\big) < \psi(\theta_X) < \psi\big(\eta_U(\boldsymbol{X})\big)\big].$$

Analogously for asymptotic confidence intervals. □

Example. Let $X_1, \ldots, X_n$ be a random sample from $\mathsf{Po}(\lambda)$ distribution. Then according to the example on page **??** we know that

$$\sqrt{n}\left(2\sqrt{\overline{X}_n} - 2\sqrt{\lambda_X}\right) \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}(0,1).$$

From this result we can easily deduce that the asymptotic confidence interval for $\sqrt{\lambda_X}$ is given as

$$\left(\sqrt{\overline{X}_n} - \frac{u_{1-\alpha/2}}{2\sqrt{n}}, \sqrt{\overline{X}_n} + \frac{u_{1-\alpha/2}}{2\sqrt{n}}\right).$$

And thus the confidence interval for λ_X is given as

$$\left(\left[\max\left\{\sqrt{\overline{X}_n} - \frac{u_{1-\alpha/2}}{2\sqrt{n}}, 0\right\}\right]^2, \left[\sqrt{\overline{X}_n} + \frac{u_{1-\alpha/2}}{2\sqrt{n}}\right]^2\right).$$

3.4.2. Construction of confidence intervals

Let $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, where $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ is a random sample from some distribution $F_X \in \mathcal{F}$. We need to estimate parameter $\theta_X = t(F_X) \in \mathbb{R}$. Let us briefly describe the general procedure for construction of two-sided confidence intervals for θ_X.

1. We will find a function $\varphi(\boldsymbol{x}, \theta_X)$ satisfying that for every $\boldsymbol{x}$ fixed it is, as a function of θ_X, injective and continuous and that the distribution of the random variable $Z_n \equiv \varphi(\boldsymbol{X}, \theta_X)$ is known at least asymptotically (it depends neither on θ_X nor on any other unknown parameters) and is non-degenerate. This random variable Z_n is called *pivotal*. For the construction of function φ it may be useful to start by calculating a point estimator of θ_X, whose distribution is usually known (at least asymptotically). Let us denote by F_Z the (exact or asymptotic) cumulative distribution function of Z_n and let $c_\alpha = F_Z^{-1}(\alpha)$ be α-quantile of the distribution given by F_Z.

2. We will use the formula

$$\mathsf{P}\big(c_{\alpha/2} < \varphi(\boldsymbol{X}, \theta_X) < c_{1-\alpha/2}\big) = 1 - \alpha \quad (\text{or } \to 1 - \alpha)$$

   and we will "isolate" θ_X. In order to do that, it is needed to invert $\varphi(\boldsymbol{x}, \theta)$ as a function of θ (for $\boldsymbol{x}$ fixed). Let $\bar{\varphi}(\boldsymbol{x}, t)$ be a function such that

$$\varphi\big(\boldsymbol{x}, \bar{\varphi}(\boldsymbol{x}, t)\big) = t \quad \text{and} \quad \bar{\varphi}\big(\boldsymbol{x}, \varphi(\boldsymbol{x}, \theta)\big) = \theta$$

   for every $\boldsymbol{x}$, t and θ. Since function $\bar{\varphi}(\boldsymbol{x}, t)$ is normally decreasing in t, we get that

$$\mathsf{P}\big(\bar{\varphi}(\boldsymbol{X}, c_{1-\alpha/2}) < \theta_X < \bar{\varphi}(\boldsymbol{X}, c_{\alpha/2})\big) = 1 - \alpha.$$

3. We obtained (asymptotic) confidence interval $\big(\eta_L(\boldsymbol{X}), \eta_U(\boldsymbol{X})\big)$ with confidence level of $1 - \alpha$, where $\eta_L(\boldsymbol{X}) = \bar{\varphi}(\boldsymbol{X}, c_{1-\alpha/2})$ and $\eta_U(\boldsymbol{X}) = \bar{\varphi}(\boldsymbol{X}, c_{\alpha/2})$.

Example (variance and standard deviation of the normal distribution)**.** Consider the problem of constructing a confidence interval for the standard deviation of the normal distribution.

Data: $X_1, \ldots, X_n \sim F_X$
Model: $F_X \in \mathcal{F} = \big\{\mathsf{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\big\}$
Estimated parameter: $\sigma_X = \sqrt{\mathsf{var}\, X_i}$
Procedure:

Let us first consider variance σ_X^2. Its unbiased and consistent estimator is S_n^2. According to theorem 2.8, part (i), we know that

$$\frac{(n-1)S_n^2}{\sigma_X^2} \sim \chi^2_{n-1}.$$

Thus, we will choose $Z_n = (n-1)S_n^2/\sigma_X^2$, $F_Z = \chi^2_{n-1}$ and $c_\alpha = \chi^2_{n-1}(\alpha)$, i.e. α-quantile of χ^2_{n-1} distribution (Table 3.2).

We will use the equality

$$\mathsf{P}\left[\chi^2_{n-1}(\alpha/2) < \frac{(n-1)S_n^2}{\sigma_X^2} < \chi^2_{n-1}(1-\alpha/2)\right] = 1-\alpha$$

and after several manipulations of the expression we will arrive at

$$\mathsf{P}\left[\frac{(n-1)S_n^2}{\chi^2_{n-1}(1-\alpha/2)} < \sigma_X^2 < \frac{(n-1)S_n^2}{\chi^2_{n-1}(\alpha/2)}\right] = 1-\alpha.$$

We obtained a confidence interval

$$\left(\frac{(n-1)S_n^2}{\chi^2_{n-1}(1-\alpha/2)}, \frac{(n-1)S_n^2}{\chi^2_{n-1}(\alpha/2)}\right) \tag{3.3}$$

for the variance σ_X^2 whose confidence level is $1-\alpha$.

Confidence interval for the standard deviation σ_X will be obtained by application of square root to both endpoints of the confidence interval for the variance

$$\left(\frac{\sqrt{n-1}\,S_n}{\sqrt{\chi^2_{n-1}(1-\alpha/2)}}, \frac{\sqrt{n-1}\,S_n}{\sqrt{\chi^2_{n-1}(\alpha/2)}}\right),$$

see also Lemma 3.2 (square root is an increasing and continuous function on $(0,\infty)$).

Example (expectation of the normal distribution with unknown variance)**.** Consider the problem of constructing a confidence interval for the expectation of the normal distribution with unknown variance.
Data: $X_1, \ldots, X_n \sim F_X$

Table 3.2.: Some values of quantiles $\chi^2_f(\kappa)$ of χ^2 distribution with f degrees of freedom.

	κ							
f	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99
5	0,554	0,831	1,145	1,610	9,236	11,070	12,833	15,086
10	2,558	3,247	3,940	4,865	15,987	18,307	20,483	23,209
15	5,229	6,262	7,261	8,547	22,307	24,996	27,488	30,578
25	11,524	13,120	14,611	16,473	34,382	37,652	40,646	44,314
100	70,065	74,222	77,929	82,358	118,498	124,342	129,561	135,807

Table 3.3.: Some values of $t_f(\kappa)$ quantiles of t distribution with f degrees of freedom.

	κ				
f	0,9	0,95	0,975	0,99	0,995
5	1,476	2,015	2,571	3,365	4,032
10	1,372	1,812	2,228	2,764	3,169
15	1,341	1,753	2,131	2,602	2,947
25	1,316	1,708	2,060	2,485	2,787
100	1,290	1,660	1,984	2,364	2,626
∞	1,282	1,645	1,960	2,326	2,576

Model: $F_X \in \mathcal{F} = \{\mathsf{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$
Estimated parameter: $\theta_X = \mathsf{E}\, X_i \equiv \mu_X$
Procedure:

The estimator $\overline{X}_n$ is unbiased and consistent for the parameter μ_X. Furthermore, S_n^2 is an unbiased and consistent estimator of $\sigma_X^2 \equiv \mathsf{var}\, X_i$. From theorem 2.10 we know that

$$T_n = \frac{\sqrt{n}\,(\overline{X}_n - \mu_X)}{S_n} \sim t_{n-1}.$$

Hence, we can take T_n as our pivotal random variable, F_Z will be cumulative distribution function of t_{n-1} distribution and $c_\alpha = t_{n-1}(\alpha)$ (α-quantile of t_{n-1} distribution). Some quantiles of t-distribution are listed in Table 3.3. Clearly, already for $n - 1 = 25$ they are only slightly larger than the corresponding quantiles of the standard normal distribution, to which they converge as the number of degrees of freedom increases above all bounds. Larger values of t-quantiles compared to the quantiles of the standard normal distribution, which were used in the introductory example, reflect increased variability of the pivotal random variable, which is caused by ignorance of the true variance.

We will use the equality

$$\mathsf{P}\left[t_{n-1}\left(\tfrac{\alpha}{2}\right) < \frac{\sqrt{n}(\overline{X}_n - \mu_X)}{S_n} < t_{n-1}\left(1 - \tfrac{\alpha}{2}\right)\right] = 1 - \alpha$$

and by the same procedure as in the case of the normal distribution with known variance we will arrive at the required confidence interval

$$\left(\overline{X}_n - t_{n-1}\left(1 - \tfrac{\alpha}{2}\right)\frac{S_n}{\sqrt{n}},\ \overline{X}_n + t_{n-1}\left(1 - \tfrac{\alpha}{2}\right)\frac{S_n}{\sqrt{n}}\right), \tag{3.4}$$

whose confidence level is exactly $1 - \alpha$.

Example (expected value of an arbitrary distribution with finite variance)**.** Consider the problem of constructing a confidence interval for the expectation without the assumption of normality.
Data: $X_1, \dots, X_n \sim F_X$
Model: $F_X \in \mathcal{F} = \mathcal{L}_+^2$ (all distributions with finite non-zero variance)
Estimated parameter: $\theta_X = \mathsf{E}\, X_i \equiv \mu_X$
Procedure: The estimator $\overline{X}_n$ is unbiased and consistent for the parameter μ_X. Furthermore, S_n^2 is an unbiased and consistent estimator of $\sigma_X^2 \equiv \mathsf{var}\, X_i$. From theorem 2.9 we know that

$$T_n = \frac{\sqrt{n}\,(\overline{X}_n - \mu_X)}{S_n} \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}(0,1).$$

We can thus choose T_n as our pivotal random variable.

We will use the following relation (which holds because T_n converges in distribution to the standard normal distribution)

$$\mathsf{P}\left[u_{\frac{\alpha}{2}} < \frac{\sqrt{n}\,(\overline{X}_n - \mu_X)}{S_n} < u_{1-\alpha/2}\right] \xrightarrow[n\to\infty]{} 1-\alpha.$$

Thus, one possible asymptotic confidence interval would be

$$\left(\overline{X}_n - u_{1-\alpha/2}\,\frac{S_n}{\sqrt{n}},\ \overline{X}_n + u_{1-\alpha/2}\,\frac{S_n}{\sqrt{n}}\right). \tag{3.5}$$

Since for $n \to \infty$ quantile $t_{n-1}(\alpha)$ converges to u_α (for arbitrary $0 < \alpha < 1$), it holds that interval (3.4), which was exact confidence interval for μ_X in case of a random sample from the normal distribution, is also a valid asymptotic confidence interval for μ_X for data coming from an arbitrary distribution with finite non-zero variance.

Note that $|t_{n-1}(\alpha)| > |u_\alpha|$ for every $n \geq 2$, therefore interval (3.4) is longer than interval (3.5). For caution, it is therefore recommended to use interval (3.4).

Example (alternative distribution)**.** Let us now present one possible way to construct an asymptotic confidence interval for the probability of success in the alternative distribution. (We will show several more confidence intervals related to this problem later.)

Data: $X_1, \dots, X_n \sim F_X$
Model: $F_X \in \mathcal{F} = \{\mathsf{Alt}(p), p \in (0,1)\}$
Estimated parameter: $p_X = \mathsf{E}\, X_i = \mathsf{P}[X_i = 1]$
Procedure:

Since we are estimating probability of an event, we will start by considering empirical relative frequency $\widehat{p}_n = \overline{X}_n$, which is an unbiased and consistent estimator of p (theorem 2.3). From the central limit theorem (theorem P.7.11) we know that $\sqrt{n}\,(\widehat{p}_n - p_X) \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}\big(0, p_X(1-p_X)\big)$. Thus,

$$\frac{\sqrt{n}\,(\widehat{p}_n - p_X)}{\sqrt{p_X(1-p_X)}} \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}(0,1).$$

Left-hand side is a non-linear function of p_X, but our situation can be simplified. From the consistency of $\widehat{p}_n$ and the continuous mapping theorem (theorem P.7.3) it follows that

$$\sqrt{\widehat{p}_n(1-\widehat{p}_n)} \xrightarrow[n\to\infty]{\mathsf{P}} \sqrt{p_X(1-p_X)}.$$

From Slutsky's theorem (theorem P.7.6) we obtain that

$$\frac{\sqrt{n}\,(\widehat{p}_n - p_X)}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}} = \frac{\sqrt{n}\,(\widehat{p}_n - p_X)}{\sqrt{p_X(1-p_X)}} \frac{\sqrt{p_X(1-p_X)}}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}} \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}(0,1). \tag{3.6}$$

Therefore, we can take $Z_n = \frac{\sqrt{n}\,(\widehat{p}_n - p_X)}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}$, $F_Z = \Phi$ and $c_\alpha = u_\alpha$ (α-quantile of the standard normal distribution).

From the following relation

$$\mathsf{P}\left[-u_{1-\alpha/2} < \frac{\sqrt{n}\,(\widehat{p}_n - p_X)}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}} < u_{1-\alpha/2}\right] \xrightarrow[n\to\infty]{} 1-\alpha$$

we get that

$$\mathsf{P}\left[\widehat{p}_n - u_{1-\alpha/2}\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}} < p_X < \widehat{p}_n + u_{1-\alpha/2}\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}\right] \xrightarrow[n\to\infty]{} 1-\alpha.$$

We obtained an asymptotic confidence interval

$$\left(\widehat{p}_n - u_{1-\alpha/2}\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}},\ \widehat{p}_n + u_{1-\alpha/2}\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}\right),$$

whose coverage probability converges to $1-\alpha$ as $n \to \infty$.

3.5. Empirical estimators

Consider a random sample $X_1, X_2, \ldots, X_n$ from a distribution F_X. We will present how to estimate some characteristics of the distribution F_X.

3.5.1. Empirical cumulative distribution function

Let us first focus on estimation of the whole distribution function $F_X(x)$ for $x \in \mathbb{R}$. We consider a model that includes all distributions on $\mathbb{R}$, i.e. we do not impose any conditions at all on the distribution function F_X.

Definition 3.6 Function $\widehat{F}_n(x) \overset{\mathsf{df}}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \le x\}$ is called the *empirical distribution function* of the random sample $X_1, X_2, \ldots, X_n$.

Remark. The value of $\widehat{F}_n$ at some point x is equal to the number of observations that do not exceed x which is then divided by the total number of observations. Function $\widehat{F}_n$ is non-decreasing, right-continuous, piecewise constant with jumps in observed values of random variables X_i, the magnitude of the jump at a point x is given by the number observations which are equal to x which is then divided by the total number of observations. Empirical distribution function has all the properties of a cumulative distribution function of some discrete distribution.

For some x fixed, is the value $\widehat{F}_n(x)$ actually equal to the relative frequency of the event $[X_i \leq x]$ calculated from n observations, while the probability of this event is equal to $F_X(x)$. From theorem 2.3 we immediately obtain the most important properties of empirical distribution functions.

Theorem 3.3 (properties of empirical distribution functions) For an arbitrary $x \in \mathbb{R}$ it holds that:

(i) $\mathsf{E}\,\widehat{F}_n(x) = F_X(x)$ (unbiasedness), $\mathsf{var}\left(\widehat{F}_n(x)\right) = \frac{F_X(x)[1-F_X(x)]}{n}$;

(ii) $\widehat{F}_n(x) \xrightarrow[n\to\infty]{\mathsf{P}} F_X(x)$ (pointwise consistency);

(iii) $\sqrt{n}\left[\widehat{F}_n(x) - F_X(x)\right] \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}\big(0, F_X(x)[1-F_X(x)]\big)$ (asymptotic normality);

(iv) $n\widehat{F}_n(x) \sim \mathsf{Bi}\big(n, F_X(x)\big)$;

(v) $\sup_{x\in\mathbb{R}} \left|\widehat{F}_n(x) - F_X(x)\right| \xrightarrow[n\to\infty]{\mathsf{P}} 0$ (uniform consistency).

Remark.

- Point (iii) of the previous theorem can be used to construct an asymptotic confidence interval for $F_X(x)$ in the same way as in the case of the parameter in the alternative distribution (see page 36).
- Point (v) is sometimes called the Glivenko-Cantelli theorem. It cannot be deduced from theorem 2.3 or from other results that are currently available. It will be proved in one of the more advanced lectures on the probability theory.

3.5.2. Idea behind empirical estimators

Estimators of many basic characteristics of the distribution F_X can be derived from the empirical distribution function. Let $\theta_X = t(F_X)$ be the parameter of interest. If it can be calculated from the true cumulative distribution function F_X, then it can also be calculated from the empirical distribution function $\widehat{F}_n$ in the same way. Thus, we obtain the estimator $\widehat{\theta}_n \overset{\mathsf{df}}{=} t\big(\widehat{F}_n\big)$. These types of estimators are called *empirical estimators*. We will see that empirical estimators often have reasonable properties.

Let us first demonstrate this procedure on the example of the empirical estimator of expectation. We have that

$$\mathsf{E}\,X_i = \int_{-\infty}^{\infty} x\, dF_X(x).$$

The empirical estimator of expectation is obtained by using $\widehat{F}_n$ instead of the unknown F_X. We will get

$$\int_{-\infty}^{\infty} x \, d\widehat{F}_n(x) = \int_{-\infty}^{\infty} x \, d\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\}\right) = \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} x \, d\, \mathbb{1}\{X_i \leq x\} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

where we used the fact that $G(x) = \mathbb{1}\{X_i \leq x\}$ is for fixed X_i actually the cumulative distribution function of a random variable that is equal to X_i with probability 1. We have, therefore, reached the conclusion that the empirical estimator of expectation is the sample mean, which we already know to be unbiased and consistent.

Remark. Let us fix $\omega \in \Omega$ and denote the observed realisations of random variables as $x_1 = X_1(\omega), \ldots, x_n = X_n(\omega)$. Then $\widehat{F}_n$ satisfies all the properties of a cumulative distribution function. If Y is some random variable whose cumulative distribution function is $\widehat{F}_n$, then the integral $\int_{-\infty}^{\infty} x \, d\widehat{F}_n(x)$ is equal to the expectation of Y. Since the distribution given by $\widehat{F}_n$ is discrete and satisfies that $\mathsf{P}(Y = x_i) = \frac{1}{n}$ for every $i = 1, \ldots, n$, then it holds that

$$\mathsf{E}\, Y = \sum_{i=1}^{n} x_i \, \mathsf{P}(Y = x_i) = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} X_i(\omega).$$

3.5.3. Empirical moment estimators

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution F_X and h be a measurable real-valued function such that $\mathsf{E}\,|h(X_i)| < \infty$. It is easy to verify that the empirical estimator of the parameter $\mathsf{E}\, h(X_i)$ is the sample mean of the observed values $h(X_i)$, i.e. $\frac{1}{n}\sum_{i=1}^{n} h(X_i)$. This estimator is unbiased and consistent.

Let us derive the *empirical estimator of the variance* $\sigma_X^2 = \mathsf{E}\, X_i^2 - (\mathsf{E}\, X_i)^2$. We know that the empirical estimator of $\mathsf{E}\, X_i$ is $\overline{X}_n$ and that the empirical estimator of $\mathsf{E}\, X_i^2$ is $\frac{1}{n}\sum_{i=1}^{n} X_i^2$. The empirical estimator of the variance is, therefore, given as

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

Remark. It holds that

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2 = \frac{n}{n-1}\widehat{\sigma}_n^2.$$

For n sufficiently large is the difference between $\widehat{\sigma}_n^2$ and S_n^2 small, because thanks to theorem 2.6(i)

$$\widehat{\sigma}_n^2 - S_n^2 = -\frac{S_n^2}{n} \xrightarrow[n\to\infty]{\mathsf{P}} 0.$$

It follows from theorem 2.6 that the sample variance S_n^2 is an unbiased and consistent estimator of σ_X^2. The empirical estimator of the variance $\widehat{\sigma}_n^2$ is consistent, however,

it is not unbiased. On the other hand, from the example on page 22 we know that in model $\mathcal{F} = \{\mathsf{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$ it holds that $\mathrm{MSE}(\widehat{\sigma}_n^2) < \mathrm{MSE}(S_n^2)$.

Similarly, we can derive empirical estimators for higher order moments. *Empirical estimators of non-central moments* $\mu_k' = \mathsf{E}\, X_i^k$ are

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

Empirical estimators of central moments $\mu_k = \mathsf{E}\,(X_i - \mathsf{E}\, X_i)^k$ are

$$\widehat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^k.$$

Empirical estimators of non-central moments are evidently unbiased as well as consistent. Empirical estimators of central moments are consistent. In general, however, they are not unbiased.

The empirical estimator of the skewness is

$$\widehat{\gamma}_3 = \frac{\widehat{\mu}_3}{(\widehat{\sigma}_n^2)^{3/2}},$$

The empirical estimator of the kurtosis is

$$\widehat{\gamma}_4 = \frac{\widehat{\mu}_4}{\widehat{\sigma}_n^4}.$$

Both of them are consistent (according to the continuous mapping theorem, theorem P.7.3).

Exercise. Prove that if $\mathsf{E}\,|X_i|^k < \infty$, then $\widehat{\mu}_k \xrightarrow[n\to\infty]{\mathsf{P}} \mu_k$.
Hint:

$$\widehat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{k}\binom{j}{k}X_i^k\big(-\overline{X}_n\big)^{k-j} = \sum_{j=0}^{k}\binom{j}{k}\Big(\frac{1}{n}\sum_{i=1}^{n}X_i^k\Big)\big(-\overline{X}_n\big)^{k-j}.$$

3.5.4. Empirical (sample) quantiles

Let α be a preselected number from the interval $(0,1)$. The *quantile function* of a given distribution F_X is defined as

$$F_X^{-1}(\alpha) = \inf\big\{x : F_X(x) \geq \alpha\big\}.$$

Then, α*-quantile* of distribution F_X is defined as $u_X(\alpha) = F_X^{-1}(\alpha)$. For α-quantile it holds that

$$F_X\big(u_X(\alpha)\big) \geq \alpha \quad \text{and} \quad F_X\big(u_X(\alpha) - h\big) < \alpha \text{ for } \forall h > 0.$$

As an empirical estimator, we use the value of α-quantile of the empirical distribution function, i.e.

$$\widehat{F}_n^{-1}(\alpha) = \inf\big\{x : \widehat{F}_n(x) \geq \alpha\big\}.$$

Definition 3.7 (Empirical quantile) For $\alpha \in (0,1)$ we define the *empirical (sample) α-quantile* as $\widehat{u}_n(\alpha) = \widehat{F}_n^{-1}(\alpha)$.

Remark.

- Recall that the empirical distribution function is piecewise constant with jumps at points $X_{(1)}, X_{(2)}, \dots, X_{(n)}$. Therefore, the empirical quantile will be (according to our definition) an appropriately chosen order statistic. Since it holds that
$$\widehat{F}_n\big(X_{(k)}\big) \geq \frac{k}{n} \quad \text{and} \quad \widehat{F}_n\big(X_{(k)} - h\big) < \frac{k}{n} \text{ for } \forall h > 0,$$
the empirical quantile will satisfy that
$$\widehat{u}_n(\alpha) = X_{(k_\alpha)}, \quad \text{where} \quad k_\alpha = \begin{cases} n\alpha & \text{for } (n\alpha) \in \mathbb{N}, \\ \lfloor n\alpha \rfloor + 1 & \text{for } (n\alpha) \notin \mathbb{N}. \end{cases}$$
Since we do not assume continuity of the distribution, the order statistics $X_{(k_\alpha)}$ must be understood in terms of the note on page **??**.
- For $\alpha = 0{,}5$ we get the *sample median*: $\widehat{m}_n = X_{(\frac{n+1}{2})}$ for n odd and $\widehat{m}_n = X_{(\frac{n}{2})}$ for n even.
- The empirical α-quantile satisfies inequalities
$$\widehat{F}_n\big(\widehat{u}_n(\alpha)\big) \geq \alpha \quad \text{and} \quad \lim_{h \searrow 0} \widehat{F}_n\big(\widehat{u}_n(\alpha) - h\big) < \alpha,$$
i.e. at least $n\alpha$ observations are less than or equal to $\widehat{u}_n(\alpha)$ and, simultaneously, for every $h > 0$ at least $n(1-\alpha)$ observation are greater than or equal to $\widehat{u}_n(\alpha) - h$.
- There are many different definitions of the empirical α-quantile (typically some linear interpolation between points $X_{(k_\alpha - 1)}$, $X_{(k_\alpha)}$ and $X_{(k_\alpha + 1)}$). For example for n even is the sample median often defined as
$$\widehat{m}_n = \frac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2}.$$

The following lemma characterises the empirical quantile as a solution of some minimization problem (compare with lemma 2.1).

Lemma 3.4 Let $\alpha \in (0,1)$. For the empirical α-quantile $\widehat{u}_n(\alpha)$ it holds that
$$\widehat{u}_n(\alpha) = \arg\min_{c \in \mathbb{R}} \sum_{i=1}^{n} \varrho_\alpha(X_i - c),$$
where $\varrho_\alpha(u) = \alpha\, u \mathbb{1}\{u \geq 0\} + (1-\alpha)(-u)\mathbb{1}\{u < 0\}$.

Note that for $\alpha = \frac{1}{2}$ we obtain that $\varrho_{1/2}(u) = \frac{1}{2}|u|$. Since the constant $\frac{1}{2}$ is for the optimization irrelevant, it holds that the sample median satisfies
$$\widehat{m}_n = \arg\min_{c \in \mathbb{R}} \sum_{i=1}^{n} \big|X_i - c\big|,$$
i.e. $\widehat{m}_n$ minimizes the sum of absolute deviations.

Remark. The minimization problem from part (ii) can be formulated as a problem of linear programming in the form

$$\arg\min_{c\in\mathbb{R}}\Big[-(1-\alpha)\sum_{i:X_i<c}\big(X_i-c\big)+\alpha\sum_{i:X_i\geq c}\big(X_i-c\big)\Big].$$

If we introduce the notation $U_i = (X_i - c)\mathbb{1}(X_i \geq c)$, $V_i = -(X_i - c)\mathbb{1}(X_i < c)$, $\boldsymbol{U} = (U_1,\ldots,U_n)^\mathsf{T}$, $\boldsymbol{V} = (V_1,\ldots,V_n)^\mathsf{T}$, $\boldsymbol{X} = (X_1,\ldots,X_n)^\mathsf{T}$, our problem can be reformulated as an optimization problem of linear programming in $(2n+1)$-dimensional space

$$\min_{\boldsymbol{U},\boldsymbol{V},c}\alpha\mathbf{1}_n^\mathsf{T}\boldsymbol{U}+(1-\alpha)\mathbf{1}_n^\mathsf{T}\boldsymbol{V}$$

subject to

$$c\mathbf{1}_n+\boldsymbol{U}-\boldsymbol{V}=\boldsymbol{X},\quad \boldsymbol{U}\geq 0,\quad \boldsymbol{V}\geq 0.$$

Naturally, this minimization problem does not have to have a unique solution. The minimum can be attained at every point from some interval.

Properties of empirical quantiles will be studied (proved) only in continuous distributions with increasing cumulative distribution functions F_X and densities f_X.

Theorem 3.5 Let $\alpha \in (0,1)$. Let $X_1,\ldots,X_n$ be a random sample from a distribution whose cumulative distribution function F_X is continuous and increasing on some neighbourhood of $u_X(\alpha)$.

(i) Then $\widehat{u}_n(\alpha)\xrightarrow[n\to\infty]{\mathsf{P}} u_X(\alpha)$.

(ii) Additionally, if there exists density f_X, which is continuous and non-zero at $u_X(\alpha)$, then

$$\sqrt{n}\big[\widehat{u}_n(\alpha)-u_X(\alpha)\big]\xrightarrow[n\to\infty]{\mathsf{d}}\mathsf{N}\big(0,V(\alpha)\big),\quad \text{where}\quad V(\alpha)=\frac{\alpha(1-\alpha)}{f_X^2(u_X(\alpha))}.$$

Proof. Part (i): Let $\varepsilon > 0$. We need to prove that

$$\mathsf{P}\big(|\widehat{u}_n(\alpha)-u_X(\alpha)|>\varepsilon\big)\xrightarrow[n\to\infty]{}0.$$

In order to do that, it is sufficient to show that

$$\mathsf{P}\big(\widehat{u}_n(\alpha)<u_X(\alpha)-\varepsilon\big)\xrightarrow[n\to\infty]{}0\quad\text{and}\quad\mathsf{P}\big(\widehat{u}_n(\alpha)>u_X(\alpha)+\varepsilon\big)\xrightarrow[n\to\infty]{}0.$$

So let us calculate

$$\begin{aligned}\mathsf{P}\big(\widehat{u}_n(\alpha)<u_X(\alpha)-\varepsilon\big)&=\mathsf{P}\big(X_{(k_\alpha)}<u_X(\alpha)-\varepsilon\big)\\&=\mathsf{P}\Big(\textstyle\sum_{i=1}^n\mathbb{1}\big\{X_i<u_X(\alpha)-\varepsilon\big\}\geq k_\alpha\Big)\\&\leq\mathsf{P}\Big(\widehat{F}_n\big(u_X(\alpha)-\varepsilon\big)-F_X\big(u_X(\alpha)-\varepsilon\big)\geq\tfrac{k_\alpha}{n}-F_X\big(u_X(\alpha)-\varepsilon\big)\Big).\end{aligned}\tag{3.7}$$

From theorem 3.3 it follows that

$$\widehat{F}_n\big(u_X(\alpha)-\varepsilon\big)-F_X\big(u_X(\alpha)-\varepsilon\big)\xrightarrow[n\to\infty]{\mathsf{P}}0, \tag{3.8}$$

and from the assumptions of this theorem we have that

$$\tfrac{k_\alpha}{n}-F_X\big(u_X(\alpha)-\varepsilon\big)\xrightarrow[n\to\infty]{}\alpha-F_X\big(u_X(\alpha)-\varepsilon\big)>0. \tag{3.9}$$

By combining (3.8) and (3.9) we obtain that the right-hand side of equality (3.7) converges to zero, thus we have proved that $\mathsf{P}\big(\widehat{u}_n(\alpha)<u_X(\alpha)-\varepsilon\big)\xrightarrow[n\to\infty]{}0$.

Similarly

$$\begin{aligned}\mathsf{P}\big(\widehat{u}_n(\alpha)>u_X(\alpha)+\varepsilon\big)&=\mathsf{P}\Big(\textstyle\sum_{i=1}^n\mathbb{1}\{X_i\le u_X(\alpha)+\varepsilon\}<k_\alpha\Big)\\&\le\mathsf{P}\Big(\widehat{F}_n\big(u_X(\alpha)+\varepsilon\big)-F_X\big(u_X(\alpha)+\varepsilon\big)<\tfrac{k_\alpha}{n}-F_X\big(u_X(\alpha)+\varepsilon\big)\Big).\end{aligned} \tag{3.10}$$

From theorem 3.3 it follows that

$$\widehat{F}_n\big(u_X(\alpha)+\varepsilon\big)-F_X\big(u_X(\alpha)+\varepsilon\big)\xrightarrow[n\to\infty]{\mathsf{P}}0, \tag{3.11}$$

and from the assumptions of this theorem we have that

$$\tfrac{k_\alpha}{n}-F_X\big(u_X(\alpha)+\varepsilon\big)\xrightarrow[n\to\infty]{}\alpha-F_X\big(u_X(\alpha)+\varepsilon\big)<0. \tag{3.12}$$

By combining (3.11) and (3.12) we obtain that the right-hand side of equality (3.10) converges to zero, thus we have proved that $\mathsf{P}\big(\widehat{u}_n(\alpha)>u_X(\alpha)+\varepsilon\big)\xrightarrow[n\to\infty]{}0$.

Part (ii): * Similarly as in the part (i) let us calculate

$$\begin{aligned}\mathsf{P}\Big(\sqrt{n}\big[\widehat{u}_n(\alpha)-u_X(\alpha)\big]\le x\Big)&=\mathsf{P}\Big(\widehat{u}_n(\alpha)\le u_X(\alpha)+\tfrac{x}{\sqrt{n}}\Big)\\&=\mathsf{P}\Big(\widehat{F}_n\big(u_X(\alpha)+\tfrac{x}{\sqrt{n}}\big)-F_X\big(u_X(\alpha)+\tfrac{x}{\sqrt{n}}\big)\ge\tfrac{k_\alpha}{n}-F_X\big(u_X(\alpha)+\tfrac{x}{\sqrt{n}}\big)\Big).\\&=\mathsf{P}\big(Z_n\ge x_n\big),\end{aligned}$$

where

$$Z_n=\frac{\sqrt{n}\big[\widehat{F}_n\big(u_X(\alpha)+\frac{x}{\sqrt{n}}\big)-F_X\big(u_X(\alpha)+\frac{x}{\sqrt{n}}\big)\big]}{\sqrt{\alpha(1-\alpha)}}$$

and

$$x_n=\frac{\sqrt{n}\big[\frac{k_\alpha}{n}-F_X\big(u_X(\alpha)-\frac{x}{\sqrt{n}}\big)\big]}{\sqrt{\alpha(1-\alpha)}}\,.$$

* This part of the proof was not done in the lecture.

From the central limit theorem for triangular arrays (e.g. Theorem 16.4 Lachout, 2004) it follows that $Z_n \xrightarrow[n\to\infty]{\mathrm{d}} Z$, where $Z \sim \mathsf{N}(0,1)$. Furthermore, from the assumptions of the theorem we get that $x_n \xrightarrow[n\to\infty]{} \frac{-x\, f_X\left(u_X(\alpha)\right)}{\sqrt{\alpha(1-\alpha)}}$. So in total we have that

$$\mathsf{P}\Big(\sqrt{n}\big[\widehat{u}_n(\alpha) - u_X(\alpha)\big] \le x\Big) \xrightarrow[n\to\infty]{} \mathsf{P}\Big(Z \ge \frac{-x\, f_X\left(u_X(\alpha)\right)}{\sqrt{\alpha(1-\alpha)}}\Big) = \mathsf{P}\Big(Z \le \frac{x\, f_X\left(u_X(\alpha)\right)}{\sqrt{\alpha(1-\alpha)}}\Big),$$

which (together with the definition of convergence in distribution) implies the statement of the theorem. □

The asymptotic variance $V(\alpha)$ of the empirical quantile is difficult to estimate because we do not have a universally applicable and reliable estimator of the density. Under the assumption that F_X is continuous at $u_X(\alpha)$, it is possible to use order statistics to construct a confidence interval.

For example *two-sided confidence interval* for $u_X(\alpha)$ with confidence level of $1 - \beta$ can be found in the form of $(X_{(k_L)}, X_{(k_U)})$. To determine numbers k_L and k_U let us observe that

$$\mathsf{P}\Big(X_{(k_L)} \ge u_X(\alpha)\Big) = \mathsf{P}\Big(\sum_{i=1}^{n} \mathbb{1}\big\{X_i < u_X(\alpha)\big\} \le k_L - 1\Big) = \mathsf{P}\Big(\mathsf{Bi}(n,\alpha) \le k_L - 1\Big),$$
$$\mathsf{P}\Big(X_{(k_U)} \le u_X(\alpha)\Big) = \mathsf{P}\Big(\sum_{i=1}^{n} \mathbb{1}\big\{X_i \le u_X(\alpha)\big\} \ge k_U\Big) = \mathsf{P}\Big(\mathsf{Bi}\big(n,\alpha\big) \ge k_U\Big).$$

Therefore, numbers k_L and k_U can be found using the binomial distribution as the largest and smallest natural numbers such that

$$\mathsf{P}\Big(\mathsf{Bi}(n,\alpha) \le k_L - 1\Big) \le \tfrac{\beta}{2}, \qquad \mathsf{P}\Big(\mathsf{Bi}\big(n,\alpha\big) \ge k_U\Big) \le \tfrac{\beta}{2}.$$

If it is not feasible to work directly with the binomial distribution, we can approximate it by the normal distribution. In this case it is good to notice that

$$\mathsf{P}\Big(\mathsf{Bi}(n,\alpha) \le k_L - 1\Big) = \mathsf{P}\Big(\mathsf{Bi}(n,\alpha) < k_L\Big) \text{ and } \mathsf{P}\Big(\mathsf{Bi}\big(n,\alpha\big) \ge k_U\Big) = \mathsf{P}\Big(\mathsf{Bi}\big(n,\alpha\big) > k_U - 1\Big).$$

Therefore, as a "compromise" before the normal approximation, we proceed from the following equations

$$\mathsf{P}\Big(X_{(k_L)} \ge u_X(\alpha)\Big) = \mathsf{P}\Big(\mathsf{Bi}(n,\alpha) < k_L - \tfrac{1}{2}\Big), \qquad \mathsf{P}\Big(X_{(k_U)} \le u_X(\alpha)\Big) = \mathsf{P}\Big(\mathsf{Bi}\big(n,\alpha\big) > k_U - \tfrac{1}{2}\Big).$$

Now, using the normal approximation

$$\mathsf{P}\Big(\mathsf{Bi}(n,\alpha) < k_L - \tfrac{1}{2}\Big) = \mathsf{P}\Big(\tfrac{\mathsf{Bi}(n,\alpha)-n\alpha}{\sqrt{n\alpha(1-\alpha)}} < \tfrac{k_L-\frac{1}{2}-n\alpha}{\sqrt{n\alpha(1-\alpha)}}\Big) \doteq \Phi\Big(\tfrac{k_L-\frac{1}{2}-n\alpha}{\sqrt{n\alpha(1-\alpha)}}\Big),$$
$$\mathsf{P}\Big(\mathsf{Bi}(n,\alpha) > k_U - \tfrac{1}{2}\Big) = \mathsf{P}\Big(\tfrac{\mathsf{Bi}(n,\alpha)-n\alpha}{\sqrt{n\alpha(1-\alpha)}} > \tfrac{k_U-\frac{1}{2}-n\alpha}{\sqrt{n\alpha(1-\alpha)}}\Big) \doteq 1 - \Phi\Big(\tfrac{k_U-\frac{1}{2}-n\alpha}{\sqrt{n\alpha(1-\alpha)}}\Big).$$

From here we can already express the approximate values k_L a k_U

$$k_L = \left\lfloor \tfrac{1}{2} + n\alpha - u_{1-\frac{\beta}{2}}\sqrt{n\alpha(1-\alpha)} \right\rfloor, \quad k_U = \left\lceil \tfrac{1}{2} + n\alpha + u_{1-\frac{\beta}{2}}\sqrt{n\alpha(1-\alpha)} \right\rceil.$$

The aforementioned "compromise" is usually called the *continuity correction*. The purpose of this "correction", however, is not to make something continuous out of something discontinuous. It is a certain caution in case that a discrete distribution (in our case binomial) is approximated by a continuous one (in our case normal).

Remark. For small sample sizes n and α close to zero or one it can happen that either $\mathsf{P}\big(\mathsf{Bi}(n,\alpha) = 0\big) > \frac{\beta}{2}$ or $\mathsf{P}\big(\mathsf{Bi}(n,\alpha) = n\big) > \frac{\beta}{2}$. In that case we choose the lower (or the upper) bound of our confidence interval to be equal to $-\infty$ (or $+\infty$).

Exercise. Show that if we omit the assumption of continuity of the cumulative distribution function at the estimated quantile $u_X(\alpha)$, then the closed interval $\langle X_{(k_L)}, X_{(k_U)}\rangle$ will have (for n sufficiently large) probability of coverage at least $1-\beta$.

3.5.5. Empirical estimators for random vectors

Empirical estimators of first two moments can be easily generalised to random vectors. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample of independent k-dimensional random vectors from a distribution F_X. Individual components of the vector $\boldsymbol{X}_i$ will be denoted by X_{ij}, $i = 1, \ldots, n$, $j \in \{1, \ldots, k\}$. Further, let us denote

$$\boldsymbol{\mu} = \mathsf{E}\,\boldsymbol{X}_i, \qquad \Sigma = \mathsf{var}\,\boldsymbol{X}_i.$$

The empirical estimator of $\boldsymbol{\mu}$ is apparently the vector of empirical estimators of its individual components, i.e. k-dimensional sample mean

$$\overline{\boldsymbol{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i.$$

The empirical estimator of the variance matrix Σ can be obtained from the following representation

$$\Sigma = \mathsf{E}\,\big(\boldsymbol{X}_i - \mathsf{E}\,\boldsymbol{X}_i\big)\big(\boldsymbol{X}_i - \mathsf{E}\,\boldsymbol{X}_i\big)^{\mathsf{T}} = \mathsf{E}\,\boldsymbol{X}_i\boldsymbol{X}_i^{\mathsf{T}} - (\mathsf{E}\,\boldsymbol{X}_i)(\mathsf{E}\,\boldsymbol{X}_i)^{\mathsf{T}} = \mathsf{E}\,\boldsymbol{X}_i^{\otimes 2} - (\mathsf{E}\,\boldsymbol{X}_i)^{\otimes 2}$$

if we replace the expected values by their empirical estimators (i.e. sample means). Thus, we obtain

$$\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i^{\otimes 2} - \overline{\boldsymbol{X}}_n^{\otimes 2} = \frac{1}{n}\sum_{i=1}^{n} \big(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n\big)\big(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n\big)^{\mathsf{T}}.$$

Nevertheless, usually so called *sample covariation matrix* is used. It is defined as a multidimensional analogy of the sample variance S_n^2:

$$\boldsymbol{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n} \big(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n\big)\big(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n\big)^{\mathsf{T}}.$$

Remark.

- Diagonal elements of $\boldsymbol{S}_n^2$ are sample variances of individual components, i.e.

$$S_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)^2,$$

  for $j \in \{1,\dots,k\}$, where $\overline{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$.

- Element (j,m) of the matrix $\boldsymbol{S}_n^2$ is given by the expression

$$S_{jm} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)(X_{im} - \overline{X}_m)$$

  for $j \in \{1,\dots,k\}$ and $m \in \{1,\dots,k\}$, $j \neq m$. This random variable estimates the covariance $\mathsf{cov}\,(X_{ij}, X_{im})$ between j-th a m-th component of $\boldsymbol{X}_i$. It is called the *sample covariance*.

- $\boldsymbol{S}_n^2$ is positive semi-definite and it holds that

$$\boldsymbol{S}_n^2 = \frac{n}{n-1}\widehat{\Sigma}_n = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i^{\otimes 2} - \overline{\boldsymbol{X}}_n^{\otimes 2}\right).$$

The following assertion shows that both $\overline{\boldsymbol{X}}_n$ and $\boldsymbol{S}_n^2$ are unbiased and consistent estimators.

Proposition 3.6

(i) If $\mathsf{E}\left|X_{ij}\right| < \infty$ for every $j \in \{1,\dots,k\}$, then $\mathsf{E}\,\overline{\boldsymbol{X}}_n = \boldsymbol{\mu}$ and $\overline{\boldsymbol{X}}_n \xrightarrow[n\to\infty]{\mathsf{P}} \boldsymbol{\mu}$.

(ii) If $\mathsf{var}\,(X_{ij}) < \infty$ for every $j \in \{1,\dots,k\}$, then $\mathsf{E}\,\boldsymbol{S}_n^2 = \Sigma$ and $\boldsymbol{S}_n^2 \xrightarrow[n\to\infty]{\mathsf{P}} \Sigma$.

Proof. Part (i): Follows directly from theorem 2.2, which we use componentwise.

Part (ii): Consistency of $\boldsymbol{S}_n^2$ can be proved analogously as in the case of S_n^2 (see theorem 2.6(i)).

Unbiasedness can be proved in the following way:

$$\begin{aligned}
\mathsf{E}\,\boldsymbol{S}_n^2 &= \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}\mathsf{E}\,\boldsymbol{X}_i^{\otimes 2} - \mathsf{E}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\right)^{\otimes 2}\right] \\
&= \frac{n}{n-1}\left(\mathsf{E}\,\boldsymbol{X}_i^{\otimes 2} - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathsf{E}\,\boldsymbol{X}_i\boldsymbol{X}_j^{\mathsf{T}}\right) \\
&= \frac{n}{n-1}\left(\mathsf{E}\,\boldsymbol{X}_i^{\otimes 2} - \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{E}\,\boldsymbol{X}_i^{\otimes 2} - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathsf{E}\,\boldsymbol{X}_i\boldsymbol{X}_j^{\mathsf{T}}\right) \\
&= \frac{n}{n-1}\left[\mathsf{E}\,\boldsymbol{X}_i^{\otimes 2}\left(1 - \frac{1}{n}\right) - \frac{n-1}{n}\left(\mathsf{E}\,\boldsymbol{X}_i\right)^{\otimes 2}\right] = \Sigma.
\end{aligned}$$

□

*Recall the definition of the correlation coefficient of the random variables X_{ij} and X_{im}:

$$\varrho(X_{ij}, X_{im}) = \frac{\operatorname{cov}(X_{ij}, X_{im})}{\sqrt{\operatorname{var} X_{ij} \operatorname{var} X_{im}}}.$$

It is logical to define the sample correlation coefficient as the empirical estimator of this parameter, composed of empirical estimators of individual components.

Definition 3.8 The *sample correlation coefficient* $\widehat{\varrho}_{jm}$ of variables X_{ij} and X_{im}, $j \in \{1, \dots, k\}$ and $m \in \{1, \dots, k\}$, $j \neq m$, is defined as

$$\widehat{\varrho}_{jm} = \frac{S_{jm}}{S_j S_m} = \frac{\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)(X_{im} - \overline{X}_m)}{\sqrt{\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)^2 \sum_{i=1}^{n}(X_{im} - \overline{X}_m)^2}}.$$

Remark.

- $-1 \leq \widehat{\varrho}_{jm} \leq 1$ (see the Cauchy-Schwarz inequality).
- $\widehat{\varrho}_{jm} = 1$ (or -1) if and only if there exist constants $a \in \mathbb{R}$ and $b > 0$ (or $b < 0$) such that $X_{ij} = a + bX_{im}$ for every $i = 1, \dots, n$.
- $\widehat{\varrho}_{jm}$ is a consistent estimator of the correlation coefficient $\varrho(X_{ij}, X_{im})$ (this follows from consistency of $\boldsymbol{S}_n^2$ and theorem 1.1). But it is not unbiased.

Exercise. Prove that $\widehat{\varrho}_{jm} \xrightarrow[n\to\infty]{\mathsf{P}} \varrho(X_{ij}, X_{im})$.

* The rest of the chapter was not lectured in 2020/21.

Sample examples for the preparation for the exam.

1. Consider a random sample $X_1, \ldots, X_n$ from a distribution given by the density $f(x;\delta) = \frac{e^{-x/\delta}}{\delta}\mathbb{1}\{x > 0\}$, where $\delta > 0$ is an unknown parameter. Consider the estimator $\widehat{\delta}_n = \overline{X}_n$. Show that it is an unbiased estimator of δ_X. Further, consider the estimator $\widetilde{\delta}_n(a) = a\overline{X}_n$, where a is a constant. Find a which minimizes the mean squared error of $\widetilde{\delta}_n(a)$.

2. Consider a random sample $X_1, \ldots, X_n$ from the alternative distribution with some parameter p_X. Estimate the parameter p_X by the method of moments and then transform this estimator to create an estimator of $\theta_X = p_X(1-p_X)$. Examine the unbiasedness and consistency of this new estimator of the variance. How is it different from the ordinary sample variance?

3. Consider a random sample $X_1, \ldots, X_n$ from the alternative distribution with some parameter p_X. From the example on page 36 we know that an asymptotic confidence interval for the parameter p_X whose confidence level is $1-\alpha$ is

$$\left(\widehat{p}_n - u_{1-\alpha/2}\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}},\ \widehat{p}_n + u_{1-\alpha/2}\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}\right).$$

   Using this information derive a confidence interval for the parameter $\theta_X = p_X(1-p_X)$.
   Suppose that the confidence interval for the parameter p_X was calculated from the data. Interval $(0.35, 0.55)$ was obtained. In that case, how does the confidence interval for the parameter $\theta_X = p_X(1-p_X)$ look?

4. Let $X_1, \ldots, X_n$ be a random sample from $\mathsf{N}(\mu_X, 9)$ distribution. How many observations do we need so that the length of the confidence interval for μ_X with the confidence level of 0,90 is at most 0,25?

5. Let $\overline{X}_n$ be the sample mean of a random sample $X_1, \ldots, X_n$ from $\mathrm{Po}(\lambda_X)$ distribution. Determine the asymptotic distribution of the sample mean $\overline{X}_n$ and based on this distribution construct an asymptotic confidence interval for the parameter $\theta_X = \exp\{-\lambda_X\}$.

6. Let $X_1, \ldots, X_n$ be a random sample from the uniform distribution $\mathsf{R}(0,1)$. Let $k_n = \lceil \sqrt{n} \rceil$. Prove that $X_{(k_n)} \xrightarrow[n\to\infty]{\mathsf{P}} 0$.

A. Appendix

A.1. χ^2 and t distribution

Definition A.1 (χ^2-distribution) Let $Y = X_1, \ldots, X_k$ be independent and identically distributed random variables with distribution $\mathsf{N}(0,1)$. Then the distribution of the random variable $\sum_{i=1}^{k} X_i^2$ is the χ^2-distribution of k degrees of freedom. We write that $Y \sim \chi_k^2$.

Definition A.2 (t-distribution) Let $X \sim \mathsf{N}(0,1)$ and $Z \sim \chi_k^2$ be independent. Then the distribution of the random variable $T \overset{\mathsf{df}}{=} \frac{X}{\sqrt{Z/k}}$ is called the [Student] *t distribution* with k degrees of freedom. We write $T \sim t_k$.

A.2. Idempotentní matice

Definition A.3 The squared matrix $\mathbb{A}$ (of dimension $n \times n$) is **idempotent**, when $\mathbb{A}\mathbb{A} = \mathbb{A}$.

Lemma A.1 Let $\boldsymbol{X} \sim \mathsf{N}_n(\mathbf{0}, \Sigma)$ and $\mathbb{A}$ be a positively semidefinite matrix of dimension $n \times n$ such that $\mathbb{A}\Sigma$ is non-null and idempotent. Then

$$\boldsymbol{X}^{\mathsf{T}} \mathbb{A} \boldsymbol{X} \sim \chi^2_{\mathsf{tr}(\mathbb{A}\Sigma)}.$$

Bibliography

Kulich, M. (2018). Základy teorie pravděpodobnosti pro předmět Matematická statistika 1. https://www.karlin.mff.cuni.cz/~kulich/vyuka/ms1/doc/pravdepodobnost_ms1.pdf.

Lachout, P. (2004). *Teorie pravděpodobnosti*. Karolinum. Skripta.