NMSA331 Mathematical statistics 1

Lecture notes

Last updated November 18, 2025.

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

This text originated as a translation of the course notes for the course Mathematical Statistics 1. The core of the course notes were originally prepared by ***Michal Kulich*** *and then modified and extended by* ***Marek Omelka****. Many thanks to* ***Filip Kulla*** *and* ***Martina Petráková*** *who helped a lot with translating the course notes to the English language. The translation of the course notes were supported by the Student Faculty Grant (SFG).*

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
$\xrightarrow{\text{a.s.}}$	convergence almost surely
$\xrightarrow{\mathsf{d}}$	convergence in distribution
$X \sim \mathcal{L}$	X has the exact distribution $\mathcal{L}$
$X \overset{\text{as.}}{\sim} \mathcal{L}$	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 variab
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 $d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\sum_{j=1}^{k}(x_j - y_j)^2}$.

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 device 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

(i) $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{a.s.}} \boldsymbol{X} \;\Rightarrow\; \boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{P}} \boldsymbol{X}$

(ii) $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{P}} \boldsymbol{X} \;\Rightarrow\; \boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{d}} \boldsymbol{X}$

Remark. The opposite implication does not hold. Nevertheless if the random vectors converge to a constant, i.e. $\boldsymbol{X}_n \xrightarrow{\text{d}} \boldsymbol{c}$ (where $\boldsymbol{c} \in \mathbb{R}^k$) then also $\boldsymbol{X}_n \xrightarrow{\text{P}} \boldsymbol{c}$.

Proposition 1.2 (**Continuous Mapping Theorem, CMT**) Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be continuous in each point of an open set $C \subseteq \mathbb{R}^k$ such that $\mathsf{P}(\boldsymbol{X} \in C) = 1$. Then

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

Proposition 1.3 (**Cramér-Slutsky, CS**) Let $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{d}} \boldsymbol{X}$, $\boldsymbol{Y}_n \xrightarrow[n\to\infty]{\text{P}} \boldsymbol{c}$, then

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

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.

Exercise. Write down the above definitions and propositions for the special case $k = 1$ and $m = 1$. Do you know any examples where these statements and definitions are used?

1.2. Basic asymptotic results

Proposition 1.4 (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 1.5 (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

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

1.3. Δ-METHOD

Let $\boldsymbol{T}_n = (T_{n1}, \dots, T_{nk})^\mathsf{T}$ be an estimator of a k-dimensional parameter $\boldsymbol{\mu} = (\mu_1, \dots, \mu_k)^\mathsf{T}$ and $\boldsymbol{g} = (g_1, \dots, g_m)^\mathsf{T}$ 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.

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

Proposition 1.6 (Δ-method) Let

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

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

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

Theorem 1.6 is most often applied for $k = m = 1$ and $T_n = \overline{X}_n$, where $X_1, \dots, X_n$ are i.i.d. random variables. Then by the central limit theorem

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

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

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

Sometimes instead of (1.1) we write shortly $g(\overline{X}_n) \overset{\mathsf{as}}{\approx} \mathsf{N}\big(g(\mu), \frac{[g'(\mu)]^2\,\mathsf{var}\,(X_i)}{n}\big)$. 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 $n \to \infty$.

Exercise. 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 $Y_n \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}(0,1)$. Thus $\mathsf{avar}\,(Y_n) = 1$. On the other hand $\mathsf{var}\,(Y_n) = 2$ for each $n \in \mathbb{N}$.

Exercise. 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. $\theta_X = \log\left(\frac{p_X}{1-p_X}\right)$. Compare the variance and the asymptotic variance of $\widehat{\theta}_X = \log\left(\frac{\overline{X}_n}{1-\overline{X}_n}\right)$.

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 $\mathcal{F}$, 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)**.**

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

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. $\overline{X}_n \overset{\text{as.}}{\sim} \mathsf{N}(\mu, \frac{\sigma^2}{n})$

Proof. (i) follows by the straightforward calculation. (ii) follow from the law of large numbers (Proposition 1.4 pro $k = 1$) and (iii) from the central limit theorem (Proposition 1.5 for $k = 1$). □

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

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

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 definition of the multivariate normal distribution (see for instance Chapter B.6 of Bickel and Doksum, 2015) it follows that $\boldsymbol{Z} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2 \mathbb{I}_n)$. Denote $\boldsymbol{c} = (\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})^\mathsf{T} \in \mathbb{R}^n$. 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) $\sqrt{n}\left(\overline{X}_n - p\right) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}(0, p(1-p))$

(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 estimate the true value of parameter p with an arbitrary precision.

The end of self-study for week 1 (29.9.-3.10.).

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

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

□

Exercise. Use Lemma 2.5 for the special case when the matrix $\mathbb{B}$ is an identity matrix.

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

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

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 1.4) 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 1.2(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-Slutsky theorem (Proposition 1.3).

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 1.5) a Cramér-Slutsky theorem (Proposition 1.3)

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

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

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

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

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 semidefinite 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 orthonormal 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 $\mathbb{U}_A$, $\mathbb{U}_B$ are orthonormal matrices and $\mathbb{D}_A$, $\mathbb{D}_B$ are diagonal matrices with non-negative elements on diagonals.

Further from the assumption of the lemma

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

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 idempotent 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-distribution 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 Definition A.1.

and thus

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

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 asymptotic 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 parameter).

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 (limit distribution of 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

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

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

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

By the central limit theorem (Proposition 1.5, for $k=1$) one has that

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

Further as $S_n^2 \xrightarrow[n\to\infty]{\mathsf{P}} \sigma^2$ (Theorem 2.6(i)) and by the continuous mapping theorem (Proposition 1.2(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-Slutsky theorem (Proposition 1.3). □

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

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

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 $\mathsf{N}(0,1)$.

Now we will consider two random samples from the normal distributions.

Definition 2.5 (F-distribution) Let $X \sim \chi^2_n$ and $Y \sim \chi^2_m$ be independent. Then the distributions of the random variables

$$Z = \frac{X/n}{Y/m}$$

is called [Fisher-Snedecor] *F-distribution* with n and m degrees of freedom. This distribution is denoted as $F_{n,m}$.

Theorem 2.11 (Theorem about F statistic) Let $X_1, \ldots, X_n$ be a random sample from the normal distribution $\mathsf{N}(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_m$ be a random sample from the normal distribution $\mathsf{N}(\mu_Y, \sigma_Y^2)$. Let the random vectors $(X_1, \ldots, X_n)^\mathsf{T}$ and $(Y_1, \ldots, Y_m)^\mathsf{T}$ be independent. Denote the sample means as $\overline{X}_n$, $\overline{Y}_m$ and the sample variances as

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \quad \text{a} \quad S_Y^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \overline{Y}_m)^2.$$

Then it holds that

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{n-1,m-1}.$$

* Viz Definition A.2.

Proof. The statistics can be rewritten as

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} = \frac{\frac{(n-1)S_X^2}{\sigma_X^2}/(n-1)}{\frac{(m-1)S_Y^2}{\sigma_Y^2}/(m-1)}.$$

Further $\frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi^2_{n-1}$ and $\frac{(m-1)S_Y^2}{\sigma_Y^2} \sim \chi^2_{m-1}$ (Theorem 2.8(ii)). Moreover these random variables are independent. The statement of the theorem now follows from the definition of F-distribution (Definition 2.5). □

The end of self-study for week 2 (6.10.-10.10.).

2.3. Ordered random sample

Suppose that the random sample $X_1, \ldots, X_n$ is from the one-dimensional distribution with the cumulative distribution function F and density f with respect to the Lebesgue measure. Let $n \geq 2$. Using the fact that $X_1, \ldots, X_n$ are independent and have continuous distributions implies

$$\mathsf{P}\big(X_i = X_j \text{ for some } i, j \in \{1, \ldots, n\}\big) = 0.$$

Definition 2.6 (The ordered random sample and ranks)

(i) By ordering the random variables $X_1, \ldots, X_n$ from the smallest to the largest we get *ordered random sample*

$$X_{(1)} < X_{(2)} < \cdots < X_{(n-1)} < X_{(n)}.$$

With the symbol $X_{(k)}$ we understand the kth smallest value among the observations $X_1, \ldots, X_n$ and we call it the kth *order statistic*.

(ii) By the rank of the random variables X_i in the random sample $X_1, \ldots, X_n$ we understand the number $R_i \in \{1, \ldots, n\}$ such that $X_i = X_{(R_i)}$.

The whole ordered random sample will be denoted as $\boldsymbol{X}_{(\cdot)}$, i.e.

$$\boldsymbol{X}_{(\cdot)} = \big(X_{(1)}, \ldots, X_{(n)}\big)^\mathsf{T}.$$

Analogously denote

$$\boldsymbol{R} = (R_1, \ldots, R_n)^\mathsf{T}.$$

Remark.

1. The original sample values $X_1, \ldots, X_n$ can be reconstructed from $\boldsymbol{X}_{(\cdot)}$ and $\boldsymbol{R}$.
2. The first order statistic is the minimum of the variables in the random sample. Analogously the nth order statistic is the maximum.
3. It holds that $R_i = \sum_{j=1}^n \mathbb{1}\{X_i \geq X_j\} = 1 + \sum_{j=1}^n \mathbb{1}\{X_i > X_j\}$.

4. Order statistics and ranks are random variables and at the same time statistics in the sense of Definition 2.3.

By $\mathcal{P}_n$ denote **set of all permutations** of the sequence of numbers $(1,\dots,n)$. The number of the elements of this set is $n!$.

Theorem 2.12 The joint density of the random vector $\boldsymbol{X}_{(\cdot)} = (X_{(1)},\dots,X_{(n)})^\mathsf{T}$ with respect to the Lebesgue measure is

$$p(y_1,\dots,y_n) = \begin{cases} n!\,f(y_1)f(y_2)\cdots f(y_n), & \text{for } y_1 < \cdots < y_n, \\ 0, & \text{otherwise.} \end{cases}$$

Remark. Random variables $X_{(1)},\dots,X_{(n)}$ are **not** independent. Analogously also random variables $R_1,\dots,R_n$ (i.e. the ranks of $X_1,\dots,X_n$) are **not** independent.

Theorem 2.13 The cumulative distribution function of the kth order statistic is given by

$$\begin{aligned} F_{(k)}(x) = \mathsf{P}\big(X_{(k)} \le x\big) &\overset{(i)}{=} \sum_{j=k}^{n} \binom{n}{j} F^j(x)\big(1-F(x)\big)^{n-j} \\ &= \frac{1}{B(k,n-k+1)} \int_0^{F(x)} t^{k-1}(1-t)^{n-k}\,dt, \end{aligned}$$

where $B(\cdot,\cdot)$ denotes Beta function (see Appendix A.4).

Proof. We will show only the inequality *(i)*. Denote $Z_i = \mathbb{1}\{X_i \le x\}$. Then $Y_n = \sum_{i=1}^n Z_i$ is the number of the random variables that are less or equal to x. Moreover $Y_n \sim \mathsf{Bi}\big(n, F(x)\big)$. Thus

$$\mathsf{P}\big(X_{(k)} \le x\big) = \mathsf{P}\big(Y_n \ge k\big) = \sum_{j=k}^{n} \mathsf{P}\big(Y_n = j\big) = \sum_{j=k}^{n} \binom{n}{j} F^j(x)\big(1-F(x)\big)^{n-j}.$$

□

Consequences.

1. Let X_i follows a uniform distribution on the interval $(0,1)$, then the random variable $X_{(k)}$ follows the distributions with the density $f(x) = \frac{1}{B(k,n-k+1)}\,x^{k-1}(1-x)^{n-k}\mathbb{1}\{x \in (0,1)\}$, i.e. Beta distribution $\mathsf{B}(k, n-k+1)$. From that it follows among others that
$$\mathsf{E}\,X_{(k)} = \frac{k}{n+1}, \quad \mathsf{var}\,\big(X_{(k)}\big) = \frac{k(n-k+1)}{(n+2)(n+1)^2}.$$
2. Let X_i is a continuous random variable with the increasing cumulative distribution function F. Then $F(X_{(k)}) \sim \mathsf{B}(k, n-k+1)$.
On the other hand let $Z \sim \mathsf{B}(k, n-k+1)$. Then
$$\mathsf{P}[X_{(k)} \le x] = \mathsf{P}[F(X_{(k)}) \le F(x)] = \mathsf{P}[Z \le F(x)] = \mathsf{P}[F^{-1}(Z) \le x],$$
i.e. $X_{(k)}$ has the same distribution as $F^{-1}(Z)$.

Theorem 2.14 The density of kth order statistic with respect to Lebesgue measure is

$$f_{(k)}(x) = n\binom{n-1}{k-1} f(x) F^{k-1}(x)[1-F(x)]^{n-k}.$$

Proof. With the help of Theorem 2.13

$$f_{(k)}(x) = F'_{(k)}(x) = \frac{1}{B(k, n-k+1)} f(x) F^{k-1}(x)\big(1-F(x)\big)^{n-k}$$

and the statement of the theorem follows from the fact that

$$\frac{1}{B(k, n-k+1)} \overset{(A.1)}{=} \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} = \frac{n!}{(k-1)!(n-k)!} = \frac{n(n-1)!}{(k-1)!(n-k)!} = n\binom{n-1}{k-1}.$$

□

Theorem 2.15 The random vector $\boldsymbol{R} = (R_1, \ldots, R_n)^\mathsf{T}$ takes values in the set $\mathcal{P}_n$ and each element of this set has the the same probability $1/n!$.

Theorem 2.16 It holds that

(i) $\mathsf{P}\big(R_i = k\big) = \frac{1}{n}$ for each $i, k \in \{1, \ldots, n\}$.
(ii) $\mathsf{P}\big(R_i = k, R_j = m\big) = \frac{1}{n(n-1)}$ for each $i \neq j, k \neq m \in \{1, \ldots, n\}$.
(iii) $\mathsf{E}\, R_i = \frac{n+1}{2}$, $\mathsf{var}\, R_i = \frac{n^2-1}{12}$ for each $i \in \{1, \ldots, n\}$.
(iv) $\mathsf{cov}\,(R_i, R_j) = -\frac{n+1}{12}$ for each $i \neq j \in \{1, \ldots, n\}$.

Proof. Part (i). Without loss of generality one can assume that $i = n$. Further let the set $\mathcal{P}^k_{n-1}$ contains the elements of $\mathcal{P}_n$, which have the number k as the last component. Now

$$\mathsf{P}\big(R_n = k\big) = \sum_{\boldsymbol{r} \in \mathcal{P}^k_{n-1}} \mathsf{P}\big(\boldsymbol{R} = \boldsymbol{r}\big) = (n-1)!\,\frac{1}{n!} = \frac{1}{n},$$

where we make use of Theorem 2.15 and that the set $\mathcal{P}^k_{n-1}$ has $(n-1)!$ elements.

Part (ii). Without loss of generality we can assume that $i = n-1$ and $j = n$. Further let $\mathcal{P}^{k,m}_{n-2}$ contain the elements $\mathcal{P}_n$, that have the number m as the last element and the number k as the second to the last element. Then

$$\mathsf{P}\big(R_{n-1} = k, R_n = m\big) = \sum_{\boldsymbol{r} \in \mathcal{P}^{k,m}_{n-2}} \mathsf{P}\big(\boldsymbol{R} = \boldsymbol{r}\big) = (n-2)!\,\frac{1}{n!} = \frac{1}{n(n-1)},$$

where we make use of Theorem 2.15 and that the set $\mathcal{P}^{k,m}_{n-2}$ have $(n-2)!$ elements.

Part (iii). Wit the help of statement (i):

$$\mathsf{E}\, R_i = \sum_{k=1}^{n} k\,\mathsf{P}\big(R_i = k\big) = \sum_{k=1}^{n} k\,\frac{1}{n} = \frac{1}{n}\,\frac{n(n+1)}{2} = \frac{n+1}{2}.$$

Analogously

$$\begin{aligned}\operatorname{var} R_i &= \mathsf{E}\, R_i^2 - \left(\mathsf{E}\, R_i\right)^2 = \sum_{k=1}^{n} k^2 \frac{1}{n} - \left(\frac{n+1}{2}\right)^2 = \frac{n(n+1)(2n+1)}{6n} - \frac{(n+1)^2}{4} \\ &= \frac{n+1}{12}\left(4n+2-3n-3\right) = \frac{(n+1)(n-1)}{12}.\end{aligned}$$

Part (iv).

$$\begin{aligned}\operatorname{cov}\left(R_i, R_j\right) &= \mathsf{E}\, R_i R_j - \mathsf{E}\, R_i\, \mathsf{E}\, R_j = \sum_{k=1}^{n} \sum_{m=1, m\neq k}^{n} km\, \frac{1}{n(n-1)} - \left(\frac{n+1}{2}\right)^2 \\ &= \frac{1}{n(n-1)}\left[\sum_{k=1}^{n} k \sum_{m=1}^{n} m - \sum_{k=1}^{n} k^2\right] - \left(\frac{n+1}{2}\right)^2 \\ &= \frac{1}{n(n-1)}\left[\left(\frac{n(n+1)}{2}\right)^2 - \frac{n(n+1)(2n+1)}{6}\right] - \left(\frac{n+1}{2}\right)^2 \\ &= \frac{n(n+1)^2}{4(n-1)} - \frac{(n+1)(2n+1)}{6(n-1)} - \frac{(n+1)^2}{4} \\ &= \frac{(n+1)}{12(n-1)}\left[3n(n+1) - 2(2n+1) - 3(n+1)(n-1)\right] \\ &= \frac{(n+1)}{12(n-1)}\,(1-n) = -\frac{(n+1)}{12}.\end{aligned}$$

□

Remark. When the random sample does not come from the continuous distribution or they contain ties because of rounding then it still makes sense to define the ordered random sample as

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n-1)} \leq X_{(n)},$$

where the kth order statistic $X_{(k)}$ is still well defined and the statement of Theorem 2.13 still holds.

But ranks cannot be uniquely defined. In practice we usually use *average ranks*, which can be calculated as

$$\tilde{R}_i = 1 + \sum_{j=1}^{n} \mathbb{1}\{X_i > X_j\} + \frac{1}{2} \sum_{j=1, j\neq i}^{n} \mathbb{1}\{X_i = X_j\}.$$

For ranks defined in this was it holds that $\mathsf{E}\, \tilde{R}_i = \frac{n+1}{2}$ (see Theorem 2.16(iii)). But all the other above statements are not true.

Alternatively one can assign the ranks to the tied observations randomly. For such randomized ranks Theorem 2.15 holds true and thus also Theorem 2.16. The disadvantage of this approach is that this approach introduce additional variability into our inference.

2.4. Transformation in statistics

2.4.1. Transformation of the observations and its impact on the parameters of interest

Let $X_1, \dots, X_n$ be a random sample from the distribution with the cumulative distribution function F_X. The corresponding density denote as f_X and the support as S_X. Consider the strictly monotone * differentiable function $g : S_X \to \mathbb{R}$ and define $Y_i = g(X_i)$. Then $Y_1, \dots, Y_n$ be a random sample from the distribution with the density f_Y that can be calculated with the help of the theorem about the density of transformed random variables.

Transformation of the observations is used in statistics quite often. The usual reason is that the original random sample $X_1, \dots, X_n$ (obviously) does not meet assumptions of the methods that we intend to use (for instance normality, symmetry of the density, ...). Thus we choose an appropriate function g such that $Y_i = g(X_i)$ seems to satisfy assumption of the intended methods and then we work with the random sample $Y_1, \dots, Y_n$ instead of the original random sample $X_1, \dots, X_n$. The most widely used transformations of the positive random variables are $g(x) = \log x$ and $g(x) = \sqrt{x}$.

Example. Let X_i have *logarithmic-normal distribution* $\mathsf{LN}(\mu, \sigma^2)$. Then $\log(X_i)$ follows normal distribution $\mathsf{N}(\mu, \sigma^2)$

When using transformation one has to keep in mind that some **parameters of the distribution** F_X of the original random sample **will be affected with the transformation** in such a way that **we will be not able to identify them**.

For instance the expected values $\mu_X = \mathsf{E}\, X_i$ changes to $\mu_Y = \mathsf{E}\, g(X_i)$. Thus if we do not know the distribution X_i, then it is in general impossible (unless g is linear) to calculate the value μ_X from μ_Y. Let g be a continuous and strictly concave function, then with the help of Jensen inequality it holds that $\mu_Y < g(\mu_X)$. Thus $g^{-1}(\mu_Y) < \mu_X$.

Thus the sample $\overline{Y}_n$ from the transformed converges in probability (see Theorem 2.2(ii)) to μ_Y. Thus $g^{-1}(\overline{Y}_n)$ converges in probability to $g^{-1}(\mu_Y) \neq \mu_X$. In general it is impossible to find a function h such that $h(\overline{Y}_n)$ converges to μ_X. When we are interested in μ_X then we have to work with the original data. Analogously when we are interested for instance in variance.

Example. Let $X_i \sim \mathsf{LN}(\mu, \sigma^2)$. Then for $g(x) = \log x$ it holds that $Y_i = g(X_i) \sim \mathsf{N}(\mu, \sigma^2)$. Thus

$$g^{-1}\left(\overline{Y}_n\right) \xrightarrow[n\to\infty]{\mathsf{P}} \mathrm{e}^{\mathsf{E}\, Y_i} = \mathrm{e}^{\mu} < \mathrm{e}^{\mu+\sigma^2/2} = \mathsf{E}\, X_i.$$

Some parameters do not have these difficulties. For instance median (or any other quantile) can be easily calculated with the help of g^{-1}. Let m_X be median of X_i and m_Y be median of Y_i. Further let g be an increasing function. Then it holds that $m_Y = g(m_X)$ and m_X can be identified as $\mu_X = g^{-1}(m_Y)$.

* We are usually avoiding non-monotone transformations as they imply that some of the information is lost.

Ranks are invariant with respect to increasing transformations. This implies that also values of the statistics calculated only from the ranks are the same for the original as well as transformed random sample.

2.4.2. Asymptotic variance-stabilization transformations

Another motivation for transforming some statistics is to stabilize (asymptotic) variance. Let the sequence of random variables $\{T_n\}$ satisfy

$$\sqrt{n}\,(T_n - \mu) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}\big(0, \sigma^2(\mu)\big).$$

The variance $\sigma^2(\mu)$ of the asymptotic normal distribution is called also the asymptotic variance of $\sqrt{n}(T_n - \mu)$.

As we see later for inference (testing and confidence intervals) about parameter μ it is good that the asymptotic variance of the used random variable does not depend on parameter μ.

Let g be a real function that is defined and differentiable in the neighbourhood of μ. Then with the help of Δ-method (Proposition 1.6) we get that

$$\sqrt{n}\,\big(g(T_n) - g(\mu)\big) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}\big(0, \big[g'(\mu)\big]^2\sigma^2(\mu)\big).$$

Thus with the help of the choice

$$g(x) = c\int \frac{1}{\sigma(x)}\,dx, \tag{2.5}$$

one gets $g'(\mu) = \frac{c}{\sigma(\mu)}$, which implies that

$$\sqrt{n}\,\big(g(T_n) - g(\mu)\big) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}\big(0, c^2\big)$$

and the influence of μ on the asymptotic variance will be eliminated.

Example. Let $X_1, \ldots, X_n$ be a random sample from Poisson distribution $\mathsf{Po}(\lambda)$. Then the statistic $T_n = \overline{X}_n$ with the help of the central limit theorem (Proposition 1.5) satisfies

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

Thus $\sigma(x) = \sqrt{x}$ and one gets $g(x) = \int \frac{1}{\sigma(x)}\,dx = \int x^{-1/2}\,dx = 2\sqrt{x}$. Thus

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

Remark. An analogous idea is sometimes used for individual random variables X_i. Let $\mathsf{E}\,X_i = \lambda$ and $\mathsf{var}\,X_i = \sigma^2(\lambda)$. Then we hope that using the transformation $Y_i = g(X_i)$, where g is calculated with the help of (2.5), we get the observation Y_i that has a distribution that is closer to the normal distribution. For instance when one assumes that $X_i \sim \mathsf{Po}(\lambda)$ than often in analysis one works with $Y_i = \sqrt{X_i}$.

Exercise. Let $\{Y_n\}$ be a sequence of random variables such that $Y_n \sim \mathsf{Po}(n\lambda)$. Show that $\sqrt{Y_n} - \sqrt{n\,\lambda} \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}(0, \frac{1}{4})$.
Hint: Note that Y_n can be represented as $\sum_{i=1}^{n} X_i$, where $X_1, \ldots, X_n$ is a random sample from $\mathsf{Po}(\lambda)$.

2.4.3. Standardization

A special type of transformation is *standardization*. Suppose we have a random sample $X_1, \ldots, X_n$ and we calculate $\overline{X}_n$ and S_n^2. Then define the random variables $Z_1, \ldots, Z_n$ as

$$Z_i = \frac{X_i - \overline{X}_n}{S_n}.$$

These variables has the sample mean 0 and the sample variance 1. But $Z_1, \ldots, Z_n$ do not form a random sample as there are not independent. Nevertheless using the facts that $\overline{X}_n \xrightarrow{\mathsf{P}} \mathsf{E}\,X_i$ and $S_n \xrightarrow{\mathsf{P}} \sqrt{\mathsf{var}\,X_i}$ as $n \to \infty$, then for large sample sizes $Z_1, \ldots, Z_n$ behave almost independent variables with zero expectation and unit variance. Often it can be proved that the dependence induced by the fact that the unknown $\mathsf{E}\,X_i$ and $\sqrt{\mathsf{var}\,X_i}$ are replaced with its sample analogs (i.e. $\overline{X}_n$ and S_n) can be safely ignored.

Standardization is used when we want to get rid off the first two moments as we are interested in other aspects of the distribution F_X (see for instance the sample correlation coefficient in the last chapter).

Sample examples for the preparation for the exam.

1. Let X_1 a X_2 be independent random variables with the uniform distribution on the interval $(0,1)$. Calculate $\mathsf{E}\,\frac{X_1^2}{X_2}$ and $\mathsf{E}\,\frac{X_{(1)}^2}{X_{(2)}}$.

2. Let $X_1,\ldots,X_4$ be independent and identically distributed random variables with the density f with respect to Lebesgue measure. Find the probability $\mathsf{P}(R_1+R_2=5)$.

3. Let $X_1,\ldots,X_n$ be independent and identically distributed random variables with the density f with respect to Lebesgue measure. Calculate $\mathsf{E}\,\frac{R_1}{R_2}$ and then also $\lim_{n\to\infty}\mathsf{E}\,\frac{R_1}{R_2}$.

The end of self-study for week 3 (13.10.-17.10.).

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 the 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

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

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

(ii) Expression

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

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

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

where $\widehat{\mathsf{var}}\left(\widehat{\theta}_n\right)$ is an estimator of $\mathsf{var}\left(\widehat{\theta}_n\right)$
- 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:

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

Proof of the aforementioned assertion is a direct calculation:

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

- 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

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

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* $\mathcal{F} = \left\{\mathsf{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\right\}$. 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 $\mathsf{var}\,(\widehat{\theta}_n) \xrightarrow[n\to\infty]{} 0$, 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:

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

Now, both terms on the right-hand side converge to zero because thanks to the assumptions of the theorem $\mathsf{var}\,(\widehat{\theta}_n) \to 0$ 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 1.4) in combination with the continuous mapping theorem (Theorem 1.2).

Example. Let $X_1, \ldots, X_n$ be a random sample from the alternative distribution $\mathsf{Be}(p)$. 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*:

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

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 1.2).
- If τ^{-1} has a continuous derivative on some neighbourhood of $\mathsf{E}\,X_i$, then thanks to the Δ-method (Theorem 1.6)

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

  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

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

  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 $\mathsf{U}(0, \theta_X)$ 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 $\sqrt{n}\,(\widehat{\theta}_n - \theta_X) \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}(0, \theta_X^2/3)$.

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

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

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 $(\mathsf{E}\, X_i, \mathsf{var}\, X_i)^\top = \boldsymbol{\tau}(\boldsymbol{\theta}_X)$, 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)

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

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 it is know that $\mathsf{E}\, X_i = \frac{p}{a}$ and $\mathsf{var}\, X_i = \frac{p}{a^2}$. 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 $\mathsf{U}(\theta_1, \theta_2)$ 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 $\tau^{-1}(x_1, x_2) = (x_1 - \sqrt{3x_2}, x_1 + \sqrt{3x_2})$ 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.

6. $X_1, \ldots, X_n$ is a random sample from $\mathsf{B}(\alpha, \beta)$ distribution, 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 τ has continuous derivative on some neighbourhood of $(\mathsf{E}\,X_i, \operatorname{var} X_i)$.

3.4. Maximum likelihood estimators

Suppose we have a random sample of random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ being distributed as the generic vector $\boldsymbol{X} = (X_1, \ldots, X_k)\mathsf{tr}$ that has a density $f(\boldsymbol{x}; \boldsymbol{\theta})$ with respect to a σ-finite measure μ and that the density is known up to an unknown p-dimensional parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\mathsf{T} \in \Theta$. Let $\boldsymbol{\theta}_X = (\theta_{X1}, \ldots, \theta_{Xp})^\mathsf{T}$ be the true value of the parameter.

Define the *likelihood function* as

$$L_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{X}_i; \boldsymbol{\theta}).$$

Note that the likelihood function is in fact a joint density viewed as a function of the parameter. This corresponds to the fact that when dealing with real data we are in fact dealing with the realizations of the random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Thus when dealing data the observed values of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are fixed. Nevertheless from the point of the theory the likelihood function can be viewed as a random function as it depends on the random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$.

The maximum likelihood estimator is usually defined as

$$\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta} \in \Theta} L_n(\boldsymbol{\theta}). \tag{3.3}$$

Very often it is much more tractable to maximize the *log-likelihood function* as

$$\ell_n(\boldsymbol{\theta}) = \log L_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\boldsymbol{X}_i; \boldsymbol{\theta}).$$

Further, in regular systems the function $\log f(\boldsymbol{x}; \boldsymbol{\theta})$ is differentiable with respect to $\boldsymbol{\theta}$ and one defines the maximum likelihood estimator as an appropriate root of the maximum likelihood equations given by

$$\frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \stackrel{!}{=} \mathbf{0}_p.$$

Example. Suppose we have a random sample $X_1, \ldots, X_n$ from the Bernoulli distribution with the parameter p. Show the maximum likelihood estimator of p and derive its asymptotic distribution.

Example. Suppose we have a random sample $X_1, \ldots, X_n$ from the exponential distribution with the density $f(x; \lambda) = \lambda \exp\{-\lambda x\} \mathbb{1}\{x > 0\}$. Find the maximum likelihood estimator of λ and derive its asymptotic distribution.

Example. Suppose we have a random sample $X_1, \ldots, X_n$ from the uniform distribution on the interval $(0, \theta)$, i.e. X_i has the density $f(x; \theta) = \frac{1}{\theta} \mathbb{1}\{x \in (0, \theta)\}$. Derive the maximum likelihood estimator of θ and show its consistency.

Example. Suppose we have a random sample $X_1, \ldots, X_n$ from the normal distribution $\mathsf{N}(\mu, \sigma^2)$. Derive the maximum likelihood estimator of $\boldsymbol{\theta} = (\mu, \sigma^2)$.

3.5. 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.

3.5.1. 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 fixed before the analysis; 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 $\left(-\infty, \eta_U(\boldsymbol{X})\right)$ is called *upper one-sided confidence interval*. We have that $\mathsf{P}\left[\theta_X < \eta_U(\boldsymbol{X})\right] = 1 - \alpha$.

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: $F_X \in \mathcal{F} = \left\{\mathsf{N}(\mu, \sigma_X^2), \mu \in \mathbb{R}, \sigma_X^2 \text{ known}\right\}$
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

$$\frac{\sqrt{n}\left(\overline{X}_n - \mu_X\right)}{\sigma_X} \sim \mathsf{N}(0,1).$$

2. We will use the equality

$$\mathsf{P}\left[u_{\frac{\alpha}{2}} < \frac{\sqrt{n}\left(\overline{X}_n - \mu_X\right)}{\sigma_X} < u_{1-\alpha/2}\right] = 1 - \alpha,$$

   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.

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

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

$$\big(-\infty, \eta_U(\boldsymbol{X})\big), \quad \text{where} \quad \eta_U(\boldsymbol{X}) = \overline{X}_n + u_{1-\alpha}\,\frac{\sigma_X}{\sqrt{n}}.$$

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 $2u_{1-\alpha/2}\,\sigma_X/\sqrt{n} \le d$. 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 29 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

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

3.5.2. 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: $F_X \in \mathcal{F} = \{\mathsf{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$
Estimated parameter: $\sigma_X = \sqrt{\operatorname{var} X_i}$
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

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

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

We will use the equality

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

and after several manipulations of the expression we will arrive at

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

We obtained a confidence interval

$$\left(\frac{(n-1)S_n^2}{\chi_{n-1}^2(1-\alpha/2)}, \frac{(n-1)S_n^2}{\chi_{n-1}^2(\alpha/2)}\right) \tag{3.4}$$

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

Table 3.2.: Some values of quantiles $\chi_f^2(\kappa)$ 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

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: $X_1, \ldots, X_n \sim F_X$
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

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

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

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

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: $X_1, \ldots, X_n \sim F_X$
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

$$\left(\overline{X}_n - u_{1-\alpha/2}\,\frac{S_n}{\sqrt{n}},\ \overline{X}_n + u_{1-\alpha/2}\,\frac{S_n}{\sqrt{n}}\right). \tag{3.6}$$

Since for $n \to \infty$ quantile $t_{n-1}(\alpha)$ converges to u_α (for arbitrary $0 < \alpha < 1$), it holds that interval (3.5), 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 (3.5) is longer than interval (3.6). For caution, it is therefore recommended to use interval (3.5).

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: $F_X \in \mathcal{F} = \{\mathrm{Be}(p), p \in (0,1)\}$
Estimated parameter: $p_X = \mathrm{E}\, X_i = \mathrm{P}[X_i = 1]$
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 $\sqrt{n}\,(\widehat{p}_n - p_X) \xrightarrow[n\to\infty]{\mathrm{d}} \mathrm{N}(0, p_X(1-p_X))$. Thus,

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

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

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

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

$$\frac{\sqrt{n}\,(\widehat{p}_n - p_X)}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}} = \frac{\sqrt{n}\,(\widehat{p}_n - p_X)}{\sqrt{p_X(1-p_X)}} \frac{\sqrt{p_X(1-p_X)}}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}} \xrightarrow[n\to\infty]{\mathrm{d}} \mathrm{N}(0,1). \tag{3.7}$$

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

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

we get that

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

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 $n \to \infty$.

The end of self-study for week 4 (27.10.-31.10.).

3.6. 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.

3.6.1. 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 $\widehat{F}_n(x) \overset{\text{df}}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\}$ 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 50).
- 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.

3.6.2. 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 $\widehat{\theta}_n \stackrel{\text{df}}{=} t(\widehat{F}_n)$. 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

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

3.6.3. 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 34 we know that in model $\mathcal{F} = \left\{\mathsf{N}(\mu,\sigma^2), \mu\in\mathbb{R}, \sigma^2 > 0\right\}$ 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:

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

3.6.4. 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

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

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

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

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

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

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 $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. Therefore, the empirical quantile will be (according to our definition) an appropriately chosen order statistic. Since it holds that

  $$\widehat{F}_n(X_{(k)}) \geq \frac{k}{n} \quad \text{and} \quad \widehat{F}_n(X_{(k)} - h) < \frac{k}{n} \text{ for } \forall h > 0,$$

  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 27.
- 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

  $$\widehat{F}_n(\widehat{u}_n(\alpha)) \geq \alpha \quad \text{and} \quad \lim_{h \searrow 0} \widehat{F}_n(\widehat{u}_n(\alpha) - h) < \alpha,$$

  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

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

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

$$\begin{aligned}\mathsf{P}\big(\widehat{u}_n(\alpha) < u_X(\alpha) - \varepsilon\big) &= \mathsf{P}\big(X_{(k_\alpha)} < u_X(\alpha) - \varepsilon\big)\\ &= \mathsf{P}\Big(\textstyle\sum_{i=1}^n \mathbb{1}\{X_i < u_X(\alpha) - \varepsilon\} \geq k_\alpha\Big)\\ &\leq \mathsf{P}\Big(\widehat{F}_n\big(u_X(\alpha) - \varepsilon\big) - F_X\big(u_X(\alpha) - \varepsilon\big) \geq \tfrac{k_\alpha}{n} - F_X\big(u_X(\alpha) - \varepsilon\big)\Big). \qquad (3.8)\end{aligned}$$

From Theorem 3.3 it follows that

$$\widehat{F}_n\big(u_X(\alpha) - \varepsilon\big) - F_X\big(u_X(\alpha) - \varepsilon\big) \xrightarrow[n\to\infty]{\mathsf{P}} 0, \qquad (3.9)$$

and from the assumptions of this theorem we have that

$$\tfrac{k_\alpha}{n} - F_X\big(u_X(\alpha) - \varepsilon\big) \xrightarrow[n\to\infty]{} \alpha - F_X\big(u_X(\alpha) - \varepsilon\big) > 0. \qquad (3.10)$$

By combining (3.9) and (3.10) we obtain that the right-hand side of equality (3.8) 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

$$\begin{aligned}\mathsf{P}\big(\widehat{u}_n(\alpha) > u_X(\alpha) + \varepsilon\big) &= \mathsf{P}\Big(\textstyle\sum_{i=1}^n \mathbb{1}\{X_i \leq u_X(\alpha) + \varepsilon\} < k_\alpha\Big)\\ &\leq \mathsf{P}\Big(\widehat{F}_n\big(u_X(\alpha) + \varepsilon\big) - F_X\big(u_X(\alpha) + \varepsilon\big) < \tfrac{k_\alpha}{n} - F_X\big(u_X(\alpha) + \varepsilon\big)\Big). \qquad (3.11)\end{aligned}$$

From Theorem 3.3 it follows that

$$\widehat{F}_n\big(u_X(\alpha) + \varepsilon\big) - F_X\big(u_X(\alpha) + \varepsilon\big) \xrightarrow[n\to\infty]{\mathsf{P}} 0, \qquad (3.12)$$

and from the assumptions of this theorem we have that

$$\tfrac{k_\alpha}{n} - F_X\big(u_X(\alpha) + \varepsilon\big) \xrightarrow[n\to\infty]{} \alpha - F_X\big(u_X(\alpha) + \varepsilon\big) < 0. \qquad (3.13)$$

By combining (3.12) and (3.13) we obtain that the right-hand side of equality (3.11) 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

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

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

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

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

From the central limit theorem for triangular arrays 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

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

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 $1-\beta$.

3.6.5. 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 covariance matrix* is used. It is defined as a multidimensional analogy of the sample variance S_n^2:

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

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 $\mathrm{cov}\,(X_{ij}, X_{im})$ 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

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

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 $\mathsf{E}\,|X_{ij}| < \infty$ 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 $\mathrm{var}\,(X_{ij}) < \infty$ 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 component-wise.

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}:

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

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 1.2). But it is not unbiased.

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

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 50 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 $X_{(k_n)} \xrightarrow[n\to\infty]{\mathsf{P}} 0$.

The end of self-study for week 5 (3.11.-7.11.).

4. Testing of statistical hypothesis

4.1. Basic notions and definitions

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample of independent k-dimensional random vectors with distribution $F_X \in \mathcal{F}$, where $\mathcal{F}$ is our model. Let $\boldsymbol{\theta} = \boldsymbol{t}(F) \in \mathbb{R}^d$ be the characteristic of the distribution (so called parameter), which is of our interest and denote by $\Theta = \{\boldsymbol{t}(F), F \in \mathcal{F}\} \subseteq \mathbb{R}^d$ the set of all possible values of this parameter in model $\mathcal{F}$ (so called *parameter space*). Denote the true value of our parameter of interest $\boldsymbol{\theta}_X = \boldsymbol{t}(F_X)$ and let $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ denote all of the observed data.

Examples. All of the new theory of this chapter will be explained on the following examples.

A. Let $X_1, \ldots, X_n$ be a random sample from the distribution $\mathsf{N}(\theta_X, \sigma_0^2)$, where $\sigma_0^2 > 0$ is known. Our model is
$$\mathcal{F}^A = \{\mathsf{N}(\theta, \sigma_0^2),\ \theta \in \mathbb{R}\}.$$

B. Let $X_1, \ldots, X_n$ be a random sample from the distribution $\mathsf{N}(\theta_X, \sigma_X^2)$, where σ_X^2 is unknown. We work with model
$$\mathcal{F}^B = \{\mathsf{N}(\theta, \sigma^2),\ \theta \in \mathbb{R},\ \sigma^2 > 0\} \supset \mathcal{F}^A.$$

C. Let $X_1, \ldots, X_n$ be a random sample from the distribution F_X with finite positive variance. Then we work with non-parametric model
$$\mathcal{F}^C = \mathcal{L}_+^2 \supset \mathcal{F}^B \supset \mathcal{F}^A.$$

The tested parameter will be the expected value $\theta = \int x\, dF(x)$, whose true value is $\theta_X = \mathsf{E}\, X_i$, the dimension d of our parameter θ is 1. Parameter space is $\Theta = \mathbb{R}$.

Choose two non-empty disjoint subsets of Θ and denote them Θ_0 a Θ_1. Assume that we are not interested in the exact value of $\boldsymbol{\theta}_X$, but we want to answer the question whether $\boldsymbol{\theta}_X \in \Theta_0$ or $\boldsymbol{\theta}_X \in \Theta_1$.

Definition 4.1 (Null hypothesis and alternative hypothesis)
- The set Θ_0 is called the *null hypothesis* and the set Θ_1 is called the *alternative hypothesis*.

- Denote
$$\mathcal{F}_0 \stackrel{\text{df}}{=} \{F \in \mathcal{F} : \boldsymbol{t}(F) \in \Theta_0\},$$
i.e. all distributions from model $\mathcal{F}$ whose parameter satisfies the null hypothesis. If $\mathcal{F}_0 = \{F_0\}$ (i.e. there is exactly one distribution from our model that satisfies the null hypothesis), the null hypothesis is called *simple null hypothesis*, otherwise we call it *composite null hypothesis*.
- Denote
$$\mathcal{F}_1 \stackrel{\text{df}}{=} \{F \in \mathcal{F} : \boldsymbol{t}(F) \in \Theta_1\},$$
i.e. all distributions from model $\mathcal{F}$ whose parameter satisfies the alternative hypothesis. If $\mathcal{F}_1 = \{F_1\}$ (i.e. there is exactly one distribution from our model that satisfies the alternative hypothesis), the alternative is called *simple alternative*, otherwise we call it *composite alternative*.

Remark.
- Null hypothesis is usually denoted by H_0, alternative by H_1. We speak about *testing* the null hypothesis
$$H_0 : \boldsymbol{\theta}_X \in \Theta_0 \quad \text{against the alternative} \quad H_1 : \boldsymbol{\theta}_X \in \Theta_1.$$
- We are in the situation of simple hypothesis, if $\Theta_0 = \{\boldsymbol{\theta}_0\}$, i.e. it contains only one point, and there exists exactly one distribution $F_0 \in \mathcal{F}$ such that $\boldsymbol{t}(F_0) = \boldsymbol{\theta}_0$.
- Simple alternative occurs, if $\Theta_1 = \{\boldsymbol{\theta}_1\}$, i.e. it contains only one point, and there exists exactly one distribution $F_1 \in \mathcal{F}$ such that $\boldsymbol{t}(F_1) = \boldsymbol{\theta}_1$.

Usually, we take $\Theta_1 = \Theta_0^c$ and $\mathcal{F}_1 = \mathcal{F}_0^c$. If this was not the case, i.e. $\Theta_0 \cup \Theta_1 \subsetneq \Theta$, our model can be narrowed to $\mathcal{F}^0 = \{F \in \mathcal{F} : \boldsymbol{t}(F) \in \Theta_0 \cup \Theta_1\}$. Therefore we can assume without loss of generality that $\Theta_1 = \Theta_0^c$ a $\mathcal{F}_1 = \mathcal{F}_0^c$.

Choice of hypothesis for one-dimensional parameter θ
- Most common choice of null hypothesis is $\Theta_0 = \{\theta_0\}$ for some chosen $\theta_0 \in \mathbb{R}$, i.e. we test the null hypothesis $H_0 : \theta_X = \theta_0$. We take $\Theta_1 = \Theta_0^c$ as the alternative, i.e. $H_1 : \theta_X \neq \theta_0$. This procedure is called *two-sided test* or *test against two-sided alternative*.
- Other possibility is to take either $\Theta_0 = (-\infty, \theta_0]$, i.e. test $H_0 : \theta_X \leq \theta_0$ against $H_1 : \theta_X > \theta_0$, or $\Theta_0 = [\theta_0, \infty)$, i.e. test $H_0 : \theta_X \geq \theta_0$ against $H_1 : \theta_X < \theta_0$. These tests are called *one-sided tests* or *tests against one-sided alternative*. Notice that **the extreme value of θ_0 is included in the null hypothesis**.

The choice of the hypothesis is given by the practical problem that we are trying to solve. In some cases, the choice can be different from the three possibilities mentioned above. However, in this lecture, we will only deal with one-sided and two-sided tests.

Examples. Consider two-sided test of parameter $\theta = t(F) = \int x\, dF(x) \in \mathbb{R}$. We test the null hypothesis $H_0 : \theta_X = \theta_0$ against the alternative $H_1 : \theta_X \neq \theta_0$.

A. Take model $\mathcal{F}^A = \{\mathsf{N}(\theta, \sigma_0^2),\ \theta \in \mathbb{R}\}$. In this case, we have $\mathcal{F}_0 = \{\mathsf{N}(\theta_0, \sigma_0^2)\}$, so we are in the situation of simple null hypothesis. Alternative is composite, $\mathcal{F}_1 = \{\mathsf{N}(\theta, \sigma_0^2),\ \theta \in \mathbb{R} \setminus \{\theta_0\}\}$.

B. In model $\mathcal{F}^B = \{\mathsf{N}(\theta, \sigma^2),\ \theta \in \mathbb{R},\ \sigma^2 > 0\}$ we have a composite null hypothesis, $\mathcal{F}_0 = \{\mathsf{N}(\theta_0, \sigma^2),\ \sigma^2 > 0\}$, and the alternative hypothesis is also composite, $\mathcal{F}_1 = \{\mathsf{N}(\theta, \sigma^2),\ \theta \in \mathbb{R} \setminus \{\theta_0\},\ \sigma^2 > 0\}$.

C. For model $\mathcal{F}^C = \mathcal{L}_+^2$, the hypothesis is composite, $\mathcal{F}_0 = \{F \in \mathcal{L}_+^2 : t(F) = \theta_0\}$, and so is the alternative, $\mathcal{F}_1 = \{F \in \mathcal{L}_+^2 : t(F) \neq \theta_0\}$.

We would like to decide, based on random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, whether H_0 holds or not. To do that, we take appropriately chosen function of our data $S_n(\boldsymbol{X})$, which is called *test statistic*, and appropriately chosen set C, called *critical region*. Test statistic is usually one-dimensional; critical region is then some subset of $\mathbb{R}$. Our decision is then based on whether test statistic lies in critical region or not.

- If $S_n(\boldsymbol{X}) \in C$, then the conclusion is that we *reject* null hypothesis H_0 and accept alternative H_1.
- If $S_n(\boldsymbol{X}) \notin C$, then the conclusion is that we *cannot reject* the null hypothesis H_0 and accept alternative H_1.

Remark. Some authors define critical region as a subset of the sample space, i.e. in our notation $S_n^{-1}(C)$. They reject the hypothesis H_0 if $\boldsymbol{X} \in S_n^{-1}(C)$.

Definition 4.2 (Test) *Statistical test* is defined by test statistic $S_n(\boldsymbol{X})$, critical region C and rule for rejecting hypothesis defined above. Two tests $(S_n(\boldsymbol{X}), C)$ and $(S_n^*(\boldsymbol{X}), C^*)$ are called *equivalent* if and only if $S_n(\boldsymbol{X}) \in C \Leftrightarrow S_n^*(\boldsymbol{X}) \in C^*$ almost surely, i.e. both tests give us the same result with probability 1.

4.2. Significance level and power of a test

There are four possible scenarios that can occur while testing hypothesis, depending on whether the null hypothesis holds or not and whether the test rejects the null hypothesis or not.

- **The null hypothesis holds, test does not reject it**, i.e. $\theta_X \in \Theta_0$ and $S_n(\boldsymbol{X}) \notin C$. In this case, the test made the right decision.
- **The null hypothesis holds, test rejects it**, i.e. $\theta_X \in \Theta_0$ and $S_n(\boldsymbol{X}) \in C$. In this case, the test made the wrong decision.
- **The null hypothesis does not hold, test does not reject it**, i.e. $\theta_X \notin \Theta_0$ and $S_n(\boldsymbol{X}) \notin C$. In this case, the test made the wrong decision.
- **The null hypothesis does not hold, test rejects it**, i.e. $\theta_X \notin \Theta_0$ and $S_n(\boldsymbol{X}) \in C$. In this case, the test made the right decision.

Definition 4.3 (Type I and II error)
(i) If test rejects true hypothesis, we call it *type I error*.
(ii) If the test does not reject hypothesis that does not hold, we call it *type II error*.

The four possible scenarios are presented in table 4.1.

Table 4.1.: Possible scenarios for testing hypothesis.

	H_0 **is not rejected**	H_0 **is rejected**
H_0 **holds**	OK	*type I error*
H_0 **does not hold**	*type II error*	OK

It is not possible to avoid type I and II errors. The standard statistical approach to testing hypothesis is to **control the probability of type I error**.

Regarding type II error, the ideal approach would be to choose such test that minimizes the probability of type II error. However, since the probability of type II error depends on the choice of alternative, we can only find these ideal tests in cases, where the alternative is not too big.

4.2.1. Significance level

Take $F \in \mathcal{F}$ and denote

$$\mathsf{P}_F\big[S_n(\boldsymbol{X}) \in B\big] = \int \mathbb{1}\big\{S_n(\boldsymbol{x}) \in B\big\}\, dF(\boldsymbol{x}_1) \cdots dF(\boldsymbol{x}_n).$$

If there exists a unique relation between the parameter $\boldsymbol{\theta} \in \Theta$ and the distribution $F \in \mathcal{F}$, then we can write

$$\mathsf{P}_{\boldsymbol{\theta}}[S_n(\boldsymbol{X}) \in B] = \int \mathbb{1}\big\{S_n(\boldsymbol{x}) \in B\big\}\, dF(\boldsymbol{x}_1) \cdots dF(\boldsymbol{x}_n), \tag{4.1}$$

where F is the distribution satisfying $\boldsymbol{t}(F) = \boldsymbol{\theta}$.

Notice that we can also work with (4.1) if the distribution of the random variable $S_n(\boldsymbol{X})$ is the same for all F such that $\boldsymbol{t}(F) = \boldsymbol{\theta}$.

Definition 4.4 (Significance level) Fix $\alpha \in (0,1)$.
(i) If the critical region C satisfies condition

$$\sup_{F \in \mathcal{F}_0} \mathsf{P}_F[S_n(\boldsymbol{X}) \in C] = \alpha,$$

we say that test $\big(S_n(\boldsymbol{X}), C\big)$ has *significance level* equal to α.

(ii) If the critical region C satisfies condition

$$\sup_{F\in\mathcal{F}_0} \lim_{n\to\infty} \mathsf{P}_F[S_n(\boldsymbol{X}) \in C] = \alpha,$$

we say that test $(S_n(\boldsymbol{X}), C)$ has significance level α *asymptotically*.

Remark.

- If the set $\mathcal{F}_0 = \{F_0\}$ has only one element, then the significance level can be written as
$$\alpha = \mathsf{P}_{\theta_0}[S_n(\boldsymbol{X}) \in C], \text{ where } \boldsymbol{\theta}_0 = \boldsymbol{t}(F_0).$$
- Roughly speaking, significance level is the probability of type I error, i.e. probability of rejecting true hypothesis. If the hypothesis contains more than one value of parameter, it is the biggest possible probability of type I error.
- Test that reaches the significance level α exactly is called *exact test*. Test that reaches the significance level α only asymptotically will be called *asymptotic test*.

Standard approach to testing hypothesis can be summarized in the following steps.

1. At first, we specify the required significance level α, which should be reached exactly or asymptotically by the test.
2. We choose appropriate test statistic $S_n(\boldsymbol{X})$.
3. We choose the critical region $C = C(\alpha)$ according to α, such that the significance level (exact or asymptotic) will be α and the probability of type II error will be the smallest possible.

Remark.

- Significance level is chosen to be small, generally we choose $\alpha = 0.05$.
- If the test statistic $S_n(\boldsymbol{X})$ has *discrete distribution*, it is not possible to reach any significance level α. If the required level α is not reachable, we choose such level $\alpha' < \alpha$, which is the closest to the originally required level α. This guarantees that the probability of rejecting true hypothesis cannot be larger then the chosen tolerance α.

Terminology.

- Test, whose real significance level is smaller then required α, is called *conservative test*. Test, whose real significance level is larger then required α, is called *liberal*.

4.2.2. Power of a test

Definition 4.5 (Power function and power of a test) Function

$$\beta_n(F) = \mathsf{P}_F[S_n(\boldsymbol{X}) \in C]$$

which maps $\mathcal{F}$ into $[0,1]$ is called *the power function* of a test.

For $F \in \mathcal{F}_1$ the value $\beta_n(F)$ is called *the power* of a test against alternative F.

Remark.

- The power of a test is *the probability, that we reject the null hypothesis, which does not hold* in the case of given alternative F. The power of a test depends on the alternative and it is equal to the complement of probability of type II error to 1. There is no non-trivial lower boundary for the power of a test; we cannot assume, that the probability of type II error is small.
- If the test has exact (resp. asymptotic) significance level α, then the following must hold
$$\sup_{F\in\mathcal{F}_0} \beta_n(F) = \alpha, \quad \text{resp.} \quad \sup_{F\in\mathcal{F}_0} \lim_{n\to\infty} \beta_n(F) = \alpha.$$
- If there exists a unique relation between $\boldsymbol{\theta} \in \Theta$ and $F \in \mathcal{F}$ then the power function is usually defined as a mapping of the parameter space Θ into $[0,1]$ given by the formula
$$\beta_n(\boldsymbol{\theta}) = \mathsf{P}_{\boldsymbol{\theta}}[S_n(\boldsymbol{X}) \in C]\,.$$

Remark (Interpretation of results of the test)**.**

- If we *reject the null hypothesis* H_0, it means that the distribution of our data is not consistent with the distribution it should have under the null hypothesis. The probability that we wrongly reject true hypothesis is bounded from above by level α, which is chosen to be small. The null hypothesis H_0 is rejected, we have proven that the alternative H_1 holds.
- If the result of the test is that *we cannot reject the null hypothesis* H_0, it means that the distribution of our data is not different enough from the distribution, which our data should have under the null hypothesis. We cannot conclude that the null hypothesis H_0 holds and the alternative does not, since the probability of a wrong decision in the case, that the hypothesis does not hold, can be considerably large. So, this result does not confirm that the hypothesis holds.
- The null hypothesis H_0 and alternative H_1 are not in symmetric positions in testing. The null hypothesis can be rejected, but it cannot be confirmed or proven.

To be able to choose a critical region $C(\alpha)$ which keeps the required significance level α, we must be able to determine the exact or asymptotic distribution of our test statistic under the null hypothesis and this distribution cannot depend on any unknown characteristics of distribution F_X.

The test statistic $S_n(\boldsymbol{X})$ is chosen so that

(i) its distribution is *sensitive* to the real value of tested parameter $\boldsymbol{\theta}_X$;

(ii) its distribution under the null hypothesis* is known (at least asymptotically) and it *does not depend on unknown parameters.*

After choosing the test statistic, **the critical region** $C(\alpha)$ is chosen so that

(i) the required significance level α *is kept*;

(ii) all values of test statistic which are *less probable* under the null hypothesis than under the alternative are included in the critical region.

* In the case of one-sided tests we should say if the true value of tested parameter is at the boundary of null hypothesis and alternative.

Example (A1)**.** TWO-SIDED TEST OF THE EXPECTED VALUE OF GAUSSIAN DISTRIBUTION WITH KNOWN VARIANCE.

Let us have random sample $X_1, \dots, X_n$ from distribution $F_X = \mathsf{N}(\theta_X, \sigma_0^2) \in \mathcal{F}^A = \left\{\mathsf{N}(\theta, \sigma_0^2),\ \theta \in \mathbb{R}\right\}$. We test

$$H_0 : \theta_X = \theta_0 \qquad \text{against} \qquad H_1 : \theta_X \neq \theta_0.$$

Our test statistic will be based on estimator of our parameter of interest θ_X, i.e. the sample mean. We know that

$$U_n = \frac{\sqrt{n}\,(\overline{X}_n - \theta_0)}{\sigma_0}$$

has under the null hypothesis H_0 the distribution $\mathsf{N}(0,1)$. If the hypothesis does not hold, i.e. $\theta_X - \theta_0 = \delta \neq 0$, then

$$U_n = \frac{\sqrt{n}\,(\overline{X}_n - \theta_X + \theta_X - \theta_0)}{\sigma_0} = \frac{\sqrt{n}\,(\overline{X}_n - \theta_X)}{\sigma_0} + \frac{\sqrt{n}\,\delta}{\sigma_0}$$

has the distribution $\mathsf{N}(\nu_n, 1)$, where $\nu_n = \frac{\sqrt{n}\,\delta}{\sigma_0}$. If the null hypothesis does not hold, then the distribution of our test statistic moves further away from zero, and this distance is larger with larger n and $|\theta_X - \theta_0|$. So, values of our test statistic far away from zero will lead to rejecting the null hypothesis.

The critical region $C(\alpha)$ is chosen as

$$\big(-\infty, c_L(\alpha)\big] \cup \big[c_U(\alpha), \infty\big).$$

Critical values $c_L(\alpha)$ and $c_U(\alpha)$ are chosen so that

$$\mathsf{P}_{\theta_0}\big[U_n \in \big(-\infty, c_L(\alpha)\big]\big] = \mathsf{P}_{\theta_0}\big[U_n \in \big[c_U(\alpha), \infty\big)\big] = \frac{\alpha}{2}.$$

This ensures that the significance level is exactly equal to α. Thanks to the symmetry of the density of Gaussian distribution we have $c_U(\alpha) = -c_L(\alpha) = u_{1-\alpha/2}$. The test works in the following way

$$\text{reject } H_0 : \theta_X = \theta_0 \quad \Longleftrightarrow \quad |U_n| = \left|\frac{\sqrt{n}\,(\overline{X}_n - \theta_0)}{\sigma_0}\right| \geq u_{1-\alpha/2},$$

i.e. reject the null hypothesis if $\overline{X}_n$ differs from hypothetical value θ_0 by more than $\frac{u_{1-\alpha/2}\,\sigma_0}{\sqrt{n}}$.

We put 1.96 as $u_{1-\alpha/2}$ for $\alpha = 0.05$ and 1.645 for $\alpha = 0.1$. The critical region and the densities of test statistic under the null hypothesis and alternative can be seen in figure 4.1.

Let us compute the *power function* of this test. Take some θ such that $\theta - \theta_0 = \delta \neq 0$. If θ is the true value of our parameter, then the distribution of U_n is $\mathsf{N}(\nu_n, 1)$ and the distribution of $U_n - \nu_n$ is $\mathsf{N}(0,1)$. We get

$$\begin{aligned}\beta_n(\theta) &= \mathsf{P}_\theta[U_n \in C(\alpha)] = \mathsf{P}_\theta\big[U_n \leq -u_{1-\alpha/2}\big] + \mathsf{P}_\theta\big[U_n \geq u_{1-\alpha/2}\big] = \\ &= \mathsf{P}_\theta\big[U_n - \nu_n \leq -u_{1-\alpha/2} - \nu_n\big] + \mathsf{P}_\theta\big[U_n - \nu_n \geq u_{1-\alpha/2} - \nu_n\big] = \\ &= \Phi\big(-u_{1-\alpha/2} - \nu_n\big) + 1 - \Phi\big(u_{1-\alpha/2} - \nu_n\big).\end{aligned}$$

Since $\Phi(-x) = 1 - \Phi(x)$, we can rewrite this and get

$$\beta_n(\theta) = \Phi\big(-u_{1-\alpha/2} - |\nu_n|\big) + 1 - \Phi\big(u_{1-\alpha/2} - |\nu_n|\big). \tag{4.2}$$

For $\theta = \theta_0$ we get that $\nu_n = 0$, so $\beta_n(\theta_0) = \alpha$. The power function of this test can be seen in figure 4.2.

Let δ be non-zero. Then $|\nu_n|$ goes to infinity with increasing n and it turns out that from certain n the value $\Phi(-u_{1-\alpha/2} - |\nu_n|)$ is negligible compared to $\Phi(u_{1-\alpha/2} - |\nu_n|)$. The power function can be approximated by $1 - \Phi\big(u_{1-\alpha/2} - \frac{\sqrt{n}\,|\delta|}{\sigma_0}\big)$, and it holds that

$$\beta_n(\theta) \geq 1 - \Phi\Big(u_{1-\alpha/2} - \tfrac{\sqrt{n}\,|\delta|}{\sigma_0}\Big). \tag{4.3}$$

By solving the equation

$$1 - \Phi\Big(u_{1-\alpha/2} - \tfrac{\sqrt{n}\,|\delta|}{\sigma_0}\Big) \stackrel{!}{=} \beta,$$

Figure 4.1.: Density of the test statistic U_n under the null hypothesis and alternative for $\nu_n = 1$ and $\alpha = 0.1$. Critical values are blue, critical region is red.

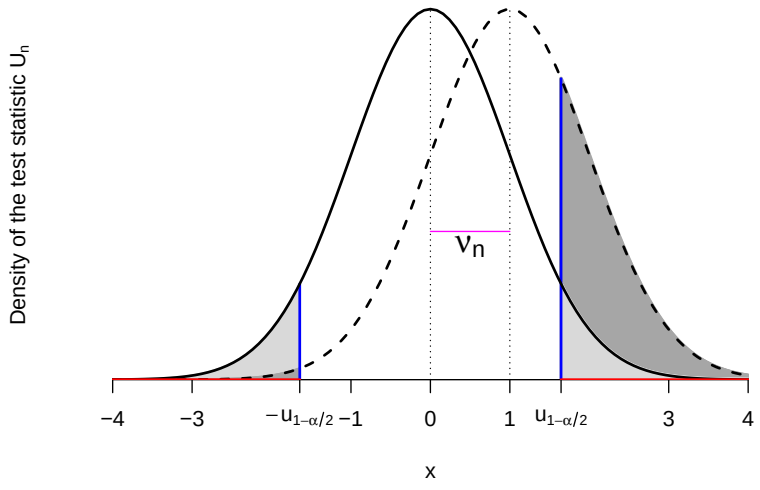
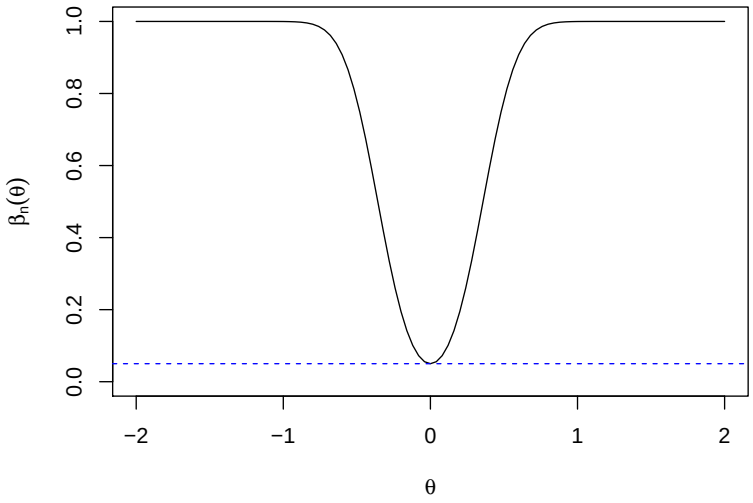


Figure 4.2.: The power function for two-sided test of expected value of Gaussian distribution with known variance for $\theta_0 = 0$, $\sigma_0^2 = 1$, $n = 30$ and $\alpha = 0.05$.



we can compute, how many observations are needed for the test to have power at least β (for example 0.8). The required sample size is

$$n \geq (u_{1-\alpha/2} - u_{1-\beta})^2 \frac{\sigma_0^2}{\delta^2} = (u_{1-\alpha/2} + u_{\beta})^2 \frac{\sigma_0^2}{\delta^2}. \tag{4.4}$$

Remark. As we have seen in the previous example, the power of the test depends on

- significance level α;
- alternative θ, respectively her distance δ from the null hypothesis θ_0;
- variance of the observations σ_0^2;
- sample size n.

Out of all of these factors, we can only influence the sample size. If we want our test to have sufficient power, we need to have at least the number of observations computed in (4.4).

Remark. Notice that the power of the previous test converges to 1 as $n \to \infty$ regardless of the alternative (see (4.3)). This property is called *consistency of the test*. Consistency is very desirable property, otherwise we might not be able to reach required power, even with large sample size.

Definition 4.6 Test $(S_n(\boldsymbol{X}), C)$ with level α is called *consistent test*, if $\forall F \in \mathcal{F}_1$ we have that $\lim_{n\to\infty} \beta_n(F) = 1$.

We will define one more useful property of statistical test: *unbiasedness*.

Definition 4.7 Test $(S_n(\boldsymbol{X}), C)$ with level α is called *unbiased test*, if $\forall F \in \mathcal{F}_1$ we have that $\beta_n(F) \geq \alpha$.

Remark.

- Beware: the notion of unbiasedness and consistency of test have only vague (if any) relation to notion of unbiasedness and consistency of estimate.
- Unbiasedness of a test requires that the power against every alternative is at least α. If it was not the case, i.e. $\exists F \in \mathcal{F}_1$ such that $\beta_n(F) < \alpha$, the test would take this F as a part of the null hypothesis.
- Test that always rejects H_0 with probability α (for whatever data) is unbiased. Especially, there exists unbiased test.
- Sometimes the notion of unbiasedness and consistency is defined with respect to specific alternative. So, for example, we would say that the test is *consistent against alternative* $F \in \mathcal{F}_1$, if we have that $\lim_{n\to\infty} \beta_n(F) = 1$.

4.2.3. Choice of critical region

The critical region $C(\alpha)$ is usually taken in one of the following forms:

- $\big[c_U(\alpha), \infty\big)$, i.e. we reject for *large* values of the test statistic $S_n(\boldsymbol{X})$;
- $\big(-\infty, c_L(\alpha)\big]$, i.e. we reject for *small* values of the test statistic $S_n(\boldsymbol{X})$;
- $\big(-\infty, c_L(\alpha)\big] \cup \big[c_U(\alpha), \infty\big)$, i.e. we reject for *too small* and for *too large* values of the test statistic $S_n(\boldsymbol{X})$;
- $\big(-\infty, -c_U(\alpha)\big] \cup \big[c_U(\alpha), \infty\big)$, i.e. we reject for *large* values of $|S_n(\boldsymbol{X})|$.

The constants $c_L(\alpha)$ a $c_U(\alpha)$, which determine the boundary of our critical region, are called *critical values*. These values are chosen so that the test has the prescribed significance level. As we will see in the following examples, critical values can be expressed using quantiles of appropriately chosen distribution function G_0.

Critical region in the form of $C(\alpha) = \big[c_U(\alpha), \infty\big)$

At first, consider for simplicity *exact test*. Then the critical value $c_U(\alpha)$ is chosen so that

$$\sup_{F\in\mathcal{F}_0} \mathsf{P}_F[S_n(\boldsymbol{X}) \geq c_U(\alpha)] = \alpha.$$

We will only work with examples where we can easily find $F_0 \in \mathcal{F}_0$ such that

$$\sup_{F\in\mathcal{F}_0} \mathsf{P}_F[S_n(\boldsymbol{X}) \geq c] = \mathsf{P}_{F_0}[S_n(\boldsymbol{X}) \geq c] \quad \forall c \in \mathbb{R}. \tag{4.5}$$

When we look for the distribution F_0, we usually look for a distribution which satisfies the null hypothesis (i.e. it lies in $\mathcal{F}_0$), but it is the closest to the alternative (i.e. it is the closest to the set $\mathcal{F}_1$, see **example** (A2) below). Let G_0 denote the cumulative

distribution function of $S_n(\boldsymbol{X})$, if the distribution of $\boldsymbol{X}_i$ is F_0. Then in the case of continuous distribution function we get

$$c_U(\alpha) = c_U(\alpha) = G_0^{-1}(1-\alpha). \tag{4.6}$$

More generally if G_0 is not continuous then one can use the open critical region

$$C(\alpha) = \big(G_0^{-1}(1-\alpha), \infty\big).$$

Note that the above critical region works also for G_0 continuous. That is why in this chapter and in Chapter 4.3 (about p-values) we will use open critical regions as they are easier to express. Nevertheless in Chapter 4.4 we will use closed critical regions so that it matches with open confidence intervals.

In the case of the *asymptotic test* we can use as G_0 the distribution function of the asymptotic distribution of our test statistic under the null hypothesis. More precisely, G_0 is a function that satisfies

$$\sup_{F\in\mathcal{F}_0} \lim_{n\to\infty} \mathsf{P}_F[S_n(\boldsymbol{X}) \geq c] = 1 - G_0(c-).$$

Since for us the function G_0 will always be continuous, the right-hand side of the last equation will be $1 - G_0(c)$.

Critical region in the form of $C(\alpha) = \big(-\infty, c_L(\alpha)\big]$

Similarly as above let $F_0 \in \mathcal{F}_0$ be a distribution of $\boldsymbol{X}_i$ satisfying

$$\sup_{F\in\mathcal{F}_0} \mathsf{P}_F[S_n(\boldsymbol{X}) \leq c] = \mathsf{P}_{F_0}[S_n(\boldsymbol{X}) \leq c] \quad \forall c \in \mathbb{R}.$$

Let G_0 denote the distribution function of $S_n(\boldsymbol{X})$, if the distribution of $\boldsymbol{X}_i$ is F_0. Then $c_L(\alpha)$ is chosen as

$$c_L(\alpha) = G_0^{-1}(\alpha). \tag{4.7}$$

Again, if G_0 is not continuous then the open critical region

$$C(\alpha) = (-\infty, G_0^{-1}(\alpha))$$

will do the job.

Critical region in the form of $C(\alpha) = (-\infty, c_L(\alpha)\big] \cup \big[c_U(\alpha), \infty)$

In this case it is common to choose the critical values $c_L(\alpha)$ and $c_U(\alpha)$ so that

$$\sup_{F\in\mathcal{F}_0} \mathsf{P}_F[S_n(\boldsymbol{X}) \leq c_L(\alpha)] = \sup_{F\in\mathcal{F}_0} \mathsf{P}_F[S_n(\boldsymbol{X}) \geq c_U(\alpha)] = \frac{\alpha}{2}, \tag{4.8}$$

resp.

$$\sup_{F\in\mathcal{F}_0} \lim_{n\to\infty} \mathsf{P}_F[S_n(\boldsymbol{X}) \leq c_L(\alpha)] = \sup_{F\in\mathcal{F}_0} \lim_{n\to\infty} \mathsf{P}_F[S_n(\boldsymbol{X}) \geq c_U(\alpha)] = \frac{\alpha}{2}. \tag{4.9}$$

Furthermore in the situations we will be dealing with, the distribution (exact or asymptotic) of the test statistic $S_n(\boldsymbol{X})$ under the null hypothesis will be always the same, for any true distribution F from $\mathcal{F}_0$ (see also **examples** (B) a (C) on page 74 and 76). Denote this distribution by G_0. This means that we can omit the supremum in the equations (4.8) and (4.9) and the condition is simplified to

$$G_0\big(c_L(\alpha)\big) = 1 - G_0\big(c_U(\alpha) - \big) = \frac{\alpha}{2}.$$

Critical values are equal to

$$c_L(\alpha) = G_0^{-1}(\alpha/2) \quad \text{and} \quad c_U(\alpha) = G_0^{-1}(1-\alpha/2). \tag{4.10}$$

Example (A2)**.** One-sided test of the expected value of Gaussian distribution with known variance.

Let us have random sample $X_1, \ldots, X_n$ from the distribution $F_X = \mathsf{N}(\theta_X, \sigma_0^2) \in \mathcal{F}^A = \big\{\mathsf{N}(\theta, \sigma_0^2),\ \theta \in \mathbb{R}\big\}$. We test

$$H_0 : \theta_X \le \theta_0 \quad \text{against} \quad H_1 : \theta_X > \theta_0.$$

Test statistic is the same as in example A1

$$U_n = \frac{\sqrt{n}\,(\overline{X}_n - \theta_0)}{\sigma_0}.$$

Its distribution for $\theta_X = \theta_0$ is $\mathsf{N}(0,1)$. For the values $\theta_X = \theta_0 + \delta$ we have $U_n \sim \mathsf{N}(\nu_n, 1)$, where $\nu_n = \frac{\sqrt{n}\,\delta}{\sigma_0}$. If the null hypothesis is violated, then the distribution of the test statistic is moving to the positive values and it is further away with larger n and δ. Too large positive values of the test statistic will lead to rejecting the null hypothesis.

The critical region will be $C(\alpha) = \big[c_U(\alpha), \infty\big)$. The critical value $c_U(\alpha)$ will be chosen so that

$$\sup_{\boldsymbol{\theta} \in \Theta_0} \mathsf{P}_\theta[U_n \in C(\alpha)] = \alpha.$$

Since

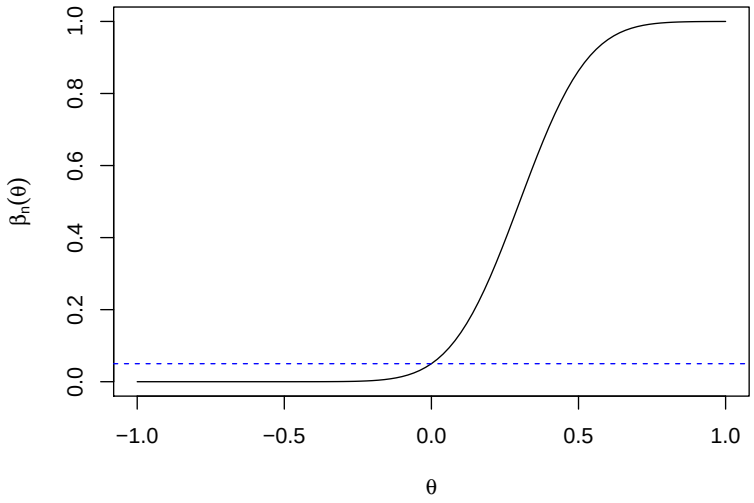
$$\mathsf{P}_\theta\big[U_n \in \big[c_U(\alpha), \infty\big)\big] = \mathsf{P}_\theta\Big[\tfrac{\sqrt{n}\,(\overline{X}_n - \theta_0)}{\sigma_0} \ge c_U(\alpha)\Big]$$

is increasing function of parameter θ, we have

$$\begin{aligned}\sup_{F \in \mathcal{F}_0} \mathsf{P}_F\big[U_n \in C(\alpha)\big] &= \sup_{\theta \in \Theta_0} \mathsf{P}_\theta\big[U_n \in C(\alpha)\big] = \sup_{\theta : \theta \le \theta_0} \mathsf{P}_\theta\big[U_n \in \big[c_U(\alpha), \infty\big)\big] \\ &= \mathsf{P}_{\theta_0}\big[U_n \in \big[c_U(\alpha), \infty\big)\big] = 1 - \Phi\big(c_U(\alpha)\big).\end{aligned}$$

So for $c_U(\alpha) = u_{1-\alpha}$ this test satisfies the condition $\sup_{\theta \in \Theta_0} \mathsf{P}_\theta[U_n \in C(\alpha)] = \alpha$ and so its significance level is α. It is worth noting that in this example the distribution F_0 from (4.5) is $\mathsf{N}(\theta_0, \sigma_0^2)$ and the function G_0, i.e. the distribution function of the test statistic U_n for X_i with distribution F_0, is the cumulative distribution function Φ of $\mathsf{N}(0,1)$.

Figure 4.3.: The power function of a test of the expected value of Gaussian distribution with known variance against right-sided alternative for $\theta_0 = 0$, $\sigma_0^2 = 1$, $n = 30$ and $\alpha = 0.05$.



Altogether we get the rule

$$\text{reject } H_0 : \theta_X \leq \theta_0 \quad \Longleftrightarrow \quad U_n = \frac{\sqrt{n}\left(\overline{X}_n - \theta_0\right)}{\sigma_0} \geq u_{1-\alpha},$$

i.e. reject the null hypothesis, if $\overline{X}_n$ is larger than θ_0 by more than $\frac{u_{1-\alpha}\sigma_0}{\sqrt{n}}$. We take 1.645 as the quantile $u_{1-\alpha/2}$ for $\alpha = 0.05$ and 1.282 for $\alpha = 0.1$. The critical value for one-sided test on level α is the same as the critical value for two-sided test on level $\alpha/2$. This follows from the fact that we reject the null hypothesis only in one of the tails of the distribution of U_n.

The computation of the *power function* is easier than before. Take some θ such that $\theta - \theta_0 = \delta$ and we get

$$\beta_n(\theta) = \mathsf{P}_\theta[U_n \geq u_{1-\alpha}] = \mathsf{P}_\theta[U_n - \nu_n \geq u_{1-\alpha} - \nu_n] = 1 - \Phi(u_{1-\alpha} - \nu_n).$$

The graph of the power function can be seen in figure 4.3. The sample size required for the test to have at least the power β against the alternative $\theta_0 + \delta$, $\delta > 0$, is

$$n \geq (u_{1-\alpha} + u_\beta)^2 \, \frac{\sigma_0^2}{\delta^2}.$$

Example (B)**.** Two-sided test of the expected value of Gaussian distribution with unknown variance.

Take random sample $X_1, \dots, X_n$ from distribution $F_X = \mathsf{N}(\theta_X, \sigma_X^2) \in \mathcal{F}^B = \left\{\mathsf{N}(\theta, \sigma^2),\ \theta \in \mathbb{R}, \sigma^2 > 0\right\}$. We test $H_0 : \theta_X = \theta_0$ against $H_1 : \theta_X \neq \theta_0$.

The test statistic from examples (A1) and (A2) cannot be used in this example, since the real variance σ_X^2 is unknown. However, we can replace it by the sample variance S_n^2 and get statistic

$$T_n = \frac{\sqrt{n}\,(\overline{X}_n - \theta_0)}{S_n},$$

which has, in this model under the null hypothesis H_0, t_{n-1} distribution (see theorem 2.10). If the null hypothesis does not hold, i.e. $\theta_X - \theta_0 = \delta \neq 0$, the test statistic can be written as

$$T_n = \frac{Z}{\sqrt{U/(n-1)}},$$

where $Z \sim \mathsf{N}(\nu_n, 1)$, $\nu_n = \frac{\sqrt{n}\delta}{\sigma_X}$, $U \sim \chi^2_{n-1}$ and U, Z are independent. Distribution of this random variable is called *non-central t distribution with $n-1$ degrees of freedom and noncentrality parameter ν_n*. Its characteristics (density, distribution function, moments) are complicated, but it is sufficient to know that it can be approximated for large n by the distribution $\mathsf{N}(\nu_n, 1)$.

As in the previous examples, if the null hypothesis does not hold, the distribution of the test statistic is moving away from zero, and this distance grows with larger n and $|\theta_X - \theta_0|$. So values of the test statistic far away from zero will lead to rejecting the null hypothesis.

The critical region is $\left(-\infty, c_L(\alpha)\right] \cup \left[c_U(\alpha), \infty\right)$. Notice that we take t_{n-1} distribution as G_0, since under H_0 it holds that $T_n \sim t_{n-1}$, for any positive σ^2. Since t_{n-1} is a symmetric distribution, we get, using (4.10), the following

$$c_L(\alpha) = t_{n-1}(\alpha/2) = -t_{n-1}(1-\alpha/2), \quad c_U(\alpha) = t_{n-1}(1-\alpha/2).$$

Let us complete this example by verifying that the test has (with the above choice of critical values) significance level α. Compute

$$\sup_{F \in \mathcal{F}_0} \mathsf{P}_F\big(T_n \in C(\alpha)\big) = \sup_{\sigma^2 > 0} \mathsf{P}_{\theta_0,\sigma^2}\big(T_n \leq -t_{n-1}(1-\alpha/2) \text{ or } T_n \geq t_{n-1}(1-\alpha/2)\big) = \alpha.$$

So the test has exact level α and we get the rule

$$\text{reject } H_0 : \theta_X = \theta_0 \quad \Longleftrightarrow \quad |T_n| = \left|\frac{\sqrt{n}\,(\overline{X}_n - \theta_0)}{S_n}\right| \geq t_{n-1}(1-\alpha/2).$$

This means that the null hypothesis will be rejected if the sample mean $\overline{X}_n$ will differ from the hypothetical value θ_0 by more than $\frac{t_{n-1}(1-\alpha/2)S_n}{\sqrt{n}}$. This test is called *one-sample t-test*.

The power function of this test can be obtained by similar process as in example (1A). Take some θ such that $\theta - \theta_0 = \delta \neq 0$. If θ is the true value of our parameter,

then the distribution of T_n is non -central t distribution with $n-1$ degrees of freedom and noncentrality parameter $\nu_n = \frac{\sqrt{n}\delta}{\sigma_X}$. Denote the distribution function of this distribution as G_{n,ν_n} and compute

$$\begin{aligned}\beta_n(\theta, \sigma_X^2) &= \mathsf{P}_{\theta,\sigma_X^2}[T_n \in C(\alpha)] \\ &= \mathsf{P}_{\theta,\sigma_X^2}[T_n \le -t_{n-1}(1-\alpha/2)] + \mathsf{P}_{\theta,\sigma_X^2}[T_n \ge t_{n-1}(1-\alpha/2)] \\ &= G_{n,\nu_n}\big(-t_{n-1}(1-\alpha/2)\big) + 1 - G_{n,\nu_n}\big(t_{n-1}(1-\alpha/2)\big).\end{aligned}$$

Non-central t distribution has non-symmetric density, hence the result cannot be simplified. If the number of observations n is large enough, we can approximate the power using the formula (4.2) or (4.3).

Using (4.3) we can get an approximation for the number of observations n needed for the test to have at least power β. The required sample size is

$$n \ge (u_{1-\alpha/2} + u_\beta)^2 \frac{\sigma_X^2}{\delta^2} + 1.$$

We add one to the left side to compensate for approximating t-distribution by Gaussian. To compute the power of our test and the required sample size, we either need to know the true value of variance σ_X^2 or it can be replaced by some preliminary estimate (since these calculations are usually done before obtaining our data).

Example (C)**.** TWO-SIDED TEST OF THE EXPECTED VALUE OF ANY DISTRIBUTION WITH FINITE VARIANCE.

Take random sample $X_1, \dots, X_n$ from distribution $F_X \in \mathcal{F}^C = \mathcal{L}_+^2$. Denote $\mathsf{E}\, X_i = \theta_X$ and $\mathsf{var}\, X_i = \sigma_X^2$. We test $H_0 : \theta_X = \theta_0$ against $H_1 : \theta_X \ne \theta_0$.

According to theorem 2.9 (limit theorem for T_n) the random variable

$$T_n = \frac{\sqrt{n}\,(\overline{X}_n - \theta_0)}{S_n}$$

has in this model under H_0 asymptotic distribution $\mathsf{N}(0,1)$. If the null hypothesis does not hold, i.e. $\theta_X - \theta_0 = \delta \ne 0$, then it can easily be shown*, that the test statistic

$$T_n = \frac{\sqrt{n}\,(\overline{X}_n - \theta_X + \theta_X - \theta_0)}{S_n} = \frac{\sqrt{n}\,(\overline{X}_n - \theta_X)}{S_n} + \sqrt{n}\,\frac{\delta}{S_n}$$

converges in probability to $+\infty$ or $-\infty$, depending on the sign of δ. So the values of the test statistic far away from zero will lead to rejecting the hypothesis.

The critical region will be $\big(-\infty, c_L(\alpha)\big] \cup \big[c_U(\alpha), \infty\big)$. Notice that

$$\sup_{F \in \mathcal{F}_0} \lim_{n\to\infty} \mathsf{P}_F\big(|T_n| \ge u_{1-\alpha/2}\big) = \mathsf{P}\big(|Z| \ge u_{1-\alpha/2}\big) = \alpha,$$

* We recommend to do this as an exercise.

where $Z \sim \mathsf{N}(0,1)$. So G_0 in (4.10) is, in this example, the distribution function of $\mathsf{N}(0,1)$. Therefore the critical values $c_U(\alpha) = -c_L(\alpha) = u_{1-\alpha/2}$ guarantee that the asymptotic level of the test is equal to α.

Instead of the critical value $u_{1-\alpha/2}$ we can use $t_{n-1}(1-\alpha/2)$, since the test is asymptotic and $t_{n-1}(1-\alpha/2) \to u_{1-\alpha/2}$ for $n \to \infty$. As $|t_{n-1}(\alpha)| \geq |u_\alpha|$ holds, the test will be more conservative, if we use the quantiles of t-distribution instead of the quantiles of Gaussian distribution.

Altogether we get the rule

$$\text{reject } H_0 : \theta_X = \theta_0 \quad \Longleftrightarrow \quad |T_n| = \left| \frac{\sqrt{n}\,(\overline{X}_n - \theta_0)}{S_n} \right| \geq t_{n-1}(1-\alpha/2).$$

It is again one-sample t-test. We have shown that, as an asymptotic test, it can be used for any data from distribution with finite variance.

Exercise.

1. In example (A1) (page 68) consider test $(U_n, C(\alpha))$, where $C(\alpha) = [u_{1/2-\alpha/2}, u_{1/2+\alpha/2}]$. Show that this test has significance level exactly α. Further show that the following holds for this test: for all θ not equal to θ_0

$$\beta_n(\theta) < \alpha \quad \text{and} \quad \lim_{n\to\infty} \beta_n(\theta) = 0.$$

2. In example (A1) (page 68) consider test $(U_n, C(\alpha))$, where $C(\alpha) = [u_{1-\alpha}, \infty)$. Show that this test has significance level exactly α. Is this test unbiased? For which θ is this test conservative? For which θ is this test consistent ?

3. Prove that the test from example (A2) (page 73) is unbiased and consistent.

4. Prove that the test from example (B) (page 74) is unbiased and consistent.
Hint: To prove unbiasedness we can use the fact that for random variable Z with non-central student distribution with ν degrees of freedom and non-zero parameter of noncentrality the following holds: $\mathsf{P}(|Z_n| \geq t_\nu(1-\alpha/2)) > \alpha$.

5. Prove that the test from example (C) (see previous page) is consistent.

6. The PR department of a certain high school would like to prove that the expected value of IQ of their students is higher then 105. They expect that the real expected value of IQ of their students is 110 and the standard deviation of the distribution of IQ of these students is 15. Find out the number of students whose IQ needs to be measured so that if we choose the significance level of 5 %, our test will prove with probability 95 % that the expected value of the student IQ is higher than 105.

The end of self-study for week 6 (10.11.-14.11.).

4.3. P-value

Deriving results of the test based on whether $S_n(\boldsymbol{X})$ lies in C or not is not the only way nor the most common way. Results of the test ale usually derived using so called **p-value**. It corresponds to the *smallest possible significance level on which we could reject the null hypothesis.*

Consider null hypothesis $H_0 : \boldsymbol{\theta}_X \in \Theta_0$ against alternative $H_1 : \boldsymbol{\theta}_X \in \Theta_1$ and, for fixed $\alpha \in (0,1)$, let $\big(S_n(\boldsymbol{X}), C(\alpha)\big)$ be a test with prescribed significance level α. For precision define $C(1) = (-\infty, \infty)$.

As we know from remark on page 66, if $S_n(\boldsymbol{X})$ has discrete distribution under the null hypothesis, it is not possible to reach any desired significance level of a test. If our desired level α is unreachable, we denote by $C(\alpha)$ the critical region of a test which has level $\alpha' < \alpha$, where α' is the closest to the desired α.

Definition 4.8 (P-value) Let $s_{\boldsymbol{x}} = S_n(\boldsymbol{x})$ be the observed value of the test statistic. Then we define *p-value* or *the obtained level of test* as

$$p(\boldsymbol{x}) = \inf\big\{\alpha \in (0,1\rangle \,:\, s_{\boldsymbol{x}} \in C(\alpha)\big\}.$$

If the test $\big(S_n(\boldsymbol{X}), C(\alpha)\big)$ is exact (resp. asymptotic), the p-value is called *exact* (resp. *asymptotic*).

If a test has prescribed level α, the following rule can be used to make our conclusion

$$\begin{aligned} H_0 \text{ is rejected, if } p(\boldsymbol{x}) \le \alpha, \\ H_0 \text{ is not rejected, if } p(\boldsymbol{x}) > \alpha. \end{aligned} \tag{4.11}$$

Therefore if we know the p-value $p(\boldsymbol{x})$, we can reject the null hypothesis on all levels $\alpha' \ge p(\boldsymbol{x})$, but we cannot reject it on levels $\alpha' < p(\boldsymbol{x})$. This is the reason for calling p-value *obtained level of test.*

If our decision is based on p-value, we **do not have to state the critical region** and we do not have to recalculate it, if we decide to change the level of a test. However, we do have to highlight that *changing the significance level after the result is known is not legitimate.*

Remark.

- P-value can be understood as *the amount of agreement of data* with the null hypothesis. If $p(\boldsymbol{x}) \ll \alpha$, the null hypothesis is rejected "safely". If $p(\boldsymbol{x})$ is close to α, we sometimes say that the result is "on the verge of statistical significance".
- P-value **cannot be** explained as a *"probability that the null hypothesis holds"*. Whether the null hypothesis holds or not is not a random event, but a deterministic one.
- If $S_n(\boldsymbol{X})$ has *discrete* distribution, then the rule (4.11) gives us a test which has the closest possible reachable level α' such that $\alpha' \le \alpha$.

4.3.1. Calculation of p-value for one-sided critical region

As can be seen from definition 4.8, p-value is a function of observed data $\boldsymbol{x}$ and her calculation depends on the used statistic $S_n(\boldsymbol{X})$ and on the way the critical region $C(\alpha)$ changes if we change α. The simplest case is the situation with one-sided critical region, i.e. we reject for too large (or too small) values of the test statistic.

Assume at first that we reject for too large values of the test statistic. For this purpose it is easier for a moment to think about the critical region in the open form $C(\alpha) = (c_U(\alpha), \infty)$. We know from Chapter 4.2.3 that $c_U(\alpha) = \lim_{h\to 0_+} G_0^{-1}(1-\alpha)$, where G_0 is a distribution function such that

$$\sup_{F\in\mathcal{F}_0} \mathsf{P}_F[S_n(\boldsymbol{X}) \geq c] = 1 - G_0(c-) \quad \forall c \in \mathbb{R}.$$

In this case we get from definition of p-value that

$$p(\boldsymbol{x}) = \inf\left\{\alpha \in (0,1\rangle : s_{\boldsymbol{x}} > G_0^{-1}(1-\alpha)\right\} = 1 - G_0(s_{\boldsymbol{x}}-). \tag{4.12}$$

We can proceed analogously for a critical region in the form of $C(\alpha) = \left(-\infty, G_0^{-1}(\alpha)\right)$, where G_0 is the distribution function such that

$$\sup_{F\in\mathcal{F}_0} \mathsf{P}_F[S_n(\boldsymbol{X}) \leq c] = G_0(c) \quad \forall c \in \mathbb{R}.$$

So from the definition of p-value

$$p(\boldsymbol{x}) = \inf\left\{\alpha \in (0,1\rangle : s_{\boldsymbol{x}} < G_0^{-1}(\alpha)\right\} = G_0(s_{\boldsymbol{x}}). \tag{4.13}$$

Remark.

- The formulas (4.12) and (4.13) can be used even for *asymptotic p-value* if G_0 is the distribution function of the asymptotic distribution of test statistic under the null hypothesis. I.e. consider critical region in the form of $C(\alpha) = \left[c_U(\alpha), \infty\right)$. Then we need that G_0 is a distribution function that satisfies

  $$\sup_{F\in\mathcal{F}_0} \lim_{n\to\infty} \mathsf{P}_F[S_n(\boldsymbol{X}) \geq c] = 1 - G_0(c-), \quad \forall c \in \mathbb{R}.$$

  Similarly for critical region $C(\alpha) = \left(-\infty, c_L(\alpha)\right]$ we need that G_0 satisfies

  $$\sup_{F\in\mathcal{F}_0} \lim_{n\to\infty} \mathsf{P}_F[S_n(\boldsymbol{X}) \leq c] = G_0(c), \quad \forall c \in \mathbb{R}.$$

- Notice that for critical region $C(\alpha) = \left[c_U(\alpha), \infty\right)$ the formula (4.12) for p-value can be rewritten into

  $$p(\boldsymbol{x}) = 1 - G_0(s_{\boldsymbol{x}}-) = \sup_{F\in\mathcal{F}_0} \mathsf{P}_F[S_n(\boldsymbol{X}) \geq s_{\boldsymbol{x}}]\,. \tag{4.14}$$

Similarly for critical region $C(\alpha) = \left(-\infty, c_L(\alpha)\right]$

$$p(\boldsymbol{x}) = G_0(s_{\boldsymbol{x}}) = \sup_{F \in \mathcal{F}_0} \mathsf{P}_F[S_n(\boldsymbol{X}) \leq s_{\boldsymbol{x}}]\,. \tag{4.15}$$

So the p-value can be also viewed as a (maximal possible) probability, that under the null hypothesis we would observe data which would be *in the same or larger disagreement* with the null hypothesis than the data we analyse.

Example (A). TEST OF EXPECTED VALUE OF GAUSSIAN DISTRIBUTION WITH KNOWN VARIANCE.

Let us have random sample $X_1, \ldots, X_n$ from distribution $F_X = \mathsf{N}(\theta_X, \sigma_0^2) \in \mathcal{F}^A = \left\{\mathsf{N}(\theta, \sigma_0^2),\ \theta \in \mathbb{R}\right\}$. We test

$$H_0 : \theta_X \geq \theta_0 \quad \text{against} \quad H_1 : \theta_X < \theta_0.$$

Test statistic is chosen as

$$U_n = \frac{\sqrt{n}\,(\overline{X}_n - \theta_0)}{\sigma_0},$$

and we reject for *small* values of the test statistic.

Notice that (see example 4.2.2) the distribution of the test statistic is $\mathsf{N}(\nu_n, 1)$, where the expected value $\nu_n = \frac{\sqrt{n}(\theta_X - \theta_0)}{\sigma_0}$ is non-negative under the null hypothesis. Let $u_{\boldsymbol{x}}$ denote the observed value of our test statistic U_n. Since the critical region is $C(\alpha) = \left(-\infty, c_L(\alpha)\right]$, where $c_L(\alpha) = \Phi^{-1}(\alpha)$, we get from definition 4.8 that

$$p(\boldsymbol{x}) = \inf\{\alpha \in (0,1) : u_{\boldsymbol{x}} \leq \Phi^{-1}(\alpha)\} = \Phi(u_{\boldsymbol{x}}).$$

Example (B). TEST OF EXPECTED VALUE OF GAUSSIAN DISTRIBUTION WITH UNKNOWN VARIANCE.

Take random sample $X_1, \ldots, X_n$, $n = 26$ from distribution $F_X \in \mathcal{F}^B$ and consider $\theta_X = \mathsf{E}\, X_i$. We test

$$H_0 : \theta_X \leq \theta_0 \quad \text{against} \quad H_1 : \theta_X > \theta_0.$$

To do that we use the test statistic

$$T_n = \frac{\sqrt{n}\,(\overline{X}_n - \theta_0)}{S_n},$$

and the null hypothesis is rejected for *large* values of our test statistic.

Suppose that we have calculated the value of our test statistic and denote this as $t_{\boldsymbol{x}}$. It was shown in the example B on page 74 that the test statistic T_n has non-central t-distribution with $n-1$ degrees of freedom and parameter of noncentrality $\nu_n = \frac{\sqrt{n}(\theta_X - \theta_0)}{\sigma_X}$. Notice that under the null hypothesis this parameter is negative or zero.

So large values of our test statistic will give evidence against the null hypothesis and critical region will be of form $C(\alpha) = \langle c_U(\alpha), \infty)$. Given all of this, we choose

the distribution function of t_{n-1}- distribution as G_0 in calculating the critical region using the formula (4.6), i.e. in our example we use the distribution function of t_{25}. So $c_U(\alpha) = t_{25}(1-\alpha)$ and therefore

$$p(\boldsymbol{x}) = \inf\left\{\alpha \in (0,1) : t_{\boldsymbol{x}} \geq t_{25}(1-\alpha)\right\} = 1 - G_{25}(t_{\boldsymbol{x}}),$$

where G_{25} is the distribution function of distribution t_{25}.

4.3.2. Calculation of p-value for a two-sided critical region

Again here it is easier to think about the critical region in the open form $C(\alpha) = (-\infty, c_L(\alpha)) \cup (c_U(\alpha), \infty)$, where $-\infty < c_L(\alpha) < c_U(\alpha) < \infty$, we get from definition of p-value that

$$p(\boldsymbol{x}) = \inf\left\{\alpha \in (0,1\rangle : s_{\boldsymbol{x}} < c_L(\alpha) \text{ or } s_{\boldsymbol{x}} > c_U(\alpha)\right\}. \tag{4.16}$$

We know from Chapter 4.2.3 that in the following text we will only encounter situations where the exact (or asymptotic) distribution of test statistic $S_n(\boldsymbol{X})$ does not depend on the choice of F from $\mathcal{F}_0$. Denote the distribution function of this (exact or asymptotic) distribution of $S_n(\boldsymbol{X})$ by G_0. Then according to (4.10) we have that

$$c_L(\alpha) = G_0^{-1}(\alpha/2), \quad \text{and} \quad c_U(\alpha) = \lim_{h\to 0_+} G_0^{-1}(1-\alpha/2).$$

So thanks to the formula (4.16) we get for the p-value that

$$\begin{aligned} p(\boldsymbol{x}) &= \inf\left\{\alpha \in (0,1\rangle : s_{\boldsymbol{x}} < G_0^{-1}(\alpha/2) \text{ or } s_{\boldsymbol{x}} > G_0^{-1}(1-\alpha/2)\right\} \\ &= 2\,\min\left\{G_0(s_{\boldsymbol{x}}), 1 - G_0(s_{\boldsymbol{x}}-)\right\}. \end{aligned} \tag{4.17}$$

The formula (4.17) can be simplified in the case that the exact (resp. asymptotic) distribution G_0 is symmetric around 0 and $c_L = -c_U$ (which is often true in practice). Then the exact (resp. asymptotic) p-value can be obtained as

$$p(\boldsymbol{x}) = \mathsf{P}_{F_0}\left[\,|S_n(\boldsymbol{X})| \geq |s_{\boldsymbol{x}}|\,\right] = 2\left(1 - G_0(|s_{\boldsymbol{x}}|-)\right). \tag{4.18}$$

Example (A). Test of expected value of Gaussian distribution with known variance.

We have random sample $X_1, \ldots, X_n$ from distribution $F_X = \mathsf{N}(\theta_X, \sigma_0^2) \in \mathcal{F}^A = \left\{\mathsf{N}(\theta, \sigma_0^2),\ \theta \in \mathbb{R}\right\}$ and we are interested in the hypothesis

$$H_0 : \theta_X = \theta_0 \quad \text{against} \quad H_1 : \theta_X \neq \theta_0.$$

Test statistic is

$$U_n = \frac{\sqrt{n}\left(\overline{X}_n - \theta_0\right)}{\sigma_0},$$

and we reject for too large, resp. too small values of test statistic. The calculation of p-value is fairly easy for this case since the hypothesis contains exactly one distribution $\mathsf{N}(\theta_0, \sigma_0^2)$, which will play the role of distribution F_0. Furthermore, the test statistic U_n has, under the null hypothesis, distribution $\mathsf{N}(0,1)$, which is symmetric around zero. So p-value can be obtained as

$$p(\boldsymbol{x}) = 2\min\left\{1-\Phi(u_{\boldsymbol{x}}), \Phi(u_{\boldsymbol{x}})\right\} = 2\left(1-\Phi(|u_{\boldsymbol{x}}|)\right).$$

Example (B)**.** Test of expected value of Gaussian distribution with unknown variance.

Take random sample $X_1, \ldots, X_n$, $n = 26$, from distribution $F_X \in \mathcal{F}^B$ and consider $\theta_X = \mathsf{E}\, X_i$. We test $H_0 : \theta_X = \theta_0$ against $H_1 : \theta_X \neq \theta_0$. To do that we use test statistic

$$T_n = \frac{\sqrt{n}\,(\overline{X}_n - \theta_0)}{S_n},$$

and the null hypothesis is rejected for too large or too small values of this test statistic.

Suppose that we have calculated the value of our test statistic and denote this value by $t_{\boldsymbol{x}}$. It was shown in example B on page 74 that the test statistic T_n has non-central t-distribution with $n-1$ degrees of freedom and parameter of noncentrality $\nu_n = \sqrt{n}(\theta_X - \theta_0)/\sigma_X$. Notice that under the null hypothesis this parameter is zero. So

$$p(\boldsymbol{x}) = 2\min\left\{1-G_{25}(t_{\boldsymbol{x}}), G_{25}(t_{\boldsymbol{x}})\right\} = 2\left(1-G_{25}(|t_{\boldsymbol{x}}|)\right).$$

We have used the fact that t-distribution with $n-1$ degrees of freedom is symmetric around zero.

Specifically for $t_{\boldsymbol{x}} = 1.37$ we get

$$p(\boldsymbol{x}) = 2\left(1-G_{25}(|1.37|)\right) \doteq 0.183.$$

Example (C)**.** Take random sample $X_1, \ldots, X_n$, $n = 26$, from distribution $F_X \in \mathcal{F}^C = \mathcal{L}_+^2$ with expected value $\mathsf{E}\, X_i = \theta_X$. We test $H_0 : \theta_X = \theta_0$ against $H_1 : \theta_X \neq \theta_0$. The test statistic T_n has under the null hypothesis approximately $\mathsf{N}(0,1)$ distribution, which is symmetric around 0. We have calculated the test statistic and the result is $t_{\boldsymbol{x}} = 1.37$. We can use (4.18) to obtain the asymptotic p-value of this test as

$$p(\boldsymbol{x}) = 2\left(1-\Phi(|1.37|)\right) \doteq 0.171. \tag{4.19}$$

We test on significance level $\alpha = 0.05$ and therefore we cannot reject the hypothesis, since $p(\boldsymbol{x}) > 0.05$. However, if we have set (before performing the test) our significance level as $\alpha' = 0.2$, we could reject the hypothesis.

Notice that in model $\mathcal{F}^B$ (i.e. the set of Gaussian distributions with unknown variance) we could use (4.18) to calculate the exact p-value as

$$p(\boldsymbol{x}) = 2\left(1-G_{25}(|1.37|)\right) \doteq 0.183, \tag{4.20}$$

where G_{25} denotes the distribution function of t-distribution t_{25}. As this p-value is higher than the asymptotic p-value (4.19), it is usual to use the p-value (4.20) calculated using the distribution t_{25} also in model $\mathcal{F}^C$ to be more careful (conservative). Since the distribution t_{n-1} converges (in distribution) to Gaussian distribution $\mathsf{N}(0,1)$, the p-value (4.20) can be viewed as an asymptotic p-value for model $\mathcal{F}^C$.

It is worth noticing that the formula (4.20) can be obtained directly from the definition of p-value, if we use the critical values $c_L(\alpha) = t_{n-1}(\alpha/2)$ a $c_U(\alpha) = t_{n-1}(1-\alpha/2)$. In this case we have

$$\begin{aligned} p(\boldsymbol{x}) &= \inf\left\{\alpha \in (0,1\rangle : 1.37 \leq t_{n-1}(\alpha/2) \text{ or } 1.37 \geq t_{n-1}^{-1}(1-\alpha/2)\right\} \\ &= 2\,\min\left\{1 - G_{25}(1.37), G_0(1.37)\right\} = 2\left[1 - G_{25}(|1.37|)\right]. \end{aligned}$$

4.3.3. Distribution of p-value under null hypothesis

Consider now p-value $p(\boldsymbol{X})$ as a random variable, i.e. statistic calculated from random sample $\boldsymbol{X}$. It can be shown that, under certain assumptions, $p(\boldsymbol{X})$ has under the null hypothesis uniform distribution on the interval $(0,1)$.

Proposition 4.1 Assume that the null hypothesis holds (i.e. $F_X \in \mathcal{F}_0$) and let the following be true

$$\sup_{F \in \mathcal{F}_0} \mathsf{P}_F[S_n(\boldsymbol{X}) \in C(\alpha)] = \mathsf{P}_{F_X}[S_n(\boldsymbol{X}) \in C(\alpha)]\,, \forall \alpha \in (0,1). \tag{4.21}$$

Assume that the test statistic $S_n(\boldsymbol{X})$ has **continuous** distribution. Then $p(\boldsymbol{X}) \sim \mathsf{U}(0,1)$.

Proof. Denote $U = G_0\big(S_n(\boldsymbol{X})\big)$, where G_0 is the distribution function of random variable $S_n(\boldsymbol{X})$, if the distribution of $\boldsymbol{X}_i$ is F_X. Notice that in this case the random variable U has uniform distribution on $(0,1)$ (see lemma A.2). The proposition will be proven separately for different forms of critical region.

(i) $C(\alpha) = \big[c_U(\alpha), \infty\big)$

In this case the formula (4.12) gives us p-value $p(\boldsymbol{x}) = 1 - G_0(s_{\boldsymbol{x}})$. So we can write for distribution function of random variable $p(\boldsymbol{X})$

$$\mathsf{P}_{F_X}[p(\boldsymbol{X}) \leq u] = \mathsf{P}_{F_X}\big[1 - G_0\big(S_n(\boldsymbol{X})\big) \leq u\big] = \mathsf{P}\big[1 - U \leq u\big] = \mathsf{P}\big[1 - u \leq U\big] = u,$$

for $\forall u \in (0,1)$. Therefore the distribution function of $p(\boldsymbol{X})$ is the distribution function of uniform distribution on $(0,1)$, which was to be proven.

(ii) $C(\alpha) = \big(-\infty, c_L(\alpha)\big]$

In this case we can use (4.13) for $\forall u \in (0,1)$ to get

$$\mathsf{P}_{F_X}[p(\boldsymbol{X}) \leq u] = \mathsf{P}_{F_X}\big[G_0\big(S_n(\boldsymbol{X})\big) \leq u\big] = \mathsf{P}\big[U \leq u\big] = u.$$

(iii) $C(\alpha) = \big(-\infty, c_L(\alpha)\big] \cup \big[c_U(\alpha), \infty\big)$

Using the formula (4.17) for $\forall u \in (0,1)$ we get

$$\begin{aligned}
\mathsf{P}_{F_X}[p(\boldsymbol{X}) \le u] &= \mathsf{P}_{F_X}\left[2\min\left\{1 - G_0\big(S_n(\boldsymbol{X})\big), G_0\big(S_n(\boldsymbol{X})\big)\right\} \le u\right] \\
&= \mathsf{P}\left[2\min\left\{1-U, U\right\} \le u\right] \\
&= \mathsf{P}\left[2\min\left\{1-U, U\right\} \le u,\ U \le \tfrac{1}{2}\right] + \mathsf{P}\left[2\min\left\{1-U, U\right\} \le u,\ U \ge \tfrac{1}{2}\right] \\
&= \mathsf{P}\left[2U \le u,\ U \le \tfrac{1}{2}\right] + \mathsf{P}\left[2(1-U) \le u,\ U \ge \tfrac{1}{2}\right] \\
&= \mathsf{P}\left[U \le \min\left\{\tfrac{u}{2}, \tfrac{1}{2}\right\}\right] + \mathsf{P}\left[U \ge \max\left\{1 - \tfrac{u}{2}, \tfrac{1}{2}\right\}\right] \\
&= \tfrac{u}{2} + 1 - (1 - \tfrac{u}{2}) = u.
\end{aligned}$$

□

Remark. The previous proposition does not hold if the distribution of the test statistic is discrete. It also would not hold if the hypothesis did hold (i.e. $F_X \in \mathcal{F}_0$), but F_X would not be "the closest" to the alternative, (i.e. we could not replace $\sup_{F\in\mathcal{F}_0} \mathsf{P}_F$ by P_{F_X} in (4.21)).

4.4. Duality between interval estimation and hypothesis testing

Consider random sample $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ from distribution $F_X \in \mathcal{F}$, where $\mathcal{F}$ is some model. Let $\theta = t(F) \in \mathbb{R}$ be a parameter and $\theta_X = t(F_X)$ its true value. In chapter 3.5 we have dealt with the problem of interval estimation of parameter θ_X, i.e. we have looked for random variables $\eta_L(\boldsymbol{X})$ and $\eta_U(\boldsymbol{X})$ such that

$$\mathsf{P}_F\left[\big(\eta_L(\boldsymbol{X}), \eta_U(\boldsymbol{X})\big) \ni \theta\right] = 1 - \alpha \quad (\text{or } \xrightarrow[n\to\infty]{} 1-\alpha) \quad \text{for } \forall F \in \mathcal{F}.$$

In this chapter we deal with hypothesis testing, specifically the hypothesis

$$H_0 : \theta_X = \theta_0 \text{ against } H_1 : \theta_X \ne \theta_0.$$

Both problems are solved by procedures that are similar in some way, even though they differ in details.

The following proposition shows that there exists certain duality between the problem of testing hypothesis about some parameter and looking for interval estimate for the same parameter. Interval estimation can be used to hypothesis testing and test of a hypothesis can be converted to interval estimation.

Proposition 4.2 (Duality of interval estimates and testing)

(i) Assume that we have two-sided confidence interval for parameter θ_X with confidence level $1-\alpha$ (exact or asymptotic), in the form $\big(\eta_L(\boldsymbol{X}), \eta_U(\boldsymbol{X})\big)$. Consider test of hypothesis $H_0 : \theta_X = \theta_0$ against $H_1 : \theta_X \ne \theta_0$ based on the rule

$$\begin{aligned}
&H_0 \text{ is rejected if } \theta_0 \notin \big(\eta_L(\boldsymbol{X}), \eta_U(\boldsymbol{X})\big) \\
&H_0 \text{ is not rejected if } \theta_0 \in \big(\eta_L(\boldsymbol{X}), \eta_U(\boldsymbol{X})\big).
\end{aligned} \tag{4.22}$$

Then the significance level of this test is α (exact or asymptotic).

(ii) Let there be, for all $\theta \in \Theta$, a test $(S_n(\boldsymbol{X}, \theta), C_\theta(\alpha))$ of the hypothesis $H_0 : \theta_X = \theta$ against $H_1 : \theta_X \neq \theta$ such that for all F satisfying $\theta = t(F)$

$$\mathsf{P}_F\big[S_n(\boldsymbol{X}, \theta) \in C_\theta(\alpha)\big] = \alpha \quad (\text{or} \quad \xrightarrow[n\to\infty]{} \alpha).$$

Denote by $B_n(\boldsymbol{X})$ the set containing all parameters $\theta \in \Theta$, such that for observed data $\boldsymbol{X}$ we do not reject the hypothesis $H_0 : \theta_X = \theta$. Then for all $F \in \mathcal{F}$

$$\mathsf{P}_F\big[B_n(\boldsymbol{X}) \ni \theta\big] = 1 - \alpha \quad (\text{or} \quad \xrightarrow[n\to\infty]{} 1 - \alpha),$$

and (if $B_n(\boldsymbol{X})$ is an interval) we have assembled confidence interval for parameter $\boldsymbol{\theta}_X$ with confidence level $1 - \alpha$ (exact or asymptotic).

Proof. Part (i) Let $\big(\eta_L(\boldsymbol{X}), \eta_U(\boldsymbol{X})\big)$ be exact confidence interval. The proof for asymptotic confidence interval would be analogous.

Confidence interval for the true value of parameter θ_X satisfies

$$\mathsf{P}_{F_X}\big[\big(\eta_L(\boldsymbol{X}), \eta_U(\boldsymbol{X})\big) \ni \theta_X\big] = 1 - \alpha.$$

So under the null hypothesis, i.e. for $\theta_X = \theta_0$, it holds for all $F \in \mathcal{F}_0 = \big\{F \in \mathcal{F} : t(F) = \theta_0\big\}$ that

$$\mathsf{P}_F\big[\big(\eta_L(\boldsymbol{X}), \eta_U(\boldsymbol{X})\big) \ni \theta_0\big] = 1 - \alpha.$$

Therefore the significance level of the test given by (4.22) is

$$\sup_{F\in\mathcal{F}_0} \mathsf{P}_F\big[\big(\eta_L(\boldsymbol{X}), \eta_U(\boldsymbol{X})\big) \not\ni \theta_0\big] = \alpha,$$

which was to be proven.

Part (ii) Let $\big(S_n(\boldsymbol{X}, \theta), C_\theta(\alpha)\big)$ be, for all $\theta \in \Theta$, the exact test of null hypothesis $H_0 : \theta_X = \theta$ against alternative $H_1 : \theta_X \neq \theta$ with significance level α. The proof for asymptotic test would be analogous.

Denote

$$B_n(\boldsymbol{X}) = \big\{\theta \in \Theta : S_n(\boldsymbol{X}, \theta) \notin C_\theta(\alpha)\big\}.$$

Then for all $F \in \mathcal{F}$, $\theta = t(F)$ we have that

$$\mathsf{P}_F[B_n(\boldsymbol{X}) \ni \theta] = \mathsf{P}_F[S_n(\boldsymbol{X}, \theta) \notin C_\theta(\alpha)] = 1 - \alpha,$$

which was to be proven. □

Proposition 4.2 says that if we can construct confidence interval for parameter, we can use it to test hypothesis about this parameter. Conversely, if we have a test for hypothesis, we can use it to construct confidence interval. However, this step requires more work, since we have to test all possible values of our parameter. Set of all values of our parameter, for which we do not reject the hypothesis, then has required confidence level, but it is not necessarily an interval.

Example. Let us have random sample $X_1, \dots, X_n$ from Gaussian distribution $F_X = \mathsf{N}(\theta_X, \sigma_X^2) \in \mathcal{F}^B = \{\mathsf{N}(\theta, \sigma^2),\ \theta \in \mathbb{R}, \sigma^2 > 0\}$.

Assume that we have calculated confidence interval (3.5) for expected value of Gaussian distribution with unknown variance. We then reject the null hypothesis $H_0 : \theta_X = \theta_0$ against alternative $H_1 : \theta_X \neq \theta_0$, if

$$\theta_0 \notin \left(\overline{X}_n - t_{n-1}(1 - \tfrac{\alpha}{2}) \frac{S_n}{\sqrt{n}},\ \overline{X}_n + t_{n-1}(1 - \tfrac{\alpha}{2}) \frac{S_n}{\sqrt{n}} \right).$$

I.e. confidence interval contains those values of our parameter for which we would not reject null hypothesis.

Conversely, if we use, for test $H_0 : \theta_X = \theta$ against alternative $H_1 : \theta_X \neq \theta$, test statistic

$$T_n(\theta) = \frac{\sqrt{n}\,(\overline{X}_n - \theta)}{S_n},$$

(see example (B) on page 74), then the above stated confidence interval can be derived as

$$\begin{aligned} \{\theta \in \mathbb{R} :\ &\text{do not reject } H_0 : \theta_X = \theta \text{ against } H_1 : \theta_X \neq \theta\} \\ &= \{\theta \in R : |T_n(\theta)| < t_{n-1}(1 - \alpha/2)\} \\ &= \left\{\theta \in R : \left|\tfrac{\sqrt{n}\,(\overline{X}_n - \theta)}{S_n}\right| < t_{n-1}(1 - \alpha/2)\right\}. \end{aligned}$$

Exercise. What would be the form of confidence interval derived from one-sided test, i.e. from testing $H_0 : \theta_X \leq \theta_0$ against alternative $H_1 : \theta_X > \theta_0$.

The end of self-study for week 7 (17.11.-21.11.).

5. One-sample and paired-problems for quantitative data

In this chapter we consider a random sample $X_1, \ldots, X_n$ of quantitative random variables with the cumulative distribution function F_X that belongs to the model $\mathcal{F}$. We are interested in the parameter $\theta_X = t(F_X)$. We want to test the hypothesis about this parameter and also to find a confidence interval for this parameter whenever possible

5.1. One-sample Kolmogorov-Smirnov test

The aim of the one-sample Kolmogorov-Smirnov test is to find if the true cumulative distribution function is the same as the given cumulative distribution function. It is a nonparametric test.

Model: $\mathcal{F}$ = {all **continuous** distributions}

The parameter being tested: The entire cumulative distribution function F_X

The hypothesis and the alternative:

$$H_0 : F_X(x) = F_0(x) \quad \forall x \in \mathbb{R}, \qquad H_1 : \exists x \in \mathbb{R} : F_X(x) \neq F_0(x),$$

where F_0 is a given continuous cumulative distribution function (without unknown parameters).

The test statistic is based on the empirical cumulative distribution function $\widehat{F}_n$, which was introduced in Chapter 3.6.1 (see page 51). Its properties are summarized in Theorem 3.3. The empirical cumulative distribution function is an unbiased and consistent estimator of the true cumulative distribution function in each of the point. Further according to Theorem 3.3(v) it is uniformly consistent, i.e.

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

The test statistic also uses this supreme norm which searches for the biggest difference between $\widehat{F}_n(x)$ and $F_0(x)$.

Test statistic:

$$K_n = \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F_0(x) \right|$$

If the (null) hypothesis is true and F_0 is true cumulative distribution function, then the value of the test statistic K_n is close to zero. The hypothesis is rejected, when the

empirical cumulative distribution function is not too different from F_0, i.e. when the value of the test statistic is too large.

Denote

$$K_n^+ = \sup_{x\in\mathbb{R}} \left(\widehat{F}_n(x) - F_0(x)\right) \quad \text{and} \quad K_n^- = \sup_{x\in\mathbb{R}} \left(F_0(x) - \widehat{F}_n(x)\right).$$

Then $K_n = \max\left(K_n^+, K_n^-\right)$.

Lemma 5.1 If F_0 is **continuous** then

$$K_n^+ = \max_{1\le i\le n}\left(\frac{i}{n} - F_0(X_{(i)})\right), \quad K_n^- = \max_{1\le i\le n}\left(F_0(X_{(i)}) - \frac{i-1}{n}\right).$$

Proof. Define $X_{(0)} = -\infty$ and $X_{(n+1)} = +\infty$. Then

$$\widehat{F}_n(x) = \frac{i}{n}, \quad \text{pro} \quad x \in \langle X_{(i)}, X_{(i+1)}), \quad i = 0, 1, \ldots, n.$$

Thus with the help of the above equation

$$\begin{aligned} K_n^+ &= \sup_{x\in\mathbb{R}} \left(\widehat{F}_n(x) - F_0(x)\right) = \max_{0\le i\le n} \sup_{X_{(i)}\le x<X_{(i+1)}} \left(\widehat{F}_n(x) - F_0(x)\right) \\ &= \max_{0\le i\le n} \left(\tfrac{i}{n} - \inf_{X_{(i)}\le x<X_{(i+1)}} F_0(x)\right) \\ &= \max_{0\le i\le n} \left(\tfrac{i}{n} - F_0(X_{(i)})\right) = \max_{1\le i\le n} \left(\tfrac{i}{n} - F_0(X_{(i)})\right), \end{aligned}$$

where in the last equality we make use of the fact that $F_0(X_{(0)}) = 0$ and that $1 - F_0(X_{(n)}) \ge 0$.

Analogously for K_n^-:

$$\begin{aligned} K_n^- &= \sup_{x\in\mathbb{R}} \left(F_0(x) - \widehat{F}_n(x)\right) = \max_{0\le i\le n} \sup_{X_{(i)}\le x<X_{(i+1)}} \left(F_0(x) - \widehat{F}_n(x)\right) \\ &= \max_{0\le i\le n} \left(F_0(X_{(i+1)}) - \tfrac{i}{n}\right) = \max_{0\le i\le n-1} \left(F_0(X_{(i+1)}) - \tfrac{i}{n}\right) \\ &= \max_{1\le i\le n} \left(F_0(X_{(i)}) - \tfrac{i-1}{n}\right), \end{aligned}$$

where in the second to last equality we make use of the fact that $F_0(X_{(n+1)}) = 1$ and that $F_0(X_{(1)}) \ge 0$. In the last equality we only shift indices. □

Remark. The above lemma has several important consequences.

- The test statistic K_n can be calculated with the help of Lemma 5.1. No that to calculate K_n it is sufficient to calculate the ordered random sample (and not $\widehat{F}_n$).
- With the help of Theorem 2.13 under the null hypothesis $F_0(X_{(i)})$ follows a beta distribution whose parameters do not depend on F_0. Thus the distribution of K_n under the hypothesis does not depend on F_0 (i.e. it is pivotal).

- With the help Lemma 5.1 one can theoretically find the exact distribution distribution of the test statistics under the null hypothesis. But to really evaluate this distribution would be a rather computationally intensive task. Thus the exact distribution of K_n is used only for small sample sizes n for which it is tabulated.

Asymptotic distribution of the test statistic under the null hypothesis is given by the following proposition which generalizes the result of Theorem 3.3(v).

Proposition 5.2 Let $X_1, \dots, X_n$ be a random sample from the continuous distribution with the cumulative distribution function F_X. Then

$$\sqrt{n} \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F_X(x) \right| \xrightarrow[n\to\infty]{\mathsf{d}} Z,$$

where the random variable Z has the cumulative distribution function given by

$$G(y) = \begin{cases} 1 - 2\sum_{k=1}^{\infty} (-1)^{k+1} \mathrm{e}^{-2k^2y^2}, & y > 0, \\ 0, & y \le 0. \end{cases} \tag{5.1}$$

The cumulative distribution function $G(y)$ gives the limiting distribution of the normalized test statistic $\sqrt{n}K_n$ under the null hypothesis, i.e. for $F_X = F_0$. Note that this distribution is not Gaussian. It is worth noting that this distribution does not depend on the choice of F_0. The proof of the Proposition 5.2 requires a very advanced methods of the theory of probability.

Now we can find the critical value for rejecting H_0 so that the test has the asymptotic level α. Let $k_\alpha = G^{-1}(\alpha)$ be the α-quantile of the distribution given by the cumulative distribution function G. Now we reject H_0 when $\sqrt{n}K_n$ exceeds $k_{1-\alpha}$.

Critical region:

$$H_0 \text{ is rejected } \Leftrightarrow \sqrt{n}K_n \ge k_{1-\alpha}. \tag{5.2}$$

With the help of Proposition 5.2 we know that the asymptotic level of the test is α.

P-value: $p = 1 - G\left(\sqrt{n}\, k_n\right)$, where k_n is observed value of the statistic K_n. Note that the above equation gives an asymptotic p-value.

Remark.

- Under the alternative

$$K_n \xrightarrow[n\to\infty]{\mathsf{P}} \sup_{x \in \mathbb{R}} |F_X(x) - F_0(x)| > 0$$

  from which one conclude that the test is consistent. The advantage of Kolmogorov-Smirnov test is its universality (it is capable to detect any difference of the true distribution of data from the the distribution given by null hypothesis) and that no parametric assumptions are made.
- On the other hand this test has a relatively small power against specific violations of H_0 (e.g. the change in the expectation). When we know what type of the violation of H_0 to expect in the given application then it is usually better to use a test that is specialized to detect this particular violation.

- It is possible to formulate this test also as one-sided, i.e. $H_1' : F_X(x) \geq F_0(x)$, $\exists x \in \mathbb{R} : F_X(x) > F_0(x)$ or $H_1'' : F_X(x) \leq F_0(x)$, $\exists x \in \mathbb{R} : F_X(x) < F_0(x)$. Then we use either K_n^+ or K_n^- as the test statistics and we reject for large values of test statistics. But one cannot use Proposition 5.2 to find critical values. For that reason one needs to derive the asymptotic distribution of $\sqrt{n}\,K_n^+$ (or $\sqrt{n}\,K_n^-$) under H_0.

CONFIDENCE INTERVALS FOR F_X

Suppose that $x \in S_X = \{x : F_X(x) \in (0,1)\}$ be given and we are interested in the confidence interval for $F_X(x)$. Then we can use Theorem 3.3(iii) and use the same construction as in the example on page 50 in Chapter 3.5.2. Then we get the confidence interval

$$IS_n(x) = \left(\widehat{F}_n(x) - \frac{u_{1-\frac{\alpha}{2}}\sqrt{\widehat{F}_n(x)(1-\widehat{F}_n(x))}}{\sqrt{n}},\ \widehat{F}_n(x) + \frac{u_{1-\frac{\alpha}{2}}\sqrt{\widehat{F}_n(x)(1-\widehat{F}_n(x))}}{\sqrt{n}} \right).$$

For this confidence interval it holds that

$$\mathsf{P}\left[IS_n(x) \ni F_X(x)\right] \xrightarrow[n\to\infty]{} 1-\alpha,\ \forall x \in S_X.$$

This interval is also called pointwise confidence interval for $F_X(x)$.

Sometimes we are not interested in a given point x but rather in set that would cover the entire cumulative distribution function. To do that one can make use of Proposition 5.2. The thing is that

$$\mathsf{P}\left[\sqrt{n}\left|\widehat{F}_n(x) - F_X(x)\right| < k_{1-\alpha},\ \forall x \in \mathbb{R}\right] = \mathsf{P}\left[\sqrt{n}\sup_{x\in\mathbb{R}}\left|\widehat{F}_n(x) - F_X(x)\right| < k_{1-\alpha}\right] \xrightarrow[n\to\infty]{} 1-\alpha$$

Thus for $x \in \mathbb{R}$ one can calculate the interval

$$B_n(x) = \left(\widehat{F}_n(x) - \frac{k_{1-\alpha}}{\sqrt{n}},\ \widehat{F}_n(x) + \frac{k_{1-\alpha}}{\sqrt{n}} \right),$$

which has the following property

$$\mathsf{P}\left[B_n(x) \ni F_X(x),\ \forall x \in S_X\right] \xrightarrow[n\to\infty]{} 1-\alpha.$$

Such intervals that creates a region that covers the entire unknown function with a given probability are called **confidence bounds**. As the boundaries of the above confidence bounds for the cumulative distribution function can be outside of the interval $\langle 0,1\rangle$ it is natural to redefine the lower bound as $\max\{0, \widehat{F}_n(x) - k_{1-\alpha}/\sqrt{n}\}$ and the upper bound as $\min\{1, \widehat{F}_n(x) + k_{1-\alpha}/\sqrt{n}\}$.*

* In fact this is only one of the possible ways how to calculate confidence bounds for F_X.

POSSIBLE VIOLATIONS OF THE ASSUMPTIONS OF THE TEST

F_0 is not continuous Also in this situation one can use statistic K_n. But one has to be careful that now the statement of Proposition 5.2 does not hold. Ignoring this fact and using the quantile $k_{1-\alpha}$ would result in an asymptotically conservative test implying a lost of power. One should be also careful that in this situation one cannot use Lemma 5.1 to calculate the test statistic.

F_0 is continuous but there ties in observed data. Strictly speaking the probability of observing ties is zero when the data comes from the continuous distribution. In applications ties can be present due to rounding. Thus formally in fact we observe $\widetilde{X}_1, \ldots, \widetilde{X}_n$, where $\widetilde{X}_i$ is a rounded value of X_i. Thus the empirical cumulative distribution function of the observed values $\widetilde{X}_1, \ldots, \widetilde{X}_n$

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\widetilde{X}_i \leq x\}$$

estimates in fact the cumulative distribution function $\widetilde{F}_0$ of rounded a random variable $\widetilde{X}_i$. Nevertheless the test can be still used as an approximate test when $\widetilde{F}_0$ is not different from F_0. More precisely when

$$\sqrt{n} \sup_{x \in \mathbb{R}} \left| \widetilde{F}_0(x) - F_0(x) \right|,$$

is not too "large". This is often satisfied in applications.

Hypothesis is not simple. Note that F_0 should not contain unknown parameters (or its estimates). Suppose that we are interested in testing the null hypothesis

$$H_0 : F_X \in \mathcal{F}_0, \qquad H_1 : F_X \notin \mathcal{F}_0,$$

where $\mathcal{F}_0 = \{F(x;\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ is a a parametric family of distributions (e.g. $\{\mathsf{N}(\mu, \sigma^2),\ \mu \in \mathbb{R}, \sigma^2 > 0\}$). Then it is natural to consider the test statistic

$$\widetilde{K}_n = \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F(x; \widehat{\boldsymbol{\theta}}_n) \right|,$$

where $\widehat{\boldsymbol{\theta}}_n$ is the estimate of the true value of the parameter $\boldsymbol{\theta}_X$. The problem is that Proposition 5.2 does not hold for the statistic $\widetilde{K}_n$. Further it has been derived that the asymptotic distribution of $\widetilde{K}_n$ is rather complex and depending on the unknown value of the parameter $\boldsymbol{\theta}_X$. Ignoring this fact and using the the quantile $k_{1-\alpha}$ would result in a test that is very conservative and thus suffers from a big loss of power.

All the above problems can be solved with the help of the bootstrap methods (the course *Mathematical Statistics 4*).

Exercise. Consider the test with the critical region

$$H_0 \text{ is rejected} \quad \Leftrightarrow \quad \sqrt{n} K_n \leq k_{\alpha/2} \text{ or } \sqrt{n} K_n \geq k_{1-\alpha/2}.$$

1. Does this test keep the level α (exactly or asymptotically)?
2. How would you calculate the p-value of this test?
3. Is this test consistent?
4. Why is this test better or worse than the test wit the critical region given by (5.2)?

5.2. One-sample t-test

One-sample t-test compares **the expected value** that is in agreement with our data with the given constant. This test was described and investigated in detail in Example B on p. 74 and also in Example C on p. 76. The only difference was only in the models that was assumed.

Model: $\mathcal{F}^B = \{\mathsf{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$ or $\mathcal{F}^C = \mathcal{L}_+^2$

The parameter being tested: The expected value $\mu_X = \mathsf{E}\, X_i$

The hypothesis and the alternative:

$$H_0 : \mu_X = \mu_0, \quad H_1 : \mu_X \neq \mu_0,$$

where μ_0 is a given constant.

Test statistic:

$$T_n = \frac{\sqrt{n}\,(\overline{X}_n - \mu_0)}{S_n},$$

where $\overline{X}_n$ is a sample mean and S_n^2 is a sample variance

Distribution of the test statistic under H_0:

$$\text{In model } \mathcal{F}^B : T_n \sim t_{n-1} \text{ (see Theorem 2.10)}$$
$$\text{In model } \mathcal{F}^C : T_n \overset{\text{as.}}{\sim} \mathsf{N}(0,1) \text{ (see Theorem 2.9).}$$

Thus the test is **exact**, when in the "smaller" model $\mathcal{F}^B$. For the "bigger" model $\mathcal{F}^C$ this test **asymptotic**. Analogously this hold true also for the p-value and the confidence interval. In model $\mathcal{F}^B$ the p-value and the confidence interval are exact. In model $\mathcal{F}^C$ only asymptotic.

Critical region:

$$H_0 \text{ is rejected } \Leftrightarrow |T_n| \geq t_{n-1}(1-\alpha/2),$$

where $t_{n-1}(1-\alpha/2)$ is the $(1-\alpha/2)$ quantile of Student t-distribution with $n-1$ degrees of freedom.

P-value: $p = 2\,(1 - F_n(|t|))$, where t is the observed value of the test statistic T_n and F_n is the cumulative distribution function distribution of t_{n-1}.

Confidence interval for μ_X: Confidence interval for the expected value is given by

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

See the formula on p. 48 and the following example.

Remark. T-test does not necessarily requires the normal distribution. It works as an asymptotic (approximate) test of the expected value for an arbitrary distribution with the finite variance.

Remark. The t-test can be also performed as an one-sample test

$$\text{Rejecting } H_0' : \mu_X \leq \mu_0 \text{ against } H_1' : \mu_X > \mu_0 \;\Leftrightarrow\; T_n \geq t_{n-1}(1-\alpha).$$

Analogously $H_0'' : \mu_X \geq \mu_0$ against $H_1'' : \mu_X < \mu_0$ is rejected, when the test statistic T_n is smaller than the critical value $-t_{n-1}(1-\alpha)$.

The end of self-study for week 8 (24.11.-28.11.).

5.3. One-sample sign test

One-sample sign test compares **the median** that is in agreement with our data with the given value. It is a non-parametric test and it works for any continuous distribution.

Model: $\mathcal{F}$ = {all **continuous** distributions}

The parameter being tested: the median $m_X = F_X^{-1}(0.5)$

The hypothesis and the alternative:

$$H_0 : m_X = m_0, \quad H_1 : m_X \neq m_0,$$

where m_0 is a given constant.

Test statistic:

$$B_n = \sum_{i=1}^{n} \mathbb{1}\{X_i > m_0\}$$

(number of observations bigger than m_0).

Theorem 5.3 Let $X_1, \ldots, X_n$ be a random sample from an arbitrary **continuous** distribution with the median m_X. Then

(i)
$$\sum_{i=1}^{n} \mathbb{1}\{X_i > m_X\} \sim \mathrm{Bi}\left(n, \tfrac{1}{2}\right),$$

(ii)
$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\mathbb{1}\{X_i > m_X\} - \frac{1}{2} \right] \xrightarrow[n\to\infty]{\mathrm{d}} \mathrm{N}\left(0, \tfrac{1}{4}\right).$$

Remark. Theorem 5.3 follows from Theorem 2.3(iii) and (iv).

The exact distribution of the test statistic under H_0:

$$B_n \sim \mathrm{Bi}\left(n, \tfrac{1}{2}\right), \quad \text{(viz Theorem 5.3(i))}$$

Critical region (exact test): The hypothesis is rejected for too small or too large values of B_n.

$$H_0 \text{ is rejected} \;\Leftrightarrow\; B_n \le c_L(\alpha) \text{ or } B_n \ge c_U(\alpha)$$

kde

$$c_L(\alpha) = \max\left\{k_1 \in \mathbb{N}_0 : \mathsf{P}\left(\mathrm{Bi}\left(n, \tfrac{1}{2}\right) \le k_1\right) \le \frac{\alpha}{2}\right\} = \max\left\{k_1 \in \mathbb{N}_0 : \frac{1}{2^n}\sum_{j=0}^{k_1}\binom{n}{j} \le \frac{\alpha}{2}\right\}$$

$$c_U(\alpha) = \min\left\{k_2 \in \mathbb{N}_0 : \mathsf{P}\left(\mathrm{Bi}\left(n, \tfrac{1}{2}\right) \ge k_2\right) \le \frac{\alpha}{2}\right\} = \min\left\{k_2 \in \mathbb{N}_0 : \frac{1}{2^n}\sum_{j=k_2}^{n}\binom{n}{j} \le \frac{\alpha}{2}\right\}$$

From the symmetry of the binomial distribution for $p = \frac{1}{2}$ it follows that $c_L(\alpha) + c_U(\alpha) = n$. This test has the level at most α (the α might not be attainable).

P-value (exact):

$$p = 2\min\left\{\mathsf{P}\left(\mathrm{Bi}\left(n, \tfrac{1}{2}\right) \le y_n\right), \mathsf{P}\left(\mathrm{Bi}\left(n, \tfrac{1}{2}\right) \ge y_n\right)\right\} = 2\min\left\{1 - G_0(y_n - 1), G_0(y_n)\right\},$$

where G_0 is cumulative distribution function $\mathrm{Bi}\left(n, \frac{1}{2}\right)$ and y_n is the observed value of B_n.

Asymptotic distributions of the test statistic under H_0:

$$Z_n = \frac{B_n - \frac{n}{2}}{\sqrt{\frac{n}{4}}} = \frac{2}{\sqrt{n}}\left(B_n - \frac{n}{2}\right) \overset{\text{as.}}{\sim} \mathsf{N}(0,1), \quad \text{(see Theorem 5.3(ii))}$$

Critical region (asymptotic test): The hypothesis is rejected for too small or too large values of B_n.

$$H_0 \text{ is rejected} \;\Leftrightarrow |Z_n| \ge u_{1-\alpha/2}.$$

P-value (asymptotic): $p = 2\left(1 - \Phi(|z_n|)\right)$, where z_n is observed value of the test statistic Z_n.

Confidence interval pro m_X: See confidence intervals for quantiles (Chapter 3.6.4).

Remark.

- Note that we do not need the exact values X_i to calculate the test statistic. All we need is to know how many of them are bigger than m_0.
- This test can be performed also as one-sided test $H_0' : m_X \ge m_0$ (or $\le m_0$).

- This test can be easily modified as a test about an arbitrary quantile. I.e. one can test the hypothesis

$$H_0 : u_X(\beta) = u_0, \quad H_1 : u_X(\beta) \neq u_0,$$

where $\beta \in (0,1)$. Then the test statistic $B_n = \sum_{i=1}^{n} \mathbb{1}\{X_i > u_0\}$ under the null hypothesis follows the binomial distribution $\mathrm{Bi}(n, 1-\beta)$. The test about the quantile is then performed as a test about the parameter of the binomial distribution. This will be in detail treated later in Chapter 7.1.

Exercise. Show that the sign test is consistent.
Hint: It might be easier to work with the asymptotic version of the sign test.

Violations of the assumptions

Although in the literature it is usually required that the distribution F_X is continuous in fact it is sufficient to assume that $\mathsf{P}[X_i = m_0] = 0$. Nevertheless in applications it might happen that due to rounding some of the observations are exactly equal to m_0. The usual practice is then to remove such observations.

5.4. One-sample Wilcoxon test (Wilcoxon signed-rank test)

This test assumes a **symmetric** distribution and it compares **the center of the symmetry** with a given constant.

Model: $\mathcal{F} = \{$continuous distribution with the density f that satisfies $\exists \delta \in \mathbb{R} : f(\delta - x) = f(\delta + x)\ \forall x \in \mathbb{R}\}$

The parameter being tested: the center of the symmetry δ_X

Remark. The model requires the **density** of X_i being **symmetric** around the point δ_X. Then it holds that $m_X = \delta_X$ and if moreover $X_i \in \mathcal{L}^1$, then also $\mathsf{E}\, X_i \equiv \mu_X = \delta_X$.

The hypothesis and the alternative:

$$H_0 : \delta_X = \delta_0, \quad H_1 : \delta_X \neq \delta_0,$$

where δ_0 is a given constant.

Remark. Provided that model $\mathcal{F}$ holds then the hypothesis H_0 is equivalent to the hypothesis $H_0^* : m_X = \delta_0$ (i.e. we are resting the median). Further if $X_i \in \mathcal{L}^1$, then the hypothesis H_0 is also equivalent to the hypothesis $H_0^{**} : \mu_X = \delta_0$ (i.e. we are testing the expected value).

Test statistic: Let $Z_i \overset{\text{df}}{=} X_i - \delta_0$. Define

$$W_n = \sum_{i \in \mathcal{I}} R_i,$$

where $\mathcal{I} = \{i \in \{1, \dots, n\} : Z_i > 0\}$ is a set of indices such that Z_i is positive and R_i is the rank of the absolute values $|Z_i|$ among all absolute values $|Z_1|, \dots, |Z_n|$.

Remark. The test statistic W_n takes values in the set $\{0, 1, \dots, \frac{n(n+1)}{2}\}$. It is calculated as follows.

1. Calculate $Z_i = X_i - \delta_0$ and find the set $\mathcal{I}$.
2. Calculate $|Z_1|, \dots, |Z_n|$.
3. Order $|Z_i|$ from the smallest ones to the largest and get the ordered random sample
$$0 < |Z|_{(1)} < |Z|_{(2)} < \cdots < |Z|_{(n)}.$$
4. Find the rank R_i of the random variable $|Z_i|$ among all random variables $|Z|_{(1)}, \dots, |Z|_{(n)}$. It holds that $|Z_i| = |Z|_{(R_i)}$.
5. Calculate the sum of the ranks R_i for $i \in \mathcal{I}$.

The cardinality of the set $\mathcal{I}$ is equal to the number of variables for which $X_i > \delta_0$. (compare this with the test statistic of the sign test).

Proposition 5.4 Let $X_1, \dots, X_n$ be a random sample from an arbitrary **continuous** distribution that belongs to $\mathcal{F}$. Further let **the null hypothesis** $H_0 : \delta_X = \delta_0$ **holds**. Then

(i)
$$\mathsf{E}_{H_0} W_n = \frac{n(n+1)}{4}, \quad \mathsf{var}_{H_0}(W_n) = \frac{n(n+1)(2n+1)}{24}.$$

(ii)
$$\frac{W_n - \mathsf{E}_{H_0} W_n}{\sqrt{\mathsf{var}_{H_0}(W_n)}} \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}(0,1).$$

Proof. Without loss of generality consider $\delta_0 = 0$ and introduce the random variables $\Delta_i = \text{sign}(Z_i)$. Note that

$$W_n = \sum_{i=1}^{n} R_i \, \mathbb{1}\{\Delta_i = 1\}.$$

The random variables $\Delta_1, \dots, \Delta_n$ are under H_0 independent and identically distributed and

$$\mathsf{P}(\Delta_i = 1) = \mathsf{P}(\Delta_i = -1) = \tfrac{1}{2}.$$

From this we easily calculate that

$$\mathsf{E}\,\Delta_i = 0, \qquad \mathsf{E}\,\Delta_i^2 = 1.$$

The proof will be divided into 3 steps.

1. *Showing that* $(R_1, \dots, R_n)^\mathsf{T}$ *and* $(\Delta_1, \dots, \Delta_n)^\mathsf{T}$ *are independent.*

First note that the random vector $(R_1, \dots, R_n)^\mathsf{T}$ is a function of the random vector $(|Z_1|, \dots, |Z_n|)^\mathsf{T}$. Thus it is sufficient to show that the random vectors $(|Z_1|, \dots, |Z_n|)^\mathsf{T}$ and $(\Delta_1, \dots, \Delta_n)^\mathsf{T}$ are independent.

In order to do that note that the random vectors $\binom{|Z_1|}{\Delta_1}, \dots, \binom{|Z_n|}{\Delta_n}$ are independent. Thus it is sufficient to show the independence of $|Z_i|$ and Δ_i.

For $\forall z > 0$ it holds that

$$\begin{aligned}\mathsf{P}\big[|Z_i| \le z, \Delta_i = 1\big] &= \mathsf{P}\big[0 \le Z_i \le z\big] = \tfrac{1}{2}\,\mathsf{P}\big[-z \le Z_i \le z\big] \\ &= \tfrac{1}{2}\,\mathsf{P}\big[0 \le |Z_i| \le z\big] = \mathsf{P}\big[\Delta_i = 1\big]\,\mathsf{P}\big[|Z_i| \le z\big],\end{aligned}$$

where in the second equation we use the fact that the distribution of Z_i is (under the null hypothesis) symmetric around zero. Thus $|Z_i|$ and Δ_i are indeed independent.

2. *Writing* W_n *as a function of* R_i *and* Δ_i.

Note that

$$\begin{aligned}\sum_{i=1}^n R_i\,\mathbb{1}\{\Delta_i = 1\} + \sum_{i=1}^n R_i\,\mathbb{1}\{\Delta_i = -1\} &= \sum_{i=1}^n R_i = \frac{n(n+1)}{2}, \\ \sum_{i=1}^n R_i\,\mathbb{1}\{\Delta_i = 1\} - \sum_{i=1}^n R_i\mathbb{1}\{\Delta_i = -1\} &= \sum_{i=1}^n R_i\,\Delta_i.\end{aligned}$$

"Averaging" the above two equations and with the help that $W_n = \sum_{i=1}^n R_i\,\mathbb{1}\{\Delta_i = 1\}$ we get

$$W_n = \frac{n(n+1)}{4} + \frac{1}{2}\sum_{i=1}^n R_i\,\Delta_i. \tag{5.3}$$

3. *Calculating* $\mathsf{E}_{H_0} W_n$ *and* $\mathsf{var}_{H_0}(W_n)$.

Using (5.3) together with the independence of R_i and Δ_i and that $\mathsf{E}\,\Delta_i = 0$ it holds

$$\mathsf{E}_{H_0} W_n = \frac{n(n+1)}{4} + \frac{1}{2}\sum_{i=1}^n \mathsf{E}\,R_i\,\mathsf{E}\,\Delta_i = \frac{n(n+1)}{4}.$$

Further

$$\mathsf{var}_{H_0}(W_n) = \frac{1}{4}\,\mathsf{var}\left(\sum_{i=1}^n R_i\,\Delta_i\right) = \frac{1}{4}\sum_{i=1}^n \mathsf{var}\left(R_i\,\Delta_i\right) + \frac{1}{4}\sum_{i=1}^n \sum_{j=1, j\neq i}^n \mathsf{cov}\left(R_i\,\Delta_i, R_j\,\Delta_j\right).$$

Next

$$\mathsf{var}\left(R_i\,\Delta_i\right) = \mathsf{E}\left(R_i\,\Delta_i\right)^2 = \mathsf{E}\,R_i^2\,\mathsf{E}\,\Delta_i^2 = \frac{1}{n}\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6n} = \frac{(n+1)(2n+1)}{6},$$

where we utilize that $\mathsf{E}\,R_i\Delta_i = 0$, $\mathsf{E}\,\Delta_i^2 = 1$ and Theorem 2.16(i) which implies $\mathsf{P}[R_i = k] = \frac{1}{n}$ for all $i, k \in \{1, \dots, n\}$.

Further for $i \neq j$ calculate

$$\operatorname{cov}\left(R_i\,\Delta_i, R_j\,\Delta_j\right) = \mathsf{E}\left(R_i\,\Delta_i\,R_j\,\Delta_j\right) = \mathsf{E}\left(R_i\,R_j\right)\mathsf{E}\,\Delta_i\,\mathsf{E}\,\Delta_j = 0,$$

where we use the independence of R_i and Δ_i.

Finally we get

$$\operatorname{var}_{H_0}(W_n) = \frac{1}{4}\sum_{i=1}^{n}\frac{(n+1)(2n+1)}{6} = \frac{n(n+1)(2n+1)}{24}\,.$$

□

Remark.

- The proof of asymptotic normality is left out. The proof is difficult because of the fact that the ranks $R_1, \ldots, R_n$ are not independent random variables variables.
- The hypothesis is rejected for too small or too large values of W_n.
- If the sample size n is not too large then **under the null hypothesis** one can derive the exact distribution of W_n (numerically or with the help of already calculated tables). The critical values are tabulated.

Asymptotic distribution of the test statistic under H_0:

$$U_n = \frac{W_n - \mathsf{E}_{H_0}W_n}{\sqrt{\operatorname{var}_{H_0}(W_n)}} = \frac{W_n - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \overset{\text{as.}}{\sim} \mathsf{N}(0,1)$$

Critical region (asymptotic test):

$$H_0 \text{ is rejected } \Leftrightarrow |U_n| \geq u_{1-\alpha/2}.$$

P-value (asymptotic): $p = 2\left(1-\Phi(|u_n|)\right)$, where u_n is observed value of the test statistic U_n.

Remark. One-sample Wilcoxon test takes into consideration also the magnitude of the differences of our observations from δ_0 (not only the sign as the sign test does). It has usually a large power for testing the median then the sign test. On the other hand the disadvantage of the one-sample Wilcoxon test is that it requires the symmetric distribution of our observations.

Violations of the assumptions

Ties due to rounding. It is rather common that due to rounding there ties in the dataset. In this situation similarly as for the sign test we first give away the observations whose values are exactly equal to δ_0. The test statistic W_n is then calculated

from the remaining observations. Further because of rounding we work with average ranks. Then one can show that under the null hypothesis

$$\frac{W_n - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - cor.}} \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}(0,1),$$

where n is the (possibly reduced) sample size and $cor.$ is a correction of the variance given by*

$$cor. = \frac{1}{48}\sum_z (t_z^3 - t_z),$$

where t_z is the number how many times one observes the value z among the values $|Z_1| \ldots, |Z_n|$. The sum $\sum_z$ then indicates one sums over all possible unique values of $\{|Z_1| \ldots, |Z_n|\}$.

It is worth noting that without this variance correction $cor.$ the test would be (asymptotically) conservative.

Asymmetry. When the density f is not symmetric, then the parameter being tested is not the median of X_i but the so called *pseudo-the median* that is the median of the random variable $\frac{X_1+X_2}{2}$. The problem of the pseudo-median is that it is difficult to interpret. Generally one can also say that its value lies between the median m_X and the expected value $\mathsf{E}\, X_i$ (provided that this expectation exists).

The next unpleasant consequence of the asymmetry of our observations is that even if view the one-sample Wilcoxon test as the test of the pseudo-median than its actual/true level (exact as well as asymptotic) is different from the prescribed level α. Nevertheless the simulation experiments show that the difference of the true level from the prescribed level is not large even for rather asymmetric distributions. Thus when the data are not obviously asymmetric then the main problem of the one-sample Wilcoxon test is the interpretation of the pseudo-median.

5.5. One-sample χ^2-test about variance

It is test about variance that requires normality of observed data. Under this normality assumption is exact without this assumption is not even asymptotic.

Model: $\mathcal{F} = \left\{\mathsf{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\right\}$

The parameter being tested: variance $\sigma_X^2 = \mathsf{var}\, X_i$.

The hypothesis and the alternative:

$$H_0 : \sigma_X^2 = \sigma_0^2, \quad H_1 : \sigma_X^2 \neq \sigma_0^2,$$

where σ_0^2 is a given constant.

* See e.g. Hollander et al. (2013), p. 42.

Test statistic:

$$\frac{(n-1)S_n^2}{\sigma_0^2},$$

where S_n^2 is a sample variance (see Definition 2.4).

The exact distribution of the test statistic under H_0:

$$\frac{(n-1)S_n^2}{\sigma_0^2} \sim \chi^2_{n-1} \qquad \text{(see Theorem 2.8(i))}.$$

Critical region: The null hypothesis is rejected when the sample variance is too different from the variance assumed under the null distribution. I.e. when the test statistic is either too small or too large

$$H_0 \text{ is rejected } \Leftrightarrow \frac{(n-1)S_n^2}{\sigma_0^2} \leq \chi^2_{n-1}(\alpha/2) \text{ or } \frac{(n-1)S_n^2}{\sigma_0^2} \geq \chi^2_{n-1}(1-\alpha/2),$$

where $\chi^2_{n-1}(\alpha/2)$ and $\chi^2_{n-1}(1-\alpha/2)$ are the $\alpha/2$ a $1-\alpha/2$ quantiles of χ^2 distribution with $n-1$ degrees of freedom.

P-value: $p = 2\min\{1 - G_{n-1}(s), G_{n-1}(s)\}$, where s is the observed value of the test statistic and G_{n-1} is the cumulative distribution function of the distribution χ^2_{n-1}.

Confidence interval for σ_X^2:

$$\left(\frac{(n-1)S_n^2}{\chi^2_{n-1}(1-\alpha/2)}, \frac{(n-1)S_n^2}{\chi^2_{n-1}(\alpha/2)}\right) \qquad \text{(see (3.4))}.$$

Exercise. Show that the one-sample χ^2-test about variance is consistent.
Hint: Consider the one-sided hypothesis and alternative and note that $\frac{\chi^2_{n-1}(\beta)}{n} \xrightarrow[n\to\infty]{} 1$ *for all* $\beta \in (0,1)$.

Remark.

- When the assumption of normality is violated then this test **does not keep the level** even asymptotically. When one is afraid that the normality assumption is violated then it is more appropriate to make use of the asymptotic distribution S_n^2, see Theorem 2.6(iii).
- This test can be also considered as one-sided test

  $$\text{rejecting } H_0' : \sigma_X^2 \leq \sigma_0^2 \text{ against } H_1' : \sigma_X^2 > \sigma_0^2 \Leftrightarrow \frac{(n-1)S_n^2}{\sigma_0^2} \geq \chi^2_{n-1}(1-\alpha)$$

  Analogously the hypothesis $H_0'' : \sigma_X^2 \geq \sigma_0^2$ against the alternative $H_1'' : \sigma_X^2 < \sigma_0^2$, is rejected when the test statistic is smaller (or equal to) $\chi^2_{n-1}(\alpha)$.

5.6. Paired tests

Consider a random sample

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$$

of bivariate random vectors with the joint cumulative distribution function $F_{X,Y}$. Usually we are interested in comparing the marginal distribution F_X (of the random variable X_i) with the marginal distribution F_Y (of the random variable Y_i). The problem is that the random variables X_i and Y_i are not independent.

The main idea of the paired test is rather simple. Consider the differences $Z_i = X_i - Y_i$ and note that these differences from a random sample. Now one can proceed by using an appropriate one-sample test. Nevertheless the crucial point is to think what hypothesis is tested in the end. I.e. whether this hypothesis has some meaningful interpretation for comparing the distributions F_X and F_Y. This is sometimes true but sometimes (for instance think about the interpretation of the paired Kolmogorov-Smirnov test).

Consider for instance the one-sample test of the expected value Z_i testuje. To be more specific consider $H_0 : \mathsf{E}\, Z_i = 0$. This hypothesis hold if and only if $\mathsf{E}\, X_i = \mathsf{E}\, Y_i$. Thus the paired test is really a test of equality of expectations of X_i and Y_i.

The above might not be true for other characteristics. For instance when we are testing the the median Z_i it does not mean (in general) that we are testing the equality of the medians of X_i and Y_i. Similarly testing the variances Z_i with the one-sample test then does not provide evidence of possible differences of distributions X_i and Y_i.

The paired are typically used on the ordered pairs of the measurements of the same quantity, for instance the left eye and the right eye, the husband and the wife, before treatment and after treatment, today and one year ago, ...

The hypothesis of the null effect

In applications the random vector $(X_i, Y_i)^\mathsf{T}$ often means a measurement (called often response) before and after treatment. The null hypothesis says that the treatment has zero effect on the response, i.e.

$$H_0 : F_X(x) = F_Y(x), \forall x \in \mathbb{R} \quad H_1 : \exists x \in \mathbb{R}\ F_X(x) \neq F_Y(x), \tag{5.4}$$

where F_X and F_Y are (marginal) cumulative distribution functions of random variables X_i and Y_i.

It is important to note that each of the tests described below is designed to detect **one specific violation** of the null hypothesis (5.4).

5.7. Paired t-test

The paired t-test is performed as one-sample t-test applied to the differences Z_i.

Model: $\mathcal{F}_n = \{Z_i = X_i - Y_i \sim \mathsf{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$ or $\mathcal{F}_{as} = \{Z_i = X_i - Y_i \in \mathcal{L}_+^2\}$
Tested parameters: Expected values $\mu_X = \mathsf{E}\, X_i$ and $\mu_Y = \mathsf{E}\, Y_i$.
The hypothesis and the alternative:

$$H_0 : \mu_X - \mu_Y = \delta_0, \quad H_1 : \mu_X - \mu_Y \neq \delta_0,$$

where δ_0 is a given constant (usually $\delta_0 = 0$).
Test statistic:

$$T_n = \frac{\sqrt{n}\,(\overline{Z}_n - \delta_0)}{S_Z},$$

where $\overline{Z}_n$ is the mean of Z_i (which is equal to $\overline{X}_n - \overline{Y}_n$) and S_Z is the sample standard deviation of Z_i.

Remark. Note that

$$S_Z^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Z_i - \overline{Z}_n\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - Y_i - \overline{X}_n + \overline{Y}_n\right)^2 = S_X^2 - 2S_{X,Y} + S_Y^2,$$

where S_X^2 and S_Y^2 are the corresponding sample variances and $S_{X,Y}$ is the sample covariance. Thus we can rewrite the test statistic as

$$T_n = \frac{\overline{X}_n - \overline{Y}_n - \delta_0}{\sqrt{S_X^2/n + S_Y^2/n - 2\,S_{X,Y}/n}},$$

which resembles the test statistic of the two-sample t-test in case of equal sample sizes (see Chapter 6.2). In our situation one has in the denominator the extra term $-2\,S_{X,Y}/n$. As usually $S_{X,Y} > 0$ (as X_i and Y_i are typically positively correlated) by using the two-sample t-test on the paired problem would result in a loss of power.

Distribution of the test statistic under H_0:

$$\text{In model } \mathcal{F}_n : T_n \sim t_{n-1}, \qquad \text{In model } \mathcal{F}_{as} : T_n \overset{\text{as.}}{\sim} \mathsf{N}(0,1).$$

Similarly as for the one-sample t-test (Chapter 5.3) is this test **exact** in the "smaller" model $\mathcal{F}_n$. In the "larger" model $\mathcal{F}_{as}$ is this test **asymptotic**. Similarly this hold true also for the p-value and confidence interval which are in model $\mathcal{F}_n$ exact and in model $\mathcal{F}_{as}$ asymptotic.

Critical region:

$$H_0 \text{ is rejected } \Leftrightarrow |T_n| \geq t_{n-1}(1-\alpha/2),$$

where $t_{n-1}(1-\alpha/2)$ is $(1-\alpha/2)$ quantile of t-distribution with $n-1$ degrees of freedom.
P-value: $p = 2\left(1 - G_{n-1}(|t|)\right)$, where t is the observed value of the test statistic and G_{n-1} is the cumulative distribution function of distribution t_{n-1}.
Confidence interval pro $\mu_X - \mu_Y$: Homework exercise.

Remark. For $\delta_0 = 0$ one can view t-test also as a test of the hypothesis of the null effect (5.4). From this point of view the test will be sensitive to detect differences in the expected values (i.e. the test is consistent for the alternatives for which the expected values are different). On the other hand the test is not consistent when H_0 v (5.4) is not true but at the same time $\mathsf{E}\, Z_i = 0$. I.e. the treatment has no effect on the expected value $\mathsf{E}\, Y_i$, but it has an effect for instance on the variance $\mathsf{var}\, Y_i$.

5.8. Paired sign test

Paired sign test is performed as a one-sample sign test on the differences Z_i. Suppose that the distribution of Z_i is continuous.

Model: $\mathcal{F} = \{Z_i \text{ has an arbitrary continuous distribution}\}$

The parameter being tested: the median m_Z of the difference $Z_i = X_i - Y_i$.

The hypothesis and the alternative:

$$H_0 : m_Z = 0, \quad H_1 : m_Z \neq 0.$$

Remark.

1. In general the median Z_i cannot be expressed as the difference of the medians X_i and Y_i. Thus the test is not a test of the difference of the medians of X_i and Y_i.
2. H_0 holds if and only if $\mathsf{P}[X_i \leq Y_i] = \mathsf{P}[X_i \geq Y_i] = 1/2$, i.e. X_i is with the probability one half smaller than Y_i but also at the same time with the same probability it is smaller than Y_i. Thus from the point of view of testing the null hypothesis of the null effect (5.4) the test is consistent when the treatment effect affects the distribution of Y_i in such a way that $\mathsf{P}[X_i \geq Y_i] \neq \mathsf{P}[X_i \leq Y_i]$.
3. Generalizing the null hypothesis and the alternative to

   $$H_0 : m_Z = m_0, \quad H_1 : m_Z \neq m_0,$$

   we are in fact testing that $\mathsf{P}[X_i \leq Y_i + m_0] = \mathsf{P}[X_i \geq Y_i + m_0] = 1/2$.
4. Further if Z_i has a finite expected value and the density **symmetric** around 0, then it holds that $\mathsf{E}\, Z_i = \mathsf{E}\, X_i - \mathsf{E}\, Y_i = 0$. Under this additional assumptions H_0 is equivalent to the hypothesis of the equality of the expectations X_i and Y_i.

Test statistic:

$$B_n = \sum_{i=1}^{n} \mathbb{1}\{Z_i > 0\}, \quad \text{(i.e. the number of pairs for which } X_i > Y_i\text{)}.$$

The exact distribution of the test statistic under H_0:

$$B_n \sim \mathsf{Bi}\left(n, \tfrac{1}{2}\right)$$

Critical region (exact test): See the one-sample sign test.
Asymptotic distributions of the test statistic under H_0:

$$\frac{B_n - \frac{n}{2}}{\sqrt{\frac{n}{4}}} \overset{\text{as.}}{\sim} \mathsf{N}(0,1)$$

Critical region (asymptotic test):

$$H_0 \text{ is rejected } \Leftrightarrow \left| \frac{B_n - \frac{n}{2}}{\sqrt{\frac{n}{4}}} \right| \geq u_{1-\alpha/2}.$$

Remark. The advantage of the paired sign test is that it does not require to enumerate the difference between X_i and Y_i. It is sufficient to know whether X_i is "better" than Y_i or if X_i is "worse" than Y_i. This test is useful for applications in which it might be problematic to enumerate the values X_i and Y_i.

5.9. The paired Wilcoxon (signed-rank test) test

The paired Wilcoxon test compares the center of the symmetry δ_Z of the distribution of Z_i with a given constant.
Model: $\mathcal{F} = \{Z_i \text{ has a continuous distribution with the density } f \text{ satisfying } \exists \delta \in \mathbb{R} : f(\delta - x) = f(\delta + x) \quad \forall x \in \mathbb{R}\}$

Remark. Note that the it is sufficient that the density of Z_i is symmetric. We do not require the symmetry of the original observations X_i a Y_i. Provided that the corresponding expected values exists then the assumption of the symmetry of Z_i implies that $\delta_Z = \mathsf{E}\, Z_i = \mathsf{E}\, X_i - \mathsf{E}\, Y_i$.

Tested parameter: the center of the symmetry δ_Z
The hypothesis and the alternative:

$$H_0 : \delta_Z = \delta_0, \quad H_1 : \delta_Z \neq \delta_0,$$

where δ_0 is a given constant (usually $\delta_0 = 0$).
Test statistic:

$$W_n = \sum_{i \in \mathcal{I}} R_i,$$

where $\mathcal{I} \subset \{1, \ldots, n\}$ is a set of indices such that $Z_i^* \overset{\text{df}}{=} X_i - Y_i - \delta_0$ is positive for $i \in \mathcal{I}$ and R_i is the rank of the random variable $|Z_i^*|$ among the all variables $|Z_1^*|, \ldots, |Z_n^*|$.
Properties of the test statistic and critical region: see the one-sample (signed-rank) Wilcoxon test.

Remark.

1. The paired Wilcoxon test can be interpreted as the test of expected values. Nevertheless the paired t-test is usually more appropriate for testing the equality of expected values as it does not require the symmetry of the difference Z_i.
2. For $\delta_0 = 0$ we can consider this test as test of the hypothesis of the null effect (5.4). In this situation it is common to assume that under the null hypothesis the joint distribution of the random vector $(X_i, Y_i)^\mathsf{T}$ is the same as the joint distribution of $(Y_i, X_i)^\mathsf{T}$. Under this additional assumption one can conclude that under the null hypothesis the distribution of the random variable $Z_i = X_i - Y_i$ is **symmetric** around zero. Thus the test will be hold the prescribed level. But it is important to realize that the test will be consistent against the alternatives for which the pseudo-the median Z_i (i.e. the median of $\frac{Z_1+Z_2}{2}$) different from zero. Thus the test is consistent against the alternatives for which

$$\mathsf{P}[Z_1 + Z_2 \leq 0] \neq \mathsf{P}[Z_1 + Z_2 \geq 0].$$

Sample examples for the preparation for the exam.

The solution of "the practical exercises" should contain the mathematical model, the null and the alternative hypothesis, the test statistic and its (either exact or asymptotic) distribution under the null hypothesis, critical region and the formula to calculate the p-value. It should be also explicitly stated if the test is exact or asymptotic.

1. It is know that the distribution of IQ in the overall population has the standard deviation equal to 15. We managed to get the values of IQ for 158 randomly selected members of a given party. Suggest a test (i.e. give the appropriate model suitable for your data, the null and the alternative hypothesis, test statistic, critical region and the formula to calculate the p-value) that aims at showing that the members of this party is a more homogeneous group in comparison to the overall population.

2. Suppose that data on gross salary of 300 randomly chosen graduates of study programe Probability, Mathematical Statistics and Econometrics. Suggest a test to prove that at least 75 % of graduates gets a gross monthly salary higher than 40 000 CZK.

3. Suppose that you know the gross monthly salaries of 500 randomly chosen employers of the given insurance company. For each of this employer we know the entry salary and the salary after two years working for the company. Suggest tests aiming to prove that during the first two years of the working for the company:
   (a) the expected increase in the salary is larger then 15 000 CZK;
   (b) with the probability at least 90 % the salary increases by at least 10 000 CZK.

   Do you think that with one dataset it is possible to prove both statements?

4. Suppose that we know the body heights of 300 randomly chosen of female students of Charles University. Further it is said that the average height of the adult women in the Czech Republic is 168 cm. We would like to show the female students of Charles University are in some sense higher than what is common in the overall population of women. Suggest an appropriate test and explain what would be proved by rejecting the null hypothesis.

5. The following table contains the number of points that 10 randomly chosen employees get from the English test before and after intensive English course.

Employer	1	2	3	4	5	6	7	8	9	10
Before the course	37	41	36	48	42	36	42	44	40	34
After the course	38	43	43	47	52	44	41	42	42	39

   Suggest a test to prove that the language test improves the language skills of the employees.

6. We are interested in finding out if the spreadsheet software has a good generator of random numbers from the uniform distribution $U(0,1)$. To do that we generated a sample of 1 000 random numbers. Suggest a test to find out if the generator is a good one.

7. Try to think whether it makes sense to consider the paired Kolmogorov-Smirnov test.

The end of self-study for week 9 (1.12.-5.12.).

6. Two-sample problems for quantitative data

Consider two *independent* random samples: let $X_1, \ldots, X_n$ be a random sample with distribution function F_X and $Y_1, \ldots, Y_m$ a random sample with distribution function F_Y. Model $\mathcal{F}$ specifies the set of considered distribution functions F_X and F_Y. We are given a parameter $\theta = t(F)$ and we would like to compare its value for both samples. Denote $\theta_X = t(F_X)$ and $\theta_Y = t(F_Y)$. Usually, we want to test the null hypothesis $H_0 : \theta_X = \theta_Y$ against the alternative $H_1 : \theta_X \neq \theta_Y$; eventually we want to construct an interval estimate for the difference $\theta_X - \theta_Y$.

The two-sample problem can be also formulated in another way. Let us have a random sample from bivariate distribution

$$\begin{pmatrix} Z_1 \\ I_1 \end{pmatrix}, \ldots, \begin{pmatrix} Z_N \\ I_N \end{pmatrix},$$

where Z_j are independent identically distributed random variables and I_j has alternative distribution with parameter $p_G \in (0,1)$. Indicator I_j determines the group of jth observation (if $I_j = 0$, then the jth observation belongs to the first group, otherwise to the second group). If we now denote the variable Z_j by X_i or Y_i based on the group it belongs to, i.e.

$$(X_1, \ldots, X_n) \stackrel{\mathrm{df}}{=} (Z_j : I_j = 0) \quad \text{and} \quad (Y_1, \ldots, Y_m) \stackrel{\mathrm{df}}{=} (Z_j : I_j = 1),$$

we get two independent random samples as in the first formulation of the problem. We would like to compare the conditional distribution of Z_j in both groups, i.e. we are interested in the conditional distribution functions $F_X(x) = \mathsf{P}\left[Z_j \leq x \,\middle|\, I_j = 0\right]$ and $F_Y(x) = \mathsf{P}\left[Z_j \leq x \,\middle|\, I_j = 1\right]$, respectively their parameters $\theta_X = t(F_X)$ and $\theta_Y = t(F_Y)$. This second formulation of the two-sample problem is the same as the first formulation with one exception - the sizes of random samples n and m are not constants, but they are random variables with binomial distribution $\left(n = \sum_{j=1}^{N}(1 - I_j) \sim \mathsf{Bi}(N, 1 - p_I)\right.$, where $\left.p_I = \mathsf{P}(I_j = 1)\right)$. **However, the analysis of our data is performed in the same way as for constant sizes of random samples**.

Data corresponding to the first formulation are obtained by determining in advance the number of observations in each group and afterwards observing the required number of values for each group separately. Data corresponding to the second formulation are obtained if we determine the total number of observations $N = n+m$, then obtain these N observations and afterwards decide for each observation the group it belongs to.

Both formulations differ a little bit in the concept of asymptotic results. With the second formulation, we only need that $N \to \infty$. For the first formulation, we need that $n \to \infty$ and $m \to \infty$ and we also have to assume that the speed of the convergence is the same for both sample sizes, i.e. that $n/m \to q$, where $0 < q < \infty$.

All methods presented in this chapter can be used for both formulations of the two-sample problem.

6.1. Two-sample Kolmogorov-Smirnov test

Two-sample Kolmogorov-Smirnov test is an extension of the one-sample test with the same name. It is a non-parametric test that can be used for any pair of continuous distributions.

Model: $\mathcal{F}$ = {all continuous distributions}

Tested parameters: distribution functions F_X and F_Y

Null hypothesis and alternative:

$$H_0 : F_X(x) = F_Y(x) \quad \forall x \in \mathbb{R}, \qquad H_1 : \exists x \in \mathbb{R} : F_X(x) \neq F_Y(x). \tag{6.1}$$

We test whether both random samples come from the same distribution. This hypothesis will be from now on called the **null-difference hypothesis**.

Test statistic:

$$K_{n,m} = \sup_{x \in \mathbb{R}} \left| \widehat{F}_X(x) - \widehat{F}_Y(x) \right|,$$

where $\widehat{F}_X$ is the empirical distribution function of the random sample $X_1, \ldots, X_n$ and $\widehat{F}_Y$ is the empirical distribution function of the random sample $Y_1, \ldots, Y_m$.

Proposition 6.1 Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be two independent random samples from continuous distribution with distribution function F_0. Then

$$\sqrt{\tfrac{nm}{n+m}}\, K_{n,m} \xrightarrow{\mathsf{d}} Z, \text{ for } m, n \to \infty,$$

where the random variable Z has a distribution function given by the formula (5.1).

Remark.

- We reject the null hypothesis if empirical distribution functions of both samples differ too much from each other, i.e. for large values of our test statistic.
- Proposition 6.1 implies that, under the null hypothesis, $\sqrt{\frac{nm}{n+m}} K_{n,m}$ converges in distribution to a random variable with distribution function $G(y)$, which is the same as for one-sample Kolmogorov-Smirnov test (see Proposition 5.2). The important thing is that this distribution function does not depend on the real (for both samples) distribution function F_0. This enables us to determine critical value for rejecting H_0.

Critical region:

$$H_0 \text{ is rejected } \Leftrightarrow \sqrt{\tfrac{nm}{n+m}}\, K_{n,m} \geq k_{1-\alpha}, \tag{6.2}$$

where $k_{1-\alpha} = G^{-1}(1-\alpha)$ is $(1-\alpha)$-quantile of the distribution with distribution function G.

According to Proposition 6.1, this test has asymptotic significance level α.

Remark.

- It is possible to compute the exact critical value for two-sample Kolmogorov-Smirnov test for continuous distributions and small sample sizes n, m.
- Notice that under the alternative for $m, n \to \infty$,

$$K_{n,m} \xrightarrow{\mathsf{P}} \sup_{x\in\mathbb{R}} \left|F_X(x) - F_Y(x)\right| > 0 \Longrightarrow \sqrt{\tfrac{nm}{n+m}}\, K_{n,m} \xrightarrow{\mathsf{P}} \infty.$$

  In particular, this test is consistent against any alternative. In other words, the test reacts to any difference in distributions of both samples. Another advantage of this test is the absence of restrictive assumptions. The disadvantage of this test is that its power is small against specific violations of hypothesis H_0. If we are interested (or we expect) only a specific type of violation of H_0 (for example difference of expected values), it is better to use a test which is focused on a specific parameter.
- It is worth noticing that the test statistic does not change if, at first, we transform all observations by some injective function g. It can be shown that two-sample Kolmogorov-Smirnov test can be reformulated as a *rank test*.

Violation of assumptions

If the samples come, under the null hypothesis, from discrete distribution (i.e. F_0 from Proposition 6.1 is not continuous), then the test with critical region (6.2) will be conservative. Similarly if the "discreetness" arises from rounding. In this case however, it is necessary to assume that the rounding is performed in the same way for both samples.

6.2. Two-sample t-test without the assumption of equality of variances

Two-sample t-test compares the **expected values** of both samples. The execution of this test differs based on whether we assume (see Chapter 6.3) or do not assume the equality of variances.

Model:

$$\mathcal{F} = \left\{F_X \in \mathcal{L}^2_+, F_Y \in \mathcal{L}^2_+\right\}.$$

Tested parameters: Expected values $\mu_X = \mathsf{E}\, X_i$ and $\mu_Y = \mathsf{E}\, Y_i$.

Null hypothesis and alternative:

$$H_0 : \mu_X = \mu_Y + \delta_0, \quad H_1 : \mu_X \neq \mu_Y + \delta_0. \tag{6.3}$$

We test whether expected values differ by δ_0 (usually we choose $\delta_0 = 0$).
Test statistic:

$$\widetilde{T}_{n,m} = \frac{\overline{X}_n - \overline{Y}_m - \delta_0}{\sqrt{S_X^2/n + S_Y^2/m}},$$

where $\overline{X}_n$, $\overline{Y}_m$ are sample means and S_X^2, S_Y^2 are sample variances of the two samples.

Remark. The test statistic $\widetilde{T}_{n,m}$ can be remembered with the help of the following observation. Notice that

$$\mathrm{var}\left(\overline{X}_n - \overline{Y}_m\right) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}.$$

Natural (and even unbiased) estimate of this variance is $S_X^2/n + S_Y^2/m$.

Theorem 6.2 Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be two independent random samples from distributions with expected values μ_X and μ_Y and finite variances. Then

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sqrt{S_X^2/n + S_Y^2/m}} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1) \text{ for } m, n \to \infty, \frac{n}{m} \to q \in (0, \infty).$$

Proof. We can rewrite

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sqrt{S_X^2/n + S_Y^2/m}} = \frac{\sqrt{m}\left(\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)\right)}{\sqrt{S_X^2 \frac{m}{n} + S_Y^2}}.$$

From the consistency of sample variance we have $S_X^2 \xrightarrow{\mathrm{P}} \sigma_X^2$, $S_Y^2 \xrightarrow{\mathrm{P}} \sigma_Y^2$ and therefore we get, with the help of the continuous mapping theorem (Proposition 1.2), that $\sqrt{S_X^2 \frac{m}{n} + S_Y^2} \xrightarrow{\mathrm{P}} \sqrt{\sigma_X^2/q + \sigma_Y^2}$. So, if we take into account the Cramér-Slutsky theorem (Proposition 1.3), it is enough to show that

$$\sqrt{m}\left(\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)\right) \xrightarrow{\mathrm{d}} \mathrm{N}\left(0, \sigma_X^2/q + \sigma_Y^2\right). \tag{6.4}$$

From the central limit theorem we get that $\sqrt{n}\left(\overline{X}_n - \mu_X\right) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \sigma_X^2)$ and therefore

$$\sqrt{m}\left(\overline{X}_n - \mu_X\right) = \sqrt{\frac{m}{n}}\sqrt{n}\left(\overline{X}_n - \mu_X\right) \xrightarrow{\mathrm{d}} \mathrm{N}\left(0, \sigma_X^2/q\right), \tag{6.5}$$

since from the assumptions of the theorem we have $\sqrt{\frac{m}{n}} \to \frac{1}{\sqrt{q}}$. Furthermore, also thanks to the central limit theorem, we have

$$\sqrt{m}\left(\overline{Y}_m - \mu_Y\right) \xrightarrow{\mathrm{d}} \mathrm{N}\left(0, \sigma_Y^2\right). \tag{6.6}$$

Now, using (6.5), (6.6) and the **independence** of $\overline{X}_n$ and $\overline{Y}_m$, we get

$$\sqrt{m}\begin{pmatrix}\overline{X}_n - \mu_X\\ \overline{Y}_m - \mu_Y\end{pmatrix} \xrightarrow{\mathsf{d}} \mathsf{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_X^2/q & 0\\ 0 & \sigma_Y^2\end{pmatrix}\right).$$

Therefore, also for all $\boldsymbol{c} \in \mathbb{R}^2$

$$\boldsymbol{c}^\mathsf{T}\sqrt{m}\begin{pmatrix}\overline{X}_n - \mu_X\\ \overline{Y}_m - \mu_Y\end{pmatrix} \xrightarrow{\mathsf{d}} \mathsf{N}(0, \boldsymbol{c}^\mathsf{T}\Sigma\boldsymbol{c}), \quad \text{where} \quad \Sigma = \begin{pmatrix}\sigma_X^2/q & 0\\ 0 & \sigma_Y^2\end{pmatrix}.$$

The convergence in (6.4) now follows from the above stated convergence for $\boldsymbol{c} = (1,-1)^\mathsf{T}$, since in that case

$$c^\mathsf{T}\Sigma\boldsymbol{c} = \sigma_X^2/q + \sigma_Y^2.$$

□

Remark.

- We will reject the hypothesis if the sample means of both random samples differ too much, i.e. if the test statistic if too large or too small.
- Theorem 6.2 implies that in model $\mathcal{F}$ and under the null hypothesis H_0 the test statistic $\widetilde{T}_{n,m}$ has asymptotic distribution $\mathsf{N}(0,1)$.

Critical region:

$$H_0 \text{ is rejected } \Leftrightarrow \left|\widetilde{T}_{n,m}\right| \geq u_{1-\alpha/2},$$

where $u_{1-\alpha/2}$ is $(1-\alpha/2)$-quantile of standard normal distribution.

P-value: $p = 2\left(1 - \Phi(|t|)\right)$, where t is the observed value of the test statistic $\widetilde{T}_{n,m}$ and Φ is the distribution function of $\mathsf{N}(0,1)$.

Confidence interval for $\mu_X - \mu_Y$: It is possible to derive an asymptotic confidence interval for the difference between expected values of both samples from Theorem 6.2. For $n, m \to \infty$ we get

$$\mathsf{P}\left[\overline{X}_n - \overline{Y}_m - u_{1-\alpha/2}\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}} < \mu_X - \mu_Y < \overline{X}_n - \overline{Y}_m + u_{1-\alpha/2}\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}\right] \to 1 - \alpha.$$

It turns out that even if we added the assumption of normality of our observations, i.e. $X_i \sim \mathsf{N}(\mu_X, \sigma_X^2)$ and $Y_i \sim \mathsf{N}(\mu_Y, \sigma_Y^2)$, the distribution of our test statistic $\widetilde{T}_{n,m}$ under the null hypothesis would still be pivotal only asymptotically. The exact distribution of the test statistic $\widetilde{T}_{n,m}$, even with the assumption of normality, will depend on the ratio σ_X^2/σ_Y^2. That's why we are content to use the asymptotic test in practical problems.*

* The problem of performing an exact test is known as Behrens-Fisher problem.

Remark. There exists a better approximation of critical values for this test, which is based on the t-distribution with number of degrees of freedom depending on the number of observations in both groups and on sample variances. There exists several of these approximations. One of these approximations, so called *Welch test*, is implemented in the software environment R as a standard method for testing equality of expected values of two samples (it is performed by function `t.test`). For this approximation, quantiles of t-distribution with f degrees of freedom are used as critical values, where f is given by the formula

$$f = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{(S_X^2)^2}{n^2(n-1)} + \frac{(S_Y^2)^2}{m^2(m-1)}}.$$

This formula was derived under the assumption of normality and is based on the approximation of the distribution of random variable $\frac{S_X^2}{n} + \frac{S_Y^2}{m}$ from the denominator of the test statistic, using a multiple of χ^2-distribution with "appropriate" degrees of freedom (details can be found in Welch, 1938).

Welch test can be understood as a variant of the two-sample t-test (without the assumption of equal variances) with improved critical values.

P-value of Welch t-test for the two-sided alternative (6.3) can be calculated using the formula $p = 2(1 - G_f(|t|))$, where t is the observed value of the test statistic $\widetilde{T}_{n,m}$ and G_f is the distribution function of t-distribution with f degrees of freedom.

6.3. Two-sample t-test with the assumption of equal variances

Similarly as in the case of one-sample t-test (see Chapter 5.3) we will derive an exact test under the assumption of normality and asymptotic test without this assumption.

Model:

$$\mathcal{F}_n = \left\{F_X = \mathsf{N}(\mu_X, \sigma^2), F_Y = \mathsf{N}(\mu_Y, \sigma^2), \mu_X, \mu_Y \in \mathbb{R}, \sigma^2 > 0\right\}$$

or

$$\mathcal{F}_{as} = \left\{F_X \in \mathcal{L}_+^2, F_Y \in \mathcal{L}_+^2, \text{ where } \mathsf{var}\,(X_i) = \mathsf{var}\,(Y_i) := \sigma^2\right\}.$$

In model $\mathcal{F}_n$ both random samples have Gaussian distribution with the same variance σ^2, i.e. they can differ only in the mean value. In model $\mathcal{F}_{as}$ it is only required that the variances are the same (i.e. the distributions can be different).

Tested parameters: Expected values $\mu_X = \mathsf{E}\,X_i$ and $\mu_Y = \mathsf{E}\,Y_j$.

Null hypothesis and alternative:

$$H_0 : \mu_X = \mu_Y + \delta_0, \quad H_1 : \mu_X \neq \mu_Y + \delta_0.$$

We test whether the expected values of our samples differ by δ_0 (usually $\delta_0 = 0$).

Test statistic:

$$T_{n,m} = \frac{\overline{X}_n - \overline{Y}_m - \delta_0}{\sqrt{S^2_{n,m}\left(\frac{1}{n} + \frac{1}{m}\right)}} = \sqrt{\frac{nm}{n+m}}\,\frac{\overline{X}_n - \overline{Y}_m - \delta_0}{S_{n,m}},$$

where $\overline{X}_n$ and $\overline{Y}_m$ are sample means of both samples and

$$S^2_{n,m} \overset{\mathrm{df}}{=} \frac{1}{n+m-2}\left[\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 + \sum_{j=1}^{m}(Y_j - \overline{Y}_m)^2\right] = \frac{n-1}{n+m-2}S^2_X + \frac{m-1}{n+m-2}S^2_Y$$

is the unbiased estimate of the common variance σ^2 calculated from both samples (weighted average of sample variances S^2_X and S^2_Y).

Remark. Test statistic $T_{n,m}$ can be remembered with the help of the following observation. Notice that

$$\mathsf{var}\left(\overline{X}_n - \overline{Y}_m\right) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right).$$

Since $S^2_{n,m}$ is (unbiased) estimate of σ^2, we have that $S^2_{n,m}\left(\frac{1}{n} + \frac{1}{m}\right)$ is natural (and even unbiased) estimate of $\sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)$.

Theorem 6.3 Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be independent random samples from distributions with expected values μ_X and μ_Y and finite variances $\sigma^2_X = \mathsf{var}\,(X_i)$ and $\sigma^2_Y = \mathsf{var}\,(Y_j)$. Then

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sqrt{S^2_{n,m}\left(\frac{1}{n} + \frac{1}{m}\right)}} \xrightarrow{\mathsf{d}} \mathsf{N}(0, \sigma^2_*), \text{ for } m, n \to \infty, \frac{n}{n+m} \to \lambda \in (0,1),$$

where

$$\sigma^2_* = \frac{(1-\lambda)\sigma^2_X + \lambda\sigma^2_Y}{\lambda\sigma^2_X + (1-\lambda)\sigma^2_Y}.$$

Proof. Proof is analogous to the proof of Theorem 6.2. At first we rewrite

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sqrt{S^2_{n,m}\left(\frac{1}{n} + \frac{1}{m}\right)}} = \frac{\sqrt{m}\left(\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)\right)}{\sqrt{S^2_{n,m}\,\frac{m+n}{n}}}.$$

Now we can show, similarly as in Theorem 6.2, that

$$\sqrt{m}\left(\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)\right) \xrightarrow{\mathsf{d}} \mathsf{N}\left(0, \tfrac{1-\lambda}{\lambda}\sigma^2_X + \sigma^2_Y\right),$$

where we have used the fact that $\frac{1}{q} = \frac{1-\lambda}{\lambda}$. Then we show that

$$\sqrt{S^2_{n,m}\,\frac{m+n}{n}} \xrightarrow[n\to\infty]{\mathsf{P}} \sqrt{\frac{\lambda\sigma^2_X + (1-\lambda)\sigma^2_Y}{\lambda}}.$$

□

Notice that under the assumption of equal variances, i.e. $\sigma_X^2 = \sigma_Y^2$ (i.e. model $\mathcal{F}_{as}$ holds), we have that $\sigma_*^2 = 1$ and so

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sqrt{S_{n,m}^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \xrightarrow{\mathsf{d}} \mathsf{N}(0,1), \text{ for } m, n \to \infty, \frac{n}{n+m} \to \lambda \in (0,1).$$

Furthermore, if we can add the assumption of *normality* (i.e. model $\mathcal{F}_n$ holds), we can derive the exact distribution of random variable $\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sqrt{S_{n,m}^2\left(\frac{1}{n} + \frac{1}{m}\right)}}$.

Theorem 6.4 Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be two independent random samples from **Gaussian** distributions with expected values μ_X and μ_Y and with **the same variance** σ^2. Then

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sqrt{S_{n,m}^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim t_{n+m-2}.$$

Proof. Rewrite

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sqrt{S_{n,m}^2\left(\frac{1}{n} + \frac{1}{m}\right)}} = \frac{U}{\sqrt{Z/(n+m-2)}},$$

where

$$U = \frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \quad \text{and} \quad Z = \frac{(n+m-2)\, S_{n,m}^2}{\sigma^2}.$$

To complete the proof, it is enough to show that $U \sim \mathsf{N}(0,1)$, $Z \sim \chi^2_{n+m-2}$ and that U is independent with Z.

1. $U \sim \mathsf{N}(0,1)$. To show this part, it is enough to realize that, because to the independence of random samples, sample means $\overline{X}_n$ and $\overline{Y}_m$ are also independent and it holds that

$$\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y) \sim \mathsf{N}\big(0, \tfrac{\sigma^2}{n} + \tfrac{\sigma^2}{m}\big).$$

So

$$U = \frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim \mathsf{N}(0,1).$$

2. $Z \sim \chi^2_{n+m-2}$. Z can be written, using S_X^2 and S_Y^2, as

$$Z = \frac{(n+m-2)\, S_{n,m}^2}{\sigma^2} = \frac{(n-1)\, S_X^2}{\sigma^2} + \frac{(m-1)\, S_Y^2}{\sigma^2}.$$

Now thanks to Theorem 2.8(i), we get that $\frac{(n-1)\, S_X^2}{\sigma^2} \sim \chi^2_{n-1}$ and $\frac{(m-1)\, S_Y^2}{\sigma^2} \sim \chi^2_{m-1}$.

Furthermore, from the independence of S_X^2 and S_Y^2 we get that the distribution of Z is the same as the distribution of a sum of two **independent** random variables with

χ^2_{n-1} and χ^2_{m-1} distributions. Now considering the definition of χ^2-distribution (distribution of the sum of squares of independent, $\mathsf{N}(0,1)$ distributed variables) we get that $Z \sim \chi^2_{n+m-2}$.

3. *Independence of U and Z.* Because of the independence of the random samples, random vectors $(\overline{X}_n, S_X^2)^\mathsf{T}$ and $(\overline{Y}_m, S_Y^2)^\mathsf{T}$ are also independent. Furthermore, from Theorem 2.8(ii) we get that random variables $\overline{X}_n$ and S_X^2 are independent and similarly random variables $\overline{Y}_m$ and S_Y^2 are also independent. Therefore, the random variables $\overline{X}_n - \overline{Y}_m$ and $S^2_{n,m}$ are independent. This implies the independence of U and Z. □

Remark.

- While Theorem 6.3 implies that, in model $\mathcal{F}_{as}$, $T_{n,m}$ has asymptotic distribution $\mathsf{N}(0,1)$, Theorem 6.4 tells us that in smaller model $\mathcal{F}_n$ it holds that $T_{n,m}$ has, under H_0, exact distribution t_{n+m-2}.
- The null hypothesis will be rejected if sample means of both samples differ too much from each other, i.e. the test statistic is too large or too small.

Critical region:

$$H_0 \text{ is rejected } \Leftrightarrow \left|T_{n,m}\right| \geq t_{n+m-2}(1-\alpha/2),$$

where $t_{n+m-2}(1-\alpha/2)$ is $(1-\alpha/2)$-quantile of t-distribution with $n+m-2$ degrees of freedom.

Similarly as in the one-sample t-test, the above described test is exact in model $\mathcal{F}_n$ and asymptotic in model $\mathcal{F}_{as}$. The same holds for the following p-value and confidence interval.

P-value: $p = 2\left(1 - F(|t|)\right)$, where t is the observed value of test statistic $T_{n,m}$ and F is the distribution function of t_{n+m-2}-distribution.

Confidence interval for $\mu_X - \mu_Y$: Using Theorem 6.4 (resp. Theorem 6.3), it is possible to derive an exact (resp. asymptotic) confidence interval for the difference of expected values of both samples. We get

$$\mathsf{P}\left[\overline{X}_n - \overline{Y}_m - t_{n+m-2}(1-\alpha/2)\, S_{n,m}\sqrt{\tfrac{1}{n}+\tfrac{1}{m}} < \mu_X - \mu_Y < \right.$$
$$\left. \overline{X}_n - \overline{Y}_m + t_{n+m-2}(1-\alpha/2)\, S_{n,m}\sqrt{\tfrac{1}{n}+\tfrac{1}{m}}\right] = 1-\alpha.$$

Exercise. Modify the critical region and the formula for p-value for the test of the null hypothesis $H_0 : \mu_X \leq \mu_Y + \delta_0$ against the alternative $H_1 : \mu_X > \mu_Y + \delta_0$.

Violation of the assumption of equal variances

According to Theorem 6.3

$$T_{n,m} \xrightarrow{\mathsf{d}} \mathsf{N}(0, \sigma_*^2), \quad \text{where} \quad \sigma_*^2 = \frac{(1-\lambda)\sigma_X^2 + \lambda\sigma_Y^2}{\lambda\sigma_X^2 + (1-\lambda)\sigma_Y^2} \quad \text{and} \quad \tfrac{n}{n+m} \to \lambda \in (0,1).$$

Therefore, the test generally does not keep the required level even asymptotically. It is also worth noticing that if we have, for example, $\sigma_X^2 > \sigma_Y^2$ and at the same time $\lambda < \frac{1}{2}$ (i.e. larger variance is in the sample with smaller sample size), then $\sigma_*^2 > 1$ and the test is asymptotically liberal.

Notice also that for $\lambda = \frac{1}{2}$ we have $\sigma_*^2 = 1$. So, for samples with roughly the same size, it still holds that

$$T_{n,m} \xrightarrow{\mathsf{d}} \mathsf{N}(0,1)$$

and our test does keep the level asymptotically.

Furthermore, if the sizes of both samples equal, i.e. $m = n$, then

$$\sqrt{S_X^2/n + S_Y^2/m} = \sqrt{\left(S_X^2/2 + S_Y^2/2\right)\frac{2}{n}} = \sqrt{S_{n,m}^2\left(\frac{1}{n} + \frac{1}{m}\right)}\,.$$

In this case, it always holds that $T_{n,n} = \widetilde{T}_{n,n}$, i.e. test statistics of two-sample t-test is the same with or without the assumption of equal variances.

t-TEST AS A TEST OF THE NULL-DIFFERENCE HYPOTHESIS

If we take $\delta_0 = 0$, this test can be understood as a test of the null-difference hypothesis (6.1). Even though we do not have the assumption of normality, we have equal variances under the null hypothesis. Therefore, the test will keep the required level asymptotically.

Regarding the power of the test, it will be consistent against the alternative for which we have $\mu_X - \mu_Y \neq 0$. However, if the distributions F_X and F_Y differ not only in expected values, but also in variances, we do not control the influence of this difference. It can both increase and decrease the power of our test. Furthermore, if we reject the null hypothesis (6.1), we can only claim that we have proven difference of distributions F_X and F_Y. The rejection of the null hypothesis cannot be attributed only to the difference in expected values, since the difference in variances could also contribute to this result.

Exercise. Prove (in detail) Theorem 6.3.

Remark. Sometimes it is recommended to test the equality of variances of our samples before using the two-sample t-test; this can be done for example by using the test from Chapter 6.5 or so called Levene's test (not presented in these lecture notes). If the equality of variances is rejected, we use Welch test, otherwise we use two-sample t-test. However, we advise against using this kind of approach. It is so called *two-phase test*, where the result depends on three different test statistics that are not independent. It is not guaranteed that the significance level of this test is equal to the required level α. If we are not sure about the assumption of normality or equal variances, we should use the Welch test. Then we do not have to verify either one of the assumptions of the two-sample t-test.

6.4. Two-sample Wilcoxon test

The two-sample Wilcoxon test (also called the Wilcoxon rank-sum test) is a non-parametric test based on ranks.

Model: $\mathcal{F} = \big\{\exists$ increasing function g and $\exists \delta \in \mathbb{R}$:

$$g(X_i) \sim \widetilde{F}_X \textbf{ continuous } \text{d.f.}, g(Y_j) \sim \widetilde{F}_Y, \widetilde{F}_X(x) = \widetilde{F}_Y(x-\delta)\ \forall x \in \mathbb{R}\big\}. \quad (6.7)$$

Tested parameter: Shift δ_{XY}.

Null hypothesis and alternative:

$$H_0 : \delta_{XY} = 0, \quad H_1 : \delta_{XY} \neq 0.$$

If we have $g(x) = x$, then model $\mathcal{F}$ is called *location model*. So model $\mathcal{F}$ will be called *generalized location model*.

Remark.

- Unlike in one-sample and paired Wilcoxon test, we **do not require symmetry** of any density.
- If both model $\mathcal{F}$ and hypothesis H_0 hold, then the distributions of X and Y are **identical**. Then it holds that $m_X = m_Y$ and $\mathsf{E}\,X = \mathsf{E}\,Y$ (if the expected values exist). In other words, if model $\mathcal{F}$ holds, then two-sample Wilcoxon test can be understood as a test of equality of expected values and medians. Usually, the two-sample Wilcoxon test is considered as a test of the equality of medians.

Test statistic:

$$W_{n,m} = \sum_{i=1}^{n} R_i,$$

where $R_1, R_2, \ldots, R_n$ are ranks of random variables X_i in the combined random sample $X_1, \ldots, X_n, Y_1, \ldots, Y_m$.

Remark. Test statistic $W_{n,m}$ can attain values from the set $\big\{\frac{n(n+1)}{2}, \ldots, mn + \frac{n(n+1)}{2}\big\}$. It can be computed in the following way:

1. Take combined random sample $(Z_1, \ldots, Z_{n+m}) \overset{\text{df}}{=} (X_1, \ldots, X_n, Y_1, \ldots, Y_m)$.
2. Order all Z_j from smallest to largest to get the ordered random sample

$$Z_{(1)} < Z_{(2)} < \cdots < Z_{(n+m)}.$$

3. Determine ranks R_i of random variables X_i between all $Z_{(1)}, \ldots, Z_{(n+m)}$. It holds that $X_i = Z_{(R_i)}$.
4. Sum ranks R_i for $i = 1, \ldots, n$.

It is possible to find an *exact distribution* of the test statistic $W_{n,m}$ under the null hypothesis for small values of n and m (numerically or in tables). This exact distribution can be derived from the fact that **under the null hypothesis** any order of random variables $Z_1, \ldots, Z_{n+m}$ has the same probability (see Theorem 2.15) and therefore

$$\mathsf{P}\big(R_1 = r_1, \ldots, R_n = r_n\big) = \frac{m!}{(n+m)!}$$

for all $r_1, \ldots, r_n \in \{1, \ldots, n+m\}$ different.

For large values of n and m the following proposition is used.

Proposition 6.5 Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be two independent random samples from model $\mathcal{F}$. Suppose that the **null hypothesis** H_0 holds, then

(i)
$$\mathsf{E}_{H_0} W_{n,m} = \frac{n(n+m+1)}{2}, \quad \mathsf{var}_{H_0}(W_{n,m}) = \frac{mn(n+m+1)}{12}.$$

(ii) If $n, m \to \infty$, then
$$\frac{W_{n,m} - \mathsf{E}_{H_0} W_{n,m}}{\sqrt{\mathsf{var}_{H_0}(W_{n,m})}} \xrightarrow{\mathsf{d}} \mathsf{N}(0,1).$$

Proof. Part (i). Under the null hypothesis, distributions of X_i and Y_j are the same, so the combined sample $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ is a random sample of size $n+m$. It follows from Theorem 2.16 that for $i \neq j$:

$$\mathsf{E}\, R_i = \frac{n+m+1}{2}, \quad \mathsf{var}\,(R_i) = \frac{(n+m)^2 - 1}{12}, \quad \mathsf{cov}\,(R_i, R_j) = -\frac{n+m+1}{12}.$$

So

$$\mathsf{E}_{H_0} W_{n,m} = \sum_{i=1}^{n} \mathsf{E}\, R_i = \frac{n\,(n+m+1)}{2}$$

and

$$\begin{aligned}
\mathsf{var}_{H_0}(W_{n,m}) &= \sum_{i=1}^{n} \mathsf{var}\,(R_i) + \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \mathsf{cov}\,(R_i, R_j) \\
&= n\,\frac{(n+m+1)(n+m-1)}{12} - n(n-1)\frac{n+m+1}{12} \\
&= \frac{n(n+m+1)}{12}\big[n+m-1-(n-1)\big] = \frac{nm(n+m+1)}{12}.
\end{aligned}$$

Part (ii). Will not be proven. The difficulty of this proof lies in the fact that the ranks $R_1, \ldots, R_n$ are not independent random variables. □

Remark.

- Hypothesis will be rejected for too large or too small values of $W_{n,m}$.

- The previous proposition gives us instructions for finding critical values that ensure asymptotic significance level α.

Critical region (asymptotic test):

$$H_0 \text{ is rejected } \Leftrightarrow \frac{\left|W_{n,m} - \frac{n(m+n+1)}{2}\right|}{\sqrt{\frac{mn(m+n+1)}{12}}} \geq u_{1-\alpha/2}.$$

Violation of assumptions

Ties due to rounding. Because of rounding, we often see ties in our data. In that situation, test statistic $W_{n,m}$ can be computed using so called average ranks. It can be shown that under the null hypothesis

$$\frac{W_{n,m} - \frac{n(n+m+1)}{2}}{\sqrt{\frac{mn(n+m+1-kor.)}{12}}} \xrightarrow{\mathsf{d}} \mathsf{N}(0,1), \text{ for } n, m \to \infty,$$

where $kor.$ is a correction of variance given by formula*

$$kor. = \frac{1}{(n+m)(n+m-1)} \sum_z (t_z^3 - t_z),$$

where t_z denotes the number of the random variables $Z_1 \ldots, Z_{n+m}$ which attain the value z. By $\sum_z$ we denote the sum over all different values from the set $\{Z_1 \ldots, Z_{n+m}\}$.

It is worth noticing that without the use of correction $kor.$ in the denominator, the test would be asymptotically conservative.

Generalized location model $\mathcal{F}$ does not hold. Notice at first that under the null-difference hypothesis, i.e. $F_X = F_Y$, the test keeps (asymptotically) the required significance level. The invalidity of this model has therefore effect on the interpretation and the power of the test.

Concerning the **interpretation of the test**, rejecting the null hypothesis outside of the generalized location model only tells us that the distributions F_X and F_Y are not identical. However in general, it is not possible to claim that the medians, resp. the expected values, of those distributions differ.

Concerning **the power** of the test, in the previously described generalized location model it holds that Wilcoxon test is consistent.

In practice however, we can never be sure that the generalized location model holds. Therefore, to better understand the two-sample Wilcoxon test, it is convenient to use the Mann-Whitney formulation of the Wilcoxon test presented in the following section.

* See for example Hollander et al. (2013), page 118.

Mann-Whitney formulation of Wilcoxon test

Test equivalent to Wilcoxon test can be obtained by the following reasoning. Consider all pairs (X_i, Y_j) for $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$ and determine how many of them satisfy $X_i < Y_j$:

$$W^*_{n,m} = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{1}\{X_i < Y_j\}.$$

The random variable $W^*_{n,m}$, called the *Mann-Whitney statistic*, can attain values from the set $\{0, \ldots, nm\}$.

The following proposition shows that there exists a deterministic linear relation between the two-sample Wilcoxon statistic $W_{n,m}$ and the Mann-Whitney statistic $W^*_{n,m}$. In particular, we can easily compute moments of $W^*_{n,m}$.

Proposition 6.6

(i) $W_{n,m} + W^*_{n,m} = nm + \frac{n(n+1)}{2}$.

(ii) If $\min(n, m) \to \infty$, then $\frac{W^*_{n,m}}{nm} \xrightarrow{\mathsf{P}} \mathsf{P}[X_i < Y_j]$.

Proof. Part (i). From the definition of a rank we have

$$R_i = \sum_{j=1}^{n+m} \mathbb{1}\{Z_j \le X_i\} = \sum_{j=1}^{n} \mathbb{1}\{X_j \le X_i\} + \sum_{j=1}^{m} \mathbb{1}\{Y_j \le X_i\}.$$

So

$$\begin{aligned} W_{n,m} + W^*_{n,m} &= \sum_{i=1}^{n} R_i + \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{1}\{X_i < Y_j\} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}\{X_j \le X_i\} + \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{1}\{Y_j \le X_i\} + \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{1}\{X_i < Y_j\} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}\{X_{(j)} \le X_{(i)}\} + \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{1}\{Y_j \le X_i \text{ or } Y_j > X_i\} \\ &= \sum_{i=1}^{n} i + nm = \frac{n(n+1)}{2} + nm. \end{aligned}$$

Part (ii). Will not be proven. The difficulty of this proof lies in the fact that the indicators $\mathbb{1}\{X_i < Y_j\}$ are not (for $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, m\}$) independent random variables. □

Let us analyse corollaries of Proposition 6.6. Part (i) tells us that tests based on the Wilcoxon test statistic and the Mann-Whitney test statistic are equivalent. Part (ii) shows that $\frac{W^*_{n,m}}{nm}$ is a consistent estimate of the parameter $\theta_{XY} = \mathsf{P}[X_i < Y_j]$. It can

easily be shown that if $F_X = F_Y$ then $\theta_{XY} = 1/2$. However, parameter θ_{XY} can be equal to $1/2$ even for two distributions that are not identical.

So, if we consider two-sample Wilcoxon test as a test of the null-difference hypothesis (6.1), then this test is consistent only against alternatives for which $\theta_{XY} \neq \frac{1}{2}$. This inequality cannot be in general (i.e. outside of the generalized location model) interpreted as a inequality of expected values or medians. There exist continuous distributions F_X and F_Y such that their expected values (resp. medians) are different and at the same time $\theta_{XY} = \frac{1}{2}$. On the other hand, there also exist continuous distributions F_X and F_Y such that their expected values (resp. medians) are the same and at the same time $\theta_{XY} \neq \frac{1}{2}$.

Considering all of the above, we could be interested in the question, whether we could regard the Mann-Whitney test as a test for the following general situation.

Model: $\mathcal{F}^* = \{X \sim F_X \text{ continuous d.f., } Y \sim F_Y \text{ continuous d.f.}\}$

Tested parameter: $\theta_{XY} = \mathsf{P}[X < Y]$

Null hypothesis and alternative:

$$H_0^* : \theta_{XY} = \tfrac{1}{2}, \quad H_1^* : \theta_{XY} \neq \tfrac{1}{2}.$$

However, the problem lies in the fact that, in this case, we cannot compute the variance of the test statistic $W_{n,m}^*$ under the null hypothesis with the help of Proposition 6.5 (since under the null hypothesis we do not have in general identically distributed random variables). So critical values computed for Wilcoxon test in model $\mathcal{F}$ do not work in general model $\mathcal{F}^*$. And it turns out that ignoring this fact can lead to both conservative and liberal tests.*

The above reasoning leads to clear conclusion: *If we want to test the equality of expected values without additional assumptions on the shape of the distributions of both samples, we use two-sample t-test without the assumption of equality of variances (Welch test), not Wilcoxon test.*

Remark. It is sometimes recommended to test the normality of both samples (e.g. by the popular Shapiro-Wilk test, which is not presented) before using two-sample t-test to compare the expected values. If the normality is rejected, we use Wilcoxon test, otherwise we use two-sample t-test. However, we strongly advise against using this approach. As we already know, these two test are testing different hypothesis, we cannot use them on the same problem. If we are uncertain of the normality of our data, we should rather use Welch test, which tests the required hypothesis but does not require the assumption of normality.

Remark. If ties are present, it is necessary to slightly modify Proposition 6.6. If we use the average ranks to compute the statistic $W_{n,m}$, then formula (i) holds, if we re-

* Standardization of the test statistic $W_{n,m}^*$ which assures that the test keeps the required level asymptotically, even in general model $\mathcal{F}^*$, can be found for example in Chung and Romano (2016).

define the statistic $W^*_{n,m}$ as

$$W^*_{n,m} = \sum_{i=1}^{n} \sum_{j=1}^{m} \left[\mathbb{1}\{X_i < Y_j\} + \tfrac{1}{2} \mathbb{1}\{X_i = Y_j\} \right].$$

Part (ii) must then be modified to

$$\frac{W^*_{n,m}}{mn} \xrightarrow{\mathsf{P}} \mathsf{P}[X_i < Y_j] + \tfrac{1}{2}\mathsf{P}[X_i = Y_j].$$

6.5. Two-sample F-test of equality of variances

Two-sample F-test of equality of variances is an exact test comparing variances of two independent random samples under the assumption of **normality**.

Model: $\mathcal{F} = \{X_i \sim \mathsf{N}(\mu_X, \sigma_X^2), Y_j \sim \mathsf{N}(\mu_Y, \sigma_Y^2), \mu_X, \mu_Y \in \mathbb{R}, \sigma_X^2 > 0, \sigma_Y^2 > 0\}$

Tested parameters: Variances $\sigma_X^2 = \mathrm{var}\, X_i$ and $\sigma_Y^2 = \mathrm{var}\, Y_j$.

Null hypothesis and alternative:

$$H_0 : \sigma_X^2 = \sigma_Y^2, \quad H_1 : \sigma_X^2 \neq \sigma_Y^2.$$

Test statistic:

$$F = \frac{S_X^2}{S_Y^2},$$

where S_X^2 is the sample variance of the random sample $X_1, \ldots, X_n$ and S_Y^2 is the sample variance of the random sample $Y_1, \ldots, Y_m$.

Remark.

- Theorem 2.11 implies that, in the above model and under the null hypothesis, the exact distribution of the test statistic is $F_{n-1,m-1}$ distribution.
- We rejected the null hypothesis if the sample variances differ too much, i.e. if the value of the test statistic is too small or too large.

Critical region:

$$H_0 \text{ is rejected } \Leftrightarrow F \leq F_{n-1,m-1}(\alpha/2) \text{ or } F \geq F_{n-1,m-1}(1-\alpha/2),$$

where $F_{n-1,m-1}(\alpha/2)$ and $F_{n-1,m-1}(1-\alpha/2)$ are $(\alpha/2)$-quantile and $(1-\alpha/2)$-quantile of the F-distribution with $n-1$ and $m-1$ degrees of freedom.

P-value: $p = 2\min\{1 - G(s), G(s)\}$, where s is the observed value of the test statistic F and G is the distribution function of the distribution $F_{n-1,m-1}$.

Confidence intervals for σ_X^2/σ_Y^2: According to Theorem 2.11 it holds that

$$\mathsf{P}\left[F_{n-1,m-1}(\alpha/2) < \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} < F_{n-1,m-1}(1-\alpha/2) \right] = 1 - \alpha.$$

So confidence interval for σ_X^2/σ_Y^2 is given by formula

$$\left(\frac{S_X^2}{S_Y^2}\,\frac{1}{F_{n-1,m-1}(1-\frac{\alpha}{2})}, \frac{S_X^2}{S_Y^2}\,\frac{1}{F_{n-1,m-1}(\frac{\alpha}{2})}\right).$$

Remark. This test can be modified to one-sided test: Null hypothesis $H_0' : \sigma_X^2 \leq \sigma_Y^2$ is rejected for large values of the test statistic, critical value is $F_{m-1,n-1}(1-\alpha)$. Null hypothesis $H_0'' : \sigma_X^2 \geq \sigma_Y^2$ is rejected for small values of the test statistic, critical value is $F_{m-1,n-1}(\alpha)$.

Violation of assumptions

If **the assumption of normality is violated**, this test **does not keep the level** even asymptotically. To construct a test without this assumption, we would need to derive an asymptotic distribution of the test statistic F under the hypothesis and work with this distribution. Alternatively, one can use the Levene's test. It can be used to compare more independent random samples. However, we do have to highlight that, in general, it does not test the equality of variances, but the equality of a slightly different parameter of variability.

Sample examples for the preparation for the exam.

1. Consider $X_i \sim \mathrm{Exp}(\lambda_1)$ and $Y_j \sim \mathrm{Exp}(\lambda_2)$. Show that in this situation X_i and Y_j satisfy the generalized location model (6.7).
   Hint. Consider $g(x) = \log x$.

2. Modify the two-sample F-test of variance so that it tests the null hypothesis $H_0 : \sigma_X^2 \leq \sigma_Y^2$ against the alternative $H_1 : \sigma_X^2 > \sigma_Y^2$. Justify your modification. What would be the formula for p-value in this modified test?

3. We are deciding whether to send our employees to a language course from the company Analfabet or from the company Buran. To make this decision, we have randomly chosen 20 employees and split them (again randomly) between those two courses (10 employees went to each course). In the following table we present the number of points received in an English test, taken by all of the employees after they have completed their respective course.

Analfabet course	37	41	36	48	42	36	42	44	40	34
Buran course	38	43	43	47	52	44	41	42	42	39

   How would you test that there is no difference between the courses?

   The point of this exercise is not to numerically calculate the test statistic, but rather to explain the test in detail (i.e. define suitable model for your data, null and alternative hypothesis, test statistic and critical region).

4. We have data about the salaries of 100 employees in a large insurance company. We also have the information whether these employees studied at MFF UK or at another school. Suggest a test (i.e. define suitable model for the data, null and alternative hypothesis, test statistic and formula for p-value), if we want to show that the graduates of MFF UK have higher salaries then the graduates of other schools.

5. We have two generators of independent numbers from two given distributions. We have obtained 500 random numbers from each generator. Suggest a test (i.e. define suitable model for the data, null and alternative hypothesis, test statistic and critical region), which can be used to test that both generators generate the random numbers from the same distribution.

The end of self-study for week 10 (8.12.-12.12.).

7. One-sample and two-samples problems for binary data

In this chapter we will be dealing with *binary variables*, i.e. variables that can take only tow values.

7.1. One-sample problem

Bernoulli distribution is the most simple for the categorical variable that takes only two possible values coded as 0 and 1. Let $p_X \in (0,1)$ be the probability that a given subject is classified in the category 1.

Let $Y_1, \ldots, Y_n$ be a random sample from the Bernoulli distribution $\mathsf{Be}(p_X)$ that represents the categories of n subjects. Denote the number of subject classified in the category 1 as $X_n = \sum_{i=1}^n Y_i$. This random variable has the binomial distribution $\mathsf{Bi}(n, p_X)$ (see Theorem 2.3(iv)).

We know that the relative frequency

$$\widehat{p}_n = \frac{X_n}{n} = \frac{\sum_{i=1}^n Y_i}{n} = \overline{Y}_n$$

is a consistent and unbiased estimator of p_X. The properties of $\widehat{p}_n$ are summarized in Theorem 2.3.

7.1.1. Clopper-Pearson method

This method makes use of $\mathsf{Bi}(n, p_X)$ which is the exact distribution of the statistic X_n.

Consider the hypothesis $H_0 : p_X = p_0$ against the alternative $H_1 : p_X \neq p_0$. The critical region is given by

$$H_0 \text{ is rejected } \Leftrightarrow X_n \leq c_L(\alpha) \text{ or } X_n \geq c_U(\alpha),$$

where $c_L(\alpha)$ is the largest integer such that

$$\mathsf{P}\Big(\mathsf{Bi}(n, p_0) \leq c_L(\alpha)\Big) = \sum_{j=0}^{c_L(\alpha)} \binom{n}{j} p_0^j (1-p_0)^{n-j} \leq \frac{\alpha}{2}$$

and $c_U(\alpha)$ is the smallest integer, such that

$$\mathsf{P}\Big(\mathsf{Bi}(n, p_0) \geq c_U(\alpha)\Big) = \sum_{j=c_U(\alpha)}^{n} \binom{n}{j} p_0^j (1-p_0)^{n-j} \leq \frac{\alpha}{2}.$$

This test (called *Clopper-Pearson*) has the level at most α (due to the discrete distribution of the test statistic not all levels are attainable). P-value of this test is given by

$$p(x_n) = 2\min\left\{\mathsf{P}\big(\mathsf{Bi}(n,p_0) \le x_n\big), \mathsf{P}\big(\mathsf{Bi}(n,p_0) \ge x_n\big)\right\} = 2\min\big\{G_0(x_n), 1 - G_0(x_n - 1)\big\},$$

where G_0 is the cumulative distribution function of $\mathsf{Bi}(n,p_0)$ and x_n is the observed value of X_n.

Now consider the task of *finding confidence interval* for p_X with the probability of coverage (at least) $1-\alpha$. Making use of the duality of the confidence intervals and testing we can find the confidence interval as the set set containing the values of the parameters $p \in (0,1)$ for which (with given X_n) we do not reject the null hypothesis $H_0 : p_X = p$ against the alternative $H_1 : p_X \neq p$. Denote G_p cumulative distribution function $\mathsf{Bi}(n,p)$. Then

$$\begin{aligned} IS_n &= \big\{p \in (0,1) : p(X_n) > \alpha, \text{ where } p(X_n) \text{ is the p-value of the test } H_0 : p_X = p\big\} \\ &= \big\{p \in (0,1) : 2\min\{G_p(X_n), 1 - G_p(X_n - 1)\} > \alpha\big\} \\ &= \left\{p \in (0,1) : \sum_{j=0}^{X_n} \binom{n}{j} p^j (1-p)^{n-j} > \frac{\alpha}{2} \quad \text{and at the same time } \sum_{j=X_n}^{n} \binom{n}{j} p^j (1-p)^{n-j} > \frac{\alpha}{2}\right\}. \end{aligned}$$

Thus the confidence interval will be of the form (p_L, p_U), where p_L and p_U are found as the solutions of the following equations

$$\sum_{j=X_n}^{n} \binom{n}{j} p^j (1-p)^{n-j} = \frac{\alpha}{2}, \qquad \sum_{j=0}^{X_n} \binom{n}{j} p^j (1-p)^{n-j} = \frac{\alpha}{2}.$$

It can be shown that p_L and p_U can be calculated explicitly as

$$\left(\frac{X_n q_L(\alpha)}{X_n q_L(\alpha) + n - X_n + 1}, \frac{(X_n+1) q_U(\alpha)}{(X_n+1) q_U(\alpha) + n - X_n}\right),$$

where $q_L(\alpha)$ is the $\alpha/2$-quantile of the distribution $F_{2X_n, 2(n-X_n+1)}$ and $q_U(\alpha)$ is $(1-\alpha/2)$-quantile $F_{2(X_n+1),2(n-X_n)}$. When $X_n = 0$ then we put the lower bound of the confidence interval to 0. Further if $X_n = n$ then the upper bound of the confidence interval is put to 1.

The above interval is called the *Clopper-Pearson confidence interval* for the parameter of the binomial distribution. The advantage of this interval is that the coverage probability is at least $1-\alpha$ for each sample size. The disadvantage is that the coverage probability can be considerable bigger than $1-\alpha$ (which implies that it is too wide).

Now we can return to Clopper-Pearson test of the hypothesis $H_0 : p_X = p_0$ against the alternative $H_1 : p_X \neq p_0$. Instead of calculating the critical values $c_L(\alpha)$ and $c_U(\alpha)$ one can calculate the Clopper-Pearson confidence interval and reject H_0 when p_0 is not included in this interval.

7.1.2. Standard asymptotic method

In Chapter 3.5.2 in the example on p. 50 we found the asymptotic confidence interval for p_X based on Theorem 2.3(iii) and Cramér-Slutsky theorem (Proposition 1.3). Using (3.7) it holds that

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

This can be used to derive the asymptotic test of the hypothesis $H_0 : p_X = p_0$ against the alternative $H_1 : p_X \neq p_0$ with the critical region

$$H_0 \text{ is rejected } \Leftrightarrow \left|\frac{\sqrt{n}\,(\widehat{p}_n - p_0)}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}\right| \geq u_{1-\alpha/2}. \tag{7.1}$$

From the duality of testing and confidence interval (Proposition 4.2(ii)) we can find the confidence interval for p_X as

$$IS_n = \left\{p \in (0,1) : \left|\frac{\sqrt{n}\,(\widehat{p}_n - p)}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}\right| < u_{1-\alpha/2}\right\} = \left(\widehat{p}_n - u_{1-\frac{\alpha}{2}}\sqrt{\tfrac{\widehat{p}_n(1-\widehat{p}_n)}{n}},\ \widehat{p}_n + u_{1-\frac{\alpha}{2}}\sqrt{\tfrac{\widehat{p}_n(1-\widehat{p}_n)}{n}}\right).$$

It is worth noting that this confidence interval is the same as the confidence interval for p_X in Chapter 3.5.2.

The disadvantage of this approach is that if p_X is close to zero or one, than one needs large samples sizes so that the asymptotic approximation is reliable. In practice it is often recommended that for the asymptotic approximation one needs that $\min\{X_n, n - X_n\} \geq 5$. It is also worth noting that this interval is not necessarily included in the interval $(0,1)$.

Exercise. As Bernoulli distribution is in $\mathcal{L}_+^2$, one can also use the the t-test (see Chapter 5.3) that is valid asymptotically. Show that

$$T_n = \frac{\sqrt{n-1}\,(\widehat{p}_n - p_0)}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}.$$

Further this test statistic would be compared with the quantiles of t_{n-1}-distribution. Thus the t-test would result in a test that is slightly more conservative than the test given in (7.1).

7.1.3. Wilsonova method

This method is based directly on Theorem 2.3(iii) which states that

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

Under the null hypothesis $H_0 : p_X = p_0$ we know p_X and thus one can perform the test as

$$H_0 \text{ is rejected } \Leftrightarrow \left| \frac{\sqrt{n}\,(\widehat{p}_n - p_0)}{\sqrt{p_0(1-p_0)}} \right| \geq u_{1-\alpha/2}.$$

This test is known as *Wilson test*.

The confidence interval can be found again with the help of duality of testing and confidence intervals as

$$IS_n = \left\{ p \in (0,1) : \left| \frac{\sqrt{n}\,(\widehat{p}_n - p)}{\sqrt{p(1-p)}} \right| < u_{1-\alpha/2} \right\}.$$

After some algebra we get the following formula for the asymptotic confidence interval

$$\left(\widehat{p}_n + \frac{u^2}{2n} \mp u \sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n} + \frac{u^2}{4n^2}} \right) \frac{1}{1+u^2/n},$$

where u denotes $u_{1-\alpha/2}$. This interval is known also as *Wilson confidence interval*. It is known that Wilson test and Wilson confidence interval provides a more precise results than the methods of Chapter 7.1.2.

It is interesting to note that the middle of the Wilson interval can be expresses ad the weighted mean $w_n \widehat{p}_n + (1 - w_n)1/2$, where $w_n = (1 + u^2/n)^{-1} \to 1$ for $n \to \infty$. When calculating the 95% confidence interval, the middle of the Wilson interval is approximately $(X_n + 2)/(n + 4)$.

7.2. Two sample problems

Let $Y_{11}, \ldots, Y_{1n}$ be a random sample from Bernoulli distribution $\mathsf{Be}(p_1)$ and $Y_{21}, \ldots, Y_{2m}$ be a random sample from $\mathsf{Be}(p_2)$. Denote $X_1 = \sum_{i=1}^{n} Y_{1i}$ and $X_2 = \sum_{i=1}^{m} Y_{2i}$. We will be interested in comparing two independent binomial random variables $X_1 \sim \mathsf{Bi}(n, p_1)$ and $X_2 \sim \mathsf{Bi}(m, p_2)$. We want to find out what is the difference in probabilities p_1 and p_2. The difference between p_1 and p_2 can be expressed in several ways.

If the random variables X_1 and X_2 give the numbers of some negative events (death, disease, defect) then the parameters p_1 and p_2 are called *risks* of events. Probabilities (risks) p_1 and p_2 can be estimated by the corresponding relative frequencies $\widehat{p}_1 = X_1/n$, $\widehat{p}_2 = X_2/m$. The properties of these relative frequencies are summarized by 2.3.

Probabilities (risks) $\widehat{p}_1$ and $\widehat{p}_2$ are usually compared by one of the following three ways:

1. difference of probabilities (risk difference, excess risk) $d_X = p_1 - p_2$, is estimated as $\widehat{d} = \widehat{p}_1 - \widehat{p}_2$;
2. ratio of probabilities (relative risk) $r_X = \frac{p_1}{p_2}$ is estimated as $\widehat{r} = \frac{\widehat{p}_1}{\widehat{p}_2}$;
3. odds ratio $o_X = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1(1-p_2)}{p_2(1-p_1)}$, is estimated as $\widehat{o} = \frac{\widehat{p}_1(1-\widehat{p}_2)}{\widehat{p}_2(1-\widehat{p}_1)} = \frac{X_1(m-X_2)}{X_2(n-X_1)}$.

For each of this way of comparing we will need to derive the asymptotic distribution of the corresponding estimator. For all the asymptotic results given below we will assume that

$$n \to \infty, \quad m \to \infty, \quad n/m \to q \in (0, \infty). \tag{7.2}$$

The results given in this results are also valid when only the number of all observations $n + m$ is fixed, while the sample sizes n and m are random (see the discussion on p. 108).

Note that with the help of the central limit theorem

$$\sqrt{n}\,(\widehat{p}_1 - p_1) \xrightarrow{\mathsf{d}} \mathsf{N}(0, p_1(1-p_1)) \text{ and } \sqrt{m}\,(\widehat{p}_2 - p_2) \xrightarrow{\mathsf{d}} \mathsf{N}(0, p_2(1-p_2)).$$

Further thanks to independence $\widehat{p}_1$ and $\widehat{p}_2$ we get in the same way as in the proof of Theorem 6.2 that

$$\sqrt{m}\begin{pmatrix}\widehat{p}_1 - p_1 \\ \widehat{p}_2 - p_2\end{pmatrix} \xrightarrow{\mathsf{d}} \mathsf{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\frac{p_1(1-p_1)}{q} & 0 \\ 0 & p_2(1-p_2)\end{pmatrix}\right). \tag{7.3}$$

7.2.1. The risk difference

The risk difference is given by $d_X = p_1 - p_2$. This difference says by how much is the risk in population 1 larger than in population 2. This parameter can take values between -1 and 1. The zero value of d_X corresponds to the situation when $p_1 = p_2$.

The consistent and unbiased estimator of parameter d_X is $\widehat{d} = \widehat{p}_1 - \widehat{p}_2$.

Proposition 7.1 Let $p_1, p_2 \in (0, 1)$ and it holds that (7.2). Then

$$\frac{\widehat{d} - d_X}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}} \xrightarrow{\mathsf{d}} \mathsf{N}(0, 1).$$

Proof. The proof is completely analogous to the proof of Theorem 6.2.

First we rewrite

$$\frac{\widehat{d} - d_X}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}} = \frac{\sqrt{m}\,(\widehat{d} - d_X)}{\sqrt{\widehat{p}_1(1-\widehat{p}_1)\frac{m}{n} + \widehat{p}_2(1-\widehat{p}_2)}}.$$

Now with the help of law of large numbers (Proposition 1.4) and continuous mapping theorem (Proposition 1.2) one can show that

$$\sqrt{\widehat{p}_1(1-\widehat{p}_1)\tfrac{m}{n} + \widehat{p}_2(1-\widehat{p}_2)} \xrightarrow{\mathsf{P}} \sqrt{\tfrac{p_1(1-p_1)}{q} + p_2(1-p_2)}\,.$$

With the help of Cramér-Slutsky theorem (Theorem 1.3) it remains to show that

$$\sqrt{m}\,(\widehat{d} - d_X) \xrightarrow{\mathsf{d}} \mathsf{N}\Big(0, \tfrac{p_1(1-p_1)}{q} + p_2(1-p_2)\Big),$$

which can be proved analogously as in the proof of Theorem 6.2 from the joint asymptotic normality of estimators $\widehat{p}_1$ and $\widehat{p}_2$ v (7.3). □

For the asymptotic test hypothesis $H_0 : d_X = 0$ against the alternative $H_1 : d_X \neq 0$ we will use the test statistic

$$\widetilde{T}_d = \frac{\widehat{d}}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}}$$

and the hypothesis will be rejected when $\left|\widetilde{T}_d\right| \geq u_{1-\alpha/2}$.

From Proposition 7.1 we get by the straightforward algebra that

$$\mathsf{P}\left[\widehat{d} - u_{1-\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}} < d_X < \widehat{d} + u_{1-\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}\right] \to 1-\alpha.$$

From this one can easily get the asymptotic confidence interval for the difference of probabilities d_X.

Remark. As the null hypothesis $H_0 : d_X = 0$ implies that $p_1 = p_2$, one can also instead of T_d use the test statistic

$$T_d = \frac{\widehat{d}}{\sqrt{\widetilde{p}(1-\widetilde{p})\left(\frac{1}{n} + \frac{1}{m}\right)}}, \tag{7.4}$$

where $\widetilde{p} = \frac{X_1+X_2}{n+m}$ is the estimate of the joint probability under the null hypothesis. The test statistic T_d has asymptotic distribution $\mathsf{N}(0,1)$ under the null hypothesis. The advantage of this test statistic is that the actual level of the corresponding test is usually closer to α than actual level of the test based on $\widetilde{T}_d$. On the other hand the disadvantage of this test statistic is that it cannot be used to to construct the confidence interval for the difference of probabilities $d_X = p_1 - p_2$.

Exercise. Alternatively one can use also the two-sample t-test (see Chapter 6.2) for testing the hypothesis $H_0 : \mu_X = \mu_Y$. Show that in this situation the test statistic $\widetilde{T}_{n,m}$ is of the form

$$\widetilde{T}_{n,m} = \frac{\widehat{d}}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n-1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m-1}}}.$$

7.2.2. Relative risk

A different way of comparing probabilities (risk) is the *relative risk* $r_X = p_1/p_2$. This parameter says how many times is the risk in population 1 bigger than in population 2 and it can take values in the interval $(0, \infty)$. The probabilities (risks) are the same if and only if $r_X = 1$.

The estimator $\widehat{r} = \widehat{p}_1/\widehat{p}_2$ is consistent (but not unbiased) estimator of the parameter r_X.

Although we can derive the asymptotic distribution of $\widehat{r} = \widehat{p}_1/\widehat{p}_2$, it is known that the normal approximation is more appropriate for the logarithm of the $\widehat{r}$.

Proposition 7.2 Let $p_1, p_2 \in (0,1)$ and it holds that (7.2). Then

$$\frac{\log \widehat{r} - \log r_X}{\sqrt{\frac{1-\widehat{p}_1}{n\widehat{p}_1} + \frac{1-\widehat{p}_2}{m\widehat{p}_2}}} \xrightarrow{\mathsf{d}} \mathsf{N}(0,1).$$

Proof. Again we will proceed analogously as in the proof of Theorem 6.2.

First rewrite

$$\frac{\log \widehat{r} - \log r_X}{\sqrt{\frac{1-\widehat{p}_1}{n\widehat{p}_1} + \frac{1-\widehat{p}_2}{m\widehat{p}_2}}} = \frac{\sqrt{m}\left(\log \widehat{r} - \log r_X\right)}{\sqrt{\frac{m}{n}\frac{1-\widehat{p}_1}{\widehat{p}_1} + \frac{1-\widehat{p}_2}{\widehat{p}_2}}}$$

Now with the help of law of large numbers (Proposition 1.4) and continuous mapping theorem (Proposition 1.2) it is straightforward to show that

$$\sqrt{\frac{m}{n}\frac{1-\widehat{p}_1}{\widehat{p}_1} + \frac{1-\widehat{p}_2}{\widehat{p}_2}} \xrightarrow{\mathsf{P}} \sqrt{\frac{1-p_1}{qp_1} + \frac{1-p_2}{p_2}}.$$

Thus with the help of Cramér-Slutsky theorem (Theorem 1.3) it remains to show that

$$\sqrt{m}\left(\log \widehat{r} - \log r_X\right) \xrightarrow{\mathsf{d}} \mathsf{N}\left(0, \frac{1-p_1}{qp_1} + \frac{1-p_2}{p_2}\right).$$

But this is implied by delta-method (Proposition 1.6) and the joint asymptotic normality (7.3), as the gradient of the function

$$g(p_1, p_2) = \log\left(\frac{p_1}{p_2}\right) = \log p_1 - \log p_2$$

is $\left(\frac{1}{p_1}, \frac{-1}{p_2}\right)$. Thus the asymptotic variance of the random variable $\sqrt{m}\left(\log \widehat{r} - r_X\right)$ is

$$\begin{pmatrix} \frac{1}{p_1} & \frac{-1}{p_2} \end{pmatrix} \begin{pmatrix} \frac{p_1(1-p_1)}{q} & 0 \\ 0 & p_2(1-p_2) \end{pmatrix} \begin{pmatrix} \frac{1}{p_1} \\ \frac{-1}{p_2} \end{pmatrix} = \frac{1-p_1}{qp_1} + \frac{1-p_2}{p_2}.$$

□

Suppose we are interested in testing $r_X = 1$. This can be also expressed as $\log r_X = 0$. Thus for the test of $H_0 : r_X = 1$ against the alternative $H_1 : r_X \neq 1$ one can use the test statistic

$$T_r = \frac{\log \widehat{r}}{\sqrt{\frac{1-\widehat{p}_1}{n\widehat{p}_1} + \frac{1-\widehat{p}_2}{m\widehat{p}_2}}}.$$

The hypothesis is rejected when $|T_r| \geq u_{1-\alpha/2}$.

Proposition 7.2 implies that

$$\mathsf{P}\left[\log \widehat{r} - u_{1-\alpha/2}\sqrt{\frac{1-\widehat{p}_1}{n\widehat{p}_1} + \frac{1-\widehat{p}_2}{m\widehat{p}_2}} < \log r_X < \log \widehat{r} + u_{1-\alpha/2}\sqrt{\frac{1-\widehat{p}_1}{n\widehat{p}_1} + \frac{1-\widehat{p}_2}{m\widehat{p}_2}}\right] \to 1-\alpha.$$

Thus the asymptotic confidence interval for r_X is of the form

$$\left(\widehat{r}\exp\left\{-u_{1-\alpha/2}\sqrt{\frac{1-\widehat{p}_1}{n\widehat{p}_1} + \frac{1-\widehat{p}_2}{m\widehat{p}_2}}\right\}, \widehat{r}\exp\left\{u_{1-\alpha/2}\sqrt{\frac{1-\widehat{p}_1}{n\widehat{p}_1} + \frac{1-\widehat{p}_2}{m\widehat{p}_2}}\right\}\right).$$

Exercise. What would be the critical region for testing the hypothesis $H_0 : r_X = 2$ against the alternative $H_1 : r_X \neq 2$?

7.2.3. Odds ratio

The another way of comparing two probabilities is with the help odds ratio

$$o_X = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1(1-p_2)}{p_2(1-p_1)}.$$

This parameter quantifies how much is is the odd in population 1 larger than in population 2. This parameter can take values in the interval $(0, \infty)$. The probabilities (risks) in both populations are equal if and only if $o_X = 1$.

The consistent estimator of the parameter o_X is given by

$$\widehat{o} = \frac{\widehat{p}_1(1-\widehat{p}_2)}{\widehat{p}_2(1-\widehat{p}_1)} = \frac{X_1(m-X_2)}{X_2(n-X_1)}.$$

Although one can derive the asymptotic distribution of the estimator $\widehat{o} = \frac{\widehat{p}_1(1-\widehat{p}_2)}{\widehat{p}_2(1-\widehat{p}_1)}$, it has been observed that the normal approximation works better for the logarithm of this estimator.

Proposition 7.3 Let $p_1, p_2 \in (0,1)$ and it holds that (7.2). Put

$$\widehat{V}_o = \frac{1}{n\widehat{p}_1} + \frac{1}{n(1-\widehat{p}_1)} + \frac{1}{m\widehat{p}_2} + \frac{1}{m(1-\widehat{p}_2)} = \frac{1}{X_1} + \frac{1}{n-X_1} + \frac{1}{X_2} + \frac{1}{m-X_2}.$$

Then

$$\frac{\log\widehat{o} - \log o_X}{\sqrt{\widehat{V}_o}} \xrightarrow{\mathsf{d}} \mathsf{N}(0,1).$$

Proof. Similarly as in the proof of Theorem 6.2 first we rewrite

$$\frac{\log\widehat{o} - \log o_X}{\sqrt{\widehat{V}_o}} = \frac{\sqrt{m}\,(\log\widehat{o} - \log o_X)}{\sqrt{m\,\widehat{V}_o}}$$

Now with the help of the law of large numbers (Proposition 1.4) and the continuous mapping theorem (Proposition 1.2) one can show that

$$\sqrt{m\,\widehat{V}_o} = \sqrt{\tfrac{m}{n\widehat{p}_1} + \tfrac{m}{n(1-\widehat{p}_1)} + \tfrac{1}{\widehat{p}_2} + \tfrac{1}{(1-\widehat{p}_2)}} \xrightarrow{\mathsf{P}} \sqrt{\tfrac{1}{qp_1} + \tfrac{1}{q(1-p_1)} + \tfrac{1}{p_2} + \tfrac{1}{(1-p_2)}}$$

Further using Cramér-Slutsky theorem (Theorem 1.3) it remains to show that

$$\sqrt{m}\,(\log\widehat{o} - \log o_X) \xrightarrow{\mathsf{d}} \mathsf{N}\Big(0, \tfrac{1}{qp_1} + \tfrac{1}{q(1-p_1)} + \tfrac{1}{p_2} + \tfrac{1}{(1-p_2)}\Big),$$

which follows from the delta-method (Proposition 1.6) from (7.3) as the gradient of the function

$$g(p_1, p_2) = \log\left(\frac{p_1}{1-p_1} / \frac{p_2}{1-p_2}\right) = \log p_1 - \log(1-p_1) - \log p_2 + \log(1-p_2)$$

is $\left(\frac{1}{p_1} + \frac{1}{1-p_1}, -\frac{1}{p_2} - \frac{1}{1-p_2}\right)$. □

The probabilities (odds) in the two populations are equal if and only if $o_X = 1$ (or alternatively if $\log o_X = 0$). For the asymptotic test of the hypothesis $H_0 : o_X = 1$ against the alternative $H_1 : o_X \neq 1$ we will use the test statistic

$$T_o = \frac{\log \widehat{o}}{\sqrt{\widehat{V}_o}}$$

and the hypothesis is rejected when $|T_o| \geq u_{1-\alpha/2}$.

Proposition 7.3 implies that

$$\mathsf{P}\left[\log \widehat{o} - u_{1-\alpha/2}\sqrt{\widehat{V}_o} < \log o_X < \log \widehat{o} + u_{1-\alpha/2}\sqrt{\widehat{V}_o}\right] \to 1 - \alpha.$$

Thus the asymptotic confidence interval for odds ratio o_X is of the form

$$\left(\widehat{o}\exp\left\{-u_{1-\alpha/2}\sqrt{\widehat{V}_o}\right\},\ \widehat{o}\exp\left\{u_{1-\alpha/2}\sqrt{\widehat{V}_o}\right\}\right).$$

Exercise. What would be the critical region for the hypothesis $H_0 : o_X \leq 2$ against the alternative $H_1 : o_X > 2$?

Sample examples for the preparation for the exam.

The solution of "the practical exercises" should contain the mathematical model, the null and the alternative hypothesis, the test statistic and its (either exact or asymptotic) distribution under the null hypothesis, critical region and the formula to calculate the p-value. It should be also explicitly stated if the test is exact or asymptotic.

1. From 100 (randomly chosen) university graduates there were 11 who support the given party. On the other hand from 200 (randomly chosen) high-school graduates there were 84 people who support that party.
   a) It is possible to say that that the party has the support at least 35 % among high-school graduates?
   b) Is it possible to say that the support among high-school graduates is at least two times larger than among university graduates?
2. The mayor of a small municipality would like to organize a new-year firework but he is not sure if the citizens are in favor of that. He has found that 61 from 100 citizens are in favor of the firework. Based on this data can the mayor be sufficiently sure that at least half of the citizens are in favor of firework?

The end of self-study for week 11 (15.12.-19.12.).

8. Multinomial distribution and contingency tables

In this chapter we will be dealing with *categorical variables*, which can take in general more than two values. The term categorical variable was explained in chapter 3.2.2. Shortly speaking, it is a discrete variable which takes values from a finite set typically denoted as $1, \dots, K$. The values from this does not have to have numerical interpretation. Usually they denote a membership in a given group (category). The parameters used in the analysis of categorical data are typically the probabilities of the categories.

8.1. Multinomial distribution

Multinomial distribution generalises binomial distribution to allow for situations where the categorical variable can take more than two different values.

8.1.1. Multinomial distribution: definition and properties

Definition 8.1 (Multinomial distribution) Let $K \geq 2$ and $n \geq 1$ are non-negative integers and $\boldsymbol{p} = (p_1, \dots, p_K)^\mathsf{T}$ is the vector of the constants such that $p_k > 0 \ \forall k$ and $\sum_{k=1}^K p_k = 1$. We say that the random vector $\boldsymbol{X} = (X_1, \dots, X_K)^\mathsf{T}$ has a multinomial distribution $\mathsf{Mult}_K(n, \boldsymbol{p})$, if his density with respect to the counting measure on $\mathbb{Z}^K$ is

$$\mathsf{P}\left[X_1 = x_1, X_2 = x_2, \dots, X_K = x_K\right] = \begin{cases} \dfrac{n!}{x_1! \cdots x_K!} p_1^{x_1} p_2^{x_2} \cdots p_K^{x_K} & \sum_{k=1}^K x_k = n \\ & x_k \in \mathbb{N}_0 \ \forall k \\ 0 & \text{otherwise.} \end{cases}$$

A multinomial distribution is the distribution of the numbers of elements in each of the K boxes (compartments) in n independent experiments, when in each of the experiments the probability of the putting of the element in the boxes is given by $\boldsymbol{p}$.

Theorem 8.1 (Representation of multinomial distribution.) Let $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ be independent random vectors with the distribution $\mathsf{Mult}_K(1, \boldsymbol{p})$. Then $\sum_{i=1}^n \boldsymbol{Y}_i \sim \mathsf{Mult}_K(n, \boldsymbol{p})$.

Proof. We will proceed by the mathematical induction.

For $n = 1$ the statement obviously holds.

Assume now that the statement holds for $n-1$, i.e. $\boldsymbol{X} = \sum_{i=1}^{n-1} \boldsymbol{Y}_i \sim \mathsf{Mult}_K(n-1, \boldsymbol{p})$. We will show that $\boldsymbol{X} + \boldsymbol{Y}_n \sim \mathsf{Mult}_K(n, \boldsymbol{p})$.

Denote $\boldsymbol{Y}_n = (Y_{n1} \dots, Y_{nK})^\mathsf{T}$ and for $\sum_{k=1}^K x_k = n$ we can make use of the induction assumption and calculate

$$\begin{aligned}
&\mathsf{P}\big[X_1 + Y_{n1} = x_1, \dots, X_K + Y_{nK} = x_K\big] \\
&\qquad = \sum_{k=1}^K \mathsf{P}\big[X_1 + Y_{n1} = x_1, \dots, X_K + Y_{nK} = x_K \mid Y_{nk} = 1\big]\, \mathsf{P}\big[Y_{nk} = 1\big] \\
&\qquad = \sum_{k=1}^K \mathsf{P}\big[X_k = x_k - 1, X_j = x_j, \forall_{j\neq k}\big]\, \mathsf{P}\big[Y_{nk} = 1\big] \\
&\qquad = \sum_{k=1}^K \frac{(n-1)!}{(x_k-1)! \prod_{j=1, j\neq k}^K x_j!}\, p_k^{x_k - 1} \left(\prod_{j=1, j\neq k}^K p_j^{x_j} \right) p_k \\
&\qquad = \frac{(n-1)!}{\prod_{j=1}^K x_j!} \left(\prod_{j=1}^K p_j^{x_j} \right) \sum_{k=1}^K x_k = \frac{n!}{x_1! \cdots x_K!}\, p_1^{x_1} \cdots p_K^{x_K}.
\end{aligned}$$

□

Theorem 8.2 (Properties of the multinomial distribution.) Let $\boldsymbol{X} \sim \mathsf{Mult}_K(n, \boldsymbol{p})$. Then

(i) $X_k \sim \mathsf{Bi}(n, p_k)$,
(ii) $\mathsf{E}\, X_k = np_k$, $\mathsf{var}\, X_k = np_k(1-p_k)$,
(iii) $\mathsf{cov}\,(X_j, X_k) = -np_jp_k$, for $j \neq k$,
(iv) the variance matrix of $\boldsymbol{X}$ is

$$\mathsf{var}\, \boldsymbol{X} = n\left[\mathsf{diag}\,(\boldsymbol{p}) - \boldsymbol{p}^{\otimes 2}\right],$$

where $\mathsf{diag}\,(\boldsymbol{p})$ is the diagonal matrix with the diagonal given by the elements of the vector $\boldsymbol{p} = (p_1, \dots, p_K)$ a $\boldsymbol{p}^{\otimes 2} = \boldsymbol{p}\boldsymbol{p}^\mathsf{T}$.

Proof. With the help of theorem 8.1 we can represent $\boldsymbol{X}$ as $\boldsymbol{X} = \sum_{i=1}^n \boldsymbol{Y}_i$, where $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ are independent random vectors with the distribution $\mathsf{Mult}_K(1, \boldsymbol{p})$.

Part (i) follows from the fact that $X_k = \sum_{i=1}^n Y_{ik}$.

Part (ii) follows from the properties binomial distribution.

Part (iii). With the help of the above representation one can calculate for $j \neq k$

$$\begin{aligned}
\mathsf{cov}\,(X_j, X_k) &= \mathsf{cov}\left(\sum_{i=1}^n Y_{ij}, \sum_{l=1}^n Y_{lk} \right) = \sum_{i=1}^n \sum_{l=1}^n \mathsf{cov}\,(Y_{ij}, Y_{lk}) \\
&= \sum_{i=1}^n \mathsf{cov}\,(Y_{ij}, Y_{ik}) = n\,\mathsf{cov}\,(Y_{ij}, Y_{ik}) \\
&= n\,(\mathsf{E}\, Y_{ij}Y_{ik} - \mathsf{E}\, Y_{ij}\, \mathsf{E}\, Y_{ik}) = -np_jp_k,
\end{aligned}$$

where we make use of the fact that $\mathsf{cov}\,(Y_{ij}, Y_{lk}) = 0$ for $i \neq j$ (by the independence of random vectors $\boldsymbol{Y}_i$ and $\boldsymbol{Y}_l$), $\mathsf{E}\, Y_{ij}Y_{ik} = 0$ (as only one element of the vector $\boldsymbol{Y}_i$ is non-zero), $\mathsf{E}\, Y_{ij} = p_j$ and $\mathsf{E}\, Y_{ik} = p_k$.

Part (iv). From the statements (ii) and (iii) it follows that

$$\operatorname{var} \boldsymbol{X} = \begin{pmatrix} np_1(1-p_1) & -np_1p_2 & \dots & -np_1p_K \\ -np_2p_1 & np_2(1-p_2) & \dots & -np_2p_K \\ \dots & \dots & \dots & \dots \\ -np_Kp_1 & -np_Kp_2 & \dots & np_K(1-p_K) \end{pmatrix} = n\left[\operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^{\mathsf{T}}\right].$$

□

Theorem 8.3 (Asymptotic properties of a multinomial distribution.)
Let $\boldsymbol{X}_n \sim \mathsf{Mult}_K(n, \boldsymbol{p})$. Then
(i)

$$\frac{1}{\sqrt{n}}(\boldsymbol{X}_n - n\boldsymbol{p}) \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}_K\left(\boldsymbol{0}, \operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}^{\otimes 2}\right),$$

(ii)

$$\sum_{k=1}^{K} \frac{(X_{kn} - np_k)^2}{np_k} \xrightarrow[n\to\infty]{\mathsf{d}} \chi^2_{K-1}.$$

Proof. Part (i). With the help of the theorem 8.1 we can represent $\boldsymbol{X}_n = \sum_{i=1}^{n} \boldsymbol{Y}_i$, where $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ are independent random vectors with the distribution $\mathsf{Mult}_K(1, \boldsymbol{p})$. Further from the theorem 8.2 we know that

$$\mathsf{E}\,\boldsymbol{Y}_i = \boldsymbol{p}, \qquad \operatorname{var} \boldsymbol{Y}_i = \operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}^{\otimes 2}.$$

Thus with the help of central limit theorem for independent identically distributed random vectors (Proposition 1.5)

$$\frac{1}{\sqrt{n}}(\boldsymbol{X}_n - n\,\boldsymbol{p}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\boldsymbol{Y}_i - \boldsymbol{p}) \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}_K\left(\boldsymbol{0}, \operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}^{\otimes 2}\right).$$

Part (ii). Note that

$$\sum_{k=1}^{K} \frac{(X_{nk} - np_k)^2}{np_k} = \boldsymbol{Z}_n^{\mathsf{T}} \boldsymbol{Z}_n,$$

where

$$\boldsymbol{Z}_n = \frac{1}{\sqrt{n}} \operatorname{diag}\left(\tfrac{1}{\sqrt{\boldsymbol{p}}}\right)(\boldsymbol{X}_n - n\boldsymbol{p}).$$

Now wit the help of part (i)

$$\boldsymbol{Z}_n \xrightarrow[n\to\infty]{\mathsf{d}} \boldsymbol{Z} \sim \mathsf{N}_K(\boldsymbol{0}, \Sigma), \tag{8.1}$$

where $\operatorname{diag}\left(\frac{1}{\sqrt{\boldsymbol{p}}}\right)$ is the diagonal matrix with the elements $\frac{1}{\sqrt{p_1}}, \dots, \frac{1}{\sqrt{p_K}}$ on the diagonal.

$$\Sigma = \operatorname{diag}\left(\tfrac{1}{\sqrt{\boldsymbol{p}}}\right)\left[\operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}^{\otimes 2}\right]\operatorname{diag}\left(\tfrac{1}{\sqrt{\boldsymbol{p}}}\right) = \mathbb{I}_K - \sqrt{\boldsymbol{p}}^{\otimes 2}.$$

Note that

$$\begin{aligned}(\mathbb{I}_K - \sqrt{\boldsymbol{p}}^{\otimes 2})(\mathbb{I}_K - \sqrt{\boldsymbol{p}}^{\otimes 2}) &= \mathbb{I}_K - 2\sqrt{\boldsymbol{p}}^{\otimes 2} + \sqrt{\boldsymbol{p}}\sqrt{\boldsymbol{p}}^{\mathsf{T}}\sqrt{\boldsymbol{p}}\sqrt{\boldsymbol{p}}^{\mathsf{T}} \\ &= \mathbb{I}_K - 2\sqrt{\boldsymbol{p}}^{\otimes 2} + \sqrt{\boldsymbol{p}}\sqrt{\boldsymbol{p}}^{\mathsf{T}} = \mathbb{I}_K - \sqrt{\boldsymbol{p}}^{\otimes 2},\end{aligned}$$

as $\sqrt{\boldsymbol{p}}^{\mathsf{T}}\sqrt{\boldsymbol{p}} = 1$. Thus the matrix $\mathbb{I}_K - \sqrt{\boldsymbol{p}}^{\otimes 2}$ is idempotent.

Further with the help of (8.1) and continuous mapping theorem (Proposition 1.5) one gets that

$$\boldsymbol{Z}_n^{\mathsf{T}}\boldsymbol{Z}_n \xrightarrow[n\to\infty]{\mathsf{d}} \boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}.$$

Let the matrix $\Sigma = \mathbb{I}_K - \sqrt{\boldsymbol{p}}^{\otimes 2}$ be idempotent. Then with the help of Lemma A.1 with $\mathbb{A} = \mathbb{I}_K$ we get that the quadratic form $\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}$ follows χ^2-distribution with the degrees of freedom given by

$$\mathsf{tr}\left(\mathbb{A}\Sigma\right) = \mathsf{tr}\left(\mathbb{I}_K - \sqrt{\boldsymbol{p}}^{\otimes 2}\right) = K - \sum_{k=1}^{K} p_k = K - 1.$$

□

8.1.2. Estimating parameters of a multinomial distribution

For estimating the parameters p_k, testing hypotheses about p_k and for the construction of the confidence intervals for p_k we can use the methods described in Chapter 7.1 as by Theorem 8.2(i) it holds that $X_k \sim \mathsf{Bi}(n, p_k)$.

The entire vector $\boldsymbol{p}$ can be estimated by $\widehat{\boldsymbol{p}}_n = \frac{\boldsymbol{X}}{n}$. The joint asymptotic distribution of the estimate $\widehat{\boldsymbol{p}}_n$ follows from Theorem 8.3(i):

$$\sqrt{n}\left(\widehat{\boldsymbol{p}}_n - \boldsymbol{p}\right) = \frac{1}{\sqrt{n}}(\boldsymbol{X} - n\boldsymbol{p}) \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}_K\left(\boldsymbol{0}, \mathsf{diag}\,(\boldsymbol{p}) - \boldsymbol{p}^{\otimes 2}\right).$$

For an arbitrary K-dimensional vector of constants $\boldsymbol{c}$ it holds that

$$\sqrt{n}\left(\boldsymbol{c}^{\mathsf{T}}\widehat{\boldsymbol{p}}_n - \boldsymbol{c}^{\mathsf{T}}\boldsymbol{p}\right) \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}\left(0, \boldsymbol{c}^{\mathsf{T}}[\mathsf{diag}\,(\boldsymbol{p}) - \boldsymbol{p}^{\otimes 2}]\boldsymbol{c}\right).$$

The **unknown** asymptotic variance $V_c = \boldsymbol{c}^{\mathsf{T}}[\mathsf{diag}\,(\boldsymbol{p}) - \boldsymbol{p}^{\otimes 2}]\boldsymbol{c}$ can be estimated as

$$\widehat{V}_c = \boldsymbol{c}^{\mathsf{T}}[\mathsf{diag}\,(\widehat{\boldsymbol{p}}_n) - \widehat{\boldsymbol{p}}_n^{\otimes 2}]\boldsymbol{c}.$$

Then $V_c \neq 0$ and moreover with the help of Cramér-Slucký theorem (Proposition 1.3)

$$\frac{\sqrt{n}\left(\boldsymbol{c}^{\mathsf{T}}\widehat{\boldsymbol{p}}_n - \boldsymbol{c}^{\mathsf{T}}\boldsymbol{p}\right)}{\sqrt{\widehat{V}_c}} \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}(0,1). \tag{8.2}$$

With the help of that one can easily derive the asymptotic test of the hypothesis

$$H_0 : \boldsymbol{c}^{\mathsf{T}}\boldsymbol{p} = \gamma_0, \quad H_1 : \boldsymbol{c}^{\mathsf{T}}\boldsymbol{p} \neq \gamma_0.$$

Consider the following test statistic

$$T_c = \frac{\sqrt{n}\,(\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{p}}_n - \gamma_0)}{\sqrt{\widehat{V}_c}}.$$

Thanks to (8.2) under the null hypothesis this statistic has asymptotically standard normal distribution. Thus we reject H_0 if only if $|T_c| \geq u_{1-\alpha/2}$.

The asymptotic confidence interval for $\boldsymbol{c}^\mathsf{T}\boldsymbol{p}$ based on (8.2) is given by

$$\left(\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{p}}_n - u_{1-\alpha/2}\sqrt{\tfrac{\widehat{V}_c}{n}},\ \boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{p}}_n + u_{1-\alpha/2}\sqrt{\tfrac{\widehat{V}_c}{n}}\right).$$

The vector $\boldsymbol{c}$ is chosen in such a way so that the linear combination $\boldsymbol{c}^\mathsf{T}\boldsymbol{p}$ represents the parameter that we are interested in. For instance if we are interested to know if the probabilities of the first and the last category is the same and we want to calculate the confidence interval for the difference of these probabilities then we take $\boldsymbol{c} = (1, 0, \ldots, 0, -1)^\mathsf{T}$ and $\gamma_0 = 0$.

8.1.3. χ^2-TEST OF GOODNESS OF FIT FOR MULTINOMIAL DISTRIBUTION

By *χ^2-test of goodness of fit* we understand the test of the hypothesis $H_0 : \boldsymbol{p} = \boldsymbol{p}^0$ based on Theorem 8.3(ii). This hypothesis states that the probabilities of categories $\boldsymbol{p} = (p_1, \ldots, p_K)^\mathsf{T}$ are equal to the given hypothetical probabilities $\boldsymbol{p}^0 = (p_1^0, \ldots, p_K^0)^\mathsf{T}$, i.e. $p_k = p_k^0$ for each $k \in \{1, \ldots, K\}$.

By Theorem 8.3(ii) under H_0

$$\chi^2 = \sum_{k=1}^{K} \frac{\left(X_k - np_k^0\right)^2}{np_k^0} \xrightarrow[n\to\infty]{\mathsf{d}} \chi^2_{K-1}. \tag{8.3}$$

Note that the test statistic compares the observed frequency X_k in the category k with the frequency np_k^0 expected under the null hypothesis. **Large values** test statistic speaks against H_0. Thus the null hypothesis H_0 is rejected when

$$H_0 \text{ is rejected } \Leftrightarrow \chi^2 \geq \chi^2_{K-1}(1-\alpha), \tag{8.4}$$

where $\chi^2_{K-1}(1-\alpha)$ stands for the $(1-\alpha)$-quantile of the distribution χ^2_{K-1}.

Let $s_{\boldsymbol{x}}$ be observed value of the test statistic χ^2. Asymptotic p-value of this test is calculated with the help of (4.12) as

$$p(\boldsymbol{x}) = 1 - G_{K-1}(s_{\boldsymbol{x}}),$$

where G_{K-1} is cumulative distribution function of χ^2-distribution with $K-1$ degrees of freedom.

Remark. The asymptotic approximation with the help of χ^2 distribution requires that the sample size n is sufficiently large. A simple rule of thumb is that the expected frequencies np_k^0 should be at least 5 for each $k \in \{1,\dots,K\}$. Otherwise the χ^2-approximation might be rather inaccurate.

Remark. For $K = 2$, $p_1^0 \equiv p_0$, $X_2 = n - X_1$, $p_2^0 = 1 - p_0$ one gets

$$\chi^2 = \frac{(X_1 - np_0)^2}{np_0} + \frac{[n - X_1 - n(1-p_0)]^2}{n(1-p_0)} = \left[\frac{\sqrt{n}\,(\widehat{p}_n - p_0)}{\sqrt{p_0(1-p_0)}}\right]^2, \quad \text{kde } \widehat{p}_n = \frac{X_1}{n}.$$

Thus the test statistic of χ^2-test for $K = 2$ categories coincides with the square of the Wilson test statistic introduced in Chapter 7.1.3.

Remark. Note that for $K > 2$ one cannot express the null hypothesis and the alternative with a one-dimensional parameter. Thus one cannot simply use the duality of confidence intervals and statistical testing (Proposition 4.2). Analogously this hold true for all the tests that follow (with the exception of Chapter 8.2.1) in this chapter. That is why no confidence intervals are given below.

Example (Is the dice regular?)**.** We throw the dice n-times. Let $X_1,\dots,X_6$ be the absolute frequencies of the numbers 1–6 on the dice. The dice is regular when $p_k^0 = 1/6$, $k = 1,\dots,6$. If the the null hypothesis H_0 is rejected then we have proved that the dice is not regular.

Example (Are child-births uniform in the calendar year?)**.** Suppose we know the number of babies $X_1,\dots,X_{12}$ born in the each of the months (from January to December). Then we put $p_k^0 = \frac{m_k}{365}$, where m_k is the number of days in the k-th month. By rejecting H_0 we prove that the child-births are not uniform in the calendar year.

Example (Follow data the distribution given by cdf F_0?)**.** Suppose we have a random sample $Z_1,\dots,Z_n$ and we are interested if this sample is from the distribution given by the cumulative distribution function $F_0(x) = F(x;\boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ is known.

Introduce the intervals $(a_{k-1}, a_k\rangle$, $k = 1,\dots,K$, where $a_0 = -\infty$ and $a_K = \infty$. The number K should be chosen in such a way that it is much smaller than n. Denote $X_k = \sum_{i=1}^n \mathbb{1}_{(a_{k-1},a_k\rangle}(Z_i)$ the number of observations in the k-th interval. Now if $F_0(x) = F(x;\boldsymbol{\theta}_0)$ is the true distribution function of Z_i, then the random vector $\boldsymbol{X} = (X_1,\dots,X_K)^\top$ follows the multinomial distribution $\mathsf{Mult}_K(n,\boldsymbol{p}^0)$, where probabilities of the categories are given by $p_k^0 = F(a_k;\boldsymbol{\theta}_0) - F(a_{k-1};\boldsymbol{\theta}_0)$.

Now we test the null hypothesis $H_0 : \boldsymbol{p} = \boldsymbol{p}^0$ with the χ^2-test of goodness of fit, see (8.4). By rejecting H_0 we prove that $F(x;\boldsymbol{\theta}_0)$ is not the true distribution function of Z_i.

8.1.4. χ^2-test of goodness of fit for multinomial distribution with estimated (nuisance) parameters

In the last example we see that the probabilities of categories p_k^0 may depend on the vector parameter $\boldsymbol{\theta}_0$. The test statistic of goodness of fit statistic (8.3) can be calculated only if this parameter is known. In practice we are often interested in situations when this parameter is not known but we can estimate it. We will show how to modify the test statistic (8.3) and the critical region (8.4) for the situation of the unknown parameter $\boldsymbol{\theta}_0$.

Consider the *model* $\mathcal{F}_0$ given as follows. Let the random vector $\boldsymbol{X} = (X_1, \dots, X_K)^\mathsf{T}$ follow multinomial distribution $\mathsf{Mult}_K(n, \boldsymbol{p}(\boldsymbol{\theta}_X))$, where $\boldsymbol{\theta}_X \in \Theta \subset \mathbb{R}^d$ is unknown d-dimensional parameter, $d < K-1$, and $\boldsymbol{p}$ is a function mapping Θ into $(0,1)^K$ so that $\boldsymbol{p}(\boldsymbol{\theta})^\mathsf{T}\mathbf{1}_K = 1$ for each $\boldsymbol{\theta} \in \Theta$ (the sum of the coordinates of $\boldsymbol{p}(\boldsymbol{\theta})$ is always 1). We are interested whether the distribution $\boldsymbol{X}$ can be described with this model or not.

Example. Suppose that in a given population there are two variants of a given gene. Denote these variants as A (e.g. dark eyes) and a (e.g. blue eyes). Let $\theta_X \in (0,1)$ be the proportion of A in the population of the given gene. Each individual has two variants of the given gene (one from the father and one from the mother). Thus each individual has one of the pairs AA or Aa or aa. If the variants of the genes are mixing independently (i.e. it holds Hardy-Weinberg equilibrium), then the following table give the probabilities of the three possible pairs.

Genotype	Probability
AA	θ_X^2
Aa	$2\theta_X(1-\theta_X)$
aa	$(1-\theta_X)^2$

Suppose now that we observe n independent individuals. Denote X_1, X_2, X_3 the number of individuals with the corresponding pair AA, Aa, aa. Provided that Hardy-Weinberg equilibrium holds then the vector $\boldsymbol{X} = (X_1, X_2, X_3)^\mathsf{T}$ follows the multinomial distribution $\mathsf{Mult}_3(n, \boldsymbol{p}(\theta_X))$, where $\boldsymbol{p}(\theta_X) = (\theta_X^2, 2\theta_X(1-\theta_X), (1-\theta_X)^2)^\mathsf{T}$. Base on the observations $\boldsymbol{X}$ we would like to show if the population is in the Hardy-Weinberg equilibrium.

The parameter $\boldsymbol{\theta}_X$ needs to be estimated. For this reason it is natural to use the maximum likelihood method. Note that the log-likelihood is of the form

$$\ell_n(\boldsymbol{\theta}) = \log\left(\tfrac{n!}{X_1!\cdots X_K!}\,[p_1(\boldsymbol{\theta})]^{X_1}\cdots[p_K(\boldsymbol{\theta})]^{X_K}\right) = \sum_{k=1}^{K} X_k \log p_k(\boldsymbol{\theta}) + \log\left(\tfrac{n!}{X_1!\cdots X_K!}\right).$$

Thus the system of the likelihood equations is given by $\frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n} = 0$, leads to the

system of d equations that determines a d-dimensional parameter $\widehat{\boldsymbol{\theta}}_n$:

$$\sum_{k=1}^{K} \frac{X_k}{p_k(\widehat{\boldsymbol{\theta}}_n)} \frac{\partial p_k(\widehat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{8.5}$$

Consider now the hypotheses

$$H_0 : \exists \boldsymbol{\theta}_X \in \Theta \quad \boldsymbol{p} = \boldsymbol{p}(\boldsymbol{\theta}_X) \qquad \text{(model } \mathcal{F}_0 \text{ holds)}$$

against the alternative

$$H_1 : \forall \boldsymbol{\theta}_X \in \Theta \quad \boldsymbol{p} \neq \boldsymbol{p}(\boldsymbol{\theta}_X) \qquad \text{(model } \mathcal{F}_0 \text{ does not hold).}$$

First we get the estimator $\widehat{\boldsymbol{\theta}}_n$ of the unknown parameter $\boldsymbol{\theta}_X$ by solving (8.5). Then we can test H_0 by the test of goodness of fit with estimated parameters (instead of unknown parameters). The asymptotic distribution of the test statistic is still χ^2. But for each estimated one-dimensional parameter we loose one degrees of freedom.

Proposition 8.4 Let the hypothesis H_0 holds. Then (under appropriate regularity assumption) the test statistic

$$\chi^2 = \sum_{k=1}^{K} \frac{\left[X_k - np_k(\widehat{\boldsymbol{\theta}}_n)\right]^2}{np_k(\widehat{\boldsymbol{\theta}}_n)}$$

has asymptotically χ^2 -distribution with $K - d - 1$ degrees of freedom, where d is the number of estimated parameters.

Note that under the null hypothesis $\mathsf{E}\, X_k = np_k(\boldsymbol{\theta}_X)$. Thus the test statistic compares the observed frequency X_k in the category k with $np_k(\widehat{\boldsymbol{\theta}}_n)$. The latter quantity can be viewed as the estimate of the expected frequencies under the null hypothesis. As **large values** of the test statistic speaks against H_0 one gets the critical region

$$H_0 \text{ is rejected } \Leftrightarrow \chi^2 \geq \chi^2_{K-d-1}(1-\alpha), \tag{8.6}$$

where $\chi^2_{K-d-1}(1-\alpha)$ denotes the $(1-\alpha)$-quantile of the distribution χ^2_{K-d-1}.

Example (Testing goodness-of-fit with a given parametric family?)**.** Suppose we have a random sample $Z_1, \ldots, Z_n$. We are interested if the distribution of Z_i is given by the cumulative distribution function $F_X(x) = F(x; \boldsymbol{\theta}_X)$, where $\boldsymbol{\theta}_X \in \Theta$ is not know (e.g. a normal distribution, a gamma distribution, a Poisson distribution).

Introduce the intervals $(a_{k-1}, a_k\rangle$, $k = 1, \ldots, K$, $a_0 = -\infty$, $a_K = \infty$, where K is small in comparison with n. Let the observed frequencies by given by $X_k = \sum_{i=1}^{n} \mathbb{1}_{(a_{k-1}, a_k\rangle}(Z_i)$.

If the distribution of Z_i is given by $F(x; \boldsymbol{\theta}_X)$, then the random vector $\boldsymbol{X} = (X_1, \ldots, X_K)^\mathsf{T}$ follows the multinomial distribution $\mathsf{Mult}_K\big(n, \boldsymbol{p}(\boldsymbol{\theta}_X)\big)$, where the probabilities of individual categories are given by $p_k(\boldsymbol{\theta}_X) = F(a_k; \boldsymbol{\theta}_X) - F(a_{k-1}; \boldsymbol{\theta}_X)$.

By solving the system of equations (8.5) we get the estimate $\widehat{\boldsymbol{\theta}}_n$ of the parameter $\boldsymbol{\theta}_X$. Now we can perform the test as in (8.6). If the null hypothesis is rejected then we have proved that the distribution of Z_i is not in a given parametric family.

Note that we need that Proposition 8.4 requires that the parameter $\boldsymbol{\theta}_X$ is estimated with the help of maximum likelihood in the model $\boldsymbol{X} \sim \mathsf{Mult}_K\big(n, \boldsymbol{p}(\boldsymbol{\theta}_X)\big)$. Proposition 8.4 is **not true** when the maximum likelihood estimator is used in the model $Z_i \sim F(\cdot; \boldsymbol{\theta}_X)$. I.e. when the estimate of $\boldsymbol{\theta}$ is found as

$$\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\Theta} \sum_{i=1}^{n} \log f(Z_i; \boldsymbol{\theta}),$$

where $f(\cdot; \boldsymbol{\theta})$ is the density of the random variable Z_i with respect to the σ-finite measure μ.

8.2. Contingency tables

Let $\binom{X}{Z}$ be a random vector whose both components are categorical. More specifically suppose that $X \in \{1, \dots, J\}$ and $Z \in \{1, \dots, K\}$. Let

$$\binom{X_1}{Z_1}, \dots, \binom{X_N}{Z_N}$$

be a random sample from the distribution given by the vector $\binom{X}{Z}$ with the fixed sample size N. Denote the number of individuals classified into the j-th category of X and the k-th category of Z as

$$n_{jk} = \sum_{i=1}^{N} \mathbb{1}\{X_i = j, Z_i = k\}, \quad j \in \{1, \dots, J\},\ k = 1, \dots, K.$$

The random variable n_{jk} is called the observed frequency for the combination of categories j and k. Denote $p_{jk} = \mathsf{P}[X = j, Z = k]$ and $\boldsymbol{p} = (p_{11}, \dots, p_{JK})^\mathsf{T}$. As the observed frequencies were classifying N independent individuals into JK categories, the random vector $\boldsymbol{n} = (n_{11}, \dots, n_{JK})^\mathsf{T}$ follows the multinomial distribution $\mathsf{Mult}_{JK}(N, \boldsymbol{p})$. As we work with the multinomial distribution, we can make use of the results presented in Chapter 8.1.

Further denote

$$n_{j+} = \sum_{k=1}^{K} n_{jk}, \quad n_{+k} = \sum_{j=1}^{J} n_{jk}, \quad n_{++} = \sum_{j=1}^{J}\sum_{k=1}^{K} n_{jk} = N,$$
$$p_{j+} = \sum_{k=1}^{K} p_{jk}, \quad p_{+k} = \sum_{j=1}^{J} p_{jk}, \quad p_{++} = \sum_{j=1}^{J}\sum_{k=1}^{K} p_{jk} = 1.$$

While the probabilities p_{jk} characterize the joint distribution of X and Z, probabilities $p_{j+} = \mathsf{P}[X = j]$ characterize the marginal distribution of X and probabilities $p_{+k} = \mathsf{P}[Z = k]$ characterize marginal distribution of Z.

Observed frequencies can be represented by the table that is called the contingency table.

	$Z=1$	...	$Z=K$	Σ
$X=1$	n_{11}	...	n_{1K}	n_{1+}
$X=2$	n_{21}	...	n_{2K}	n_{2+}
...	...	...	...	...
$X=J$	n_{J1}	...	n_{JK}	n_{J+}
Σ	n_{+1}	...	n_{+K}	N

Analogously one can put together the table of probabilities that describes the joint distribution of the vector $(X,Z)^\mathsf{T}$ and the corresponding marginal distribution of the random variables X and Z.

	$Z=1$	...	$Z=K$	Σ
$X=1$	p_{11}	...	p_{1K}	p_{1+}
$X=2$	p_{21}	...	p_{2K}	p_{2+}
...	...	...	...	...
$X=J$	p_{J1}	...	p_{JK}	p_{J+}
Σ	p_{+1}	...	p_{+K}	1

Finally denote the conditional probabilities as

$$\mathsf{P}\left[X = j \mid Z = k\right] = p_{j(k)} = \frac{p_{jk}}{p_{+k}},$$
$$\mathsf{P}\left[Z = k \mid X = j\right] = p_{(j)k} = \frac{p_{jk}}{p_{j+}}.$$

Testing independence χ^2-testem

Random variables X and Z are independent if and only if for each $j \in \{1,\dots,J\}$ and $k \in \{1,\dots,K\}$ it holds that

$$\mathsf{P}[X = j, Z = k] = \mathsf{P}[X = j]\,\mathsf{P}[Z = k] \quad \text{neboli} \quad p_{jk} = p_{j+}p_{+k}.$$

If the null hypothesis holds, then X and Z are independent random variables and the joint probabilities $\boldsymbol{p} = (p_{11},\dots,p_{JK})^\mathsf{T}$ can be written as functions of $d = J + K - 2$ parameters

$$\boldsymbol{\theta}_X = (p_{1+},\dots,p_{(J-1)+},p_{+1},\dots,p_{+(K-1)})^\mathsf{T}.$$

Maximum likelihood estimator of the parameter θ_X under the null hypothesis of independence can be found as the solution of the system of equations (8.5) which is now of the form

$$\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{n_{jk}}{p_{jk}(\widehat{\boldsymbol{\theta}}_n)}\frac{\partial p_{jk}(\widehat{\boldsymbol{\theta}}_n)}{\partial\boldsymbol{\theta}}=\mathbf{0}.$$

Note that differentiating with respect to the parameter p_{j+} gives

$$\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{n_{jk}}{p_{jk}(\boldsymbol{\theta})}\frac{\partial p_{jk}(\boldsymbol{\theta})}{\partial p_{j+}}=\sum_{k=1}^{K}\frac{n_{jk}}{p_{j+}p_{+k}}p_{+k}-\sum_{k=1}^{K}\frac{n_{Jk}}{p_{J+}p_{+k}}p_{+k}=\sum_{k=1}^{K}\left(\frac{n_{jk}}{p_{j+}}-\frac{n_{Jk}}{p_{j+}}\right)=\frac{n_{j+}}{p_{j+}}-\frac{n_{J+}}{p_{j+}}.$$

Thus we get the equations

$$\frac{n_{j+}}{\widehat{p}_{j+}}=\frac{n_{J+}}{\widehat{p}_{J+}},\qquad j=1,\dots,J-1.$$

Analogously for differentiating with respect to the parameter p_{+k}

$$\tfrac{n_{+k}}{\widehat{p}_{+k}}-\tfrac{n_{+K}}{\widehat{p}_{+K}}=0,\qquad k=1,\dots,K-1.$$

Solving the above system of equations gives $\widehat{p}_{j+}=\frac{n_{j+}}{N}$ and $\widehat{p}_{+k}=\frac{n_{+k}}{N}$ yielding that

$$\widehat{\boldsymbol{\theta}}_n=\left(\widehat{p}_{1+},\dots,\widehat{p}_{(J-1)+},\widehat{p}_{+1},\dots,\widehat{p}_{+(K-1)}\right)^{\mathsf{T}}=\left(\tfrac{n_{1+}}{N},\dots,\tfrac{n_{(J-1)+}}{N},\tfrac{n_{+1}}{N},\dots,\tfrac{n_{+(K-1)}}{N}\right)^{\mathsf{T}}.$$

Maximum likelihood estimator of the vector parameter $\boldsymbol{p}$ under the hypothesis independence has components

$$p_{jk}(\widehat{\boldsymbol{\theta}}_n)=\widehat{p}_{j+}\widehat{p}_{+k}=\frac{n_{j+}n_{+k}}{N^2},\quad j\in\{1,\dots,J\},k=\{1,\dots,K\}.$$

Thus the estimated expected frequencies in the contingency table under the null hypothesis of independence are

$$Np_{jk}(\widehat{\boldsymbol{\theta}}_n)=N\widehat{p}_{j+}\widehat{p}_{+k}=\frac{n_{j+}n_{+k}}{N}.$$

So the test statistic of Proposition 8.4 is (for the test of independence) of the form

$$\chi^2=\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{\left(n_{jk}-\frac{n_{j+}n_{+k}}{N}\right)^2}{\frac{n_{j+}n_{+k}}{N}}.\tag{8.7}$$

By Proposition 8.4 under the null hypothesis the asymptotic distribution of this statistic is χ^2-distribution with the degrees of freedom equal to $JK-d-1$, where $d=J+K-2$. i.e. $\chi^2_{(J-1)(K-1)}$. The hypothesis of independence is rejected when $\chi^2\geq\chi^2_{(J-1)(K-1)}(1-\alpha)$.

Remark. The test described above is called the χ^2*-test of independence in the contingency table*. It can be summarized as follows.

Model (for the conting. table): $\binom{X_1}{Z_1}, \dots, \binom{X_N}{Z_N}$ be a random sample from the distribution given by the random vector $\binom{X}{Z}$. Thus for the frequency in the contingency table it holds that

$$\boldsymbol{n} \sim \mathrm{Mult}_{JK}\big(N, (p_{11}, \dots, p_{JK})\big), \text{ where } p_{jk} = \mathsf{P}[X = j, Z = k]. \tag{8.8}$$

This model will be called *the joint multinomial model.*

Hypothesis and alternative:

$$\begin{aligned} &H_0 : X \text{ and } Z \text{ are independent, i.e. } p_{jk} = p_{j+}p_{+k}\ \forall j \in \{1, \dots, J\},\ \forall k \in \{1, \dots, K\} \\ &H_1 : X \text{ and } Z \text{ are not independent, i.e. } \exists_{j\in\{1,\dots,J\}} \exists_{k\in\{1,\dots,K\}} p_{jk} \neq p_{j+}p_{+k} \end{aligned} \tag{8.9}$$

Test statistic: χ^2 given by (8.7)

Distribution of the test statistic under H_0: $\chi^2 \overset{\text{as.}}{\sim} \chi^2_{(J-1)(K-1)}$

Critical region: H_0 is rejected $\Leftrightarrow \chi^2 \geq \chi^2_{(J-1)(K-1)}(1-\alpha)$.

χ^2-TEST TEST AS A TEST OF THE HOMOGENEITY OF MULTINOMIAL DISTRIBUTIONS

Sometimes is natural to view the contingency table column-wise as the realizations of K independent multinomial distributions. But before formulating the model formally note that the components of the random vector $\binom{X}{Z}$ are independent, if and only if for all $j \in \{1, \dots, J\}$ and $k \in \{1, \dots, K\}$ it holds that

$$\mathsf{P}\left[X = j \mid Z = k\right] = \mathsf{P}[X = j] \quad \text{neboli} \quad p_{j(k)} = p_{j+}\,.$$

I.e. the null hypothesis of independence holds, if and only if

$$p_{j(1)} = p_{j(2)} = \ldots = p_{j(K)} \text{ for each } j \in \{1, \dots, J\}.$$

Denote $\boldsymbol{p}_{(k)} = \big(p_{1(k)}, \dots, p_{J(k)}\big)^\mathsf{T}$. From the above thoughts one can conclude that independence X of Z is equivalent to the fact that the the vectors of conditional probabilities $\boldsymbol{p}_{(1)}, \dots, \boldsymbol{p}_{(K)}$ are equal.

Let us now formalize the columns-wise view on the contingency table. Denote $\boldsymbol{n}_{(k)} = (n_{1k}, \dots, n_{Jk})^\mathsf{T}$ the vector of frequencies in the k-th column.

Model (for conting. table):

$$\boldsymbol{n}_{(k)} \sim \mathrm{Mult}_J(n_{+k}, \boldsymbol{p}_{(k)}),\ k \in \{1, \dots, K\}, \text{ where } \boldsymbol{n}_{\cdot 1}, \dots, \boldsymbol{n}_{\cdot K} \text{ are independent.} \tag{8.10}$$

This model will be called *the column-wise multinomial model.*

Hypothesis and alternative:

$$H_0 : \boldsymbol{p}_{(1)} = \ldots = \boldsymbol{p}_{(K)}, \quad H_1 : \exists_{k,l\in\{1,\dots,K\}}\, \boldsymbol{p}_{(k)} \neq \boldsymbol{p}_{(l)}, \tag{8.11}$$

From the above considerations we know that the above null hypothesis is equivalent to independence X of Z (8.9) in the joint multinomial model (8.8). From this one can conclude that the test statistic χ^2 given by (8.7) is also a suitable statistic for testing the null hypothesis (8.11) in the column-wise model (8.10). Further it can be proved that under the null hypothesis and appropriate assumptions on the column sizes $(n_{+1}, \ldots, n_{+K})$ it holds that $\chi^2 \overset{\text{as.}}{\sim} \chi^2_{(J-1)(K-1)}$. Thus the test of the hypothesis of *homogeneity of multinomial distributions* (8.11) in the column-wise multinomial model can be performed in the completely same way as a test of independence in the joint multinomial model.

Remark. The above considerations can be summarized as follows. The test statistic (8.7) can be used for testing hypotheses independence (8.9) in the model (8.8) as well as for testing the hypothesis of homogeneity multinomial distributions (8.11) in the model (8.10). The choice of the model and the corresponding hypothesis depends on the given applications.

Further it is also worth noting that in the column-wise multinomial model and hypothesis (8.11) we perform in fact a K-sample test.*

8.2.1. Contingency tables 2×2

Consider now the special situation when $J = 2$ and $K = 2$, i.e. both components of the random vector $\binom{X}{Z}$ can take only two values. The corresponding contingency table is

	$Z=1$	$Z=2$	Σ
$X=1$	n_{11}	n_{12}	n_{1+}
$X=2$	n_{21}	n_{22}	n_{2+}
Σ	n_{+1}	n_{+2}	N

	$Z=1$	$Z=2$	Σ
$X=1$	p_{11}	p_{12}	p_{1+}
$X=2$	p_{21}	p_{22}	p_{2+}
Σ	p_{+1}	p_{+2}	1

The test statistic is given by

$$\chi^2 = \sum_{j=1}^{2} \sum_{k=1}^{2} \frac{\left(n_{jk} - \frac{n_{j+}n_{+k}}{N}\right)^2}{\frac{n_{j+}n_{+k}}{N}}. \tag{8.12}$$

Under the null hypothesis of independence the test statistic has asymptotically χ^2_1 distribution. The hypothesis of independence is rejected when $\chi^2 \geq \chi^2_1(1-\alpha)$.

χ^2-test as a test of homogeneity of two binomial distributions

Suppose that the variable Z stands for the number of the sample. Then we have one sample consisting of random variables X representing individuals for which $Z = 1$. The second sample consists of random variables X of individuals satisfying $Z = 2$.

In the first sample consisting of n_{+1} observations there are n_{11} individuals for which $X = 1$ (a success) and n_{21} values with $X = 2$ (a failure). The probability of success in

* K-sample tests for quantitative data will be considered in Chapter 9.

the first sample can be denoted as $p_{1(1)} = p_{11}/p_{+1}$. In the second sample consisting of n_{+2} observations there are n_{12} individuals for which $X = 1$ (success) and n_{22} values with $X = 2$ (a failure). The probability of success in the second sample is $p_{1(2)} = p_{12}/p_{+2}$.

From the considerations on the previous pages we know that χ^2-test can be also viewed as a test of the equality of the parameters $p_{1(1)}$ and $p_{1(2)}$ of two independent binomial distribution $\mathrm{Bi}(n_{+1}, p_{1(1)})$ and $\mathrm{Bi}(n_{+2}, p_{1(2)})$. This problem was already treated in Chapter 7.2.

Notation used in Chapter 7.2 can be easily transformed to the notation used here (and otherwise around). The contingency table rewritten in the notation of Chapter 7.2 is given by:

	$Z = 1$	$Z = 2$	Σ
$X = 1$	X_1	X_2	$X_1 + X_2$
$X = 2$	$n - X_1$	$m - X_2$	$n + m - X_1 - X_2$
Σ	n	m	$n + m$

The only difference is that while in Chapter 7.2 we consider two independent random samples here we consider one random sample from the multinomial distribution with 2x2 possible values. While in Chapter 7.2 the sample sizes n, m are considered as fixed, now the sample size are binomial random variables and only the total number of observations $N = n + m$ is fixed. Thus we are again facing two possible formulations of the two-sample problem as discussed at the beginning of Chapter 6 about two-sample tests for quantitative data. Similarly as there it does not matter which of the formulations is chosen and which of the two models is more appropriate for the given contingency table. All the methods presented here are valid for both of the models.

Chapter 7.2 explains how to compare probabilities (risks) of the event $[X = 1]$ for different values of Z. Basically we can make us of one of the three methods of comparison:

- *difference of probabilities* $d_X = p_{1(1)} - p_{1(2)}$ is estimated by $\widehat{d} = \frac{n_{11}}{n_{+1}} - \frac{n_{12}}{n_{+2}}$;
- *ratio of probabilities* $r_X = p_{1(1)}/p_{1(2)}$ is estimated by $\widehat{r} = \frac{n_{11}n_{+2}}{n_{12}n_{+1}}$;
- *odds ratio* $o_X = \frac{p_{1(1)}(1-p_{1(2)})}{p_{1(2)}(1-p_{1(1)})}$ is estimated by $\widehat{o} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$ (that is why the odds ratio is also called *cross ratio*).

The methods for testing and confidence intervals for these parameters are described in Chapter 7.2.

Note that the independence of random variables X and Z are equivalent to one of the equalities below

$$d_X = 0, \quad r_X = 1, \quad o_x = 1.$$

Thus the test of null risk difference, the unit relative risk or odds ratio is equivalent to the test of independence of X and Z.

Remark. It can be shown that for the test statistic of χ^2-test of independence (8.12) it holds that

$$\chi^2 = T_d^2,$$

where T_d is a test statistic for the difference of probabilities give by (7.4).

8.2.2. Contingency table $2 \times K$

Now consider the special situation when $J = 2$ and $K \geq 2$. The contingency table consists of $2 \times K$ frequencies:

	$Z=1$	$Z=2$	...	$Z=K$	Σ
$X=1$	n_{11}	n_{12}	...	n_{1K}	n_{1+}
$X=2$	n_{21}	n_{22}	...	n_{2K}	n_{2+}
Σ	n_{+1}	n_{+2}	...	n_{+K}	N

	$Z=1$	$Z=2$	...	$Z=K$	Σ
$X=1$	p_{11}	p_{12}	...	p_{1K}	p_{1+}
$X=2$	p_{21}	p_{22}	...	p_{2K}	p_{2+}
Σ	p_{+1}	p_{+2}	...	p_{+K}	N

One can view the table column-wise as having K independent samples from the binomial distributions with potentially different probabilities of success p_{1k}/p_{+k}. This can be viewed as a generalization of the two-sample problem treated Chapter 7.2 to more then two samples.

Alternatively one can also view the table row-wise as two samples from the multinomial distribution with potentially different vectors of probabilities

$$\left(\frac{p_{11}}{p_{1+}}, \frac{p_{12}}{p_{1+}}, \ldots, \frac{p_{1K}}{p_{1+}}\right)^\mathsf{T} \quad \text{a} \quad \left(\frac{p_{21}}{p_{2+}}, \frac{p_{22}}{p_{2+}}, \ldots, \frac{p_{2K}}{p_{2+}}\right)^\mathsf{T}.$$

Testing independence by the χ^2-test

X and Z are independent, if and only if $p_{1(1)} = p_{1(2)} = \ldots = p_{1(K)}$. This requires that for each pair of the groups $Z = k_1$ and $Z = k_2$ the difference of the risks 0 (alternatively the relative risk or odds ratio is 1).

When the null hypothesis of independence of X and Z holds then the probabilities $\boldsymbol{p} = (p_{11}, p_{21}, \ldots, p_{1K}, p_{2K})^\mathsf{T}$ specifying the multinomial distribution vector $\boldsymbol{n}$ are functions of K parameters (p_{1+} a $p_{+1}, \ldots, p_{+(K-1)}$). Thus we get that the test statistic is

$$\chi^2 = \sum_{j=1}^{2} \sum_{k=1}^{K} \frac{\left(n_{jk} - \frac{n_{j+}n_{+k}}{N}\right)^2}{\frac{n_{j+}n_{+k}}{N}}.$$

Under the null hypothesis this test statistic has asymptotically χ^2_{K-1} distribution. The null hypothesis is rejected when $\chi^2 \geq \chi^2_{K-1}(1-\alpha)$.

Analogously as in Chapter 8.2.1 one can view the χ^2-test of independence as a test of homogeneity of binomial distributions (i.e. a K-sample tests for binomial distributions).

Alternatively one can view the test also a test that two multinomial distributions have the same vectors of probabilities (i.e. a two-sample test in the multinomial distribution).

Example. Suppose that we observe data about the highest gained education (primary, high-school, university) and whether the given person is a regularly smoker or not. Suppose that we are interested in the relationship of smoking and gained education.

The null hypothesis that the smoking is independent of the gained education can be viewed in two equivalent ways:

- for each of the groups (according to the gained education) the probability of smoking is the same (i.e. we compare three binomial distributions);
- the structure of gained education is the same in the group of smokers as in the group of non-smokers (i.e. we compare two multinomial distributions).

Remark. Suppose we observe K independent random variables $X_1, \dots, X_K$, where $X_k \sim \mathrm{Bi}(n_k, p_k)$ for each $k \in \{1, \dots, K\}$. We want to test the hypotheses

$$H_0 : p_1 = \cdots = p_K, \quad H_1 : \exists_{k \neq j}\, p_k \neq p_j.$$

For this situation in statistical textbooks one can often find that the null hypothesis should be rejected when

$$Q \geq \chi^2_{K-1}(1-\alpha), \qquad \text{where } Q = \frac{1}{\widetilde{p}(1-\widetilde{p})} \sum_{k=1}^{K} n_k (\widehat{p}_k - \widetilde{p})^2,$$

with $\widehat{p}_k = \frac{X_k}{n_k}$, $\widetilde{p} = \frac{1}{N}\sum_{k=1}^{K} X_k$ and $N = \sum_{k=1}^{K} n_k$.

It can be proved that

$$Q = \chi^2,$$

where χ^2 is test statistic of χ^2-test of independence calculated from the following contingency table

	$Z=1$	$Z=2$	$\dots$	$Z=K$	Σ
$X=1$	X_1	X_2	$\dots$	X_K	$\sum_{k=1}^{K} X_k$
$X=2$	$n_1 - X_1$	$n_2 - X_2$	$\dots$	$n_K - X_K$	$N - \sum_{k=1}^{K} X_k$
Σ	n_1	n_2	$\dots$	n_K	N

.

Thus the test based on test statistic Q is the same as the approach based on the χ^2-test of independence.

Sample examples for the preparation for the exam.

The solution of "the practical exercises" should contain the mathematical model, the null and the alternative hypothesis, the test statistic and its (either exact or asymptotic) distribution under the null hypothesis, critical region and the formula to calculate the p-value. It should be also explicitly stated if the test is exact or asymptotic.

1. The target is divided into 4 segments. They were n_j shots in the j-th segment ($j = 1, \ldots, 4$).
   (a) Suggest a test of the hypothesis that the probabilities of hitting the first and the second segment are equal.
   (b) Suggest a test of the hypothesis that the probability of hitting the first segment is at least two times larger than the probability of hitting the fourth segment.

2. In a large shopping centre there are 3 elevators. The management of the shopping centre would like to know if the customers have some preferences regarding these elevators. Suggest a way what data to collect and how to statistically test the hypothesis that the customers have no preferences.

3. Four universities have decided to compare how many left-handed students they have. Each of the university taken a sample of 100 randomly sampled students. Suggest a test of the hypothesis that there is no difference among the universities in the proportions of left-handed students.

9. K-SAMPLE PROBLEM FOR QUANTITATIVE DATA

Two-sample tests verify whether two groups of independent samples differ in some characteristic, usually in the expected value. The question is, how to compare more than two groups at the same time. The problem of comparing several groups of categorical data (binomial or multinomial distribution) was addressed in the previous chapter. In this chapter, we will study this problem for the case of quantitative random variables.

Let us have $K \geq 2$ independent random samples (groups)

$$\begin{aligned} &Y_{11}, \ldots, Y_{1n_1} \text{ from the distribution } F_1, \\ &Y_{21}, \ldots, Y_{2n_2} \text{ from the distribution } F_2, \\ &\vdots \\ \text{and} \quad &Y_{K1}, \ldots, Y_{Kn_K} \text{ from the distribution } F_K. \end{aligned}$$

Individual observations are denoted by Y_{ki}, where the index k stands for the number of the sample the observation belongs to and it attains values from 1 to K, while i is the index of the observation within said sample and it attains values from 1 to n_k, where n_k is the size of kth sample. Denote $N = \sum_{k=1}^{K} n_k$ and $\boldsymbol{n} = (n_1, \ldots, n_K)^\mathsf{T}$. Then we have that $\mathbf{1}_K^\mathsf{T} \boldsymbol{n} = \sum_{k=1}^{K} n_k = N$.

K-sample problem tests the *null-difference hypothesis*

$$H_0 : F_1(x) = F_2(x) = \ldots = F_K(x), \quad \forall x \in R,$$

against the alternative that there exists at least one pair of different groups, i.e.

$$H_1 : \exists_{k \neq j}\, \exists x \in \mathbb{R} : F_k(x) \neq F_j(x).$$

9.1. ANALYSIS OF VARIANCE (ANOVA)

We will assume a model that requires all the distributions $F_1, \ldots, F_K$ to have **the same variance**.

Similarly as in the case of one-sample and two-sample t-test with the assumption of equality of variances (see Sections 5.3 and 6.3) the further described test will be exact under the assumption of normality and asymptotic without this assumption.

Model:

$$\mathcal{F}_n = \left\{ F_k = \mathsf{N}(\mu_k, \sigma^2), \mu_k \in \mathbb{R}, k \in \{1, \ldots, K\}, \sigma^2 > 0 \right\} \tag{9.1}$$

or

$$\mathcal{F}_{as} = \left\{F_k \in \mathcal{L}_+^2, k \in \{1,\dots,K\}, \text{ where } \mathsf{var}\,(Y_{11}) = \mathsf{var}\,(Y_{21}) = \mathsf{var}\,(Y_{K1}) := \sigma^2\right\}.$$

Notice that in the normal model $\mathcal{F}_n$, individual groups can only differ in expected values.

Let μ_k denote the expected value of the kth group, i.e. $\mu_k = \mathsf{E}\,Y_{ki}$. We will deal with the question whether all groups have the same expected value.

Tested parameters: Expected values $\mu_k = \mathsf{E}\,Y_{ki}$.

Null hypothesis and alternative

$$H_0 : \mu_1 = \cdots = \mu_K, \quad H_1 : \exists_{k\neq j}\, \mu_k \neq \mu_j.$$

Notation. Let

- $Y_{k+} \overset{\text{df}}{=} \sum_{i=1}^{n_k} Y_{ki}$ and $\overline{Y}_{k+} \overset{\text{df}}{=} n_k^{-1} \sum_{i=1}^{n_k} Y_{ki}$ be the sum and sample mean of each group
- $Y_{++} \overset{\text{df}}{=} \sum_{k=1}^{K}\sum_{i=1}^{n_k} Y_{ki}$ be the total sum and $\overline{Y}_{++} \overset{\text{df}}{=} N^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_k} Y_{ki}$ be the total sample mean.

Notice that $\overline{Y}_{++}$ is the weighted mean of all group means $\overline{Y}_{k+}$ with weights n_k, i.e.

$$\overline{Y}_{++} = \frac{\sum_{k=1}^{K} n_k \overline{Y}_{k+}}{\sum_{k=1}^{K} n_k}.$$

Furthermore, denote the observations in the groups $\boldsymbol{Y}_k = (Y_{k1},\dots,Y_{kn_k})^\mathsf{T}$, $k \in \{1,\dots,K\}$ and all of the observations $\boldsymbol{Y} = (\boldsymbol{Y}_1^\mathsf{T},\dots,\boldsymbol{Y}_K^\mathsf{T})^\mathsf{T}$.

Our approach will be based on several kinds of sums of squares presented in the following definition.

Definition 9.1 The sum of squares in the analysis of variance:

- $SS_C \overset{\text{df}}{=} \sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(Y_{ki} - \overline{Y}_{++}\right)^2$ is called *the total sum of squares*,
- $SS_A \overset{\text{df}}{=} \sum_{k=1}^{K} n_k\left(\overline{Y}_{k+} - \overline{Y}_{++}\right)^2$ is called *the between group sum of squares*,
- $SS_e \overset{\text{df}}{=} \sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(Y_{ki} - \overline{Y}_{k+}\right)^2$ is called *the residual sum of squares* or *the error sum of squares*.

Theorem 9.1 It holds that

$$SS_C = SS_A + SS_e.$$

Proof.

$$\begin{aligned}
SS_C &= \sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(Y_{ki}-\overline{Y}_{++}\right)^2 = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(Y_{ki}-\overline{Y}_{k+}+\overline{Y}_{k+}-\overline{Y}_{++}\right)^2 \\
&= \sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(Y_{ki}-\overline{Y}_{k+}\right)^2 + \sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(\overline{Y}_{k+}-\overline{Y}_{++}\right)^2 + 2\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(Y_{ki}-\overline{Y}_{k+}\right)\left(\overline{Y}_{k+}-\overline{Y}_{++}\right) \\
&= SS_e + \sum_{k=1}^{K} n_k\left(\overline{Y}_{k+}-\overline{Y}_{++}\right)^2 + 2\sum_{k=1}^{K}\left(\overline{Y}_{k+}-\overline{Y}_{++}\right)\sum_{i=1}^{n_k}\left(Y_{ki}-\overline{Y}_{k+}\right) \\
&= SS_e + SS_A + 0.
\end{aligned}$$

We have used the fact that

$$\sum_{i=1}^{n_k}\left(Y_{ki}-\overline{Y}_{k+}\right) = Y_{k+} - n_k\overline{Y}_{k+} = 0, \qquad \text{for } k \in \{1,\dots,K\}.$$

□

Remark. SS_C measures the total variability of our data. This variability can be decomposed into two parts, the variability between individual groups expressing their difference (SS_A) and the variability within each group SS_e.

$\overline{Y}_{k+}$ is an estimate of μ_k and $\overline{Y}_{++}$ is an estimate of the total expected value (under H_0), therefore SS_A should be small compared to SS_e under the null hypothesis. If SS_A is too large compared to SS_e, it implies that the means of the individual groups differ too much from each other and we should reject the hypothesis of equal expected values.

The test statistic will compare *the variability of the sample means* (SS_A) and *the variability within individual groups* (SS_e). In the following part, we will examine properties of statistics SS_e and SS_A.

Lemma 9.2 Suppose that model $\mathcal{F}_{as}$ holds.

(i) Then it holds that

$$\mathsf{E}\, SS_e = (N-K)\,\sigma^2.$$

(ii) Furthermore, if model $\mathcal{F}_n$ holds, then $\dfrac{SS_e}{\sigma^2} \sim \chi^2_{N-K}$.

Proof. Part (i) Notice that

$$SS_e = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(Y_{ki}-\overline{Y}_{k+}\right)^2 = \sum_{k=1}^{K}(n_k-1)\,S_k^2, \tag{9.2}$$

where $S_k^2 = \frac{1}{n_k-1}\sum_{i=1}^{n_k}(Y_{ki} - \overline{Y}_{k+})^2$ is the sample variance of the kth group. According to Theorem 2.6(ii), S_k^2 is an unbiased estimate of the variance σ^2. Therefore

$$\mathsf{E}\, SS_e = \sum_{k=1}^{K}(n_k - 1)\,\sigma^2 = (N-K)\,\sigma^2.$$

Part (ii) Using (9.2), we can write

$$\frac{SS_e}{\sigma^2} = \sum_{k=1}^{K}\frac{(n_k-1)\,S_k^2}{\sigma^2}.$$

For $k \in \{1,\dots,K\}$, the random variables $\frac{(n_k-1)\,S_k^2}{\sigma^2}$ have, according to Theorem 2.8(i), χ^2-distribution with $n_k - 1$ degrees of freedom. Furthermore, these random variables are independent. Therefore, the random variable $\frac{SS_e}{\sigma^2}$ has χ^2-distribution with $\sum_{k=1}^{K}(n_k-1) = N - K$ degrees of freedom.

□

The following lemma summarises the properties of SS_A. At first, let us denote

$$\overline{\mu} = \mathsf{E}\,\overline{Y}_{++} = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\mathsf{E}\,Y_{ki} = \frac{1}{N}\sum_{k=1}^{K} n_k\mu_k.$$

Lemma 9.3 Assume that the model $\mathcal{F}_{as}$ holds.

(i) Then

$$\mathsf{E}\, SS_A = \sum_{k=1}^{K} n_k(\mu_k - \overline{\mu})^2 + (K-1)\sigma^2.$$

(ii) Furthermore, if the model $\mathcal{F}_n$ holds, then SS_A and SS_e are independent.

(iii) Furthermore, if the model $\mathcal{F}_n$ and the **null hypothesis** H_0 hold, then $\frac{SS_A}{\sigma^2} \sim \chi^2_{K-1}$.

Proof. To prove this theorem, we use the following fact from Theorem 9.1:

$$SS_A = SS_C - SS_e. \tag{9.3}$$

Part (i). Let us at first compute $\mathsf{E}\, SS_C$. Similarly as in Theorem 2.4(ii) we can write

$$SS_C = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(Y_{ki} - \overline{Y}_{++})^2 = \boldsymbol{Y}^{\mathsf{T}}\mathbb{A}_C\boldsymbol{Y}, \quad \text{where} \quad \mathbb{A}_C = \mathbb{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^{\mathsf{T}}. \tag{9.4}$$

So, with the help of Lemma 2.5, we have that

$$\mathsf{E}\, SS_C = \mathsf{E}\,\boldsymbol{Y}^{\mathsf{T}}\mathbb{A}_C\,\mathsf{E}\,\boldsymbol{Y} + \mathsf{tr}\left(\mathbb{A}_C\,\mathsf{var}\,(\boldsymbol{Y})\right) = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(\mu_k - \overline{\mu})^2 + \sigma^2\,\mathsf{tr}\left(\mathbb{A}_C\right)$$
$$= \sum_{k=1}^{K} n_k(\mu_k - \overline{\mu})^2 + \sigma^2(N-1).$$

Furthermore, we know from Lemma 9.2 that $\mathsf{E}\, SS_e = (N-K)\,\sigma^2$. Using (9.3), we can write

$$\mathsf{E}\, SS_A = \mathsf{E}\, SS_C - \mathsf{E}\, SS_e = \sum_{k=1}^{K} n_k(\mu_k - \overline{\mu})^2 + \sigma^2(K-1).$$

Part (ii). Notice that $\boldsymbol{Y} \sim \mathsf{N}_N(\cdot, \sigma^2 \mathbb{I}_N)$. Since our aim is to use Lemma 2.7(ii), we have to, at first, express SS_e and SS_A as quadratic forms of all observations $\boldsymbol{Y}$.

Notice that by using (9.2) we get, similarly as in (9.4), that

$$SS_e = \sum_{k=1}^{K}(n_k - 1)\, S_k^2 = \sum_{k=1}^{K} \boldsymbol{Y}_k^{\mathsf{T}}\big(\mathbb{I}_{n_k} - \tfrac{1}{n_k}\mathbf{1}_{n_k}\mathbf{1}_{n_k}^{\mathsf{T}}\big)\boldsymbol{Y}_k = \boldsymbol{Y}^{\mathsf{T}}\big(\mathbb{I}_N - \mathbb{B}\big)\boldsymbol{Y}, \tag{9.5}$$

where

$$\mathbb{B} = \begin{pmatrix} \frac{1}{n_1}\mathbf{1}_{n_1}\mathbf{1}_{n_1}^{\mathsf{T}} & \mathbb{0}_{n_1\times n_2} & \cdots & \mathbb{0}_{n_1\times n_K} \\ \mathbb{0}_{n_2\times n_1} & \frac{1}{n_2}\mathbf{1}_{n_2}\mathbf{1}_{n_2}^{\mathsf{T}} & \cdots & \mathbb{0}_{n_2\times n_K} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{0}_{n_K\times n_1} & \mathbb{0}_{n_K\times n_2} & \cdots & \frac{1}{n_K}\mathbf{1}_{n_K}\mathbf{1}_{n_K}^{\mathsf{T}} \end{pmatrix}.$$

Moreover, using (9.3), (9.4) and (9.5), we get that

$$\begin{aligned} SS_A = SS_C - SS_e &= \boldsymbol{Y}^{\mathsf{T}}\big(\mathbb{I}_N - \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^{\mathsf{T}}\big)\boldsymbol{Y} - \boldsymbol{Y}^{\mathsf{T}}\big(\mathbb{I}_N - \mathbb{B}\big)\boldsymbol{Y} \\ &= \boldsymbol{Y}^{\mathsf{T}}\big(\mathbb{B} - \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^{\mathsf{T}}\big)\boldsymbol{Y}. \end{aligned} \tag{9.6}$$

Since we have $\Sigma = \mathsf{var}\,(\boldsymbol{Y}) = \sigma^2\mathbb{I}_N$, it is now enough to verify, thanks to Lemma 2.7(ii), that the product of matrices $\big(\mathbb{B} - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^{\mathsf{T}}\big)\big(\mathbb{I}_N - \mathbb{B}\big)$ is a null matrix. We can compute

$$\begin{aligned} \big(\mathbb{B} - \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^{\mathsf{T}}\big)\big(\mathbb{I}_N - \mathbb{B}\big) &= \mathbb{B} - \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^{\mathsf{T}} - \mathbb{B}\mathbb{B} + \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^{\mathsf{T}}\mathbb{B} \\ &= \mathbb{B} - \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^{\mathsf{T}} - \mathbb{B} + \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^{\mathsf{T}} = \mathbb{0}_{N\times N}, \end{aligned}$$

where we have used the fact that

$$\mathbb{B}\mathbb{B} = \mathbb{B} \quad \text{and} \quad \mathbf{1}_N^{\mathsf{T}}\mathbb{B} = \mathbf{1}_N^{\mathsf{T}}. \tag{9.7}$$

Part (iii). Notice at first that the statistic SS_A is invariant under translations, i.e. the value SS_A does not change, if we compute it from $\widetilde{\boldsymbol{Y}} = \boldsymbol{Y} - c\mathbf{1}_N$, for any $c \in \mathbb{R}$.

So, using (9.6), we get that

$$\frac{SS_A}{\sigma^2} = \Big(\tfrac{\boldsymbol{Y}-\mu\mathbf{1}_N}{\sigma}\Big)^{\mathsf{T}}\Big(\mathbb{B} - \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^{\mathsf{T}}\Big)\Big(\tfrac{\boldsymbol{Y}-\mu\mathbf{1}_N}{\sigma}\Big),$$

where μ is the common value of parameters $\mu_1, \ldots, \mu_K$ under the null hypothesis.

We have that $\frac{\boldsymbol{Y}-\mu\mathbf{1}_N}{\sigma} \sim \mathsf{N}_N(\mathbf{0}, \mathbb{I}_N)$. So we are in the situation of Lemma A.1 with $\mathbb{A} = \mathbb{B} - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\mathsf{T}$ and $\Sigma = \mathbb{I}_N$. It remains to verify that the matrix $\mathbb{A}\Sigma = \mathbb{B} - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\mathsf{T}$ is idempotent. Let us compute

$$\begin{aligned}\left(\mathbb{B} - \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^\mathsf{T}\right)\left(\mathbb{B} - \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^\mathsf{T}\right) &= \mathbb{B}\mathbb{B} - \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^\mathsf{T}\mathbb{B} - \mathbb{B}\tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^\mathsf{T} + \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^\mathsf{T}\tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^\mathsf{T} \\ &= \mathbb{B} - \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^\mathsf{T} - \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^\mathsf{T} + \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^\mathsf{T} = \mathbb{B} - \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^\mathsf{T},\end{aligned}$$

where we have used (9.7), the symmetry of the matrix $\mathbb{B}$ and the fact that $\frac{1}{N}\mathbf{1}_N^\mathsf{T}\mathbf{1}_N = 1$. So, according to Lemma A.1, $\frac{SS_A}{\sigma^2}$ has χ^2-distribution with the following degrees of freedom

$$\mathsf{tr}\left(\mathbb{B} - \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^\mathsf{T}\right) = K - 1.$$

□

Notice that, according to Lemma 9.2 (i), the statistic $\frac{SS_e}{N-K}$ is also an unbiased estimate of σ^2. On the other hand, by Lemma 9.3 (i), we have that $\frac{SS_A}{K-1}$ is an unbiased estimate of σ^2 **only under the null hypothesis**, while under the alternative we have that $\mathsf{E}\,\frac{SS_A}{K-1} > \sigma^2$. This brings us to the following test.

Test statistic:

$$F_A = \frac{SS_A/(K-1)}{SS_e/(N-K)}.$$

The null hypothesis will be rejected for **too large** values of F_A.

Theorem 9.4 Suppose that the model $\mathcal{F}_n$ and also the null hypothesis H_0 hold, then $F_A \sim F_{K-1,N-K}$.

Proof. The statistic F_A can be rewritten as

$$F_A = \frac{\frac{SS_A}{\sigma^2}/(K-1)}{\frac{SS_e}{\sigma^2}/(N-K)}.$$

From Lemma 9.3(ii) and Lemma 9.2(iii) we have that $\frac{SS_A}{\sigma^2} \sim \chi^2_{K-1}$ and $\frac{SS_e}{\sigma^2} \sim \chi^2_{N-K}$. The independence of random variables $\frac{SS_A}{\sigma^2}$ and $\frac{SS_e}{\sigma^2}$ follows from Lemma 9.3(ii). The theorem then follows from the definition of F-distribution (see Definition 2.5). □

Using Theorem 9.4 and the reasoning before it, we get the following.

Critical region:

$$H_0 \text{ is rejected } \Leftrightarrow F_A \geq F_{K-1,N-K}(1-\alpha),$$

where $F_{K-1,N-K}(1-\alpha)$ is $(1-\alpha)$-quantile of F-distribution with $K-1$ and $N-K$ degrees of freedom.

P-value: $1 - F^*(s)$, where s is the observed value of the test statistic F_A and F^* is the distribution function of the distribution $F_{K-1,N-K}$.

Remark.

- The above described method is called the *analysis of variance* or *ANOVA* due to the way the test statistic is constructed (we essentially compare two estimates of σ^2). However **the purpose of ANOVA is not to analyse the variance**. The test itself is called the *F-test of analysis of variance*.
- In Gaussian model with equal variances (i.e. in the model $\mathcal{F}_n$) we have that the F-test of analysis of variance is an exact test of equality of expected values in $K \geq 2$ independent samples.
- It can be shown that without the assumption of normality but with the assumption of equal variances (i.e. in the model $\mathcal{F}_{as}$), F-test of analysis of variance keeps the significance level at least asymptotically.

Remark. The results of analysis of variance are usually given in a table

Source of variation	Sum of squares	Degrees of freedom	Quotient	F
Group	SS_A	$K-1$	$\frac{SS_A}{K-1}$	$\frac{SS_A}{K-1} / \frac{SS_e}{N-K}$
Residual	SS_e	$N-K$	$\frac{SS_e}{N-K}$	
Total	SS_C	$N-1$		

Proposition 9.5 For $K = 2$ we have that

$$F_A = T^2_{n_1,n_2},$$

where F_A is the test statistic in analysis of variance and $T^2_{n_1,n_2}$ is the square of the test statistic of two-sample t-test for the case of equal variances (see Chapter 6.3).

Proof. Using (9.2), the numerator of the test statistic F_A can be rewritten as

$$\frac{SS_e}{N-K} = \frac{1}{n_1+n_2-2}\left((n_1-1)S_1^2 + (n_2-1)S_2^2\right) = S^2_{n_1,n_2}, \tag{9.8}$$

where S_k^2 is the sample variance in kth group.

After this, we notice that

$$\overline{Y}_{1+} - \overline{Y}_{++} = \overline{Y}_{1+} - \frac{n_1\overline{Y}_{1+} + n_2\overline{Y}_{2+}}{n_1+n_2} = \frac{n_2\left(\overline{Y}_{1+} - \overline{Y}_{2+}\right)}{n_1+n_2}.$$

Similarly

$$\overline{Y}_{2+} - \overline{Y}_{++} = \frac{n_1\left(\overline{Y}_{2+} - \overline{Y}_{1+}\right)}{n_1+n_2}.$$

So

$$\begin{aligned}\frac{SS_A}{K-1} &= n_1\left(\overline{Y}_{1+} - \overline{Y}_{++}\right)^2 + n_2\left(\overline{Y}_{2+} - \overline{Y}_{++}\right)^2 = \frac{\left(\overline{Y}_{1+} - \overline{Y}_{2+}\right)^2}{(n_1+n_2)^2}\left(n_1n_2^2 + n_2n_1^2\right)\\ &= \frac{n_1n_2\left(\overline{Y}_{1+} - \overline{Y}_{2+}\right)^2}{n_1+n_2} = \frac{\left(\overline{Y}_{1+} - \overline{Y}_{2+}\right)^2}{\frac{1}{n_1}+\frac{1}{n_2}}.\end{aligned} \tag{9.9}$$

Now, with the help of (9.8) and (9.9) we get that

$$F_A = \frac{SS_A/(K-1)}{SS_e/(N-K)} = \left(\frac{\overline{Y}_{1+} - \overline{Y}_{2+}}{\sqrt{S^2_{n_1,n_2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right)^2 = T^2_{n_1,n_2},$$

which was to be proven. □

So, if we compare only two groups, the analysis of variance is equivalent to the two-sample t-test with the assumption of equal variances (see Chapter 6.3). In this case, i.e. for $K = 2$, it is usually preferred to use the t-test, since it allows us to test one-sided hypothesis and we are able to easily derive a confidence interval from it.

On the other hand, if $K > 2$, then we are not able to talk about one-sided hypothesis or deal with this problem with the help of one confidence interval.

Remark. The analysis of variance is further generalised into multi-way analysis of variance. This generalisation is discussed in the class *Linear regression*. For example, two-way analysis of variance is based on dividing observations into KJ groups according to two categorical variables with K and J possible values. We are interested in whether one of those categorical variables influences the mean value of our observations.

Violation of assumptions

Violation of equality of variances. In this case, F-test of analysis of variance does not keep the exact or asymptotic significance level. However, published simulation studies show that if the number of observations is roughly the same in all groups, then the true significance level of F-test of analysis of variance is close to the required level.

For the case of unequal variances, a generalization of the test statistic and approximation of its distribution has been proposed already in Welch (1951). It is a generalization of the two-sample Welch test for a situation with more samples. The test statistic of this test takes into account the potentially different variances and it is given by the formula

$$F_w = \frac{\sum_{k=1}^K w_k \left(\overline{Y}_{k+} - \overline{Y}_w\right)^2}{K-1} \frac{1}{1 + 2\Lambda(K-2)},$$

where $w_k = \frac{n_k}{S_k^2}$ is a weight assigned to the kth group, $\overline{Y}_w = \frac{\sum_{k=1}^K w_k \overline{Y}_{k+}}{\sum_{k=1}^K w_k}$ is an estimate of the common mean value and

$$\Lambda = \frac{\sum_{k=1}^K \frac{1}{n_k - 1}\left(1 - \frac{w_k}{\sum_{j=1}^K w_j}\right)^2}{K^2 - 1}$$

is a certain correction, which is close to zero if we have large sample sizes in all of the groups.

It can be shown that under the null hypothesis, even without the assumption of equal variances (and also without the assumption of normality), it holds that

$$(K-1)F_w \xrightarrow{\mathsf{d}} \chi^2_{K-1}$$

where the sample sizes of all samples grow to infinity, i.e.

$$\min\{n_1,\dots,n_K\} \to \infty \text{ and simultaneously } \tfrac{n_k}{N} \to \lambda_k > 0,\ k \in \{1,\dots,K\}. \tag{9.10}$$

However, similarly as for the two-sample Welch t-test (see page 113), it is recommended, out of caution, to compare the test statistic F_w with quantiles of the F-distribution with $K-1$ and $1/(3\Lambda)$ degrees of freedom.

9.2. Multiple comparisons

In the analysis of variance we compare expected values of K groups. If the F-test of analysis of variance rejects the null hypothesis that all groups have the same expected value, we conclude that at least two groups differ in their expectations. However, we do not know, how many and what groups actually differ in their expectations.

If we wanted to compare the expectations of two groups, for example groups k and j, we would use two-sample t-test. We could perform two-sample t-tests for all $\frac{K(K-1)}{2}$ possible pairs of groups and test all hypotheses $H_0^{kj} : \mu_k = \mu_j$ on level α. However, the probability that at least one of these hypotheses will be rejected, under the condition that all of them hold, is not equal to α, it is in fact higher.

The problem of simultaneous testing of several hypotheses is usually called *multiple comparisons* or *multiple testing*.

The general problem of multiple testing can be formulated as follows. We want to test m null hypotheses $H_0^1,\dots,H_0^m$. To test hypothesis H_0^j we use the test statistic T_j with critical region C_j chosen such that each test has level α_0. Then we have that for all $j \in \{1,\dots,m\}$

$$\mathsf{P}_{H_0^j}\left[T_j \in C_j\right] = \alpha_0.$$

The probability of rejecting at least one hypothesis, under the condition that all hypotheses hold, is then

$$\mathsf{P}_{\bigcap_{j=1}^m H_0^j}\left(\bigcup_{j=1}^m [T_j \in C_j]\right) = \alpha_C.$$

Naturally, α_C is larger than α_0, usually distinctly. Our aim is to find, for a prescribed level α, tests $\widetilde{T}_j$ with critical region $\widetilde{C}_j$, so that

$$\mathsf{P}_{\bigcap_{j=1}^m H_0^j}\left(\bigcup_{j=1}^m \left[\widetilde{T}_j \in \widetilde{C}_j\right]\right) \le \alpha.$$

The situation is similar for confidence intervals. Let $B_1, \ldots, B_m$ be the confidence intervals for parameters $\theta_X^{(1)}, \ldots, \theta_X^{(m)}$, that satisfy

$$\mathsf{P}\left(B_j \ni \theta_X^{(j)}\right) = 1 - \alpha, \quad j \in \{1, \ldots, m\},$$

where $1 - \alpha$ is the prescribed probability of coverage.

Then typically

$$\mathsf{P}\left(B_1 \ni \theta_X^{(1)}, \ldots, B_m \ni \theta_X^{(m)}\right) < 1 - \alpha.$$

Our aim is to construct such confidence intervals $\widetilde{B}_1, \ldots, \widetilde{B}_m$, for which we will have

$$\mathsf{P}\left(\widetilde{B}_1 \ni \theta_X^{(1)}, \ldots, \widetilde{B}_m \ni \theta_X^{(m)}\right) \geq 1 - \alpha.$$

Such intervals $\widetilde{B}_1, \ldots, \widetilde{B}_m$ are called *simultaneous confidence intervals*.

In the following chapter, we will introduce one universal approach to this problem and after that a special method for comparing expected values of several independent random samples.

9.2.1. Bonferroni correction

We are given the total required level α and we want to guarantee that $\alpha_C \leq \alpha$. To do that, we use the following lemma.

Lemma 9.6 (Boole's inequality) For any random events $A_1, \ldots, A_m$ we have that

$$\mathsf{P}\left(\bigcup_{j=1}^{m} A_j\right) \leq \sum_{j=1}^{m} \mathsf{P}(A_j).$$

This inequality is trivial for $m = 2$ and it can easily be proven for higher m by mathematical induction.

We have that

$$\alpha_C = \mathsf{P}_{\bigcap_{j=1}^{m} H_0^j}\left(\bigcup_{j=1}^{m} [T_j \in C_j]\right) \leq m\alpha_0.$$

If we choose $\alpha_0 = \alpha/m$, then it must hold that $\alpha_C \leq \alpha$. Therefore, if me want to perform m tests and keep the total level of all tests (the probability of rejecting at least one hypothesis under the condition that they all hold) to be at least α, we perform individual tests on level α/m.

Similarly, if we want to construct m confidence intervals, which satisfy that all of them cover their respective parameters with probability at least $1 - \alpha$, it is enough to choose the individual intervals with probability of coverage at least $1 - \alpha/m$. This approach to multiple testing and construction of simultaneous confidence intervals is called the *Bonferroni correction*.

The advantage of Bonferroni correction is its simplicity and universality. On the other hand its disadvantage is that the correction of level α to α/m is almost always

too strict. Therefore, this method produces tests with low power and overly wide confidence intervals. Special methods of multiple testing, derived for specific problems (for example Tukey method described below) try to overcome these disadvantages of Bonferroni correction.

Application of Bonferroni correction to multiple comparisons in the analysis of variance looks as follows: we perform all $\frac{K(K-1)}{2}$ two-sample t-tests for all possible pairs of groups and test all hypotheses $H_0^{kj} : \mu_k = \mu_j$ on level $\frac{2\alpha}{K(K-1)}$. If at least one of these hypotheses is rejected, we proclaim the expected values of these two groups as significantly different on the total level α.

Imagine that we have chosen $\alpha = 0{,}05$ and we have $K = 6$ groups, then we perform 15 tests of equality of expected values for 15 different pairs of groups on level $0{,}05/15 \doteq 0{,}0033$. This significance level is so low that it may be difficult to find two significantly different groups, even though the F-test of analysis of variance rejects the hypothesis that the expected values of all groups are the same.

Remark. While using a method, which takes into account the problem of multiple testing, we sometimes define the so called *p-value adjusted for multiple comparisons*. For Bonferroni correction, this adjusted p-value can be easily computed as

$$\widetilde{p}_j = \min\{mp_j, 1\}, \quad j \in \{1, \dots, m\},$$

where p_j is the standard (non-adjusted) p-value of the jth test.

The end of self-study for week 12 (5.1.-9.1.).

9.2.2. Tukey method

This method is derived from normal (homoscedastic) model (9.1) assumed for the analysis of variance. Under the assumptions of this model, this new method has higher power and it produces shorter confidence intervals compared to Bonferroni correction.

Rem.: This part was not presented during 2020/21.

Let us have independent random variables $Z_j \sim \mathsf{N}(\mu, \sigma^2)$, where $j \in \{1, \dots, m\}$. Let S^2 be an estimate of the variance σ^2 such that S^2 is independent with $Z_1, \dots, Z_m$ and for some ν natural we have that $\frac{\nu S^2}{\sigma^2} \sim \chi^2_\nu$.

Let us define the *studentized range* as

$$Q = \frac{\max_{j\in\{1,\dots,m\}} Z_j - \min_{j\in\{1,\dots,m\}} Z_j}{S}.$$

It can be shown that the random variable Q has distribution which depends only on the values m and ν. Denote by $q_{m,\nu}(\alpha)$ the quantile function of this distribution. (We will not present formulas for density and cumulative distribution function here.)*

* Sometimes, studentized range is defined as $Q/\sqrt{2}$. One needs to be aware of this while using values of $q_{m,\nu}(\alpha)$ from the tables or software. To check correctness of our values, we can compare distribution Q with $m = 2$ with distribution $|T|$, where $T \sim t_k$. For our definition, these two distributions are the same.

Studentized range can be used to construct simultaneous confidence intervals for differences of expected values. This approach is called the *Tukey method*, the Tukey's range test or Tukey's HSD (honest significant difference) test.

Theorem 9.7 (Tukey) Let $Z_1, \ldots, Z_m$ be independent random variables with distributions $Z_j \sim \mathsf{N}(\mu_j, \sigma^2)$. Let S^2 be an estimate of the variance σ^2 such that S^2 is independent with $Z_1, \ldots, Z_m$ and for some ν natural it holds that $\frac{\nu S^2}{\sigma^2} \sim \chi^2_\nu$. Then

$$\mathsf{P}\Big[Z_k - Z_j - Sq_{m,\nu}(1-\alpha) \le \mu_k - \mu_k \le Z_k - Z_j + Sq_{m,\nu}(1-\alpha),\ \forall k \ne j \in \{1,\ldots,m\}\Big] = 1-\alpha.$$

The above theorem can be used for hypothesis testing as well. The hypothesis H_0^{kj} : $\mu_k = \mu_j$ is rejected, if $\big|Z_k - Z_j\big| > Sq_{m,\nu}(1-\alpha)$. The null hypothesis $H_0 : \mu_1 = \ldots = \mu_m$ is rejected on total significance level α, if for at least one pair $k \ne j$ we have that $\big|Z_k - Z_j\big| > Sq_{m,\nu}(1-\alpha)$.

Tukey theorem can be directly used for the problem of multiple comparison in the analysis of variance, if the sample sizes of all groups are the same, i.e. $n_1 = \cdots = n_K \equiv n$. Then it holds that $\overline{Y}_{1+}, \ldots, \overline{Y}_{K+}$ are independent random variables with distributions $\overline{Y}_{k+} \sim \mathsf{N}(\mu_k, \frac{\sigma^2}{n})$. We take $\frac{SS_e}{n(N-K)}$ as S^2, the estimate of $\frac{\sigma^2}{n}$. We have $\nu = N - K$. The null hypothesis $H_0^{kj} : \mu_k = \mu_j$ is rejected, if

$$\Big|\overline{Y}_{k+} - \overline{Y}_{j+}\Big| \ge \sqrt{\frac{SS_e}{N-K}}\sqrt{\frac{1}{n}}\, q_{K,N-K}(1-\alpha). \tag{9.11}$$

If the sample sizes of all groups are not the same, we cannot use Tukey theorem directly, since its assumptions do not hold. However, it can be shown that, if we replace the formula $\sqrt{\frac{1}{n}}$ in (9.11) by $\sqrt{\frac{1}{2n_k} + \frac{1}{2n_j}}$, then the total probability of rejecting one of the true hypotheses H_0^{kj} does not exceed α. So, Tukey method still works after this adjustment, however, it does become somewhat more conservative.

9.3. Kruskal-Wallis test

Kruskal-Wallis test is a generalization of the two-sample Wilcoxon test to compare $K \ge 2$ samples. The notation used in this section is the notation for K-sample problem defined at the beginning of this chapter.

Model: $\mathcal{F} = \big\{\exists g(\cdot)$ increasing function $\exists F$ continuous CDF $\exists \delta_1, \ldots, \delta_K \in \mathbb{R}$:
$$g(X_{k1}) \sim F_k, F_k(x) = F(x - \delta_k)\ \forall x \in \mathbb{R}, k \in \{1,\ldots,K\}\big\}$$

It is a model for K continuous distributions which are, after a suitable transformation g, mutually shifted in location.

The null hypothesis and alternative:

$$H_0 : \delta_1 = \cdots = \delta_K, \quad H_1 : \exists_{k \ne j}\ \delta_k \ne \delta_j.$$

Remark. If both model $\mathcal{F}$ and hypothesis H_0 hold, then the distributions of all groups are the same. In that case all K groups share coinciding characteristics.

Test statistic:

It can be shown that the test statistic of two-sample Wilcoxon test is equivalent to the numerator of the test statistic of two-sample t-test (i.e. difference of sample means), if, instead of the original observations, we use their ranks. We can try to proceed with the same logic and use the construction of the F-test of analysis of variance, where instead of using the observations in joint random sample $\boldsymbol{Y} = (Y_{11}, \dots, Y_{Kn_K})^\mathsf{T}$, we use their ranks $R_{11}, \dots, R_{Kn_K}$.

Then

$$\widetilde{SS}_A = \sum_{k=1}^{K} n_k \left(\overline{R}_{k+} - \overline{R}_{++}\right)^2,$$

where $\overline{R}_{k+} = n_k^{-1} \sum_{i=1}^{n_k} R_{ki}$ is the mean rank in kth group and

$$\overline{R}_{++} = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{n_k} R_{ki} = \frac{N+1}{2}$$

is the total mean rank.

Notice that in the standard analysis of variance the random variable $\frac{SS_e}{N-K}$ estimates the unknown variance $\mathsf{var}\,(Y_{ki}) = \sigma^2$. However, in the case of ranks we know, thanks to Theorem 2.16(iii), that under the null hypothesis $\widetilde{\sigma}^2 = \mathsf{var}\,(R_{ki}) = \frac{N^2-1}{12}$. Therefore, the candidate for our test statistic seems to be

$$\widetilde{Q} = \frac{\widetilde{SS}_A}{\widetilde{\sigma}^2} = \frac{12}{(N-1)(N+1)} \sum_{k=1}^{K} n_k \left(\overline{R}_{k+} - \tfrac{N+1}{2}\right)^2. \tag{9.12}$$

It can be shown that an asymptotic analogy of Lemma 9.3(iii) holds, i.e. under the null hypothesis and with increasing number of observations, see (9.10), we have that

$$\widetilde{Q} \xrightarrow{\mathsf{d}} \chi^2_{K-1}.$$

As we will show below, it holds that $\mathsf{E}\,\widetilde{Q} = (K-1)\frac{N}{N-1}$. However, since the expected value of the asymptotic distribution χ^2_{K-1} is $K-1$, we use the following test statistic (to improve on the asymptotic approximation)

$$Q = \frac{N-1}{N}\,\widetilde{Q} = \frac{12}{N(N+1)} \sum_{k=1}^{K} n_k \left(\overline{R}_{k+} - \tfrac{N+1}{2}\right)^2.$$

Critical region: Since large values of the test statistic indicate against the null hypothesis, we get for our asymptotic test the following rule

$$H_0 \text{ is rejected } \Leftrightarrow Q \geq \chi^2_{K-1}(1-\alpha).$$

The above stated test is called the *Kruskal-Wallis test*. Similarly as for the (two-sample) Wilcoxon test, it is possible to use exact critical values, which are tabulated, for small sample sizes (if there are no ties in our data).

Remark. Let $R_{k+} = \sum_{i=1}^{n_k} R_{ki}$. Then

$$\sum_{k=1}^{K} n_k \left(\overline{R}_{k+} - \tfrac{N+1}{2}\right)^2 = \sum_{k=1}^{K} \frac{1}{n_k}\left(R_{k+} - n_k \tfrac{N+1}{2}\right)^2$$
$$= \sum_{k=1}^{K} \frac{1}{n_k}\left(R_{k+}^2 - R_{k+} n_k (N+1) + n_k^2 \tfrac{(N+1)^2}{4}\right) = \sum_{k=1}^{K} \frac{R_{k+}^2}{n_k} - \frac{N(N+1)^2}{4}.$$

Hence the test statistic Q is often given in a computationally easier formula

$$Q = \frac{12}{N(N+1)} \sum_{k=1}^{K} \frac{R_{k+}^2}{n_k} - 3(N+1). \tag{9.13}$$

Remark. We will use the formula (9.13) to calculate the expected value of the test statistic Q **under the null hypothesis**. To do that, we will at first carry out the following calculation, using Theorem 2.16

$$\mathsf{E}\, R_{k+}^2 = \mathsf{var}\left(R_{k+}\right) + \left(\mathsf{E}\, R_{k+}\right)^2$$
$$= \sum_{i=1}^{n_k} \mathsf{var}\left(R_{ki}\right) + \sum_{i=1}^{n_k} \sum_{i'=1, i' \neq i}^{n_k} \mathsf{cov}\left(R_{ki}, R_{ki'}\right) + \left(\sum_{i=1}^{n_k} \frac{N+1}{2}\right)^2$$
$$= \frac{n_k(N^2-1)}{12} - \frac{n_k(n_k-1)(N+1)}{12} + \frac{n_k^2(N+1)^2}{4}.$$

Therefore (under the null hypothesis)

$$\mathsf{E}\, Q = \frac{12}{N(N+1)} \sum_{k=1}^{K} \frac{\mathsf{E}\, R_{k+}^2}{n_k} - 3(N+1)$$
$$= \frac{12}{N(N+1)} \sum_{k=1}^{K} \left[\frac{(N^2-1)}{12} - \frac{(n_k-1)(N+1)}{12} + \frac{n_k(N+1)^2}{4}\right] - 3(N+1)$$
$$= \frac{K(N-1)}{N} - \frac{(N-K)}{N} + 3(N+1) - 3(N-1) = K-1.$$

This corresponds to the expected value of the distribution χ^2_{K-1} and it is the reason why instead of the test statistic $\widetilde{Q}$ given by the formula (9.12) we use the test statistic Q.

Violation of assumptions

Ties due to rounding. We often see ties in our data because of rounding. The test statistic Q is then calculated using the so called average ranks. It can be shown that, under the null hypothesis, we have

$$\frac{Q}{1 - kor.} \xrightarrow[(9.10)]{d} \chi^2_{K-1},$$

where $kor.$ is the variance adjusting correction, given by the formula*

$$kor. = \frac{1}{N(N^2-1)} \sum_y (t_y^3 - t_y),$$

where t_y denotes the number of the random variables $Y_{11} \ldots, Y_{Kn_K}$ which attain the value y. It is worth noticing that, without this adjustment, the test would be (asymptotically) conservative.

The generalized location model does not hold. Notice at first that the test keeps the significance level (asymptotically), if the observations $Y_{11} \ldots, Y_{Kn_K}$ are independent and identically distributed. Therefore, the fact that the model does not hold has, similarly as for the two-sample Wilcoxon test (see Chapter 6.4), two unpleasant consequences regarding the behaviour of the test under the alternative:

1. **Interpretation problem** - if the generalized location model does not hold, then we are only able to conclude that the distributions are not the same in individual groups from the rejection of the null hypothesis. However, we generally cannot conclude that their expected values or medians are different.

2. **The power of the test** - similarly as for Mann-Whitney formulation of two-sample Wilcoxon test (see page 121), it can be shown that the Kruskal-Wallis test tests whether $\mathsf{P}[Y_{k1} < Y_{j1}] = 1/2$ holds for all $k, j \in \{1, \ldots, K\}$. In the generalized location model, if we have $\delta_k \neq \delta_j$, then indeed $\mathsf{P}[Y_{k1} < Y_{j1}] \neq 1/2$. However, if under the alternative we have some additional changes and not only the change in the location parameters, then it is not clear what will be the consequence of this on the power of the test.

* See for example Hollander et al. (2013), page 205.

Sample examples for the preparation for the exam.

Your solution to the "practical problem" should include a mathematical model, a null hypothesis, a test statistic and its exact (or asymptotic) distribution under the null hypothesis. It should also include a critical region or a formula for p-value and you should state whether the test is exact or asymptotic.

1. We have data about the height of 500 adult women and about the colour of their eyes, where we distinguish between brown, blue and green. Propose a suitable test to find out whether the height is connected with the colour of the eyes.

2. We have data about the salaries of 2 000 employees from the IT domain and about the region (8 possibilities) they live in. Propose appropriate method which will find, while keeping the required significance level, two regions whose salaries can be considered as different.

A. Appendix

A.1. χ^2- and t-distribution

Definition A.1 (χ^2-distribution) Let $X_1, \dots, X_k$ 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. Idempotent matrices

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

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

A.3. Transformation of the random variable with its cumulative distribution function

Lemma A.2 Let the random variable X have **continuous** distribution function F. Then the random variable $F(X)$ follows the uniform distribution on the interval $(0,1)$.

Proof. For $u \in (0,1)$ calculate

$$\mathsf{P}\big[F(X) \le u\big] = \mathsf{P}\big[X \le F^{-1}(u)\big] = F\big(F^{-1}(u)\big) = u,$$

where in the last equality we use the continuity of F. □

The following lemma is inverse to the lemma above. It is used for generating random variables. Note that it does not require the continuity of F.

Lemma A.3 Let the random variable U follows the uniform distribution on $(0,1)$ and F is a cumulative distribution function. Then the random variable $F^{-1}(U)$ follows the distribution given by F.

Proof. Let $x \in \mathbb{R}$. Calculate

$$\mathsf{P}\big[F^{-1}(U) \leq x\big] = \mathsf{P}\big[U \leq F(x)\big] = F(x).$$

□

A.4. Gama function and beta function

Gama function is for $z > 0$ definied as

$$\Gamma(z) = \int_0^\infty x^{z-1}\,\mathrm{e}^{-x}\,dx.$$

From the properties of the gama function it is often used that $\Gamma(n) = (n-1)!$.

Beta function is for $a, b > 0$ defined as

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx.$$

It holds that

$$B(a,b) = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}\,. \tag{A.1}$$

Bibliography

Bickel, P. J. and K. A. Doksum (2015). *Mathematical statistics: basic ideas and selected topics, volume I*. Chapman and Hall/CRC.

Chung, E. and J. P. Romano (2016). Asymptotically valid and exact permutation tests based on two-sample U-statistics. *Journal of Statistical Planning and Inference 168*, 97–105.

Hollander, M., D. A. Wolfe, and E. Chicken (2013). *Nonparametric statistical methods*. John Wiley & Sons, New York.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika 29*(3/4), 350–362.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika 38*, 330–336.