The rules of estimating.

Assessment method	**Proportion**	**Assessment and evaluation rules**
Work in semester	45	Consists of homework and two test papers. The maximum score for homework and practice grade is 25 points. The maximum score for each test paper is 10 points.
Final test	55	Written form, 10 questions.

The use of lecture notes, textbooks, as well as electronic devices for storing, processing, or transmitting information during control works and the final test is strictly prohibited. If the use of unauthorized materials (devices) is detected, an incident report is issued, and the student is dismissed from the auditory. As noted above, the final grade for the course is determined by a weighted sum of the results from in-class tests, computer practice, and the final examination. The total score is then converted into a grade according to the following scale:

Total Score	Traditional Grade
60–100	pass
less 60	fail

Homework and practice grades rules. Each task in a homework assignment is graded from 0 to 2 points. A score of 2 points is given for a completely correct solution, 1 point is awarded for minor calculation errors or slightly inaccurate reasoning that does not affect the correctness of the final result, and 0 points is given in all other cases. Each practice task is graded from 0 to 1. 1 point is given for a completely correct solution and 0 points otherwise. The final homework and practice grade, up to 25 points, is calculated proportionally to the ratio of the total number of points earned for all assignments submitted during the semester to the maximum possible score. Homework must be submitted to the instructor at the end of the lecture corresponding to the deadline. Late submission results in a 50% deduction from the earned points. Resubmission of homework is not allowed.

1 Introduction to Mathematical Statistics.

1.1 Problems of mathematical statistics.

Preface

Usually, a problem in probability theory is formulated as follows: *Given a random variable X and its distribution, compute the distribution of another random variable $f(X)$.*

From the point of view of mathematical statistics, something similar occurs: we are given a random variable X whose distribution is unknown, but we want to draw certain conclusions about this unknown distribution and about some functions of X.

In other words, we aim to make inferences about the surrounding world based on observations.

Remark

Mathematical statistics deals with the analysis of data that has a probabilistic nature.

Examples of Statistical Problems

1. Estimating the probability of success. Let $X_1, \ldots, X_n$ be independent trials.

$$P(X_i = 0) = p, \qquad P(X_i = 1) = 1 - p.$$

How can we estimate p?

2. Testing biological, medical, and other hypotheses. A classic example is Mendel's experiments on pea plant hybridization. In the second generation, he obtained 8,506 yellow peas and 23,174 green peas. His hypothesis stated that the ratio of phenotypes should be $1 : 3$.

In this case, the task of mathematical statistics is to determine whether the observed deviations from the expected ratio are statistically significant.

3. GWAS — Genome-Wide Association Study. It is known that there exists a relationship between genotype and phenotype. Moreover, a phenotype is determined largely (though not entirely) by genotype. The task (informally) is to learn how to predict a phenotype from a genotype.

4. Authorship attribution of texts. For example, one possible approach is to analyze stylistic or lexical features of the text to determine the most likely author.

1.2 Mathematical formulation of statistical problems

Definition 1.1. $(\Omega, \mathcal{F}, P)$ is a probability space. $X_1, \ldots, X_n$ is a set of identically distributed random variables. $\{P_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$ is a family of probability distributions.

It is assumed that the X_i are defined on the same set. Each X_i is distributed according to P_θ (but we do not know which θ).

Usually, X_i are called *observations*, and the set $X_1, \ldots, X_n$ is called a *sample*.

Definition 1.2. (Parameter estimation problem) $X_1, \ldots, X_n$ are observations, $X_i \sim P_\theta$, $\theta \in \Theta$.

It is required to construct:

(a) a function $T(X_1, \ldots, X_n)$ that would approximate θ well;

(b) intervals $(T_1(X_1,\dots,X_n), T_2(X_1,\dots,X_n))$ such that

$$P(T_1 < \theta < T_2) \geq \gamma \quad \text{(assuming } \theta \in \mathbb{R}\text{)}.$$

(assuming $\theta \in \mathbb{R}$).

Definition 1.3. (Statistical hypothesis testing problem)
$\Theta = \Theta_1 \cup \Theta_0$

$$H_0 \text{ and } H_1 \text{ are hypotheses}$$

$$H_0 : \theta \in \Theta_0, \qquad H_1 : \theta \in \Theta_1$$

It is required to distinguish between H_0 and H_1 based on $X_1,\dots,X_n$, i.e., to determine which hypothesis is true.

In general, $\Theta = \bigsqcup_{i=1}^{n} \Theta_i$, and $H_i : \theta \in \Theta_i$.

Example 1. $X_1,\dots,X_n$ are observations, $X_i \sim F$.

$$F_n(t) = \frac{1}{n}\#\{i : X_i < t\}$$

is the *empirical distribution function*: the proportion of observations taking values less than t.

Equivalently, we can write:

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i<t\}}.$$

Hence, we obtain the expectation:

$$\mathbb{E}F_n(t) = \frac{1}{n}\sum_{i=1}^{n} P(X_i < t) = F(t).$$

Thus, on average, we do not miss the true value.

Compute the variance:

$$DF_n(t) = \frac{1}{n^2}\sum_{i=1}^{n} D\mathbf{1}_{\{X_i<t\}} = \frac{1}{n^2}\sum_{i=1}^{n} F_X(t)(1-F_X(t)) = \frac{1}{n}F(t)(1-F(t)).$$

Therefore, by the law of large numbers, we obtain:

$$F_n(t) \xrightarrow{P} F(t).$$

Theorem 1.1. (Glivenko–Cantelli)

$$P\left(\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \to 0\right) = 1.$$

Proof. Omitted.

Remark. We have obtained that the empirical distribution function converges in probability to the distribution function of the considered random variable.

Thus, if we recall the main goal of mathematical statistics, we might think that we have learned how to solve it.

However, in reality, the convergence rate of the empirical distribution function is quite poor, which explains why statistical problems are not so easy to solve.

Example 2. Let us learn how to estimate the expectation.
$X_1, \dots, X_n$ are observations, and $m = \mathbb{E}X_i$.
How can we estimate m?
A natural idea is to estimate it by the sample mean:

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}.$$

Compute the expectation of $\overline{X}$:

$$\mathbb{E}\overline{X} = \frac{m + \cdots + m}{n} = m.$$

Thus, on average, we are not biased.
If, moreover, $\mathbb{E}X_i^2 < +\infty$, then

$$\overline{X} \xrightarrow{P} m$$

(by the law of large numbers).

Furthermore, if $\mathbb{E}X_i^2 < +\infty$, then asymptotic normality holds, i.e.

$$\sqrt{n}(\overline{X} - m) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\cdot)).$$

More generally, if we denote $m_k = \mathbb{E}X_i^k$, then we can consider

$$\frac{X_1^k + \cdots + X_n^k}{n}.$$

Example 3. Lets try to make the same for $S^2 = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i - \overline{X})^2$. And learn how to estimate the variance. Find $E_\theta(S^2)$ and show $E_\theta(S^2) = \frac{n-1}{n}\sigma^2$. In the future we show $D_\theta(S^2) = \frac{2\sigma^4}{n-1}$

Reminder. Convergence in distribution $X_n \xrightarrow{d} X$ means the pointwise convergence of the distribution functions F_{X_n} to F_X (in the case when the latter is continuous).

Remark to Glivenko-Cantelli Theorem The same question can be applied to **empirical measure**: $P_n(B) = \frac{1}{n}\sum\limits_{i=1}^{n} \mathbb{1}\{X_i \in B\}$. So the question is $\sup\limits_{B\in\mathfrak{B}} |P_n(B) - P_\theta(B)| \to 0$, as $n \to \infty$. with probability 1?
The answer depends on how "rich" the set $\mathfrak{B}$ is — the poorer it is, the more likely the answer is positive.
Moreover will be correct the next statement:

$$\sup_{f\in\mathfrak{F}} \left| \int\limits_{\mathfrak{X}} f dP_n - \int\limits_{\mathfrak{X}} f dP_\theta \right| \to 0 \quad \text{as} \quad n \to \infty \quad \text{with probability 1?}$$

1.3 Empirical moments.

In probability theory, all moments can be represented as $\int \psi(x) dF(x)$; of course in mathematical statistics we will match it to $\int \psi(x) dF_n(X) = \frac{1}{n}\sum\limits_i \psi(X_i)$.

- $\psi(x) = x, E(X) \leftrightarrow \overline{X} = n^{-1}\sum_i X_i$;
- $E(X) = a, \psi(x) = (x-a)^2 \quad D(X) \leftrightarrow S^2 = n^{-1}\sum(X_i - \overline{X})^2$.
- $\psi(x) = x^k, E(X^k) = \alpha_k \leftrightarrow a_k = \overline{a}_k = n^{-1}\sum X_i^k$;
- $\psi(x) = (x-a)^k, \mu_k = E(X-a)^2 \leftrightarrow m_k = n^{-1}\sum(X_i - \overline{X})^k$.

2 Statistical estimators

2.1 Requirements

Definition 4.1

$X_1, \ldots, X_n$ are a sample, $X_i \sim P_\theta, \theta \in \Theta$

$T_n = T(X_1, \ldots, X_n)$ is an arbitrary estimator.

Define the criteria for the "goodness" of our estimator:

1. Unbiasedness: $\forall\theta \quad \mathbb{E}_\theta T = \theta$
2. Asymptotic unbiasedness: $\forall\theta \quad \mathbb{E}_\theta T_n \to \theta$
3. Consistency: $\forall\theta \quad T_n \xrightarrow{P_\theta} \theta$
4. Strong consistency: $\forall\theta \quad T_n \xrightarrow{a.s.} \theta \quad with \quad P_\theta - prob. \quad 1$
5. Asymptotic normality: $\sqrt{n}(T_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta))$
6. Efficiency: T_1 is "more efficient" than T_2, if $\forall\theta \quad \mathbb{E}_\theta(T_1 - \theta)^2 \leq \mathbb{E}_\theta(T_2 - \theta)^2$

An estimate is called efficient if it is more efficient than all other estimates.

Note.

Nevertheless, it is not necessary for our estimators to satisfy all the criteria described above. Even the absence of unbiasedness usually does not make an estimator "bad".

Examples Check the properties for $\overline{X}$ and S^2.

End of the 2-nd Seminar.

Problem(Lemma). Lets T_n be an asymptotic unbiased estimator of some parameter θ and $DT_n \to 0$ as $n \to \infty$ $\forall\theta \in \Theta$. Prove that T_n is consistent.

• Firstly, consider the case of unbiased estimator.

Order Statistics and Related Concepts

- **Variational series:** After arranging a sample in ascending order, we obtain the variational (ordered) series:
$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$
where $x_{(k)}$ denotes the k-th order statistic.
- **Order statistic:** The k-th order statistic of a sample $x_1, x_2, \ldots, x_n$ is the k-th smallest value in the ordered sample:
$$X_{(k)} = \text{the } k\text{-th smallest value among } X_1, X_2, \ldots, X_n.$$
- **Minimum:** The smallest sample value:
$$X_{(1)} = \min(X_1, X_2, \ldots, X_n)$$
- **Maximum:** The largest sample value:
$$X_{(n)} = \max(X_1, X_2, \ldots, X_n)$$

- **Median:** The middle order statistic (for an odd sample size):

$$\text{Median} = X_{\left(\frac{n+1}{2}\right)}$$

For even n:

$$\text{Median} = \frac{1}{2}\left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}\right)$$

- **General k-th order statistic:** The value that occupies the k-th position in the ordered sample:

$$X_{(k)}, \quad k = 1, 2, \ldots, n$$

Probability density function of the k-th order statistic:

If $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables with cumulative distribution function $F(x)$ and probability density function $f(x)$, then the cumulative distribution function (CDF) of $X_{(k)}$ is given by:

$$G_{X_{(k)}}(x) = P(X_{(k)} \leq x) = \sum_{j=k}^{n} \binom{n}{j} [F(x)]^j [1 - F(x)]^{n-j}$$

This expression represents the probability that at least k of the sample values do not exceed x.

and the probability density function of the k-th order statistic $X_{(k)}$ is:

$$g_{X_{(k)}}(x) = k \cdot C_n^k [F(x)]^{k-1} [1 - F(x)]^{n-k} f(x) = \frac{n!}{(k-1)!(n-k)!} [F(x)]^{k-1} [1 - F(x)]^{n-k} f(x)$$

And

$$E(X_{(k)}) = \int_{-\infty}^{\infty} t g_k(t) dt.$$

Example. $F(t) = t, t \in [0, 1]$ Find the $E(X_{(k)})$.

Problems. 1. a) Find the joint distribution function of the minimum and maximum order statistics for a sample from a distribution with distribution function F.

b) For a sample with distribution function F, find the joint distribution function of $X_{(k)}$ and $X_{(m)}$, where $1 \leq k < m \leq n$.

2. Let $X_1, \ldots, X_n$ be a sample from a uniform distribution on the interval $[0, \theta]$. Verify the consistency and unbiasedness of the following estimators of the parameter θ:

(a) $2\overline{X}$;

(b) $\frac{n+1}{n} X_{(n)}$;

End of the 3-rd Seminar.

(c) $(n+1)X_{(1)}$;

(d) $X_{(1)} + X_{(n)}$;

(e) $\overline{X} + X_{(n)}/2$.

Problems. 1. a) Find the joint distribution function of the minimum and maximum order statistics for a sample from a distribution with distribution function F.

b) For a sample with distribution function F, find the joint distribution function of $X_{(k)}$ and $X_{(m)}$, where $1 \le k < m \le n$.

2. Let $X_1, \ldots, X_n$ be a sample from a uniform distribution on the interval $[0, \theta]$. Verify the consistency and unbiasedness of the following estimators of the parameter θ:

3. If $X_1, ..., X_n \sim F(t) = 1 - e^{-t}, \quad t \ge 0$. Find the distribution of $X_n - a_n$ where $a_n \to \infty \quad$ as $\quad n \to \infty$.

End of the 4-th Seminar.

Example: A consistent but not asymptotically unbiased estimator

Let $X_1, \ldots, X_n$ be i.i.d. random variables with finite mean μ. Define the estimator

$$T_n = \overline{X}_n + n \cdot \mathbf{1}_{A_n},$$

where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, and A_n is an event independent of the sample such that $\Pr(A_n) = \frac{1}{n}$.

1. Consistency

We show that $T_n \xrightarrow{p} \mu$.

For any $\varepsilon > 0$,

$$\Pr(|T_n - \mu| > \varepsilon) \le \Pr(|\overline{X}_n - \mu| > \varepsilon) + \Pr(A_n).$$

Justification: Let

$$B = \{|T_n - \mu| > \varepsilon\}, \quad C = \{|\overline{X}_n - \mu| > \varepsilon\}, \quad A = A_n.$$

If $\omega \notin A$, then $T_n(\omega) = \overline{X}_n(\omega)$; hence if $\omega \in B$, we must have $\omega \in C$. Thus $B \subset C \cup A$, implying

$$\Pr(B) \le \Pr(C \cup A) \le \Pr(C) + \Pr(A).$$

By the Law of Large Numbers, $\Pr(|\overline{X}_n - \mu| > \varepsilon) \to 0$, and since $\Pr(A_n) = 1/n \to 0$,

$$\Pr(|T_n - \mu| > \varepsilon) \to 0.$$

Therefore, T_n is **consistent** for μ.

2. Asymptotic Unbiasedness

Compute the expectation:

$$\mathbb{E}[T_n] = \mathbb{E}[\overline{X}_n] + n \Pr(A_n) = \mu + n \cdot \frac{1}{n} = \mu + 1.$$

Hence the bias is

$$\mathbb{E}[T_n] - \mu = 1,$$

which does not vanish as $n \to \infty$. Therefore, T_n is **not asymptotically unbiased**, even though it is consistent.

3. Intuition

Consistency concerns convergence in probability to the true parameter, not convergence of expectations. Here, T_n coincides with the sample mean (a good estimator) with high probability, but occasionally "jumps" by n with small probability $1/n$. These rare but large deviations do not affect convergence in probability but keep the bias fixed at 1.

2.2 Quantiles

2.2.1 Definition of a quantile

Definition 2.1.

X is a random variable, $X \sim F$.

$p \in (0,1)$, $\exists t : F(t) = p$, and around t, the function F is continuous and strictly increasing.

The quantile of order p is $\zeta_p = F^{-1}(p)$.

Note.

The question arises what to do if the distribution function is not continuous (which is exactly what we face in mathematical statistics).

Hence, the goal is to learn to estimate quantiles.

Definition 2.2.

$X_1, \ldots, X_n$ are observations.

Consider a random variable ξ taking values $X_1, \ldots, X_n$ with probability $\frac{1}{n}$.

Then our sample moments are just the expectation for ξ.

Indeed, earlier we understood that EX can be estimated by the sample mean $\overline{X}$, which exactly matches $E\xi$.

Similarly, we learned to estimate $F(t)$ by the empirical distribution function $F_n(t)$, which coincides with the distribution function F_ξ.

Then the sample quantile of order p is $Z_p = X_{[(n+1)p]}$ (or sometimes $X_{[np]+1}$), [..]-integer part i.e., the order statistic with index $[np]+1$.

2.2. Estimation of population quantiles

Theorem 2.1.

$X_1, \ldots, X_n$ are observations, $X_i \sim F$, $f = F'$.

$p \in (0,1)$, ζ_p is the population quantile of order p, and $f(\zeta_p) \neq 0$.

Then the sample quantile is asymptotically normal, i.e.,

$$\sqrt{n}(Z_p - \zeta_p) \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{f^2(\zeta_p)}\right).$$

2.2. Estimation of Population Quantiles

Theorem 2.1.

Let $X_1, \ldots, X_n$ be observations, $X_i \sim F$, $f = F'$.

Let $p \in (0,1)$, and ζ_p be the population quantile of order p, with $f(\zeta_p) \neq 0$.

Then the sample quantile is asymptotically normal, i.e.,

$$\sqrt{n}(Z_p - \zeta_p) \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{f^2(\zeta_p)}\right)$$

Proof.

Let for convenient$k := [np] + 1$.

$$\begin{aligned}
\mathbb{P}\left(\sqrt{n}(Z_p - \zeta_p) < t\right) &= \mathbb{P}\left(Z_p < \zeta_p + \frac{t}{\sqrt{n}}\right) = \mathbb{P}\left(X_{(k)} < \zeta_p + \frac{t}{\sqrt{n}}\right) \\
&= \mathbb{P}\left(\text{at least } k \text{ observations } X_i < \zeta_p + \frac{t}{\sqrt{n}}\right) \\
&= \mathbb{P}\left(\sum_{i=1}^{n} \mathbf{1}\{X_i < \zeta_p + \frac{t}{\sqrt{n}}\} \geq k\right)
\end{aligned}$$

On the left, we have the sum of independent identically distributed random variables. We can use the Central Limit Theorem. Let's compute the mean and variance:

$$\mathbb{E}\left[\mathbf{1}\left\{X_i < \zeta_p + \frac{t}{\sqrt{n}}\right\}\right] = \mathbb{P}\left(X_i < \zeta_p + \frac{t}{\sqrt{n}}\right) = F\left(\zeta_p + \frac{t}{\sqrt{n}}\right) = F(\zeta_p) + f(\zeta_p)\frac{t}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right)$$

Recall that $F(\zeta_p) = p$ and $F'(\zeta_p) = f(\zeta_p)$:

$$\mathbb{E}\left[\mathbf{1}\left\{X_i < \zeta_p + \frac{t}{\sqrt{n}}\right\}\right] = p + f(\zeta_p)\frac{t}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right)$$

$$\text{Variance:} \quad \mathbb{D}\left[\mathbf{1}\left\{X_i < \zeta_p + \frac{t}{\sqrt{n}}\right\}\right] = F\left(\zeta_p + \frac{t}{\sqrt{n}}\right)\left[1 - F\left(\zeta_p + \frac{t}{\sqrt{n}}\right)\right] = p(1-p)(1+o(1))$$

Let $S_n = \sum_{i=1}^n \mathbf{1}\{X_i < \zeta_p + \frac{t}{\sqrt{n}}\}$. Introducing $\Phi(t)$, the distribution function for $\mathcal{N}(0,1)$, by the CLT we have:

$$\begin{aligned}\mathbb{P}(S_n \geq k) &= 1 - \mathbb{P}(S_n < k) = 1 - \mathbb{P}\left(\frac{S_n - \mathbb{E}S_n}{\sqrt{\mathbb{D}S_n}} < \frac{k - \mathbb{E}S_n}{\sqrt{\mathbb{D}S_n}}\right) \\ &= 1 - \Phi\left(\frac{k - np - \sqrt{n}tf(\zeta_p) + o(\sqrt{n})}{\sqrt{np(1-p)(1+o(1))}}\right) + o(1)\end{aligned}$$

For the index,

$$k - np = 1 + np - np = 1$$

and for large n:

$$\frac{1 - tf(\zeta_p) + o(\sqrt{n})}{\sqrt{np(1-p)}} = -\frac{tf(\zeta_p)}{\sqrt{p(1-p)}}$$

So

$$\mathbb{P}\left(\sqrt{n}(Z_p - \zeta_p) < t\right) = \Phi\left(\frac{tf(\zeta_p)}{\sqrt{p(1-p)}}\right) = \Phi\left(\frac{t}{\sigma}\right) \text{ where } \sigma = \frac{\sqrt{p(1-p)}}{f(\zeta_p)}$$

This is exactly what was required to prove. □

Statement 2.2.

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \implies T_n - \theta \xrightarrow{P} 0$$

Proof.

Let's write the condition for convergence in probability:

$\forall \varepsilon > 0 : \mathbb{P}(|T_n - \theta| > \varepsilon) \to 0$

$\mathbb{P}(|T_n - \theta| > \varepsilon) = \mathbb{P}(\sqrt{n}|T_n - \theta| > \sqrt{n}\varepsilon) = 1 - \mathbb{P}(-\sqrt{n}\varepsilon \leq \sqrt{n}(T_n - \theta) \leq \sqrt{n}\varepsilon)$

By convergence in distribution:

$\mathbb{P}(|T_n - \theta| > \varepsilon) = 1 - \left(\Phi\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right) - \Phi\left(\frac{-\sqrt{n}\varepsilon}{\sigma}\right) + o(1)\right) \to 0$ as $n \to \infty$.

3 Fisher's Lemma.

Random gaussian vector (normal)

Reminder.

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$$

where:

- $\boldsymbol{\mu} \in \mathbb{R}^n$ is the mean vector
- $\Sigma \in \mathbb{R}^{n \times n}$ is the covariance matrix

The probability density function is:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Let A be a matrix of size $m \times n$ and $\mathbf{X} \in \mathbb{R}^n$, then:

$$\mathbf{Y} = A\mathbf{X} + \mathbf{b}$$

is also a random vector, with:

$$\mathbb{E}[\mathbf{Y}] = A\mathbb{E}[\mathbf{X}] + \mathbf{b} \quad \text{and} \quad \mathrm{Cov}(\mathbf{Y}) = A\mathrm{Cov}(\mathbf{X})A^T$$

then

$$Y \sim N(b + A\mu, A\Sigma A^T).$$

Characteristic function.

Definition 3.2. The characteristic function of a random vector is

$$\varphi_\xi(t) = E\, e^{i(t,\xi)}.$$

Proposition 3.1. Let us find the characteristic function of a standard Gaussian vector:

$$\varphi_X(t) = E\, e^{i(t,X)} = E\, e^{i\sum_{j=1}^n t_j X_j} = \prod_{j=1}^n E\, e^{it_jX_j} = \prod_{j=1}^n e^{-\frac{1}{2}t_j^2} = e^{-\frac{1}{2}(t,t)}.$$

For an arbitrary vector with a Gaussian distribution, we have:

$$\varphi_Y(t) = E\, e^{i(t,Y)} = E\, e^{i(t,a+LX)} = e^{i(t,a)}\, E\, e^{i(t,LX)} = e^{i(t,a)}\, E\, e^{i(L^Tt,X)} = e^{i(t,a)}\, E\, e^{-1/2(L^Tt,L^Tt)}.$$

Then, using the form of the characteristic function for the standard Gaussian vector, we obtain the characteristic function for an arbitrary Gaussian vector Y:

$$\varphi_Y(t) = e^{i(t,a)}\, e^{-\frac{1}{2}(L^Tt,L^Tt)} = e^{i(t,a)}\, e^{-\frac{1}{2}(t,LL^Tt)},$$

where $Y = a + LX$.

End of the 5-th Seminar.

3.1 Fisher's Lemma.

Reminder. A matrix C is orthogonal if $C^TC = CC^T = E$, i.e., if $C^T = C^{-1}$.

Proposition 4.1. Let $X = (X_1, \ldots, X_n)^T$, where X_i are independent components, $X_i \sim \mathcal{N}(0, \sigma^2)$. Let C be an orthogonal matrix. Then $CX \stackrel{d}{=} X$.

Proof.

$$E\, e^{i(t,X)} = e^{-\frac{\sigma^2}{2}(t,t)},$$

$$E\, e^{i(t,CX)} = e^{-\frac{\sigma^2}{2}(C^Tt,C^Tt)} = e^{-\frac{\sigma^2}{2}(t,CC^Tt)} = e^{-\frac{\sigma^2}{2}(t,t)}.$$

Thus, the characteristic functions of X and CX coincide. Hence, by the uniqueness theorem, their distributions coincide as well. □

Definition 4.2 (Chi-squared distribution). Let $X_1, \ldots, X_n$ be independent random variables, $X_i \sim \mathcal{N}(0, 1)$. Then the chi-squared distribution is defined as

$$\chi^2_n = X_1^2 + \ldots + X_n^2.$$

If $X \sim \chi^2_k$, then its probability density function is

$$f_{\chi^2_k}(x) = \begin{cases} \dfrac{1}{2^{k/2}\Gamma(k/2)}\, x^{\frac{k}{2}-1}e^{-x/2}, & x > 0, \\ 0, & x \le 0. \end{cases}$$

Definition 4.3 (Student's t-distribution). Let $X, X_1, \ldots, X_n$ be independent random variables, $X_i \sim \mathcal{N}(0, 1)$. Then the Student's t-distribution is defined as

$$T_n = \frac{X}{\sqrt{\frac{1}{n}(X_1^2 + \ldots + X_n^2)}}.$$

If $T \sim t_k$, then its probability density function is

$$f_{t_k}(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}}, \quad t \in \mathbb{R}.$$

Fisher's Lemma.

Let $X_1, \ldots, X_n$ be independent observations, $X_i \sim \mathcal{N}(\theta, \sigma^2)$.

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n}$$

$$S^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})^2$$

Then:

1. $\sqrt{n}\frac{\overline{X}-\theta}{\sigma} \sim \mathcal{N}(0, 1)$
2. $\frac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$
3. $\overline{X}$ and S^2 are independent
4. $\sqrt{n-1}\frac{\overline{X}-\theta}{S} \sim t_{n-1}$

Proof.
1. $\overline{X}$ is a linear combination of independent normal random variables, hence it is normally distributed. This can be easily seen by considering the characteristic function of $\overline{X}$.
In particular, $\overline{X} \sim \mathcal{N}\left(\mathbb{E}X, \mathbb{D}X\right) = \mathcal{N}\left(\theta, \frac{\sigma^2}{n}\right)$.
Therefore,

$$\frac{\overline{X} - \theta}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

2, 3. Assume $\theta = 0$.
Shifting all observations by θ does not change S^2.
However, $\overline{X}$ does change, but only by a constant. This means the independence of S^2 and $\overline{X}$ is not affected.
Consider the vector $X = (X_1, \ldots, X_n)^T$, where $X_i \sim \mathcal{N}(0, \sigma^2)$, and the orthogonal matrix C of the form:

$$C = \begin{pmatrix} \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ * & \cdots & * \end{pmatrix}$$

It is important that the first row of this matrix has this exact form. Such a matrix always exists.
Then define $Y = CX$.
Attempt to express sample mean and sample variance via Y_k:

$$Y_1 = \frac{1}{\sqrt{n}}(X_1 + \ldots + X_n) = \sqrt{n}\ \overline{X}$$

Expanding the sample variance, utilizing the fact that C preserves the vector length:

$$S^2 = \frac{1}{n}\sum_{k=1}^{n} X_k^2 - \overline{X}^2 = \frac{1}{n}\sum_{k=2}^{n} Y_k^2$$

Thus both claims are proved.
The second follows from the elementary substitution instead of S^2 of the obtained result.
The third follows because $\overline{X}$ and S^2 are expressed via different Y_k, hence independent.
4.

$$T_n = \sqrt{n}\frac{\overline{X}}{S} = \frac{X}{\sqrt{\frac{1}{n-1}(X_1^2 + \ldots + X_n^2)}},$$

where $X, X_i \sim \mathcal{N}(0,1)$ and independent.
Let

$$X = \frac{\overline{X} - \theta}{\sqrt{\frac{\sigma^2}{n}}}.$$

Then

$$T_{n-1} \sim \frac{\sqrt{n}\frac{\overline{X}-\theta}{\sigma}}{\sqrt{\frac{1}{n-1}n\frac{S^2}{\sigma^2}}} = \sqrt{n-1}\frac{\overline{X} - \theta}{S},$$

and the distribution is the Student's t.

Confidence intervals.

Let $\{F_\theta,\ \theta \in \Theta\}$ be some parametric family of distributions, $\Theta \subseteq \mathbb{R}$, and let $X_1, \ldots, X_n$ be a sample from the distribution F_θ.

Let $\theta_n^- = \theta_n^-(X_1, \ldots, X_n)$ and $\theta_n^+ = \theta_n^+(X_1, \ldots, X_n)$ be some statistics. The random interval (θ_n^-, θ_n^+) is called a *confidence interval of level* $1 - \varepsilon$ if

$$P_\theta\{\theta \in (\theta_n^-, \theta_n^+)\} = 1 - \varepsilon.$$

The random interval (θ_n^-, θ_n^+) is called an *exact confidence interval of level* $1 - \varepsilon$ if, for all $\theta \in \Theta$,

$$P_\theta\{\theta \in (\theta_n^-, \theta_n^+)\} = 1 - \varepsilon.$$

To construct an exact confidence interval, the following approach is usually used. A function $G(X_1, \ldots, X_n, \theta)$ is chosen such that the distribution of $P_\theta\{G(X_1, \ldots, X_n, \theta) \le t\}$ does not depend on the parameter θ (the distribution is free of θ). The function G must be monotonic and invertible with respect to the argument θ for any fixed sample values $X_1, \ldots, X_n$.

Let, for definiteness, the function G be increasing. Denote by $t(X_1, \ldots, X_n, y)$ the inverse of the function $G(X_1, \ldots, X_n, \theta)$ with respect to θ. Then the confidence interval of level $1 - \varepsilon$ has the form

$$(t(X_1, \ldots, X_n, y^-),\ t(X_1, \ldots, X_n, y^+)),$$

where the values y^- and y^+ are found (generally speaking, not uniquely) from the equation

$$P_\theta\{y^- < G(X_1, \ldots, X_n, \theta) < y^+\} = 1 - \varepsilon.$$

Confidence Intervals for Parameters of the Normal Distribution

Let

$$X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$$

be a sample of a normal distribution. and α is a significant level.

Denote:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2, \qquad \widehat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

1. Confidence interval for the mean μ when σ is known

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1),$$

$$\boxed{\mu \in \left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right),}$$

2. Confidence interval for the mean μ when σ is unknown

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

$$\boxed{\mu \in \left(\bar{X} - t_{1-\alpha/2,\,n-1}\frac{\widehat{S}}{\sqrt{n}},\ \bar{X} + t_{1-\alpha/2,\,n-1}\frac{\widehat{S}}{\sqrt{n}}\right),}$$

End of the 6-th Seminar.

3. Confidence interval for variance σ^2 when μ is unknown

$$\frac{nS^2}{\sigma^2} \sim \chi^2_{n-1},$$

$$\boxed{\sigma^2 \in \left(\frac{nS^2}{\chi^2_{1-\alpha/2,\, n-1}},\ \frac{nS^2}{\chi^2_{\alpha/2,\, n-1}} \right),}$$

4. Confidence interval for variance σ^2 when μ is known

$$\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2} \sim \chi^2_n,$$

$$\boxed{\sigma^2 \in \left(\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{1-\alpha/2,\, n}},\ \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{\alpha/2,\, n}} \right),}$$

5. Confidence interval for standard deviation σ

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

$$\boxed{\sigma \in \left(\sqrt{\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{1-\alpha/2,\, n}}},\ \sqrt{\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{\alpha/2,\, n}}} \right),}$$

Problems.

1. A factory monitors the weight of packages of grain. It is assumed that the package weight X (in grams) follows a normal distribution.

A random sample of $n = 16$ packages was taken, and the following data were obtained:

$$\bar{x} = 1002 \text{ g}, \quad \widehat{s} = 4.05 \text{ g}.$$

1. Construct a 95% confidence interval for the true mean weight μ if the population variance is **unknown**.
2. Construct a 95% confidence interval for μ if the population standard deviation is **known** and equal to $\sigma = 4$ g.

Answers:

2. Construct 95% confidence intervals for the population variance σ^2 in the following two cases:

1. when the population mean μ is **unknown**;
2. when the population mean μ is **known** and =1002.

Answers:

Remark. If our sample is not from normal distribution, than we can use the next:
1. Asymptotic approach – for a sufficiently large sample (n), the sampling distribution of the statistic (e.g., the sample mean) is approximately normal by the Central Limit Theorem, allowing the use of standard confidence intervals.

2. Using "statistics whose distribution does not depend on the parameter" – select a statistic (e.g., the median, rank-based, or bootstrap statistics) whose distribution does not depend on the unknown parameter, and construct the interval based on it.
Def A random interval (θ_n^-, θ_n^+) is called an *asymptotic confidence interval of level* $1-\varepsilon$ if for all θ

$$\liminf_{n\to\infty} \mathbb{P}_\theta\{\theta \in (\theta_n^-, \theta_n^+)\} \geq 1-\varepsilon.$$

A random interval (θ_n^-, θ_n^+) is called an *asymptotically exact confidence interval of level* $1-\varepsilon$ if for all θ

$$\lim_{n\to\infty} \mathbb{P}_\theta\{\theta \in (\theta_n^-, \theta_n^+)\} = 1-\varepsilon.$$

Examples. 1. $Bin(\theta, n)$.
2. $Exp(\theta)$
a) using $\overline{X}$,
b) using $X_{(n)}$.
3. $Unif[0, \theta]$ **End of the 7-th Seminar.**

σ Rule: 1, 2, and 3 σ

Sigma (σ) is the *standard deviation*, a measure of how much the data deviate from the mean (μ).

- Small $\sigma \rightarrow$ data are tightly grouped around the mean.
- Large data $\sigma \rightarrow$ are widely distributed.

Empirical (68–95–99.7) Rule

The empirical rule describes how much of the data in a **normal distribution** fall within 1, 2, and 3 standard deviations of the mean.

Range	Interval	Percent of Data	Explanation
1σ	$\mu \pm 1\sigma$	$\approx 68.3\%$	About two-thirds of all data near the mean
2σ	$\mu \pm 2\sigma$	$\approx 95.4\%$	Almost all data, except for extreme 5%
3σ	$\mu \pm 3\sigma$	$\approx 99.7\%$	Nearly all data, almost no outliers

Example

Assume human height is normally distributed:

$$\mu = 170\,\text{cm}, \quad \sigma = 10\,\text{cm}$$

Then:

$$1\sigma : 160\text{–}180\,\text{cm} \Rightarrow 68\%$$
$$2\sigma : 150\text{–}190\,\text{cm} \Rightarrow 95\%$$
$$3\sigma : 140\text{–}200\,\text{cm} \Rightarrow 99.7\%$$

Construction of Estimators. Method of Moments

Motivation

So far, we have tried to recover information about the distribution from the observations we have.

For example, we have already found that the distribution function is well approximated by the empirical distribution function, and the expectation — by the sample mean.

Moreover, in the same way, one can try to estimate various functionals (for example, integral ones):

$$\int g(x)\,dF(x) = \int g(x)\,dF_n(x) = \frac{1}{n}\sum_{k=1}^{n} g(X_k)$$

We would like to use the current knowledge to construct some method that would allow us to approximate well the parameters of interest.

Problem

Let us return to the general problem. We will consider the simple case when all observations and parameters exist in $\mathbb{R}$.

Let $X_1, \ldots, X_n \sim P_\theta$ be observations, $\theta \in \Theta$.

As always, the parameter θ is unknown to us, but we want to estimate it.

Solution

Consider some function $g : \mathbb{R} \to \mathbb{R}$, such that the population moment $E_\theta g(X) =: h(\theta)$ is a sufficiently nice function. We will require that h be invertible.

Assume that our sample moment $E_\theta g(X)$ is well approximated by the sample mean. In fact, in the ideal case, we would have the following equality:

$$\frac{1}{n}\sum_{k=1}^{n} g(X_k) = \int g(x)\,dF_n(x) = \int g(x)\,dF_\theta(x) = E_\theta g(X) = h(\theta)$$

Hence arises the idea of estimating the parameter as

$$\hat{\theta} = h^{-1}\left(\frac{1}{n}\sum_{k=1}^{n} g(X_k)\right).$$

This idea indeed seems quite reasonable. Nevertheless, it is necessary to study the properties possessed by such an estimator of parameters, as well as to check how well it works on some known distributions.

Remark.

Usually, $g(x)$ is chosen as some simple and convenient function. The most typical option is $g(x) = x^k$, since in this case we have to work with the moments EX^k, which we already know a lot about.

Remark.

If the parameter lies in a higher-dimensional space (i.e., $\theta \in \mathbb{R}^d$, where $d > 1$), such an approach still has the right to exist. However, now to obtain $\hat{\theta}$, we will need at least d different equations.

Example.

1. Consider the normal distribution $\mathcal{N}(\theta, \sigma^2)$.

In this case, we have two unknown parameters. To estimate them, we take the first two moments, i.e., we choose $g_1(x) = x$ and $g_2(x) = x^2$:

$$\begin{cases} E_{(\theta,\sigma)}X = \frac{1}{n}\sum_{k=1}^{n} X_k & \Rightarrow \hat{\theta} = \bar{X} \\ E_{(\theta,\sigma)}X^2 = \frac{1}{n}\sum_{k=1}^{n} X_k^2 & \Rightarrow \hat{\sigma}^2 + \hat{\theta}^2 = \frac{1}{n}\sum_{k=1}^{n} X_k^2 \end{cases} \iff \begin{cases} \hat{\theta} = \bar{X} \\ \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n} X_k^2 - \bar{X}^2 \end{cases}$$

We obtained that the parameters of the normal distribution can be estimated as $\bar{X}$ and S^2. But we also know that these estimators approximate well the expectation and variance of the underlying random distribution.

And since the parameters of the normal distribution are exactly its expectation and variance, we actually obtained a good estimate of these parameters.

Hence, there is hope that this method really works well.

2. Another example — the Cauchy distribution.

   Recall that the Cauchy distribution is a class of absolutely continuous distributions.

   The density of such distributions has the form

$$f(x) = \frac{1}{\pi}\left[\frac{1}{(x-\theta)^2 + 1}\right],$$

   where θ is a location parameter.

   This distribution is remarkable in that it does not have a mathematical expectation. Therefore, estimating this distribution using the first moment is simply impossible.

   Now, we only need to check the properties possessed by the estimators we derived. But before we proceed to consider them, we prove the following statement:

Proposition 4.2.

$$\sqrt{n}\frac{T_n - a}{\sigma} \to \mathcal{N}(0, 1)$$

if f is a continuous function and $f'(a) \neq 0$. Then $f(T_n)$ is asymptotically normal.

Proof.

$$f(T_n) = f(a) + (T_n - a)f'(a) + o(T_n - a)$$
$$\sqrt{n}\,(f(T_n) - f(a)) = \sqrt{n}(T_n - a)f'(a) + o(\sqrt{n}(T_n - a))$$
$$\sqrt{n}(T_n - a) \xrightarrow{d} N(0, \sigma^2)$$
$$o(\sqrt{n}(T_n - a)) = o(1) \cdot \sqrt{n}(T_n - a) \xrightarrow{p} 0$$

That is,

$$\sqrt{n}\,(f(T_n) - f(a)) \xrightarrow{d} N(0, \sigma^2 \cdot (f'(a))^2)$$

Remark. It is important that $f'(a)$ does not vanish, since otherwise it would be necessary to expand further terms in the Taylor series.

Properties.

Let's examine what happens with the properties of such estimates:

1. Unbiasedness.

Suppose we wish to check the equality $\mathbb{E}\hat{\theta} \stackrel{?}{=} \theta$.

Let $Z := \frac{1}{n}\sum_{k=1}^{n} g(X_k)$. Then,

$$\mathbb{E}\hat{\theta} = \mathbb{E}h^{-1}(Z)$$

$$\theta = h^{-1}(h(\theta)) \stackrel{\text{def}}{=} h^{-1}(\mathbb{E}g(X)) = h^{-1}\left(\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} g(X_k)\right]\right) = h^{-1}\left(\mathbb{E}Z\right)$$

Thus, it is required to verify the equality $h^{-1}(\mathbb{E}Z)$ and $\mathbb{E}h^{-1}(Z)$.

But if h^{-1} is strictly convex or strictly concave (which is a quite common situation), then by Jensen's inequality we obtain a strict inequality.

Therefore, in general, unbiasedness may not exist.

2. Consistency.

If $\mathbb{D}g(X) < \infty$ and h^{-1} is continuous, then we have:

$$\frac{1}{n}\sum_{k=1}^{n} g(X_k) \xrightarrow{p} \mathbb{E}g(X) = h(\theta)$$

(by the law of large numbers)

$$\hat{\theta} = h^{-1}\left(\frac{1}{n}\sum_{k=1}^{n} g(X_k)\right) \xrightarrow{p} h^{-1}(h(\theta)) = \theta$$

End of the 8-th Seminar.

---

3. Asymptotic normality.

If $\mathbb{D}g(X) < \infty$, and h^{-1} is continuously differentiable and does not vanish, then

$$\frac{\frac{1}{n}\sum_{k=1}^{n} g(X_k) - \mathbb{E}g(X)}{\sqrt{\mathbb{D}g(X)/n}} \xrightarrow{d} N(0,1)$$

(central limit theorem)

Applying h^{-1} to our estimator, from what is shown above we get:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \mathbb{D}g(X) \cdot ((h^{-1})'(h(\theta)))^2)$$

Problems.

Find the estimators of unknown parameter θ for $X_1, ..., X_n \sim F_\theta(\cdot)$ using the moment's method, if:
1. $Unif[-\theta, \theta]$;
2. $Unif[\theta, \theta + 1]$;
3. $Unif[a, b]$;
4. $Exp(1/\theta)$; $\quad g(x) = x^k$;
5. $Unif[0, \theta]$, $\quad g(x) = x^k$.

End of the 9-th Seminar.

Maximum likelihood estimation.

We have already found that the method of moments is quite capable of providing reasonable estimates. However, it requires us to compute the inverse function, which, for an arbitrary distribution, is very difficult and inconvenient. In some cases, it cannot provide any estimate at all (for example, when moments do not exist).

Therefore, we would like to consider a method of obtaining estimates that is free from at least some of these problems.

Moreover, we will also show that the method considered below provides asymptotically efficient estimates, while for the efficiency of the method of moments, in general, it is difficult to say anything.

Reminder.

The distribution P_θ is absolutely continuous with respect to a measure μ (denoted as $P_\theta \ll \mu$) if $\mu(A) = 0 \Rightarrow P_\theta(A) = 0$.

Definition.

Let $X_1, \ldots, X_n \sim P_\theta,\ \theta \in \Theta$.

$P_\theta \ll \mu$, where μ is a σ-finite measure (for example, the Lebesgue measure or the counting measure).

According to the Radon–Nikodym theorem, P_θ can be represented as

$$P_\theta(A) = \int_A f(x;\theta)\,\mu(dx)$$

for some function f.

In this case, $f(x;\theta)$ is the density function of the distribution (in the case of the Lebesgue measure), or the probability of the event $P_\theta(X = x)$ (in the case of the counting measure).

Then, for an arbitrary given sample $\mathbf{X} = (X_1, \ldots, X_n)$, we can define the likelihood function L as

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

$$\hat{\theta} = \text{maximum likelihood estimate (MLE)}, \quad \text{if } L(\hat{\theta}) = \max_\theta L(X;\theta).$$

Accordingly, the main idea of the maximum likelihood method is to choose such a parameter $\hat{\theta}$ for which the probability of obtaining the given sample is maximal, i.e., the maximum of $L(X;\theta)$ is achieved.

Moreover, it is often convenient to maximize $\ln L(X;\theta)$ instead of $L(X;\theta)$, since one can move from the product to a sum.

Remark.

In fact, when speaking about the likelihood function, we mean the function of the variable θ for a fixed sample X. However, in the literature, the sample X is often specified as an argument of the function L, and therefore we will also use this notation.

Example.

1. Let $X_1, \ldots, X_n \sim \text{Poisson}(\theta),\ \theta \in \Theta$.

2. a) $Norm(\theta, 1)$ or $Norm(\theta, 4)$

b) $Norm(a, \sigma^2)$, a is knowm

c)$Norm(\theta^2, \sigma^2)$ for both unknown parameters.

3.a) $Unif[0, \theta]$ and $Unif[\theta, \theta + 1]$.

3.b) $Exp(1/\theta)$ and $Exp(\theta)$.

4)*. For Multivariate Normal Law.

Lets $X_1, ..., X_n \in \mathbb{R}^d$ and $\theta \in \mathbb{R}^d$–vector of means and Σ-symmetric positive difinite covariance matrix, both are unknown with the following density

$$f(\mathbf{x}, \theta, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

Suggestions: $\widehat{\theta_n} = \overline{X}$ and $\widehat{\Sigma_n} = \frac{1}{n}\sum(X_j - \overline{X})(X_j - \overline{X})^T$.

End of the 10-th Seminar.

Definitions.

Let us have observations $X = (X_1, \ldots, X_n)$ with distribution P_θ, depending on a parameter $\theta \in \Theta \subseteq \mathbb{R}$. The experiment is called regular if the following conditions hold:

1. $\exists$ a dominating measure μ and the density $p_\theta(x)$, differentiable with respect to θ :

$$P_\theta(dx) = p_\theta(x)\mu(dx), \quad p_\theta(x) > 0$$

2. The likelihood function $L(\theta) = \prod_{i=1}^n p_\theta(X_i)$ is continuously differentiable with respect to θ.
3. Differentiation under the integral sign is allowed:

$$\frac{\partial}{\partial\theta}\int p_\theta(x)\mu(dx) = \int \frac{\partial}{\partial\theta} p_\theta(x)\mu(dx) = 0$$

4. The Fisher information is finite and positive:

$$I(\theta) = \int \left(\frac{\partial}{\partial\theta}\ln p_\theta(x)\right)^2 p_\theta(x)\mu(dx) \in (0, \infty)$$

Definition. Let $X_1, \ldots, X_n \sim P_\theta,\ \theta \in \Theta$.

The *Fisher information* is defined as the function

$$I_n(\theta) = E_\theta\left(\frac{\partial \ln f(X_1, \ldots, X_n; \theta)}{\partial\theta}\right)^2.$$

Remark. Since the observations are independent,

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

Hence, the Fisher information can be written as

$$I_n(\theta) = E\left(\frac{\partial \ln L(X; \theta)}{\partial\theta}\right)^2.$$

Properties of Fisher information.

1. The Fisher information for a sample of size n can be represented as

$$I_n(\theta) = \int \frac{(f'_\theta)^2}{f}\, d\mu.$$

Proof.

$$I_n(\theta) = E_\theta\left(\frac{\partial \ln f(X_1, \ldots, X_n; \theta)}{\partial\theta}\right)^2 = E\left(\frac{f'_\theta}{f}\right)^2 = \int\left(\frac{f'_\theta}{f}\right)^2 f\, d\mu = \int \frac{(f'_\theta)^2}{f}\, d\mu.$$

2. Suppose that for the density $f(X_1, \ldots, X_n; \theta)$ the following equality holds:

$$\int_X \frac{\partial}{\partial\theta} f(X_1, \ldots, X_n; \theta)\, d\mu = \frac{\partial}{\partial\theta}\int_X f(X_1, \ldots, X_n; \theta)\, d\mu.$$

Then

$$I_n(\theta) = nI_1(\theta).$$

Proof. Since f is a probability density, its integral equals 1 and therefore does not depend on the parameter θ. Hence we have

$$E\left(\frac{\partial \ln f}{\partial \theta}\right) = \int \frac{f'_\theta}{f} \cdot f \, d\mu = \int f'_\theta \, d\mu = \frac{\partial}{\partial \theta} \int f \, d\mu = 0.$$

We will use this property later in the proof. For now, we would like to rewrite the definition of Fisher information in a slightly different form.

Let us expand in more detail:

$$\frac{\partial \ln f}{\partial \theta}.$$

$$\frac{\partial \ln f(X_1, \ldots, X_n; \theta)}{\partial \theta} = \frac{\partial}{\partial \theta} \ln \prod_{i=1}^{n} f(X_i; \theta) = \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \ln f(X_i; \theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln f(X_i; \theta).$$

From this expression, it is clear how to simplify $I_n(\theta)$ in terms of $I_1(\theta)$. Expanding the square in the definition of Fisher information and using properties of expectation, we obtain:

$$I_n(\theta) = E\left(\frac{\partial \ln f(X_1, \ldots, X_n; \theta)}{\partial \theta}\right)^2 = E\left(\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln f(X_i; \theta)\right)^2$$

$$= E\left[\sum_{i=1}^{n} \left(\frac{\partial}{\partial \theta} \ln f(X_i; \theta)\right)^2\right] + E\left[\sum_{i \neq j} \frac{\partial}{\partial \theta} \ln f(X_i; \theta) \cdot \frac{\partial}{\partial \theta} \ln f(X_j; \theta)\right].$$

$$= nI_1(\theta) + \sum_{i \neq j} E\left[\frac{\partial}{\partial \theta} \ln f(X_i; \theta)\right] \cdot E\left[\frac{\partial}{\partial \theta} \ln f(X_j; \theta)\right] = nI_1(\theta).$$

Remark. From basic mathematical analysis, we know that if $\frac{\partial^k}{\partial \theta^k} f(x, \theta)$ exists and is continuously differentiable with respect to θ, then it is true that:

$$\frac{\partial^{k+1}}{\partial \theta^{k+1}} \int_X f(x; \theta) \, d\mu = \int_X \frac{\partial^{k+1}}{\partial \theta^{k+1}} f(x; \theta) \, d\mu. \tag{1}$$

However, instead of verifying continuous differentiability directly, it is sometimes more convenient to check this equality manually.

Theorem (Rao–Cramér Inequality).
$X_1, \ldots, X_n \sim P_\theta$ — random sample, $\theta \in \Theta$
$0 \neq I_1(\theta) < \infty$

And let the following condition of regularity for the distribution be satisfied:

$$\frac{d}{d\theta} \int T(x) f(x, \theta) d\mu = \int T(x) f'_\theta(x, \theta) d\mu$$

Consider any estimator T for the parameter θ.
$\mathbb{E}_\theta T = \theta + b(\theta)$, where $b(\theta)$ is the bias.

Then

$$\mathbb{E}_\theta (T - \theta)^2 \geq \frac{(1 + b'(\theta))^2}{I_n(\theta)} + b^2(\theta)$$

Proof.

Note that to obtain $1 + b'(\theta)$, it is enough to differentiate our expectation with respect to the parameter.

Let us write this value in a slightly different way, making use of the ability to swap the integral and derivative:

$$1 + b'(\theta) = (\mathbb{E}_\theta T)' = \frac{\partial}{\partial\theta}\int_X T \cdot f(x;\theta)\,d\mu = \int_X T \cdot f'_\theta(x;\theta)\,d\mu$$

$$= \int_X (T - \mathbb{E}_\theta T) \cdot f'_\theta(x;\theta)\,d\mu = \int_X T \cdot f'_\theta(x;\theta)\,d\mu - \mathbb{E}_\theta T \cdot 0 = \int_X (T - \mathbb{E}_\theta T) \cdot f'_\theta(x;\theta)\,d\mu$$

Let us now try to obtain some upper bound estimate for the value $(1+b'(\theta))^2$. For this, we rewrite the obtained integral in a slightly different way, and then use the Cauchy–Bunyakovsky inequality:

$$(1 + b'(\theta))^2 = \left[\int_X (T - \mathbb{E}_\theta T) \cdot f'_\theta d\mu\right]^2 = \left[\int_X (T - \mathbb{E}_\theta T) \cdot \sqrt{f} \cdot \frac{f'_\theta}{\sqrt{f}} d\mu\right]^2$$

$$\leq \left(\int_X \left((T - \mathbb{E}_\theta T) \cdot \sqrt{f}\right)^2 d\mu\right) \cdot \left(\int_X \left(\frac{f'_\theta}{\sqrt{f}}\right)^2 d\mu\right)$$

Note that the left integral is exactly the variance $\mathrm{D}_\theta(T - \mathbb{E}_\theta T) = \mathrm{D}_\theta T$, the variance of our estimator. The right integral is the Fisher information. Hence we have:

$$(1 + b'(\theta))^2 \leq \mathrm{D}_\theta(T - \mathbb{E}_\theta T)^2 \cdot I_n(\theta)$$

$$\frac{(1 + b'(\theta))^2}{I_n(\theta)} \leq \mathrm{D}_\theta(T - \mathbb{E}_\theta T)^2 = \mathrm{E}_\theta\left((T - \theta) - b(\theta)\right)^2 = \mathrm{E}_\theta(T - \theta)^2 - 2b(\theta)\mathrm{E}_\theta(T - \theta) + b^2(\theta)$$

Recall that $\mathrm{E}_\theta(T - \theta) = b(\theta)$, thus

$$\frac{(1 + b'(\theta))^2}{I_n(\theta)} \leq \mathrm{E}_\theta(T - \theta)^2 - b^2(\theta) \implies \mathrm{E}_\theta(T - \theta)^2 \geq \frac{(1 + b'(\theta))^2}{I_n(\theta)} + b^2(\theta)$$

Corollary.

If in the conditions of the theorem $b(\theta) \equiv 0$, then

$$\mathrm{E}_\theta(T - \theta)^2 \geq \frac{1}{I_n(\theta)}$$

Proof.

$$\mathrm{E}_\theta(T - \theta)^2 \geq \frac{(1 + b'(\theta))^2}{I_n(\theta)} + b^2(\theta), \quad b, b' \equiv 0 \Rightarrow \frac{1}{I_n(\theta)}$$

Corollary.

If under the conditions of the theorem T is an unbiased maximum likelihood estimator, and the condition of regularity is satisfied up to the second derivative, then the estimator T is asymptotically efficient.

Proof.

We know that our estimator is unbiased. Therefore, $\mathrm{D}(T - \theta)$ is bounded below by $I_n^{-1}(\theta)$.

For convenience, consider now the variance of the quantity $\sqrt{n}(T - \theta)$:

$$\mathrm{D}(\sqrt{n}(T - \theta)) = \mathrm{E}(\sqrt{n}(T - \theta))^2 = n\mathrm{E}(T - \theta)^2 \geq n \cdot \frac{1}{I_n(\theta)} = \frac{1}{I_1(\theta)}$$

Recall that the maximum likelihood estimator is asymptotically normal:

$$\sqrt{n}(T-\theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I_1(\theta)}\right)$$

Thus, we have obtained that the variance of deviations of our estimator converges to its lower bound. Therefore, the unbiased maximum likelihood estimator is asymptotically efficient.

Def. If unbiased estimator T_n such as

$$\mathrm{E}(\sqrt{n}(T-\theta))^2 = \frac{1}{I_1(\theta)},$$

then it is called R-efficient estimator.

End of the 11-th Seminar.

Property of maximal likelihood estimator.

Theorem 4.3. Let $X_1, \ldots, X_n \sim P_\theta$ be a random sample, where $\theta \in \Theta \subset \mathbb{R}$, and

$$dP_\theta = f(x,\theta)\, d\mu, \quad 0 < f(x,\theta) < \infty.$$

Assume the following conditions hold:

1. The support of $f(x,\theta)$ does not depend on θ;
2. $f(x,\theta)$ satisfies condition (1) for the first two derivatives;
3. $\forall\theta \in \Theta \quad \left|\frac{d^3}{d\theta^3} ln f(x,\theta)\right| \le M(X)$ and $E_{\theta_0} M(X) < \infty$, where θ_0 is the true value of parameter θ.

Then the following statements are true:

1. The maximum likelihood estimator (MLE) $\hat{\theta}_n$ is consistent, i.e. $\hat{\theta}_n \xrightarrow{P} \theta_0$;
2. The MLE is asymptotically normal, i.e.

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I_1(\theta_0)}\right).$$

Proof. 1. Without proof.

2. Expand $\dfrac{\partial \ln L}{\partial \theta}(X;\theta)$ in a Taylor series:

$$\frac{\partial \ln L}{\partial \theta}(X;\theta) = \frac{\partial \ln L}{\partial \theta}(X;\theta_0) + (\theta-\theta_0)\frac{\partial^2 \ln L}{\partial \theta^2}(X;\theta_0) + \frac{1}{2}(\theta-\theta_0)^2\frac{\partial^3 \ln L}{\partial \theta^3}(X;\tilde{\theta}), \quad \tilde{\theta} \in (\theta_0, \theta).$$

Now substitute the maximum likelihood estimate $\hat{\theta}$ instead of θ. Since the likelihood function attains its maximum at $\hat{\theta}$, its derivative at this point equals zero. Hence, we obtain:

$$0 = \frac{\partial \ln L}{\partial \theta}(X;\theta_0) + (\hat{\theta}-\theta_0)\frac{\partial^2 \ln L}{\partial \theta^2}(X;\theta_0) + \frac{1}{2}(\hat{\theta}-\theta_0)^2\frac{\partial^3 \ln L}{\partial \theta^3}(X;\tilde{\theta}), \quad \tilde{\theta} \in (\theta_0, \hat{\theta}).$$

Let us now find an expression for $\hat{\theta}-\theta_0$ from this equality, and multiply both sides by $\sqrt{n}$:

$$\sqrt{n}(\hat{\theta}-\theta_0) = -\frac{\frac{1}{\sqrt{n}}\frac{\partial \ln L}{\partial \theta}(X;\theta_0)}{\frac{1}{n}\frac{\partial^2 \ln L}{\partial \theta^2}(X;\theta_0) + \frac{1}{2n}(\hat{\theta}-\theta_0)\frac{\partial^3 \ln L}{\partial \theta^3}(X;\tilde{\theta})}, \quad \tilde{\theta} \in (\theta_0, \hat{\theta}).$$

We can now analyze each part of this expression separately. Start with the first derivative:

$$\frac{\partial \ln L}{\partial \theta}(X;\theta_0) = \frac{\partial}{\partial \theta}\sum_{i=1}^{n} \ln f(X_i;\theta)\Big|_{\theta=\theta_0} = \sum_{i=1}^{n} \frac{f'_\theta}{f}(X_i;\theta_0).$$

Thus, we obtained a sum of independent random variables. Moreover, the normalization factor $\frac{1}{\sqrt{n}}$ suggests the use of the Central Limit Theorem. We only need to compute the mean and variance of the summands:

$$E\left[\frac{f'_\theta}{f}(X_i;\theta_0)\right] = \int \frac{f'_\theta}{f} f \, d\mu = \int f'_\theta \, d\mu = 0,$$

$$D\left[\frac{f'_\theta}{f}(X_i;\theta_0)\right] = E\left[\left(\frac{f'_\theta}{f}(X_i;\theta_0)\right)^2\right] = E\left[\left(\frac{\partial \ln f(X_i;\theta_0)}{\partial \theta}\right)^2\right] = I_1(\theta_0).$$

Applying the Central Limit Theorem, we obtain:

$$\frac{1}{\sqrt{n}}\frac{\partial \ln L}{\partial \theta}(X;\theta_0) \xrightarrow{d} N(0, I_1(\theta_0)).$$

We now know the distribution of the numerator. It remains to determine the behavior of the denominator. Compute first the second derivative:

$$\frac{\partial^2 \ln L}{\partial \theta^2}(X;\theta_0) = \frac{\partial^2}{\partial \theta^2}\sum_{i=1}^{n} \ln f(X_i;\theta)\Big|_{\theta=\theta_0} = \sum_{i=1}^{n} \frac{f''_\theta f - (f'_\theta)^2}{f^2}(X_i;\theta_0).$$

Again, we have a sum of independent random variables. The normalization factor $\frac{1}{n}$ suggests the use of the Law of Large Numbers. To apply it, we compute the expected value:

$$E\left[\frac{f''_\theta f - (f'_\theta)^2}{f^2}(X_i;\theta_0)\right] = E\left[\frac{f''_\theta}{f}(X_i;\theta_0)\right] - E\left[\frac{(f'_\theta)^2}{f^2}(X_i;\theta_0)\right].$$

tex

$$= \int_X \frac{\partial^2}{\partial \theta^2} f(x;\theta_0) d\mu - I_1(\theta_0) = \frac{\partial^2}{\partial \theta^2}\int_X f(x;\theta_0) d\mu - I_1(\theta_0) = -I_1(\theta_0)$$

Thus, the first term in the denominator converges in probability to $-I_1(\theta_0)$, i.e.

$$\frac{1}{n}\frac{\partial^2}{\partial \theta^2} \ln L(\mathbf{X};\theta_0) \xrightarrow{P_{\theta_0}} -I_1(\theta_0)$$

It remains to consider the second term of the denominator. On one hand, from the first item of the theorem, we know that the estimator $\hat{\theta}$ is consistent; i.e., $\hat{\theta} - \theta_0 \xrightarrow{P_{\theta_0}} 0$. On the other hand, the remaining factors are bounded by the condition. Therefore, the whole term converges in probability to 0.

As a result, we find that the numerator is asymptotically normal: $N(0, I_1(\theta_0))$, and the denominator converges in probability to $-I_1(\theta_0)$. Thus, we obtain:

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} \frac{1}{I_1(\theta_0)} N\left(0, I_1(\theta_0)\right) = N\left(0, \frac{1}{I_1(\theta_0)}\right)$$

Remark.

The theorem cited above is also valid for $\Theta \subset \mathbb{R}^k$. Its proof only differs by using Taylor's formula for the multivariate case.

Application.

1. The maximum likelihood estimator (MLE) is asymptotically normal, meaning that, under regularity conditions and for large sample size n, the distribution of the scaled estimation error converges to a normal distribution:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right),$$

Using the asymptotic normality property, an approximate confidence interval for the parameter θ with confidence level $1 - \alpha$ is given by:

$$\hat{\theta} \pm z_{1-\alpha/2} \cdot \sqrt{\frac{1}{nI_{obs}(\hat{\theta})}}$$

where $z_{1-\alpha/2}$ is the quantile of the standard normal distribution and $I_{obs}(\hat{\theta})$ is the observed Fisher information at the MLE.

2. Let T_n and V_n be asymptotically normal estimators of the parameter θ with variances $\sigma^2_{T_n}$ and $\sigma^2_{V_n}$, respectively. The asymptotic efficiency of estimator T_n with respect to estimator V_n is defined as the ratio of their variances:

$$Eff(T_n, V_n) := \text{efficiency}(T_n, V_n) = \frac{\sigma^2_{V_n}}{\sigma^2_{T_n}}.$$

In case of the Rao-Cramer theorem, we can get the following expression and call it the exact efficiency:

$$\text{Efficiency}(\hat{\theta}_n) = \frac{1/I(\theta)}{\text{Var}(\hat{\theta}_n)} \leq 1.$$

Examples. 1. $Caughy(\theta, 1)$.
$\overline{\theta_n} = med_n = \widehat{Z}_{\frac{1}{2}}$; $\quad \sqrt{n}(\widehat{Z}_{\frac{1}{2}} - \theta) \Rightarrow N(0, \frac{\pi^2}{4})$.
Then for MLE using the Rao-Cramer inequality we can get: $I(\theta) = \frac{1}{2} \Rightarrow E_\theta(\widehat{\theta} - \theta)^2 \geq nI(\theta) = \frac{n}{2} \Rightarrow Eff^F_{med,MLE} = \frac{8}{\pi^2}$.
End of the 12-th Seminar.

Sufficient statistic.

Reminder. Conditional expectation

Definition. Let X be a random variable, $E(X) < \infty$ and $\mathfrak{F}$ - sigma-algebra, then $Z := E(X|\mathfrak{F})$ is such random variable, that
1. Z – $\mathfrak{F}$-measurable.
2. $\forall B \in \mathfrak{F} \quad E(X\mathbf{1}_B) = E(Z\mathbf{1}_B)$.

$E(X|Y) =: g(Y)$, where
in discrete case $g(y) = \frac{E(X \cdot \mathbf{1}(Y=y))}{P(Y=y)} = \sum_i x_i \cdot P(X = x_i|Y = y)$;
in abs. continuous case $g(y) = \int\limits_R x \frac{f_{x,y}(x,y)}{f_Y(y)} dx$.

Properties.
1. $E(const|\mathfrak{F}) = const$;
2. $E(aX|\mathfrak{F}) = aE(X|\mathfrak{F})$;
3. $E(X + Y|F) = E(X|\mathfrak{F}) + E(Y|\mathfrak{F})$;

4. $E(E(X|\mathfrak{F})) = E(X)$;
5. if $X \le Y \Rightarrow E(X|\mathfrak{F}) \le E(Y|\mathfrak{F})$;
6. If X is $\mathfrak{F}$-measurable, $E(X|\mathfrak{F}) = X$;
7. If X is indepemdent of $\mathfrak{F}$, than $E(X|\mathfrak{F}) = E(X)$.
8. If $E|X| < \infty$ then $|E(X|\mathfrak{F}) \le E(|X||\mathfrak{F})$;
9. If Y is $\mathfrak{F}$-measurable, then $E(XY|\mathfrak{F}) = YE(X|\mathfrak{F})$.
10. If $\mathfrak{A} \subset \mathfrak{B} \subset \mathfrak{F}$, then $E(\xi|\mathfrak{A}) = E\left(E(\xi|\mathfrak{B})|\mathfrak{A}\right)$.

Back to statistics.

Definition.

Let there be a sample $X_1, \ldots, X_n \sim P_\theta$, where $\theta \in \Theta$.

A statistic T is said to be *sufficient* for $\{P_\theta \mid \theta \in \Theta\}$ if

$$P_\theta(X \in A \mid T = t)$$

does not depend on θ.

Remark. The equivalent option: for measurable function ϕ, $E_\theta(\phi(X)|T)$ doesn't depend on θ.

Remark.

Such a statistic T essentially contains all the relevant information about the parameter. Accordingly, if we know the value of this statistic, we also know the distribution of our sample.

Example.

Let the elements of the sample $X_1, \ldots, X_n$ be distributed according to the Bernoulli distribution with parameter θ, i.e.

$$P(X_i = 1) = \theta, \quad P(X_i = 0) = 1 - \theta.$$

We will show that the statistic

$$T = \sum_{k=1}^{n} X_k$$

is sufficient for our family.

Let us verify the sufficiency condition for elementary events:

$$P_\theta(X_1 = x_1, \ldots, X_n = x_n \mid T = t) = \frac{P_\theta(X_1 = x_1, \ldots, X_n = x_n; \sum_{i=1}^n X_i = t)}{P\left(\sum_{i=1}^n X_i = t\right)}.$$

Note that if $\sum_{i=1}^n x_i \neq t$, then the probability of such an elementary event is equal to 0.

If, however, $\sum_{i=1}^n x_i = t$, then our conditional probability can be written as follows:

$$P_\theta(X_1 = x_1, \ldots, X_n = x_n \mid T = t) = \frac{P_\theta(X_1 = x_1, \ldots, X_n = x_n)}{P_\theta\left(\sum_{i=1}^n X_i = t\right)}.$$

$$= \frac{\prod_{i=1}^n P_\theta(X_i = x_i)}{P_\theta\left(\sum_{i=1}^n X_i = t\right)} = \frac{\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}}$$

Thus, when the value of the statistic T is known, our conditional probability does not depend on the parameter θ, which is exactly what we needed to prove.

Note that checking the sufficiency condition directly from the definition can often be very difficult. Therefore, we would like to obtain some simple method for verifying such a condition. To this end, let us prove the following theorem, known as the *factorization theorem.*

Theorem (Fisher, Neyman).

Let $X_1, \ldots, X_n$ have a joint density $f(x;\theta)$ with respect to some σ-finite measure μ.

Then the statistic T is sufficient $\iff$

$$f(x_1, \ldots, x_n; \theta) = g_\theta(T(x_1, \ldots, x_n)) \cdot h(x_1, \ldots, x_n).$$

Remark.
In fact, this theorem not only allows us to check the sufficiency of statistics but also provides a way to find them: it is enough to represent f as a product of g_θ and h.

Examples.
1. Let us again consider the example with the Bernoulli distribution.

$$f(x_1, \dots, x_n; \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} = \theta^T(1-\theta)^{n-T} =: g_\theta(T)$$

Thus, if we set $h \equiv 1$, we obtain precisely the factorization of the joint density as the product of two functions $g_\theta(T)$ and h. Hence, our statistic T is sufficient.

Moreover, in this example, we can see how one could derive the sufficient statistic T even if we did not already know it.
2. $Pois(\theta)$
3. $Norm(\theta, \sigma^2)$.
4. $Unif[0, \theta]$.
5.
Let us now consider a less pleasant case — the Cauchy distribution with parameter θ.

Its joint density has the following form:

$$f(X_1, \dots, X_n; \theta) = \frac{1}{\pi^n} \prod_{i=1}^{n} \frac{1}{1 + (X_i - \theta)^2}.$$

Proof. Below is a proof for the special case when the probability distribution is discrete. Then $f_\theta(x) = \mathbb{P}(X = x \mid \theta)$ is the probability mass function.

Suppose this function can be factorized as in the theorem statement, and let $T(x) = t$.

Then we have:

$$\mathbb{P}(X = x \mid T(X) = t, \theta) = \frac{\mathbb{P}(X = x \mid \theta)}{\mathbb{P}(T(X) = t \mid \theta)} = \frac{h(x)g(\theta, T(x))}{\sum_{x:T(x)=t} h(x)g(\theta, T(x))} = \frac{h(x)}{\sum_{x:T(x)=t} h(x)}.$$

From this we see that the conditional probability of the vector X given the statistic $T(X)$ does not depend on the parameter θ, and therefore $T(X)$ is a sufficient statistic.

Conversely, we can write:

$$\mathbb{P}(X = x \mid \theta) = \mathbb{P}(X = x \mid T(X) = t, \theta) \cdot \mathbb{P}(T(X) = t \mid \theta).$$

From the above we see that the first factor on the right-hand side does not depend on the parameter θ, and it can be taken as the function $h(x)$ from the theorem statement. The other factor is a function of θ and $T(X)$, and it can be taken as the function $g(\theta, T(x))$. Thus, we obtain the required decomposition, which completes the proof of the theorem.

End of the 13-th Seminar.

Theorem (Rao, Blackwell, Kolmogorov).
Let T be a sufficient statistic, and let S be any unbiased estimator of the parameter θ. Then there exists an unbiased estimator S_1 of θ, which depends on T and is more efficient than S. In other words, S_1 satisfies the condition

$$E_\theta(S_1 - \theta)^2 \le E_\theta(S - \theta)^2 \quad \forall \theta.$$

Proof.
Let us define $S_1 := E_\theta(S \mid T)$.

Note that, since T is a sufficient statistic, S_1 does not depend on θ. Hence, S_1 is indeed some estimator.

Furthermore,

$$E_\theta S_1 = E_\theta(E_\theta(S \mid T)) = E_\theta S = \theta,$$

so S_1 is an unbiased estimator.

It remains to show that S_1 is more efficient than S. Indeed,

$$E_\theta(S-\theta)^2 = E_\theta(S - S_1 + S_1 - \theta)^2 = E_\theta(S-S_1)^2 + E_\theta(S_1-\theta)^2 + 2E_\theta[(S-S_1)(S_1-\theta)].$$

Let us analyze the third term in more detail:

$$E_\theta[(S-S_1)(S_1-\theta)] = E_\theta\left(E_\theta[(S-S_1)(S_1-\theta) \mid T]\right) = E_\theta\left((S_1-\theta)E_\theta(S-S_1 \mid T)\right) = E_\theta\left((S_1-\theta)(E_\theta(S \mid T) - S_1)\right)$$

In the second equality we used the property of conditional expectation, which allows measurable factors to be taken outside the expectation.

Thus, since the third term equals zero, and the second term is nonnegative and bounded below by 0, we obtain precisely the desired inequality. □

In a sense, we have shown that an efficient estimator exists. The only thing missing is uniqueness.

Definition.

A sufficient statistic T is said to be *complete* if for any measurable function φ the following condition holds:

$$(\forall\theta:\ E_\theta\varphi(T) = 0) \ \Rightarrow\ \varphi(T) = 0.$$

Examples.

1. Bernoulli (p).
2. $Norm(\theta, 1)$.
3. Let $X_1, \ldots, X_n$ be a sample from the uniform distribution on the interval $[0, \theta]$, where $\theta \in \Theta$. Prove that the statistic $X_{(n)}$ is complete for the parameter θ when $\Theta = (0, \infty)$. Is $X_{(n)}$ a complete statistic for θ when $\Theta = (1, \infty)$?

End of the 14-th Seminar.

Theorem (Lehmann, Scheffé).

Let T be a complete sufficient statistic.

Then there exists at most one (up to a function of T) unbiased estimator depending on T.

Moreover, if such an estimator exists, it is efficient.

Proof.

Let S be an unbiased estimator depending on T. Assume that at least one such estimator exists.

We will show that if there are two unbiased estimators $S_1(T)$ and $S_2(T)$, then they coincide. Suppose not. Then we have:

$$E_\theta S_1(T) = E_\theta S_2(T) = \theta \ \Rightarrow\ E_\theta(S_1(T) - S_2(T)) = 0 \quad \forall\theta.$$

Now, if we recall that our statistic T is complete, we obtain:

$$E_\theta\left[(S_1 - S_2)(T)\right] = 0 \implies (S_1 - S_2)(T) \equiv 0 \implies S_1(T) \equiv S_2(T)$$

Let us understand why S will be efficient. Again, we proceed by contradiction. Suppose there exists an unbiased estimator R that is more efficient than S. Then we know that there exists an unbiased estimator

$$R_1 = E_\theta(R \mid T),$$

which depends on T and is more efficient than S. This leads to a contradiction with the uniqueness of the unbiased estimator depending on T.

Hence, S is an efficient estimator, as was to be proved.

Definition (Ancillary Statistic). A statistic $A = A(X_1, \ldots, X_n)$ is called *ancillary* if its probability distribution does not depend on the parameter θ. Formally,

$$P_\theta(A \in B) = P(A \in B) \quad \text{for all } \theta.$$

Basu's Theorem. If $T(X)$ is a *complete and sufficient* statistic for the parameter θ, and $A(X)$ is an *ancillary* statistic for θ, then $T(X)$ and $A(X)$ are *independent*.
Proof. Consider $\gamma(B) = P_\theta(A \in B)$ and $\beta(B|T) = P_\theta(A \in B|T) = E_\theta(\mathbb{1}\{A \in B\}|T)$ then $E_\theta(\gamma(B) - \beta(B|T)) = 0$ then $\gamma(B) = \beta(B|T)$.

Bayesian and minimax estimators.

Up to this point, we have been discussing Fisherian statistics. This model assumes that errors in estimating parameters arise solely because data collection occurs on a sample rather than on the entire population. Essentially, the only type of error allowed in this model is the sampling error, since the sample may turn out to be insufficiently balanced or corrected.

In contrast to this model, there exists Bayesian statistics, which attempts to predict in advance which parameter values of the distribution are more probable, and then adjusts its representation of them in accordance with subsequent observations.

Both of these models, depending on the specific task, can perform either well or poorly. Accordingly, we will now discuss how to estimate the parameters of distributions in terms of Bayesian statistics.

To begin, let us introduce several important definitions for Bayesian statistics.

Definition
A **prior distribution** of a parameter θ is called a distribution $q(\theta)$, which reflects our representation of the distribution of the parameter before taking experimental data into account.

A **posterior distribution** of a parameter θ is called the conditional distribution $q(\theta|X)$, which takes into account the data obtained as a result of the experiment. In particular, it is defined as

$$q(\theta|X) = \frac{f(X|\theta) \cdot q(\theta)}{f(X)} = \frac{f(X|\theta) \cdot q(\theta)}{\int_\Theta f(X|\theta) \cdot q(\theta)\, d\theta}.$$

In the discrete case, f and q can be interpreted as probabilities of corresponding events.

definition
A **loss function** is a function $W(T, \theta)$ that determines how well a statistic T estimates the parameter θ for a given sample. Essentially, this function must satisfy the following conditions:

1. $W(T, \theta) \geq 0$, and $W(T, \theta) = 0$ if $T(X) = \theta$;
2. If $W(T_1, \theta) < W(T_2, \theta)$, then for the current sample, statistic T_1 provides a better estimate of the parameter than T_2.

Thus, by defining W, we determine a certain rule according to which we decide how satisfactory our estimate is.

Remark.
As a rule, the loss function is taken as either $|T - \theta|$ or $(T - \theta)^2$. It is with functions of this type that we will continue to work.

Definition
The **risk function** $R(T, \theta)$ is defined as the expected loss:

$$R(T, \theta) = \mathbb{E}_\theta W(T, \theta),$$

that is, the risk of the estimator T at the point θ.

Remark.
It is quite natural to wish to find such an estimator T that minimizes the risk for any possible parameter value θ. However, it can be shown that this is not possible.

Indeed, suppose such an estimator T exists. Then the following condition would hold:

$$R(T, \theta) \leq R(S, \theta), \quad \forall S, \forall \theta.$$

In particular, we could choose $S = \theta$. Then

$$R(T, \theta) \leq R(\theta, \theta) = \mathbb{E}_\theta W(\theta, \theta) = 0.$$

But this is possible only if $W(T, \theta) = 0$ for all θ, which means $T \equiv \theta$ for all θ, which never happens.

We have shown that, although the risk function allows us to evaluate the error of our estimator, it provides no information about which estimator is truly better in a given situation.

Nevertheless, this issue can be addressed if we define in advance which type of risk we wish to minimize. To do this, let us introduce the concepts of the **minimax** and **Bayesian** estimators.

Definition
An estimator T is called **minimax** if it minimizes the maximum risk, i.e.

$$\sup_\theta R(T, \theta) = \inf_S \sup_\theta R(S, \theta).$$

Definition
Let a probability distribution Q be defined on our parameter set Θ. We define the **Bayesian risk** of the estimator T with respect to the prior distribution Q as

$$R(T) := \int_\Theta R(T, \theta)\, dQ(\theta).$$

An estimator T is called a **Bayesian estimator** with respect to the prior distribution Q if it minimizes the Bayesian risk, that is,

$$R(T) = \inf_S R(S).$$

As always, we would now like to understand how to construct such estimators.

Proposition
Let $X_1, \ldots, X_n$ have densities $f(x; \theta)$ with respect to a σ-finite measure μ. Let a prior distribution Q with the corresponding density q be given on Θ. Then, the Bayesian estimator in the case of a quadratic loss function has the following form:

$$T(X) = \int_\Theta \theta\, q(\theta|X)\, d\theta,$$

where $q(\theta|X)$ is the density of the corresponding posterior distribution of Q.

Proof. We know that the Bayesian estimator must minimize the Bayesian risk. That is, we want

$$R(T,\theta) = \int_\Theta W(T,\theta)\, f(x|\theta)\, q(\theta)\, d\theta\, dx = \int_{\mathbb{X}} \int_\Theta (T(x)-\theta)^2 f(x|\theta)\, q(\theta)\, d\theta\, dx \to \min\,.$$

Since we are integrating a nonnegative function, we can interchange the order of integration. Therefore, we can minimize the integrand for each x separately:

$$\int_\Theta (T(x)-\theta)^2 f(x|\theta)\, q(\theta)\, d\theta \to \min \quad \Longleftrightarrow \quad \int_\Theta (T(x)-\theta)^2 q(\theta|x)\, d\theta \to \min\,.$$

Expanding the square in this expression, we obtain:

$$T(x)^2 \int_\Theta f(x|\theta) q(\theta)\, d\theta - 2T(x) \int_\Theta \theta f(x|\theta) q(\theta)\, d\theta + \int_\Theta \theta^2 f(x|\theta) q(\theta)\, d\theta \to \min\,.$$

We have obtained a quadratic function. To find the minimum of a quadratic function, we differentiate and equate to zero:

$$T(x) = \frac{\int_\Theta \theta f(x|\theta) q(\theta)\, d\theta}{\int_\Theta f(x|\theta) q(\theta)\, d\theta} = \int_\Theta \theta\, q(\theta|x)\, d\theta.$$

Remark. We have found that, in the case of a quadratic loss function, the Bayesian estimator is the posterior mean. It is also easy to verify that if the loss function is given by the absolute value of the difference, then the estimator will be the posterior median.

Now let us move on to minimax estimators.

Theorem. Let T_k be the Bayesian estimator with respect to the prior distribution Q_k. Let T be any estimator of the parameter θ. Then, if

$$\sup_\theta R(T,\theta) \le \lim_{k\to\infty} R(T_k),$$

the estimator T is minimax.

Proof.

Let us consider an arbitrary estimator S. Then, it holds that:

$$\sup_\theta R(S,\theta) \ge \int R(S,\theta)\, dQ_k(\theta) \ge \int R(T_k,\theta)\, dQ_k(\theta) = R(T_k)$$

Note that this is true for any k. We can then pass to the limit as $k \to \infty$ and use the theorem's assumption:

$$\sup_\theta R(S,\theta) \ge \lim_{k\to\infty} R(T_k) \ge \sup_\theta R(T,\theta)$$

Thus, we obtain the definition of a minimax estimator. □

Corollary.

Let T be a Bayes estimator with constant risk, i.e., $R(T,\theta)$ does not depend on θ. Then T is a minimax estimator. □

Proof.

In the assumptions of the theorem, take the sequence $T_k = T$. Since $R(T,\theta)$ does not depend on θ, we have:

$$\sup_\theta R(T,\theta) = \int R(T,\theta)\, dQ = R(T)$$

Thus, the theorem's condition is satisfied and T is a minimax estimator. □

Example.1. Consider the sample $X_1, ..., X_n \sim Norm(\theta, 1), \quad \theta \in \mathbb{R}$. Show that $\overline{X}$ is minimax for the square

loss function. Construct the Bayessian estimator T_k with respect to the prior distribution N(0,k).
Solution.

$$T_k = \frac{n\overline{X}}{n + \frac{1}{k}};$$

$$R(T_k) = \int E_\theta \left(\frac{n\overline{X}}{n + \frac{1}{k}} - \theta \right)^2 dN(0,k) = \frac{1}{n + \frac{1}{k}} \to \frac{1}{n}, \quad k \to +\infty;$$

$$\sup_\theta R(\overline{X}, \theta) = \frac{1}{n} = \lim_{k \to \infty} R(T_k)$$

them by the theorem $\overline{X}$ is minimax.
2. $Ber(\theta)$. Show that $\overline{X}$ is not minimax.
End of the 17-th Seminar.

Linear regression.

History of direction

The projection onto the space of functions of X:

$$\widehat{Y} = \arg\min_{g(X)} \mathbb{E}\big[(Y - g(X))^2\big].$$

Conditional expectation as the solution:

$$\widehat{Y} = \mathbb{E}[Y \mid X].$$

Measurability:

$$\mathbb{E}[Y \mid X] = g(X) \quad \text{for some function } g.$$

Orthogonality of the error:

$$\mathbb{E}\big[(Y - \mathbb{E}[Y \mid X])\, h(X)\big] = 0 \quad \forall\, h(X).$$

Error minimization:

$$\mathbb{E}\big[(Y - \mathbb{E}[Y \mid X])^2\big] \le \mathbb{E}\big[(Y - g(X))^2\big].$$

Definition 1. Let $\vec{X} = (X_1, \ldots, X_k)^T$ be a set of independent random variables, and Y an arbitrary random variable depending on $\vec{X}$. Then the function $f(\vec{x}) = E(Y \mid \vec{X} = \vec{x})$ is called the *regression* of Y on $\vec{X}$.

Obviously, in the general case Y may behave in any way depending on the values of the independent variables $\vec{X}$. Therefore, to begin with, it makes sense to consider some simple regression model.
Definition 2. We shall assume that the behavior of Y is described by:

$$Y \mid \vec{X} = \vec{X}^T \vec{\theta} + \varepsilon,$$

where $\vec{\theta}$ is a set of k parameters, which are unknown but fixed beforehand. These are the parameters we want to find.

The random variable ε here is an uncontrolled random perturbation (noise), which has the following properties:

$$E\varepsilon = 0, \quad D\varepsilon = \sigma^2,$$

where σ^2 is known.

The regression itself in this case has the form:

$$f(\vec{x}) = E(Y \mid \vec{X} = \vec{x}) = \vec{x}^T\vec{\theta}.$$

Such a regression model in which the expected value of the function f is linearly dependent on the values of $\vec{X}$ is called *linear regression.*

Remark. We will continue with an apartment example. In this case, the values of X_i are some quantitative parameters of an apartment: area, floor, year of construction, etc.

Accordingly, in the linear regression model we assume that the price of the apartment has a certain linear dependence on the values of these parameters. But noise is allowed to deviate the value to one side or the other.

Remark. From now on we will work with parameters and independent variables as with vectors. Clearly, we will no longer use the boldface "vector nature" notation.

We are not yet discussing linear regression itself and are solving a simpler task: to find a vector of parameters $\vec{\theta}$ that describes the behavior of the dependent variable Y. Later we will examine the nature of the dependence of the parameters on each other.

As always, we will try to restore the model parameters based on observations.

Statement. Let $\{(Y_i, X_i)\}$ be observations.

Then in our linear model the vector of observed values Y is represented as follows:

$$Y = X \cdot \vec{\theta} + \varepsilon,$$

where

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} \vec{X}_1^T \\ \vdots \\ \vec{X}_n^T \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad E(\varepsilon_i\varepsilon_j) = 0 \text{ for } i \neq j.$$

Note that even if we could "forget" about the existing noise and not take it into account, it would still be optimal to use the method of least squares, since in this case it coincides with the method of Gauss.

However, if the number of observations n turns out to be relatively small compared to the number of parameters k, then the system in general will not have a unique solution.

Formally, we want to find such a set of parameters θ that minimizes the function

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

An estimate θ^* such that $S(\theta^*) = \min_\theta S(\theta)$ is called the least squares estimate (LSE).

It remains to understand how we can obtain this estimate. First, let us introduce the following notation and prove a couple of related statements.

Notation. We introduce the following notation for the vector of partial derivatives:

$$\frac{df}{d\theta} = \begin{pmatrix} \frac{\partial f}{\partial \theta_1} \\ \vdots \\ \frac{\partial f}{\partial \theta_k} \end{pmatrix}.$$

Statement. Let α be an arbitrary vector from $\mathbb{R}^k$. Then the following equality holds:

$$\frac{d}{d\theta}\left(\alpha^T\theta\right) = \frac{d}{d\theta}\left(\theta^T\alpha\right) = \alpha.$$

Proof. Since $\alpha^T\theta = \sum_{i=1}^k \alpha_i\theta_i = \theta^T\alpha$, the partial derivative with respect to θ_i is exactly α_i. □

Statement. Let A be a symmetric matrix, i.e. $A^T = A$. Then the following equality holds:

$$\frac{d}{d\theta}\left(\theta^T A\theta\right) = 2A\theta.$$

Proof. Expand the quadratic form:

$$\theta^T A\theta = \sum_{i,j} A_{ij}\theta_i\theta_j.$$

Now compute its partial derivative with respect to θ_i and verify that it coincides with the i-th row of the vector $2A\theta$:

$$\frac{\partial}{\partial\theta_i}\left(\sum_{i,j} A_{ij}\theta_i\theta_j\right) = \sum_{j=1}^{k}(A_{ij} + A_{ji})\theta_j = 2\sum_{j=1}^{k} A_{ij}\theta_j = (2A\theta)_i.$$

□

Now let us proceed to finding the least squares estimate.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be our observations. Then the least squares estimate θ^* is obtained as any solution of the system $X^T X\,\theta = X^T Y$. In particular, if the matrix $X^T X$ is invertible, then the LSE has the form

$$\theta^* = (X^T X)^{-1} X^T Y.$$

Proof. Let us expand our function:

$$S(\theta) = \|Y - X\theta\|^2 = (Y - X\theta)^T(Y - X\theta) = \theta^T X^T X\theta - 2X^T Y\cdot\theta + Y^T Y.$$

This function always reaches its minimum, because the resulting quadratic form is non-negative definite. In particular, the derivative at the minimum equals zero.

First, compute how the derivative looks in general:

$$\frac{dS(\theta)}{d\theta} = 2X^T X\,\theta - 2X^T Y.$$

Thus it is clear that we are interested in solving the system

$$X^T X\,\theta = X^T Y.$$

If the matrix $X^T X$ is invertible, then the LSE is

$$\theta^* = (X^T X)^{-1} X^T Y.$$

It remains to show that if the matrix $X^T X$ is not invertible, then any solution of this system gives the same value of $S(\theta)$.

Consider some solution θ^*. Show that for any other parameter value θ the result will be no better:

$$\begin{aligned} S(\theta) &= \|Y - X\theta\|^2 = \|Y - X\theta^* + X\theta^* - X\theta\|^2 \\ &= \|Y - X\theta^*\|^2 + \|X(\theta^* - \theta)\|^2 + 2(Y - X\theta^*)^T X(\theta^* - \theta). \end{aligned}$$

Since $X^T(Y - X\theta^*) = 0$ (because $X^T Y - X^T X\theta^* = 0$), we get

$$(Y - X\theta^*)^T X(\theta^* - \theta) = 0,$$

so

$$S(\theta) = \|Y - X\theta^*\|^2 + \|X(\theta^* - \theta)\|^2 \geq S(\theta^*).$$

□

Remark. Since the matrix X in our case is formed from the observations, we can always ensure that X^TX is invertible. Even if in practice the obtained solution differs slightly from the true one (for example, because the matrix is nearly singular), the solution still exists and can be found (using Gaussian elimination or other methods).

Thus, we obtain some estimate of the parameters for our linear regression model. The question remains: how accurate this estimate is in reality?

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two estimates of the parameter vector θ. We say that $\hat{\theta}_1$ is *more efficient* than $\hat{\theta}_2$ if

$$\operatorname{cov}(\hat{\theta}_1) \leq \operatorname{cov}(\hat{\theta}_2).$$

Here $\operatorname{cov}(\theta)$ is the covariance matrix of the random vector θ.

Remark. This definition of efficiency is fully consistent with the corresponding definition for estimators of scalar parameters.

Only now the random variable whose variance we compare is a random vector, so we now compare covariance matrices instead of numbers.

We introduce the natural order on covariance matrices: if $A \leq B$, then $B - A$ is a non-negative definite matrix. *Note.*

This definition of efficiency is a generalization of the corresponding definition for the univariate case.

Previously, we only tried to estimate the variance of our estimator, but now we can try to estimate its "multidimensional version", that is, the covariance matrix.

Note.

We introduce a partial order on matrices as follows: $A \leq B$ if $B - A$ is positive semidefinite. Then the estimator θ_1 is more efficient than the estimator θ_2 if $\operatorname{cov}(\theta_1) \leq \operatorname{cov}(\theta_2)$.

Theorem (Gauss, Markov).

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be our observations, and suppose that $(X^\top X)^{-1}$ exists. Moreover, we consider a model in which Y has the distribution $Y|X = X\theta + \varepsilon$.

Then $\theta^* = (X^\top X)^{-1}X^\top Y$ is an unbiased linear estimator. In English literature, it is known as the best linear unbiased estimator, or simply BLUE.

Proof.

Linearity of our estimator is obvious (we make linear transformations with the elements of Y). Let us check the unbiasedness of our estimator:

$$\begin{aligned}\mathbb{E}_\theta\theta^* &= \mathbb{E}_\theta\left((X^\top X)^{-1}X^\top Y\right) = \mathbb{E}_\theta\left((X^\top X)^{-1}X^\top(X\theta+\varepsilon)\right)\\ &= \mathbb{E}_\theta\left((X^\top X)^{-1}X^\top X\theta + (X^\top X)^{-1}X^\top\varepsilon\right)\\ &= \theta + \mathbb{E}_\theta\left((X^\top X)^{-1}X^\top\varepsilon\right) = \theta + (X^\top X)^{-1}X^\top\mathbb{E}_\theta\varepsilon = \theta\end{aligned}$$

(The last transition here is legitimate by the linearity of expectation and because the distribution of X does not depend on θ.)

It remains to check efficiency. For this, we take any other linear unbiased estimator $\tilde{\theta}$ and show that it is not better.

First, in general, any linear estimator has the form LY. But as we want this estimator to be unbiased, the matrix L must satisfy the condition $\mathbb{E}_\theta(LY) = \theta$ for all θ, that is,

$$\mathbb{E}_\theta(LY) = \mathbb{E}_\theta(L(X\theta+\varepsilon)) = LX\theta$$

So $LX = I_k$, where I_k is the identity matrix of size k.

Now let us look at the covariance matrix of θ^*:

$$\begin{aligned}\operatorname{cov}(\theta^*) &= \mathbb{E}_\theta\left[(\theta^*-\theta)(\theta^*-\theta)^\top\right]\\ &= \mathbb{E}_\theta\left[\left((X^\top X)^{-1}X^\top(X\theta+\varepsilon)-\theta\right)\left((X^\top X)^{-1}X^\top(X\theta+\varepsilon)-\theta\right)^\top\right]\\ &= \mathbb{E}_\theta\left[(X^\top X)^{-1}X^\top\varepsilon\left((X^\top X)^{-1}X^\top\varepsilon\right)^\top\right]\\ &= (X^\top X)^{-1}X^\top\mathbb{E}_\theta(\varepsilon\varepsilon^\top)X(X^\top X)^{-1}\end{aligned}$$

Notice that in the last expression the random variable is only $\varepsilon\varepsilon^\top$, and the whole rest can be taken out of the expectation as it depends only linearly on ε. Note that $\mathbb{E}(\varepsilon\varepsilon^\top)$ is, in essence, the covariance matrix of the "noise". Moreover, the diagonal of this matrix will contain the noise variances, and in the other positions—zeros, because the different noises are uncorrelated with each other.

$$(X^\top X)^{-1}X^\top\cdot\sigma^2E_n\cdot X(X^\top X)^{-1} = \sigma^2\cdot(X^\top X)^{-1}$$

Thus, we were able to calculate the covariance matrix for the estimator θ^*. It remains to calculate the covariance matrix for $\tilde\theta = LY$, given that $LX = E_k$.

Let us introduce the following notation:

$$C := L-(X^\top X)^{-1}X^\top \implies L = C+(X^\top X)^{-1}X^\top$$

Now let us expand the covariance matrix $\operatorname{cov}(\tilde\theta)$:

$$\begin{aligned}\operatorname{cov}(\tilde\theta) &= \mathbb{E}_\theta\left[(\tilde\theta-\theta)(\tilde\theta-\theta)^\top\right] = \mathbb{E}_\theta\left[(LY-\theta)(LY-\theta)^\top\right]\\ &= \mathbb{E}_\theta\left[((C+(X^\top X)^{-1}X^\top)(X\theta+\varepsilon)-\theta)\cdot((C+(X^\top X)^{-1}X^\top)(X\theta+\varepsilon)-\theta)^\top\right]\\ &= \mathbb{E}_\theta\left[(CX\theta+C\varepsilon+(X^\top X)^{-1}X^\top X\theta+(X^\top X)^{-1}X^\top\varepsilon-\theta)\cdots\right]\\ &= \mathbb{E}_\theta\left[(C\varepsilon+(X^\top X)^{-1}X^\top\varepsilon)\cdots(C\varepsilon+(X^\top X)^{-1}X^\top\varepsilon)^\top\right]\\ &= \mathbb{E}_\theta\left[C\varepsilon\varepsilon^\top C^\top+(X^\top X)^{-1}X^\top\varepsilon\varepsilon^\top C^\top+C\varepsilon\varepsilon^\top X(X^\top X)^{-1}+(X^\top X)^{-1}X^\top\varepsilon\varepsilon^\top X(X^\top X)^{-1}\right]\\ &= \mathbb{E}_\theta\left[C\varepsilon\varepsilon^\top C^\top\right]+(X^\top X)^{-1}X^\top\mathbb{E}_\theta\left[\varepsilon\varepsilon^\top\right]C^\top+C\mathbb{E}_\theta\left[\varepsilon\varepsilon^\top\right]X(X^\top X)^{-1}+(X^\top X)^{-1}X^\top\mathbb{E}_\theta\left[\varepsilon\varepsilon^\top\right]X(X^\top X)^{-1}\end{aligned}$$

From the definition of C and the condition on L, observe that $CX = 0$. Therefore,

$$\mathbb{E}_\theta\left[(C\varepsilon+(X^\top X)^{-1}X^\top\varepsilon)(C\varepsilon+(X^\top X)^{-1}X^\top\varepsilon)^\top\right] = \mathbb{E}_\theta\left[C\varepsilon\varepsilon^\top C^\top+(X^\top X)^{-1}X^\top\varepsilon\varepsilon^\top(X^\top X)^{-1}X^\top\right]$$

Using linearity of expectation, let us evaluate each term separately:

$$= \sigma^2\cdot CC^\top+[\sigma^2\cdot(X^\top X)^{-1}X^\top CX(X^\top X)^{-1}]+\sigma^2(X^\top X)^{-1}$$

$$= \sigma^2\cdot CC^\top+0+\sigma^2(X^\top X)^{-1} = \operatorname{cov}(\theta^*)+\sigma^2\cdot CC^\top\geq\operatorname{cov}(\theta^*)$$

Thus, we have shown that our estimator θ^* is efficient, as claimed. □

Finally, let us obtain one more small result. Namely, let us try to estimate the variance of the noise, which was originally unknown.

Proposition.

$$\frac{1}{n-k}S(\theta^*)$$

is an unbiased estimator for σ^2.

Proof.
Let us evaluate the expectation of $S(\theta^*)$:

$$\mathbb{E}_\theta S(\theta^*) = \mathbb{E}_\theta \left[(Y - X\theta^*)^\top (Y - X\theta^*)\right]$$

$$= \mathbb{E}_\theta \left[(X\theta + \varepsilon - X(X^TX)^{-1}X^T(X\theta+\varepsilon))^\top (X\theta + \varepsilon - X(X^TX)^{-1}X^T(X\theta+\varepsilon))\right]$$

$$\mathbb{E}_\theta \left(\left(\epsilon - X(X^\top X)^{-1}X^\top \epsilon\right)^\top \left(\epsilon - X(X^\top X)^{-1}X^\top \epsilon\right)\right)$$

$$= \mathbb{E}_\theta \left[\epsilon^\top \left(E_n - X(X^\top X)^{-1}X^\top\right)^\top \left(E_n - X(X^\top X)^{-1}X^\top\right)\epsilon\right] =: \mathbb{E}_\theta \left(\epsilon^\top B^\top B \epsilon\right)$$

Since our errors are independent and their variances coincide, we can rewrite the obtained expectation as follows:

$$\mathbb{E}_\theta \left(\epsilon^\top B^\top B\epsilon\right) = \sum_{i,j} (B^\top B)_{ij}\mathbb{E}_\theta(\epsilon_i\epsilon_j) = \sum_{i=1}^{n} (B^\top B)_{ii}\mathbb{E}_\theta\epsilon_i^2 = \sigma^2 \cdot \mathrm{tr}(B^\top B)$$

It remains to determine what the trace of our matrix is equal to: $n - k$. Indeed,

$$B^\top B = E_n - X(X^\top X)^{-1}X^\top + X(X^\top X)^{-1}X^\top - X(X^\top X)^{-1}X^\top = E_n - X(X^\top X)^{-1}X^\top$$

$$\mathrm{tr}(B^\top B) = \mathrm{tr}(E_n) - \mathrm{tr}(X(X^\top X)^{-1}X^\top) = n - \mathrm{tr}((X^\top X)^{-1}X^\top X) = n - k$$

End of the 18-th Seminar.

Hypothesis testing.

Problem Statement

Let us recall the formulation of the hypothesis testing problem in its simplest form.

Let $X_1, \ldots, X_n \sim P_\theta$ be a random sample whose distribution depends on a parameter $\theta \in \Theta$.

Assume that the parameter set Θ can be represented as $\Theta = \Theta_0 \cup \Theta_1$.

Then the statements H_0 and H_1 are formally called hypotheses, expressing to which of the sets Θ_0 or Θ_1 the parameter θ belongs.

Usually, the hypothesis H_0 is called the null (or basic) hypothesis, while H_1 is called the alternative hypothesis.

Accordingly, our task is to use the sample to determine which of the two hypotheses should be accepted and which should be rejected.

Since we do not initially know whether the parameter θ lies in Θ_0 or Θ_1, accepting one of the hypotheses may lead either to a correct decision or to an error. Thus, we may distinguish four possible outcomes:

	accept H_0	accept H_1
true H_0	✓	Type I error
true H_1	Type II error	✓

All we want now is to learn how to choose between the hypotheses in some "optimal" way.

Let $X_1, \ldots, X_n \sim P_\theta$ be a random sample, $\theta \in \Theta$, and suppose we consider the hypotheses $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$, where $\Theta = \Theta_0 \cup \Theta_1$.

A *test* is a function $\varphi : \mathbb{R}^n \to \{0, 1\}$ that, given the sample X, outputs the probability of rejecting H_0.

Based on this test, we will decide which hypothesis should be accepted.

Remark. It is assumed that if the test φ evaluated at our sample yields 0 or 1, then we make an unambiguous decision between the two hypotheses. Otherwise, we would need to perform some independent experiment and base the final decision on it.

For example, in such a situation one may flip a "biased coin" whose one outcome occurs with probability $\varphi(X)$.

Remark. We may also observe that the error probabilities are expressed as follows:

$$P(\text{Type I error}) = E_0\,\varphi(X), \qquad P(\text{Type II error}) = E_1(1-\varphi(X)).$$

Now let us describe in detail a possible formulation of our problem.

Problem. Suppose we are given some number $\alpha \in (0,1)$. Then, for a given significance level α, we must construct a test φ such that
1) $P(\text{Type I error}) \le \alpha$
2) $P(\text{Type II error}) \to \min$

The power of such a test is defined as $1 - P(\text{Type II error})$.

Remark. As a consequence of the previous remark, we may obtain an alternative expression for the power of a test:

$$\text{Power of the test} = E_1\,\varphi(X).$$

Let us consider a pair of examples to better understand how tests can be constructed. We will restrict ourselves to the construction of some level–α tests without delving into their full power analysis.

Examples.

1. Let $X_1, \dots, X_n \sim \mathcal{N}(\theta, \sigma^2)$ be a sample from a normal distribution with known variance σ^2.

We consider the hypotheses:

$$H_0 : \theta = \theta_0, \qquad H_1 : \theta \neq \theta_0.$$

Note that if H_0 is true, then by Fisher's lemma the following holds:

$$Y := \frac{\sqrt{n}\,(\overline{X} - \theta_0)}{\sigma} \sim \mathcal{N}(0,1).$$

We now learn how to construct a level–α test using this fact. We make use of the same "trick" as when constructing confidence intervals:

$$\alpha = P(|Y| > z_\alpha) = 2(1 - \Phi(z_\alpha)) \quad \Longrightarrow \quad z_\alpha = \Phi^{-1}\Big(1 - \frac{\alpha}{2}\Big).$$

Hence we define the test:

$$\varphi(X) := \begin{cases} 1, & \left|\dfrac{\sqrt{n}\,(\overline{X} - \theta_0)}{\sigma}\right| > z_\alpha, \\ 0, & \left|\dfrac{\sqrt{n}\,(\overline{X} - \theta_0)}{\sigma}\right| \le z_\alpha. \end{cases}$$

We chose z_α in such a way that

$$P\left(\left|\frac{\sqrt{n}\,(\overline{X} - \theta_0)}{\sigma}\right| > z_\alpha\right) = \alpha,$$

so that the probability of a Type I error (i.e., an error when the true value of θ is exactly θ_0) is exactly equal to our significance level α.

We will discuss the power of this test later.

2. Suppose in the previous example the variance σ^2 is unknown.

Then we can again use Fisher's lemma, estimating σ^2 by the sample variance S, and obtain the following form of the criterion:

$$\varphi(\mathbf{X}) = \begin{cases} 1, & \sqrt{n}\frac{|\overline{X}-\theta_0|}{S} \geq z_\alpha \\ 0, & \sqrt{n}\frac{|\overline{X}-\theta_0|}{S} < z_\alpha \end{cases}$$

where $z_\alpha = F^{-1}_{T_{n-1}}\left(1-\frac{\alpha}{2}\right)$

End of the 19-th Seminar.

Most Powerful Criteria

Usually, we want not only to control the probability of a Type I error, but, all else equal, to minimize the probability of a Type II error.

To address this problem, let's consider a more general version of hypothesis testing and see what the criterion will look like in this case.

Note. If a hypothesis specifies only one distribution, it is called simple; otherwise, it is called complex.

Let the hypotheses be

$$H_0: \ \theta \in \Theta_0, \qquad H_1: \ \theta \in \Theta_1,$$

both simple. A test $\phi(X)$ of level α is called *most powerful* (MP) at level α if:

1. It has level α:
$$\mathbb{E}_{\theta_0}[\phi(X)] \leq \alpha,$$
2. For any other test $\psi(X)$ of the same level,
$$\mathbb{E}_{\theta_1}[\phi(X)] \geq \mathbb{E}_{\theta_1}[\psi(X)].$$

Definition (Uniformly Most Powerful Test). Let the alternative hypothesis be composite,

$$H_1: \ \theta \in \Theta_1.$$

A test $\phi(X)$ of level α is called *uniformly most powerful* (UMP) if:

1. It has level α:
$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\phi(X)] \leq \alpha,$$
2. For all $\theta \in \Theta_1$ and for any other test $\psi(X)$ of level α,
$$\mathbb{E}_\theta[\phi(X)] \geq \mathbb{E}_\theta[\psi(X)].$$

Lemma (Neyman-Pearson).

Let $X_1, \ldots, X_n \sim P$ be a random sample from some distribution.

Consider the simple hypotheses: $H_0 : P = P_0$ and $H_1 : P = P_1$.

Let f_0 and f_1 be the densities of P_0 and P_1 with respect to some measure μ, and consider the family of criteria of the following form:

$$\varphi(\mathbf{X}) = \begin{cases} 1, & f_1(\mathbf{X}) > k \cdot f_0(\mathbf{X}) \\ p(k), & f_1(\mathbf{X}) = k \cdot f_0(\mathbf{X}) \\ 0, & f_1(\mathbf{X}) < k \cdot f_0(\mathbf{X}) \end{cases} \qquad (*)$$

where

$$p(k) = \begin{cases} 0, & \mathbb{P}\left(f_1(\mathbf{X}) = k \cdot f_0(\mathbf{X})\right) = 0 \\ \frac{\alpha_0 - \alpha(k)}{\alpha(k_0) - \alpha(k)}, & \mathbb{P}\left(f_1(\mathbf{X}) = k \cdot f_0(\mathbf{X})\right) > 0 \end{cases}$$

with $\alpha_0 \in (0,1),\ \alpha(k) = \mathbb{P}_0\left(f_1(\mathbf{X}) > k \cdot f_0(\mathbf{X})\right)$.

Then the following statements hold: 1) For every $\alpha_0 \in (0,1)$ there exists a criterion of level exactly α_0 of the form $(*)$. 2) This criterion is the most powerful among all criteria of level $\leq \alpha_0$. 3) Any most powerful criterion of level α_0 has the form $(*)$ (up to the values $p(k)$).

Proof.

1. Examine the function $\alpha(k)$ carefully, which corresponds to the probability of a Type I error and defines the criterion's level. It is monotonically decreasing and right-continuous, with $\alpha(-\infty) = 1$ and $\alpha(+\infty) = 0$. Thus, there are two cases. In the first case, there exists a k such that $\alpha(k) = \alpha_0$.

In this situation, by right-continuity, we have $\mathbb{P}(f_1(\mathbf{X}) = k \cdot f_0(\mathbf{X})) = 0$. Then the criterion has the form:

$$\varphi(\mathbf{X}) := \begin{cases} 1, & f_1(\mathbf{X}) > k \cdot f_0(\mathbf{X}) \\ 0, & f_1(\mathbf{X}) < k \cdot f_0(\mathbf{X}) \end{cases}$$

Otherwise, by right-continuity, there is a k such that $\alpha(k-0) > \alpha_0 > \alpha(k)$.

Let's see that in this case the criterion will have exactly the form $(*)$.

Indeed, calculate the probability of a type I error:

$$\mathbb{E}_0\varphi = \mathbb{P}_0\left(f_1(\mathbf{X}) > k \cdot f_0(\mathbf{X})\right) + \mathbb{P}_0\left(f_1(\mathbf{X}) = k \cdot f_0(\mathbf{X})\right) \cdot \frac{\alpha_0 - \alpha(k)}{\alpha(k-0) - \alpha(k)} = \alpha_0$$

Consider any other criterion $\tilde{\varphi}$ of level $\leq \alpha_0$, i.e. $\mathbb{E}_0\tilde{\varphi} \leq \alpha_0$. Show that this criterion is not more powerful, i.e.

$$\mathbb{E}_1\varphi \geq \mathbb{E}_1\tilde{\varphi}$$

Indeed,

$$\mathbb{E}_1(\varphi - \tilde{\varphi}) \geq \mathbb{E}_1(\varphi - \tilde{\varphi}) - k \cdot \mathbb{E}_0(\varphi - \tilde{\varphi}) = \int_{\mathbb{X}^n} (\varphi - \tilde{\varphi})(f_1(x) - kf_0(x))d\mu \geq 0$$

Let's show the last inequality. Consider three cases: 1) $f_1(x) > k \cdot f_0(x) \implies \varphi(x) = 1 \geq \tilde{\varphi}(x)$, so $(\varphi(x) - \tilde{\varphi}(x))(f_1(x) - kf_0(x)) \geq 0$ 2) $f_1(x) < k \cdot f_0(x) \implies \varphi(x) = 0 \leq \tilde{\varphi}(x)$, so $(\varphi(x) - \tilde{\varphi}(x))(f_1(x) - kf_0(x)) \geq 0$ 3) $f_1(x) = k \cdot f_0(x) \implies (\varphi(x) - \tilde{\varphi}(x))/\cdots \geq 0$

In all three cases, the integrand is non-negative, so the integral itself is non-negative, as required.

Let $\tilde{\varphi}$ be any most powerful criterion of level $\leq \alpha_0$.

By the result above, we have:

$$\int_{\mathbb{X}^n} (\varphi - \tilde{\varphi})(f_1(x) - kf_0(x))d\mu = 0$$

Since the integrand is non-negative, we could only obtain zero if the function itself is zero. Then it follows that $\varphi = \tilde{\varphi}$ almost everywhere where $f_1(x) \neq kf_0(x)$, which was required.

Nevertheless, in real life, it is quite rare that we need to choose between a simple null hypothesis and a simple alternative.

Let's see how one can proceed when the alternative is composite.

Statement

Let $X_1, \ldots, X_n$ be a sample from the distribution P_θ.

Suppose for any pair of parameters θ_1 and θ_2 there is a function $\psi_{\theta_0,\theta}$ such that:

$$\frac{f_\theta(\mathbf{X})}{f_{\theta_0}(\mathbf{X})} = \psi_{\theta_0,\theta}(T(\mathbf{X}))$$

where $T(\mathbf{X})$ is some statistic, and for all $\theta_1 < \theta_2$ the function $\psi_{\theta_0,\theta}$ increases and is continuous.

Thus, we learn how to distinguish between the hypotheses:

$$H_0 : \theta = \theta_0, \quad H_1 : \theta > \theta_0$$

Let's consider any parameter $\theta_1 > \theta_0$.
To test the null hypothesis, we would like to have a criterion of the form $\psi_{\theta_0,\theta_1}(T(\mathbf{X})) > k$. That is:

$$\frac{f_{\theta_1}(\mathbf{X})}{f_{\theta_0}(\mathbf{X})} > k \iff \psi_{\theta_0,\theta_1}(T(\mathbf{X})) > k \iff T(\mathbf{X}) > c$$

As before, if the given condition is fulfilled, the null hypothesis is rejected; otherwise, it is accepted.
Now, if we want to obtain a criterion for the significance level α, we must choose c so that

$$P_{\theta_0}\left(T(\mathbf{X}) > c\right) = \alpha,$$

since, under the null hypothesis, we want to make a mistake with probability no greater than α.

Thus, it is evident that c depends only on θ_0 and α, and does not depend on the value of the alternative parameter θ_1. That is, we obtain a criterion for testing a simple null hypothesis against a composite alternative (for the situation described above).

Note.
If $\psi_{\theta_0,\theta}$ decreases for $\theta_1 < \theta_2$, then it becomes possible to test the hypotheses:

$$H_0 : \theta = \theta_0, \quad H_1 : \theta < \theta_0$$

Example.
Suppose there is a sample $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$, and σ^2 is known.
Test the hypotheses:

$$H_0 : \theta = \theta_0, \quad H_1 : \theta > \theta_0$$

By the Neyman–Pearson lemma, to build the criterion, we need to consider the likelihood ratio:

$$\frac{f_1(\mathbf{X})}{f_0(\mathbf{X})} = \exp\left(\frac{1}{2\sigma^2}\sum_{k=1}^{n}\left((X_k-\theta_0)^2 - (X_k-\theta_1)^2\right)\right) = \exp\left(\frac{\theta_1-\theta_0}{2\sigma^2}\sum_{k=1}^{n}(2X_k-\theta_1-\theta_0)\right) > k$$

Then

$$\left(\frac{\theta_1-\theta_0}{2\sigma^2}\sum_{k=1}^{n}(2X_k-\theta_1-\theta_0)\right) > \ln(k) \Leftrightarrow \sum_{k=1}^{n}(2X_k-\theta_1-\theta_0) > \frac{2\sigma^2\ln(k)}{\theta_1-\theta_0} =: c = const \Leftrightarrow \overline{X} > \overline{c} := \frac{c}{n}.$$

Finding the constant $\overline{C}$ from the Type I error constraint
We require the test to satisfy

$$P_\theta(\text{reject } H_0 \mid \theta = \theta_0) = \alpha.$$

Rejecting H_0 when $\overline{X} > \overline{C}$ gives

$$P_{\theta_0}(\overline{X} > \overline{C}) = P\left(\frac{\sqrt{n}(\overline{X}-\theta_0)}{\sigma} > \frac{\sqrt{n}(\overline{C}-\theta_0)}{\sigma}\right) = \alpha.$$

Let $Z \sim N(0,1)$. Then

$$P(Z > z_\alpha) = \alpha \quad \Longleftrightarrow \quad z_\alpha = \Phi^{-1}(1-\alpha).$$

So we must choose

$$\frac{\sqrt{n}(\overline{C}-\theta_0)}{\sigma} = z_\alpha.$$

Solving for $\overline{C}$:

$$\boxed{\overline{C} = \theta_0 + \frac{\sigma}{\sqrt{n}}\, z_\alpha}.$$

Thus, the NP most powerful level–α test is

$$\boxed{\varphi(\mathbf{X}) = \mathbf{1}\left(\overline{X} > \theta_0 + \frac{\sigma}{\sqrt{n}}\, z_\alpha\right).}$$

Power of the test
Under the alternative $\theta > \theta_0$,

$$\overline{X} \sim N\left(\theta, \frac{\sigma^2}{n}\right).$$

Hence the power is

$$\beta(\theta) = P_\theta(\text{reject } H_0) = P_\theta\left(\overline{X} > \theta_0 + \frac{\sigma}{\sqrt{n}}\, z_\alpha\right) = P\left(Z > \frac{\sqrt{n}(\theta_0 - \theta)}{\sigma} + z_\alpha\right).$$

Since $\theta > \theta_0$, the numerator $\theta_0 - \theta < 0$, and thus

$$\frac{\sqrt{n}(\theta_0 - \theta)}{\sigma} \longrightarrow -\infty \qquad (n \to \infty).$$

Therefore

$$\beta(\theta) = 1 - \Phi\left(z_\alpha + \frac{\sqrt{n}(\theta_0 - \theta)}{\sigma}\right) \longrightarrow 1.$$

$$\boxed{\text{The power of the test tends to 1 for every } \theta > \theta_0.}$$

This shows that the test is *consistent*.

Lemma.
If $\{f(x,\theta), \quad \theta \in R\}$ and $\psi_{\theta_0,\theta}$ is monotonous, then there exists the UMP criteria for testing the hypothesis $H_0 : \theta \le \theta_0, \quad H_1 : \theta > \theta_0$

Likelihood Ratio Statistic

Let observations

$$X_1, \ldots, X_n$$

have a density (or probability mass function)

$$f(x;\theta), \qquad \theta \in \Theta \subset \mathbb{R}^p,$$

where Θ is the parameter space of the full model.

We want to test the hypothesis

$$H_0 : \theta \in \Theta_0 \quad \text{against} \quad H_1 : \theta \in \Theta \setminus \Theta_0,$$

where $\Theta_0 \subset \Theta$ is a subspace of lower dimension, $\dim(\Theta_0) = p - k$. Thus the number of restrictions is

$$k = p - \dim(\Theta_0).$$

The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f(X_i;\theta) \quad \text{or} \quad \sup_\theta \prod_{i=1}^{n} f(X_i;\theta)$$

The *likelihood ratio statistic* is

$$\Lambda = \frac{\sup\{P_\theta(x): \quad \theta \in \Theta_0\}}{\sup\{P_\theta(x): \quad \theta \in \Theta\}} =: \frac{L(\hat{\theta}_0)}{L(\hat{\theta})}, \qquad 0 \leq \Lambda \leq 1.$$

The commonly used transformed statistic is

$$-2\ln\Lambda = -2\left(\ln L(\hat{\theta}_0) - \ln L(\hat{\theta})\right).$$

Wilks' Theorem.

Assume that standard regularity conditions of maximum likelihood theory hold (smoothness, identifiability, existence of Fisher information, etc.). Then, under the null hypothesis H_0 and as $n \to \infty$,

$$-2\ln\Lambda \xrightarrow{d} \chi^2_k,$$

where

$$k = \dim(\Theta) - \dim(\Theta_0)$$

is the number of independent restrictions imposed by H_0.
Example. $X_1, ..., X_n \sim N(\theta, \sigma^2)$, σ^2 is unknown. $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$.

Tests for Interval-Type Numerical Data

Definition.
Suppose $X_1, \ldots, X_n \sim P$, Hypothesis $H_0 : P \sim P_0$ is called a simple goodness-of-fit hypothesis.
Nevertheless, sometimes we want to test not the equality to a specific distribution, but the membership in an entire family of distributions P_θ In such a case, the goodness-of-fit hypothesis is called composite
definition
Let $X_1, \ldots, X_n \sim P$, $Y_1, \ldots, Y_m \sim Q$.
Hypothesis $H_0 : P \sim Q$ is called hypothesis of homogency.
Definition.
Let $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$, Hypothesis:

$$H_0 : \exists P_1, P_2 P(x, y) = P_1(x) \cdot P_2(y)$$

is called a hypothesis of statistical independence.
In interval-type data, when we want to determine whether two distributions coincide, the empirical distribution functions become extremely useful.
These functions form the basis for many statistical tests. Essentially, they allow us to assess whether the observed sample supports H_0, or whether deviations are large enough to reject it.
Definition.
The empirical distribution function uses the following statistic:

$$\sup_x |F_n(x) - F_0(x)|,$$

where F_n and F_m are empirical distribution functions of the two samples.
We will consider the following statistics:
1. $\sup_x |F_n(x) - F_0(x)|$
2. $\sup_x (F_n(x) - F_0(x))$

3. $\sup_x(F_0(x) - F_n(x))$
4. $\int_0^1 (F_n(x) - F_0(x))^2 \, dF_0(x)$

The first of these measures is called the Kolmogorov statistic; the second and the third are the Smirnov statistics; and the fourth is the Cramér–von Mises statistic (its one-sided version).

Together, these three metrics are often referred to as the Kolmogorov–Smirnov metrics.

Additionally, note that these metrics behave quite well.

Theorem. If F_0 is continuous and H_0 holds, then the distribution of the statistic does not depend on F_0.

Proof.

We prove the theorem using the Kolmogorov theorem.

Assume H_0 holds. Then F_0 is continuous. Hence we can write:

$$T = \sup_x |F_n(x) - F_0(x)| - \sup_u |U_n(u) - u|.$$

Make the substitution $u = F_0(x)$.

Thus, $F_0(F_0^{-1}(u)) = u$ for $u \in (0, 1)$, and

$$F_n(F_0^{-1}(u)) = \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\{X_k \le F_0^{-1}(u)\}}.$$

But this quantity has the same distribution as

$$\frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\{U_k \le u\}}, \quad U_k \sim U[0, 1].$$

Hence the statistic becomes

$$T = \sup_u \left| \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\{U_k \le u\}} - u \right|,$$

which clearly does not depend on the original distribution F_0, as required.

□

Thus, the distribution of these metrics does not depend on the underlying F_0. Moreover, the following results are also known:

Statement.

1. $n \sup_x |F_n(x) - F_0(x)| \xrightarrow{d} K$, where K has the Kolmogorov distribution.
2. $\sqrt{n} \sup_x (F_n(x) - F_0(x)) \xrightarrow{d} \text{Rayleigh}(\frac{1}{2})$
4. $n \int_0^1 (F_n(x) - F_0(x))^2 \, dF_0(x) \xrightarrow{d} \omega^2$, where ω^2 has the one-sided Cramér–von Mises distribution.

Kolmogorov Distribution

The limiting distribution of the Kolmogorov–Smirnov statistic

$$D_n = \sup_x |F_n(x) - F(x)|$$

as $n \to \infty$ has the Kolmogorov distribution with the cumulative distribution function

$$K(t) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2 t^2} = \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2 t^2}, \qquad t > 0.$$

Omega-Square (Cramér–von Mises) Distribution

The limiting distribution of the Cramér–von Mises statistic

$$\omega_n^2 = \int_{-\infty}^{+\infty} \left(F_n(x) - F(x)\right)^2 \, dF(x)$$

as $n \to \infty$ has the following cumulative distribution function:

$$F_{\Omega^2}(x) = \sum_{k=1}^{\infty} (-1)^{k-1} \exp\left(-\frac{k^2\pi^2}{8x}\right), \qquad x > 0.$$

A more precise classical expansion (Anderson–Darling form) is

$$F_{\Omega^2}(x) = \sum_{k=1}^{\infty} \left(1 - 2k^2\pi^2 x\right) \exp\left(-k^2\pi^2 x\right).$$

These results give us the basis for constructing hypothesis tests — specifically, testing whether two distributions are equal.

Let us consider the hypothesis of homogency $H_0 : F = G$.

Let $X_1, \ldots, X_n \sim F$ and $Y_1, \ldots, Y_m \sim G$ be independent.

Under H_0 we want to check whether there exists a difference between the empirical distribution functions of the two samples.

To this end, one can introduce the two-sample Kolmogorov–Smirnov statistic:

$$\sqrt{\frac{nm}{n+m}} \sup_x |F_n(x) - G_m(x)|.$$

It is known that under H_0 this statistic converges in distribution to the Kolmogorov distribution.

Similarly,

$$\sqrt{\frac{nm}{n+m}} \sup_x (F_n(x) - G_m(x)) \quad \xrightarrow{d} \quad \text{Rayleigh}(\frac{1}{2}),$$

and

3. $\dfrac{nm}{n+m} \displaystyle\int_0^1 (F_n(x) - G_m(x))^2 \, dF_0(x) \xrightarrow{d} \omega^2$

We prove the second statement in the case when the sample sizes are equal.

Statement. Let samples $X_1, \ldots, X_n \sim F$ and $Y_1, \ldots, Y_n \sim G$ be given. Then if $F = G$ and F, G are continuous, we have:

$$\mathbb{P}\left(\sqrt{\frac{n}{2}} \sup_t \big(F_n(t) - G_n(t)\big) < z\right) \longrightarrow F_{\text{Rayleigh}(1/2)}(z) = 1 - e^{-2z^2}.$$

Proof. Rewrite the probability:

$$\mathbb{P}\left(\sqrt{\frac{n}{2}} \sup_t \big(F_n(t) - G_n(t)\big) > z\right) = \mathbb{P}\left(n \sup_t \big(F_n(t) - G_n(t)\big) > z\sqrt{2n}\right).$$

Let us understand how we can express the left-hand side of the obtained inequality differently. For this, consider the combined order statistics of both samples

$$Z_1 < Z_2 < \cdots < Z_{2n},$$

and introduce the following random variable:

$$\eta_k := \begin{cases} 1, & \text{if } Z_k \in X, \\ -1, & \text{if } Z_k \in Y, \end{cases} \qquad S_k := \sum_{i=1}^{k} \eta_i.$$

Then:

$$n \sup_t \left(F_n(t) - G_n(t)\right) = \sup_t \left(\sum_{k=1}^{n} \mathbf{1}\{X_k < t\} - \sum_{k=1}^{n} \mathbf{1}\{Y_k < t\} \right) = \sup_t \left(\sum_{Z_k < t} \eta_k \right) = \max_k S_k.$$

Hence, the original probability can be rewritten as:

$$\mathbb{P}\Big(\max_k S_k > z\sqrt{2n}\Big) = \mathbb{P}\Big(\max_k S_k > r\Big), \qquad \text{where } r = \Big\lfloor z\sqrt{2n} \Big\rfloor + 1.$$

In essence, we have reduced our problem to a problem about random walks.

In particular, any sequence of values η_k corresponds to some diagonal path on the coordinate grid from point $(0,0)$ to point $(2n,0)$. And among such paths we need to compute the probability that this path intersects the line $y = r$.

More precisely, since under $F = G$ each value η_k equals 1 or -1 with probability $1/2$, the problem is equivalent to counting such paths. And we can resolve this using the method of reflection.

Namely, consider any path that touches or crosses the line $y = r$. If we reflect the portion of the path after the moment it crosses $y = r$ with respect to the line $y = r$, then we obtain a path from $(0,0)$ to $(2n, 2r)$. It is easy to see that a one-to-one correspondence exists between the paths reaching the line $y = r$ and the paths from $(0,0)$ to $(2n, 2r)$.

Thus the task is reduced to counting the number of diagonal paths from $(0,0)$ to $(2n, 2r)$. The desired probability can then be written as:

$$\mathbb{P}\left(\max_k S_k \geq r\right) = \mathbb{P}(S_{2n} = 2r) = \frac{\#\{\text{paths from } (0,0) \text{ to } (2n,2r)\}}{\binom{2n}{n}} = \frac{\binom{2n}{n+r}}{\binom{2n}{n}}.$$

All that remains is to compute the asymptotics and show that the resulting expression tends to e^{-2z^2}.

$$\ln\left(\frac{\binom{2n}{n+r}}{\binom{2n}{n}}\right) = \ln\left(\frac{(2n)!}{(n+r)!(n-r)!}\right) - \ln\left(\frac{(2n)!}{(n!)^2}\right) = \ln\left(\frac{(n!)^2}{(n+r)!(n-r)!}\right).$$

Using Stirling's approximation:

$$\ln\left(\frac{(n!)^2}{(n+r)!(n-r)!}\right) = \ln\left(\frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{\sqrt{2\pi(n+r)}\left(\frac{n+r}{e}\right)^{n+r}\sqrt{2\pi(n-r)}\left(\frac{n-r}{e}\right)^{n-r}}\right) + o(1)$$

$$= \ln\left(\frac{2\pi n\left(\frac{n}{e}\right)^{2n}}{2\pi\sqrt{(n+r)(n-r)}\left(\frac{n+r}{e}\right)^{n+r}\left(\frac{n-r}{e}\right)^{n-r}}\right) + o(1)$$

$$= \ln\left(\frac{n}{\sqrt{(n+r)(n-r)}} \cdot \frac{n^{2n}}{(n+r)^{n+r}(n-r)^{n-r}}\right) + o(1)$$

$$= \ln\left(\left(\frac{n}{n+r}\right)^{n+r}\left(\frac{n}{n-r}\right)^{n-r}\right) + o(1)$$

$$= (n+r)\ln\left(1-\frac{r}{n+r}\right) + (n-r)\ln\left(1+\frac{r}{n-r}\right) + o(1).$$

Expand the logarithms via Taylor series, noting additionally that $r \sim \sqrt{n}$:

$$= -(n+r)\left(\frac{r}{n} + \frac{r^2}{2n^2} + o\left(\frac{r^2}{n^2}\right)\right) + (n-r)\left(\frac{r}{n} - \frac{r^2}{2n^2} + o\left(\frac{r^2}{n^2}\right)\right)$$

$$= -\left(r + \frac{r^2}{2n} + o(1)\right) - \left(r - \frac{r^2}{2n} + o(1)\right) = -\frac{2r^2}{2n} + o(1) = -\frac{r^2}{n} + o(1) = -2z^2 + o(1).$$

Thus the required asymptotics is:

$$\frac{\binom{2n}{n+r}}{\binom{2n}{n}} \longrightarrow e^{-2z^2}.$$

The proof of the first statement is given for the case when the sample sizes are equal.

Let us now move on to discussing various tests for different statistical hypotheses. As a rule, the applicability of these tests often depends on the type of data, so we will begin by considering the most common case of interval-type observations.

To make the discussion clearer, let us first look at several simple examples of tests where the underlying distribution is assumed to be normal.

Example.

Let us assume that we have one sample $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ and we know how to test the hypotheses $H_0 : \theta = \theta_0$, when σ^2 is known and for unknown.

Now let $X_1, ..., X_n \sim N(\theta_1, \sigma_1^2)$ and $Y_1, ..., Y_m \sim N(\theta_2, \sigma_2^2)$. Test the hypothesis $H_0 : \quad \theta_1 = \theta_2$. We will consider three cases depending on what is known about σ_1, and σ_2.

Case 1. σ_1 and σ_2 are known.

Let

$$T = \overline{X} - \overline{Y} \sim (0, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}).$$

Then

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0,1).$$

Thus, using the standard normal distribution and knowing σ, we can determine the distribution of this statistic and construct a test based on it.

Case 2. $\sigma_1^2 = \sigma_2^2 = \sigma^2$ are unknown.

Consider the variance

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \bar{Y} = \frac{1}{m}\sum_{j=1}^{m} Y_j$$

$$\tilde{S}_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad \tilde{S}_Y^2 = \frac{1}{m}\sum_{j=1}^{m}(Y_j - \bar{Y})^2$$

$$\tilde{S}_p^2 = \frac{n\tilde{S}_X^2 + m\tilde{S}_Y^2}{n+m}$$

$$SE = \sqrt{\tilde{S}_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}, \quad T = \frac{\bar{X} - \bar{Y}}{SE} = \frac{\bar{X} - \bar{Y}}{\tilde{S}_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

$$n\frac{\tilde{S}_X^2}{\sigma^2} \sim \chi^2_{n-1}, \quad m\frac{\tilde{S}_Y^2}{\sigma^2} \sim \chi^2_{m-1}, \quad n\frac{\tilde{S}_X^2}{\sigma^2} + m\frac{\tilde{S}_Y^2}{\sigma^2} \sim \chi^2_{n+m-2}$$

$$T = \frac{Z}{\sqrt{V/(n+m-2)}}, \quad Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0,1),\ V \sim \chi^2_{n+m-2}$$

Note that the distribution of this statistic does not depend on θ, making it correctly defined.

Thus, we have obtained a new useful statistic, which allows us to build various nonparametric tests (of the "gold standard" type, as they are sometimes called).

Case 3. σ_1^2 and σ_2^2 are unknown.

In this case, we can use the fact that the difference between the sample means follows a normal distribution (The Behrens–Fisher Problem (he Lindberg–Feller theorem)). Thus, using the central limit theorem, we can construct the required test statistic.

Nevertheless, in practice we rarely know the underlying distribution exactly. Most of the time we only assume that the sample is close to normal. In this case, statistical tests differ both in terms of hypotheses tested and in terms of the statistics used.

The Lindeberg–Feller Central Limit Theorem

Let $X_1, X_2, \ldots$ be independent (but not necessarily identically distributed) random variables with

$$S_n = \sum_{k=1}^{n} X_k, \qquad B_n^2 = \sum_{k=1}^{n} \mathrm{Var}(X_k).$$

The Lindeberg condition is the requirement that for every $\varepsilon > 0$,

$$\frac{1}{B_n^2} \sum_{k=1}^{n} \mathbb{E}\big[X_k^2 \mathbf{1}_{\{|X_k| > \varepsilon B_n\}}\big] \longrightarrow 0, \qquad n \to \infty.$$

Lindeberg–Feller CLT. If the Lindeberg condition holds, then

$$\frac{S_n - \mathbb{E}S_n}{B_n} \xrightarrow{d} N(0,1).$$

Appearance of the Lindeberg–Feller Theorem in the Behrens–Fisher Problem

Consider two independent samples

$$X_1, \ldots, X_n \sim N(\mu_X, \sigma_X^2), \qquad Y_1, \ldots, Y_m \sim N(\mu_Y, \sigma_Y^2),$$

with both variances σ_X^2 and σ_Y^2 unknown and not assumed to be equal.

We test the hypotheses

$$H_0 :\ \mu_X = \mu_Y, \qquad H_1 :\ \mu_X \neq \mu_Y.$$

Let

$$\overline{X} = \frac{1}{n} \sum_{k=1}^{n} X_k, \qquad \overline{Y} = \frac{1}{m} \sum_{j=1}^{m} Y_j,$$

and the empirical variances

$$s_X^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \overline{X})^2, \qquad s_Y^2 = \frac{1}{m-1} \sum_{j=1}^{m} (Y_j - \overline{Y})^2.$$

To obtain the asymptotic distribution of the statistic

$$T_n = \sqrt{\frac{mn}{m+n}}\,\frac{\overline{X}-\overline{Y}}{\sqrt{s_X^2+s_Y^2}},$$

we first analyze the numerator. Under H_0,

$$\sqrt{\frac{mn}{m+n}}\,(\overline{X}-\overline{Y}) = \sum_{k=1}^{n} a_{k,n}(X_k-\mu_X) + \sum_{j=1}^{m} b_{j,n}(Y_j-\mu_Y),$$

where the coefficients

$$a_{k,n} = \sqrt{\frac{m}{n(m+n)}}, \qquad b_{j,n} = -\sqrt{\frac{n}{m(m+n)}},$$

produce a weighted sum of independent but non-identically distributed random variables.

Because the terms in this sum have different variances ($\sigma_X^2 \neq \sigma_Y^2$ in general), the classical Lyapunov CLT does not directly apply. Instead, one verifies that the Lindeberg condition holds, which yields

$$\sqrt{\frac{mn}{m+n}}\,(\overline{X}-\overline{Y}) \;\xrightarrow{d}\; N(0,\ \sigma_X^2+\sigma_Y^2).$$

Finally, since $s_X^2 \to \sigma_X^2$ and $s_Y^2 \to \sigma_Y^2$ in probability, Slutsky's theorem gives

$$T_n = \sqrt{\frac{mn}{m+n}}\,\frac{\overline{X}-\overline{Y}}{\sqrt{s_X^2+s_Y^2}} \;\xrightarrow{d}\; N(0,1).$$

Thus, the Lindeberg–Feller theorem is the key tool that ensures the asymptotic normality of the Behrens–Fisher test statistic when the variances are unknown and unequal.

Tests for Categorical and Discrete Numerical Data

We have discussed some tests that work well in situations when our data have a continuous nature.

However, such tests perform poorly for data that take values from some discrete set. For example, these may be numerical data taking only integer values, or non-numerical data where each element corresponds to some category (label).

Nevertheless, for data of this type there also exist many tests. One of the most powerful among them is Pearson's chi-squared test.

Theorem (Pearson's chi-squared test)

Let a sample $X_1, \ldots, X_n \sim P$ be given, where the X_i are independent.

Assume the null hypothesis $H_0 : P \sim P_0$, where

$$P_0 : \begin{array}{c|c|c|c} x_1 & x_2 & \ldots & x_s \\ \hline p_1 & p_2 & \ldots & p_s \end{array}$$

Then the following statement holds:

$$\chi^2 := \sum_{i=1}^{s} \frac{(\nu_i - np_i)^2}{np_i} \;\xrightarrow[n\to\infty]{H_0}\; \chi^2_{s-1}, \qquad \text{where } \nu_i = \#\{k : X_k = x_i\}.$$

Remark. The fact is not necessary about finiteness of the space. However, in fact, it can also be used when our space is infinite (in particular, when it is continuous). For this, it is enough to choose an arbitrary partition of the space $\mathcal{X} = \mathcal{X}_1 \sqcup \ldots \sqcup \mathcal{X}_s$ and define $p_i := \mathbb{P}(X \in \mathcal{X}_i)$.

Remark. Moreover, in this theorem we do not use the structure of our space at all. All that interests us is the probabilities p_i of the event falling into each of the blocks. Therefore, the resulting test also works in the case of categorical data.

We now know how to use Pearson's test to check simple goodness-of-fit hypotheses. For composite hypotheses, the following theorem holds.

Theorem 6.7. Let a sample $X_1, \dots, X_n \sim P$ be given. And let a family of distributions $\{P_\theta \mid \theta \in \Theta \subset \mathbb{R}^d\}$ be given such that under P_θ:

	x_1	x_2	$\dots$	x_s
$P_\theta:$	$p_1(\theta)$	$p_2(\theta)$	$\dots$	$p_s(\theta)$

Define $\chi^2(\theta)$ as follows:

$$\chi^2(\theta) := \sum_{i=1}^{s} \frac{(\nu_i - np_i(\theta))^2}{np_i(\theta)}, \qquad \text{where } \nu_i = \#\{k : X_k = x_i\}.$$

And choose $\hat{\theta}$ such that $\chi^2(\hat{\theta}) = \inf_\theta \chi^2(\theta)$.

Then, if the null hypothesis $H_0 : P \in \{P_\theta\}$ is true, the following statement holds:

$$\chi^2(\hat{\theta}) = \sum_{i=1}^{s} \frac{(\nu_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})} \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2_{s-1-d}.$$

Remark. In fact, it is sufficient to use not the exact value $\hat{\theta}$ but any approximate estimate, since for a small change of the parameter the value of the function $\chi^2(\theta)$ also changes only slightly.

Remark. The number of degrees of freedom of the chi-square distribution should be interpreted as follows.

We will assume that our distribution is uniquely determined by $s-1$ parameters, since all probabilities except one are chosen arbitrarily, while the last one is determined uniquely. And since we estimate d parameters, the number of random parameters decreases to $s - d - 1$.

This theorem is useful not only because it allows us to construct tests for composite goodness-of-fit hypotheses, but also because with its help one can rather easily construct tests for homogeneity and independence.

Statement. Let samples $X_1, \dots, X_{n_1} \sim F$ and $Y_1, \dots, Y_{n_2} \sim G$ be given, where F and G are discrete distributions taking values $x_1, \dots, x_s$.

Suppose we test the hypothesis of homogeneity $H_0 : F = G$.

For each outcome we count the number of corresponding events:

	x_1	x_2	$\dots$	x_s
X	$\nu_{1,1}$	$\nu_{1,2}$	$\dots$	$\nu_{1,s}$
Y	$\nu_{2,1}$	$\nu_{2,2}$	$\dots$	$\nu_{2,s}$
	$\nu_{\cdot,1}$	$\nu_{\cdot,2}$	$\dots$	$\nu_{\cdot,s}$

where $\nu_{\cdot,i} = \nu_{1,i} + \nu_{2,i}$.

And we construct estimates of the probabilities as

$$\hat{p}_i = \frac{\nu_{\cdot,i}}{n_1 + n_2}.$$

Then, if the hypothesis H_0 is true, the following statement holds:

$$\sum_{\substack{1\le i\le s\\ 1\le j\le 2}} \frac{\left(\nu_{i,j} - n_i\hat{p}_j\right)^2}{n_i\hat{p}_j} \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2_{s-1}.$$

Remark. An informal proof can be described as follows: inside the obtained table, each of the rows contains $s-1$ independent parameters, and since we give an estimate for $s-1$ parameters, as a result only $2(s-1) - (s-1) = s-1$ random parameters remain.

Statement Let a joint sample $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$ be given.
And suppose we test the hypothesis of independence

$$H_0 : \ (\exists\, P_1, P_2 : \ P(x, y) = P_1(x) \cdot P_2(y)).$$

We count, for each outcome, the number of corresponding events:

$X \backslash Y$	y_1	y_2	$\ldots$	y_r	
x_1	$\nu_{1,1}$	$\nu_{1,2}$	$\ldots$	$\nu_{1,r}$	$\nu_{\cdot,1}$
x_2	$\nu_{2,1}$	$\nu_{2,2}$	$\ldots$	$\nu_{2,r}$	$\nu_{\cdot,2}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_s	$\nu_{s,1}$	$\nu_{s,2}$	$\ldots$	$\nu_{s,r}$	$\nu_{\cdot,r}$
	$\nu_{1,\cdot}$	$\nu_{2,\cdot}$	$\ldots$	$\nu_{s,\cdot}$	

where

$$\nu_{i,\cdot} = \sum_{j=1}^{r} \nu_{i,j}, \qquad \nu_{\cdot,j} = \sum_{i=1}^{s} \nu_{i,j}.$$

And based on the obtained table we construct estimates for the probabilities as

$$\hat{p}_{i,j} = \frac{\nu_{i,\cdot}\, \nu_{\cdot,j}}{n^2}.$$

Then, if the hypothesis H_0 is true, the following statement holds:

$$\sum_{\substack{1 \le i \le s \\ 1 \le j \le r}} \frac{(\nu_{i,j} - n\hat{p}_{i,j})^2}{n\hat{p}_{i,j}} \ \xrightarrow[n\to\infty]{\mathcal{D}} \ \chi^2_{(s-1)(r-1)}.$$

Proof. Without proof.
Remark. We again give an informal proof of this statement: initially the table contains $rs - 1$ independent parameters, while we give estimates for $(r-1)+(s-1)$ parameters (namely, for $\nu_{i,\cdot}$ and $\nu_{\cdot,j}$, on the basis of which $\hat{p}_{i,j}$ were computed). Hence we obtain that the number of degrees of freedom of the chi-square distribution is exactly $rs - 1 - (r-1) - (s-1) = (r-1)(s-1)$.

Ranks. The Most Powerful Rank-Based Tests.

Assume that our data are continuous but structured in such a way that we can only compare observations with one another and determine which one is "smaller". Moreover, we assume that this relation is transitive.

Such data are called *ranked data*, and these are precisely the type of data we will discuss now. However, before doing so, we must first become a bit more familiar with order statistics.

Reminder.
Let $X_1, \ldots, X_n$ be some observations possessing the property described above. Denote by $X_{(i)}$ the i-th order statistic, that is, the value of the i-th smallest observation among $X_1, \ldots, X_n$.

Notation.
Let $X_1, \ldots, X_n$ be observations possessing the property described above. Introduce the following notation:

$$X := (X_1, \ldots, X_n) \quad \text{— the vector of observations,}$$

$$X_{(\cdot)} := (X_{(1)}, \ldots, X_{(n)}) \quad \text{— the vector of order statistics (the variational series).}$$

Definition.

Let $X_1, \dots, X_n \sim F$ be continuously distributed observations. We require continuity to ensure that the probability of ties is zero.

Define the rank R_i as the position of the element X_i in the variational series $X_{(\cdot)}$, i.e. $X_i = X_{(R_i)}$.

$$R := (R_1, \dots, R_n) \quad \text{— the vector of ranks.}$$

Remark.

From the definition it is clear that the vector R is a permutation of the numbers $\{1, \dots, n\}$, i.e. $R \in S_n$.

Theorem.

Let f be the joint density of the random vector X. Then the variational series $X_{(\cdot)}$ has a distribution with density

$$f\big(x_{(\cdot)}\big) := \sum_{r \in S_n} f\big(x_{(r_1)}, \dots, x_{(r_n)}\big).$$

Moreover,

$$\mathbb{P}(R = r \mid X_{(\cdot)} = x_{(\cdot)}) = \frac{f\big(x_{(r_1)}, \dots, x_{(r_n)}\big)}{f\big(x_{(\cdot)}\big)}.$$

When working with real data, however, we usually assume that our observations are, ideally, independent and identically distributed. As we will now see, in this case the distribution of the rank vector and the variational series takes a rather simple and pleasant form.

Theorem.

Let $X_1, \dots, X_n$ be independent and identically distributed random variables with joint density q. Then the random vectors R and $X_{(\cdot)}$ are independent, and moreover,

$$\mathbb{P}(R = r) = \frac{1}{n!}, \qquad q\big(x_{(\cdot)}\big) = n!\, q(x_{(\cdot)}).$$

Definition.

Rank tests are tests whose critical region depends only on the vector of ranks R.

Theorem.

Let a sample $X_1, \dots, X_n$ be given such that $X \sim P$. Assume we are testing the following simple hypotheses:

$$H_0 : P = P_0, \qquad H_1 : P = P_1.$$

If the distribution P_0 corresponds to the joint distribution of independent and identically distributed random variables, then the most powerful rank test has the following structure:

$$\varphi(X) = \begin{cases} 1, & P_1(R = r) \geq k, \\ 0, & P_1(R = r) < k, \end{cases}$$

where r is the vector of ranks of the elements of X.

Proof.

We use the Neyman–Pearson lemma. By the definition of rank tests, the critical region depends only on the vector of ranks. Therefore, we are interested only in the distribution of the rank vector.

But if the observations are independent, then

$$P_0(R = r) = \frac{1}{n!},$$

that is, this quantity is a constant.

Definition.

Let us be given a family of probability distributions $\{P_\delta\}$, parametrized by some parameter δ. Assume that the joint distribution of the sample $X_1, \ldots, X_n$ belongs to this family. We consider the hypotheses:

$$H_0 : \delta = 0, \qquad H_1 : \delta > 0.$$

A test is called *locally most powerful* if there exists $\varepsilon > 0$ such that for every $\delta_0 \in (0, \varepsilon)$ the test φ is the most powerful test against the alternative $H_1 : \delta = \delta_0$.

Remark.

From now on, we will assume that the densities of the sample elements belong to the family of densities $\{f(x, \theta)\}$, and that the joint density of the distribution has the following form:

$$q_\delta(X) = \prod_{i=1}^{n} f(X_i,\, \delta\, c_i), \qquad \delta > 0.$$

Furthermore, in order to define the subsequent tests, we will require that the densities satisfy certain regularity conditions:

Regularity conditions.

1. $f(x, \theta)$ is absolutely continuous in θ (i.e. $f(x, \theta) \in C^1$);
2. $f'(x, 0)$ exists, where
$$f'(x, 0) = \lim_{\theta \to 0} \frac{f(x, \theta) - f(x, 0)}{\theta};$$
3. $$\lim_{\theta \to 0} \int_{-\infty}^{\infty} |f'(x, \theta)|\, dx = \int_{-\infty}^{\infty} |f'(x, 0)|\, dx < \infty.$$

We now proceed to the main hypotheses and the tests used to verify them.

Rank tests. The randomness hypothesis

We begin with the definition of the randomness hypothesis.

Definition 6.10. Let $X_1, \ldots, X_n$ be an arbitrary sample with joint density q. The hypothesis

$$H_0 : q(X_1, \ldots, X_n) = \prod_{i=1}^{n} f(X_i)$$

is called the *randomness hypothesis*. In other words, this hypothesis simply asserts that all observations are independent and identically distributed.

We already know the distributions of R and $X^{(\cdot)}$ in the case when the randomness hypothesis holds. We now derive the locally most powerful rank test for verifying this hypothesis, and then use it to construct other tests.

Notation. We call "scores" the functions

$$a_n(i, f) = E_0 \left[\frac{f'(X_{(i)}, 0)}{f(X_{(i)}, 0)} \right],$$

where n is the sample size, and E_0 denotes expectation with respect to the density $f(x, 0)$.

Theorem 6.13. Assume the regularity conditions hold. Then the locally most powerful rank test for the randomness hypothesis has a critical region of the form

$$\sum_{i=1}^{n} c_i\, a_n(R_i, f) \geq k.$$

Proof. We expand the probability of observing a specific rank vector r under P_δ:

$$P_\delta(R=r) = \int_{R=r} q_\delta(x_1,\dots,x_n)\,dx_1\cdots dx_n = \int_{R=r} \prod_{i=1}^n f(x_i,\delta c_i)\,dx_1\cdots dx_n.$$

$$= \int_{R=r} \prod_{i=1}^n f(x_i,0)\,dx_1\cdots dx_n + \delta \int_{R=r} \frac{1}{\delta}\left(\prod_{i=1}^n f(x_i,\delta c_i) - \prod_{i=1}^n f(x_i,0)\right) dx_1\cdots dx_n.$$

The first integral is the probability of a given rank vector under H_0, and we already know it equals

$$\frac{1}{n!}.$$

Rewrite the second integral using the identity

$$\prod_{i=1}^n a_i - \prod_{i=1}^n b_i = \sum_{i=1}^n (a_i - b_i)\left(\prod_{j=1}^{i-1} a_j\right)\left(\prod_{j=i+1}^n b_j\right).$$

Hence,

$$P_\delta(R=r) = \frac{1}{n!} + \delta \int_{R=r} \sum_{i=1}^n \left(c_i\,\frac{f(x_i,\delta c_i) - f(x_i,0)}{\delta c_i} \prod_{j=1}^{i-1} f(x_j,\delta c_j) \prod_{j=i+1}^n f(x_j,0)\right) dx_1\cdots dx_n.$$

Using the regularity conditions and taking the limit as $\delta \to 0$, we obtain

$$= \frac{1}{n!} + \delta \sum_{i=1}^n \int_{R=r} \left(c_i f'(x_i,0) \prod_{j\neq i} f(x_j,0)\right) dx_1\cdots dx_n + o(\delta)$$

$$= \frac{1}{n!} + \delta \sum_{i=1}^n c_i E_0\left[\frac{f'(x_{(r_i)},0)}{f(x_{(r_i)},0)} \;\middle|\; R=r\right] + o(\delta).$$

Since $x_i = x_{(r_i)}$, this becomes

$$P_\delta(R=r) = \frac{1}{n!} + \delta \sum_{i=1}^n c_i\, a_n(r_i,f) + o(\delta).$$

For fixed $\delta > 0$, the critical region has the form

$$P_\delta(R=r) \geq \tilde{k}.$$

Thus one can choose k such that for δ close to 0,

$$\frac{1}{n!} + \delta \sum_{i=1}^n c_i\, a_n(r_i,f) + o(\delta) \geq \tilde{k} \quad \Longrightarrow \quad \sum_{i=1}^n c_i\, a_n(r_i,f) \geq k.$$

This gives the general form of the locally most powerful test for the randomness hypothesis.

Now we use it to construct several tests for checking this hypothesis against a shift alternative. Assume we have two samples $X_1,\dots,X_m$ and $X_{m+1},\dots,X_{m+n}$, and we test

$$H_0: \quad q_0(X) = \prod_{i=1}^{m+n} f(X_i), \qquad H_1: \quad q_\delta(X) = \prod_{i=1}^m f(X_i-\delta) \prod_{i=1}^n f(X_{m+i}).$$

1. The normal scores test. Suppose that the samples come from the normal distribution,

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad f(x,\theta) := f(x+\theta).$$

We compute the scores:

$$\frac{f'(x,0)}{f(x,0)} = (\ln f(x,0))' = \left(-\frac{x^2}{2}\right)' = -x.$$

Thus

$$a_N(i,f) = E_0\left[\frac{f'(X_{(i)},0)}{f(X_{(i)},0)}\right] = E_0(-X_{(i)}) = -E\big(\Phi^{-1}(U_{(i)})\big).$$

Since $c_1,\dots,c_m = -1$ and $c_{m+1},\dots,c_{m+n} = 0$, the critical region becomes

$$\sum_{i=1}^{m} -a_{m+n}(R_i,f) = \sum_{i=1}^{m} E\big(\Phi^{-1}(U_{(R_i)})\big) \ge k.$$

2. Van der Waerden test

The main problem of the normal scores test is the complexity of computing $a_n(i,f)$. However, note that, since $E[\Phi^{-1}(U_{(i)})] \approx \Phi^{-1}(E[U_{(i)}])$, it can be rewritten in a more convenient form:

$$\sum_{i=1}^{m} a_{n+m}(R_i,f) = \sum_{i=1}^{m} \Phi^{-1}\Big(\frac{R_i}{n+m+1}\Big) \ge k.$$

3. Wilcoxon test

Now consider the case of the logistic distribution with density

$$f(x) = \frac{e^{-x}}{(1+e^{-x})^2}.$$

Compute the scores in this case:

$$\frac{f'(x,0)}{f(x,0)} = (\ln f(x,0))' = \big(-x - 2\ln(1+e^{-x})\big)' = -1 + \frac{2e^{-x}}{1+e^{-x}} = \frac{e^{-x}-1}{e^{-x}+1}.$$

The cumulative distribution function is

$$F(x) = \frac{1}{1+e^{-x}}, \qquad F^{-1}(y) = -\ln\Big(\frac{1}{y} - 1\Big) = \ln\frac{y}{1-y}.$$

Hence,

$$\frac{f'}{f}\big(F^{-1}(y)\big) = \frac{e^{-x}-1}{e^{-x}+1} \circ \ln\frac{y}{1-y} = \frac{1-y}{y} - \frac{y}{1-y} = 1-2y.$$

Thus,

$$a_N(i,f) = E_0\Big[\frac{f'}{f}(X_{(i)},0)\Big] = E\Big[\frac{f'}{f}(F^{-1}(U_{(i)}),0)\Big] = E[1-2U_{(i)}] = 1 - \frac{2i}{N+1}.$$

As before, some c_i are -1 and some are 0. The critical region of the test can now be written as:

$$\sum_{i=1}^{m} R_i \ge k.$$

Rank tests. Symmetry hypothesis

Now we turn to the symmetry hypothesis.

Definition Let $X_1, \ldots, X_n$ be an arbitrary sample with joint density q. The hypothesis

$$H_0 : q(X_1, \ldots, X_n) = \prod_{i=1}^{n} f(X_i), \quad f(x) = f(-x)$$

is called the *symmetry hypothesis*. In essence, this is the same as the randomness hypothesis, but additionally it tests whether the distribution is symmetric about zero.

Notation. Let $X = (X_1, \ldots, X_n)$ be the original sample. Then define:

$$|X| := (|X_1|, \ldots, |X_n|) \quad \text{— the vector of absolute values,}$$

$$R^+ := R(|X|) \quad \text{— the vector of ranks of the absolute values,}$$

$$\operatorname{sign}(X) := (\operatorname{sign}(X_1), \ldots, \operatorname{sign}(X_n)) \quad \text{— the vector of signs of the original values.}$$

Using the introduced notation, we can attempt to understand the distributions of the corresponding vectors. For this, we state the following theorem:

Theorem. Let $X_1, \ldots, X_n$ be an arbitrary sample with joint density q. If the symmetry hypothesis holds, then the vectors $\operatorname{sign}(X)$, R^+, and $|X|_{(\cdot)}$ are independent, with

$$P(\operatorname{sign}(X) = s) = \frac{1}{2^n}, \quad P(R^+ = r) = \frac{1}{n!}, \quad q(|X|_{(\cdot)}) = n!\, 2^n \prod_{i=1}^{n} f(X_i).$$

Proof. Independence follows from the fact that the original vector X can be uniquely recovered from this triple of vectors, and vice versa.

Next, consider the probability of observing a specific sign vector. Since the distribution is symmetric about zero, each observation is equally likely to be positive or negative, and because the observations are independent, the probability of a vector is the product of probabilities for each coordinate.

As for the distributions of R^+ and $|X|_{(\cdot)}$, the proof was given above.

Symmetry vs shift alternative. Suppose we want to test the symmetry hypothesis against a shift alternative, i.e., for a sample $X_1, \ldots, X_n$ we test

$$H_0 : q_0(X) = \prod_{i=1}^{n} f(X_i), \qquad H_1 : q_\delta(X) = \prod_{i=1}^{n} f(X_i - \delta).$$

Analogous to the randomness hypothesis, we obtain the locally most powerful rank test for testing against the shift alternative.

Notation. Introduce an alternative definition of scores to obtain the form of the locally most powerful test:

$$a_n^+(i, f) := E_0 \left[\frac{f'(|X|_{(i)})}{f(|X|_{(i)})} \right].$$

Theorem. The locally most powerful rank test for testing the symmetry hypothesis against a shift alternative has a critical region of the form

$$\sum_{i=1}^{n} a_n^+(R_i^+, f) \operatorname{sign}(X_i) \geq k.$$

Proof. Left as an exercise; it follows exactly the proof for the randomness hypothesis.

Remark. The critical region can be rewritten in a more convenient form:

$$\sum_{X_i>0} a_n^+(R_i^+, f) = \frac{1}{2}\left(\sum_{i=1}^n a_n^+(R_i^+, f)\operatorname{sign}(X_i) + \sum_{i=1}^n a_n^+(i, f)\right) \geq \frac{1}{2}(k + C) = \tilde{k},$$

so that the critical region of the locally most powerful test can be written as

$$\sum_{X_i>0} a_n^+(R_i^+, f) \geq k.$$

1. One-sample Wilcoxon test. Let $X_1, \ldots, X_n \sim f$, where $f(x)$ is the density of the logistic distribution. Then the critical region can be written as

$$\sum_{X_i>0} R_i \geq k,$$

and is proved similarly to the Wilcoxon test for the randomness hypothesis.

2. Sign test. Let $X_1, \ldots, X_n \sim f$, where

$$f(x) = \frac{1}{2}e^{-|x|}.$$

Compute the scores:

$$\frac{f'}{f}(x) = \operatorname{sign}(x) \implies a_n^+(R_i, f) = 1,$$

so that the critical region becomes

$$\sum_{X_i>0} 1 = \#\{X_i > 0\} \geq k.$$

Rank tests. Independence hypothesis

Consider the setup for testing the independence hypothesis. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a joint sample with $X_i \sim f$ and $Y_i \sim g$. Let R and Q be the ranks of X and Y, respectively.

We test the hypotheses:

$$H_0 : q_0(X, Y) = \prod_{i=1}^n f(X_i)g(Y_i), \qquad H_1 : q_\delta(X, Y) = \prod_{i=1}^n \left[\int_{-1}^1 f(X_i - \delta z)g(Y_i - \delta z)\, dH(z)\right],$$

where $H(z)$ is the distribution function of some random variable Z. In essence, the alternative hypothesis states that the sample elements are generated as

$$X_i = X_i' + \delta Z_i, \qquad Y_i = Y_i' + \delta Z_i.$$

Theorem 6.16. The locally most powerful rank test for testing independence has a critical region of the form

$$\sum_{i=1}^n a_n(R_i, f)\, a_n(Q_i, g) \geq k.$$

Proof. Omitted.

Specific tests for independence:

1. Spearman correlation coefficient. Let f and g be logistic densities. Then the critical region can be written as

$$\sum_{i=1}^n R_i Q_i \geq k.$$

This is proved similarly as before. Note that the Spearman correlation coefficient is defined as

$$\rho = \frac{12}{n^3} \sum_{i=1}^{n} (R_i - ER_i)(Q_i - EQ_i) = \frac{12}{n^3} \sum_{i=1}^{n} \left(R_i - \frac{n+1}{2} \right) \left(Q_i - \frac{n+1}{2} \right).$$

This coefficient reflects the correlation between ranks. Expanding the brackets shows that the critical region above is equivalent to $\rho \geq \tilde{k}$ for some $\tilde{k}$.