Basic of Statistics

Margus Pihlak

2021
October–November

Chapter 1

Assumptions of classical Statistics

Definition 1.0.1. The set about which will make decision is called as population.

Definition 1.0.2. The set on base of which will make deciosion is called as sample.

Let us denote population as U and sample as u. From Definition 1.0.2 follows that $u \subseteq U$. That means sample is a subset of population.

Definition 1.0.3. Number of elements in sample u is called as sample size.

Let us denote sample size as n.

Statistics can call as mathematical if $u \subset U$. Mathematical Statistics apply tools of probability theory.

Statistical analysis can devide into 4 parts.

1) Experimental design. To formulate problems which we are interested in,

2) To collect data. That means we have to compose the sample which characterizes which best characterizes the population,

3) To compose the statistical model. Estimators obtained from the sample are found on base of which we have to make decision about all population,

4) To interpret results of statistical analysis. The problems about economy, engineering or life sciences.

Assumptions of classical statistics have formulated British statistician and biologist Roland Aylmer Fisher (1890–1962).

1° Let population U be infinite and sample be its finite subset u, that means sample size $|u| = n < \infty$.

2° Let us have a sample $\mathbf{X} = (X_1, X_2, ..., X_n)^\intercal$. And let us assume that elements of sample are independent, that means selection is with replacement. Every element has equal probability to be selected $\frac{1}{n}$.

3° Parametrical assumption. Let us assume that $X_i \sim F(\Theta)$, $i = 1, 2, ..., n$, where distribution F describes population U. Vector

$$\Theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{pmatrix}$$

is called as parameter vector of this distribution. The purpose is to estimate parameters θ_j, $j = 1, 2, ..., k$.

Let us present vector Θ for some distributions.

Examnple 1.0.1. Let X_i has normal distribution which is denoted as $X_i \sim \mathcal{N}(\mu, \sigma)$. Its density function

$$f(x_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{(x_i - \mu)^2}{2\sigma^2} \right), \quad \mu > 0.$$

Thus parameter vector of normal distribution

$$\Theta = \begin{pmatrix} \mu \\ \sigma \end{pmatrix}.$$

Examnple 1.0.2. Let X_i has binomila distribution which is denoted as $X_i \sim B(n, p)$. Its probability function

$$P(X_i = k) = C_n^k p^k (1 - p)^{n-k}, \quad p \in (0; 1).$$

The parameter vector of binomial

$$\Theta = \begin{pmatrix} n \\ p \end{pmatrix}.$$

Examnple 1.0.3. let X_i has Poisson distribution which is denoted as $X_i \sim Po(\lambda)$. Its probability function

$$P(X_i = k) = \frac{\lambda^k}{k!} \exp(-\lambda), \quad \lambda > 0.$$

The parameter vector of Poisson distribution $\Theta = \lambda$.

Examnple 1.0.4. Let X_i has exponential distribution which is denoted as $X_i \sim \mathcal{E}(\nu)$. Its density function

$$f(x) = \begin{cases} 0, \text{ if } x < 0, \\ \nu \exp(-\nu x), \text{ if } x \geq 0 \end{cases}, \quad \nu > 0.$$

The parameter vector of exponential distribution $\Theta = \nu$.

Examnple 1.0.5. Let X_i has geometric distribution which is denoted as $X_i \sim Geo(p)$. Its probability function

$$P(X_i = k) = p(1-p)^{k-1}, \quad p \in (0;1)$$

and parameter vector $\Theta = p$.

Chapter 2

Statistics as Estimators

Let us study how to find estimators of parameters.

Let us have a sample $\mathbf{X} = (X_1, X_2, ..., X_n)^\intercal$ where $X_i \sim F$, $i = 1, 2, ..., n$. Let this be presented with random variable X which is a copy which have the same distribution with every elemt of sample of sample $\mathbf{X}$. Let us characterize a sample by means of representative X. Let $\mathbf{x} = (x_1, x_2, ..., x_n)^\intercal$ be a realization of sample $\mathbf{X}$.

Definition 2.0.1. The Statistic is random variable composed on the base of sample

By means of Statistic will be estimated parameter θ_j, $j = 1, 2, ..., k$.

Let $\widehat{\theta_j}$ be estimator of parameter θ_j. Let us denote this Statistic as $\widehat{\theta_j} = T(\mathbf{X})$. The realization of this Statistic $T(\mathbf{x})$ can be a rational number, interval or set. Our in discrete case problem can be formulated as follows

$$\alpha = P(T(\mathbf{X}) = \theta_j).$$

That means with which probability α takes the parameter value θ_j. In continuous case we are interested in probability

$$\alpha = P(T(\mathbf{X}) \in \mathcal{H}_\alpha)$$

where $\mathcal{H}_\alpha$ denotes a set where value of parameter θ_j belongs with probability α. This can denote also as follws:

$$\alpha = P(\widehat{\theta_j} = \theta_j) \text{ või } \alpha = P(\widehat{\theta_j} \in \mathcal{H}_\alpha).$$

2.1 Point estimator

Point estimator is the simplest estimator. This estimator characterizes the best replacement of parameter θ_j.

Point estimator can said to be very good if this has 2 properties:

1) Unbiased

Definition 2.1.1. Point estimator $\widehat{\theta}_j = T(\mathbf{X})$ is said to be unbiased if

$$E(\widehat{\theta}_j) = \theta_j$$

on every sample size n.

2) Consistent

Definition 2.1.2. Point estimator $\widehat{\theta}_j = T(\mathbf{X})$ is said to be consistent if

$$\lim_{n\to\infty} V(\widehat{\theta}_j) = 0.$$

The error of estimator can be divided into two parts: systematic (or directed) error and random error.

Definition 2.1.3. The quantity

$$b = E(\widehat{\theta}_j) - \theta_j$$

is called as systematic error or biase.

Definition 2.1.4. The quantity

$$s_{error} = \frac{\sigma_{X_i}}{\sqrt{n}}$$

is called as standard error.

Standard error characterized random error of estimator. Let us define mean square error of parameter θ_j estimator $\widehat{\theta}_j$.

Definition 2.1.5. Mean square error of point estimator $\widehat{\theta}_j$ is defined as

$$\mathrm{MSE}(\widehat{\theta}_j) = E(\widehat{\theta}_j - \theta_j)^2.$$

We can express MSE as follows:

$$\mathrm{MSE}(\widehat{\theta}_j) = E(\widehat{\theta}_j - E(\widehat{\theta}_j) + E(\widehat{\theta}_j) - \theta_j)^2 =$$

$$= E\{(\widehat{\theta}_j - E(\widehat{\theta}_j))^2 + 2(\widehat{\theta}_j - E(\widehat{\theta}_j))(E(\widehat{\theta}_j) - \theta_j) + (E(\widehat{\theta}_j) - \theta_j)^2\}$$

$$= E\{\widehat{\theta}_j - E(\widehat{\theta}_j)\}^2 + \{\theta_j - E(\widehat{\theta}_j)\}^2 = V(\widehat{\theta}_j) + b^2$$

because

$$2E\{(\widehat{\theta}_j - E(\widehat{\theta}_j))(E(\widehat{\theta}_j) - \theta_j)\} = 0.$$

Thus MSE consist of

1) random error component $V(\widehat{\theta}_j)$

and

2) systematic error component b^2.

Let us give a classical example about estimator which is unbiased and consistent. Let us have a sample where $X_i \sim \mathcal{N}(\mu, \sigma)$, $i = 1, 2, ..., n$. Our purpose is to constract the Statistic $T(\mathbf{X})$ to estimate parameters μ and σ.

1) Let

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2.1}$$

be point estimator for expectation μ. Let us ensure that this estimator is unbiased. Applying properties of expectation we get

$$E(\overline{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) =$$

$$= \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}n\mu = \mu.$$

We have gotten $E(\overline{x}) = \mu$ which says that estimator (2.1) is unbiased.

Let us ensure that estimator (2.1) is also consistent. Applying propositions of variance we get

$$V(\overline{x}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}V\left(\sum_{i=1}^{n} X_i\right) =$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

WE have gotten that $V(\overline{x}) = \dfrac{\sigma^2}{n}$. Thus

$$MSE(\overline{x}) = \frac{\sigma^2}{n}.$$

From limit value

$$\lim_{n\to\infty} \frac{\sigma^2}{n} = 0$$

concludes that estimator (2.1) is consistent.

Conclusion 2.1.1. Let random variable $X_i \sim \mathcal{N}(\mu, \sigma)$, $i = 1, 2, ..., n$. Then random variable $\overline{x} \sim \mathcal{N}\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$.

2) Let us study point estimator of variance σ^2 which is defined as

$$\overline{s^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{x})^2. \tag{2.2}$$

But estimator (2.2) is with biase. Because

$$E(\overline{s^2}) = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{x})^2\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i - \mu + \mu - \overline{x})^2 =$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} E((X_i - \mu)^2 - 2E(\overline{x} - \mu)(X_i - \mu) + E(\overline{x} - \mu)^2\right) =$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} E(X_i - \mu)^2 - 2E(\overline{x} - \mu)\sum_{i=1}^{n}(X_i - \mu) + nE(\overline{x} - \mu)^2\right) =$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} E(X_i - \mu)^2 - nE(\overline{x} - \mu)^2\right).$$

From definition of variance follows $E(X_i - \mu)^2 = \sigma^2$. From Conclusion 2.1.1 we get that $E(\overline{x} - \mu)^2 = \dfrac{\sigma^2}{n}$. Thus

$$E(\overline{s^2}) = \frac{1}{n}(n\sigma^2 - \sigma^2) = \sigma^2 - \frac{\sigma^2}{n},$$

which says that point estimator (2.2) has biase

$$b = -\frac{\sigma^2}{n}.$$

How to remove this biase?

Proposistion 2.1.1. Let random variable $X_i \sim \mathcal{N}(\mu, \sigma)$, $i = 1, 2, ..., n$. Then Statistic

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{x})^2 \tag{2.3}$$

is unbiased point estimator for variance σ^2.

2.1.1 Maximum Likelihood Estimation (MLE)

MLE consists of all distribution information. Likelihood is characterized by function L which is called as likelihood function.

In continuous case

$$L(\Theta, \mathbf{x}) = \prod_{i=1}^{n} f(\Theta, x_i) \tag{2.4}$$

and in discrete case

$$L(\Theta, \mathbf{x}) = \prod_{i=1}^{n} P(X = x_i). \tag{2.5}$$

The purpose is to maximize function $L(\Theta, \mathbf{x})$. That means we have to solve the system of equations

$$\begin{aligned} \frac{\partial}{\partial \theta_1} L(\theta_1, \theta_2, ..., \theta_k, \mathbf{x}) &= 0 \\ \frac{\partial}{\partial \theta_2} L(\theta_1, \theta_2, ..., \theta_k, \mathbf{x}) &= 0 \\ &\vdots \\ \frac{\partial}{\partial \theta_k} L(\theta_1, \theta_2, ..., \theta_k, \mathbf{x}) &= 0. \end{aligned} \tag{2.6}$$

The solution of this system is calls as MLE $\Theta^* = (\theta_1^*, \theta_2^*, ..., \theta_k^*)^\intercal$. That means we have found such values of parameters which maximize likelihood function L value in the case of realization $\mathbf{x}$.

AS we see we have to find derivatives of products. But that is technically veri difficult. Much easier is to find derivatives of sums. Because of this will be performed logarithmization

$$l(\Theta, \mathbf{x}) = \ln(L(\Theta, \mathbf{x})) = \ln\left(\prod_{i=1}^{n} f(\Theta, x_i)\right) = \sum_{i=1}^{n} \ln(f(\Theta, x_i)).$$

Frequently is solving of system (2.6) very complex. Often we can meet situations when analytical solutions don't exist. Let us give examples where finding MLE is simple.

Examnple 2.1.1. Let $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $i = 1, 2, ..., n$. From probability theory is known that $E(X_i) = \mu$ and $V(X_i) = \sigma^2$. The likelohood function of normal distribution

$$L(\Theta, \mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(x_i - \mu)^2}{2\sigma^2}\right).$$

With logarithmization we get

$$l(\Theta, \mathbf{x}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Let us find MLE for μ. Partial derivative by μ

$$\frac{\partial l(\mu, \sigma^2, \mathbf{x})}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu).$$

Equating right hand side to zero

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0$$

we get that

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

Thus MLE for parameter μ is arithmetical mean $\overline{x}$.

Let us find partial derivative by parameter σ^2 replacing MLE of μ. We get

$$\frac{\partial l(\overline{x}, \sigma^2, \mathbf{x})}{\partial \sigma^2} = \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \overline{x})^2 - \frac{n}{2\sigma^2} = 0.$$

After some transformation we get

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Thus MLE for σ^2 is estimator s_*^2. But this estimator is with biase. In the case of normal distribution MLE $\Theta^* = (\overline{x}, s_*^2)^\intercal$.

Examnple 2.1.2. Let us have a sample where X_i, $i = 1, 2, ..., n$, has Poisson distribution with parameter λ. From probability theory we know that in this case $E(X_i) = D(X_i) = \lambda$. Likelihood function

$$l(\lambda, \mathbf{x}) = \ln(\lambda) \sum_{i=1}^{n} x_i - \ln(x_i!) - n\lambda.$$

Equation partial derivative of l by λ to zero we get

$$\frac{\partial l(\lambda, \mathbf{x})}{\partial \lambda} = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0$$

from which follows that

$$\lambda = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Thus MLE for λ is arithmetical mean $\overline{x}$.

Examnple 2.1.3. Let us have a sample where X_i, $i = 1, 2, ..., n$, has exponential distribution with parameter $\nu > 0$. In this case expectation and standard deviation equal with standardhälve suurusega $\frac{1}{\nu}$. Likelihood function

$$l(\nu, \mathbf{x}) = n \ln(\nu) - \nu \sum_{i=1}^{n} x_i.$$

Equation partial derivative of l by ν to zero we get

$$\frac{\partial l(\nu, \mathbf{x})}{\partial \nu} = \frac{n}{\nu} - \sum_{i=1}^{n} x_i = 0,$$

from which follows that

$$\nu = \frac{n}{\sum_{i=1}^{n} x_i}.$$

Thus MLE for parameter ν is inverse value of arithmetical mean $\overline{x}$.

Examnple 2.1.4. Let us have a sample where X_i, $i = 1, 2, ..., n$, has Reileigh distribution with parameter $h > 0$. Then density function

$$f(x) = \begin{cases} 0, \text{ if } x < 0, \\ 2h^2 x \exp(-h^2 x^2), \text{ if } x \geq 0. \end{cases}$$

Rayleigh' distribution is widely applied in life sciences and technics, for example on tomography. Also distributions of incomes are frequently similar

with Reileigh' distributione. Expeaction and variance of this distribution are as

$$E(X) = \frac{\sqrt{\pi}}{2h} \text{ and } D(X) = \frac{4-\pi}{4h^2}.$$

Likelihood function

$$l(h, \mathbf{x}) = n\ln(2) + 2n\ln(h) + \sum_{i=1}^{n} \ln(x_i) - h^2 \sum_{i=1}^{n} x_i^2.$$

Equation partial derivative of l by h to zero we get

$$\frac{\partial l(h, \mathbf{x})}{\partial h} = \frac{2n}{h} - 2h \sum_{i=1}^{n} x_i^2 = 0,$$

from which follows that

$$h = \frac{1}{\sqrt{\overline{x^2}}},$$

where

$$\overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2}{n}.$$

2.2 Interval estimation

Interval estimation takes into account expectation and standard deviation. This kind of estimation is mostly used in practice. Our purpose is construct α-confidence interval to parameter θ_j.

Definition 2.2.1. The set I_α is aclled as α-usaldusintervalliks for parameter θ_j if

$$P(\theta_j \in I_\alpha) = 1 - \alpha$$

That means the set I_α covers parameter θ_j with probability $1 - \alpha$.

Probability $1 - \alpha$ in Definition 2.2.1 is called as confidence level. Let $I_\alpha = [u, U]$. Then quantity u is called as lower confidence bound and quantity U as upper confidence bound. From Definition Definitsioonist 2.2.1 follows that

$$P(\theta_j < u) = P(\theta_j > U) = \frac{\alpha}{2}.$$

If distribution of Statistic is symmetric then $u = -U$.

Interval estimation is based on α-quantile.

Definition 2.2.2. Value x_α is called as α-quantile of random variable X if

$$P(X \leq x_\alpha) = \alpha.$$

Function $F^{-1}(\alpha)$ is called as quantile function of random variable X. Quantile function is inverse function of distribution function. Thus

$$x_\alpha = F^{-1}(\alpha).$$

Definition 2.2.3. Value $\overline{x}_\alpha$ is called as supplement quantile of random variable X if

$$P(X > \overline{x}_\alpha) = \alpha.$$

Thus α-kvantiil tema $1-\alpha$-supplement quantile equals with $x_\alpha = \overline{x}_{1-\alpha}$. Figure 2.1 shows quantile and supplement quantile of Rayleigh' distribution.

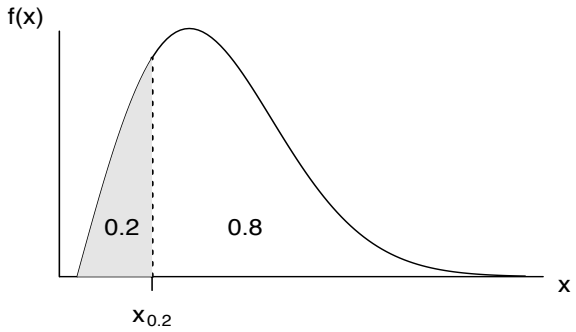

Figure 2.1. juhusliku suuruse 0.2-quantile and 0.8-supplement quantile of Rayleigh' distribution ($h = 0.3$)

Let us construct confidence interval which is based on Central Limit Theorem (CLT). Let $X_1, X_2, ..., X_n$ are i.i.d. let $E(X_i) = \mu$ and $V(X_i) = \sigma^2$. Let

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then random variable

$$Z = \frac{\overline{x} - \mu}{\sigma}\sqrt{n}$$

has approximately normal distribution $\mathcal{N}(0, 1)$.

Let us find α-confidence interval for expecatation μ applying CLT.Our purpose is find ϵ_α when

$$P(\mu \in [\overline{x} - \epsilon_\alpha; \overline{x} + \epsilon_\alpha]) = 1 - \alpha.$$

We get

$$P(\overline{x} - \epsilon_\alpha \leq \mu \leq \overline{x} + \epsilon_\alpha) = 1 - \alpha \Leftrightarrow P(-\epsilon_\alpha \leq \overline{x} - \mu \leq \epsilon_\alpha) = 1 - \alpha \Leftrightarrow$$

$$\Leftrightarrow P\left(-\frac{\epsilon_\alpha}{\sigma}\sqrt{n} \leq \frac{\overline{x} - \mu}{\sigma}\sqrt{n} \leq \frac{\epsilon_\alpha}{\sigma}\sqrt{n}\right) = 1-\alpha = F\left(\frac{\epsilon_\alpha}{\sigma}\sqrt{n}\right) - F\left(-\frac{\epsilon_\alpha}{\sigma}\sqrt{n}\right)$$

where F denotes distribution function of

$$Z = \frac{\overline{x} - \mu}{\sigma}\sqrt{n}.$$

According to CLT

$$F\left(\frac{\epsilon_\alpha}{\sigma}\sqrt{n}\right) - F\left(-\frac{\epsilon_\alpha}{\sigma}\sqrt{n}\right) \approx 2\Phi\left(\frac{\epsilon_\alpha}{\sigma}\sqrt{n}\right) = 1 - \alpha.$$

We get that

$$\epsilon_\alpha \approx \Phi^{-1}\left(\frac{1-\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}.$$

If $\alpha = 0.05$ then $\Phi^{-1}\left(\frac{1-\alpha}{2}\right) \approx 1.96$ and α-confidence interval

$$I_\alpha \approx \left[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}; \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right].$$

If $\alpha = 0.01$ then $\Phi^{-1}\left(\frac{1-\alpha}{2}\right) \approx 2.58$ and

$$I_\alpha \approx \left[\overline{x} - 2.58\frac{\sigma}{\sqrt{n}}; \overline{x} + 2.58\frac{\sigma}{\sqrt{n}}\right].$$

The larger sampler size n, the narrower α-confidence interval.If we decrease α then length of α-confidence interval increases.

In the next graphic is demonstrated how is gotten number 1.96.

Examnple 2.2.1. let us have a sample $\mathbf{X}$. Let 9 measurements will be performed and let

$$\mathbf{x} = (1.1, 1.8, 1.9, 2.1, 2.2, 2.5, 1.3, 1.9, 1.8)^\intercal.$$

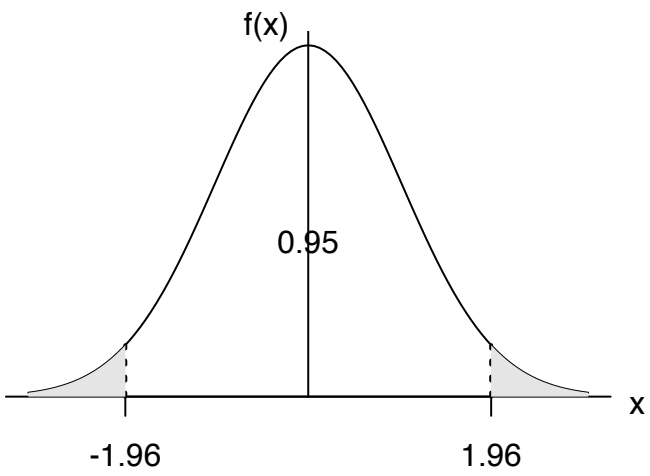

Figure 2.2. Finding quantity $\Phi^{-1}\left(\frac{0.95}{2}\right)$

Let standard deviation $\sigma = 0.5$. On the base of samples realization we get for arithmetical mean $\overline{x} \approx 1.8$. For real expectation we get 0.05-confidence interval

$$I_{0.05} \approx [1.5; 2.2]$$

and 0.01-confidence interval

$$I_{0.01} \approx [1.4; 2.3].$$

Chapter 3

Statistical Tests

Testing of hypothesis is the main goal of mathematical statistics. On base of these tests will be interpreted main results of scientific experiments.

Definition 3.0.1. Statistical test is called as assumption about random variables distributions and its parameters

3.1 The ideological essence of statistical hypothesis

Let us give an overview about main terms of statistical hypothesis. The aim of statistical hypothesis is to reject a previously known truth. This truth is called as zero hypothesis. Our aim is to prove statement which is called as meaningful hypothesis. Let us denote zero hypothesis as H_0 and meaningful hypothesis as H_1. These hypothesis cal formulate as follows:

$$\begin{cases} H_0: \text{ Statement,} \\ H_1: \text{ Objection .} \end{cases}$$

Zero hypothesis and meaningful hypothesis have to formulate so that

$$P(H_0 \cup H_1) = 1 \text{ and } P(H_0 \cap H_1) = 0.$$

Zero hypotheses in the different areas could be as follows:

In law, the presumption of innocence ;

In medicine, synthesized chemical compound is not a drug;

In environmental science, rivers pollution loads are not changed.

Hypothesis H_0 means in general present situation or *status quo*. Meaningful hypothesis H_1 is like change or discovery.

Since control of hypothesis is based on random variable we will inevitably make mistakes in this. These mistakes are two types: I type of error and II type of error.

Let us denote I and II type or error as γ_1 and γ_2 respectively. Thus

$$\gamma_1 = P(\text{to prove } \ H_1 \mid \ \ H_0 \text{ is true})$$

and

$$\gamma_2 = P(\text{to stay on } \ H_0 \ \ \mid \ \ H_1 \text{ is true}).$$

The next table summarizes control of statistical hypothesis.

Is really / Believe to be	Hypothesis H_0	Hypothesis H_1
Hypothesis H_0	True	II type of error γ_2
Hypothesis H_1	I type of error γ_1	True

The question immediately arises as to which of the errors is worse, type I or type II. The answer to this question is that a type I error is much worse than a type II error. Thus, the test of statistical hypotheses must be based on the principle of justice, so that the perpetrator is not punished before the innocent is convicted. With regard to the risk of a type I error, the probability called as p-value is introduced.

Definition 3.1.1. Risk to do I type of error calculated on the base of sample realization is called as p-value.

In Definition 3.1.1 is very important the phrase calculated on the base of sample realization. Let value of $T(\mathbf{X}) = t$ and let us define on the base of value t set $\mathcal{H}_t$ as follows:

1) If graphic of density function is symmetric then $\mathcal{H}_t = (-\infty; -t)$, $\mathcal{H}_t = (t; \infty)$ or $\mathcal{H}_t = (-\infty; -t) \cup (t; \infty)$,

2) If values of test statistic are positive and Its graphic density function graphic is not symmetric then $\mathcal{H}_t = (0; t)$ or $\mathcal{H}_t = (t; \infty)$.

Let us define p-value by the set $\mathcal{H}_t$. By means of that set we get

$$p\text{-}value = P(T(\mathbf{X}) \in \mathcal{H}_t \mid \ \ H_0 \text{ is true }). \tag{3.1}$$

The relation (3.1) gives us essence of p-value.

Next graphics illustrates p-values in the symmetric (Figure 3.1) and non-symmetric (Figures 3.2–3.3) case.

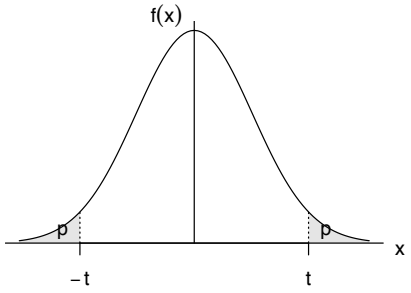

Figure 3.1. Value p of significance level corresponding to the set $\mathcal{H}_t = (-\infty; -t) \cup (t; \infty)$

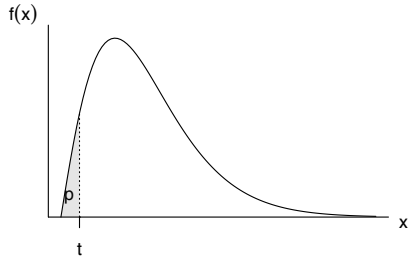

Figure 3.2. Value p of significance level corresponding to the set $\mathcal{H}_t = (0; t)$

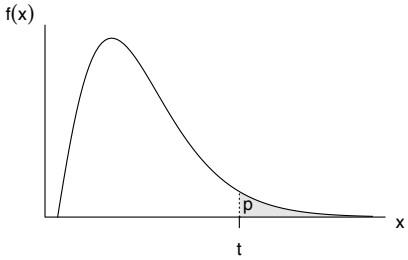

Figure 3.3. Value p of significance level corresponding to the set $\mathcal{H}_t = (t; \infty)$

But what is the greatest tolerated p-value? This limit is called as significance level.

Definition 3.1.2. Maximum allowed value of *p-value* is called as significance level.

Significance level is mostly denoted as α. Decision criterion can be formulated as follows:

1) If *p-value* $< \alpha$ then to prove hypothesis H_1,

2) If *p-value* $\geq \alpha$ then we have to stay on zero hypothesis H_0.

What is the best significance level α? It depends on study area. For example

1) In relatively error-prone areas, such as sociological surveys, the significance level $\alpha = 0.1$,

2) In Life sciences and engineering we can take $\alpha = 0.05$,

3) In tests on which human destiny or human life depends, the significance level of α should be 0.01.

Mostly value of significance level belongs to the set $[0.01; 0.1]$.

Let us study method of statistical tests which is based on finding of critical value. Let us denote this value as t_{crit}. If test statistics density function graphic is not symmetric then can exist two critical values: t_{crit_1} and t_{crit_2}. Olgu $t_{crit_1} < t_{crit_2}$. The criterion choosing the critical value is that

$$P(T(\mathbf{X}) \in \mathcal{H}_{crit} \mid \quad H_0 \text{ is true }) = \alpha$$

where set $\mathcal{H}_{crit}$ depend on critical value t_{crit} or on values t_{crit_1} and (or) t_{crit_2}.

On the base of critical value test statistics values will be divided into 2 parts:

1) Part of zero hypothesis $\mathcal{H}_0$,

2) Part of meaningful hypothesis $\mathcal{H}_1$.

In division into sets $\mathcal{H}_0$ and $\mathcal{H}_1$ we have to take into account that

$$P(T(\mathbf{X}) \in \mathcal{H}_1) = \alpha \text{ and } P(T(\mathbf{X}) \in \mathcal{H}_0) = 1 - \alpha$$

which means that

$$\mathcal{H}_1 = \mathcal{H}_{crit} \text{ and } \mathcal{H}_0 = \mathcal{H}_{crit}^c.$$

The sets $\mathcal{H}_0$ and $\mathcal{H}_1$ are generally constructed as follows:

1) If test statistics density function is paired then

$$\mathcal{H}_0 = [-t_{crit}; 0], \quad \mathcal{H}_0 = [-t_{crit}; t_{crit}] \text{ or } \mathcal{H}_0 = [0; t_{crit}]$$

ning

$$\mathcal{H}_1 = (-\infty; -t_{crit}), \quad \mathcal{H}_1 = (t_{crit}; \infty) \text{ või } \mathcal{H}_1 = (-\infty; -t_{crit}) \cup (t_{crit}; \infty);$$

2) If density function graphic is not symmetric then

$$\mathcal{H}_0 = [t_{crit_1}; t_{crit_2}], \quad \mathcal{H}_0 = [t_{crit_1}; \infty] \text{ or } \mathcal{H}_0 = [0; t_{crit_2}]$$

and

$$\mathcal{H}_1 = [0; t_{crit_1}), \quad \mathcal{H}_1 = (t_{crit_2}; \infty) \text{ or } \mathcal{H}_1 = [0; t_{crit_1}) \cup (t_{crit_2}; \infty).$$

If test statistics value t which is found on the base of sample realization belongs to the set $\mathcal{H}_1$ then we can say that H_1 is proven. If to the set $\mathcal{H}_0$ then we have to stay on hypothesis H_0.

3.2 Different statistical tests

Let us divide statistical tests into 4 parts.

3.2.1 Normal approximation

Normal approximation means that test statistics distribution is approximated with normal distribution. In general, this approximation works if sample size $n \geq 30$. The method is based on Central Limit Theorem which states that

$$Z = \frac{\overline{x} - \mu_0}{\sigma}\sqrt{n} \sim \mathcal{N}(0, 1).$$

Let us call this Statistic as Z-statistic.

Two-sided hypothesis

In this case pair of hypotheses

$$\begin{cases} H_0 : \mu = \mu_0, \\ H_1 : \mu \neq \mu_0. \end{cases}$$

Let us find critical value z_{crit}. We get that on significance level

$$\alpha = P(\mid Z \mid > z_{crit}) = 1 - P(\mid Z \mid \leq z_{crit}) =$$

$$= 1 - P(-z_{crit} \leq Z \leq z_{crit}) = 1 - 2\Phi(z_{crit}),$$

from which follows

$$z_{crit} = \Phi^{-1}\left(\frac{1-\alpha}{2}\right).$$

If $z \leq z_{crit}$ then we have to stay on H_0 but if $z > z_{crit}$ then we can say that hypothesis H_1 is proven. Let us present critical values which correspond to significance levels α.

α	z_{crit}
0.1	1.64
0.05	1.96
0.01	2.58

Probability p-value which corresponds to the value z is found as follows:

$$p\text{-}value = 2(0.5 - \Phi(\mid z \mid)).$$

Left hand sided hypothesis

Pair of hypotheses is

$$\begin{cases} H_0 : \mu = \mu_0, \\ H_1 : \mu < \mu_0. \end{cases}$$

In this case significance level

$$\alpha = P(Z < -z_{crit}) = 0.5 + \Phi(-z_{crit})$$

from whic follows

$$-z_{crit} = \Phi^{-1}(\alpha - 0.5).$$

The hypothesis H_1 is said to be proven if $z < -z_{crit}$. In the other case we have to stay on hypothesis H_0. In corresponding case probability

$$p\text{-}value = \Phi(z) + 0.5.$$

Right hand sided hypothesis

Pair of hypothesis is

$$\begin{cases} H_0 : \mu = \mu_0, \\ H_1 : \mu > \mu_0. \end{cases}$$

In corresponding case significance level

$$\alpha = P(Z > z_{crit}) = 1 - P(Z \leq z_{crit}) = 1 - (0.5 + \Phi(z_{crit}))$$

from which follows that

$$z_{crit} = \Phi^{-1}(0.5 - \alpha).$$

Hypothesis H_1 is said to be proven if $z > z_{crit}$. In the other case we have to stay on hypothesis H_0. Probability

$$p\text{-}value = 0.5 - \Phi(z).$$

Let us apply normal approximation on different distributions

Application on binomial distribution

Let us try n times independently an event A. Our aim is to test probability $p = P(A)$. Let us have sample $\mathbf{X} = (X_1, X_2, ..., X_n)$ where

$$X_i = \begin{cases} 1, \text{ if } A, \\ 0, \text{ if } \overline{A} \end{cases}$$

and pair of hypotheses

$$\begin{cases} H_0 : p = p_0, \\ H_1 : p \neq p_0. \end{cases}$$

In this case $E(X_i) = p_0$ and $V(X_i) = p_0(1 - p_0)$. Thus test statistic

$$Z = \frac{\widehat{p} - p_0}{\sqrt{p_0(1 - p_0)}}\sqrt{n},$$

where $\widehat{p} = \dfrac{k}{n}$ and k denotes number of cases when $X_i = 1$.

Application on Poisson distribution

Let us have sample where $X_i =\sim Po(\lambda)$ $i = 1, 2, ..., n$. Let us study frequency of rare event and test Poisson distribution parameter λ. In this case pair of hypothesis is

$$\begin{cases} H_0 : \lambda = \lambda_0, \\ H_1 : \lambda \neq \lambda_0. \end{cases}$$

Under zero hypothesis H_0 expectation $E(X_i) = V(X_i) = \lambda_0$. According to the MLE for λ we get test statistic

$$Z = \frac{\overline{x} - \lambda_0}{\sqrt{\lambda_0}}\sqrt{n}.$$

Application on exponential distribution

In this case w have sample where $X_i \sim \mathcal{E}(\nu)$, $i = 1, 2, ..., n$, $\nu > 0$. Let under zero hypothesis $\nu = \nu_0$. Thus we have pair of hypotheses

$$\begin{cases} H_0 : \mu = \dfrac{1}{\nu_0}, \\ H_1 : \mu \neq \dfrac{1}{\nu_0}. \end{cases}$$

In the case of hypothesis H_0 expectation $E(X_i) = \sqrt{V(X_i)} = \dfrac{1}{\nu_0}$ and MLE for parameter ν equals with arithmetical mean $\overline{x}$. The test Statistic

$$Z = \frac{\overline{x} - \dfrac{1}{\nu_0}}{\dfrac{1}{\nu_0}} \sqrt{n}.$$

3.2.2 Chi-squared tests

Definition 3.2.1. Let us have independent random variables

$$X_1, X_2, ..., X_n.$$

Let $X_i \sim \mathcal{N}(0, 1)$, $i = 1, 2, ..., n$. Then random variable

$$Y_n = \sum_{i=1}^{n} X_i^2$$

has χ^2-squared distribution with degrees of freedom n.

This is denoted as $Y_n \sim \chi^2(n)$.

Proposistion 3.2.1. Let us have independent events $X_i \sim \mathcal{N}(\mu, \sigma)$, $i = 1, 2, ..., n$. Let

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then random variable

$$H = \frac{\sum_{i=1}^{n} (X_i - \overline{x})^2}{\sigma^2}$$

has χ^2-squared distribution with degrees of freedom $n - 1$.

Let us study 3 different types of chi-squared tests

Variance test for normal distribution

From 3.2.1 follows that

$$H = \frac{s^2}{\sigma^2}(n-1) \sim \chi^2(n-1).$$

Let $X_i \sim \mathcal{N}(\mu, \sigma)$, $i = 1, 2, ..., n$ and we have pair of hypotheses

$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2, \\ H_1 : \sigma^2 \neq \sigma_0^2. \end{cases}$$

Under H_0 test statistic

$$H = \frac{s^2}{\sigma_0^2}(n-1) \sim \chi^2(n-1).$$

In our case

$$p\text{-}value = \min\{P(H > h); P(H \leq h)\}.$$

On the base of random variable H we can construct α-confidence interval for variance of normal distribution. Let α-quantile be $h_{\alpha;n}$ if degrees of freedom is n. Then we get

$$P\left(h_{\frac{\alpha}{2};n-1} \leq \frac{s^2}{\sigma^2}(n-1) \leq h_{1-\frac{\alpha}{2};n-1}\right) = 1-\alpha \Longleftrightarrow$$

$$\Longleftrightarrow P\left(\frac{h_{\frac{\alpha}{2};n-1}}{s^2(n-1)} \leq \frac{1}{\sigma^2} \leq \frac{h_{1-\frac{\alpha}{2};n-1}}{s^2(n-1)}\right) = 1-\alpha \Longleftrightarrow$$

$$\Longleftrightarrow P\left(\frac{s^2(n-1)}{h_{1-\frac{\alpha}{2};n-1}} \leq \sigma^2 \leq \frac{s^2(n-1)}{h_{\frac{\alpha}{2};n-1}}\right) = 1-\alpha.$$

We have gotten α-confidence interval

$$I_\alpha = \left[\frac{s^2(n-1)}{h_{1-\frac{\alpha}{2};n-1}}; \frac{s^2(n-1)}{h_{\frac{\alpha}{2};n-1}}\right].$$

Goodness of fit test

Let us divide our sample to size classes:

Size class	Freqency
$[x_1; x_2)$	n_1
$[x_2; x_3)$	n_2
$\vdots$	$\vdots$
$[x_i; x_{i+1})$	n_i
$\vdots$	$\vdots$
$[x_m; x_{m+1})$	n_m

Every sample element belongs to one size class and

$$\sum_{i=1}^{m} n_i = n.$$

Let us have pair of hypotheses

$$\begin{cases} H_0 : X \sim F, \\ H_1 : X \text{ has another distribution .} \end{cases}$$

For rejecting of H_0 the next statistic will be composed:

$$H = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i}, \tag{3.2}$$

where

$$p_i = P(x_i \leq X \leq x_{i+1}) = F(x_{i+1}) - F(x_i).$$

Statistic described by equation 3.2 is called as Pearson chi-squared criterion. This statistic has approximately chi-squared distribution with degrees of freedom $m - k - 1$ where m is number of size classes and k number of distribution F parameters. For example, in the case of normal distribution $k = 2$, but F is exponential distribution then $k = 1$.

Test of dependence

Let us have random variables $X = X_1, X_2, ..., X_l$ and $Y = Y_1, Y_2, ..., Y_r$. Let symbol $\perp$ denote independence. Let us have pair of hypotheses

$$\begin{cases} H_0 : X \perp Y, \\ H_1 : X, Y \text{ are not independent.} \end{cases}$$

By means of frequency table we can compose test statistic

$$H = n \sum_{i=1}^{l} \sum_{j=1}^{r} \frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}} \sim \chi^2((l-1)(r-1)).$$

In both tests (goodness of fit and dependence test) we have following criterion of decision:

If $h > h_{crit}$ then we can prove H_1,

If $h \leq h_{crit}$ then we have to stay on H_0.

Value h_{crit} is in both tests chosen so that

$$P(H > h_{crit}) = \alpha.$$

Probability

$$p\text{-}value = P(H > h)$$

where h is value of test static found on the base of sample.

3.2.3 Student t-tests

Student was a nickname of William Sealy Gosset (1876 — 1937).

Definition 3.2.2. Let us have random variable $Z \sim \mathcal{N}(0,1)$ and random variable $Y_n \sim \chi^2(n)$. Then random variable

$$T = \frac{Z}{\sqrt{\frac{Y_n}{n}}}$$

has t-distribution with degrees of freedom n.

This is denoted as $T \sim t(n)$.

Proposistion 3.2.2. Let random variables $X_i \sim \mathcal{N}(\mu, \sigma)$, $i = 1, 2, ..., n$. Then

$$T = \frac{\overline{x} - \mu}{s}\sqrt{n} \sim t(n-1),$$

where s denotes unbiased estimator for standard deviation of X_i.

On statement 3.2.2 is based Student t-test.

Let us have sample $\mathbf{X} = (X_1, X_2, ..., X_n)^\intercal$ where $X_i \sim \mathcal{N}(\mu, \sigma)$, $i = 1, 2, ..., n$. We compose to the parameter μ (expectation of X_i) pair of hypotheses.

$$\begin{cases} H_0 : \mu = \mu_0, \\ H_1 : \mu \neq \mu_0. \end{cases}$$

To this pair correspond value $\mu = \mu_0$ in the Proposition 3.2.2.

Let us compose t-test testing two populations. Let population $\mathbf{X} = (X_1, X_2, ..., X_n)^\intercal$ and population $\mathbf{Y} = (Y_1, Y_2, ..., Y_m)^\intercal$. It is essential to assume that $X_i \sim \mathcal{N}(\mu_x, \sigma_x)$, $i = 1, 2, ..., n$ and $Y_j \sim \mathcal{N}(\mu_y, \sigma_y)$, $j = 1, 2, ..., m$. Let us compose pair of hypotheses

$$\begin{cases} H_0 : \mu_x = \mu_y, \\ H_1 : \mu_x \neq \mu_y. \end{cases}$$

This pair can control with statistic

$$T = \frac{\overline{x} - \overline{y}}{\sqrt{(n-1)s_x^2 + (m-1)s_y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}} \sim t(n+m-2)$$

where s_x^2 and s_y^2 denote unbiased estimators of samples $\mathbf{X}$ and $\mathbf{Y}$ variances.

Advantages and failures of normal approximation and t-tests are described in the following table:

	Student t-test	Normal approximation
Advantages	Sample size n can be small	Large choice of distributions
Failures	We have to assume $X \sim \mathcal{N}(\mu, \sigma)$	Sample size $n \geq 30$

3.2.4 Fisher F-tests

This test was found by Ronald Aylmer Fisher (1890 — 1962).

Definition 3.2.3. Let $Y_n \sim \chi^2(n)$ and $Y_m \sim \chi^2(m)$. Then random variable

$$G = \frac{\frac{Y_n}{n}}{\frac{Y_m}{m}}$$

has F-distribution with degrees of freedom n and m.

Let us denote that as $G \sim F(n, m)$.

Let us have samples $\mathbf{X} = (X_1, X_2, ..., X_n)^\intercal$ and $\mathbf{Y} = (Y_1, Y_2, ..., Y_m)^\intercal$. Let $X_i \sim \mathcal{N}(\mu_x, \sigma_x)$, $i = 1, 2, ..., n$ ja $Y_j \sim \mathcal{N}(\mu_y, \sigma_y)$, $j = 1, 2, ..., m$. Let us have pair of hypotheses

$$\begin{cases} H_0 : \sigma_x^2 = \sigma_y^2, \\ H_1 : \sigma_x^2 \neq \sigma_y^2. \end{cases}$$

For controlling of these hypotheses let us find unbiased estimators of variances

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{x})^2$$

and

$$s_y^2 = \frac{1}{m-1} \sum_{j=1}^{m} (Y_j - \overline{y})^2.$$

The statistic

$$H_x = \frac{s_x^2}{\sigma_x^2}(n-1) \sim \chi^2(n-1)$$

and statistic

$$H_y = \frac{s_y^2}{\sigma_y^2}(m-1) \sim \chi^2(m-1).$$

Thus we can construct from ratio of H_x and H_y the test Statistic

$$G = \frac{s_x^2}{s_y^2} \sim F(n-1, m-1), \tag{3.3}$$

which parameters $n-1$ and $m-1$ are called as degrees of freedom.

Chapter 4

General Linear models

Let us apply tools of Probability and Statistics on real data. Methods of statistical analyses can divide into 3 parts:

1) Regression analysis,

2) Analyse of Variance (ANOVA),

3) Factor Analysis.

4.1 Structure of General Linear Model (GLM)

In this course we handle with General Linear models in regression analysis. Data analyses starts from data matrix which has n rows and $k+1$ columns. The rows are for objects and columns for variables. Data matrix can present as following table:

$\mathbf{Y}$	$\mathbf{X}_1$	$\cdots$	$\mathbf{X}_j$	$\cdots$	$\mathbf{X}_k$
y_1	x_{11}	$\cdots$	x_{1j}	$\cdots$	x_{1k}
$\vdots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\vdots$
y_i	x_{i1}	$\cdots$	x_{ij}	$\cdots$	x_{ik}
$\vdots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\vdots$
y_n	x_{n1}	$\cdots$	x_{nj}	$\cdots$	x_{nk}

The first column of data matrix $\mathbf{Y} = (Y_1, Y_2, ..., Y_i, ..., Y_n)^\intercal$ is called as

response variable other k columns

$$\mathbf{X}_j = (X_{1j}, X_{2j}, ..., X_{ij}, ..., X_{nj})^\intercal, \quad j = 1, 2, ..., k,$$

can call as factors.

In the case of general linear models we suppose that response variable $Y_i \sim \mathcal{N}(\mu, \sigma)$, $i = 1, 2, ..., n$. Basic form of general linear model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \epsilon_i, \tag{4.1}$$

where models residuals $\epsilon_i \sim \mathcal{N}(0, \sigma)$. The matrix form of model (4.1) is as follows:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{i1} & \cdots & X_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Briefly we can write it as

$$\mathbf{Y} = \mathbf{Z}\beta + \epsilon$$

where$\mathbf{Y} : n \times 1$ is vector of response variable, $\mathbf{Z} : n \times (k+1)$ denotes design matrix where first column consists of values 1,$\beta : (k+1) \times 1$ is called as parameters vector and $\epsilon : n \times 1$ denotes models residuals vector. In modelling we find expectation of Y. Assuming that $E(\epsilon_i) = 0$ we get

$$E(Y_i) = \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}$$

or in the vector form

$$\mu_i = \mathbf{z}_i \beta, \tag{4.2}$$

where $\mu_i = E(Y_i)$ and $\mathbf{z}_i : 1 \times k + 1$ denotes row vector of matrix $\mathbf{Z}$.

4.2 Estimation of models parameters with Least Square method

Let us find the better estimator for vector

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}.$$

This estimation is based on Least Square method.

Our aim is to minimize function

$$Q = \sum_{i=1}^{n} (y_i - \mathbf{z}_i \beta)^2.$$

That means we have to find such values of parameters $\beta_0, \beta_1, ..., \beta_k$ when squared difference between empirical and modelled values is minimal. Let us assume that graphic of function Q is concave. Then we have to solve system

$$\frac{\partial Q}{\partial \beta_j} = 0, \quad j = 0, 1, ..., k. \tag{4.3}$$

Let us present solution of system (4.3) in 3 steps.

1) Let us have so-called zero model. That means in the model is only parameter β_0. We get that

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (y_i - \beta_0)^2 = 0$$

from which follows

$$\beta_0 = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

Thus the best estimator for parameter β_0 is arithmetical mean of the measurements.

2) Let us have a simple model

$$y_i = \beta_0 + \beta_1 \cdot x_i.$$

Let measurements of the factor $x_1, x_2, ..., x_n$. Then we get system (4.3) in the form like

$$\begin{cases} \dfrac{\partial}{\partial \beta_0} \displaystyle\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = 0 \\ \dfrac{\partial}{\partial \beta_1} \displaystyle\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = 0. \end{cases}$$

Finding partial derivatives we get the linear system for finding parameters β_0 and β_1

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \end{cases}$$

Solving that system we get

$$\beta_0 = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}$$

and

$$\beta_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}.$$

Let $\overline{xy} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i$ and $\overline{x^2} = \frac{1}{n} \sum_{i=1}^{n} x_i^2$. Then

$$\beta_0 = \frac{\overline{y}\ \overline{x^2} - \overline{x}\ \overline{xy}}{\overline{x^2} - (\overline{x})^2}$$

and

$$\beta_1 = \frac{\overline{xy} - \overline{x}\ \overline{y}}{\overline{x^2} - (\overline{x})^2}.$$

Let $\overline{s^2}_x$ and $\overline{s^2}_y$ be unbiased estimators for variances of random variables X and Y. Statistic $\overline{s}_{xy} = \overline{xy} - \overline{x}\ \overline{y}$ is unbiased estimator of covariation between X and Y. Then we get point estimator for Pearson linear correlation coefficient

$$r_{xy} = \frac{\overline{s}_{xy}}{\sqrt{\overline{s^2}_x} \sqrt{\overline{s^2}_y}}.$$

Since $\overline{s^2}_x = \overline{x^2} - \overline{x}^2$ ja $\overline{s^2}_y = \overline{y^2} - \overline{y}^2$ the parameter

$$\beta_1 = \frac{\overline{s}_{xy}}{\overline{s^2}_x} = \frac{r_{xy} \sqrt{\overline{s^2}_x} \sqrt{\overline{s^2}_y}}{\overline{s^2}_x} = r_{xy} \sqrt{\frac{\overline{s^2}_y}{\overline{s^2}_x}}.$$

We have gotten

$$\beta_1 = r_{xy}\frac{\overline{s}_y}{\overline{s}_x}. \tag{4.4}$$

3) Let us study least square estimation in general case. We have to find derivative of function Q be parameters vector β. We get

$$\frac{\partial Q}{\partial \beta} = \frac{\partial}{\partial \beta}(\mathbf{y} - \mathbf{Z}\beta)^\intercal(\mathbf{y} - \mathbf{Z}\beta) = -2\mathbf{Z}^\intercal(\mathbf{y} - \mathbf{Z}\beta) = \mathbf{0}_{k+1},$$

where $\mathbf{0}_{k+1}$ denotes zero vector with $k+1$ elements. After some matrix operations we get that the best estimator in the sense of least square

$$\widehat{\beta} = (\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal\mathbf{y}. \tag{4.5}$$

Not that not always we can find unique inverse matrix for $\mathbf{Z}^\intercal\mathbf{Z}$.

4.3 Diagnostics of GLM

The main part of GLM diagnostics consists of 3 stages.

4.3.1 Is the model significant?

Let us define pair of hypotheses:

$$\begin{cases} H_0 : \text{Model is not significant}, \\ H_1 : \text{Model is significant}. \end{cases}$$

Let us have a realization of empirical values Y

$$y_1, y_2, ..., y_n$$

and theoeretical values $\widehat{Y}$ or realization of the model (4.2

$$\widehat{y}_1, \widehat{y}_2, ..., \widehat{y}_n.$$

Let us divide total variance S_y into two parts: regression variance S_{reg} and residual variance S_{res}. We get

$$S_y = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i + \widehat{y}_i - \overline{y})^2 =$$

$$= \sum_{i=1}^{n} \{(y_i - \widehat{y}_i)^2 + 2(y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) + (\widehat{y}_i - \overline{y})^2\} =$$

$$= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 + 2\sum_{i=1}^{n} (y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) + \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2.$$

Supposing that $E(\epsilon_i) = 0$ we get

$$\sum_{i=1}^{n} (y_i - \widehat{y}_i) = 0 \Rightarrow \sum_{i=1}^{n} (y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) = 0.$$

Let

$$S_{reg} = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$$

regression variance and

$$S_{res} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

residula variance. Thus we have gotten

$$S_y = S_{reg} + S_{res}.$$

Which variance S_{reg} or S_{res} proposition is greater in total variance S_y? For controlling of hypothesis models significance will be composed test statistic

$$G = \frac{S_{reg}(n-k-1)}{S_{res}k} \sim F(k, n-k-1).$$

If in model is 1 factor then

$$G \sim F(1, n-2).$$

Model is said to be significant in the level α if $g > g_{1-\alpha}$. If model is significant then the next question will be posed.

4.3.2 Which models parameters are significant?

Let us separate significant and non-significant parameters of GLM. For controlling significance of parameters $\beta_0, \beta_1, ..., \beta_k$ will be composed pair of hypotheses

$$\begin{cases} H_0 : \beta_j = 0, \\ H_1 : \beta_j \neq 0, \end{cases}$$

where $j = 0, 1, ..., k$. This pair of hypotheses is controlled bey means of t-distribution. Let $\widehat{\beta}_j$ be estimator of β_j calculated on the base of (4.5). Let deviation of jth factor

$$S_{x_j} = \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2$$

where

$$\overline{x}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}$$

$j = 1, 2, ..., k$. Then test statistic

$$T_{\beta_j} = \frac{\widehat{\beta}_j}{\sqrt{\frac{S_y}{(n-2)S_{x_j}}}} \sim t(n-2), \quad j = 1, 2, ...k.$$

If $j = 0$ then

$$T_{\beta_j} = \frac{\widehat{\beta}_j}{\sqrt{\frac{S_y}{(n-2)}}} \sim t(n-2).$$

Statistic T_{β_0} is for testing of linear models intercept.

In the model will be retained these parameters for which hypothesis H_1 is proven. For other parameters we can say that they don't differ significantly from zero.

4.3.3 The descriptive ability of the model

Finally we ask how much of the response variable Y variance is described by the model and how much by the residuals. Descriptive ability of the model is characterize by

$$R^2 = \frac{S_{reg}}{S_y} \in [0; 1]$$

which is called as Rsquare coefficient.

Variance of the response variable describes $R^2 \cdot 100\%$ the model (4.1) and $(1 - R^2) \cdot 100\%$ residulas ϵ_i. The next statement is valid

Proposistion 4.3.1. Let $r^2_{y\widehat{y}}$ denotes estimator of Pearson linear correlation coefficient between measured value Y and modelled value $\widehat{Y}$. Then

$$R^2 = r^2_{y\widehat{y}}.$$