Statistics

1. Choosing a model
2. Estimating its parameter(s)
    1. point estimates
    2. interval estimates
3. Testing hypotheses

Distributions used in statistics: χ^2_n-distribution

Let $X_1, X_2 \ldots, X_n$ be independent identically distributed random variables with distribution $N(0, 1)$. Then the random variable

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

has so called χ^2_n-distribution ("χ-square distribution with n degrees of freedom").

Let X be a random variable with distribution $N(0,1)$ and Y be a random variable with distribution χ^2_n. Then the random variable

$$Z = \frac{X}{\sqrt{Y}}\sqrt{n}$$

has so called t_n-distribution (called also Student t-distribution with n degrees of freedom).

Definition

Random vector $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$ of independent identically distributed random variables with distribution function F_θ dependent on a parameter θ is called the random sample.

Definition

The function

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

of the random sample $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$ is called the sample mean and the function

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

is called the sample variance. $S_n = \sqrt{S_n^2}$ is then called the sample standard deviation.

Sample mean and sample variance for normal distribution 6

Theorem

Let $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$ *be a random sample from distribution* $N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0$. *Then*

1. *the sample mean* $\bar{X}_n$ *and the sample variance* S_n^2 *are independent random variables,*
2. *the distribution of the sample mean* $\bar{X}_n$ *is* $N(\mu, \sigma^2/n)$,
3. *the random variable* $(n-1)S_n^2/\sigma^2$ *has* $\chi^2_{(n-1)}$*-distribution,*
4. *random variable* $T = \frac{\bar{X}_n - \mu}{S_n}\sqrt{n}$ *has* $t_{(n-1)}$*-distribution.*

Definition

Let the distribution function F be continuous, monotone and let $0 < \beta < 1$. Then the number z_β so that $F(z_\beta) = \beta$ is called β-quantile of this distribution.

Basic quantiles

- u_β ... β-quantile of the standard normal distribution,
- $t_{\beta,n}$ ... β-quantile of the t_n-distribution,
- $\chi^2_{\beta,n}$ ... β-quantile of the χ^2_n distribution.

Remark

If the random variable X has the distribution function F and the quantiles z_β, then

$$P(z_{\alpha/2} < X < z_{1-\alpha/2}) = F(z_{1-\alpha/2}) - F(z_{\alpha/2}) = 1 - \alpha.$$

Definition

Let $(x_1, x_2 \ldots, x_n)^T$ be a realisation of the random sample $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$. Then

$$F_{emp}(x) = \frac{\#\{x_i : x_i \leq x\}}{n},$$

where $\#$ denotes the number of elements, is called the empirical distribution function.

Definition

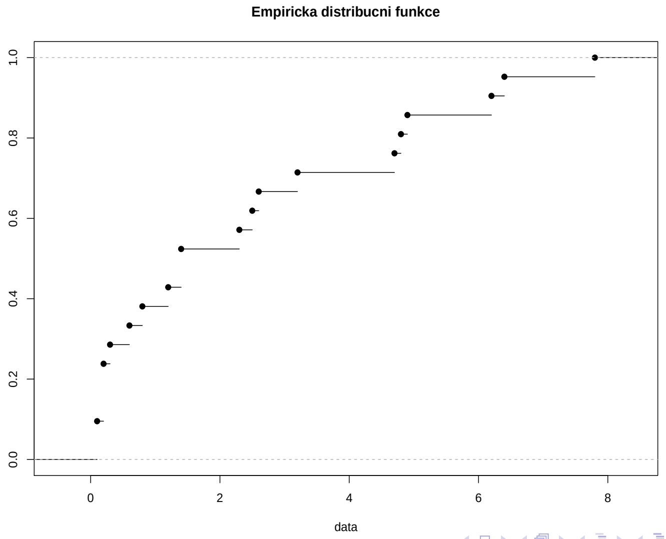
Let $(x_1, x_2 \ldots, x_n)^T$ be a realisation of the random sample $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$. Then

- $z = min(x_i : F_{emp}(x_i) \geq 1/4)$ is called the first quartile,
- $z = min(x_i : F_{emp}(x_i) \geq 3/4)$ is called the third quartile,
- $z = min(x_i : F_{emp}(x_i) \geq 1/2)$ is called median (the second quartile),
- the most occurring element is called the modus.

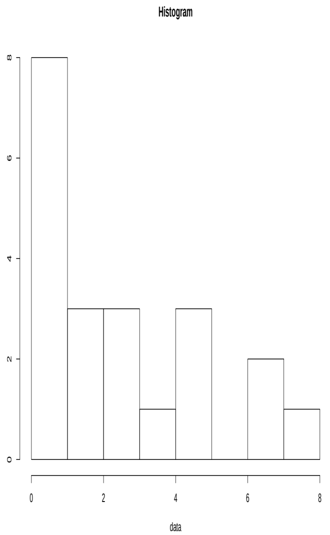
Durations till the breakdown of an instrument were observed 21×. The values are:
4.9, 6.2, 2.6, 0.6, 0.3, 2.3, 3.2, 1.4, 6.4, 4.8, 1.2
2.5, 0.2, 0.2, 0.8, 0.1, 0.1, 1.4, 7.8, 0.2, 4.7.

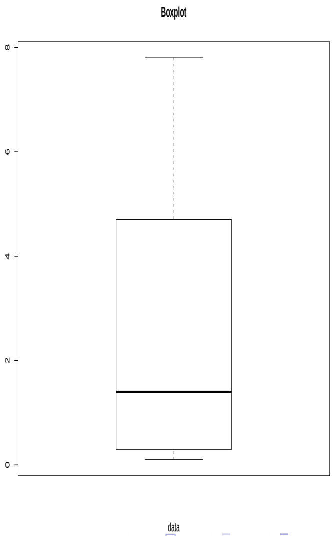
For better summary, order the data from the smallest value to the largest one. We get:
0.1, 0.1, 0.2, 0.2, 0.2, **0.3**, 0.6, 0.8, 1.2, 1.4, **1.4**,
2.3, 2.5, 2.6, 3.2, **4.7**, 4.8, 4.9, 6.2, 6.4, **7.8**.

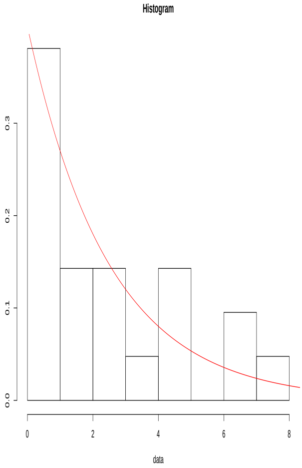
We have:
sample mean $\bar{X}_{21} = 2.471$,
sample variance $S_{21}^2 = 5.81$,
sample standard deviation $S_{21} = \sqrt{5.81} = 2.21$,
1st quartile $= 0.3$, median (i.e. 2nd quartile) $= 1.4$ and 3rd quartile $=$ 4.7,
minimum $= 0.1$, maximum $= 7.8$, modus $= 0.2$.


Empiricka distribucni funkce
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
data

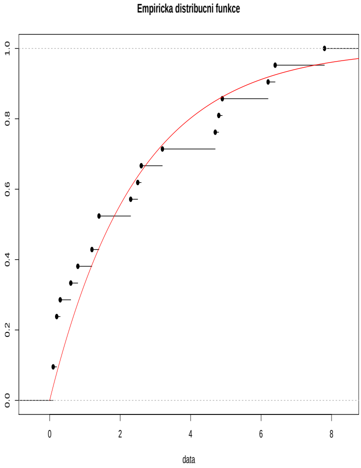
Histogram and boxplot







Histogram
0.0
0.1
0.2
0.3
0
2
4
6
8
data



Empiricka distribucni funkce
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
data

Definition

Consider a random sample $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$, where the distribution (i.e. the distribution function etc.) of the random variables $X_1, \ldots, X_n$ depends on a parameter θ. Point estimate of the parameter θ is an arbitrary function $\theta^*(\mathbb{X})$ of the random sample $\mathbb{X}$, whose formula does not depend on θ. If $\mathrm{E}\theta^*(\mathbb{X}) = \theta$, then the estimate is called unbiased.

Remark

For simplicity, we can imagine the estimate $\hat{\theta}$ *as the number obtained from the realisation* $(x_1, x_2 \ldots, x_n)^T$ *of the random sample* $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$, *where this number corresponds to the parameter* θ *as best as possible.*

Consider $(x_1, x_2 \ldots, x_n)^T$ a realisation of the random sample $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$. The distribution of the random variables $X_1, \ldots, X_n$ depends on the parameters $\theta_1, ..., \theta_k \in \Theta$, where Θ is a parameter set (e.g. positive real numbers).

Assumptions: $\mathrm{E}X_1^i < \infty \ \forall i = 1, ...k$ and $\mathrm{E}X_1^i$ depend on $\theta_1, ..., \theta_k$.

Method: Set

$$\mathrm{E}X_1^i = m_i,$$

where m_i is $i-$th sample moment obtained as

$$m_i = \frac{1}{n} \sum_{j=1}^{n} x_j^i$$

for all $i = 1, ...k$. In this way, we get system of k equations of k variables $\theta_1, ..., \theta_k$, whose solutions are the required estimates $\hat{\theta}_1, ..., \hat{\theta}_k$.

Alternative: If $k = 2$, then instead of i-th moments, $i = 1, 2$, we can take $\mathrm{E}X_1 = \bar{X}_n$ and $varX_1 = S_n^2$.

Disadvantage: This estimate has large variance.

Consider $(x_1, x_2 \ldots, x_n)^T$ a realisation of the random sample $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$ from the distribution with probabilities $P_\theta(X_1 = .)$ or with a density f_θ, respectively, and let this probabilities or density, respectively, depend on a parameter $\theta \in \Theta$.

Definition

The estimate $\hat{\theta}$ is called the maximum likelihood estimate, if

$$\prod_{i=1}^{n} P_{\hat{\theta}}(X_1 = x_i) = \max_{\theta \in \Theta} \prod_{i=1}^{n} P_\theta(X_1 = x_i),$$

$$\text{or} \quad \prod_{i=1}^{n} f_{\hat{\theta}}(x_i) = \max_{\theta \in \Theta} \prod_{i=1}^{n} f_\theta(x_i), \quad \text{respectively.}$$

Maximum likelihood method for the sample from discrete distribution

1. Construct the likelihood function $L(\theta) = \prod_{i=1}^{n} P_\theta(X_1 = x_i)$.
2. Construct the log-likelihood function $l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log P_\theta(X_1 = x_i)$.
3. Set $\frac{\partial l(\theta)}{\partial \theta} = 0$.
4. The solution of $\frac{\partial l(\theta)}{\partial \theta} = 0$ is the required maximum likelihood estimate $\hat{\theta}$.

1. Construct the likelihood function $L(\theta) = \prod_{i=1}^{n} f_\theta(x_i)$.
2. Construct the log-likelihood function $l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f_\theta(x_i)$.
3. Set $\frac{\partial l(\theta)}{\partial \theta} = 0$.
4. The solution of $\frac{\partial l(\theta)}{\partial \theta} = 0$ is the required maximum likelihood estimate $\hat{\theta}$.

Definition

Consider a random sample $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$ and a number $\alpha \in (0, 1)$.

1. The couple $(\theta_L^*(X_1, \ldots, X_n), \theta_U^*(X_1, \ldots, X_n))$ is called the $(1-\alpha)$ confidence interval estimate (denoted as $(1-\alpha)$-CI or $(1-\alpha) \cdot 100\%$-CI) of the parameter θ, if

$$P(\theta_L^*(X_1, \ldots, X_n) < \theta < \theta_U^*(X_1, \ldots, X_n)) = 1 - \alpha.$$

2. $(\theta_D^*(X_1, \ldots, X_n))$ is called the lower $(1-\alpha)$-CI if

$$P(\theta_D^*(X_1, \ldots, X_n) < \theta) = 1 - \alpha.$$

3. $(\theta_H^*(X_1, \ldots, X_n))$ is called the upper $(1-\alpha)$-CI if

$$P(\theta_H^*(X_1, \ldots, X_n) > \theta) = 1 - \alpha.$$

Theorem

Let $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$ be a random sample from the distribution $N(\mu, \sigma^2), \mu \in \mathbb{R}$ is unknown parameter, $\sigma^2 > 0$ is known constant. Then

1. $(\bar{X}_n - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}_n + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}})$ *is the* $(1-\alpha)$*-CI of the parameter* μ*,*
2. $\bar{X}_n - u_{1-\alpha}\frac{\sigma}{\sqrt{n}}$ *is the lower* $(1-\alpha)$*-CI of the parameter* μ*,*
3. $\bar{X}_n + u_{1-\alpha}\frac{\sigma}{\sqrt{n}}$ *is the upper* $(1-\alpha)$*-CI of the parameter* μ*,.*

Theorem

Let $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$ *be a random sample from the distribution* $N(\mu, \sigma^2), \mu \in \mathbb{R}$, $\sigma^2 > 0$, *oba parameters neznámé. Then*

1. $(\bar{X}_n - t_{1-\alpha/2,n-1}\frac{S_n}{\sqrt{n}}, \bar{X}_n + t_{1-\alpha/2,n-1}\frac{S_n}{\sqrt{n}})$ *is the* $(1-\alpha)$*-CI of the parameter* μ,
2. $\bar{X}_n - t_{1-\alpha,n-1}\frac{S_n}{\sqrt{n}}$ *is the lower* $(1-\alpha)$*-CI of the parameter* μ,
3. $\bar{X}_n + t_{1-\alpha,n-1}\frac{S_n}{\sqrt{n}}$ *is the upper* $(1-\alpha)$*-CI of the parameter* μ.
4. $(\frac{(n-1)S_n^2}{\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)S_n^2}{\chi^2_{\alpha/2,n-1}})$ *is the* $(1-\alpha)$*-CI of the parameter* σ^2,
5. $\frac{(n-1)S_n^2}{\chi^2_{1-\alpha,n-1}}$ *is the lower* $(1-\alpha)$*-CI of the parameter* σ^2,
6. $\frac{(n-1)S_n^2}{\chi^2_{\alpha,n-1}}$ *is the upper* $(1-\alpha)$*-CI of the parameter* σ^2.

Theorem

Let $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$ *be a random sample from an arbitrary distribution for which* $0 < \sigma^2 < \infty$. *Then the* $(1-\alpha)$*-CI of the parameter* $\mu = \mathrm{E}X$ *is*

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

Proof

Recall that for a large n it holds that $\frac{S_n}{\sigma} \to 1$, i.e. S_n is an approximation of σ. From CLT, we know that $\frac{\sum X_i - n\mu}{\sqrt{n\sigma^2}}$ has approximately standard normal distribution, i.e.

$$P(u_{\frac{\alpha}{2}} \leq \frac{\sum X_i - n\mu}{\sqrt{n\sigma^2}} \leq u_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

$$P(u_{\frac{\alpha}{2}} \leq \frac{\sum X_i - n\mu}{\sqrt{n}S_n} \leq u_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

$$P(\frac{\sum X_i}{n} + u_{1-\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}} \geq \mu \geq \frac{\sum X_i}{n} + u_{\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}}) = 1 - \alpha$$

$$P(\bar{X}_n + u_{1-\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}} \geq \mu \geq \bar{X}_n - u_{1-\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}}) = 1 - \alpha,$$

which is the definition of the $(1-\alpha)$-CI of the parameter μ.

Theorem

1. *Let* $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$ *be a random sample from alternative distribution with parameter* $0 < p < 1$*. Then the* $(1-\alpha)$*-CI of the parameter* p *is*

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

2. *Let* $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$ *be a random sample from Poisson distribution with parameter* $0 < \lambda < \infty$*. Then the* $(1-\alpha)$*-CI of the parameter* λ *is*

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

Proof

The proof comes from the forms of the characteristics of the introduced distribution: for alternative distribution, we have $\mathrm{E}X = p$ and $varX = p(1-p)$ and for Poisson distribution it holds that $\mathrm{E}X = varX = \lambda$.

- Let $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$ be a random sample from distribution depending on a parameter $\theta \in \Theta$.
- The assumption that θ belongs to a set Θ_0, is called null hypothesis (denote $H_0 : \theta \in \Theta_0$).
- Based on the sample $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$, we test the null hypothesis again the alternative hypothesis $H_A : \theta \in \Theta \setminus \Theta_0$. In order to test it, we establish a set W (so called rejection region) so that we reject H_0 if $\mathbb{X} \in W$, and we accept H_0 otherwise.

Remark

Mostly, we test $H_0 : \theta = \theta_0$, where θ_0 is a concrete value. Natural alternative is then $H_A : \theta \neq \theta_0$. However sometimes, it is more meaningful to consider $H_A : \theta > \theta_0$ (even when it is theoretically possible that $\theta < \theta_0$, it has no sense for us).

The following situations may occur:

- H_0 holds and the test accepts it $\surd$
- H_0 does not hold and the test rejects it $\surd$
- H_0 holds and the test rejects it $\rightarrow$ error of the first kind
- H_0 does not hold and the test accepts it $\rightarrow$ error of the second kind

Significance level:
Choose a number α (usually 0.05, sometimes 0.01 or 0.1). W is constructed so that the probability of the error of the first kind is not larger (usually equal to) α. Such α is called the significance level.

Let $(X_1, X_2 \dots, X_n)^T$ be a random sample from $N(\mu, \sigma^2)$, where $\sigma^2 > 0$, and assume that both the parameters are unknown. Testing $H_0 : \mu = \mu_0$ against $H_A : \mu \neq \mu_0$ follows:

1. Calculate the value $T_0 = \frac{\bar{X}_n - \mu_0}{S_n}\sqrt{n}$.
2. If $|T_0| \geq t_{1-\alpha/2, n-1}$, reject H_0, otherwise accept H_0.

Testing $H_0 : \mu = \mu_0$ against $H_A : \mu > \mu_0$ is analogous:

1. Calculate the value $T_0 = \frac{\bar{X}_n - \mu_0}{S_n}\sqrt{n}$.
2. If $T_0 \geq t_{1-\alpha, n-1}$, reject H_0, otherwise accept H_0.

The null hypothesis $H_0 : \mu = \mu_0$ against $H_A : \mu < \mu_0$ is then rejected in case of $T_0 \leq t_{\alpha, n-1} = -t_{1-\alpha, n-1}$.

- Used in the case of observing two characteristics on one object (e.g. diopters on both eyes, yields of two branches of the same company etc.).
- Consider a random sample $(Y_1, Z_1), (Y_2, Z_2) \ldots, (Y_n, Z_n)^T$ and test $H_0 : \mathrm{E}Y_i - \mathrm{E}Z_i = \mu_0$ (most often $\mu_0 = 0$, i.e. equality of two expected values) against some of the alternative hypotheses introduced above.
- Construct the differences

$$X_1 = Y_1 - Z_1, \ldots, X_n = Y_n - Z_n$$

and if $X_1, \ldots, X_n$ come from normal distribution, we use the one-sample t-test introduced above.

- Consider two independent random samples $(X_1, X_2 \dots, X_m)^T$ from $N(\mu_1, \sigma^2)$ and $(Y_1, Y_2 \dots, Y_n)^T$ from $N(\mu_2, \sigma^2)$, where $\sigma^2 > 0$.
- Denote $\bar{X}$ the sample mean of the sample $(X_1, X_2 \dots, X_m)^T$, $\bar{Y}$ the sample mean of the sample $(Y_1, Y_2 \dots, Y_n)^T$, S_X^2 the sample variance of the sample $(X_1, X_2 \dots, X_m)^T$ and S_Y^2 the sample variance of the sample $(Y_1, Y_2 \dots, Y_n)^T$.
- Under the null hypothesis,

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{(m-1)S_X^2 + (n-1)S_Y^2}} \sqrt{\frac{mn(m+n-2)}{m+n}}$$

has t_{m+n-2} distribution.

- Thus in order to test $H_0 : \mu_1 - \mu_2 = \mu_0$ against $H_A : \mu_1 - \mu_2 \neq \mu_0$, we work as follows:
  1. Calculate the value $T_0 = \frac{\bar{X} - \bar{Y} - \mu_0}{\sqrt{(m-1)S_X^2 + (n-1)S_Y^2}} \sqrt{\frac{mn(m+n-2)}{m+n}}$.
  2. If $|T_0| \geq t_{1-\alpha/2, m+n-2}$, reject H_0, otherwise accept H_0.

Multinomial distribution

Consider a trial. Suppose that only one of the results $A_1, A_2 \ldots, A_k$ can occur in this trial, and denote $p_i = P(A_i)$. Repeat the trial $n-$times and denote X_i the number of the result A_i in these n trials. Then

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k}, \quad \sum_{i=1}^{k} p_i = 1, \quad \sum_{i=1}^{k} x_i = n$$

and the distribution of the vector $(X_1, X_2 \ldots, X_k)^T$ is called multinomial.

We test H_0 : marginal probabilities of the results are $p_1, \ldots, p_k$, against H_A : at least one p_i is different. We work as follows:

1. Calculate the value $\chi^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$.
2. If $\chi^2 > \chi^2_{1-\alpha,k-1}$, reject H_0, otherwise accept H_0.

- Consider a sample $(Y_1, Z_1), (Y_2, Z_2) \dots, (Y_n, Z_n)$, where Y_k takes the values $1, \dots, r$ and Z_k takes the values $1, \dots, c$ for all $k = 1, \dots, n$.
- We test H_0: "Y and Z are independent" against H_A: "Y and Z are not independent".
- Denote n_{ij} the number of couples $(Y_k = i, Z_k = j)$. Then the matrix of dimension $r \times c$ with elements n_{ij} is called contingency table and the elemenst n_{ij} are called joint frequencies. Marginal frequencies are

$$n_{i.} = \sum_j n_{ij}, \quad n_{.j} = \sum_i n_{ij}.$$

- Test of independency is following:
  1. Calculate
  $$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \frac{n_{i.} n_{.j}}{n})^2}{\frac{n_{i.} n_{.j}}{n}}.$$
  2. If $\chi^2 \geq \chi^2_{1-\alpha,(r-1)(c-1)}$, reject H_0, otherwise accept H_0.