Chapter 2

Processing the statistical material

2.1 Plots of frequencies

Definition 2.1 *We have data file of size n:* $x_1, \ldots, x_n$. *Let* a *be minimal value,* b *maximal value:* $x_{min} = a$, $x_{max} = b$.

1. *Interval* $< a, b >$ *is called* **variational field**

2. $x = b - a$ *is called* **variational range**.

3. *Variational field* $< a, b >$ *is decomposed into smaller parts, called* **classes**

$< a_k, b_k >$.

4. The width h *of a class is* $h = b_k - a_k$.

5. *The value* x_k*, which is usually the center of the class, stands for all values belonging to the class, is called* **class mark.**

During the decomposition we will think of this rules:

1. If the data file contains only little different values, all of them will be considered as a class x_k.
If the data file contains plenty of different values, we esteblish classes. The width h can be $h \approx \frac{8}{100} \cdot (b - a)$.
Or we can choose the number of classes k usually between 8 and 20. $k \approx 3,3log(n)$ or $k \approx \sqrt{n}$.

2. If more than one value fall on the border of two classes, we give half to first class and half to second class. If one remains we decide by coin.

3. If there are only little values in the edge classes, we can combine those classes in one wider.

Definition 2.2 *Types of frequencies:*

1. *The number of data in a class is called* **absolute frequency** f_k.

2. a) $\frac{f_k}{n}$ *is* **relative frequency**,

b) $100 * \frac{f_k}{n}$ *is* **percentage relative frequency**.

3. Cumulative absolute frequency

$$F_k = \sum_{j=1}^{k} f_j.$$

4. Cumulative relative frequency R_k

$$R_k = \sum_{j=1}^{k} \frac{f_j}{n} = \frac{F_k}{n}.$$

Remark 2.1 *If we have* r *classes, then*

1.

$$\sum_{k=1}^{r} f_k = n$$

2.

$$F_r = n$$

3.

$$\sum_{k=1}^{r} \frac{f_k}{n} = 1$$

Definition 2.3 Tabel of frequency distribution *is a table where all absolute or relative frequencies are summarized.*

Example 2.1 *The numbers of phone calls during 1 minute recorded at phone centrale is shown in the table. The* $n = 60$.

3,2,2,3,1,1,0,4,2,1
1,4,0,1,2,3,1,2,5,2

3,0,2,4,1,2,3,0,1,2
1,3,1,2,0,7,3,2,1,1
4,0,0,1,4,2,3,2,1,3
2,2,3,1,4,0,2,1,1,5.

The number of phone during 1 min	Absolute frequency	Relative frequency
0	8	0.133
1	17	0.283
2	16	0.266
3	10	0.166
4	6	0,1
5	2	0,033
7	1	0,016
Total	60	1

Table 2.1: **The table of frequency distribution.**

The data file represents n realizations of a random variable X. From law of large numbers, $\frac{f_k}{n}$ estimates probability, that X falls in k-th class, thus $p_k =$

$P(a_k \leq X \leq b_k) \approx \frac{f_k}{n}$.

Definition 2.4 *Types of frequency plots:*

1. **Histogram** *contains of rectangles with base* $< a_k, b_k >$ *on axis* x *and with hight coresponding to apropriate frequency. Histogram of relative freguencies aproximate density of random variable* X.

2. **Polygon** *is pointed line with points at* (x_k, f_k) *(for absolute frequency.*

3. **Ogive** *is polygon for cumulative relative frequencies, it aproximates disribution function of* X.

2.2 Characteristics of location

Definition 2.5 *Consider data file with values* $x_1, x_2, ..., x_n$, *which is devided into* r *classes.*

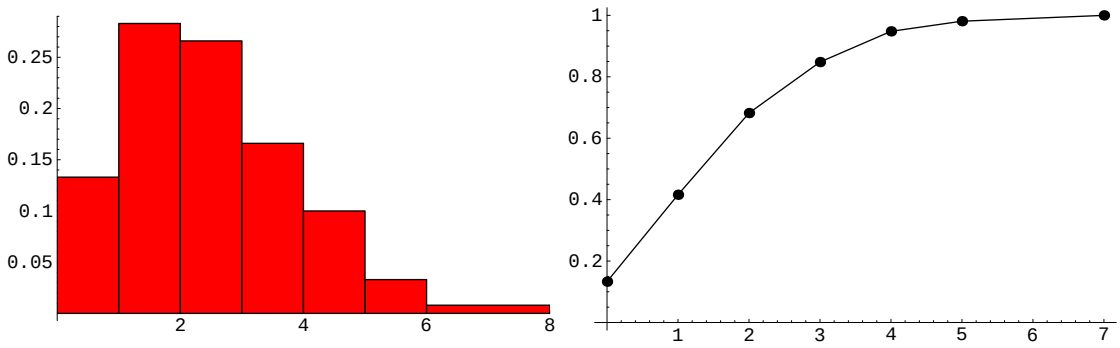

Figure 2.1: Histogram and ogive for data file from Example 2.1

1. Arithmetical average $\bar{X}$

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{1}{n}\sum_{i=1}^{r} f_i x_i. \tag{2.1}$$

2. Geometrical average $\bar{X}_g$

$$\bar{X}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_1} \tag{2.2}$$

3. Harmonical average $\bar{X}_h$

$$\bar{X}_h = \frac{1}{A}, where A = \frac{1}{n}\sum_{k=1}^{n}\frac{1}{x_k} = \frac{1}{n}\sum_{i=1}^{r}\frac{f_i}{x_i}. \tag{2.3}$$

In Formulas 2.1, 2.3 are two expressions. First for unsorted file, second for soted one.

Theorem 2.1

$$\bar{X}_h \leq \bar{X}_g \leq \bar{X}.$$

The sorted data file

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

consists of $X_{(1)}$ the smallest value, $X_{(2)}$ second smallest value, ... , $X_{(n)}$ the highest value.

Definition 2.6 Median *is designated with the dependence on number of data. If n is odd, then the median $\tilde{x}$ is the middle value*

$$\tilde{x} = X_{\left(\left[\frac{n}{2}\right]+1\right)}.$$

If n is even, then the median $\tilde{x}$ is the mean of two middle values

$$\tilde{x} = \frac{X_{([\frac{n}{2}])} + X_{([\frac{n}{2}]+1)}}{2}.$$

Median is special case of *sample kvantil*.

Definition 2.7 Modus *is the value with highest absolute frecuency. It need not be defined unequivocally.*

2.3 Characteristics of variability

Definition 2.8 *Characteristics of variability:*

1. Variance *of data file is*

$$S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{r} f_i(x_i - \bar{X})^2. \tag{2.4}$$

2. Standard deviation *is*

$$\sqrt{S^2} = S \geq 0. \tag{2.5}$$

3. Average deviation *is*

$$\bar{d} = \frac{1}{n}\sum_{k=1}^{n}|x_k - \bar{X}| = \frac{1}{n}\sum_{i=1}^{r} f_i|x_i - \bar{X}|. \tag{2.6}$$

4. Variational coefficient *is*

$$v = \frac{S}{\bar{X}}. \tag{2.7}$$

Remark 2.2 *The variance is defined by 2.4, but for computation we can also use*

$$S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k^2) - \frac{n}{n-1}\bar{X}^2 = \frac{1}{n-1}\sum_{i=1}^{r} f_i x_i^2 - \frac{n}{n-1}\bar{X}^2. \tag{2.8}$$

Remark 2.3 *The data file represents n realzations of a random variable X. All characteristics of location aproximates expectation of X. The variance of data file aproximates the variance of X.*

Chapter 3

Random sample

Types of random samples

a) *Simple random sample with returning*

b) *Simple random sample without returning*

c) *Stratified random sample* deviding the space into disjunct sets and making the random sample within each set.

d) *Systematic random sample* every k-th element.

Range of random sample

a) *Small*: $n < 30$.

b) *Big*: $n \geq 30$.

We will be interested only in simple random sample with returning. It is:

Definition 3.1 *The statistical sample* $\mathbf{Z}$*, represents a random variable* X*.* **Random sample** *from distribution of* X*, is* n *independent realizations of* X*, which are given by independent random variables* $X_1, X_2, \ldots, X_n$*, with same distribution as* X*.*

Definition 3.2 *The charakteristics of statistical sample* $\mathbf{Z}$ *(random variable* X*) will be called* **theoretical***. Charakteristics obtained from empirical random sample is called* **sample)***.*

Definition 3.3 *Let* $X_1, \ldots, X_n$ *is a random sample from distribution with expectation* μ *and finite variance* σ^2*.*

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

$\bar{X}$ is **sample mean** *and S^2 is* **sample variance**.

Theorem 3.1

$$\mathbb{E}\bar{X} = \mu, \qquad \mathrm{Var}\bar{X} = \frac{\sigma^2}{n}, \qquad \mathbb{E}S^2 = \sigma^2.$$

Theorem 3.2 Strong law of large numbers

$$\bar{X} \to \mu \qquad \text{almost surely.}$$

Convergence almost surely means, that exists only a set ($A \subset \Omega$) of probability 0 (P(A)=0), for which the expression does not converge.

Theorem 3.3 Random sample from normal distribution *Let $X_1, \ldots, X_n$ is random sample from N(μ, σ^2), where $\sigma^2 > 0$. Then*

- $\bar{X} \sim$ *N($\mu, \frac{\sigma^2}{n}$).*
- *If $n \geq 2$, then $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.*
- *If $n \geq 2$, then $\bar{X}$ and S^2 are independent.*
- *If $n \geq 2$, then $\frac{\bar{X}-\mu}{S}\sqrt{n} \sim t_{n-1}$.*

Chapter 4

Parametres estimates

Point estimates of mean and variance: Theorem 3.1 says that
$\bar{X}$ is unbiased estimate of mean μ ($\mathbb{E}\bar{X} = \mu$),
S^2 is unbiased estimate of σ^2.

Interval estimate
(Coeficient of reliability) $q = 1 - \alpha$. α is usually chosen 0.05, 0.01.

Definition 4.1 *Let B_1, B_2 be such that for $\alpha \in (0,1)$ holds*

$$P(B_1 \leq \beta \leq B_2) = 1 - \alpha.$$

Then the interval $[B_1, B_2]$ is called **confidence interval** *for parameter β with reliability $1 - \alpha$.*
Let B_2 be such that for $\alpha \in (0,1)$ holds

$$P(\beta \leq B_2) = 1 - \alpha.$$

Then the interval B_2 is called **upper confidence bound** *for parameter β with reliability $1 - \alpha$.*
Let B_1 be such that for $\alpha \in (0,1)$ holds

$$P(\beta \geq B_1) = 1 - \alpha.$$

Then the interval B_1 is called **lower confidence bound** *for parameter β with reliability $1 - \alpha$.*

4.1 Confidence intervals for parametrs of normal distr.

Let $X_1, \ldots, X_n$ be random sample from $\mathrm{N}(\mu, \sigma^2)$, parameter $\sigma^2 > 0$ is not known. Then by Theorem 3.3

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

thus by definition of critical values of students distr. is

$$P\left[t_{n-1}(\alpha/2) \leq \frac{\bar{X} - \mu}{S}\sqrt{n} \leq t_{n-1}(1-\alpha/2)\right] = 1 - \alpha,$$

By reordering we get bothsided confidence interval for mean μ of normal distr. with reliability $1 - \alpha$

$$\left[\bar{X} - t_{n-1}(1-\alpha/2)\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1}(1-\alpha/2)\frac{S}{\sqrt{n}}\right]. \tag{4.1}$$

confidence interval for variance σ^2 can be derived similarly.

$$(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}.$$

$$P\left[\chi^2_{n-1}\left(\frac{\alpha}{2}\right) \leq (n-1)S^2/\sigma^2 \leq \chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)\right] = 1-\alpha,$$

By reordering we get bothsided confidence interval for variance σ^2 of normal distr. with reliability $1-\alpha$

$$\left[\frac{S^2(n-1)}{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)}, \frac{S^2(n-1)}{\chi^2_{n-1}\left(\frac{\alpha}{2}\right)}\right]. \tag{4.2}$$

4.2 Confidence interval for mean by use of CLT

If the variables do not have the normal distribution, we can not use the previous. But if there are more vaiables, at least 20, we can use CLT.

Let $X_1, \ldots, X_n$ be random sample from distribution with finite mean μ and finite variance σ^2. Then by Central limit theorem

$$\frac{\bar{X} - \mu}{S}\sqrt{n} \to_{n\to\infty} \Phi \sim N(0,1)$$

has asymptoticaly normed normal distribution. By the definition of critical value of the normed normal distribution is

$$P\left[-u(1-\frac{\alpha}{2}) \leq \frac{\bar{X} - \mu}{S}\sqrt{n} \leq u(1-\frac{\alpha}{2})\right] = 1-\alpha,$$

by reordering we have bothsided confidence interval for mean μ with reliability $1-\alpha$

$$\left[\bar{X} - u(1-\frac{\alpha}{2})\frac{S}{\sqrt{n}}, \bar{X} + u(1-\frac{\alpha}{2})\frac{S}{\sqrt{n}}\right]. \tag{4.3}$$

Chapter 5

Parametrical tests

Hypothesis: H_0 (zero) against alternative H_1.
Assume that distribution of the random variable depends on parametr θ.
The zero hypothesis is then

$$H_0 : \theta = \theta_0.$$

Alternative hypothesis can be either both sided alternative

$$H_1 : \theta \neq \theta_0,$$

or one sided alternative

$$H_1 : \theta > \theta_0 \text{ or } H_1 : \theta < \theta_0.$$

Two mistakes are possible:
We reject H_0, but H_0 is true - *error of first kind.*
We do not reject H_0, but H_0 is false - *error of second kind.*
We want both mistake to keep down!

Test:
Construct statistic T from random sample $X_1, X_2, ..., X_n$.
If T is in critical range W, we reject H_0.
If T is not in critical range W, we do not reject H_0.
We can control error of first kind only. We set the probability of this error to be $\alpha \in (0, 1)$. Usually $\alpha = 0.05$ or $\alpha = 0.01$. α is called significance level of test.

Remark 5.1 *Nowadays the statistical software (Statistica, S+, SAS, R, Excel) gives reached level (called P-value, significance value). It is the smallest level for which we still reject* H_0.

Thus if $\alpha = 0.05$ *is chosen and P-value is less than 0.05 (or equal), then we reject* H_0 *with significance level* $\alpha = 0.05$. *If P-value is greater than 0.05, then we do not reject* H_0 *with significance level* $\alpha = 0.05$.

5.1 One sample t test

Let $X_1, \ldots, X_n$, be random sample from $\mathrm{N}(\mu, \sigma^2)$, where $n > 1$. Parametr $\sigma^2 > 0$ is not known. We want to test

$$H_0 : \mu = \mu_0,$$

where μ_0 is given number, against the alternative

$$H_1 : \mu \neq \mu_0$$

. Hypothesis H_0 is rejected if $\bar{X}$ is rather far from μ_0. Under H_0 the statistic T has

$$T = \frac{(\bar{X} - \mu_0)\sqrt{n}}{S} \sim t_{n-1}$$

students distribution with n-1 degree of freedom. From definition of the critical values, we have

$$P\left[|T| \geq t_{n-1}(1 - \alpha/2)\right] = \alpha.$$

Thus H_0 is rejected with significance level α, if

$$|T| \geq t_{n-1}(1 - \alpha/2).$$

p-value is computed from the distribution function of t:

$$p = 2(1 - F_{t_{n-1}}(|T|))$$

In case of one sided alternative $H_1 : \mu > \mu_0$, resp. $H_1 : \mu < \mu_0$ is H_0 rejected, if

$$T \geq t_{n-1}(1-\alpha), \quad \text{resp.} \quad T \leq -t_{n-1}(1-\alpha).$$

p-value of the one sided test is:

$$p = 1 - F_{t_{n-1}}(T), \quad \text{resp.} \quad p = F_{t_{n-1}}(T)$$

5.2 Test about variance of normal distribution.

Let $X_1, \ldots, X_n$, be random sample from $\mathrm{N}(\mu, \sigma^2)$, where $n > 1$. We want to test

$$H_0 : \sigma^2 = \sigma_0^2,$$

where σ_0^2 is given number, against the alternative

$$H_1 : \sigma^2 \neq \sigma_0^2$$

. H_0 is rejected, if S^2 is rather far from σ_0^2. Under H_0 the statistic T has

$$T = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

χ^2 distribution with o n-1 degree of freedom. From definition of the critical values, we have

$$P\left[\chi^2_{n-1}\left(\frac{\alpha}{2}\right) \leq (n-1)S^2/\sigma_0^2 \leq \chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)\right] = 1-\alpha,$$

Thus H_0 is rejected with significance level α, if

$$T \leq \chi^2_{n-1}\left(\frac{\alpha}{2}\right) \quad \text{or} \quad T \geq \chi^2_{n-1}\left(1-\frac{\alpha}{2}\right).$$

p-value is computed from the distribution function of χ^2:

$$p = \min\left(2(1 - F_{\chi^2_{n-1}}(T)), 2(F_{\chi^2_{n-1}}(T))\right)$$

In case of one sided alternative $H_1 : \sigma^2 > \sigma_0^2$, resp. $H_1 : \sigma^2 < \sigma_0^2$ is H_0 rejected, if

$$T \geq \chi^2_{n-1}(1-\alpha), \quad \text{resp.} \quad T \leq \chi^2_{n-1}(\alpha).$$

p-value of the one sided test is:

$$p = 1 - F_{\chi^2_{n-1}}(T), \quad \text{resp.} \quad p = F_{\chi^2_{n-1}}(T)$$

5.3 Paired t test

Consider random sample $(Y_1, Z_1), (Y_2, Z_2), \ldots, (Y_n, Z_n)$ from two-dimensional normal distribution with expectation (μ_1, μ_2). We want to test

$$H_0 : \mu_1 - \mu_2 = \Delta$$

against alternative $H_1 : \mu_1 - \mu_2 \neq \Delta$, where Δ is given number (usually $\Delta = 0$). We set

$$X_1 = Y_1 - Z_1, X_2 = Y_2 - Z_2, \ldots, X_n = Y_n - Z_n.$$

The variables $X_1, X_2, ..., X_n$ are independent. Assume, that $X_i \sim N(\mu, \sigma^2), i = 1, 2, \ldots, n$, $\mu = \mu_1 - \mu_2$. Thus the task is transferred to one sample t test. From variables $X_1, X_2, ..., X_n$ we calculate $\bar{X}$ and S^2. Then H_0 is rejected with significance level α, if

$$|T| = \left| \frac{(\bar{X} - \Delta)\sqrt{n}}{S} \right| \geq t_{n-1}(1 - \alpha/2).$$

p-value is:

$$p = 2(1 - F_{t_{n-1}}(|T|))$$

Paired t test is used in situations, when we have, for every object, from n measured objects, 2 measuranments. The objects are independent but not measuranments on one object. The paired t test is used for example, when we test effectiveness of the medicanment on n patients, and Y_i are measuranments before and Z_i after medicanment is taken.

5.4 Two sample t test

Let $X_1, X_2, \ldots, X_n$ be random sample from $N(\mu_1, \sigma^2)$ and $Y_1, Y_2, \ldots, Y_m$ sample from $N(\mu_2, \sigma^2)$. Assume that the two random samples are independent. Assume that $n \geq 2, m \geq 2, \sigma^2 > 0$ and σ^2 is not known. We want to test

$$H_0 : \mu_1 - \mu_2 = \Delta$$

against $H_1 : \mu_1 - \mu_2 \neq \Delta$, where Δ is given number (usually $\Delta = 0$). Denote $\bar{X}, S_X^2$ and $\bar{Y}, S_Y^2$ charakteristics of appropriate samples. Then H_0 is rejected with significance level α if

$$\mid T \mid = \left| \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{(n-1)S_X^2 + (m-1)S_Y^2}} \cdot \sqrt{\frac{nm(n+m-2)}{n+m}} \right| \geq t_{n+m-2}(1 - \alpha/2).$$

p-value is:

$$p = 2(1 - F_{t_{n+m-2}}(|T|))$$

two sample t test is used, when n patients try medicanment A and other m patients try medicanment B.

Often one uses two sample instead of paired and vica verca. Remember that two sample can be used only when independence of samples is satisfied.

Assumptions: Independence - most important.
Normality - Greater sample - Since CLT - OK
Small sample - nonparametric tests must be used.
Homogeneity of variances in twosamle t test - is tested by the following test. If it is not satisfied use weighted least squares methods.

5.5 Test about homogeneity of two variances

Let $X_1, X_2, \ldots, X_n$ is a sample from $N(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_m$ is a sample from $N(\mu_2, \sigma_2^2)$. Assume independance of samples and that $n \geq 2, m \geq 2, \sigma_1^2 > 0, \sigma_2^2 > 0$. We test the

$$H_0 : \sigma_1^2 = \sigma_2^2$$

against

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

. Since S_X^2 is unbiased estimate of σ_1^2 and S_Y^2 is unbiased estimate of σ_2^2, we can expect, that under H_0 the fraction $\frac{S_X^2}{S_Y^2}$ will be close to one. Thus H_0 is rejected, if

$$\frac{S_X^2}{S_Y^2} \leq k_1 \quad \text{or} \quad \frac{S_X^2}{S_Y^2} \geq k_2,$$

while

$$k_1 = F_{n-1,m-1}(\frac{\alpha}{2}) = \frac{1}{F_{m-1,n-1}(1-\alpha/2)}, \quad k_2 = F_{n-1,m-1}(1-\frac{\alpha}{2}),$$

where $F_{n-1,m-1}(\alpha/2)$ is critical value of Fisher-Snedecor distribution with n-1 and m-1 degree of freedom. p-value is:

$$p = \min\left(2(1 - F_{F_{n-1,m-1}}(\frac{S_X^2}{S_Y^2})), 2(F_{F_{n-1,m-1}}(\frac{S_X^2}{S_Y^2}))\right)$$

5.6 Test about mean with use of CLT

We do not assume normality, but we need at least 20 or 30 data.

Let $X_1, \ldots, X_n$ be random sample from a distribution with finite mean μ and finite variance σ^2. We test

$$H_0 : \mu = \mu_0,$$

where μ_0 is given number, against alternative

$$H_1 : \mu \neq \mu_0.$$

Hypothesis H_0 is rejected, if $\bar{X}$ rather far from μ_0. By CLT under H_0 T has asymptoticaly

$$T = \frac{\bar{X} - \mu_0}{\sigma_0}\sqrt{n} \rightarrow_{n\rightarrow\infty} \Phi \sim N(0,1)$$

normalized narmal distribution.

Again by definition of critical values is asyptoticaly

$$P\left[|T| \leq u(1 - \frac{\alpha}{2})\right] = 1 - \alpha.$$

Thus H_0 is rejected with significance level α, if

$$|T| \geq u(1 - \frac{\alpha}{2}).$$

p-value is computed from distribution function of normalised normal distribution:

$$p = 2(1 - \Phi(|T|))$$

In case of one sided alternative $H_1 : \mu > \mu_0$, resp. $H_1 : \mu < \mu_0$ is H_0 rejected, if

$$T \geq u(1 - \alpha), \quad \text{resp.} \quad T \leq -u(1 - \alpha).$$

p-value of the one sided test is:

$$p = 1 - \Phi(T), \quad \text{resp.} \quad p = \Phi(T)$$

In case when σ_0^2 is not known, we use unbiased estimate S^2 in computation of T.

Chapter 6

ANOVA

6.1 One-way ANOVA

This is extension of twosample t-test for more than 2 samples. Assume I independent samples,

$$Y_{11}, ..., Y_{1n_1} \quad \text{from } N(\mu_1, \sigma^2)$$

$$\vdots$$

$$Y_{I1}, ..., Y_{In_I} \quad \text{from } N(\mu_I, \sigma^2).$$

We will test $H_0 : \mu_1 = \ldots = \mu_I$ against, that exist at least two means which are

not equal.

We can write the model - H_1:

$$Y_{ij} = \mu + \alpha_i + e_{ij},$$

where $\mu + \alpha_i = \mu_i$ and $e_{ij} \sim \mathrm{N}(0, \sigma^2)$ is the error. H_0 can be rewritten to simplified model:

$$Y_{ij} = \mu + e_{ij}.$$

Test is done in following way. Denote

$$\overline{Y}_i = \frac{Y_{i1} + ... + Y_{in_i}}{n_i} \quad \text{pro} \quad i = 1, ..., I$$

$$\overline{Y} = \frac{\sum_i \sum_j Y_{ij}}{n},$$

where $n = n_1 + \ldots + n_I$. Total sum of squares S_T is total square error under H_0.

$$S_T = \sum_i \sum_j (Y_{ij} - \overline{Y})^2 = \sum_i \sum_j Y_{ij}^2 - n\overline{Y}^2.$$

Residual sum of square S_e is square error under H_1.

$$S_e = \sum_i \sum_j (Y_{ij} - \overline{Y}_i)^2 = \sum_i \sum_j Y_{ij}^2 - \sum_i n_i \overline{Y}_i^2.$$

$S_A = S_T - S_e$ is sum of squares contained in difference of samples. If S_A is small, then the models are simillar and H_0 will not be rejected.

Under H_0

$$F_A = \frac{(n-I)S_A}{(I-1)S_e} \sim F_{I-1,n-I}$$

has F distribution with $I-1$ and $n-I$ degree of freedom. Thus H_0 is rejected with significance α if

$$F_A \geq F_{I-1,n-I}(1-\alpha).$$

Characteristic $s^2 = S_e/(n-I)$ is called residual variance and it is an unbiased estimator of the true variance σ^2.

Assumption Independence, normality (can be broken for a lot observations) Homogeneity of all variances - Bartlett test. (Statistica - Anova - Assumtions)

Variability	sum of squares S	degree of freedom f	fraction S/f	F
differences	S_A	$f_A = I - 1$	S_A/f_A	F_A
residual	S_e	$f_e = n - I$	S_e/f_e	-
total	S_T	$f_t = n - 1$	-	-

Table 6.1: ANOVA table

Remark 6.1 *One can think of applying a set of $I(I-1)/2$ two sample t-tests, instead of one way anova. But if every test would have significance α, the common significance of all tests together would me much higher.*

In the case of rejection H_0, it is usually needed to decide which samples are different. **Tukey multiple comparison method** (Statistica - Anova - Post-hoc).

6.2 Anova - two ways

Model:

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \quad \text{kde} \quad i = 1, \ldots, I, \quad j = 1, \ldots, J \tag{6.1}$$

where μ, α_i for $i = 1, \ldots, I$ and β_j for $j = 1, \ldots, J$ are not known parameters and $e_{ij} \sim \mathrm{N}(0, \sigma^2)$ are errors. This means that Y_{ij} depend both on column and row. Furthermore in each row we have same number of event. Two parameters are needless therefore we set

$$\sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0.$$

We want to test $H_0 : \alpha_1 = \ldots = \alpha_I = 0$ (e.g. data does not depend on rows), thus the simplified H_0 model is one way anova:

$$Y_{ij} = \mu + \beta_j + e_{ij}.$$

(Statistica - Anova - Main Effects Anova)

In the case of rejection H_0, it is usually needed to decide which samples are different. **Tukey multiple comparison method** (Statistica - Anova - Post-hoc).

Remark 6.2 *More models:*
More data for each group (Statistica - Anova - Main Effects Anova)

$$Y_{ijp} = \mu + \alpha_i + \beta_j + e_{ijp}, \quad kde \quad i = 1, \ldots, I, \quad j = 1, \ldots, J, \quad p = 1, \ldots, P.$$

Variability	sum of squares S	degree freedom f	ratio S/f	F
rows	S_A	$f_A = I - 1$	S_A/f_A	F_A
columns	S_B	$f_B = J - 1$	S_B/f_B	F_B
residual	S_e	$f_e = n - I - J - 1$	S_e/f_e	-
celkov	S_T	$f_t = n - 1$	-	-

Table 6.2: Anova Table two-way anova

(Statistica - Anova - Main Effects Anova)

Interactions (Statistica - Anova - Factorial Anova):

$$Y_{ijp} = \mu + \alpha_i + \beta_j + \lambda_{ij} + e_{ijp}.$$

Repeated measuranments - Here we assume dependence between observations in k-th level $(Y_{ijp1}, ..., Y_{ijpK})$ *- (Statistica - Anova - Repeated Measures Anova):*

$$Y_{ijpk} = \mu + \alpha_i + \beta_j + \gamma_k + \lambda_{ij} + e_{ijpk}.$$

Chapter 7

Nonparametrics

In Statistica - nonparametrics statistics we can find nonparametric variations of tests which were described in parametric chapters. Instead of expectations we compare the whole distributions, but the tests are mainly sensitive for medians only.

Two sample t-test $\sim$ Comparing two independent samples - usually Mann-Whitney U test or Wilcoxon test is preferred.

One way ANOVA $\sim$ Comparing multiple independent samples - usually Kruskal-Wallis test is preferred. Multiple comparison can be done here also.

Paired t-test $\sim$ Comparing two dependent samples - usually Wilcoxon test is preferred. When only data of character +1, -1 (better, worse) is available, then sign test is appropriate.

2 way ANOVA $\sim$ Comparing multiple dependent samples - Friedmann test is preferred.

Correlation $\sim$ Spearman correlation is preferred.

Chapter 8

Correlation analysis

Independence $\Rightarrow$ the uncorrelatness $\rho = 0$.
Thus if $H_0 : \rho = 0$ is rejected, then we can reject also the hypothesis of unbiasedness.

8.1 Sample correlation coefficient

We have random sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ from two dimensional distribution. The correlation coefficient is

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}X \ \mathrm{Var}Y}}.$$

For estimates of Var X and Var Y we use sample variances

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

$\mathbb{E}S_X^2 = \mathrm{Var}X$ and $\mathbb{E}S_Y^2 = \mathrm{Var}Y$. Similarly we define sample covariance

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}),$$

here $\mathbb{E}S_{XY} = \mathrm{Cov}(X, Y)$. Thus if $S_X^2 > 0$ and $S_Y^2 > 0$, we define sample correlation coefficient r as

$$r = \frac{S_{XY}}{\sqrt{S_X^2 S_Y^2}}.$$

From Schwarz unequality $-1 \leq r \leq 1$.

r is not unbiased.

For $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ - twodimensional random sample from normal distribution and Var $X > 0$, Var $Y > 0$, $|\rho| < 1$, we have that

$$\mathbb{E}r = \rho - \frac{1-\rho^2}{n} + o(n^{-1}),$$

where $o(n^{-1})$ denotes function $f(n)$, for which $\lim_{n\to\infty} \frac{f(n)}{n} = 0$.

We test now $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$. Under H_0 and assumption of normal random samples has

$$T = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \sim t_{n-2}$$

Students distribution with $n-2$ degree of freedom. Thus H_0 is rejected with significance level α, in case that

$$|T| \geq t_{n-2}(1-\alpha/2).$$

Here the normality is important assumption. If normality is not satisfied use nonparametric Spearman correlational coefficient.

Chapter 9

Linear regression

9.1 Liner regression with one explanatory variable

Regression model

$$Y = f(x)$$

explains dependence of Y on values x through f. The aim of regression is to find function f, from n observed pairs

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$$

where x_i are independent values, explanatory variable x and y_i are dependent values, ascribed variable Y. Assume that y_i are measured with error e_i.

For point estimates no further assumption is needed.
For interval estimates and test we assume normality of error - N(0, σ^2).

Thus we have n independent equations

$$Y_i = f(x_i) + e_i, \quad i = 1, 2, \ldots, n.$$

Linear regression

$$f(x) = \beta_0 + \beta_1 \cdot x.$$

We want to estimate parameters β_0 and β_1. This is done by least square method. We search β_0 and β_1 for which the sum of squares of errors is minimal. Thus we are looking for minimum of

$$g(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 \cdot x_i))^2.$$

Thus we solve the set of equations

$$\frac{\delta g(\beta_0, \beta_1)}{\delta \beta_0} = 0, \quad \frac{\delta g(\beta_0, \beta_1)}{\delta \beta_1} = 0.$$

The result is

$$b_1 = \frac{\sum (x_i - \overline{x}) \cdot Y_i}{(x_i - \overline{x})^2} = \frac{\sum x_i Y_i - n\overline{x}\overline{Y}}{\sum x_i^2 - n\overline{x}^2}, \quad b_0 = \overline{Y} - b_1 \cdot \overline{x}, \tag{9.1}$$

where $\overline{x} = \frac{1}{n}\sum x_i$ and $\overline{Y} = \frac{1}{n}\sum Y_i$. Estimates b_0, b_1 are best unbiased estimates, e.g. b_0, b_1 are unbiased ($\mathbb{E}b_0 = \beta_0, \quad \mathbb{E}b_1 = \beta_1$) and have the smallest variance from all unbiased estimators.

Minimum of g

$$S_e = g(b_0, b_1) = \sum (Y_i - (b_0 + b_1 \cdot x_i))^2 = \sum Y_i^2 - b_0 \sum Y_i - b_1 \sum x_i Y_i$$

is called **residual sum of squares**. The estimator of variance of errors σ^2 is

$$s^2 = \frac{S_e}{n-2}.$$

Total sum of squares

$$S_T = \sum (Y_i - \overline{Y})^2$$

express total square error of regression model.

The appropriateness of the model is expressed in coefficient of determination

$$R^2 = 1 - \frac{S_e}{S_T} = \frac{S_T - S_e}{S_T},$$

which express which part of total error S_T is explained by the regression model. (S_e contains, what the regression model does not explain). It can be calculated also by

$$R^2 = \frac{\sum (\widehat{Y}_i - \overline{Y})^2}{\sum (Y_i - \overline{Y})^2},$$

where $\widehat{Y}_i = \widehat{f}(x_i) = b_0 + b_1 \cdot x_i$ is regression estimate in x_i. As close is R^2 to 1, as better the model is.

The most common question is, if the model can be simplified, so that Y_i do not depend on x_i. We test

$$H_0 : \beta_1 = 0 \quad \text{against} \quad H_1 : \beta_1 \neq 0$$

Under H_0 has

$$T = \frac{b_1}{s} \cdot \sqrt{\sum x_i^2 - n\overline{x}^2} \sim t_{n-2}$$

students distribution with $n-2$ degree of freedom. Thus if $|T| \geq t_{n-2}(1-\alpha/2)$ we reject H_0 with significance α. If this H_0 is rejected we confirm the linear dependence of Y_i on x_i, which is masked by random errors e_i.

Confidence interval By standard approach we make confidence interval of β_1 with reliability $1-\alpha$:

$$\left(b_1 - \frac{t_{n-2}(1-\alpha/2)s}{\sqrt{\sum x_i^2 - n\overline{x}^2}}, b_1 + \frac{t_{n-2}(1-\alpha/2)s}{\sqrt{\sum x_i^2 - n\overline{x}^2}} \right).$$

Often we need confidence interval for $\beta_0 + \beta_1 x$:

$$\left(b_0 + b_1 x - t_{n-2}(1-\alpha/2)s\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum x_i^2 - n\overline{x}^2}}, b_0 + b_1 x + t_{n-2}(1-\alpha/2)s\sqrt{\frac{1}{n} + \frac{}{\sum}} \right.$$

This interval cover the value $\beta_0 + \beta_1 x$ with probability $1-\alpha$. Such intervals constructed for all $x \in [\min x_i, \max x_i]$, is called **belt of reliability around regression function**.

Example 9.1 *The number of hours in use in month* (x_i) *the expenses* (Y_i).

x_i	275	350	250	325	375	400	300
Y_i	149	170	140	164	192	200	165

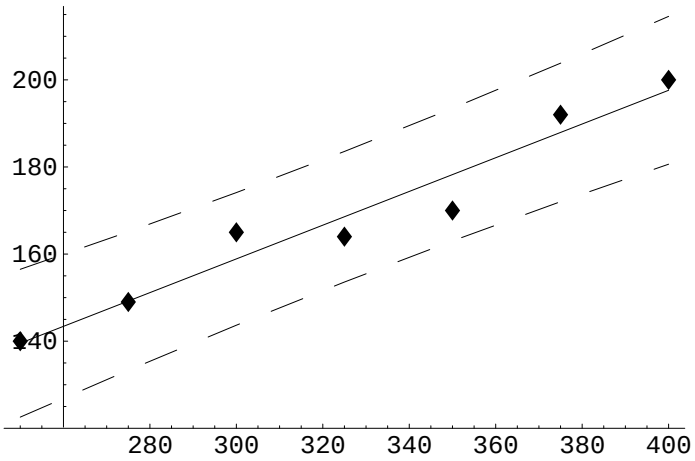

Figure 9.1:

Interpretation of model: Absolute term b_0 *- estimates fix expenses, independent of length of use. Linear term* $b_1 x$ *estimates variable expenses per hour of use.*

9.2 Linear regression with more explanatory variables

We assume model:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k. \tag{9.2}$$

For n observations, we have n equations with *k+1* unknowns:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + e_i, \quad \text{where } i = 1, 2, ..., n, \tag{9.3}$$

Here e_i are random errors. For point estimates no further assumption is needed. For interval estimates and test we assume normality of error - N(0, σ^2).

In matrix form we have

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e},$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & X_{11} & \dots & X_{1k} \\ 1 & X_{21} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \dots & X_{nk} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}. \tag{9.4}$$

The aim is to estimate parameters $\beta_0, \beta_1, \beta_2, ..., \beta_k$. Again by *least square method*. We minimize

$$g(\beta_0, \beta_1, ..., \beta_k) = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik}))^2. \tag{9.5}$$

Thus we set the partial derivative to be zero

$$\frac{\partial g(\beta_0, \beta_1, ..., \beta_k)}{\partial \beta_0} = 2 \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik}))(-1) = 0$$

a

$$\frac{\partial g(\beta_0, \beta_1, ..., \beta_k)}{\partial \beta_j} = 2 \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik}))(-X_{ij}) = 0,$$

where $j = 1, 2, ..., k$.

$$\vdots$$

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} X_{i1} + \beta_2 \sum_{i=1}^{n} X_{i2} + ... + \beta_k \sum_{i=1}^{n} X_{ik} = \sum_{i=1}^{n} Y_i$$

$$\beta_0 \sum_{i=1}^{n} X_{i1} + \beta_1 \sum_{i=1}^{n} X_{i1}^2 + \beta_2 \sum_{i=1}^{n} X_{i2}X_{i1} + ... + \beta_k \sum_{i=1}^{n} X_{ik}X_{i1} = \sum_{i=1}^{n} Y_i X_{i1}$$

$$\vdots$$

$$\beta_0 \sum_{i=1}^{n} X_{ik} + \beta_1 \sum_{i=1}^{n} X_{ik}X_{i1} + \beta_2 \sum_{i=1}^{n} X_{ik}X_{i2} + ... + \beta_k \sum_{i=1}^{n} X_{ik}^2 = \sum_{i=1}^{n} Y_i X_{ik}. \quad (9.6)$$

Solving this set of equations gives estimates $b_0, b_1, ..., b_k$ of parameters $\beta_0, \beta_1, \beta_2,$ In matrix form

$$(\mathbf{X}^T\mathbf{X}) \cdot \beta = \mathbf{X}^T\mathbf{Y}. \quad (9.7)$$

If $(\mathbf{X}^T\mathbf{X})$ is regular (e.g. there exists an inverse $(\mathbf{X}^T\mathbf{X})^{-1}$), then the estimator of parameters $\beta = \beta_0, \beta_1, \beta_2, ..., \beta_k$ is

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \quad (9.8)$$

The minimum of g is called **residual sum of squares**

$$S_e = g(\mathbf{b}) = \sum (Y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + + b_k x_{ik}))^2 = \sum (Y_i - \widehat{Y}_i)^2,$$

where $\widehat{Y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik}$ is regression estimate of value Y_i. The estimator of variance of errors σ^2 is $s^2 = \frac{S_e}{n-k-1}$. s^2 is called residual variance.

Total sum of squares

$$S_T = \sum (Y_i - \overline{Y})^2$$

express total square error of regression model.

The appropriateness of the model is expressed in coefficient of determination

$$R^2 = 1 - \frac{S_e}{S_T} = \frac{S_T - S_e}{S_T},$$

which express which part of total error S_T is explained by the regression model. (S_e contains, what the regression model does not explain). It can be calculated also by

$$R^2 = \frac{\sum (\widehat{Y}_i - \overline{Y})^2}{\sum (Y_i - \overline{Y})^2},$$

where $\widehat{Y}_i = \widehat{f}(x_i) = b_0 + b_1 \cdot x_i$ is regression estimate in x_i. As close is R^2 to 1, as better the model is.

Confidence interval with reliability $1-\alpha$ for parameters β_i is interval

$$\left(b_i - t_{n-k-1}(1-\alpha/2)\cdot s\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}, b_i + t_{n-k-1}(1-\alpha/2)\cdot s\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}\right), \tag{9.9}$$

where $(\mathbf{X}^T\mathbf{X})^{-1}_{ii}$ is element of matrix $(\mathbf{X}^T\mathbf{X})^{-1}$, in i-th line and i-th column.

The most common question is if we can simplify the model so that the values Y_i do not depend on x_{ij} for certain j. We test

$$H_0 : \beta_j = 0 \quad \text{against} \quad H_1 : \beta_j \neq 0$$

Under H_0

$$T = \frac{b_j}{s\cdot\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}} \sim t_{n-k-1} \tag{9.10}$$

has students distribution with $n-k-1$ degree of freedom. Thus if $|T| \geq t_{n-k-1}(1-\alpha/2)$ we reject H_0 with significance α.

Sometimes we ask, if more than one explanatory variable can be release. It is not possible to use two previous tests, because its common significance level would not be α.

We test

$$H_0 : \beta_{j_1} = \beta_{j_2} = \ldots = \beta_{j_l} = 0, \quad 1 \leq j_1, \ldots, j_l \leq k$$

against alternative that simplified model is not true (e.g. that at least one $\beta_{j_i} \neq 0$). Number l is the number of parameters to be released. Matrix form of the simplified model is

$$\mathbf{Y} = \widetilde{\mathbf{X}}\widetilde{\beta} + \widetilde{\mathbf{e}},$$

where matrix $\widetilde{X}$ is constructed from X by releasing of columns appropriate to $\beta_{j_1}, \beta_{j_2}, \ldots, \beta_{j_l}$. Vector $\widetilde{\beta}$ is constructed from $\beta$by releasing $\beta_{j_1}, \beta_{j_2}, \ldots, \beta_{j_l}$. Similarly $\widetilde{e}$.

Parameters of simplified model $\widetilde{\beta}$ are estimated by

$$\widetilde{\mathbf{b}} = (\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\mathbf{Y}. \tag{9.11}$$

Then the residual sum of squares is calculated for simplified model

$$\widetilde{S}_e = \sum (Y_i - \widetilde{\widehat{Y}}_i)^2,$$

where $\widetilde{\widehat{Y}}_i$ is regression estimator of Y_i in simplified model. It is obvious, that $\widetilde{S}_e \geq S_e$.

Under H_0

$$F = \frac{(n-k-1)(\widetilde{S}_e - S_e)}{lS_e} \sim F_{l,n-k-1}$$

has $F_{l,n-k-1}$ distribution. Thus if $F \geq F_{l,n-k-1}(1-\alpha)$ we reject H_0 with significance α and we can not simplify the model.

9.3 Polynomial regression

Quadratic regression:

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot X_i^2 + e_i, \quad i = 1, 2, ..., n,$$

kde $e_i \sim N(0, \sigma^2), n \geq 4$. Here Y_i depends quadraticly on X_i.

If we set $Z_i = X_i^2, \quad i = 1, 2, ..., n$ we have model

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot Z_i + e_i, \quad i = 1, 2, ..., n.$$

Here Y_i depends linearly on X_i and Z_i. So the quadratic task were replaced by linear.

Similarly the regression of higher order.

9.4 Non-linear regression

$$Y_i = f(X_i, \beta) + e_i, \quad i = 1, 2, ..., n,$$

where f is regression function and β is vector of unknown parameters. The estimate of β can be found again by least square method, by minimizing

$$S(\beta) = \sum_{i=1}^{n} (Y_i - f(X_i, \beta))^2.$$

This can be solved by statististical software iteratively.

Starting approximation can be found for **linearizable models** e.g. models which can be transformed to linear. As an example we look at exponential function:

$$Y_i = \beta_0 e^{\beta_1 X_i} + e_i, \quad i = 1, 2, ..., n.$$

In starting approximation we forget the errors e_i and make logarithm

$$\ln Y_i = \ln \beta_0 + \beta_1 X_i, \quad i = 1, 2, ..., n.$$

Now with new parameters $\alpha_0 = \ln \beta_0$ and $\alpha_1 = \beta_1$, we have linear regression

$$\ln Y_i = \alpha_0 + \alpha_1 X_i, \quad i = 1, 2, ..., n.$$

Some examples of linearizable models:

1. $Y = e^{\beta_0 + \beta_1 X}$
2. $Y = \beta_0 X^{\beta_1}$
3. $Y = \beta_0 + \beta_1 \ln x$
4. $Y = \ln(\beta_0 + \beta_1 X)$
5. $Y = \frac{1}{\beta_0 + \beta_1 X}$

9.5 Transformation of the data

Logarithm of the data makes multiplicative model. The data are influenced by covariates in a multiplicative way.

$$\ln Y = \beta_0 + \beta_1 X + e_i$$
$$Y = e^{\beta_0 + \beta_1 X + e_i}$$
$$Y = e^{\beta_0} e^{\beta_1 X} e^{e_i}$$

Power transformations Y^λ can help to achieve normality.

Chapter 10

Distribution tests - goodness of fit tests

10.1 Testing normality

For test of normality - use: Distribution fitting - normal - Plot of observed and expected distribution.

There are two common tests of normality - χ^2 test and Kolmogorov-Smirnov test. Use option. I prefer χ^2 test for its generality and power.

Here can be fitted also other distributions.

When one perform an residual analysis, usually the informative plot - Normal

probability plot of residuals is use to see the deviation of residuals from normal distribution. The deviation here detects either wrong model selection or the non-normality of the residuals.

When we are comparing means of variables and normality is not satisfied and we have just few data, then nonparametric statistics must be used.

Checking serial correlation: Durbin-watson test in residuals analysis in regression detect the serial correlation between data. If it detects serial correlation, the data are not independent and the time series analysis has to be performed. Unfortunately in Statistica is computed only d statistics. As far is d from 2 as bigger the serial correlations are.

10.2 Pearsons χ^2 test

Use observed versus expected frequencies when you want to perform χ^2 test with known theoretical frequencies. The description of this test follows.

It can be used for example for controlling of random numbers generator, controlling the dice, controlling if you catch same number of fishes in day and in

night, ...

Let $Z_1, \ldots, Z_n$ be random sample, where $Z_j, \quad j = 1, \ldots n$ can have values $1, \ldots, k$. The variables X_i, which gives the number of accurences of the result i, will be called empirical frequencies. Random vector $X_1, \ldots, X_k$ has multinomial distribution. We wil test the hypothesis H_0, that theoretical probabilitie of multinomial distribution are equal to the numbers $p_1, ..., p_k$. The variables np_i will be called theoretical frequencies. Under H_0 the statistics

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \sim \chi^2_{k-1}$$

has asymptotically χ^2 distribution with $k-1$ degree of freedom. We reject H_0 when

$$\chi^2 \geq \chi^2_{k-1}(1-\alpha),$$

with significance α.

The χ^2 is asymptotical, therefore it can be done only when n is big enough. Usually the test is accepted when

$$np_i \geq 5q \text{ for all } i = 1, ..., k \text{ and } k \geq 3,$$

where q is ration of classes, for which $np_i < 5$, and n.

Chapter 11

Contingency tables - Analyzing discrete data

What is contingency table:

Consider random vector $Z = (X, Y)$, which has discrete distribution. X has values $1, ..., r$ and Y has values $1, ..., c$. X and Y corresponds to certain property (for example sex, education...).

The properties can be

- qualitative
- discrete quantitative

- continuous quantitative with values in classes

Denote probabilties of distribution of $Z = (X, Y)$:

$$p_{ij} = P(X = i, Y = j), p_{i.} = P(Y = i) = \sum_{j=1}^{c} p_{ij}, \quad p_{.j} = P(Z = j) = \sum_{i=1}^{r} p_{ij}.$$

Consider a random sample of range n from upper distribution. Number of cases, when (i, j) appered in the sample will be denoted by n_{ij} (absolute, empirical frequency). Random variables n_{ij} have multinomial distribution with parameters n and p_{ij}. Contingency table is then the matrix (n_{ij}). Contingency table is in Table 11.1 with matrix of probabilities (p_{ij}), while

$$n_{i.} = \sum_{j=1}^{c} n_{ij}, \quad n_{.j} = \sum_{i=1}^{r} n_{ij}, \quad n = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}$$

and

$$n = \sum_{j=1}^{c} n_{.j} = \sum_{i=1}^{r} n_{i.} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}.$$

Y	Z 1...c	$\sum$
1	$p_{11}...p_{1c}$	$p_{1.}$
...	.....	...
r	p_{r1} ... p_{rc}	$p_{r.}$
$\sum$	$p_{.1}$ ... $p_{.c}$	1

Y	Z 1...c	$\sum$
1	$n_{11}...n_{1c}$	$n_{1.}$
...	.....	...
1	n_{r1} ... n_{rc}	$n_{r.}$
$\sum$	$n_{.1}$ ... $n_{.c}$	n

Table 11.1: Left: matrix of probabilities, right: Contingency table

The following hypothesis can be tested

- hypothesis of independence of properties X and Y
- hypothesis of homogenity
- hypothesis of symetry
- hypothesis of homogenity for repeated measurenment - McNemars test

11.1 Test of independence

H_0: X (1. property) and Y (2. property) are independent
H_1: X and Y are not independent.

Theorem 11.1 *X and Y are independent if and only if $p_{ij} = p_{i.}p_{.j}$, $i = 1, ..., r; j = 1, ..., c$.*

Thus hypothesis of independence can be rewritten into

$$H_0 : p_{ij} = p_{i.}p_{.j}, \quad i = 1, ..., r; j = 1, ..., c.$$

Under H_0 statistic

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}} \tag{11.1}$$

has asymptotically distribution χ^2 with $(r-1)(c-1)$ degree of freedom. H_0 is rejected if $\chi^2 \geq \chi^2_{(r-1)(c-1)}(1-\alpha)$.

The χ^2 is asymptotical, therefore it can be done only when n is big enough. Usually the test is accepted when

$$\frac{n_{i.}n_{.j}}{n} \geq 5q \text{ for all } i = 1, ..., k \text{ and } k \geq 3,$$

where q is ration of classes, for which $\frac{n_{i.}n_{.j}}{n} < 5$, and n. If this is not satisfied some raws or columns must be combined. That is not possible for 2×2 contingency table. In such case Fishers factorial test can be used.

11.2 Test of homogenity

Or test about sameness of structure. This test sameness of one property under different condition, which are expressed by second property. For example if length distribution of catched fish is same in night and day.

H_0: probability $p_{i1}, ..., p_{ic}$ do not depend on the raw index i

(e.g. all raws of matrix p_{ij} are same)

Probabilities $p_{i1}, ..., p_{ic}$ corespond relative frequencies in i-th raw of contingency table $\frac{n_{i1}}{n_{i.}}, ..., \frac{n_{ic}}{n_{i.}}$, here $p_{i1} + ... + p_{ic} = 1$ and furthermore we assume that

raw frequencies n_i are set before experiment.

For testing the homogenity we again use statistic χ^2 equation 11.1. Under H_0 has χ^2 asymptoticaly χ^2 distribution with $(r-1)(c-1)$ degree of freedom. H_0 is then rejected if $\chi^2 \geq \chi^2_{(r-1)(c-1)}(1-\alpha)$.

Chapter 12

Further modeling

General linear models - union of ANOVA and regression. The factors can be either continuous and discrete.
Generalized linear models - Generalization of general linear models, the data can have different then normal distribution.
For count data Poisson distribution is used.
For discrete data Multinomial distribution is used.
For continuous data one can often use Gamma distribution.