Testing the Equality of Means and Variances across Populations and Implementation in XploRe [1]

Michal Benko
Wirtschaftwissenschaftliche Fakultät
Humboldt Universität zu Berlin [2]

1st March 2001

[1]prepared to obtain Bsc. degree in Statistic
[2]Supervised by **Prof. Dr. Bernd Rönz**

Contents

1 Introduction to the Testing Theory **7**
1.1 General Hypothesis Construction . . . 7
1.1.1 Two sided versus one sided hypotheses . . . 7
1.2 Tests . . . 8
P-Value . . . 9

2 Exploratory data analysis **11**
2.1 Histogram . . . 11
2.1.1 Implementation in XploRe . . . 11
2.1.2 Example . . . 12
2.2 Average shifted histograms . . . 13
2.2.1 Implementation in the XploRe . . . 13
2.2.2 Example . . . 14
2.3 Boxplot . . . 15
2.3.1 Implementation in XploRe . . . 16
2.3.2 Example . . . 17
2.4 Spread&level-Plot . . . 18
2.4.1 Implementation in XploRe . . . 18

3 Testing the Equality of Means and Variances **23**
3.1 Testing the equality of Variances across populations . . . 23
3.1.1 F-test . . . 23
Implementation in XploRe . . . 25
Example . . . 25
3.1.2 Levene Test . . . 26
Implementation . . . 26
Example . . . 27
3.2 Testing the equality of Means across populations . . . 27
3.2.1 T-test . . . 27
3.2.2 T-test under equal variances . . . 28
3.2.3 T-test with unequal variance . . . 29
3.2.4 Implementation . . . 29
3.2.5 Example . . . 29
3.2.6 Simple Analysis of Variance ANOVA . . . 30

4 Appendix **35**
4.1 Distributions 35
4.2 XploRe list 37
4.2.1 f-test 37
4.2.2 t-test 38
4.2.3 ANOVA 40
4.2.4 Levene 43
4.2.5 Spread and level Plot 45
Bibliography 51

Preface

People in statistical and Data-analytical practice often face to the problem of comparing characteristics across populations, e.g., they have to investigate the influence of environmental-changes on the certain variables. The mean and variance are interesting characteristics of a random variables from the statistical and also from the practical point of view. Hence, this paper will focus on these two basic characteristics. After discussing the theoretical background in the first chapter, we will introduce and explain fundamental methods and procedures, which solves this problematic by using statistical inference approach. In addition to the theory, this work will comment on the use of some existing procedures and methods of Exploratory data analysis and statistical inference in computing environment XploRe, and implement new procedures (quantlets) to this statistical language.

Michal Benko

Chapter 1

Introduction to the Testing Theory

1.1 General Hypothesis Construction

Suppose that a sample of $X_1, X_2, \ldots, X_n$ is generated by random variable $\boldsymbol{X}$, which depends on some abstract parameter θ, which belongs to some known parameter space Θ, the real value of the parameter is often unknown, we know only some class of possible values for θ, let us denote this class as parameter space Θ. However we can construct set of two Hypotheses about this parameter (e.q. split the parameter space into some subspaces):

Null hypothesis is an assumption about the parameter θ, which we want to "test":

$$H_0 : \theta \in \omega \text{ , where } \omega \subseteq \Theta$$

Situation is completely specified only when we know what are other alternatives for θ besides values from ω. This is the so-called **alternative hypothesis**. One of the most common examples is the alternative hypothesis that is **complementary** to the null hypothesis:

$$H_1 : \theta \in \Theta - \omega$$

1.1.1 Two sided versus one sided hypotheses

In the following text we will implicitly assume one dimensional parameter, one point hypothesis ($\omega \in \Omega$) and $\Theta \subseteq \mathbb{R}$. This assumption split our abstract situation to two basic Hypothesis types:

- **Two-sided Hypothesis($\Theta = \mathbb{R}$):**

  Null Hypothesis:
  $$H_0 : \ \theta = \theta_0$$
  against alternative Hypothesis:
  $$H_1 : \ \theta \neq \theta_0$$
  where $\theta_0 \ \in \ \mathbb{R}$

- **One sided Hypothesis**($\Theta \subseteq \mathbb{R}$), in this type we distinguish two cases:

  - $\Theta = \{\theta \geq \theta_0; \theta, \theta_0 \ \in \ \mathbb{R}\}$

    with corresponding Hypothesis:
    $$H_0 : \ \theta = \theta_0$$
    against alternative
    $$H_1 : \ \theta \geq \theta_0$$
  - $\Theta = \{\theta \leq \theta_0; \theta, \theta_0 \ \in \ \mathbb{R}\}$

    with corresponding Hypothesis:
    $$H_0 : \ \theta = \theta_0$$
    against alternative
    $$H_1 : \ \theta \leq \theta_0$$

Example:
Assume that a $X \sim N(\mu, \sigma)$. The two-sided Hypothesis would be:

Null Hypothesis:
$$H_0 : \ \mu = 0$$
against alternative Hypothesis:
$$H_1 : \ \mu \neq 0$$

1.2 Tests

DEFINITION 1.1 ***Testing*** H_0 ***against*** H_1 *is a decision process based on our sample* $X_1, X_2, \ldots, X_n$*, witch leads to rejection or no rejection of* H_0

After the testing four situations may occur:

1. H_0 is true and our decision is not to reject H_0 – correct decision

2. H_0 is true, but our decision is to reject H_0 – wrong decision
3. H_1 is true, but our decision is not to reject H_0 – wrong decision
4. H_1 is true and our decision is to reject H_0 – correct decision

Hence, there are two ways of making wrong decision, in the case (2) we make the so-called **first type error**, in the case (3), we make so-called **second type error**. For the better understanding we will discus this problematic parallel to two other concepts:

We can describe our Test by a subspace of the possible values for our sample $\boldsymbol{X}$ (in our case hold: $\boldsymbol{W} \subset \mathbb{R}^n$) — the so-called **Critical area** in following way:

$$(X_1, X_2, \ldots, X_n) \in \boldsymbol{W} \longrightarrow \text{reject } H_0$$

$$(X_1, X_2, \ldots, X_n) \notin \boldsymbol{W} \longrightarrow \text{do not reject } H_0$$

The goal is to choose the critical area so that first type error is less or equal than some a priori chosen number $\alpha > 0$, for all θ corresponding to our H_0 Hypothesis:

$$P_\theta((X_1, X_2, \ldots, X_n) \in \boldsymbol{W}) \leq \alpha \, \forall \, \theta \in \omega \tag{1.1}$$

This value $sup_{\theta \in \omega} P_\theta((X_1, X_2, \ldots, X_n) \in \boldsymbol{W})$ is called **significance level**, in our simplified one-point situation it is equal to the probability of first type error for $\theta = \theta_0$

It is convenient to say, that we are testing on the significance level α, or in the case of rejecting the H_0 hypothesis, rejecting the H_0 at the significance level α.

However, in practice, the n-dimensional critical area is usually transformed to a one-dimensional real critical area, by a function called **test statistic**: $T = T(X_1, X_2, \ldots, X_n)$. Because it is a function of a random sample, it is also a one-dimensional random variable. Consequently, the critical area is then just an interval or a set of intervals. Such intervals are mostly of the form $\langle a, b \rangle$ or (a, b), where a and b are certain quantiles of the distribution of T under the validity of H_0. Thus we have to know (at least asymptotically) the distribution of T, in order to construct the critical area with the property (1.1) and to run the test.

Example:

Assume a random sample: $(X_1, X_2, \ldots, X_n)$
The possible Test statistic would be e.g.:
Sample mean: $\overline{X} = \frac{1}{n}(\sum_{i=1}^{n} X_i)$

P-Value, Sig.value

The tests in XploRe produce as result **P-value**, which is sometimes called

Significance value. P-value is equal to the probability that a random variable with the same distribution as the test statistics T under the validity of the hypothesis H_0 is greater or equal than the value of the statistics T of the given sample. In other words, it corresponds to the biggest significance level, at which the null hypothesis H_0 cannot be rejected.

We will explain this concept in practice more precisely: Let us assume sample $\boldsymbol{X}$ and that the test-statistic T follows under H_0 $N(0,1)$ distribution. We want to test a one-sided hypothesis for some general parameter θ, e.g. $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. We can directly see from the definitions, that $\alpha = P(T > \phi_{1-\alpha} = P(T > T_{crit}))$, where $\phi_{1-\alpha}$ is a $(1-\alpha)$-quantile of the standardized normal distribution - $N(0,1)$ (see 4.1), and α is the significance level. Hence, the interval (T_{crit}, ∞) is the Critical area with the property (1.1). From the test procedure, we will obtain certain value for T let say T_{sample} (depending on the sample $\boldsymbol{X}$). It is now possible to compute the probability that the random variable T is bigger than T_{sample}: $P = P(T > T_{sample})$. The test-procedure is the following: If $P < \alpha$, implies $P(T > T_{sample}) < P(T > T_{crit})$, from the monotony of probability measure, we will obtain: $T_{sample} > T_{crit}$, so $T_{sample} \in$ Critical area, so we can reject the hypothesis H_0 at significance level α. In the case of $\alpha \leq P$ we will obtain that $T_{sample} \notin$ Critical area so we can not reject H_0.

We will also discuss the two-sided hypothesis:

$$H_0 : \theta = \theta_0$$

against

$$H_0 : \theta \neq \theta_0$$

using the same notation we obtain: $\alpha = \alpha/2 + \alpha/2 = P(T < -T_{crit}) + P(T > T_{crit})$, where $T_{crit} = \phi_{1-\alpha/2}$. We can also denote $P = P(T < -T_{sample}) + P(T > T_{sample})$. If $P < \alpha$ implies
$P = P(T < -T_{sample}) + P(T > T_{sample}) < P(T < -T_{crit}) + P(T > T_{crit})$,the monotony of probability measure and the symmetry of the normal distribution imply that $T < -T_{crit}$ or $T > T_{crit}$ so $T \in$ Critical area , so we can reject H_0. If $P \geq \alpha$ we can similar obtain that $T \notin$ Critical area so we can not reject H_0.

Chapter 2

Exploratory data analysis

In this chapter we will discuss some of exploratory methods which can be used to show the differences across samples. This analysis should help us to construct hypothesis about mean and variance for further testing. We will focus on two most common graphic tools: boxplots, histograms, and spread-level-plots — exploratory tool for investigating the homogenity of variances.

2.1 Histogram

The histogram is the most common method of one dimensional density estimation. It is useful for continuous distribution or for discrete distribution with big numbers of expression. The idea of histogram is the following: Construct the disjunct serie of intervals B_j, where $B_j(x_0, h) = (x_0 + (j+1)h, x_0 + jh],\ j \in \mathbb{Z}$ correspond with the bins of length h and origin point x_0. The histogram is then defined by:

$$\widehat{f}_h(x) = n^{-1}h^{-1} \sum_{j \in \mathbb{Z}} \sum_{i=1}^{n} I\{x \in B_j(x_0, h)\}$$

where I means Identification function. Parameter h is a smoothing parameter, that means, if we use smaller h, we get smaller intervals (bins) $B_j(x_0, h)$ and so more structure of data is visible in our estimation. The optimal choice of this parameter is described in (Härdle, W., Müller, M., Sperlich, S., & Werwatz, A., 1999)

2.1.1 Implementation in XploRe

```
gr=grhist (x, h, o, col)
     grhist generates graphical object histogram
```

with following parameters

`x`
: is a $n \times 1$ data vector

`h`
: bindwidth, scalar, default is $h = \sqrt{var(x)}/2$

`o`
: origin (x_0), scalar, default is $x = 0$

`col`
: color, default is black

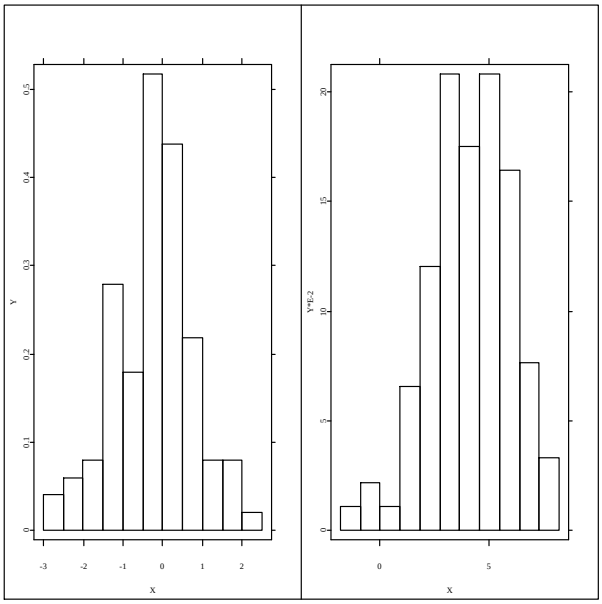
`gr`
: graphical object

2.1.2 Example

`exhist.xpl`

We simulate 100 observations with standard Normal distribution,and 100 observations with $N(2, 4)$, we can obtain histograms by following sequence:

```
library("graphic")
x1=normal(10)
x2=(normal(100)+2).*2
gr1=grhist(x1)
gr2=grhist(x2)
di=createdisplay(1,2)
show(di,1,1,gr1)
show(di,1,2,gr2)
```



In this figure, we can see the estimates of the distribution of the populations (histograms). The sample from the standard normal distribution in the left display and the sample from $N(2, 4)$ in the right display. However this simple principle is quite sensitive to the choice of the parameters x_0 and h. By the comparing to histograms one has also take care about scaling factors of the plots. To solve this problems partially we can use average shifted histograms, which we will discussed in the next chapter.

2.2 Average shifted histograms

Average shifted histograms are based on an idea of averaging several histograms with different origins, to obtain density estimation independent on the choice of x_0.

2.2.1 Implementation in the XploRe

```
gr=grash (x, h, o, col)
      grash generates graphical object histogram
```

`x`
: is a $n \times 1$ data vector

`h`
: bindwidth, scalar, defaults is $h = \sqrt{var(x)}/2$

`k`
: number of shifts, scalar, default is $k = 50$

`col`
: color, default is black

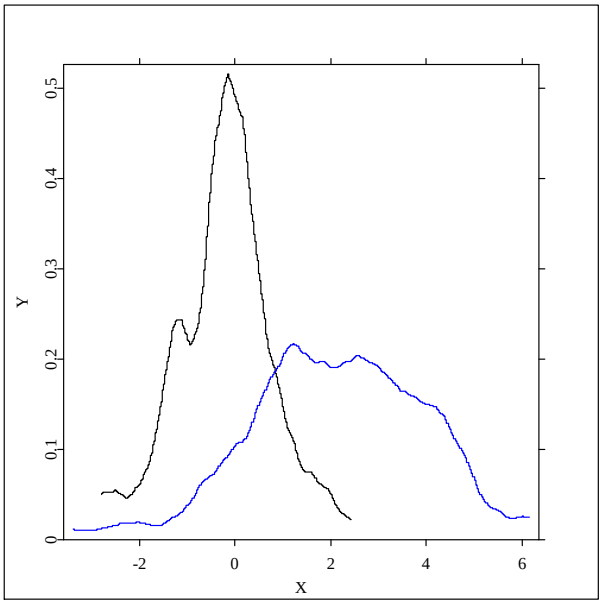
`gr`
: graphical object

2.2.2 Example

`exash.xpl`

We simulate 100 observations with standard Normal distribution,and 100 observations with $N(2, 4)$, we can obtain Average Shifted Histograms by typing:

```
library("graphic")
randomize(0)
x1=normal(100)
x2=2*(normal(100))+2
mean(x2)
gr1=grash(x1,sqrt(var(x1))/2,30,0)
gr2=grash(x2,sqrt(var(x2))/2,30,1)
di=createdisplay(1,1)
show(di,1,1,gr1,gr2)
```



In this case we can observe the differences in the density estimations, the different location and spread of our estimators. The estimation of the generating density of first sample is black and the estimation of the generating density of second sample is blue. We can see that the black line is located left to the blue line, so we could assume inequality of means and test it. It is also visible, that the spread of the blue line is bigger than the spread of the black line. Hence we can also assume (and test) the variance inequality. However in this example we know the true parameter (if we assume that the random generator works fully stochastic), this example should only show the usage of the averaged shifted histograms.

2.3 Boxplot

Boxplot is also a common graphical tool to display characteristics of a distribution. It is a representation of the so-called Five Number Summary, namely upper quartile (F_U) and lower quartile (F_L), median and extremes. To define this characteristics we have to consider order-statistics $x_{(1)}, x_{(2)} \ldots, x_{(n)}$ as ordered sequence of variables
$x_1, x_2, \ldots, x_n$, where $x_{(i)} \leq x_{(j)}$, for $i \leq j$. Now we will introduce characteristics used in the Boxplot:

median median "cuts" the observations in to two equal parts

$$M = \begin{cases} X_{\frac{n+1}{2}} & \text{for } n \text{ odd,} \\ \\ \frac{1}{2}(X_{\frac{n}{2}} + X_{\frac{n}{2}+1}) & \text{for } n \text{ even.} \end{cases}$$

quartiles quartiles cuts the observations into four equal parts, we can introduce the depth of the data value $x_{(i)}$ as a $\min\{i, n-i+1\}$ (Depth can be also a fraction, e.g. depth of median for n even $\frac{n+1}{2}$ is a fraction, then we compute the value with this depth as a average of $x_{\frac{n}{2}}, x_{\frac{n}{2}+1}$.)Now we can calculate

$$\text{depth of fourth } = \frac{[\text{depth of median}] + 1}{2}$$

so the upper and lower quartile are the values with this depth.

IQR Interquartile Range (also-called F-spread) is defined as $d_F = F_U - F_L$ is a robust estimator of spread

outside bars

$$F_U + 1.5 d_F$$

$$F_L - 1.5 d_F$$

are the borders for outliers identification, the points outside these boarders are regarded as outliers.

extremes are minimum and maximum

mean (arithmetic mean) $\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, is a common estimator for the mean parameter

Boxplot is no density estimator (in compare to the Histograms), but graphically shows the most important characteristics of density in order to investigate the location and spread of densities.

2.3.1 Implementation in XploRe

```
plotbox(x {,Factor})
      plotbox draws boxplot in a new display
```

`x`
is a $n \times 1$ data vector

`Factor`
$n \times 1$ string vector specifying groups within X

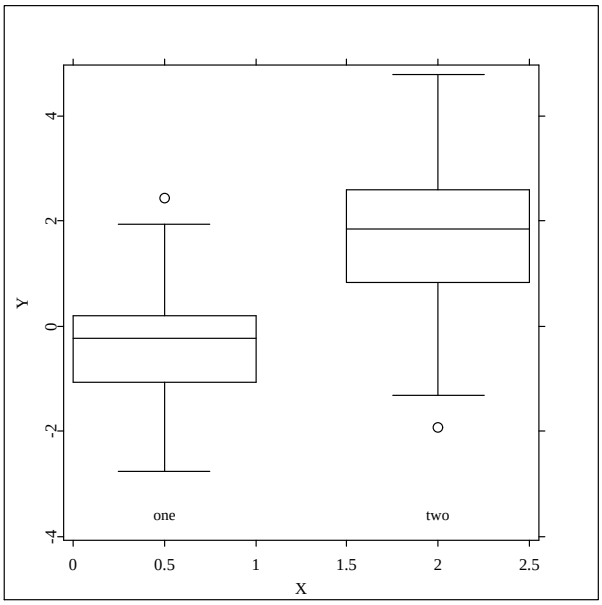
Factor is a optional parameter.

2.3.2 Example

In this example we will show the usage of box-plots as a tool of visualization of sample differences. Once again we will simulate two samples $\boldsymbol{X}_1 \sim N(0,1)$ and $\boldsymbol{X}_2 \sim N(2,2)$, we will draw boxplots of these samples to observe differences by typing following list: `explotbox.xpl`

```
library("graphic")
library("plot")
randomize(0)
x1=normal(50)
x2=sqrt(2).*normal(50)+2
x=x1|x2
f=string("one",1:50)|string("two",1:50)
plotbox(x,f)
```

In the output window we obtain:



We can visually compare the location and the height of boxes, we can see that the location of box (the solid line in the middle means median) is higher as in the first sample. The second box is higher than the first one, hence also the spreads of the boxes differs. Because the high of the box corresponds with some estimations of variance, and the location of the boxes corresponds with the estimations of means, we can also assume the differences (and run the tests) in these two distributions.

2.4 Spread&level-Plot

The Spread&level-Plot shows a plot for median of each sample against their IQR. Median and Inter Quartile Range are robust estimators for mean and standard deviation ($= \sqrt{(}Var(X))$). This plot helps to explore the homogenity of variances across populations, if the differences are low, there are only small differences on y-axes, so we can observe more or less horizontal line.

In addition to this plot quantlet `plotspleplot` computes also the slope of the line, given by :

$$\text{Slope } = \frac{\sum_{j=1}^{m}(m_j - \overline{m})(s_j - \overline{s})}{\sum_{j=1}^{m}(m_j - \overline{m})^2}$$

where

- s_j denotes IQR (spread) of the j-th sample, $\overline{s} = m^{-1}\sum_j = 1^m s_j$
- m_j denotes median (level) of the j-th sample, $\overline{l} = m^{-1}\sum_{j=1}^{m} l_j$

Optionally we can get also estimation of power transformation to obtain a data set with equal variances. To obtain this estimation we make plot and compute slope with the log of data set. The value of estimation is equal to the $1 - slope$ rounded to the nearest 0.5. If the estimation is equal to the p we should run the x^p transformation in order to obtain the data set with equal variances.

2.4.1 Implementation in XploRe

grspleplot

```
gr=grspleplot(data)
```
grspleplot generates a graphic-object with spread and level plot

`data`
is a $n \times p$ data set

`gr`
graphical object

dispspleplot

```
dispspleplot(dis,x,y,data)
```
dispspleplot draws a spread and level plot into specific display

`dis`
: display

`x`
: scalar, x-position in display dis

`y`
: scalar, y-position in display dis

`data`
: is a $n \times p$ data set

plotspleplot

`plotspleplot(data)`
: plotspleplot runs spread and level plot

`data`
: is a $n \times p$ data set

Example

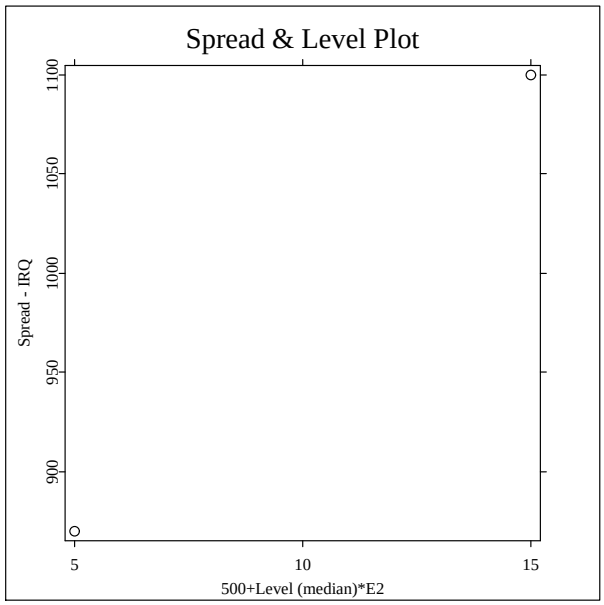
`exspleplot.xpl`

Let us compare the monthly income of people, factorized by the variable sex.The data set `allbus` from: Wittenberg,R.(1991): Computergestützte Datenanalyse have been used. This dataset contains monthly income of men and women in Germany. We can run the spread & level plot by typing:

```
library("plot")
x=read("allbus.dat")
man=paf(x,x[,1]==1)[,2]
woman=paf(x,x[,1]==2)[,2]
woman=woman|NaN.*matrix(rows(man)-rows(woman),1)
x=man~woman
plotspleplot(x)
```

We can chose if we want to have power estimation or not. We will show both outputs.

First we will get the following graphical output display



Without selecting power estimation we get following output text:

```
[1,] " --- Spread-and-level Plot--- "
[2,] "-----------------------------"
[3,] " Slope =                0.230"
```

So we can see, that there are quite big differences on y-axes, and we have the $slope = 0.230$. With selecting power estimation we will obtain:

```
[1,] " ------- Spread-and-level Plot------- "
[2,] "  slope of LN of level and LN spread  "
[3,] "--------------------------------------"
[4,] " Slope =                0.338"
[5,] "Power transf. est.      0.662"
```

In this case, we have data transformed by log-transformation, so the slope is not equal to the slope in the first case. However the plot have been plotted with data without transformation. We have obtained the power estimation $= 0.688$ so we should use power estimation $= 0.5$ We can test this with levene test (see 3.1.2). After running the tests for original data and for data transformed by power transformation $p = 0.5$, we obtained following result:

```
[1,] "-------------------------------------------------"
[2,] "Levene Test for Homogenity of Variances          "
[3,] "-------------------------------------------------"
[4,] "   Statistic     df1     df2    Signif.          "
[5,] "     16.4835       1     714    0.0001           "
```

for original data, that means it is highly significant (significance=0.001)

```
[1,] "-------------------------------------------------"
[2,] "Levene Test for Homogenity of Variances          "
[3,] "-------------------------------------------------"
[4,] "   Statistic     df1     df2    Signif.          "
[5,] "      0.0913       1     714    0.7626           "
```

for transformed data, that means this variance inequality have been strongly corrected.

Chapter 3

Testing the Equality of Means and Variances

In this chapter, we want to test the differences of distributions across populations. These question is, however very complex, so we will focus on the differences of two distribution-characteristics: first moment or mean (EX) and second central moment or Variance ($var(X) = E(X - EX)^2$). This two are the characteristics, which describe the location and spread of distribution. This two characteristics also characterize uniquely the Normal distribution. We will start with the testing for equality of variances (F-test and Levene-test) because the equality of variances is a common assumption in mean equality tests: ANOVA and T-test which we will discus later.

3.1 Testing the equality of Variances across populations

3.1.1 F-test

Let us consider two samples $X_{1,1}$, $X_{1,2}$, ..., $X_{1,n_1} \sim N(\mu_1, \sigma_1^2)$ and $X_{2,1}, X_{2,2}$, $\ldots, X_{2,n_2} \sim N(\mu_2, \sigma_2^2)$, and let the underlying random variables X_1 and X_2 be stochastically independent. Under this assumptions we can test the following hypothesis that the variances are equal:

$$H_0 \ : \ \sigma_1 = \sigma_2$$

against the two-tailed alternative

$$H_1 \ : \ \sigma_1 \neq \sigma_2.$$

Under H_0 the test statistic

$$F = \frac{s_1^2}{s_2^2} = \frac{\frac{1}{n_1-1}\sum\limits_{i=1}^{n_1}(X_{1,i} - \overline{X_1})^2}{\frac{1}{n_2-1}\sum\limits_{i=1}^{n_2}(X_{2,i} - \overline{X_2})^2}.$$

follows $F(n_1-1, n_2-1)$ distribution. Hence, the hypothesis H_0 is to be rejected if $F < F_{n_1-1,n_2-1}(\alpha/2)$ or $F > F_{n_1-1,n_2-1}(1-\alpha/2)$, where $F_{m,n}(\alpha)$ represents the α-quantile of the F distribution with m and n degrees of freedom.

Let us prove this assumption. Denote

$$\begin{array}{l} S_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_{1,i} - \overline{X_1})^2 \text{ where } \overline{X_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1,i} \\ S_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(X_{1,i} - \overline{X_1})^2 \text{ where } \overline{X_2} = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2,i} \end{array}$$

Thus the random variables $\chi_1 = \frac{(n_1-1)S_1^2}{\sigma_1^2}$ and $\chi_2 = \frac{(n_2-1)S_2^2}{\sigma_2^2}$ are sums of squares of independent, standard normal distributed variables divided by the degrees of freedom, so these variables follow the Chi-square distribution with $n_1 - 1$ or $n_2 - 1$ degrees of freedom (see 4.2). Let us construct the test statistic F:

$$F = \frac{\frac{\chi_1^2}{n_1-1}}{\frac{\chi_2^2}{n_2-1}} = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}},$$

Under the H_0 is

$$F = \frac{S_1^2}{S_2^2},$$

and T follows the F-distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

Without loss of generality, assume that s_1, the nominator of the F-statistic, is greater or equal to s_2 (which implies $F > 1$). Then we can alternatively test

$$H_0 \ : \ \sigma_1 = \sigma_2$$

against

$$H_1 \ : \ \sigma_1 > \sigma_2$$

and reject the hypothesis H_0 if

$$F \ > \ F_{n_1-1,n_2-1,1-\alpha}.$$

This test is (according to the used s_1) very sensitive to outliers and the violation of the Normality assumption.

Implementation in XploRe

```
text=ftest(d1,d2)
```
ftest runs the F-test on the samples in vectors d1 and d2

The meaning of parameters is following:

`d1`
is a $n_1 \times 1$ vector corresponding to the first sample

`d2`
is a $n_2 \times 1$ vector corresponding to the second sample

`text`
text vector—text output

Example

exftest.xpl

Consider two samples:

$$-1.02, -1.96, -0.94, 0.39, 0.33, 0.98, 0.74, -0.2, -0.64$$

and

$$0.79, 1.28, 1.65, -3.02, 0.52, 0.39, -0.93, 0.41, -0.78$$

These two samples correspond with the deviation from the exact size of product of two industrial cutting machines (Assume that the setups of these two machines are independent). We are asked to compare these two machines according to the spread of the errors.

Let assume that these two samples are produced by independent Normal distributed random variables, we want to test the equivalence of the spreads of this two sample on the confidence level 0.95, F-test can be computed by typing:

```
library("stats")
x=#(-1.02,-1.96,-0.94,0.39,0.33,0.98,0.74,-0.2,-0.64)
y=#(0.79,1.28,1.65,-3.02,0.52,0.39,-0.93,0.41,-0.78)
ftest(x,y)
```

The output, in the output window is following:

```
[1,] "------------- F test -------------"
[2,] "----------------------------------"
[3,] "testing s2>s1"
[4,] "----------------------------------"
[5,] "F value: 2.1877     Sign. 0.2890"
[6,] "dg. fr. =    9,   9"
```

According to this output, we can see that $s_2 > s_1$, and that our statistic $F \sim F_{9,9}$ equals 2.1877. Significance equals the probability that this statistic F is greater than our computed value 2.1877 — see F-value entry in the output. In our case $0.2890 > 0.05$, where 0.05 was the chosen α in our confidence level $1-\alpha$ so we cannot reject the hypothesis H_0 (equivalence of spreads) on the confidence level 0.05.

There is no significant difference between the spreads of errors of this two machines on the confidence level of 0.95

3.1.2 Levene Test

In comparison with the F-test, Levene test is less sensitive to the outliers and the violation of the normality assumption. This is caused by using the absolute deviation measure instead of squared measure. In addition, Levene test also allows to test in general $m \geq 2$ samples at once. The normality of random variables is still requested. Let us denote the samples as $X_{j,1}, \ldots, X_{j,n_j}$, $j = 1, \ldots, m$, produced by continuous random variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$, where $\boldsymbol{X}_i \sim N(\mu_i, \sigma_i^2)$. We want to test

$$H_0 : \sigma_1, = \ldots, = \sigma_m$$

against

$$H_1 : \exists \sigma_j \neq \sigma_i \text{ for } i \neq j$$

Let us construct new variable D

$$D_{j,i} = \mid X_{j,i} - \overline{X_j} \mid \quad j = 1, \ldots, m, \quad i = 1, \ldots, n_j \text{ where } \overline{X_j} = n_j^{-1} \sum_{i=1}^{n_j} x_j$$

and the test statistic L:

$$L = \frac{n-m}{m-1} \frac{\sum_{j=1}^{m} n_j (\overline{D_j} - \overline{D})^2}{\sum_{j=1}^{m} \sum_{i=1}^{n_j} (D_{j,i} - \overline{D_j})^2}$$

where $n = \sum n_j$ This statistic corresponds to the ANOVA on the variable D — Absolute deviations, which we will discuss in the next section. Hence, $L \sim F(m-1, n-m)$. So we have to reject H_0 if $L > F_{m-1,n-m,1-\alpha}$, where $F_{m-1,n-m}(1-\alpha)$ is a $(1-\alpha)$ quantile of F-distribution with $m-1, n-1$ degrees of freedom. .

Implementation

```
out=levene(datain)
      levene runs Levene test on the dataset in datain
```

The meaning of parameters is following:

`datain`
: is a $n \times p$ array, data set, NaN allowed

`out`
: is a $n_2 \times 1$ text vector, output text

Example

exlevene.xpl

Let us compare the monthly income of people, factorized by the variable sex. The data set `allbus` from: Wittenberg,R.(1991): Computergestützte Datenanalyse have been used. This dataset contains monthly income of men and women in Germany. We want to test the equality of the spreads of this two sample on the confidence level 0.95, under the assumption, that these samples have been produced by the normal random variables. Levene-test can be computed by typing:

```
library("stats")
x=read("allbus.dat")
man=paf(x,x[,1]==1)[,2]
woman=paf(x,x[,1]==2)[,2]
woman=woman|NaN.*matrix(rows(man)-rows(woman),1)
x=man~woman
levene(x)
```

As output we can see the result of Levene test:

```
[1,] "-------------------------------------------------"
[2,] "Levene Test for Homogenity of Variances          "
[3,] "-------------------------------------------------"
[4,] "   Statistic     df1     df2    Signif.          "
[5,] "     16.4835       1     714    0.0001           "
```

According to this output we can see that the significance (or P-Value) is smaller than our level 0.05 so we can reject the hypothesis, that both variances are equal.

3.2 Testing the equality of Means across populations

3.2.1 T-test

In this section, we will test the equality of the means of two populations, based on the independent samples. Under the normality assumption, we can use the so-called **t-test**, which uses two different approaches depending on the equality or inequality of sample variances of underlying samples.

Assume two samples: $X_{1,1}, X_{1,2}, \ldots, X_{1,n_1}$ being distributed according to $N(\mu_1, \sigma_1^2)$ and $X_{2,1}, X_{2,2}, \ldots, X_{2,n_2}$ being $N(\mu_2, \sigma_2^2)$ distributed. These samples

should be independent. We want to find out whether the means of the two populations (from which the samples are drawn) are equal, that is to test

$$H_0 : \mu_1 = \mu_2$$

against

$$H_1 : \mu_1 \neq \mu_2.$$

Let us first investigate the location and the spread of difference $\overline{X}_1 - \overline{X}_2$, which is a natural estimate of $\mu_1 - \mu_2$:

$$E(\overline{X}_1 - \overline{X}_2) \ = E(\overline{X}_1) - E(\overline{X}_2) \ = \ \mu_1 - \mu_2,$$

$$Var(\overline{X}_1 - \overline{X}_2) = Var(\overline{X}_1) + Var(\overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Hence,

$$N = \frac{(\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2))}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

Under H_0, we can simplify the N variable to

$$N^* = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

3.2.2 T-test under equal variances

Under the assumption of variance equality, $\sigma_1 = \sigma_2 = \sigma$, we can simplify the variable N^* and build the test statistic

$$T = \frac{\overline{X}_1 - \overline{X}_2}{S^*} = N^* \frac{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}{S^*} \sim \frac{N(0,1)}{\sqrt{\chi_f^2/f}} \sim t_{n_1+n_2-2},$$

where S^* represents an estimate of $Var(\overline{X}_1 - \overline{X}_2)$

$$S^* = \frac{((n_1-1)s_1^2 + (n_2-2)s_2^2)}{n_1+n_2-2}$$

and $f = n_1 + n_2 - 2$. Hence

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{n_1+n_2}{n_1 \cdot n_2} \cdot \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}} \sim t_{n_1+n_2-2},$$

which follows t-distribution with $n_1 + n_2 - 2$ degrees of freedom (see 4.3), under H_0. Then, we reject H_0 if $|T| > t_{n_1+n_2-2}(1-\alpha/2)$, where $t_n(\alpha)$ represents the α-quantile of the t-distribution with n degrees of freedom.

3.2.3 T-test with unequal variance

Whenever the variances are not equal, we face the Behrens-Fisher problem—we cannot construct the exact test statistic in this case. The solution is to approximate the ditribution of the test statistic

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_1^2}{n_2}}}$$

by the t-distribution with

$$d = \left\lceil \frac{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})^2}{\frac{(\frac{S_1^2}{n_1})^2}{n_1-1} + \frac{(\frac{S_2^2}{n_2})^2}{n_2-1}} \right\rceil$$

degrees of freedom (symbol $\lceil x \rceil$ represents the smallest integer greater or equal to x). Then we reject the H_0 if $|T| > t_d(1-\alpha/2)$, where $t_d(\alpha)$ means α-quantile of t-distribution with d degrees of freedom.

3.2.4 Implementation

In XploRe, both tests are implemented by one quantlet `ttest`:

```
text=ttest(x1,x2)
      ttest runs T test on x1, x2
```

The explanation of the parameters is following:

`x1`
: is a $n_1 \times 1$ vector corresponding to the first sample

`x2`
: is a $n_2 \times 1$ vector corresponding to the second sample

`text`
: text vector—text output

3.2.5 Example

`exttest.xpl`

Consider two samples

$$-1.02, -1.96, -0.94, 0.39, 0.33, 0.98, 0.74, -0.2, -0.64$$

and

$$0.79, 1.28, 1.65, -3.02, 0.52, 0.39, -0.93, 0.41, -0.78.$$

These two samples describe deviations from the exact size of a product of two industrial cutting machines (assume that the setups of these two machines are independent). We are asked to compare these two machines according to the means of the errors.

Let us assume that the underlying distributions for these two samples are normal and that the corresponding random variables are independent. To create vectors x and y containing these samples, type

```
x=#(-1.02,-1.96,-0.94,0.39,0.33,0.98,0.74,-0.2,-0.64)
y=#(0.79,1.28,1.65,-3.02,0.52,0.39,-0.93,0.41,-0.78)
```

We want to test now, whether the mean sizes (or equivalently mean deviations from the exact size) of the product produced by the two machines are the same. As the `ttest` quantlet performs the t-test both under assumption of equal and unequal variance, we can postpone testing for the equivalence of spreads to Section (3.1)

Now, we can run the t-test by typing

```
library("stats")
x=#(-1.02,-1.96,-0.94,0.39,0.33,0.98,0.74,-0.2,-0.64)
y=#(0.79,1.28,1.65,-3.02,0.52,0.39,-0.93,0.41,-0.78)
ttest(x,y)
```

The output is following:

```
[1,] " -------- t-test (For equality of Means) -------- "
[2,] "-------------------------------------------------"
[3,] "            t-value    d.f.     Sig.2-tailed "
[4,] "Equal var.: -0.5110     16       0.6163"
[5,] "Uneq. var.: -0.5110     15       0.6168"
```

We can see, that under assumption of spread equivalence our test statistic $T \sim t_{16}$ equals -0.5110 (line 4 in the output, the degrees of freedom are to be found in column 'd.f'). The significance equals 0.6163 (see 'Sig.2-tailed'), which is greater than 0.05. Thus, we cannot reject H_0 hypothesis saying that these two samples have the same mean on the confidence level 0.95.

More interestingly, we obtained almost the same result under the assumption of unequal variances (see line 5), which might suggest that variances in both samples are equal. That indicates that the use of t-test under assumption of equivalent spreads was correct. Nevertheless, such an assumption has to be statistically verified—(see Section 3.1 for the proper test.

3.2.6 Simple Analysis of Variance ANOVA

Assume p independent samples

$$X_{1,1}, \ldots, X_{1,n_1} \sim N(\mu_1, \sigma)$$
$$X_{2,1}, \ldots, X_{2,n_1} \sim N(\mu_2, \sigma)$$

$$\dots$$

$$X_{p,1}, \dots, X_{1,n_p} \sim N(\mu_p, \sigma)$$

We want to test

$$H_0 : \ \mu_1 = \mu_2 \cdots = \mu_p$$

against

$$H_1 : \ \exists \mu_i \neq \mu_j \ \text{ for } i \neq j$$

Let us denote:

$$\begin{aligned} n &= \sum_{i=1}^{p} n_i \\ \overline{X}_j &= \frac{1}{n_j} \sum_{i=1}^{n_j} X_{j,i} \\ \overline{X} &= \frac{1}{n} \sum_{j=1}^{p} n_j \overline{X}_j \end{aligned}$$

Using this notation, we can decompose sum of square (SS) in the following way:

$$\begin{aligned} SS &= \sum_{j=1}^{p} \sum_{i=1}^{n_j} (X_{j,i} - \overline{X})^2 \\ &= \sum_{j=1}^{p} \sum_{i=1}^{n_j} ((X_{j,i} - \overline{X}_j) + (\overline{X}_j - \overline{X}))^2 \\ &= \sum_{j=1}^{p} \sum_{i=1}^{n_j} (X_{j,i} - \overline{X}_j)^2 + 2 \sum_{j=1}^{p} ((\overline{X}_j - \overline{X}) \sum_{i=1}^{n_j} (X_{j,i} - \overline{X}_j)) + \sum_{j=1}^{p} \sum_{i=1}^{n_j} (\overline{X}_j - \overline{X})^2 \\ &= \sum_{j=1}^{p} \sum_{i=1}^{n_j} (X_{j,i} - \overline{X}_j)^2 + \sum_{j=1}^{p} \sum_{i=1}^{n_j} (\overline{X}_j - \overline{X})^2 \\ &= SSI + SSB \end{aligned}$$

We can interprete this decomposition as a decomposition to the "Sum of Squares within groups" and "Sum of square between groups". Under the H_0 should the variance between groups be relatively small and under the H_1 greater than certain value. In the following part we will derive from this intuitive assumption a test statistic.

Under the H_0 and the assumption of equality of Variances, follows $\frac{SSI}{\sigma^2} \sim \chi^2_{n-m}$ and $\frac{SSB}{\sigma^2} \sim \chi^2_{m-1}$, hence the test statistic

$$F = \frac{\frac{SSB}{m-1}}{\frac{SSI}{n-m}} \sim F_{m-1,n-m}$$

Where $F_{m-1,n-m}$ means Fischer-Snedecor distribution with $m-1$ and $n-m$ degrees of freedom. (see 4.4)

Hence the H_0 will be rejected on significance level α if $F > F_{m-1,n-m}(1-\alpha)$, where $F_{m-1,n-m}(1-\alpha)$ means $(1-\alpha)$ quantile of F-distribution with $m-1$ and $n-m$ degrees of freedom.

Implementation in XploRe

```
text=anova(datain)
      ttest runs ANOVA test on datain
```

The explanation of the parameters is following:

`datain`
: is a $n_1 \times p$ data set

`text`
: output text

In the output window we will with the ANOVA values also get levene test output and the description of groups. In this description we will get the number of elements in the each group, arithmetic mean, standard deviation and the 95% confidence interval for mean. So we have point estimations for mean and variance for each group, the confidence intervals can be used as intuitive, "pre-test" for mean-equality (if some intervals are disjunct, we can assume that there is "relevant" difference between the means, the problem is that, we can not just compare all these intervals, because we would got bigger probability of first error than our underlying significance level α, so we have to construct another tests as ANOVA to solve our problem.

$$I_i = (\overline{X_i} - t_{0.975,n-1}\frac{S_i}{\sqrt{n_i}}, \overline{X_i} + t_{0.975,n-1}\frac{S_i}{\sqrt{n_i}}) \text{ for } 1 \leq i \leq p$$

where $t_{0.975,n}$ means 0.975 quantile of the t-distribution with n degrees of freedom.

Example

`exanova.xpl`

We have following data set `gas` :

i	1.Group	2.Group	3.Group	4.Group	5.Group
1	91.7	91.7	92.4	91.8	93.1
2	91.2	91.9	91.2	92.2	92.9
3	90.9	90.9	91.6	92.0	92.4
4	90.6	90.9	91.0	91.4	92.4

We want to test if the gas additions have some impact at gas-anti-knocking properties . This data set (taken from (Rönz, B., 1997)) , hence we have 5

groups (5 different additions) with 4 observations in each group. We can solve our problem by testing the equality of means of these groups, let say at the significance level 5% . So from the statistical point of view we must test

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

against alternative hypothesis:

$$H_1 : \exists i, j,\ 1 \leq i \neq j \leq 5 : \mu_i \neq \mu_j$$

Let as assume, that these samples are independent and normaly distributed. The variance-equality assumption will be tested by the Levene test automatically. Hence we can run the ANOVA test, by typing:

```
library("stats")
x=read("gas.dat")
anova(x)
```

We get following output in output window:

```
"Groups description"
"--------------------------------------------------"
"count       mean     st.dev.  95% conf.i. for mean"
"--------------------------------------------------"
"    4    91.1000     0.4690    90.3489,   91.8511"
"    4    91.3500     0.5260    90.5077,   92.1923"
"    4    91.5500     0.6191    90.5585,   92.5415"
"    4    91.8500     0.3416    91.3030,   92.3970"
"    4    92.7000     0.3559    92.1301,   93.2699"
"--------------------------------------------------"
"               ANALYSIS OF VARIANCE               "
"--------------------------------------------------"
"Source of Variance  d.f.  Sum of Sq.              "
"--------------------------------------------------"
"Between Groups       4       6.1080"
"Within Groups       15       3.3700"
"Total               19       9.4780"
"--------------------------------------------------"
"F value      6.7967"
"sign.        0.0025"
"--------------------------------------------------"
"Levene Test for Homogenity of Variances           "
"--------------------------------------------------"
"   Statistic     df1     df2    Signif.           "
"      0.7385       4      15    0.5802            "
```

The third part of output window - Levene Test, have been explained above, so we will only take the results ($\alpha_{sig} = 0.5802 > 0.05 = \alpha$). So we have no reason

to reject equality of variances-hypothesis at the significance level 5%. So we can assume that also this condition for ANOVA is fulfilled.

We will focus on second part of the output window(ANALYSIS OF VARIANCE). we can see that the Total sum of squares = 9.4780 can be decomposed into Sum of Squares Within Groups = 3.3700 and Sum of Squares Between Groups = 6.1080. The F value is equal to $6.7967 = \frac{\frac{6.1080}{4}}{\frac{3.370}{15}}$, what is the value of our test statistic F, what corresponds to the significance = 0.0025, $0.0025 < 0.05$, where 0.05 is our significance level 5%. So H_0 can reject at the significance level 5%. So we can assume that the usage of gas addition have no influence to the anti-knocking properties.

Chapter 4

Appendix

4.1 Distributions

In this part we will define random distributions, which were used in the paper, and note important properties of these distributions.

DEFINITION 4.1 ***Normal distribution*** $N(\mu, \sigma^2)$ *is defined by density:*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ for } x \in \mathbb{R} \tag{4.1}$$

THEOREM 4.1 *If a random variable X follows $N(\mu, \sigma^2)$, then $EX = \mu$, $Var(X) = \sigma^2$.*

DEFINITION 4.2 χ^2_n ***distribution with*** n***-degrees of freedom*** *is defined by density:*

$$f_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2} \text{ for } x > 0 \tag{4.2}$$

where

$$\Gamma(t) = \int_0^\infty t^{a-1} e^{-t} dx \text{ for } a > 0$$

THEOREM 4.2 *If a random variable X follows χ^2_n, then $EX = n$, $Var(X) = 2n$.*

THEOREM 4.3 *Assume $X_1, X_2, \dots X_n$, n-independent random variables, where $X_i \sim N(0,1)$. Then*

$$Y = X_1^2 + X_2^2 + \cdots + X_n^2$$

follows χ^2-distribution with n degrees of freedom.

DEFINITION 4.3 ***t-distribution (Student distribution) with n- degrees of freedom*** *is defined by density:*

$$f_n(x) = \frac{\Gamma(\frac{n+1}{2}}{\Gamma(\frac{n}{2})\sqrt{\pi n}}(1 + \frac{x^2}{n})^{-(n+1)/2} \ for \ -\infty < x < \infty \tag{4.3}$$

where

$$\Gamma(t) = \int_0^\infty t^{a-1}e^{-t}dx \ for \ a > 0$$

THEOREM 4.4 *If a random variable X follows t_n, then $EX = 0$, $Var(X) = n/(n-2)$.*

THEOREM 4.5 *Assume X, Z, $X \sim N(0,1)$, $Z \sim \chi_n^2$ independent random variables, then random variable*

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

follows t-distribution with n degrees of freedom.

DEFINITION 4.4 ***F-distribution (Fisher-Snedecor distribution) with p, q degrees of freedom*** *is defined by density:*

$$f_{p,q} = \frac{\Gamma(\frac{p+q}{2})}{\Gamma(\frac{p}{2})\Gamma(\frac{q}{2})}(\frac{p}{q})^{p/2}x^{p/2-1}(1 + \frac{p}{q}x)^{-\frac{p+q}{2}} \tag{4.4}$$

THEOREM 4.6 *Assume $X \sim \chi_m^2, Y \sim \chi_n^2$, two independent random variables, implies that:*

$$Z = \frac{\frac{1}{m}X}{\frac{1}{n}Y}$$

follows F-distribution with m, n degrees of freedom.

4.2 XploRe list

4.2.1 f-test

```
proc(out)=ftest(d1,d2)
; ---------------------------------------------------------------------
; Library     stats
; ---------------------------------------------------------------------
; See_also    levene
; ---------------------------------------------------------------------
; Macro       ftest
; ---------------------------------------------------------------------
; Description ftest runs ftest
; ---------------------------------------------------------------------
; Usage       (out)=ftest(d1,d2)
; Input
; Parameter   d1
; Definition  n1 x 1 vector
; Parameter   d2
; Definition  n2 x 1 vector
; Output
; Parameter   out
; Definition  text output (string vector)
; ---------------------------------------------------------------------
; Example
; library("stats")
; x=normal(290,1)
; y=normal(290,1)
; ftest(x,y)
; ---------------------------------------------------------------------
; Result
; [1,] "------ F test ------"
; [2,] "--------------------"
; [3,] "testing s1>s2"
; [4,] "--------------------"
; [5,] "F value:     1.0801"
; [6,] "Sign.        0.5131"
; ---------------------------------------------------------------------
; Keywords    f-test, variance equality
; ---------------------------------------------------------------------
; Author      MB 010130
; ---------------------------------------------------------------------

    s1=var(d1)
    s2=var(d2)
```

```
    if (s1>s2)
        F=s1/s2
        t="testing s1>s2"
        n1=rows(d1)
        n2=rows(d2)
    else
        F=s2/s1
        t="testing s2>s1"
        n1=rows(d2)
        n2=rows(d1)
    endif
    sig=2*(1-cdff(F,n1-1,n2-1))

;constructing the text output
    out="------ F test ------"
    out=out|"--------------------"
    out=out|t
    out=out|"--------------------"
    out=out|string("F value: %10.4f",F)
    out=out|string("Sign.    %10.4f",sig)
endp
```

4.2.2 t-test

```
proc(tout)=ttest(d1,d2)
; ---------------------------------------------------------------------
; Library     stats
; ---------------------------------------------------------------------
; See_also    ANOVA
; ---------------------------------------------------------------------
; Macro       ttest
; ---------------------------------------------------------------------
; Description ttest runs t-test
; ---------------------------------------------------------------------
; Usage       (tout)=ttest(d1,d2)
; Input
; Parameter   d1
; Definition  n1 x 1 vector
; Parameter   d2
; Definition  n2 x 1 vector
; Output
; Parameter   tout
; Definition  text output (string vector)
; ---------------------------------------------------------------------
; Example
; library("stats")
```

```
; x=read("allbus.dat")
; man=paf(x,x[,1]==1)[,2]
; woman=paf(x,x[,1]==2)[,2]
; woman=woman|NaN.*matrix(rows(man)-rows(woman),1)
; x=man~woman
; ttest(man,woman)
; ----------------------------------------------------------------------
; Result
; [1,] " -------- t-test (For equality of Means) -------- "
; [2,] "-----------------------------------------------"
; [3,] "                t-value    d.f.     Sig.2-tailed "
; [4,] "Equal var.:      14.4144    714       0.0000"
; [5,] "Uneq. var.:      17.0589    685.27    0.0000"
; ----------------------------------------------------------------------
; Keywords   ttest, mean equality
; ----------------------------------------------------------------------
; Author      MB 010130
; ----------------------------------------------------------------------

    error(sum(isInf(d1))>0,"ttest:Inf detected in first vector")
    error(sum(isInf(d2))>0,"ttest:Inf detected in second vector")
    if(rows(d1)<>rows(d2));corection for levene input
        if(rows(d1)>rows(d2))
            d1l=d1
            d2l=d2|NaN.*matrix(rows(d1)-rows(d2),1)
        else
            d2l=d2
            d1l=d1|NaN.*matrix(rows(d2)-rows(d1),1)
        endif
    else ;no correction necessery
        d2l=d2
        d1l=d1
    endif

;   l=levene(d1l~d2l) ;levene test for var. eq.

; mean, var computation

    n1=sum(isNumber(d1))
    n2=sum(isNumber(d2))
    mean1=(1/n1).*(sum(replace(d1,NaN,0)))
    mean2=(1/n2).*(sum(replace(d2,NaN,0)))
    s1=var(replace(d1,NaN,mean1))
    s2=var(replace(d2,NaN,mean2))

; unequal variances
```

```
    T=(mean1-mean2)/(sqrt((s1/n1)+(s2/n2)))
    f1=((s1/n1)+(s2/n2))^2  ;df for T statistic
    f2=(((s1/n1)^2)/(n1-1)+((s2/n2)^2)/(n2-1))
    f=f1/f2

    if(f==floor(f)) ;next integer
        fl=f
    else
        fl=floor(f+1)
    endif
    s=2*(1-cdft(abs(T),fl))


;equal unknow variances

    Teq=(mean1-mean2)/sqrt(((n1+n2)/(n1*n2))
    *(((n1-1)*s1+(n2-1)*s2)/(n1+n2-2)))
    feq=n1+n2-2
    seq=2*(1-cdft(abs(Teq),feq))

; constructing output text

s0=" -------- t-test (For equality of Means) -------- "
st="--------------------------------------------------"
s1="               t-value    d.f.     Sig.2-tailed "
s2=string("Equal var.:  %10.4f",Teq)+string("   %4.0f",feq)
+string("   %10.4f",seq)
s3=string("Uneq. var.:  %10.4f",T)+string("    %6.2f",f)
+string("%10.4f",s)

out=s0|st|s1|s2|s3
;out=s0|st|s1|s2|s3|l
out
endp
```

4.2.3 ANOVA

```
proc(out)=anova(datain)
; ----------------------------------------------------------------------
; Library    stats
; ----------------------------------------------------------------------
; See_also   levene
; ----------------------------------------------------------------------
; Macro       anova
; ----------------------------------------------------------------------
```

```
; Description anova runs Simple Analysis of Variance
; ------------------------------------------------------------------
; Usage       (out)=anova(datain)
; Input
; Parameter   datain
; Definition  n x p data set
; Output
; Parameter   out
; Definition  text output (string array)
; ------------------------------------------------------------------
; Example
; library("stats")
; x=read("gas.dat")
; re=anova(x)
; re
; ------------------------------------------------------------------
; Result
; [ 1,] "Groups description"
; [ 2,] "-------------------------------------------------"
; [ 3,] "count       mean     st.dev.  95% conf.i. for mean"
; [ 4,] "-------------------------------------------------"
; [ 5,] "    4    91.1000     0.4690    90.3489,   91.8511"
; [ 6,] "    4    91.3500     0.5260    90.5077,   92.1923"
; [ 7,] "    4    91.5500     0.6191    90.5585,   92.5415"
; [ 8,] "    4    91.8500     0.3416    91.3030,   92.3970"
; [ 9,] "    4    92.7000     0.3559    92.1301,   93.2699"
; [10,] "-------------------------------------------------"
; [11,] "               ANALYSIS OF VARIANCE              "
; [12,] "-------------------------------------------------"
; [13,] "Source of Variance  d.f.  Sum of Sq.             "
; [14,] "-------------------------------------------------"
; [15,] "Between Groups       4       6.1080"
; [16,] "Within Groups       15       3.3700"
; [17,] "Total               19       9.4780"
; [18,] "-------------------------------------------------"
; [19,] "F value      6.7967"
; [20,] "sign.        0.0025"
; [21,] "-------------------------------------------------"
; [22,] "Levene Test for Homogenity of Variances          "
; [23,] "-------------------------------------------------"
; [24,] "   Statistic     df1     df2    Signif.          "
; [25,] "      0.7385       4      15    0.5802           "
; ------------------------------------------------------------------
; Keywords    ANOVA
; ------------------------------------------------------------------
; Author      MB 010130
```

```
; ---------------------------------------------------------------------

;input control
    error((exist(datain)<>1),"ANOVA:first argument must be numeric")
    error(dim(dim(datain))<>2,"ANOVA:invalid data format")
    error(sum(sum(isInf(datain)),2)>0,"ANOVA:
    Inf detected, quantlet stoped")
    nmcol=sum(isNumber(datain))
    nmtot=sum(nmcol,2)

    datacnt=datain
    ;means
    meancold=sum(replace(datacnt,NaN,0))/nmcol
    meantotd=sum(sum(replace(datacnt,NaN,0)),2)/nmtot

    ;variances
    i=1
    datactmp=datacnt[,i]-meancold[,i].*matrix(rows(datacnt),1)
    ssclt=replace(datactmp,NaN,0)'*replace(datactmp,NaN,0)
    ; ss of first column

    i=i+1
    while(i<=dim(datacnt)[2])
    x=datacnt[,i]-meancold[,i].*matrix(rows(datacnt),1)
    datactmp=datactmp~x
    ssclt=ssclt~(replace(x,NaN,0)'*replace(x,NaN,0))
    ;ss i-th column
    i=i+1
    endo

    ;sum of squares
    ssig=sum(ssclt,2) ;ss in groups
    ssbgc=nmcol.*(meancold-meantotd).*(meancold-meantotd)
    ;ss between group
    ssbg=sum(ssbgc,2)

    ;F value
    df1=cols(datain)-1
    df2=nmtot-cols(datain)
    error(ssig==0,"ANOVA:constant columns")
    F=(df2/df1)*(ssbg/ssig)
    sig=1-cdff(F,df1,df2)

varcol=sqrt(ssclt./(nmcol-1))
qf=qft(0.975*matrix(rows(nmcol),cols(nmcol)),nmcol-1)
cicol=(meancold-qf.*((varcol)/sqrt(nmcol))
```

```
|meancold+qf.*((varcol)/sqrt(nmcol)))'
out="Groups description"
out=out|"-------------------------------------------------"
out=out|"count       mean     st.dev.  95% conf.i. for mean"
out=out|"-------------------------------------------------"
out=out|string(" %4.0f",nmcol')+string(" %10.4f",meancold')
+string(" %10.4f",(varcol)')+string(" %10.4f",cicol[,1])
+string(",%10.4f",cicol[,2])

s0="-------------------------------------------------"
s1="                ANALYSIS OF VARIANCE              "
s11="Source of Variance  d.f.  Sum of Sq.             "
s12="Between Groups  "+string("  %4.0f",df1)+string(" %12.4f",ssbg)
s13="Within Groups   "+string("  %4.0f",df2)+string(" %12.4f",ssig)
dt=df1+df2
sst=ssbg+ssig
s14="Total           "+string("  %4.0f", dt)+string(" %12.4f",sst)
s3=string("F value  %10.4f",F)
s31=string("sign.    %10.4f",sig)
le=levene(datain)
text=out|s0|s1|s0|s11|s0|s12|s13|s14|s0|s3|s31|le
out=text
endp
```

4.2.4 Levene

```
proc(out)=levene(datain)
; ---------------------------------------------------------------------
; Library     stats
; ---------------------------------------------------------------------
; See_also    ANOVA
; ---------------------------------------------------------------------
; Macro       levene
; ---------------------------------------------------------------------
; Description levene runs Levene-test
; ---------------------------------------------------------------------
; Usage       (out)=levene(datain)
; Input
; Parameter   datain
; Definition  n x p data set
; Output
; Parameter   out
; Definition  text output (string array)
; ---------------------------------------------------------------------
; Example
; library("stats")
```

```
; x=read("gas.dat")
; levene(x)
; ----------------------------------------------------------------------
; Result
; [1,] "-------------------------------------------------"
; [2,] "Levene Test for Homogenity of Variances          "
; [3,] "-------------------------------------------------"
; [4,] "   Statistic     df1     df2    Signif.          "
; [5,] "      0.7385       4      15    0.5802           "
; ----------------------------------------------------------------------
; Keywords    levene-test, variance-equality
; ----------------------------------------------------------------------
; Author      MB 010130
; ----------------------------------------------------------------------

;input control
    error((exist(datain)<>1),"LEVENE:first argument must be numeric")
    error(dim(dim(datain))<>2,"LEVENE:invalid data format")
    error(sum(sum(isInf(datain)),2)>0,"LEVENE:Inf detected,
    quantlet stoped")

;construction of absolute deviation

    nmcol=sum(isNumber(datain))
    nmtot=sum(nmcol,2)

    meancol=sum(replace(datain,NaN,0))/nmcol
    meantot=sum(sum(replace(datain,NaN,0)),2)/nmtot
    datacnt=datain-meancol.*matrix(rows(datain),cols(datain))
    datacnt=abs(datacnt)


;running ANOVA on datacnt

    ;means
    meancold=sum(replace(datacnt,NaN,0))/nmcol
    meantotd=sum(sum(replace(datacnt,NaN,0)),2)/nmtot

    ;variances
    i=1
    datactmp=datacnt[,i]-meancold[,i].*matrix(rows(datacnt),1)
    ssclt=replace(datactmp,NaN,0)'*replace(datactmp,NaN,0)
    ; ss of first column

    i=i+1
    while(i<=dim(datacnt)[2])
```

```
    x=datacnt[,i]-meancold[,i].*matrix(rows(datacnt),1)
    datactmp=datactmp~x
    ssclt=ssclt~(replace(x,NaN,0)'*replace(x,NaN,0)) ;ss i-th column
    i=i+1
    endo

    ;sum of squares
    ssig=sum(ssclt,2) ;ss in groups
    ssbgc=nmcol.*(meancold-meantotd).*(meancold-meantotd)
    ;ss between group
    ssbg=sum(ssbgc,2)

    ;F value
    df1=cols(datain)-1
    df2=nmtot-cols(datain)
    error(ssig==0,"LEVENE:constant columns")
    F=(df2/df1)*(ssbg/ssig)
    sig=1-cdff(F,df1,df2)


s0="--------------------------------------------------"
s1="Levene Test for Homogenity of Variances          "
s2="   Statistic     df1     df2    Signif.          "
s3=string("  %10.4f",F)+string("    %4.0f",df1)
    +string("    %4.0f",df2)+string("%10.4f",sig)+"           "

text=s0|s1|s0|s2|s3
out=text

endp
```

4.2.5 Spread and level Plot

grspleplot

```
proc(sple)=grspleplot(data)
; ---------------------------------------------------------------------
; Library     graphic
; ---------------------------------------------------------------------
; See_also    dispspleplot
; ---------------------------------------------------------------------
; Macro       grspleplot
; ---------------------------------------------------------------------
; Description grspleplot generates a graphic-object with spread and level plot
; ---------------------------------------------------------------------
; Usage       (sple)=grspleplot(data)
```

```
; Input
; Parameter   data
; Definition  n x p dataset
; Output
; Parameter   sple
; Definition  graphical object
; ---------------------------------------------------------------------
; Example
; library("graphic")
; x=read("allbus.dat")
; man=paf(x,x[,1]==1)[,2]
; woman=paf(x,x[,1]==2)[,2]
; woman=woman|NaN.*matrix(rows(man)-rows(woman),1)
; x=man~woman
; gr=grspleplot(x)
; di=createdisplay(1,1)
; show(di,1,1,gr)
; ---------------------------------------------------------------------
; Result     there is new display with spread and level plot
; ---------------------------------------------------------------------
; Keywords    spread and level plot
; ---------------------------------------------------------------------
; Author      MB 010130
; ---------------------------------------------------------------------

error(cols(data)<=1,"GRSPLEPLOT:min 2 columns expected")
error(sum(sum(isInf(data),2),1)>0,"GRSPLEPLOT: inf detected")

n1=sum(isNumber(data),1)+1
iqr=matrix(1,cols(data)) ;int.quart. range
med=matrix(1,cols(data))

i=1

while(i<=cols(data))
    irqv=paf(data[,i],isNumber(data[,i]))
    med[,i]=quantile(irqv,1/2)
    iqr[,i]=quantile(irqv,3/4)-quantile(irqv,1/4)
    i=i+1

endo
sple=trans(med|iqr)
endp
```

dispspleplot

```
proc()=dispspleplot(dis,x,y,data)
; ---------------------------------------------------------------------
; Library     graphic
; ---------------------------------------------------------------------
; See_also    grspleplot, plotspleplot
; ---------------------------------------------------------------------
; Macro       dispspleplot
; ---------------------------------------------------------------------
; Description dispspleplot draws a spread and level plot into specific
    display
; ---------------------------------------------------------------------
; Usage       ()=dispspleplot(dis,x,y,data)
; Input
; Parameter   dis
; Definition  display
; Parameter   x
; Definition  scalar
; Parameter   y
; Definition  scalar
; Parameter   data
; Definition  n x p data set
; Output
; ---------------------------------------------------------------------
; Example
; library("graphic")
; di=createdisplay(1,1)
; x=read("allbus.dat")
; dispspleplot(di,1,1,x)
; ---------------------------------------------------------------------
; Result      there is spread and level plot in the display di
; ---------------------------------------------------------------------
; Keywords    spread and level plot
; ---------------------------------------------------------------------
; Author      MB 010130
; ---------------------------------------------------------------------

    gr=grspleplot(data)
    show(dis,x,y,gr)

endp
```

plotspleplot

```
proc()=plotspleplot(data)
```

```
; ---------------------------------------------------------------------
; Library     plot
; ---------------------------------------------------------------------
; See_also    grspleplot, dispspleplot
; ---------------------------------------------------------------------
; Macro       plotspleplot
; ---------------------------------------------------------------------
; Description plotspleplot runs spread and level plot
; ---------------------------------------------------------------------
; Usage       ()=plotspleplot(data)
; Input
; Parameter   data
; Definition  n x p dataset
; Output
; ---------------------------------------------------------------------
; Example
; library("plot")
; x=read("allbus.dat")
; man=paf(x,x[,1]==1)[,2]
; woman=paf(x,x[,1]==2)[,2]
; woman=woman|NaN.*matrix(rows(man)-rows(woman),1)
; x=man~woman
; plotspleplot(x)
; ---------------------------------------------------------------------
; Result   there is a new window with spread and level plot
; and following output:
; [1,] " ------- Spread-and-level Plot------- "
; [2,] "  slope of LN of level and LN spread  "
; [3,] "--------------------------------------"
; [4,] " Slope =                0.338"
; [5,] "Power transf. est.      0.662"
; ---------------------------------------------------------------------
; Keywords    spread and level plot
; ---------------------------------------------------------------------
; Author      MB 010130
; ---------------------------------------------------------------------

  i=selectitem("Power estimation ?",#("power estimation",
  "no power estimation"),"single")

    di=createdisplay(1,1)
    gr=grspleplot(data)
    show(di,1,1,gr)
    setgopt(di,1,1,"title","Spread & Level Plot","xlabel","
    Level (median)","ylabel","Spread - IRQ")
```

```
    ;computing the slope
    m=mean(gr)
    l=gr[,1]-m[,1]
    s=gr[,2]-m[,2]

if(i[1,1]==0) ;no power estimation
 error((l'*l)==0,"PLOTSPLEPLOT:means always equal")
    slope=(l'*s)/(l'*l) ;slope

  ;constructing the text output
    out=    " --- Spread-and-level Plot--- "
    out=out|"------------------------------"
    out=out|string(" Slope =               %6.3f",slope)

 out
else
gr=log(gr)
    m=mean(gr)
    l=gr[,1]-m[,1]
    s=gr[,2]-m[,2]

 error((l'*l)==0,"PLOTSPLEPLOT:means always equal")
    slope=(l'*s)/(l'*l) ;slope
    out=    " ------- Spread-and-level Plot------- "
    out=out|"  slope of LN of level and LN spread  "
    out=out|"-------------------------------------"
    out=out|string(" Slope =               %6.3f",slope)
        out=out|string("Power transf. est.     %6.3f",1-slope)
out
endif
endp
```

Bibliography

Anděl, J., (1985). *Matematická statistika*, Alfa-Prag

Dupač, V., Hušková, M., (1999). *Pravděpodobnost a Matematická statistika*, Karolinum, Prag

Härdle, W., Klinke, S. & Müller, M., (1999). *XploRe : Learning Guide*, Springer-Verlag.

Härdle, W., Hlávka, Z. & Klinke, S.,, (2000). *XploRe : Application Guide*, Springer-Verlag.

Härdle, W. & Simar, L., (2000). *Applied Multivariate Statistical Analysis*, Springer-Verlag.

Härdle, W., Müller, M., Sperlich, S., & Werwatz, A., (1999). *Non- and Semiparametric Modelling*, Humboldt-Universität zu Berlin.

Rönz, B., (1997). *Computergestützte Statistik I*, Humboldt-Universität zu Berlin.

Rönz, B., (1999). *Computergestützte Statistik II*, Humboldt-Universität zu Berlin.