



**ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ**

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

Elements of Statistics and Probability

Data Science

LECTURE 4 – Hypothesis testing

Xanthi Pedeli

Assistant Professor, xpedeli@aueb.gr

Department of Statistics, AUEB

Notes by Ioannis Ntzoufras, Professor

Department of Statistics, AUEB

4. Hypothesis Tests & Pairwise associations

In this module we will examine

- Exploratory data analysis
- Testing for normality
- General principles for hypothesis testing
- Standard errors and p-values
- Hypothesis tests for one sample (mean, variance, kurtosis, skewness)
- Association between one quantitative and one qualitative variable
- Association between two categorical variables
- Association between two ordinals
- Correlation between two quantitative variables

4.2. Testing for normality

- QQ plots
- PDF plots
- CDF Plots
- Testing for symmetry & skewness
- Testing for kurtosis
- Kolmogorov-Smirnov test
- Shapiro-Wilks test
- Other tests

4.2. Testing for normality



Salary dataset from Norusis

This is a data file containing information on 474 employees hired by a Midwestern bank between 1969 and 1971.

It was created for an Equal Employment Opportunity (EEO) court case involving wage discrimination.

4.2. Testing for normality

Salary dataset from Norusis

ID	Employee code
SALBEG	Beginning salary
SEX	Gender of employee (0=Female; 1=Male)
TIME	Job Seniority (months)
AGE	Age of Employee (years and fraction)
SALNOW	Current salary
EDLEVEL	Educational Level (years)
WORK	Work Experience (years x 100)
JOBCAT	Employment category (1 Clerical, 2 Office trainee, 3 Security officer, 4 College trainee, 5 Exempt employee, 6 MBA trainee 7 Technical)
MINORITY	Minority Classification (0=White; 1=Nonwhite)

4.2. Testing for normality

Dataset salary from Norusis

```
library(foreign)
salary<-read.spss('salary.sav', to.data.frame=T)
names(salary)
```

```
> names(salary)
[1] "id"      "salbeg"  "sex"     "time"    "age"     "salnow"
[7] "edlevel" "work"    "jobcat"  "minority" "sexrace"
> |
```

```
head(salary)
```

```
sapply(salary, class)
```

```
      id      salbeg      sex      time      age      salnow      edlevel      work
"numeric" "numeric" "factor" "numeric" "numeric" "numeric" "numeric" "numeric"
      jobcat  minority  sexrace
"factor"  "factor"  "factor"
```

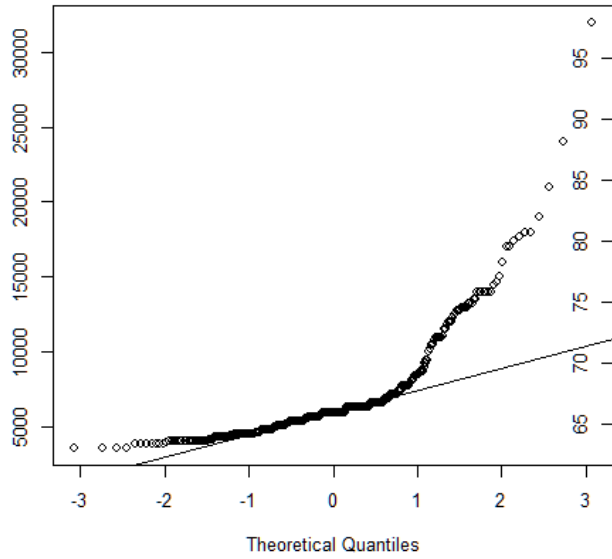
```
sal.num <- salary[,which(sapply(salary, class)=="numeric")]
```

4.2. Testing for normality

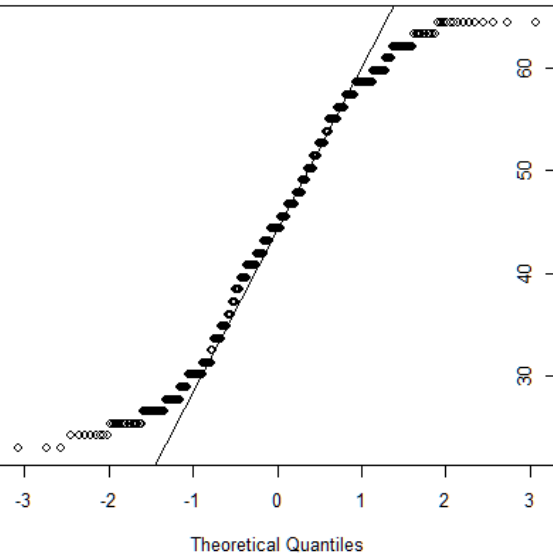
QQ-plots

```
y<-sal.num[,-1]
p<-ncol(y)
par(mfrow=c(2,3))
for (i in 1:p){
    qqnorm(y[,i])
    qqline(y[,i])
}
```

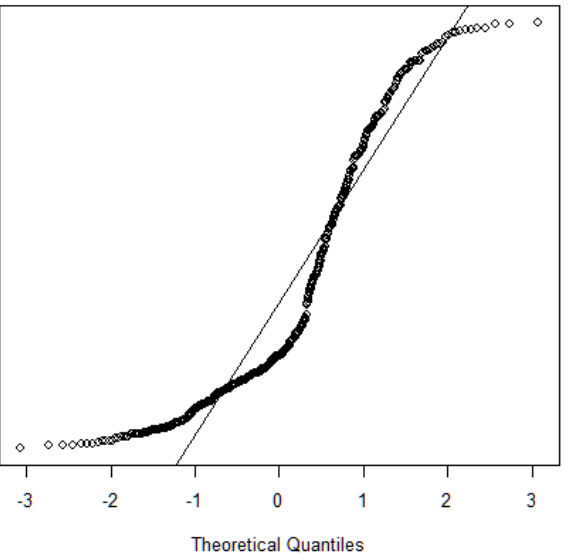
Normal Q-Q Plot



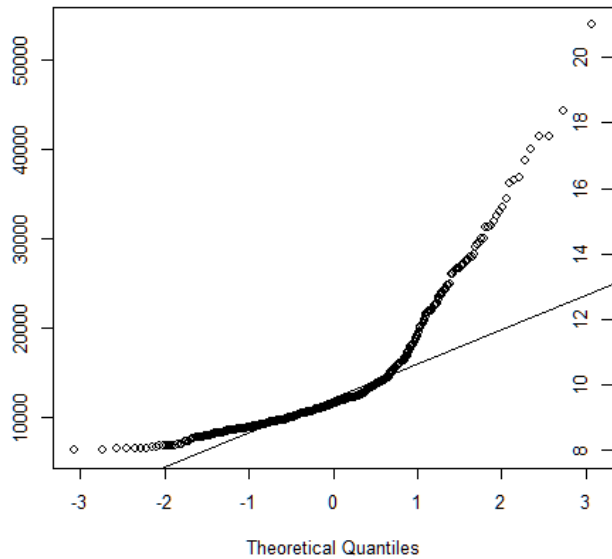
Normal Q-Q Plot



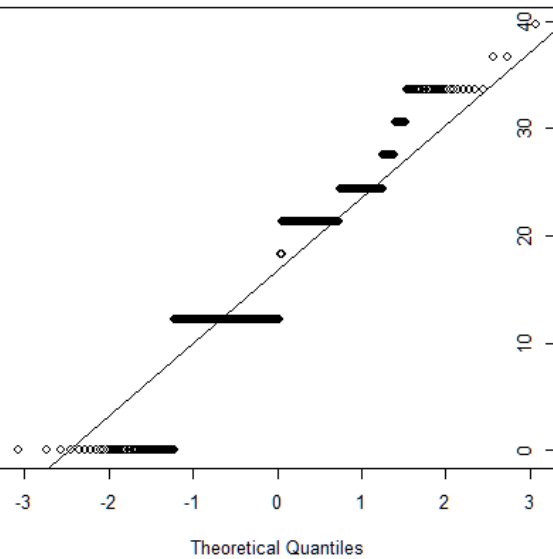
Normal Q-Q Plot



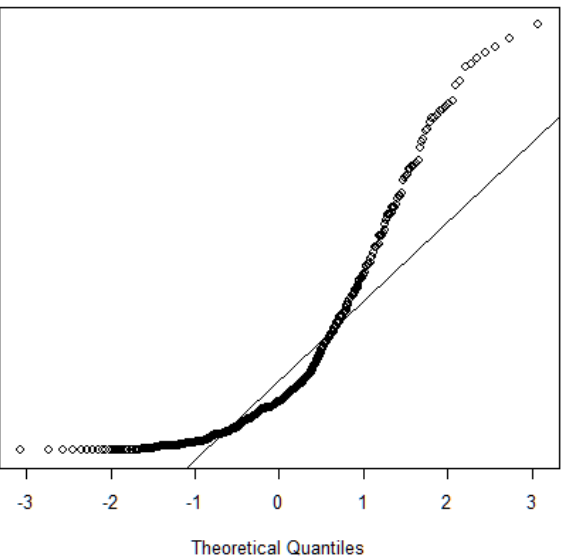
Normal Q-Q Plot



Normal Q-Q Plot



Normal Q-Q Plot



4.2. Testing for normality

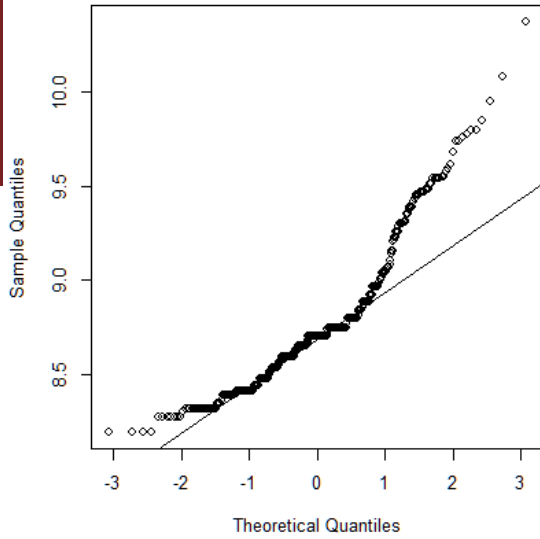
QQ-plots

```
sapply(y,min)
```

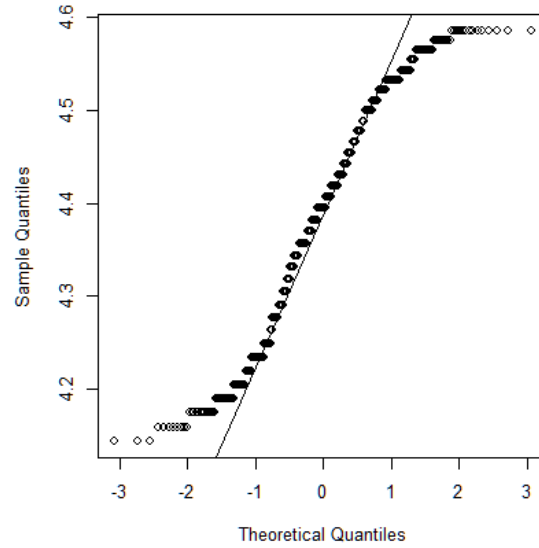
```
salbeg      time      age      salnow  edlevel      work  
  3600         63       23      6300         8         0
```

```
y[,6] <- y[,6]+0.5  
y<-log(y)  
p<-ncol(y)  
par(mfrow=c(2,3))  
for (i in 1:p){  
    qqnorm(y[,i])  
    qqline(y[,i])  
}
```

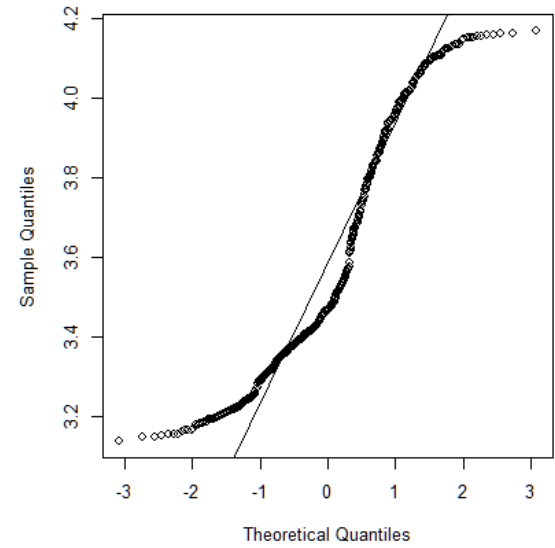
Normal Q-Q Plot



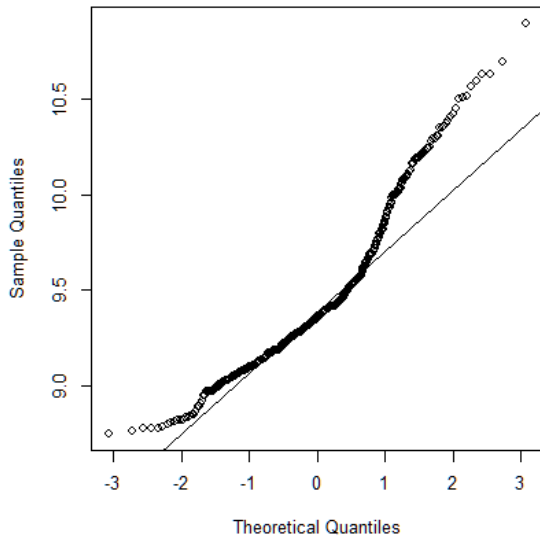
Normal Q-Q Plot



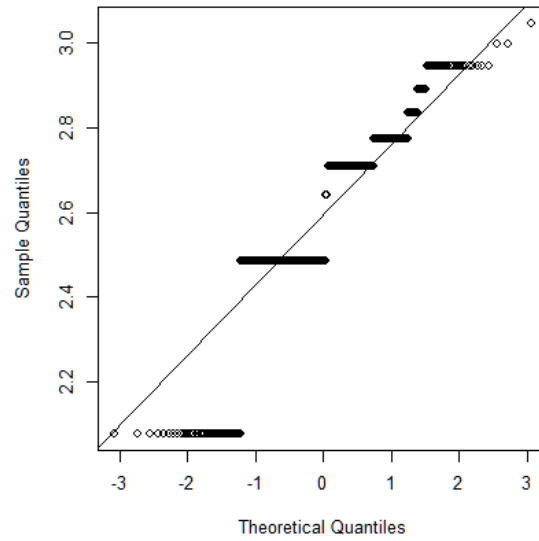
Normal Q-Q Plot



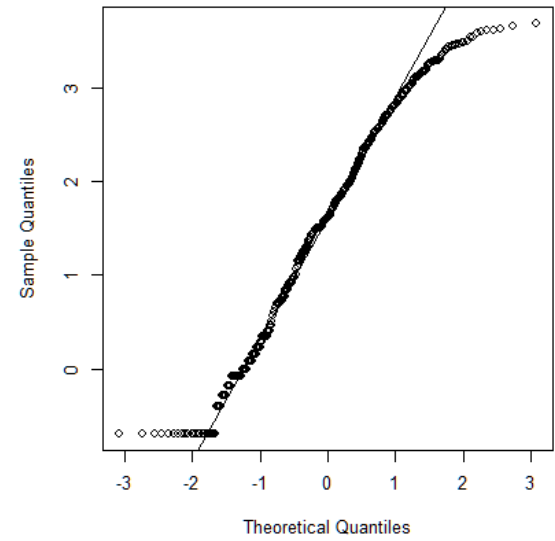
Normal Q-Q Plot

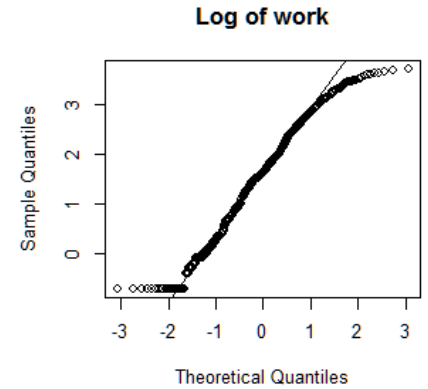
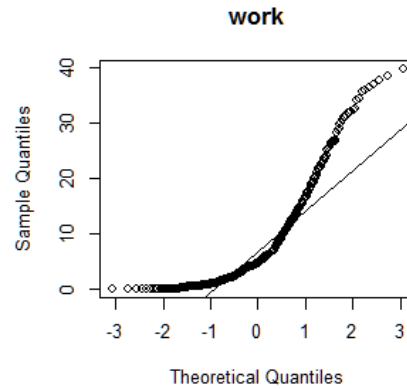
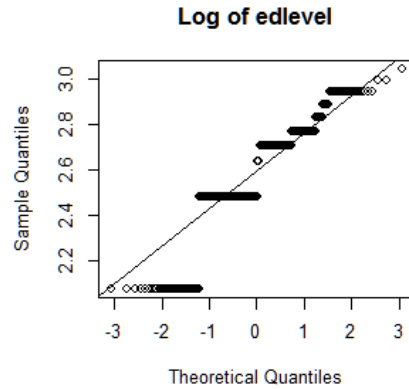
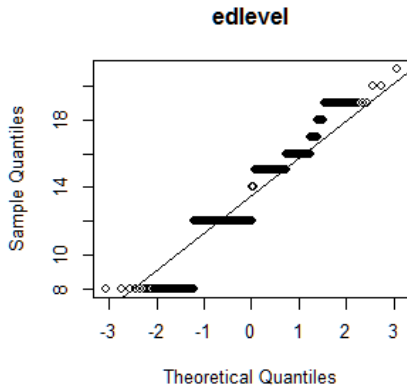
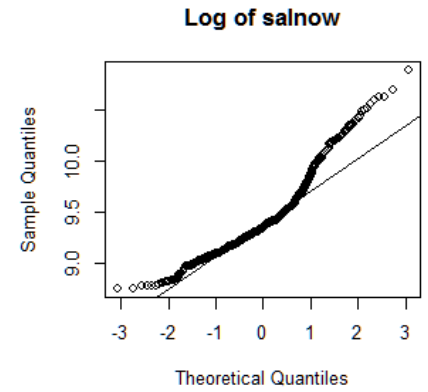
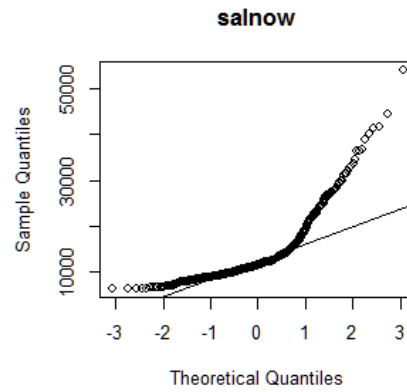
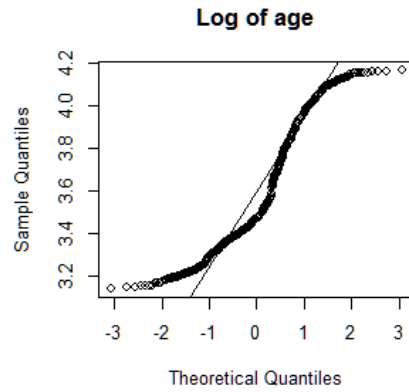
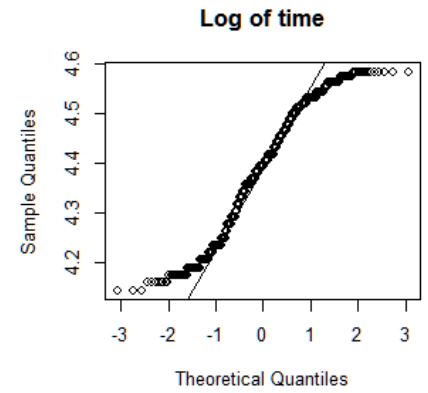
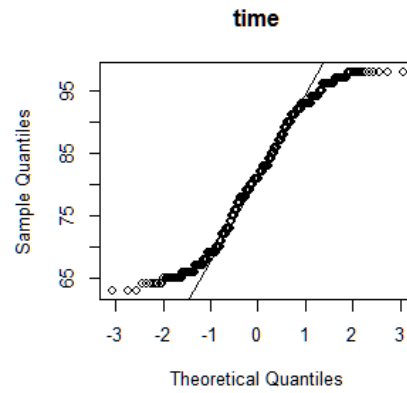
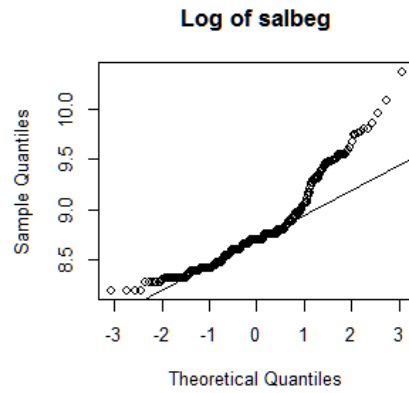
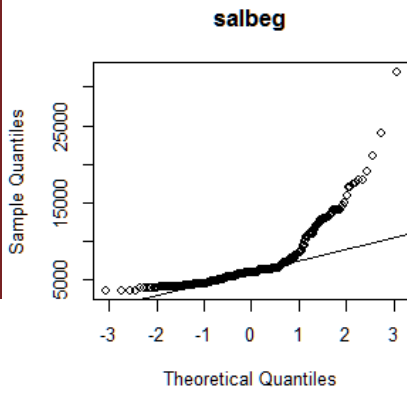


Normal Q-Q Plot



Normal Q-Q Plot





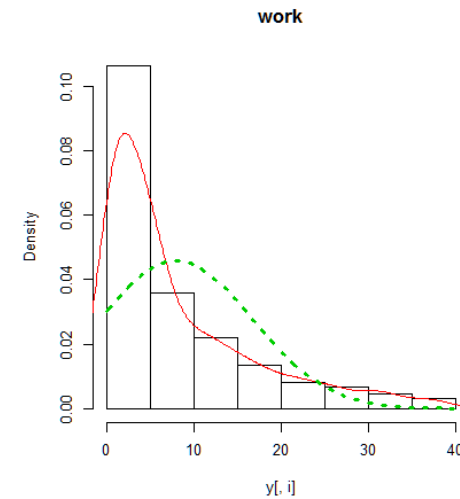
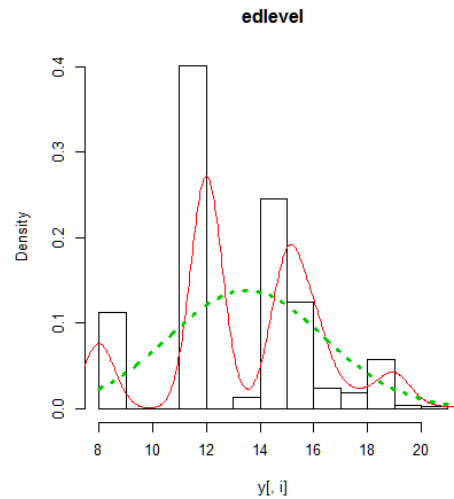
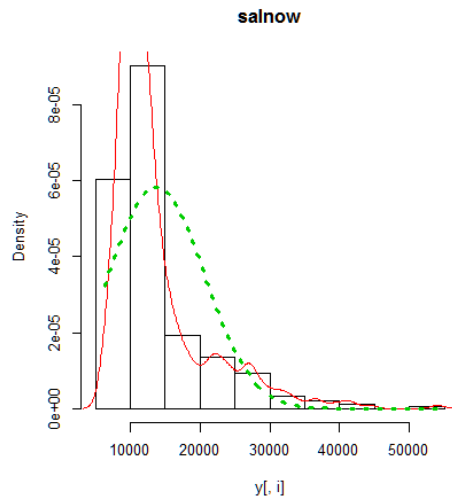
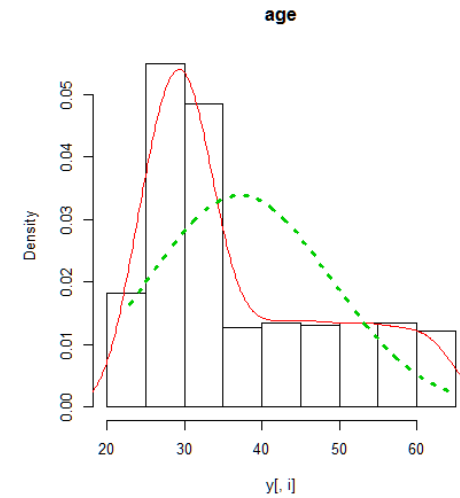
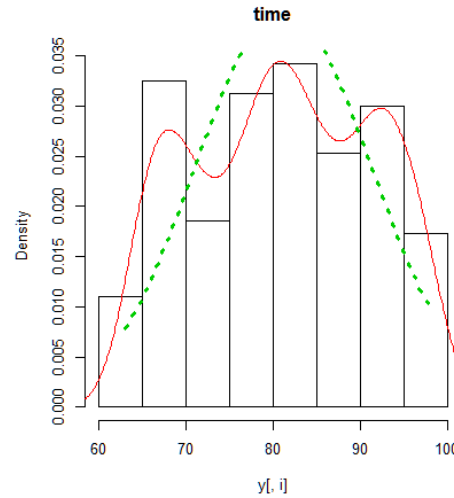
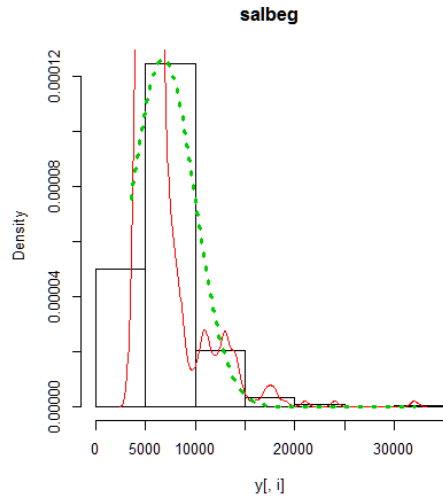
4.2. Testing for normality

Density plots

```
y<-sal.num[,-1]
p<-ncol(y)
par(mfrow=c(2,3))
for (i in 1:p){
    hist(y[,i], main=names(y)[i], probability=TRUE)
    lines(density(y[,i]), col=2)
    index <- seq( min(y[,i]), max(y[,i]),
length.out=100)
    ynorm <- dnorm( index, mean=mean(y[,i]),
sd(y[,i]) )
    lines( index, ynorm, col=3, lty=3, lwd=3 )
}
```

4.2. Testing for normality

Density plots



4.2. Testing for normality

Kolmogorov-Smirnov test

The KS test quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution. In the special case of testing for normality of the distribution, samples are standardized and compared with a standard normal distribution.

```
> for(i in 1:length(y)){  
+ print(ks.test(y[,i], y='pnorm'))  
+ }
```

One-sample Kolmogorov-Smirnov test

```
data: y[, i]  
D = 1, p-value < 2.2e-16  
alternative hypothesis: two-sided
```

One-sample Kolmogorov-Smirnov test

```
data: y[, i]  
D = 1, p-value < 2.2e-16  
alternative hypothesis: two-sided
```

One-sample Kolmogorov-Smirnov test

```
data: y[, i]  
D = 1, p-value < 2.2e-16  
alternative hypothesis: two-sided
```

One-sample Kolmogorov-Smirnov test

```
data: y[, i]  
D = 1, p-value < 2.2e-16  
alternative hypothesis: two-sided
```

One-sample Kolmogorov-Smirnov test

```
data: y[, i]  
D = 1, p-value < 2.2e-16  
alternative hypothesis: two-sided
```



4.2. Testing for normality

Kolmogorov-Smirnov test



```
> lapply(y,ks.test, y='pnorm')
$salbeg
```

```
One-sample Kolmogorov-Smirnov test
```

```
data: X[[1L]]
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
$time
```

```
One-sample Kolmogorov-Smirnov test
```

```
data: X[[2L]]
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
$age
```

```
One-sample Kolmogorov-Smirnov test
```

```
data: X[[3L]]
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
$salnow
```

```
One-sample Kolmogorov-Smirnov test
```

```
data: X[[4L]]
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
$edlevel
```

```
One-sample Kolmogorov-Smirnov test
```

4.2. Testing for normality

Lilliefors Kolmogorov-Smirnov test



- Modification of the Kolmogorov-Smirnov test
- More conservative (less likely to show that data is normally distributed)

```
> library('nortest')
> sapply(y,lillie.test)
      salbeg                time
statistic 0.2518788          0.08255521
p.value   7.444209e-85        3.377694e-08
method    "Lilliefors (Kolmogorov-Smirnov) normality test" "Lilliefors (Kolmogorov-Smirnov) normality test"
data.name "X[[1L]]"          "X[[2L]]"
      age
statistic 0.1898205          0.2078544
p.value   3.475671e-47        5.44953e-57
method    "Lilliefors (Kolmogorov-Smirnov) normality test" "Lilliefors (Kolmogorov-Smirnov) normality test"
data.name "X[[3L]]"          "X[[4L]]"
      edlevel
statistic 0.2100917          0.1845442
p.value   2.849403e-58        1.709136e-44
method    "Lilliefors (Kolmogorov-Smirnov) normality test" "Lilliefors (Kolmogorov-Smirnov) normality test"
data.name "X[[5L]]"          "X[[6L]]"
> lapply(v,lillie.test)
```


4.2. Testing for normality

Lilliefors Kolmogorov-Smirnov test



```
> lapply(y,lillie.test)
```

```
$salbeg
```

```
Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data: X[[1L]]
```

```
D = 0.2519, p-value < 2.2e-16
```

```
$time
```

```
Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data: X[[2L]]
```

```
D = 0.0826, p-value = 3.378e-08
```

```
$age
```

```
Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data: X[[3L]]
```

```
D = 0.1898, p-value < 2.2e-16
```

```
$salnow
```

```
Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data: X[[4L]]
```

```
D = 0.2079, p-value < 2.2e-16
```

```
$edlevel
```

```
Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data: X[[5L]]
```

```
D = 0.2101, p-value < 2.2e-16
```

```
$work
```

```
Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data: X[[6L]]
```

```
D = 0.1845, p-value < 2.2e-16
```

```
> |
```

4.3. *General principles for hypothesis testing*

Hypothesis tests:

- Use a null hypothesis expressed with an equality concerning a set of parameters $H_0: \theta = \theta_0$
- The alternative hypothesis is usually the complementary statement e.g. $H_1: \theta \neq \theta_0$
- We use a **Statistic** which usually measures the distance between the **observed data** and the **null hypothesis**
- We use the distribution under the null hypothesis to find thresholds for rejecting the null
- So the reasoning is that if the Statistic is large i.e. the distance between the observed data and the null hypothesis is large, then something went wrong with our reasoning and probably this is the original hypothesis

4.3. *General principles for hypothesis testing*



Hypothesis tests:

- We do reject the null hypothesis
- We do not accept the alternative (one study cannot “prove” anything, but it can provide evidence for (or against) a hypothesis)

4.3. *Standard errors & p values*

What is a p-value?

The P - Value Interpretation Table

Less than 0.01	Strong evidence against Null Hypothesis, Very statistically significant
0.01 to 0.05	Some evidence against Null Hypothesis, statistically significant
Greater than 0.05	Insufficient evidence against Null Hypothesis

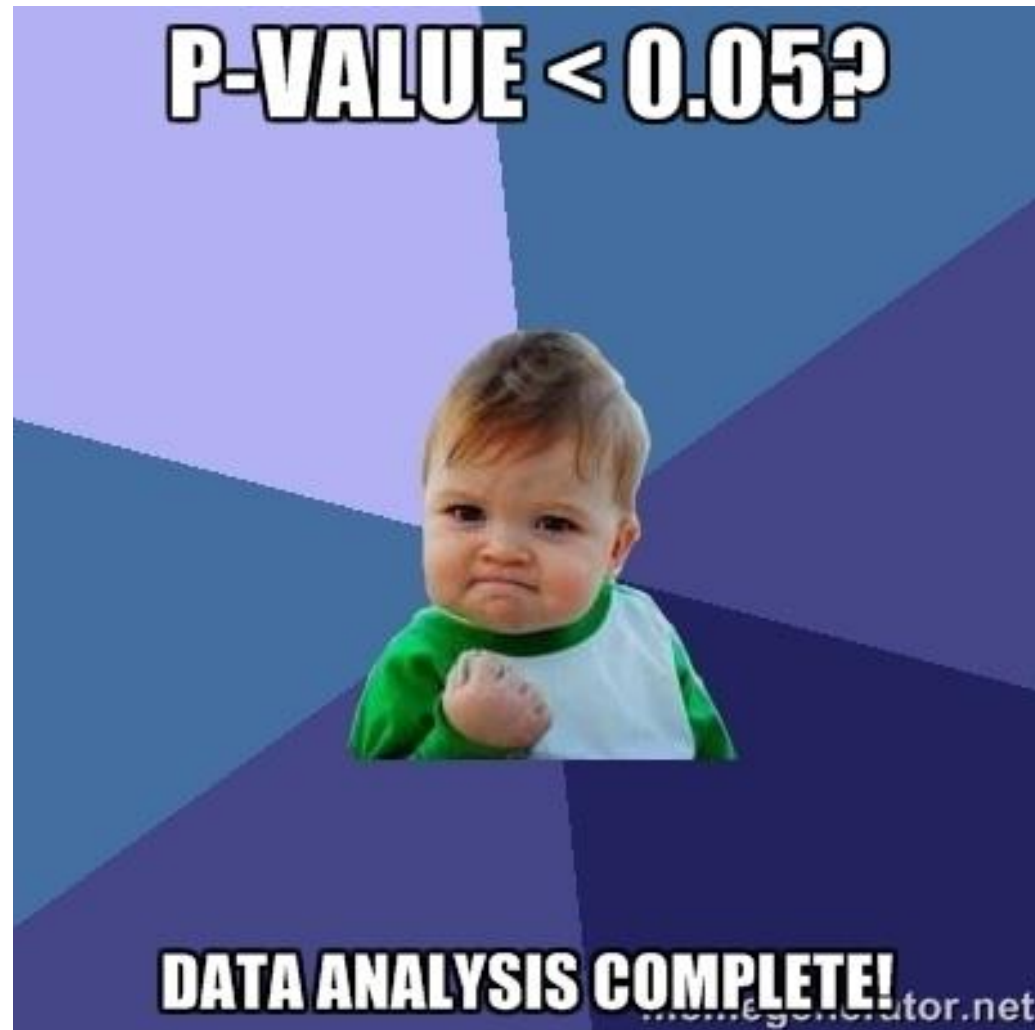
$0.05 < p < 0.10$ Not significant at $\alpha=5\%$

Significant at $\alpha=10\%$

Often reported as marginal significant result

4.3. *Standard errors & p values*

What is a p-value?



4.3. *Standard errors & p values*



What is a p-value?

https://www.youtube.com/watch?time_continue=17&v=9jW9G8MO4PQ

4.3. *Standard errors & p values*

What is a p-value?

- It is not the probability that the null hypothesis is not correct!
- It is the probability to observe a dataset with the same or more distant statistic than the one already observed if the null is true
- Hence if the p-value is small it means that not a lot of datasets are worse than the one you observed under the null
- So either the null is wrong and should be rejected or we are unlucky
- How much unlucky? The probability to reject the null when it is true is equal to α (usually equal to 5% or 1%)

4.3. *Standard errors & p values*

What is a standard error?

- Standard deviation or standard error
- Is it equal to $s/n^{1/2}$?
- What about estimators?

4.3. *Standard errors & p values*

What is a standard error?

- Estimators are functions of random variables (before observing the data)
- Estimates are the specific observed value for a specific sample
- Estimators are random variables. Therefore:
 - for each sample we have different value (estimate)
 - They have a mean
 - They have a variance

Standard error of an estimator is simply the standard deviation of an estimator

It is called error because it measures how close we are in the true value (i.e. the mean in the unbiased estimators)

4.3. *Standard errors & p values*

What is a standard error?

Good estimators

- They are unbiased
(i.e. if we consider a lot of samples, the mean of the estimates will be equal to the true value)
- They have small standard error
(i.e. the estimates are close to the mean/true value)

4.4. Hypothesis tests for a single continuous variable

Hypothesis test for the mean

- $H_0: \mu = \mu_0$ vs $H_1: \mu \neq \mu_0$
- Z-test for normal data with known variance
- T-test for normal data with unknown variance
- Assumption: Normality or large sample $n > 30$ or > 50

Hypothesis test for the median

- $H_0: M = M_0$ vs $H_1: M \neq M_0$
- Wilcoxon sign rank test

4.4. Hypothesis tests for a single continuous variable



Hypothesis test for the median

- $H_0: M = M_0$ vs $H_1: M \neq M_0$
- Wilcoxon sign rank test

Assumptions. There is no requirement that the individual X_i and Y_i be independent, only that the pairs $(X_1, Y_1), \dots, (X_n, Y_n)$, and therefore the resulting differences Z_1, \dots, Z_n , be mutually independent. Indeed, in most applications, the individual X_i and Y_i are dependent. For paired replicates data, the symmetry part of Assumption A2 is often inherently satisfied. In particular, if each X_i and $Y_i, i = 1, \dots, n$, arise from populations differing only in location (i.e., the only treatment “effect” is a change in location), then the $(Z_i - \theta)$ ’s come from populations that are symmetric about zero. (This is, in fact, true under more general conditions.)

Source: Hollander, Wolfe & Chicken (2014). *Nonparametric Statistical Methods*. Wiley & sons.

4.4. Hypothesis tests for a single continuous variable

Έλεγχος για ένα δείγμα
(1 ποσοτική μεταβλητή)

Ναι

Είναι η μεταβλητή μας κανονική;
(Έλεγχος Κανονικότητας)
[SW αν $n \leq 50$ - KS+SW αν $n > 50$]

Όχι

Ναι

Είναι το δείγμα μεγάλο
($n > 50$)

Είναι ο μέσος κατάλληλο μέτρο
περιγραφής της κεντρικής θέσης;

Ναι (?)

Όχι

Όχι

**Έλεγχος για τη μέση τιμή
t-test για ένα δείγμα**

**Έλεγχος για τη διάμεσο
Wilcoxon test για ένα δείγμα**

4.4. Hypothesis tests for a single continuous variable

Hypothesis test for the mean

```
> x1<-rnorm(100)
> t.test(x1)
```

One Sample t-test

```
data: x1
t = 1.7548, df = 99, p-value = 0.08238
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.01969387  0.32103009
sample estimates:
mean of x
0.1506681
```

```
> temp<-t.test(x1)
> names(temp)
[1] "statistic" "parameter" "p.value" "conf.int" "estimate"
[6] "null.value" "alternative" "method" "data.name"
> temp$sp.val
[1] 0.08238023
> temp$estim
mean of x
0.1506681
```

4.4. Hypothesis tests for a single continuous variable

Hypothesis test for the mean

```
> shapiro.test(x1)
```

```
Shapiro-Wilk normality test
```

```
data: x1  
W = 0.9954, p-value = 0.9844
```

```
> library('nortest')  
> lillie.test(x1)
```

```
Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data: x1  
D = 0.0495, p-value = 0.7915
```

```
> ks.test(x1, 'pnorm')
```

```
One-sample Kolmogorov-Smirnov test
```

```
data: x1  
D = 0.1238, p-value = 0.09332  
alternative hypothesis: two-sided
```

4.4. Hypothesis tests for a single continuous variable

```
> y<-salary
> logdiff<-log(y$salnow/y$salbeg)
>
> #ks.test(logdiff,'pnorm')
> library('nortest')
> lillie.test(logdiff)
```

Lilliefors (Kolmogorov-Smirnov) normality test

```
data: logdiff
D = 0.0543, p-value = 0.001931
```

```
> shapiro.test(logdiff)
```

Shapiro-Wilk normality test

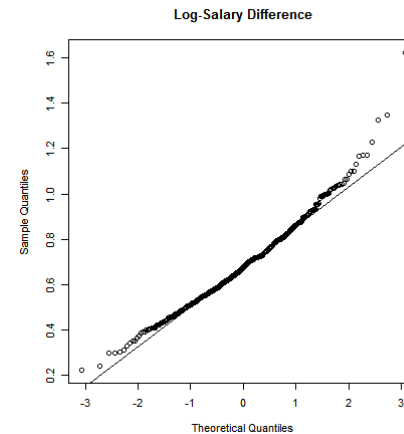
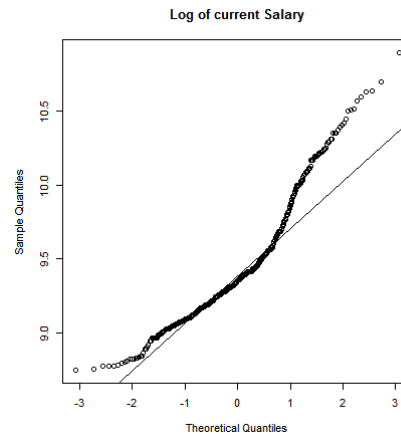
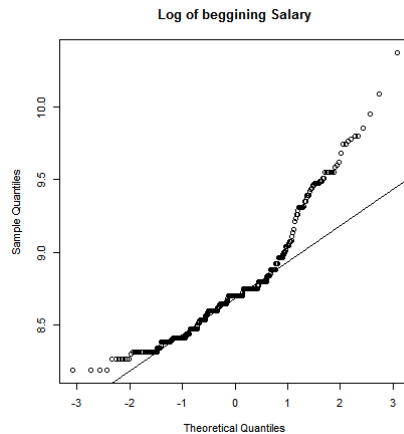
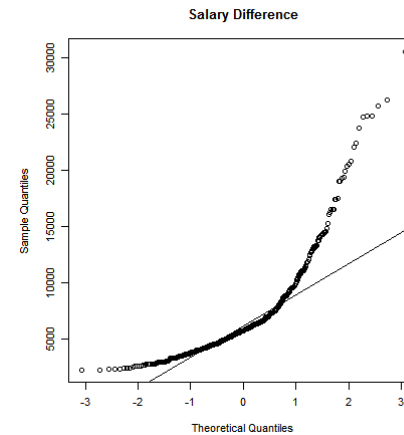
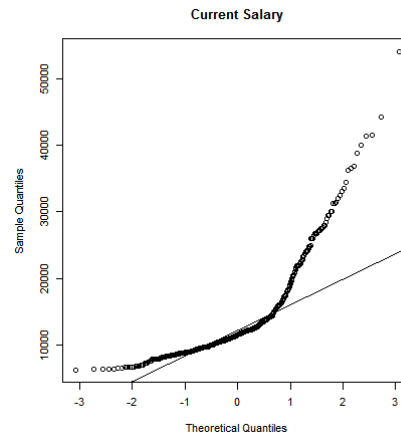
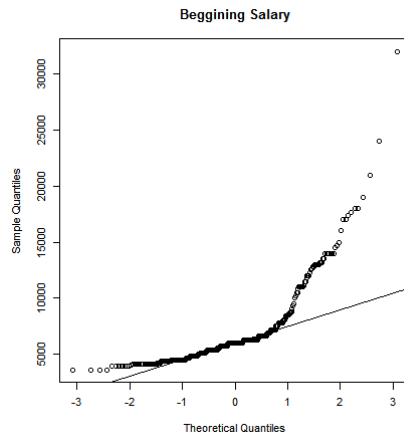
```
data: logdiff
W = 0.9779, p-value = 1.343e-06
```

```
>
> mean(logdiff)
[1] 0.6873881
> median(logdiff)
[1] 0.6751261
```


4.4. Hypothesis tests for a single continuous variable



Hypothesis test for the mean



4.4. Hypothesis tests for a single continuous variable

```
> t.test(logdiff, mu=0)
```

```
One Sample t-test
```

```
data: logdiff  
t = 81.3521, df = 473, p-value < 2.2e-16  
alternative hypothesis: true mean is not equal to 0  
95 percent confidence interval:  
 0.6707848 0.7039913  
sample estimates:  
mean of x  
0.6873881
```

```
>
```

```
> t.test(logdiff, mu=0.7)
```

```
One Sample t-test
```

```
data: logdiff  
t = -1.4926, df = 473, p-value = 0.1362  
alternative hypothesis: true mean is not equal to 0.7  
95 percent confidence interval:  
 0.6707848 0.7039913  
sample estimates:  
mean of x  
0.6873881
```

```
> |
```

4.4. Hypothesis tests for a single continuous variable



```
> y<-salary
> diff<- y$salnow - y$salbeg
>
> #ks.test(diff,'pnorm')
> library('nortest')
> lillie.test(diff)

      Lilliefors (Kolmogorov-Smirnov) normality test

data:  diff
D = 0.186, p-value < 2.2e-16

> shapiro.test(diff)

      Shapiro-Wilk normality test

data:  diff
W = 0.7817, p-value < 2.2e-16

>
> mean(diff)
[1] 6961.392
> median(diff)
[1] 5700
>
> wilcox.test(diff, mu=0)

      Wilcoxon signed rank test with continuity correction

data:  diff
V = 112575, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 0
```

4.4. Hypothesis tests for a single continuous variable

```
> wilcox.test(diff, mu=6000)
```

```
Wilcoxon signed rank test with continuity correction
```

```
data: diff  
V = 56813.5, p-value = 0.5075  
alternative hypothesis: true location is not equal to 6000
```

```
>
```

```
> t.test(diff, mu=6000)
```

```
One Sample t-test
```

```
data: diff  
t = 4.8386, df = 473, p-value = 1.773e-06  
alternative hypothesis: true mean is not equal to 6000  
95 percent confidence interval:  
 6570.963 7351.822  
sample estimates:  
mean of x  
 6961.392
```

```
> |
```

4.5. Hypothesis tests for two dependent samples



- Examples
 - Grades in 2 different courses
 - Cholesterol measurements **before and after** a treatment
 - Sales before and after a marketing policy
 - Treatment effects in left and right hand of a patient
 - Behavior study of twins under different environmental conditions
- Examines the association between:
 - Grades + course
 - Cholesterol + treatment
 - Sales + marketing policy
 - Disease + treatment
 - Behavior + environmental conditions

4.5. Hypothesis tests for two dependent samples

- We are interested to test for difference between the two dependent values – measurements
- We eliminate correlation by using the difference of each pair: $\Delta_i = X_{1i} - X_{2i}$
- We test if the mean of the difference is zero or not
- Hence we implement a one-sample t-test for the difference

$$H_0: \mu_{\Delta} = 0 \text{ vs } H_1: \mu_{\Delta} \neq 0 .$$

- This is the paired t-test or the dependent samples t-test
- **ASSUMPTION**
 - The difference should follow the normal distribution or the sample size to be large ($n > 50$)
- If normality assumption is rejected then we use the non-parametric test

$$H_0: M_{\Delta} = 0 \text{ vs } H_1: M_{\Delta} \neq 0$$

4.5. Hypothesis tests for two dependent samples

Έλεγχοι για 2 εξαρτημένα δείγματα
(2 ποσοτικές μεταβλητές στην R)

Είναι η διαφορά τους κανονική;
(Έλεγχος Κανονικότητας)
[SW αν $n \leq 50$ - KS+SW αν $n > 50$]

Ναι

Όχι

Είναι το δείγμα μεγάλο
($n > 50$)

Ναι

Ναι (?)

Είναι ο μέσος κατάλληλο μέτρο
περιγραφής της κεντρικής θέσης
της διαφοράς;

Όχι

Όχι

Έλεγχος για μηδενική
μέση διαφορά
t-test ανά ζεύγη

Απόρριψη H_0

Απόρριψη H_0

Έλεγχος για μηδενική
διάμεσο διαφορά
Wilcoxon test ανά ζεύγη

Error-bar διαφοράς

Box-plot διαφοράς

4.6. Hypothesis tests for two samples

Testing for the association between a continuous and a categorical variable



ΟΠΑ
ΑΥΕΒ

Hypothesis test for the equality of the means between two independent samples/groups

- Two variables:
 - One continuous
 - One binary-categorical specifying the groups
- $H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 \neq \mu_2$
- T-test for normal data with unknown variance
- **Assumption:** Normality or large sample for each group $n > 30$ or > 50
- Different degrees of freedom for equal and unequal variances

4.6. Hypothesis tests for two samples

Testing for the association between a continuous and a categorical variable



Hypothesis test for the equality of the medians between two independent samples/groups

- Two variables:
 - One continuous
 - One binary-categorical specifying the groups
- $H_0: M_1=M_2$ vs $H_1: M_1 \neq M_2$
- Wilcoxon rank-sum test (or Mann-Whitney)
- No assumptions
- Can be used when the normality assumption is not valid or the mean cannot be used as representative of the “usual” observation

4.6. Hypothesis tests for two samples

Testing for the association between a continuous and a categorical variable



ΟΠΑ
ΑΥΕΒ

- Measurements of the **same variable** in **two groups** of **different** individuals/research units
- For example: 2 groups of randomly selected patients
 - 1st group receives placebo (virtual of fake treatment)
 - 2nd group receives new treatment
- Response: Cholesterol
- We study the association/relationship between a quantitative and a binary qualitative (treatment) variable
- In R data frame => 2 columns = 1 numeric (cholesterol) and a factor with two levels (treatment)

4.6. Hypothesis tests for two samples

Testing for the association between a continuous and a categorical variable



- Examples
 - Performance/Grades in the same course in two subsequent years
 - Cholesterol for two groups of patients receiving different treatment
 - Sales in two different groups of shops in which different business strategy was implemented
 - Study of the psychological scales of two groups of individuals tested under different conditions of stress.
- Examines the association between
 - Performance/Grades + year
 - Cholesterol + treatment
 - Sales + business strategy
 - Psychological state + stress conditions

4.6. Hypothesis tests for two samples

Testing for the association between a continuous and a categorical variable



ΟΠΑ
ΑΥΕΒ

- We are interested to test for differences on the values of the quantitative variable for the two groups (are the means of the medians equal?)
- Testing for the equality of means:

$$H_0: \mu_1 = \mu_2 \text{ vs } H_1: \mu_1 \neq \mu_2 .$$

- This is the independent samples t-test
- In practice, it examines the relationship between the binary and the numeric variable since if the means on average are the same then the state of the binary does not influence the (conditional?) mean
- The hypothesis test can be re-written as

$$H_0: E(Y|X=1) = E(Y|X=2) \text{ vs } H_1: E(Y|X=1) \neq E(Y|X=2) .$$

Y is the numeric variable, X is the factor

4.6. Hypothesis tests for two samples

Testing for the association between a continuous and a categorical variable



- We are interested to test for differences on the values of the quantitative variable for the two groups (are the means of the medians equal?)
- Testing for the equality of means:
$$H_0: \mu_1 = \mu_2 \text{ vs } H_1: \mu_1 \neq \mu_2 .$$
- This is the independent samples t-test

BASIC ASSUMPTION:

- Normality in each group or the size to be high ($n_1, n_2 > 50$)
- If the previous assumption is not valid then we use a non-parametric test (the WILCOXON SUM-RANKED TEST)

$$H_0: M_1 = M_2 \text{ vs } H_1: M_1 \neq M_2$$

4.6. Hypothesis tests for two samples

Testing for the association between a continuous and a categorical variable



ΟΠΑ
ΑΥΕΒ

Example 4-5 [dataset1.dat]

- The education manager of a company wants to evaluate the efficiency of two different teaching methods of the employees.
- For this reason, he selects randomly 24 employees and divides them in two groups of equal size.
- Every group is educated in a new software by a different teaching method
- 2 employees of the 2nd group did not complete the seminar
- In the end, all employees were tested in a common examination and graded with a scale from 0 to 100.

4.6. Hypothesis tests for two samples

Testing for the association between a continuous and a categorical variable



Example 4-5 [dataset1.dat]

Group A	70	93	82	90	77	86	79	84	98	73	81	85
Group B	89	78	94	83	88	80	91	92	87	97		

- **AIM:** To identify which method is more efficient to follow it in the next seminars.

4.6. Hypothesis tests for two samples

Testing for the association between a continuous and a categorical variable



Example 4-5 [dataset1.dat]

- Study unit:
 - employee
 - Sample size $n=22$ employees (rows)
- Characteristics – variables:
 - Performance – Grades (numeric) + teaching method (binary factor)
 - $p=2$ (columns)
- Size of each group
 - $n_1=12$
 - $n_2=10$

```
R Console
> dataset1
  grades method
1     70      A
2     93      A
3     82      A
4     90      A
5     77      A
6     86      A
7     79      A
8     84      A
9     98      A
10    73      A
11    81      A
12    85      A
13    89      B
14    78      B
15    94      B
16    83      B
17    88      B
18    80      B
19    91      B
20    92      B
21    87      B
22    97      B
```

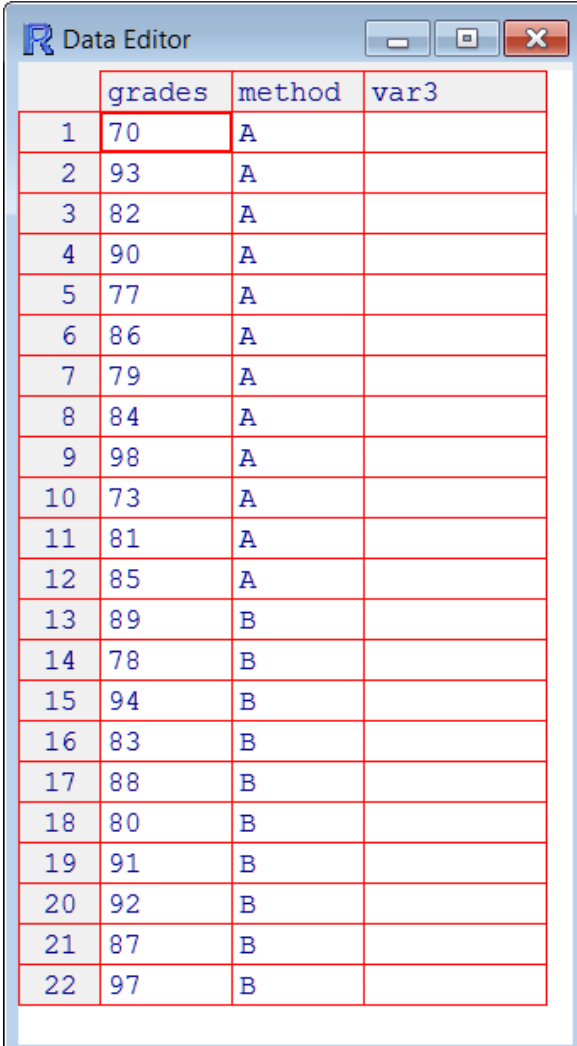

4.6. Hypothesis tests for two samples

Testing for the association between a continuous and a categorical variable

Example 4-5 [dataset1.dat]

- Study unit:
 - employee
 - Sample size $n=22$ employees (rows)
- Characteristics – variables:
 - Performance – Grades (numeric) + teaching method (binary factor)
 - $p=2$ (columns)
- Size of each group
 - $n_1=12$
 - $n_2=10$

```
dataset1 <- edit(dataset1)
```



	grades	method	var3
1	70	A	
2	93	A	
3	82	A	
4	90	A	
5	77	A	
6	86	A	
7	79	A	
8	84	A	
9	98	A	
10	73	A	
11	81	A	
12	85	A	
13	89	B	
14	78	B	
15	94	B	
16	83	B	
17	88	B	
18	80	B	
19	91	B	
20	92	B	
21	87	B	
22	97	B	

4.6. Hypothesis tests for two samples

Testing for the association between a continuous and a categorical variable



```
groupA <- c(70 ,93 ,82 ,90 ,77 ,86 ,79 ,84 ,98 ,73 ,81 ,85)
groupB <-c(89 ,78 ,94 ,83 ,88 ,80 ,91 ,92 ,87 ,97)
n1<-length(groupA)
n2<-length(groupB)

dataset1 <- data.frame( grades=c(groupA, groupB),
                        method=factor( rep(1:2, c(n1,n2)), labels=c('A','B') ) )
```

```
dataset1 <- edit(dataset1)
```

```
dput(dataset1, 'dataset1.dat')
dput(dataset1)
```

```
structure(list(grades = c(70, 93, 82, 90, 77, 86, 79, 84, 98, 73, 81, 85, 89, 78, 94, 83, 88, 80, 91, 92, 87, 97), method = structure(c(1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, 2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L, 2L), .Label = c("A", "B"), class = "factor")), .Names = c("grades", "method"), row.names = c(NA, 22L), class = "data.frame")
```

```
> |
```

4.6. Hypothesis tests for two samples

Testing for the association between a continuous and a categorical variable



Steps for implementing hypothesis tests in two independent samples

1. We test for normality in each group
2. We implement the appropriate test
 - a) If normality is not rejected
 - i. We test for the equality of variances (homoscedasticity)
 - ii. We select the appropriate t-test (with equal or unequal variances)
 - iii. If there are differences, we visualize them using ERROR BARS
 - b) If normality is rejected
 - i. We implement the WILCOXON rank-sum test
 - ii. If differences are traced, then we visualize them using BOX PLOTS

4.6. Hypothesis tests for two samples

Testing for the association between a continuous and a categorical variable



Example 4-5 [dataset1.dat]

1. We test for the normality in each group

```
> library(nortest)
> by( dataset1$grades, dataset1$method, lillie.test)
dataset1$method: A
```

Lilliefors (Kolmogorov-Smirnov) normality test

```
data: dd[x, ]
D = 0.1125, p-value = 0.944
```

```
> by( dataset1$grades, dataset1$method, shapiro.test)
dataset1$method: A
```

Shapiro-Wilk normality test

```
data: dd[x, ]
W = 0.9893, p-value = 0.9996
```

```
dataset1$method: B
```

Lilliefors (Kolmogorov-Smirnov) normality test

```
data: dd[x, ]
D = 0.1412, p-value = 0.8278
```

```
dataset1$method: B
```

Shapiro-Wilk normality test

```
data: dd[x, ]
W = 0.9697, p-value = 0.8876
```

```
> |
```

All p-values > 0.05 => we do not reject normality in either of the two groups

4.6. Hypothesis tests for two samples

Testing for the association between a continuous and a categorical variable



ΟΠΑ
ΑΥΕΒ

Example 4-5 [dataset1.dat]

2. α) independent samples t-test

i. Checking for the equality of variances

```
> var.test( grades~method, data=dataset1 )
```

F test to compare two variances

```
data: grades by method  
F = 1.754, num df = 11, denom df = 9, p-value = 0.4079  
alternative hypothesis: true ratio of variances is not equal to 1  
95 percent confidence interval:  
 0.4483574 6.2932011  
sample estimates:  
ratio of variances  
 1.754007
```

P-value=0.41 > 0.05 => we do not reject H_0 =>
we can assume equal variances

```
var.test(x$A, x$B)  
var.test(x[[1]], x[[2]])  
var.test( grades~method, data=dataset1 )
```

4.6. Hypothesis tests for two samples

Testing for the association between a continuous and a categorical variable



Example 4-5 [dataset1.dat]

2. α) independent samples t-test

ii. Using the t-test for equal variances

```
t.test(x$A, x$B, var.equal=T)
t.test(x[[1]], x[[2]], var.equal=T)
t.test( grades~method, data=dataset1 ,
       var.equal=T)
```

```
> t.test( grades~method, data=dataset1 , var.equal=T)
```

```
Two Sample t-test
```

```
data: grades by method
t = -1.5282, df = 20, p-value = 0.1421
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.194276  1.727609
sample estimates:
mean in group A mean in group B
      83.16667      87.90000
```

P-value=0.14>0.05 => we do not reject the H_0 => no differences between the performance of the employees with different training

4.6. Hypothesis tests for two samples

Testing for the association between a continuous and a categorical variable



ΟΠΑ
ΑΥΕΒ

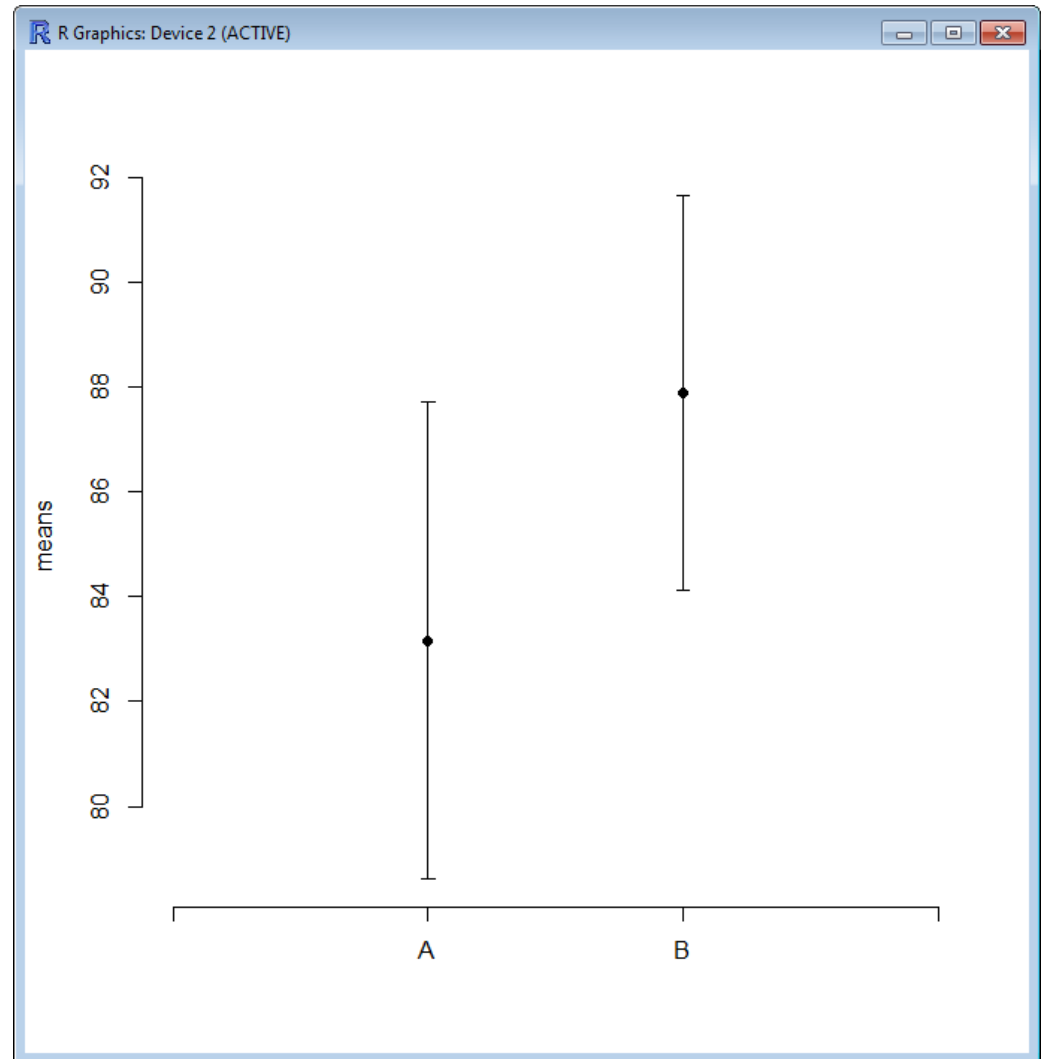
Example 4-5

[dataset1.dat]

2. α) Independent samples t-test

iii. Visualization using error bars

```
library(Hmisc)
myerrorbar( dataset1$grades,
            dataset1$method )
```



4.6. Hypothesis tests for two samples

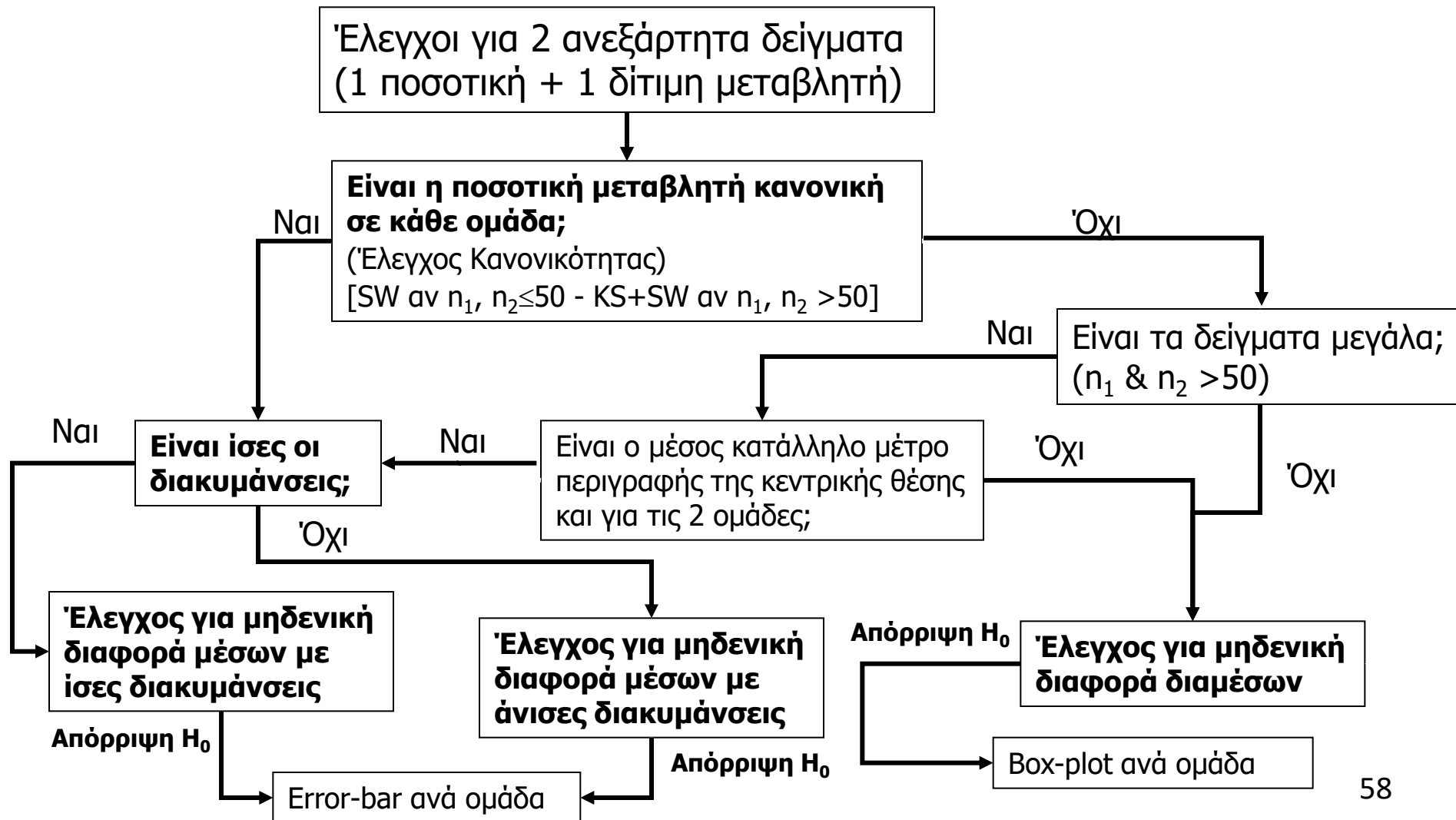
Testing for the association between a continuous and a categorical variable



```
myerrorbar<-function(x,y, horizontal=F){
  a<-0.05
  sdata <- split(x,y)
  means <- sapply( sdata,mean )
  sds <- sapply( split(x,y), sd )
  ns <- table(y)
  LB <- means + qnorm( a/2 ) * sds /sqrt(ns)
  UB <- means + qnorm( 1-a/2 ) * sds /sqrt(ns)
  nlev <- nlevels(y)
  if (horizontal) { errbar( levels(y), means, UB, LB )
                    } else {
    errbar( 1:nlev, means, UB, LB,
           xlim=c(0,nlev+1), axes=F, xlab="" )
    axis(2)
    axis(1, at=0:(nlev+1), labels=c("",levels(y),""))
  }
}
```


4.6. Hypothesis tests for two samples

Testing for the association between a continuous and a categorical variable



4.7. Hypothesis tests for multiple samples

Testing for the association between a continuous and a categorical variable



- Introduction to analysis of variance (ANOVA)
- Assumptions
- Multiple comparisons
- Nonparametric hypothesis test
- The link between t-test and ANOVA

4.7. Hypothesis tests for multiple samples

4.7.1. Analysis of variance



- Let us assume that we have measurement of the **same quantitative variable** in **$k(>2)$ groups of different individuals**
- For example, 3 groups of patients are randomly selected
 - $1n$ group with virtual/placebo treatment
 - $2n$ group receives a standard treatment
 - $3n$ group received a new treatment
- We examine for a possible relationship between a quantitative and a categorical variable
- In R data frame => 2 columns = 1 numeric and a factor with 3 levels (or k levels in a more general setup)

4.7. Hypothesis tests for multiple samples

4.7.1. Analysis of variance



- Examples
 - Grades/student performance in 4 different academic years
 - Sales in 3 different groups of sale points with different business strategies
- Examines the association
 - Grades/performance + year
 - Sales + strategy

4.7. Hypothesis tests for multiple samples

4.7.1. Analysis of variance



- We are interested to test for the differences between the quantitative variable and the groups i.e., are the means or the medians equal across groups
- We test for
$$H_0: \mu_1 = \mu_2 = \dots = \mu_k \text{ vs}$$
$$H_1: \mu_k \neq \mu_j \text{ for some } k \neq j \in \{1, 2, \dots, K\}.$$
- This hypothesis test is called analysis of variance (ANOVA: **A**nalysis of **V**ariance)
- **ASSUMPTIONS:**
 - ✓ Residuals' normality or the sample size to be large ($n > 50$)
 - ✓ Equal variances
- If the above hypotheses are rejected, then we may use a non-parametric method (KRUSKAL-WALLIS TEST)
$$H_0: M_1 = M_2 = \dots = M_K \text{ vs}$$
$$H_1: M_i \neq M_j \text{ for some } i \neq j = 1, 2, \dots, K.$$

4.7. Hypothesis tests for multiple samples

4.7.1. Analysis of variance



- Let us assume Y quantitative variable and X categorical with κ levels.
- n_j : is the sample size of group j , $j=1,2,\dots,\kappa$
- $n = \sum_{j=1}^{\kappa} n_j$: total sample size
- Y_{ij} is the i -th observation of group j
- \bar{Y}_j is the sample mean of Y for group j
- $e_{ij} = Y_{ij} - \bar{Y}_j$ is the residual values of i observation of group j

4.7. Hypothesis tests for multiple samples

4.7.1. Analysis of variance



Sum of squares

- $TSS = \sum_{j=1}^{\kappa} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y})^2$ is the total sum of squares (variance of Y?)
- $RSS = \sum_{j=1}^{\kappa} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2$ is the residual sum of squares and expresses the variability within groups [within groups sum of squares]
- $BSS = TSS - RSS = \sum_{j=1}^{\kappa} n_j (\bar{Y}_j - \bar{Y})^2$ is the between groups sum of squares measuring the variance of the samples means of different groups

– We use the test Statistic:

$$F^* = \frac{BSS / (\kappa - 1)}{RSS / (n - \kappa)} \sim F_{\kappa-1, n-\kappa}$$

4.7. Hypothesis tests for multiple samples

4.7.1. Analysis of variance



- So, we compare variances for testing for the equality of means.
WHY???
- **BE CAREFULL:** Rejection of H_0 implies only that some mean differ. The test does not provide any information concerning which means differ.
 - We implement multiple pairwise comparisons
 - We visualize differences using error-bars

4.7. Hypothesis tests for multiple samples

4.7.1. Analysis of variance



Example 4-7:

- Let us reconsider example 4-6 with the different training methods.
- Now we consider 15 employees divided in 3 different groups of training
- The data are given in the following table:

Training method	Grade/performance				
A	86	79	81	70	84
B	90	76	88	82	89
C	82	68	73	71	81

4.7. Hypothesis tests for multiple samples

4.7.1. Analysis of variance



Example 4-7:

- Unit of study: the employee
- Sample size: $n=15$
- Characteristics $p=2$
 - Training method
 - Grade/performance

```
grades<-c(86,79,81,70,84,90,76,88,82,89,82,  
          68,73, 71,81)  
method<-rep(1:3,rep(5,3))  
method<-factor(method, labels=paste('Method',  
                                     LETTERS[1:3]))  
  
ex4.7<-data.frame( grades=grades,  
                  method=method )  
rm(grades)  
rm(method)
```

> ex4.7

	grades	method
1	86	Method A
2	79	Method A
3	81	Method A
4	70	Method A
5	84	Method A
6	90	Method B
7	76	Method B
8	88	Method B
9	82	Method B
10	89	Method B
11	82	Method C
12	68	Method C
13	73	Method C
14	71	Method C
15	81	Method C

4.7. Hypothesis tests for multiple samples

4.7.1. Analysis of variance



Example 4-7:

- Unit of study: the employee
- Sample size: $n=15$
- Characteristics $p=2$
 - Training method
 - Grade/performance

```
ex4.7<-edit(ex4.7)
```

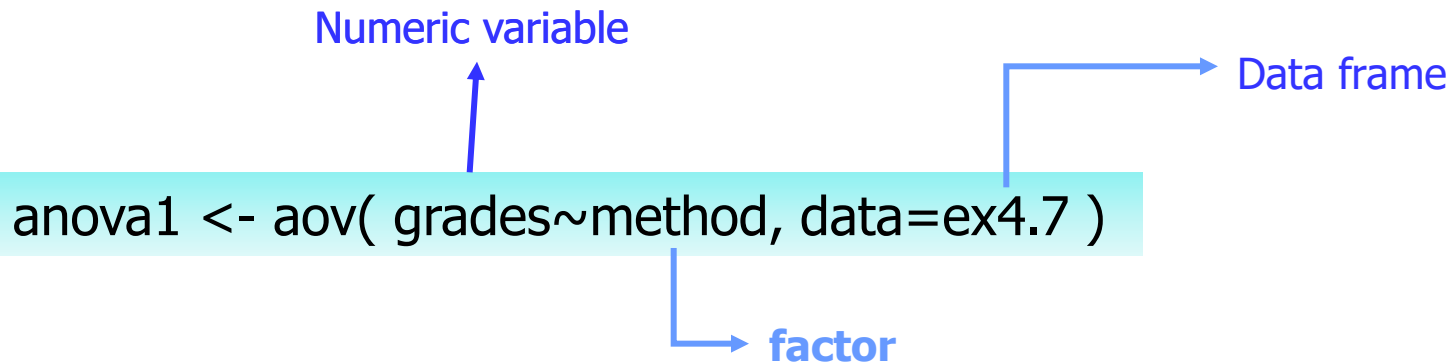
	grades	method	var3
1	86	Method A	
2	79	Method A	
3	81	Method A	
4	70	Method A	
5	84	Method A	
6	90	Method B	
7	76	Method B	
8	88	Method B	
9	82	Method B	
10	89	Method B	
11	82	Method C	
12	68	Method C	
13	73	Method C	
14	71	Method C	
15	81	Method C	
16			
17			

4.7. Hypothesis tests for multiple samples

4.7.1. Analysis of variance



Example 4-7: ANOVA using aov function in R



also see

```
anova2 <- oneway.test( grades~method, data=ex4.7 )
```

4.7. Hypothesis tests for multiple samples

4.7.1. Analysis of variance



Example 4-7: ANOVA RESULTS IN R

```
> anova
Call:
  aov(formula = grades ~ method, data = ex4.7)

Terms:
              method Residuals
Sum of Squares      250       448
Deg. of Freedom       2       12

Residual standard error: 6.110101
Estimated effects may be unbalanced
> names(anova)
 [1] "coefficients"  "residuals"      "effects"
 [4] "rank"          "fitted.values"  "assign"
 [7] "qr"           "df.residual"    "contrasts"
[10] "xlevels"       "call"           "terms"
[13] "model"
```

4.7. Hypothesis tests for multiple samples

4.7.1. Analysis of variance



Example 4-7: Anova table using the summary function in R

```
> summary(anova1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
method	2	250	125.00	3.348	0.0699
Residuals	12	448	37.33		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Δεν απορρίπτεται η H_0
για $\alpha=5\%$

4.7. Hypothesis tests for multiple samples

4.7.1. Analysis of variance



Example 4-7: ANOVA RESULTS IN R

```
> names(anova1)
 [1] "coefficients"  "residuals"      "effects"
 [4] "rank"          "fitted.values"  "assign"
 [7] "qr"           "df.residual"    "contrasts"
[10] "xlevels"       "call"           "terms"
[13] "model"
```

4.7. Hypothesis tests for multiple samples

4.7.1. Analysis of variance



Example 4-7: ANOVA RESULTS IN R

- `coefficients` model coefficients
- `fitted.values` the fitted mean values.
in one-way anova fitted values = the group mean
- `residuals` the residuals: observed value – fitted value
- `df.residual` the residual degrees of freedom.
- `call` the exact expression used to generate this object
- `contrasts` The parametrization (constraints) used
- `xlevels` the levels of the factors used
- `y` the response used (only if `y=TRUE`)
- `x` the model matrix used (only if `x=TRUE`)
- `model` the model frame used (only if `model=TRUE` – default value)
- `rank` rank of data/design matrix $X \Rightarrow$ i.e. the number of parameters \Rightarrow in one-way anova the number of groups ⁷³

4.7. Hypothesis tests for multiple samples

4.7.1. Analysis of variance



Example 4-7: ANOVA RESULTS IN R

```
> anova1$fitted.values
```

```
1  2  3  4  5  6  7  8  9 10 11 12 13 14 15  
80 80 80 80 80 85 85 85 85 85 75 75 75 75 75
```

```
> anova1$coef
```

```
(Intercept) methodMethod B methodMethod C  
            80                5             -5
```

```
> anova1$residuals
```

```
1  2  3  4  5  6  7  8  9 10 11 12 13 14 15  
6 -1  1 -10  4  5 -9  3 -3  4  7 -7 -2 -4  6
```

```
> anova1$call
```

```
aov(formula = grades ~ method, data = ex4.7)
```

4.7. Hypothesis tests for multiple samples

4.7.2. Checking for the assumptions of ANOVA

Example 4-7: Checking the assumptions of anova

1) Normality of the residuals

```
> library(nortest)
> lillie.test(anova1$residuals)
```

Lilliefors (Kolmogorov-Smirnov) normality test

```
data: anova1$residuals
D = 0.1687, p-value = 0.299
```

```
> shapiro.test(anova1$residuals)
```

Shapiro-Wilk normality test

```
data: anova1$residuals
W = 0.9196, p-value = 0.1897
```

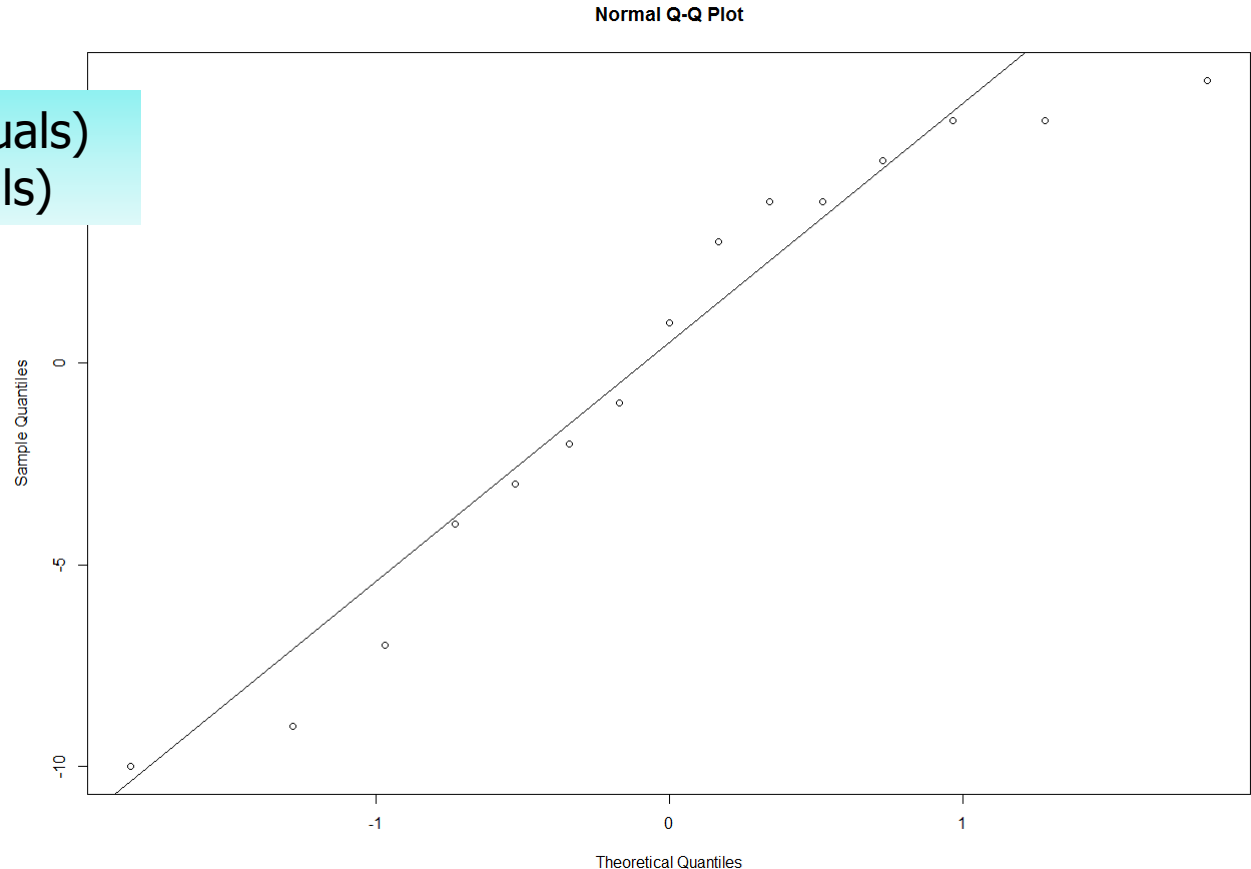
4.7. Hypothesis tests for multiple samples

4.7.2. Checking for the assumptions of ANOVA

Example 4-7: Checking the assumptions of anova

1) Normality of the residuals

```
qqnorm(anova1$residuals)  
qqline(anova1$residuals)
```



4.7. Hypothesis tests for multiple samples

4.7.2. Checking for the assumptions of ANOVA



Example 4-7: Checking the assumptions of anova

2) Homogeneity of variances

$$H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 \quad \text{vs} \quad H_1: \sigma_k^2 \neq \sigma_j^2 \text{ for some } j, k \in \{1, 2, \dots, k\}$$

There are many ways of testing data for homogeneity of variance. Three methods are shown here.

- **Bartlett's test** - If the data is normally distributed, this is the best test to use. It is sensitive to data which is not non-normally distribution; it is more likely to return a "false positive" when the data is non-normal.
- **Levene's test** - this is more robust to departures from normality than Bartlett's test. It is in the car package.
- **Fligner-Killeen test** - this is a non-parametric test which is very robust against departures from normality.

4.7. Hypothesis tests for multiple samples

4.7.2. Checking for the assumptions of ANOVA



Example 4-7: Checking the assumptions of anova

2) Homogeneity of variances

```
> bartlett.test(grades~method, data=ex4.7)
```

```
Bartlett test of homogeneity of variances
```

```
data: grades by method
```

```
Bartlett's K-squared = 0.0108, df = 2, p-value = 0.9946
```

```
> fligner.test(grades~method, data=ex4.7)
```

```
Fligner-Killeen test of homogeneity of variances
```

```
data: grades by method
```

```
Fligner-Killeen:med chi-squared = 0.0247, df = 2, p-value = 0.9877
```

```
> library(car)
```

```
> leveneTest(grades~method, data=ex4.7)
```

```
Levene's Test for Homogeneity of Variance (center = median)
```

```
      Df F value Pr(>F)  
group 2  0.0318 0.9688  
      12
```

4.8. *Two categorical variables*

- Testing for the equality of proportions/probabilities in independent groups/samples
 - ✓ Testing for the independence between binaries => comparison of success probabilities/proportions for two groups
 - t-test for the difference of proportions
 - t-test for the log-ratio of proportions (relative risk)
 - t-test for the log-odds ratio
 - Testing for independence in 2x2 contingency tables
 - ✓ Testing for independence in IxJ contingency tables
- Equality of proportions in dependent samples/paired values
 - ✓ Testing for the equality of marginal proportions using the McNemar test for 2x2 contingency tables

4.8. *Two categorical variables*

4.8.1. *Testing for the association between two categorical variables (independent samples)*



Example 4-11: Masticha shop customer satisfaction survey

The effect of gender on the knowledge for products

- ✓ Alcoholic drinks
- ✓ Bakery products
- ✓ Desserts and sweets
- ✓ Cosmetics

4.8. Two categorical variables

4.8.1. Testing for the association between two categorical variables (independent samples)



Example 4-11: Masticha shop customer satisfaction survey

The effect of gender on the knowledge for alcoholic drinks

- We are interested to test whether there is a difference between males and females concerning the knowledge of the existence of some products
⇒
- We would like to test for the equality of proportions of males or females that were informed for the existence of a type of product (for example alcoholic drinks) ⇒
- $H_0: \pi_{\text{males}} = \pi_{\text{females}}$ vs $H_1: \pi_{\text{males}} \neq \pi_{\text{females}}$ ⇔
- H_0 : “Independence between gender and knowledge of a product” vs
- H_1 : “there is association between gender and product knowledge”

4.8. Two categorical variables

4.8.1. Testing for the association between two categorical variables (independent samples)



Example 4-11: Masticha shop customer satisfaction survey

```
> tab1<-table(masticha.all$gender, masticha.all$a4_drinks )  
> tab1
```

```
      OXI NAI  
Ανδρας   7  27  
Γυναίκα  8  62
```

```
> prop.table(tab1)
```

```
      OXI      NAI  
Ανδρας 0.06730769 0.25961538  
Γυναίκα 0.07692308 0.59615385
```

```
> prop.table(tab1,1)
```

```
      OXI      NAI  
Ανδρας 0.2058824 0.7941176  
Γυναίκα 0.1142857 0.8857143
```

```
> prop.table(tab1,2)
```

```
      OXI      NAI  
Ανδρας 0.4666667 0.3033708  
Γυναίκα 0.5333333 0.6966292
```

Total table proportions

Row proportions

Column proportions

4.8. *Two categorical variables*

4.8.1. *Testing for the association between two categorical variables (independent samples)*

Example 4-11: Masticha shop customer satisfaction survey

```
> round(100*prop.table(tab1,1),1)
```

	OXI	NAI
Άνδρας	20.6	79.4
Γυναίκα	11.4	88.6

4.8. Two categorical variables

4.8.1. Testing for the association between two categorical variables (independent samples)



Example 4-11: Masticha shop customer satisfaction survey

```
> prop.test( tab1 )
```

```
2-sample test for equality of proportions with continuity correction
```

```
data: tab1
```

```
X-squared = 0.902, df = 1, p-value = 0.3423
```

```
alternative hypothesis: two.sided
```

```
95 percent confidence interval:
```

```
-0.08525954 0.26845282
```

```
sample estimates:
```

```
prop 1 prop 2
```

```
0.2058824 0.1142857
```

```
Warning message:
```

```
In prop.test(tab1) : Chi-squared approximation may be incorrect
```

prop.test implements the Pearson's chi-square statistics for independence

4.8. Two categorical variables

4.8.1. Testing for the association between two categorical variables (independent samples)



Example 4-11: Masticha shop customer satisfaction survey

```
> chisq.test(tab1)
```

```
Pearson's Chi-squared test with Yates' continuity correction
```

```
data: tab1
```

```
X-squared = 0.902, df = 1, p-value = 0.3423
```

```
Warning message:
```

```
In chisq.test(tab1) : Chi-squared approximation may be incorrect
```

```
< |
```

Same also for `chisq.test`

4.8. Two categorical variables

4.8.1. Testing for the association between two categorical variables (independent samples)



Example 4-11: Masticha shop customer satisfaction survey

```
> xtabs(~gender+a4_drinks,data=masticha.all)
      a4_drinks
gender  OXI NAI
Ανδρας    7  27
Γυναίκα   8  62
> summary(xtabs(~gender+a4_drinks,data=masticha.all))
Call: xtabs(formula = ~gender + a4_drinks, data = masticha.all)
Number of cases in table: 104
Number of factors: 2
Test for independence of all factors:
  Chisq = 1.5556, df = 1, p-value = 0.2123
  Chi-squared approximation may be incorrect
> chisq.test( tab1, correct=F )

      Pearson's Chi-squared test

data:  tab1
X-squared = 1.5556, df = 1, p-value = 0.2123

Warning message:
In chisq.test(tab1, correct = F) :
  Chi-squared approximation may be incorrect
```

4.8. Two categorical variables

4.8.1. Testing for the association between two categorical variables (independent samples)



Example 4-11: Masticha shop customer satisfaction survey

```
> chisq.test(tab1)
```

```
Pearson's Chi-squared test with Yates' continuity correction
```

```
data: tab1  
X-squared = 0.902, df = 1, p-value = 0.3423
```

Warning message:

```
In chisq.test(tab1) : Chi-squared approximation may be incorrect
```

```
> chisq.test(tab1, correct=FALSE, simulate.p.value = TRUE)
```

```
Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)
```

```
data: tab1  
X-squared = 1.5556, df = NA, p-value = 0.2279
```

chisq.test is more flexible than prop.test

If small expected values (<5) arise then use simulate.p.value to obtain a Monte Carlo estimate of the p-values which is more accurate

Otherwise use the Fisher's exact test

4.8. Two categorical variables

4.8.1. Testing for the association between two categorical variables (independent samples)



Example 4-11: Masticha shop customer satisfaction survey

```
> chisq.test(tab1, correct=FALSE, simulate.p.value = TRUE)
```

```
Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)
```

```
data: tab1  
X-squared = 1.5556, df = NA, p-value = 0.2459
```

```
> fisher.test(tab1)
```

```
Fisher's Exact Test for Count Data
```

```
data: tab1  
p-value = 0.2415  
alternative hypothesis: true odds ratio is not equal to 1  
95 percent confidence interval:  
 0.5551755 7.0295991  
sample estimates:  
odds ratio  
 1.994795
```

MC estimated p-value and the Fisher's exact p-value are very close as expected and more accurate than approximate p-value in chi-sq test

4.8. Two categorical variables

4.8.1. Testing for the association between two categorical variables (independent samples)



Example 4-11: Masticha shop customer satisfaction survey

Total Observations in Table: 104

		masticha.all\$a4_drinks		
masticha.all\$gender		OXI	NAI	Row Total
Ανδρας		7	27	34
		0.896	0.151	
		0.206	0.794	0.327
		0.467	0.305	
Γυναίκα		0.067	0.260	
		8	62	70
		0.435	0.073	
		0.114	0.886	0.673
	0.533	0.697		
	0.077	0.596		
Column Total		15	89	104
		0.144	0.856	

```
CrossTable(masticha.all$gender, masticha.all$a4_drinks)
```


4.8. Two categorical variables

4.8.1. Testing for the association between two categorical variables (independent samples)

Example 4-11: Masticha shop customer satisfaction survey

Total Observations in Table: 104

		masticha.all\$a4_drinks		
masticha.all\$gender	OXI	NAI	Row Total	
Ανδρας	7	27	34	
	0.9	0.2		
	20.6%	79.4%	32.7%	
	46.7%	30.3%		
	6.7%	26.0%		
Γυναίκα	8	62	70	
	0.4	0.1		
	11.4%	88.6%	67.3%	
	53.3%	69.7%		
	7.7%	59.6%		
Column Total	15	89	104	
	14.4%	85.6%		

```
CrossTable(masticha.all$gender, masticha.all$a4_drinks, digits=1, format='SPSS')
```

4.8. Two categorical variables

4.8.1. Testing for the association between two categorical variables (independent samples)

Example 4-11: Masticha shop customer satisfaction survey

```
Cell Contents  
-----  
Count  
Row Percent  
-----
```

```
CrossTable(masticha.all$gender,masticha.all$a4_drinks,  
digits=1, format='SPSS', expected=FALSE,  
prop.r=TRUE, prop.c=F, prop.t=F,  
prop.chisq=F, chisq = T, fisher=T,  
mcnemar=FALSE)
```

Total Observations in Table: 104

masticha.all\$gender	masticha.all\$a4_drinks		Row Total
	OXI	NAI	
Ανδρας	7 20.6%	27 79.4%	34 32.7%
Γυναίκα	8 11.4%	62 88.6%	70 67.3%
Column Total	15	89	104

4.8. Two categorical variables

4.8.1. Testing for the association between two categorical variables (independent samples)



Example 4-11: Masticha shop customer satisfaction survey

```
CrossTable(masticha.all$gender,masticha.all$a4_drinks,  
digits=1, format='SPSS', expected=FALSE,  
prop.r=TRUE, prop.c=F, prop.t=F,  
prop.chisq=F, chisq = T, fisher=T,  
mcnemar=FALSE)
```

Statistics for All Table Factors

Pearson's Chi-squared test

Chi^2 = 1.555565 d.f. = 1 p = 0.2123158

Pearson's Chi-squared test with Yates' continuity correction

Chi^2 = 0.9019683 d.f. = 1 p = 0.3422545

4.8. Two categorical variables

4.8.1. Testing for the association between two categorical variables (independent samples)



Example 4-11: Masticha shop customer satisfaction survey

Fisher's Exact Test for Count Data

fisher=T

Sample estimate odds ratio: 1.994795

Alternative hypothesis: true odds ratio is not equal to 1
p = 0.2415254
95% confidence interval: 0.5551755 7.029599

Alternative hypothesis: true odds ratio is less than 1
p = 0.9360206
95% confidence interval: 0 5.864602

Alternative hypothesis: true odds ratio is greater than 1
p = 0.1704272
95% confidence interval: 0.6695709 Inf

4.8. *Two categorical variables*

4.8.1. *Testing for the association between two categorical variables (independent samples)*



Example 4-11: Masticha shop customer satisfaction survey

Warning for the validity of chi-square test. Expected values should be > 5 in order to accurately calculate the (approximate) p-value

```
Minimum expected frequency: 4.903846  
Cells with Expected Frequency < 5: 1 of 4 (25%)
```

Warning messages:

```
1: In chisq.test(t, correct = TRUE, ...) :  
  Chi-squared approximation may be incorrect  
2: In chisq.test(t, correct = FALSE, ...) :  
  Chi-squared approximation may be incorrect
```

4.8. Two categorical variables

4.8.1. Testing for the association between two categorical variables (independent samples)

Example 4-11: Masticha shop customer satisfaction survey

```
library(sjPlot)
sjt.xtab(masticha.all$gender, masticha.all$a4_drinks)
```

<i>var.row</i>	<i>var.col</i>		<i>Total</i>
	OXI	NAI	
Ανδρας	7	27	34
Γυναίκα	8	62	70
Total	15	89	104

Fisher's $p=0.242 \cdot df=1 \cdot \Phi=0.122$

observed values · expected values · % within var.row · % within var.col · % of total

4.8. Two categorical variables

4.8.1. Testing for the association between two categorical variables (independent samples)



Example 4-11: Masticha shop customer satisfaction survey

```
sjt.xtab(masticha.all$gender, masticha.all$a4_drinks, show.cell.prc= FALSE,  
show.row.prc= TRUE, show.col.prc = FALSE, show.exp= FALSE,  
var.labels = c('Φύλο', 'Γνώση Ύπαρξης Ποτών στο Κατάστημα'),  
encoding = "UTF-8")
```

Φύλο	Γνώση Ύπαρξης Ποτών στο Κατάστημα		Total
	OXI	NAI	
Ανδρας	7 20.6 %	27 79.4 %	34 100.0 %
Γυναίκα	8 11.4 %	62 88.6 %	70 100.0 %
Total	15 14.4 %	89 85.6 %	104 100.0 %

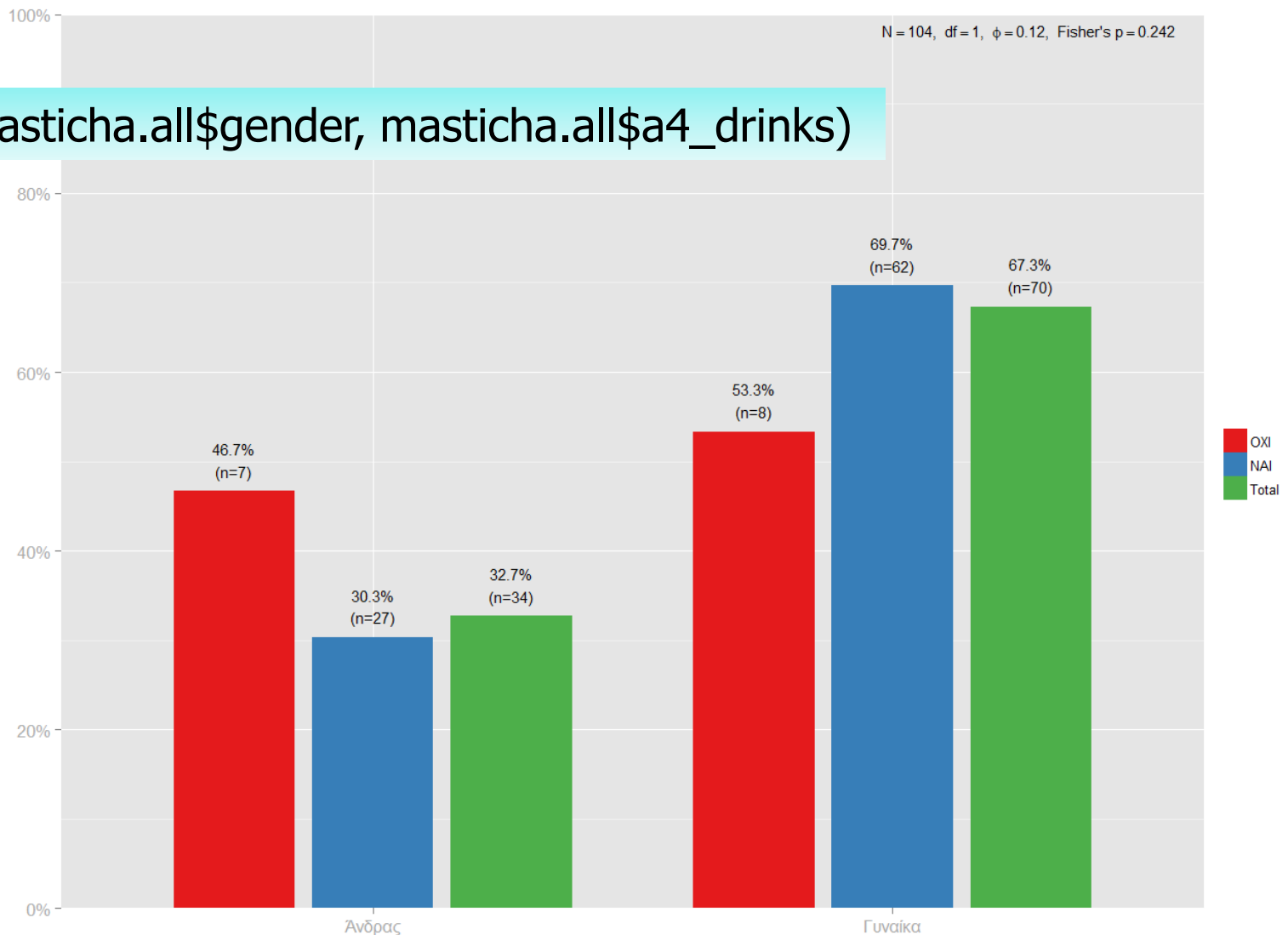
Fisher's $p=0.242 \cdot df=1 \cdot \Phi=0.122$

4.8. Two categorical variables

4.8.1. Testing for the association between two categorical variables (independent samples)



```
sjp.xtab(masticha.all$gender, masticha.all$a4_drinks)
```



4.8. Two categorical variables

4.8.1. Testing for the association between two categorical variables (independent samples)



Example 4-11: Masticha shop customer satisfaction survey

```
library(MASS)  
loglm( ~ 1+2, tab1)
```

Likelihood ratio test for
independence

```
> library(MASS)
```

```
> loglm( ~ 1+2, tab1)
```

```
Call:
```

```
loglm(formula = ~1 + 2, data = tab1)
```

```
Statistics:
```

	X ²	df	P(> X ²)
Likelihood Ratio	1.486339	1	0.2227854
Pearson	1.555565	1	0.2123158

4.8. Two categorical variables

4.8.1. Testing for the association between two categorical variables (independent samples)



Assumptions for tests

- Pearson's χ^2 independence test
 - Good approximation when expected values > 5
 - [Less strict assumption according to Cochran (1954, Biometrics): 80% of expected values > 5 and all of them > 1]
- χ^2 test with Yates correction
 - Only for 2x2 tables [approximates better the Fisher Exact test]
- Likelihood ratio test of independence
 - Bad approximation for $n/IJ < 5$ (Agresti, 1990, p. 49)
- Fisher's exact test
 - It is the best since it is based on the exact distribution
- Monte Carlo estimates of p-values – it should be used when the χ^2 approximation is not valid and/or the Fisher exact test cannot be obtained 99

4.8. Two categorical variables

4.8.1. Testing for the association between two categorical variables (independent samples)



Example 4-11: Masticha shop customer satisfaction survey

```
index<-6:9
pvalues<-matrix(nrow=4,ncol=3)
for (i in 1:4){
  var<-index[i]
  tab <- table( masticha.all$gender, masticha.all[,var] )
  pvalues[i,1]<-chisq.test(tab, correct=F)$p.value
  pvalues[i,2]<-summary(loglm( ~ 1+2, tab))$tests[1,3]
  pvalues[i,3]<-fisher.test(tab)$p.value
}
pvalues
colnames(pvalues) <- c('Chisq', 'LRT', "Fisher's")
rownames(pvalues) <- c('Alcoholic drinks', 'Bakery products', 'Sweets',
  'Cosmetics')
round(pvalues,3)
```

4.8. Two categorical variables

4.8.1. Testing for the association between two categorical variables (independent samples)



Example 4-11: Masticha shop customer satisfaction survey

P-values for the association of the gender with various type of products

```
> round(pvalues, 3)
```

	Chisq	LRT	Fisher's
Alcoholic drinks	0.212	0.223	0.242
Bakery products	0.495	0.496	0.527
Sweets	0.492	0.485	0.595
Cosmetics	0.040	0.035	0.048

```
~
```

4.8. *Two categorical variables*

4.8.2. *Testing for the equality of proportions in dependent samples*



- Here we will consider the corresponding test of the paired t-test but for categorical variables.
- Hence, we may have the same binary or categorical measurements in two different time-point on the same study units (subjects or individuals).
- For binary variables, the resulted table is of 2x2 dimension
- We are interested to test the hypothesis:
- $H_0: P(\text{success in 1}^{\text{st}} \text{ time point}) = P(\text{success in 2}^{\text{nd}} \text{ time point}) \Rightarrow$
- $H_0: \pi_{i.} = \pi_{.i}$ vs. $H_1: \pi_{i.} \neq \pi_{.i} \Rightarrow$
- H_0 : “There is no time effect in the categorical variable” vs H_1 : “there is a time effect on the categorical variable”

4.8. *Two categorical variables*

4.8.2. *Testing for the equality of proportions in dependent samples*



Example 4-13: Prime minister's approval

- [Agresti, 2002, 2nd ed, p. 409; Table 10.1].
- Random sample of 1600 voting-age British citizens
- Two satisfaction/approval surveys with difference of 6 months
- Is the proportion of voters that approve the Prime minister's similar?

4.8. Two categorical variables

4.8.2. Testing for the equality of proportions in dependent samples



Example 4-13: Prime minister's approval

- H_0 : “There is no difference in the approval of the Prime minister within the last 6 months” vs. H_1 : “There is a difference in the approval of the Prime minister within the last 6 months” \Rightarrow
- H_0 : $P(\text{Approval} | \text{First Survey}) = P(\text{Approval} | \text{Second Survey})$ vs. H_1 : $P(\text{Approval} | \text{First Survey}) \neq P(\text{Approval} | \text{Second Survey}) \Rightarrow$
- H_0 : $\pi_{2.} = \pi_{.2}$ vs της εναλλακτικής H_1 : $\pi_{2.} \neq \pi_{.2}$

4.8. Two categorical variables

4.8.2. Testing for the equality of proportions in dependent samples



Example 4-13: Prime minister's approval the last six months

	Second Survey		
First Survey	Approve	Disapprove	Total
Approve	794	150	944
Disapprove	86	570	656
Total	880	720	1600

4.8. Two categorical variables

4.8.2. Testing for the equality of proportions in dependent samples



Example 4-13: Prime minister's approval Setting up the data directly as a table

```
tabex4.13b <- as.table( matrix(c(794, 150, 86, 570),2,2, byrow=TRUE) )
```

```
tabex4.13b
```

```
rownames(tabex4.13b)<-c( 'Approve', 'Disapprove' )
```

```
colnames(tabex4.13b)<-c( 'Approve', 'Disapprove' )
```

```
tabex4.13b
```

```
dimnames(tabex4.13b)
```

```
names(dimnames(tabex4.13b)) <- c('First Survey','Second Survey')
```

```
dimnames(tabex4.13b)
```

```
tabex4.13b
```

```
> dimnames( tabex4.13b)
```

```
[[1]]
```

```
[1] "Approve" "Disapprove"
```

```
[[2]]
```

```
[1] "Approve" "Disapprove"
```

```
> names( dimnames( tabex4.13b) ) <- c( 'First Survey', 'Second Survey' )
```

```
> dimnames( tabex4.13b)
```

```
$`First Survey`
```

```
[1] "Approve" "Disapprove"
```

```
$`Second Survey`
```

```
[1] "Approve" "Disapprove"
```

```
> tabex4.13b
```

```
Second Survey
```

```
First Survey Approve Disapprove
```

```
Approve 794 150
```

```
Disapprove 86 570
```

4.8. *Two categorical variables*

4.8.2. *Testing for the equality of proportions in dependent samples*



Example 4-13: Prime minister's approval Using the function Crosstable

```
library(gmodels)
CrossTable(tabex4.13b, prop.r=T, prop.c=T, prop.t=T, prop.chisq=F,
           chisq = T, fisher=T,
           mcnemar=TRUE, format='SPSS', digits=1)
```

4.8. Two categorical variables

4.8.2. Testing for the equality of proportions in dependent samples



Example 4-13: Prime minister's approval Using the function Crosstable

Total Observations in Table: 1600

First Survey	Second Survey		Row Total
	Approve	Disapprove	
Approve	794 84.1% 90.2% 49.6%	150 15.9% 20.8% 9.4%	944 59.0%
Disapprove	86 13.1% 9.8% 5.4%	570 86.9% 79.2% 35.6%	656 41.0%
Column Total	880 55.0%	720 45.0%	1600

4.8. Two categorical variables

4.8.2. Testing for the equality of proportions in dependent samples



Example 4-13: Prime minister's approval Using the function Crosstable

Total Observations in Table: 1600

First Survey	Second Survey		Row Total
	Approve	Disapprove	
Approve	794 84.1% 90.2% 49.6%	150 15.9% 20.8% 9.4%	944 59.0%
Disapprove	86 13.1% 9.8% 5.4%	570 86.9% 79.2% 35.6%	656 41.0%
Column Total	880 55.0%	720 45.0%	1600

4.8. Two categorical variables

4.8.2. Testing for the equality of proportions in dependent samples



Example 4-13: Prime minister's approval Using the function Crosstable

Statistics for All Table Factors

Pearson's Chi-squared test

Chi² = 788.3197 d.f. = 1 p = 1.868886e-173

Pearson's Chi-squared test with Yates' continuity correction

Chi² = 785.4536 d.f. = 1 p = 7.847587e-173

McNemar's Chi-squared test

Chi² = 17.35593 d.f. = 1 p = 3.099293e-05

McNemar's Chi-squared test with continuity correction

Chi² = 16.8178 d.f. = 1 p = 4.114562e-05

Chi square test significant as expected since the two variables are dependent by definition

Statistically significant => difference in the marginal proportions => there is a difference in the approval of the Prime minister

4.8. Two categorical variables

4.8.2. Testing for the equality of proportions in dependent samples



Example 4-13: Prime minister's approval Using the function `mcnemar.test`

```
> chisq.test(tabex4.13b)
```

```
Pearson's Chi-squared test with Yates' continuity correction
```

```
data: tabex4.13b
```

```
X-squared = 785.4536, df = 1, p-value < 2.2e-16
```

```
> mcnemar.test(tabex4.13b)
```

```
McNemar's Chi-squared test with continuity correction
```

```
data: tabex4.13b
```

```
McNemar's chi-squared = 16.8178, df = 1, p-value = 4.115e-05
```

Statistically significant
=> difference in the
marginal proportions =>
there is difference in the
side effects before and
after the treatment

4.8. Two categorical variables

4.8.2. Testing for the equality of proportions in dependent samples



Example 4-13: Prime minister's approval

Changing the table to a data.frame (with counts)

```
tabex4.13b <- as.table(matrix(c(794, 150, 86, 570),2,2, byrow=TRUE) )
rownames(tabex4.13b)<-c( 'Approve', 'Disapprove' )
colnames(tabex4.13b)<-c( 'Approve', 'Disapprove' )
dimnames(tabex4.13b)
names(dimnames(tabex4.13b)) <- c('First Survey','Second Survey')
```

```
> ex4.13b <- as.data.frame(tabex4.13b)
```

```
> ex4.13b
```

	First.Survey	Second.Survey	Freq
1	Approve	Approve	794
2	Disapprove	Approve	86
3	Approve	Disapprove	150
4	Disapprove	Disapprove	570

4.8. Two categorical variables

4.8.2. Testing for the equality of proportions in dependent samples



Example 4-13: Prime minister's approval Changing the table to a data.frame (fully expanded)

```
index<-1:nrow(ex4.13b)
dfex4.13b <- ex4.13b[rep(index,ex4.13b$Freq),-3]
head(dfex4.13b)
```

```
  First.Survey Second.Survey
1      Approve      Approve
1.1    Approve      Approve
1.2    Approve      Approve
1.3    Approve      Approve
1.4    Approve      Approve
1.5    Approve      Approve
```

```
table(dfex4.13b)
```

```
> table(dfex4.13b)
      Second.Survey
First.Survey Approve Disapprove
Approve      794     150
Disapprove   86     570
```


4.8. *Two categorical variables*

4.8.2. *Testing for the equality of proportions in dependent samples*



Example 4-13: Prime minister's approval Changing the table to a data.frame

```
sjt.xtab(ex4.13b[,1],ex4.13b[,2], weight.by=ex4.13b[,3],  
var.labels = names(ex4.13b),  
show.cell.prc = T, show.row.prc = T,  
encoding = "UTF-8")
```

OR

```
sjt.xtab(dfex4.13b[,1],dfex4.13b[,2], var.labels= names(ex4.13b),  
show.cell.prc= T, show.row.prc = T,  
encoding = "UTF-8")
```

4.8. Two categorical variables

4.8.2. Testing for the equality of proportions in dependent samples



Example 4-13: Prime minister's approval Using the function Crosstable

<i>First.Survey</i>	<i>Second.Survey</i>		<i>Total</i>
	Approve	Disapprove	
Approve	794 84.1 % 49.6 %	150 15.9 % 9.4 %	944 100.0 % 59 %
Disapprove	86 13.1 % 5.4 %	570 86.9 % 35.6 %	656 100.0 % 41 %
<i>Total</i>	880 55 % 55 %	720 45 % 45 %	1600 100.0 % 100.0 %

$$X^2=785.454 \cdot df=1 \cdot \Phi=0.702 \cdot p=0.000$$

4.8. Two categorical variables

4.8.2. Testing for the equality of proportions in dependent samples



Example 4-13: Prime minister's approval Using the function Crosstable

<i>First.Survey</i>	<i>Second.Survey</i>		<i>Total</i>
	Approve	Disapprove	
Approve	794 84.1 %	150 15.9 %	944 100.0 %
	49.6 %	9.4 %	59 %
Disapprove	86 13.1 %	570 86.9 %	656 100.0 %
	5.4 %	35.6 %	41 %
<i>Total</i>	880 55 %	720 45 %	1600 100.0 %
	55 %	45 %	100.0 %

$$X^2=785.454 \cdot df=1 \cdot \Phi=0.702 \cdot p=0.000$$

4.8. *Two categorical variables*

4.8.2. *Testing for the equality of proportions in dependent samples*



The mcnemar.test for more than two levels

In the case of paired values with more than two levels, `mcnemar.test` implements the generalization of McNemar test which does not test the equality of the marginal distributions but the symmetry of the table which is a special case of the marginal homogeneity