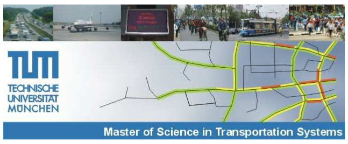
M.Sc. in ‚Transportation Systems‘



Applied Statistics in Transport
Statistical Tests, Models

Regine Gerike
Technische Universität München, mobil.TUM
regine.gerike@tum.de

Munich, 10/01/2011

Last Week: Statistical Tests on the Mean

- One sample problem
- Comparison of two independent samples
- Comparison of two paired samples
- Variance known/unknown

Plan for Today's Lecture:
Statistical Tests on the Mean, Statistical Models

- Tests on normal/non-normal data

- Checking the assumptions:
- Comparison of variances
- Test for normality of the data

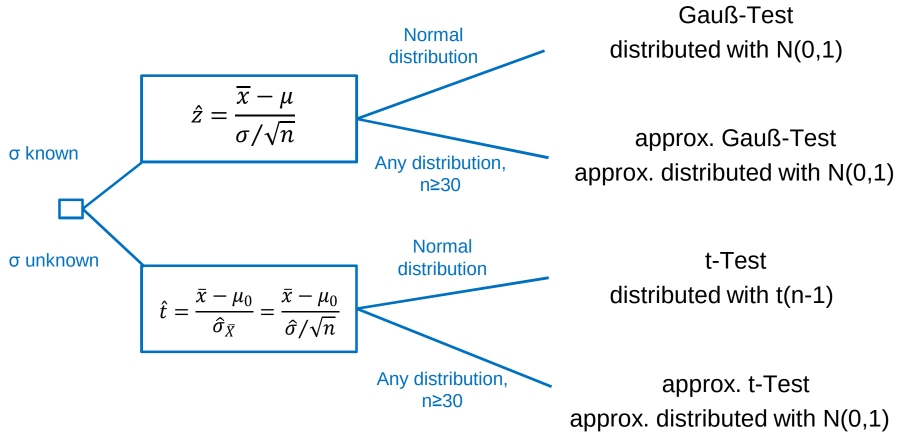
- Statistical Models - Overview

Outlook

9	11.01.2011	2 lectures, room 0790/N1090, Rest tests, statistical models, ANOVA
	18.01.2011	Guest: Vasco Paul Kolmorgen, bahnconcept
10	**25.01.2011**	Exercise confidence, hypothesis testing + lecture regression, room 0790/N1090
11	**01.02.2011**	Repetition
12	08.02.2011	Exam, room N1090

One-Sample-Tests: Tests for the Mean

a) H_0: $\mu = \mu_0$ H_1: $\mu \neq \mu_0$
b) H_0: $\mu \geq \mu_0$ H_1: $\mu < \mu_0$
c) H_0: $\mu \leq \mu_0$ H_1: $\mu > \mu_0$



Tests for a Difference in Means, two Samples

Condition	Test Statistic	Distribution
$X_1 \sim N(\mu_1, \sigma_{X_1}^2)$ $X_2 \sim N(\mu_2, \sigma_{X_2}^2)$ $\boxed{\sigma_1^2, \sigma_2^2\ known}$	$\frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$N(0,1)$
$X_1 \sim N(\mu_1, \sigma_{X_1}^2)$ $X_2 \sim N(\mu_2, \sigma_{X_2}^2)$ $\boxed{\sigma_1^2 = \sigma_2^2\ unknown}$	$\frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{(n_1 - 1) + (n_2 - 1)} * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	$t(n_1 + n_2 - 2)$
$X_1 \sim N(\mu_1, \sigma_{X_1}^2)$ $X_2 \sim N(\mu_2, \sigma_{X_2}^2)$ $\boxed{\sigma_1^2, \sigma_2^2\ unknown}$	$\frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$	$t(k)$ $for\ n_1, n_2 \geq 30$ $appr. N(0,1)$
$\boxed{X_1, X_2\ any\ distribution \quad n_1, n_2 \geq 30}$	$\frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$	$appr. N(0,1)$

$k = (\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2)^2 / \left(\frac{1}{n_1 - 1}\left(\frac{\hat{\sigma}_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{\hat{\sigma}_2^2}{n_2}\right)^2\right)$ = degrees of freedom for unequal variances, $\sigma_1^2, \sigma_2^2\ unknown$

Comparison of Two Samples for their Central Tendency: Wilcoxon-Test

Wilcoxon-test: Non-parametric alternative to Student's t test.

- Can be used for ordinal variables; if interval data is non-normal; when sample size is small

Procedure unpaired test:

- Put both samples in one array, with their sample names clearly attached.
- Sort the aggregate list, taking care to keep the sample labels with their respective values.
- Assign a rank to each value, with ties getting the appropriate average rank (two way ties get: (rank i+ (rank i+1))/2, three-way ties: (rank i+(rank i+1)+(rank i+2))/3 and so on.
- Finally the ranks are added up for each of the two samples, and significance is assessed on size of the smaller sum of ranks.
- Pairs with differences equal to zero are not considered for this calculation; is the share of such pairs high, this is a strong indication that the null hypothesis is true.

Procedure for paired samples:

- Compute the differences in the ranks (similar to the paired t-test) and sort the absolute values of the differences, compare the sum of the differences for each sample.

Wilcoxon-Test on Paired Samples - Example

```
> (solvedQuestionsE<-c(40,60,30,55,55,35,30,35,40,35,50,25,10,40,55))
 [1] 40 60 30 55 55 35 30 35 40 35 50 25 10 40 55
> (solvedQuestions<- c(48,55,44,59,70,36,44,28,39,50,64,22,19,53,60))
 [1] 48 55 44 59 70 36 44 28 39 50 64 22 19 53 60
> t.test(solvedQuestionsE,solvedQuestions,paired=T)

        Paired t-test

data:  solvedQuestionsE and solvedQuestions
t = -3.156, df = 14, p-value = 0.007008
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.749414  -2.050586
sample estimates:
mean of the differences
                   -6.4

> wilcox.test(solvedQuestionsE,solvedQuestions,paired=T)

        Wilcoxon signed rank test with continuity correction

data:  solvedQuestionsE and solvedQuestions
V = 17, p-value = 0.01564
alternative hypothesis: true location shift is not equal to 0

Warnmeldung:
In wilcox.test.default(solvedQuestionsE, solvedQuestions, paired = T) :
  kann bei Bindungen keinen exakten p-Wert Berechnen
> (Wilcox<-(solvedQuestionsE-solvedQuestions))
 [1]  -8   5 -14  -4 -15  -1 -14   7   1 -15 -14   3  -9 -13  -5
> (absWilcox<-abs(Wilcox))
 [1]  8  5 14  4 15  1 14  7  1 15 14  3  9 13  5
> (rgWilcox<-rank(absWilcox))
 [1]  8.0  5.5 12.0  4.0 14.5  1.5 12.0  7.0  1.5 14.5 12.0  3.0  9.0 10.0  5.5
> (rgWilcox[Wilcox>0])
[1] 5.5 7.0 1.5 3.0
> sum(rgWilcox[Wilcox>0])
[1] 17
```

Wilcoxon-Test on Paired Samples - Example

```
> (solvedQuestionsE<-c(40,60,30,55,55,35,30,35,40,35,50,25,10,40,55))
 [1] 40 60 30 55 55 35 30 35 40 35 50 25 10 40 55
> (solvedQuestions<- c(48,55,44,59,70,36,44,28,39,50,64,22,19,53,60))
 [1] 48 55 44 59 70 36 44 28 39 50 64 22 19 53 60
> wilcox.test(solvedQuestionsE,solvedQuestions,paired=T)

        Wilcoxon signed rank test with continuity correction

data:  solvedQuestionsE and solvedQuestions
V = 17, p-value = 0.01564
alternative hypothesis: true location shift is not equal to 0

Warnmeldung:
In wilcox.test.default(solvedQuestionsE, solvedQuestions, paired = T) :
  kann bei Bindungen keinen exakten p-Wert Berechnen
> (Wilcox<-(solvedQuestionsE-solvedQuestions))
 [1]  -8   5 -14  -4 -15  -1 -14   7   1 -15 -14   3  -9 -13  -5
> (absWilcox<-abs(Wilcox))
 [1]  8  5 14  4 15  1 14  7  1 15 14  3  9 13  5
> (rgWilcox<-rank(absWilcox))
 [1]  8.0  5.5 12.0  4.0 14.5  1.5 12.0  7.0  1.5 14.5 12.0  3.0  9.0 10.0  5.5
> (rgWilcox[Wilcox>0])
[1] 5.5 7.0 1.5 3.0
> sum(rgWilcox[Wilcox>0])
[1] 17
```

For n>20 the test statistic $\sum \text{rank}|d_i|$ for $d_i > 0$ is approx. normally distributed with $N(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24})$.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
d_i	-8	5	-14	-4	-15	-1	-14	7	1	-15	-14	3	-9	-13	-5
$\|d_i\|$	8	5	14	4	15	1	14	7	1	15	14	3	9	13	5
rank$\|d_i\|$	8	5.5	12	4	14.5	1.5	12	7	1.5	14.5	12	3	9	10	5.5

Ties for rank of difference=14: (11+12+13)/3; for 15: (14+15)/2

Sum of ranks for positive differences: (5.5+7+1.5+3)=17

Comparison of Two Samples for their Central Tendency: Wilcoxon-Test

- For the above example we get:

```
> t.test(solvedQuestionsE,solvedQuestions,paired=T)

        Paired t-test

data:  solvedQuestionsE and solvedQuestions
t = -3.156, df = 14, p-value = 0.007008
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.749414  -2.050586
sample estimates:
mean of the differences
                   -6.4

> wilcox.test(solvedQuestionsE,solvedQuestions,paired=T)

        Wilcoxon signed rank test with continuity correction

data:  solvedQuestionsE and solvedQuestions
V = 17, p-value = 0.01564
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(solvedQuestionsE, solvedQuestions, paired = T) :
  kann bei Bindungen keinen exakten p-Wert Berechnen
```

- The p-value of 0.01564 is less than 0.05, so we reject the null hypothesis, and conclude that the means are significantly different.
- The t-test gives the lower p value, so the Wilcoxon-test is said to be conservative: if a difference is significant under a Wilcoxon-test it would have been even more significant under a t test.

Chi-Squared Contingency Tables

- For count data (whole numbers or integers)
- Definition of contingency: a thing that depend on an uncertain event.
- In statistics: the contingencies are all the events that could possibly happen: a contingency table shows the counts of how many times each of the contingencies actually happen in a particular sample.
- We test the independency of the two events: (Pearson's) Chi-squared test.

Chi-Squared Contingency Tables: Example from Lecture Probability

Samples of emissions from three suppliers are classified for conformance to air quality specifications. The results from 100 samples are summarised as follows:

		Conforms		
		Yes	No	C
	1	22	8	30
Supplier	2	25	5	30
	3	30	10	40
Total (R)		77	23	100

R = Row total

C = Column total

G = 100 = Grand total

How should the table look like for independent events? E=(R*C)/G

		Conforms		
		Yes	No	C
	1	23.1	6.9	30
Supplier	2	23.1	6.9	30
	3	30.8	9.2	40
Total R		77	23	100

There are differences between the observed (O) and the expected frequencies (E). The Chi-squared test tests whether it is significant.

Chi-Squared Contingency Tables: Example

Test statistic X^2 for the Chi-squared test: $X^2 = \sum \frac{(O-E)^2}{E}$

Degrees of freedom: d.f.=(r-1)*(c-1)

Example: d.f.=(r-1)*(c-1)=(3-1)*(2-1)=2

	Observed frequency (O)	Expected frequency for independent events (E)	(O-E)^2	(O-E)^2/E
S1,CY	22	23.1	1.21	0.05
S2,CY	25	23.1	3.61	0.16
S3,CY	30	30.8	0.64	0.02
S1,CN	8	6.9	1.21	0.18
S2,CN	5	6.9	3.61	0.52
S3,CN	10	9.2	0.64	0.07
Total				0.998

Chi-Squared Contingency Tables: Example

```
> #Chi-Squared Contingency Tables,Crawely 301, ex from lecture probabilities
> (o<-c(22,25,30,8,5,10))  #observed frequencies
[1] 22 25 30  8  5 10
> (e<-c(23.1,23.1,30.8,6.9,6.9,9.2)) #expected frequencies for independent events
[1] 23.1 23.1 30.8  6.9  6.9  9.2
> (o_minus_e_squared<-(o-e)^2)
[1] 1.21 3.61 0.64 1.21 3.61 0.64
> (o_minus_e_squared_divided_by_e<-((o-e)^2)/e)
[1] 0.05238095 0.15627706 0.02077922 0.17536232 0.52318841 0.06956522
> sum(o_minus_e_squared_divided_by_e)   #Test statistic chi-squared: 0.9975532
[1] 0.9975532
> # d.f.=(r-1)*(c-1)=(3-1)*(2-1)=2
> #qchisq(p, df), p=vector of probabilities
> qchisq(0.95,2)  #5.991465, cuts 5% of the right hand tail
[1] 5.991465
> #dchisq(x, df)
> 1-pchisq(0.9975532,2) #0.6072731
[1] 0.6072731
> #With R-command chisq.test:
> (count<-matrix(c(22,25,30,8,5,10),nrow=3))
     [,1] [,2]
[1,]   22    8
[2,]   25    5
[3,]   30   10
> chisq.test(count)

        Pearson's Chi-squared test

data:  count
X-squared = 0.9976, df = 2, p-value = 0.6073
```

Conclusion: The test statistic is inside the critical region, lower than the critical value, so we cannot reject the null hypothesis, the differences between the observed and the expected values are not significant at the 5% level.

Checking the Conditions for Applying the T-Test

We look at homogeneity of variances and normality of the data.

Fisher's F test: Comparing two Variances

- Fisher's F test: compares two sample variances.
- Should be done before comparing two sample means: test whether two variances are significantly different (homo-/heteroscedasticity).
- Procedure: divide the larger by the smaller variance
- In order to be different, the ratio will need to be significantly bigger than 1 (because the larger variance goes on top, in the numerator).
- Decide on the significance of the variance ratio with the help of the critical value of the variance ratio: the critical value of the Fisher's F test.

- R-function for getting critical values: “qf(p, df1, df2)”: quantiles of the F distribution
- var.test(x, y, ratio = 1,
  alternative = c("two.sided", "less", "greater"), conf.level = 0.95, ...)

Fisher's F test: Comparing two Variances, Example

- F-test for the above example:

```
#F-test
(solvedQuestionsE<-c(40,60,30,55,55,35,30,35,40,35,50,25,10,40,55))
(solvedQuestions<- c(48,55,44,59,70,36,44,28,39,50,64,22,19,53,60))
length(solvedQuestionsE) #15
length(solvedQuestions) #15
var(solvedQuestionsE) #183.8095
var(solvedQuestions)  #228.6381
var(solvedQuestions)/var(solvedQuestionsE)  #1.243886
#qf(p, df1, df2)
qf(0.95,14,14) #2.483726
pf(2.483726,14,14)  #0.95
2*(1-pf(1.2439,14,14))  #0.6886401
#We double the probability to allow for the two-tailed test:
#The probability that the variances are the same is p<0.68.
var.test(solvedQuestions,solvedQuestionsE)
#        F test to compare two variances
#data:  solvedQuestions and solvedQuestionsE
#F = 1.2439, num df = 14, denom df = 14, p-value = 0.6887
#alternative hypothesis: true ratio of variances is not equal to 1
#95 percent confidence interval:
# 0.4176094 3.7050234
#sample estimates:
#ratio of variances
#          1.243886
```

You need not to compute the CI but you should be able to interpret it, the number one (equal variances is included here).

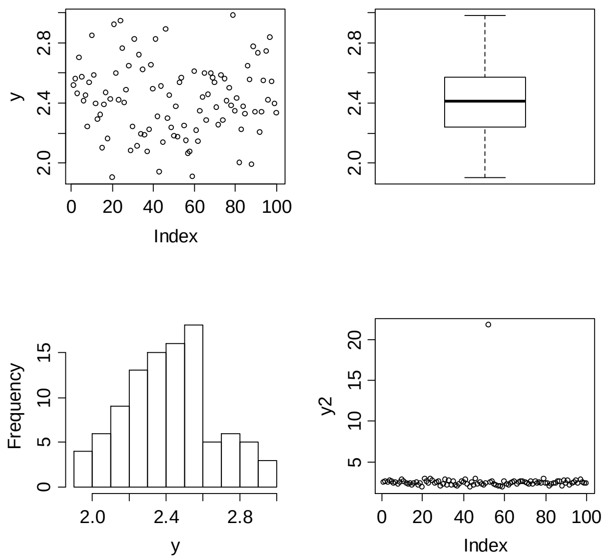
There are no significant differences between the two variances.

Test for Normality of Your Data – Data Summary

```
> setwd("c:\\Users\\Regine\\Documents\\Lacie\\Daten\\R\\")
> data<-read.table("CrawleyTheRBook\\das.txt",header=T)
> attach(data)

> str(data)
'data.frame':   100 obs. of  1 variable:
 $ y: num  2.51 2.56 2.46 2.7 2.57 ...
> par(mfrow=c(2,2))
> plot(y,cex.axis=1.5,cex.lab=1.5)  #index plot
> boxplot(y,cex.axis=1.5,cex.lab=1.5) #box-and-whisker-Plot
> hist(y,main="",cex.axis=1.5,cex.lab=1.5)  #default title without main="": Histogram of y
> y2<-y
> y2[52]<-21.75
> plot(y2,cex.axis=1.5,cex.lab=1.5) #index plot with outlier
> summary(y) #min,1.qu.,median,mean,3.qu.,max
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.904   2.241   2.414   2.419   2.568   2.984
```

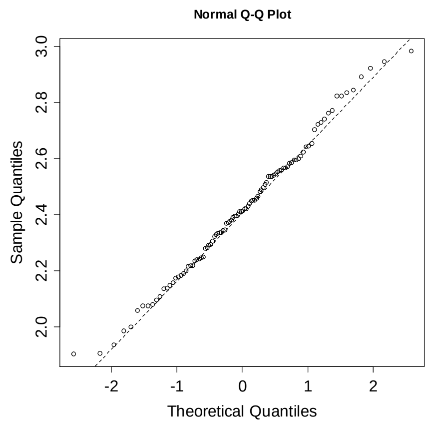
Test for Normality of Your Data – Data Summary



Quantile-Quantile-Plots for Testing Normality

- Simplest way (and in many cases the best) is the 'quantile-quantile-plot'
- Plots the ranked samples from our distribution against a similar number of ranked quantiles taken from a normal distribution.

- If our sample is normally distributed, then the line will be straight.
- Departures from normality show up as various sorts of non-linearity (e.g. S-shapes or banana-shapes).
- This shows a slight S-shape, but there is no compelling evidence of non-normality (our distribution is somewhat skew to the left; see histogram above)




```
> par(mfrow=c(1,1))
> qqnorm(y,cex.axis=1.5,cex.lab=1.5) #plots the data against the normal distribution points
> qqline(y,lty=2) #plots the ideal line, works only as second command, needs the plot before
```

Summary: Statistical Tests on the Mean

- One sample problem
- Comparison of two independent samples
- Comparison of two paired samples
- Variance known/unknown
- Normal/non-normal data

- Checking the assumptions:
- Comparison of variances
- Test for normality of the data

Next step: Statistical Modelling - Overview

Statistical Modelling - Introduction

- Choosing the right kind of statistical analysis is an important step of your analysis
- Choice depends on the nature of your data and on the particular question you are trying to answer

Key:

- Type of response variable = variable whose variation you are attempting to understand, y-axis of the graph
- Type of explanatory variables = which variation in the response variable is associated with variation in the explanatory variable

Questions:

- Which of your variables is the response variable?
- Which are the explanatory variables?
- Are the explanatory variables continuous or categorical, or a mixture of both?
- What kind of response variable do you have: is it a continuous measurement, a count, a proportion, a time at death, a category?

- These simple keys will lead you to the appropriate statistical model.

Statistical Modelling - Introduction

- Objective: determine the values of the parameters in a specific model that lead to the best fit of the model to the data
- The model is fitted to the data, not the other way round
- Best model: produces the least unexplained variation (the minimal residual deviance), subject to the constraint that all parameters in the model should be statistically significant.
- There is not one model.

- The model should be minimal: principle of parsimony (Occam's Razor):
- Given a set of equally good explanations for a given phenomenon, the correct explanation is the simplest explanation.
- Einstein: The model should be as simple as possible. But not simpler.
- Oscar Wilde: Truth is rarely pure, and never simple.

Principle of Parsimony for Statistical Modelling

- Models should have as few parameters as possible.
- Linear models should be preferred to non-linear models.
- Experiments relying on few assumptions should be preferred to those relying on many.
- Models should be simplified until they are minimal adequate.
- Simple explanations should be preferred to complex explanations.
- Prefer explanatory variables that are easy to measure.
- Prefer models that are based on a sound mechanistic understanding of the process over purely empirical functions.

Model Formulae in R

- Response variable ~ explanatory variable(s)
- The right hand side of the model formulae shows:
- The number of explanatory variables and their identities
- The interactions between the explanatory variables
- (Non-linear terms in the explanatory variables)

- As with the response variable, the explanatory variables can appear as transformations, or as powers, or polynomials.

Model Formulae in R

- Important: symbols are used differently in model formulae than in arithmetic expressions:
- \+ indicates inclusion of an explanatory variable in the model (not addition)
- \- indicates deletion of an explanatory variable from the model (not subtraction)
- * indicates inclusion of explanatory variables and interactions (not multiplication)
- | indicates conditioning (not 'or'),
  so that y~x|z is read as 'y as a function of x given z'
- The colon denotes an interaction.

Model Formulae in R

- Important: symbols are used differently in model formulae than in arithmetic expressions:
- A*B*C is the same as A+B+C+A:B+A:C+B:C+A:B:C
- (A+B+C)^3 is the same as A*B*C
- (A+B+C)^2 is the same as A*B*C-A:B:C

- Interactions between explanatory variables
- Number of interaction effects: (a-1)(b-1)
  with a and b being the number of levels of the two factors
- Interaction between a categorical and a continuous variable are interpreted as an analysis of covariance: a separate slope and intercept are fitted for each level of the categorical variable.

Choosing the Appropriate Statistical Model

The explanatory variables:

- All explanatory variables continuous: **Regression**
- All explanatory variables categorial: **Analysis of Variance (ANOVA)**
- Explanatory variables both continuous and categorial:
  Analysis of covariance (ANCOVA)

The response variable:

- Continuous Normal regression, ANOVA, ANCOVA
- Proportion Logistic regression (glm(y~x,family=binomial)
- Count Log-linear models (glm(y~x,family=poisson)
- Binary Binary logistic analysis (glm(y~x,family=binomial)

Multivariate models: More than one response variable

- Multivariate ANalysis Of VAriance - MANOVA
- Principal Component Analysis - PCA
- Cluster analysis

Regression

- Regression and correlation analysis:
  describe and analyse the relationship between random variables
- Regression: describes the type of directional relationship between mainly ratio/interval scaled variables (the more … the more/less …)
- Correlation: describes the intensity of the non-directional relationship
- Example:
- Relation between the cubic capacity and the fuel consumption of a car
- Regression of the response variable PS as a function of household type, income, spatial location of the household, usage of the car

ANOVA

- Response variable: ratio/interval scaled
- Explanatory variables: categorical (ordinal, nominal, grouped interval data), are called factors
- Each factor has two or more levels
- For one single factor we can use the t-test

- ANOVA can be classified by the number of explanatory variables
- One-way ANOVA: one explanatory variable
- Two- and more way ANOVA: interaction effects can be included.

- We test whether the differences between the means of the different groups are high enough to conclude on differences in the populations; is the variation between the groups higher than within the groups?

Analysis of Covariance – ANCOVA

- Combines elements from regression and ANOVA
- Response variable is continuous
- At least one continuous and at least one categorical explanatory variable

- Approach:
- Fit two or more linear regressions of y against x (one for each level of the factor)
- Estimate different slopes and intercepts for each level
- Use model simplification (deletion tests) to eliminate unnecessary parameters

Analysis of Covariance – ANCOVA, Example

- Medical experiment: response variable: days to recovery; explanatory variables: smoker or not (categorical) and blood cell count (continuous)
- Economics: response variable: local unemployment rate; explanatory variables: country (categorical) and population size (continuous)

- Weight: response variable: weight; explanatory variables: sex (categorical) and age (continuous)
- Maximal model: four parameters: two slopes (one for males and one for females) and two intercepts (one for males and one for females):
- $Weight_{male} = a_{male} + b_{male} * age$
- $Weight_{female} = a_{female} + b_{female} * age$
- Try to simplify the model (principle of parsimony):
- Possible models: two intercepts and a common slope, one intercept and two slopes, ...

Factor Analysis / Principal Component Analysis

- Factor Analysis is a familiy of approaches, PCA is one of them
- Basic aim of PCA: describe variation in a set of correlated variables, x_1, x_2, ..., x_n, in terms of a new set of uncorrelated variables, y_1, y_2, ..., y_n, each of which is a linear combination of the x variables.
- The new variables are derived in decreasing order of “importance” in the sense that 1 accounts for as much of the variation in the original data amongst all linear combinations of x_1, x_2, ..., x_n.
- Then y_2 is chosen to account for as much as possible of the remaining variation, subject to being uncorrelated with y_1, etc.
- The new variables defined by this process, y_1, y_2, ..., y_n, are the principal components.
- PCA is often done as a first step of cluster analysis.

Cluster Analysis

- Cluster analysis is a generic term for a range of numerical methods for examining multivariate data
- Goal: uncover groups/clusters of observations that are homogeneous and separated from other groups
- Clusters are identified by the relative distances between points
- Examples for methods for finding the clusters:
- Agglomerative hierarchical clustering: start with each object being a cluster, compute the distances, merge the two objects with the smallest distance, compare the remaining clusters, merge the two with the smallest distance, etc.
- k-means: consider every possible partition of the n objects into k groups, select the one with the lowest within-sum-of-squares, search algorithms are necessary, initial partition can be found with hierarchical clustering techniques
- Example: Clustering of test-drive-data for finding typical clusters of kinematic characteristics (e.g. average speed, acceleration, share stop-and-go)