M.Sc. in ‚Transportation Systems'

Applied Statistics in Transport
Statistical Tests

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

Munich, 21/12/2010

Your Data Set: Mobilität in Deutschland 2008

- One diary day
- Weekdays and weekend days during a whole year
- All region types
- Covering the whole household (including children, complete households in 81% of the cases)

Your Data Set: Mobilität in Deutschland 2008

	households	persons	cars	trips
in general	• household size • vehicle ownership • telephone • telephone number • income • cell phone, computer, internet • residential area • profile household members • email-adress	• socio-demographics • school/occupation • driving licences • long distance trips last quarter • duration of residence • accessibility local public transport • mobility handicap • car availability • local public transport subscription/season ticket • bike availability • use of means of transport in general • accessibility of normal destinations	• vehicle data • usual parking space at home • annual mileage • main driver	
diary day		• being out • normal day • car availability • weather	• odometer reading	• purpose/destination • means of transport • distance • duration (departure/arrival) • destination address • number of persons • additional module • business trips • use of household vehicle

red: expansion compared to previous surveys of the KONTIV type (since 2002)

grey: abbreviation 2008

Your Data Set: Mobilität in Deutschland 2008

- You get:

Rdata-files:

- H2008.Rdata, P2008.Rdata, W2008.Rdata
- load("...\\w2008.Rdata")
- save(w2008, file="...\\W2008.Rdata")

Textfiles:

- MiD2008_PUF_Wege.dat, MiD2008_PUF_Personen.dat, MiD2008_PUF_Haushalte.dat
- Each of the data sets contains information from other data sets

Codeplan:

- MiD2008 English Codeplan_20100521.xlsx

Your Data Set: Mobilität in Deutschland 2008

- Relevant variables:
- Number of trips on the diary day: wege1
- Travel time diary day: anzmin
- Travel distances diary day: anzkm
- Yearly distances travelled by car [km/year]: fahrlj_h
- CO_2-Emissions: co2tag_h (Household, diary day), co2tag_p (Person, diary day), co2weg (per trip)
- Household income: hheink

- Household variables: household type and household size, number of employed household members, number of driving licences, number of cars/bikes/motor cycles, reasons for not owning a car, type of region
- Person variables: car/bike availability, frequency usage car/bike/PT/airplane, gender, age, distances to and accessibility of facilities, employment status, education, type of PT-tickets, driving licence, availability handies/computers, mobility restrictions, type of region

Examples for research questions from last year

- Does the number of kilometers driven depend on the accessibility of specific facilities?
- Does the number of cars in a household depend on …
- Does the satisfaction with public transport affect the km driven?
- Does the distance travelled depend on the household income?

Project Work

- Data Set: “Mobilität in Deutschland 2008” (MiD 2008) (Mobility in Germany)
- 6/7 groups à max. 6 students
- Task: Write one paper per group
- Max. 7,000 words
- Deadline: 13/03/2011
- One note per group
- Paper (30%) + exam (70%) = Note Applied Statistics in Transport

- Structure of the paper:
- Title, abstract, introduction, main part, conclusions, R-code

Project Work

Steps for project:

- Data Preparation, validation
- Descriptive statistics, visualization
- Research questions
- T-tests
- Optional: other tests, ANOVA, regression analysis

- No literature research!

Main Goals:

- Work with real data (read and prepare the data)
- Get familiar with exploratory data analysis (descriptive measures, graphics)
- Formulate research questions
- Write a paper: with a clear line of argument, a consistent layout, an introduction and conclusions

Last Week: Hypothesis Testing

- General procedure for hypothesis testing
- Example
- Your Project
- Hints for your paper (see last handout)

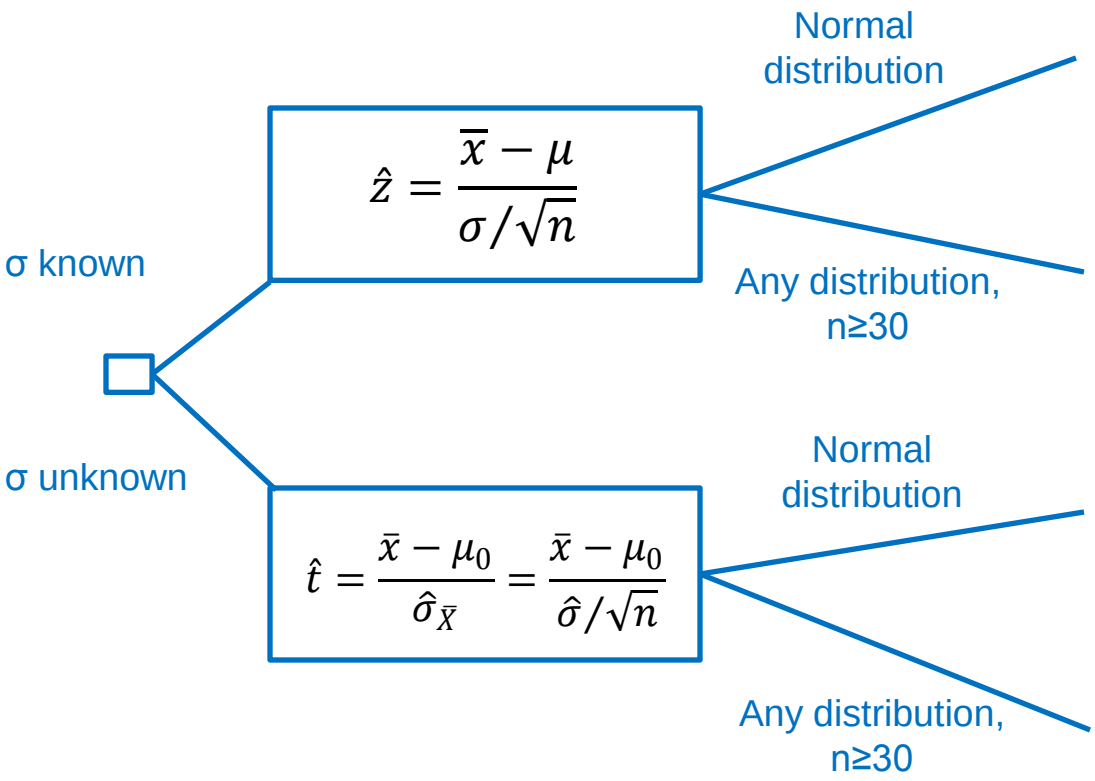
Plan for Today's Lecture: Statistical Tests on the Mean

- 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

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) $H_0: \mu \leq \mu_0$ $H_1: \mu > \mu_0$



Gauß-Test
distributed with N(0,1)

approx. Gauß-Test
approx. distributed with N(0,1)

t-Test
distributed with t(n-1)

approx. t-Test
approx. distributed with N(0,1)

From Last Lecture: One Sample Problem

General Procedure for Hypotheses Tests: Tests for the Mean of a Normal Distribution, Variance known

- Step 1:
  Formulate the null hypothesis H_0, and the alternative hypothesis H_1.
  H_0: $\mu_0=\mu_1$; H_1: $\mu_0\neq\mu_1$
- Step 2:
  Choose a significance level α.
- Step 3:
  Determine the appropriate test statistic. $\hat{z} = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$ Gauß-Test
- Step 4:
  State the rejection region for the test statistic.
- Step 5:
  Compute the necessary sample quantities, substitute these into the equation for the test statistic, and compute that value.
- Step 6:
  Decide whether or not H_0 should be rejected. Formulate your conclusion in terms of original problem.

General Procedure for Hypotheses Tests: Tests for the Mean of a Normal Distribution, Variance unknown

- Step 1:
  Formulate the null hypothesis H_0, and the alternative hypothesis H_1.
  H_0: $\mu_0=\mu_1$; H_1: $\mu_0\neq\mu_1$
- Step 2:
  Choose a significance level α.
- Step 3:
  Determine the appropriate test statistic.

t-test

$$\hat{t} = \frac{\bar{x} - \mu_0}{\hat{\sigma}_{\bar{X}}} = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$$ (with d.f.=n-1)

- Step 4:
  State the rejection region for the test statistic.
- Step 5:
  Compute the necessary sample quantities, substitute these into the equation for the test statistic, and compute that value.
- Step 6:
  Decide whether or not H_0 should be rejected. Formulate your conclusion in terms of original problem.

Two-Sample-Tests

- So far we compared a sample mean with a population mean.
- Example:
- Households of type „two adults with children“ and an average income of x
- own on average 1.8 cars.
- Households of the same type in a car-free neighborhood
- own on average 1.6 cars.
- Are there significant differences between the two means?

- Now we compare the means of two independent samples:
- Do you remember from first class?
- The city of Munich offers welcome information packages
- with information on public transport services to new citizens.
- For monitoring purposes two groups of people who recently moved to Munich were surveyed: one group with and the control group without the welcome package.
- What statistical methods are suitable for analysing whether the mean distance travelled by public transport differs significantly between the two groups?

Tests for a Difference in Means, two Samples

- Given what we know about the within-sample-variance, how likely is it that our two sample means were drawn from populations with the same averages?
- If it is highly likely we say that the two sample means are not significantly different.
- If it is rather unlikely we say that the two sample means are significantly different.
- As before we use probabilities to decide what is called likely/unlikely.

- Two important tests for comparing two sample means:
- Gauß-test/student's t-test for independent/paired samples and normally distributed data/large sample sizes.
- Wilcoxon's rank-sum test when the samples are independent/paired but the data is not normally distributed.

Tests for a Difference in Means, two Samples

Condition	Test Statistic	Distribution
$X_1 \sim N(\mu_1, \sigma_{X_1}^2)$ $X_2 \sim N(\mu_2, \sigma_{X_2}^2)$ σ_1^2, σ_2^2 *known*	$\dfrac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	$N(0,1)$
$X_1 \sim N(\mu_1, \sigma_{X_1}^2)$ $X_2 \sim N(\mu_2, \sigma_{X_2}^2)$ $\sigma_1^2 = \sigma_2^2$ *unknown*	$\dfrac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\dfrac{(n_1-1)\hat{\sigma}_1^2 + (n_2-1)\hat{\sigma}_2^2}{(n_1-1)+(n_2-1)} * (\dfrac{1}{n_1} + \dfrac{1}{n_2})}}$	$t(n+m-2)$
$X_1 \sim N(\mu_1, \sigma_{X_1}^2)$ $X_2 \sim N(\mu_2, \sigma_{X_2}^2)$ σ_1^2, σ_2^2 *unknown*	$\dfrac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\dfrac{\hat{\sigma}_1^2}{n_1} + \dfrac{\hat{\sigma}_2^2}{n_2}}}$	$t(k)$ $for\ n_1, n_2 \geq 30$ $appr.N(0,1)$
X_1, X_2 *any distribution* $n_1, n_2 \geq 30$	$\dfrac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\dfrac{\hat{\sigma}_1^2}{n_1} + \dfrac{\hat{\sigma}_2^2}{n_2}}}$	$appr.N(0,1)$

$k = (\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2)^2 / \left(\frac{1}{n_1-1}\left(\frac{\hat{\sigma}_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{\hat{\sigma}_2^2}{n_2}\right)^2\right)$ = degrees of freedom for unequal variances, σ_1^2, σ_2^2 *unknown*

Tests for a Difference in Means, Variances Known

Consider hypothesis testing on the difference in the means $\bar{x}_1 - \bar{x}_2$ of two normal populations.

H_0: $\bar{x}_1 - \bar{x}_2 = 0$ (or $= \Delta_0$ = some defined distance)

H_1: $\bar{x}_1 - \bar{x}_2 \neq 0$

Take two samples many times and compute the difference between $\bar{x}_1$ and $\bar{x}_2$.

You get the distribution of the differences.

This distribution has the expected value of zero when H_0 is true.

The variance of this distribution is a linear combination of $\sigma^2_{\bar{X}_1}$ and $\sigma^2_{\bar{X}_2}$.

With $\sigma^2_{\bar{X}_1} = \sigma_1^2/n_1$ and $\sigma^2_{\bar{X}_1} = \sigma_2^2/n_2$ we get the standard error of the difference of the means:

$$\sigma_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

For equal population variances we get:

$$\sigma_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\sigma^2 * (\frac{1}{n_1} + \frac{1}{n_2})}$$

Test statistic: $z = \frac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$.

Tests for a Difference in Means, Variances Unknown

Equal variances:

If the population variance is unknown we can estimate it with the variances of the two samples of the normally distributed variables:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n_1}(x_{i1}-\bar{x}_1)^2+\sum_{i=1}^{n_1}(x_{i2}-\bar{x}_2)^2}{(n_1-1)+(n_2-1)}$$

For the standard error with variance unknown we get:

$$\hat{\sigma}_{(\bar{X}_1-\bar{X}_2)} = \sqrt{\frac{\sum_{i=1}^{n_1}(x_{i1}-\bar{x}_1)^2+\sum_{i=1}^{n_1}(x_{i2}-\bar{x}_2)^2}{(n_1-1)+(n_2-1)} * (\frac{1}{n_1}+\frac{1}{n_2})}$$

Or we can work directly with the two estimated population variances $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$:

$$\hat{\sigma}_{(\bar{X}_1-\bar{X}_2)} = \sqrt{\frac{(n_1-1)\hat{\sigma}_1^2+(n_2-1)\hat{\sigma}_2^2}{(n_1-1)+(n_2-1)} * (\frac{1}{n_1}+\frac{1}{n_2})}$$

Unequal variances:

$$\sigma_{(\bar{X}_1-\bar{X}_2)} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1}+\frac{\hat{\sigma}_2^2}{n_2}}$$

Test statistic: $t = \frac{(\bar{x}_1-\bar{x}_2)-\Delta_0}{\hat{\sigma}_{(\bar{X}_1-\bar{X}_2)}}$

For $\Delta_0 = 0$ the equation reduces to: $t = \frac{\bar{x}_1-\bar{x}_2}{\hat{\sigma}_{(\bar{X}_1-\bar{X}_2)}}$

Tests for a Difference in Means

According to the central limit theorem, the two random variables $\bar{X}_1$ and $\bar{X}_2$ are normally distributed for n≥30. Therefore the difference $\bar{X}_1 - \bar{X}_2$ is also normally distributed.

For smaller sample sizes (n1+n2<50) the differences (standardized with the standard error) follow a t-distribution with d.f.=n1+n2-2 if the observations are normally distributed in the original populations.

The relevance of the observed difference $\bar{x}_1 - \bar{x}_2$ from the difference Δ_0 is calculated in relation to $\hat{\sigma}_{(\bar{X}_1 - \bar{X}_2)}$:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\hat{\sigma}_{(\bar{X}_1 - \bar{X}_2)}}$$

When we formulate the null hypothesis with $\Delta_0 = 0$ the equation reduces to:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{(\bar{X}_1 - \bar{X}_2)}}$$

This equation defines a random variable that is for small sample sizes t-distributed with d.f.=n1+n2-2. For bigger sample sizes (n1+n2<50) it is approximately normally distributed. We can always use the t-distributions also for bigger sample sizes.

Tests for a Difference in Means, two Samples

Condition	Test Statistic	Distribution
$X_1 \sim N(\mu_1, \sigma_{X_1}^2)$ $X_2 \sim N(\mu_2, \sigma_{X_2}^2)$ σ_1^2, σ_2^2 *known*	$\dfrac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	$N(0,1)$
$X_1 \sim N(\mu_1, \sigma_{X_1}^2)$ $X_2 \sim N(\mu_2, \sigma_{X_2}^2)$ $\sigma_1^2 = \sigma_2^2$ *unknown*	$\dfrac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\dfrac{(n_1-1)\hat{\sigma}_1^2 + (n_2-1)\hat{\sigma}_2^2}{(n_1-1)+(n_2-1)} * (\dfrac{1}{n_1} + \dfrac{1}{n_2})}}$	$t(n+m-2)$
$X_1 \sim N(\mu_1, \sigma_{X_1}^2)$ $X_2 \sim N(\mu_2, \sigma_{X_2}^2)$ σ_1^2, σ_2^2 *unknown*	$\dfrac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\dfrac{\hat{\sigma}_1^2}{n_1} + \dfrac{\hat{\sigma}_2^2}{n_2}}}$	$t(k)$ $for\ n_1, n_2 \geq 30$ $appr.\ N(0,1)$
X_1, X_2 *any distribution* $n_1, n_2 \geq 30$	$\dfrac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\dfrac{\hat{\sigma}_1^2}{n_1} + \dfrac{\hat{\sigma}_2^2}{n_2}}}$	$appr.\ N(0,1)$

$k = (\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2)^2 / \left(\frac{1}{n-1}\left(\frac{\hat{\sigma}_1^2}{n_1}\right)^2 + \frac{1}{n-1}\left(\frac{\hat{\sigma}_2^2}{n_2}\right)^2\right)$ = degrees of freedom for unequal variances, σ_1^2, σ_2^2 *unknown*

Confidence Interval on a Difference in Means, Variances Known

The 100(1-α)% confidence interval on the difference in two means Δ_0 (variances known) can be computed similarly to the confidence interval of point estimators.

The difference in sample means $\bar{X}_1 - \bar{X}_2$ is a point estimator of Δ_0, and

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sigma_{(\bar{X}_1 - \bar{X}_2)}} = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has a standard normal distribution if the two populations are normal or is approximately normal if the conditions of the central limit theorem apply.

We can define the 100(1-α)% confidence interval for $\Delta_0$as follows:

If $\bar{x}_1$ and $\bar{x}_2$ are the means of independent random samples of sizes n_1 and n_2 from two independent normal populations with known variances σ_1^2 and σ_2^2, respectively, a 100(1-α)% confidence interval for μ_1-μ_2 is:

$$\bar{x}_1 - \bar{x}_2 - z_{(\alpha/2)} * \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \Delta_0 \leq \bar{x}_1 - \bar{x}_2 - z_{(\alpha/2)} * \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Where $z_{(\alpha/2)}$ is the upper $\alpha/2$ percentage point of the standard normal distribution.

The confidence level (1-α) is exact when the populations are normal. For non-normal populations, the confidence level is approximately valid for large sample sizes.

Tests for a Difference in Means - Example

- The city of Munich offers welcome information packages with information on public transport services to new citizens.
- For monitoring purposes two groups of people who recently moved to Munich were surveyed: one group with and the control group without the welcome package.
- The control group (50 individuals) travelled on average 14,400 km/person&year, the treatment group (50 individuals) travelled on average 14,052 km/person&year.
- Was the campaign successful at the 5%-level?
- The standard deviation is known and with 355 km the same in the two samples.

Tests for a Difference in Means – Command t.test in R

- t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95, ...)
- R always assumes unknown variances, with the argument var.equal=T you indicate whether you have equal variances.

```
> t.test(1:10,7:20)

        Welch Two Sample t-test

data:  1:10 and 7:20
t = -5.4349, df = 21.982, p-value = 1.855e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.052802  -4.947198
sample estimates:
mean of x mean of y
      5.5      13.5
```

```
> t.test(1:10)

        One Sample t-test

data:  1:10
t = 5.7446, df = 9, p-value = 0.0002782
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 3.334149 7.665851
sample estimates:
mean of x
      5.5
```

T-Tests on Paired Samples

- A special case of the two-sample t-tests occurs when the observations on the two populations of interest are collected in pairs.
- Each pair of observations is taken under homogeneous conditions, but these conditions may change from one pair to another.

- Examples:
- Comparison of travel behaviour of spouses
- Compare the travel behaviour of the same individual at different points of time/before and after an intervention (repeated sampling).

- In paired t-tests we consider the possible influence of the variance of one sample on the variance of the paired sample and vice versa.

T-Tests on Paired Samples - Procedure

Compute the difference d_i for each pair of observations: $d_i = x_{i1} - x_{i2}$.

Compute the arithmetic mean of the differences d_i: $\bar{x}_d = \frac{\sum_{i=1}^{n} d_i}{n}$

With n being the **number of pairs** of observations.

We test the distribution of the mean of the differences in the pairs.

(Tests for independent samples test the differences of the means of the two samples.)

Standard error $\hat{\sigma}_{\bar{X}_d}$ of the mean of the differences in the pairs: $\hat{\sigma}_{\bar{X}_d} = \frac{\hat{\sigma}_d}{\sqrt{n}}$

The standard deviation of the differences in the population $\hat{\sigma}_d$ we estimate similar to the standard error of the arithmetic mean:

$$\hat{\sigma}_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{x}_d)^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} {d_i}^2 - \left(\left(\sum_{i=1}^{n} d_i\right)^2 / n\right)}{n-1}}$$

We compute the t-value as follows: $t = \frac{\bar{x}_d - \mu_d}{\hat{\sigma}_{\bar{X}_d}}$

For the null hypothesis $\mu_d = 0$ we get: $t = \frac{\bar{x}_d}{\hat{\sigma}_{\bar{X}_d}}$

Compare the t-value with the critical t-value for your level of significance and the degrees of freedom: d.f.=n-1 (n = number of pairs).

For larger sample sizes the t-values can be approximated by the normal distribution.

T-Tests on Paired Samples - Example

- We want to check whether students are able to anticipate their results in the exam.
- 15 candidates should indicate before the exam how many of the questions they will solve correctly.
- The number of actually correctly solved questions is then compared to the ex-ante estimation.
- Does the estimated number of solved questions significantly differ from the number of actually solved questions?
- We assume the differences between the estimated and the actual results to be normally distributed.
- We are not sure whether the students under- or overestimate their results, so we use a two-sided test at the alpha=5% level.

- Estimated results:
- (solvedQuestionsE<-c(40,60,30,55,55,35,30,35,40,35,50,25,10,40,55))
- Actual results:
- (solvedQuestions<- c(48,55,44,59,70,36,44,28,39,50,64,22,19,53,60))

T-Test on Paired Samples - Example

	solvedQuestionsE	solvedQuestions	d_i	d_i^2
1	40	48	-8	64
2	60	55	5	25
3	30	44	-14	196
4	55	59	-4	16
5	55	70	-15	225
6	35	36	-1	1
7	30	44	-14	196
8	35	28	7	49
9	40	39	1	1
10	35	50	-15	225
11	50	64	-14	196
12	25	22	3	9
13	10	19	-9	81
14	40	53	-13	169
15	55	60	-5	25
		Sum:	-96	1478

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
Supplier	1	22	8	30
	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
Supplier	1	23.1	6.9	30
	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
```

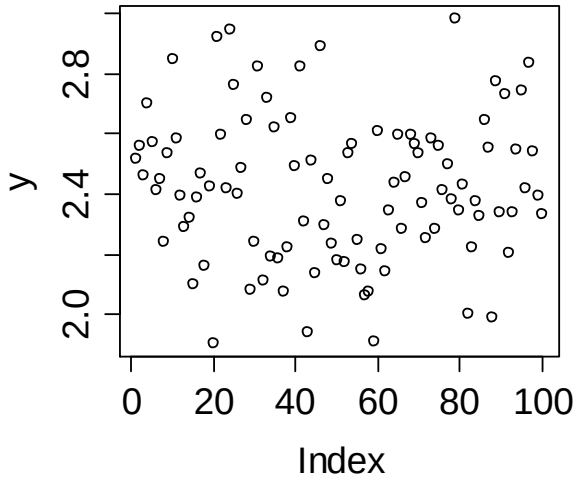
You need not to compute the CI but you should be able to interpret it, the number one (equal variances is included here).

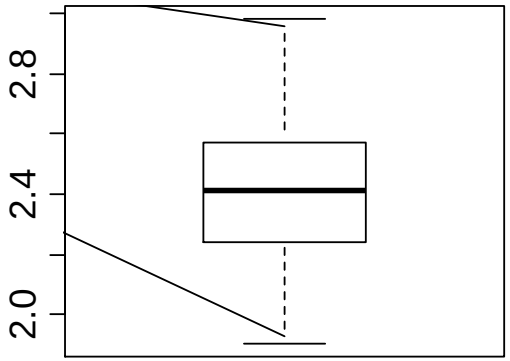
There are no significant differences between the two variances.

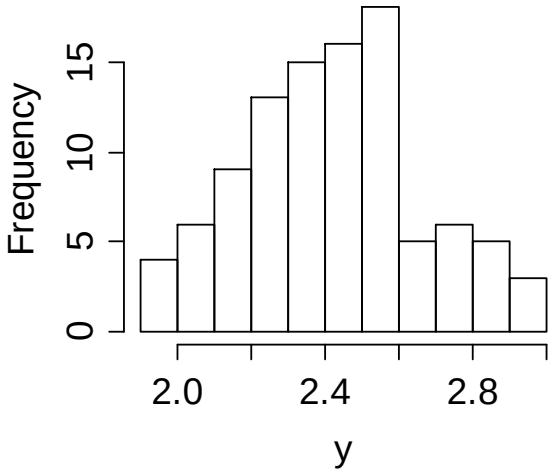
Test for Normality of Your Data – Data Summary

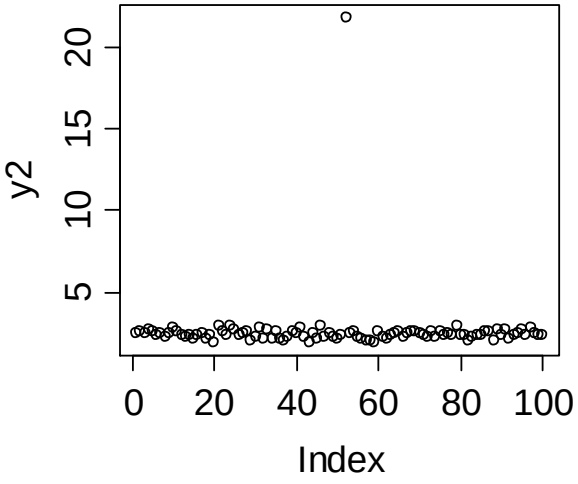
```
> setwd("c:\\Users\\Regine\\Documents\\Lacie\\Daten\\R\\")
> data<-read.table("CrawleyTheRBook\\das.txt",header=T)
> attach(data)

> 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





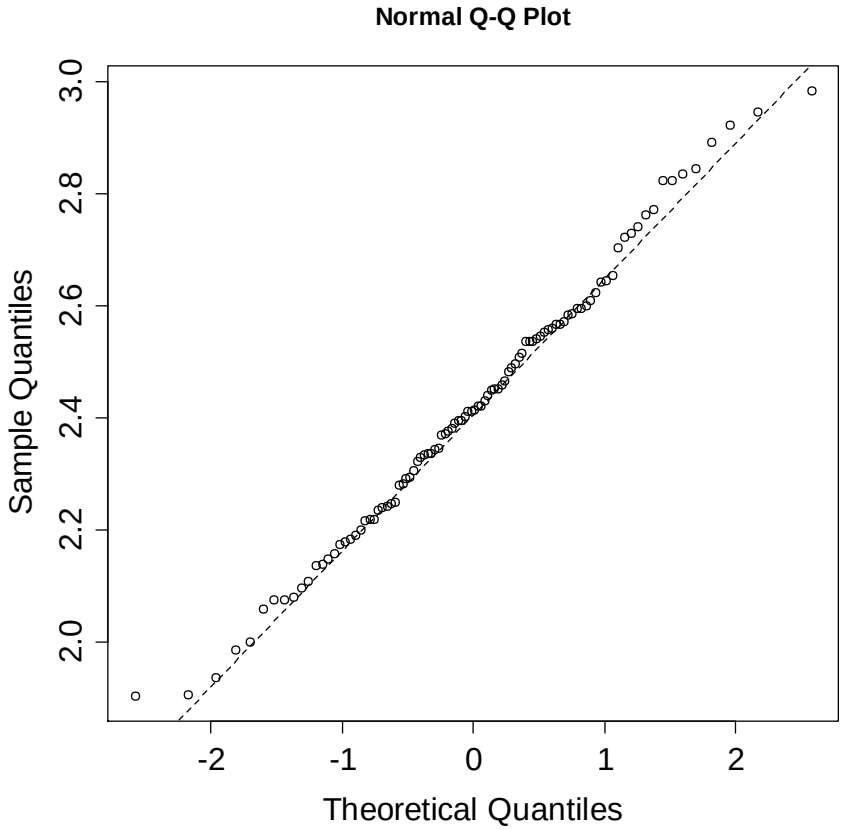




Quantile-Quantile-Plots for Testing Normality

- 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:
Plan for Today's Lecture: 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

Reserve

- #

Comparison of Two Samples for their Central Tendency: unpaired 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:
- 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.
- 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.

Summary, From Crawley

Two samples: classical test for two samples include – summary?!:

- comparing two variances (Fisher's F test, var.test)
- comparing two sample means with normal errors (Student's t test)
- comparing two means with non-normal errors (Wilcoxon's rank test, wilcox.test)
- comparing two proportions (the binomial test, prop.test)
- (correlating two variables (Pearson's or Spearman's rank correlation, cor.test)) later
- testing for independence of two variables in a contingency table (chi-squared, chisq.test, or Fisher's exact test, fisher.test).

# #

- #