

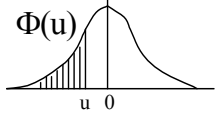
Biomedical Statistics

Version 03 / 2012 (Skript 1367)
Dr. Stefan von Weber, HS Furtwangen University
Fachbereich Maschinenbau und Verfahrenstechnik

Table of Contents Part

0. Table of critical values
1. Introduction
2. Probability calculus
3. Basics of statistics
 - 3.1 Types of errors, random numbers
 - 3.2 Distributions
 - 3.3 Estimation of distribution parameters
4. Data acquisition
 - 4.1 Data input, data check
 - 4.2 Transformation of data
5. Diagrams
6. Statistical numbers
7. Test of hypotheses
8. Test of frequency numbers
 - 8.1 Comparison of one observed relative frequency number
 - 8.2 Comparison of two observed frequency numbers
9. Contingency tables
 - 9.1 Test of contingency or test of homogeneity
 - 9.2 Configural Frequency Analysis (CFA) of Lienert and Victor
 - 9.3 χ^2 -analysis of Lancaster
 - 9.4 Variable selection - seeking the most significant table
 - 9.5 2x2-Tables: Association measures, searching types
10. χ^2 -test of fit for a distribution
11. Comparison of means
 - 11.1 one-sample t-Test
 - 11.2 Comparison of two normally distributed populations
 - 11.3 Mann-Whitney-Test (Comparison of two means, Rank test)
 - 11.4 Paired t-Test
 - 11.5 Paired Rank-test of Wilcoxon
12. Correlation und Regression
 - 12.1 Bravais-Pearson correlation
 - 12.2 Linear correlation coefficient r
 - 12.3 Simple linear regression, least square straight line fit
 - 12.4 Time series
 - 12.5 Nonlinear regression
 - 12.6 Multiple regression
13. Analysis of Variance (ANOVA)
 - 13.1 One way Analysis of Variance
 - 13.2 Two way ANOVA with comparison of means
14. Classification
 - 14.1 Linear discriminant analysis
 - 14.2 Cluster analysis
 - 14.3 Logistic regression
15. Survival analysis
16. References and Dictionary
17. Practical exercises
18. Examples of old exams
19. List of examples from lecture

0. Table of critical values of the t-, χ^2 -, F- und $\Phi(u)$ -distribution with $\alpha=0.05$ (5%)

df	t		χ^2	F							df1 / df2	 $\Phi(u)$	
	1-tail	2-tail		1	2	3	4	5	10	20		u	$\Phi(u)$
1	6,31	12,71	3,84	161	200	216	225	230	242	248	1		
2	2,92	4,30	5,99	18,5	19,0	19,2	19,2	19,3	19,4	19,4	2	-0,1	0,4602
3	2,35	3,18	7,81	10,1	9,55	9,28	9,12	9,01	8,79	8,66	3	-0,2	0,4207
4	2,13	2,78	9,49	7,71	6,94	6,59	6,39	6,26	5,96	5,80	4	-0,3	0,3821
5	2,02	2,57	11,07	6,61	5,79	5,41	5,19	5,05	4,74	4,56	5	-0,4	0,3446
6	1,94	2,45	12,59	5,99	5,14	4,76	4,53	4,39	4,06	3,87	6	-0,5	0,3085
7	1,89	2,36	14,07	5,59	4,74	4,35	4,12	3,97	3,64	3,44	7	-0,6	0,2742
8	1,86	2,31	15,51	5,32	4,46	4,07	3,84	3,69	3,35	3,15	8	-0,7	0,2420
9	1,83	2,26	16,92	5,12	4,26	3,86	3,63	3,48	3,14	2,93	9	-0,8	0,2119
10	1,81	2,23	18,31	4,96	4,10	3,71	3,48	3,33	2,98	2,77	10	-0,9	0,1841
11	1,80	2,20	19,68	4,84	3,98	3,59	3,36	3,20	2,85	2,65	11	-1,0	0,1587
12	1,78	2,18	21,03	4,75	3,89	3,49	3,26	3,11	2,75	2,54	12	-1,1	0,1357
13	1,77	2,16	22,36	4,67	3,81	3,41	3,18	3,03	2,67	2,46	13	-1,2	0,1151
14	1,76	2,14	23,68	4,60	3,74	3,34	3,11	2,96	2,60	2,39	14	-1,3	0,0968
15	1,75	2,13	25,00	4,54	3,68	3,29	3,06	2,90	2,54	2,33	15	-1,4	0,0808
16	1,75	2,12	26,30	4,49	3,63	3,24	3,01	2,85	2,49	2,28	16	-1,5	0,0668
17	1,74	2,11	27,59	4,45	3,59	3,20	2,96	2,81	2,45	2,23	17	-1,6	0,0548
18	1,73	2,10	28,87	4,41	3,55	3,16	2,93	2,77	2,41	2,19	18	-1,7	0,0446
19	1,73	2,09	30,14	4,38	3,52	3,13	2,90	2,74	2,38	2,15	19	-1,8	0,0359
20	1,72	2,09	31,41	4,35	3,49	3,10	2,87	2,71	2,35	2,12	20	-1,9	0,0287
21	1,72	2,08	32,67	4,32	3,47	3,07	2,84	2,68	2,32	2,09	21	-2,0	0,0227
22	1,72	2,07	33,92	4,30	3,44	3,05	2,82	2,66	2,30	2,07	22	-2,1	0,0179
23	1,71	2,07	35,17	4,28	3,42	3,03	2,80	2,64	2,27	2,04	23	-2,2	0,0139
24	1,71	2,06	36,42	4,26	3,40	3,01	2,78	2,62	2,25	2,02	24	-2,3	0,01072
25	1,71	2,06	37,65	4,24	3,39	2,99	2,76	2,60	2,24	2,00	25	-2,4	0,00820
26	1,71	2,06	38,89	4,23	3,37	2,98	2,74	2,59	2,22	1,99	26	-2,5	0,00621
27	1,70	2,06	40,11	4,21	3,35	2,96	2,73	2,57	2,20	1,97	27	-2,6	0,00466
28	1,70	2,05	41,34	4,20	3,34	2,95	2,71	2,56	2,19	1,96	28	-2,7	0,00347
29	1,70	2,05	42,56	4,18	3,33	2,93	2,70	2,55	2,18	1,94	29	-2,8	0,00255
30	1,70	2,04	43,77	4,17	3,32	2,92	2,69	2,53	2,16	1,93	30	-2,9	0,00187
34	1,69	2,03	48,60	4,13	3,28	2,88	2,65	2,49	2,12	1,89	34	-3,0	0,001350
40	1,68	2,02	55,76	4,08	3,23	2,84	2,61	2,45	2,08	1,84	40	-3,1	0,000967
44	1,68	2,02	60,48	4,06	3,21	2,82	2,58	2,43	2,05	1,81	44	-3,2	0,000688
50	1,68	2,01	67,50	4,03	3,18	2,79	2,56	2,40	2,03	1,78	50	-3,3	0,000484
60	1,67	2,00	79,08	4,00	3,15	2,76	2,53	2,37	1,99	1,75	60	-3,4	0,000337
70	1,67	1,99	90,53	3,98	3,13	2,74	2,50	2,35	1,97	1,72	70	-3,5	0,000233
80	1,66	1,99	101,88	3,96	3,11	2,72	2,49	2,33	1,95	1,70	80	-3,6	0,000159
90	1,66	1,99	113,15	3,95	3,10	2,71	2,47	2,32	1,94	1,69	90	-3,7	0,0001080
100	1,66	1,98	124,34	3,94	3,09	2,70	2,46	2,31	1,93	1,68	100	-3,8	0,0000723
150	1,66	1,98	179,58	3,90	3,06	2,66	2,43	2,27	1,89	1,64	150	-3,9	0,0000480
200	1,65	1,97	233,99	3,89	3,04	2,65	2,42	2,26	1,88	1,62	200	-4,0	0,0000317
∞	1,65	1,96	∞	3,84	3,00	2,60	2,37	2,21	1,83	1,57	∞		

N=22 values of systolic blood pressure to the example 3 in the lecture:

122 127 131 104 138 139 148 115 137 144
113 132 134 142 119 127 133 111 141 134
129 131

1. Introduction

Statistics is descriptive	or inferential (hypotheses proving)
Means, standard deviations, regression-coefficients, correlation-coefficients, probability estimations	Hypothesis → sample → test → statement together with its error probability concerning the population
Example: Inquiry in Munich: "Would you like to test a new diet?" 23 of 100 probands answered with "Yes" → probability $p = 23/100 = 0.23$ in the sample, i.e. 23% of the 100 inquired. This number is only an estimation of the p-value of all people in Munich.	Hypothesis (example): "Less than 20% of all inhabitants of Munich want a new diet → Inquiry see left box → asymptotical Binomial test 0.23 with 0.2 and $n=100$ → $u=0.75$ → hypothesis rejected, i.e.. no significant deviation found from value of 20%

Statistics summarizes, visualizes und analyses data. Aim of **descriptive** statistics is information and forecast/predict future data, aim of **inferential** statistics is the proof of hypotheses using samples. From the data of a **sample** one draws conclusions concerning the whole **Population**. Example: One draws conclusions from a study with 15 patients. These conclusions shall be valid for all patients suffering under the same disease. Important is here the **error probability**. A sample is, e.g., a set of 10 randomly chosen trees in a forest. The **population** is the forest. Persons we call also **patient, proband, case**, objects we call also **case, point, measuring point**.

Types of studies

- Therapy studies (cross-/longitudinal studies, retrospective / prospective studies, retrospective with archive (Historical **Case Control** study, prospective cohort study)
- Observational studies
- Physiological and neurophysiological investigations
- Experiments with animals, micro-organisms or techniques

Clinical Therapy studies: Proof of the efficacy of a therapy. Important the placebo group, since also placebos have a (psychological) efficacy. The color of the placebo is also important (efficacy increases with the colors yellow-green-blue-red). By Gasser&Seifert is important:

- Randomisation of patients (random grouping)
- Stratification by disturbance variables (age, degree of disease, ...) and recording them (as covariables)
- simple blind (only patient) or double blind (doctor and patient dont know whether placebo or not)
- standardized regime (replicable): e.g., when how many pills in what pots ...
- including and excluding criterions (overweight, underweight, age-limit, ...)
- informed consent document of the patient

Observational studies: no therapy or other influences, estimation of probabilities.

Example 1: Estimation of the prevalence = percentage of cases of a disease

Example 2: Search for factors of disease, e.g., for Myocardial infarct (heart attack)

Experimental design: Experimental design means:

- to choose representative samples
- to get a significant result with low cost
- to exclude disturbance factors, or to measure them as covariates

Preliminary experiments: Often one needs preliminary experiments to estimate the variances of the data. Example: Can bacterium tribe 2 take the hurdle of $d=3\%$ improvement of production of insuline, made with bacterium tribe 1 at present ? From preliminary experiments we know that measurements of the insuline production scatter with standard deviation $s=4.3$.

Statistical issue: Estimation of the **number n of measurements** needed from each group (bacterium tribe) to confirm the wished effect $d = \mu_2 - \mu_1 \geq 3.0$ in the averages with $\alpha=5\%$ error probability.

Solution (here, for example): One uses the t-test statistic from paragraph 11.2 *comparison of two normally distributed populations*, $t=(d/s)*((n*n)/(n+n))^{0.5}$ or $t=(d/s)*(n/2)^{0.5}$ with $df=2n-2$ degrees of freedom. By varying the number n one seeks the smallest n delivering $t>t_a$. Group size $n=19$ is giving $t=2.15 > t_\alpha=2.11$, i.e. significance. Recommended group size is $n=19$ measurements for each tribe.

2. Probability calculus

For what? Quality control, calculation of chances, simulation of stochastic models with a PC, base for some test distributions

The possible outcomes of a **random experiment** are called **elementary events** (e.g. number 4 on dice). The set is called **sample space R** (1-6 on dice). The *certain event* (a number $1 \leq x \leq 6$) occurs always, the *impossible event* (e.g. number 0 or 7) occurs never. The **Probability P** of an event is a number $0 \leq P \leq 1$, or $0 \leq P\% \leq 100\%$, respectively

The **expectation value E** of the **frequency** of the occurrence of an event is $E = N \cdot P$
 N =total number of drawings, P =Probability of the occurrence of the event

Example: Manufacturing errors of pills $R=\{1,2\}$, $N=1.000.000$

Elementary event	N_i	$P_i = N_i / N$	$P_i \% = P_i \cdot 100$
A_1 (underweight)	632	0,000632	0,0632
A_2 (overweight)	869	0,000869	0,0869

Multiplication theorem: The probability $P(A \wedge B)$ of the joint occurrence of stochastically independent events **A and B**: $P(A \wedge B) = P(A) \cdot P(B)$. The probability to get 2 times number 6 throwing twice the dice is $P(6 \wedge 6) = (1/6) \cdot (1/6) = (1/36)$. **Stochastic independency** means that the occurrence of A_1 is not dependent on the occurrence of A_2 , A_2 not on A_1 , and no hidden dependency exists.

Addition theorem: The probability $P(A \vee B)$ of the occurrence either of event **A or exclusively** of event **B**. **A and B are disjunct**, i.e., they exclude one another: $P(A \vee B) = P(A) + P(B)$. Related to the pill manufacturing errors is $P(A_1 \vee A_2) = 0,000632 + 0,000869 = 0,001501$, i.e., we find with this probability an underweighted or an overweighted pill.. A pill is never underweighted and overweighted at the same time.

Conditional Probability $P[A|B]$ is the probability that an event **A** occurs, provided that event **B** has occurred before. Examples: The probability that a fermenter charge fails provided that the fermenter charge before has failed too. The probability that a male proband will collapse on the top of the Zugspitze mountain provided that he is 80 years old or elder.

$$P[A|B] = \frac{P[A \cap B]}{P[B]}$$

Probability of a metric measuring value: A temperature of exactly 20°C has as mathematical point probability 0. We had to define an interval, e.g. $P[19.5 < x < 20.5]$ to get a finite probability.

Permutations: The number of possible orders of N distinguishable objects is $N_p = N!$ (speak N-factorial) with $0! = 1$, $1! = 1$, $2! = 1 \cdot 2 = 2$, $3! = 1 \cdot 2 \cdot 3 = 6$, $N! = 1 \cdot 2 \cdot 3 \cdots N$.

Example 1: We can reorder 3 objects A, B, C into 6 orders ABC, ACB, BAC, BCA, CAB, CBA. If the set of the N objects does consist of of k groups (with group sizes N_1, N_2, \dots, N_k), and inside a group the objects are not distinguishable, then the number of distinguishable permutations is $N_p = \frac{N!}{N_1! \cdot N_2! \cdots N_k!}$. For example, a female and a male pair of twins are giving

the 6 distinguishable permutations WWmm, WmWm, ..., mmWW.

3. Basics of statistics

3.1 Types of errors, random numbers :

Systematical errors: We produce systematical errors e.g. by false experimental design (e.g. very different sizes of groups, great differences in age, ...), bad calibrated measuring instruments, not operationalized approach by different doctors. One can avoid systematical errors or can partially correct them, if one does follow the → recommendations of the *experimental design*.

Random errors / random numbers: Measuring values or observed quantities we understand in statistics as random numbers. The averaged noon-temperature at summertime is 21,3 °C. Deviations from the average we don't explainate by weather research, but produced by chance. A woman has 1.4 children at average. The real number of kids of a woman we do'nt explainate by the circumstances of her, but produced by chance. A **random variable** is a function giving a real number to the outcome of a random experient. A random value x of the random variable X is called **realization** or **outcome**.

Discrete random numbers can realize only some few, mostly integer values (Realization, outcome, symptom, category) (e.g., number of kids 1, 2, ..., or volume of wine bottles 0.5, 0.75, 1.0, 1.5, 2.0, 3.0, 5.0 and so on). **Continuous random numbers** can realize inside of their definition range all values (Realizations), e.g., $T = 21.3 \text{ °C}$ or $T = 21.297 \text{ °C}$.

Nominal (qualitative) data are always discrete. We use them only for sorting or dividing data in groups. For example, the postage code in addresses of patients is of the nominal type. To calculate sums or averages of nominal variables is nonsense. **Categorical Data** we treat as nominal if they are not **ordinal**, i.e. no rank order is hidden in the categories.(e.g. "white", "red", "black" mycelium colonies, also if they are coded as 1,2,3 in the data.).

Metric (quantitative, continuous) data we can arrange on a straight line. Between two values always a relation exists " $<$ ", " $=$ ", " $>$ ". With metric data it is allowed to calculate (sums, averages, ...) **Rank ordered categorial data** we treat often as metric data, e.g. marks (grades) 1,2,...,5 or categories of cars 1="small", 2="middle", 3="great". **Binary data** (with two outcomes only) we can also treat very often as metric data, e.g. female=1, male=2 or healthy=0, ill=1).

3.2 Distributions

The **distribution** says, how many data we expect for a given deviation from average. The representation of the distribution of **discrete data** we do with the **bar diagram** or **pie diagram**. Each bar is corresponding to one outcome of the discrete variable. The representation of the distribution of **observed continuous data** we do with the bar diagram too (**histogram** of absolute or relative frequencies). The representation of continuous distribution functions we do with the **line diagram**. In the case of a histogram we find the number of classes, or the class breadth, respectively, by trial and error. It depends on the total number of N of objects. Great $N \Rightarrow$ many classes, small $N \Rightarrow$ few classes. There is no exact rule. The **cumulative histogram** of observed data is a staircase function with ascending values from 0 to 1 (see *sum distribution*).

Discrete distribution (pill errors A_1, A_2, A_3)	Histogram absolute frequencies (trunk diameter of trees)	Histogram relative frequencies (trunk diameter of trees)	Line diagram of a normal density distri- bution

- For a discrete distribution is $\sum P_i = 1$ or $\sum P_i \% = 100\%$, respectively
- For a histogram of absolute frequencies is $N = \sum N_i$ (N =total number of objects)
- For a histogram of relative frequencies is $\sum P_i = 1$ or $\sum P_i \% = 100\%$, respectively
- For a density distribution the total area under the curve $f(x)$ equals the value 1 always.

Theoretical distributions follow from a **process model**. The **density function** $f(x)$ yields the probability P_{ab} with its area above the interval $[a,b]$. P_{ab} is the probability that a x -value occurs from the interval $[a,b]$. X is a continuous random variable.

	Probability	Normalization to area 1
	$P_{ab} = \int_a^b f(x) dx$	$\int_{-\infty}^{+\infty} f(x) dx = 1$

The **distribution function** (sum curve, cumulative distribution) $F(x)=P$ yields the probability P that a x -value occurs from the interval $[-\infty, x]$.

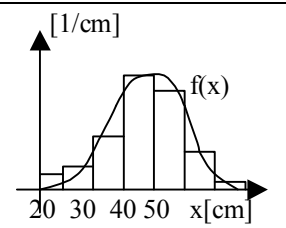
Density distribution	Distribution function	Formula distribution function
		$F(x) = \int_{-\infty}^x f(u) du$

When I use the density function, or when I use the distribution function, respectively? There is no rule. The information is in both curves. We can describe a distribution by its **moments** μ_i without any exact picture of the curve, similarly to the Taylor series of a function. The

moments $\mu_1-\mu_4$ we call mean (average, center, expected value), variance, skewness and kurtosis. The higher moments (μ_2, μ_3, \dots) one calculates relatively to the mean, i.e., $(x-E)^i$, and call them central moments.

1. Moment: Expected value (Mean, arithmetic mean, center of mass)	2. Moment: Variance. Under Normal distribution is $\mu_2=\sigma^2/2$ with σ = Standard deviation	3. Moment: Skewness $\mu_3>0$: Summit left from E $\mu_3<0$: Summit right from E
$\mu_1 = E(x) = \int_{-\infty}^{+\infty} x \cdot f(x)dx$	$\mu_2 = \int_{-\infty}^{+\infty} (x - E)^2 f(x)dx$	$\mu_3 = \int_{-\infty}^{+\infty} (x - E)^3 f(x)dx$

We *impute* **Theoretical data distributions** to observed or measured data sets. One says *the data are normally distributed*, e.g., or *the data are binomial distributed*, respectively. There is no proof that the data are distributed so indeed. But one can evaluate the deviations between observed and theoretical distribution imputed. A concluding statement could be: "There is no significant deviation from Normal distribution".



Poisson distribution, *Binomial* distribution, *multinomial* distribution, and *hypergeometrical* distribution are the most **important theoretical distributions for discrete random data**. All these four distributions we use also as **test distributions** for the proof of hypotheses, but more seldom than u-, t-, χ^2 - and F-distribution.

The **Poisson** distribution is a discrete distribution with parameter $\lambda=Np >0$. N is the number of drawings with $N \rightarrow \infty$ and probability $p \rightarrow 0$. It yields the probability that a rare event (death by the impact of a meteorite, e.g. with N meteorites per anum, or the decay of some radioactive material with N atoms in the time interval of t seconds) is occurring k times exactly. The distribution has expectation $\lambda=Np$ and variance $\sigma^2=\lambda=Np$

$$P_{N,k} = \frac{\lambda^k}{k!} e^{-\lambda}$$

The model of the **Binomial** distribution is an urn with a part p of black balls and a part $q=1-p$ of white balls. $P_{n,k}$ is the probability to draw exactly k black balls (with replacement) with n drawings. p is called parameter of the Binomial distribution. Expectation is $E= n p$, variance is $\sigma^2 = pq n$.

$$P_{n,k} = \binom{n}{k} p^k q^{n-k} \quad \text{mit} \quad \binom{n}{0} = 1, \quad \binom{n}{k} = \frac{n(n-1)\dots(n-k+1)}{1 \cdot 2 \cdot \dots \cdot k} \quad (\text{spek "n over k"})$$

Example: An industrial process is running out of control with probability $p=0,068$ (statistical average). What is the probability that 3 charges from 10 are running out of control?

$$P_{10,3} = \binom{10}{3} 0.068^3 \cdot 0.932^7 = \frac{10 \cdot 9 \cdot 8}{1 \cdot 2 \cdot 3} \cdot 0.00031 \cdot 0.611 = 0.023 \quad \text{or} \quad 2,3\%$$

If the plant is running 10 charges per week, then one expects all 40 weeks one week with 3 bad charges. The sum P of the 11 probabilities $P = P_{10,0} + P_{10,1} + \dots + P_{10,10}$ is $P=1$ exactly.

<p>The Hypergeometrical distribution follows from the same urn process, as the binomial distribution, but we draw without replacement. So, each drawing changes p and q (if the number of balls in the urn is small enough). We calculate the probability to draw exactly k black balls with n drawings from initially N black and white balls in the urn (with M=pN). For great N we find the Binomial distribution.</p>	$\varphi = \frac{\binom{n}{k} \binom{N-M}{n-k}}{\binom{N}{n}}$
--	--

Normal distribution (u- or Gauss-distribution), Log-normal distribution, t- or Student-distribution, χ^2 - or Chi-square-distribution, and the F- or Fisher-distribution are the most important **continuous distributions**. The Normal- or u-distribution and the Log-normal distribution are often found as data distributions. The t-, χ^2 - and F-distributions are rarely data distributions, but are used very often as **test distributions** to test hypotheses. The Normal distribution (u-distribution) is both - data distribution and test distribution.

<p>Density curve of the Normal distribution: μ (expectation) and σ^2 (variance) are called the parameters of the normal distribution. One estimates them by a sample, using the mean of the sample as an estimation of μ, and using the square σ_{n-1}^2 of the standard deviation as an estimation of σ^2. Normal distribution arises if many random influences are added.</p>	$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
--	---

Under Normal distribution imputed to the data, we find for a sample of a population the following relations between sample, estimations and parameters μ and σ^2 of the population:

Sample statistics	→ estimation	→	parameters of population
Mean $\bar{x} = \sum x_i / n$	→ $\hat{\mu}$	→	μ
Variance $\sigma_{n-1}^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$	→ $\hat{\sigma}^2$	→	σ^2

Example: n=10 diameters of trunks: 36 41 39 52 48 53 55 61 54 49 cm. The sample mean 48.8 cm is an estimator of the unknown mean of μ of the population. The standard deviation $\sigma_{n-1}=7,91$ cm is an estimator of the standard deviation σ of the population. The **true parameters** μ and σ^2 of the population we can get only for $n \rightarrow \infty$. Estimated values (marked by the hat ^) have always errors.

Normal distribution with **mean** μ und **variance** σ^2 is abbreviated by **N(μ ; σ^2)**. N(0;1) is the **Standard-Normal distribution** with mean 0 und variance 1. The distribution function (sum curve) **$\Phi(u)$** of the Normal distribution is called also **Gaussian error integral**, and is tebled in all statistic books. **$\Phi(u)$** and the inverse function **$u(\Phi)$** are important test distributions. The Normal distribution is important because of the **Central Limit Theorem**: *The distribution of the sum S of any randomly distributed numbers z_i approximates the Normal distribution with growing number of summands.* Indeed, the quantity $S=z_1+z_2+\dots+z_n$ is already well normally distributed for a value of $n=5$, e.g., each sample mean with sample size $n \geq 5$.

Density curve of the Lognormal distribution: M (expectancy) and S^2 (variance) are the parameters. One calculates the logarithms of the data, from this mean and variance, and takes them as M and S . Lognormal random numbers arise if random influences are multiplied. The distribution is unsymmetrical.	$f(x) = \frac{1}{S \cdot x \sqrt{2\pi}} e^{-\frac{(\ln x - M)^2}{2S^2}}$
t-distribution (also Student-distribution, <i>Student</i> was the pen name of W. P. Gosset) is the distribution of the quotient $t = u / \chi^2$. Here u is $N(0;1)$ -distributed and χ^2 is χ^2 -distributed with df degrees of freedom. The distribution is symmetrical.	$t = \frac{u}{\chi} \sqrt{df}$
χ^2-distribution (Chi-Square distribution by F.R. Helmert and K. Pearson) is the distribution of the sum $\chi^2 = u_1^2 + \dots + u_k^2$. The u_i are $N(0;1)$ -distributed and stochastically independent. Number of degrees of freedom is $df=k$. The distribution is unsymmetrical.	$\chi^2 = u_1^2 + \dots + u_k^2$ with $df=k$ degrees of freedom
F-distribution (by R. A. Fisher) is the distribution of the quotient $F = \chi^2_1 / \chi^2_2$. Here χ^2_1 is χ^2 -distributed with df_1 degrees of freedom, χ^2_2 is χ^2 -distributed with df_2 degrees of freedom.	$F = \chi^2_1 / \chi^2_2$ with df_1 and df_2 degrees of freedom

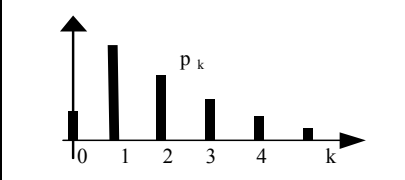
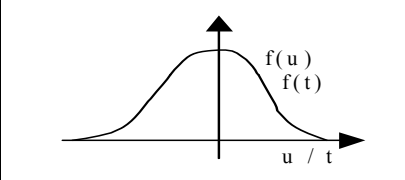
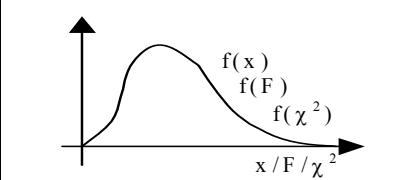
The F-distribution is of interest also, because it contains the t- and the χ^2 -distribution.

We find $t^2(df) = F$ with $df_1=1$ and $df_2=df$.

We find $\chi^2(df) = df_2 F$ with $df_1 \rightarrow \infty$ and $df_2=df$.

The **degree of freedom df** is the number of the „free data points“, which are used for the computation of a standard deviation. Example: Deviation of the points from a best fit straight line. In the case of 2 points the line goes exactly through both points. No point is free ($df=0$). In the case of three points is one point free ($df=1$). Generally is in the case of a straight line $df = n - 1$.

The three figures below are showing the typical view of the following distributions:

Poisson-, Binomial-, Hypergeometrical distribution	Normal distribution, t-distribution	lognormal, χ^2 -, F-distribution
		

3.3 Estimation of parameters of distributions

An Estimation (or estimator) is a number calculated by a determined formula fitting well some parameter of the population, e.g. the mean. We find good, very good and the **best estimator**. A general method to find the best estimator is the **Maximum Likelihood** method. The observed data have *maximum likelihood* if we use the best estimators as **parameters of the distribution imputed for the data**. Imputing Normal distribution we find that *Maximum Likelihood* and the *Method of Least Squares* equals asymptotically, i.e. they are identical for $n \rightarrow \infty$).

Point estimators and confidence intervals: We call the calculation of a (single) value from a sample a *point estimation*, e.g. the sample average as an estimation of the unknown mean of the population. In the descriptive statistics point estimations are oftenly used. In inferential statistics (hypotheses proving) point estimators are the base for the construction of *confidence*

intervals. Confidence intervals: Repeating a study very oftenly to estimate a parameter θ , we would get similar, but slightly other values for the parameter θ . That is the effect of chance - other probands, other season of the year, ... What is the true value of the parameter θ in the population? Here the confidence interval is giving a typically statistical answer:

An $(1-\alpha)$ -confidence interval $[\theta_U, \theta_O]$ for parameter θ is a random interval containing with probability $(1-\alpha)$ the wished value θ .

Some examples of confidence intervals are:

Confidence interval for μ with known σ_0^2 , here sample mean \bar{x} estimated from n values, Normal distribution assumed.	$\bar{x} \pm \frac{\sigma_0}{\sqrt{n}} u(1-\alpha/2)$
Confidence interval for μ . Both values, \bar{x} and σ_{n-1}^2 , are estimated from a sample with size n , Normal distribution assumed.	$\bar{x} \pm \frac{\sigma_{n-1}}{\sqrt{n}} t(\alpha, df = n-1, \text{zweis.})$
Exact confidence interval $[p_U, p_O]$ for relative frequency $\hat{p} = k/n$. F_1 is the F-distribution with the degrees of freedom $df_{11}=2(k+1)$, $df_{12}=2(n-k)$, F_2 is the F-distribution with the degrees of freedom $df_{21}=2(n-k+1)$, $df_{22}=2k$.	$p_U = \frac{k}{k + (n-k+1) \cdot F_1(\alpha/2, df_{11}, df_{12})}$ $p_O = \frac{(k+1) \cdot F_2(\alpha/2, df_{21}, df_{22})}{n-k + (k+1) \cdot F_2(\alpha/2, df_{21}, df_{22})}$
Approximative ($n \rightarrow \infty$) confidence interval $[p_U, p_O]$ for relative frequency $\hat{p} = k/n$.	$\hat{p} \pm u(1-\alpha/2) \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

4. Data acquisition

4.1 Data input, data check

Data have mostly the shape of a table. The best way for input is the use of a simple text editor giving an ASCII-File. Please avoid input of characters, but code them as numbers, e.g. female $f=1$, male $m=2$ or blood group $0=0$, $A=1$, $B=2$, $AB=3$. For example, EXCEL does not accept missing values. Other programs (e.g. DASY) accept missing values. In the worstest case one had to delete all lines of the data table containing a missing value before using EXCEL. In no case use a zero!! DASY interprets all sequences of tokens (e.g. — or A or AA) as missing values if they are not interpretable as numbers.

Data check is an important step before we are working with the data:

Decimal error during input, e.g. the weights of probands 84.3 77.1 59.0 ... **820.** ...

Permuted digits, e.g. the age of pupils 12 17 14 ... **71** ...

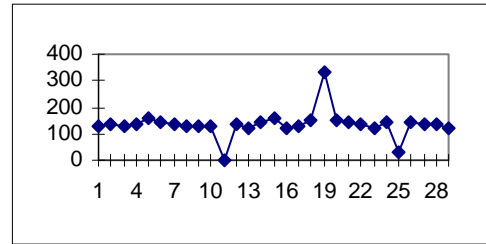
Zero instead of a missing value, e.g. the weights of probands 84.3 77.1 59.0 ... **0** ...

False group, e.g. 2=disease instead of 1=healthy 1 1 1 1 ... **2** ...

A **first data check** we do with the simple statistical numbers as mean, standard deviation, maximum, minimum, sample size, number of correct values. EXCEL and each statistical package is delivering these numbers.

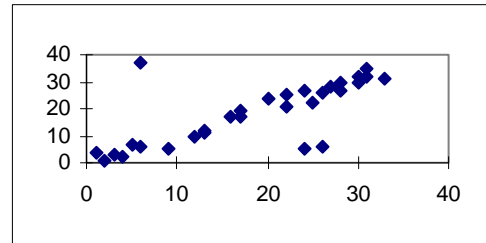
Visual outlier check with the X-N-Plot (Excel)

The data are in a row or in a column → mark the data → *Insert* → *Diagram* → On this sheet → draw a rectangle for the plot → *continue* → *Lines* → *continue* → 1 → *continue* → *continue* → *end*



Visual outlier check in the X-Y-Plot (Excel)

The data are in two rows or in two columns → Mark the data → *Insert* → *Diagram* → On this sheet → draw a rectangle for the plot → *continue* → *Points* → *continue* → 1 → *continue* → *continue* → *end*



Outlier check with the 3-sigma rule: If the value of $u = (|X - \bar{X}|) / \sigma_{n-1}$ is extending the value of 3 for some data value X then we assume X to be an outlier. \bar{X} is the mean and σ_{n-1} is the standard deviation of the data. After deletion of the outlier X from the data, we had to perform another check, because σ_{n-1} and \bar{X} are changed by the deletion of X.

4.2 Transformation of data

Mostly data transformation has the aim to produce Normal distribution in the data. Why? The best statistical tests require Normal distribution. But many distributions have positive skew (they are **right-skewed**), i.e. both values, the modal value (highest point of the distribution curve) and the median (divides the area under the density curve 50% to 50%), respectively, are positioned at the left side of the arithmetic mean.

The aims of a data transformation are:

- We get nearly a Normal distribution
- Stabilization of the variance: E.g. in the case of the χ^2 -distribution is $\sigma^2=2\mu$, i.e. the variance changes with the mean. Instead, in the case of Normal distribution μ und σ^2 are independent. Performing a group mean test, e.g., such variance effects are disturbing.

Between others, the following transformations are used: $\sqrt{x}, \sqrt[3]{x}, \ln(x), \log(x), -\frac{1}{\sqrt{x}}, -\frac{1}{x}$

The first is the weakest, the last, $-1/x$, is the strongest. $\ln(x)$ and $\log(x)$ differ only by a constant factor. Recommendations:

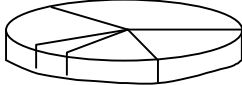
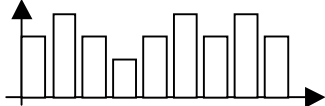

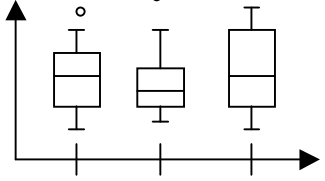
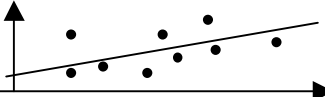
Body weights	Transformation $-1/\sqrt{x}$ (Gasser&Seifert)
Percents, relative frequencies, e.g. $x=h/n$	Transformation $\arcsin(\sqrt{x/n})$
Absolute frequencies, counting values	Transformationen $\sqrt{x}, \sqrt{x+3/8}$
Retention time	Transformation $-1/x$

Indexing a data row X_1, X_2, \dots, X_n to a common starting point of 100%: Indexed data are handy because they allow an observer to quickly determine rates of growth by looking at a chart's vertical axis. They also allow for comparison of variables with different magnitudes. Each curve is starting at 100% and changes only relatively to this starting point. Formula is

$$X'_i = (X_i \cdot 100) / X_1$$

One divides each number X_i by the first number X_1 of the row and multiplies by 100.

5. Diagrams

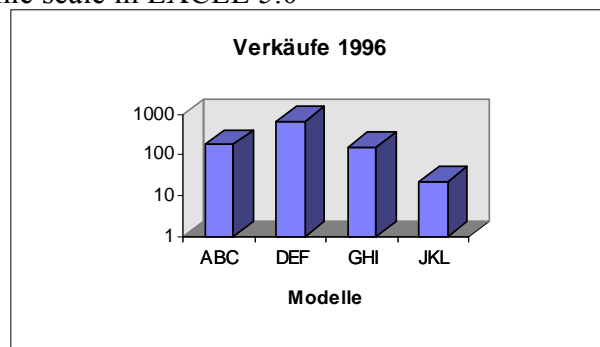
<p>Pie charts on dividing up a cake (100%) into different groups, e.g. market shares of the European vitamine production.</p>	
<p>Bar diagrams for the presentation of summarized data (Histograms, comparison of quarter sums, comparison of group means, comparison of the cash flow in the years 2005, 2006, ...)</p>	
<p>Line diagrams for the presentation of point data (course of the averaged daily temperatures, weight curves, fever curves, ...)</p>	
<p>Boxplots show the data distribution. The box itself is giving the interval from the 25%- to 75%-percentile. with the median (50% percentile) inside the box. The "whiskers" at the ends show the 10% and 90%-percentiles. Some boxplots show extreme values as 'o's or '*'. Example: Comparison of 3 groups by their boxplots</p>	
<p>Scatterplots (x-y-Diagrams) show the measuring values as points in a coordinate system, oftenly coupled with a line diagram.</p>	

Coordinate axes have a **ruler**. The ruler has

- a **ruler range** with **start** and **end**, the start must not be zero always
- a **scale division**, which can be fine or big
- a **scale**, which can be **linear** or **logarithmic**

Example 3-D-Column diagram with logarithmic scale in EXCEL 5.0

	A	B
Row 1	ABC	177
Row 2	DEF	672
Row 3	GHI	154
Row 4	JKL	22



Mark the cells A1 to B4 → *Insert* → *Diagram* → *On the same sheet* → draw a box → *continue* → *3-D-Columns* → *continue* → *1* → *continue* → *continue* →

Legende no → Titel, x-axis text
double click the diagram → double click
double click *logarithmic*

→ *End* → click some cell →
vertical axis → *Scale* →

6. Statistical numbers

Estimated class probability $\hat{p}_i = h_i / n$

h_i = Number of values found in class i , n = sample size

Arithmetic mean

x_i = The i-th value of the sample
 n = Sample size

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Weighted arithmetic mean

g_i = Gewicht zum Wert x_i
The weights must be positive (>0).

$$\bar{x} = \left(\sum_{i=1}^n g_i x_i \right) / \left(\sum_{i=1}^n g_i \right)$$

Example: The values given are the class averages and frequencies (number of trunks in the class) from 7 diameter classes of spruces. Class 1, e.g., contains the data of all trunks with diameters from 25 to 30 cm.

class average x_i :	27.5	32.5	37.5	42.5	47.5	52.5	57.5
class size g_i :	41	84	207	213	156	47	9

$G = \sum g_i = 757$, $\sum g_i x_i = 31067.5$, weighted mean = $31067.5 / 757 = 41.04$ cm

geometric mean is the n-th root of the product of the n measured values

$$\bar{x}_G = e^{(\sum \ln(x_i)) / n}$$

$$\bar{x}_G = \sqrt[n]{\prod x_i} = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

alternatively, in case of large n, with $\ln(x)$ as Natural Logarithm and e^x as Exponential function

Example: A stock fund changed in the last four years by yearly values of +3,6%, -7,2%, +1.6%, +13.4%. Because of the negative sign of the value -7.2% we had to use absolute percent values: 103,6%, 92,8%, 101,6%, 113,4%. The geometric mean of the absolute percents is $\sqrt[4]{103.6 \cdot 92.8 \cdot 101.6 \cdot 113.4} = 102.59$. We go back to the relative growth rate, and get a yearly growth of the value of 2,59% averaged over 4 years.

Median: Firstly one has to sort the n values of the sample. In the case of an odd n the median is the value in the middle, in case of an even n the median is the arithmetic mean of the two values in the middle.

Example: Sorting the 10 diameters 54 46 61 47 43 59 38 44 49 41, one gets the sequence 38 41 43 44 **46 47** 49 54 59 61. The mean of the two values in the middle, 46.5, is the median.

Mode, Modal value: The measuring value most frequently found in a sample (with great sample size) and unimodal distribution (only one maximum) is the modal value. It is seldomly used in statistics.

When I take what mean?

- The Median, if either the *typical value* is giving the best statement, or one is seeking a mean robust against outliers. One millionaire and 100 poor farmers in a village have a total income of 1.000.000 + 100 x 1000 Euro. Arithmetic mean 11.000, but the median=1000 Euro income is typical for the village.
- The arithmetic mean, if some kind of balance sheet is sought. A small river with 1000 Gramm pollution per m^3 and 100 other waters with a pollution of 1 Gramm per m^3 pollute the Lake Constance in average with 11 Gramm per m^3 . The arithmetic mean is sensible against outliers.
- The mode to characterize additionally a distribution.
- The weighted arithmetic mean to average condensed data (e.g., one calculates the total mean from class averages, because the original data are missed)

Standard deviation of the sample, i.e., more exactly only for the n values of the sample

$$\sigma_n = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} \quad \sigma_n = \sqrt{\frac{(\sum x_i^2) - n \cdot \bar{x}^2}{n}}$$

Standard deviation of the population estimated from a sample with size n

$$\sigma_{n-1} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \quad \sigma_{n-1} = \sqrt{\frac{(\sum x_i^2) - n \cdot \bar{x}^2}{n-1}}$$

Variance: The square of the standard deviation is called variance, e.g., σ^2 , $\hat{\sigma}_{n-1}^2$

If not further specified or one is speaking only about σ^2 , then always σ_{n-1}^2 is meant.

Covariance of the population estimated from a sample with size n

$$\text{cov}(x, y) = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1} \quad \text{cov}(x, y) = \frac{(\sum x_i y_i) - n \bar{x} \bar{y}}{n-1}$$

Standard error of the mean $\sigma_{\bar{x}}$: Drawing a lot of samples of size n from a population, one finds that the calculated means scatter around the unknown mean μ of the population. The standard error of the mean estimates the error of the estimation in relation to the expectation value μ (unknown mean) of a population using a sample of size n. A mean calculated from n measuring values has so the error or standard deviation $\sigma_{\bar{x}}$, i.e., we write $\bar{x} \pm \sigma_{\bar{x}}$.

$$\sigma_{\bar{x}} = \frac{\sigma_{n-1}}{\sqrt{n}}$$

When I take what error?

- σ_{n-1} (s, standard deviation, SD) in all cases, where one wishes to describe the variability of the data measured, e.g., the height of 12-year-old pupils in Germany 143 ± 6 cm. The height scatters around the mean of 143 cm by 6 cm.
- The interquartile distance (75%-25%-Quartiles) instead of σ_{n-1} in the case of highly skewed data.
- $\sigma_{\bar{x}}$ (SE, Standard Error of Mean) if one wants to give the accuracy of an estimation. For example, one calculated from a representative sample of 1600 German 12-year-old boys the mean height to 143.6 ± 0.15 cm. The accuracy of the estimation of the unknown population mean is 0.15 cm.
- σ_n in the extremely seldom cases, in which one wants to describe the variability of the sample itself, e.g., *our testing group of 12-year-old boys has mean and standard deviation of 147.8 ± 3.6 cm*. Here the standard deviation is describing the variability **only** of the boys of the testing group.

95%-confidence interval, in which we find the true mean μ of a population with a probability of 95% using an estimation from a sample with size n. $t(df, tws, \alpha)$ is the two-sided critical value of the t-distribution with $\alpha=0.05$.

$$\bar{x} \pm \sigma_{\bar{x}} \cdot t(df, tws., \alpha)$$

df = n-1

Example: 11 diameters of a wire measured in mm: 0,141 0,138 0,143 0,142 0,145
0,141 0,142 0,144 0,143 0,139 0,144

$\bar{x} = 0,1420$ mm

Arithmetic mean, estimation of μ in the population

$\sigma_{n-1} = 0,00214$ mm

Standard deviation, estimation of σ in the population

$\sigma_{\bar{x}} = 0,000645$ mm

Standard error of the mean (for n=11 measurements)

$0,1420 \pm 2.23 \cdot 0.000645$

95%-confidence interval of the true mean μ with $t_{\alpha}=2.23$, two-sided and df=10 degrees of freedom.

Quantiles and Percentiles: Quantile X_p for probability P is the number x on the x -axis with the property that exactly the part P of the population has smaller values than X_p . If one is giving the probability P in percents, then we are speaking from percentiles. What is the probability P that random numbers x are less than some quantile X_p , if x is normally distributed with some mean \bar{x} and standard deviation σ_{n-1} ?

Calculate $u = (X_p - \bar{x}) / \sigma_{n-1}$ and find in table $\Phi(u)$ the wished value of P .

Which quantile X_p we find for probability $P\%$ for normally distributed values of a population?

P is given, seek u -value in $\Phi(u)$ and calculate $X_p = \bar{x} + u \cdot \sigma_{n-1}$

Please pay attention that table $\Phi(u)$ is given for negative u -values sometimes only. Positive u -values we find for probabilities $P > 0.5$. Because of the symmetry of the Normal Distribution we find $\Phi(u) = 1 - \Phi(-u)$

The **median** of a distribution is the 50%-percentile. The 25%- and 75%-percentiles we call also **quartiles**. Their distance on x -axis is called **interquartile distance**.

Index numbers: P are prices, g are weights (piece numbers, e.g.) 0 is the base year index, 1 the index of actual year, n is the number of products in the ware basket (Used in economy)

Price index by Paasche	Quantity index by Laspeyres
$I_P = \left(\sum_{i=1}^n (g_{1i} P_{1i}) \right) / \left(\sum_{i=1}^n (g_{0i} P_{0i}) \right)$	$I_Q = \left(\sum_{i=1}^n (g_{1i} P_{0i}) \right) / \left(\sum_{i=1}^n (g_{0i} P_{0i}) \right)$

7. Test of Hypotheses

A **scientific hypothesis** is a statement concerning a population. Example: *After taking our new ACE-inhibitor the blood pressure of hypertonic patients will drop.* With *hypertonic patients* the statement means the total population of all hypertonic patients. A proof can be done only with a sample. We extrapolate the results of the (mostly small) sample on a (mostly great) population. Here **random errors** can arise. They buoy up with false hopes of a blood pressure decrease why we have chosen by chance more probands, which react positively on the new ACE-inhibitor. The next testing group could show the opposite behaviour. To handle such errors, the inferential statistics determine a maximal **error probability**. A test proves whether or not the results of a sample exceed the determined error probability. Nearly all tests of hypotheses follow this **scheme**.

Step 0: **Determination of the Null Hypothesis H_0** , which denies all real effects (e.g., all effects on blood pressure found are only by chance). Against that we determine the **Alternative Hypothesis H_A** , which postulates a significant effect (e.g., a blood pressure decrease for most of the patients). Calling, e.g., the differences $P_1 - P_2$ of the blood pressure measurements before and after therapy with d , then the hypotheses in a case of a **two-sided test** are $H_0: d=0$ and $H_A: d \neq 0$. The **Error Type I, α** , is the probability of rejection of a true null hypothesis H_0 , i.e. the probability for a false decision for hypothesis H_A . Commonly used values of α are 0.05, or 0.01, respectively (5% or 1%). The more seldom controlled **Error Type II, β** , is the

probability of rejection of a true alternative hypothesis H_A . Commonly used values of β are 10-30%.

Step 1: Choice of an appropriate **Test Statistic (Method)**. A test statistic is a quantity, which under H_0 follows some test distribution, e.g., the **u-, t-, χ^2 -, F-distribution**. (There more test distributions exist.). Step 1 needs a lot of experience and empathy in the statistics. In cases of doubt it is better to consult an experienced statistician.

Step 2: **Calculate the test statistic**, e.g., **t**, and perhaps degrees of freedom from the data of the sample. Thereby it can be that one has to perform a data transformation before using the data, if the original data distribution does not fit the demands of the test, e.g., the data are definitely not normally distributed, what is demanded by the test.

Step 3: Decision over **acceptation** or **rejection** of **H_0** . If the absolute value of the test statistic, e.g. $|t|$, is so great that the probability of its occurrence decreases below α under the assumption H_0 is true, then doubts arise that H_0 is further true, and we decide for H_A . Extremely great values of the test statistic are namely much more probably under H_A than under H_0 . The **critical values**, e.g. t_α , of the test distribution depend on α , and we find them in tables (see last page). The critical values mark those points at the t-axis, e.g., from which t-values occur only with a low probability under H_0 . In the case of a two-sided test the decision is done performing the simple comparisons of, e.g., $|t|$ and t_α :

$$\begin{aligned} &\text{Accept } H_0, \text{ if } |t| < t_\alpha, \text{ and} \\ &\text{accept } H_A, \text{ if } |t| \geq t_\alpha. \end{aligned}$$

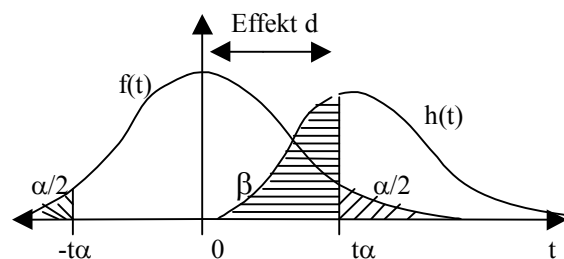
The **relationship between the error type I, α , and the error type II, β** , is illustrated by the following graphic: Curve $f(t)$ is the t-distribution by Gosset, and shall be valid under H_0 . Curve $h(t)$ is some example of any distribution of "t-values" valid under H_A . (This distribution does not interest really.) The critical value in the case of a two-sided test is t_α for $f(t)$. It appears symmetrically as $+t_\alpha$ and $-t_\alpha$. Each corner (gusset) of the $f(t)$ -curve has probability $\alpha/2$, together α .

Case 1: H_0 is valid, i.e. $h(t)$ does not exist. $f(t)$ is the valid distribution of the t-values. We are right to accept hypothesis H_0 if we find a t-value from the sample with $|t| < t_\alpha$.

Case 2: H_0 is valid. We make the error type I, α , if we reject hypothesis H_0 finding a t-value from the sample with $|t| \geq t_\alpha$.

Case 3: H_A is valid, i.e. curve $h(t)$ is the valid distribution of the t-values from the samples. We make the error type II, β , if we accept hypothesis H_0 finding a t-value from the sample with $|t| < t_\alpha$.

Case 4: H_A is valid. We are right to accept H_A finding a t-value with $|t| \geq t_\alpha$.

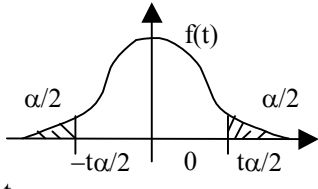
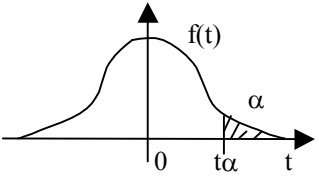
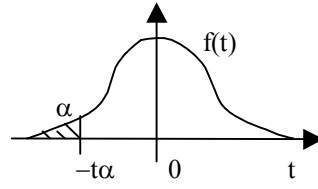


Is the **effect** small, then the two distributions $f(t)$ and $h(t)$ cover one another very strongly, and the error type II, β , is growing. One can confirm the effect statistically only if it is great enough. In general we find: **Great α \longleftrightarrow small β** and reversely. One has to find the compromise which is oftenly a financial optimization (\rightarrow experimental design).

p-Value of a test is the probability of the occurrence of the test value (or still greater values than occurred), all under the condition that H_0 is valid. A p-value of $p \leq 0.05$ means significance at the 5%-level, a p-value of $p \leq 0.01$ means significance at the 1%-level, and so on. **Power of a test** is defined as $1-\beta$, i.e. the probability to confirm a right alternative hypothesis. **Optimal Tests** have maximal power if all presuppositions are fulfilled (right data distribution, e.g., ...).

- The power is increasing with \sqrt{n} . For a fixed α and increasing the sample size n , one can decrease the β to any low amount, if one has time and money enough, and some effect is existing really.
- The power decreases if the α is lowered, i.e., one should work with the highest α allowed (5% in biology and medicine, 1% or sometimes 0.1% in pharmacy).
- The power increases with improved measuring methodology (smaller variances inside groups, e.g., ...).
- The power is better in the case of the one-sided test (but caution! You must prove the one-sided hypothesis very well.)

One-sided or two-sided test: If one does not know anything about the direction of the effect then the two-sided test is correct always. If one has **a-priori knowledge** from former scientific investigations, or logic cogently dictates a positive or a negative effect, then a one-sided hypothesis is allowed. One is rewarded by a lower amount of the critical value, i.e., one gets (with fewer probands) a significant answer.

two-sided hypothesis $H_0: \mu_1 = \mu_2, H_A: \mu_1 \neq \mu_2$	one-sided positive hypothesis $H_0: \mu_1 \leq \mu_2, H_A: \mu_1 > \mu_2$	one-sided negative hypothesis $H_0: \mu_1 \geq \mu_2, H_A: \mu_1 < \mu_2$
		
α divides up into a left and a right corner (gusset). Accordingly, the critical value $t_{\alpha/2}$ has a great amount.	The whole amount of the error type I, α , is in the right corner. Accordingly, the one-sided critical value t_{α} is smaller.	The whole amount of the error type I, α , is in the left corner. Accordingly, the one-sided critical value t_{α} is smaller.

Take the one-sided hypothesis only if you are able to prove it very well. The table above uses as an example the t-statistic and the question whether the population means μ_1 and μ_2 of two populations do differ.

Degree of freedom (df): The nomen comes from mechanics, and is giving there the number of translations and rotations which a collection of objects can perform. In statistics it is the number of independent values in a sum of squares. This number of independent values is $df = N - N_p$. Here N is the total number of squared values, and N_p is the number of independent sample parameters, which are involved. Sample parameters are parameters, which are calculated from the N sample data.

Example: Total- χ^2 of a 4x2-Contingency table	Example: Variance from n data
$\chi_{ij}^2 = \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}, \hat{e}_{ij} = \frac{n_i \cdot n_{.j}}{n}, \chi^2 = \sum_{i=1}^I \sum_{j=1}^J \chi_{ij}^2$	$\sigma_{n-1} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$
8 observed independent frequencies n_{ij} . We have 5 independent parameters for the calculation of the expectations e_{ij} : Total number n , row sum $n_{1.}$ and the 3 sums of columns $n_{.1}, n_{.2}, n_{.3}$.	n independent measuring values x_i exist. We have only one parameter, which we calculate from the data: Mean \bar{x}
df = 8-5 = 3	df = n-1

Multiple Tests and Alpha-Adjusting

If one is performing with one sample more than one test, each of them with error type I, α , e.g. $\alpha=5\%$, then one will find five significant answers between 100 tests, although in reality always the Null-hypothesis is valid. How we deal with this issue?

1. We make independent hypotheses only. We know that about α % of them are falsely evaluated. We accept this fact, because the greater part is correctly evaluated.
2. We demand that the full set of our hypotheses must be treated as a single multiple hypothesis. This multiple hypothesis is allowed to be falsely evaluated with probability α . That means that also in the case of 100 single hypotheses the total probability of an error shall not exceed α . Therefore we adjust the α of the single tests.

The **Bonferroni Adjustment** divides α by the number of hypotheses n_H , i.e. $\alpha^* = \alpha / n_H$, and tests the single hypotheses with α^* instead of α . **Holmes Sequential Procedure** computes first the p-values of all n_H single tests. Then it sorts the p-values ascending by their amount. Then Holmes procedure compares the smallest p-value with $\alpha_0 = \alpha / n_H$, the next greater with $\alpha_1 = \alpha / (n_H - 1)$, and so on until the largest p-value, which is compared with α . If a p-value is greater than his α_i , then this test and all following tests are not significant. The Bonferroni Adjustment is simpler to perform, but yields in some cases less significant tests than Holmes procedure.

8. Test of frequency numbers

8.1 Test of an observed relative frequency with a constant

We compare an observed relative frequency number \hat{p} with a given constant probability p_0 . p is the "unknown" probability of the population.

Step 0: Hypothesis $H_0: p = p_0$ $H_A: p \neq p_0$ (2-sided Test) $\alpha=0.05$ (5%)
or e.g. $H_A: p > p_0$ (1-sided >Test)

Step 1: Method asymptotical Binomial-Test: u is under H_0 asymptotically normally distributed

Step 2: Compute $\hat{p} = h / n$

h =number of yes-answers,
 n =total number of answers

$$u = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}} \sqrt{n}$$

Step 3: Decision: The critical values for $u(\alpha)$ are identical with these of $t(\alpha, df \rightarrow \infty)$ and identical with these of the standard normal distribution integral $\Phi(u)$.

With 2-sided test and $\alpha=0.05$ is $u(\alpha)=1.96$, with 1-sided test is $u(\alpha)=1.65$.

2-sided test: If $u > u(\alpha)$, then is significant $p > p_0$

If $u < -u(\alpha)$, then is significant $p < p_0$

1-sided test: If $u > u(\alpha)$, then is significant $p > p_0$ with $H_A: p > p_0$

In all other cases we accept $H_0: p = p_0$.

Numerical example: The bio-plant *Laktozar* wishes to start its new campaign in Munich if the quote of diet-friends will exceed significantly the amount of $p=20\%$. An inquiry with 100 probands gave 23 "yes" votes for the new diet.

$$p=23/100=0.23, p_0=0.2, u = ((0.23-0.2)/(0.2*0.8)) * 100^{0.5} = 0.75$$

$0.75 < 1.96$, i.e. we accept H_0 . No significant deviation from value $p=20\%$ was found. The campaign will not be started.

8.2 Comparison of two observed relative frequency numbers

(More exactly the comparison of the estimated probabilities p_1 and p_2 in two populations.) We have two samples with size n_1 and size n_2 , respectively, and h_1 , respectively h_2 "yes" answers.

Step 0: Hypothesis $H_0: p_1 = p_2$ $H_A: p_1 \neq p_2$ (2-sided test) $\alpha=0.05$ (5%)

or e.g. $H_A: p_1 > p_2$ (1-sided >test)

Step 1 : Method t-test for frequencies

Step 2 : Compute

$\hat{p}_1 = h_1 / n_1$	$\hat{p}_2 = h_2 / n_2$	$df = n_1 + n_2 - 2$
$p = \frac{h_1 + h_2}{n_1 + n_2}$	$q = 1 - p$	$t = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$

Step 3 : Decision : Find critical value $t(\alpha, df)$ from table (be cautious for 1- or 2-sided test)

2-sided test: If $t < -t(\alpha, df)$, then is significant $p_1 < p_2$

If $t > t(\alpha, df)$, then is significant $p_1 > p_2$

1-sided test : If $t > t(\alpha, df)$, then is significant $p_1 > p_2$ with $H_A: p_1 > p_2$

in all other cases we accept $H_0: p_1 = p_2$ (no significant deviation)

Numerical example: The bio-plant *Laktozar* will start its new campaign for the new diet in Germany if France has not significantly more diet-friends. Two inquiries were done - one in Germany, and one in France.

Germany: $h_1=127$ from $n_1=500$ probands voted for the new diet

France: $h_2=64$ from $n_2=300$ probands voted for the new diet

$H_0: p_1=p_2, H_A:p_1 \neq p_2$ (2-sided test), $\alpha=0.05$, i.e. $t_\alpha=1.96$

$\hat{p}_1 = \frac{127}{500} = 0.254$	$\hat{p}_2 = \frac{64}{300} = 0.213$	$df=500+300-2=798$
$p = \frac{127 + 64}{500 + 300} = 0.239$	$q = 1 - 0.239 = 0.761$	$t = \frac{0.254 - 0.213}{\sqrt{0.239 \cdot 0.761}} \sqrt{\frac{500 \cdot 300}{500 + 300}} = 1.316$

Because of $t < t_{\alpha}$, i.e., $1.316 < 1.96$, we accept H_0 . There is no significant difference in the number of diet-friends in Germany and France. The campaign will be started in Germany.

9. Contingency Tables

Contingency tables arise from analyzing categorical variables. The number of variables designates the dimension of the table (e.g. two variables build a matrix with rows and columns, 3 variables build a 3-dimensional array, and so on). Example: An inquiry of 100 probands concerning their smoking habit. Variable *sex* has two categories (*female / male*). Variable *smoking habit* has 3 categories here (*never / moderate / strong*). The frequencies n_{ij} is termed here **configuration** or **cell**. Configurations or cells we identify by their indexes i, j, k, \dots

	smoke never	moderate	strong
f	$n_{11}=22$	$n_{12}=17$	$n_{13}=11$
m	$n_{21}=26$	$n_{22}=16$	$n_{23}=8$

In DASYS we can read a contingency table as matrix, or we can read rough data, and compute then the contingency table from the data. We can use only categorical data to build a table. But one can also transform a metric variable into a categorical variable. The simplest transformation makes a dichotomous (or binary) 0/1-variable. The categories of the variable must be coded as numbers 1, 2, 3,

What does contingency table analysis?

- Contingency test (Chi-Square-Test of independence of categorical variables)
- Search for types with **Configuration Frequency Analysis** by G.A.Lienert and N. Victor
- Chi-Square-decomposition of a contingency table by Lancaster (dependence structure)
- Selection of variables by stage-wise reduction of a n-dimensional table
- Analysis of 2x2-tables (dependence measures and association measures, search for types with the zero-order model by A. von Eye)

9.1 Test of contingency or test of homogeneity

Tests association or independence otherwise of two categorical variables. We start with a **contingency table** built from $k \geq 2$ categorical variables. Hypothesis H_0 of the global test: The variables are independent - there is no association (or contingency). Hypothesis H_A of the global test: The variables are dependent - there an association or contingency exists. The test statistic is χ^2 , and we test 1-sided for exceeding the upper critical value of the χ^2 -distribution with df degrees of freedom.

Step 0: Hypothesis H_0 : „No association“, H_A : „Significant association“
 $\alpha=0,05$ (5%)

Step 1: Method global χ^2 -test in contingency tables

Step 2: n_{ij} = frequency of symptom configuration (i, j) (example is with $k=2$)

$n_{i.}$ = sum of row i $n_{.j}$ = sum of column j

I = number of rows J = number of columns

n = total sum (number of all probands or cases)

Compute degree of freedom, expectations, χ^2 -components and total χ^2

$$df = I J - (I-1) - (J-1) - 1 \qquad \hat{e}_{ij} = \frac{n_{i.} n_{.j}}{n}$$

$$\chi_{ij}^2 = \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \quad \chi^2 = \sum_{i=1}^I \sum_{j=1}^J \chi_{ij}^2$$

Step 3: Find the critical value $\chi^2(\alpha, df)$. If $\chi^2 \geq \chi^2(\alpha, df)$ then assume a significant association or a significant contingency between the variables, otherwise accept H_0 : No significant association (no contingency).

9.2 Configural Frequency Analysis (CFA) of Lienert and Victor

Contingency tables (or cross classification tables) are used to analyze the relations between categorical variables. The Configural Frequency Analysis (CFA) founded by G. A. Lienert (1969) has been shown to be an universal method for analyzing contingency tables.

Simplified search for the strongest type or antitype by Krauth and Lienert

We start with a contingency table (see above). Make the χ^2 -test for contingency! If it is significant then a **contingency type** exists, otherwise not. It may be a type or antitype. A contingency type is significantly over-frequented in relation to its expectation. An antitype is significantly under-frequented in relation to its expectation. To find the strongest contingency type or antitype in the case of a significant contingency in the table, please scan the cells for that cell with the greatest χ^2 -component. This cell is the type or antitype searched. The χ^2 -components are computed already for the global χ^2 -test, so that you will have no additional effort.

Numerical example of search for types: By analyzing video tapes made in a central pharmacy one wants to find Lienert's contingency type or antitype. That means that type of customer which appears much more frequently than expected under hypothesis of independence. Analyzing the tapes the sex of the customers (female, male) is recorded and the shopping time (forenoon, noon, afternoon, evening). The contingency table with row sums and column sums and expectations \hat{e}_{ij} and χ^2 -components, both already computed, is:

	forenoon	noon	afternoon	evening	sum
f	$n_{11}=132$ $\hat{e}_{11}=119.4$ $\chi^2_{11}=\mathbf{1.32}$	$n_{12}=92$ $\hat{e}_{12}=105.6$ $\chi^2_{12}=1.75$	$n_{13}=74$ $\hat{e}_{13}=58.2$ $\chi^2_{13}=\mathbf{4.29}$	$n_{14}=179$ $\hat{e}_{14}=193.8$ $\chi^2_{14}=1.13$	$n_{1.}=477$
m	$n_{21}=67$ $\hat{e}_{21}=79.6$ $\chi^2_{21}=1.99$	$n_{22}=84$ $\hat{e}_{22}=70.4$ $\chi^2_{22}=\mathbf{2.63}$	$n_{23}=23$ $\hat{e}_{23}=38.8$ $\chi^2_{23}=6.43$	$n_{24}=144$ $\hat{e}_{24}=129.2$ $\chi^2_{24}=\mathbf{1.70}$	$n_{2.}=318$
sum	$n_{.1}=199$	$n_{.2}=176$	$n_{.3}=97$	$n_{.4}=323$	$n=795$

For example is $n_{1.}=132+92+74+179=477$ or $\hat{e}_{11} = 477*199 / 795 = 119.4$

For cells with $n_{ij} > \hat{e}_{ij}$ the χ^2 -value is typed in bold letters. This cells are candidates for the type searched. The other cells are candidates for antitypes.

Sum of the χ^2 -values is $\chi^2 = \chi^2_{11} + \chi^2_{12} + \dots + \chi^2_{24} = 1.32 + 1.75 + \dots + 1.70 = 21.24$

Number of the degrees of freedom is $df=2*4-1-3-1 = 3$

The critical value of the χ^2 -distribution with $\alpha=0.05$ and 3 degrees of freedom is $\chi^2_{\alpha} = 7.81$

We reject H_0 , because of $\chi^2 > \chi^2_{\alpha}$, i.e., $21.24 > 7.81$.

At least one significant type or antitype does exist at the α -level $\alpha=5\%$.

We find the highest χ^2 -value with an amount of 6.43 in cell (2,3), an antitype cell.

Result: Male customers appear more seldomly at afternoon than expected - a reason to show PR-videos specially for women in this pharmacy at afternoon.

Full search for types with the CFA

Type/Antitype by G.A.Lienert: If the observed cell frequency n_{ijk} is significantly greater than the expectation e_{ijk} , then we speak from a contingency type. If it is significantly less, i.e. $n_{ijk} < e_{ijk}$, we speak from an antitype. But the definition of antitypes is disputed. Typen nach Victor: If the observed cell frequency n_{ijk} is significantly greater than a specially computed expectation V_{ijk} (Victor-expectation), then we speak from a Victor-type. We estimate the V_{ijk} from the marginal sums also, but these sums are calculated without the surplus-frequencies of the type-cells. The "surplus" shall not falsify the expectations under independence. The task can be solved by iteration only, since the type cells are a priori unknown.

H₀ in the local cell test (single test): The cell is not a type - deviations of the observed frequency from the expectation under independence are by chance only. **H_A** for the single test: The cell is a type or an antitype - deviations are caused by a real effect and reproducibly. The confirmation of the multiple hypothesis we make with Holme's procedure. Local cell tests evaluate by a test statistic the difference $n_{ijk} - e_{ijk}$ of each cell. For example, the Chi-Square component test computes the statistics $\chi^2_{ijk} = (n_{ijk} - e_{ijk})^2 / e_{ijk}$ for each cell (i,j,k). The degree of freedom is (by Perli et al.) $df=1$ for each χ^2_{ijk} . If the distance is significant then a type (or antitype) is found.

1-sided test: One tests 1-sided for types if $n_{ijk} > e_{ijk}$, or $n_{ijk} > V_{ijk}$, respectively. One tests 1-sided for antitypes if $n_{ijk} < e_{ijk}$, or $n_{ijk} < V_{ijk}$, respectively. The total amount of the given error probability α is put into one side of the distribution of the test statistic, and decreases so the amount of the test statistic needed for the significant identification of a type/antitype. A possible motivation for 1-sided testing: A visible deviation of the frequency n_{ijk} from its expectation is a pre-information in the sense of Bayes, allowing us 1-sided testing. **2-sided test:** For each cell both hypotheses are allowed (type or antitype). Error probability α is divided into two equal parts. The critical value is higher compared with the 1-sided test, but otherwise the 2-sided test needs no motivation: Its use is always correct.

Continuity correction: Small cell probabilities ($e_{ijk} < 5$) are producing oftenly anti-conservative results. One reduces the test statistic on purpose (acting in all cases of small frequencies especially strong), and reduces so the probability of errors. Lautsch and v. Weber showed in 2001 that each test can have both - conservative (yields too much types) and anti-conservative (yields too less types) behaviour. A correction constant K computed individually for each table ensures in DASY that asymptotically the given α is not exceeded, but also is exhausted. Only in this way we can minimize the type-II error, β .

A paper of v.Eye, Lautsch and v.Weber (2004) recommends the following local tests in the given order:

1. Combinatoric Search by Dunkl, von Eye, Lautsch, Victor, von Weber
2. Gradient Method by Lautsch and von Weber, if the computing time of the Combinatoric Search is too long.
3. Chi-Square-Test by Lienert (for 2^d -tables and $df > 20$)
4. Alternatively to 3. the asymptotical test by Perli et al. . (for 2^d -Tafeln and $df > 20$)

Numerical example CFA: The famous LSD-Data of G.A.Lienert from 1970 are showing the psychotoxic syndrome found by Leuner in 1962. 65 studentens took voluntary lysergacid diethylamide (LSD), and made, if still able, different tests. Leuner's syndrome is a combination from

M01 = clouded consciouness

M02 = disturbed thinking

M03 = altered affectivity

Die Typensuche mit der Combinatoric Search:
 65 Probanden, 8 Zellen, mBl= 8.13 mittlere Belegung
 37.92 Chi-Quadrat-Gesamt mit FG=4
 6.346E-06 (***) einseitige Irrtumswahrscheinlichkeit
 6.00 geschaetztes maximales Typgewicht
 Test: Combinatoric Search (Weber et al.) Zweiseitig
 Geschaetzter Korrekturwert= -1.11
 Geschaetztes Beta = 34.03 %
 Sie arbeiten mit Alpha = 0.05

Nr.	i	j	k	l	m	N _{ijk}	E _{ijk}	V _{ijk}	koTw	KIW	T/AT	Signif
001	1	1	1	.	.	20	12.51	0.69	4.47	0.00000	1	***
002	1	1	2	.	.	1	6.85	2.12	-0.26	0.39743	0	
003	1	2	1	.	.	4	11.40	3.65	0.08	0.47006	0	
004	1	2	2	.	.	12	6.24	11.24	0.10	0.45887	0	
005	2	1	1	.	.	3	9.46	2.92	0.02	0.49232	0	
006	2	1	2	.	.	10	5.18	8.97	0.15	0.43883	0	
007	2	2	1	.	.	15	8.63	15.44	-0.05	0.47935	0	
008	2	2	2	.	.	0	4.73	47.51	-3.24	0.00060	-1	***

ijklm sind Zellindizes

E_{ijk} Unabhaengigkeits-Erwartungswerte aus den Randsummen berechnet

V_{ijk} VICTOR-Erwartungswerte bei der kombinatorischen Suche und beim Gradientenverfahren. Sonst ist E_{ijk}=V_{ijk} gesetzt.

Ein E_{ijk} bzw. V_{ijk} kleiner 3 wird im Test auf 3 hochgesetzt

koTw Testwerte mit Stetigkeitskorrektur nach Lautsch und v. Weber

KIW Einseitige Irrtumswahrscheinlichkeiten zum Testwert koTw

A/AT Eine 1 bedeutet Typ, eine -1 Antityp, eine 0 weder/noch.

* bedeutet ein KIW um 0.05, ** um 0.01, *** um 0.001

The Combinatoric Search finds the type (1,1,1) and antitype (2,2,2). Surprisingly great the Victor expectation $V_{ijk}=47.51$ of the antitype (2,2,2), and otherwise, the small amount of $V_{ijk}=0.69$ of the type (1,1,1). The sum of the V_{ijk} must not yield the number $N=65$ of probands, as we expect from the sum of the E_{ijk} . The 6 cells 002-007 define an average level of action of LSD. One can see that the Victor expectations equals very exactly the observed frequencies n_{ijk} . Cell 001 is an outlier in the sense that the action of the drug has reinforced extremely, so that we can find no normal reaction in the test. Cell 008 is an outlier in the sense that we expect much more probands not affected by the drug. The two outliers show that the action of LSD does not follow a simple log-linear model.

9.3 χ^2 -analysis of Lancaster

In 1973 Krauth and Lienert made the proposal - in connection with the association and contingency structural analysis of 2^d -tables - to use the χ^2 -analysis of Lancaster (1951) for the proof of **interactions** between the variables (symptoms). The Chi-square of a 2×2 -CFA is assumed to be the measure of interaction of the first order, i.e. the measure of the deviation from the basic model (interaction of 0. order). the basic model. If we use lower case letters a, b, c, ... as indices of the parts of chi-square, then we find the three formal identities for the 3 variables A, B, and C and interaction of the 1. order:

$$\chi^2 ab = \chi^2 (A,B) \quad \chi^2 ac = \chi^2 (A,C) \quad \chi^2 bc = \chi^2 (B,C)$$

In the case of missing interaction of 2. order one can explain the total χ^2 of the 2^3 - CFA with the interactions of 1.order alone. Any residuum is assumed to be an interaction of 2. order (interaction of 3 variables):

$$\chi^2_{abc} = \chi^2(A,B,C) - \chi^2_{ab} - \chi^2_{ac} - \chi^2_{bc}.$$

We give one degree of freedom to each χ^2 -part. If our table is not of type 2^d , then we had to seek the interesting configurations (types) by means of the CFA. The association structure of a special type we analyse by a collapsed table. For the LSD-data of Krauth and Lienert with its variables C= clouded consciousness, D= disturbed thinking, and A= altered affectivity we find the following χ^2 -value, degrees of freedom and χ^2 -parts, i.e. interactions:

χ^2	df	χ^2 -part
$\chi^2(C,D,A) = 37,92$	4	$\chi^2_{cda} = 36,95$ ***
$\chi^2(C,D) = 0,68$	1	$\chi^2_{cd} = 0,68$
$\chi^2(C,A) = 0,00$	1	$\chi^2_{ca} = 0,00$
$\chi^2(D,A) = 0,29$	1	$\chi^2_{da} = 0,29$

The *** marked interaction is significant at level $\alpha=0.001$, so that we had to assume an interaction of 2. order. The number of χ^2 -parts produced by the algorithm will increase rapidly with the number d of variables. In the case d=5 we will find 26 terms already.

9.4 Variable selection - seeking the most significant table

If one has rough data with more than 5 categorical variables then (at least in DASYS) we had to select proper variables. Otherwise we can not make local CFA-tests or a χ^2 -analysis. How we can find a group of variables with much information? One possibility is the evaluation of a contingency table by its chi-square. We start with the full table of all variables, and then alternately exclude on trial one variable. With the variables remained we compute the table, the chi-square, the degrees of freedom, and the probability P. That variable is selected and excluded from the variable set whose rejection results in the lowest increase of probability P. That set of variables is wanted which results in a table of ultimately low probability. This steps of selection are repeated until still only two variables are in the set.

9.5 2x2-Tables: Association measures, searching types

	nonsmokers	smokers	row sums
female	a= 37	b= 13	a+b= 50
male	c= 29	d= 21	c+d= 50
column sums	a+c= 66	b+d= 34	total N=100

2x2-tables arise if we analyse two binary variables. In most cases, we ask for the independence of the variables or for its association. More seldomly, we ask for types here. Example:: Variable M1 is the smoking habit (nonsmoker / smoker), variable M2 is the sex (female / male).

The four numbers $a=N_{11}=37$, $b=N_{12}=13$, $c=N_{21}=29$, $d=N_{22}=21$ are the cell frequencies. The marginal sums $N_{i.}$ and $N_{.j}$, and the total N are computed from the frequencies. From the marginal sums und the total N we compute the cell expectancies E_{ij} . Depending on the chosen model we have different estimations of the cell probabilities: The contingency model calculates $E_{ij} = N_{i.}N_{.j} / N$, von Eye's cluster model (CFA of 0-th order) calculates $E_{ij}=P_i P_j$ (P_i , P_j given marginal probabilities)

A research paper from 2003 (v. Eye, Lautsch and v. Weber) recommends especially the following 6 association measures (contingency measures). The formulas we find in DASYS's help.

Friedrich Vogel's Z_v (F. Vogel, Bamberg)	Normal approximated Z
log-odds ratio θ Teta	log-linear interaction λ Lambda
Goodman's Lambda λ_{Good}	Binomial test K_r (J. Krauth, Düsseldorf)

Association measures do two jobs:

1. To evaluate and to make comparable the association between two variables, similarly, as the correlation coefficient works in the case of two metric variables. Unfortunately, this measures fulfill the conditions only partially, because they are not normed to yield values from the interval $[-1; +1]$, or they use only a part of this interval.
2. To be a test statistic for the "test of independence of the two variables". This task is fulfilled well by the measures recommended above.

Exampel to Vogel's Z_v : One of the best contingency measures for 2x2-tables

$$Z_v = \frac{d(U, G)}{d(U, M)} \quad \text{with} \quad d(U, G) = \sum_{ij} |\hat{e}_{ij} - n_{ij}| \quad \text{and} \quad d(U, M) = \sum_{ij} |\hat{e}_{ij} - m_{ij}|$$

Here is:

n_{ij} = observed frequencies

m_{ij} = maximal-minimal-values of the table under conservation of the marginal sums

The sign of Z_v ist that of the deteminant Ist $D=ad-bc$ of the table. If $D=ad-bc>0$, then is also $Z_v>0$.

Given the following observed 2x2-Table::

	smoker	nonsmoker	sum
High pressure	49	10	59
Low pressure	20	39	59
sum	69	49	118

We build the maximal-minimale table. The smallest frequency iss et to zero (here frequency 10 is set to 0). Under conservation of the marginal sums we find:

	smoker	nonsmoker	sum
High pressure	59	0	59
Low pressure	10	49	59
sum	69	49	118

With the expectations e_{ij} : e.g.. $e_{11} = 59 \cdot 69 / 118 = 34,5$ or $e_{12} = 59 \cdot 49 / 118 = 24,5$

	smoker	nonsmoker
High pressure	34,5	24,5
Low pressure	34,5	24,5

Weg et $d(U, G) = |49 - 34,5| + |10 - 24,5| + \dots = 58$

and $d(U, M) = |59 - 34,5| + |0 - 24,5| + \dots = 98$

an with determinant $D = ad - bc = 59 \cdot 49 - 10 \cdot 20 = 2691 > 0$.

We compute $Z_v = 58 / 98 = +0,59$.

The critical values of Z_V one can only find by the simulation of many tables with this marginal sums. The author has found using the DASY software a value of $Z_V(\alpha=5\%) = 0,267$.

H_0 : There is no significant association between blood pressure and smoking habit.

H_A : There is a significant association between blood pressure and smoking habit.

We accept H_0 if $Z_V < Z_V(\alpha=5\%)$.

We accept H_A if $Z_V \geq Z_V(\alpha=5\%)$.

Here is $Z_V \geq Z_V(\alpha=5\%)$. We postulate a significant association between blood pressure and smoking habit.

Search for types in 2x2-tables: Another result of the paper cited above is the fact that any search for contingency types makes no sense in 2x2-tables. We can never realize an error type I of $\alpha < 0.5$ ($\alpha < 50\%$). The contingency model estimates the cell probabilities from the marginal sums. That means the lost of 3 degrees of freedom. Only one independent hypothesis remains. The "Zero-Order model" of A. von Eye (also called Cluster model or **CFA of Order 0**), works without the marginal sums. The Zero-Order-Model is working with fixed probabilities of all 4 cells as basic model (e.g., $P_{ij}=0.25$ for all 4 cells in the case of assumed uniform distribution of both variables). Because one uses only the total proband number N to compute the expectations, one has 3 degrees of freedom. That are two more than in the case of the contingency- or independence-model. That means more "redundant information", which allows a better identification of a type cell.

Example: A study concerning the smoking habit of men and women is assuming uniform distribution of the sexes in the observed district of the survey (50% female, 50% male) and a ratio of 40% of smokers to 60% of non-smokers. These numbers (40%, 60%) we could have found in another survey of the whole country. The actual survey is done in a single plant or a single college. The probabilities of the 4 cells are the products of the marginal probabilities given above, i.e., $0.5*0.4$, $0.5*0.6$. One tests the deviation from the average of the country, i.e., a type found with this study has local meaning only, perhaps in the sense "Here in this plant more men are smoking than in average of the country". Local cluster-type would be "the male smoking worker of factory Xyz". The user has a choice between 3 type tests:

1. Small group test by von Eye / Dunkl
2. Chi-Component test by G. A. Lienert
3. Exact Binomial test by J. Krauth

10. χ^2 -test of fit for a distribution

Test for Normal distribution in a population: We have a sample with n values x_1, x_2, \dots, x_n . The number of values should be 25 at least, since otherwise the test can fail. You will test whether the data are normally distributed.

Step 0: Hypothesis H_0 : „There is no objection against assumption of Normal distribution in the population“. Hypothesis H_A : „The data of the population have not the Normal distribution“.

Step 1: Method χ^2 -test of class frequencies

Step 2: Compute \bar{x} , σ_{n-1} , n of the sample, build $k=5$ classes symmetrically to the mean according to the table below, count the frequencies h_i of the values in each class, compute the class expectations \hat{e}_i and the χ^2 -components χ^2_i of all classes. (σ_{n-1} is called σ in the table)..

class	FROM _i	TO _i	h _i	\hat{e}_i	χ^2_i
1	$-\infty$	$\bar{x} - 0,84 \sigma$	h ₁	n·0,20	χ^2_1
2	$\bar{x} - 0,84 \sigma$	$\bar{x} - 0,25 \sigma$	h ₂	n·0,20	χ^2_2
3	$\bar{x} - 0,25 \sigma$	$\bar{x} + 0,25 \sigma$	h ₃	n·0,20	χ^2_3
4	$\bar{x} + 0,25 \sigma$	$\bar{x} + 0,84 \sigma$	h ₄	n·0,20	χ^2_4
5	$\bar{x} + 0,84 \sigma$	$-\infty$	h ₅	n·0,20	χ^2_5

$\chi^2 = \sum \chi^2_i$

We put value x in class i , if $FROM_i \leq x < TO_i$. h_i = frequency of values counted in class i ($i=1,2,\dots,k$). (Factor 0,2 in column \hat{e}_i of expectations arises from $\hat{e}_i = 1/k$, class boundaries FROM and TO arises from distribution curve $\Phi(u)$ of the Normal distribution, e.g., $\Phi(+0,25) - \Phi(-0,25) = 0,2$). The χ^2 -components χ^2_i we compute with relation $\chi^2_i = (h_i - \hat{e}_i)^2 / \hat{e}_i$. From the components we compute the total χ^2 .

Step 3: Statement: Critical value in the case of 5 classes is $\chi^2_\alpha = \chi^2(\alpha=0.05, df=3)=7.81$. If $\chi^2 < 7.81$, then accept hypothesis H_0 : „There is no objection against assumption of Normal distribution in the population“, otherwise accept hypothesis H_A : „The data of the population have significantly not the Normal distribution“.

11. Comparison of means

11.1 One-sample t-Test (Tests the population mean against a constant)

We have a sample x_1, x_2, \dots, x_n . The population mean is μ . We test μ against a constant μ_0 . The constant μ_0 can be a norm of the government, a value from literature given without error bar, or some other theoretically founded number without error bar.

Step 0: Hypothesis $H_0: \mu = \mu_0$ $H_A: \mu \neq \mu_0$ (2-sided test) $\alpha = 0.05$ (5%)
or e.g. $H_A: \mu > \mu_0$ (1-sided ">test")

Step 1 : Method t-test

Step 2: $t = \frac{\bar{x} - \mu_0}{\sigma_{n-1}} \sqrt{n}$, $df = n - 1$

Step 3 : Statement : Take critical value $t(\alpha, df)$ from t-table (pay attention to 1- or-2-sided)

2-sided test: If $t < -t(\alpha, df, 2\text{-sided})$ then is significantly $\mu < \mu_0$

 If $t > t(\alpha, df, 2\text{-sided})$ then is significantly $\mu > \mu_0$

1-sided test, e.g. : If $t > t(\alpha, df, 1\text{-sided})$ then is significantly $\mu > \mu_0$

in all other cases accept $H_0: \mu = \mu_0$ (no significant difference)

11.2 Comparison of two normally distributed populations

We have two samples (measurements, observations) $x_{11}, x_{12}, \dots, x_{1n_1}$ and $x_{21}, x_{22}, \dots, x_{2n_2}$ with sample sizes n_1 and n_2 . The first index is the number of the sample, i.e. 1 or 2, the second index numbers the values inside the sample by 1,2,3,... One wants to test whether the difference of the population means is significant.

Step 0: Hypothesis $H_0: \mu_1 = \mu_2$ $H_A: \mu_1 \neq \mu_2$ (2-sided test) $\alpha = 0.05$ (5%)
or e.g. $H_A: \mu_1 > \mu_2$ (1-sided ">Test")

Step 1 : Method t-test with pooled standard deviation

Step2 : Calculate for each sample \bar{x}_i , $SQD_i = n_i \sigma_{in}^2 = (\sum x_{ij}^2) - n_i (\bar{x}_i)^2$, $i=1,2$

Calculate
 $df = n_1 + n_2 - 2$

$$\bar{\sigma} = \sqrt{\frac{SQD_1 + SQD_2}{n_1 + n_2 - 2}} \quad t = \frac{\bar{x}_1 - \bar{x}_2}{\bar{\sigma}} \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}$$

Step 3 : Statement : Take critical value $t(\alpha, df)$ from t-table (pay attention to 1- or-2-sided)

2-sided test: If $t < -t(\alpha, df, 2\text{-sided})$ then is significantly $\mu_1 < \mu_2$

If $t > t(\alpha, df, 2\text{-sided.})$ then is significantly $\mu_1 > \mu_2$

1-sided test, e.g. : e.g. $H_A: \mu_1 > \mu_2$

If $t > t(\alpha, df, 1\text{-sided})$ then is significantly $\mu_1 > \mu_2$

in all other cases accept $H_0: \mu_1 = \mu_2$ (no significant difference)

Numerical example for the comparison of two normally distributed populations

A photoreactor destroys in water solved organic substances by the means of UV-radiation. Its effectiveness is measured in [mg/KWh] for a standardized testing substance. We have two rows of measurements with different sizes (lamp A and lamp B).

Effectiveness of lamp A in mg/KWh	3.6	2.9	3.0	4.1	---
Effectiveness of lamp B in mg/KWh	3.9	4.4	3.2	3.8	4.3

Question: *Can we state generally a significant difference between effectiveness of typy-A lamps and type-B lamps?*

First the calculation scheme for the sums:

	Lamp A= x_1	x_1^2	Lamp B= x_2	x_2^2
1	3.6	12.96	3.9	15.21
2	2.9	8.41	4.4	19.36
3	3.0	9.00	3.2	10.24
4	4.1	16.81	3.8	14.44
5	---	---	4.3	18.49
Σ	13.6	47.18	19.6	77.74

$n_1=4$, $\bar{x}_1=3.40$ [mg/l], $SQD_1 = n_1\sigma_{1,n}^2 = (47.18 - 4 \cdot 3.40^2) = 0.940$ size, mean, SQD
 $n_2=5$, $\bar{x}_2=3.92$ [mg/l], $SQD_2 = n_2\sigma_{2,n}^2 = (77.74 - 5 \cdot 3.92^2) = 0.908$ size, mean, SQD

Hypothesis $H_0: \mu_1 = \mu_2$, $H_A: \mu_1 \neq \mu_2$ (2-sided test), $\alpha=0.05$ (5%) Pair of hypotheses

$\bar{\sigma} = ((0.940 + 0.908) / (4 + 5 - 2))^{0.5} = 0.5138$ [mg/l] pooled σ

$df = 4 + 5 - 2 = 7$ degrees of freedom

$t_\alpha = t(\alpha=0.05, df=7, 2\text{-sided}) = 2.36$ critical value

$t = ((3.40 - 3.92) / 0.5138) * ((4 \cdot 5) / (4 + 5))^{0.5} = -1.509$ test-statistic

We accept H_0 because of $|t| < t_\alpha$ choice of hypothesis

There is no significant difference between effectiveness of type-A lamps and type-B lamps.

One takes the averaged standard deviation $\bar{\sigma}$ for the 2-sample-t-test for independent samples if it is supposed, that the variances of the two populations are equal (homoscedasticity). In the case of different variances (heteroscedasticity) this inequality doesn't harm if the sample sizes are both $n_1 > 30$ and $n_2 > 30$. But if it isn't so, then we should perform the Welch-test or a similarly constructed test. The Welch-test yields degrees of freedom which are in general not integers. We had to round them.

Comparison of two normally distributed populations with unequal variances and either $n_1 \leq 30$ or $n_2 \leq 30$ or both $n \leq 30$ (Welch-Test).

Hypotheses $H_0: \mu_1 = \mu_2$ $H_A: \mu_1 \neq \mu_2$ (2-sided test) $\alpha = 0.05$ (5%)

Method: Welch-test with averaged standard deviation and adjusted degrees of freedom.

Compute for each sample the mean \bar{x}_i , standard deviation $\sigma_{i, n-1}$ with $i=1,2$

$$\bar{\sigma} = \sqrt{\frac{\sigma_{1, n-1}^2}{n_1} + \frac{\sigma_{2, n-1}^2}{n_2}}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\bar{\sigma}}$$

With degrees of freedom $df = \frac{(g_1 + g_2)^2}{\frac{g_1^2}{n_1 - 1} + \frac{g_2^2}{n_2 - 1}}$ and $g_1 = \frac{\sigma_{1, n-1}^2}{n_1}$ and $g_2 = \frac{\sigma_{2, n-1}^2}{n_2}$.

Compute by Excel function TINV(...) critical value $t(\alpha, df, 2\text{-sided})$

or round the degrees of freedom into an integer and use the table at page 2.

2-sided test: If $t \leq -t(\alpha, df, 2\text{-sided})$, then is significantly $\mu_1 < \mu_2$

If $t \geq t(\alpha, df, 2\text{-sided})$, then is significantly $\mu_1 > \mu_2$

in all other cases accept $H_0: \mu_1 = \mu_2$ (no significant difference).

F-test for the decision, whether we have equal or significant unequal variances in our two populations.

If $\sigma_{1, n-1}^2$ and $\sigma_{2, n-1}^2$ are the variances of the samples $x_{11}, x_{12}, \dots, x_{1n_1}$ and $x_{21}, x_{22}, \dots, x_{2n_2}$ with size n_1 and n_2 then the test statistic F with

$$F = \frac{\sigma_{1, n-1}^2}{\sigma_{2, n-1}^2} \text{ is under } H_0 \text{ F-distributed with } df_1 = n_1 - 1 \text{ and } df_2 = n_2 - 1 \text{ degrees of freedom.}$$

$H_0: \sigma_1^2 = \sigma_2^2$ Equality of variances (homoscedasticity) in the two populations.

$H_A: \sigma_1^2 \neq \sigma_2^2$ Unequality of variances (heteroscedasticity) in the two populations.

We accept H_A if $F \geq F(\alpha, df_1, df_2)$ (critical point of the F-distribution).

We accept H_0 if $F < F(\alpha, FG_1, FG_2)$.

If $F < 1$, then we take the reverse value $1/F$, and we test with that value instead. Hereby the two degrees of freedom are changed. Now is $df_1 = n_2 - 1$ and $df_2 = n_1 - 1$.

The F-table at page two gives the critical values of a 1-sided test (5% error probability at the right side of the distribution curve).

We use the simple t-Test in the case of equal variances (Hypothesis H_0 by the F-test.)

We use the Welch-Test in the case of unequal variances (Hypothesis H_A by the F-test.)

Numerical example of a F-test and the Welch-test in the case of unequal variances

Two independent samples are given. Two groups of patients (both with a light demency) have made an IQ-test. Group 1 without a drug, group 2 with a drug. Question: Does the drug influence the IQ significantly?

Group 1	102	89	97	88	94	100	91	97	105	102	95	93	99	90	95
Group 2	82	116	104	87	98	74	114	79	98	84	113	117	114	123	

$$\sigma_{1,n-1} = 5,12974519 \quad \sigma_{2,n-1} = 16,4652481 \quad n_1=15 \text{ (df=14)} \quad n_2=14 \text{ (df=13)}.$$

Because $\sigma_2 > \sigma_1$ we change nominator and denominator, i.e., our F is $F = \sigma_2^2 / \sigma_1^2$.
 $F = (16,4652481)^2 / (5,12974519)^2 = 10,30255575$ with $df_1=14$ and $df_2=13$.

We find the critical point $F(\alpha=5\%, df_1=14, df_2=13) = 2,554$

- By interpolation in our F-table at page 2
- Or by use of the EXCEL function =FINV(0,05 ; 14 ; 13)

Here, we accept H_A , because $F \geq F(\alpha, df_1, df_2)$. There exists a significant difference between the two variances of the two populations (without drug and with drug).

The Welch-test is recommended.

$$\bar{\sigma} = \sqrt{\frac{\sigma_{1,n-1}^2}{n_1} + \frac{\sigma_{2,n-1}^2}{n_2}} = 4,5955 \quad \text{and}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\bar{\sigma}} = (95,8 - 100,214) / 4,5955 = -0,9606$$

$$g_1 = \frac{\sigma_{1,n-1}^2}{n_1} = 1,754 \quad , \quad g_2 = \frac{\sigma_{2,n-1}^2}{n_2} = 19,365 \quad , \quad FG = \frac{(g_1 + g_2)^2}{\frac{g_1^2}{n_1-1} + \frac{g_2^2}{n_2-1}} = 15,345.$$

We round $df=15$. Critical t-value is $t(\alpha=5\%, df=15, 2\text{-sided}) = 2,13$.

Because $|t| < t(\alpha=5\%, df=15, 2\text{-sided}) = 2,13$ we accept hypothesis H_0 .

The drug has no significant influence on the IQ mean value (but on the variance, as the F-test has shown).

11.3 Mann-Whitney-Test (Comparison of two means, Rank test)

We have two samples (measurements, observations) x_1, x_2, \dots, x_n and y_1, y_2, \dots, y_m with sample sizes n and m . You want to test the difference of the population means. The data **are not normally distributed**, and we can not find a proper transformation to make them normally distributed, or we do not want to perform such a transformation. The Mann-Whitney-test is a rank test for arbitrarily distributed, independent variables and comparison of just two groups.

Step 0: Hypothesis $H_0: \mu_1 = \mu_2$ $H_A: \mu_1 \neq \mu_2$ (2-sided test) $\alpha = 0.05$ (5%)
 or e.g. $H_A: \mu_1 > \mu_2$ (1-sided ">test")

Step 1: Rank test of Mann-Whitney

Step 2: One sorts the merged data getting the rank order. An odd number of equal values has the same rank, e.g., ...25, **27, 27, 27**, 29 ... An even number of equal values has the averaged rank, e.g., ...,25, **26.5, 26.5**, 28, ...

We calculate the rank sum R_x to the x-values and the rank sum R_y to y-values.

From these sums we calculate the test statistic U_X and U_Y :

$$U_X = nm + \frac{n(n+1)}{2} - R_X, \quad U_Y = nm + \frac{m(m+1)}{2} - R_Y,$$

If $n \leq 10$ or $m \leq 10$, then one calculates $U = \min(U_X, U_Y)$, and is ready.

Otherwise one calculates u from U , i.e.,
$$u = \frac{U - (nm/2)}{\sqrt{nm(n+m+1)/12}}$$

Step 3: We find critical values $U_\alpha = U(\alpha, n, m)$ e.g. in E. Weber, Tab. 19 ff.

2-sided test: If $U = U_X > U_\alpha$, then is significantly $\mu_X > \mu_Y$

If $U = U_Y > U_\alpha$, then is significantly $\mu_Y > \mu_X$

1-sided test: If $U = U_X > U_{\alpha/2}$, then is significantly $\mu_X > \mu_Y$

in all other cases we accept $H_0: \mu_X = \mu_Y$ (no significant difference)

If $n > 10$ and $m > 10$ then we compare u with the 2-sided critical value of the Normal distribution, $u_\alpha = 1.96$, (1-sided $u_\alpha = 1.65$, both for $\alpha = 0.05$)

11.4 Paired t-Test

The paired t-test focuses on the difference between the paired data and reports the probability that the actual mean difference is consistent with zero. The data should be normally distributed. We have a sample of n correlated pairs of values $(y_1, x_1), (y_2, x_2), \dots, (y_n, x_n)$. *Correlated* means that we have measured or observed both values, y_i and x_i , at the same object, e.g., y is the blood pressure before medication, and x is the same blood pressure after medication. Which quantity is named x and which is named y is arbitrary. One has only to consider the sign of the effect $d = y - x$.

Step 0: Hypothese $H_0: d = 0$ $H_A: d \neq 0$ (2-sided test) $\alpha = 0.05$ (5%)
 or e.g. $H_A: d > 0$ (1-sided "> test")

Step 1: Method paired t-Test

Step 2: Calculate all differences $d_i = y_i - x_i$, from them mean and standard deviation.

The left-side s_d -formula yields a more exact result, whereas the right-side s_d -formula

is more convenient. Then calculate teststatistic t and degrees of freedom df .

$\bar{d} = \frac{\sum d_i}{n}$	$s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{(\sum d_i^2) - n \cdot \bar{d}^2}{n-1}}$
$t = (\bar{d} / s_d) \cdot \sqrt{n}$	$df = n - 1$

Step 3: Statement: Seek critical value $t(\alpha, df)$ from t-table (pay attention to 2- or 1-sided)

2-sided test: If $t < -t(\alpha, df)$, then is significantly $\mu_y < \mu_x$ or $d < 0$

If $t > t(\alpha, df)$, then is significantly $\mu_y > \mu_x$ or $d > 0$

1-sided test: e.g. $H_A: \mu_y > \mu_x$ or $H_A: d > 0$

If $t > t(\alpha, df)$, then is significantly $\mu_y > \mu_x$ or $d > 0$

in all other cases we accept $H_0: \mu_y = \mu_x$ (no significant difference)

11.5 Paired Rank-test of Wilcoxon

(Matched-pairs signed-ranks test) Comparison of the means using a **not normally distributed** correlated sample. We have the same data situation as in the case of the paired t-test (11.4).

Step 0: Hypothesis $H_0: \mathbf{d} = 0$ $H_A: \mathbf{d} \neq 0$ (2-sided test) $\alpha=0.05$ (5%)
 or e.g. $H_A: \mathbf{d} > 0$ (1-sided ">test")

Step 1: Wilcoxon-test (Rank test)

Step2: The differences $d_i=y_i-x_i$ are calculated and then ascendingly ordered and ranked by their absolute values. Two or more numbers d_i of equal absolute value get the average of their ranks, e.g., ...25, **27, 27, 27**, 29 ... , or e.g., ...,25, **26.5, 26.5**, 28, ... Differences $d_i=0$ we eliminate and we decrease the number n of pairs accordingly. Now each rank number gets the sign of its original difference d_i , and we sum separately for each sign. R_N is the sum of the negatively signed ranks, R_P is the sum of the positively signed ranks.

If $n \leq 25$ one calculates $U = \text{Min}(R_N, R_P)$

If $n > 25$ one calculates $u = \frac{U - (n(n+1)/4)}{\sqrt{n(n+1)(2n+1)/24}}$

Step 3: Critical value $U_\alpha = U(\alpha, n)$ we find e.g. in E. Weber, Tab. 25.

2-sided test: If $U = R_P < U_\alpha$, then is significantly $\mu_Y > \mu_X$

If $U = R_N < U_\alpha$, then is significantly $\mu_X > \mu_Y$

1-sided test : If $U = R_P < U_{\alpha/2}$, then is significantly $\mu_Y > \mu_X$

in all other cases accept $H_0: \mu_X = \mu_Y$ (no significant difference).

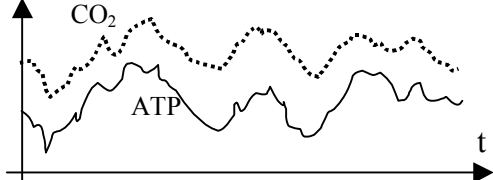
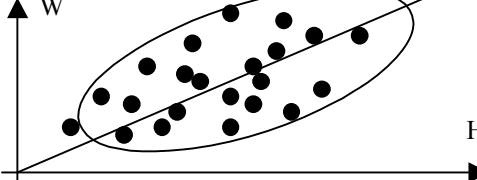
In the case of $n > 25$ we compare u with the 2-sided critical value of the Normal distribution, $u_\alpha = 1.96$, (1-sided $u_\alpha = 1.65$: both under $\alpha = 0.05$).

12. Correlation and Regression Analysis

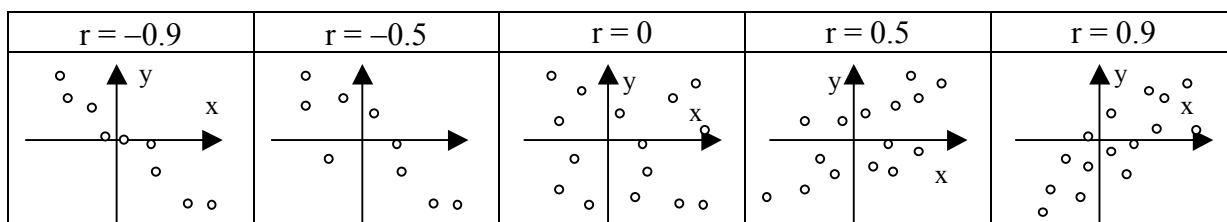
Correlation analysis and simple regression analysis are important methods for a pair of two metric variables. We use the **correlation analysis** if a pair of observed or measured variables are in some relation to one another, and none of the variables can be qualified to be directly dependent on the other. For example: Blood pressure P_1 measured on wrist and blood pressure measured at the aorta. One can not say that one of the variables P_1 or P_2 depends on the other, but both variables depend surely on one or more third variables (**factors**). We use simple **regression analysis** if definitely a dependent and an independent variable exists. For example: The blood pressure P_1 measured at the wrist depends surely on the dose x of some anti-hypertensive drug administered to the patient.

12.1 Correlation Coefficient by Bravais-Pearson

Equal or similar behaviour of two variables is termed *correlation*, whereat a direct dependency of the variables is not a presupposition. Tidal correlation can be imagined without any dependency if one is considering the social and cultural evolutions at isolated continents.

Tidal correlation between ATP- and CO ₂ -production of <i>Candida saccharomyces</i>	Product-Moment Correlation between height H and weight W
	
The ATP-production of yeast cells and the CO ₂ -outcome of the fermentor have a similar course if one records it over the time. High values of ATP correlate with high values of the CO ₂	Great probands have, in average, a greater weight than small probands, but there does not exist a slavish dependency. The correlation ellipse is a line of altitude of the 2-dimensional density distribution of the points.

The correlation coefficient r is normed so that it falls in the interval $[-1, +1]$. The value $r = 1$ has the meaning that we have an exact linear dependency of the form $y = a + b x$ (or $x = c + d y$) between the two variables x and y without any deviation. A value of $r = -1$ has the meaning that this exact dependency has the same linear form, but with a negative slope, i.e., $y = a - b x$ or $y = c - d y$. The quantities a, b, c, d are constants. The figures below show data points with different values of the correlation coefficient.



12.2 Linear correlation coefficient r

The Product-Moment-Correlation coefficient by Bravais and Pearson describes the linear correlation between two metric variables in a population. A sample is given with n matched pairs of data, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. We can freely decide which variable is called x , or which is called y , respectively. The first one calculates the sums of squares, SAQ_{xx} and SAQ_{yy} , and the sum of products, SAP_{xy} . Here the left hand formula is more exact, the right hand formula is faster to compute. One uses the following scheme if the pocket calculator is not programmed to yield the correlation coefficient. Calculate the 5 sums $\sum x_i, \sum y_i, \sum x_i^2, \sum x_i y_i$, and $\sum y_i^2$, and from these sums one calculates the values of $SAP_{xy}, SAQ_{xx}, SAQ_{yy}$ with the right hand formulas. Attention, please!! Don't round the means. 6 significant digits are needed to avoid strong rounding errors.

Nr	x_i	y_i	x_i^2	$x_i y_i$	y_i^2
1	x_1	y_1	x_1^2	$x_1 y_1$	y_1^2
2	x_2	y_2	x_2^2	$x_2 y_2$	y_2^2
...
n	x_n	y_n	x_n^2	$x_n y_n$	y_n^2
	$\sum x_i$	$\sum y_i$	$\sum x_i^2$	$\sum x_i y_i$	

$$SAP_{xy} = \sum_{i=1}^n ((x_i - \bar{x})(y_i - \bar{y})) \quad \text{bzw.} \quad SAP_{xy} = \left(\sum_{i=1}^n x_i y_i \right) - n \cdot \bar{x} \bar{y}$$

$$SAQ_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

$$SAQ_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$\hat{r} = \frac{SAP_{xy}}{\sqrt{SAQ_{xx} \cdot SAQ_{yy}}}$$

$$\text{bzw. } SAQ_{xx} = \left(\sum_{i=1}^n x_i^2 \right) - n \cdot \bar{x}^2$$

$$\text{bzw. } SAQ_{yy} = \left(\sum_{i=1}^n y_i^2 \right) - n \cdot \bar{y}^2$$

$$t = \frac{\hat{r}}{\sqrt{1 - \hat{r}^2}} \sqrt{n - 2} \quad \text{df} = n - 2$$

\hat{r} is an estimation of the correlation coefficient r of the population. The Null Hypothesis is $H_0: r=0$ (no linear correlation in the population), $H_A: r \neq 0$ (linear correlation is assumed in the population). The linear correlation r in the population is significantly different from zero if $|t| \geq t(\alpha, df, 2\text{-sided})$ for a two-sided test. Otherwise, one accepts $H_0: r = 0$, i.e., "no significant linear correlation in the population."

Numerical example for the correlation coefficient: The oxygen concentration y [mg/l] was measured in a fermenter together with the air flow x [m³/h]. Our scheme for the sums is:

Nr	x	y	x ²	xy	y ²
1	50	1.3	2500	65	1.60
2	110	1.9	12100	209	3.61
3	110	2.1	12100	231	4.41
4	300	3.7	90000	1110	13.69
5	370	5.1	136900	1887	26.01
Σ	940	14.1	253600	3502	49.41

$$\bar{x} = 188, \quad \bar{y} = 2.82, \quad SAQ_{xx} = 253600 - 5 \cdot 188^2 = 76880, \quad SAQ_{yy} = 49.41 - 5 \cdot 2.82^2 = 9.648,$$

$$SAP_{xy} = 3502 - 5 \cdot 188 \cdot 2.82 = 851.2,$$

$$\hat{r} = 851.2 / (76880 \cdot 9.648)^{0.5} = 0.98834$$

$$H_0: r=0, \quad H_A: r \neq 0, \quad \alpha=0.05$$

$$t = (0.98834 / (1 - 0.98834^2)^{0.5}) \cdot 3^{0.5} = 11.22$$

$$df = 5 - 2 = 3$$

$$t_\alpha = t(\alpha=0.05, df, 2\text{-sided}) = 3.18$$

correlation coefficient

hypotheses

t-statistic

degrees of freedom

critical values of t-distribution

We accept H_A , i.e., oxygen concentration and air flow are (highly) correlated in the fermenter.

12.3 Simple linear regression analysis, fitted line

Linear regression analysis attempts to model the relationship between two variables by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y . The values of the variable x are assumed to be error free. The values of y are assumed to scatter with a normal distribution $N(0; \sigma_{\text{Rest}}^2)$ around the fitted line independently on x . If variable x is the time then we say also **trend analysis**.

$$\text{Regression model} \quad y_i = \mathbf{a} + \mathbf{b} x_i + e_i$$

Sought are estimations of the regression constant \mathbf{a} and regression coefficient \mathbf{b} in the population. A sample is given with the n matched pairs $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ similarly to the correlation coefficient data. Dependent variable \mathbf{y} is called also *target variable*, independent variable \mathbf{x} also *predictor variable*. The deviation e_i is called also *residuum* or *error*. Index i is the case number (point number, patient number). Regression constant \mathbf{a} is the expectation

of the dependent variable y at point $x=0$. Regression coefficient b is the *slope* of the straight line. The coefficients a and b are estimated by the *Least Squares Method*, i.e., the sum of the error squares is minimized, $\sum e_i^2 = \text{Minimum}$. The first we compute SAP_{xy} , SAQ_{xx} , SAQ_{yy} as in the case of the correlation coefficient, then we compute

$$\begin{aligned} \hat{b} &= SAP_{xy} / SAQ_{xx} && \text{estimates the regression coefficient } \mathbf{b} \\ \hat{a} &= \bar{y} - \hat{b} \cdot \bar{x} && \text{estimates the regression constant } \mathbf{a} \\ \hat{y}_i &= \hat{a} + \hat{b} \cdot x_i = \bar{y} + \hat{b} \cdot (x_i - \bar{x}) && \text{estimates } \mathbf{y} \text{ at point } x_i \text{ (expectation value)} \\ \hat{e}_i &= y_i - \hat{y}_i && \text{estimates the error } \mathbf{e}_i \text{ at point } x_i \\ \hat{S}_R &= \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SAQ_{yy} - \hat{b} \cdot SAP_{xy}}{n-2}} = \sqrt{\frac{\sum \hat{e}_i^2}{n-2}} \end{aligned}$$

\hat{S}_R estimates the mean error σ_R in the population (residual standard deviation of the points around the straight line (parallel to y-axis measured)). The formula using SAQ_{yy} and SAP_{xy} is the best for pocket calculators.

$$\begin{aligned} df &= n-2 && \text{Degrees of freedom of } \hat{S}_R \\ S_b &= \hat{S}_R / \sqrt{SAQ_{xx}} && \text{Estimation error of regression coefficient } \mathbf{b} \end{aligned}$$

Estimation error of regression constant \mathbf{a}	Estimation error of expectation value \hat{y}_i
$S_a = \hat{S}_R \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SAQ_{xx}}}$	$S_{\hat{y}} = \hat{S}_R \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{SAQ_{xx}}}$

$$t_a = \hat{a} / S_a \quad \text{with } df = n-2 \text{ tests } H_0: \mathbf{a}=0 \text{ and } H_A: \mathbf{a} \neq 0 \text{ (2-sided)}$$

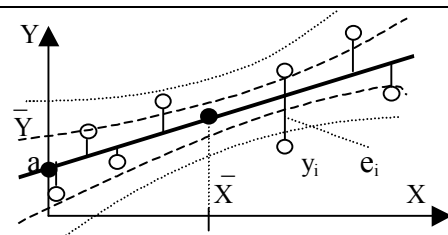
$$t_b = \hat{b} / S_b \quad \text{with } df = n-2 \text{ tests } H_0: \mathbf{b}=0 \text{ and } H_A: \mathbf{b} \neq 0 \text{ (2-sided)}$$

Using pre-knowledge, we can perform both tests also 1-sided. A significant $\mathbf{a} \neq 0$ means, that target variable y has a value $y \neq 0$ at point $x=0$. A significant slope $\mathbf{b} \neq 0$ means that predictor variable x has some direct or indirect influence at the target variable y , i.e., the slope is not by chance.

$$\hat{y}_i \pm t(\alpha, df, 2\text{-tailed}) \cdot \hat{S}_R \cdot \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{SAQ_{xx}}} \quad \text{Confidence interval of the "true regression line"}$$

$$\hat{y}_i \pm t(\alpha, df, 2\text{-tailed}) \cdot \hat{S}_R \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SAQ_{xx}}} \quad \text{Confidence interval of prediction}$$

The graphik shows the regression line in the X-Y-coordinate system. It crosses point \mathbf{a} at the Y-axis and the point (\bar{x}, \bar{y}) . The measurements y_i are given by small circles, the residuals e_i by short slashes. The confidence interval of the true regression line is dashed, that of the prediction is dotted.



Drawing a great number of samples of size n , one gets n fitted lines with n different slopes. The confidence interval of the "true regression line" contains with probability $1-\alpha$ the true regression line of the population. For the **prediction** of y -values at x -points not used in the

sample we are interested in the expected error of the prediction. The confidence interval of prediction contains with probability $1-\alpha$ all observations, i.e., those of the given sample and those in future. Outside the x -range given by the sample the confidence intervals grow dramatically. That forbids bold predictions into future, e.g.

The regression model assumes that the data fulfill the following conditions:

1. The simple linear model $y_i = a + b x_i + e_i$ is valid in the population
2. The measuring points scatter around the fitted curve for each x -value with normal distribution $N(\mu=0; \sigma=\hat{S}_R)$.

Numerical example fitted straight line: The oxygen concentration y [mg/l] was measured inside a fermenter depending on the air flow x [m³/h]. The computing scheme is identical with the numerical example for the correlation coefficient. There we find the numbers:

$$\bar{x}=188, \quad \bar{y}=2.82, \quad SAQ_{xx}=253600-5*188^2=76880, \quad SAQ_{yy}=49.41-5*2.82^2=9.648, \\ SAP_{xy}=3502-5*188*2.82=851.2,$$

$\hat{b}=851.2/76880=0.0110718$ [mg/l / m ³ /h]	Slope of the straight line
$\hat{a}=2,82-0.01107*188=0.7388$ [mg/l]	Regression constant
$\hat{S}_R=(9.648-0.0110718*851.2)/(5-2)^{0.5}=0.273$ [mg/l]	Residual standard deviation
df=5-2=3	Degrees of freedom of \hat{S}_R
t ($\alpha=0.05$, df, 2-sided) = 3.18	Critical values of the t-distribution

$\hat{y}_{x=500}=0.7388+0.01107*500=6.2738$ [mg/l]	Expectation at $x=500$
[m ³ /h]	
$6.27 \pm 0.273 * (1/5 + (500-188)^2/76880)^{0.5} * 3.18 =$	95%-confidence interval of the
6.27 ± 1.055 , gerundet 6.27 ± 1.0	"true regression line" at $x=500$
$6.27 \pm 0.273 * (1 + 1/5 + (500-188)^2/76880)^{0.5} * 3.18 =$	95%-confidence interval for the
6.27 ± 1.368 , gerundet 6.27 ± 1.4	predictions at $x=500$

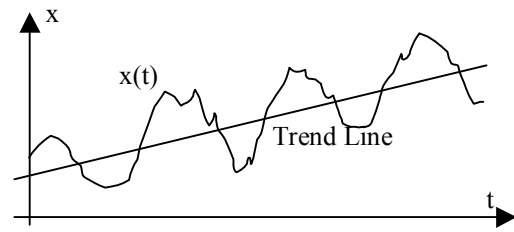
$H_0: \mathbf{b}=0$ and $H_A: \mathbf{b}\neq 0$ (2-sided), $\alpha=0.05$	Pair of hypotheses for slope \mathbf{b}
$t=0.0110718*(76880)^{0.5}/0.273=11.24$	t-statistic for slope \mathbf{b}
Da $ t \geq 3.18$, akzeptieren wir H_A	Choice of H_0 or H_A
Slope \mathbf{b} of the regression line of the population is significantly different from zero. We assume a significant interdependence between oxygen concentration y [mg/l] and air flow x [m ³ /h].	

$H_0: \mathbf{a}=0$ and $H_A: \mathbf{a}\neq 0$ (2-sided), $\alpha=0.05$	Pair of hypotheses for constant \mathbf{a}
$t=0.7388 / (0.274 * (1/5 + 188^2/76880)^{0.5}) = 3,320$	t-statistic for constant \mathbf{a}
Because $ t \geq 3.18$, we accept H_A	Choice of H_0 or H_A
The regression constant \mathbf{a} of the population is significantly different from zero (at 5%-level). Also if air flow is zero ($x=0$), we can expect an oxygen concentration $y \neq 0$.	

This example shows clearly how important can be the significance level: The test result for constant \mathbf{a} is rubbish!! Each biologist knows we will find no free oxygen in the fermenter for zero air flow. Because of $t(\alpha=0.01, df=3, 2-sided)=5.84$, the constant \mathbf{a} is not significant at the $\alpha=1\%$ -level. One should remove the constant from the model, and work with the simpler model $y_i = b x_i + e_i$. This is the model of a regression line through the origin of the coordinate system. EXCEL and DASYS, e.g., offer this model. One wins one degree of freedom of the residual standard deviation, since only the slope \mathbf{b} if the line is estimated from the data.

12.4 Time series

The graphic shows a trend line superposed by a simple periodical oscillation. Such curves we find, e.g., studying the plankton growth under the influence of the tides (ebb and flood caused by Sun and Moon). We model periodical oscillations, e.g., by Sine waves with a phase shift.



If time is the influencing independent variable then we speak from time series. The influence of other variables exist, but we don't model it directly. Mostly, one divides the time series into a linear trend and a number of periodical oscillations (saisonal pattern around the trend line). The theories differ in the modelling of the periodical oscillations, caused by daily rhythms, monthly rhythms, Lunar rhythms and so on. Each theory fills a book.

Autocorrelation function

The autocorrelation function $Ac(LAG)$ is produced if one correlates a time function $x(t)$ with itself. The “copy” of $x(t)$ is shifted more and more in relation to the original curve $x(t)$. For each single lag the linear correlation coefficient r is computed anew. All values of r make together the autocorrelation curve over the lag, $Ac(LAG)$. Hereby LAG is the temporal lag between $x(t)$ and the shifted copy, $x(t-LAG)$.

If $x(t)$ is given in the time interval $[t_1, t_2]$, then the LAG can be maximally $t_2 - t_1$, since otherwise one will find no pairs of values. Practically, one computes $Ac(LAG)$ in the interval $[-(t_2 - t_1)/2, +(t_2 - t_1)/2]$. $Ac(LAG)$ is an even function, that is we have the relation $Ac(LAG) = Ac(-LAG)$. The LAG of the maximum value is the period τ .

Cross correlation function

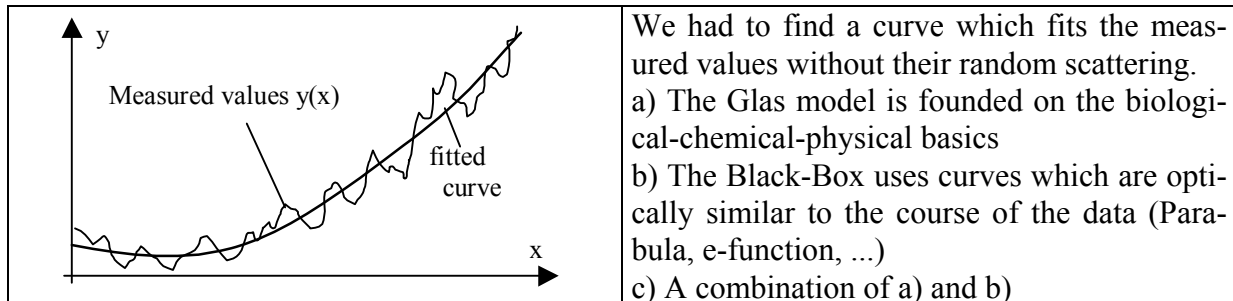
The cross correlation function $Cc(LAG)$ is produced if one correlates two time functions, $x(t)$ and $y(t)$, respectively, and one shifts $y(t)$ laterally in relation to $x(t)$. For each single lag the linear correlation coefficient r is computed anew. All values of r make together the cross correlation curve over the lag, $Cc(LAG)$. Hereby LAG is the temporal lag between $x(t)$ and the shifted curve $y(t-LAG)$.

If $x(t)$ and $y(t)$ are given in the time interval $[t_1, t_2]$, then the LAG can be maximally $t_2 - t_1$ (positive LAG or shift to the right of $y(t)$ in relation to $x(t)$) or otherwise the LAG can be minimally $t_1 - t_2$ (negative LAG or shift to the left of $y(t)$), since otherwise one will find no pairs of values. Practically, one computes $Cc(LAG)$ in the interval $[-(t_2 - t_1)/2, +(t_2 - t_1)/2]$. $Cc(LAG)$ is in general not an even function, that is we have nearly always the relation $Cc(LAG) \neq Cc(-LAG)$.

One can find from the position of the maximum value of the function $Cc(LAG)$ the temporal displacement between $x(t)$ and $y(t)$. If a pattern of the $y(t)$ curve repeats laterally displaced by the time difference dt in the $x(t)$ curve, then the correlation coefficient grows. The lag of the maximum value of the $Cc(LAG)$ curve is the seeked displacement dt . A positive value of dt (i.e. a $(LAG > 0)$) means that $y(t)$ runs temporally behind $x(t)$. One has to shift $y(t)$ to the left to

bring $y(t)$ and $x(t)$ in coincidence. Otherwise, a negative dt (i.e. a $(LAG < 0)$) means that $y(t)$ runs temporally before $x(t)$.

12.5 Nonlinear Regression



Computing the curve fit we distinguish between linearized model, quasi-linear model, and nonlinear model

Example Linearization of the Exponential Function: The growth of yeast, or growth in general, can be described in its starting phase by the Exponential function $Z(t) = Z_0 e^{\alpha t}$.

The growth coefficient α has the dimension $[h^{-1}]$. Z_0 is the starting mass of yeast at $t=0$. Logarithmic transformation of the modelling equation yields $\ln(Z) = \ln(Z_0) + \alpha t$. By renaming $y = \ln(Z)$, $a = \ln(Z_0)$ and $b = \alpha$ we get the simple linear regression model $y = a + bt$. One estimates the two coefficients a and b as described in the chapter simple linear regression, and gets by backward transformation $Z_0 = e^a$ and $\alpha = b$ the wished estimations of the coefficients of the nonlinear model. The error of the slope, s_b , we can interpret directly (with some restrictions) as error s_α of the growth coefficient α . The error of the constant, s_a , of the logarithmic transformed model $y = a + bt$ is now a multiplier of the original coefficient Z_0 , i.e., $Z_0 + s_Z = Z_0 * e^{s_a}$, and $Z_0 - s_Z = Z_0 / e^{s_a}$, respectively. But pay attention:

- The curve found in this way does minimize the sum of squared errors only in the transformed model, but not in the original data. Otherwise, in general, the differences are minimal.
- The test of the hypotheses is correct only in the transformed model. Otherwise, in general, the differences are minimal.

Quasilinear Models: One replaces predictor variable x by one or more functions of x . Each function is represented in the new quasilinear model by an own variable. All new variables construct the multiple quasilinear regression model:

The polynomial, e.g.,	$y = a + b t + c t^2$	we replace by
the quasilinear model	$y = b_0 + b_1 X_1 + b_2 X_2$	with $X_1 = t$ and $X_2 = t^2$

Another model of a quasilinear model is:

	$y = b_0 + b_1 \sin(c * x) + b_2 \cos(c * x) + b_3 x + b_4 e^{dx}$
with	$X_1 = \sin(x)$
and	$X_2 = \cos(x)$ and so on,

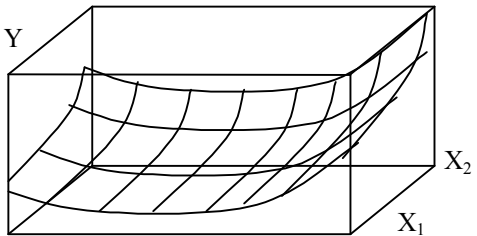
with the fixed constants c and d . Their values must be known. The multiple linear regression analysis can estimate only the values of the so called "linear" coefficients b_0, b_1, b_2, \dots . Also here the problem arises that the sum of squared errors is not minimized in all cases. One reason is, for example, that the terms x and x^2 of an quadratic model are highly correlated.

Pay attention!! The use of polynomials of higher order is dangerous!! If a sample is given with n points then a polynomial of order $n-1$ fits *always* all points exactly. A straight line fits always 2 points, a parabola 3 points, a polynomial of 9-th order fits 10 points. But do'nt ask what the polynomial is making between the points. Very oftenly you will find there mad values. Therefore the following **recommendation**: If n is the number of measured points in the sample, then the number, p , of coefficients used in the model should not exceed $n/2$, i.e., the number of points, n , should be twice as large as the number of coefficients, p . Using stepwise regression, or CWA-regression in DASY, the initial number of coefficients can be any number. Here the algorithm finds the proper number of coefficients (and variables) used in the finally resulting model

Nonlinear Models: Solver, as working in EXCEL, e.g., can fit any curve to any data. Here we must not dicriminate between linear and nonlinear coefficients. But in some cases the solver can fail in finding a correct solution. In this cases the user should change the starting values of the coefficients. Most of the programs yield also the standard errors of the coefficients. These estimations one should use cautiously, only as rough estimations.

<p>The Haldane kinetics is a good example of a nonlinear model. The kinetics describe the relative growth, μ, depending on the substrate concentration, c_s. The model uses the three coefficients μ^*, K_S and K_{SI}. The solver estimates the three coefficients, seeking the best fit for the given data, μ and c_s. The best fit is characterized by a minimum sum of squared errors (squared residuals).</p>	$\mu(c_s) = \frac{\mu^* \cdot c_s}{K_S + c_s + \frac{c_s^2}{K_{SI}}}$
--	---

12.6 Multiple Regression

<p>Multiple Regression estimates the target variable, Y, from p descriptor variables, X_1, X_2, \dots, X_p. The model can or can not include a regression constant b_0. The common model is:</p> $Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \dots + b_p \cdot X_p + e$ <p>Die geometrical interpretation is a function</p>	
--	--

over the space put up by the variables X_1, X_2, \dots, X_p . The software (e.g. Excel, DASY, ...) estimates the regression coefficients b_1, b_2, \dots, b_p (and b_0 if demanded) using the Least-Squares method ($\Sigma e^2 = \text{Minimum}$). Residuuum e is assumed to be a random error (deviation between measuring value and expectation). Exampel **Parametric Price fixing**: A company wants to estimate the achievable price of a new electrical motor. For this task one needs technical data of the new motor, as weight, electrical power, revolutions per minute, voltage. Additionally, one needs the data of already existing motors together with their market prices. Target variable is the market price. Influencing variables (descriptor variables) are the technical data. The multiple linear model is, e.g.:

$$\text{Price} = b_0 + b_1 \cdot \text{Weight} + b_2 \cdot \text{Power} + b_3 \cdot \text{Revolutions} + b_4 \cdot \text{Voltage} + e$$

The multiple regression analysis estimates the regression coefficients from the given data and prices. To get the estimated market price of the new motor one calculates $\text{Price} = b_0 + b_1 \cdot \text{Weight} + b_2 \cdot \text{Power} + b_3 \cdot \text{Revolutions} + b_4 \cdot \text{Voltage}$ with the technical data of the new motor.

For **Quasilinear Models** see "nonlinear regression". One speaks from **Weighted Regression** if each data point has its own weight G_i . The number of degrees of freedom will in most statistical programs not be changed by the weights.

Multiple regression can solve three main tasks:

1. **Prediction** (forecasting) of Y-values outside the region given by the X-values of the sample and/or for new points inside the given X-region. For example, we want to estimate the energy costs for the next year on the base of planned production numbers.
2. **Reproduction** of the Y-values exactly at the points of the sample (base points). The effect here is a pure data reduction (only a few regression coefficients instead of a greater number of base points). Example: The formula $t = b_0 + b_1*(1/FG) + b_2*(1/FG^2)$ needs only three coefficients, b_0, b_1, b_2 , to yield all critical values of the t-distribution for a fixed alpha and $df > 10$.
3. **Search for significant predictor variables X**: Example: Which factors are important for the yield of a newly cultivated sort of rapeseed (temperature? quantity of rain water? lime? nitrogen? ...?)

We recommend different regression algorithms accordingly to the three main tasks of regression analysis:

- The **CWA algorithm** in DASYS for prediction (forecasting) with a great number of predictors, only few data points, and strongly scattering values of the target variable Y.
- The **Stepwise algorithm** (forward or backward) for prediction (forecasting) with a moderate number of predictors, and/or few data points, and/or moderate scattering values of the target variable Y.
- **Regression with all predictors** for the exact reproduction of the Y-values at the base points of the sample.
- The **Stepwise algorithm** (forward or backward) for the search for significant predictors.

In the case of **prediction** (forecasting) one is interested in the precision of the predicted Y-values. Not the smallest residual standard deviation is the question, but the smallest error of prediction. To estimate this error of prediction one uses **Lachenbruch-Jackknife**- and/or **Bootstrap**-methods and/or working samples (in DASYS Lachenbruch-Jackknife and working sample only). The significance of the predictors is not of interest here.

In the case of **exact reproduction** (data reduction) one needs an especially well defined regression model which minimizes the residual standard deviation with only few predictors. An unhappy choice of the model can cause phantasy-values for Y if one deviates only a bit from the base points (X-values of the sample). The error of forecast or the significance of the predictors is not of interest here.

The exploration of **significant predictor variables (influencing variables)** has oftenly some scientific or practical meaning. The result can outline causal associations between the dependent variable and predictor variables, or, otherwise, it can minimize the work load of data collection for prediction. Issues during the search for significant predictor variables are:

1. The algorithms select only one variable from each subset of highly correlated influencing variables. The other variables of the subset can have about the same influence, as the selected variable. The existence of such subsets can be an indication for hidden factors controlling the measured variables. More seldom, the factor itself appears as a variable. It would make sense to run a factor analysis before regression analysis (in DASYS not able), and then to use the factors as influencing variables.

- In the case of many influencing variables one can use an Alpha-adjustment to ensure that the selected variables are significant indeed, e.g. the Bonferroni Adjustment or Holmes Sequential Procedure (DASY allows for variable selection in regression analysis the Bonferroni- adjustment only). Otherwise, random variables without information would have a chance to be selected as significant influencing variables. E.g., without Alpha-adjustment, we would select about 5 random variables as significant in the case of 100 variables and a given Alpha of $\alpha=5\%$.

CWA-Regression is a method computing the regression coefficients iteratively by a special gradient algorithm minimizing the residual variance of prediction (Cierzynski / v.Weber 1989). Advantages are:

- Highly correlated variables does not exclude one another from selection, but are averaged to a kind of factor (one avoids factorial regression)
- The iteration stops if the error of prediction has reached its minimum
- We get a "robust solution" which holds also under moderate changes of the data base (changes of the X-values)

Stepwise Regression (Forward Selection, Backward Selection or combined) : A significance test (t-test, F-test) makes the decision whether a variable is selected or not for the regression model. Advantages are:

- Predictor variables with a significant influence on the reduction of the residual variance only are selected for the model (Exception in DASY: If no variable is significant then that variable is selected with the highest t-value)
- A group of highly correlated variables is represented in the model by one variable only.
- We get a "robust solution" which holds also under moderate changes of the data base (changes of the X-values)

Regression with all predictor variables" is a method which removes predictor variables from the model only if strong linear dependence between predictor variables causes numerical difficulties. The advantage of the regression with all variables is that for the base points (and only for them) the residual variance can be minimized. It depends strongly on the regression model whether the estimation of the target variable yields still reasonable numbers outside the base points. Check your model in DASY with the jackknife feature. If the jackknife-based estimation of the residual standard deviation is much higher than the common estimation of the residual standard deviation then the model is not appropriate.

Polynomial Standard Models: *Simple Polynomial model:* If the wished order of the polynomial is 2 then for each given predictor variable X_i of the model we calculate an additional variable X_i^2 . If the wished order is 3 then additional we calculate X_i^2 and X_i^3 , and so on (in DASY up to order of 5). *Full Polynomial model:* Works as simple polynomial, but additional are calculated all products (interactions) of the predictor variables of the model. A full polynomial model with a order of 2 and original two predictor variables X_1, X_2 includes additional the new variables $X_1^2, X_2^2, X_1 * X_2$, with a order of 3 the new variables $X_1^2, X_1^3, X_2^2, X_2^3, X_1 * X_2, X_1^2 * X_2, X_1 * X_2^2$.

Explanation of quantities and results of the multiple regression:

Y	Target variable, dependent variable, response variable
X _j	Predictor variables (influencing variables) (j = 1, 2, 3,...p) p=number of predictors
n	Number of data points (cases, probands, rows) without missing value rows
B = R ²	R-square, R ² , or the coefficient of determination B, a measure of the improve of prediction by knowledge of X_1, X_2, \dots, X_p . B=SAQ _{Reg} / SAQ _{Rest} has values of $0 \leq B \leq 1$. Here SAQ _{Reg} is the sum of squares

	$(\sum (\hat{y}_i - \bar{y})^2)$ and SAQ_{Rest} is the sum of squares Σe^2 .
F	F-test statistic to test coefficient B. The null hypothesis $H_0: B=0$ (No relationship between dependent variable Y and predictors X_j) with $F=B(n-k)/(1-B)$ and $df_1=p$ and $df_2=n-k$, k =number of coefficients in the model including coeff. b_0 .
df	$df=N-k$, degrees of freedom of the residual standard deviation S_R
KIW(B)	Error probability rejecting the null hypothesis $H_0: B=0$ (one-sided F-test)
b_j	Coefficient Numerical value of the estimated regression coefficient
s_{b_j}	Standard error of the estimated regression coefficient
t_j	t-statistic t-distributed test statistic to test the null hypothesis $H_0: b_j = 0$ (the coefficient b_j is zero in the population)
p-value	Error probability rejecting the null hypothesis $H_0: b_j = 0$ (2-sided t-test)
S_R	Residual standard deviation or mean residuum (mean error e)
S_j	Mean error of prediction e estimated by the Lachenbruch-Jackknife Method
S_w	Mean error of prediction e estimated by a working sample

Multiple linear Regression with EXCEL: Model $Y = b + m_1X_1 + m_2X_2 + \dots + m_qX_q + e$

Here Y is the target variable, and X_1, \dots, X_q are the q predictor variables, e is the residuum (deviation, error), b ist die regression constant, m_1, \dots, m_q the regression coefficients.

The right-side table shows a part of an EXCEL-table with columns A,B,C,... and rows 1,2,... We want to start a regression without regression constant b. The model equation is:
 $P = m_1C + m_2V + m_3W$

We block out (select, make black) a matrix with (always) 5 rows (here starting with row 9) and so many columns as coefficients to calculate (here 3). If selected, we had to count constant b too. We type the following statement using the keyboard (German version) :

`=rgp(a2:a7;b2:d7:falsch;wahr)` and the triple key combination Strg-Shift-Enter.

	A	B	C	D
Row1	Price	Cyc/s.	Voltage	Weight
Row2	1400	1400	380	240
Row3	3800	2000	600	900
Row4	1850	2800	380	180
Row5	4450	12000	380	95
Row6	5900	1200	600	1800
Row7	22500	600	15000	3250
Row8				
Row9	2,796	0,881	0,323	
Row10	0,0549	0,0139	0,008	
Row11	0,999	112,6		
Row12	3		
Row13		

Cells *a2:a7* contain the *Price* data, cells *b2:d7* contain the data of the 3 predictor variables, *falsch* (false) says that the model includes no constant b, *wahr* (true) says that EXCEL calculates besides the coefficientsl also additional statistical quantities (errors of the coefficients, e.g.) See also the HELP-feature of EXCEL. EXCEL computes the coefficients in row 9 in the order m_3, m_2, m_1 . In row 10 we find the errors s_{m3}, s_{m2}, s_{m1} of the coefficients. In row 11 we find the coefficient of determination B (or R^2) and the residual standard deviation (mean error) S_R . In row 12 we find the degree of freedom df of S_R . We can use S_R and df for testing. The other numbers are not of interest here.

13. ANOVA – Analysis of Variance

13. One way Analysis of Variance with comparison of means

Analysis of Variance (ANOVA) is a statistical method to analyze grouped data, which can be metric or rank ordered. Grouping is done by a categorical or nominal variable, oftenly termed as **factor** in literature. The *levels of the factor* variable are termed also as *treatments*, and are mostly integers in the computer (treatment 1, treatment 2, ...). The numbers of the target variable (metrical or rank ordered) for one factor level (for one treatment) construct one group. Literature (**EISENHART**) discriminates between two models of analysis of variance.

Model 1 (Fixed Model): The groups are determined by an experimental design. One is interested in differences between group means. For example: The tensile strength of a fabric [N/m] exposed to UVB radiation for a time of 14 days depending on different surface coverings with aluminum oxide. Target quantity is the tensile strength, factor is the density of covering with the levels 1="uncovered", 2="8 g Al/m²", 3="16 g Al/m²".

Model 2 (Random Model): Grouping is observed, not controlled, i.e., by chance. One is interested in the variability of the observed data, i.e., in the question whether the data inside the groups have a lower or higher variance than the group means. For example: A cattle of cows is divided in groups. Grouping factor is the "father". Cows from the same father construct a group. The factor "father" has the levels 1="Anton", 2="Bogumil", 3="Caesar, ... Target quantity (items of target variable) is the annual milk production of each cow. If the numbers scatter less within groups than between group means, then a significant genetic influence of the fathers is supposed. One speaks of the estimation of heritability of the "milk-production genome".

Global Test: We ask if the averaged variance between groups is significantly greater than the pooled variance within groups, i.e., is there a significant influence of the factor?

X_{tot} = Total mean of the target variable calculated from all n cases (items)

X_i = Group means, $i=1,..g$, g =number of groups

SAQ_{tot} = Sum of squared deviations of target data from X_{tot}

SAQ_{in} = Sum of squared deviations of target data from their group mean X_i summarized over all g groups

SAQ_{bet} = $SAQ_{tot} - SAQ_{in}$, sum of squared deviations between groups

MQ_{bet} = SAQ_{bet}/FG_{bet} Averaged sum of squares with degree of freedom $df_{bet} = g - 1$

MQ_{in} = SAQ_{in}/FG_{in} Averaged sum of squares with degree of freedom $df_{in} = n - g$

F_{gl} = MQ_{bet}/MQ_{in} global F-statistic with degrees of freedom (df_{bet} , df_{in})

KIW_{gl} = Error probability rejecting the global null hypothesis. The null hypothesis supposes no influence of the factor on the population. The F-test is performed one-sided.

Comparisons of means are performed pairwise, i.e., we compare each X_i with each X_j . Here we can test one- or two-sided. The number of pairwise single hypotheses is $h = g(g - 1)/2$. Each single test statistic is $t_{ij}^2 = F_{ij} = ((X_i - X_j)^2 * n_1 * n_2) / (MQ_{in} * (n_1 + n_2))$ with degrees of freedom df_{in} if t-test is used, and (1, df_{in}) if F-test is used.

KIW_{ij} = Error probability rejecting the single null hypothesis $\mu_i = \mu_j$. We compare the KIW_{ij} values with an adjusted α^* . DASY uses here Holmes Sequential Procedure. Mostly (also in DASY) the F-test is replaced by a two-sided t-test with degree of freedom df_{in} .

A special case: PERLI proved that one can replace the first pairwise test (Maximal $F = F_1$ or maximal $t = t_1$) by the global F-test. If the global F-test is significant then the most highly evaluated mean difference is assumed to be significant also, also if Holmes procedure would accept the null hypothesis here.

13.2 ANOVA – Cross classification with comparison of means

We consider here only the cross-classification model with fixed effects. The observations are classified here by 2 variables. In our example we have $r = 4$ rows (4 A-classes) and $s = 3$ columns (3 B-classes). The data were collected without repetitions, i.e., each cell is occupied by one number only. The 4 A-classes are 4 agar plates which were inoculated with bacterium subtilis (a gram-positive bacterium). The 3 B-classes are 3 assays of Penicillin solution of different origin. A measured value x_{ij} is the width in mm of the area of activity surrounding a little hole punched out from the agar plate. The 3 holes of each agar plate were filled with the 3 different Penicillin solutions.

Firstly we compute the marginal sums S_i and S_j , then the total sum $S_{..}$.

	Assay 1	Assay 2	Assay 3	Sum	Means
Plate 1	$x_{11} = 27$	$x_{12} = 21$	$x_{13} = 34$	$S_{1.} = 82$	27,33
Plate 2	$x_{21} = 26$	$x_{22} = 19$	$x_{23} = 31$	$S_{2.} = 76$	25,33
Plate 3	$x_{31} = 27$	$x_{32} = 18$	$x_{33} = 34$	$S_{3.} = 79$	26,33
Plate 4	$x_{41} = 29$	$x_{42} = 22$	$x_{43} = 33$	$S_{4.} = 84$	28,00
Sum	$S_{.1} = 109$	$S_{.2} = 80$	$S_{.3} = 132$	$S_{..} = 321$	Total = 26,75
Square sum	$\sum x_{i1}^2 = 2975$	$\sum x_{i2}^2 = 1610$	$\sum x_{i3}^2 = 4362$	$\sum x_{ij}^2 = 8947$	
Means	27,25	20,00	33,00		

From the marginal sums we can compute the mean values of the rows, the mean values of the columns and the total mean value. The summation of squares for the computation of the total sum of squares over all measured values x_{ij} we perform here using the columns (but the same we could do with the rows). The model of this ANOVA is demanding here that each x_{ij} must follow the distribution $N(\mu_{ij}, \sigma_e^2)$, i.e. distributed with mean value μ_{ij} and variance σ_e^2 . The μ_{ij} is the expectation value of the width of the area of activity at plate i surrounding the hole filled with solution j . To illustrate the effects (here the interesting effects are the differences of the widths), one can make a decomposition of the measured value in the following way:

$$\mu_{ij} = \mu_{..} + (\mu_{i.} - \mu_{..}) + (\mu_{.j} - \mu_{..}) + (\mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..})$$

or

$$\mu_{ij} = \mu_{..} + \alpha_i + \beta_j + e_{ij}$$

The sums of the effects α_i and β_j are separately and also together always zero. The experimental errors e_{ij} are normally distributed with $N(0, \sigma_e^2)$, i.e., with mean value 0 and variance σ_e^2 . Our intention is to find the experimental error σ_e and then to test the linear contrasts $\mu_1 - \mu_2$, $\mu_1 - \mu_3$ and $\mu_2 - \mu_3$ whether they are $\neq 0$ (are there significant differences between the 3 Penicillin assays?).

Mathematically, we can make a decomposition of the sum of squares, e.g. :

$$\sum_{j=1}^s \sum_{i=1}^r (x_{ij} - \bar{x}_{..})^2 = \sum_i s (\bar{x}_{i.} - \bar{x}_{..})^2 + \sum_j r (x_{.j} - \bar{x}_{..})^2 + \sum_j \sum_i (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})^2$$

or

$$SAQT = SAQA + SAQB + SAQR$$

with

$$SAQT = \sum_j \sum_i x_{ij}^2 - \frac{S_{..}^2}{n} \quad \text{with} \quad S_{..} = \sum_j \sum_i x_{ij} \quad \text{and} \quad n = r \cdot s ,$$

$$SAQA = \sum_{i=1}^r \frac{S_{i.}^2}{s} - \frac{S_{..}^2}{n} \quad \text{with} \quad S_{i.} = \sum_j x_{ij} ,$$

$$SAQB = \sum_i \frac{S_{.j}^2}{r} - \frac{S_{..}^2}{n} \quad \text{with} \quad S_{.j} = \sum_i x_{ij} ,$$

$$SAQR = SAQT - SAQA - SAQB$$

Now we can establish the table of the ANOVA with the decomposition of the sums of squares (*SAQ*), their degrees of freedom (*DF*), their mean squares (*MQ*) and their expectations of the mean squares (*E(MQ)*) :

Cause of Variance	SAQ	FG	MQ	E(MQ)
Variance between A-classes	SAQA	FGA = r-1	$MQA = \frac{SAQA}{r-1}$	$\sigma_e^2 + \frac{s}{r-1} \sum_i \alpha_i^2$
Variance between B-classes	SAQB	FGB = s-1	$MQB = \frac{SAQB}{s-1}$	$\sigma_e^2 + \frac{r}{s-1} \sum_j \beta_j^2$
Rest variance	SAQR	FGR = (r-1)(s-1)	$MQR = \frac{SAQR}{(r-1)(s-1)}$	σ_e^2

We can use the F-test as a global test of homogeneity, e.g. of the homogeneity of the B-means (here of the Penicillin assays):

$$F_{\alpha, FG1=s-1, FG2=(s-1)(r-1)} = \frac{MQB}{MQR}$$

We accept homogeneity of the B-means (hypothesis H_0 or *no significant differences of the means between the B-classes*), if $F < F\alpha$. We accept the existence of at least one significant difference of means, if $F \geq F\alpha$.

A linear contrast is a linear combination of a set of means μ_i and constants c_i of the form

$$K = c_1\mu_1 + c_2\mu_2 + \dots + c_k\mu_k \quad \text{with} \quad \sum_{i=1}^k c_k = 0.$$

We will consider here only the simplest contrasts possible, namely those for $k = 2$. These contrasts are of the form $K_{ij} = \mu_i - \mu_j$ with $c_i = 1$ and $c_j = -1$. We can use a **F-test** but also a **t-test** as a test of significant contrasts of the B-classes:

$$t_{ij} = \frac{(\bar{x}_i - \bar{x}_j)}{\sqrt{MQR}} \sqrt{\frac{r \cdot r}{r + r}} \quad \text{with DFR degrees of freedom and } \sqrt{MQR} = \sigma_e.$$

If we want to test the A-means, then we use the formula

$$t_{ij} = \frac{(\bar{x}_i - \bar{x}_j)}{\sqrt{MQR}} \sqrt{\frac{s \cdot s}{s + s}} \quad \text{with DFR degrees of freedom as well.}$$

The table of ANOVA for our numerical example of the $r = 4$ agar plates and $s = 3$ Penicillin assays is:

SAQT = 360,25	DFT = $n - 1 = 12 - 1 = 11$	
SAQA = 12,25	DFA = $r - 1 = 4 - 1 = 3$	MQA = SAQA/DFA = 4,083
SAQB = 339,50	DFB = $s - 1 = 3 - 1 = 2$	MQB = SAQB/DFB = 169,75
SAQR = 8,50	DFR = $(r-1)(s-1) = 6$	MQR = SAQR/DFR = 1,4166 = σ_e^2

The first t-test is testing the contrast between the Penicillin assays 1 and 2.

$$t_{12} = \frac{(27,25 - 20)}{\sqrt{1,4166}} \sqrt{\frac{4 \cdot 4}{4 + 4}} = 8,61 \quad \text{with degrees of freedom } DF = 6.$$

We accept hypothesis H_0 (no significant difference), if $|t_{ij}| < t_\alpha$.

We accept hypothesis H_A (a significant difference), if $|t_{ij}| \geq t_\alpha$.

Because of $t_{\alpha=5\%, DF=6, 2\text{-sided.}} = 2,45$ we accept here the hypothesis H_A for the first contrast. Assay 1 has with a width of 27,25 mm of the area of activity a significantly better effect as assay 2 with a width of 20 mm only. In the same kind we perform the tests of the assays 1 and 3 and of the assays 2 and 3. The t-values are $t_{13} = -6,83$ and $t_{23} = -15,44$. This differences are significant as well.

If one does not want to test 3 single hypotheses, but a multiple hypothesis, then one can use the Bonferroni adjustment of the error probability α , or otherwise Holm's sequential procedure. In both cases one needs t-tables of the critical points for the adjusted α . In the case of three mean values the Bonferroni adjustment is $\alpha^* = \alpha/3 = 0,05 / 3 = 0,01667$. For such an α our little table of critical points at page 2 of the script is not arranged. But the Excel-function =TINV(error probability ; degrees of freedom) can yield the 2-sided critical points t_α of the t-distribution for an arbitrary α (not only for $\alpha = 0,05$ or 5%).

14. Classification

The Discriminant Analysis can classify new objects if we have a learning sample (training set) of objects with known class. Cluster Analysis tries to find some classification in a set of data without any knowledge.

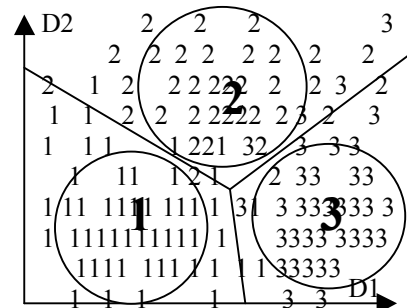
14.1 Linear Discriminant Analysis

Discriminant analysis has the following main tasks:

1. Construction of classification rules for objects using a learning sample (training set) of objects with known class, and to estimate the classification error expected

2. Classification of new objects (working objects from a working sample) using the constructed rules, and graphical or tabular representation of the results
3. Search of significant discriminating variables, e.g. for the reduction of data acquisition effort of classification data
4. Test of multiple mean differences
5. Test for isolation of object classes, especially in combination with cluster analysis.

The right-side figure shows the linear discriminant space spanned by the discriminating variables D1 and D2. Classes (here 1, 2, 3) have theoretically the shape of a circle in such a projection. Dividing lines are the border lines between class areas. The objects (here the digits) are not always inside of their class area (incorrectly classified objects). A main goal of the user is the minimisation of the classification error. Discriminant space has dimension $d=k-1$ if number of classes is k (here $d=2$).



Example for Classification: We want that a computer software can discriminate the GC-curves of 10 solution mediums for colors (a task of Gas-Chromatography). One feeds the program with 10-30 values taken from the characteristic points of each chromatogram. To improve the redundancy of the data one takes at least 10 samples (10 GC-curves) for each solution medium. Now, the linear discriminant analysis computes one or more discriminating variables from the 10-30 input variables, and additionally the border lines between classes. The border lines separate the classes in the discriminant space, spanned by the discriminant variables. Each past or future chromatogram is represented by a point in the discriminant space. One assumes now that the point has the class of that area where it is positioned. So the computer can classify solution mediums by their GC-curves.

Example for selection of variables: By reasons of time one wishes to minimize the number of measuring points needed for the classification of manufacturing errors. One performs a lot of measurements (much too much) on components with known manufacturing errors. Now we make an automatic reduction of the variable set using the forward or backward stepwise method of discriminant analysis yielding the set of essential variables (essential for correct classification). The "stepwise method" selects variables only with a significant discriminating power. If some variables with the same power are highly correlated, the method selects more or less by chance only one member of this group of variables.

Example for multivariate comparison of means: Do differences exist between newborns from cities and newborns from country? One acquires data from newborn babies, e.g. weight, length, temperature and so on, but also data concerning descent (e.g., city or country). The program computes the Mahalanobis distance of both classes (city / country). The Mahalanobis distance is a kind of averaged distance of the means **of all measured variables**. One can test this distance by a multivariate test.

Due to the main tasks we recommend different methods of the discriminant analysis:

- In cases with a high number of variables and a low number of training objects we recommend the **stepwise algorithm** (forward or backward). We recommend this method also if one wishes to select the best discriminating variables..

- In cases with only few variables and many training objects we recommend analysis **with all variables..** (Only extremely high correlated subsets of variables will shed some of them.)

Classification of new objects implies also the question of classification errors. We are not interested in the smallest error of re-classification, but in the minimal classification error for new objects. One uses jackknife-methods (Lachenbruch) or bootstrap-methods (in DASY jackknife only. The Lachenbruch-method uses division of the training set into 10 parts. 9/10-th of the objects are used for learning, 1/10-th to test correct classification. This is repeated until each part was used for testing. Another jackknife-method is to divide a great learning sample into two parts only - a learning part und a working part. The learning part is used to test the classification.) The question which variables are significant is not important here.

Multiple multivariate comparision of means: A global F-test is performed (Ahrens/Laeuter p.106, equ.7.12). The test tests multivariate distances of the class centroids. Simultaneous comparision of each class i with each other class j yields a matrix F_{ij} of F-values. They all have the same critical value F_{sim} . Together with the test of centroid distances we can perform the pair-wise test for isolation (Ahrens/Laeuter p.138, equ. 7.73). If classes are not isolated we expect a high classification error. In connection with cluster analysis we can decide by the test for isolation wheather two classes should be better united or not.

Strategies of classification: Without a-priori probability: Classification of an object into a class depends only on the squared distance k of its Euclidean distance from the nearest class centroid in the discriminant space, and on a factor $N_j/(N_j+1)$, which is always nearly 1. Here N_j is the class size (number of objects of the class in the training sample). **With a-priori probability:** Classification depends on distance k , on factor $N_j/(N_j+1)$, and additionally on the probability P_j of the target class (Ahrens/Laeuter p. 131, equ. 7.63). A-priori probability of the target class is $P_j=N_j/N$, i.e. the relative frequency of objects of class j in the training sample. A great class has automatically a great a-priori probability. There is no rule when we should work with and when we should work without a-priori probabilty. Criterion is only the quality of classification measured by the jackknife classification error.

Estimation of classification error: If an object is assigned to a false class, we have a classification error. We discriminate between:

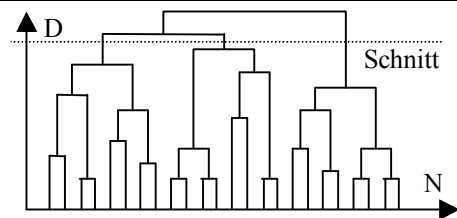
- **Error of re-classification:** The objects of the learning sample are re-classified, i.e., we assign them to some class. If this class is not correct, we make errors. This errors decreases with increasing number p of variables. But you must not let deceive by this decrease of the re-classification errors. The classification of objects not in the learning sample grows poorer and poorer, the greater the number of unneeded variables.
- **Jackknife-error estimation (Lachenbruch):** The training set is divided randomly into parts (most 10 parts) Nine parts are used for learning, the tenth is used to prove the correct classification, and to estimate the classification error. Then we take another 9 parts of the 10 for learning, and the remaining part for testing. This change is repeated 10 times until each part was used one time for error estimation. This method of error estimation works very realistically.
- **Jackknife-error estimation (working sample):** If the data set used is great enough, one can divide it into a learning sample and a working sample. The learning sample is used to find the discriminant variables and to construct the discriminating rules. The working sample is used to prove the correct classification, and to estimate the classification error. This method id the most realistic method for classification error estimation.

Data structure for a linear discriminant analysis: One needs a categorical **target variable** Y with class numbers and one or more **discriminating variables** X_j. These can be metric, binary or rank ordered. Additionally, we can calculate new variables from the input variables by raising to power or by multiplication (**polynomial models**). A categorical variable with k categories the user had to transform into k-1 binary variables (Example: Color of hair with 3 categories K1=black, K2=red, K3=blonde the user must transform into two new variables M1=black/non-black and M2=red/non-red)

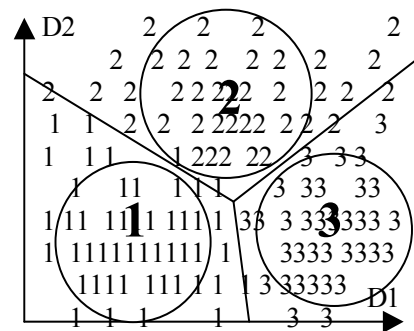
14.2 Cluster analysis

If one has no idea how his collected or measured data are structured, one uses cluster analysis to find a first class structure. This method is comparable with a view to the starry sky. One tries to find figures. If the class structure found by the cluster analysis has some meaning one has to prove by further scientific analyses of the properties of objects *thrown* in one class. We have two principally different clustering strategies:

Hierarchical Methods: They produce a dendrogram (tree structure) by ordering the N objects using the distance D between objects in the multidimensional variable space. Objects with a small distance between one another are thrown into one class. We create k classes by a cut in the appropriate height (here k=3).



Partitional Methods use starting objects (seed points) as centres of future classes. By exchange of objects between neighbouring clusters one tries to improve the isolation and compactness of the clusters found, i.e., classes without outliers and of spherical shape. The graphical representation uses the features of the discriminant analysis. Cluster analysis does not know a-priori the notion classification error, because the classes are defined by the clustering process the first time.



Common to all methods is that the user should have an idea of the number of classes he wishes to find. Another common property of clustering algorithms is that they need a lot of computer time. The **input data** are row vectors - always one row for each object. Concerning the variables, we find the same restrictions as in discriminant analysis: we can use only metrical, binary or rank ordered categorical variables. The **result** of cluster analysis is a class number for each object, and some statistics concerning the clusters (mean, size, ...). Here one uses the features of discriminant analysis again.

14.3 Logistic Regression

Logistic regression divides a set of objects into two classes exactly ($y=0$ and $y=1$). The class number 0 or 1 is estimated similarly to regression analysis using p predictor variables x_1, x_2, \dots, x_p .

Example: y =Dental caries 0/1, x_1 =Water Fluoridation 0/1, x_2 =Percentages of Sugar in Food.

Variable y is theoretically **Bernoulli-distributed** with $P(y=r) = p^r (1-p)^{1-r}$ and $r = 0 / 1$. Expectation is $E(y)=p$, Variance $Var(y) = \sigma_y^2 = p(1-p)$. To

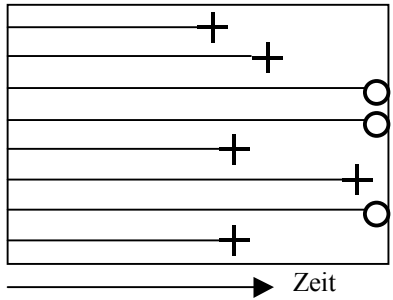
$$p(x) = \frac{\exp(b_0 + b_1 x_1 + \dots + b_p x_p)}{1 + \exp(b_0 + b_1 x_1 + \dots + b_p x_p)}$$

make a model for probability p depending on the x -variables one uses the Logistic Distribution $p(x)$. Because of $p(1-p) = \exp(b_0 + b_1x_1 + \dots + b_px_p)$ is the model $g(x) = \log(p/(1-p)) = b_0 + b_1x_1 + \dots + b_px_p$.	logistic regression model: $g(x) = \log\left(\frac{p(x)}{1-p(x)}\right) = b_0 + b_1x_1 + \dots$
---	--

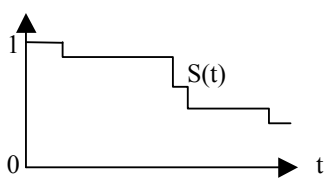
The predictor variables x_1, x_2, \dots can be metrical, categorial rank ordered, or binary. Logistic regression estimates the probability $p(x)$ that y has a value of $y=1$. To classify a new object (e.g. a patient into class "ill" or class "healthy") one has to compute $p(x)$ using the estimated regression coefficients b_0, b_1, \dots and values for the variables x_1, x_2, \dots of the object wished to be classified. The estimation of the coefficients is done with a training set (learning sample) of known classification. Then we compare p with a **threshold** (default is 0.5, or better, the threshold is given by the minimum of classification errors in the learning sample). The estimation of the regression coefficients b_0, b_1, \dots is a time-consuming iterative procedure for maximization of the **Maximum-Likelihood Function**. The procedure is giving also the standard errors s_{b_j} if the coefficients b_0, b_1, \dots . The **Wald-Test** (by Abraham Wald), $W = b_j / s_{b_j}$, tests the significance of the coefficients (and so the influence of the variables), whereas W is assumed to be approximately normally distributed. The **Likelihood-ratio-Test** is a global test to compare different logistic regression models.

Explanation of action of a binary predictor variable ($x_j=0/1$) leads to the Odds-Ratio OR with $\log(OR) = g(1) - g(0) = b_j$. $OR = e^{b_j}$ is the probability which $x_j = 1$ is adding to the disease risk.	$OR = \frac{p(1)}{1-p(1)} \bigg/ \frac{p(0)}{1-p(0)}$
In the case of continuous predictor variables x_j one computes the increase of the risk probability if x_j is increased by 1. All other predictor variables $x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_p$ remain unchanged.	$\frac{p(x_1, \dots, x_j + 1, \dots, x_p)}{p(x_1, \dots, x_j, \dots, x_p)}$

15. Survival Analysis

Estimation of Survival Time after a treatment is the major variable of interest of the survival analysis. Survival time depends on factors as age, sex and so on. This fact leads to stratified analyses. During the course of the study patients can die (+), or leave the study (o = censored). The end of the study, removal, or death by accident are reasons for leaving the study (censoring). Such data are called Censored Observations (Censored Data).	
---	--

The Survival Function $S(t)$ is $S(t) = 1 - F(t)$, with $F(t)$ as cumulative sum curve of all survival times t . $S(t)$ is giving the percentage of patients which survived after time t . The Kaplan-Meier estimation of $S(t)$ is the cumulative product of the survival probabilities p_t (survival from day t to day $t+1$).	$S(t) = \prod_t p_t$
--	----------------------

The survival probabilities p_t are estimated in this way (n_0 =Starting number of patients): Nobody dies / is censored: $p_t = (n_t - 0) / n_t$ and $n_{t+1} = n_t$ Anyone dies: $p_t = (n_t - 1) / n_t$ and $n_{t+1} = n_t - 1$ Anyone is censored: $p_t = (n_t - 0) / n_t$ and $n_{t+1} = n_t - 1$	
--	---


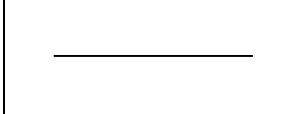


The **Median of Survival Time** (50% survived) is given by time t with $S(t)=0.5$. An approximative **confidence interval** of the *true survival function* is

$$\hat{S}(t) \pm u_{\alpha, zweis.} \hat{S}(t) \sqrt{(1 - \hat{S}(t)) / n_t}$$

with the two-sided critical value of Normal distribution, e.g., $u_{\alpha}=1.96$ under $\alpha=0.05$.

The comparison of two survival functions $S_1(t)$ and $S_2(t)$ of two groups of patients one can perform, e.g., with the Log-rank Trend Test. One builds the 2x2-tables, $a=d_{1i}$, $b=d_{2i}$, $c=n_{1i}$, $d=n_{2i}$, at each point of time $i=1,2,\dots,I$ with a failure, regardless of group 1 is referred to, or group 2, or both groups. Here the d_{ji} are the numbers of failures at time i in group j . The n_{ji} are the number of patients directly before time i in group j . We calculate the expectations $E(a)=(a+b)(a+c)/(a+b+c+d)$ and $E(b)=(a+b)(b+d)/(a+b+c+d)$. From the expectations we calculate the four sums $O_1= \Sigma a$, $O_2= \Sigma b$, $E_1= \Sigma E(a)$, $E_2= \Sigma E(b)$. Summation goes over $i=1,2,\dots,I$. Then we compute the TREND-test, $\chi^2_{Trend} = T^2/V_T$, using $T=(O_1+O_2)-(E_1+E_2)$ and $V_T=(E_1+E_2)^2/(E_1+E_2)= E_1+E_2$. The statistic χ^2_{Trend} is χ^2 -distributed under H_0 with $df=1$. Null hypothesis H_0 assumes equal risk in both groups. We compare χ^2_{Trend} one-sided with the critical value χ^2_{α} . If $\chi^2_{Trend} \geq \chi^2_{\alpha}$, then H_0 is rejected. We assume a significantly different risk.

The **Hazard function** (hazard rate, force of mortality, failure rate (in Quality Assurance) $h(t)$ is the probability for a patient to die at day t / for a device to fail. The estimation of the function is a hard job (Statistical packages BMDP, SPSS, SAS, SPSSX, StatView, e.g., contain programs for survival analysis to do it) The definition of the Hazard function is $h(t) = f(t)/S(t)$, with density distribution $f(t)$ of the survival times. The following 4 figures show typical shapes of Hazard functions:

negative aging, e.g., after operation	no aging (good repair service)	positive aging with increasing age	bathtub curve from newborn to old man
			

The **Model of Cox** is a typical nonlinear regression model using the variable t , but also predictor variables x_1, x_2, \dots, x_p . The "proportional hazards model" of Cox is:

$$h(t) = h_0(t) \cdot \exp(b_1 x_1 + b_2 x_2 + \dots + b_p x_p).$$

Here $h_0(t)$ is called Baseline Hazard. It is a non-parametric unknown function, independent on the linear regression $b_1 x_1 + b_2 x_2 + \dots + b_p x_p$. Each term $b_j x_j$ of the linear regression acts by virtue of the exponential function e^x as a factor of $h_0(t)$.

16. References and Dictionary

allgemeine Biostatistik	Spezialliteratur
Sachs., Lothar. (1997). <i>Angewandte Statistik: Anwendung statistischer Methoden</i> . 8th ed., Springer Verlag (one of the mostly used statistical books in Germany)	Lautsch, Erwin & Stefan von Weber (1995). <i>Methoden und Anwebdungen der Konfigurationsfrequenzanalyse (KFA)</i> , Beltz Psychologische Verlags Union
Gasser, Theo & Burkhardt Seifert (1999). <i>Biostatistik</i> , 3. Auflage, Universität Zürich, Abteilung Biostatistik (recommended, please contact the authors)	Ahrens, Heinz. & Jürgen Läuter: "Mehrdimensionale Varianzanalyse", Akademie Verlag Berlin 1981.

Weber, Erna (1967). <i>Grundriß der biologischen Statistik</i> , Gustav Fischer Verlag Jena	Mucha, Hans-Jo.: Clusteranalyse mit Mikrocomputern, Akademie-Verlag Berlin, 1992
Altmann, D.G. (1991): Practical Statistics for medical research, Chapman and Hall	Josef Puhani, Statistik, Bayrische Verlagsanstalt, Bamberg 1995 (for Economists)

Dictionary German-English

abhängig dependent Abhängigkeit dependence ablehnen reject Abweichungsquadrat square of residuals Achse (geometrisch) axis adjustiertes Alpha adjusted alpha akzeptieren accept Alternativhypothese alternative hypothesis arithmetisches Mittel arithmetic mean asymptotisch asymptotical Ausreißer outlier Balkendiagramm (horizontal) bar chart beschreibend descriptive Binomialtest binominal test bivariat bivariate Bogenmaß radian measure Chi-Quadrat-Verteilung chi-square-distribution Clusteranalyse cluster analysis Diagramm chart, diagram, plot Dichtekurve density function Diskriminanzanalyse discriminant analysis Einflußgröße independent variable einseitig one-sided empirisch empirical, observed Ereignis event Erhebung survey, data collection Erwartungswert expectation value F-Verteilung Fishers distribution, F-distribution Faktoranalyse factor analysis Fall case Fehler 1. Art error type I Fehlerquadratsumme error sum of squares Fragestellung hypothesis Freiheitsgrad degree of freedom geometrisches Mittel geometric mean Geradengleichung straight line function gewogenes Mittel weighted average Gleichverteilung uniform distribution Gradmaß degree Grenze limit Grundgesamtheit population Gruppe group, class, category Häufigkeitsdiagramm frequency histogram Häufigkeitsverteilung frequency distribution Histogramm histogram Hypothese hypothesis hypothesenprüfend confirmatory Irrtumswahrscheinlichkeit error probability Kartogramm cartogram kategorial categorical Klasse class, group, category Klassenbreite width of class interval	Korridor interval Kreisdiagramm pie chart, circular chart Kreisfrequenz angular frequency Liniendiagramm line plot Maßstab scale Median median Merkmal variable, variate, symptom, factor Merkmalskombination syndrom metrisch metric Mittelwertvergleich comparison of group centroids mittlerer Fehler mean error multivariat multivariate Normalverteilung normal distribution Nullhypothese null hypothesis parameterfrei parameter free, nonparametric Piktogramm pictograph, pictogram Prognose prediction, forecasting Prüfgröße test statistic Prüfverteilung test distribution Punktwolke scatter plot Randsumme marginal sum, marginal total ranggeordnet rank ordered Regressionsanalyse regression analysis Regressionskoeffizient regr. coefficient, slope Regressionskonstante regr. constant, intercept saisonale Schwankung seasonal variation Säulendiagramm (vertical) bar chart Schätzwert estimated value Sicherheit significance Sicherheitspunkt critical value, significance limit Signifikanzniveau confidence level, error type I level Standardabweichung standard deviation Statistik statistics Stichprobe sample Streubreite range streuen scatter Streumaß measure of spread Summenkurve cumulative distribution function t-Verteilung students distribution, t-distribution Teilung (Maßstab) graduation, division Teststatistik test statistic theoretische Verteilung theoretical distribution Typ type Umfang (einer Stichprobe) size, sample size unabhängig independent Unterschied difference unvollständig incomplete Varianz variance Varianzanalyse analysis of variance Vergleich comparison
---	--

Klasseneinteilung classification	Voraussetzung assumption
Kleinste-Quadrate-Schätzung least squares estimation	Vorhersage prediction, forecasting
Konfidenzintervall confidence interval	Wahrscheinlichkeit probability (likelihood)
Kontingenztafel contingency table	Zeitreihe time series
kontinuierlich continuous	Zielgröße dependent variable
Korrelationsanalyse correlation analysis	Zufallsvariable random variable
korreliert correlated	Zusammenhang relationship, association
	zweiseitig two-sided

17. Practical Exercises

The lecture is without Practical exercises. But if one wants to get some numerical practice then he can do the following exercises at his own PC:

Some EXCEL-Functions (*German version of Excel !!!*)

EXCEL has good help-features. This table can be only a small introduction.

Function and Parameters	Example of call	What is delivered?
geomittel(xwerte)	=geomittel(a1:a5);	geometric mean
häufigkeit(x;klassengrenzen);	=häufigkeit(a2:a35;b7:b8)	class frequencies
norminv(p;mittelwert;sigma)	=norminv(b5;c1;d1)	Quantile X_p Normal Distrib.
normvert(x;mittel;sigma;typ)	=normvert(a8:a12;b1;c1;1)	$\Phi(u)$ with $u=(x-mittel)/sigma$
rgp(y;x;konst;zusatzstatistik)	=rgp(a2:a7;b2:d7;1;0)	(multiple) linear Regression
stabw(xwerte)	=stabw(c1:k1)	σ_{n-1} variance of population
tvert(t;df;s)	=tvert(d8;b9;2)	Error probability α to degree of freedoms df, 2-sided
trend(y;x;x*;k)	=trend(a2:a7;b2:b7;b8:b12;1)	y-values on straight line (k=1 with constant)
ttest(g1;g2;s;typ)	=ttest(a2:a9;b2:b14;2;2)	Comparison of means under normal distribution
*	=(a1:a5)*(b1:b5)	pairwise multiplication
potenz(x;y)	=potenz(((a1:a5)-a6);2)	$(A_i - A_6)^2$ with $i=1,\dots,5$

Different mean values and standard deviation: Run Excel. Type in cell A1 some data name, for example the word *data*. Type 7 numbers in cells A2:A8, for example 7 temperatures or 7 weights.

Type in cell A10 =Mittelwert(A2:A8) and then press ENTER.

A2:A8 means the selection of the field A2:A8 with the mouse, or you can type A2:A8.

Type in cell B10 the word "average" or "mean value" as an explanation.

In cell A11 the standard deviation σ_{n-1} by =Stabw(A2:A8), in B11 the word „Sigma“

In cell A12 the median by =Median(A2:A8), in B12 the word „Median“

In cell A13 the geometric mean by =Geomittel(A2:A8), in B13 „Geomean“

Outlier control with 3-Sigma-Rule:

Type in cell C1 the word "u-values".

Type in cell C2 the formula =(A2-\$A\$10)/\$A\$11 and press Enter.

Go in the right below corner of the cell and "draw" until cell C8.

The dollar characters prevent the change of the address (fixed addresses)

Type in cell D1 the word “Absolut u”.

Type in cell D2 the formula `=abs(D2)` and Enter and draw until cell D8.

Compute in cell D10 the maximum u-value by `=max(D2:D8)`.

Type in cell E10 the word “Maximum u”.

Decide whether there is an outlier in the data ($\text{Max } u > 3$?).

Make a chart of the columns P1 and P2 and seek visually for outliers.

Select A1 to A8 → Diagram assistant → Points (x,y) → Only points → make ready

Computing quartiles: Compute in cell A15 the first quartile by `=Quartile(A2:A8 ; 1)` (boundary of the lower 25% of the data), then in cell A16 the second quartile, in A17 the third quartile. Compare with the median. What is interesting?

Moments of the distribution of the data: Compute the first 4 moments.

In cell A20 the mean value `=Mittelwert(A2:A8)` In cell B20 the words “mean value”

In cell A21 the variance `=Varianz(A2:A8)` In cell B21 “Variance”

In cell A22 the skewnes `=Schiefe(A2:A8)` In cell B22 “Skewnes”

In cell A23 the kurtosis `=Kurt(A2:A8)` In cell B23 “Kurtosis”

Kurtosis or Excess is a thickening of the tails of a distribution related to Normal distribution.

Illustrate data distribution by a histogram: Go to table 2 in column A. Type in cell A1 the word *data*. Copy your 7 numbers from table 1 and insert in A2:A8. Type another 18 numbers until cell A26 (total $n=25$ numbers).

Type in cell B1 the word “class boundaries”. Type below 5 numbers in cells B2:B6. This class boundaries must be sorted ascending. The first boundary should be greater than the minimum data value, the last boundary should be less than the maximum boundary.

Type in cell C1 the word “Frequencies”

Select by mouse the field C2:C7. Type in cell C2 (still white coloured) the formula `=Häufigkeit(A2:A26 ; B2:B6)` and press the triple key `Strg-Shift-Enter`. The six cells C2:C7 are now filled by the counted frequencies. The first frequency is the number of datas in class 1 (minimum data until including the first boundary). The last frequency is for the class including and beyond the last boundary.

Type in cell D1 the word „class“. If your class boundaries were 10, 20, 30, 40, 50, for example, then type in cell D2 the text “*til including 10*”, in D3 “*from 11 til including 20*”, and so on, and in D7 type “*up including 50*”.

Select C1:C7 → Diagram assistant → columns → weiter → Reihe → click in the box right from „Beschriftung der Rubrikenachse (x)“ and select by the mouse D2:D7 → Fertigstellen.

Indexing a data row X_1, X_2, \dots, X_n to a common starting point of 100%: Please, run the example “Indexing” from the lecture (formulas in chapter 4.2) with own data. Please, make two line graphs. The first with the original data, the second with the indexed data.

Indexierung auf gemeinsamen Startwert 100%: Spielen Sie das Beispiel Indexierung aus der Vorlesung mit eigenen Daten durch. Machen Sie eine Liniengraphik der beiden Datenreihen vor und nach der Indexierung.

Simple linear Regression with Tests in Excel

One uses the simple linear regression analysis for the following challenges:

- One wishes a best-fit line of the data
- One wants to know the error of the fit
- One wants to test for a significant slope
- One wants to test for a significant constant, or whether a model without constant would be the better choice
- One wants to extend the straight line for a forecast, an one wants to know how precise are the predicted .

The function `=trend(y-values ; x-values)` computes the expectations \hat{y}_i of the best-fit line defined by the y- and x-values. The function `=rgp(y-values ; x-values ; wahr ; wahr)` computes the regression coefficients, the errors of the coefficients, the residual standard deviation, the coefficient of determination $B=r^2$, the degree of freedom and so on. The first *wahr* stands for a "model with regression constant", the second *wahr* for "besides the coefficients Excel will compute another statistical numbers", as mentioned above. The abbreviation **SSE** is the three-fold key Strg-Shift-Enter. First, press the both left keys Strg and $\hat{\uparrow}$, then additionally ENTER. To perform a regression analysis type the characters "x", "y" and the word "y-Dach" as column headers, then type the x-numbers, the y-numbers in the cells A2:A7, or B2:B7, respectively. Then follow the scheme:

	S1=A	S2=B	S3=C	
Z1	x	y	y-Dach	Select cells C2:C7 and type: <code>=trend(sel. B2:B7 ; sel. A2:A7) SSE</code>
Z2	1,7	3,3		(y-value x-values)
Z3	2,3	4,1		At C2 to C7 the expectations y-Dach appear. Now we want to compute the regression coefficients and other statistical numbers: Select A9:B13 and type: <code>=rgp(sel.B2:B7;sel.A2:A7;wahr;wahr) SSE</code>
Z4	2,1	4,5		In column A and B appear the coefficients b1=slope, bo=Regressions constant of the straight line $y = bo + b1 x$
Z5	2,4	4,7		Computing the t-values: Sel. A15:B15 <code>= sel.A9:B9 / sel. A10:B10 SSE</code>
Z6	3,9	8,3		The t-values t1 and t0 appear.
Z7	1,6	3,3		
Z8				
Z9	2,20	-0,45	b1,bo	
Z10	0,18	0,45	sb1,sbo	
Z11	0,97	0,34	r2, sR	
Z12	144	4	F, df	
Z13	16,8	0,47	ssreg,ssres	
Z14				
Z15	12	-1.0	t1, t0	

Linear regression analysis attempts to model the relationship between two variables by fitting a linear equation to observed data. If variable x is the time then we say also **trend analysis**.

$$\text{Regression model } y_i = \mathbf{a} + \mathbf{b} x_i + e_i$$

Seeked are estimations of the regression constant **a** and regression coefficient **b** in the population. A sample is given with the *n* matched pairs $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ similarly to the correlation coefficient data. Dependent variable **y** is called also *target variable*, independent variable **x** also *predictor variable*. The deviation e_i is called also *residuum* or *error*. Index **i** is the case number (point number, patient number). Regression constant **a** is the expectation of the dependent variable **y** at point $x=0$. Regression coefficient **b** is the *slope* of the straight

line. The coefficients **a** and **b** are estimated by the *Least Squares Method*, i.e., the sum of the error squares is minimized, $\sum e_i^2 = \text{Minimum}$.

$$\hat{y}_i = \hat{a} + \hat{b} \cdot x_i = \bar{y} + \hat{b} \cdot (x_i - \bar{x}) \quad \text{estimates } y \text{ at point } x_i \text{ (expectation value)}$$

$$\hat{e}_i = y_i - \hat{y}_i \quad \text{estimates the error } e_i \text{ at point } x_i$$

$$\hat{S}_R = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SAQ_{yy} - \hat{b} \cdot SAP_{xy}}{n-2}} = \sqrt{\frac{\sum \hat{e}_i^2}{n-2}}$$

\hat{S}_R estimates the mean error σ_R in the population (residual standard deviation of the points around the straight line (parallel to y-axis measured)). The formula using SAQyy and SAPxy is the best for pocket calculators.

$$df = n-2 \quad \text{Degrees of freedom of } \hat{S}_R$$

$$S_b = \hat{S}_R / \sqrt{SAQ_{xx}} \quad \text{Estimation error of regression coefficient } \mathbf{b}$$

$$S_a = \hat{S}_R \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SAQ_{xx}}} \quad \text{Estimation error of regression constant } \mathbf{a}$$

$$t_a = \hat{a} / S_a \quad \text{with } df = n-2 \text{ tests } H_0: \mathbf{a}=0 \text{ and } H_A: \mathbf{a} \neq 0 \text{ (2-sided)}$$

$$t_b = \hat{b} / S_b \quad \text{with } df = n-2 \text{ tests } H_0: \mathbf{b}=0 \text{ and } H_A: \mathbf{b} \neq 0 \text{ (2-sided)}$$

Using pre-knowledge, we can perform both tests also 1-sided. A significant $\mathbf{a} \neq 0$ means, that target variable y has a value $y \neq 0$ at point $x=0$. A significant slope $\mathbf{b} \neq 0$ means that predictor variable x has some direct or indirect influence at the target variable y, i.e., the slope is not by chance.

The meaning of the numbers is:

sb1, sb0 the standard deviations (errors) of the coefficients b₁ and b₀
 r^2 coefficient of destination B (in a simple linear regression it equals the square of the correlation coefficient, that is $B=r^2$)

sR Residual standard deviation (mean error)

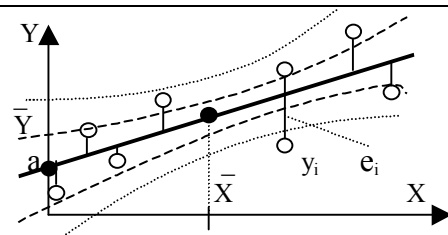
F F-statistic for hypothesis $H_0: b_1=0$ with degrees of freedom $df_1=1$ and $df_2=df$. In a simple linear regression analysis holds $F=t_1^2$, with t_1 as t-statistic for b₁ with df degrees of freedom.

t_1, t_2 are the both test statistics for the coefficients b₁ and b₀. One tests the hypotheses $H_0: b_1=0$ or $H_0: b_0=0$

ssreg = $\sum (y_i - \bar{y})^2$, also called sum of squares of the y (SAQyy)

ssresid = $\sum (\hat{y}_i - y_i)^2 = \sum e_i^2$, als called sum of squared errors.

The graphik shows the regression line in the X-Y-coordinate system. It crosses point **a** at the Y-axis and the point (\bar{x}, \bar{y}) . The measurements y_i are given by small circles, the residuals e_i by short slashes. The confidence interval of the true regression line is dashed, that of the prediction is dotted.

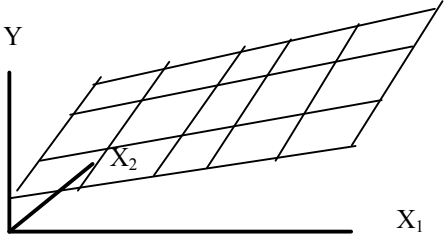


We want to store the graph of the regression line together with the data points in the gif-format or in the jpeg-format. Such a format we can insert without problems in a WORD-document:

1. Select A1:C7 that is, that the column headers are included

2. Click the diagram assistant
3. Chose type of diagram *Points (X,Y)*, there click "*Lines with points*"
4. Click *Ende* and then click the graph (it gets a black frame). Click *Bearbeiten*, then *Kopieren*, then minimize Excel (the minus sign in the right upper corner)
5. ---> *Alle Programme* ---> *Zubehör* ---> *Paint* ---> *insert graph (Einfügen)* ---> *store as (speichern unter)* ---> type *.gif, name *Trend*, don't forget the folder, store, close Paint.

Multiple linear Regression with Excel

<p>Multiple regression connects p predictor variables X_1, X_2, \dots, X_p with a target variable Y.</p> <p>The model may be with or without constant b_0</p> $Y = b_0 + b_1 * X_1 + b_2 * X_2 + \dots + b_p * X_p + e$ <p>The geometrical interpretation is a plane</p>	
--	--

over the space stretched by X_1, X_2, \dots . The regression coefficients b_1, b_2, \dots, b_p (and b_0) are estimated by the method of least squares ($\Sigma e^2 = \text{Minimum}$). e is the statistical error or residuum (deviation). Multiple regression is used for the following issues for example:

- One wishes the best-fit plane through a cloud of data points, that is, one wishes to explain the influence of the predictor variables X_1, X_2, \dots on the target variable Y by a linear formula. With such a formula one can predict values (forecast), or one can interpolate between data points.
- One wishes to know whether the linear formula fits the target variable with sufficient exactness. One can evaluate the total influence of all predictor variables by a global test.
- One wishes to identify these predictor variables out of the set of all predictor variables, which have a significant influence on the target variable. That means, one evaluates each single predictor variable.

The plane formula is the mostly used regression model:

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \dots + b_p X_{ip} + e_i$$

Y_i is an observed value of the target variable, X_{ij} is the i -th value of the j -th predictor variable, b_0 is the regression constant, b_1, b_2, \dots, b_p are regression coefficients, e_i is the error at the data point i (or deviation or residuum).

The software (e.g. Excel) estimates the regression coefficients b_1, b_2, \dots, b_p (and b_0 if demanded) using the Least-Squares method ($\Sigma e^2 = \text{Minimum}$). Residuum e is assumed to be a random error (deviation between measuring value and expectation).

Exampel Parametric Price fixing: A company wants to estimate the achievable price of a new electrical motor. For this task one needs technical data of the new motor, as weight, electrical power, revolutions per minute, voltage. Additionally, one needs the data of already existing motors together with their market prices. Target variable is the market price. Influencing variables (descriptor variables) are the technical data. The multiple linear model is, e.g.:

$$\text{Price} = b_0 + b_1 * \text{Weight} + b_2 * \text{Power} + b_3 * \text{Revolutions} + b_4 * \text{Voltage} + e$$

The multiple regression analysis estimates the regression coefficients from the given data and prices. To get the estimated market price of the new motor one calculates $\text{Price} = b_0 + b_1 * \text{Weight} + b_2 * \text{Power} + b_3 * \text{Revolutions} + b_4 * \text{Voltage}$ with the technical data of the new motor.

Example Growth of crops: The harvest is estimated depending on different parameters. The data are in the following table:

row\col	A	B	C	D	E	F
1	Soil quality	Sprinkling	Fertilisation	Temperature	Soil density	Y=Harvest
2	2	2	0,10	17	1320	1,1
3	2	3	0,15	19	1410	1,5
4	4	2	0,10	22	1190	1,8
5	3	4	0,20	20	1240	2,0
6	2	1	0	18	1240	0,80
7	1	3	0,10	18	1350	1,20
8	4	4	0	21	1270	1,95
9	2	3	0,20	15	1300	1,15

We mark the cells A11:F15 and type a regression instruction. Firstly, we type the function name „=rgp”, then the parentheses, then the cells containing the data of the target variable Y, then the cells containing all data of the predictor variables X. Then, the first „wahr“ chooses a model „with constante“, the second „wahr“ says that we want besides the coefficients still other results, for example the errors s_{b_i} , the coefficient of destination R^2 , residual standard deviation s_R , and so on. You had to mark always 5 rows. The number of columns depends on the number of coefficients in the regression model (if demanded, then b_0 is counted among).

=rgp(F2:F9; A2:E9; wahr; wahr) Strg-Shift-Enter

row 11	$b_5=0,000129$	$b_4=0,0995$	$b_3=1,379$	$b_2=0,185$	$b_1=0,137$	$b_0=-1,597$
12	$s_{b_5}=0,00036$	$s_{b_4}=0,0119$	$s_{b_3}=0,253$	$s_{b_2}=0,0214$	$s_{b_1}=0,0322$	$s_{b_0}=0,514$
13	$R^2=0,997$	$s_R=0,0431$				
14	$F=147,44$	$df=2$				
15	$ssreg=1,37$	$ssresid=0,003$	ssreg	ssresid		

In row 16 we write the names of the predictor variables and the constant:

16 Soil density Temperature Fertilisation Sprinkling Soil quality b_0

Excel makes the order reverse (b_5, b_4, \dots, b_0). In the computed statistics in the rows 12 to 15 mean:

s_{b_5}, \dots, s_{b_0} the estimated standard errors of the coefficients b_5, \dots, b_0
 R^2 the coefficient of destination B (in a simple linear regression it equals the square of the correlation coefficient, that is $B=r^2$) $R^2=0$ means that there is no linear relation between the set of predictor variables X and the target variable Y. $R^2=1$ means that the predictor variables X reproduce the target variable Y with absolute exactness without any error.

s_R Residual standard deviation of the points from the computed plane (mean error)

F F-test statistic to test coefficient B. The null hypothesis $H_0: B=0$ (No relationship between dependent variable Y and predictors X_j) with $F=B(n-k)/(1-B)$ and $df_1=p$ and $df_2=n-k$, k =number of coefficients in the model including coefficient.

b_0 . Hypothesis H_a : „There is a significant influence of the predictor variables on the target variable. The value of $R^2>0$ is not by chance“. The error probability p in the case of rejection of H_0 (or acceptance of H_a) one can compute by the function $FVERT(F ; n - df - 1 ; df)$ in the case „with b_0 “ and $FVERT(F ; n - df ; df)$ in the case „without b_0 “. The error probability p is also named p-value.

The meaning of the sums $ssreg$ and $ssresid$ were also mentioned in the chapter *simple linear regression*. To evaluate each single predictor variable one takes for each variable the pair of hypotheses H_0 and H_a . H_0 says: “This predictor variable has no linear influence on the target variable. The value $b_j \neq 0$ of the coefficient is by chance.“. Hypothesis H_a says: „This predictor variable has a significant linear influence on the target variable.” Practically, one computes a test statistic for each coefficient. Mostly used is the t-statistic. We find $t_i = |b_i / s_{bi}|$.

In the EXCEL sheet we divide with one instruction all coefficients by ist errors and make the absolute values. We mark the cells A18:F18 and type the following instruction:

=ABS(A11:F11/A12:F12) Strg-Shift-Enter

row 18	$t_5=0,351$	$t_4=8,33$	$t_3=5,43$	$t_2=8,65$	$t_1=4,27$	$t_0=3,10$
--------	-------------	------------	------------	------------	------------	------------

The t-distribution has a similar shape as the Gaussian distribution (bell shaped). Function TVERT computes from the t-value, the degree of freedom df from above and the number 2 the two-sided error probability (p-value) for the case of rejection of H_0 for the coefficient under consideration. This p-value (error probability) should be small, for example $p < 0,05$. In this case we evaluate the predictor variable as significant. To compute the p-values we mark the cells A20:F20 and we type the following instruction:

=TVERT(A18:F18; B14; 2) Strg-Shift-Enter

row 20	$p_5 = 0,75$	$p_4 = 0,014$	$p_3 = 0,032$	$p_2 = 0,013$	$p_1 = 0,0506$	$p_0 = 0,09$
--------	--------------	---------------	---------------	---------------	----------------	--------------

In research one mostly uses the error probability $\alpha = 5\%$ (0,05). If the computed p-value is greater than α , then we accept the hypothesis H_0 (the predictor variable is not significant). If $p \leq \alpha$, then we accept hypothesis H_a (the predictor variable is significant). We mark the cells A22:F22 and type the following instruction:

=wenn(A20:F20 > 0,05 ; „Ho“ ; „Ha“) Strg-Shift-Enter

row 22	Ho	Ha	Ha	Ha	Ho	Ho
23	Soil density	Temperature	Fertilisation	Sprinkling	Soil quality	b_0

The badest p-value (highest error probability) has the predictor variable X_5 =Soil density. If we want to eliminate not significant predictor variables in our regression model, then we should eliminate firstly this variable (stepwise backwards). Please, eliminate only one predictor variable (or the constant) in one step, because the p-values can change dramatically by correlations between the predictor variables already by the deletion of only one variable. To

eliminate the regression constant you have to replace the first „wahr“ in the regression instruction by „falsch“.

The global F-test tests for a significant linear relation of the whole set of the predictor variables with the target variable (see above). The following three instructions compute the number of data rows, the p-value and the hypothesis Ho or Ha, respectively. Each instruction is terminated by the ENTER-key.

	A =ANZAHL(A2:A9)	B	C =FVERT(A14; A24-B14-1; B14)	D	E	F =wenn(...
row 24	8		0,0067			Ha
row 25	Anzahl		p-Value			Hypothese

The full wenn-instruction is: =wenn(C24 > 0,05 ; „Ho“ ; „Ha“)

18. Examples of old exams

Written exam **medizinisch-biologische Statistik** (Wahlfach) SS 2006 Name:
Prof. Dr. Stefan von Weber, FB MuV, HS Furtwangen, 80 Punkte

P₁, 1. (10 P) **Conditional Probability and Multiplication of Probabilities:** a) A patient is of

leptosome body constitution, which have 27 % of all patients. He is one of those 7 % of patients which are leptosome **and** suffering under scoliosis. Please, calculate the conditional probability P₁ that a leptosome patient has scoliosis.

b) Please, compute probability P₂ of a patient to be female **and** with scoliosis. 57% of the patients are women.

Answer: a) 25,9% b) 15,4%

2. (15 P) Please, **find outliers** using the **3-σ rule** with the n=17 LDL-Cholesterine values::

145 132 178 138 127 152 157 147 163 204 144 153 166 158 149
128 151

Answer: $\bar{x}=152,47$ $\sigma_{n-1}=18,87$ $u_{\max}=2,73$ $u_{\min}=-1,34$ no outliers

3. (15 P) **Statistical numbers:** Please, calculate from the LDL-Cholesterine values of task 2 \bar{x} , σ_{n-1} , $\sigma_{\bar{x}}$ and the Median and the 95%-confidence interval of the true mean. If you have found an outlier then do not omit it.

Answer: $\bar{x}=152,47$ $\sigma_{n-1}=18,87$ $\sigma_{\bar{x}}=4,58$ Median=151 FG=16
 $t_{\alpha}=2,12$ confidence interval $152,47 \pm 9,71$

4. (20 P) Please, make the χ^2 - test of homogeneity with the following contingency table, giving the frequencies of the *Morbus Scheuermann* depending on hair color and sex.:

hair color	fair	brown	black	
female	$n_{11} = 27$	$n_{12} = 43$	$n_{13} = 30$	
male	$n_{21} = 13$	$n_{22} = 61$	$n_{23} = 76$	

Please, answer the question for independence of the variables "hair color" and "sex"

Answer: $e_{11}=16$ $e_{12}=41,6$ $\chi^2_{11}=7,56$ $\chi^2_{12}=0,05$... $\chi^2_{ges}=18,73$
 $FG=2$ $\chi^2_{\text{alfa}}=5,99$ H_A Color of hair and sex are not
independent variables concerning the frequency of Morbus Scheuermann.

5. (20 P) Please, make the **Mann-Whitney Test**, wheather there is a significant difference in the Alpha2-Globuline-values of electrophoresis of Group1-Patients (purely vegetarian cuisine) and Group2-Patients (mixed food) . The values are assumed to be not normal distributed. The values are already ordered inside the groups:

Group 1: 5.3 5.5 5.5 6.1 6.7 6.8 6.8 7.3 7.5 7.9 7.9 8.4

Group 2: 6.1 6.4 6.9 8.3 8.4 8.5 8.9 8.9 9.8 12.3

Please, answer the question for the significant difference..

Answer: Sum of ranks Gr. 1 is 100 $m=12$ $u_x=98$
Sum of ranks Gr. 2 is 153 $n=10$ $u_y=22$ $U=22$
 $u=-2,51$ $p=0,008 (<0,05)$ and so we accept H_A
We find a significant difference between the Alpha2-Globuline-values

Written exam **medizinisch-biologische Statistik** (Wahlfach) WS 2006/07
Prof. Dr. Stefan von Weber, FB MuV, HS Furtwangen, 80 Punkte

1. (17 P) Experimentant design: One enterprise has to prove for the government that *Hepu-ramol* has a significant influence on the blood pressure. From former investigations one knows that the mean resting error s_R in the regression analysis of blood pressure data is $s_R=5,7$ mmHg . The total variance of all blood pressure data of hypertonic patients was investigated as $\sigma_x^2 = \left[\sum_{i=1}^n (x_i - \bar{x})^2 \right] / (n-1) = \sigma_x^2 = 193,6$ [mm²Hg]. Find the **minimum number n** of probands needed to confirm a 10%-lowering of blood pressure (Regressioncoefficient $b_1 = -0,1$) with $\alpha= 5\%$ error probability.

a) Write down the formula which is giving the sum of squares SAQxx if σ_x^2 is given (See in the script under "Correlation Coefficient" for the formula of SAQxx)

Answer: $SAQ_{xx}=193,6 (n-1)$ (with $n=70$ according to part b) we find) $SAQ_{xx}=193,6 \cdot 69$

b) Please, insert the above values in the formula of the t-test for regression coefficient b_1 and find with different numbers n of probands the significance. Find the smallest n delivering a significant t-value. (Please, start with $n=70$ and decrease n successively. Notify at each step n , t , t_α)

Answer: $n=70 \quad t = \frac{0,1}{5,7} \sqrt{193,6 \cdot 69} = 2,03 \quad t_\alpha = 1,99 \quad \text{significant}$
 $n=68 \quad t=2,00 \quad t_\alpha = 1,99 \quad \text{significant}$
 $n=67 \quad t=1,98 \quad t_\alpha = 1,99 \quad \text{not significant}$
 One needs 68 probands

2. (21 P) Test the Normal distribution with χ^2 -godness-of-fit test: We have 25 HDL-values

174	157	133	107	148	144	140	154	137	150
116	147	136	156	130	182	167	128	163	162
121	161	202	159	191					

a) Please, compute \bar{x} and σ_{n-1} Answer: $\bar{x} = 150,6 \quad \sigma_{n-1} = 22,74$

b) Make the χ^2 -godness-of-fit test, i.e., compute the boundaries of the histogram, count the class frequencies using the data above, compute the expectations and the χ^2_i -values and χ^2_{gesamt} , df, and find the critical value χ^2_{α} . Accept one hypothesis and formulate a sentence concerning the distribution of the data.

Answer: boundaries	$-\infty$	131,5	144,9	156,3	169,7	∞	
Frequencies	5	5	5	6	4		
Expectations	5	5	5	5	5		
χ^2_i	0	0	0	0,2	0,2		$\chi^2_{\text{ges}}=0,04$
df=3		$\chi^2_\alpha=7,81$		Ho			Normal distribution accepted

3. (20 P) Mann-Whitney-Test: We have to groups A and B of times of blood coagulation. The type of data distribution is unknown. The group-B-patients were treated with *Fibrilanol*. Make the Mann-Whitney-Test for mean difference:

a) A: 4,3 4,9 3,7 2,3 1,7 4,3 5,1 2,4 4,4 6,3
 B: 0,9 2,2 1,8 2,7 1,8 1,8 1,4 2,3 2,5 3,7
 Firstly, make the pooled ranks for all values in one long row

Answer: 1 2 3 5 5 5 7 8.5 8.5 10 11 12 13.5 13.5 15.5 15.5 17 18 19 20

b) Write down the pair of hypotheses and then compute the two sums of ranks of the groups

Answer: $H_0: \mu_1 = \mu_2 \quad H_A: \mu_1 \neq \mu_2 \quad R_x=140 \quad R_y=70$

c) Test with the u-formula for large n and m using $\Phi(u)$ -table, accept your hypothesis and formulate a sentence concerning the mean difference of the blood coagulation times.

Answer: $m=10 \quad n=10 \quad u_x=15 \quad u_y=85 \quad U=15 \quad u=-2,65 \quad u_\alpha = 1,96 \quad p=\Phi(u) = 0,001$

Because of $p < 0,05$ we accept H_A . The blood coagulation times after treatment with Fibrilanolgabe are significantly smaller.

4. (22 P) Test of the mean difference of normal distributed populations: We have blood pressure data of rats measured before and after loud music:

Before: 114 117 116 121 119 122 118 --- 126 123
 After: 122 119 115 124 --- 123 121 121 129 ---
 (2 rats died by heart attack, one rat lost the measuring pipe during first measuring)

a) Make the t-test for unpaired values as in example 19 of the lecture, i.e., give the two hypotheses, the two means, averaged Sigma, t-test, your decision and a short sentence

Answer: $\bar{x}_1 = 119,55$ $\sigma_{n1}^2 = 12,69$ $n_1 = 9$ $\bar{x}_2 = 121,75$ $\sigma_{n2}^2 = 14,1875$ $n_2 = 8$
 $\bar{\sigma} = 3,89$ $t = -1,162$ $FG = 15$ $t_\alpha = 2,13$ We accept H_0
 There is no significant difference of the blood pressure data of rats measured before and after loud music

b) Make the t-test for paired data (example 21) with the 7 complete pairs, i.e., give the two hypotheses, the 7 differences, mean value of difference, sigma, t-test, your decision and a short sentence.

Answer: $n = 7$ Differences: 8 2 -1 3 1 3 3 $\bar{d} = 2,714$ $\sigma_{n-1} = 2,752$ $t = 2,609$
 $FG = 6$ $t_\alpha = 2,45$ We accept H_A
 There is a significant difference of the blood pressure data of rats measured before and after loud music

c) Which of the two test is giving here the better result? Answer: the t-test for paired data

Written Exam Biomedical Statistics SS 2009
 Doz. Dr. S. v. Weber, FB MuV, HFU, Total 73 points

Name, Semester:
 Matrikelnummer:

1.) (10 P) From an inquiry concerning the drinking behaviour of patients the following question arose: Is there a difference between men and women concerning the consumption of alcohol, juices or hot beverages? (Drinking type) The contingency table is:

		Drinking type		
		Alkohol	Juices	hot beverages
Sex	m	84	23	42
	w	27	82	54

Please, test if there is a significant association between the variables *sex* and *drinking type*. (Hypotheses, e_{ij} , χ_{ij}^2 , χ_{Gesamt}^2 , chose a hypothesis, give an answering sentence)

Answer: $e_{11} = 53,0$ $e_{12} = 50,1$ $\chi^2_{11} = 18,13$ $\chi^2_{12} = 14,66$... $\chi^2_{ges} = 63,3$

FG=2 $\chi^2_{\text{alfa}}=5,99$ H_A We have a significant association between the variables *sex* and *drinking type*.

2.) (8 P) Please, make the comparison of relative frequency numbers for the pair of numbers from the table above, column 1 (alcohol). Here n_1 is the marginal sum of row 1, n_2 is the marginal sum of row 2 of the table (Hypotheses, p, q, t, chose a hypothesis, give an answering sentence)

Answer: $H_0: p_1=p_2$ $H_A: p_1 \neq p_2$ $h_1=84$ $h_2=27$ $n_1=149$ $n_2=163$
 $\hat{p}_1 = 84/149=0,564$ $\hat{p}_2 = 27/163=0,166$ $\bar{p} = 111/312=0,356$ $q=0,644$
 $df=310$ $t=7,33$ $t_\alpha = 1,96$ We accept H_A
 Men drink significantly more often alcohol than women.

3. (15 P) There are given $n=28$ β -globulin values of 10 female and 18 male patients.

137	162	182	279	191	187	244	143	169	172	336	233	155	175
174	183	88	151	306	191								
					102	161	206	274	167	173	183	241	

Please, compute from all numbers (all 28) the mean \bar{x} , the standard deviation σ_{n-1} , the error of the mean $\sigma_{\bar{x}}$, the 95%-confidence-intervall of the true mean, the median and the range (MaxMin). (Examples 11, 7, 8)

Answer: $n=28$ $\bar{x}=191,6$ $\sigma_{n-1}=56,8$ $\sigma_{\bar{x}}=10,73$ Median=178,5 FG=27
 $t_\alpha = 2,06$ confidence interval $191,6 \pm 22,1$ range =336-88 =248

4. (15 P) Please, make using the mean \bar{x} , the standard deviation σ_{n-1} and the data from task 3 the test for normal distribution (Example 17). (Hypotheses, class boundaries, class frequencies, expectations, χ^2_i , χ^2_{Ges} , chose hypothesis, give an answering sentence)

Answer: boundaries $-\infty$ 143,9 177,4 205,8 239,3 ∞
 Frequencies 4 10 6 2 6
 Expectations 5,6 5,6 5,6 5,6 5,6
 χ^2_i 0,457 3,457 0,029 2,314 0,029 $\chi^2_{ges}=6,286$
 $df=3$ $\chi^2_\alpha=7,81$ H_0 Normal distribution acceptet

5. (10 P) How many of 5.000 patients are expected to have a β -globulin value of $x>250$, if one uses mean \bar{x} , standard deviation σ_{n-1} from task 3 ? With which β -globulin value the 25%-quartile of the „few loaded“ patients is ending? (quantile X_{25} , example 11)

Answer: a) $u=1,028$ $\Phi(-u)=0,1587$ $E= N p =5000 \cdot 0,1587= 793$ Patients
 b) $p=0,25$ $u=-0,6$ $x=157,5$

6. (15 P) Please, make the test for a different load of β -globulin of women and men using the data of task 3 (example 19) with the t-test from chapter 11.2 (Comparison of two normally distributed populations).

Answer: $H_0 : \mu_1 = \mu_2$ $H_A : \mu_1 \neq \mu_2$

$$\bar{x}_1 = 185,3 \quad \sum x^2 = 380405 \quad n_1 = 10 \quad \text{SAQ1} = 37044,1$$

$$\bar{x}_2 = 195,1 \quad \sum x^2 = 734640 \quad n_2 = 18 \quad \text{SAQ2} = 49409,6$$

$$\bar{s} = 57,66 \quad t = -0,43 \quad \text{FG} = 26 \quad t_\alpha = 2,06 \quad \text{We accept } H_0$$

There is no significant difference in the β -Globulin data of women and men.

19. List of the examples given in the lecture

1. Experimental design	15. χ^2 -analysis by Lancaster
2. Conditional probabilities	16. Contingency measures
3. Histograms	17. χ^2 -test of fit for a distribution
4. Moments of distributions	18. One-sample t-Test
5. Poisson distribution	19. a) Comparison of two normally distributed populations (equal variances)
6. Binomial distribution	19 b) Welch-Test (unequal variances)
7. Normal distribution of data	20. Mann-Whitney-Test
8. Confidence intervalls	21. Paired t-Test
9. 3-Sigma-rule	22. Wilcoxon-Test
10. Indexing	23. Product-Moment Correlation r
11. Statistical numbers	24. Linear Regression
12. Test frequency against constant	25. VA cross classification
13. Test two relative frequencies	
14. Contingency- or Homogeneity test	

Pay attention: The numbers used in the examples are mostly fictious