EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

FACULTY OF
SCIENCE

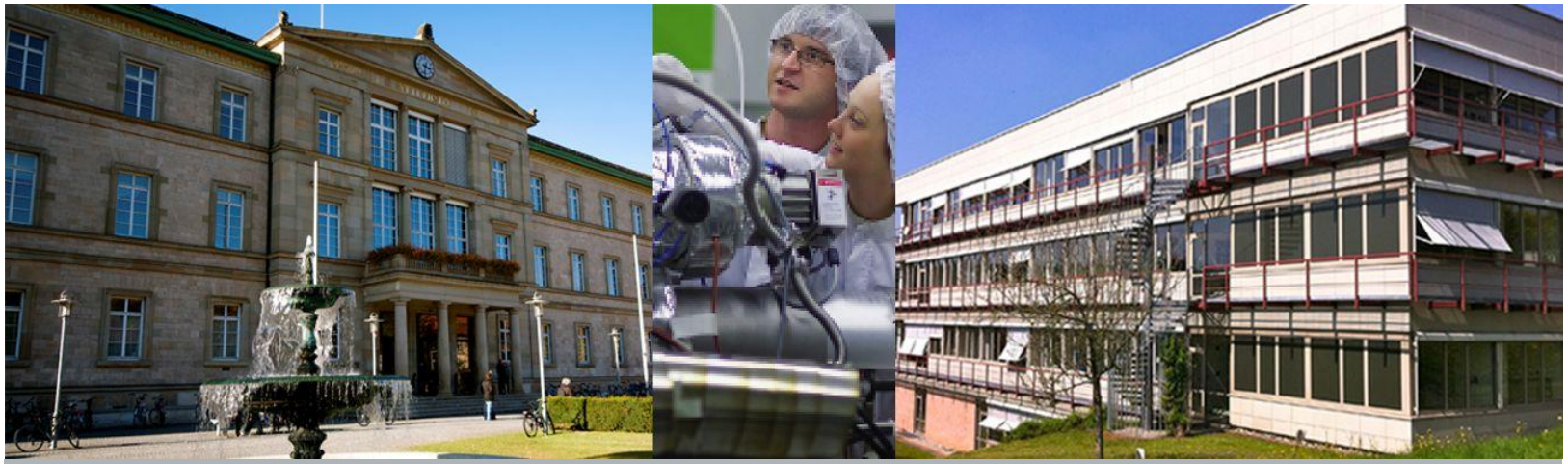
Institute of Physical and Theoretical Chemistry



Laboratory Techniques and Medical Device Approvals I

Part 01: Statistics

 © 2012 University of Tuebingen



Statistical terms

Basics and repetition

 © 2012 University of Tuebingen

Table of Contents

- Statistical terms (Part 01)
  - Basics
  - Gaussian distribution as an example
  - Density and distribution functions in general
  - Point estimators
  - Interval estimators
  - Robust / median statistics
  - Significance tests

- Multivariate Data Analysis (Classification) (Part 02)
  - Polar plot
  - Cluster analysis
  - Principle Component Analysis (PCA)

© 2012 University of Tuebingen

Table of Contents

- Mathematical Basics (Part 03)
  - Matrix calculations

- Multivariate Data Analysis (Quantification) (Part 04)
  - Multi Linear Regression (MLR)
  - Principle Component Regression (PCR)

© 2012 University of Tuebingen

Why statistics?

Why statistics?

- It helps to analyse the measured data
- Get to know similarity based classification methods
- Check the possibility of data reduction, which is the basis for validation (e. g. process validation)
- Optimization of experimental effort: Experimental planning
- Obtain relevant and reliable data
- It enables inter-laboratory comparisons
- Quality control

© 2012 University of Tuebingen

Definition of chemometrics

„Chemometrics can be defined as the chemical discipline that uses mathematical, statistical, and other methods employing formal logic

- to design or select optimal measurement procedures and experiments, and
- to provide maximum relevant chemical information by analyzing chemical data“ [FRANK and KOWALSKI 1982]

Linking elements between chemistry, mathematics, and hardware and software are used

- in sampling
- quality control
- experimental planning
- to optimize and model processes and experiments
- to evaluate and to interpret data
- to extract a maximum of chemical information from experimental data (cluster analysis, factor analysis, pattern recognition)
- spectral libraries

 © 2012 University of Tuebingen

Terms in statistics

sorted data: trends

17	96	74	42	82	56
37	7	35	33	64	19
69	73	22	39	96	52
17	45	58	19	99	51
40	47	93	7	67	46

histograms

7	22	40	64	82
7	33	42	67	93
17	35	45	69	96
17	37	46	73	96
19	39	47	74	99
19		51		
		52		
		56		
		58		

Data sorting has

- advantages: Appearance of accumulations
- disadvantages: loss of trends

© 2012 University of Tuebingen

Sorted tables displayed as histograms



 © 2012 University of Tuebingen

Terms in statistics

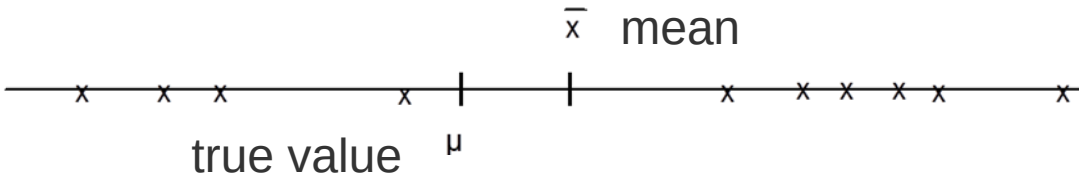
- Analytically measured quantities are **random variables**
- Random variables are belonging to an infinite amount of possible random variables, which is named **population**
- Out of this population a certain set of random variables is named **sample.**



 © 2012 University of Tuebingen

Terms in statistics

- The probability of occurrence of a random variable can be described by a **distribution function**. This is the distribution of the random variable around the **expected value** of a population. The expected value should be close to the **true value (μ)**.
- One tries to represent the population with an **unbiased (representative) sample** in order to reduce the data collection time and costs
- The probability of appearance of random variables is following a normal (gaussian) distribution

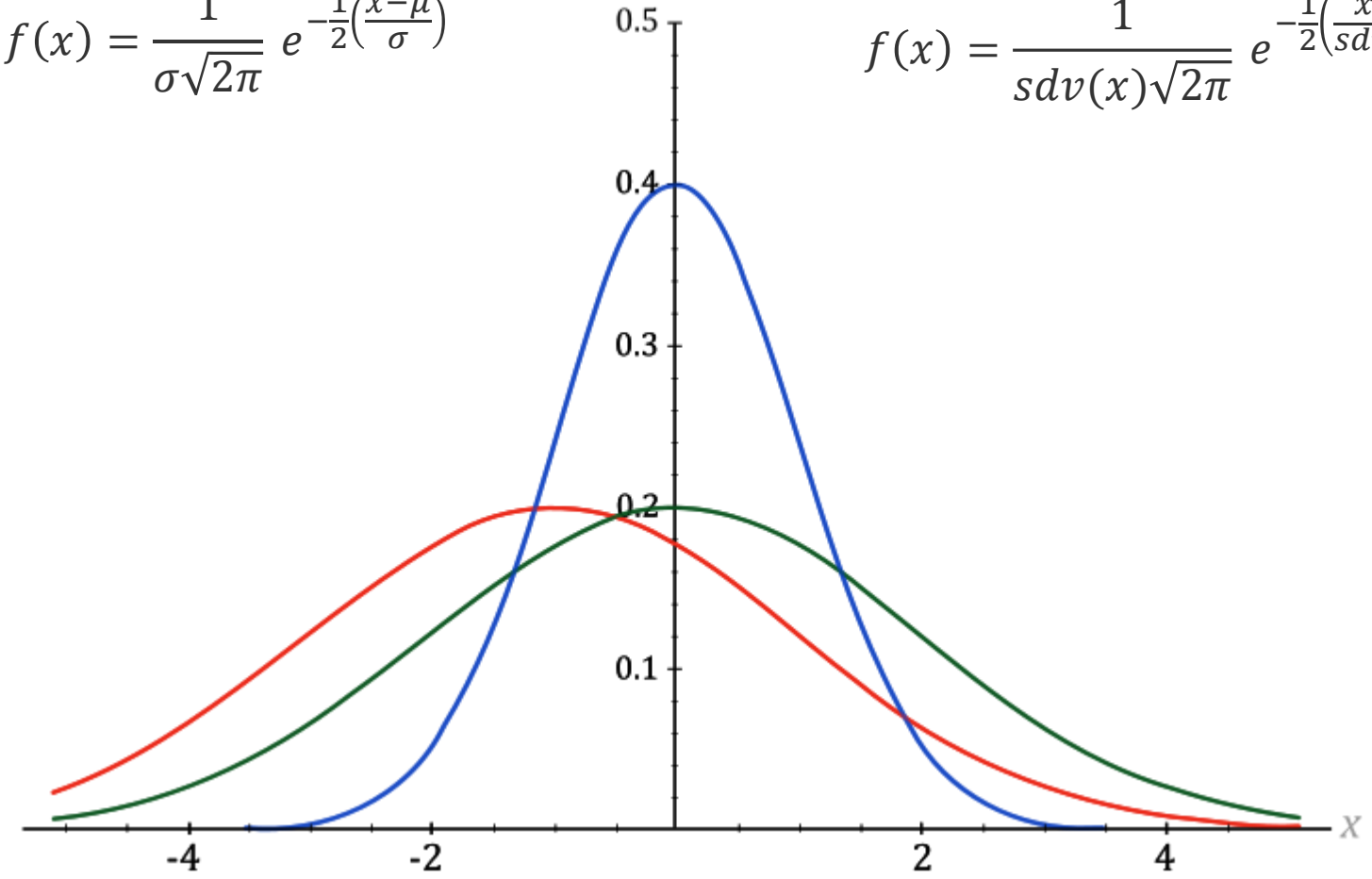


 © 2012 University of Tuebingen

Normal or Gaussian distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$f(x) = \frac{1}{sdv(x)\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{sdv(x)}\right)^2}$$



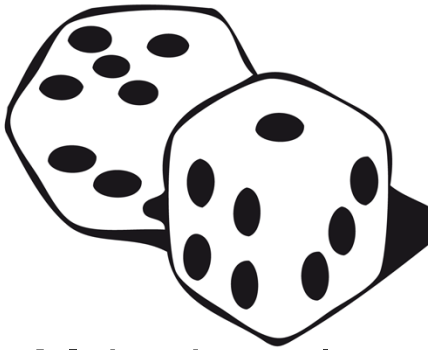
 © 2012 University of Tuebingen

Random variable

A random variable is a numerical value out of measured value or result. The value of the random variable can vary from trial to trial as the experiment is repeated.

A random variable can have

1. discrete values (**discrete random variable**)
   (E. g. dice game, display of a digital watch)



2. continuous value (**continuous random variable**). The values vary continuously within one or more real intervals.

© 2012 University of Tuebingen

Random variable

3. **Continuous fluctuating values** around a certain value. The probability of occurrence decreases with increasing distance, but the value 0 is never reached (theoretically). (Example: Readout of an analog measuring instrument.



 © 2012 University of Tuebingen

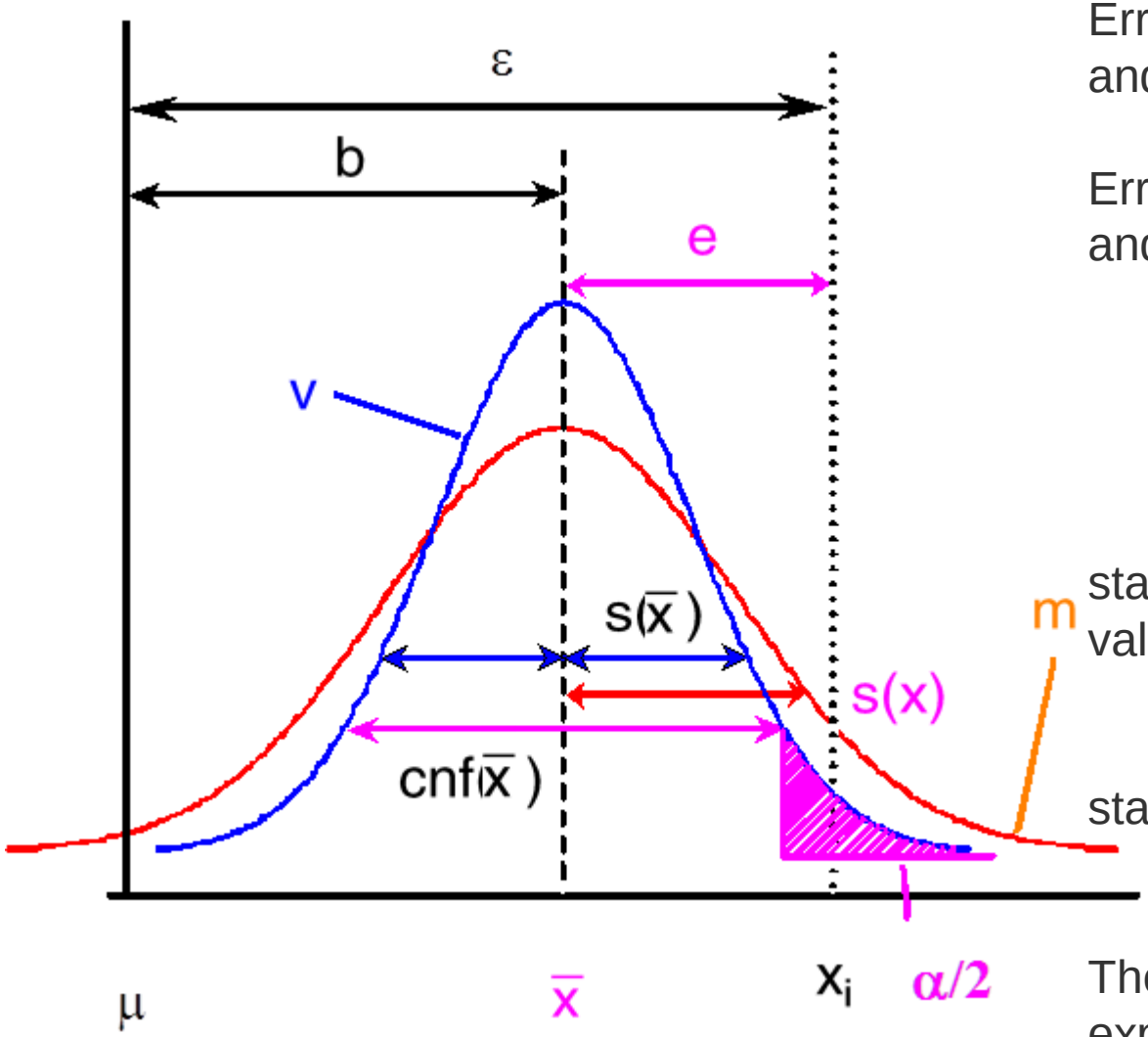
Random variable

- **Measured value:** measured, observed or read out value
- **Result**: A result is obtained after finishing an analysis. This analysis includes the completion of all measurements and all succeeding evaluation steps.
- **Mean value** $(\overline{x})$ is used for the characterization of discrete distributions.
- **Expected value** is used for the characterization of discrete and continuous distributions.

- The principal problem is, that no measured value has to be “true”. Because of this the statistically calculated mean does not have to be equal to the **true value μ**.

© 2012 University of Tuebingen

Normal or Gaussian distribution



Error between a random value x_i and the mean value $\bar{x}$

$$e_i = x_i - \bar{x}$$

Error between a random value and the true value μ

$$\varepsilon_i = x_i - \mu$$

$$\varepsilon_i = e_i + \left(\bar{x} - \mu\right)$$

$$\varepsilon_i = x_i - \bar{x} + \left(\bar{x} - \mu\right)$$

standard deviation of the single value

$$\mathrm{sdv}(x)$$

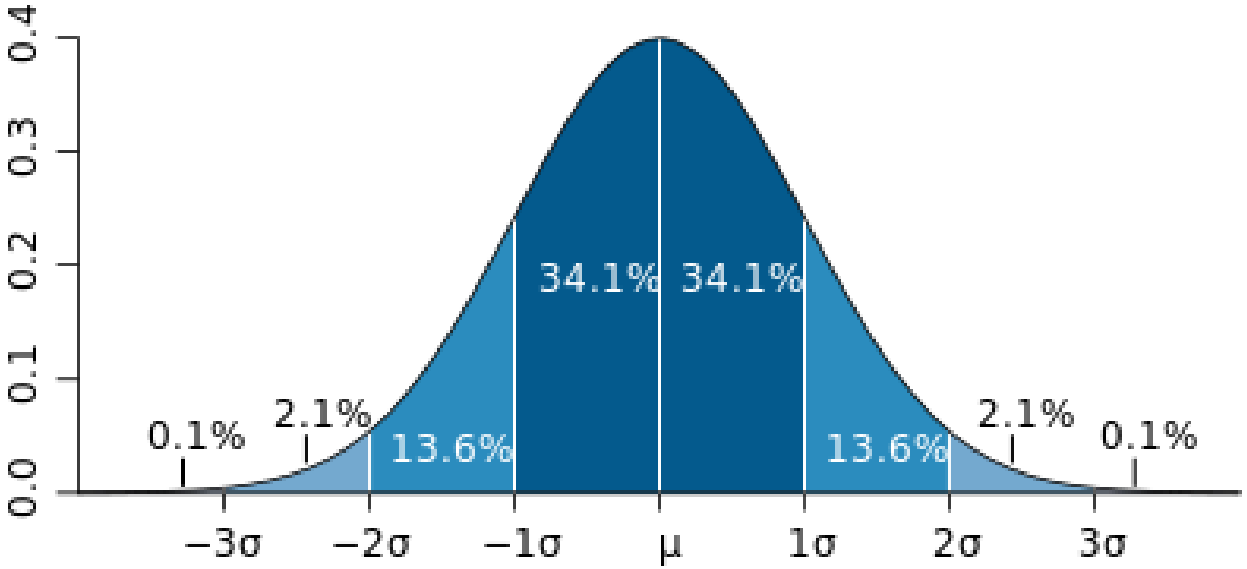
standard deviation of the mean

$$\mathrm{sdv}(\bar{x})$$

The other parameters will be explained later on.

 © 2012 University of Tuebingen

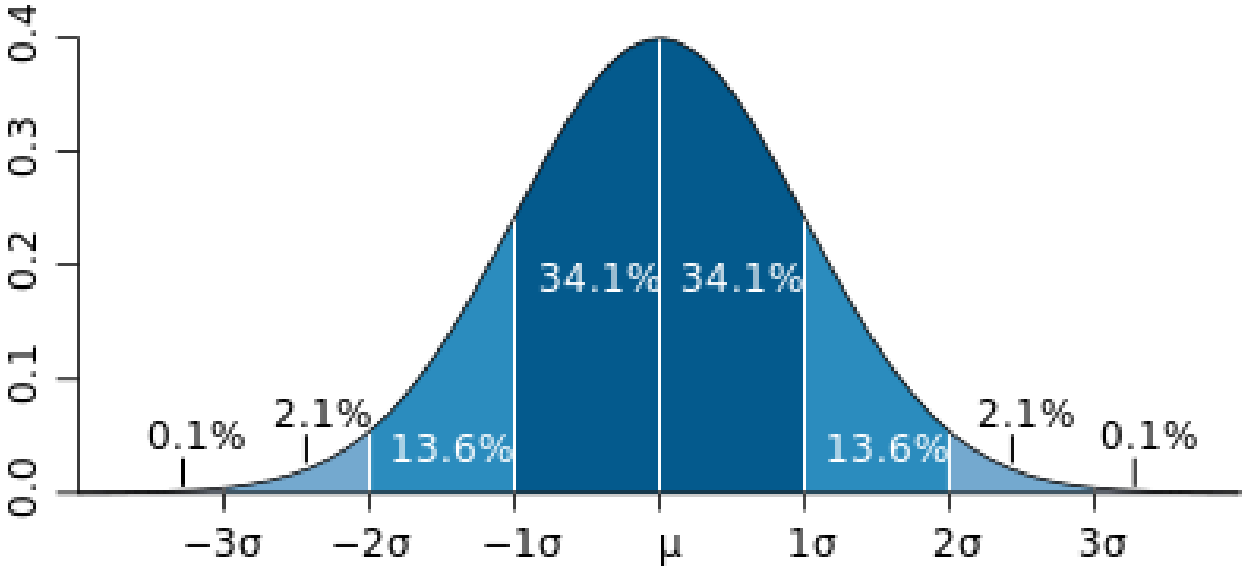
Normal or Gaussian distribution



- Dark blue is less than **one standard deviation (± σ)** away from the mean. For the normal distribution, this accounts for about **68,27%** of the set
- while **two standard deviations (± 2σ)** from the mean (medium and dark blue) account for about **95,45%**
- and **three standard deviations (± 3σ)** (light, medium, and dark blue) account for about **99.73%**

 © 2012 University of Tuebingen

Normal or Gaussian distribution



For a given probability you can also calculate the maximum deviation from the mean:

- 50 % of all measured values have a deviation from the mean value of maximum $0.675\,\sigma$.
- 90 % of all measured values have a deviation from the mean value of maximum $1.645\,\sigma$.
- 95 % of all measured values have a deviation from the mean value of maximum $1.960\,\sigma$.
- 99 % of all measured values have a deviation from the mean value of maximum $2.575\,\sigma$.

 © 2012 University of Tuebingen

Types of errors

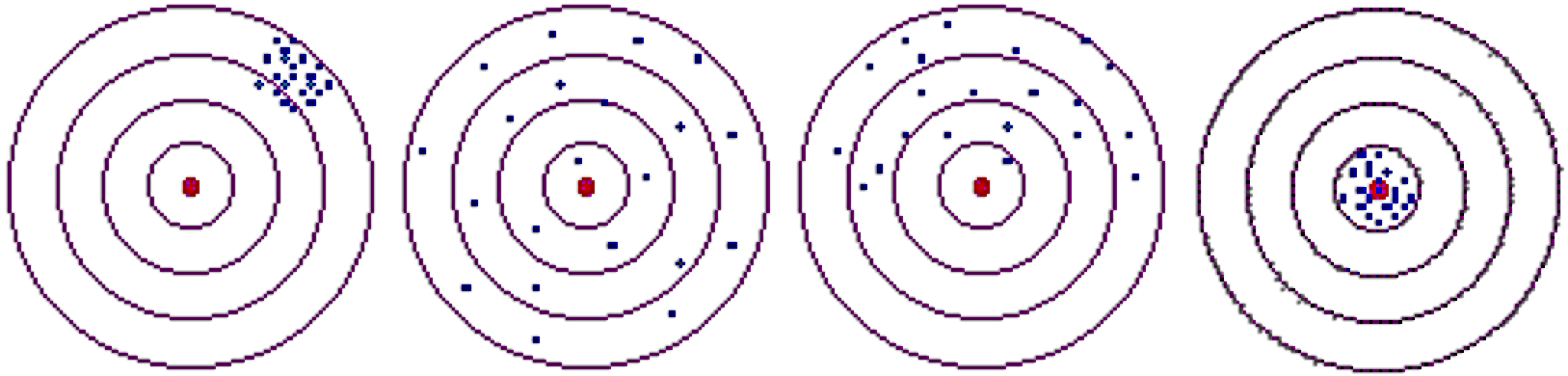
Error:
The results are distinguished from the true value by an **error**. In contrast to the error, the **deviation** is the difference between a random variable and the mean within the same sample.

Range:
The range is defined as the difference between the largest and smallest value within a sample.
Besides the **gross errors** (e.g. wrong type of measuring tape) two types of errors are important:

- **Random (statistical) error**
- **Systematic error**

© 2012 University of Tuebingen

Types of errors



Statistical:	Small	Large	Large	Small
Systematic:	Large	Small	Large	Small

- **Random error:** The statistical error effects the **precision (reproducibility)** of an analytical process. It reflects the fact that every time you make a measurement, you measure a slightly different quantity each time. They are nonavoidable and can not be characterized.
- **Systematic error (bias):** It gives the accuracy of an analytical process. It is the deviation of the results from the true value in the *same sense*. The systematic error is caused by e.g. defective instruments, wrong measurement methods and gross errors. They can be observed by **interlaboratory tests**, but not by repeated measurements.

 © 2012 University of Tuebingen

Types of errors

- **Gross errors**
  - should be avoided

- **Systematic errors** ***(bias)***
  - influence the accuracy of the mean and the trueness
  - can be observed by interlaboratory tests, only.

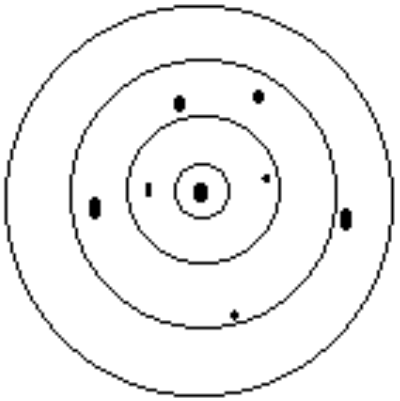
- **Statistical** ***(random)*** **errors** have influence on the
  - precision
  - reproducibility

© 2012 University of Tuebingen

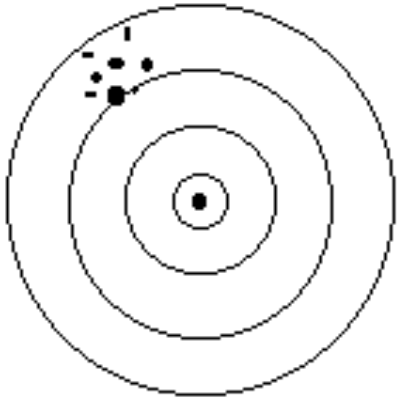
Trueness, precision, accuracy



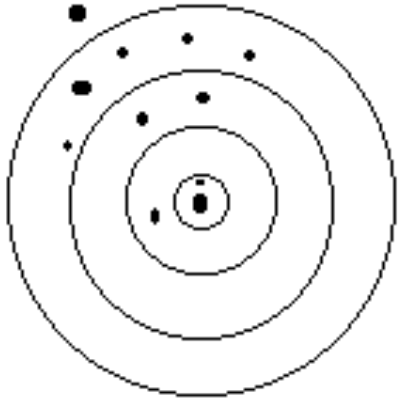
High trueness
High precision
=> accurate



High trueness
Low precision



Low trueness
High precision



Low trueness
Low precision

According to ISO 5725-1:

- **Trueness** and **precision** are used to describe the **accuracy** of a measurement.
- **Trueness** refers to the closeness of the mean of the measurement results to the actual (true) value
- **Precision** refers to the closeness of agreement within individual results.
- Therefore, according to the ISO standard, the term "**accuracy**" refers to both trueness and precision.

© 2012 University of Tuebingen

Terms used in the probability calculation

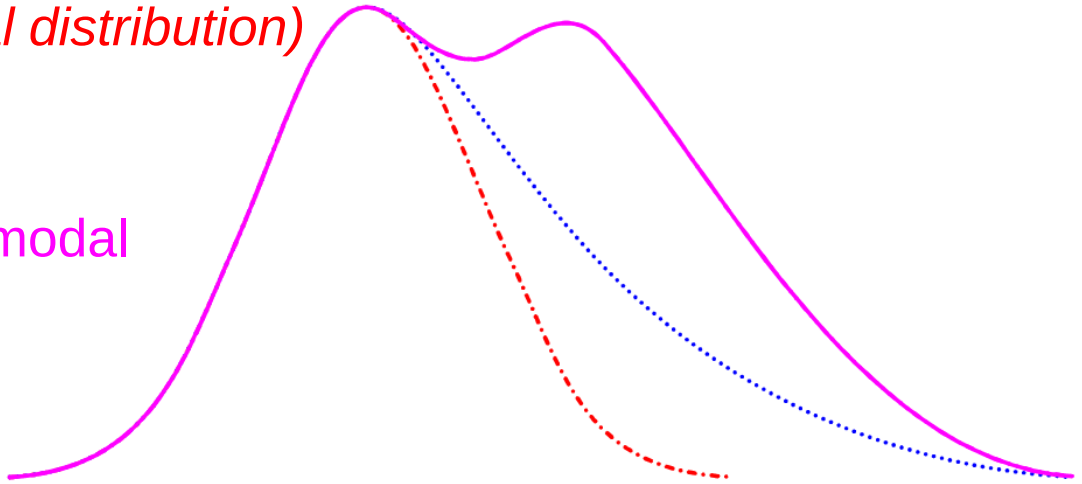
- The obtained data are listed in a data matrix
- Out of the data matrix, the **frequency** (absolute number of a certain occurrence) and the **relative frequency** (absolute number of a certain occurrence in relation to all occurrences) can be calculated.
- The **absolute frequency** is the number of data. It requires **stochastically independent** measurement data (one measurement must not depend on the previous measurement).
- If all occurrences have the same probability, then it is called **equiprobability** (e. g. a certain number on a dice).

© 2012 University of Tuebingen

Density and Distribution function

- The **probability density function** gives the frequency of equal occurrences. In general they are also called distribution curves.
- In the case of discrete distributions, the probability density function is also named **probability function**.
- In contrast to the probability density function, the **cumulative distribution function** is giving the relative frequency.

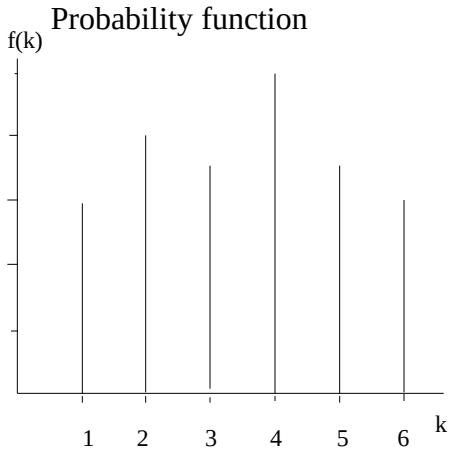
- Types of distribution curves:
  - *Symmetric (normal distribution)*
  - *Skew*
    - Unimodal
    - Bimodal/multimodal

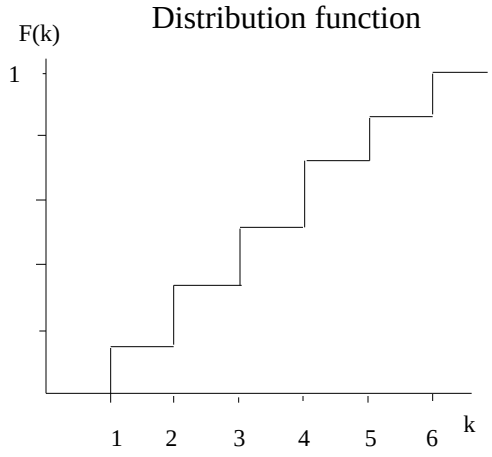


 © 2012 University of Tuebingen

Density and Distribution function

- Example: discrete uniform distributions: Probability and distribution functions





- Transformations (linear and nonlinear)
- Convolution: triangular distribution

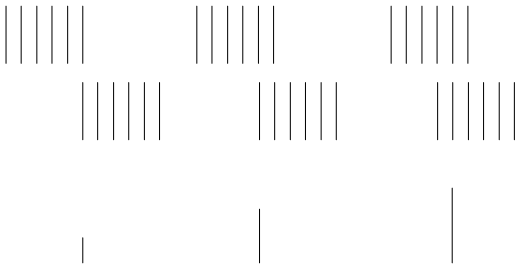
$$h(k) = \sum_{i,\,j} f(i) \cdot g(j)$$

$$i = 1, 2, 3 \ldots n$$

$$j = n, n-1, n-2 \ldots 3, 2, 1$$

$$i + j = n + 1$$

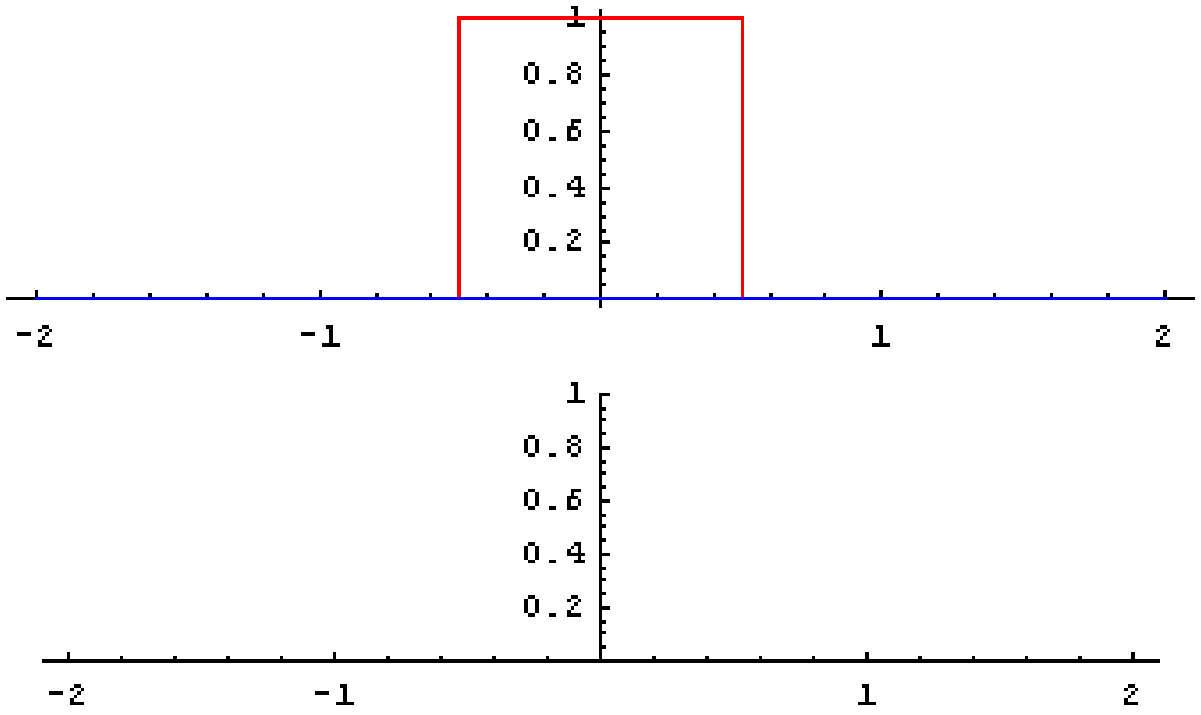
The triangular distributions is typically observed in the case of a monochromator. It is caused by the superposition of single slit diffraction.



 © 2012 University of Tuebingen

Convolution

Convolution is a mathematical operation on two functions *f* and *g*, producing a third function:



 © 2012 University of Tuebingen

Binomial or Bernoulli distribution

- The Binomial or Bernoulli distribution is giving the probability that a certain event occurs x-times within n independent conducted experiments. The probability of a single event is p.
- The binomial distribution is the most important discrete distribution. For p = 0.5 it is symmetrical, otherwise skew.
- In every experiment only two types of results are possible. The total probability of these results has to be one.

- Example: Head or tail of a coin

$$f(x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$$

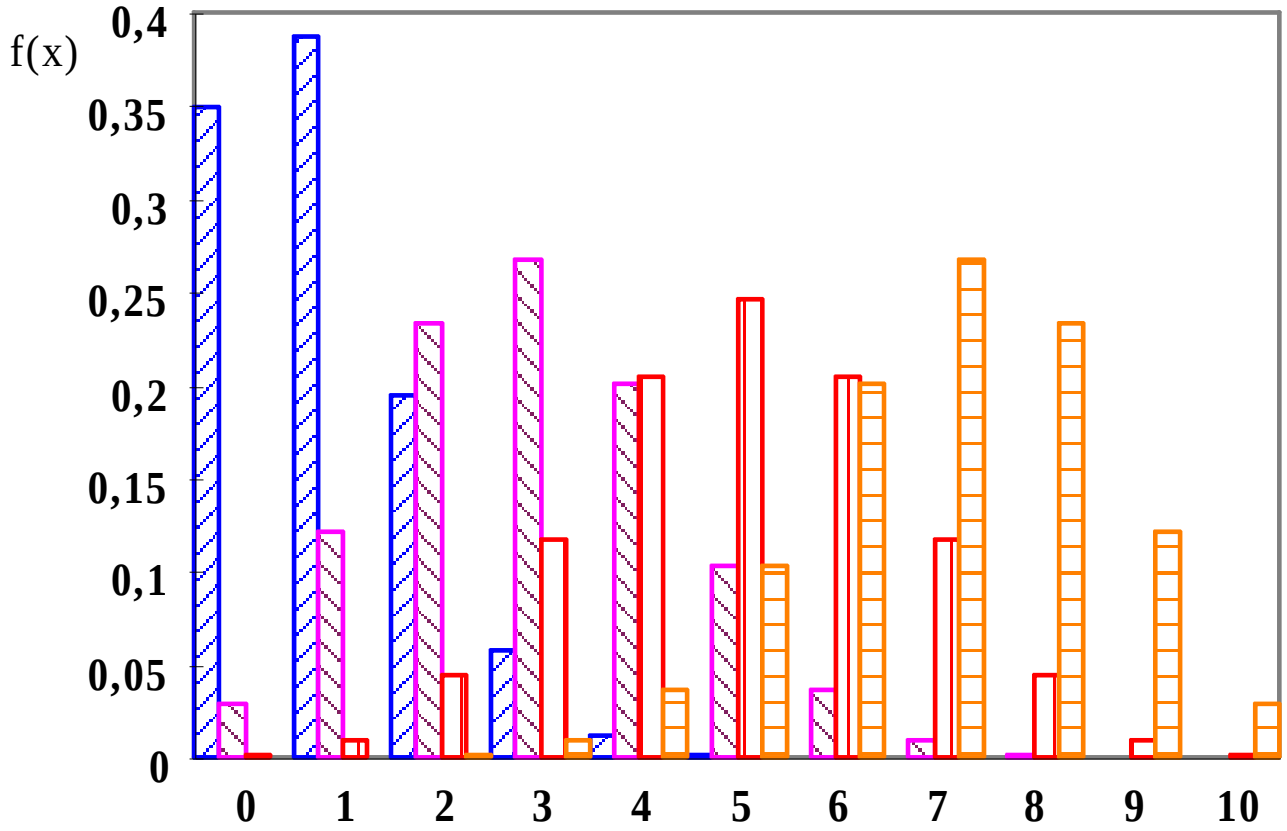
n = independent experiments
x = number of positive events
p = probability (of a single event)

Binomial coefficient: $h_i = \binom{i+1}{i-1}$

© 2012 University of Tuebingen

Binomial or Bernoulli distribution

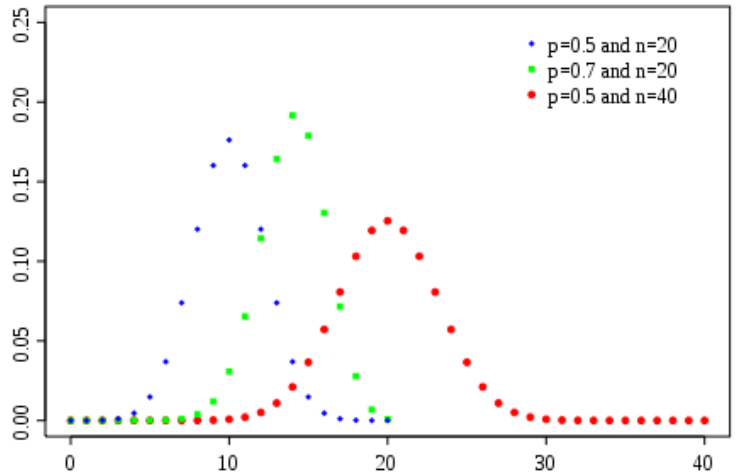
Binomial distribution for different probabilities p



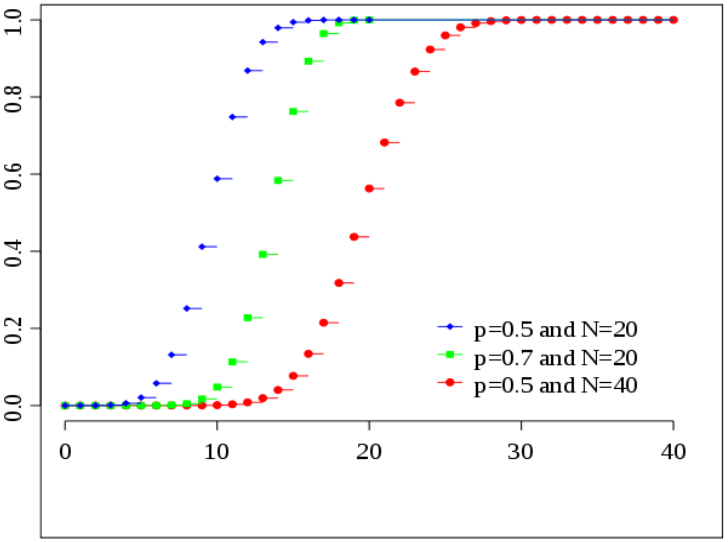
$$f(x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$$

 © 2012 University of Tuebingen

Binomial or Bernoulli distribution



Probability density function

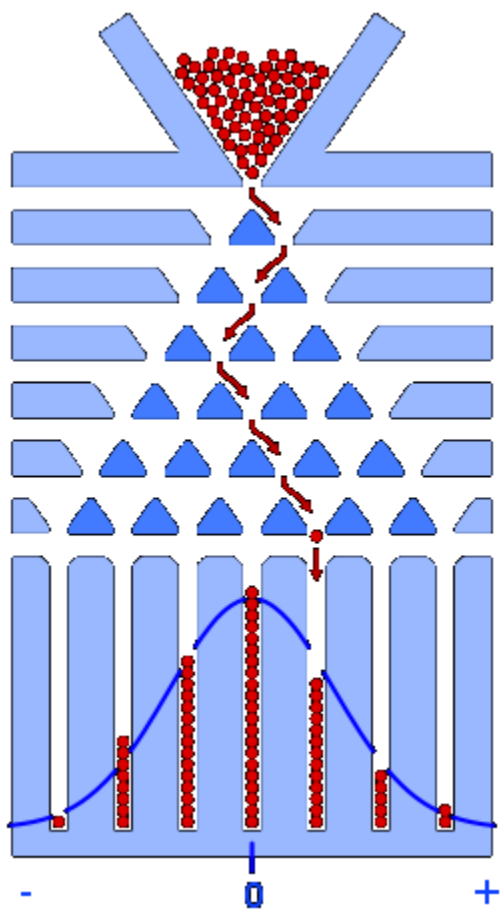


Distribution function

Source: Wikipedia

© 2012 University of Tuebingen

Demonstration of the binomial distribution: Galton box

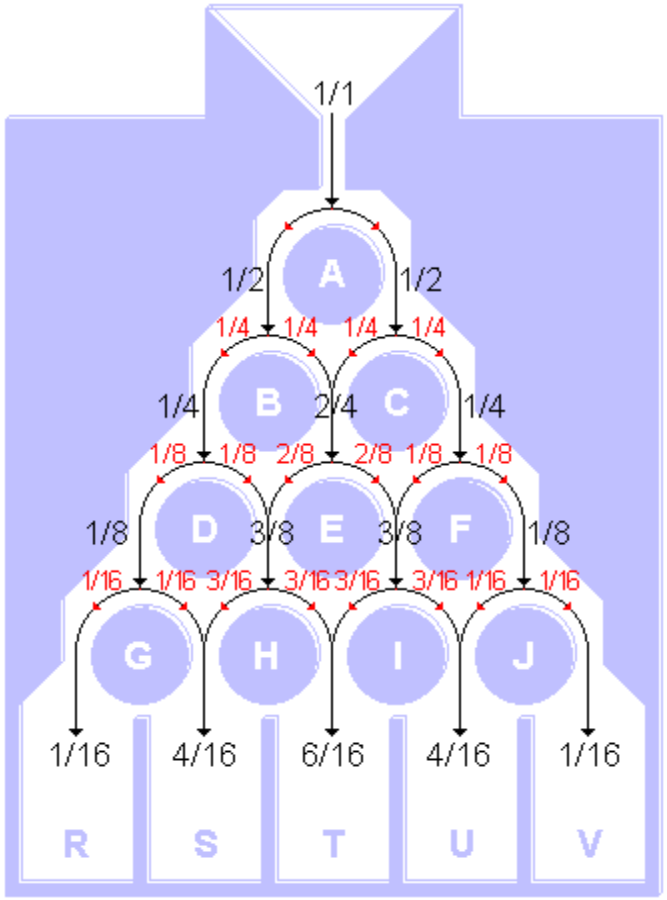




Source: wikipedia

 © 2012 University of Tuebingen

Demonstration of the binomial distribution: Galton box



p=0.5
Pascal's triangle

Source: wikipedia

 © 2012 University of Tuebingen

Demonstration of the binomial distribution: Galton box



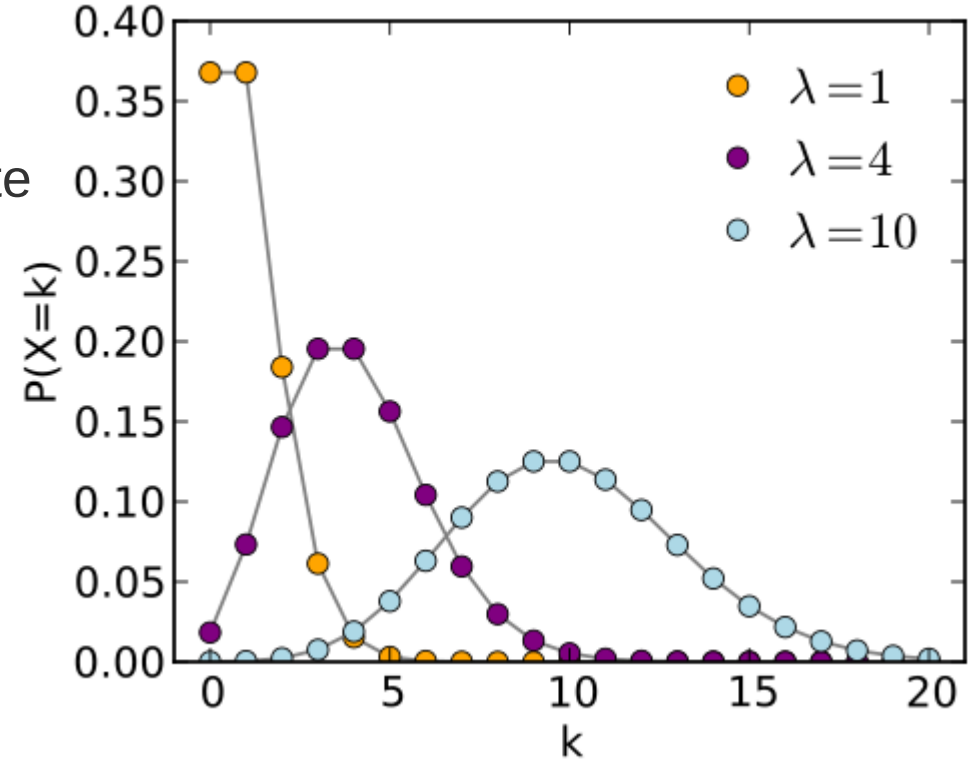
 © 2012 University of Tuebingen

Poisson distribution

- The Poisson distribution can be applied to systems with a large number of possible occurrences (k > 100), each of which has a small probability (p < 0.01)
- The product of possible events and probability should be constant.
- If n is large and p = 0.5, then the Poisson distribution is a normal distribution.
- λ is equal to the expected value ($E(X)$) of the discrete random variable X

$$\lambda = E(X)$$

$$p(\mathrm{k}, \lambda) = \frac{e^{-\lambda} \cdot \lambda^{k}}{\mathrm{k!}}$$

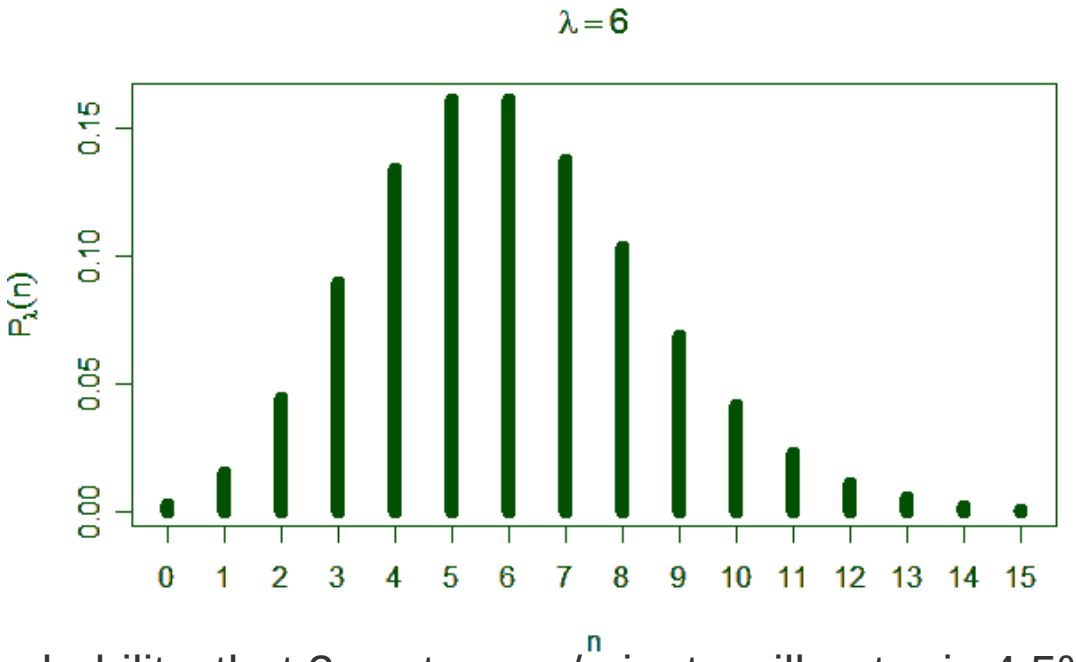


 © 2012 University of Tuebingen

Poisson distribution: Example 1

On Saturdays a store is entered every 10 seconds (t_1) by a customer. The number of customers entering in a one minute period is 6 (λ = 6 customers/minute). $P_6(n)$ is the probability for n customers in the next minute (t_2).

$P_6(n)$			
n	probability	in %	sum in %
0	0,25		0,25
1	1,49		1,74
2	4,46		6,20
3	8,92		15,12
4	13,39		28,51
5	16,06		44,57
6	16,06		60,63
7	13,77		74,40
8	10,33		84,72
9	6,88		91,61
10	4,13		95,74
11	2,25		97,99
12	1,13		99,12
13	0,52		99,64
14	0,22		99,86
15	0,09		99,95



The probability, that 2 customers/minute will enter is 4,5%.
The probability, that 0 to 9 customers will enter is 92% (Sum).
The probability, that more than 9 customers will enter is 8%

Quelle: Wikipedia

 © 2012 University of Tuebingen

Poisson distribution: Example 2

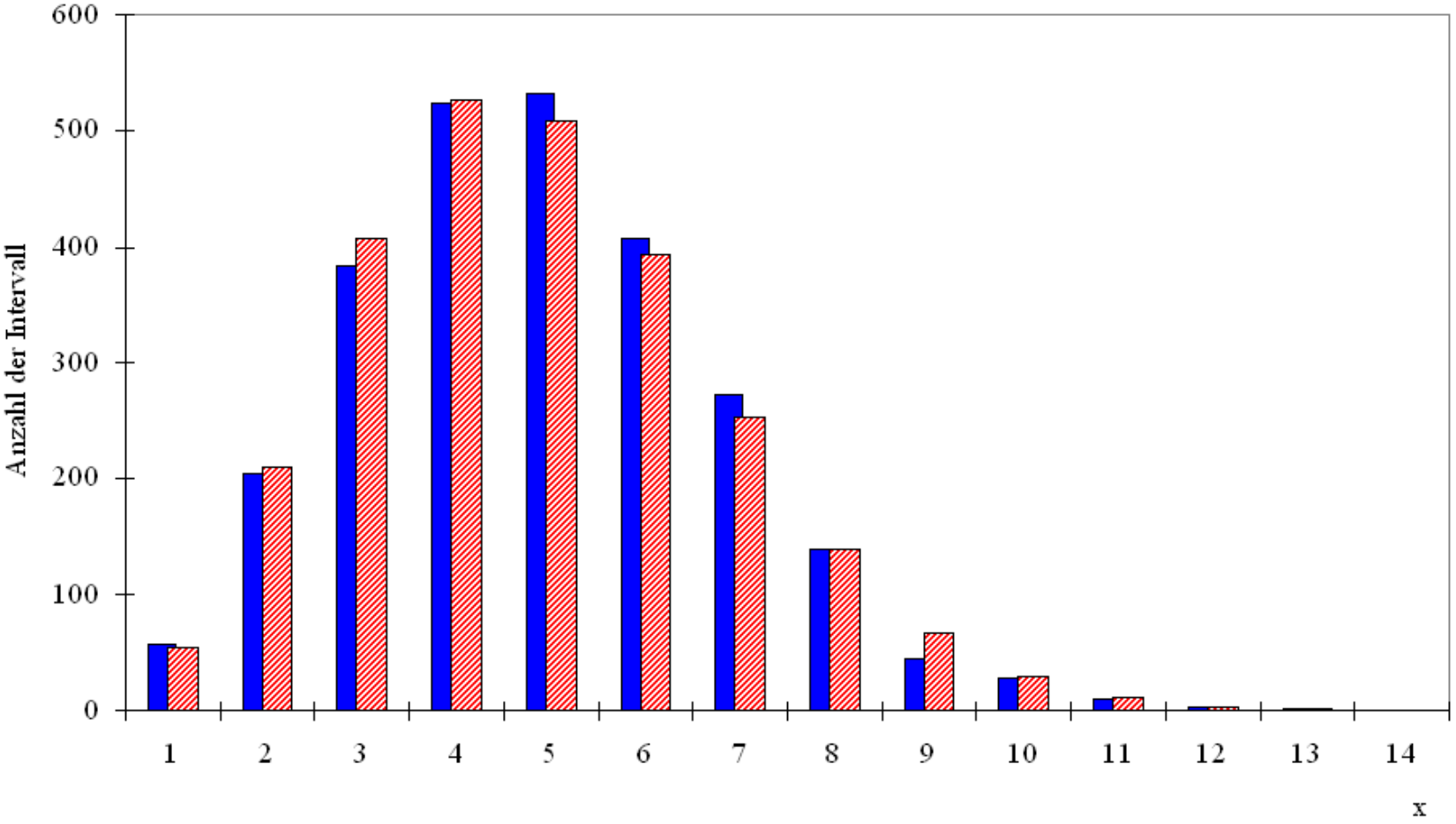
Rutherford: α-decay

x	Delta	Calculated	
0	57	54,37693908	0
1	203	210,4604383	203
2	383	407,2829111	766
3	525	525,4490931	1575
4	532	508,4243813	2128
5	408	393,5610203	2040
6	273	253,873015	1638
7	139	140,3699723	973
8	45	67,91097105	360
9	27	29,20472656	243
10	10	11,30339378	100
11	4	3,977149221	44
12	2	1,282762789	24
13	0	0,381907963	0
Sum	2608		10094
Weighted mean			3,870398773

© 2012 University of Tuebingen

Poisson distribution: Example 2

Poisson-distribution:α-decay



 © 2012 University of Tuebingen

Derived statistics of a distribution: Estimator

An estimator is any quantity calculated from the sample data which is used to give information about an unknown quantity in the population. For example, the sample mean is an estimator of the population mean.

Estimators of population parameters are sometimes distinguished form the true value by using the symbol “hat”. For example,
σ = true population standard deviation
$\hat{\sigma}$ = estimated (from a sample) population standard deviation

- **Point estimator**:
  to calculate a single value: **mean, variance**
- **Interval estimator**:
  to calculate several values: **confidence interval**

© 2012 University of Tuebingen

Derived statistics of a distribution: Point estimator

Point estimator:

- Mean

$$(\mu =)\ \hat{x} = \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Variance

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 = \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - n\mu^2\right]$$

- Variance of the single value

$$\mathrm{var}(x) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right]$$

- Variance of the mean

$$\mathrm{var}(\overline{x}) = \frac{\mathrm{var}(x)}{n}$$

© 2012 University of Tuebingen

Derived statistics of a distribution: Point estimator

Point estimator:

- standard deviation $\mathrm{s}(x) = \mathrm{sdv}(x) = \sqrt{\mathrm{var}(x)}$
- standard deviation of the mean $\mathrm{sdv}(\bar{x}) = \sqrt{\mathrm{var}(\bar{x})}$
- relative standard deviation $\mathrm{relsdv}(x) = \frac{\mathrm{sdv}(x)}{\bar{x}}$

© 2012 University of Tuebingen

Derived statistics of a distribution: Interval estimator

Interval estimator:

- Measurement uncertainty

$$u_i = statistical\,error + systematic\,error$$

$$u_i = \frac{t_{\alpha/2,\,f} \cdot \mathrm{sdv}(\mathrm{x})}{\sqrt{n}} + b$$

- Confidence interval (single value)

$$\mathrm{erw}(x) = \bar{x} \pm t_{\alpha/2,\,f} \cdot \mathrm{sdv}(x)$$

- Confidence interval (mean)

$$\mathrm{cnf}\,(x) = \bar{x} \pm t_{\alpha/2,\,f} \cdot \mathrm{sdv}(\bar{x})$$

$$= \bar{x} \pm t_{\alpha/2,\,f} \cdot \frac{\mathrm{sdv}(x)}{\sqrt{n}}$$

Remark:
t = student-factor t

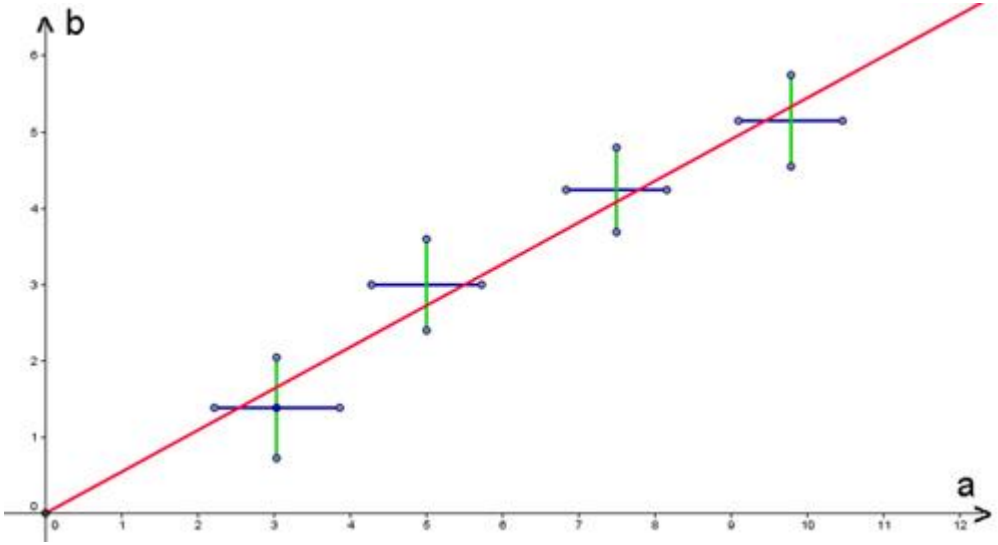
$$t_{\alpha/2,\,f} \cdot \mathrm{sdv}(\bar{x}) = \frac{t_{\alpha/2,\,f} \cdot \mathrm{sdv}(\mathrm{x})}{\sqrt{n}}$$

© 2012 University of Tuebingen

Propagation of uncertainty

Propagation of uncertainty (or propagation of error) is the effect of **variables uncertainties** (or errors) on the **uncertainty of a function** based on them.

For example: When the variables are the values of experimental measurements they have uncertainties due to measurement limitations (e. g. instrument precision) which propagate to the combination of variables in the function.



 © 2012 University of Tuebingen

Propagation of uncertainty

Uncertainty in one variable:

The influence of uncertainty in one variable can be estimated by Taylor propagation:

$y = y(x)$

$$y(x + \Delta x) = y(x) + \frac{1}{1!}\frac{\partial y(x)}{\partial x}\Delta x + \frac{1}{2!}\frac{\partial^2 y(x)}{\partial x^2}(\Delta x)^2 + \cdots + \frac{1}{n!}\frac{\partial^n y(x)}{\partial x^n}(\Delta x)^n$$

If Δx is relatively small, than the Taylor propagation can be stopped after the linear term.

Approximation:

$$y(x + \Delta x) - y(x) = \Delta y = \frac{\partial y(x)}{\partial x}\Delta x$$

Allowed approximation in the case of absolute errors.

Proportionality:

$y = c \cdot x$ => $\Delta y = c \cdot \Delta x$

© 2012 University of Tuebingen

Propagation of uncertainty

Uncertainty in more than one variable:

$$y = y(x_1, x_2, x_3, x_4, \cdots)$$

$$\Delta y = \frac{\partial y(x_1)}{\partial x_1}\Delta x_1 + \frac{\partial y(x_2)}{\partial x_2}\Delta x_2 + \frac{\partial y(x_3)}{\partial x_3}\Delta x_3 + \cdots$$

Simplification in the case of the four basic arithmetic operations:

- Addition $y = x_1 + x_2 \Rightarrow \Delta y = \Delta x_1 + \Delta x_2$
- Subtraction $y = x_1 - x_2 \Rightarrow \Delta y = \Delta x_1 - \Delta x_2$
- Multiplication $y = x_1 \cdot x_2 \Rightarrow \frac{\Delta y}{y} = \frac{\Delta x_1}{x_1} + \frac{\Delta x_2}{x_2}$
- Division $y = x_1 : x_2 \Rightarrow \frac{\Delta y}{y} = \frac{\Delta x_1}{x_1} - \frac{\Delta x_2}{x_2}$

© 2012 University of Tuebingen

Propagation of uncertainty

If the uncertainty is an approximation errors, then the errors (Δ-values) are unsigned and defined as the modulus.

$$\Delta y = \left|\frac{\partial y(x_1)}{\partial x_1}\right| \Delta x_1 + \left|\frac{\partial y(x_2)}{\partial x_2}\right| \Delta x_2 + \left|\frac{\partial y(x_3)}{\partial x_3}\right| \Delta x_3 + \cdots$$

© 2012 University of Tuebingen

Propagation of uncertainty

Gaussian error propagation:

$$\sigma^2(F) = \left(\frac{\partial F}{\partial p_i}\right)^2 \cdot \sigma^2(\mathrm{p}_i) + \left(\frac{\partial F}{\partial p_j}\right)^2 \cdot \sigma^2(\mathrm{p}_j) + 2 \cdot \underbrace{\frac{\partial F}{\partial p_i} \cdot \frac{\partial F}{\partial p_j}}_{\rho} \cdot \sigma(p_i) \cdot \sigma(p_j)$$

Covariance $\rho \cdot \sigma(p_i) \cdot \sigma(p_j) = \mathrm{cov}(p_i, p_j)$

$$\sigma^2 = \sigma^2(F) = \sum \left(\frac{\partial F}{\partial p_i}\right)^2 \cdot \sigma^2(p_i) + \sum_{i \neq j} \frac{\partial F}{\partial p_i} \cdot \frac{\partial F}{\partial p_j} \cdot \mathrm{cov}(p_i p_j)$$

or

$$\sigma = \sqrt{\left(\frac{\partial f}{\partial x_1}\right)^2 \sigma_1^2 + \left(\frac{\partial f}{\partial x_2}\right)^2 \sigma_2^2 + \cdots + \left(\frac{\partial f}{\partial x_n}\right)^2 \sigma_n^2}$$

© 2012 University of Tuebingen

Robust statistics

An estimation procedure or test procedure is called **robust** when it is not reacting sensitive to outliers (An outlier is an observation that is numerically distant from the rest of the data).

The mean value is influenced by fluctuations in the data set and by outliers. Especially in the case of numerically large outliers.

An example for the robust statistic is the **median statistics**.

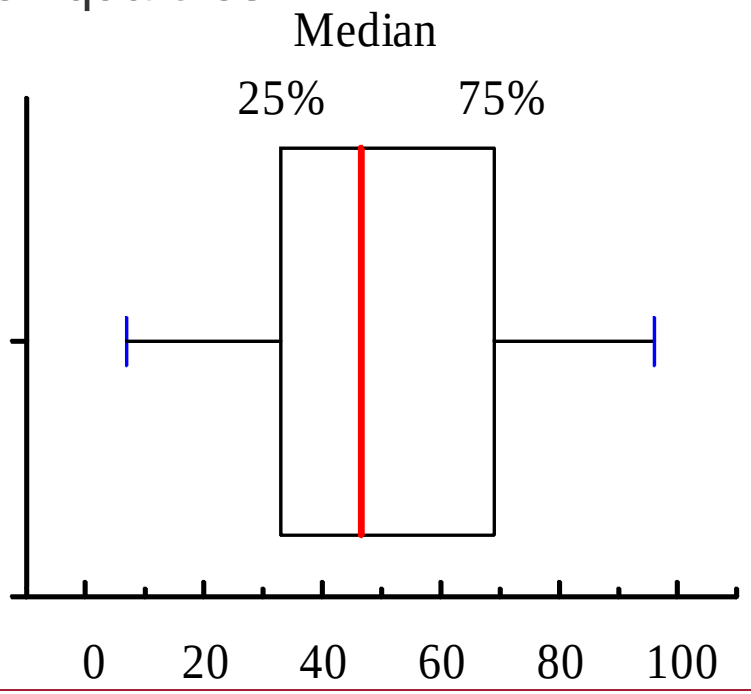
© 2012 University of Tuebingen

Robust statistics: Median statistics

The median is used primarily for skewed distributions, which it summarizes differently than the arithmetic mean.

One way to represent the data is the **Box plot**: The box plot (or box-and-whisker plot) is a convenient way of graphically depicting groups of numerical data through their quartiles.

It shows
- the median,
- the quartiles and
- the symmetry of the distribution.



 © 2012 University of Tuebingen

Box plot: Example

Draw a box-and-whisker plot for the following data set:
9, 6, 7, 7, 3, 9, 10, 1, 8, 7, 9, 9, 8, 10, 5, 10, 10, 9, 10, 8.

First step: Order the data set
1, 3, 5, 6, 7, 7, 7, 8, 8, 8, 9, 9, 9, 9, 9,10, 10, 10, 10, 10

Take the number or mean lying in the middle of the data set. For an uneven (odd) amount of values, the middle value is taken (**Median**). In the case of an even amount, the average of the neighboured values in the middle is taken (**Pseudomedian**).
Here: Average of the tenth and eleventh value: 8.5 = Pseudomedian (Q_2)

The Quartiles (Q_1, Q_3) are containing one fourth and three fourth of the data set: $Q_1 = 7$; $Q_3 = 9.5$

 © 2012 University of Tuebingen

Box plot: Example

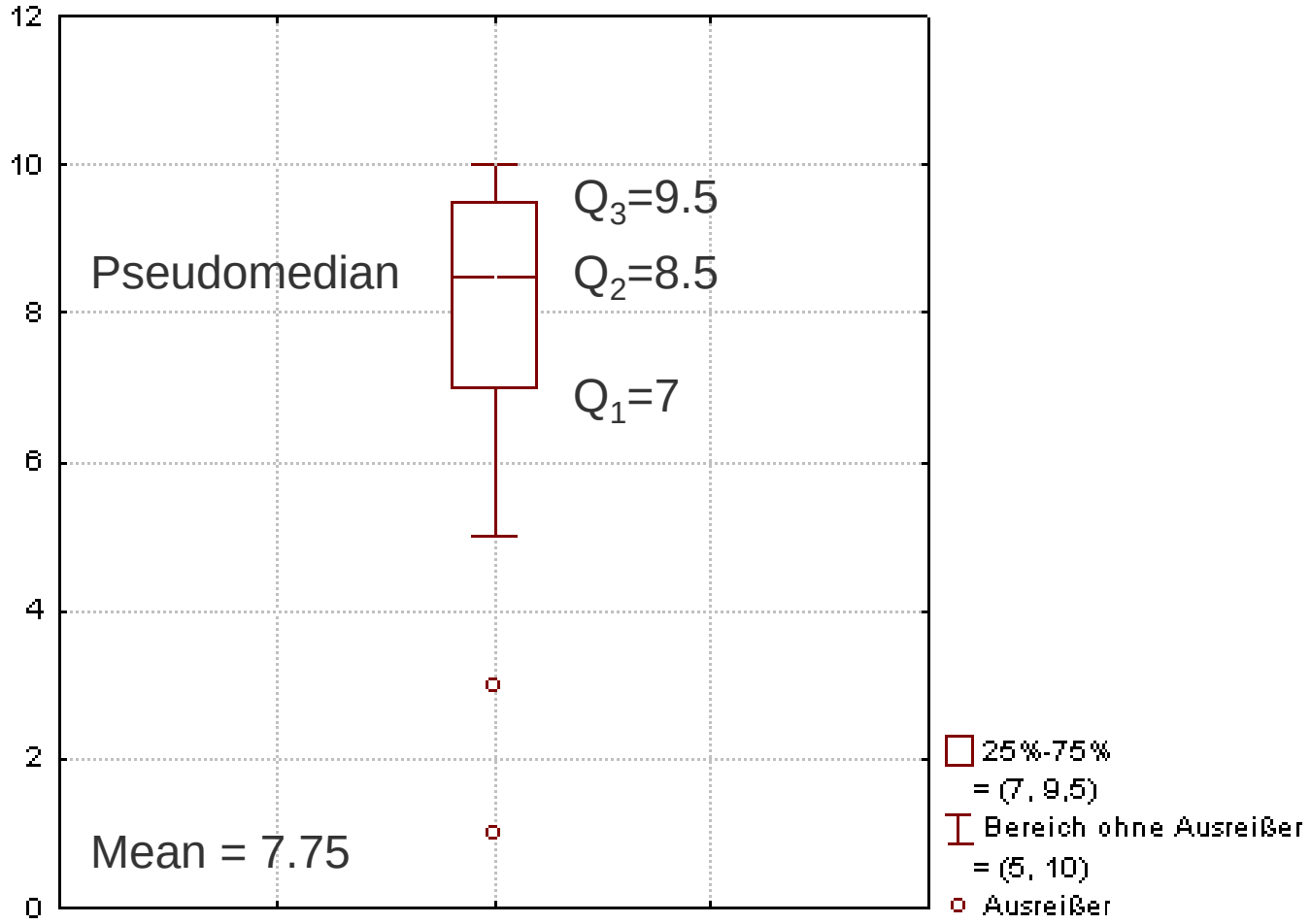
Draw a box from Q_1 to Q_3. The box contains 50 % of all measured values with an interquartile range of 2.5. Q_2 is giving you the median.

Drawing the **whiskers**: In order to identify outliers, the whiskers are drawn 1.5 times the interquartile range => 3.75
Drawing the whiskers from Q_3 => The highest value of this data set is 10.
Drawing the whiskers from Q1 => The lowest value would be 4.75. The whisker is only drawn to the value 5, because the smallest in the data set existing value is taken.

Any data not included between the whiskers should be plotted as an outlier with a dot. Here: The values 1 and 3.

© 2012 University of Tuebingen

Box plot: Example



Source: Wikipedia

 © 2012 University of Tuebingen





Significance tests

 © 2012 University of Tuebingen

Steps of testing

- Analytical methods should be free of systematic errors!
- In order to decide whether the difference between the measured and standard values is significant or random, a statistical test - known as a significance test - can be employed.
- In making a significance test we are testing the truth of a hypothesis (**Null hypothesis H_0**).

- Hypothesis can be formulated **one-sided** or **two-sided.**
- Both hypotheses allow for the significance level α but the decisions are made with distinct tabulated values:
  - One sided test with: $\alpha = 0.05$ use $q = 1 - \alpha$
  - Two sided test with: $\alpha = 0.05$ use $q = 1 - \alpha/2$

 © 2012 University of Tuebingen

Steps of testing

1. Write the null hypothesis (H_0).
   For example: With a significance level (probability of an error) of 5 %, the true value and the mean do not differ significantly.

2. Calculate the test by using the particular equation.

3. Compare the calculated value with the tabulated value.

4. Decision:
   a) the calculated value is less than the tabulated value
      → accept H_0
   b) the calculated value is larger than the tabulated value
      → reject H_0

 © 2012 University of Tuebingen

Student's t-distribution: t-test

- Student-factor t
- Corrects the finite number of data values to a big number of the normal distribution
- Dependent on the degree of freedom
- Tabulated
- With the two sided test, the interval becomes broader than $\mu \pm 1{,}96$. (for P = 0,95, infinite degrees of freedom)
- With 5 degrees of freedom:
  - $\pm$ 2,57 (two sided, 0,95)
  - $\pm$ 4,03 (two sided, 0,99)

$$t = \frac{error\ of\ the\ mean}{standard\ error\ of\ the\ mean}$$

$$t = \frac{|\bar{x} - \mu|}{sdv/\sqrt{n}}$$

F	P=0,50	0,75	0,90	0,95	0,98	0,99
1	1,00	2,41	6,31	12,7	31,82	63,7
2	0,816	1,60	2,92	4,30	6,97	9,92
3	0,765	1,42	2,35	3,18	4,54	5,84
4	0,741	1,34	2,13	2,78	3,75	4,60
5	0,727	1,30	2,01	2,57	3,37	4,03
10	0,700	1,22	1,81	2,23	2,76	3,17
15	0,691	1,20	1,75	2,13	2,60	2,95
20	0,687	1,18	1,73	2,09	2,53	2,85
25	0,684	1,18	1,71	2,06	2,49	2,79
30	0,683	1,17	1,70	2,04	2,46	2,75
40	0,681	1,17	1,68	2,02	2,42	2,70
60	0,679	1,16	1,67	2,00	2,39	2,66
120	0,677	1,16	1,66	1,98	2,36	2,62
∞	0,674	1,15	1,64	1,96	2,33	2,58

© 2012 University of Tuebingen

Student's t-distribution: t-test

- **Comparison of an experimental mean $\bar{x}$ with a known value μ**
- In order to decide whether the difference between $\bar{x}$ and μ is significant, the statistic t-value is calculated:

$$t = \frac{error\ of\ the\ mean}{standard\ error\ of\ the\ mean}$$

$$t = \frac{|\bar{x} - \mu|}{sdv/\sqrt{n}}$$

F	P=0,50	0,75	0,90	0,95	0,98	0,99
1	1,00	2,41	6,31	12,7	31,82	63,7
2	0,816	1,60	2,92	4,30	6,97	9,92
3	0,765	1,42	2,35	3,18	4,54	5,84
4	0,741	1,34	2,13	2,78	3,75	4,60
5	0,727	1,30	2,01	2,57	3,37	4,03
10	0,700	1,22	1,81	2,23	2,76	3,17
15	0,691	1,20	1,75	2,13	2,60	2,95
20	0,687	1,18	1,73	2,09	2,53	2,85
25	0,684	1,18	1,71	2,06	2,49	2,79
30	0,683	1,17	1,70	2,04	2,46	2,75
40	0,681	1,17	1,68	2,02	2,42	2,70
60	0,679	1,16	1,67	2,00	2,39	2,66
120	0,677	1,16	1,66	1,98	2,36	2,62
∞	0,674	1,15	1,64	1,96	2,33	2,58

© 2012 University of Tuebingen

F-test

- For the comparison of standard deviations.
- In order to test whether the difference between two sample variances is significant, the statistic F is calculated:

$$F = \frac{s_1^2}{s_2^2} \qquad \left(s_1^2 > s_2^2\right)$$

- The number of degrees of freedom of the numerator and denominator are n_1-1 and n_2-1 respectively.
- Requirement: populations from which the samples are taken are normal distributed.
- The test can take two forms, for example:
  - Test whether a method 1 is more precise than method 2 (one sided test). H_0: “Variance s_1 does not exceed variance s_2”.
  - Test whether the methods 1 and 2 differ in their precision (two sided test). H_0: “The variances s_1 and s_2 are equal”.

 © 2012 University of Tuebingen

F-distribution

Example:

Find the F-value, under which 95% of all values of a F-distributed random variable lies.

- df_1=2 numerator degree of freedom and
- df_2=4 denominator degree of freedom

Answer:

In the row for df_2=4 and column df_1=2 you can find the value F=6,94.

F-distribution for $\alpha = 0{,}05$:

df_2 (numerator)	1	2	3	4	5	10	20	30	100	1000000
1	161,45	199,50	215,71	224,58	230,16	241,88	248,02	250,10	253,04	254,32
2	18,51	19,00	19,16	19,25	19,30	19,40	19,45	19,46	19,49	19,50
3	10,13	9,55	9,28	9,12	9,01	8,79	8,66	8,62	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	5,96	5,80	5,75	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,74	4,56	4,50	4,41	4,37
10	4,96	4,10	3,71	3,48	3,33	2,98	2,77	2,70	2,59	2,54
15	4,54	3,68	3,29	3,06	2,90	2,54	2,33	2,25	2,12	2,07
20	4,35	3,49	3,10	2,87	2,71	2,35	2,12	2,04	1,91	1,84
30	4,17	3,32	2,92	2,69	2,53	2,16	1,93	1,84	1,74	1,62
40	4,08	3,23	2,84	2,61	2,45	2,08	1,84	1,74	1,64	1,51
50	4,03	3,18	2,79	2,56	2,40	2,03	1,78	1,69	1,58	1,44
100	3,94	3,09	2,70	2,46	2,31	1,93	1,68	1,57	1,45	1,28
200	3,89	3,04	2,65	2,42	2,26	1,88	1,62	1,52	1,39	1,19
500	3,86	3,01	2,62	2,39	2,23	1,85	1,59	1,48	1,35	1,11
1000	3,85	3,00	2,61	2,38	2,22	1,84	1,58	1,47	1,33	1,08
1000000	3,84	3,00	2,60	2,37	2,21	1,83	1,57	1,46	1,32	

© 2012 University of Tuebingen

Using the F-Test

- Null hypothesis not true: $F(calculated) > F(tab)$
  - Inhomogeneous
  - Probability of error different
  - No decision possible
  - Population cannot be pooled!

- Null hypothesis true : $F(calculated) < F(tab)$
  - This means that the variances only randomly differ from each other
  - Population can be pooled!

© 2012 University of Tuebingen

Consequences of the F-test

Example:

Method	sdv	Degree of freedom
1	s_1 = 4.3 %	f_1 = 11
2	s_2 = 2.1 %	f_2 = 11

$$F = \frac{4.3^2}{2.1^2} = \frac{18.49}{4.41} = 4.19$$

for α = 0.05 results 2.82 (tabulated value)

$$F_{calc} > F_{tab}$$

=> Null hypothesis has to be waived!

the calculated value is larger than the tabulated value:

- The variances are not homogeneous. **They cannot be pooled**.
- Systematic difference, no statistical difference!!!

© 2012 University of Tuebingen

Paired $\hat{t}$-test

- $\hat{t}$-Test: **Comparison of two means**

Requirement:
F-test in advance!

- It is tested, whether the differences in the means is caused by random error.
- The standard deviation of the difference of the means is calculated by propagation of uncertainty
- If the differences $|\bar{x}_a - \bar{x}_b|$ are random variables
  → follow t-distribution

$$\bar{x}_a \qquad \bar{x}_b$$
$$n_a \qquad n_b$$

$$s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \overline{\mathrm{sdv}}(x) \cdot \sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}}$$

- Normalized with $\overline{\mathrm{sdv}}(x)$

- To test whether n paired results are drawn from the same population, calculate the statistical value $\hat{t}$

$$\hat{t} = \frac{|\bar{x}_a - \bar{x}_b|}{\overline{\mathrm{sdv}}(x)} \sqrt{\frac{n_a n_b}{n_a + n_b}}$$

- If $\hat{t} > t_{tab}$, there is a significant difference between the two means.

© 2012 University of Tuebingen

Example 1

- Two groups measured one sample
  (3 degrees of freedom: f = 3)
- F-test:
  - Table: $\alpha = 0.05$ shows a value of 9.28
  - No difference: pooling is possible
- $\hat{t}$-test:

$(s_1)^2 = 34.25$ and $(s_2)^2 = 77.66$

$$\hat{F} = \frac{77.66}{34.25} = 2.27$$

$$\overline{\mathrm{var}}(x) = \frac{(n_a - 1)\,\mathrm{var}(\bar{x}_a) + (n_b - 1)\,\mathrm{var}(\bar{x}_b)}{n_a + n_b - 2} = \frac{3 \cdot 77.66 + 3 \cdot 34.25}{4 + 4 - 2} = 55.96$$

$$\hat{t} = \frac{|\bar{x}_a - \bar{x}_b|}{\overline{\mathrm{sdv}}(x)} \sqrt{\frac{n_a n_b}{n_a + n_b}} = \frac{21.32}{7.48} \sqrt{\frac{16}{8}} = 4.03$$

From the table: $\hat{t} = 2.45$ for $\alpha = 0.05$ and f = 6
$\hat{t} = 3.71$ for $\alpha = 0.01$

⇒ There is a significant difference of the means.
For at least one series of measurements, there is a systematic error.

© 2012 University of Tuebingen

Example 2

- var(1) = 0.01 and var(2) = 0.02.
- $n_1 = 5$, $n_2 = 4$.
- Calculation of F gives: F = 2.
- From the table F(0.05;3;4) = 6.59.
- The variances are homogeneous. They can be pooled.

- The mean variance is:

$$\overline{\mathrm{var}}(x) = \frac{(5-1)0.01 + (4-1)0.02}{5+4-2} = 0.014$$

- The calculated means are 1.2 and 1.3:

$$\hat{t} = \frac{|\bar{x}_a - \bar{x}_b|}{\overline{\mathrm{sdv}}(x)} \sqrt{\frac{n_a n_b}{n_a + n_b}} \qquad \hat{t} = \frac{0.1}{\sqrt{0.014}} \cdot \sqrt{\frac{5 \cdot 4}{5+4}} = 0.845 \cdot 1.49 = 1.26$$

- This is significantly smaller than the two sided tabulated value
  t (0.05;7)= 2.36

© 2012 University of Tuebingen

Example 3

Laboratory 1	0.470	0.448	0.463	0.449	0.482	0.454	0.477	0.409
Laboratory 2	0.529	0.490	0.489	0.521	0.486	0.502		

The concentration of titan in steel samples was measured with atom absorption spectroscopy (AAS). Can the differences of both laboratories be explained by random error?

$$\bar{x}_1 = 0.467 \qquad \bar{x}_2 = 0.503$$

$$\text{sdv}_1 = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = 0.0229 \qquad \text{sdv}_2 = 0.0189$$

$$\hat{F} = \frac{0.0229^2}{0.0182^2} = 1.583$$

The differences in variances are caused by random error: **pooling possible**

$$\overline{\text{var}}(x) = \frac{(n_1 - 1)\,\text{var}(\bar{x}_1) + (n_2 - 1)\,\text{var}(\bar{x}_2)}{n_1 + n_2 - 2} = \frac{7 \cdot 0.0229^2 + 5 \cdot 0.0182^2}{12} = 0.0211^2$$

$$\hat{t} = \frac{|\bar{x}_1 - \bar{x}_2|}{\overline{\text{sdv}}(x)} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{|0.467 - 0.503|}{0.0211} \sqrt{\frac{48}{14}} = 3.16$$

$$t(0.975;12) = 2.18$$

Calculated value is larger than the tabulated one;
Two sided test is significant. The differences of the means of the two laboratories cannot be explained by random error.

 © 2012 University of Tuebingen

X^2- test

The χ^2-test is very often used for the validation of crossbreed experiments. With this method it can be verified with a certain probability whether a result (which is deflecting from the ideal values) is corresponding with the theoretically expected experiments.

Critical Values of the χ^2 Distribution

df \ p	0.995	0.975	0.9	0.5	0.1	0.05	0.025	0.01	0.005	df
1	.000	.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879	1
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597	2
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838	3
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860	4
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750	5
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548	6
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278	7
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955	8
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589	9
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188	10

© 2012 University of Tuebingen

X^2- test: Example

Mendel made experiments on plant hybridization. He was crossbreeding different races of peas, namely round, yellow and green, edged seeds. In the F_2-generation he obtained hybrids with an ratio of 9:3:3:1 (total = 16) (see table).

	round/ yellow	edged/ yellow	round / green	edged / green	total
Observed (O)	315	101	108	32	556
Expected (E)	9/16 *556 = 312.75	104.25	104.25	34.75	556
O-E	2.25	-3.25	3.75	2.75	
Squared	5.06	10.56	14.06	7:56	
Square / E	5.06/312.75 = 0.016	0.101	0.134	0.217	
$\chi^2 = \sum \frac{\left((O-E) - \frac{1}{2}\right)^2}{E}$	0.016 +	0.101 +	0.134 +	0.217	= 0.468

© 2012 University of Tuebingen

X^2- test: Example

Take the χ^2- value (χ^2 = 0.468) and look in the table for a degree of freedom of three (f = 3). With a probability of 90 % he can assume that the obtained ratio of hybrids is correct.

Critical Values of the χ^2 Distribution

df \ p	0.995	0.975	0.9	0.5	0.1	0.05	0.025	0.01	0.005	df
1	.000	.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879	1
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597	2
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838	3
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860	4
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750	5
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548	6
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278	7
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955	8
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589	9
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188	10

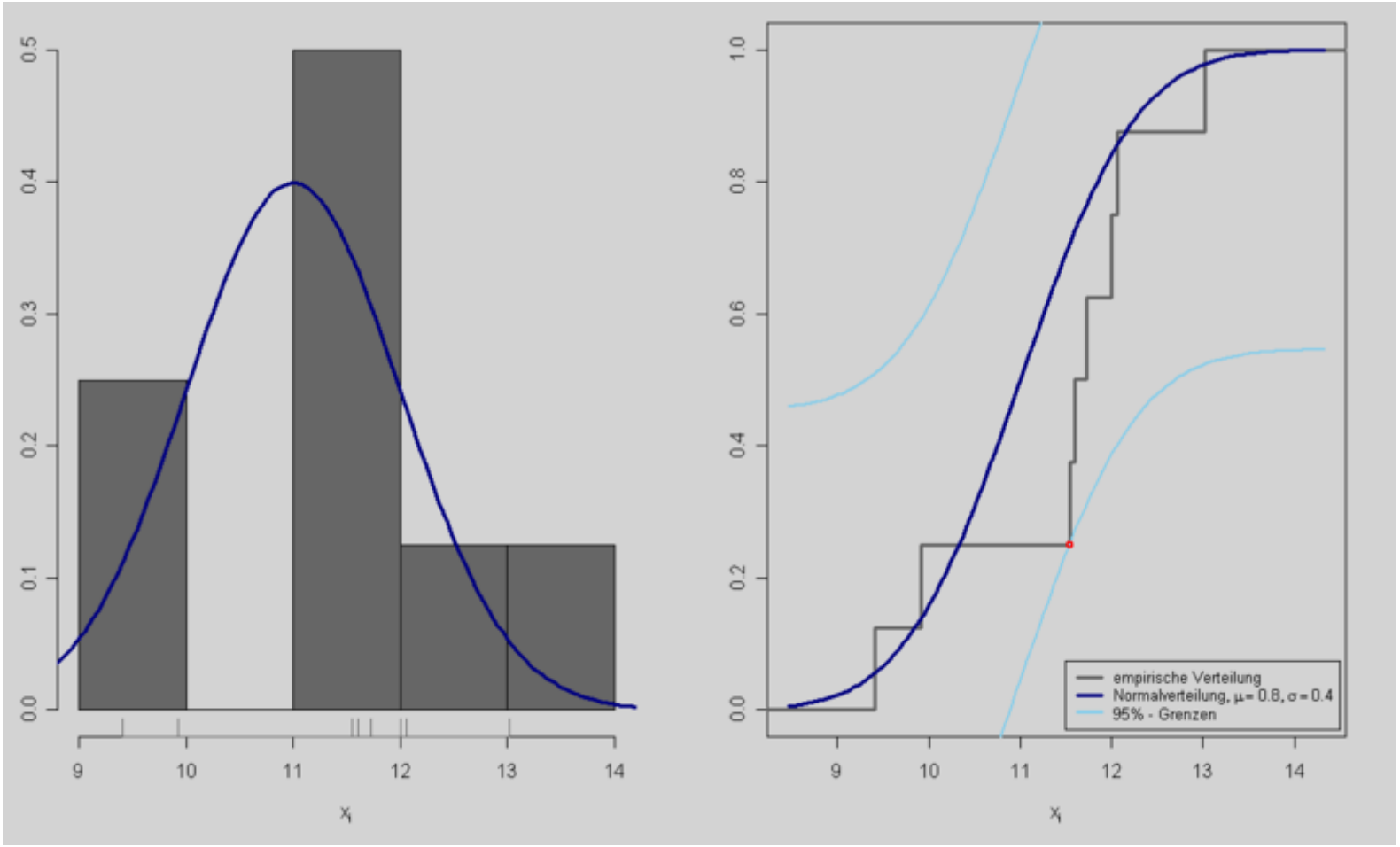
© 2012 University of Tuebingen

Kolmogorov–Smirnov test

Test whether the data are following a normal distribution:

Histogram

Comparison of the distribution functions



 © 2012 University of Tuebingen

Kolmogorov and Smirnov test: Example

	Comp. A(x) %	A(x) sorted
	10.3	9.8
	9.8	10.1
	10.1	10.3
	17.0	10.3
	11.5	10.8
	10.8	11.0
	10.3	11.5
	11.0	17.0
$\bar{x}$	11,35	
$\pm s$	2,35	

The concentration of a component A is determined with the following %-values (feature) (comp. A(x)%). In the first step the values are sorted in ascending order. (A(x) sorted).

In order to find the area under the standardized normal distribution, the feature is z-transformed in the second step.

z-transformation: $$z = \frac{(x - \bar{x})}{s}$$

x: feature

$\bar{x}$: mean

s: standard deviation

	Comp. A(x) %	A(x) sorted	z-transformed
	10.3	9.8	-0.661
	9.8	10.1	-0,533
	10.1	10.3	-0.447
	17.0	10.3	-0.447
	11.5	10.8	-0.234
	10.8	11.0	-0.149
	10.3	11.5	0.064
	11.0	17.0	2.408
$\bar{x}$	11.35		
$\pm s$	2.35		

© 2012 University of Tuebingen

Kolmogorov and Smirnov test: Example

... and for example calculating the area with the help of standard function in a table sheet:

	A	B
1	z-transformed	area of the standardized normal distribution
2	0,661	=NORM.S.DIST(A2)

	Comp. A(x) %	A(x) sorted	z-transformed	Area of the standardized normal distribution
	10.3	9.8	-0.661	0.254
	9.8	10.1	-0,533	0.297
	10.1	10.3	-0.447	0.327
	17.0	10.3	-0.447	0.327
	11.5	10.8	-0.234	0.407
	10.8	11.0	-0.149	0.441
	10.3	11.5	0.064	0.525
	11.0	17.0	2.408	0.992
$\bar{x}$	11.35			
$\pm s$	2.35			

© 2012 University of Tuebingen

Kolmogorov and Smirnov test: Example

In the third step the focus is on the area distances, which should be equal for normal distribution. Therefore, in this example their value should be 1/8 (1/n). This “ideal” distances are summarized in the column (i/n):

	Comp. A(x) %	A(x) sorted	z-transformed	Area of the standardized normal distribution	i/n
	10.3	9.8	-0.661	0.254	0.125
	9.8	10.1	-0,533	0.297	0.250
	10.1	10.3	-0.447	0.327	0.375
	17.0	10.3	-0.447	0.327	0.500
	11.5	10.8	-0.234	0.407	0.625
	10.8	11.0	-0.149	0.441	0.750
	10.3	11.5	0.064	0.525	0.875
	11.0	17.0	2.408	0.992	1.000
$\bar{x}$	11.35				
$\pm s$	2.35				
n	8				

© 2012 University of Tuebingen

Kolmogorov and Smirnov test: Example

In the fourth step the difference between the calculated area under the standardized normal distribution and the ideal distances is calculated:

	A	B	C	D	E	F
	Comp. A(x) %	**A(x) sorted**	**z-transformed**	**Area of the standardized normal distribution**	**i/n**	**Absolute difference between column D and E**
	10.3	9.8	-0.661	0.254	0.125	0.129
	9.8	10.1	-0,533	0.297	0.250	0.047
	10.1	10.3	-0.447	0.327	0.375	0.048
	17.0	10.3	-0.447	0.327	0.500	0.173
	11.5	10.8	-0.234	0.407	0.625	0.218
	10.8	11.0	-0.149	0.441	0.750	0.309
	10.3	11.5	0.064	0.525	0.875	**0.350**
	11.0	17.0	2.408	0.992	1.000	0,008
$\bar{x}$	11.35					
$\pm s$	2.35		**Max-value of column F**			**0.350**
n	8					

 © 2012 University of Tuebingen

Kolmogorov and Smirnov test: Example

The maximum absolute difference, which is **0.350**, has to be tested**.** If there is a significant difference from the normal distribution, this value will be larger than the tabulated value.

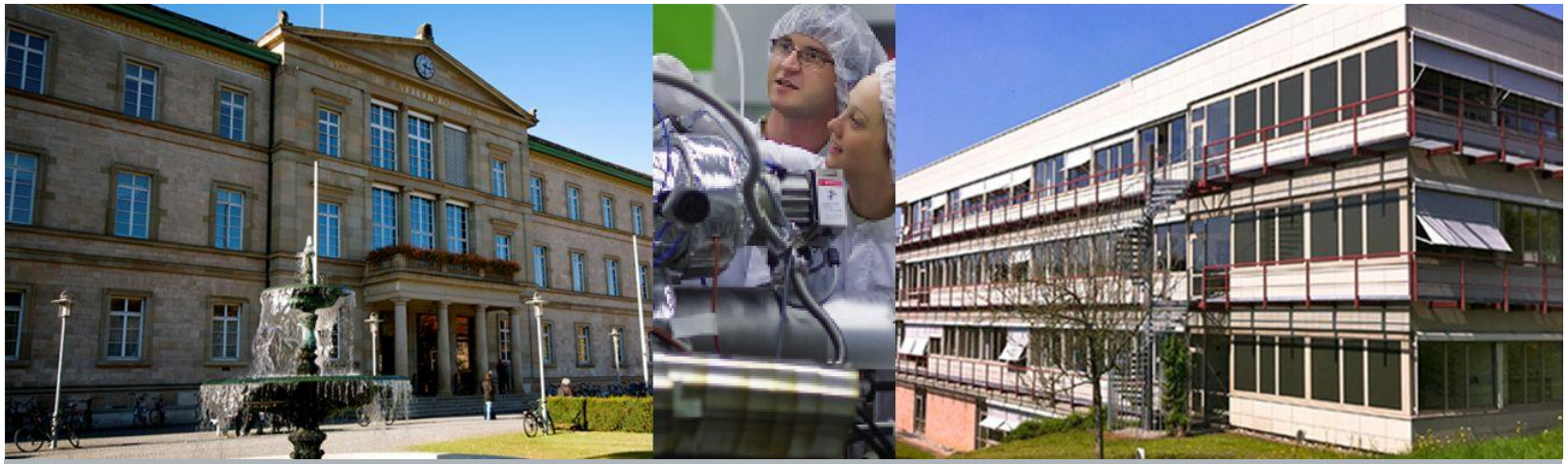
0.350 < tabulated 0.454

Therefore, there is no significant difference from the normal distribution!

Table for the Kolmogorov-Smirnov-test:

n	value	n	value
3	0,708	20	0294
4	0624	21	0,287
5	0,563	22	0,281
6	0,519	23	0,275
7	0,483	24	0,269
8	0,454	25	0,264
9	0,43	26	0,259
10	0,409	27	0,254
11	0,391	28	0,25
12	0,375	29	0,246
13	0,361	30	0,242
14	0,349	31	0,238
15	0,338	32	0,234
16	0,327	33	0,231
17	0,318	34	0,227
18	0,309	35	0,224
19	0,301		

© 2012 University of Tuebingen



Analysis of Variance

 © 2012 University of Tuebingen

Analysis of Variance (ANOVA): Introduction

- Analysis of Variance (ANOVA)
- Describes the possibility to compare different random samples.
- Random samples can be
  - from different methods
  - measured at different times
- Basis is a set of data consisting of columns and rows.
  - In the columns: The results of the different methods or times are presented.
  - In the rows: The different measurements of the random samples (the different replicas)
- Analysis of variance until now,
  - For example t-test: The means are compared by variance analysis,
  - Or F-test: Comparison of variances.
  - These were the easiest form of analysis of variance

© 2012 University of Tuebingen

Analysis of Variance (ANOVA): Hartley's test

- Ratio of the largest variance and the smallest variance
- Parameters
  - Number of k groups
  - Number of the degrees of freedom (n-1) for the different groups
- If the calculated value for a given confidence level (for expample 0.05) exceeds the tabulated one, the random samples are not homogeneous.

$$\hat{F}_{max} = \frac{largest\ variance}{smallest\ variance}$$

- Example:
  - n = 8
  - $s_1^2 = 6.21$; $s_2^2 = 1.12$; $s_3^2 = 4.34$
  - $\alpha = 0.05$
  - Degree of freedom: f = n - 1 = 7,
  - Number of groups k = 3,
  - Largest variance: 6.21,
  - Smallest Variance:1.12.

$$\hat{F}_{\max} = 5.54 < 6.94 = F_{\max}(tab)$$

 © 2012 University of Tuebingen

Analysis of Variance (ANOVA): Hartley's test

- Table for more variances (F_{max}) (Hartley's test) **($\alpha = 0{,}05$)**

ν k	2	3	4	5	6	7	8	9	10	11	12
2	39,0	87,5	142	202	266	333	403	475	550	626	704
3	15,4	27,8	39,2	50,7	62,0	72,9	83,5	93,9	104	114	124
4	6,60	15,5	20,6	25,2	29,5	33,6	37,5	41,1	44,6	48,0	51,4
5	7,15	10,8	13,7	16,3	18,7	20,8	22,9	24,7	26,5	28,2	29,9
6	5,82	8,38	10,4	12,1	13,7	15,0	16,3	17,5	18,6	19,7	20,7
7	4,99	**6,94**	8,44	9,70	10,8	11,8	12,7	13,5	14,3	15,1	15,8
8	4,43	6,00	7,18	8,12	9,03	9,78	10,5	11,1	11,7	12,2	12,7
9	4,03	5,34	6,31	7,11	7,80	8,41	8,95	9,45	9,91	10,3	10,7
10	3,72	4,85	5,67	6,34	6,92	7,42	7,87	8,28	8,66	9,01	9,34
12	3,28	4,16	4,79	5,30	5,72	6,09	6,42	6,72	7,00	7,25	7,48
15	2,86	3,54	4,01	4,37	4,68	4,95	5,19	5,40	5,59	5,77	5,93
20	2,46	2,95	3,29	3,54	3,76	3,94	4,10	4,24	4,37	4,49	4,59
30	2,07	2,40	2,61	2,78	2,91	3,02	3,12	3,21	3,29	3,36	3,39
60	1,67	1,85	1,96	2,04	2,11	2,17	2,22	2,26	2,30	2,33	2,36
∞	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

© 2012 University of Tuebingen

Analysis of Variance (ANOVA): Cochran's test

Equality of several variances

$$\hat{G}_{\max} = \frac{s^2_{\max}}{s_1^2 + s_2^2 + \cdots + s_k^2}$$

Variances of groups (s_{max}^2) are significantly larger than the remaining

Example:

- There are five variances:
- $s_1^2 = 26$, $s_2^2 = 51$, $s_3^2 = 40$, $s_4^2 = 24$ und $s_5^2 = 28$.
- 9 degrees of freedom for each variance. Tested for 95% confidence.

$$\hat{G}_{\max} = \frac{51}{26 + 51 + 40 + 24 + 28} = 0{,}302$$

$\alpha = 0{,}05$

k \ ν	1	2	3	4	5	6	7	8	9	10	16	36	144	∞
2	0,9985	0,9750	0,9392	0,9057	0,8772	0,8534	0,8332	0,8159	0,8010	0,7880	0,7341	0,6602	0,5813	0,5000
3	0,9669	0,8709	0,7977	0,7457	0,7071	0,6771	0,6530	0,6333	0,6167	0,6025	0,5466	0,4748	0,4031	0,3333
4	0,9065	0,7679	0,6841	0,6287	0,5895	0,5598	0,5365	0,5175	0,5017	0,4884	0,4366	0,3720	0,3093	0,2500
5	0,8412	0,6838	0,5981	0,5441	0,5065	0,4783	0,4564	0,4387	0,4241	0,4118	0,3645	0,3066	0,2513	0,2000
6	0,7808	0,6161	0,5321	0,4803	0,4447	0,4184	0,3980	0,3817	0,3682	0,3568	0,3135	0,2612	0,2119	0,1667
7	0,7271	0,5612	0,4800	0,4307	0,3974	0,3726	0,3535	0,3384	0,3259	0,3154	0,2756	0,2278	0,1833	0,1429
8	0,6798	0,5157	0,4377	0,3910	0,3595	0,3362	0,3185	0,3043	0,2926	0,2829	0,2462	0,2022	0,1616	0,1250
9	0,6385	0,4775	0,4027	0,3584	0,3286	0,3067	0,2901	0,2768	0,2659	0,2568	0,2226	0,1820	0,1446	0,1111
10	0,6020	0,4450	0,3733	0,3311	0,3029	0,2823	0,2666	0,2541	0,2439	0,2353	0,2032	0,1655	0,1308	0,1000
12	0,5410	0,3924	0,3264	0,2880	0,2624	0,2439	0,2299	0,2187	0,2098	0,2020	0,1737	0,1403	0,1100	0,0833
15	0,4709	0,3346	0,2758	0,2419	0,2195	0,2034	0,1911	0,1815	0,1736	0,1671	0,1429	0,1144	0,0889	0,0667
20	0,3894	0,2705	0,2205	0,1921	0,1735	0,1602	0,1501	0,1422	0,1357	0,1303	0,1108	0,0879	0,0675	0,0500
24	0,3434	0,2354	0,1907	0,1656	0,1493	0,1374	0,1286	0,1216	0,1160	0,1113	0,0942	0,0743	0,0567	0,0417
30	0,2929	0,1980	0,1593	0,1377	0,1237	0,1137	0,1061	0,1002	0,0958	0,0921	0,0771	0,0604	0,0457	0,0333
40	0,2370	0,1576	0,1259	0,1082	0,0968	0,0887	0,0827	0,0780	0,0745	0,0713	0,0595	0,0462	0,0347	0,0250
60	0,1737	0,1131	0,0895	0,0765	0,0682	0,0623	0,0583	0,0552	0,0520	0,0497	0,0411	0,0316	0,0234	0,0167
120	0,0998	0,0632	0,0495	0,0419	0,0371	0,0337	0,0312	0,0292	0,0279	0,0266	0,0218	0,0165	0,0120	0,0083
∞	0	0	0	0	0	0	0	0	0	0	0	0	0	0

 © 2012 University of Tuebingen

Analysis of Variance (ANOVA): One-Way ANOVA

- Several determining factors are included. Observations are investigated more systematically.
- Comparison of more than two means: One-Way ANOVA.
- **Requirement**: normal distributed data.
- Observed total variance is fragmented in parts, which are related to the same roots / attributes:
  - Random influence (for example: fluctuations within the random samples, methods)
  - Influence, which has to be investigated (Differences between laboratories, methods)
- Considers not only a pair of variances or means, which are compared individually.
- Instead, it is considered that the events are not individual and isolated.
- Therefore, the total probabilities have to be considered.
- The variance analysis allows us to identify
  - the influence of the single variation for each measure or method respectively
  - the influence of the measurement series on the total variance.
- For this reason, we can much better identify, which experimental conditions lead to optimal results.

 © 2012 University of Tuebingen

Analysis of Variance (ANOVA): One-Way ANOVA

Instead of many t-tests for comparison of the means of several experimental series, methods (M) or random samples (=groups), just one F-test between two variances is needed.

These variances are:

- The summation of the differences between the means of the several groups (= column; mean of the group) and the total mean: SS_F
- The sum of squares of the single value from the mean of the group: SS_{res}
- The squared sum of the deviation of the column mean (average over the row, we will write this as a column) from the total mean divided by the sum over all deviations from the column means.

© 2012 University of Tuebingen

Analysis of Variance (ANOVA): One-Way ANOVA

If these groups originate from the same population, this ratio will have nearly the same value.
If the ratio is larger than the tabulated value of the F-distribution, which is defined by α and the degrees of freedom, there are groups with different means. Then, the null hypothesis is rejected, if:

$$\rightarrow \hat{F} > F_{(M-1;n-M;\alpha)} \quad \text{(inhomogenous)}$$

$$n = \sum_{j=1} n_j = \sum_{j=1}^{M} R_j = R \cdot M$$

© 2012 University of Tuebingen

Analysis of Variance (ANOVA): One-Way ANOVA

- $R \cdot M$ is applied, if the random sample size is equal for all methods. Otherwise, in the following equations one has to distinguish between R and R_j.
- n is the total number of all values of the overall M columns, which contain n_i *rows. Two* parts:
  1. The standard deviation of the single values around the means of the groups.
  2. The standard deviation of the means of the groups around the total mean.

© 2012 University of Tuebingen

Analysis of Variance (ANOVA): One-Way ANOVA

- The mean squares over all groups are called random sample error and the mean squares within a group are called experimental error.
- Essential is, that the sum of the squared deviations of the random samples can be fragmented into two parts:
  - Sum of squares of the single values around the group means (SS within the group)
  - Sum of squares of the group means around the total mean (SS between the groups):

$$SS_{\text{total}} = SS_{\text{within the group}} + SS_{\text{between the group}} = SS_T = SS_{res} + SS_F$$

 © 2012 University of Tuebingen

Analysis of Variance (ANOVA): One-Way ANOVA

- ANOVA-scheme with R (rows), M (number of methods, columns).

$$ST$$

$$SS_T = \sum_i^{Rj}\sum_j^{M}(x_{ij} - \hat{E}(x))^2 = \sum_i^{Rj}\sum_j^{M} x_{ij}^2 - R \cdot M \cdot \hat{E}(x)^2$$

$$SS_M = R \cdot M \cdot \hat{E}(x)^2$$

$$SS_{res} = \sum_i^{Rj}\sum_j^{M}(x_{ij} - \hat{E}_j(x))^2$$

$$SS_F = \sum_{j=1}^{M} R_j \cdot \left(\hat{E}_j(x) - \hat{E}(x)\right)^2 = R \cdot \sum_{j=1} \hat{I}_j(x)^2 - M \cdot R \cdot \hat{I}(x)^2$$

$$= ST - R \cdot \sum_j \hat{E}_j(x)^2$$

$$\hat{F} = \frac{\frac{1}{M-1}\sum_j R_j\left(\hat{E}_j(x) - \hat{E}(x)\right)^2}{\frac{1}{n-M}\sum_{i,j}\left(x_{ij} - \hat{E}_j(x)\right)^2} = \frac{\frac{1}{M-1}\left[\sum_j \frac{\left(\hat{E}_j(x)\right)^2}{R_j} - \frac{\left(\hat{E}(x)\right)^2}{n}\right]}{\frac{1}{n-M}\left[\sum_{i,j}\left(x_{ij}\right)^2 - \sum_j \frac{\left(\hat{E}_j(x)\right)^2}{R_j}\right]} = \frac{\frac{1}{M-1}\left[\sum_j \frac{\left(\sum_i x_{ij}\right)^2}{R_j} - \frac{\left(\sum\sum x_{ij}\right)^2}{n}\right]}{\frac{1}{n-M}\left[\sum_{i,j}\left(x_{ij}\right)^2 - \sum_j \frac{\left(\sum_i x_{ij}\right)^2}{R_j}\right]}$$

 © 2012 University of Tuebingen

ANOVA: Overview formulas

Formula	Name / Explanation
SS	Sum of squares (Sum of squares of the deviations)
M	Number of methods
R	Replica (Number of repeated measurements)
R_j	Replica (samples) (Number of repeated measurements for different sized samples)
$E_j(x) = \frac{1}{R_j}\sum_{i=1}^{R_j} x_{ij}$	Mean of the columns (Mean of the measured values of one method)
$E(x) = \frac{\sum_{i=1}^{R_j}\sum_{j=1}^{M} x_{ij}}{\sum_{j=1}^{M} R_j}$	Total mean (Mean over all values in the table) $f_{E(x)} = 1$
$SS_T = \sum_{i=1}^{R_j}\sum_{j=1}^{M}\left(x_{ij} - E(x)\right)^2$	Sum of squares of all measured values in relation to the total mean
$SS_T = \sum_{i=1}^{R_j}\sum_{j=1}^{M}\left(x_{ij} - \hat{E}(x)\right)^2 = \sum_{i=1}^{R_j}\sum_{j=1}^{M}\left(x_{ij} - \hat{E}_j(x)\right)^2 + \sum_{j=1}^{M} R_j\left(\hat{E}_j(x) - \hat{E}(x)\right)^2$	$f_{SST} = R \cdot M - 1$
$ST = \sum_{i=1}^{R_j}\sum_{j=1}^{M}\left(x_{ij}\right)^2$	Sum of squares of all measured values $f_{ST} = R \cdot M$
$ST = SS_T - SS_M$	Sum of squares related to the variance 0 around the total mean
$SS_M = R \cdot M \cdot E(x)^2$	Sum of squares of the total mean according to the number of variables
$SS_F = \sum_{j=1}^{M} R_j\left(\hat{E}_j(x) - \hat{E}(x)\right)^2$	Sum of squares: Deviation of the column mean to the total mean (F: factors) $f_F = M - 1$
$SS_{res} = \sum_{i=1}^{R_j}\sum_{j=1}^{M}\left(x_{ij} - \hat{E}_j(x)\right)^2$ $SS_{res} = ST - R\sum_{j=1}^{M} E_j(x)^2$	Sum of squares of the single value from the mean of the group (of residuals) $f_{res} = R \cdot M - M$

 © 2012 University of Tuebingen

Analysis of Variance (ANOVA):

$i \setminus j$	1	2	3	$\sum_{i,j}\left(x_{ij}\right)^2 = 252$
1	3	4	8	9+16+64
2	7	2	4	49+4+16
3		7	6	49+36
4		3		9
$\sum_j x_{ij}$	10	16	18	$\sum_i \sum_j x_{ij} = 44$
Rj	2	4	3	$\sum R_j = n = 9$
$\left(\sum_j x_{ij}\right)^2 \quad \left(\sum_j x_{ij}\right)^2 / R_j$	100 50	256 64	324 108	$\sum_j \frac{\left(\sum_i x_{ij}\right)^2}{R_j} = 222$
$\hat{E}_j(x) = \frac{1}{R_j}\sum_i x_{ij}$	5	4	6	$\hat{E}(x) = \sum_{i=1}^{R_j}\sum_{j=1}^{M} x_{ij} \Big/ \sum_{j=1}^{M} R_j = 44/9 = 4{,}889$
				$\left(\hat{E}(x)\right)^2 / n = 44^2/9 = 215{,}111$

$$\hat{F} = \frac{\frac{1}{3-1}\left[\left(\frac{10^2}{2}+\frac{16^2}{4}+\frac{18^2}{3}\right)-\frac{44^2}{9}\right]}{\frac{1}{9-3}\left[\left(3^2+7^2+4^2+2^2+7^2+3^2+8^2+4^2+6^2\right)-\left(\frac{10^2}{2}+\frac{16^2}{4}+\frac{18^2}{3}\right)\right]}$$

$$= \frac{\frac{1}{2}\cdot[222-215{,}11]}{\frac{1}{6}\cdot[252-222]} = \frac{\frac{1}{2}\cdot 6{,}89}{\frac{1}{6}\cdot 30} = 0{,}689$$

F-table for
F(2;6;0.05) = 5.14,
Therefore, all three experiments are sums of random samples of the same population.

 © 2012 University of Tuebingen

Analysis of variance (ANOVA): identical sample size

$i \setminus j$	1	2	3	$\sum_{i,j}\left(x_{ij}\right)^2 = 450$
1	6	5	7	36+25+49
2	7	6	8	49+36+64
3	6	4	5	3616+25
4	5	5	8	25+25+64
$\sum_j x_{ij}$	24	20	28	$\sum_i \sum_j x_{ij} = 72$
Rj	4	4	4	12
$\left(\sum_j x_{ij}\right)^2 \quad \left(\sum_j x_{ij}\right)^2 / R_j$	576 144	400 100	784 196	$\sum_j \frac{\left(\sum_i x_{ij}\right)^2}{R_j} = 432$
$\hat{E}_j(x) = \frac{1}{R_j}\sum_i x_{ij}$	5	4	6	$\hat{E}(x) = \sum_{i=1}^{R_j}\sum_{j=1}^{M} x_{ij} \Big/ \sum_{j=1}^{M} R_j = 72/12 = 6$
				$\left(\hat{E}(x)\right)^2 / n = 72^2/12 = 432$

$$\hat{F} = \frac{\frac{1}{3-1}\left[\frac{1}{4}\left(24^2+20^2+28^2\right)-\frac{72^2}{12}\right]}{\frac{1}{12-3}\left[\left(6^2+\cdots+8^2\right)-\frac{1}{4}\left(24^2+20^2+28^2\right)\right]} = \frac{\frac{1}{2}\cdot[440-432]}{\frac{1}{9}\cdot[450-440]} = \frac{4}{10/9} = 3{,}60$$

F(2;9;0.05) = 4.26: Null hypothesis not to be waived

 © 2012 University of Tuebingen

$i \setminus j$	1	2	3	4	5	6	7	$\sum_{i,j}\left(x_{ij}\right)^2 =$
1	1,6	4,6	1,2	1,5	6,0	6,2	3,3	
2	2,9	2,8	1,9	2,7	3,9	3,8	3,8	
3	3,5	3,0	2,9	3,4	4,3	5,5	5,5	
4	1,8	4,5	1,1	2,0	5,8	4,2	4,9	
5	2,2	3,1	2,9	3,4	4,0	5,3	4,5	
$\sum_j x_{ij}$								$\sum_i \sum_j x_{ij} =$
Rj								12
$\left(\sum_j x_{ij}\right)^2$ $\left(\sum_j x_{ij}\right)^2 / R_j$								$\sum_j \frac{\left(\sum_i x_{ij}\right)^2}{R_j} =$
$\hat{E}_j(x) = \frac{1}{R_j}\sum_i x_{ij}$								$\hat{E}(x) = \sum_{i=1}^{R_j}\sum_{j=1}^{M} x_{ij} \Big/ \sum_{j=1}^{M} R_j =$
								$\left(\hat{E}(x)\right)^2 / n =$

$$\hat{F} = \frac{\frac{1}{M-1}\sum_j R_j\left(\hat{E}_j(x)-\hat{E}(x)\right)^2}{\frac{1}{n-M}\sum_{i,j}\left(x_{ij}-\hat{E}_j(x)\right)^2} = \frac{\frac{1}{M-1}\left[\sum_j \frac{\left(\hat{E}_j(x)\right)^2}{R_j} - \frac{\left(\hat{E}(x)\right)^2}{n}\right]}{\frac{1}{n-M}\left[\sum_{i,j}\left(x_{ij}\right)^2 - \sum_j \frac{\left(\hat{E}_j(x)\right)^2}{R_j}\right]} = \frac{\frac{1}{M-1}\left[\sum_j \frac{\left(\sum_i x_{ij}\right)^2}{R_j} - \frac{\left(\sum\sum x_{ij}\right)^2}{n}\right]}{\frac{1}{n-M}\left[\sum_{i,j}\left(x_{ij}\right)^2 - \sum_j \frac{\left(\sum_i x_{ij}\right)^2}{R_j}\right]} \qquad \hat{F} = \frac{\frac{1}{7-1}\left[\frac{1}{-}(\ ) - \right]}{\frac{1}{35-7}\left[(\ ) - \frac{1}{-}(\ )\right]} = \frac{\frac{1}{6}\cdot[\ ]}{\frac{1}{28}\cdot[\ ]} = - - =$$

 © 2012 University of Tuebingen

Examination relevant topics

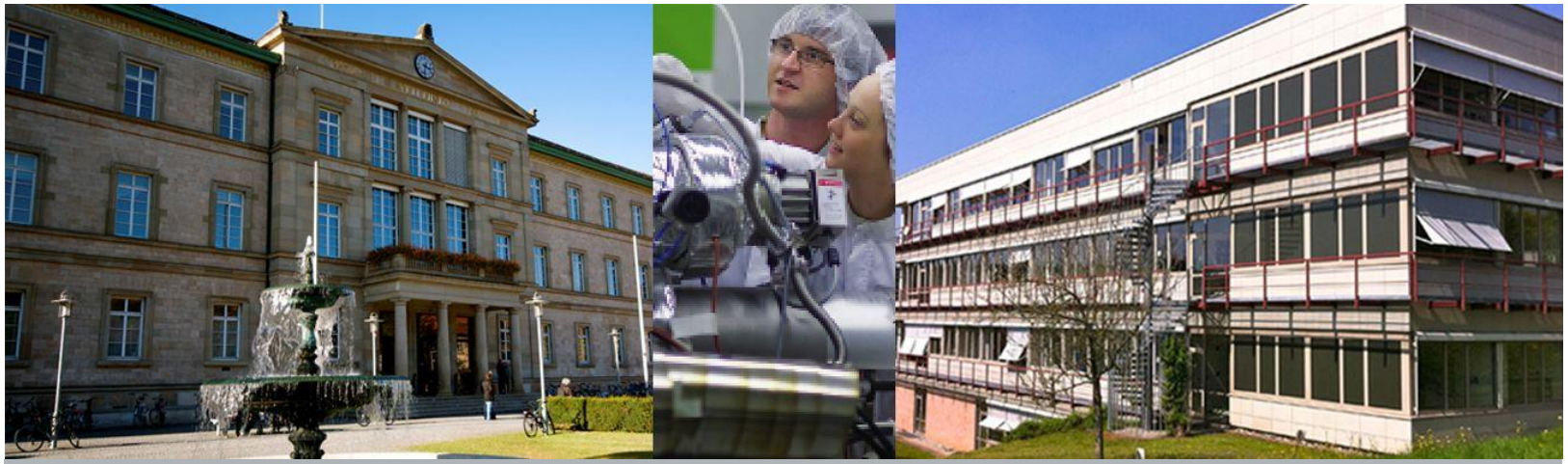
- Terms and definitions in statistics
- Normal / Gaussian distribution
  - Mathematical basics
  - Standard deviation and the meaning
  - Graphical representation
- Types of errors
  - Gross error
  - Systematic error (bias)
  - Statistical error (random)
- Density and distribution functions
- Other types of distributions: Binominal, Bernoulli, Poisson
- Point estimators
- Propagation of uncertainty
- Robust statistics: Box-Whisker Plots
  - Application and graphical representation

© 2012 University of Tuebingen

Examination relevant topics

- Statistical tests: Very important: Steps of testing!
- Student t-Test
  - Calculate for a simple case
- F-Test
  - Calculate for a simple case
- Simple ANOVA

 © 2012 University of Tuebingen



END of Part 01

 © 2012 University of Tuebingen