Statistics for User Studies

A Practical Approach

MMI 1 – WS 07/08
LFE Medieninformatik, LMU München

Accuracy vs. Precision

Accuracy	Precision
is determined by	is determined by
measurement errors	**measurement noise**
needed:	needed:
good study design	**enough data**
verified by:	verified by:
thorough description of study setup	**rigorous statistical analysis**

Types of Data

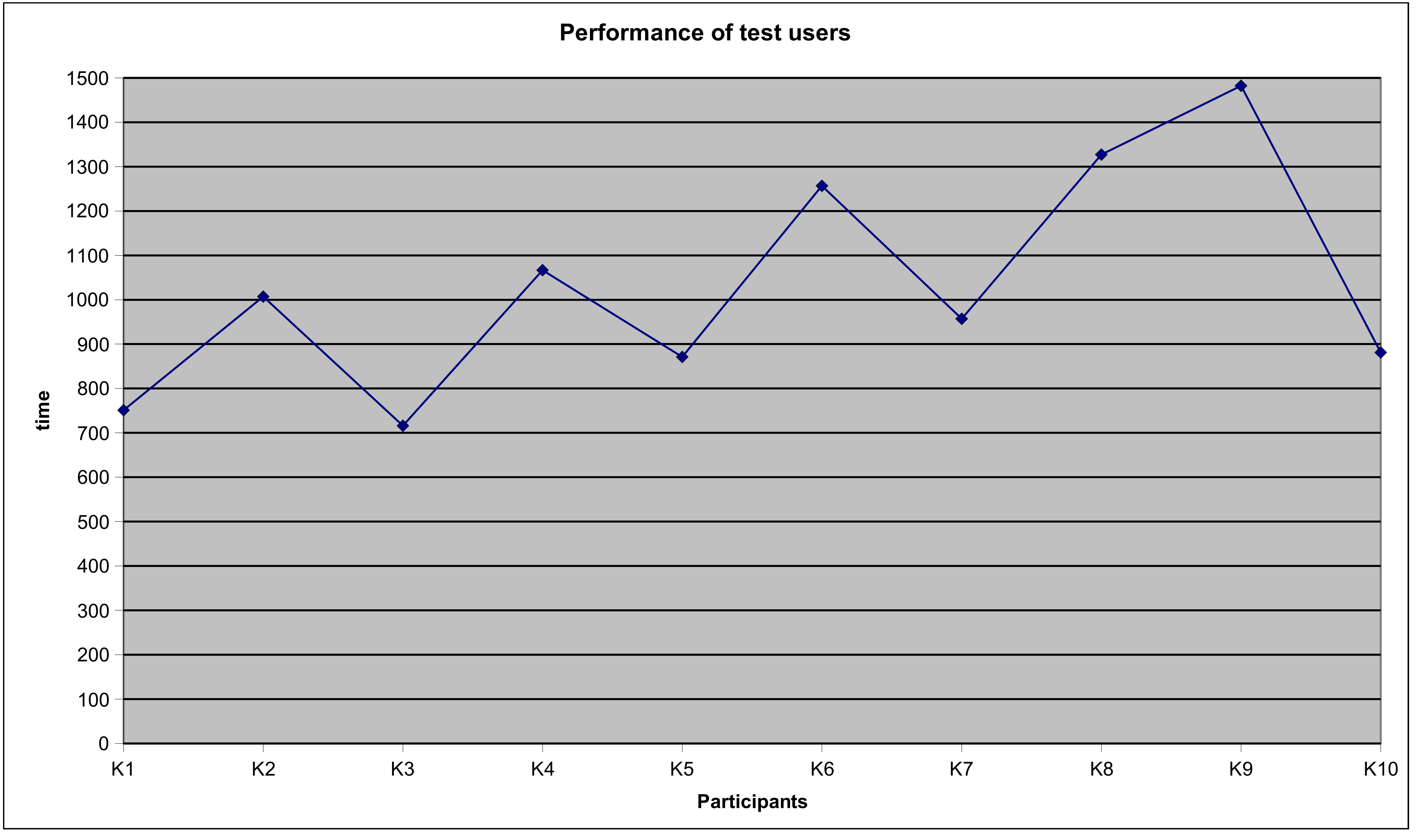
- **Categorical / Nominal Data**
  (alternatives in non-overlapping subsets, A=B, A!=B)
  - Gender: male/female, Handedness: left/right
- **Ordinal Data**
  (ranking/ordering A>B, A<B, A=B)
  - Marks in school: 1, 2, 3, 4, 5, 6
  - Type of education: school, high school, university
- **Interval Scale Data**
  (zero point is arbitrary, A-B)
  - tide
  - temperature (°C/°F),
- **Ratio Scale Data**
  (fixed zero point A / B)
  - weight
  - time

Try to get this!

Types of Variables

- Discrete Data
  - distinct and separate
  - can be counted
- Continuous Data
  - any value within a finite or infinite interval
  - always have a order

Don't Do This



Performance of test users
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
time
K1
K2
K3
K4
K5
K6
K7
K8
K9
K10
Participants

Frequency Tables

Data can be summarized in form of a frequency table

- well suited for discrete data
- continuous data have to be divided in groups

Example: days needed to answer my email

Data: 5 2 2 3 4 4 3 2 0 3 0 3 2 1 5 1 3 1 5 5 2 4 0 0 4 5 4 4 5 5

Days	*Frequency*	*Frequency (%)*
0	4	13%
1	3	10%
2	5	17%
3	5	17%
4	6	20%
5	7	23%

Likert Scales

Examples:

PowerPoint presentations are the best way to teach. State your opinion.

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

This year I will buy a new computer.

○ No ○ Uncertain ○ Yes

- ordinal data

  ⇒ actually not valid for statistical analysis
- use median, not mean
- you can force the user to make a commitment to one direction by offering an even number of choices.
- use 3 to 7 options

Likert, R. (1932). "A Technique for the Measurement of Attitudes" Archives of Psychology 140, 55

Mean, Median, Mode (I)

Mean

If $x_1, x_2, \ldots x_n$ are the data in a sample, the mean is $\frac{1}{n}\sum_{i=1}^{n} X_i$

Median

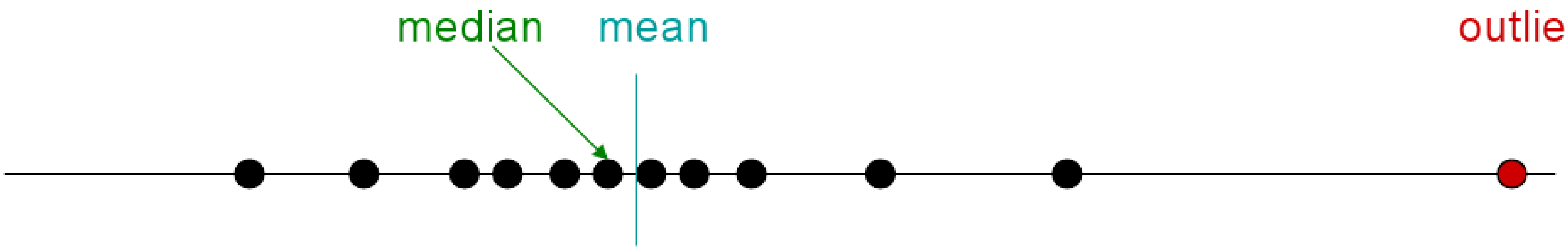
If $x_1, x_2, \ldots x_n$ are the **ordered** data in a sample, the median is $x_{(n+1)/2}$ if n is odd, and $(x_{n/2} + x_{n/2+1}) / 2$ if n is even. It is the value halfway through the ordered data set.

Mode

The mode is the value that occurs most often in a sample. There may be more than one mode in a sample.

Mean, Median, Mode (II)

Median is less sensitive on outliers

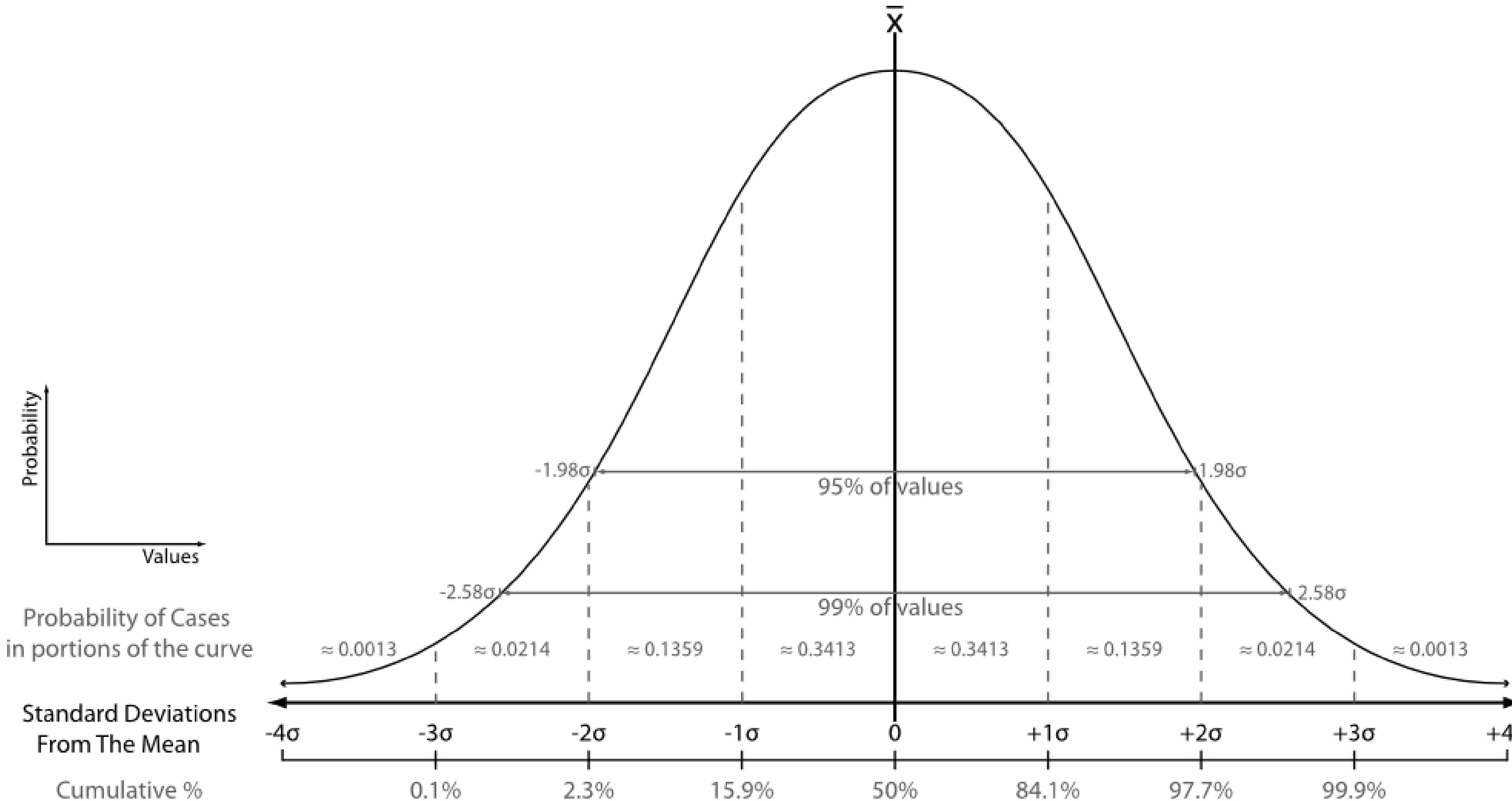


Mode works on all types of data

Median works on ordinal, interval, ratio data

Mean works on interval or ratio data

Normal Distribution



Source: http://en.wikipedia.org/wiki/Image:The_Normal_Distribution.svg

Variance and Standard Deviation

Variance

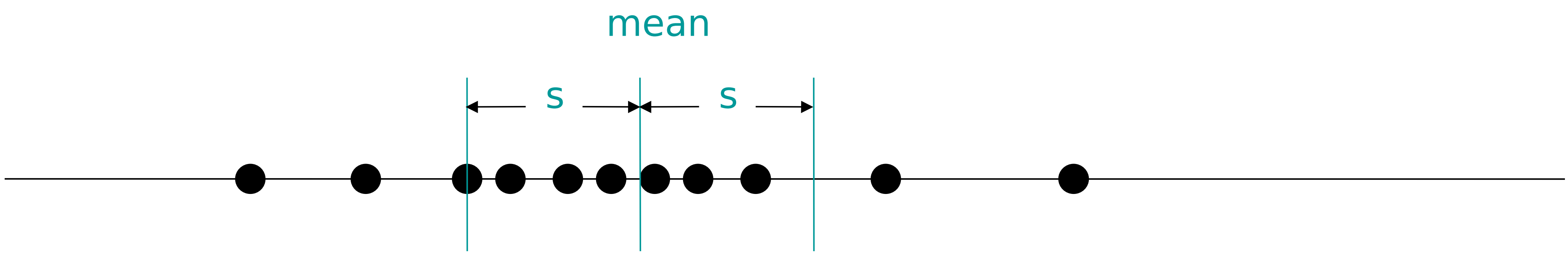
If $E(X)$ is the expected value of the random variable X then the variance $Var(X)$ is defined as: $Var(X) = E(X^2) - E(X)^2$.

If $x_1, x_2, \dots x_n$ are the data in a sample with mean m, then the sample variance s^2 is: $s^2 = (\sum(x_i - m)^2) / n$

The larger the variance, the more scattered the observations on average.

Standard Deviation

The standard deviation s is the square root of the variance: s = √Var(X)



Quantile, Quartile, Percentile

Quantile

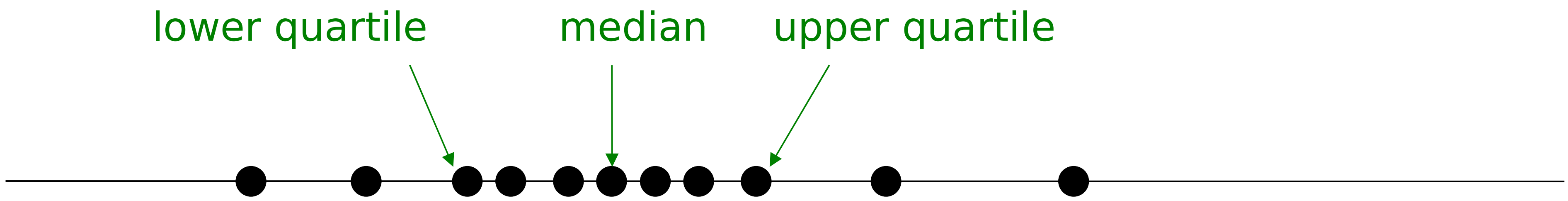
Quantiles are a set of 'cut points' that divide a sample of data into groups containing (as far as possible) equal numbers of observations.

Quartile

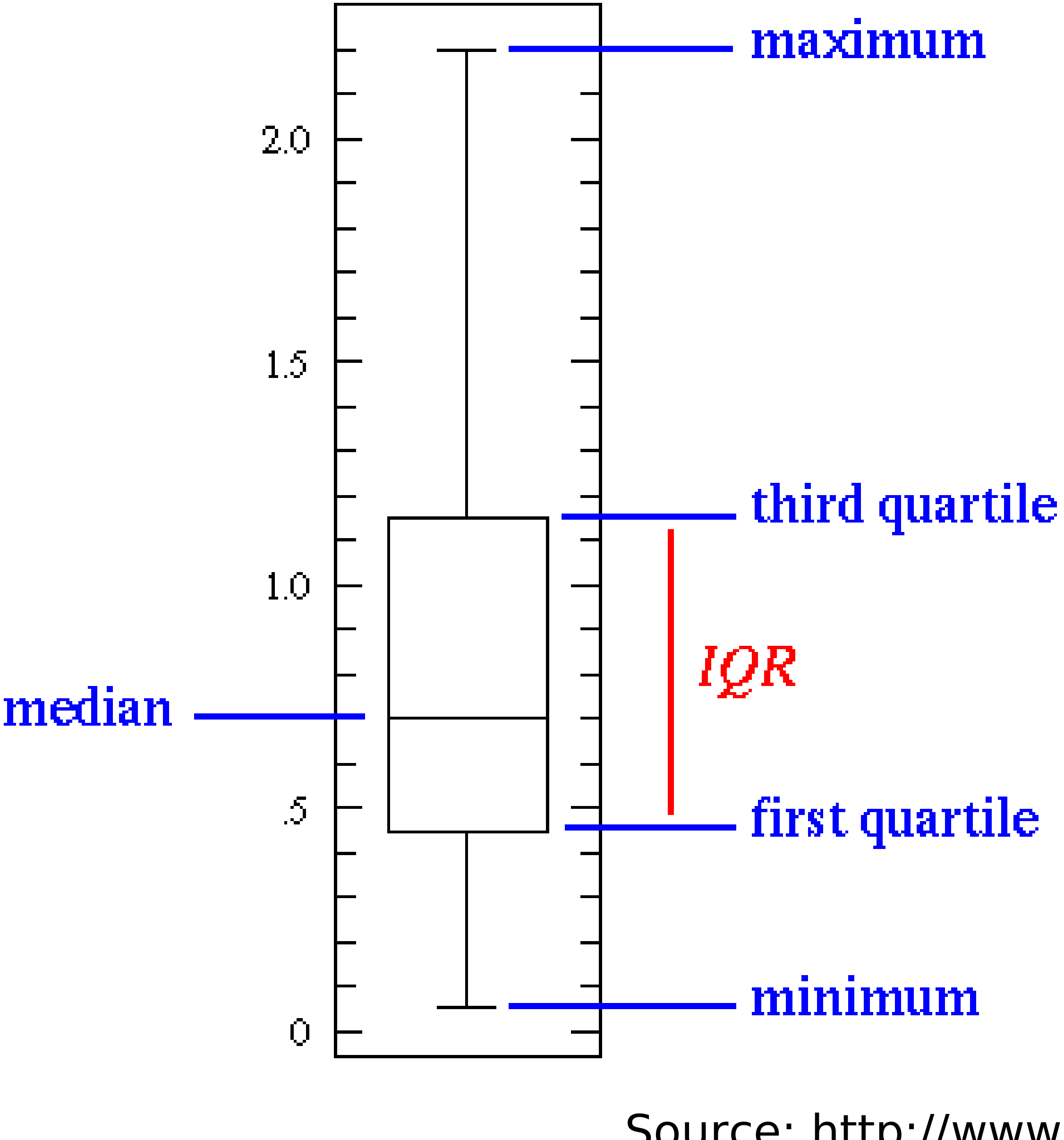
Quartiles are values that divide a sample of data into four groups containing (as far as possible) equal numbers of observations

Percentile

Quartiles are values that divide a sample of data into hundred groups containing (as far as possible) equal numbers of observations



Boxplot



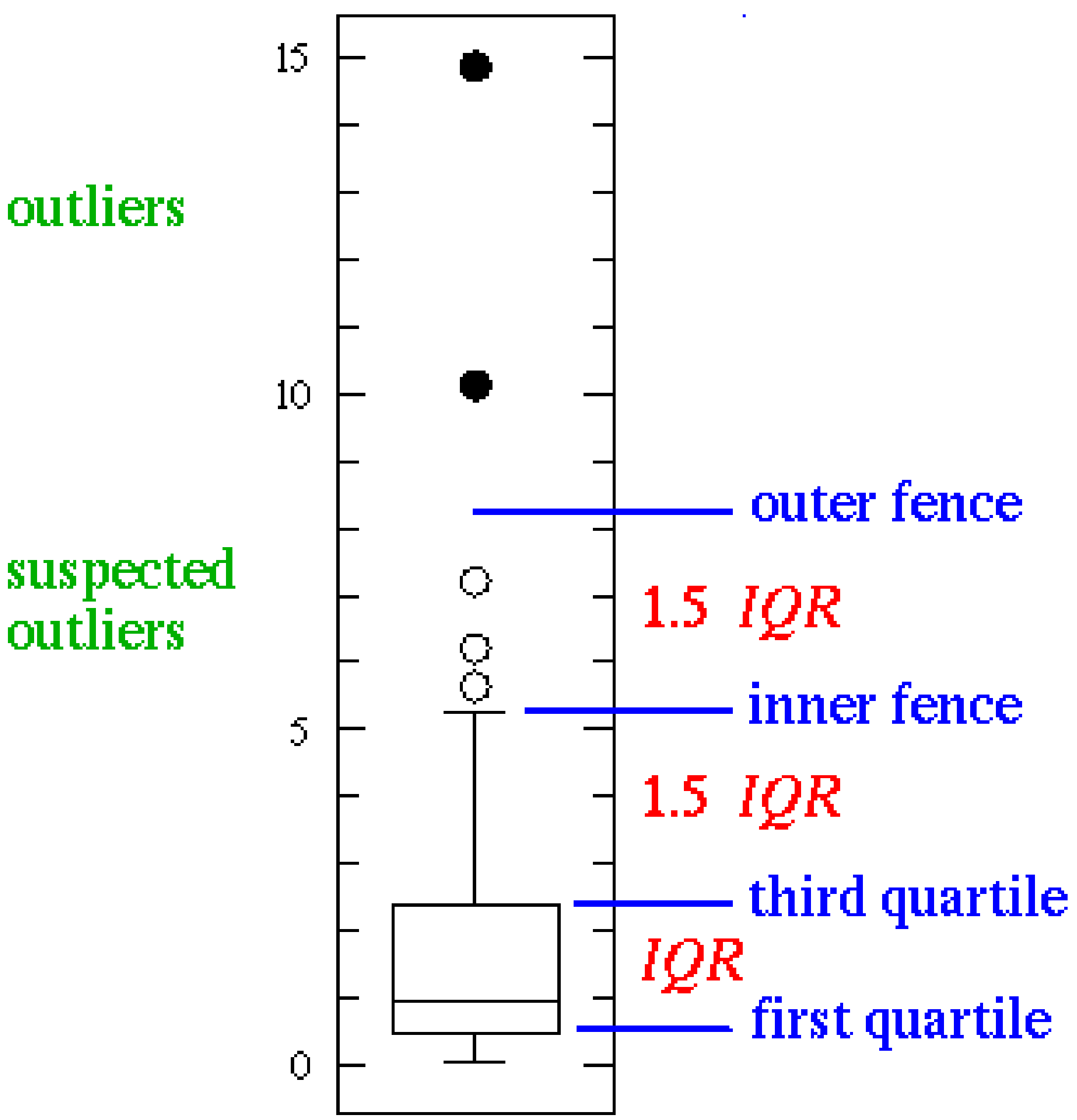
Also known as **box-and-whisker diagram** or **candlestick chart**.

Source: http://www.physics.csbsju.edu/stats/box2.html

Outliers

Try to avoid outliers

- Improve your test equipment
- Eliminate sources of disturbances
- Repeat parts of your experiment in case of disturbance



Outliers are not generally bad – they give you valuable information

With large data sets outliers can often not be avoided

Some Excel Functions

MEDIAN(Matrix)

- Matrix Data row

QUARTILE(Matrix; Quartil)

- Matrix Data row
- Quartil 0 = min, 1=lower quartile, 2 = median, 3 = upper quartile, 4 = max.

QUANTIL(Matrix; Alpha)

- Matrix Data row
- Alpha value from 0 to 1.

Box Plots with Excel 2007

http://blog.immeria.net/2007/01/box-plot-and-whisker-plots-in-excel.html

Don't Do This (II)

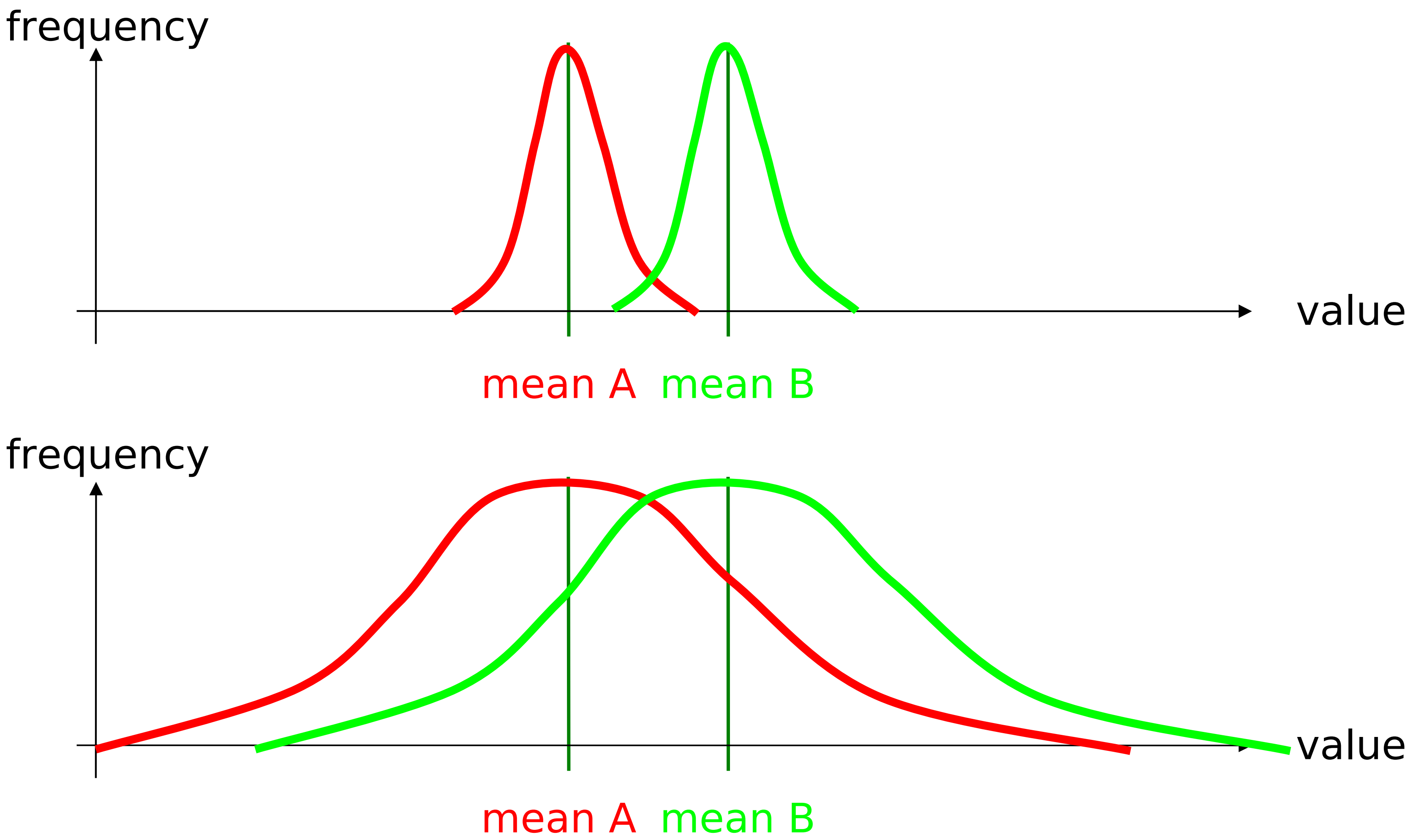
“With version A the test users needed 25 seconds in average to complete the task, but with version B it took only 21 seconds. Thus, our user study showed that version B is the better way to solve the task.”

Is the difference significant?

What does 'significant' mean?

Comparing Values

Significant differences between measurements?



Significance

In statistics, a result is called significant if it is unlikely to have occurred by chance. It does not mean that the result is of practical significance!

In the case of hypothesis testing the **significance level** is the probability that the null hypothesis ('no correlation') will be rejected in error when it is true.

Popular levels of significance are 5%, 1% and 0.1%

The t-test gives the probability that both populations have the same mean (and thus their differences are due to random noise).

A result of 0.05 from a t-test is a 5% chance for the same mean.

Student's t-Test

The t statistic was introduced by William Sealy Gosset for cheaply monitoring the quality of beer brews. "Student" was his pen name. Gosset was a statistician for the Guinness brewery in Dublin.

The t-test is a test of the null hypothesis that the means of **two normally distributed** populations are equal. The t-test gives the probability that both populations have the same mean.

(Mostly from wikipedia.org)

Student [William Sealy Gosset] (March 1908). "The probable error of a mean". Biometrika 6 (1): 1–25.

Excel: t-Test

Real data from a user study

	A	B
K1	751	1097
K2	1007	971,5
K3	716	1121
K4	1066,5	1096,5
K5	871	932
K6	1256,5	926,5
K7	957	1111
K8	1327	1211,5
K9	1482	1062
K10	881	976
Mean	**1031,5**	**1050,5**

T-test	**0,8236863**

	A	B
K1	826,5	1382
K2	806	1066
K3	791	1276,5
K4	896,5	1352
K5	696	1191
K6	1121	1066
K7	891	1217
K8	1327	1412
K9	1277	1266,5
K10	656	1101
Mean	**928,8**	**1233**

T-test	**0,0020363**

Excel functions used:

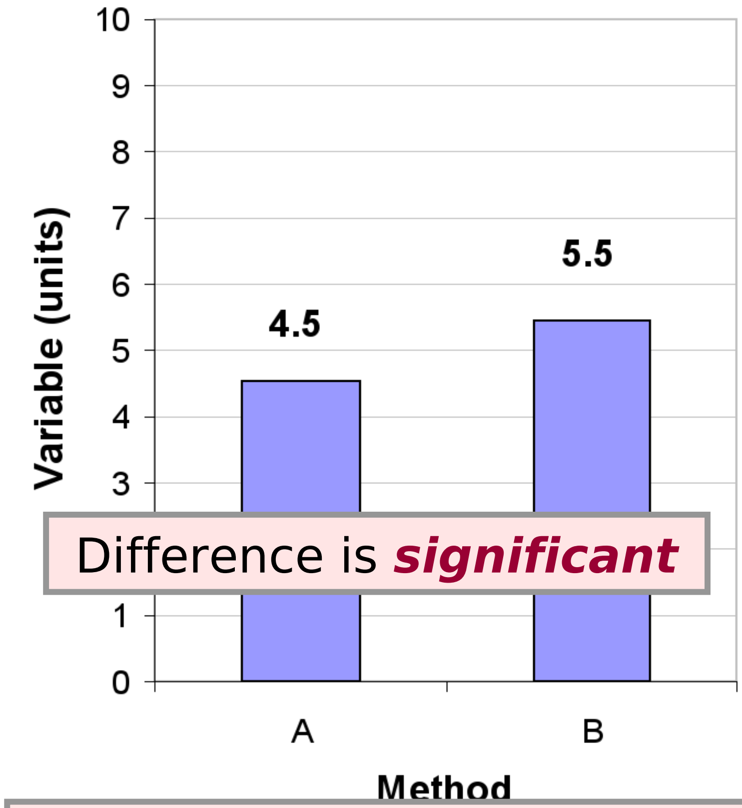
=MITTELWERT(C4:C13)
=TTEST(C4:C13;D4:D13;2;1)

(function names are localized)
Menu: Tools>Data Analysis

TTEST(...) Parameters:

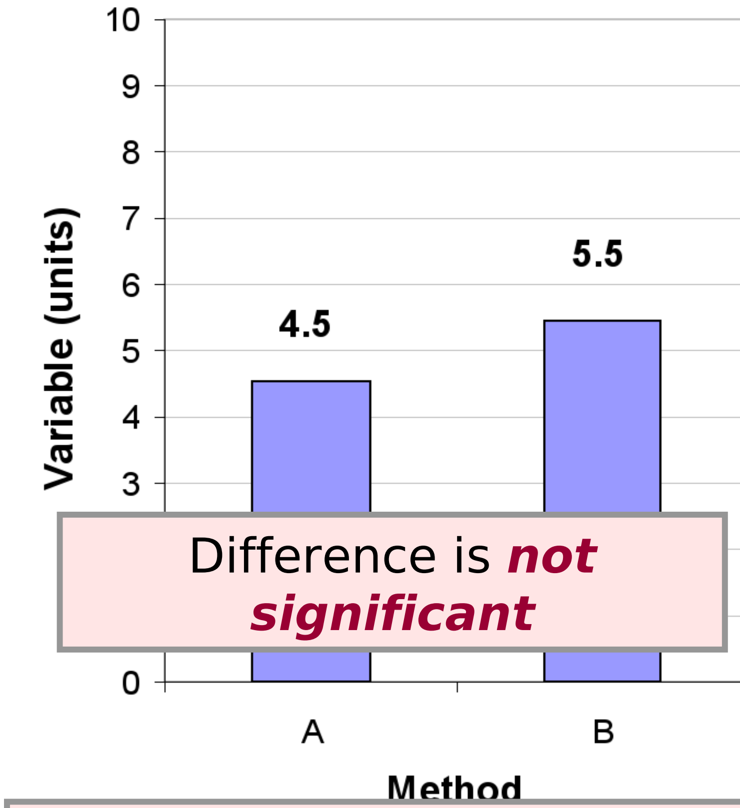
- Data row 1
- Data row 2
- Ends (1 or 2)
- Type (1=paired, 2=same variance, 3=different variance)

Example #1



“Significant” implies that in all likelihood the difference observed is due to the test conditions (Method A vs. Method B).

Example #2



“Not significant” implies that the difference observed is likely due to chance.

Source: MacKenzie, Empirical Research in HCI:What? Why? How?

Analysis of Variance (ANOVA)

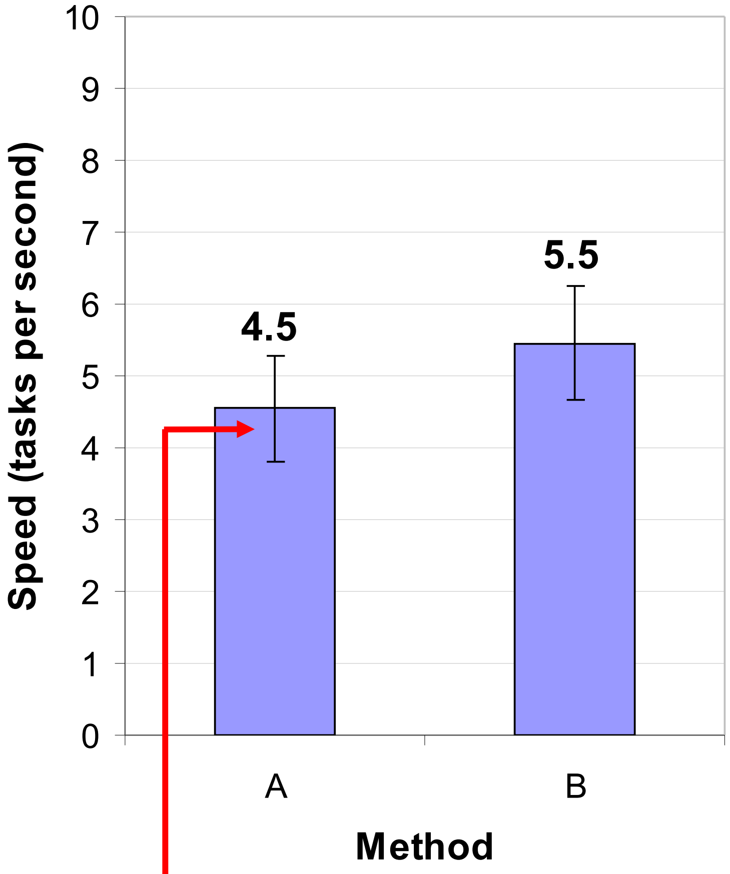
Determine if there is a significant difference between different series of measurements.

“Can the difference be explained by statistical noise?”

General Concept:

- Calculate the variance within each measurement.
- Calculate the variance in relation to the mean of all series.
- If the variance within a measurement series is much smaller than the variance in relation to the overall mean => significant!

Example #1 - Details



Example #1		
Participant	Method	
	A	B
1	5,3	5,7
2	3,6	4,6
3	5,2	5,1
4	3,3	4,5
5	4,6	6,0
6	4,1	7,0
7	4,0	6,0
8	5,0	4,6
9	5,2	5,5
10	5,1	5,6
Mean	4,5	5,5
SD	0,73	0,78

Error bars show ±1 standard deviation

Note: *SD* is the square root of the variance

Source: MacKenzie, Empirical Research in HCI:What? Why? How?

Example #1 - Anova

ANOVA Table for Speed

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	5.839	.649				
Method	1	4.161	4.161	8.443	.0174	8.443	.741
Method * Subject	9	4.435	.493				

Probability that the difference in the means is due to chance

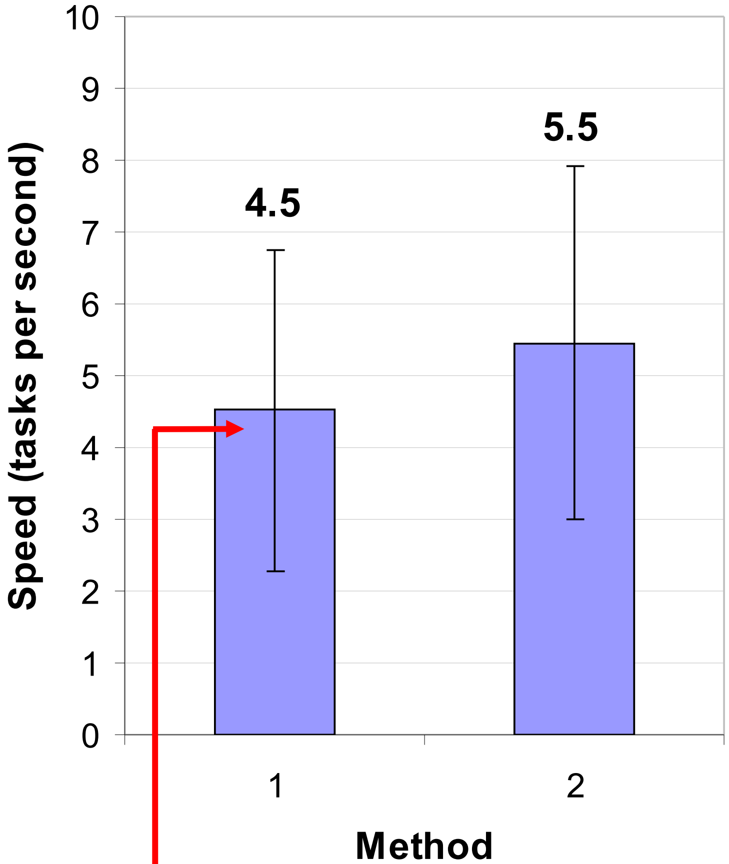
Reported as…

$F_{1,9} = 8.443$, $p < .05$

Thresholds for "p"
- .05
- .01
- .005
- .001
- .0005
- .0001

Source: MacKenzie, Empirical Research in HCI:What? Why? How?

Example #2 - Details



Example #1		
Participant	Method	
	A	B
1	5,3	5,7
2	3,6	4,6
3	5,2	5,1
4	3,3	4,5
5	4,6	6,0
6	4,1	7,0
7	4,0	6,0
8	5,0	4,6
9	5,2	5,5
10	5,1	5,6
Mean	4,5	5,5
SD	0,73	0,78

Error bars show ±1 standard deviation

Source: MacKenzie, Empirical Research in HCI:What? Why? How?

Example #2 – Anova

ANOVA Table for Speed

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	37.017	4.113				
Method	1	4.376	4.376	.634	.4462	.634	.107
Method * Subject	9	62.079	6.898				

Probability that the difference in the means is due to chance

Reported as…

$F_{1,9} = 0.634$, ns

Note: For non-significant effects, use “ns” if $F < 1.0$, or “$p > .05$” if $F > 1.0$.

Source: MacKenzie, Empirical Research in HCI:What? Why? How?

Excel: ANOVA

Anova: Single Factor

Which Bowler is Best?

SUMMARY

Groups	*Count*	*Sum*	***Average***	***Variance***
Pat	6	922	153.6667	92.26667
Mark	6	1070	178.3333	116.6667
Sheri	6	937	156.1667	54.96667

ANOVA

Source of Variation	*SS*	*df*	*MS*	***F***	*P-value*	***F crit***
Between Groups	2212.111	2	1106.056	**12.57358**	0.000621	**3.682317**
Within Groups	1319.5	15	87.96667			
Total	3531.611	17				

Tools Menu
→ Data Analysis
→ One-Way ANOVA

Source: http://www.isixsigma.com/library/content/c021111a.asp

ANOVA test online: http://www.physics.csbsju.edu/stats/anova.html

This Lecture is not Enough!

We strongly recommend to teach yourself.
There is plenty of materials on the WWW.

Further Literature

- Jürgen Bortz: Statistik für Sozialwissenschaftler, Springer
- Christel Weiß: Basiswissen Medizinische Statistik, Springer
- Lothar Sachs, Jürgen Hedderich: Angewandte Statistik, Springer
- various books by Edward R. Tufte