Statistics for User Studies

A Practical Approach
Accuracy vs. Precision

**Accuracy**
- is determined by measurement errors
- needed: good study design
- verified by: thorough description of study setup

**Precision**
- is determined by measurement noise
- needed: enough data
- verified by: rigorous statistical analysis
Types of Data

- **Categorical / Nominal Data**
  (alternatives in non-overlapping subsets, $A=B$, $A\neq B$)
  - 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 ($^\circ K/^\circ C/^\circ F$),

- **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 an order
Don't Do This
(in most cases)
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

<table>
<thead>
<tr>
<th>Days</th>
<th>Frequency</th>
<th>Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
<td>13%</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>10%</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>17%</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>17%</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>20%</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>23%</td>
</tr>
</tbody>
</table>
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
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 ⇒do not calculate mean!
• use sum-of-ranks
• you can force the user to make a commitment to one direction by offering an even number of choices.
• use 3 to 7 options

Normal Distribution

Probability of Cases in portions of the curve:
- $\approx 0.0013$ for $4\sigma$
- $\approx 0.0214$ for $3\sigma$
- $\approx 0.1359$ for $2\sigma$
- $\approx 0.3413$ for $1\sigma$
- $\approx 0.3413$ for $-1\sigma$
- $\approx 0.1359$ for $-2\sigma$
- $\approx 0.0214$ for $-3\sigma$
- $\approx 0.0013$ for $-4\sigma$

Cumulative %:
- 0.1% for $-4\sigma$
- 2.3% for $-3\sigma$
- 15.9% for $-2\sigma$
- 50% for $-1\sigma$
- 84.1% for 0
- 97.7% for $+1\sigma$
- 99.9% for $+2\sigma$
- 99.9% for $+3\sigma$
- 99.9% for $+4\sigma$

95% of values is between $-1.98\sigma$ and $1.98\sigma$

99% of values is between $-2.58\sigma$ and $2.58\sigma$

Variance and Standard Deviation

**Variance**

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

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

*Caveat: many statistical software packages calculate the "bias-corrected sample variance", dividing by $(n-1)$*

**Standard Deviation**

The standard deviation $s$ is the square root of the variance:  
$$s = \sqrt{\text{Var}(X)}$$
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.

lower quartile  median  upper quartile
Boxplot

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 form 0 to 1.

Box Plots with Excel 2007
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?
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.

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! A statistically significant speed difference of 0.1% between two text-entry methods may have little practical importance.

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

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

Different variants of the t-test are used for paired (each sample in population A has a counterpart in population B) and unpaired samples.
Examples:
  - **Paired**: speed of persons before and after treatment,
  - **Unpaired**: the reading speed of two different groups of people are compared

(mostly from wikipedia.org)

## Excel: t-Test

Real data from a user study

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>K1</td>
<td>751</td>
<td>1097</td>
</tr>
<tr>
<td>K2</td>
<td>1007</td>
<td>971,5</td>
</tr>
<tr>
<td>K3</td>
<td>716</td>
<td>1121</td>
</tr>
<tr>
<td>K4</td>
<td>1066,5</td>
<td>1096,5</td>
</tr>
<tr>
<td>K5</td>
<td>871</td>
<td>932</td>
</tr>
<tr>
<td>K6</td>
<td>1256,5</td>
<td>926,5</td>
</tr>
<tr>
<td>K7</td>
<td>957</td>
<td>1111</td>
</tr>
<tr>
<td>K8</td>
<td>1327</td>
<td>1211,5</td>
</tr>
<tr>
<td>K9</td>
<td>1482</td>
<td>1062</td>
</tr>
<tr>
<td>K10</td>
<td>881</td>
<td>976</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td><strong>1031,5</strong></td>
<td><strong>1050,5</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>K1</td>
<td>826,5</td>
<td>1382</td>
</tr>
<tr>
<td>K2</td>
<td>806</td>
<td>1066</td>
</tr>
<tr>
<td>K3</td>
<td>791</td>
<td>1276,5</td>
</tr>
<tr>
<td>K4</td>
<td>896,5</td>
<td>1352</td>
</tr>
<tr>
<td>K5</td>
<td>696</td>
<td>1191</td>
</tr>
<tr>
<td>K6</td>
<td>1121</td>
<td>1066</td>
</tr>
<tr>
<td>K7</td>
<td>891</td>
<td>1217</td>
</tr>
<tr>
<td>K8</td>
<td>1327</td>
<td>1412</td>
</tr>
<tr>
<td>K9</td>
<td>1277</td>
<td>1266,5</td>
</tr>
<tr>
<td>K10</td>
<td>656</td>
<td>1101</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td><strong>928,8</strong></td>
<td><strong>1233</strong></td>
</tr>
</tbody>
</table>

**T-test** 0,8236863  
**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) (usually 2)
- Type (1=paired, 2=same variance, 3=_different variance)
T-Test Caveats

• can only be applied to two populations
• do not add significance levels from different tests
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

<table>
<thead>
<tr>
<th>Participant</th>
<th>Method</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>5,3</td>
<td>5,7</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>3,6</td>
<td>4,6</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>5,2</td>
<td>5,1</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>3,3</td>
<td>4,5</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>4,6</td>
<td>6,0</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>4,1</td>
<td>7,0</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>4,0</td>
<td>6,0</td>
</tr>
<tr>
<td>8</td>
<td>A</td>
<td>5,0</td>
<td>4,6</td>
</tr>
<tr>
<td>9</td>
<td>A</td>
<td>5,2</td>
<td>5,5</td>
</tr>
<tr>
<td>10</td>
<td>A</td>
<td>5,1</td>
<td>5,6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Method</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Example #1</td>
<td>Mean</td>
<td>4,5</td>
<td>5,5</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0,73</td>
<td>0,78</td>
</tr>
</tbody>
</table>

Error bars show ±1 standard deviation

Example #1 - Anova

ANOVA Table for Speed

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F-Value</th>
<th>P-Value</th>
<th>Lambda</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>9</td>
<td>5.839</td>
<td>.649</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>1</td>
<td>4.161</td>
<td>4.161</td>
<td>8.443</td>
<td>.0174</td>
<td>8.443</td>
<td>.741</td>
</tr>
<tr>
<td>Method * Subject</td>
<td>9</td>
<td>4.435</td>
<td>.493</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Probability that the difference in the means is due to chance

Reported as...

\[ F_{1,9} = 8.443, p < .05 \]

\[ F = \frac{\text{explained variance}}{\text{unexplained variance}} \]

Thresholds for “p”
- .05
- .01
- .005
- .001
- .0005
- .0001

Example #2 - Details


![Bar Chart]

### Example #1

<table>
<thead>
<tr>
<th>Participant</th>
<th>Method A</th>
<th>Method B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5,3</td>
<td>5,7</td>
</tr>
<tr>
<td>2</td>
<td>3,6</td>
<td>4,6</td>
</tr>
<tr>
<td>3</td>
<td>5,2</td>
<td>5,1</td>
</tr>
<tr>
<td>4</td>
<td>3,3</td>
<td>4,5</td>
</tr>
<tr>
<td>5</td>
<td>4,6</td>
<td>6,0</td>
</tr>
<tr>
<td>6</td>
<td>4,1</td>
<td>7,0</td>
</tr>
<tr>
<td>7</td>
<td>4,0</td>
<td>6,0</td>
</tr>
<tr>
<td>8</td>
<td>5,0</td>
<td>4,6</td>
</tr>
<tr>
<td>9</td>
<td>5,2</td>
<td>5,5</td>
</tr>
<tr>
<td>10</td>
<td>5,1</td>
<td>5,6</td>
</tr>
</tbody>
</table>

**Mean:**

<table>
<thead>
<tr>
<th>Method</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>4,5</td>
<td>5,5</td>
</tr>
<tr>
<td>SD</td>
<td>0,73</td>
<td>0,78</td>
</tr>
</tbody>
</table>

**Error bars show ±1 standard deviation**
**Example #2 – Anova**

**ANOVA Table for Speed**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F-Value</th>
<th>P-Value</th>
<th>Lambda</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>9</td>
<td>37.017</td>
<td>4.113</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>1</td>
<td>4.376</td>
<td>4.376</td>
<td>.634</td>
<td>.4462</td>
<td>.634</td>
<td>.107</td>
</tr>
<tr>
<td>Method * Subject</td>
<td>9</td>
<td>62.079</td>
<td>6.898</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Probability that the difference in the means is due to chance

Reported as...

\[ F_{1,9} = 0.634, \text{ ns} \]

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

Excel: ANOVA

<table>
<thead>
<tr>
<th>Groups</th>
<th>Count</th>
<th>Sum</th>
<th>Average</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pat</td>
<td>6</td>
<td>922 153.6667</td>
<td>92.26667</td>
<td></td>
</tr>
<tr>
<td>Mark</td>
<td>6</td>
<td>1070 178.3333</td>
<td>116.6667</td>
<td></td>
</tr>
<tr>
<td>Sheri</td>
<td>6</td>
<td>937 156.1667</td>
<td>54.96667</td>
<td></td>
</tr>
</tbody>
</table>

**ANOVAS**

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P-value</th>
<th>F crit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>2212.111</td>
<td>2</td>
<td>1106.056</td>
<td>12.57358</td>
<td>0.000621</td>
<td>3.682317</td>
</tr>
<tr>
<td>Within Groups</td>
<td>1319.5</td>
<td>15</td>
<td>87.9667</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3531.611</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: [http://www.isixsigma.com/library/content/c021111a.asp](http://www.isixsigma.com/library/content/c021111a.asp)

ANOVA test online: [http://www.physics.csbsju.edu/stats/anova.html](http://www.physics.csbsju.edu/stats/anova.html)
Java API for ANOVA: [http://www.yorku.ca/mack/RN-Anova.html](http://www.yorku.ca/mack/RN-Anova.html)
This Lecture is not Enough!

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

Further Literature
Andy Field & Graham Hole: How to design and report experiments, Sage
• 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