Statistics for User Studies

A Practical Approach

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

Participants

K1
K2
K3
K4
K5
K6
K7
K8
K9
K10

0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500

time

Don't Do This
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>
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

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
Variance and Standard Deviation

**Variance**
If \( E(X) \) is the expected value of the random variable \( X \) then the variance \( \text{Var}(X) \) is defined as: \( \text{Var}(X) = E(X^2) - E(X)^2 \).
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.

**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.
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
“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)

# 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>
</tbody>
</table>

**Mean**: 1031.5 | 1050.5

**T-test**: 0.8236863

Excel functions used:

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

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

<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>
</tbody>
</table>

**Mean**: 928.8 | 1233

**T-test**: 0.0020363
Example #1

Difference is **significant**

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

Example #2

Difference is **not significant**

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

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

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

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...

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

Thresholds for “p”

.05
.01
.005
.001
.0005
.0001

Example #2 - Details

- 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>

Reported as...

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

Probability that the difference in the means is due to chance

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

Excel: ANOVA

### Anova: Single Factor

#### Which Bowler is Best?

<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 92.26667</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mark</td>
<td>6</td>
<td>1070 178.3333 116.6667</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sheri</td>
<td>6</td>
<td>937 156.1667 54.96667</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### ANOVA

<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></td>
<td>1587.9667</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)
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 Weißl: Basiswissen Medizinische Statistik, Springer
• Lothar Sachs, Jürgen Hedderich: Angewandte Statistik, Springer
• various books by Edward R. Tufte