PV182
Human Computer Interaction

Lecture 6

Evaluating Controlled Experiments

Fotis Liarokapis
liarokap@fi.muni.cz

22nd October 2018

Controlled Experiments

- What is experimental design?
- What is an experimental hypothesis?
- How do I plan an experiment?
- Why are statistics used?
- What are the important statistical methods?

Quantitative Evaluation of Systems

- Quantitative:
  - Precise measurement, numerical values
  - Bounds on how correct our statements are
- Methods
  - User performance data collection
  - Controlled experiments

Collecting User Performance Data

- Data collected on system use (often lots of data)
- Exploratory:
  - Hope something interesting shows up
  - But difficult to analyze

Collecting User Performance Data .

- Targeted
  - Look for specific information, but may miss something
    - Frequency of request for on-line assistance
      - What did people ask for help with?
    - Frequency of use of different parts of the system
      - Why are parts of system unused?
    - Number of errors and where they occurred
      - Why does an error occur repeatedly?
    - Time it takes to complete some operation
      - What tasks take longer than expected?

Controlled Experiments

- Traditional scientific method
- Reductionist
  - Clear convincing result on specific issues
- In HCI:
  - Insights into cognitive process, human performance limitations, ...
  - Allows system comparison, fine-tuning of details ...

Controlled Experiments .

- Strives for:
  - Lucid and testable hypothesis
  - Quantitative measurement
  - Measure of confidence in results obtained (statistics)
  - Replicability of experiment
  - Control of variables and conditions
  - Removal of experimenter bias

Controlled Experiments ..

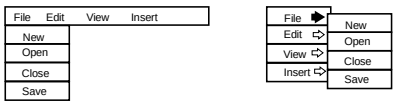
- Subjects in experiments:
  - Between-subjects (randomized design of measurement, each participant is assigned to a different condition)
  - Within-subjects (repeated measures, each user performs under each different condition)

Controlled Experiments Example

https://www.youtube.com/watch?v=D3ZB2RTylR4

Clear and Testable Hypothesis

- State a clear, testable hypothesis
  - This is a precise problem statement
- Example:
  - There is no difference in user performance (time and error rate) when selecting a single item from a pop-up or a pull down menu of 4 items, regardless of the subject's previous expertise in using a mouse or using the different menu types"



Independent Variables

- Hypothesis includes the <u>independent variables</u> that are to be altered
  - The things you manipulate independent of a subject's behaviour
  - Determines a modification to the conditions the subjects undergo
  - May arise from subjects being classified into different groups

Independent Variables .

- Menu experiment
  - Menu type: pop-up or pull-down
  - Menu length: 3, 6, 9, 12, 15
  - Subject type (expert or novice)

Dependent Variables

- Hypothesis includes the dependent variables that will be measured
  - Variables dependent on the subject's behaviour / reaction to the independent variable
  - The specific things you set out to quantitatively measure / observe

Dependent Variables .

- Menu experiment
  - Time to select an item
  - Selection errors made
  - Time to learn to use it to proficiency

Independent and Dependent Variables

https://www.youtube.com/watch?v=aeH1Fzqd0Q0

Scales of Measurements

- Four major scales of measurements
  - Nominal
  - Ordinal
  - Interval
  - Ratio

Nominal Scale

- Classification into named or numbered unordered categories
  - Country of birth, user groups, gender…
- Allowable manipulations
  - Whether an item belongs in a category
  - Counting items in a category
- Statistics
  - Number of cases in each category
  - Most frequent category
  - No means, medians…

Nominal Scale .

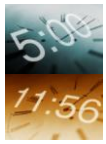
- Sources of error
  - Agreement in labelling, vague labels, vague differences in objects
- Testing for error
  - Agreement between different judges for same object

Ordinal Scale

- Classification into named or numbered ordered categories
  - No information on magnitude of differences between categories
    - i.e. Preference, social status, gold/silver/bronze medals
- Allowable manipulations
  - As with interval scale, plus
  - Merge adjacent classes
  - Transitive: if A > B > C, then A > C
- Statistics
  - Median (central value)
  - Percentiles, e.g., 30% were less than B
- Sources of error
  - As in nominal

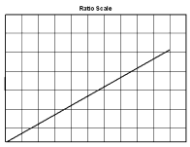
Interval Scale

- Classification into ordered categories with equal differences between categories
  - Zero only by convention
    - i.e. Temperature (C or F), time of day
- Allowable manipulations
  - Add, subtract
  - Cannot multiply as this needs an absolute zero
- Statistics
  - Mean, standard deviation, range, variance
- Sources of error
  - Instrument calibration, reproducibility and readability
  - Human error, skill…



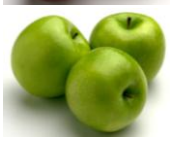
Ratio Scale

- Interval scale with absolute, non-arbitrary zero
  - i.e. temperature (K), length, weight, time periods
- Allowable manipulations
  - Multiply, divide



Example: Apples

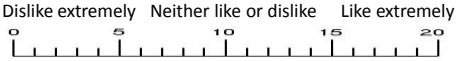
- Nominal:
  - Apple variety
  - Macintosh, Delicious, Gala…
- Ordinal:
  - Apple quality
  - US. Extra Fancy
  - U.S. Fancy
  - U.S. Combination Extra Fancy / Fancy
  - U.S. No. 1
  - U.S. Early
  - U.S. Utility
  - U.S. Hail



Example: Apples .

- Interval:
  - Apple 'Liking scale'
    - Marin, A. Consumers' evaluation of apple quality. Washington Tree Postharvest Conference 2002.

After taking at least 2 bites how much do you like the apple?
Dislike extremely Neither like or dislike Like extremely



- Ratio:
  - Apple weight, size, …

Subject Selection

- Judiciously select and assign subjects to groups
  - Ways of controlling subject variability
  - Reasonable amount of subjects
  - Random assignment
  - Make different user groups an independent variable
  - Screen for anomalies in subject group
    - Superstars versus poor performers

Controlling Bias

- Unbiased instructions
- Unbiased experimental protocols
  - Prepare scripts ahead of time
- Unbiased subject selection

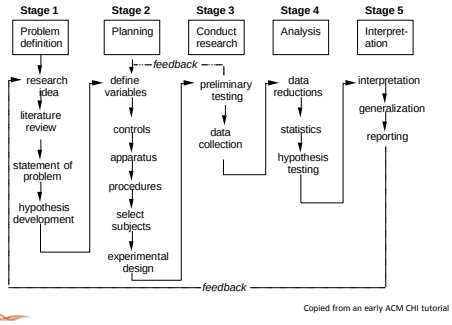
Statistical Analysis

- Apply statistical methods to data analysis
- Confidence limits:
  - The confidence that your conclusion is correct
  - "the hypothesis that computer experience makes no difference is rejected at the .05 level" means:
    - A 95% chance that your statement is correct
    - A 5% chance you are wrong

Interpretation

- What you believe the results really mean
- Their implications to your research
- Their implications to practitioners
- How generalizable they are
- Limitations and critique

Planning Flowchart for Experiments



Copied from an early ACM CHI tutorial

Statistical Analysis

- Calculations that tell us
  - Mathematical attributes about our data sets
    - Mean, amount of variance, ...
  - How data sets relate to each other
    - Whether we are "sampling" from the same or different distributions
  - The probability that our claims are correct
    - "statistical significance"

Statistical vs Practical Significance

- When n is large, even a trivial difference may show up as a statistically significant result
  - i.e. menu choice:

mean selection time of menu a is 3.00 sec
menu b is 3.05 sec

- Statistical significance does not imply that the difference is important!
  - A matter of interpretation
  - Statistical significance often abused and used to misinform

Problem with Visual Inspection of Data

- Will almost always see variation in collected data
- Differences between data sets may be due to:
  - Normal variation
    - i.e. two sets of ten tosses with different but fair dice
      - Differences between data and means are accountable by expected variation
  - Real differences between data
    - i.e. two sets of ten tosses for with loaded dice and fair dice
      - Differences between data and means are not accountable by expected variation

T-test

- A simple statistical test
  - Allows one to say something about differences between means at a certain confidence level
- Null hypothesis of the T-test:
  - No difference exists between the means of two sets of collected data
- Possible results:
  - I am 95% sure that null hypothesis is rejected
    - There is probably a true difference between the means
- I cannot reject the null hypothesis
  - The means are likely the same

Wikipedia

- The t statistic was introduced in 1908 by William Sealy Gosset, a statistician working for the Guinness brewery in Dublin, Ireland ("Student" was his pen name). Gosset had been hired due to Claude Guinness's innovative policy of recruiting the best graduates from Oxford and Cambridge to apply biochemistry and statistics to Guinness' industrial processes. Gosset devised the t-test as a way to cheaply monitor the quality of beer. He published the test in Biometrika in 1908, but was forced to use a pen name by his employer, who regarded the fact that they were using statistics as a trade secret. In fact, Gosset's identity was known to fellow statisticians.
- Today, the t-test is more generally applied to the confidence that can be placed in judgments made from small samples.

Different Types of T-tests

- Comparing two sets of independent observations
  - Usually different subjects in each group
  - Number per group may differ as well

Condition 1	Condition 2
S1–S20	S21–43

- Paired observations
  - Usually a single group studied under both experimental conditions
  - Data points of one subject are treated as a pair

Condition 1	Condition 2
S1–S20	S1–S20

Different Types of T-tests .

- Non-directional vs directional alternatives
  - Non-directional (two-tailed)
    - No expectation that the direction of difference matters
  - Directional (one-tailed)
    - Only interested if the mean of a given condition is greater than the other

T-test Assumptions

- Assumptions of t-tests
  - Data points of each sample are normally distributed
    - But t-test very robust in practice
  - Population variances are equal
    - t-test reasonably robust for differing variances
    - Deserves consideration
  - Individual observations of data points in sample are independent
    - Must be adhered to
- Significance level
  - Decide upon the level before you do the test!
  - Typically stated at the .05 or .01 level

Two-tailed Unpaired T-test

N: number of data points in the one sample
ΣX: sum of all data points in one sample
X: mean of data points in sample
Σ(X²): sum of squares of data points in sample
s²: unbiased estimate of population variation
t: t ratio
df = degrees of freedom = N1 + N2 – 2

Formulas:

$$s^2 = \frac{\Sigma(X_1^2) - \frac{(\Sigma X_1)^2}{N_1} + \Sigma(X_2^2) - \frac{(\Sigma X_2)^2}{N_2}}{N_1 + N_2 - 2} \qquad t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s^2}{N_1} + \frac{s^2}{N_2}}}$$

Level of Significance for Two-tailed test

df	.05	.01
1	12.706	63.657
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
6	2.447	3.707
7	2.365	3.499
8	2.306	3.355
9	2.262	3.250
10	2.228	3.169
11	2.201	3.106
12	2.179	3.055
13	2.160	3.012
14	2.145	2.977
15	2.131	2.947
16	2.120	2.921
18	2.101	2.878
20	2.086	2.845
22	2.074	2.819
24	2.064	2.797

Two-tailed Unpaired T-test

- Or, use a statistics package (e.g., Excel has simple stats)
  - Condition one: 3, 4, 4, 4, 5, 5, 5, 6
  - Condition two: 4, 4, 5, 5, 6, 6, 7, 7

Unpaired t-test

DF:	Unpaired t Value:	Prob. (2-tail):
14	-1.871	.0824

Group:	Count:	Mean:	Std. Dev.:	Std. Error:
one	8	4.5	.926	.327
two	8	5.5	1.195	.423

ANOVA

- Analysis of variance (ANOVA) is a collection of statistical models used to analyze the differences among group means and their associated procedures (such as "variation" among and between groups)
  - Developed by statistician and evolutionary biologist Ronald Fisher
- In the ANOVA setting, the observed variance in a particular variable is partitioned into components attributable to different sources of variation

https://en.wikipedia.org/wiki/Analysis_of_variance

ANOVA

- In its simplest form, ANOVA provides a statistical test of whether or not the means of several groups are equal, and therefore generalizes the t-test to more than two groups
- ANOVAs are useful for comparing (testing) three or more means (groups or variables) for statistical significance
  - It is conceptually similar to multiple two-sample t-tests, but is more conservative (results in less type I error) and is therefore suited to a wide range of practical problems

https://en.wikipedia.org/wiki/Analysis_of_variance

Significance Levels and Errors

- Type 1 error
  - Reject the null hypothesis when it is, in fact, true
- Type 2 error
  - Accept the null hypothesis when it is, in fact, false
- Effects of levels of significance
  - High confidence level (e.g. p<.0001)
    - Greater chance of Type 2 errors
  - Low confidence level (e.g. p>.1)
    - Greater chance of Type 1 errors
- You can 'bias' your choice depending on consequence of these errors

Type I and Type II Errors

- Type 1 error
  - Reject the null hypothesis when it is, in fact, true
- Type 2 error
  - Accept the null hypothesis when it is, in fact, false

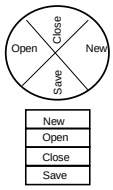
"Reality"	Decision: ***False***	Decision: ***True***
True	Type I error	✓
False	✓	Type II error

Which is Worse?

- Type I errors are considered worse
  - Because the null hypothesis is meant to reflect the incumbent theory
- BUT
  - You must use your judgement to assess actual risk of being wrong in the context of your study

Significance Levels and Errors

- There is no difference between Pie and traditional pop-up menus
- What is the consequence of each error type?
  - Type 1:
    - Extra work developing software
    - People must learn a new idiom for no benefit
  - Type 2:
    - Use a less efficient (but already familiar) menu
- Which error type is preferable?
  - Redesigning a traditional GUI interface
    - Type 2 error is preferable to a Type 1 error
  - Designing a digital mapping application where experts perform extremely frequent menu selections
    - Type 1 error preferable to a Type 2 error

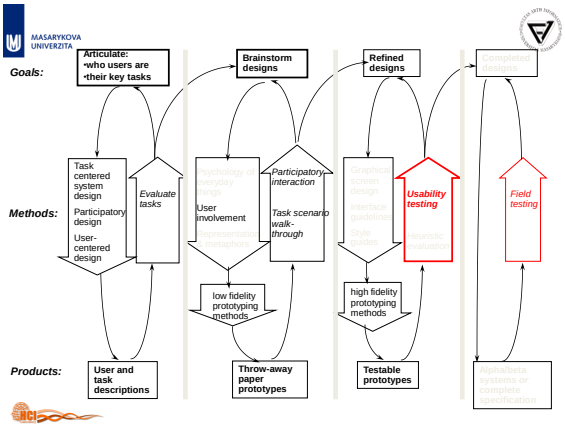


You Know Now

- Controlled experiments can provide clear convincing result on specific issues
- Creating testable hypotheses are critical to good experimental design
- Experimental design requires a great deal of planning
- Statistics inform us about
  - Mathematical attributes about our data sets
  - How data sets relate to each other
  - The probability that our claims are correct

You Know Now .

- Statistics inform us about
  - Mathematical attributes about our data sets
  - How data sets relate to each other
  - The probability that our claims are correct
- There are many statistical methods that can be applied to different experimental designs
  - T-tests



Questions

Acknowledgements

- Prof. Ing. Jiří Sochor