Quantitative Evaluation

Research Questions
Quantitative Data
Controlled Studies
Experimental Methods
Role of Statistics

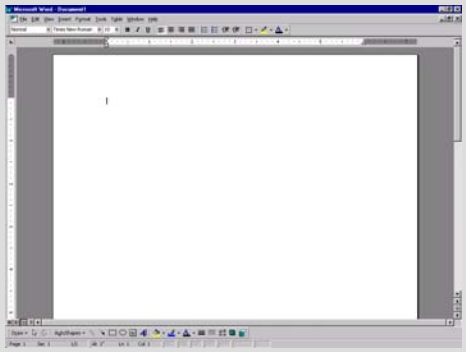
Quantitative Evaluation

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?

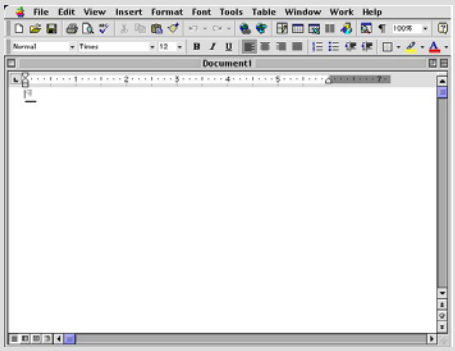
Research Question

Which menu placement system is better?

Top of Window



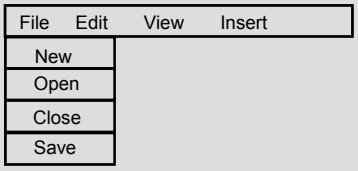
Top of Screen

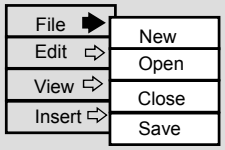


What problems would exist if we attempt to answer this research question with these screens?

Research Question

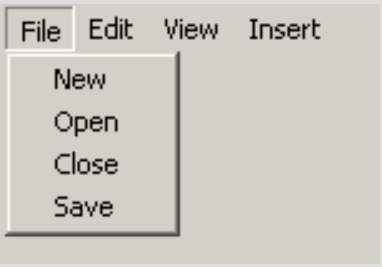
Which menu layout's design is better?

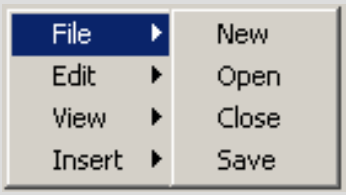




What problems would exist if we attempt to answer this research question with menus with this appearance?

Build realism, even if not “real”





https://youtu.be/wFWbdxicvK0?t=123 ← Early, mixed methods, work on touchscreen toggle concepts. Pay note at around 3:05 and 4:40 and to some of the researcher’s reflections on the impact of the visual designs used in this experiment.

Quantitative Methods

User performance data collection:

- data can be collected on system use
  - 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?
- tends to collect ***much*** data (sometimes just in the hope that something interesting shows up)
- can be difficult to sift through data unless specific aspects are targeted (as in list above)

Quantitative Methods Experiments

Controlled experiments

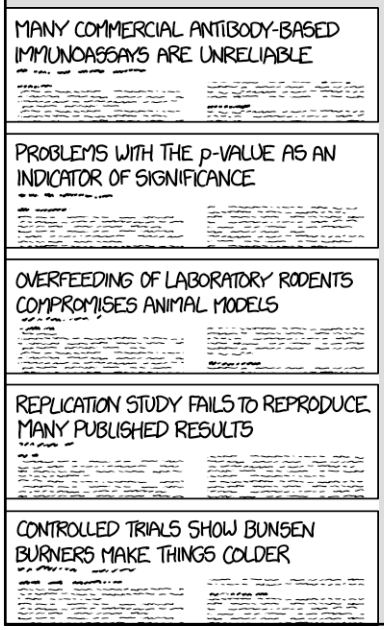
- A “traditional” scientific method which is said to provide clear and convincing results on specific issues (though we’ve seem some questions on this).
- In HCI research this approach can provide insights into human cognitive processes, performance limitations, etc. and also allows comparison of systems / fine-tuning of details.

Experimental Design

Strives to have…

- lucid and testable hypothesis
- quantitative measurement
- measure of confidence in results obtained (statistics)
- repeatability of experiment
- control of variables and conditions
- removal of experimenter bias

Experimental Methods



https://www.explainxkcd.com/wiki/index.php/1574:_Trouble_for_Science

Experimental Methods: Clear Hypothesis

Begin with a lucid, testable hypothesis.

"there is no difference in the number of cavities in children and teenagers using ***Crest*** and ***Our*** toothpaste"

"there is no difference in user performance (time, error rate, and subjective satisfaction) when selecting a single item from a ***pop-up*** or a ***pull-down*** menu, regardless of the subject's previous expertise in using a mouse or using the different menu types"

Experimental Methods: Independent Variables

Explicitly state the **independent variables** that are to be altered / controlled. These variables…

- are the things you manipulate/control independent of how a subject behaves
- determines a modification to the conditions the subjects undergo
- may arise from subjects being classified into different groups

– *in the toothpaste experiment example, could be*

- toothpaste type: uses **Crest** or uses **Our** toothpaste
- age: ≤11 years old *or* >11 years old

– *in the menu experiment example, could be*

- menu type: **pop-up** or **pull-down**
- menu length: 3, 6, 9, 12, 15
- participant type (expert or novice)

Experimental Methods: Dependent Variables

Carefully choose the **dependent variables** that will be measured. These are the variables dependent on the subject's behavior / reaction to the independent variable

– *in the toothpaste experiment example, could be*

- number of cavities
- frequency of brushing

– *in the menu experiment example, could be*

- time to select an item
- selection errors made
- subjective satisfaction as reported in a questionnaire

Experimental Methods: Subject Assignments

Judiciously select and assign subjects to groups. Consider ways of controlling subject variability…

- recognize classes (novice/expert, age ranges, etc.) and make them an independent variable
- minimize unaccounted anomalies in subject group (such as superstar users versus poor performers)
- use a reasonably large number of participants and random assignment to groups (the standard for "reasonably" large can vary based on domain and study type)

Experimental Methods: Bias

Control for biasing factors as much as possible.

Recall concerns such as the Hawthorne Effect, Pygmalion Effect, and Clever Hans Effect from earlier in the semester…

- Design unbiased instructions and experimental protocols that are prepared, reviewed, and then practiced ahead of time.
- Consider approaches such as double-blind experiments where the person running the study doesn't know what's be studied either.

Within-Subject and Between-Subject Tests

For **within-subject** testing, you have each participant try all treatments/variations.

For **between-subject** testing, each participant only tries a single treatment/variation.

For example: MenuA –vs- MenuB for speed

- Within-subject: Person does experimental tasks using MenuA and then again using MenuB (vary the order so half use MenuA first and half use MenuB first to address any learning curve on the problem itself).

- Between-subject: Person does experimental tasks using ***EITHER*** MenuA or MenuB (not both).

Which to use? Between or Within?

There are pros and cons to choosing **within** or **between** subject testing approaches.

An example of a "pro" of using the within-subject approach is that you can have ***relative speeds*** on same user. This can minimize the effect of some users being atypically fast or slow.

An example of a "con" of within-subject is that there can be a significant learning effect between the participant experiencing the different versions. Varying the order of presentation can help with this.

Experimental Methods: How many variables?

What if there are more than two independent variables that you want to test?

What if there are more than two variations of an independent variable?

CMSC250 time…

- how many orders of treatments if there are 3 variations?
- how many orders if there are two variables, each having 2 variations?

Example: Within-Subject, Four Approaches

We can work to remove the learning effect as a confounding variable if we counterbalance the order in which participants experience things. Some suggest incomplete "Latin Squares" counterbalancing, such as:

	Approach 1	**Approach 2**	**Approach 3**	**Approach 4**
Ordering 1	First	Second	Third	Fourth
Ordering 2	Second	Third	Fourth	First
Ordering 3	Third	Fourth	First	Second
Ordering 4	Fourth	First	Second	Third

Note that this is NOT every ordering possible (which would be 24 different ones).

You would want each ordering used multiple times, evenly across your population sample, which is why using all 24 would make for a LARGE number of participants.

Experimental Methods: Statistics

You will need to apply the appropriate statistical methods to data analysis and interpret your results…

- "The hypothesis that menu design choice makes no difference is rejected at the .05 level."
- "Users can select option from pull-down menus 15% faster than pop-out menus, and that result is statistically significant."

Recall things like "0.05 p-values means there's at most a 95% chance that your statement is correct" from your statistics courses and keep in mind that this means there is a 5% chance you are wrong… https://xkcd.com/1478/

Statistics Analysis

These are 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"

Beware though… https://xkcd.com/882/

Visual inspection of data

There can be problems with attempting to rely on a visual inspection of data (as we've discussed earlier). There is almost always variation in collected data. Differences between data sets may be due to normal and expected variations or represent actual differences.

- Normal variation such as two sets of ten rolls with different but fair dice. The differences between data and means are accountable by expected variation.
- True differences between data, such as two sets of ten rolls but one set with loaded dice and the other with fair dice, can be found because the differences between data and means will ***not*** be accountable by expected variation.

In brief, take STAT 400 seriously!

Statistical vs Practical significance

When the number of participants in a study is large, even a trivial difference may be large enough to produce a "statistically significant" result, but is it of practical significance?

- Imagine a statistically significant result showing an average selection time of 3 seconds for menu style A and 3.05 seconds for menu style B.
- Statistical significance does not imply that the difference is important! This ends up being a matter of interpretation…

Averages

Given two data sets measuring a conditions (cavities based on which toothpaste, time to select an item based on which menu style) we could ask whether the difference between the averages of the data sets is statistically significant

Null hypothesis would be that there is no difference between the two means.

- statistical analysis can only reject the hypothesis at a certain level of confidence

t-test (brief version)

A statistical test that can be applied to fairly small ($n<30$) data sets which follow a normal distribution and equal variances and then allows one to say something about differences between means at a certain confidence level.

- Can use independent (unpaired) samples as long as they each follow the same distribution as each other.
- Can use paired samples (each participant gets measured or two things for example).

The null hypothesis of the t-test is that no difference exists between the average of two data sets.

Correlation (brief version)

Measures the extent to which two concepts are related to each other.

- obtain the two sets of measurements
- calculate correlation coefficient
  - +1: positively correlated
  - 0: no correlation (no relation)
  - -1: negatively correlated

Don't…

- attribute ***causality*** just based on correlation
- try to draw strong conclusion from small data sets

Regression Tests (brief version)

Calculate a "best fit" line for existing data plotted onto an x,y coordinate system and then try to use a value of one variable to predict a value of the other in data you don't have.

ANOVA tests (brief version)

- Compares the relationships between multiple (rather than two) factors.
- Provides you with more “informed” results since it considers the interactions between the different factors.
- Imagine having typists at different skills levels using different keyboard layouts (alphabetic, querty, dvorak) and then being able to conclude things like:
  - beginners type at the same speed on all keyboards
  - touch-typists type fastest using the qwerty layout

Error Types **(and the Boy Who Cried Wolf)**

The “null hypothesis" is essentially a default position that there is ***not*** a relationship between two variables.

`Type I error`: You reject the null hypothesis when it is, in fact, true. (We say there is something there when there actually is not.)

`Type II error`: You accept the null hypothesis when it is, in fact, false. (We say there is not something there when there is.)

Error Causes and Consequences

Effects of levels of significance in brief are that using very low confidence levels (eg: 0.1) gives greater chance of Type I errors and going for very high confidence level (eg: 0.0001) gives a greater chance of Type II errors.

Consequences in HCI could be…

- Type I: extra work developing software and having people learn a new idiom for no benefit
- Type II: people keep using a less efficient (but already familiar) interface

Some resources…

Industry Testing suggested participant size calculator versus a more science-oriented one: http://blinkux.com/usability-sample-size/ and http://www.calculator.net/sample-size-calculator.html

Discussion of A/B Testing: https://www.nytimes.com/2015/09/27/upshot/a-better-government-one-tweak-at-a-time.html

Practice quiz about variables: https://www.proprofs.com/quiz-school/quizshow.php?title=independent-vs-dependent-variables&q=1

Iterative Design Case Studies: Chapter 6 of the optional textbook…

Summary

- Controlled experiments can provide you with clear convincing results when looking at specific issues.
- Creating testable hypotheses are critical to good experimental design.
- Experimental design requires a great deal of planning and elements to consider.
- 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…