Quantitative Evaluation

What is experimental design?

What is an experimental hypothesis?

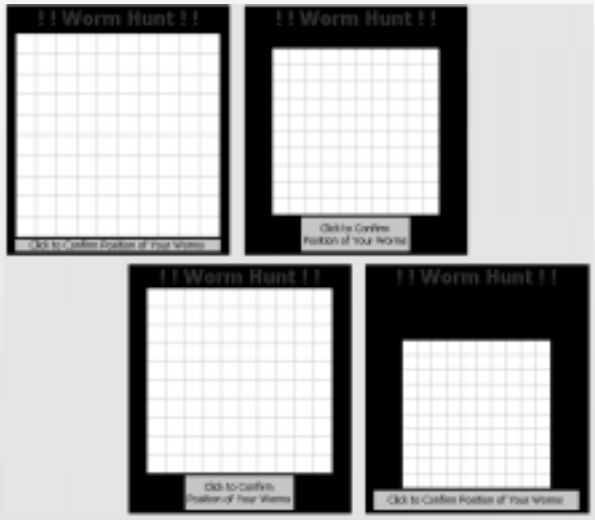
How do I plan an experiment?

Why are statistics used?

What are the important statistical methods?

Ben Bederson / Saul Greenberg

Question: Which size grid is better?



Evan Golub

Question: Which menu placement system is better?

Top of Window

Top of Screen

Evan Golub

Quantitative methods

1. User performance data collection

- data is 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?
- collects heaps of data in the hope that something interesting shows up
- often difficult to sift through data unless specific aspects are targeted
  - as in list above

Ben Bederson / Saul Greenberg

Quantitative methods ...

2. Controlled experiments

The traditional scientific method

- reductionist
  - clear convincing result on specific issues
- In HCI:
  - insights into cognitive process, human performance limitations, ...
  - allows comparison of systems, fine-tuning of details ...

Strives for

- 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

Ben Bederson / Saul Greenberg

The experimental method

a) Begin with a lucid, testable hypothesis

- Example 1:

" there is no difference in the number of cavities in children and teenagers using crest and no-teeth toothpaste"

Ben Bederson / Saul Greenberg

The experimental method

a) Begin with a lucid, testable hypothesis

- Example 2:

" 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"

Ben Bederson / Saul Greenberg

The experimental method...

b) Explicitly state the independent variables that are to be altered

independent variable

- the things you manipulate 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 toothpaste experiment

- toothpaste type: uses Crest or No-teeth toothpaste
- age: <= 11 years *or* > 11 years

in menu experiment

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

Ben Bederson / Saul Greenberg

The experimental method...

c) Carefully choose the dependent variables that will be measured

Dependent variables

- variables dependent on the subject's behaviour / reaction to the independent variable

in toothpaste experiment

- number of cavities
- frequency of brushing

in menu experiment

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

Ben Bederson / Saul Greenberg

The experimental method...

d) Judiciously select and assign subjects to groups

Ways of controlling subject variability

- recognize classes and make them an independent variable
- minimize unaccounted anomalies in subject group
  - superstars versus poor performers
- use reasonable number of subjects and random assignment

Novice

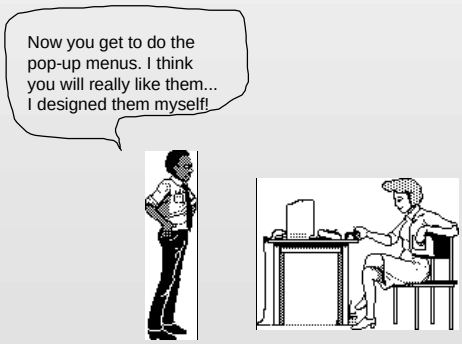
Expert

Ben Bederson / Saul Greenberg

The experimental method...

e) Control for biasing factors

- unbiased instructions + experimental protocols
  - prepare ahead of time
- double-blind experiments, ...



Ben Bederson / Saul Greenberg

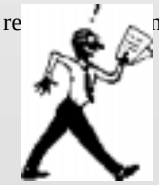
The experimental method...

f) Apply statistical methods to data analysis

- Confidence limits: the confidence that your conclusion is correct
  - "The hypothesis that mouse experience makes no difference is rejected at the .05 level"
  - "Expert mouse users can use pull-down menus 15% faster than novice mouse users, and that result is statistically significant"
  - means:
    - a 95% chance that your statement is correct
    - a 5% chance you are wrong

g) Interpret your results

- what you believe the re nd their implications



Ben Bederson / Saul Greenberg

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"

Ben Bederson / Saul Greenberg

Statistical significance vs Practical significance

when *n* is large, even a trivial difference may be large enough to produce a statistically significant result

- eg menu choice:
  mean selection time of menu a is 3 seconds;
  menu b is 3.05 seconds

Statistical significance does not imply that the difference is important!

- a matter of interpretation

Ben Bederson / Saul Greenberg

Example: Differences between means

Given: two data sets measuring a condition

- eg height difference of males and females
  time to select an item from different menu styles ...

Question:

- is the difference between the means of the data statistically significant?

Null hypothesis:

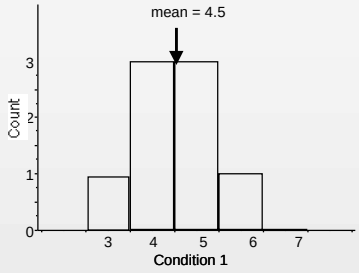
- there is no difference between the two means
- statistical analysis can only reject the hypothesis at a certain level of confidence

Ben Bederson / Saul Greenberg

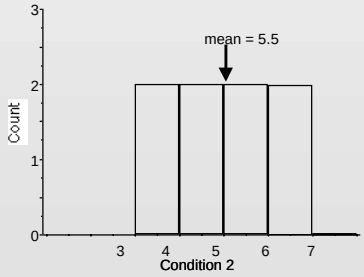
Example:

Is there a significant difference between the means?

Condition one: 3, 4, 4, 4, 5, 5, 5, 6



Condition two: 4, 4, 5, 5, 6, 6, 7, 7



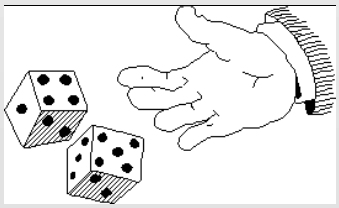
Ben Bederson / Saul Greenberg

The problem with visual inspection of data

There is almost always variation in the collected data

Differences between data sets may be due to:

- normal variation
  - eg two sets of ten tosses with different but fair dice
    differences between data and means are accountable by expected variation
- real differences between data
  - eg two sets of ten tosses for with loaded dice and fair dice
    differences between data and means are not accountable by expected variation



Ben Bederson / Saul Greenberg

Choice of significance levels and two types of 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

- very high confidence level (eg .0001) gives greater chance of Type 2 errors
- very low confidence level (eg .1) gives greater chance of Type 1 errors
- choice often depends on effects of result

Ben Bederson / Saul Greenberg

Choice of significance levels and two types of errors

There is no difference between Pie menus and traditional pop-up menus

- Type 1: extra work developing software and having people learn a new idiom for no benefit
- Type 2: use a less efficient (but already familiar) menu

- Case 1: Redesigning a traditional GUI interface
  - a Type 2 error is preferable to a Type 1 error
- Case 2: Designing a digital mapping application where experts perform extremely frequent menu selections
  - a Type 1 error is preferable to a Type 2 error

Ben Bederson / Saul Greenberg

Other Tests: Correlation

Measures the extent to which two concepts are related

- eg years of university training vs computer ownership per capita

How?

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

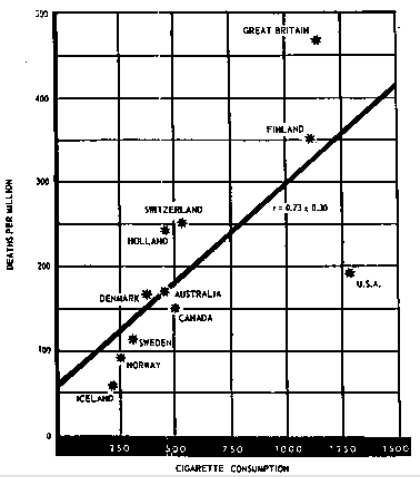
Dangers

- attributing causality
  - a correlation does not imply cause and effect
  - cause may be due to a third "hidden" variable related to both other variables
  - eg (above example) age, affluence
- drawing strong conclusion from small numbers
  - unreliable with small groups
  - be wary of accepting anything more than the direction of correlation unless you have at least 40 subjects

Ben Bederson / Saul Greenberg

Sample Study: Cigarette Consumption

Crude Male death rate for lung cancer in 1950 per capita consumption of cigarettes in 1930 in various countries.

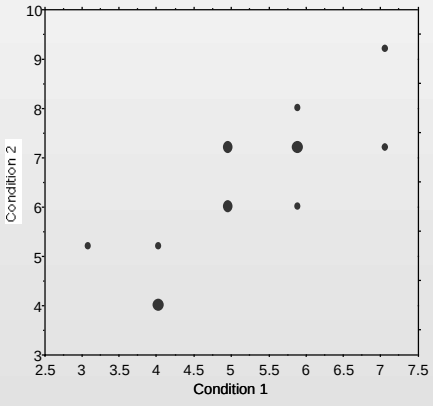


Ben Bederson / Saul Greenberg

Correlation

$r^2 = .668$

condition 1	condition 2
5	6
4	5
6	7
4	4
5	6
3	5
5	7
4	4
5	7
6	7
6	6
7	7
6	8
7	9



Ben Bederson / Saul Greenberg

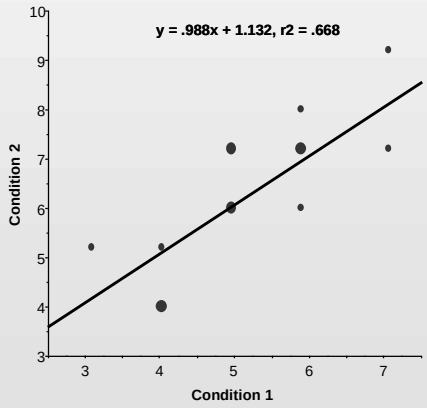
Other Tests: Regression

Calculate a line of "best fit"

use the value of one variable to predict the value of the other

- e.g., 60% of people with 3 years of university own a computer

condition 1	condition 2
5	6
4	5
6	7
4	4
5	6
3	5
5	7
4	4
5	7
6	7
6	6
7	7
6	8
7	9



Ben Bederson / Saul Greenberg

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

There are many statistical methods that can be applied to different experimental designs - one example is the use of correlation and regression.

Ben Bederson / Saul Greenberg