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?

Saul Greenberg

Quantitative ways to evaluate systems

Quantitative:

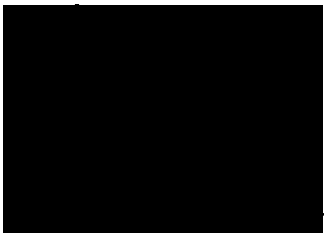
- precise measurement, numerical values
- bounds on how correct our statements are

Methods

- User performance
- Controlled Experiments
- Statistical Analysis

Saul Greenberg

Quantitative methods

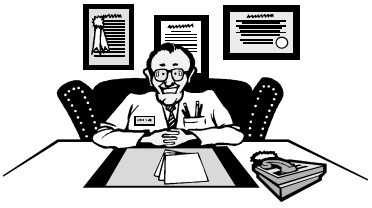


1. User performance data collection

- 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

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)
- replicability of experiment
- control of variables and conditions
- removal of experimenter bias

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"

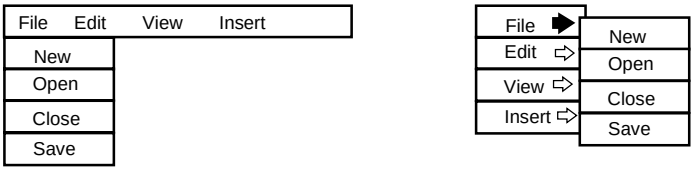
Saul Greenberg

The experimental method

a) Begin with a lucid, testable hypothesis

- Example 2:

" 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, regardless of the subject's previous expertise in using a mouse or using the different menu types"



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)

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 menu experiment

- time to select an item
- selection errors made

in toothpaste experiment

- number of cavities
- frequency of brushing

Saul Greenberg

The experimental method...

d) Judiciously select and assign subjects to groups

Ways of controlling subject variability

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

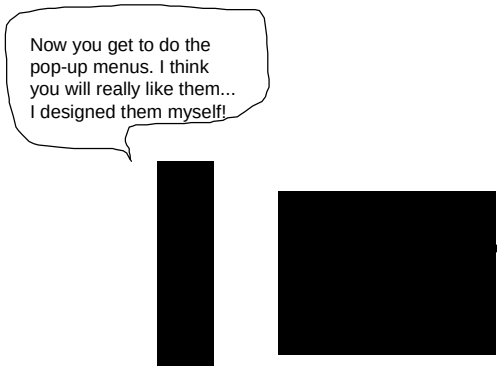
Novice Expert

Saul Greenberg

The experimental method...

e) Control for biasing factors

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



Saul Greenberg

The experimental method...

f) Apply statistical methods to data analysis

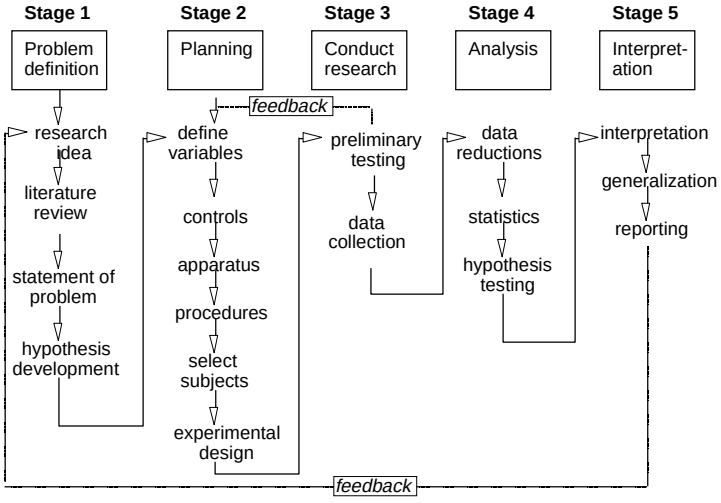
- Confidence limits: the confidence that your conclusion is correct
  - "The hypothesis that mouse experience makes no difference is rejected at the .05 level"
  - means:
    - a 95% chance that your statement is correct
    - a 5% chance you are wrong

g) Interpret your results

- what you believe the results mean, and their implications

Saul Greenberg

The Planning Flowchart



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"

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

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:

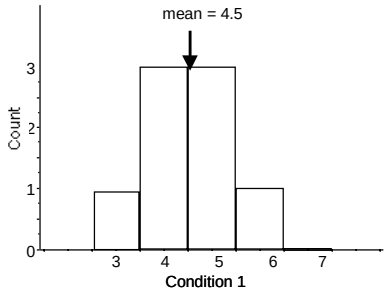
- there is no difference between the two means
- statistical analysis can only reject the hypothesis at a certain level of confidence

Saul Greenberg

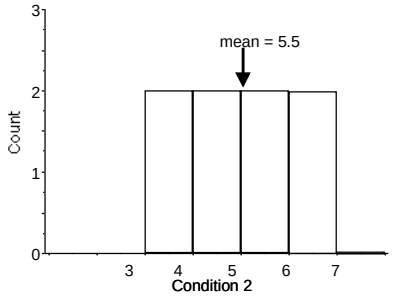
Example:

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



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

Saul Greenberg

T-test

A 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

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

Saul Greenberg

Different types of T-tests

Comparing two sets of independent observations

- usually different subjects in each group (number may differ as well)

Condition 1	Condition 2
S1–S20	S21–43

Paired observations

- usually single group studied under separate experimental conditions
- data points of one subject are treated as a pair

Condition 1	Condition 2
S1–S20	S1–S20

Non-directional vs directional alternatives

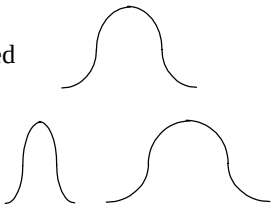
- non-directional (two-tailed)
  - no expectation that the direction of difference matters
- directional (one-tailed)
  - Only interested if the mean of a given condition is greater than the other

Saul Greenberg

T-test...

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

Saul Greenberg

Two-tailed unpaired T-test

- N: number of data points in the one sample
- ΣX: sum of all data points in one sample
- $\overline{X}$: mean of data points in sample
- $\Sigma(X^2)$: sum of squares of data points in sample
- s^2: unbiased estimate of population variation
- t: t ratio
- df = degrees of freedom = N1 + N2 – 2

Formulas

$$s^2 = \frac{\sum(X_1^2) - \frac{(\sum_{X_1})^2}{N_1} + \sum(X_2^2) - \frac{(\sum_{X_2})^2}{N_2}}{N_1 + N_2 - 2} \qquad t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s^2}{N_1} + \frac{s^2}{N_2}}}$$

Saul Greenberg

Level of significance for two-tailed test

df	***.05***	***.01***	***df***	***.05***	***.01***
1	12.706	63.657	16	2.120	2.921
2	4.303	9.925	18	2.101	2.878
3	3.182	5.841	20	2.086	2.845
4	2.776	4.604	22	2.074	2.819
5	2.571	4.032	24	2.064	2.797
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			

Saul Greenberg

Example Calculation

x_1 = 3 4 4 4 5 5 5 6
x_2 = 4 4 5 5 6 6 7 7

Hypothesis: there is no significant difference between the means at the .05 level

Step 1. Calculating s^2

	1	2
N	8	8
Σx	36	44
$\bar{x}$	4.5	5.5
$\Sigma(x^2)$	168	252
$(\Sigma x)^2$	1296	1936

$df = 14$

$$s^2 = \frac{\Sigma x_1^2 - (\Sigma x_1)^2/N_1 + \Sigma x_2^2 - (\Sigma x_2)^2/N}{N_1 + N_2 - 2}$$

$$= \frac{168 - 1296/8 + 252 - 1936/8}{8+8-2}$$

$$= 1.1429$$

Saul Greenberg

Example Calculation

Step 2. Calculating *t*

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2/N_1 + s^2/N_2}}$$

$$= \frac{4.5 - 5.5}{\sqrt{2 \cdot (1.1429/8)}}$$

$$= \frac{-1}{.5345}$$

$$= -1.871$$

Step 3: Looking up critical value of *t*

- Use table for two-tailed *t*-test, at p=.05, df=14
- critical value = 2.145
- because t=1.871 < 2.145, there is no significant difference
- therefore, we cannot reject the null hypothesis
  i.e., there is no difference between the means

Saul Greenberg

Two-tailed Unpaired T-test

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

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

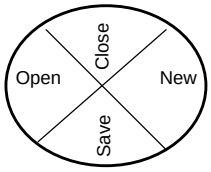
- very high confidence level (eg .0001) gives greater chance of Type 2 errors
- very low confidence level (eg .1) gives greater chance of Type 1 errors
- choice often depends on effects of result

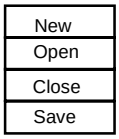
Saul Greenberg

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

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

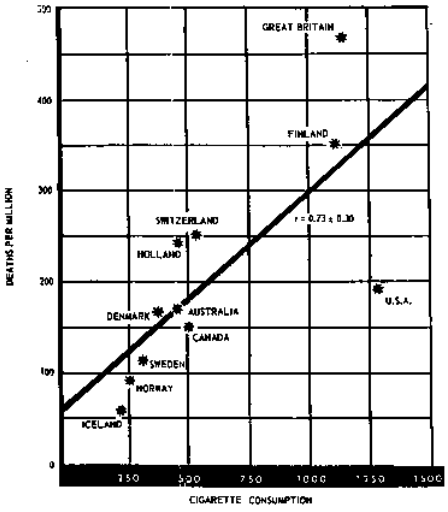
Dangers

- 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

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.

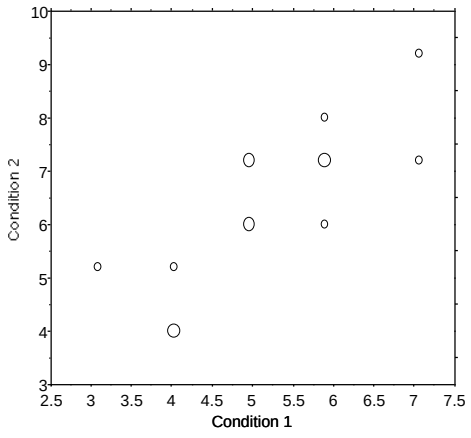


Saul Greenberg

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



Saul Greenberg

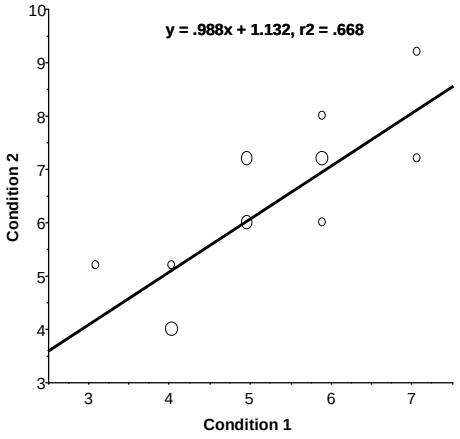
Other Tests: Regression

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



Saul Greenberg

Other Tests: Single Factor Analysis of Variance

Compare three or more means

Comparing three keyboards: example results:

- mouse-typing speed is
  - fastest on a qwerty keyboard
  - the same on an alphabetic & dvorak keyboards

Qwerty	Alphabetic	Dvorak
S1-S10	S11-S20	S21-S30

Saul Greenberg

Other Tests: Analysis of Variance (Anova)

Compares the relationships between many factors

Provides more informed results
considers the interactions between factors

Examples
- beginners type at the same speed on all keyboards,
- touch-typist type fastest on the qwerty

	Qwerty	Alphabetic	Dvorak
cannot touch type	S1-S10	S11-S20	S21-S30
can touch type	S31-S40	S41-S50	S51-S60

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
- T-tests
- Correlation and regression
- Single factor Anova
- Anova

Saul Greenberg