CS 554m:

controlled experiments

today: part I

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?
How to choose the right statistic?

Acknowledgement: Some of the material in this lecture is based on material prepared for similar courses by Saul Greenberg (University of Calgary)

2

a good portion of the material in these lectures on experimental design should be familiar from ugrad stats class, although perhaps presented here from a slightly different perspective

also, most of this material is well covered in today's reading:

Newman & Lamming, Ch 10

3

material I assume you already know and will not be covered (some additional slides at end)

types of variables
samples & populations
normal distribution
variance and standard deviation

4

quantitative methods

1. user performance data collection
   - data is collected on system use

descriptive statistics:
     - 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?
   - collect 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)

5

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 (usually a causal inference)
- quantitative measurement
- measure of confidence in results obtained (inferencial statistics)
- replicability of experiment
- control of variables and conditions
- removal of experimenter bias

6

desired outcome of a controlled experiment

statistical inference of an event or situation's probability:

"Design A is better <*in some specific sense*> than Design B"

or, Design A meets a target:

"90% of incoming students who have web experience can complete course registration within 30 minutes"

7

steps in the experimental method

step 1: begin with a lucid, testable hypothesis

Example 1:

H_0: there is no difference in the number of cavities in children and teenagers using crest and no-teeth toothpaste

H_1: children and teenagers using crest toothpaste have fewer cavities than those who use no-teeth toothpaste

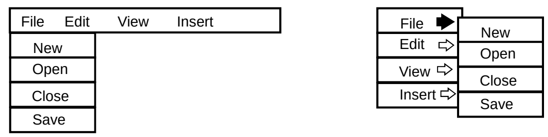
9

step 1: begin with a lucid, testable hypothesis

Example 2:

H_0: 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

H_1: selecting from a pop-up menu will be faster and less error prone than selecting from a pull down menu



10

general: hypothesis testing

hypothesis = **prediction** of the outcome of an experiment.

framed in terms of **independent** and **dependent** variables:

a variation in the independent variable will cause a difference in the dependent variable.

aim of the experiment: prove this prediction

do by: *disproving* the "null hypothesis"

H_0: experimental conditions **have no effect** on performance (to some degree of **significance)** → **null hypothesis**

H_1: experimental conditions **have an effect** on performance (to some degree of **significance)** → **alternate hypothesis**

11

step 2: explicitly state the independent variables

Independent variables

- things you control/manipulate (independent of how a subject behaves) to produce different conditions for comparison
- two different kinds:
  - treatment manipulated (can establish cause/effect, true experiment)
  - subject individual differences (can never fully establish cause/effect)

in toothpaste experiment

- toothpaste type: Crest or No-teeth toothpaste *(treatment)*
- age: <= 12 years *or* > 12 years *(subject)*

in menu experiment

- menu type: pop-up or pull-down *(treatment)*
- menu length: 3, 6, 9, 12, 15 *(treatment)*
- expertise: expert or novice *(often subject, but can train an expert)*

12

step 3: carefully choose the dependent variables

Dependent variables
- things that are measured
- expectation that they depend on the subject's behaviour / reaction to the independent variable (but unaffected by other factors)

in toothpaste experiment
- number of cavities
- frequency of brushing

in menu experiment
- time to select an item
- selection errors made

13

step 4: consider possible nuisance variables & determine mitigation approach

- undesired variations in experiment conditions which **cannot be eliminated**, but which **may affect** dependent variable
  - critical to know about them
- experiment design & analysis must generally accommodate them:
  - treat as an additional experiment independent variable (if they can be controlled)
  - randomization (if they cannot be controlled)
- common nuisance variable: *subject* (individual differences)

in toothpaste experiment
- brushing time of day: when does a subject brush their teeth
- type of food eaten during day: healthy or sugar laden

in menu experiment
- time of day subject is run: poorest performance may be after lunch
- motor ability: any motor impairments would dominate menu conditions

14

step 5: design the task to be performed

tasks must:

be externally valid

external validity = do the results generalize?

… will they be an accurate predictor of how well users can perform tasks as they would in real life?

for a large interactive system, can probably only test a small subset of all possible tasks.

exercise the designs, bringing out any differences in their support for the task

e.g., if a design supports website **navigation**, test task should **not** require subject to work within a **single page**

be feasible - supported by the design/prototype, and executable within experiment time scale

15

step 5: design the task to be performed

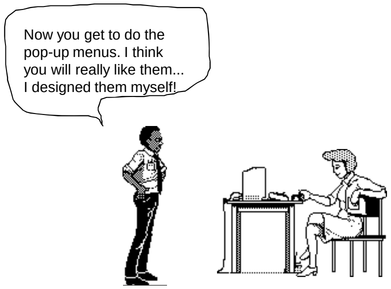
in toothpaste experiment
- use new brand of toothpaste for X number of days/weeks/months
- brush at least once a day

in menu experiment
- for each menu length, prompt user with a stream of X menu items, one at a time, and have her/him select the matching menu item. Force user to select the correct one before advancing to the next item (i.e., any errors must be corrected).

16

step 6: design experiment protocol

- steps for executing experiment are prepared well ahead of time
- includes unbiased instructions + instruments (questionnaire, interview script, observation sheet)
- double-blind experiments, ...



17

step 7: make formal experiment design explicit

simplest: 2-sample (2-condition) experiment

based on comparison of **two sample means**:
- performance data in response to Designs A, B
  - compare performance of new design with old
  - compare performance of 2 new designs

or, comparison of **one sample mean with a constant**:
- performance data in response to Design A, compared to performance requirement
  - determine whether single new design meets key design requirement

18

step 7: make formal experiment design explicit

more complex: factorial design

in toothpaste experiment

2 toothpaste types (crest, no-teeth)
x 2 age groups (<= 12 years *or* > 12 years)

in menu experiment :

2 menu types (pop-up, pull down)
x 5 menu lengths (3, 6, 9, 12, 15)
x 2 levels of expertise (novice, expert)

(more on this later)

19

step 8: judiciously select/recruit and assign subjects to groups

subject pool: similar issues as for informal studies
- match expected user population as closely as possible
- age, physical attributes, level of education
- general experience with systems similar to those being tested
- experience and knowledge of task domain

sample size: perhaps more critical here
- going for "statistical significance"
- should be large enough to be "representative" of population
- guidelines exist based on statistical methods used & required significance of results
- pragmatic concerns may dictate actual numbers
- "10" is often a good place to start

20

step 8: judiciously select/recruit 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

21

step 9: apply statistical methods to data analysis

examples: t-tests, ANOVA, correlation, regression (more on these later)

confidence limits: the confidence that your conclusion is correct

- "The hypothesis that mouse experience makes no difference is rejected at the .05 level" (i.e., null hypothesis rejected)
- this means:
  - a 95% chance that your finding is correct
  - a 5% chance you are wrong

22

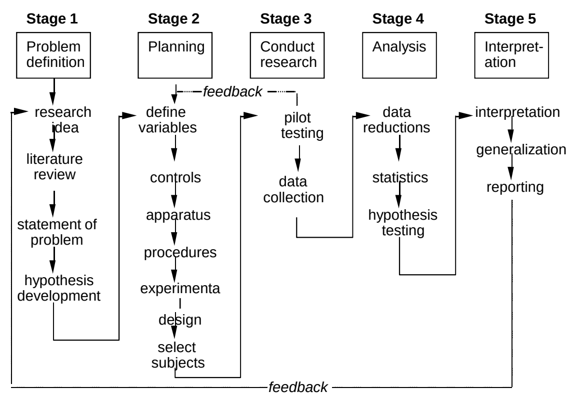
step 10: interpret your results

what *you* believe the results mean, and their implications

yes, there can be a subjective component to quantitative analysis

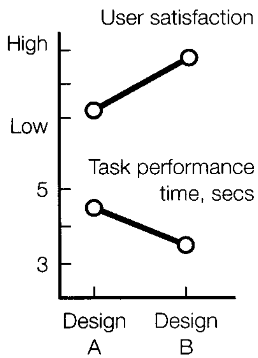
23

the planning flowchart



24

goal of experiment design



guard against ambiguous or misleading results

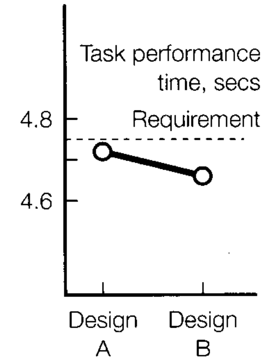
← a good (definitive) result

25

poor experiment design or results

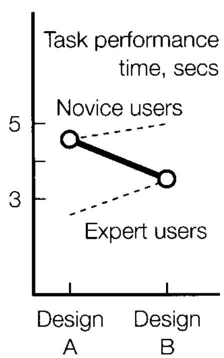
less distinguishable results:

perhaps task was poorly chosen – or there's really no difference



26

poor experiment design



misleading results

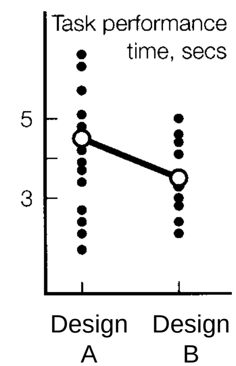
e.g. subject assignment not controlled: one design tested on novices, other on experts, disguising actual trend

27

poor experiment design or results

large spread in values

perhaps conditions were not well controlled

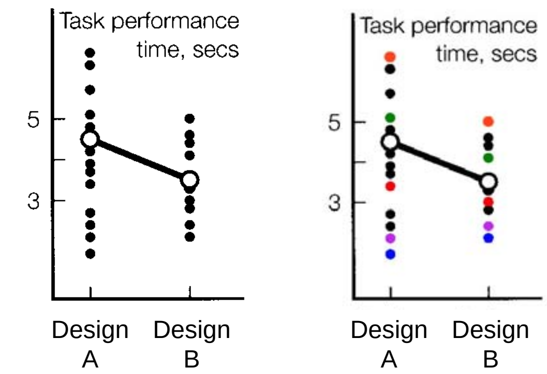


28

as we have seen

individual (subject) differences may pose a **nuisance variable:**

variation in individual abilities can mask real differences in test conditions, if not analyzed properly



most common way to deal with:

subtract each individual's mean performance at two factor levels from overall score, before combining with other individuals (paired t-test)

within/between subject comparisons

within-subject comparisons:

- **subjects exposed to multiple treatment conditions**
- → primary comparison internal to each subject
- allows control over subject variable
- greater statistical power, fewer subjects required
- not always possible (exposure to one condition might "contaminate" subject for another condition; or session too long)

between-subject comparisons:

- **subjects only exposed to one condition**
- → primary comparison is from subject to subject
- less statistical power, more subjects required
- why? because greater variability due to more individual differences

within/between subject comparisons

in toothpaste experiment

2 toothpaste types (crest, no-teeth) *between or within*
x 2 age groups (<= 12 years *or* > 12 years) *must be between*

in menu experiment :

2 menu types *(pop-up, pull down) between or within*
x 5 menu lengths (3, 6, 9, 12, 15*) should be within*
x 2 levels of expertise (novice, expert) *must be between*

to summarize so far: how a controlled experiment works

1. formulate an **alternate** and a **null** hypothesis:
   H_1: experimental conditions **have an effect** on performance
   H_0: experimental conditions **have no effect** on performance
2. through **experiment task**, try to demonstrate that the **null hypothesis is false** (reject it),
   for a particular level of **significance**
3. if successful, we can **accept** the alternate hypothesis,
   and state the probability ***p*** that we are wrong (the null hypothesis is true after all) → this is the result's **confidence level**

   e.g., selection speed is significantly faster in menus of length 5 than of length 10 ($p<.05$)

   → 5% chance we've made a mistake, 95% confident

33

statistical analysis

what is a statistic?
- a number that describes a sample
- sample is a subset (hopefully representative) of the population we are interested in understanding

statistics are calculations that tell us
- mathematical attributes about our data sets (sample)
  - mean, amount of variance, ...
- how data sets relate to each other
  - whether we are "sampling" from the same or different populations
- the probability that our claims are correct
  - "statistical significance"

34

example: differences between means

given: two data sets measuring a condition
- e.g., 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
- we never actually prove the hypothesis true

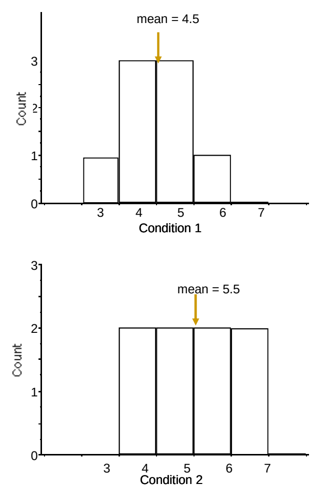
35

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



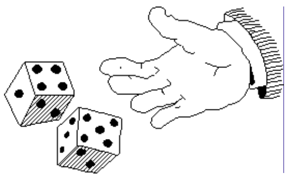
36

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
  - e.g., two sets of ten tosses with different but fair dice
    - differences between data and means are accountable by expected variation
- real differences between data
  - e.g., two sets of ten tosses with loaded dice and fair dice
    - differences between data and means are not accountable by expected variation



37

t-test

a statistical test

allows one to say something about differences between two 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

38

different types of t-tests

comparing two sets of independent observations ***(between subjects)***

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

Condition 1	Condition 2
S1–S20	S21–S43

paired observations ***(within subjects)***

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

39

different types of t-tests

non-directional vs directional alternatives

non-directional (two-tailed)

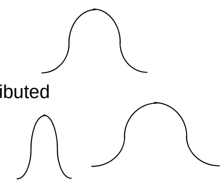
- no expectation that the direction of difference matters

directional (one-tailed)

- only interested if the mean of a given condition is greater than the other

40

t-tests



Assumptions of t-tests

- data points of each sample are normally distributed
  - but t-test very robust in practice
- sample 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 *(can you think of examples where they are not?)*

Significance level

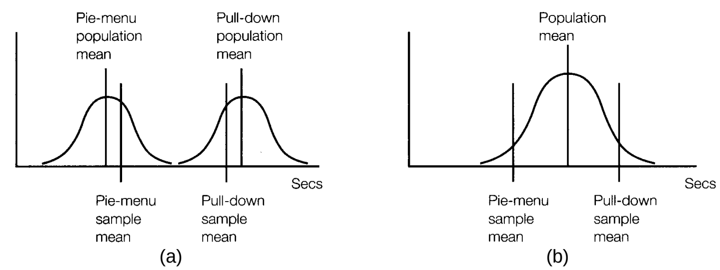
- decide upon the level before you do the test!
- typically stated at the .05 or .01 level

41

what the t-test is testing

(a) the two samples come from two different populations;

(b) the two samples are part of the same population.



Which represents H_0 and which represents H_1?

42

two-tailed unpaired t-test

n: number of data points in the one sample ($N = n_1 + n_2$)

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

N&L shows derivation of formula

How to maximize t?

Formulas

$$s^2 = \frac{\Sigma(X_1^2) - \frac{(\Sigma X_1)^2}{n_1} + \Sigma(X_2^2) - \frac{(\Sigma X_2)^2}{n_2}}{n_1 + n_2 - 2}$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

43

<N&L derivation>

mean & sum of squares

mean $= \overline{X} = \frac{\sum X_i}{N}$

sum of squares $= SS = \sum (X_i - \overline{X})^2$

(same, faster) $= \sum X_i^2 - \frac{(\sum X_i)^2}{N}$

error in N&L pg. 231

44

degrees of freedom (df)

freedom of a set of values to vary independently of one another:

$X = \{21, 20, 24\}$ N=3

$\bar{X} = \frac{65}{3} = 21.6$: $\leftarrow \bar{X}$ has N-1=2 df

once you know the mean of N values, only N-1 can vary independently

45

sample variance & standard deviation

sample variance $= s^2 = \frac{SS}{N-1}$

standard deviation $= sd = \sqrt{s^2}$

46

</N&L derivation>

calculating *t*

compute **combined variance** for the two samples:

$$s^2 = \frac{SS_1 + SS_2}{N_1 + N_2 - 2}$$

← *note df computation*

compute **standard error of difference**, s_{ed} :

$$s_{ed} = \sqrt{s^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

no, you won't have to memorize the formula for exams. but you *should* know how / when to use it.

compute *t*:

$$t = \frac{|\bar{X}_1 - \bar{X}_2|}{s_{ed}}$$

47

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

df	.05	.01
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

Critical value (threshold) that t statistic much reach to achieve significance.

How does critical value change based on *df* and confidence level?

48

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

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

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

$$= 1.1429$$

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

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

What the results would look like in stats software.

Unpaired t-test

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

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

How does the outcome change for a confidence level of 0.10?

summary of the t-test

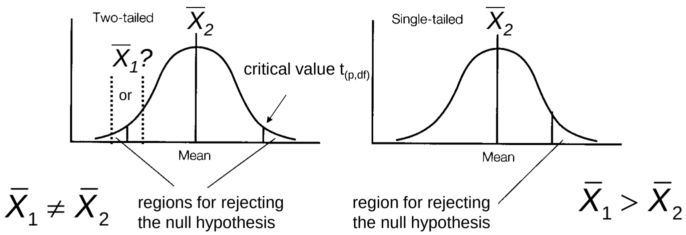
the point: establish a confidence level in the difference we've found between 2 sample means.

the process:

1. compute df
2. choose desired **significance, *p*** (aka α)
3. calculate value of the ***t* statistic**
4. compare it to the **critical value** of *t* given *p*, df: $t_{(p,df)}$
5. if $t > t_{(p,df)}$, can **reject null hypothesis at *p***

significance *p*

measure of the area of the normal distribution occupied by the null hypothesis = the chance you might be wrong



null hypothesis rejection area:

- two-tailed: divided equally between left/right
- single-tailed: all on one side

53

today: part II

learning goals:

what is an analysis of variance (ANOVA)?

what is the important terminology in ANOVA?

what are the different types of ANOVA?

when would one choose to use an ANOVA?

what is the difference between statistical and practical significance?

other tests: what are correlation & regression?

54

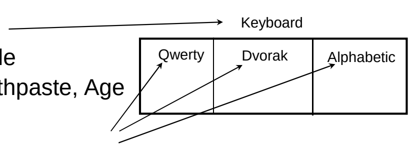
analysis of variance (ANOVA)

A Workhorse

- allows moderately complex experimental designs (relative to t-test)

Terminology

- Factor
  - independent variable
  - i.e., Keyboard, Toothpaste, Age
- Factor level
  - specific value of independent variable
  - i.e., Qwerty, Crest, 5-10 years old



55

ANOVA terminology

Between subjects

- a subject is assigned to only one factor level of treatment
- problem: greater variability, requires more subjects

Keyboard

Qwerty	Dvorak	Alphabetic
S1-20	*S21-40*	*S41-60*

Within subjects

- subjects assigned to all factor levels of a treatment
- requires fewer subjects
- less variability as subject measures are paired
- problem: order effects (e.g., learning)
- partially solved by counter-balanced ordering

Keyboard

Qwerty	Dvorak	Alphabetic
S1-20	*S1-20*	*S1-20*

56

F statistic

Within group variability (WG)
- individual differences
- measurement error

Keyboard

Qwerty	Dvorak	Alphabetic
5, 9, 7, 6, … 3, 7	3, 9, 11, 2, … 3, 10	3, 5, 5, 4, … 2, 5

Between group variability (BG)
- treatment effects
- individual differences
- measurement error

Keyboard

Qwerty	Dvorak	Alphabetic
5, 9, 7, 6, … 3, 7	3, 9, 11, 2, … 3, 10	3, 5, 5, 4, … 2, 5

These two variabilities combine to give total variability

We are mostly interested in between group variability because we are trying to understand the effect of the treatment

57

F statistic

$$F = \frac{BG}{WG} = \frac{\text{treatment + id + m.error}}{\text{id + m.error}} = ?$$

= 1, if there are no treatment effects

> 1, if there are treatment effects

Within-subjects design: the id component in numerator and denominator factored out, therefore a more powerful design

58

F statistic

Similar to the t-test, we look up the F value in a table, for a given α and degrees of freedom to determine significance

Thus, F statistic sensitive to sample size
- Big N → Big Power → Easier to find significance
- Small N → Small Power → Difficult to find significance

What we (should) want to know is the effect size
- Does the treatment make a big difference (i.e., large effect)?
- Or does it only make a small difference (i.e., small effect)?
- Depending on what we are doing, small effects may be important findings

59

statistical significance vs *practical* significance

when *N* is large, even a trivial difference (small effect) may be large enough to produce a statistically significant result
- e.g., 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, i.e., subjective opinion
- should always report means to help others make their opinion

There are measures for effect size, regrettably they are not widely used in HCI research

60

single factor analysis of variance

Compare means between two or more factor levels within a single factor

e.g.:

- dependent variable: typing speed
- independent variable (factor): keyboard
- between subject design

also called a one-way ANOVA

Qwerty	Alphabetic	Dvorak
S1: 25 secs S2: 29 … S20: 33	S21: 40 secs S22: 55 … S40: 33	S51: 17 secs S52: 45 … S60: 23

61

ANOVA terminology

- Factorial design
  - cross combination of levels of one factor with levels of another
  - e.g., keyboard type (3) x expertise (2)

2-way factorial ANOVA

- Cell
  - unique treatment combination
  - e.g., qwerty x non-typist

expertise		Keyboard		
		Qwerty	Dvorak	Alphabetic
	non-typist			
	typist			

62

ANOVA terminology

Mixed factor

- contains both between and within subject combinations

expertise		Keyboard		
		Qwerty	Dvorak	Alphabetic
	non-typist	*S1-20*	*S1-20*	*S1-20*
	typist	*S21-40*	*S21-40*	*S21-40*

63

ANOVA

Compares the relationships between many factors

Provides more informed results

- considers the interactions between factors
- e.g.,
  - typists type faster on Dvorak, than on alphabetic and Qwerty
  - non-typists are fastest on alphabetic

expertise		Keyboard		
		Qwerty	Dvorak	Alphabetic
	non-typist	*S1-20*	*S1-20*	*S1-20*
	typist	*S21-40*	*S21-40*	*S21-40*

64

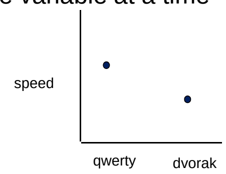
ANOVA

In reality, we can rarely look at one variable at a time

Example:

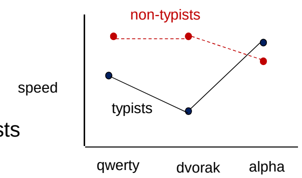
- t-test:

  subjects faster on dvorak than qwerty



- anova: keyboard x expertise

  alphabetic fastest for non-typists
  dvorak fastest for typists



65

ANOVA case study

WIMP (GUI) vs. HYBRID (graphical command line)

Motivation:

- WIMP interfaces are slow because of the mouse
- Can we create a hybrid interface that is graphical but can be fully operated through the keyboard? (sort of like a command line)
- Assume that one has been designed
- How should it be evaluated?

66

ANOVA case study

WIMP (GUI) vs. HYBRID (graphical command line)

Independent variables:

- Interface: WIMP, hybrid
- Expertise: novice, expert
- Command parameters: zero, one, two
  - E.g., bold (zero), font ariel (one), print –copies 2 –color greyscale (two)
  - Note: zero parameter commands can be done using shortcuts keys

Dependent variables:

- Performance: speed, error
- Satisfaction

67

ANOVA case study

Possible hypotheses:

H1: experts will perform better than novices (not that interesting)

H2: novices will perform better with WIMP than hybrid

H3: experts will perform better with hybrid than WIMP, but only for commands with one or more parameters

2 level (interface) x
2 level (expertise) x
3 level (parameters)

mixed factor design

		WIMP	hybrid
zero	novice	S1-8	S1-8
	expert	S9-16	S9-16
one	novice	S1-8	S1-8
	expert	S9-16	S9-16
two	novice	S1-8	S1-8
	expert	S9-16	S9-16

68

task

assume that the task is to enter a whole series of commands, one after the other

there is an equal number of 0, 1, and 2 parameter commands used

the identical commands are used in both interface conditions

69

statistical results: speed

	F-ratio.	*p*
Interface (I)	0.4	
Expertise (E)	5.5*	<0.05
Parameters (P)	31.0**	<0.01
IxE	15.2*	<0.05
IxP	8.0*	<0.05
ExP	5.0	
IxExP	14.1*	<0.05

main effects (Interface, Expertise, Parameters)

interactions (IxE, IxP, ExP, IxExP)

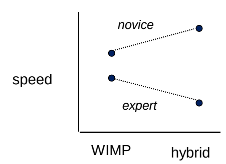
main effect: the effect of the variable averaging over all level of other variables in the experiment

interaction effect: the effect of one variable differs depending on the level of another (other) variable(s)

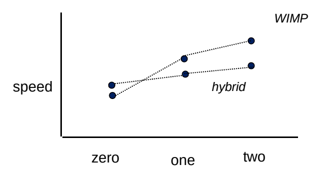
70

statistical results: speed

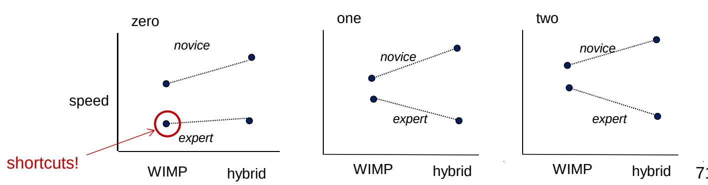
Interface x Expertise (IxE)



Interface x Parameters (IxP)



Interface x Expertise x Parameters (IxExP)



71

summary of results

Assuming same results for errors as speed…

H1: experts will perform better than novices (not that interesting)

Supported: main effect of expertise, showing experts better

H2: novices will perform better with WIMP than hybrid

Supported: 2-way interaction effect of interface and expertise, showing novices overall better with WIMP

H3: experts will perform better with hybrid than WIMP, but only for commands with one or more parameters

Supported: 3-way interaction effect of interface, expertise, and number of parameters, showing experts better with hybrid, but only with one and two parameters

72

case study conclusions

- expertise makes a big difference
- WIMP interaction should be kept for novices
- hybrid interaction should be available for experts

73

choice of significance levels and two types of errors

Type I error: reject the null hypothesis when it is, in fact, true (α = .05)
Type II error: accept the null hypothesis when it is, in fact, false (β)

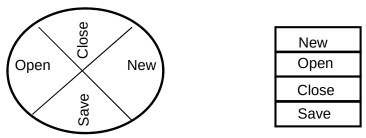
	H_0 True	H_0 False
Reject H_0	α (Type I error)	1 - β (Power)
Not Reject H_0	1 - α	β (Type II error)

Effects of levels of significance
- very high confidence level (eg .0001) gives greater chance of Type II errors
- very low confidence level (eg .1) gives greater chance of Type I errors
- tradeoff: choice often depends on effects of result

74

choice of significance levels and two types of errors

H_0 There is no difference between Pie menus and traditional pop-up menus



Type I: (reject H_0, believe there is a difference, when there isn't)
- extra work developing software and having people learn a new idiom for no benefit

Type II: (accept H_0, believe there is no difference, when there is)
- use a less efficient (but already familiar) menu

75

choice of significance levels and two types of errors

Type I: (reject H_0, believe there is a difference, when there isn't)
- extra work developing software and having people learn a new idiom for no benefit

Type II: (accept H_0, believe there is no difference, when there is)
- use a less efficient (but already familiar) menu

Case 1: Redesigning a traditional GUI interface
- Type II error is preferable to a Type I error , Why?

Case 2: Designing a digital mapping application where experts perform extremely frequent menu selections
- Type I error is preferable to a Type II error, Why?

76

other tests: correlation

Measures the extent to which two concepts are related
- e.g., years of university training vs computer ownership per capita

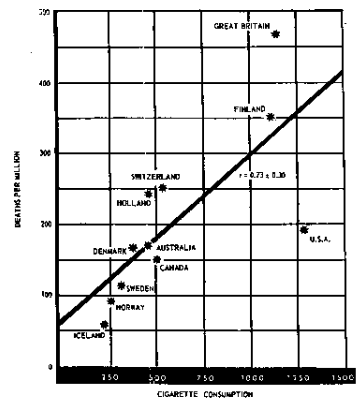
How?
- obtain the two sets of measurements
- calculate correlation coefficient
  - +1: positively correlated
  - 0: no correlation (no relation)
  - –1: negatively correlated

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
  - e.g., (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

77

non-HCI sample study: cigarette consumption

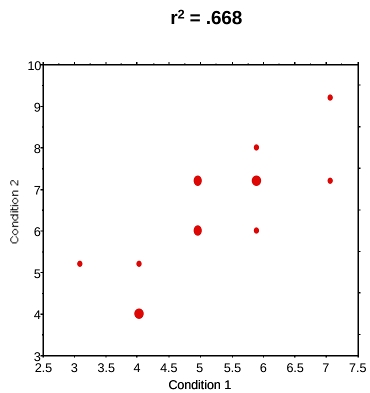
Crude Male death rate for lung cancer in 1950 per capita consumption of cigarettes in 1930 in various countries.



78

correlation

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



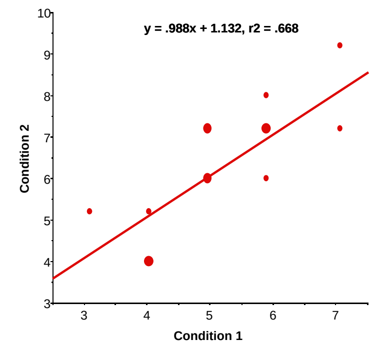
79

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



80

you now know

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

81

you now know

There are many statistical methods that can be applied to different experimental designs

- T-tests
- Single factor ANOVA
- Factorial ANOVA (case study)
- Correlation and regression

Significance levels and 2 types of errors

ANOVA terminology

- factors, levels, cells
- factorial design
  - between, within, mixed designs

82

additional slides: material I assume you know

types of variables

samples & populations

normal distribution

variance and standard deviation

83

types of variables
(independent or dependent)

discrete: can take on **finite** number of levels

- e.g. a 3-color display can only render in red, green or blue;
- a design may be version A, or version B

continuous: can take any value (usually within bounds)

- e.g. a response time that may be any positive number (to resolution of measuring technology)

normal: one particular **distribution** of a continuous variable

84

populations and samples

statistical sample =
approximation of total possible set of, e.g.

- **people** who will ever use the system
- **tasks** these users will ever perform
- **state** users might be in when performing tasks

} ← the population

"**sample**" a representative fraction

- draw **randomly** from population
- if large enough and representative enough, the **sample mean** should lie somewhere near the **population mean**

85

confidence levels

"the **sample mean** should lie somewhere near the **population mean"**

how close?

how sure are we?

a confidence interval provides an **estimate of the probability** that the statistical measure is valid:

"We are **95%** certain that selection from menus of five items is faster than that from menus of seven items"

how does this work?
important aspect of experiment design

86

establishing confidence levels:
normal distributions

fundamental premise of statistics:

predict behavior of a **population** based on a **small sample**

validity of this practice depends on the **distribution**

of the population and of the sample

many populations are **normally distributed**:

many statistical methods for **continuous dependent variables** are based on the assumption of normality

if **your sample is normally distributed**,
your **population is likely to be**,

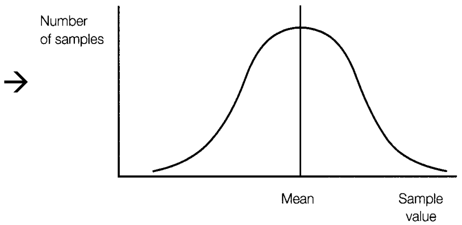
and these statistical methods are valid,

and everything is a lot easier.

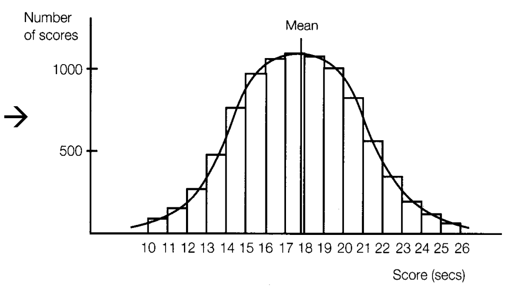
87

what's a normal distribution?

population →

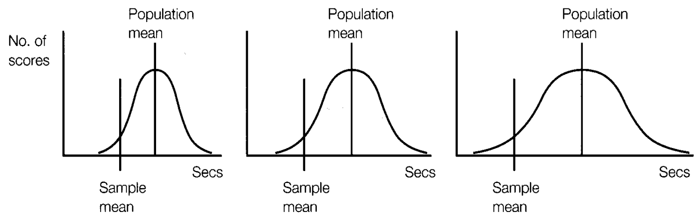


sample →



variance and standard deviation

all normal distributions are not the same:



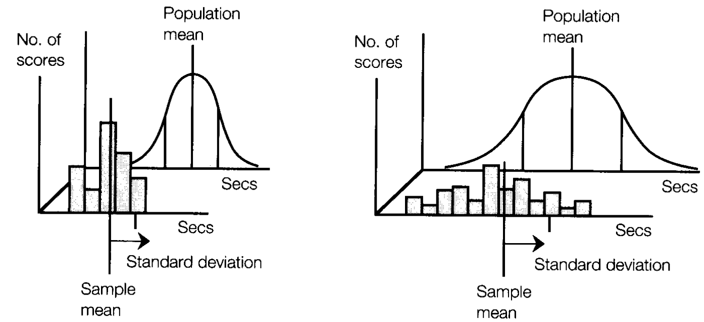
population variance is a measure of the distribution's "spread"

all normal population distributions still have the same shape

89

how do you get the population's variance?

estimate the population's (true) **variance**
from the (measured) sample's **standard deviation:**



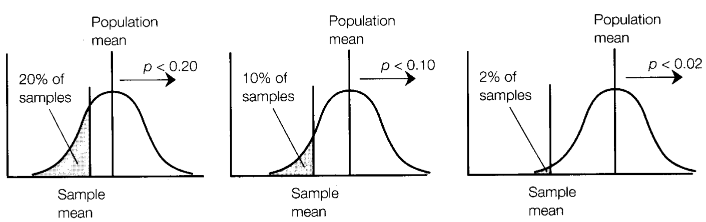
90

what's the big deal?

if you know you're dealing with samples from a normal distribution,

and you have a good estimate of its variance
(i.e. your sample's std dev)

then, you know the **probability** that a given sample came from that population (vs. a different one).



EXTRAS

quantitative ways to evaluate systems

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

methods
- controlled experiments
- statistical analysis

measures
- objective: user performance (e.g., speed & accuracy)
- subjective: user satisfaction (e.g., rated on a Likert scale)

93

statistical measures

allow answering questions like:
- is there a difference? → "hypothesis testing"
  - e.g., is one system better than the other one?
  - answers of form "we are 99% certain that selection from menus of five items is faster than that from menus of seven items"
- how big is the difference?
  - e.g., selection from five items is 260 ms faster than seven items.
- how accurate is the estimate?
  - e.g., "we are 95% certain that the difference in response time is faster by 260 ± 30 ms"
  - standard deviation or confidence intervals; probabilistic

94

statistical measures also good for…

just **looking** at data:
- some phenomena are not obvious from inspection of **raw** (completely unprocessed) data:
- statistical measures (and/or judicious plotting) can make them clear

e.g. **outliers**: single data items which are very different from the rest
- may be result of an experiment error
  or, a subject who had a bad day
  → if so, should remove from analysis
- or, it might be really important. EXERCISE CAUTION!

95

what are some tools for comparing two means?

variable types: which accurately describe the test situation

population sampling: can't study every possibility
- → statistical methods are based on an approximation from a small representative set

confidence levels: quantitative limit on the probability that our assessment is correct

normal distributions: many statistical techniques
- (e.g. to establish confidence levels) are based on a key assumption about the test population's structure

96

process of planning an experiment

any controlled experiment plan has a basic form of:

1. state hypothesis to test (the point of the experiment)
   e.g. measure some attribute of subject behavior
2. choose experimental conditions
   which vary only in values of certain "controlled" variables
   → **any change in measures can be attributed to Δ in conditions**
3. then, choose
   - subject pool to test
   - factors to manipulate, and their test values
   - size and form of the actual test (many choices)

97

variables

independent variable: *manipulated / controlled*
to produce different conditions for comparison

- each independent variable given a range of different values
- each value used in experiment = **level (also called a treatment)**

dependent variable: *measured*

- expectation that it is affected by the independent variable
- should be unaffected by other factors

some **subjective measures** can be applied against predetermined scales and analyzed quantitatively

98

example of controlled variables

an experiment will test
whether performance **improves**
as the **number of menu items decreases**.

independent variable: *number of menu items*

- test values: 5, 7, and 10 items (**3 levels tested**)

dependent variable: *speed of menu selection*

a more complex experiment:

- 2^{nd} independent variable
  = function names displayed on menu
  (dependent variable might depend on both)

99

simplest (and very common) design: the 2-sample experiment

based on comparison of **two sample means**:

- performance data in response to Designs A, B
  - compare performance of new design with old
  - compare performance of 2 new designs

or, comparison of **one sample mean with a constant**:

- performance data in response to Design A, compared to performance requirement
  - determine whether single new design meets key design requirement

100

hypothesis testing for your project

3 possibilities *(implications for prototype planning)*:

1. compare performance of new design with old
2. compare performance of 2 new designs
3. determine whether single new design meets key design requirement

   e.g. 'Telereg', where an essential performance requirement is given without reference to any past system:

   "95% of undergraduates should take no more than 5 minutes to register over the phone"

101