CPEN441: user interface design

controlled experiments - I

Controlled Studies

- eventually will want to determine whether interface is:
  - Better or worse
  - Meets spec
- controlled experiment is typical approach

This section: controlled experimentation
(use for your primary Pass 2 eval)

- process
- experiment components
- hypothesis testing
- simplest experiment design
- basic statistical methods
- normal distributions

read supplementary reading!

this material is well covered in the readings:

Newman & Lamming, Ch 10

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 ∆conditions**
3. then, choose
   - subject pool to test
     - match to target pool
     - online resources like *prolific.com*
   - factors to manipulate, and their test values
   - size and form of the actual test (many choices)

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”

experiment components

- hypothesis(es)
- variables:
  - independent
  - dependent
  - "nuisance"
- subjects
- experiment task
- experiment design
- statistical measures for results analysis

7

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

8

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)

9

nuisance variables

- 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:
  usually treat as an additional experiment independent variable (easier when they can be controlled)
- a common nuisance variable: *subject* (individual differences)

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**

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:

   "the maximum amount of time it should take an undergraduate to register over the phone is 5 minutes"

subjects

pool: similar issues as for uncontrolled studies

- match expected user population as closely as possible
- age, physical attributes, level of education
- general experience w/ systems similar to those being tested
- experience and knowledge of task domain

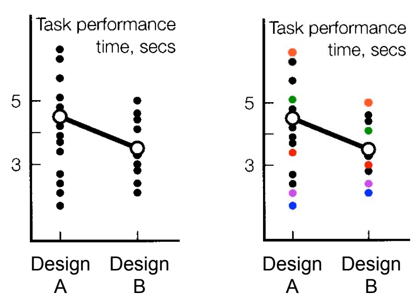
sample size: perhaps more critical here

- going for “statistical significance”
- should be large enough to be “representative” of population
- guidelines exist based on statistical methods used & required significance of results

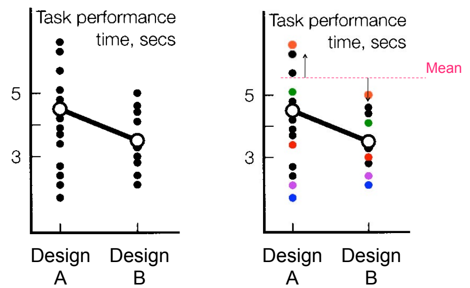
subjects

individual 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:



most common way to deal with:

– subtract each individual's mean performance at two factor levels from overall mean score, before combining with other individuals

– paired test, used with within-subject protocol

subjects, cont.

within-subject comparisons:

- **all subjects exposed to every condition**

→ 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)

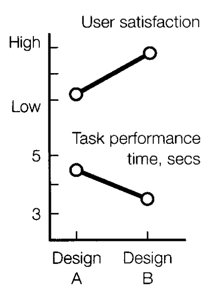
between-subject comparisons:

- **subjects only exposed to one condition**

→ primary comparison is from subject to subject

- less statistical power, more subjects required

point of experiment design



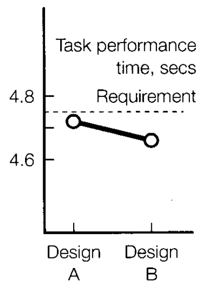
guard against ambiguous or misleading results

← a good (definitive) result

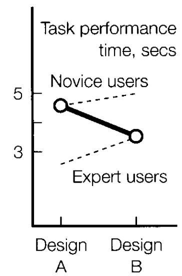
poor experiment design

less distinguishable results:

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



poor experiment design



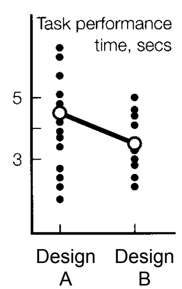
misleading results

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

poor experiment design

large spread in values

perhaps conditions were not well controlled



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**

22

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

23

statistical measures also good for…

just **looking** at data:

some phenomena are not obvious from inspection of **raw** (completely unprocessed) data:

statistic 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!

24

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?

can probably only test a small subset of all possible tasks.

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

if the design modification supports website **navigation**, test task should **not** require subject to work within a **single page**

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

25

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

26

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

27

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**

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

establishing confidence levels: normal distributions

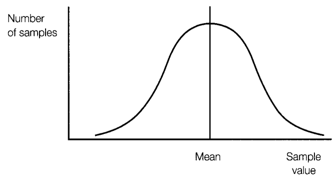
fundamental premise of statistics:
predict behaviour of a **population** based on a **small sample**

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

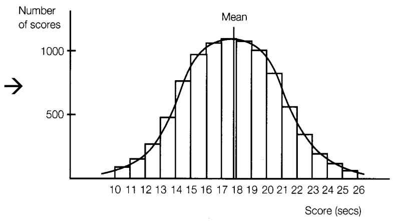
if **your sample is normally distributed**,
your **population is likely to be**,
and these statistical methods are valid,
and everything is a lot easier.

population →



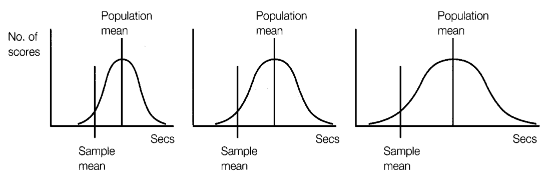
what's a normal distribution?

sample →



variance and standard deviation

all normal distributions are not the same:

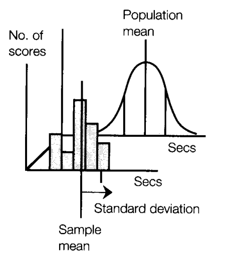


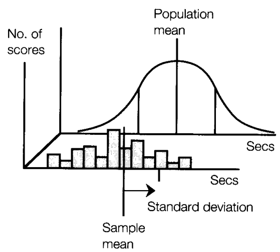
population variance is a measure of the distribution's "spread"

all normal population distributions still have the same shape

how do you get the population's variance?

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



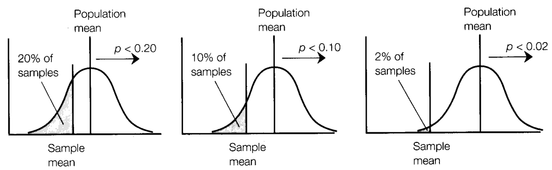


what's the big deal?

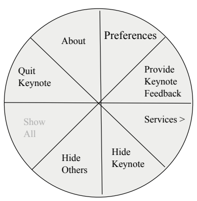
if you know you're dealing with samples from a normal distribution,

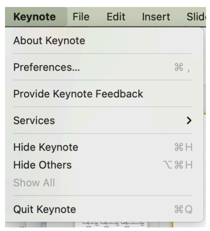
and you have a good estimate of its variance (i.e. your sample's std dev)

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

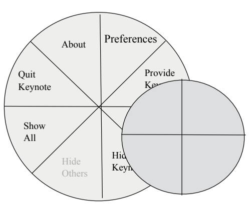


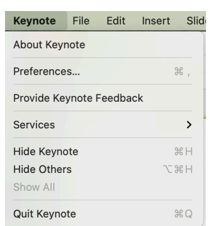
Example: Pie vs Pull-down menus





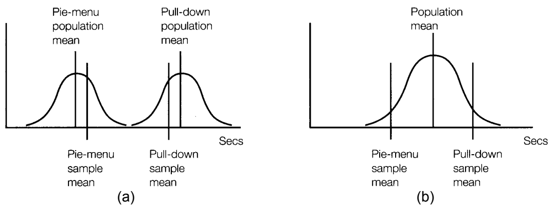
Example: Pie vs Pull-down menus





back to comparing means: for example,

(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?

Let's look further…

How can we mathematically tell:
- which distribution was data from?
- how likely is it not from that distribution?