CPEN441: user interface design

controlled user studies - II

analyzing the experiment

basic calculations

the t-test:

- estimate of confidence level on difference between 2 means
- types of t-test

example

χ^2 test

basic calculations

essential ingredients of most statistical computations:

- mean
- sum of squares of difference from mean
- degrees of freedom

- **sample** variance & standard deviation

mean & sum of squares

$$\text{mean} = \bar{X} = \frac{\sum X_i}{N}$$

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

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

error in N&L pg. 231

degrees of freedom (df)

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

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

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

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

sample variance & standard deviation

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

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

the t-test

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

the process (with lookup tables*):

1. compute df
2. choose desired **significance, *p*** (aka α)
3. calculate value of the ***t* statistic**

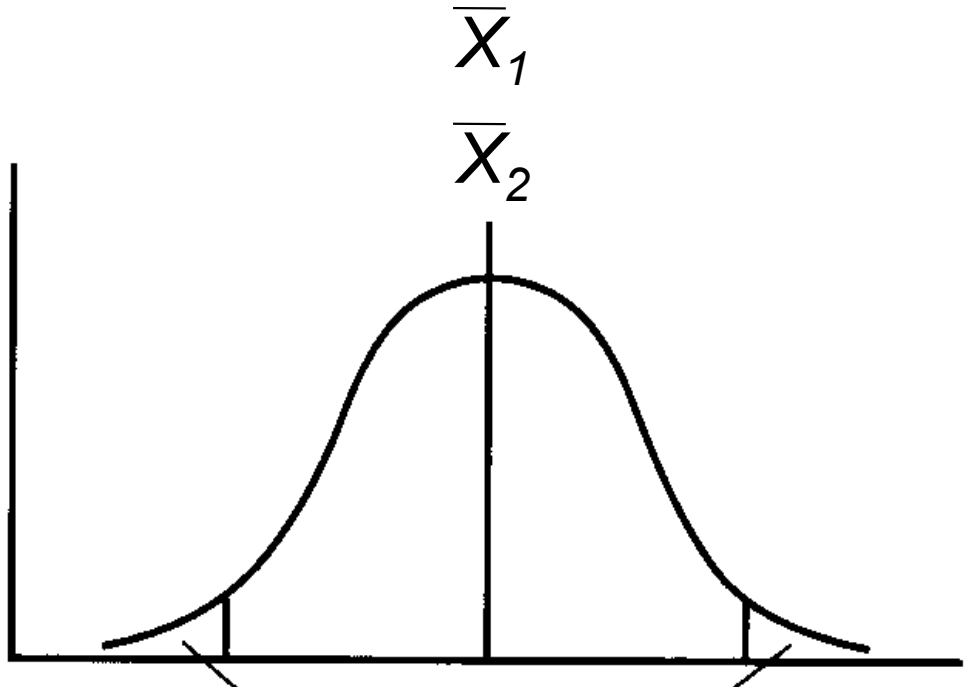
- compare it to the **critical value** of *t* given *p*, df: $t_{(p,df)}$

1. if $t > t_{(p,df)}$, can **reject null hypothesis at *p***

* with stat pack, you can compute p exactly

Two-tailed t-test: significance *p*

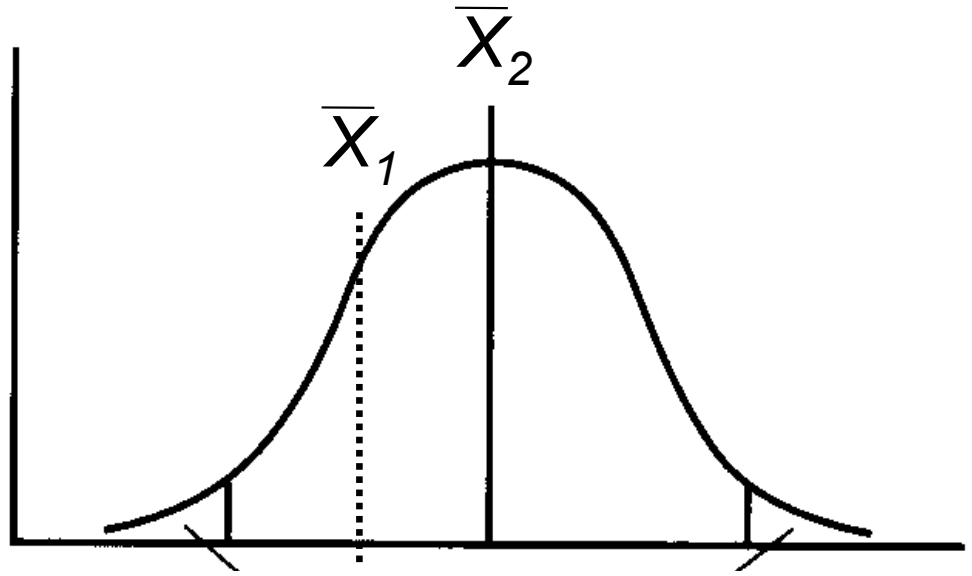
$H_0: \overline{X}_2 = \overline{X}_1$



Case 1 - what if there really isn't a difference?

Two-tailed t-test: significance *p*

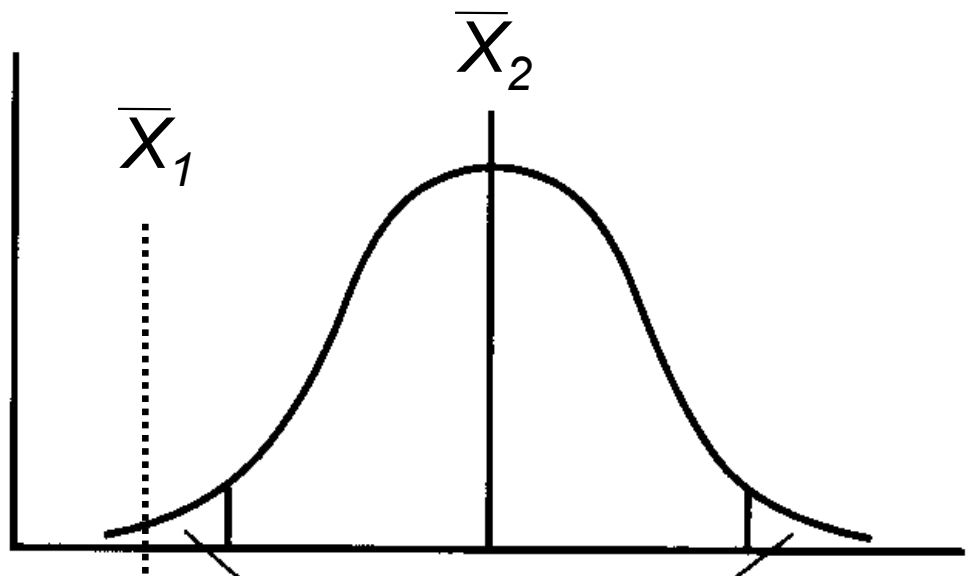
$H_0: \overline{X}_2 = \overline{X}_1$



Case 2 - what if there's a difference?

Two-tailed t-test: significance *p*

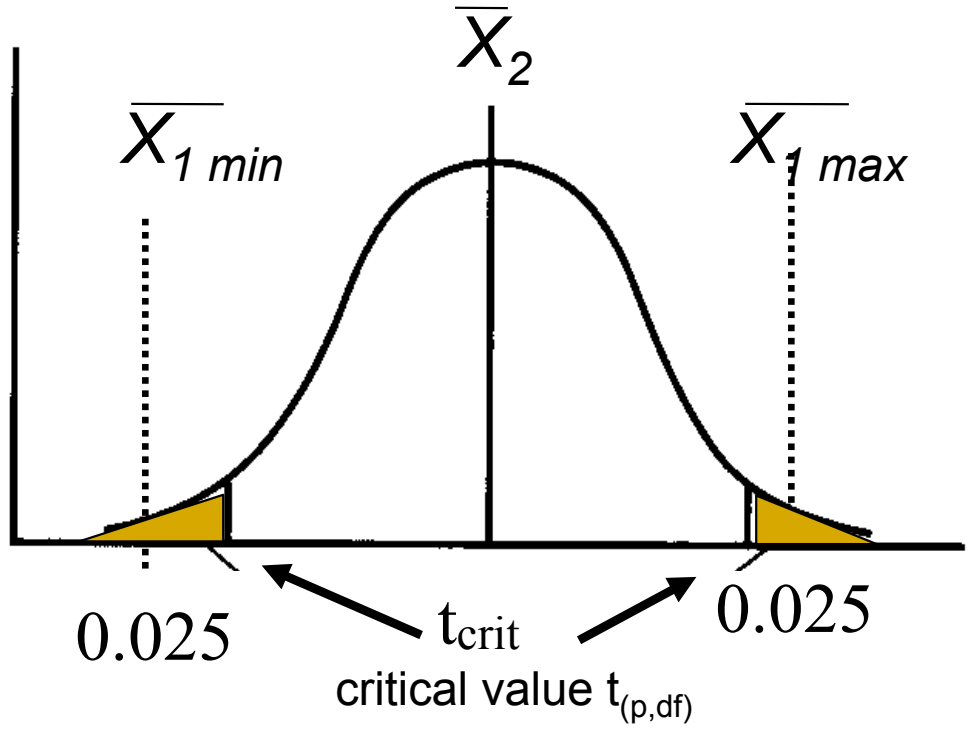
$H_0: \overline{X}_2 = \overline{X}_1$



Case 3 - what if there's a big difference?

Two-tailed t-test: significance *p*

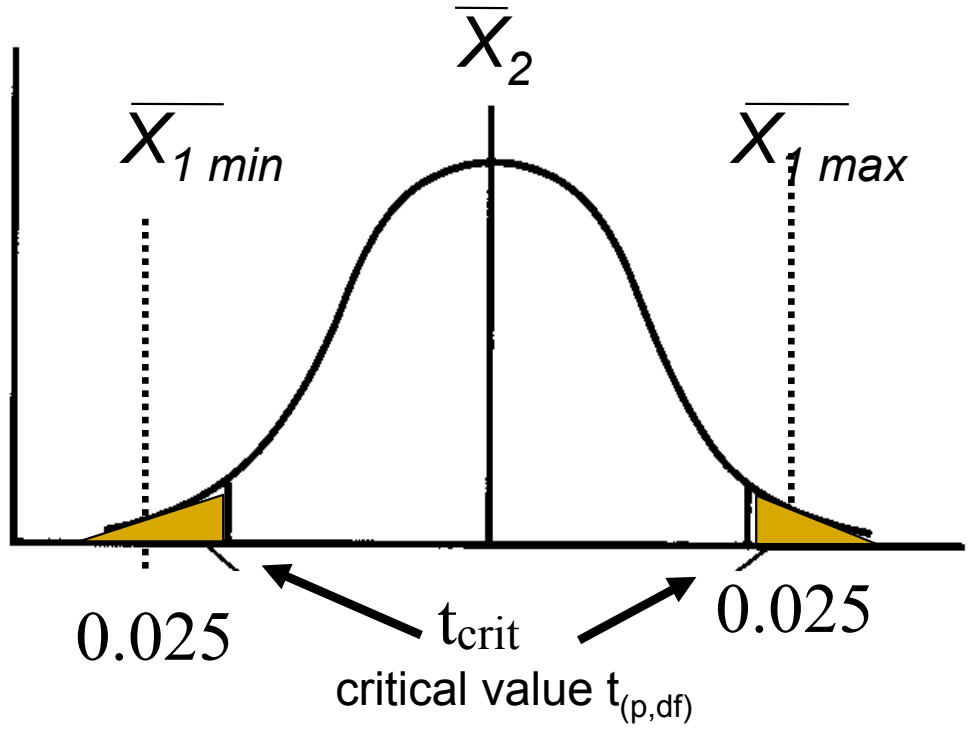
$H_0: \overline{X}_2 = \overline{X}_1$



What if we want to make sure the p <= 0.05?

Two-tailed t-test: significance *p*

$H_0: \overline{X}_2 = \overline{X}_1$

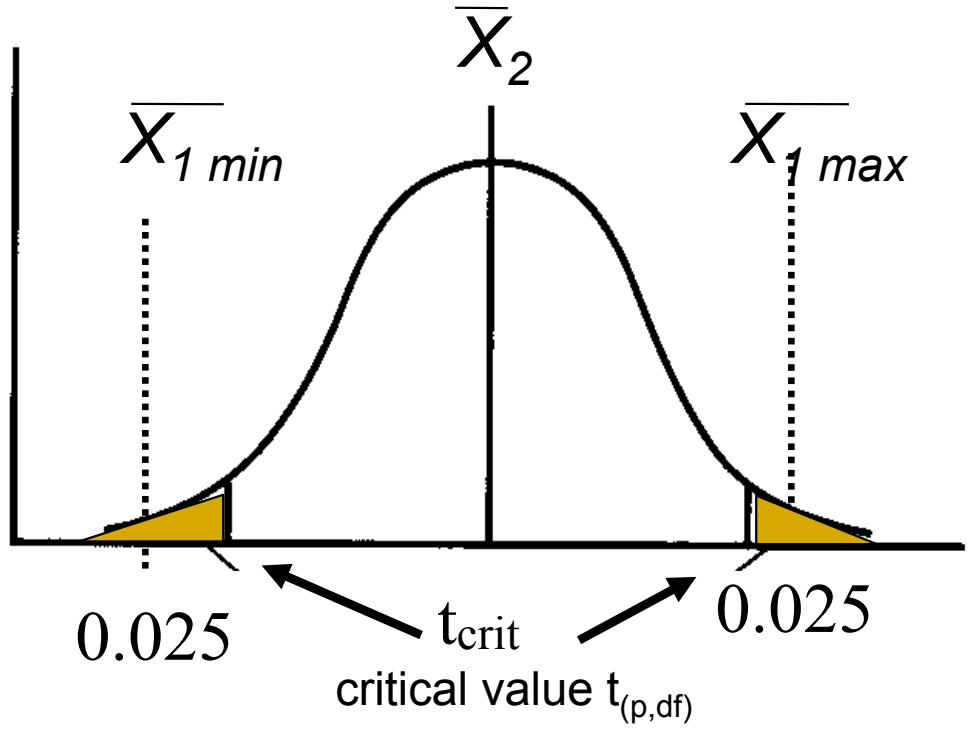


What if we want to make sure the p <= 0.05?

Calculate t and compare to t_{crit}

Two-tailed t-test: significance *p*

$H_0: \overline{X}_2 = \overline{X}_1$



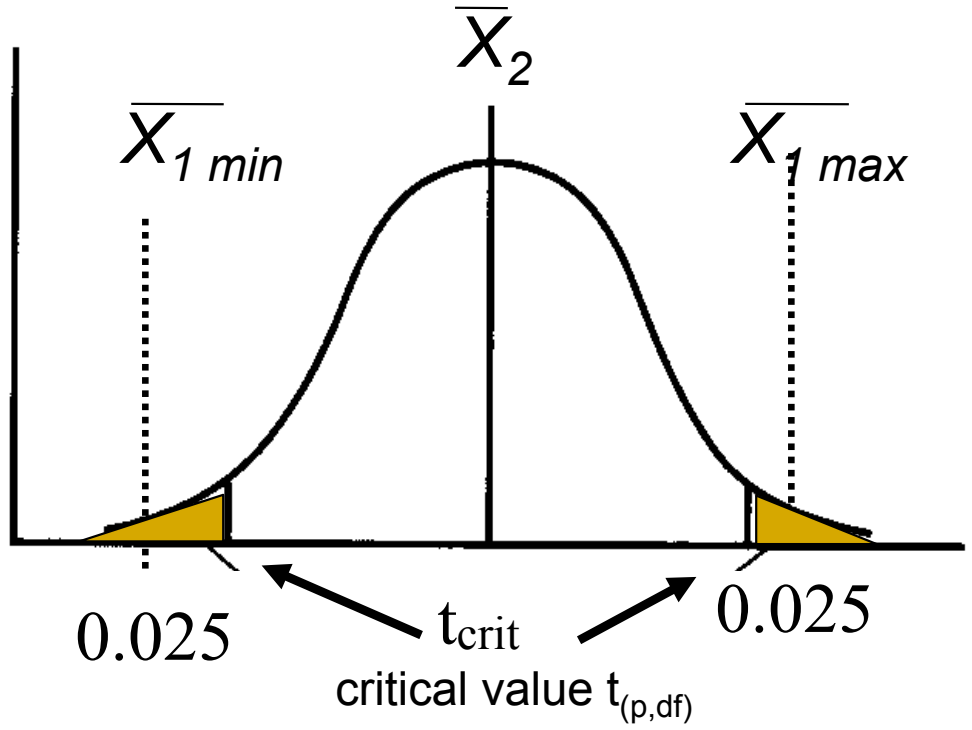
What if we want to make sure the p <= 0.05?

Calculate t and compare to t_{crit}

In this case, we can reject the H_0

Two-tailed t-test: significance *p*

$H_0: \overline{X}_2 = \overline{X}_1$



You can also calculate t and use excel/stat pack to calculate the area to give exact value of p and see if it less than your pre-determined significance test.

two-tailed and one-tailed t-test

Two-tailed test when it matters whether A > B or A < B

- which design is fastest
- Most common for HCI controlled studies

Sometime, you only care in one direction

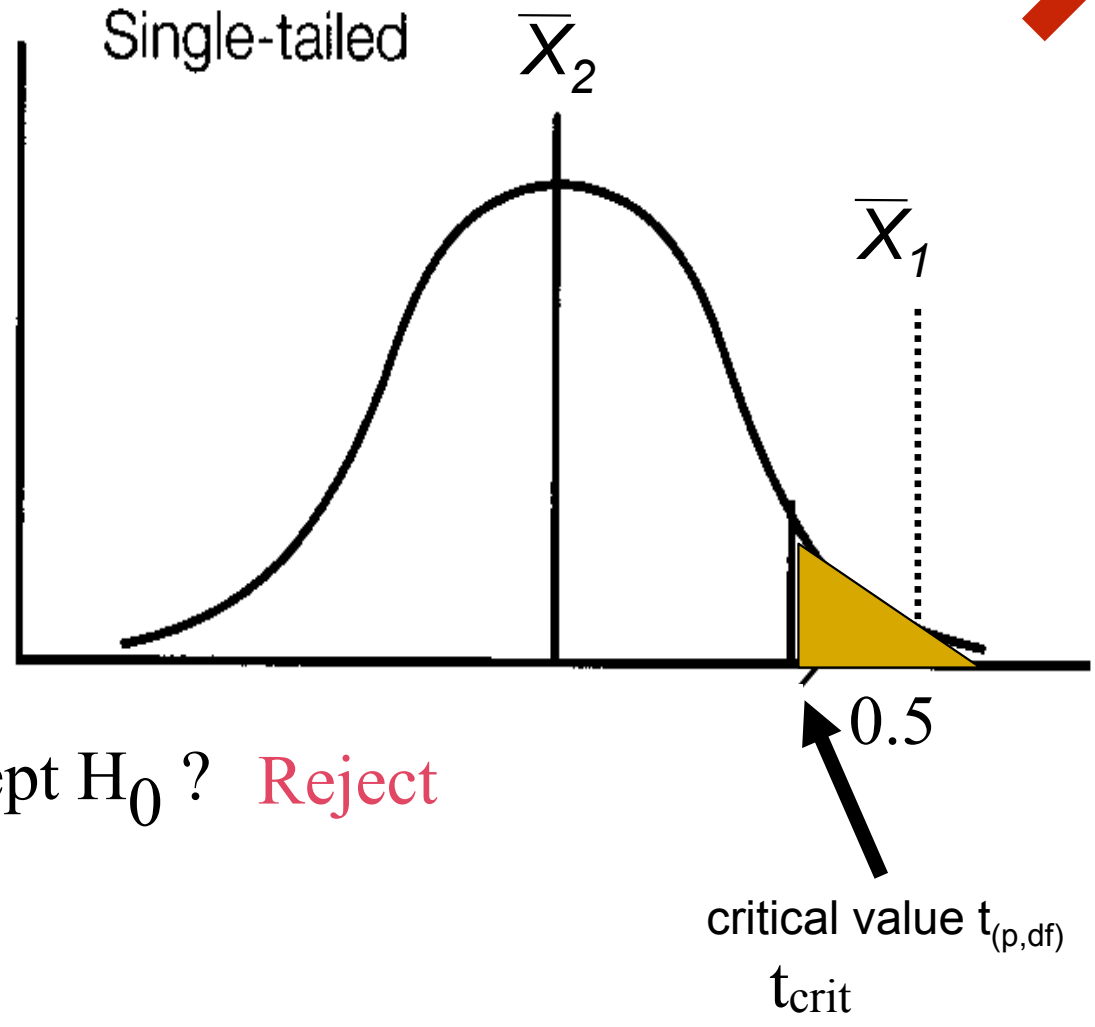
- A is not slower than B (and it's the same price!)
- A has no more errors than B
- A is no worse than 500msec
- be careful with your alternative hypothesis

» Use one-tailed test

one-tailed t-test: significance *p*

H_1: X1 is not significantly slower than X2

H_0: $\overline{X}_2 = \overline{X}_1$



Reject or accept H_0 ? Reject

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} :

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

compute *t*:

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

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

comparing *t* with critical value:

look $t_{(p,df)}$ up in a pre-computed table:

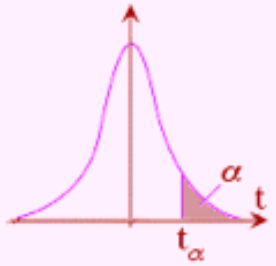
- in back of any statistics textbook
- on web, e.g.

http://www.medcalc.be/manual/mpage13-04b.html

http://www.statsoft.com/textbook/stathome.html

or (now most common): compute the *p* for your $t_{(df)}$ directly:

- Matlab or Excel functions
- web calculators (e.g. search on "student's t-test")



T-Distribution Table

two tailed: α 0.2 0.1 0.05 0.02 0.01 0.002 0.001

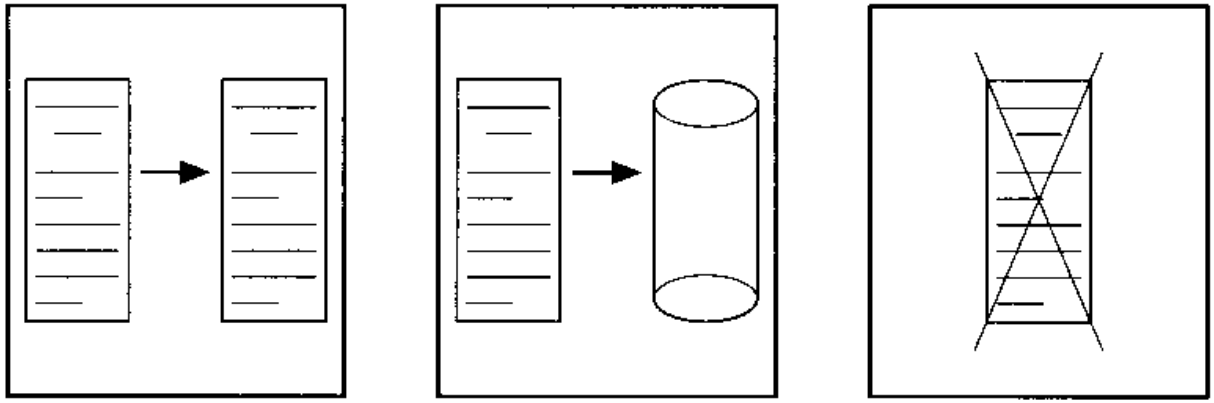
df	α = 0.1	0.05	0.025	0.01	0.005	0.001	0.0005
∞	t_α=1.282	1.645	1.960	2.326	2.576	3.091	3.291
1	3.078	6.314	12.706	31.821	63.656	318.289	636.578
2	1.886	2.920	4.303	6.965	9.925	22.328	31.600
3	1.638	2.353	3.182	4.541	5.841	10.214	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.894	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725

example: evaluating icon designs

two styles:

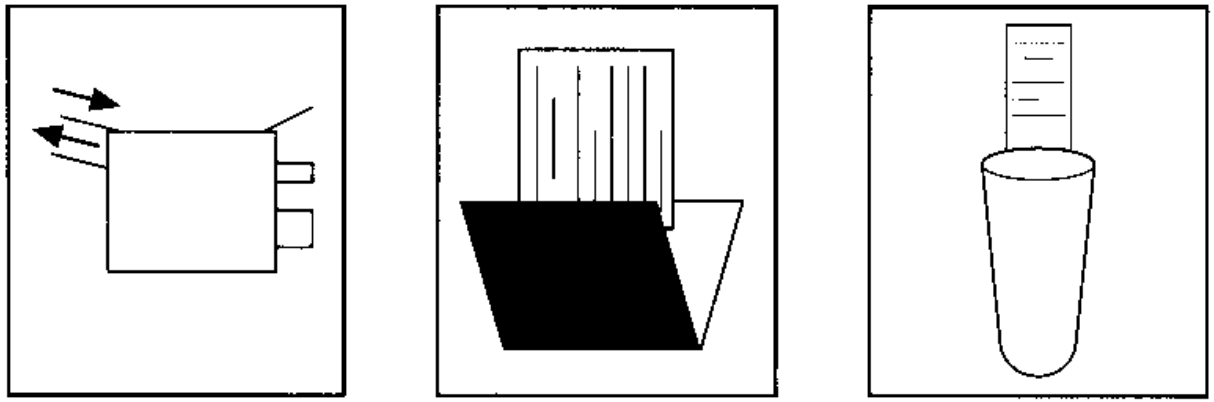
1. which will be easiest for users to remember?
2. must be faster than 750msec to use
3. which is preferred?

abstract →



Copy Save Delete

naturalistic →



hypotheses?

H_1: Natural icons are easier to recall than abstract icons

H_{o-1}: There is no difference in recall between the natural icons and the abstract icons

H_2: Natural (or Abstract) icons require less than 750msec to select.

H_{0-2}: Natural (or Abstract) icons require 750msec or more to select.

H_3: Natural icons are preferred to Abstract icons.

H_{0-3}: There is no difference in preference between Natural and Abstract icons.

variables

independent variable(s)?

- icon style (2 levels: naturalistic, abstract)

dependent variable(s)?

- users able to remember “more easily”:
  how will we measure this?
  … accurate recall, speed, user preference?

for this example, let’s assume:

1. **number of mistakes** in selection (error rate)
2. **time taken to select icon**
3. **preference specified**

variables

independent variable(s)?

- icon style (2 levels: naturalistic, abstract)

dependent variable(s)?

- users able to remember “more easily”:
  how will we measure this?
  … accurate recall, speed, user preference?

for this example, let’s assume:

1. **number of mistakes** in selection (error rate)
2. **time taken to select icon ** OUR ANALYSIS**
3. **preference specified ** OUR ANALYSIS**

experiment task

2 interfaces which are identical in every way except icon design

selection task which can be repeated for both conditions (natural, abstract) - ???

- produce a document
  a natural task: results may be more transferable
- select appropriate icon in response to a prompt
  an artificial task: may be easier to control

first, will have to give subject time to **learn** icons to avoid learning effects in data

experiment task, cont.
for this experiment, let's choose:

- **prompt user:** *e.g. 'save a document'*,
  then require user to select the proper icon
- **one task** = randomly generate a series of X prompts for each of the 3 different task types (copy, save, delete)
  for a given set of icons (either natural or abstract)
- **for each subject**: one task with **each** set of icons
  - Measure
    - speed
    - preference

how many times? (what is X?)

→10-30 reasonable for this (relatively easy-to-learn) task

consider desired session duration, expected learning curves, effect of boredom on performance.

Is this a within-subject or a between-subject design?

data and analysis

iconDataAnalysis.xls

Another example: Testing to Requirement

Requirement:

- must be not be slower than 750msec for selection using menus

Approach:

- find confidence intervals
  - max and min extremes of population mean are above (below, within) requirement (or vice-versa)

$X_{min} = \overline{X} - (t_{p,df} * s_{em})$; or Rdiff-= 750 - $(t_{p,df} * s_{em})$

$X_{max} = \overline{X} + (t_{p,df} * s_{em})$; or Rdiff+ = 750 + $(t_{p,df} * s_{em})$

$s_{em} = sqrt(s^2/N)$

(NOTE: use one sided $t_{p,df}$ since we only care if it is more than 750msec; it won't matter if is the same or less than 750msec since it won't impact our design either way)

Preference Example

Questionnaire asked for preference

- categorical data

I.e. either natural or abstract

Use χ^2 (CHI-squared) statistic: $\chi^2 = \text{SUM}((f_o - f_e)^2/f_e)$

- evaluate difference between observed and expected frequency of categories

- compare using χ^2 statistic and p value

dof = number of categories -1

see iconDataAnalysis.xls

Preference Example

$$\chi^2 = \text{SUM}((f_o - f_e)^2/f_e)$$

f_o = frequency ; f_e = equal frequency

DF	alpha=0.05	alpha=0.01
1	3.84	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21

how do you make sure your data is good?

threats to validity

construct validity

- are we measuring what we think we are measuring?
- e.g., create a questionnaire to assess early "adopter-ness", but in fact it assesses financial ability to buy new technology instead

internal validity

- is there a causal relation between independent & dependent variables?
- e.g., nuisance variable causing the change in the dependent variable
- e.g., Hawthorne effect – subjects change their behaviour because they know they are being studied

how do you make sure your data is good?

threats to validity, continued

statistical validity

- could the results be a fluke?
- e.g., were the statistical tests used appropriate? (e.g., many tests assume a normal distribution)

external validity

- do the results generalize?
- e.g., sample not representative of true population
- e.g., insufficient description of experiment protocol

ecological (face) validity (form of external validity)

- e.g., tasks in experiment not representative of real tasks

Summary

Need to do controlled studies when:

- need to know when design makes things better or worse
- or you are meeting specifications

Covered

- how to run a typical experiment
- how to do simple t-test (1 and 2 sided)
  - hypothesis testing
  - estimating probability data came from same normal distribution
- test to criteria
- χ^2 test

Summary

Careful experimental design needed to:

- address:
  - construct, internal, external, statistical and face validity
- beware of nuisance variables
- choose representative subjects
  - reasonable numbers
- choose representative tasks
- choose appropriate statistical technique