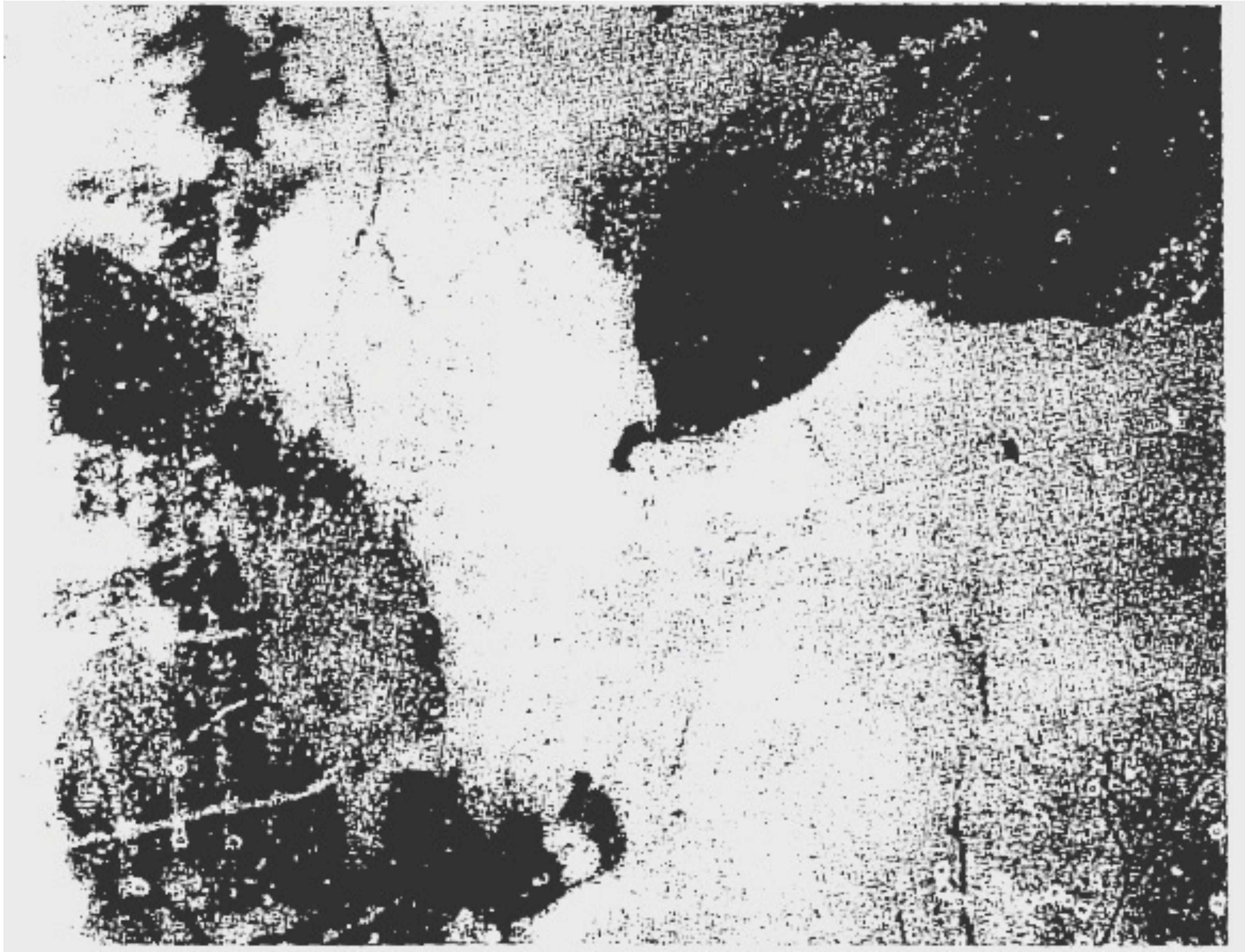
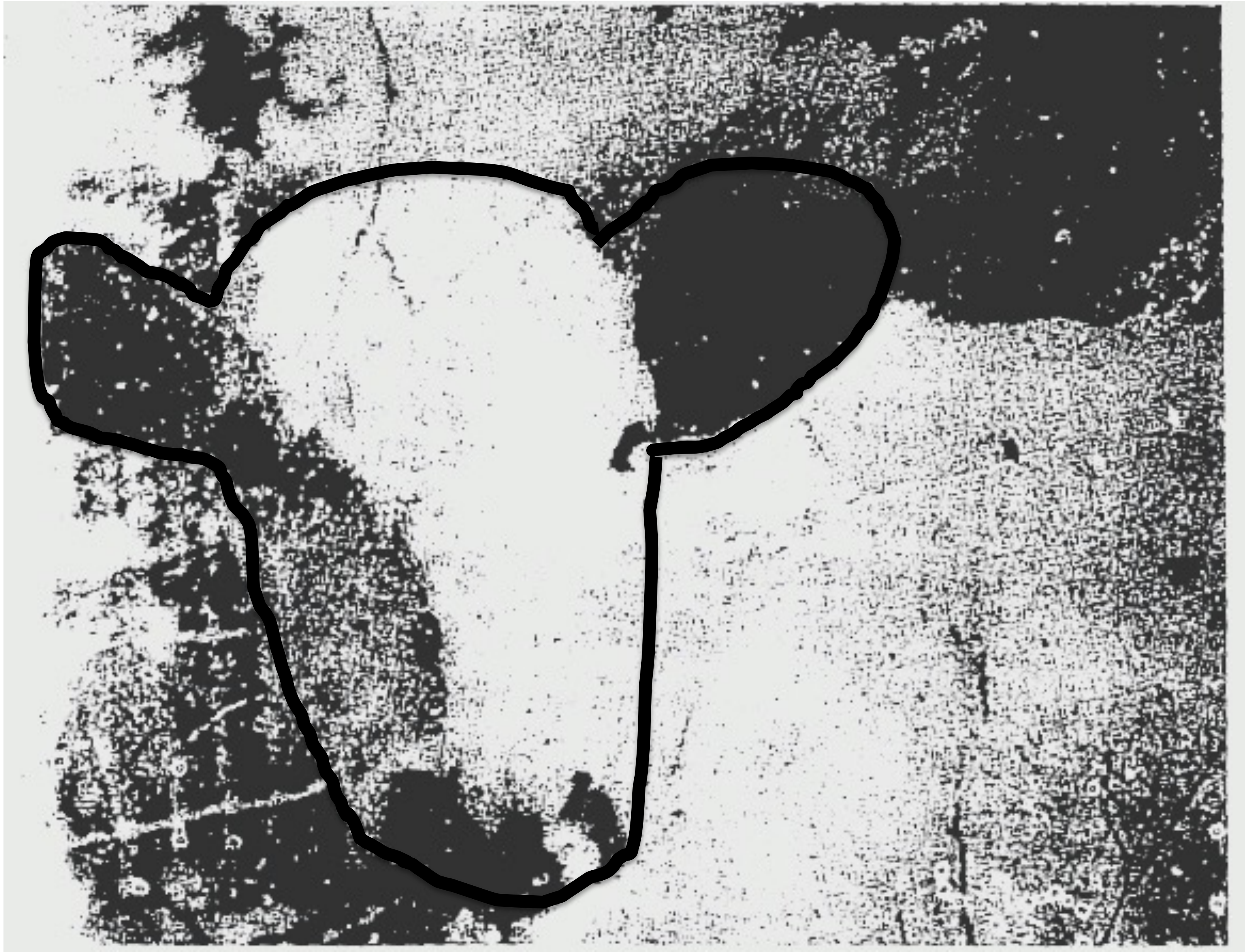
Evaluation

Evaluation

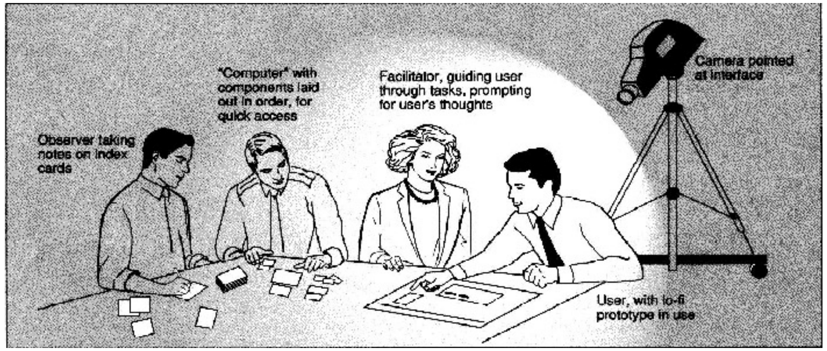
- Designers are blind to their designs
- They are uniquely unqualified to assess usability
- **Problem:** how to detect mismatch between user’s model and designer’s model?
- **Answer**: record realistic interaction
- Requires structure: simple observation is insufficient





“Think Aloud” Evaluation

- Subjects continually prompted to verbalise their thoughts
  - What they are trying to do
  - Why they took an action
  - How they interpret feedback


"Computer" with components laid out in order, for quick access
Facilitator, guiding user through tasks, prompting for user's thoughts
Camera pointed at interface
Observer taking notes on index cards
User, with lo-fi prototype in use

- One way communication from user (except prompts)
- Gives insights into user’s model
- *Hard* to talk and concentrate; awkward
- Often uncomfortable for subjects

Cooperative Evaluation

- Two subjects (sometimes one a confederate)
- Natural two-way communication
- More natural, more comfortable
- Criticism more likely
- Use Hawthorne effect to advantage

Interviews

- Good for probing particular issues
- Often leads to constructive suggestions
- Prone to post-hoc rationalisation

- Plan a central set of questions
  - Focuses the interview; base consistency
  - Be willing to follow interesting leads

Questionnaires

- Expensive to prepare; cheap to administer
- Doesn’t require presence of evaluator
- Quantitative and qualitative
- Only as good as questions asked
  - Know the purpose!
  - Know how you will analyse results
  - Know dissemination method (web, surface mail, etc.)

Questionnaires: Question Types

- Open-ended comments: important insights
- Closed questions: restrict responses
  - Take care with ambiguity
- Ranked
  - Good for forcing comparison
- Likert items: level of agreement

It is easy to recover from mistakes:

Disagree □ □ □ □ □ Agree

Continuous Evaluation

- Monitoring actual system use
  - – Field studies
  - – Diary studies
  - – Logging and ‘Customer Experience Programs’
  - – User feedback and gripe lines

Crowd-Sourced Experiments

- Crowdworkers complete “Human Intelligence Tasks” (HITs) for payment
- Disseminated on the web
- Pay *at least* US minimum wage
- Problems with noisy data and criteria for exclusion
- Include “attention check” questions
- Workers have a HIT approval rating; used as filter
- Amazon Mechanical Turk: www.mturk.com

User performance data collection

- Key loggers, customer experience improvement programs, diary studies, etc.
- Exploratory: collect loads of data and hope something interesting shows up
- Difficult to analyze

- Targeted
  - Frequency of use (e.g., hotkeys, scrollwheel)
  - Characterise activities (e.g., scrolling patterns, web use)
- (Aside: in controlled experiments, log *everything*)

Formal Empirical Evaluation

- Controlled experiments (coming up)
- Strict statistically testable hypothesis:
  better, worse, no difference
- Measure participants’ response to manipulation of experimental conditions
- Repeatable results through rigorous method
- Time-consuming, low-level UI issues, expensive

Ethics

- Testing can be distressing
  - Pressure to perform; errors inevitable
  - Feeling of inadequacy
  - Competition with other subjects
- Golden rule:
  - Subjects should be treated with respect!

https://www.youtube.com/watch?v=iktqSLt1Kes

Ethics – Before the test

- Don’t waste the user’s time
  - use pilot tests to debug experiments, questionnaires etc
  - have everything ready before the user shows up
- Make users feel comfortable
  - emphasize that it is the system that is being tested, not the user
  - acknowledge that the software may have problems
  - let users know they can stop at any time
- Maintain privacy
  - tell user that individual test results will be completely confidential
- Inform the user
  - explain any monitoring that is being used
  - answer all user’s questions (but avoid bias)
- Only use volunteers
  - user must sign an informed consent form

Ethics – During the test

- Don’t waste the user’s time: no unnecessary tasks

- Make users comfortable
  - try to give user an early success experience
  - keep a relaxed atmosphere in the room
  - coffee, breaks, etc
  - hand out test tasks one at a time
  - never indicate displeasure with the user’s performance
  - avoid disruptions
  - stop the test if it becomes too unpleasant

- Maintain privacy
  - do not allow the user’s management to observe the test

Ethics – After the test

- Make the users feel comfortable
  - state that the user has helped you find areas of improvement
- Inform the user
  - answer particular questions about the experiment that could have biased the results
- Maintain privacy
  - never report results in a way that individual users can be identified
  - only show videotapes outside the research group with the user's permission

http://www.canterbury.ac.nz/humanethics/hec/apply.shtml

Controlled Experiments

Controlled experiments

- Characteristics
  - lucid and testable hypothesis
  - quantitative measurement
  - measure of confidence in results (statistics)
  - replicability of experiment
  - control of variables and conditions
  - removal of experimenter bias

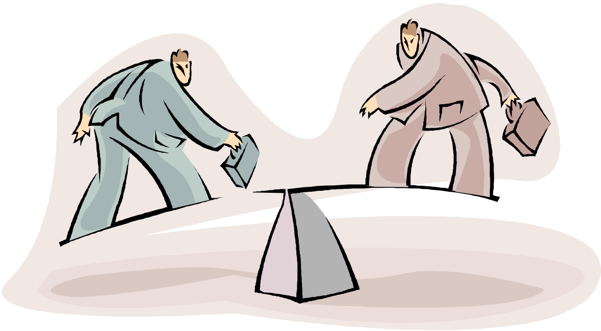
Research Question/Hypothesis

Having invented *gizmo*

- ~~Lets do a user study of *gizmo*!~~
- Is *gizmo* any good?
- Does *gizmo* beat the competition?
- Is *gizmo* faster than the competition?
- Is *gizmo* faster than *de facto* after 10 mins use?
- Is *gizmo* faster and less error prone than *de facto* after 10 mins use?

Research Question Tradeoff: Internal versus External Validity

- External validity: findings are broad/real
  "Is *gizmo* any good?"
- Internal validity: findings are due to conditions
  "Is *gizmo* faster and less error prone than *de facto* after 10 mins use?"



- Tradeoff
- Often addressed with multiple experiments

Research Question

- In HCI, most experimental research questions are comparative
  - Faster, more accurate, preferred (etc.) to baseline(s)
  - Is there a difference?
  - How big (and is this practical)?
  - How likely is it due to chance (statistics)?

Research Question (cont.)

- Hypothesis is lucid and testable
- Normally expressed in negative (“null hypothesis”)
  - “no difference” between …
  - Scientists are conservative
- Statistics may lead to rejection of null hypothesis (when $P(D|H_0)$ is low)

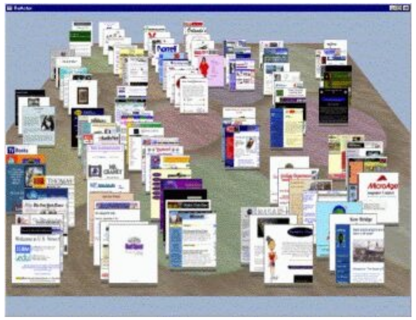
Research Question (cont.)

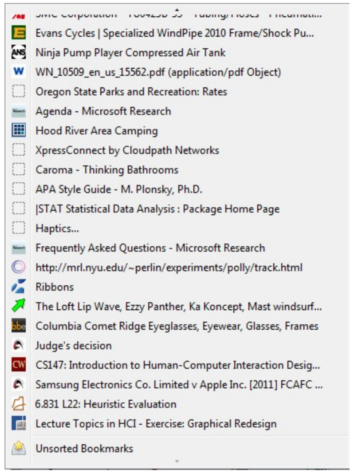
“There is no difference in user performance (time and error rate) when selecting a single item from a pull-down or pop-up menu”

“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 of 4 items, regardless of the subject’s previous expertise in using a mouse or using the different menu types”

Research Question (cont.)
Cause of comparative difference?

- A vs B comparisons can be good





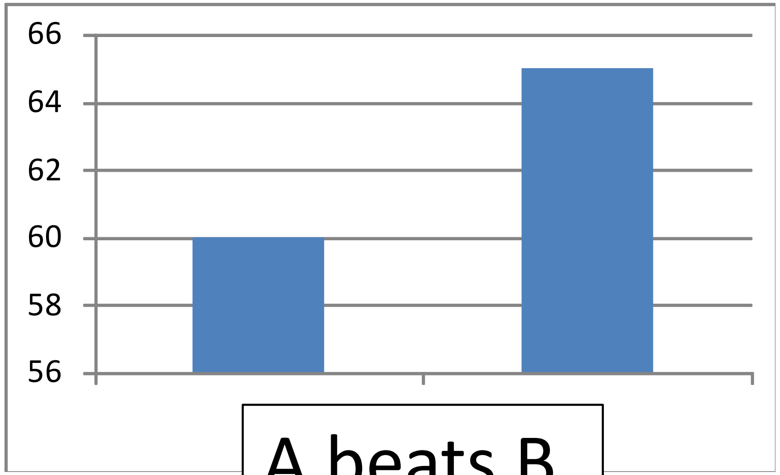
SMC Corporation - T0425B-33 - Tubing/Hoses - Pneumati...
Evans Cycles | Specialized WindPipe 2010 Frame/Shock Pu...
Ninja Pump Player Compressed Air Tank
WN_10509_en_us_15562.pdf (application/pdf Object)
Oregon State Parks and Recreation: Rates
Agenda - Microsoft Research
Hood River Area Camping
XpressConnect by Cloudpath Networks
Caroma - Thinking Bathrooms
APA Style Guide - M. Plonsky, Ph.D.
|STAT Statistical Data Analysis : Package Home Page
Haptics...
Frequently Asked Questions - Microsoft Research
http://mrl.nyu.edu/~perlin/experiments/polly/track.html
Ribbons
The Loft Lip Wave, Ezzy Panther, Ka Koncept, Mast windsurf...
Columbia Comet Ridge Eyeglasses, Eyewear, Glasses, Frames
Judge's decision
CS147: Introduction to Human-Computer Interaction Desig...
Samsung Electronics Co. Limited v Apple Inc. [2011] FCAFC ...
6.831 L22: Heuristic Evaluation
Lecture Topics in HCI - Exercise: Graphical Redesign

Unsorted Bookmarks

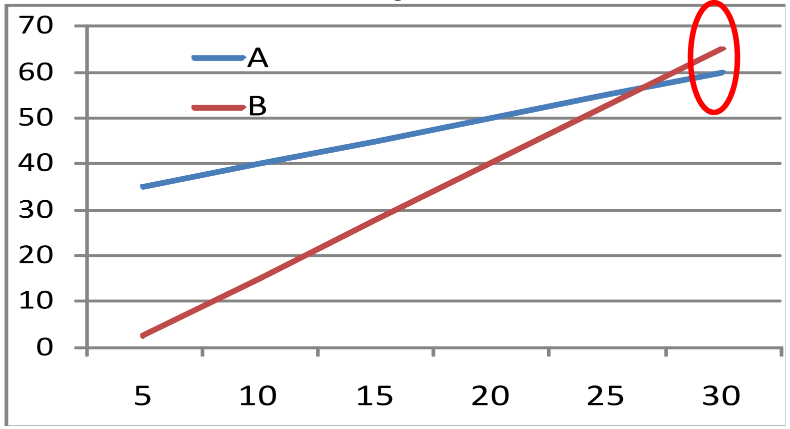
- But how to generalise?
- Know the human factor(s) underlying A/B
- Rephrase the experiment as: HF_A vs HF_B

Research Question (cont.)
Point analysis versus depth/theory/model

- A beats B or HF_A beats HF_B… nice
- Generally true, or just the tested condition?
- Identify & include salient secondary factors



A beats B



Oops. Only if n > 25

Experimental Terminology

- Independent variable
- Dependent variable
- Within versus between subjects
- Counterbalancing

Independent variables

- Controlled conditions
- Manipulated independent of behaviour
- May arise from participant classification
  e.g., males/females; gamers/non-gamers
- Discrete values are independent variable *levels*
  e.g., *Friction type* $\in$ *{high, low, variable}*
- ‘Independent variable’ ≡ ‘Factor’ with ANOVA

Dependent variables

- Measured
- Values depend on participant’s response to manipulation of the independent variable(s)
- Task time, error rate, speed, accuracy, overshoots, etc...

Within Subjects, Between Subjects

- Each independent variable is administered either within subjects or between subjects
- Within subjects: each participant tested on all levels

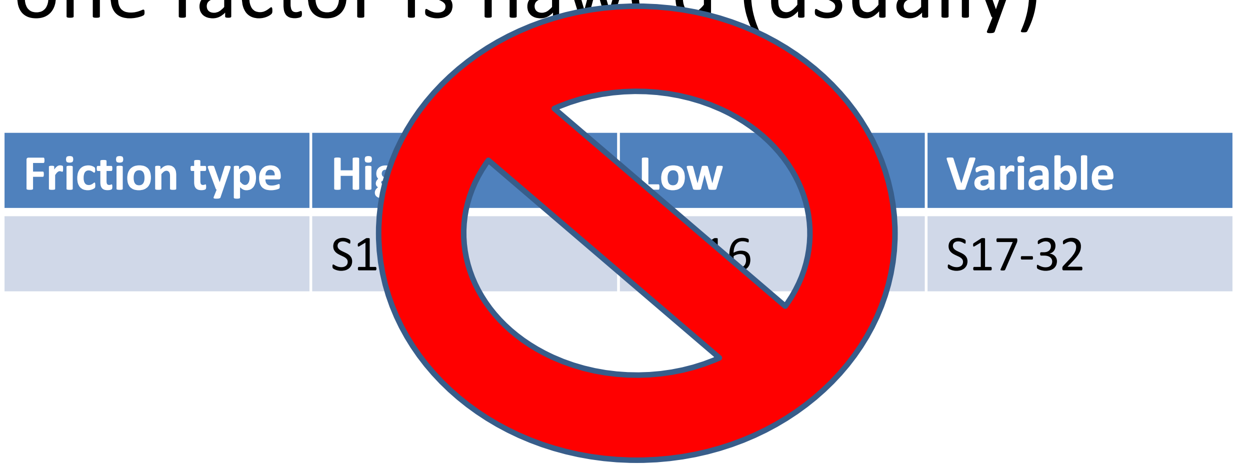
Friction type	High	Low	Variable
	S1-16	S1-16	S1-16

- Between subjects: each participants tested on one level

Friction type	High	Low	Variable
	S1-16	S17-32	S33-48

Within Subjects, Between Subjects

- Mixing within and between subjects treatment within one factor is flawed (usually)



Friction type	Hig	Low	Variable
	S1	6	S17-32

- (Mixing within subjects factors with between subjects factors is fine... multi-factor analysis, beyond 368)

Within Subjects or Between Subjects?

Within subjects:

+ Participants act as their own control

+ Fewer participants

- Need control for learning/fatigue effects

Within Subjects or Between Subjects?

Between subjects:

- Sometimes necessary (e.g., male/female)

+ No learning/fatigue effect

- Unmoderated variability

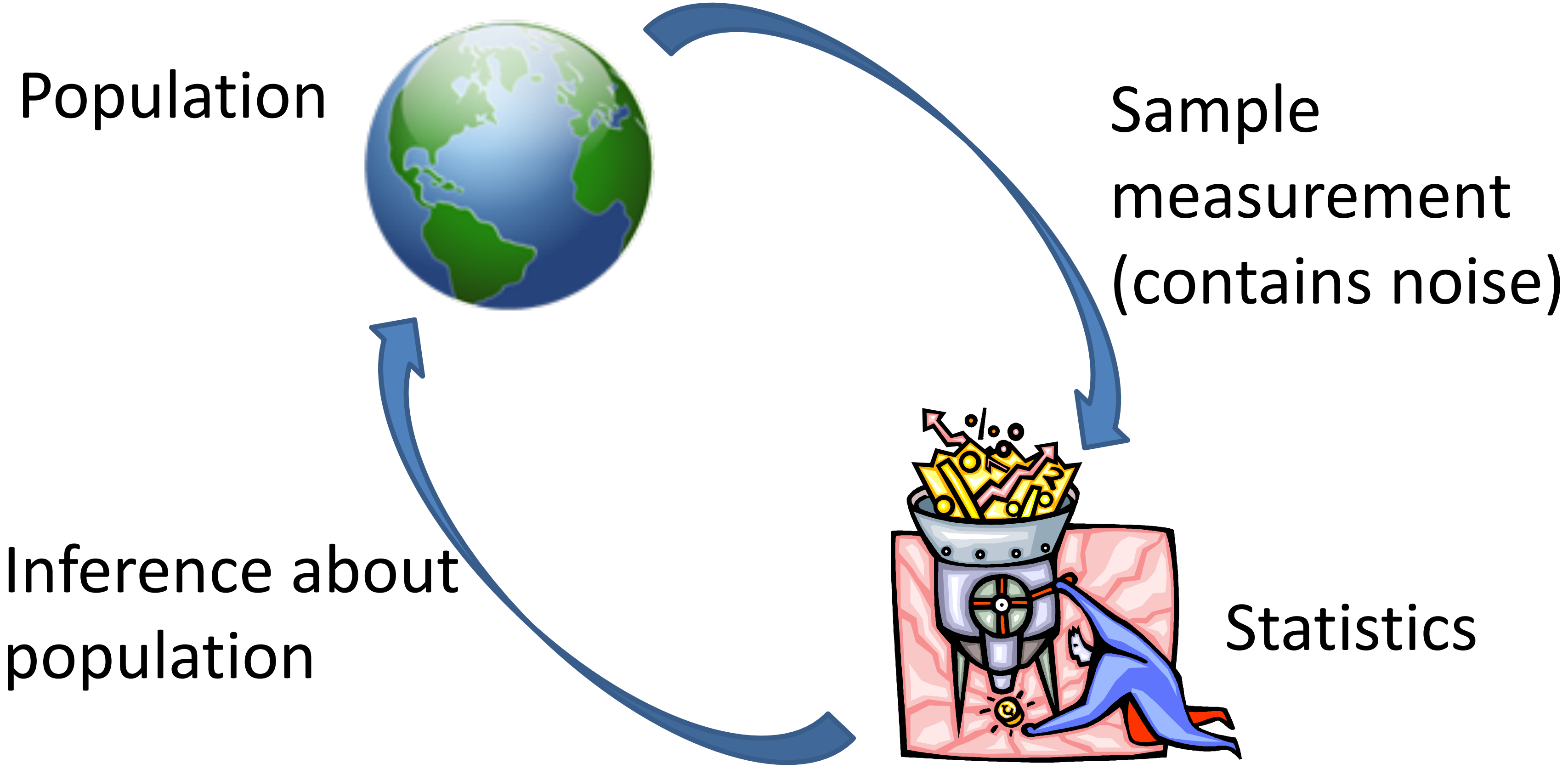
- More participants

Counterbalancing

- Within-subjects factors need control for learning/fatigue effects
- Participants divided into groups
- Different order for each group
- Group becomes a between subjects factor
  (ideally checked for asymmetric skill transfer[1], but often ignored)

[1] Poulton, E. C., & Freeman, P. R. (1966).

Statistics give confidence in answers



Population

Sample measurement (contains noise)

Statistics

Inference about population

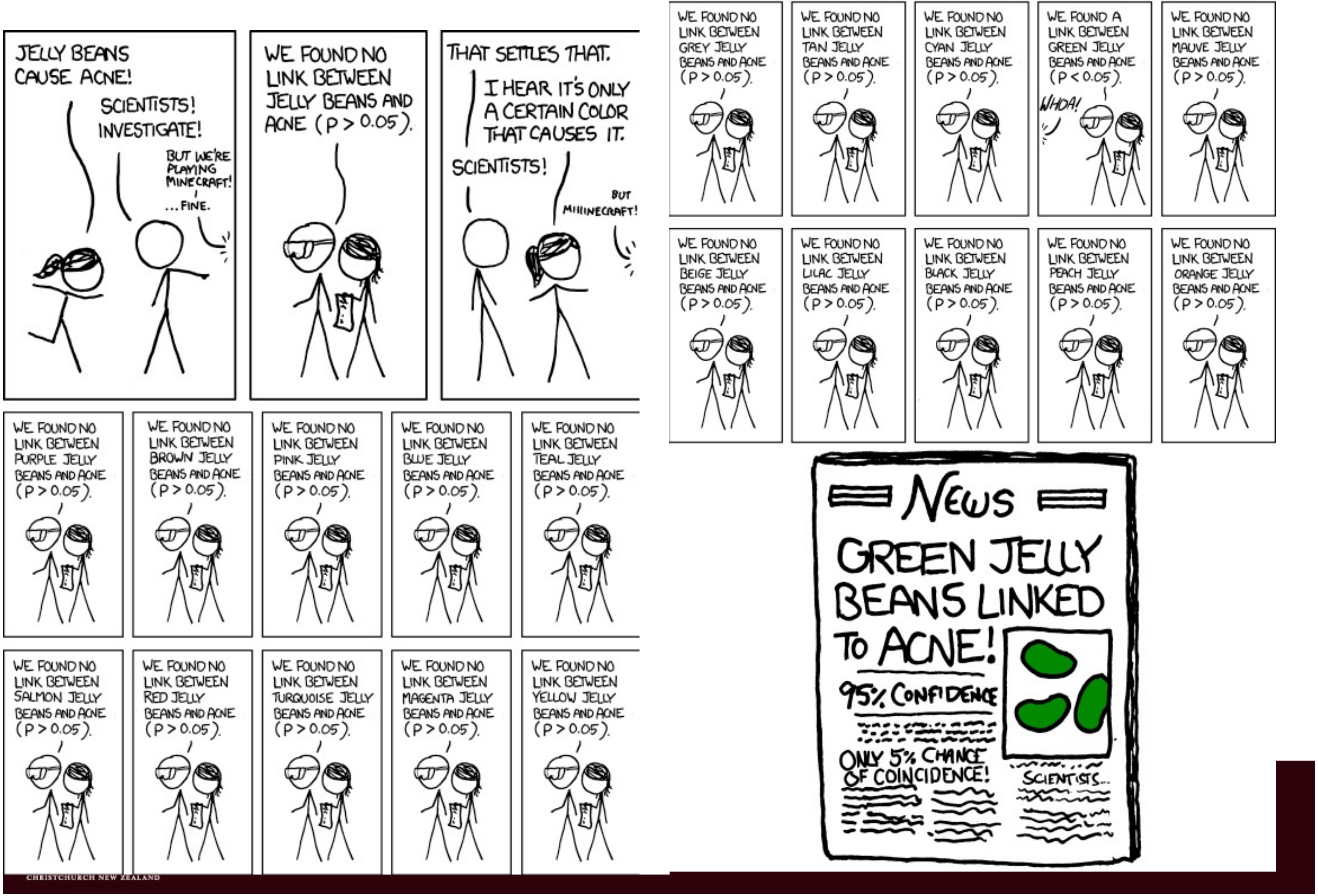
Comparative experiments (most)

- Null Hypothesis Significance Testing (NHST): widely used set of techniques for dichotomous testing
- Test the null hypothesis (H_0) of no difference $H_0: \mu_1=\mu_2$
- Reject H_0 when $p < \alpha$ (α is normally .05)
  - *p:* Assuming the null hypothesis is true, how likely (*p*) is it that we'd observe data at least as extreme as our sample?
  - $P(D|H_0) < .05$
- Failure to reject does not mean "they are the same"
  - Perhaps they are the same
  - OR your experiment wasn't good enough
- So, reject or <u>fail to reject</u> (*not* reject or accept)

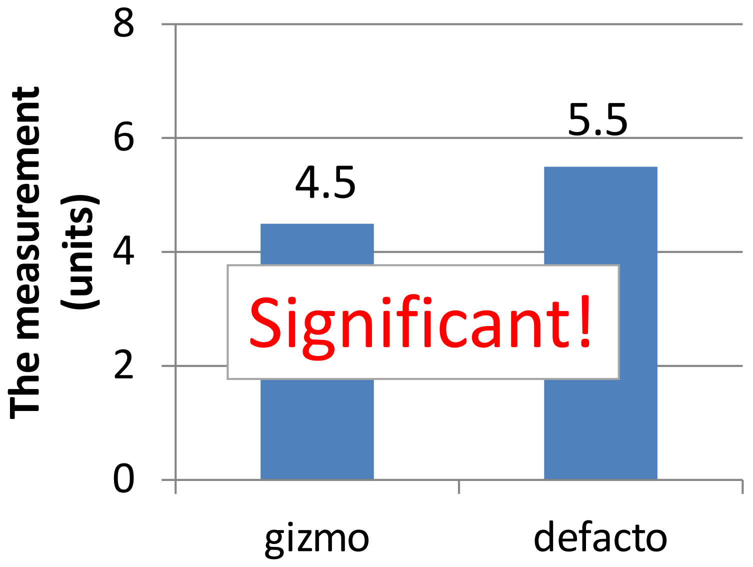
(Aside… The ‘file drawer’ effect)

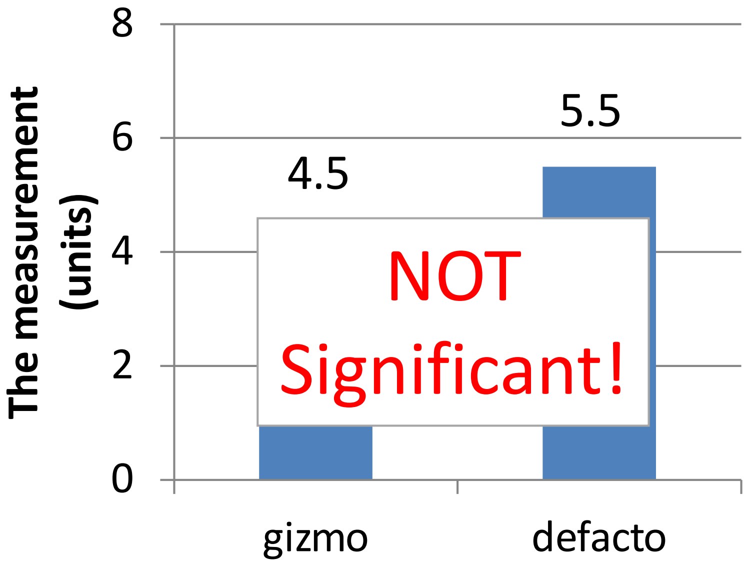
- ‘Unsuccessful’ experiments, which fail to reject the null hypothesis, tend to go unpublished
- They go into ‘the file drawer’
- But statistics are about chance; .05 means 1 in 20 chance of erroneously claiming a difference
- E.g., 19 studies correctly claiming “no significant effect” go in the file drawer; 1 incorrectly claiming a “significant effect” gets famous

(http://xkcd.com/1478/, Randall Munroe, Creative Commons Attribution-NonCommercial 2.5 License)


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

e.g., Gizmo versus de facto

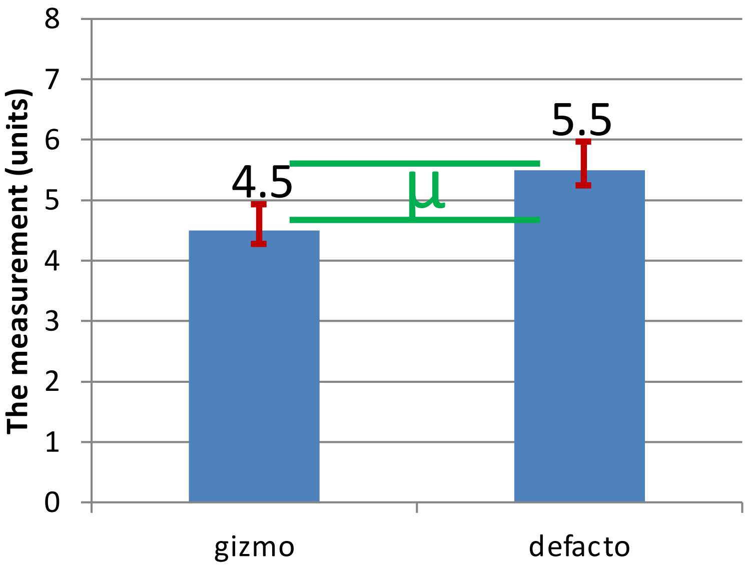




Statistics: Signal to Noise analogy

- Signal: magnitude of the difference
- Noise: random variation

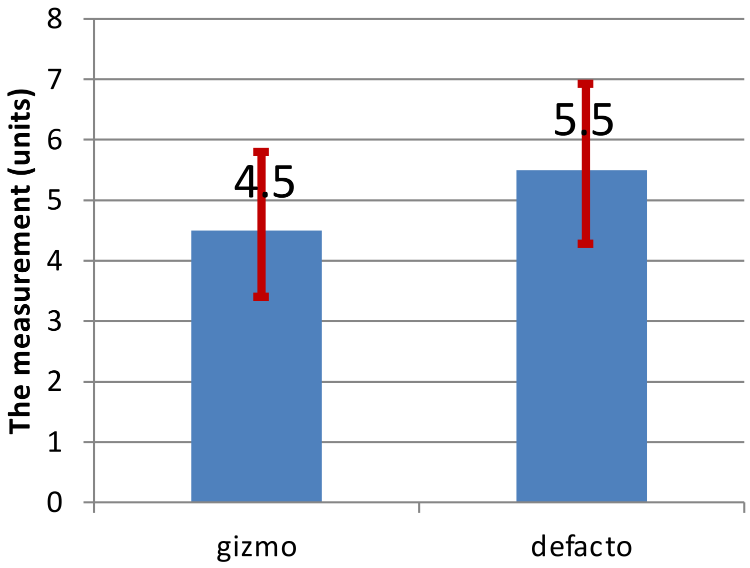
$$\frac{\hat{\mu}}{\hat{\sigma}/\sqrt{n}}$$



Significant!

Statistics: Signal to Noise analogy

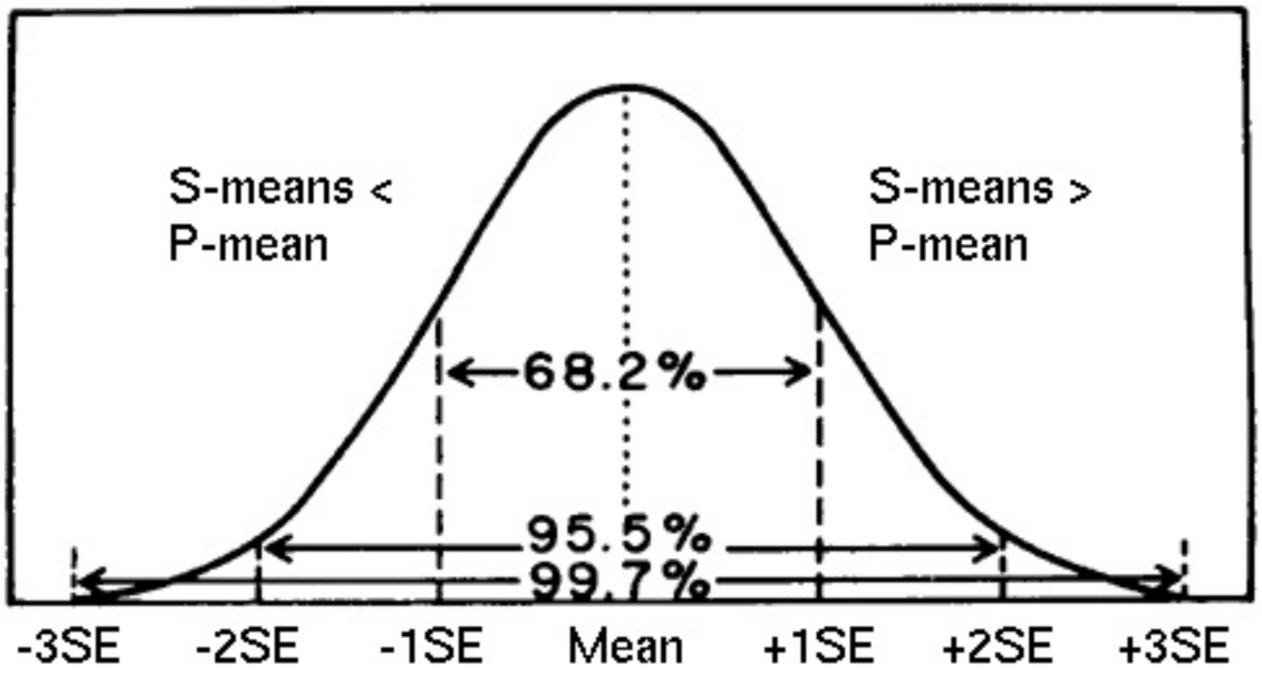
- Signal: magnitude of the difference
- Noise: random variation



NOT Significant!

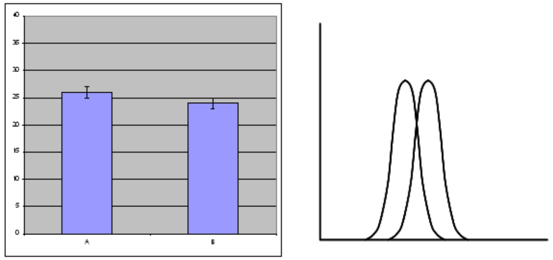
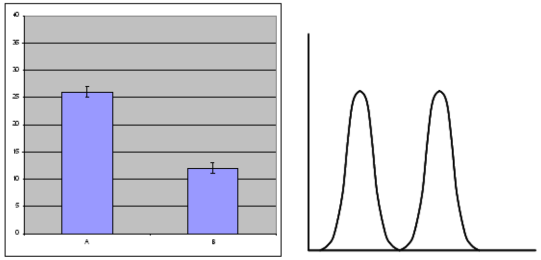
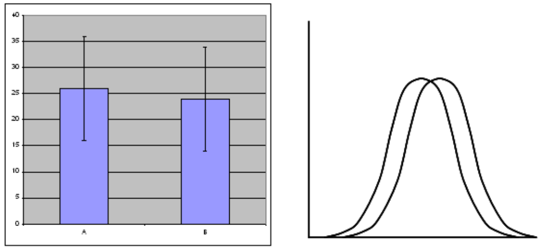
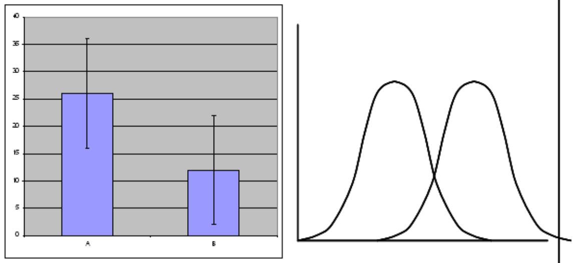
$$\frac{\hat{\mu}}{\hat{\sigma}/\sqrt{n}}$$

Parametric Statistics



$$\frac{\hat{\mu}}{\hat{\sigma}/\sqrt{n}}$$

Parametric Statistics

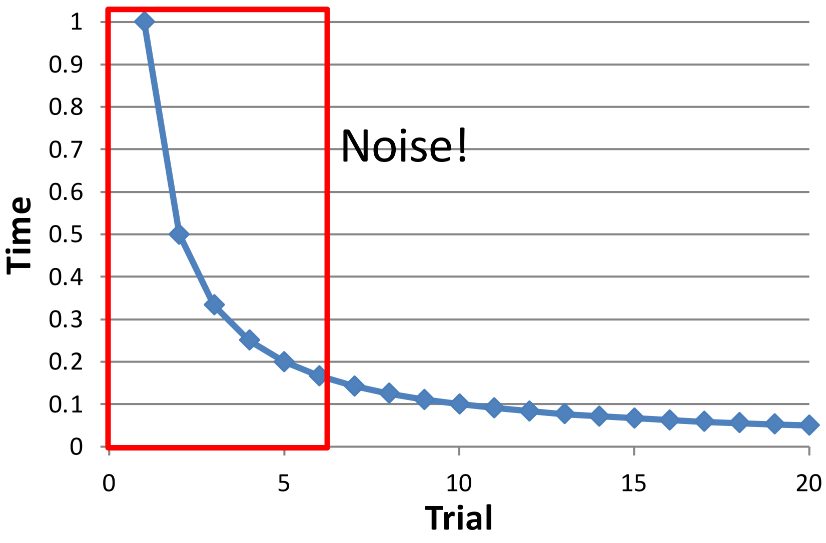
		Signal	
		Small	Large
Noise	Small	 e.g. mean max flight speeds of 747s and A380s	 e.g. mean max flight speeds of 747s and Cessnas
	Large	 e.g. mean max speeds of all cars and motorbikes	 e.g. mean max speeds of all cars and bicycles

$$\frac{\hat{\mu}}{\hat{\sigma}/\sqrt{n}}$$

Reducing the denominator

$$\frac{\hat{\mu}}{\hat{\sigma}/\sqrt{n}}$$

- Reduce *σ*
  - Better training
  - Outlier removal
  - Log transformation
- Increase *n*
  - Easy, but diminishing returns



Power law of practice

Type I and Type II Errors

		In Reality	
		H_0 true (No difference)	H_0 false (Different)
We Conclude	Reject H_0	Type I error False positive Falsely claim a difference Protected via confidence (α)	Correct decision True positive
	Do not reject H_0	Correct decision True negative	Type II error False negative Fail to identify difference Protected via power (1-β)

Can use confidence level (α) to change probability of Type I and II errors

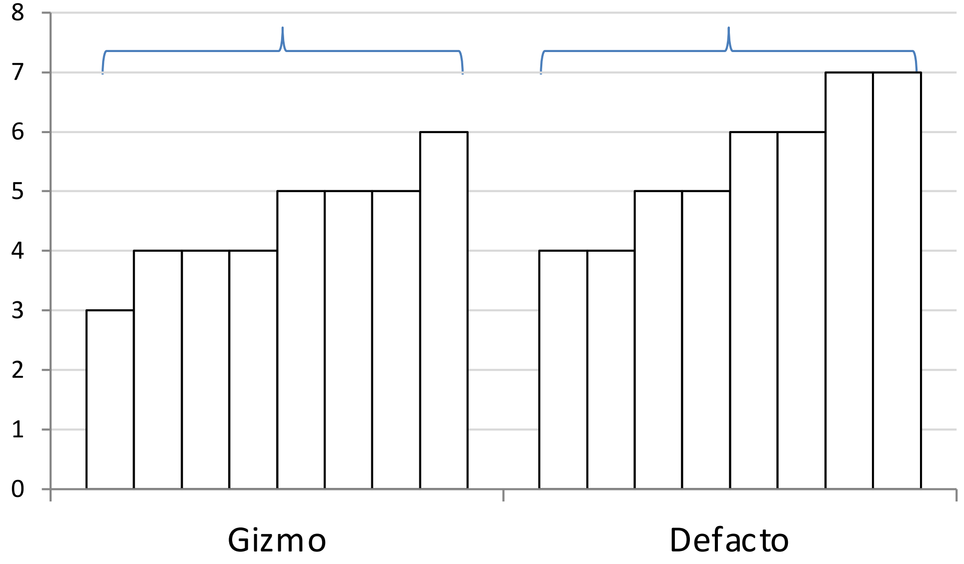
R

- https://www.r-project.org/
- Free, GNU general public license
- Trusted
- Advanced; entire language
- Lots of packages
- Great graphics facilities

- Good books: e.g., “R in Action” by Kabacoff
- Lots of online tutorials and resources (use them!)

For example:
Gizmo versus de facto

Mean 4.5 Mean 5.5



```
> more ttest-data.txt
gizmo    defacto
3        4
4        4
4        5
4        5
5        6
5        6
5        7
6        7
```

8 data points for each condition

T-Tests

- Are two samples from different populations?
- Paired T-Test (≡within subjects)
  - E.g., participants 1-8 use Gizmo *and* de facto
  - Each participant’s data is paired
- Unpaired T-Test (≡between subjects)
  - E.g., participants 1-8 use Gizmo, and 9-16 de facto
  - Independent samples

Unpaired T-test: R

```
> more ttest-eg-unpaired.R
#!/usr/bin/env Rscript

data <- read.table("ttest-data.txt", header=TRUE)
t.test(data$gizmo, data$defacto)

> ./ttest-eg-unpaired.R

        Welch Two Sample t-test

data:  gizmo and defacto
t = -1.8708, df = 13.176, p-value = 0.08374
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.1531955  0.1531955
sample estimates:
mean of x mean of y
      4.5       5.5
```

T-ratio (signal to noise)
Absolute value: bigger is better

Degrees of freedom (scale of the experiment)

Likelihood of observing this data
(or more extreme) if the null H were true.
Only reject null hypothesis if p < .05

"…. no significant difference between mean task time with Gizmo (4.5 s, sd 0.9) and de facto (5.5 s, sd 1.2): $T_{13.2}$ = 1.87, p = .08."

Paired T-test: R

```
> more ttest-eg-paired.R
#!/usr/bin/env Rscript

data <- read.table("ttest-data.txt", header=TRUE)
t.test(data$gizmo, data$defacto, paired=TRUE)

> ./ttest-eg-paired.R


        Paired t-test

data:  data$gizmo and data$defacto
t = -5.2915, df = 7, p-value = 0.001134
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.446872 -0.553128
sample estimates:
mean of the differences
                     -1
```

T-ratio

Degrees of freedom
#Pairs -1

"…. significant difference between mean task time with Gizmo (4.5 s, sd 0.9) and de facto (5.5 s, sd 1.2): $T_7 = 5.29$, p =.001."

Why significant only when paired?

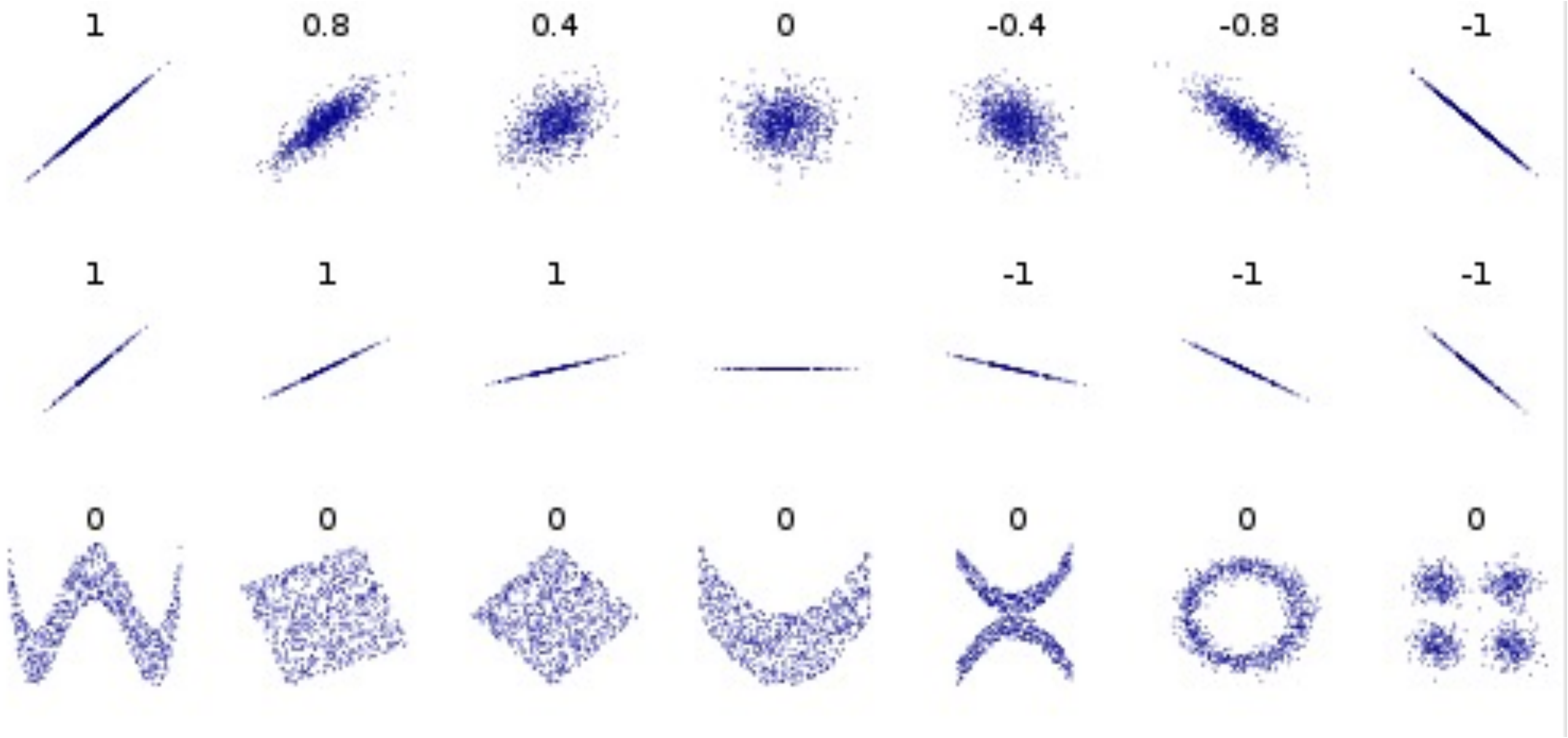
```
> more ttest-data.txt
gizmo   defacto
3       4
4       4
4       5
4       5
5       6
5       6
5       7
6       7
```

- Lots of extra information through pairing
- Col 1 (Gizmo) < Col 2 (de facto) for all but one
- Within subjects designs: participants act as their own control
- Increases experimental sensitivity

Correlation: Relating datasets

- Strength of relationship between variables
  e.g., typing and menu selection speeds
- Various models possible: linear, power, exponential, logistic…
  - Always, eyeball the data for conformance with the model
- For linear correlation, Pearson’s r:
  - Correlation coefficient -1 to 1
  - Both variables are continuous
  - Cohen: 0.1 – 0.3 ‘small’, 0.3 – 0.5 ‘med’; 0.5 – 1.0 ‘large’
- Spearman’s *rho* for ranked data
- Correlation is *not* causation

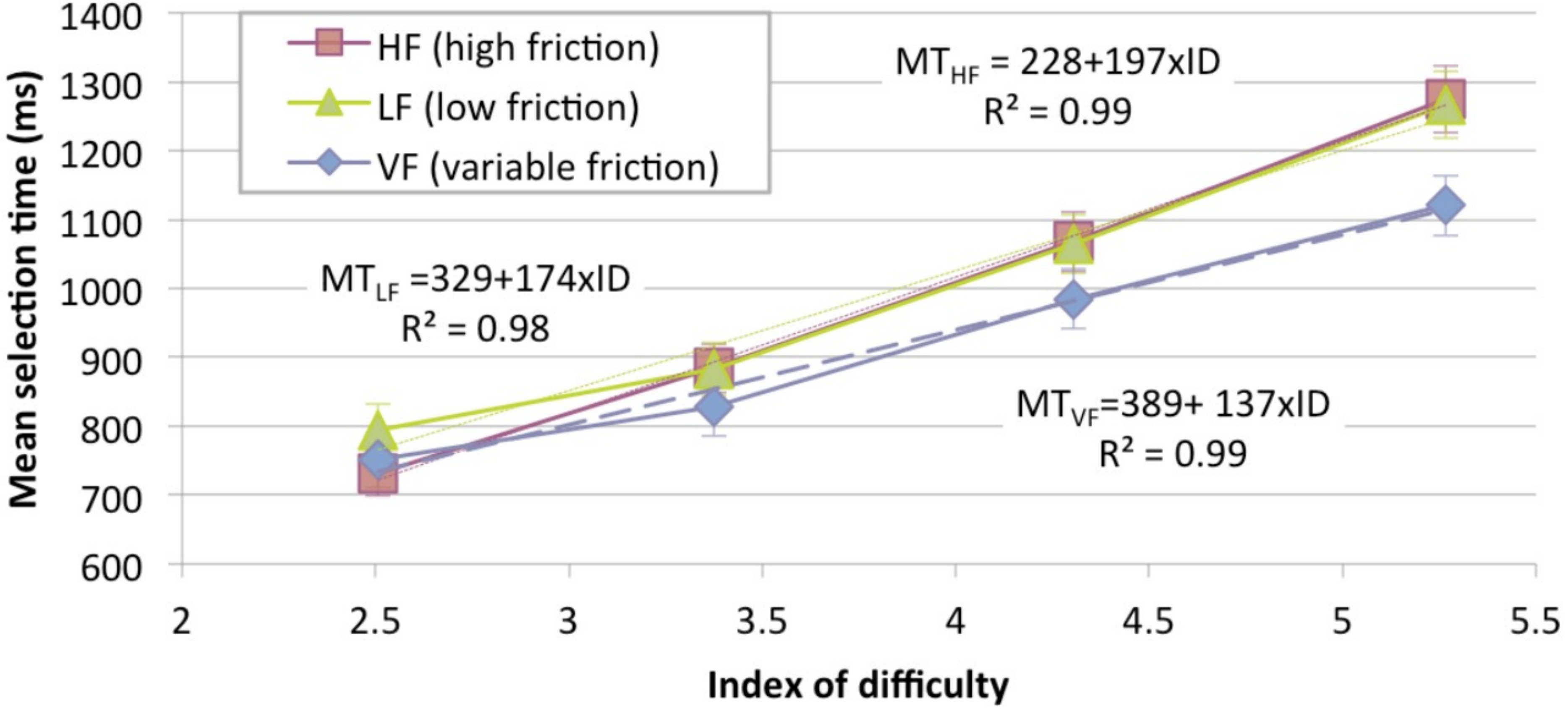
Correlation: Relating datasets



Regression: Relating datasets

- Predicting one value from another
  e.g., calculating pointing time from target distance/width
- Line of best fit
- R^2:
  - Coefficient of determination: 0 to 1
  - Proportion of the variability explained by the model
  - > 0.8 is good for human performance
- Fitts' Law: expect $R^2 > 0.95$
- Easy with Excel's 'Add Trendline'

Regression: Relating datasets



Analysis of Variance (ANOVA) Main statistical workhorse

- Independent variable with more than two levels
  e.g., *Friction type* $\in$ *{high, low, variable}*
- Compare all pairs with T-Tests?
  - # comparisons for *n* levels = $(n^2-n)/2$
  - Increased likelihood of finding a difference by chance (Type 1 error)
- ANOVA: are all conditions from the same population? H_0: $\mu_1=\mu_2=\ldots=\mu_n$

Analysis of Variance (ANOVA)

- Independent variables now called ‘factors’
  - One factor → ‘one way ANOVA’
  - More than one... (COSC411)
  - Each factor *either* within *or* between subjects

Within ✓

Friction type	Low	High	Variable
	S1-15	S1-15	S1-15

Between ✓

Friction type	Low	High	Variable
	S1-15	S16-30	S31-45

Messed up ✗

Friction type	Low	High	Variable
	S1-15	S16-30	S16-30

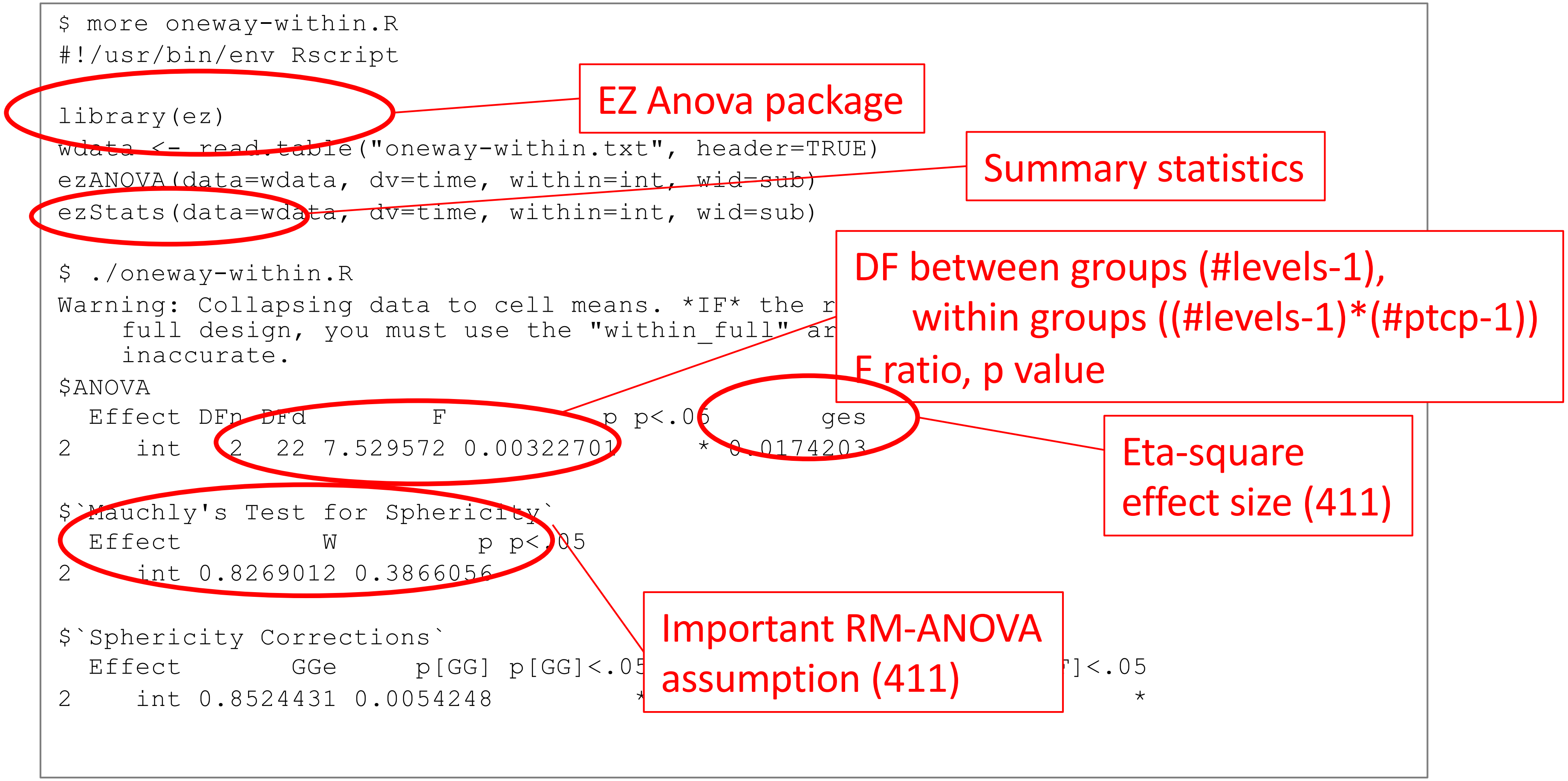
Analysis of Variance (ANOVA) e.g., one way within subjects

- Data file:
  - One datum per line (usually a trial)
  - At least one trial for every participant in every cell
  - Several trials are fine (replicated trials; averaged)
  - First column: participant identifier
  - Second column: level of the factor
  - Third column: dependent measure

Analysis of Variance (ANOVA)
e.g., one way within subjects

```
>  cat oneway-within.txt
sub int time
S1 HF  0.456
S1 LF  1.224
S1 VF  0.775
S1 VF  0.655
S2 VF  1.445
S2 VF  1.224
S2 LF  0.788
S2 HF  1.334
S3 HF  0.443
S3 LF  0.786
...
```

Analysis of Variance (ANOVA) e.g., one way within subjects with R



```
$ more oneway-within.R
#!/usr/bin/env Rscript

library(ez)
wdata <- read.table("oneway-within.txt", header=TRUE)
ezANOVA(data=wdata, dv=time, within=int, wid=sub)
ezStats(data=wdata, dv=time, within=int, wid=sub)

$ ./oneway-within.R
Warning: Collapsing data to cell means. *IF* the r
    full design, you must use the "within_full" ar
    inaccurate.
$ANOVA
  Effect DFn DFd        F          p p<.05       ges
2    int   2  22 7.529572 0.00322701     * 0.0174203

$`Mauchly's Test for Sphericity`
  Effect         W         p p<.05
2    int 0.8269012 0.3866056

$`Sphericity Corrections`
  Effect       GGe      p[GG] p[GG]<.05                       ]<.05
2    int 0.8524431 0.0054248         *                             *
```

EZ Anova package

Summary statistics

DF between groups (#levels-1),
within groups ((#levels-1)*(#ptcp-1))
F ratio, p value

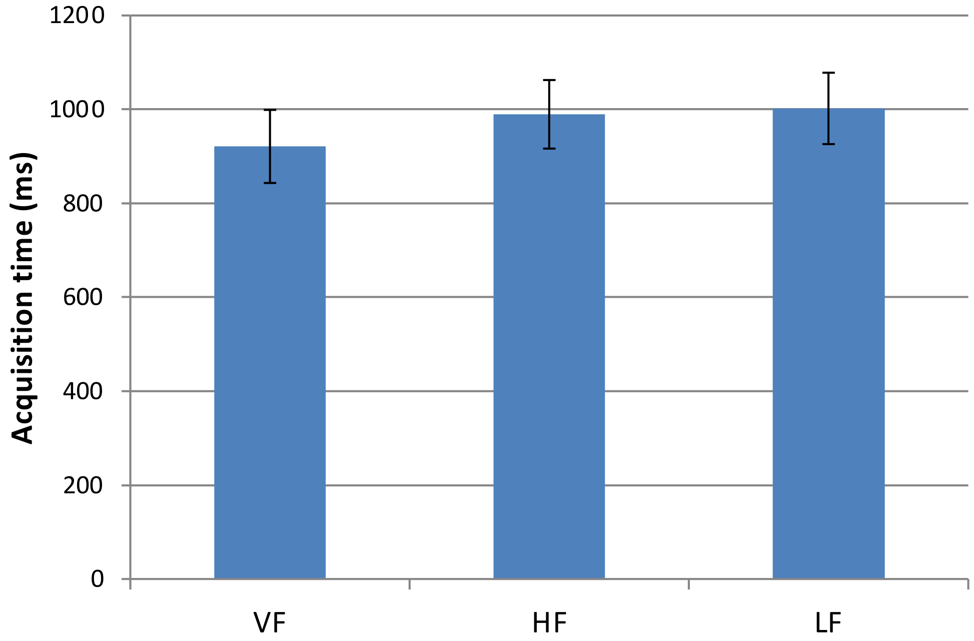
Eta-square
effect size (411)

Important RM-ANOVA
assumption (411)

Analysis of Variance (ANOVA)
e.g., one way within subjects with R

```
Warning: Collapsing data to cell means. *IF* the requested effects are a subset of the
    full design, you must use the "within_full" argument, else results may be
    inaccurate.
  int  N      Mean       SD     FLSD
1  HF 12  982.7292 255.4934 43.56331
2  LF 12 1002.2917 263.4246 43.56331
3  VF 12  923.9792 264.1256 43.56331
```

Analysis of Variance (ANOVA) e.g., one way within subjects



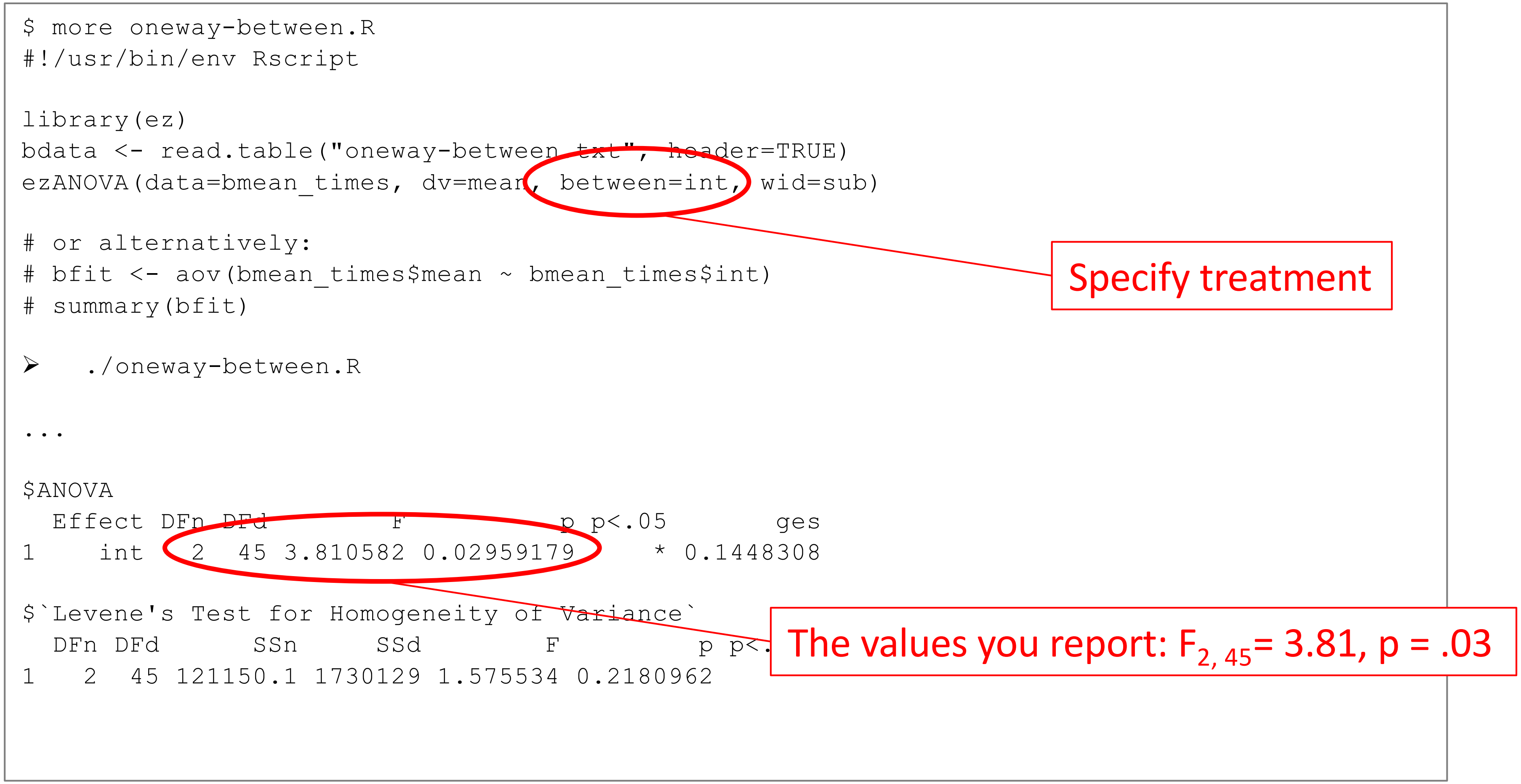
"…. significant difference between mean acquisition times, with VF fastest (924ms, sd 264) followed by HF (983ms, sd 255) and LF (1002ms, sd 263): $F_{2,22} = 7.53$, p = .003."

Analysis of Variance (ANOVA)
e.g., one way *between* subjects

- Exactly same procedure, except:
- Data file:
  - Each participant has one or more trial datum for exactly one level

```
>  cat onewaybetweeneg.txt
sub cond  time
S1  VF    0.775
S1  VF    0.655
S2  HF    1.445
S2  HF    1.224
S3  LF    1.455
S4  LF    1.25
S5  VF    1.444
S6  HF    1.222
...
```

Analysis of Variance (ANOVA)
e.g., one way *between* subjects with R

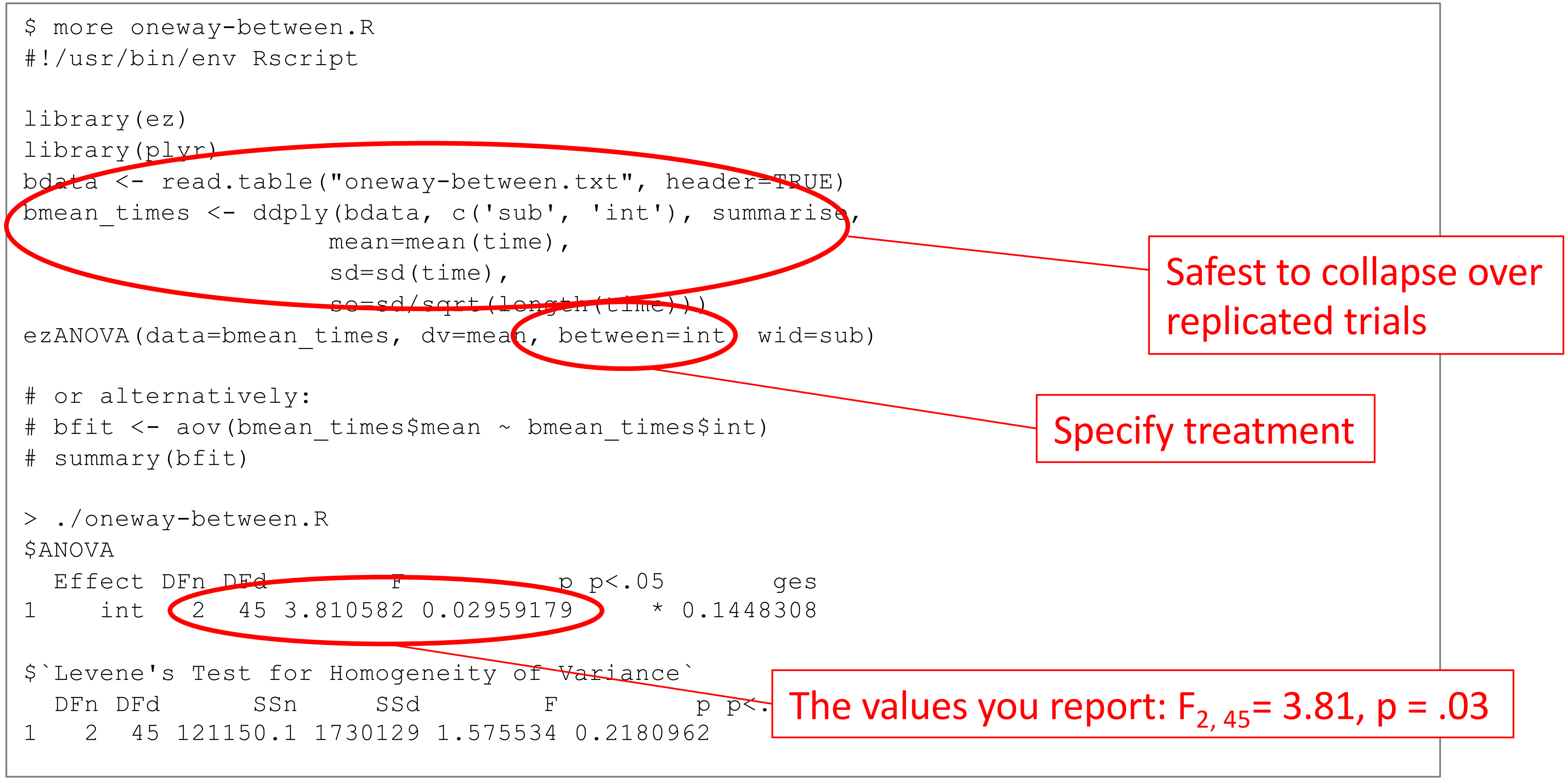


```
$ more oneway-between.R
#!/usr/bin/env Rscript

library(ez)
bdata <- read.table("oneway-between.txt", header=TRUE)
ezANOVA(data=bmean_times, dv=mean, between=int, wid=sub)

# or alternatively:
# bfit <- aov(bmean_times$mean ~ bmean_times$int)
# summary(bfit)

➢   ./oneway-between.R

...

$ANOVA
  Effect DFn DFd        F          p p<.05       ges
1    int   2  45 3.810582 0.02959179     * 0.1448308

$`Levene's Test for Homogeneity of Variance`
  DFn DFd      SSn     SSd        F         p p<.
1   2  45 121150.1 1730129 1.575534 0.2180962
```

Specify treatment

The values you report: $F_{2, 45}$= 3.81, p = .03

Analysis of Variance (ANOVA)
e.g., one way *between* subjects with R



```
$ more oneway-between.R
#!/usr/bin/env Rscript

library(ez)
library(plyr)
bdata <- read.table("oneway-between.txt", header=TRUE)
bmean_times <- ddply(bdata, c('sub', 'int'), summarise,
                     mean=mean(time),
                     sd=sd(time),
                     se=sd/sqrt(length(time)))
ezANOVA(data=bmean_times, dv=mean, between=int, wid=sub)

# or alternatively:
# bfit <- aov(bmean_times$mean ~ bmean_times$int)
# summary(bfit)

> ./oneway-between.R
$ANOVA
  Effect DFn DFd        F          p p<.05       ges
1    int   2  45 3.810582 0.02959179     * 0.1448308

$`Levene's Test for Homogeneity of Variance`
  DFn DFd      SSn     SSd        F         p p<.05
1   2  45 121150.1 1730129 1.575534 0.2180962
```

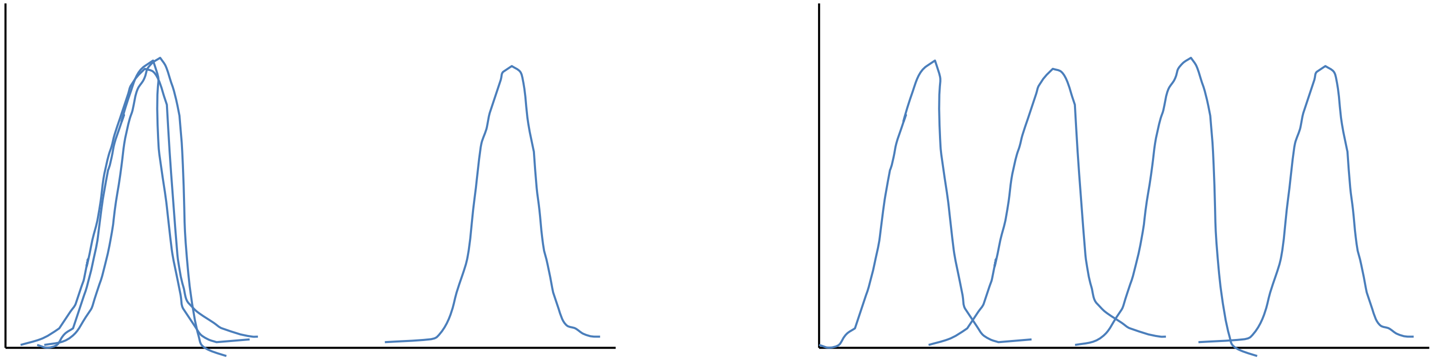
Safest to collapse over replicated trials

Specify treatment

The values you report: $F_{2, 45}$= 3.81, p = .03

What does this tell us?

- We can reject H_0: i.e., we should rarely observe data as extreme as our sample if $\mu_1=\mu_2=\ldots=\mu_n$



- What’s different to what?

Posthoc comparisons

- When the ‘main effect’ for a factor is significant we can do ‘posthoc’ pairwise comparisons
- They are conservative (reducing type 1 errors)
  - Not uncommon to reject H_0, but find no significant posthoc differences
- Bonferroni correction, Tukey test, …
- (Beyond us for now)

Subjective responses

- User opinions can amplify raw data
- Likert scales (levels of agreement)
- NASA-TLX (in Hancock and Meshkati 1998)
- Rankings
- Preference counts
- User comments

Non-continuous measures
Non-parametric analysis

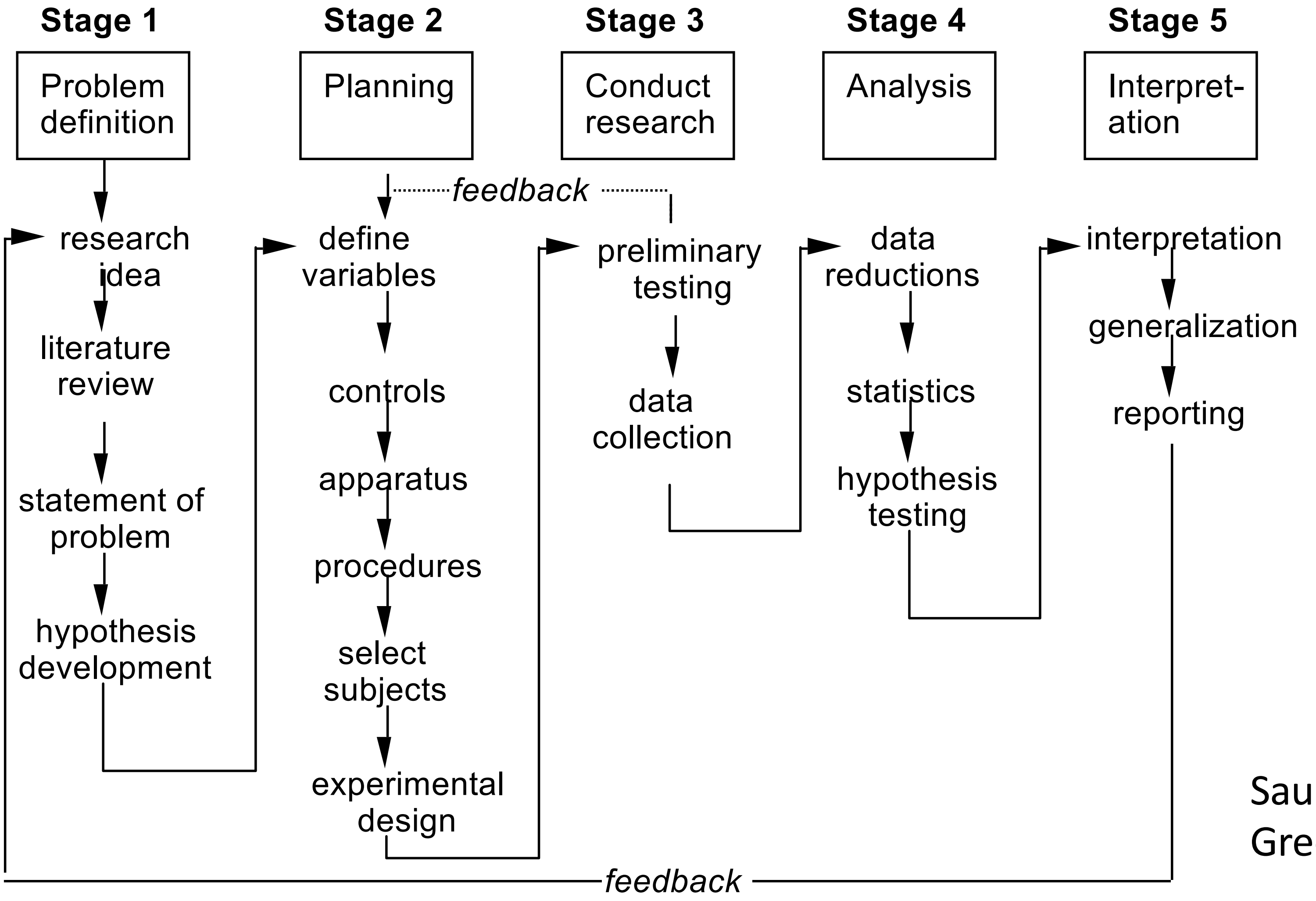
Non-parametric statistics
(In one slide)

- Likert scale responses, ranks, etc.

	Within subjects	Between subjects
2 levels	Wilcoxon Signed Ranks Test	Mann-Whitney U Test
> 2 levels	Friedman Test	Kruskal-Wallace H Test

- Frequencies and proportions
  - Chi-square test
  - Independence of samples (one datum/ptcpt)

Planning Experiments!



Saul Greenberg

What’s in the exam?

- Look at past papers.

- Solutions please?
  - Nope; come see me with your attempts

Thanks All!

- Acknowledgements:
  - Saul Greenberg, Scott MacKenzie, Alan Dix