P_hacking
and replication crisis

20

Probability and Statistics

COMS10011
Luluah Albarrak

Prepared by Dr. Anne Roudaut

ego depletion story

1998

ROY Baumeister

do people have a **limited amount** of **will power**?

1998

so he conducted a study …

the results show that humans **do have a limited pool of self-control**

once we have had to **resist temptation** it is a lot **harder to do it again**

1998

they called this

Ego Depletion::

has had huge influence on psychological research, been incorporated into dieting tactics, training techniques, and even advertisments used today

ads telling how we deserve a product, causing mental fatigue and frustration, leading us to buy

but today it seems that **this phenomenon does not exist at all**

1998

ROY Baumeister's hypothesis:

self-control is **a limited resource**, and it takes **energy and motivation to maintain self restraint**

every time use your self-control, your draw on that strength and it takes some time for you to recover it

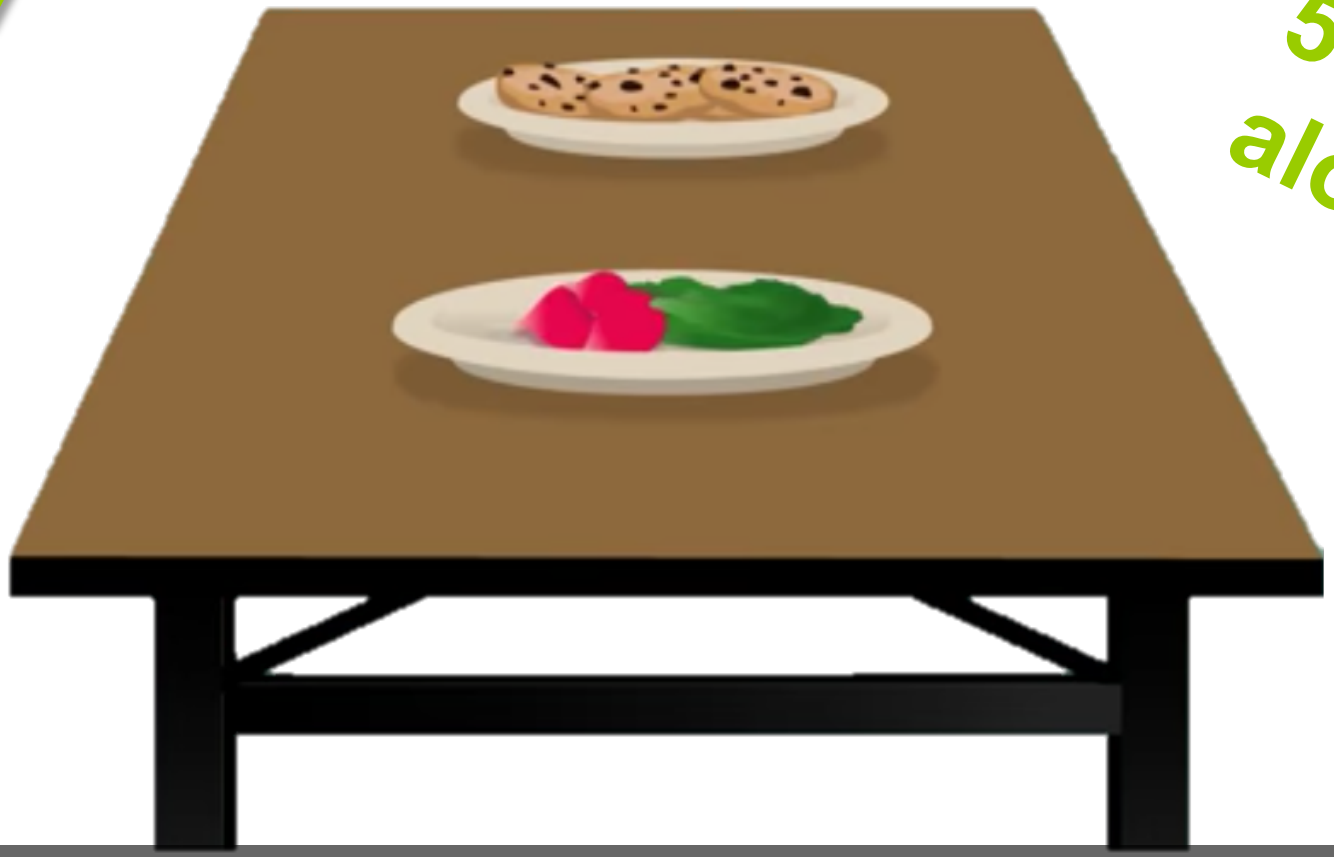
1998

ROY Baumeister’s task:

two acts of self-control back-to-back

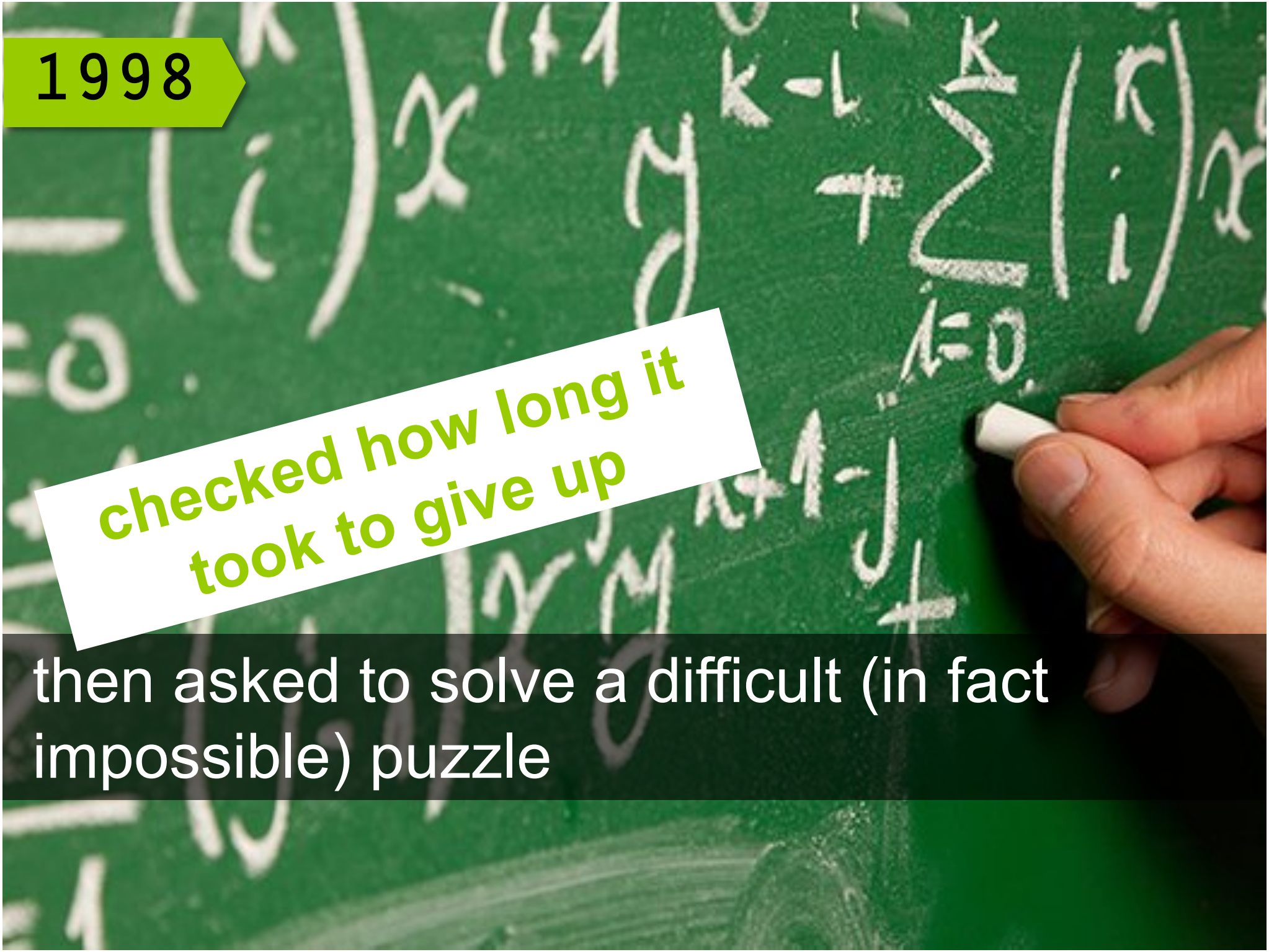

1998
cooked delicious cookies and invited 67 students with empty stomach to sit in the baking room

1998



5 min
alone

half the students told to **eat the cookies**
half the students **the radishes**


1998
checked how long it took to give up
then asked to solve a difficult (in fact impossible) puzzle

1998

would resisting the cookies make it harder for participants to keep trying?

those who ate **radishes** = **8 minutes** of trying
those who ate **cookies** = **19 minutes** of trying

<brainstorming: could you already see something wrong with their studies?>

1998

confounding variable::

eating cookies (high in sugar) could be the reason of the longer effort provided

1998

would resisting the cookies make it harder for participants to keep trying?

those who ate **radishes** = **8 minutes** of trying
those who ate **cookies** = **19 minutes** of trying
3rd group with no cookies encounter = 21 minutes

and of course all these comparison with $p<0.05$ = strong evidences for ego depletion (which later became a subfield of psychology with many more studies done to confirm it)

2007

in 2007, researchers figured out what *seemed* to be happening biologically:

as people used up their self-control, their blood sugar levels were dropping

they made subjects watch emotional videos without showing emotions / while others did not have to hold back


2007
subjects who used willpower = lower blood glucose, and when they replenished that glucose it restored their self-control

2010

evidences even more in 2010 when group of researchers led by Martin Hagger, examined **83 published studies** on ego depletion to conclude that **effect was real**

2012

but in 2012 researchers **casted some doubts**

e.g. subjects did not have to drink lemonade to replenish their will power, **tasting it** was enough

e.g. subjects who **believe in willpower** could affect their performance

2014

in 2014, researchers tried to replicate the original studies and **could not find the effect**

they also looked at the meta analysis of 2010 (the 83 papers) and found a lot of issues, e.g. they re run the analysis with newer methods … and the **ego depletion effect disappeared**

this started a wave a concern about **replicability** hitting the field of psychology

2016

in 2016, the Association for Psychological Science opened a Registered Replication Report on ego depletion:

one official experiment would be conducted by researchers in many different labs

Baumeister (original study) even helped design the experiment and Martin Hagger (meta analysis) led the project

A Multilab Preregistered Replication of the Ego-Depletion Effect

2016

24 labs (different language and culture)

only **2 found the effect**

and **1 group found the opposite effect**

what does it mean?

could argue that it is just this particular task where ego depletion does not show up but that willpower is still a limited resource

or may be ego depletion only happens under very specific circumstances … if at all

this has created a lot of **movements in the scientific community** and a lot of researches are working on it to find ways to better analyze and report studies

p hacking

p crisis

replication crisis

(veritasium channel)


(veritasium channel)

mmm does it mean that all I have learned in this lecture is wrong?

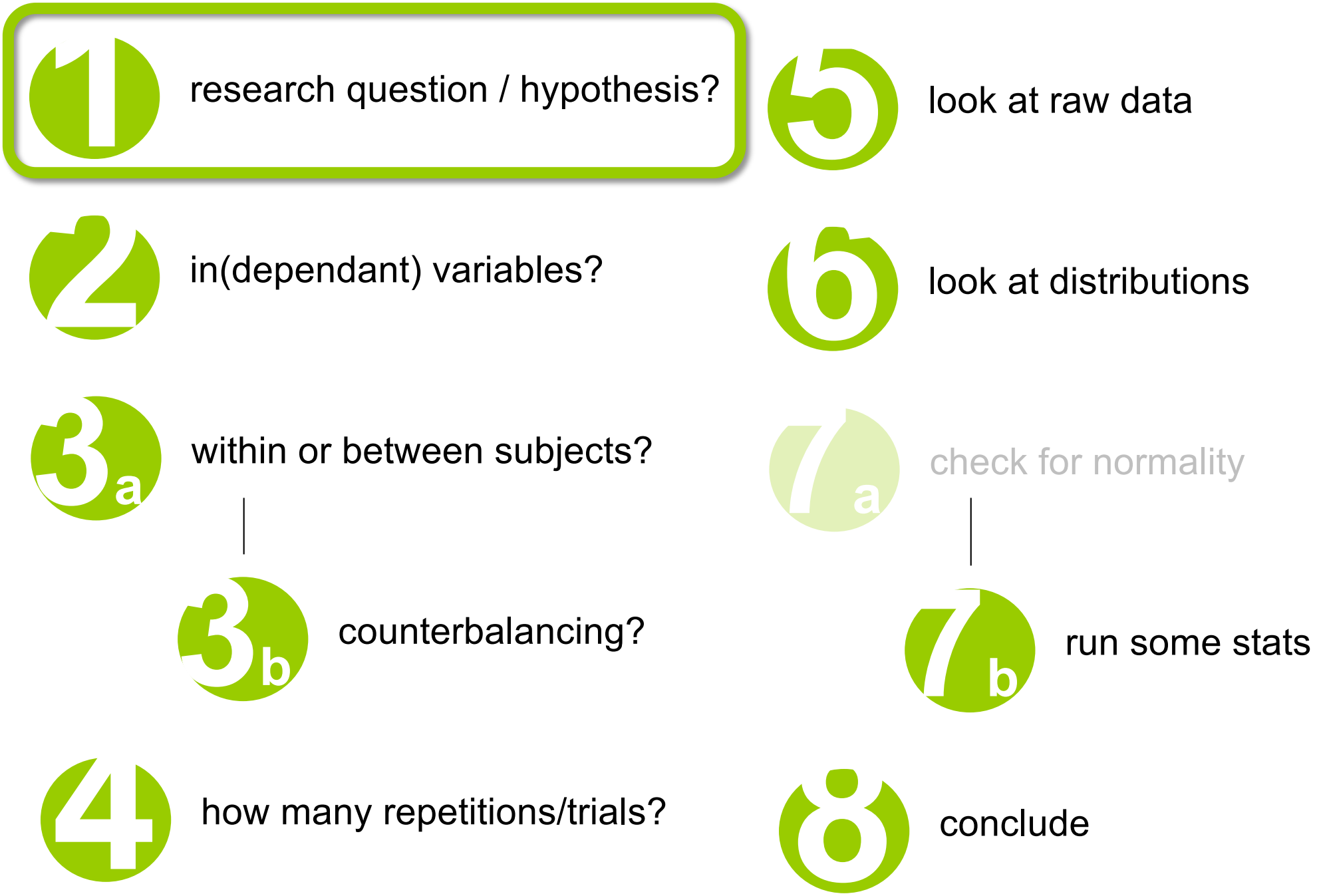
no fortunately there are a many things you can do to analyse stats and report data correctly as well as minimise these problems …

good
statistics


1
design a study properly and don’t go fishing!

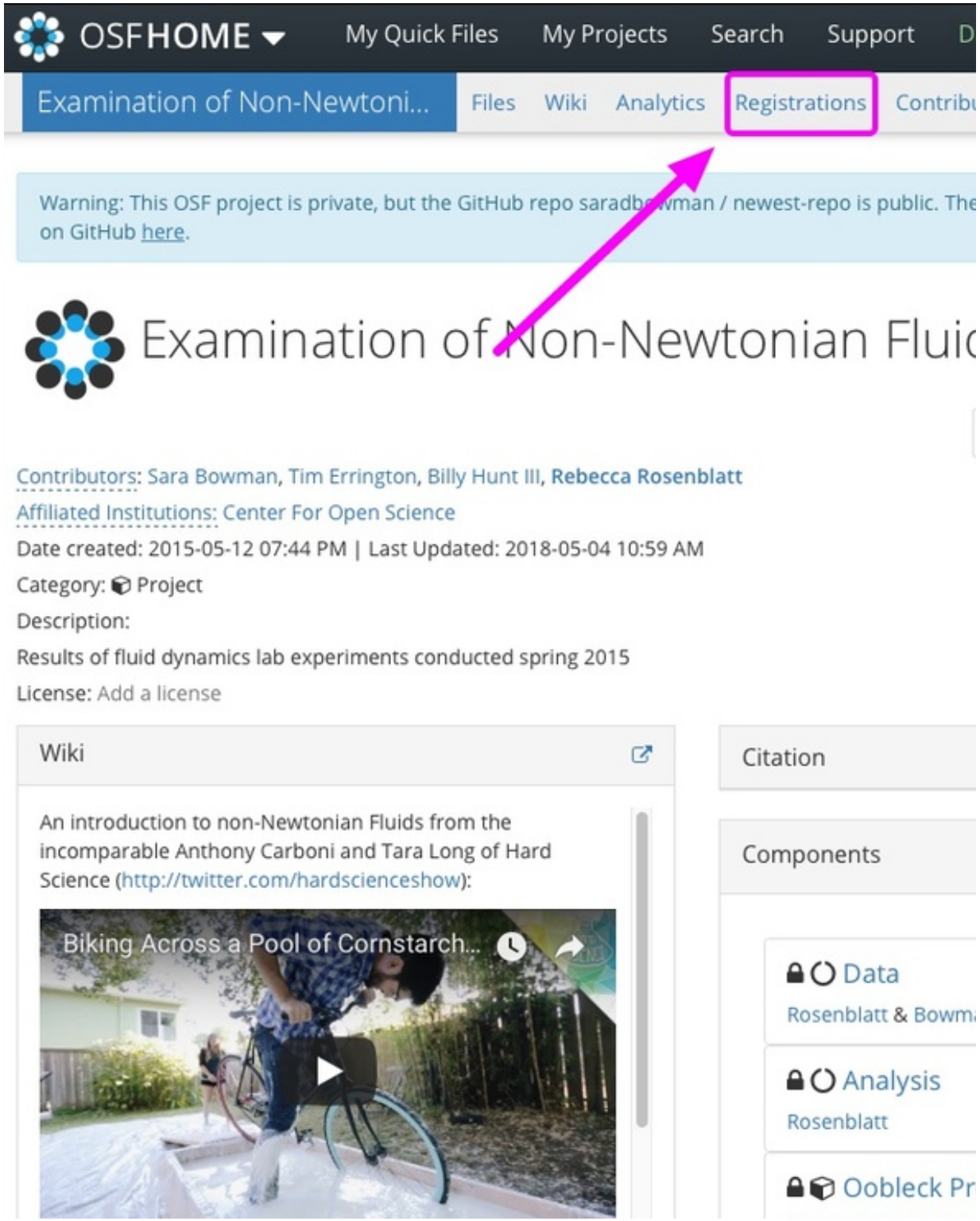
fishing = gathering as many data as you can, then try to find something statistically significant in it and report it = **NO**

rather have clear hypothesis to start with (remember one hypothesis = logical sentence that can be directly tested) and **just test for your hypothesis**


1 research question / hypothesis?
2 in(dependant) variables?
3a within or between subjects?
3b counterbalancing?
4 how many repetitions/trials?
5 look at raw data
6 look at distributions
7a check for normality
7b run some stats
8 conclude


Richard Feynman (1964)

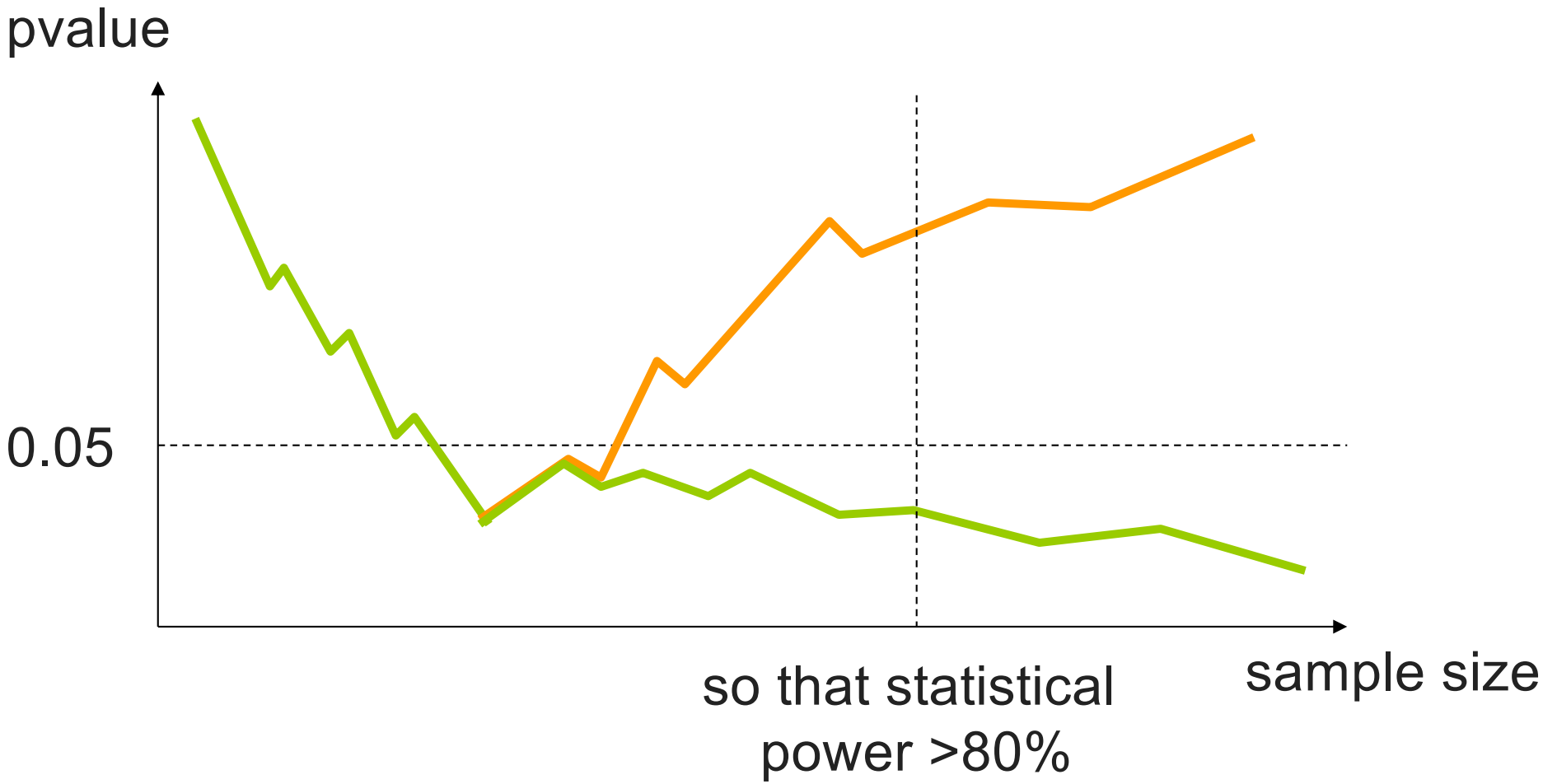
you could also decide to **pre-register your study** (more and more frequent in scientific venues)



and of course design your experiment to remove as much noise as possible

i.e. pilot it!

increase sample size or better compute the required sample size



report p value with effect size

e.g. a new hair loss shampoo is statistically better than existing shampoo

but does not say that subjects who took it only grew 5 hairs more than control group …

effect size matters more!

report confidence intervals and non misleading graphs

confidence intervals::

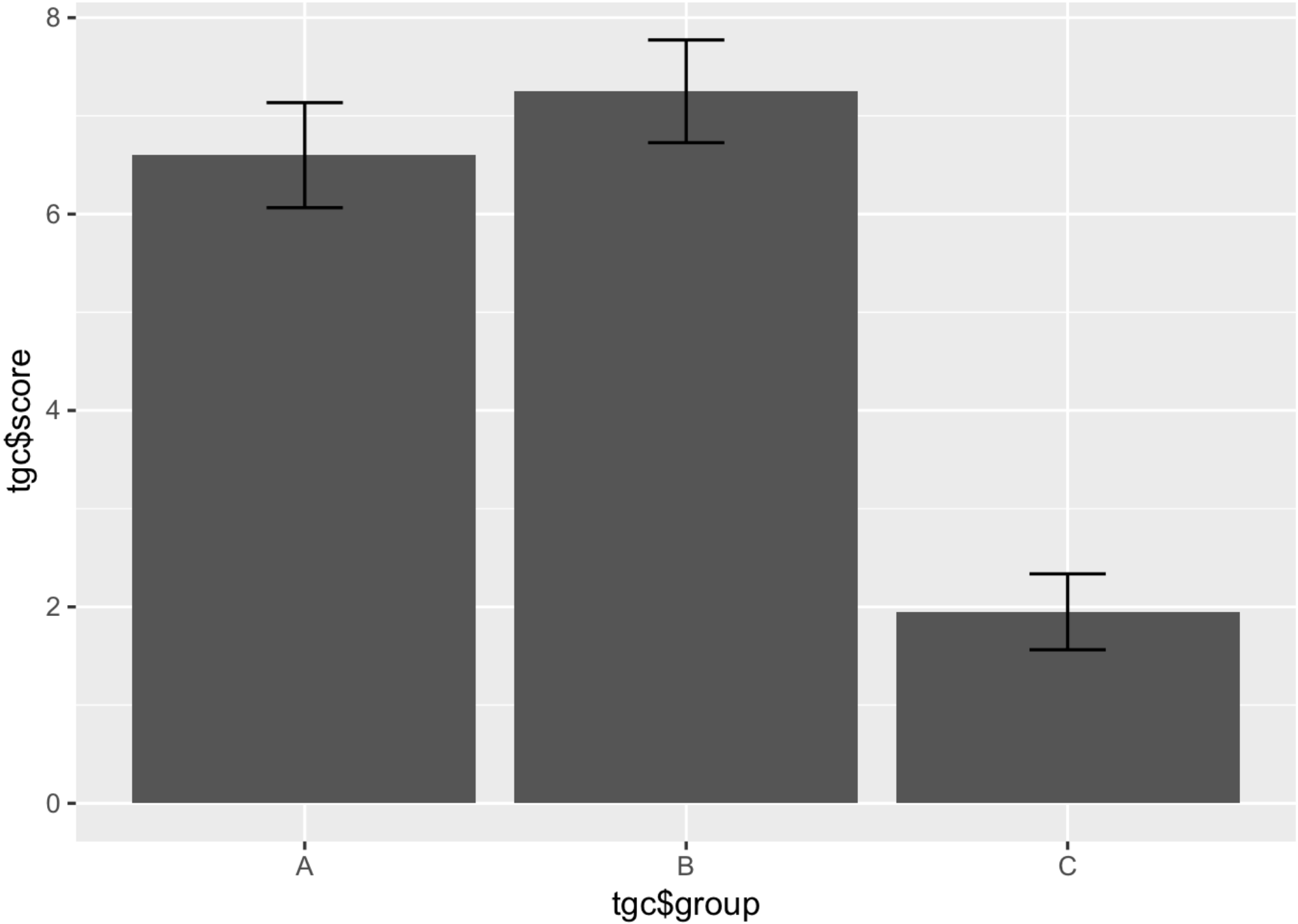
a 95% confidence interval is a range of values that you can be 95% certain contains the true mean of the population

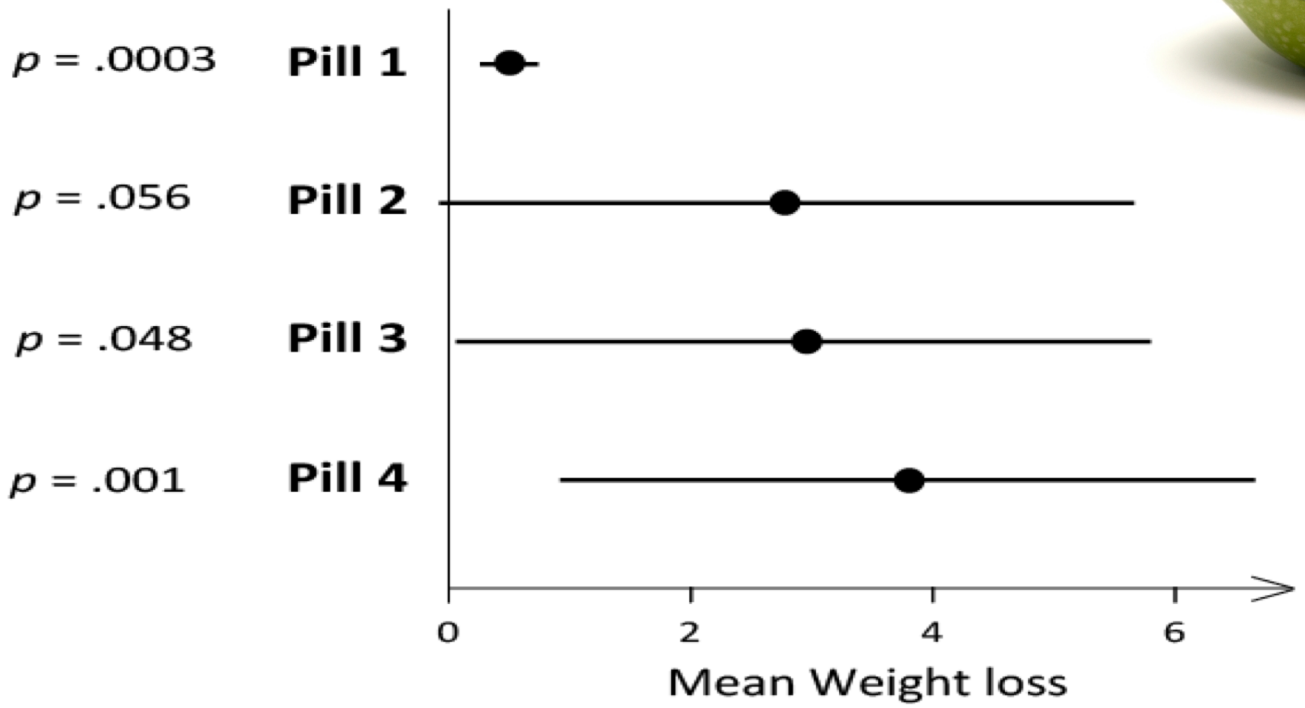
a range of plausible values for the mean (values outside relatively implausible)

```
# first we run the one-way anova
library(Rmisc)
tgc <- summarySE(dat, measurevar="score",
groupvars=c("group"))
tgc
```

```
  group  N score        sd        se        ci
1     A 20  6.60 1.1424811 0.2554665 0.5346976
2     B 20  7.25 1.1180340 0.2500000 0.5232560
3     C 20  1.95 0.8255779 0.1846048 0.3863824
```

```
ggplot(data = tgc, aes(x = tgc$group, y = tgc$score)) +
geom_bar(stat = 'identity', position = 'dodge') +
geom_errorbar(aes(ymin= tgc$score - ci, ymax= tgc$score  +
ci), width=.2, position=position_dodge(.9))
```

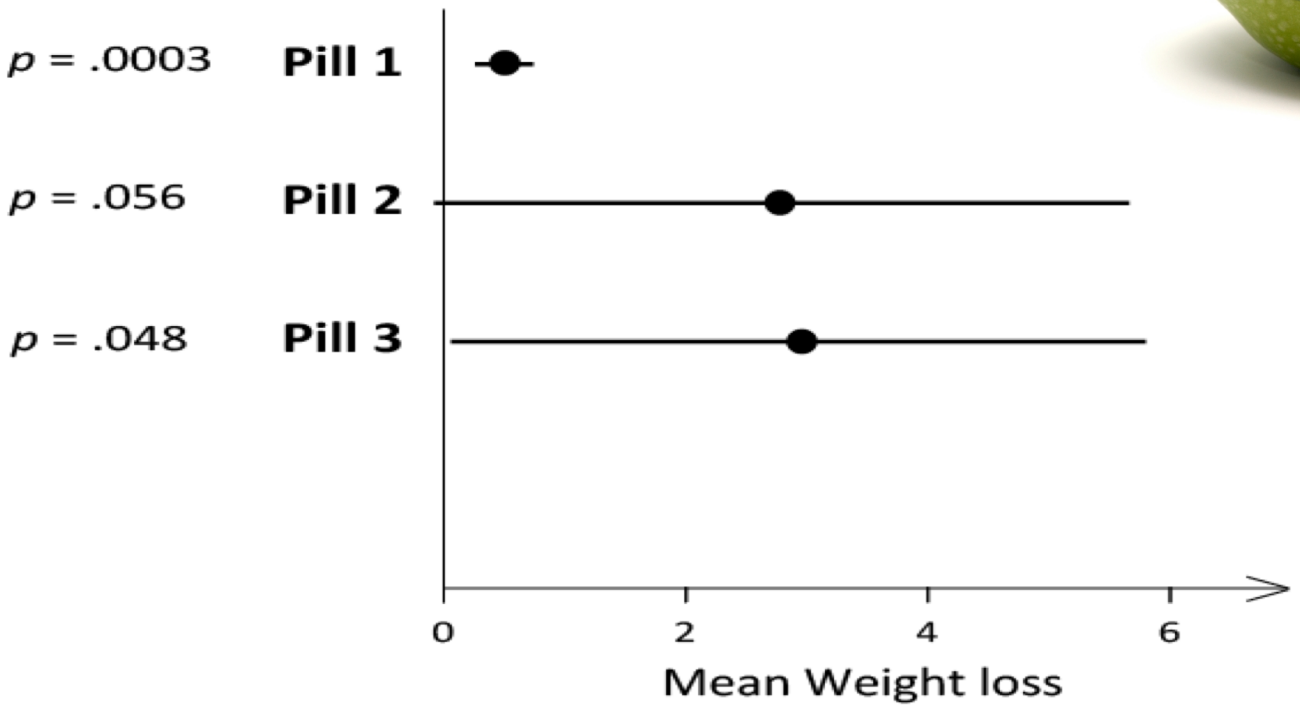

8
6
4
2
0
tgc$score
A
B
C
tgc$group



Error bars are 95% CIs
p-values are based on a null hypothesis of no effect

Adapted from
(Ziliak and McCloskey, 2009)

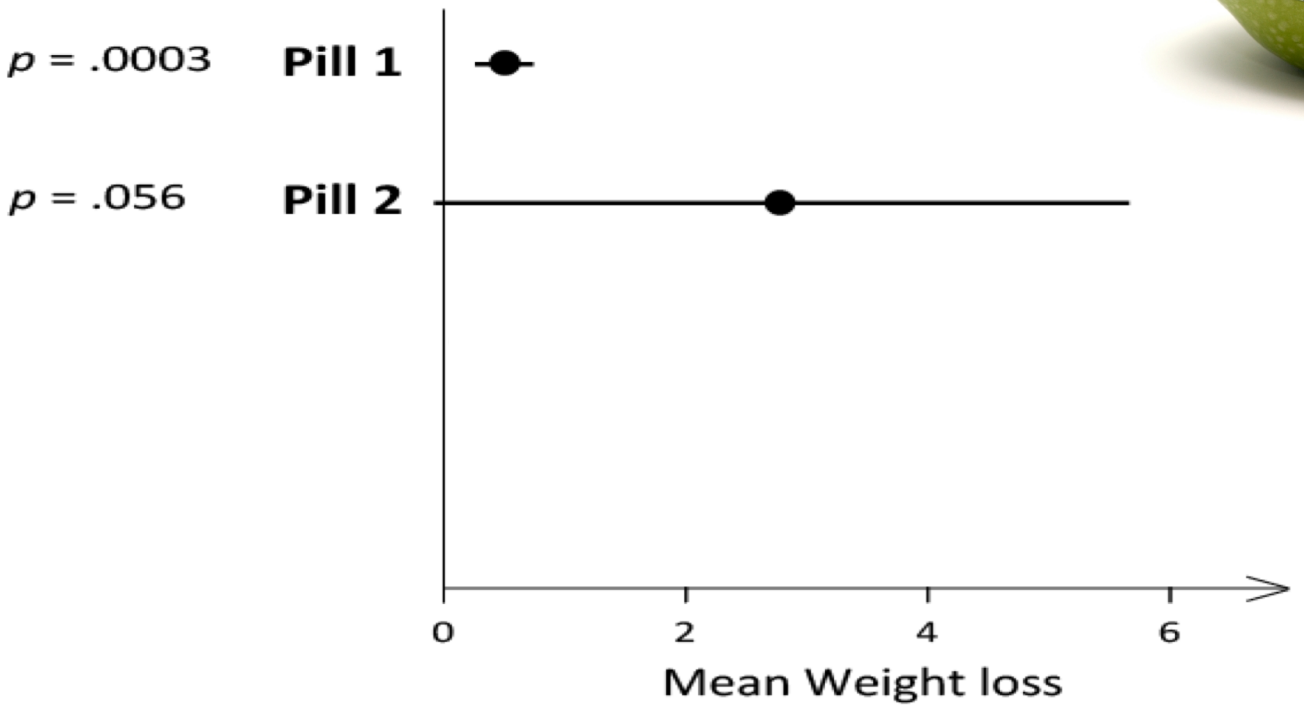
which weight-loss pill would you recommend?



Error bars are 95% CIs
p-values are based on a null hypothesis of no effect

Adapted from
(Ziliak and McCloskey, 2009)

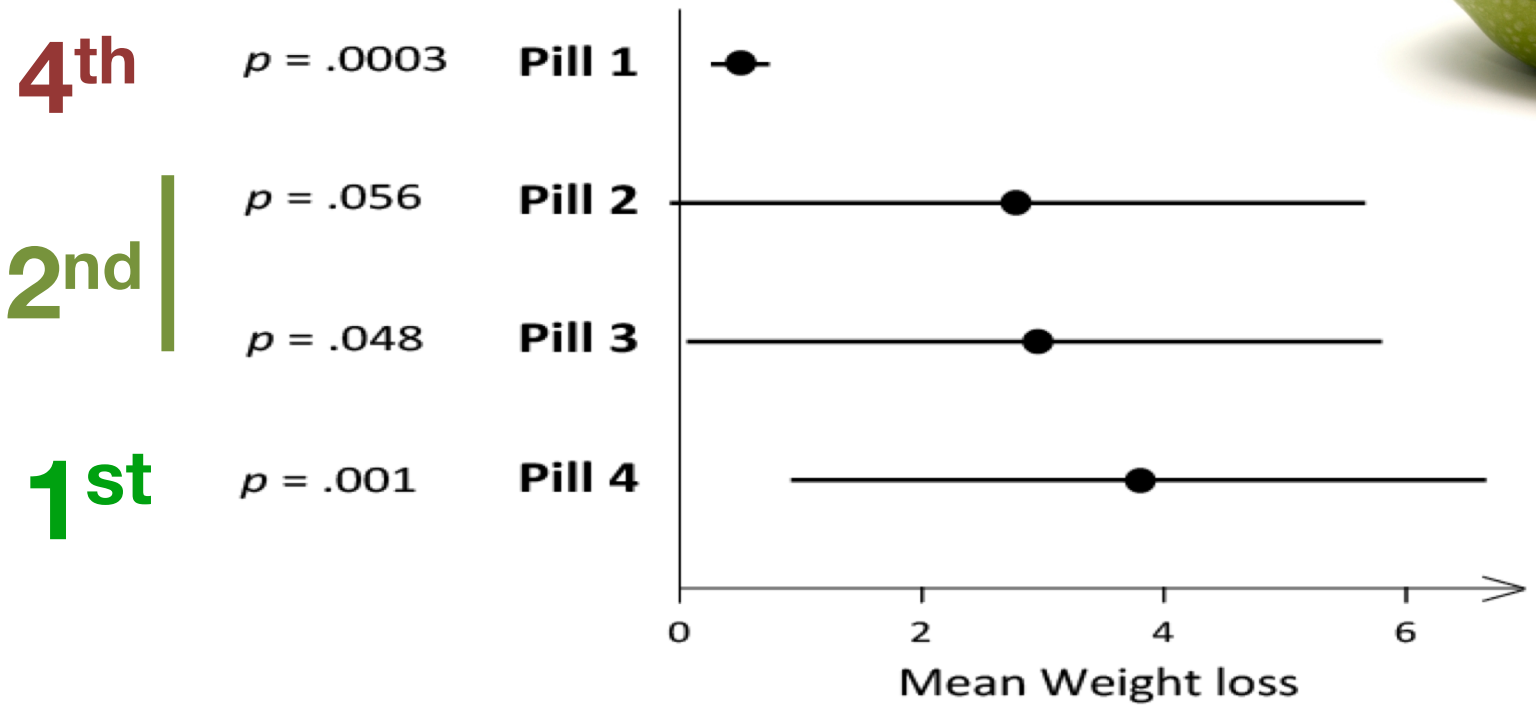
which weight-loss pill would you recommend?



Error bars are 95% CIs
p-values are based on a null hypothesis of no effect

Adapted from
(Ziliak and McCloskey, 2009)

which weight-loss pill would you recommend?

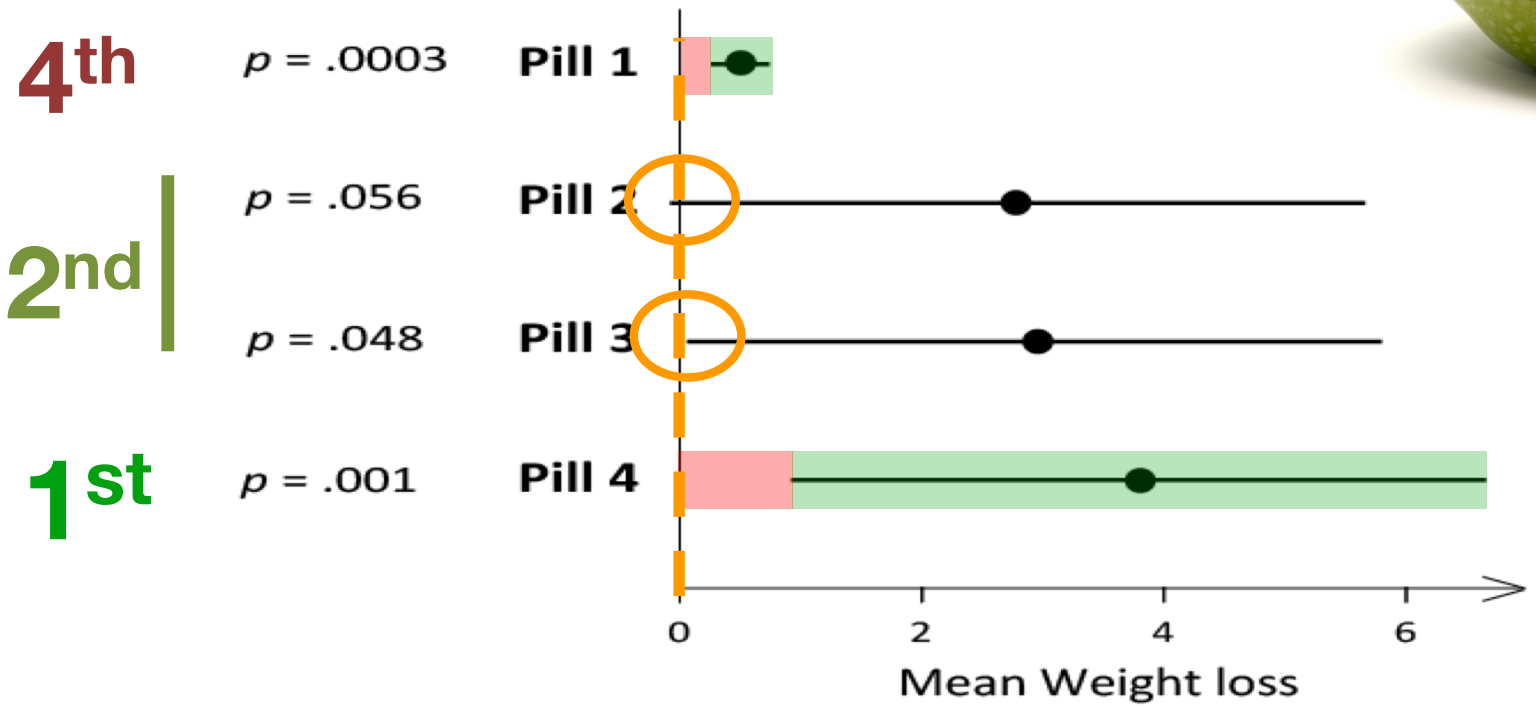


Error bars are 95% CIs
p-values are based on a null hypothesis of no effect

Adapted from
(Ziliak and McCloskey, 2009)

which weight-loss pill would you recommend?

0 kilo lost



Error bars are 95% CIs

p-values are based on a null hypothesis of no effect

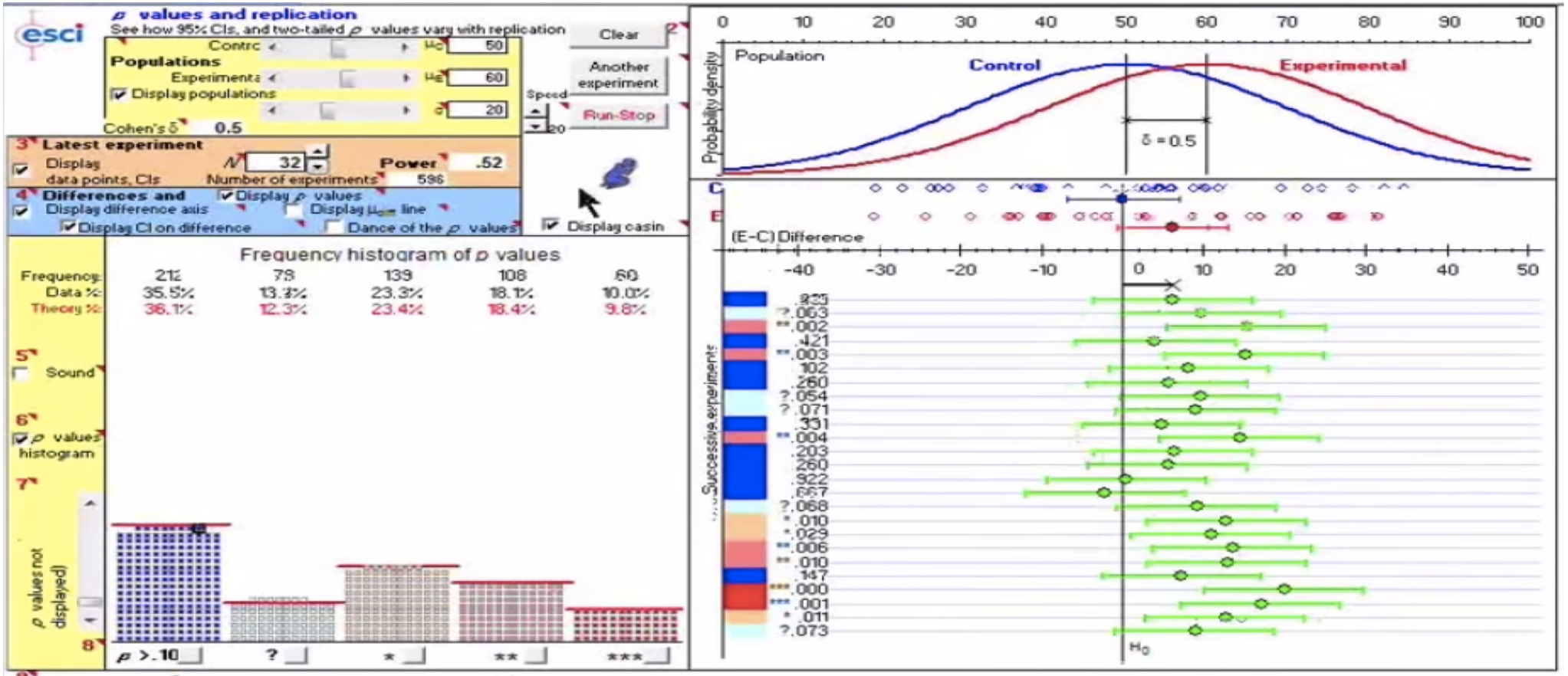
Adapted from
(Ziliak and McCloskey, 2009)

which weight-loss pill would you recommend?

“Statistical significance is perhaps the least important attribute of a good experiment; it is never a sufficient condition for claiming that a theory has been usefully corroborated, that a meaningful empirical fact has been established, or that an experimental report ought to be published” (Likken, 1968)

“We have the duty of communicating our conclusions in intelligible form, in recognition of the right of other free minds to utilize them in making their own decisions” (Fisher, 1955)

“no confidence interval should be interpreted as a a significance test” (Schmidt and Hunter, 1997)



Geoff Cumming's dance of p-values

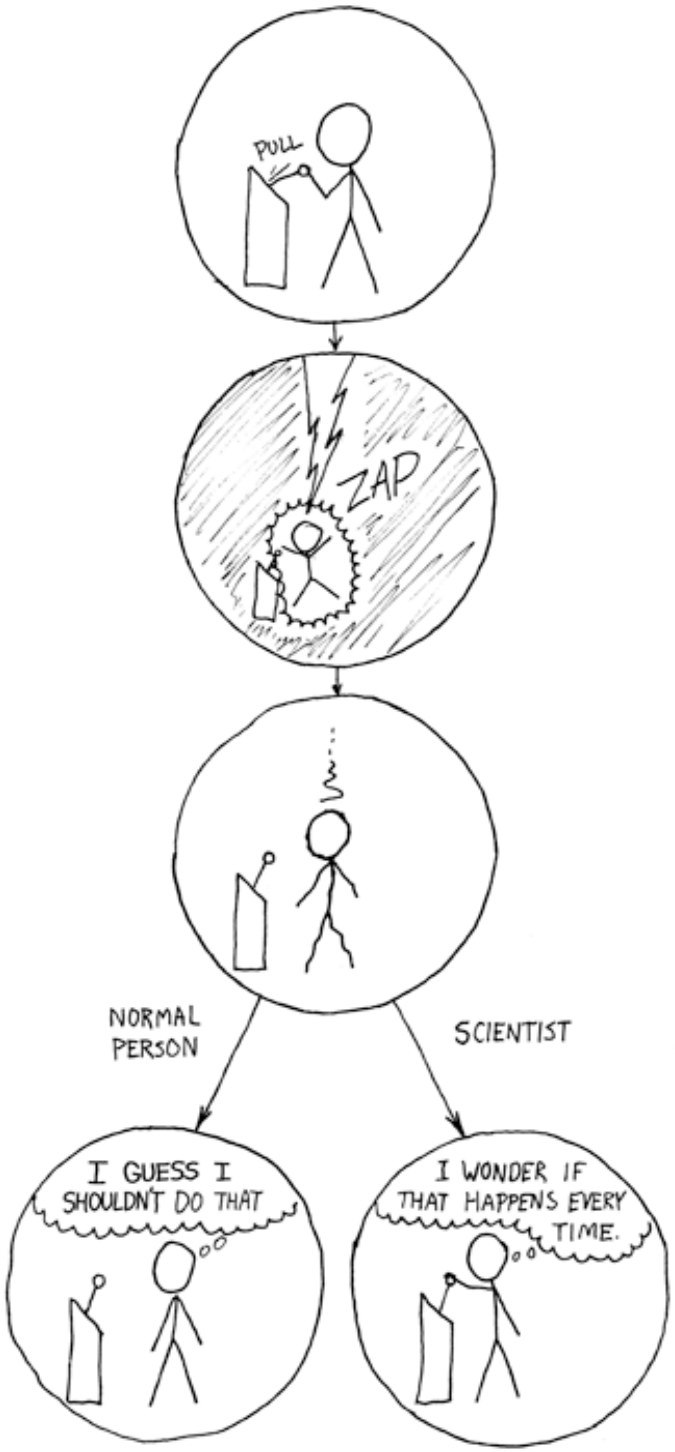
https://www.youtube.com/watch?v=ez4DgdurRPg

spot misleading graphs

replicate!

and fight publication standards,

e.g. in certain fields there are conferences to publish replication studies



rather use artificial data

certain tasks are **very sensitive to human variability** (e.g. ego depletion on will power but also anything that related to preferences)

tasks involving participants but relying on motor skills (e.g. tapping on a key) **suffer less from human variability**

or **use data without involving human** (e.g. algorithms comparisons)

remember this is an active field, always look up for new statistical methods

e.g. at the moment there is a strong tendency to push for **Bayesian testing,** although it also has drawbacks

need **prior data**

simple for AB testing but **could become quickly complex**

unclear **how it compares** to pvalue testing

(still some research to do on this so keep your eyes open!)

for the curious: tutorial on GitHub to do a simple comparison of two groups with Bayesian methods

be ethical

i.e. moral principles that govern a person's behaviour or the conducting of an activity

why are you doing a study, intrinsically because you want to learn something, not just publishing

of course be also ethical with your study design

research goes wrong (Standford Experiment)
… use ethical boards (in each university)

summary

1. Explain what is the replication crisis
2. Give the steps seen in class to avoid phacking and do good statistics
3. Understand that this is a hot topic of research and know that you need to keep your eye open if you ever encounter stats later in your career

take away

please take 5 minute **now** to give us anonymous feedback to improve the lecture

https://tinyurl.com/y65grs5e

to go further

www.biglab.co.uk

BristolIG lab (Youtube)

example of what we do
https://www.youtube.com/watch?v=liPzZIex54M

BIG Bristol Interaction Group

end