Hypothesis testing

NHST concepts

Null and alternative hypotheses

- Tell me. . .

Null and alternative hypotheses

- Tell me. . .
- **Null hypothesis**: there is no difference between groups.

Null and alternative hypotheses

- Tell me. . .
- **Null hypothesis**: there is no difference between groups.
- **Alternative hypothesis**: groups are different.

In ecology, everything is somewhat different

Are there any differences? A non-sensical question in ecology

Alejandro Martínez-Abraín

IMEDEA (CSIC-UIB), C/Miquel Marquès 21, 07190 Esporles, Majorca, Spain

ARTICLE INFO

Article history:
Received 19 December 2006
Accepted 27 April 2007
Published online 13 June 2007

Keywords:

ABSTRACT

One of the main questions that ecologists pose in their investigations includes the analysis of differences in some trait between two or more populations. I argue here that asking whether there are differences or not between populations is biologically irrelevant, since no two livings things are ever equal. On the contrary the appropriate question to pose is how large differences are between populations. That is, we urge a shift in interest from statistical significance to biological relevance for proper knowledge accumulation. I empha-

What is the p-value?

https://pollev.com/franciscorod726

P value

- Very complicated concept: even statisticians fail to describe it well.

P value

- Very complicated concept: even statisticians fail to describe it well.
- Probability of observing data as or more extreme than these *if H0 was true*.

P value

- Very complicated concept: even statisticians fail to describe it well.
- Probability of observing data as or more extreme than these *if H0 was true*.
- Low P-value: data unlikely if H0 was true.

P value

- Very complicated concept: even statisticians fail to describe it well.
- Probability of observing data as or more extreme than these *if H0 was true*.
- Low P-value: data unlikely if H0 was true.
- Large P-value: data not unusual if H0 was true.

Are differences *significant*?

- If $p < 0.05$, we **reject** H0.

Are differences *significant*?

- If $p < 0.05$, we **reject** H0.
- If $p > 0.05$, we **fail to reject** H0

Are differences *significant*?

- If $p < 0.05$, we **reject** H0.
- If $p > 0.05$, we **fail to reject** H0
- (which is **NOT** the same as 'H0 is true')

Are differences *significant*?

- If p $<$ 0.05, we **reject** H0.
- If p $>$ 0.05, we **fail to reject** H0
- (which is **NOT** the same as 'H0 is true')
- **CAUTION:**

Are differences *significant*?

- If p $<$ 0.05, we **reject** H0.
- If p $>$ 0.05, we **fail to reject** H0
- (which is **NOT** the same as 'H0 is true')
- **CAUTION:**
- This is **very widespread, but incorrect** practice.

Are differences *significant*?

- If p < 0.05, we **reject** H0.
- If p > 0.05, we **fail to reject** H0
- (which is **NOT** the same as 'H0 is true')
- **CAUTION:**
- This is **very widespread, but incorrect** practice.
- P-value is continuous. We must **avoid binary decisions** based on **arbitrary thresholds**.

Are differences *significant*?

- If p $<$ 0.05, we **reject** H0.
- If p $>$ 0.05, we **fail to reject** H0
- (which is **NOT** the same as 'H0 is true')
- **CAUTION:**
- This is **very widespread, but incorrect** practice.
- P-value is continuous. We must **avoid binary decisions** based on **arbitrary thresholds**.
- More on this later.

Let's do the test

```
t.test(h.sevi, h.out)
```

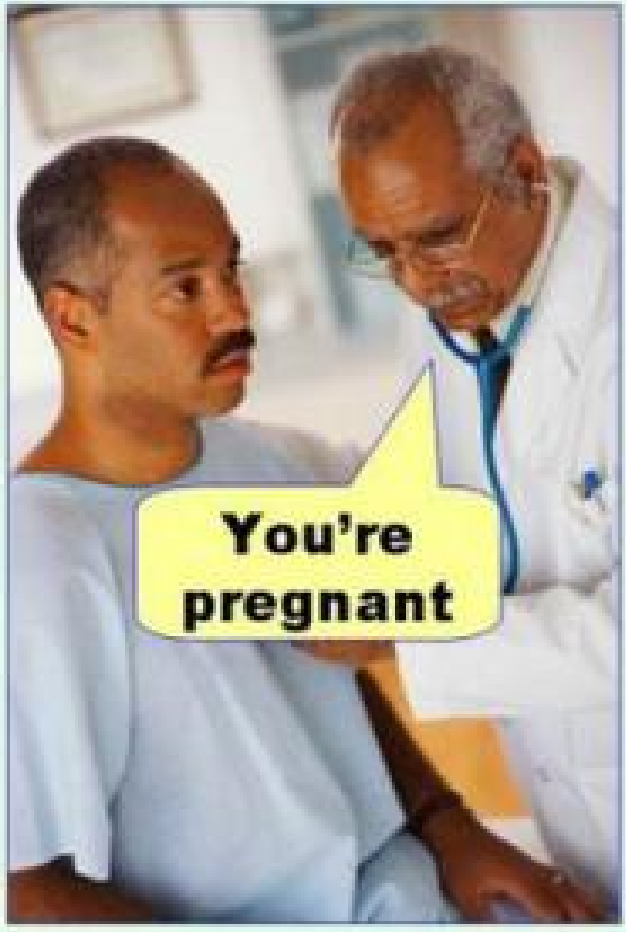
```
    Welch Two Sample t-test

data:  h.sevi and h.out
t = -0.67636, df = 8.9167, p-value = 0.516
alternative hypothesis: true difference in means is not equal to
95 percent confidence interval:
 -14.353024   7.753024
sample estimates:
mean of x mean of y
    174.4     177.7
```

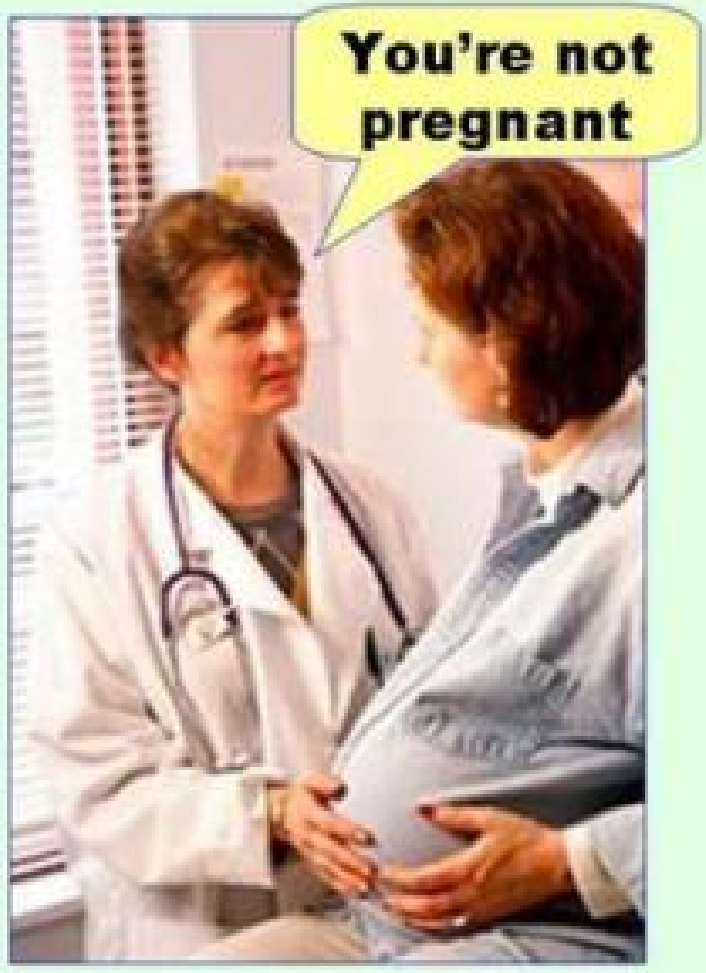
Are heights different then?

Rejecting hypotheses: two types of error

Type I error
(false positive)



Type II error
(false negative)



Rejecting hypotheses: two types of error

Statistics: Hypothesis Test	Null Hypothesis is True	Null Hypothesis is False
Reject Null Hypothesis	Type I Error	Correct
Fail to Reject Null Hypothesis	Correct	Type II Error

Power: Probability of detecting true difference (rejecting H0 when it's false).

Understanding NHST

http://rpsychologist.com/d3/NHST/

Example: biased coin

```
 [1] 0 1 1 1 0 1 0 1 1 1


    1-sample proportions test with continuity correction

data:  sum(coin) out of ntrials, null probability 0.5
X-squared = 0.9, df = 1, p-value = 0.3428
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3536707 0.9190522
sample estimates:
  p
0.7
```

Correlation between variables

http://rpsychologist.com/d3/correlation/

Common pitfalls and good practice

A must read

Eur J Epidemiol (2016) 31:337–350
DOI 10.1007/s10654-016-0149-3

ESSAY

Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations

Sander Greenland[1] · Stephen J. Senn[2] · Kenneth J. Rothman[3] · John B. Carlin[4] · Charles Poole[5] · Steven N. Goodman[6] · Douglas G. Altman[7]

https://doi.org/10.1007/s10654-016-0149-3

Good read

esa

ECOSPHERE

Applied statistics in ecology: common pitfalls and simple solutions

E. Ashley Steel,[1,†] Maureen C. Kennedy,[2] Patrick G. Cunningham,[3] and John S. Stanovick[4]

https://doi.org/10.1890/ES13-00160.1

Also http://www.statisticsdonewrong.com/

Good read

Twenty tips for interpreting scientific claims

https://doi.org/10.1038/503335a

Visualisation of data and models is key

First things first

- Always

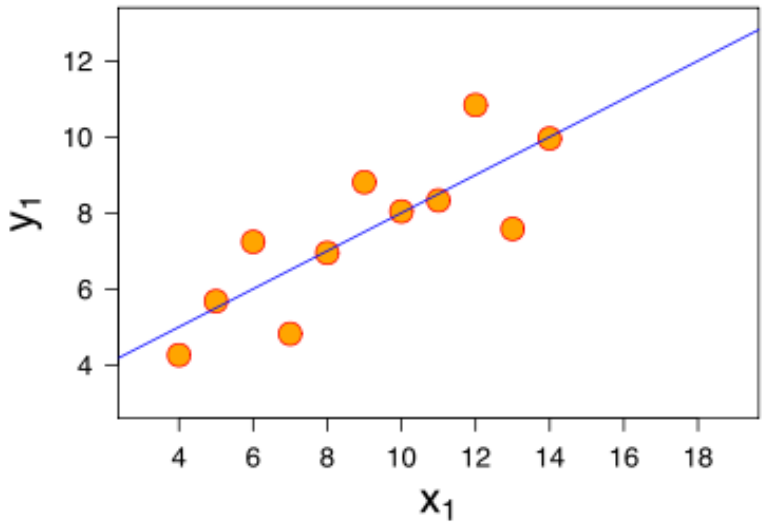
First things first

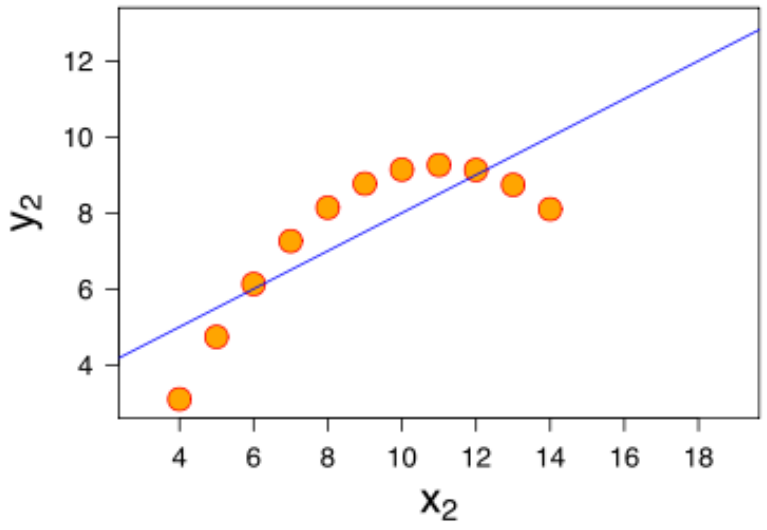
- Always
- Always

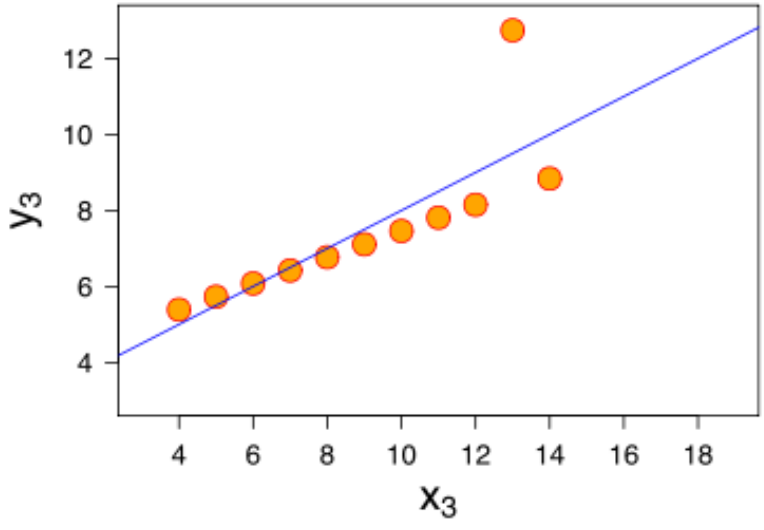
First things first

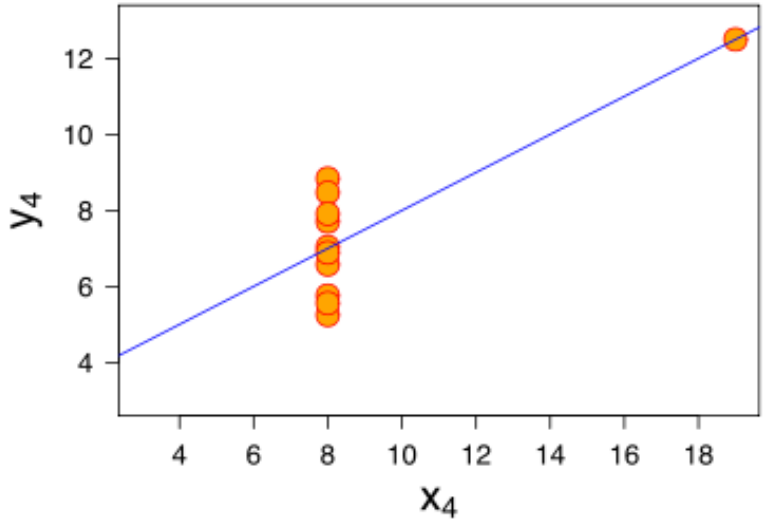
- Always
- Always
- Always

Plot data and models

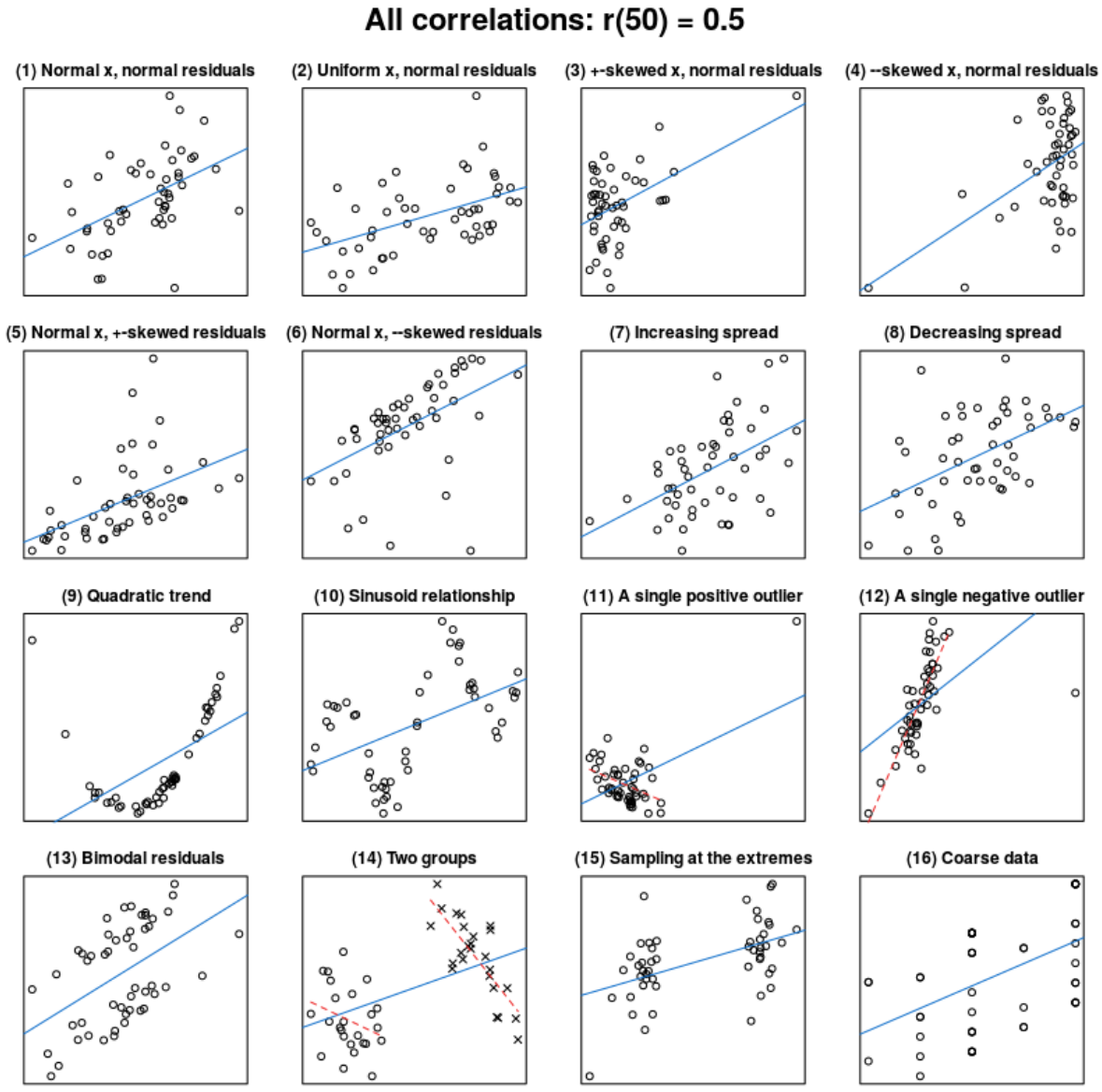






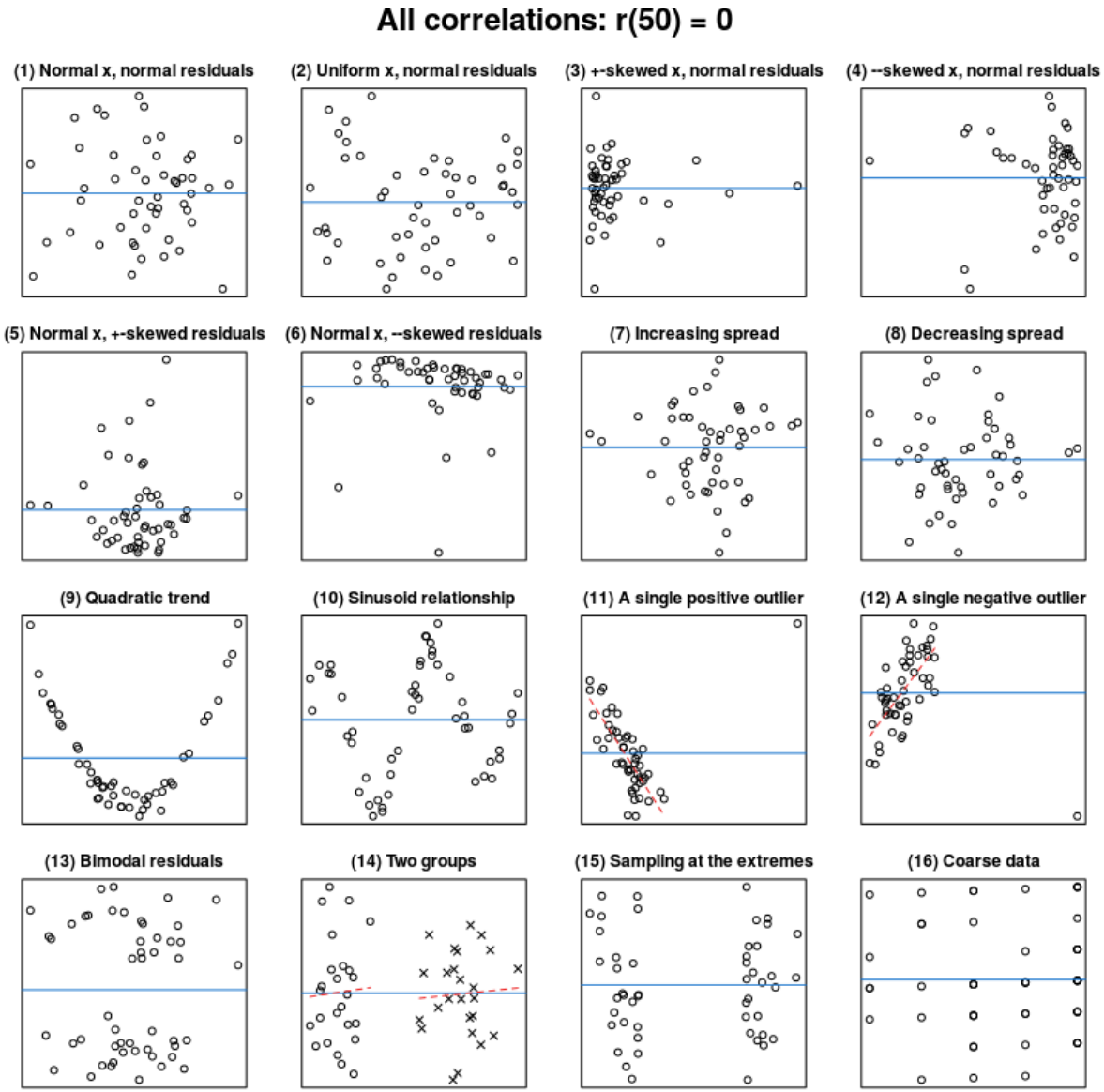


Don't use statistics blindly: *Visualise*



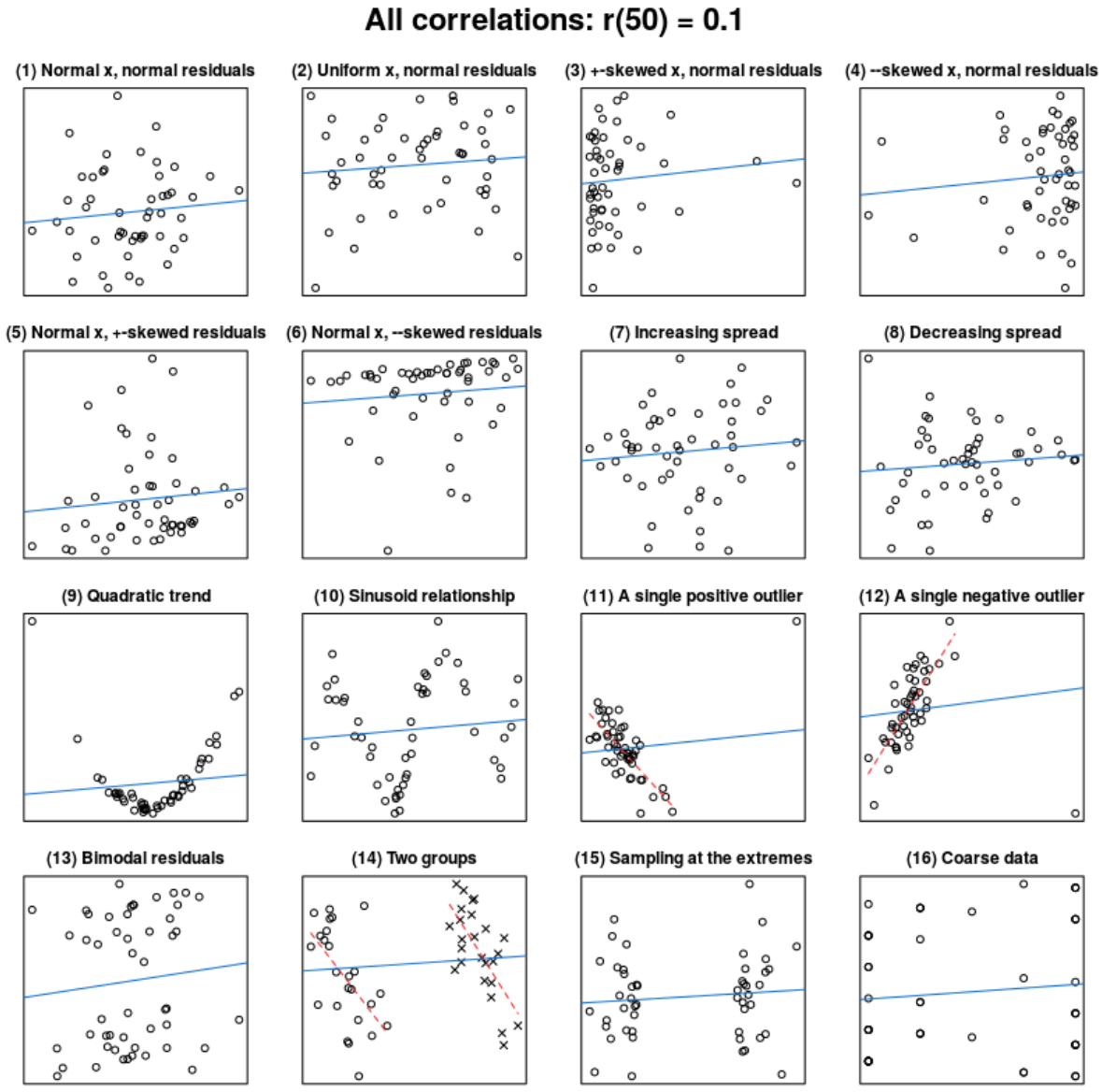
https://janhove.github.io/teaching/2016/11/21/what-correlations-look-like

Don't use statistics blindly: *Visualise*



https://janhove.github.io/teaching/2016/11/21/what-correlations-look-like

Don't use statistics blindly: *Visualise*



https://janhove.github.io/teaching/2016/11/21/what-correlations-look-like

Plot. Check models. Plot. Check assumptions. Plot.

Lavine 2014 *Ecology*

Inference from observational studies

News: Hamburgers increase risk of heart attack

- In a sample of 10,000 people, it was found that people eating >2 hamburgers a week had 20% higher probability of heart attack.

News: Hamburgers increase risk of heart attack

- In a sample of 10,000 people, it was found that people eating >2 hamburgers a week had 20% higher probability of heart attack.
- **Do hamburgers increase heart attacks?**

News: Hamburgers increase risk of heart attack

- In a sample of 10,000 people, it was found that people eating >2 hamburgers a week had 20% higher probability of heart attack.
- **Do hamburgers increase heart attacks?**
- https://pollev.com/franciscorod726

Bigger flowers increase reproductive success

- We found that plants with big flowers produced 30% more seeds...

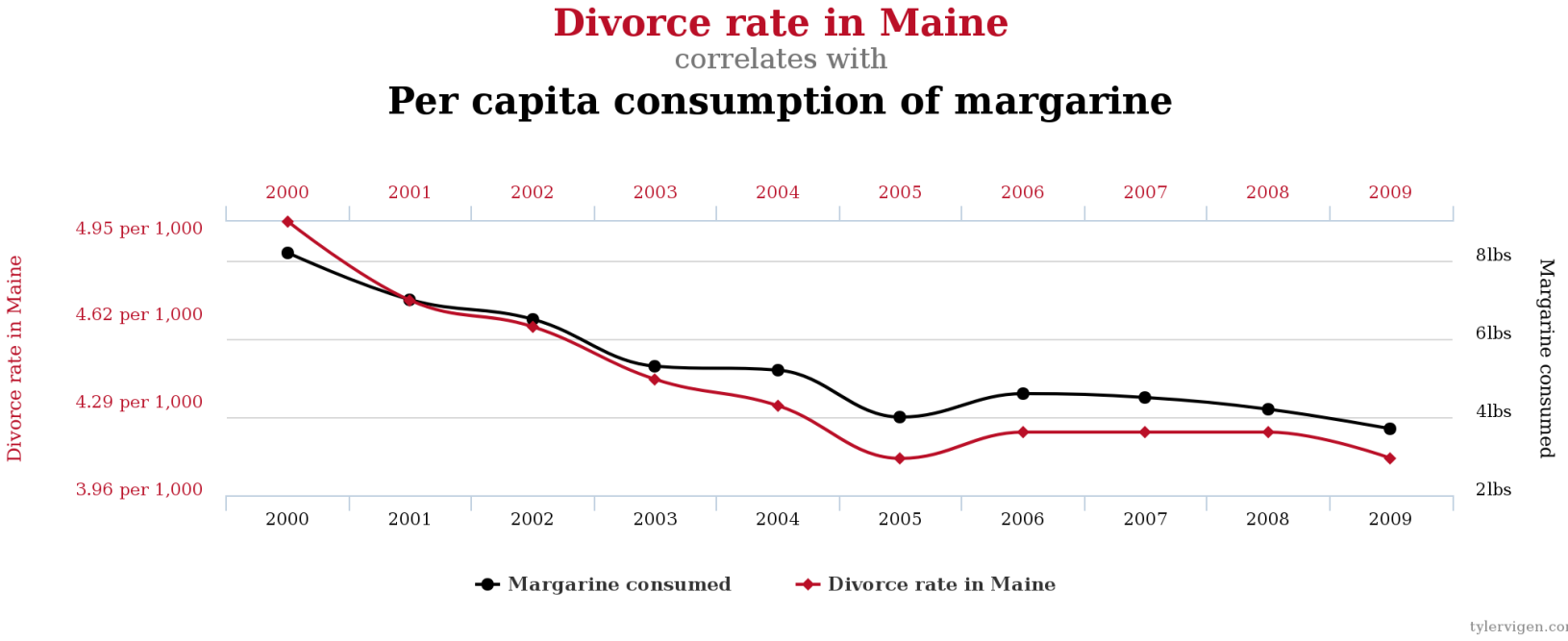
Bigger flowers increase reproductive success

- We found that plants with big flowers produced 30% more seeds...
- **Do big flowers increase reproductive success?**

Bigger flowers increase reproductive success

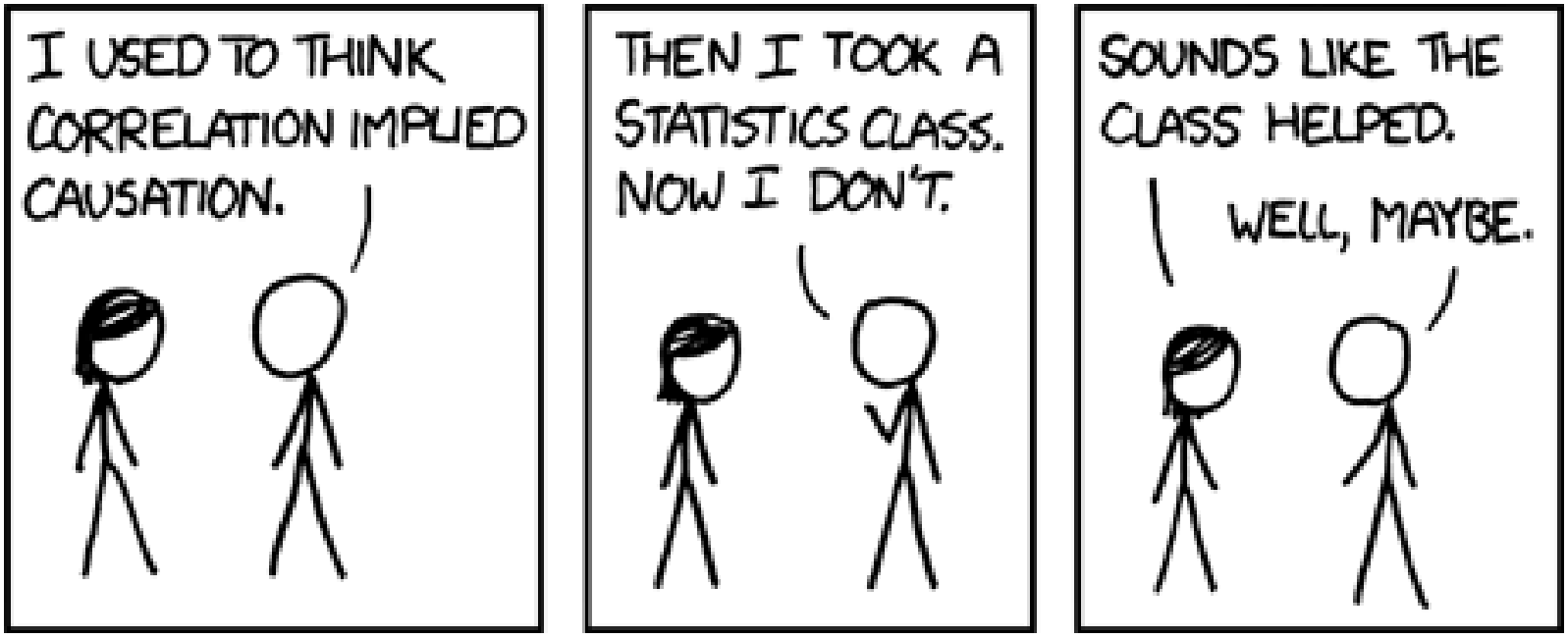
- We found that plants with big flowers produced 30% more seeds...
- **Do big flowers increase reproductive success?**
- https://pollev.com/franciscorod726

Correlation vs Causation



http://tylervigen.com/spurious-correlations

Learning statistics through xkcd



NHST and p-values

In ecology, everything is somewhat different

Are there any differences? A non-sensical question in ecology

Alejandro Martínez-Abraín

IMEDEA (CSIC-UIB), C/Miquel Marquès 21, 07190 Esporles, Majorca, Spain

ARTICLE INFO

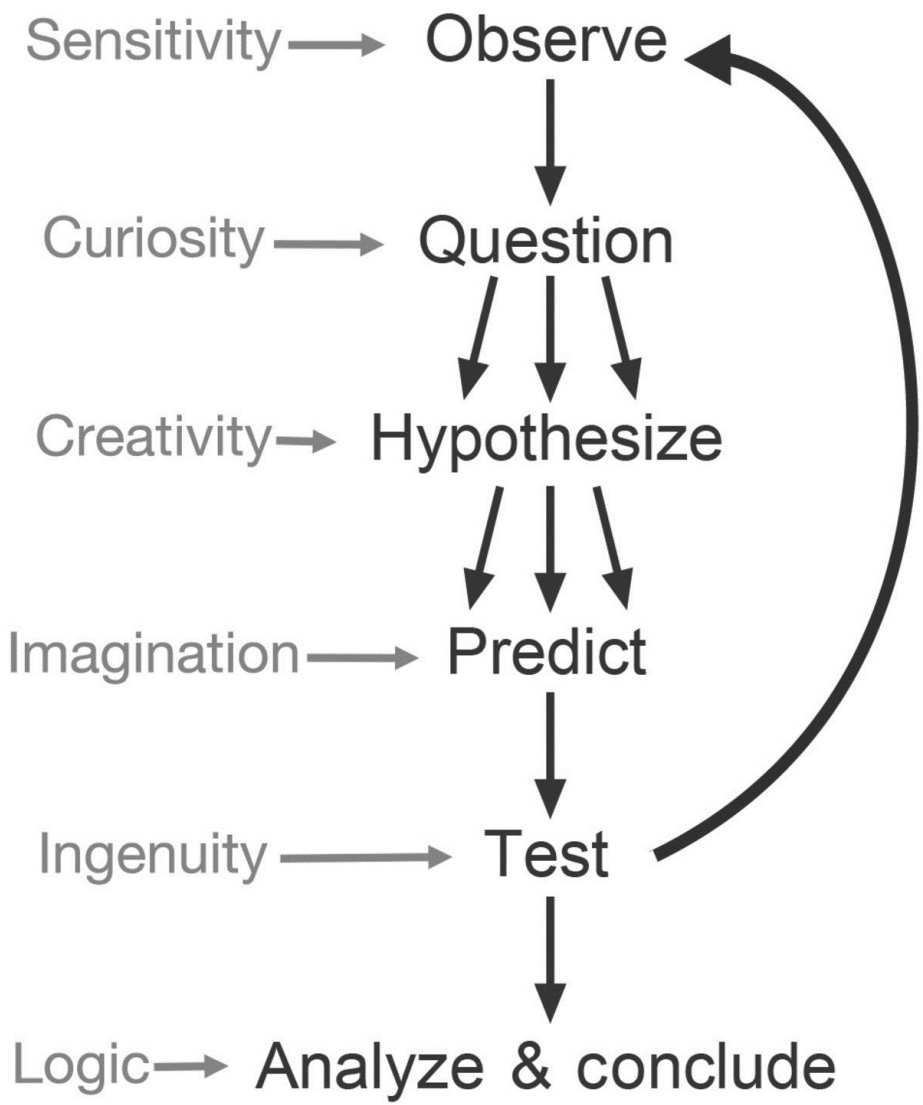
Article history:
Received 19 December 2006
Accepted 27 April 2007
Published online 13 June 2007

Keywords:

ABSTRACT

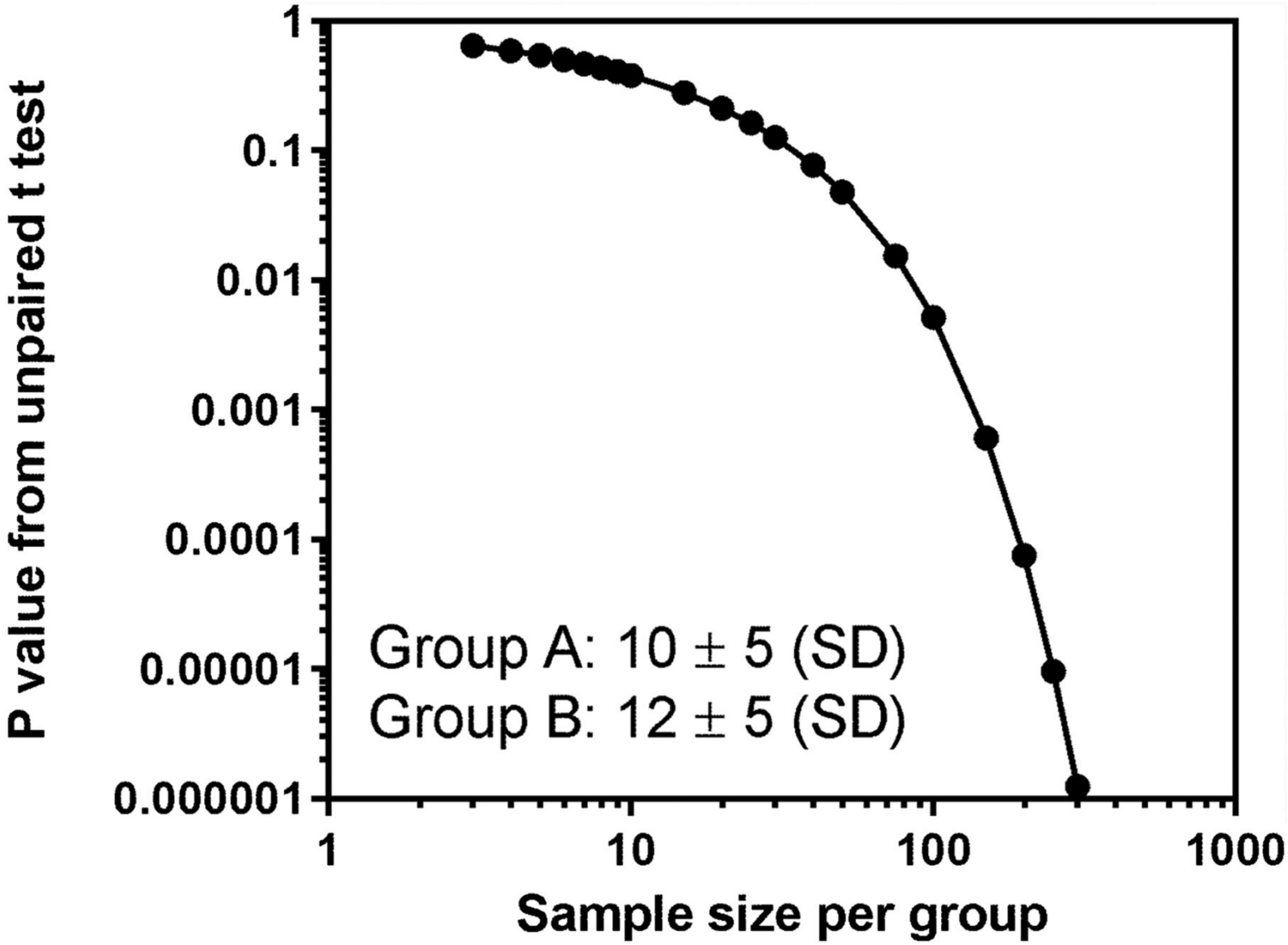
One of the main questions that ecologists pose in their investigations includes the analysis of differences in some trait between two or more populations. I argue here that asking whether there are differences or not between populations is biologically irrelevant, since no two livings things are ever equal. On the contrary the appropriate question to pose is how large differences are between populations. That is, we urge a shift in interest from statistical significance to biological relevance for proper knowledge accumulation. I empha-

Instead of falsifying a null model, estimate effects and compare meaningful models



https://doi.org/10.1242/jeb.104976

P-value depends on sample size

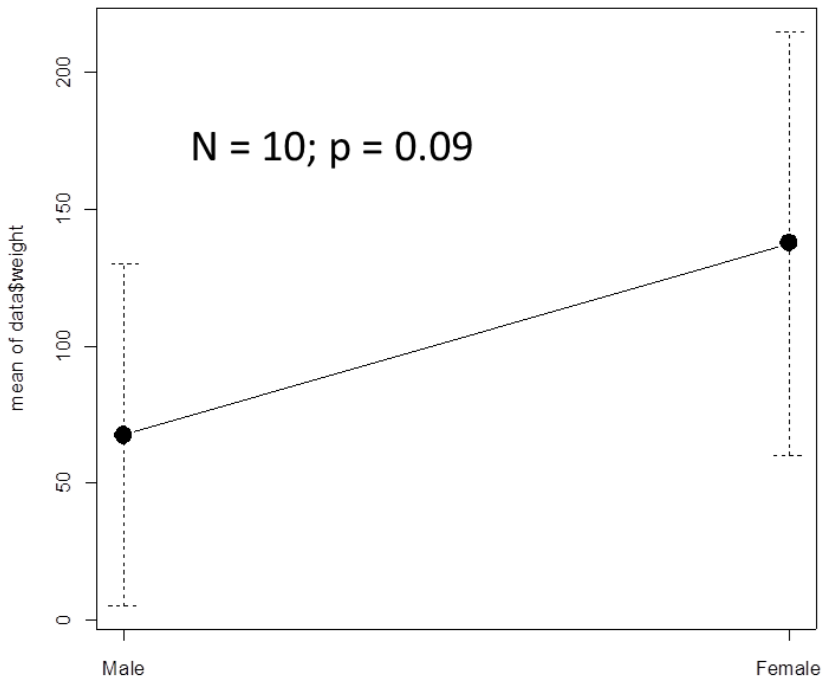


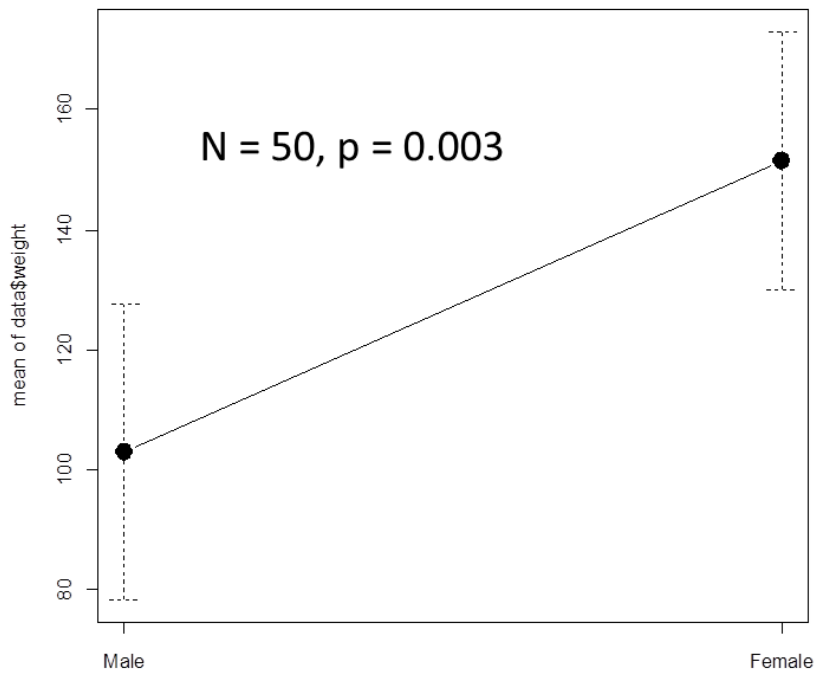
https://doi.org/10.1002/prp2.93

P-value depends on sample size

- Same real difference is detected as significant or not depending on sample size:

Real difference = 40 g

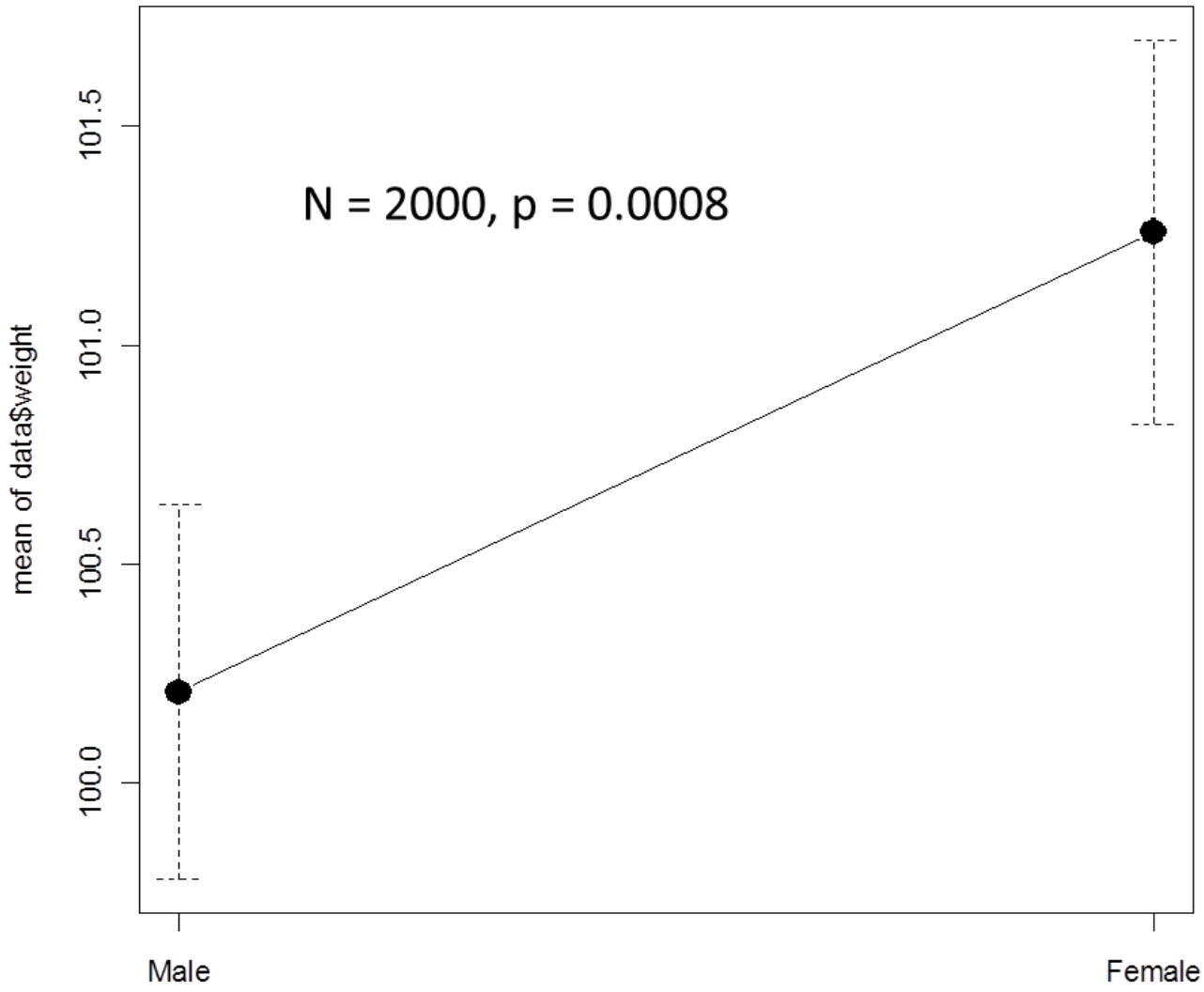




Statistically significant != biologically important

- With big sample size, we can find **highly significant but biologically unimportant** differences.

Real difference = 1 g



Statistically significant != biologically important

- Statistically significant = unlikely to be zero

Statistically significant != biologically important

- Statistically significant = unlikely to be zero
- Good read: *significantly misleading*

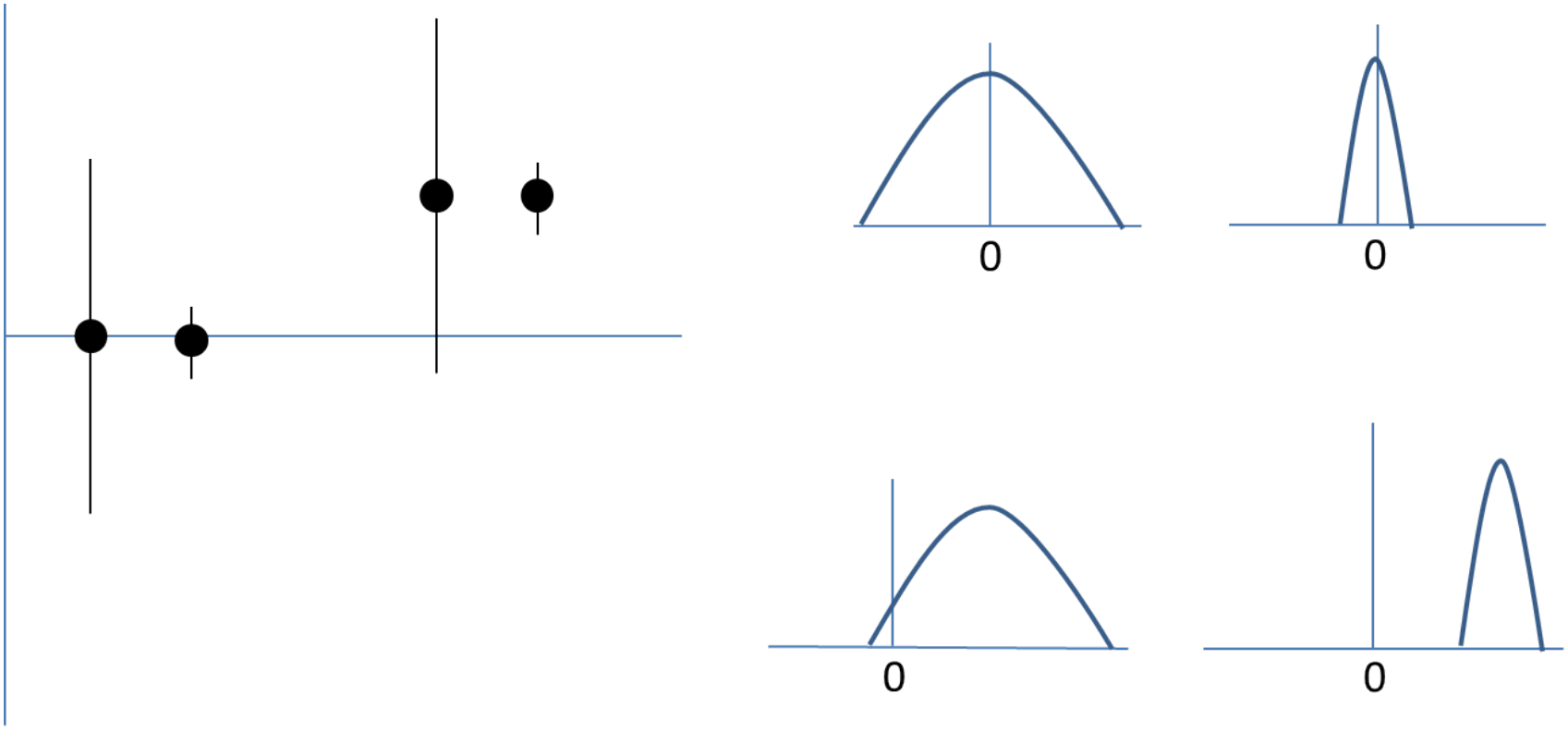
Statistically significant != biologically important

- Statistically significant = unlikely to be zero
- Good read: *significantly misleading*
- My suggestion: avoid significant/not significant (and maybe p-values too)

Statistically significant != biologically important

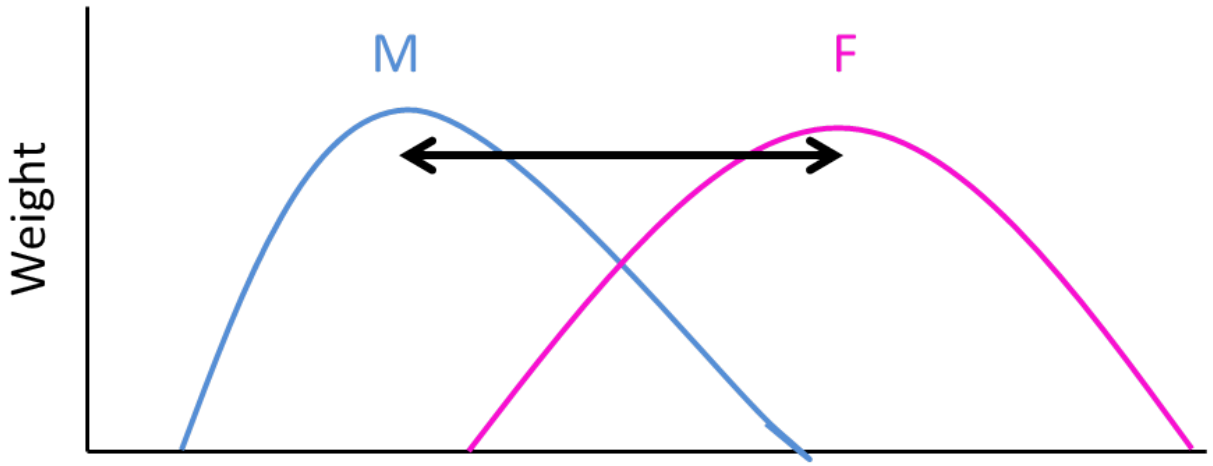
- Statistically significant = unlikely to be zero
- Good read: *significantly misleading*
- My suggestion: avoid significant/not significant (and maybe p-values too)
- Beyond significance, look at *effect sizes*.

‘Not significant’ does NOT mean ‘there is no effect’

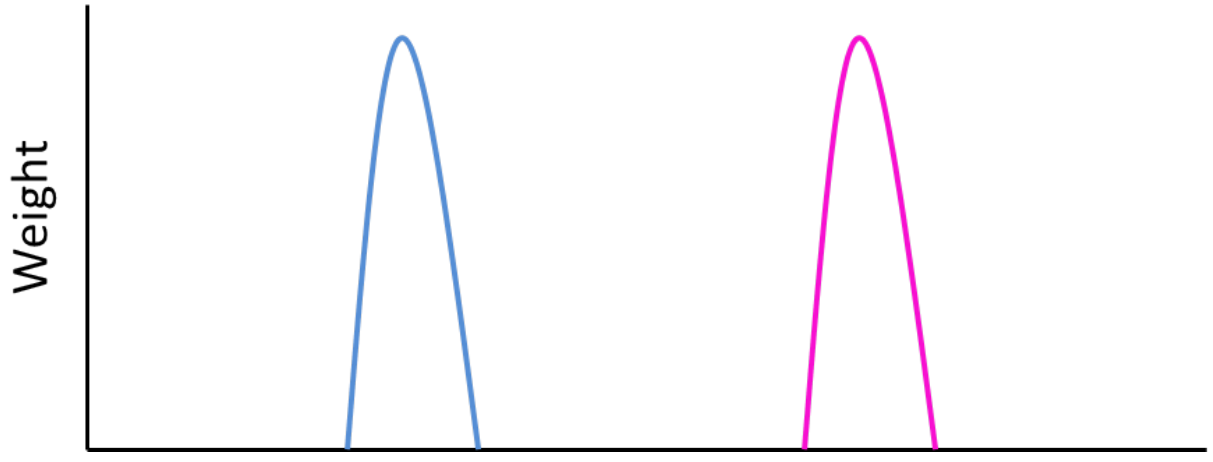


- **Absence of evidence != Evidence of absence**

Failure to reject H0 != H0 is true



P >> 0.05



P << 0.05

p-value > 0.05?

- "We were unable to find evidence against the hypothesis that A = B with the current sample size" (Harrell)

p-value $>$ 0.05?

- "We were unable to find evidence against the hypothesis that A $=$ B with the current sample size" (Harrell)
- "Differences between groups were not statistically clear" (Dushoff et al)

Is it safe to allow right turn with red lights?

- Right turn not allowed: 308 accidents

https://www.statisticsdonewrong.com/power.html#the-wrong-turn-on-red

Is it safe to allow right turn with red lights?

- Right turn not allowed: 308 accidents
- Right turn allowed: 337 accidents

https://www.statisticsdonewrong.com/power.html#the-wrong-turn-on-red

Is it safe to allow right turn with red lights?

- Right turn not allowed: 308 accidents
- Right turn allowed: 337 accidents
- No *significant* difference, hence safe

https://www.statisticsdonewrong.com/power.html#the-wrong-turn-on-red

Is it safe to allow right turn with red lights?

- Right turn not allowed: 308 accidents
- Right turn allowed: 337 accidents
- No *significant* difference, hence safe
- Misinterpretation of underpowered study cost lives

https://www.statisticsdonewrong.com/power.html#the-wrong-turn-on-red

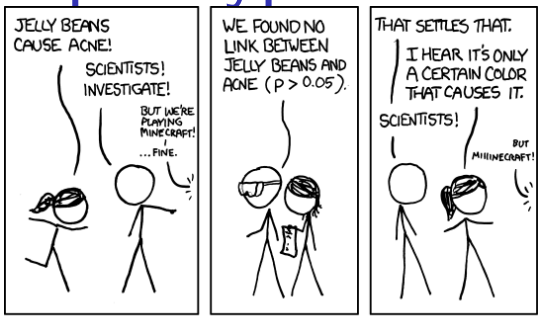
0.05 is an arbitrary threshold

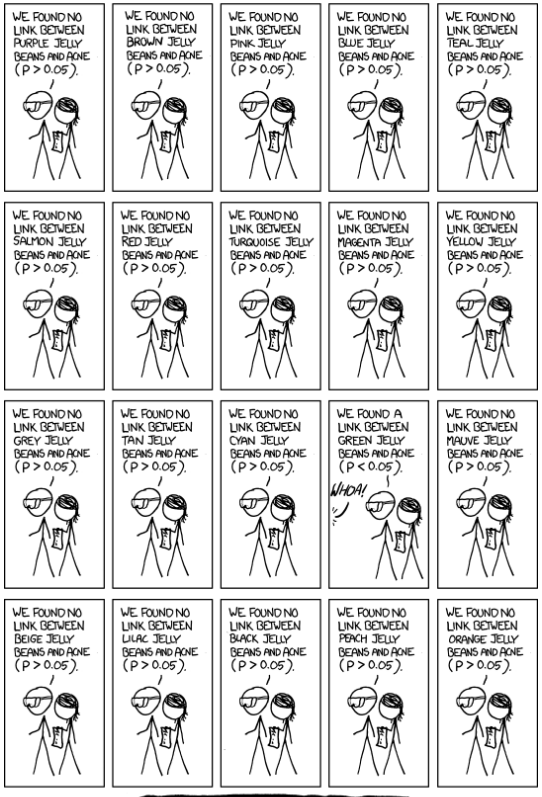
The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant

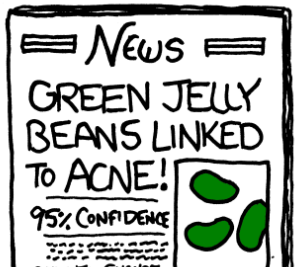
Andrew GELMAN and Hal STERN

http://dx.doi.org/10.1198/000313006X152649

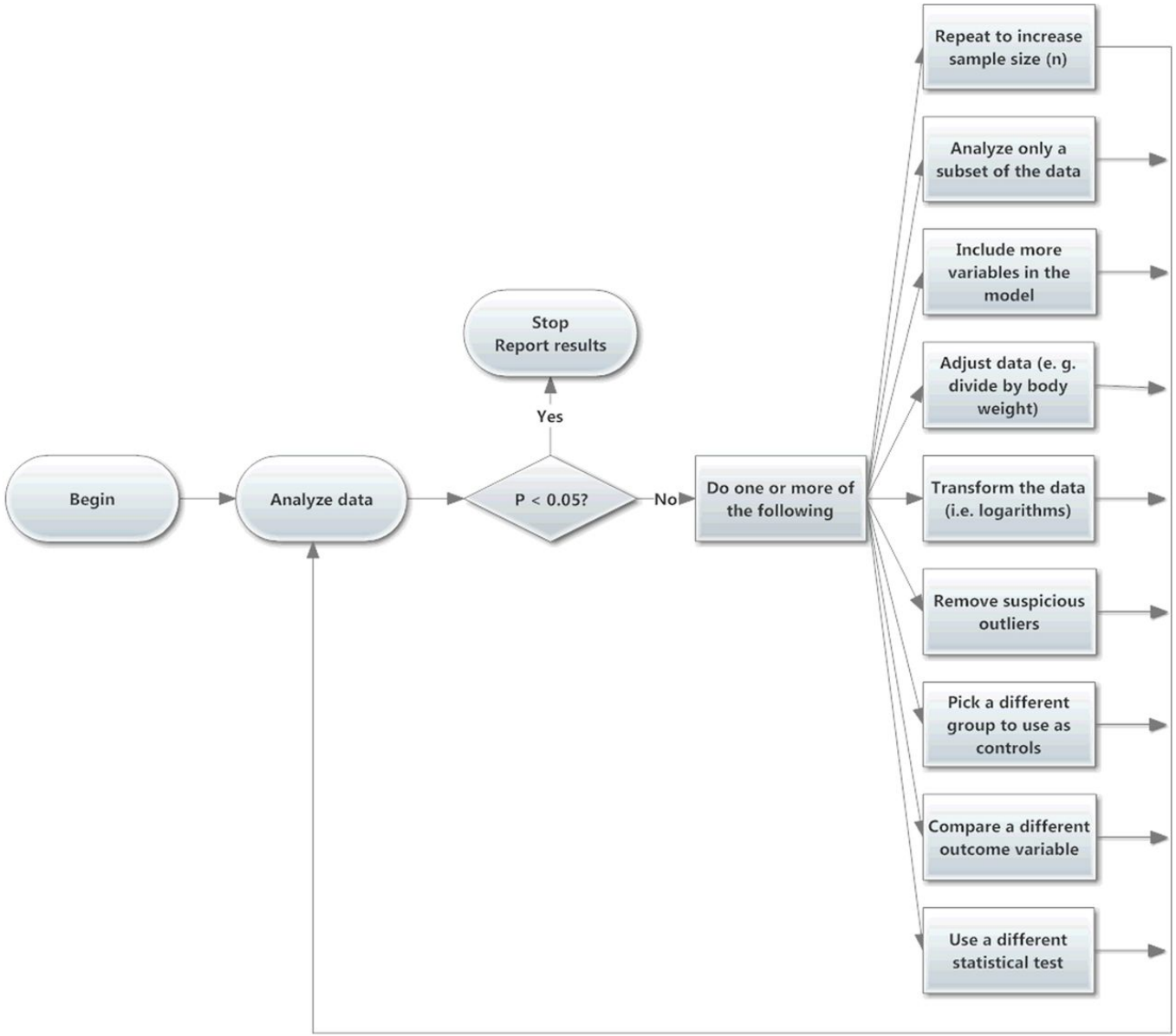
Multiple hypothesis testing







How to make your results significant: *p-hacking*



http://dx.doi.org/10.1002/prp2.93

How to make your results significant: *p-hacking*

1. Test multiple variables, then report the ones that are significant.

How to make your results significant: *p-hacking*

1. Test multiple variables, then report the ones that are significant.
2. Artificially choose when to end your experiment.

How to make your results significant: *p-hacking*

1. Test multiple variables, then report the ones that are significant.
2. Artificially choose when to end your experiment.
3. Add covariates until effects are significant.

How to make your results significant: *p-hacking*

1. Test multiple variables, then report the ones that are significant.
2. Artificially choose when to end your experiment.
3. Add covariates until effects are significant.
4. Test different conditions (e.g. different levels of a factor) and report the ones you like.

How to make your results significant: *p-hacking*

1. Test multiple variables, then report the ones that are significant.
2. Artificially choose when to end your experiment.
3. Add covariates until effects are significant.
4. Test different conditions (e.g. different levels of a factor) and report the ones you like.

- To read more: Simmons et al 2011

How to make your results significant: *p-hacking*

https://www.youtube.com/watch?v=ZaNtz76dNSI

ASA statement on p-values

- P-values do not measure the **probability of hypothesis** being true, or the probability that the data were produced by **random chance** alone.

https://doi.org/10.1080/00031305.2016.1154108

ASA statement on p-values

- P-values do not measure the **probability of hypothesis** being true, or the probability that the data were produced by **random chance** alone.
- Scientific conclusions or policy decisions should NOT be based only on **whether a p-value passes a specific threshold**.

https://doi.org/10.1080/00031305.2016.1154108

ASA statement on p-values

- P-values do not measure the **probability of hypothesis** being true, or the probability that the data were produced by **random chance** alone.
- Scientific conclusions or policy decisions should NOT be based only on **whether a p-value passes a specific threshold**.
- P-value, or statistical significance, does not measure the **size of an effect** or the **importance** of a result.

https://doi.org/10.1080/00031305.2016.1154108

ASA statement on p-values

- P-values do not measure the **probability of hypothesis** being true, or the probability that the data were produced by **random chance** alone.
- Scientific conclusions or policy decisions should NOT be based only on **whether a p-value passes a specific threshold**.
- P-value, or statistical significance, does not measure the **size of an effect** or the **importance** of a result.
- By itself, a p-value does NOT provide a good **measure of evidence** regarding a model or hypothesis.

https://doi.org/10.1080/00031305.2016.1154108

The New Statistics

Aim for estimation of effects and their uncertainty (SE, CI. . .)

General Article

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
2014, Vol. 25(1) 7–29

DOI: 10.1177/0956797613504966
pss.sagepub.com
SAGE

The New Statistics: Why and How

Geoff Cumming
La Trobe University

http://dx.doi.org/10.1177/0956797613504966

How many types of errors?

- **Type I**: False positive (incorrect rejection of null hypothesis).

How many types of errors?

- **Type I**: False positive (incorrect rejection of null hypothesis).
- **Type II**: False negative (failure to reject false null hypothesis).

How many types of errors?

- **Type I**: False positive (incorrect rejection of null hypothesis).
- **Type II**: False negative (failure to reject false null hypothesis).
- **Type S (Sign)**: estimating effect in opposite direction.

How many types of errors?

- **Type I**: False positive (incorrect rejection of null hypothesis).
- **Type II**: False negative (failure to reject false null hypothesis).
- **Type S (Sign)**: estimating effect in opposite direction.
- **Type M (Magnitude)**: Misestimating magnitude of the effect (under or overestimating).

How many types of errors?

- **Type I**: False positive (incorrect rejection of null hypothesis).
- **Type II**: False negative (failure to reject false null hypothesis).
- **Type S (Sign)**: estimating effect in opposite direction.
- **Type M (Magnitude)**: Misestimating magnitude of the effect (under or overestimating).
- Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors