# Hypothesis testing

# NHST concepts

# Null and alternative hypotheses

- Tell me...

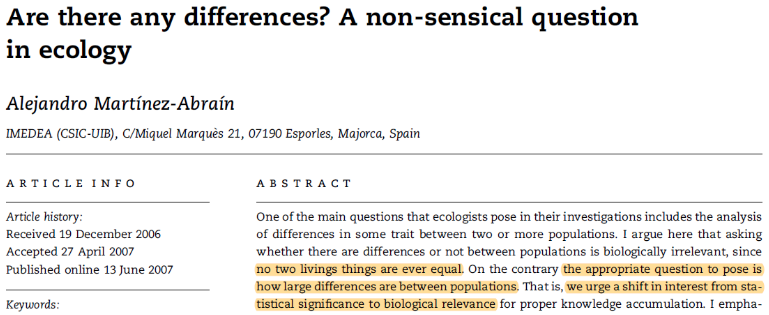
# Null and alternative hypotheses

- 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?

- The probability that the data were produced by random chance alone

https://pollev.com/franciscorod726

# What is the p-value?

- The probability that the data were produced by random chance alone
- The probability of getting results at least as extreme as the ones you observed if H0 was true

https://pollev.com/franciscorod726

# What is the p-value?

- The probability that the data were produced by random chance alone
- The probability of getting results at least as extreme as the ones you observed if H0 was true
- The probability of null hypothesis being true

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.

# If p-value > 0.05

- the null hypothesis is false, i.e. the alternative hypothesis must be true

https://pollev.com/franciscorod726

# If p-value > 0.05

- the null hypothesis is false, i.e. the alternative hypothesis must be true
- the alternative hypothesis is false, i.e. the null hypothesis must be true

`https://pollev.com/franciscorod726`

# If p-value > 0.05

- the null hypothesis is false, i.e. the alternative hypothesis must be true
- the alternative hypothesis is false, i.e. the null hypothesis must be true
- it’s unclear if there are differences between groups

`https://pollev.com/franciscorod726`

## If p-value > 0.05

- the null hypothesis is false, i.e. the alternative hypothesis must be true
- the alternative hypothesis is false, i.e. the null hypothesis must be true
- it’s unclear if there are differences between groups
- there is no difference between groups

`https://pollev.com/franciscorod726`

# 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*?

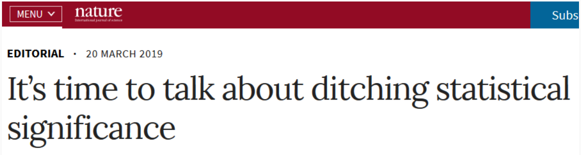
- If $p < 0.05$, we **reject** H0.
- If $p > 0.05$, we **fail to reject** H0
- (which is **NOT** the same as ‘H0 is true’)
- **CAUTION:**
- P-value is continuous. We must **avoid binary decisions** based on **arbitrary thresholds**.

# 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:**
- 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:**
- P-value is continuous. We must **avoid binary decisions** based on **arbitrary thresholds**.



**EDITORIAL** · 20 MARCH 2019

It’s time to talk about ditching statistical significance

https://doi.org/10.1038/d41586-019-00857-9

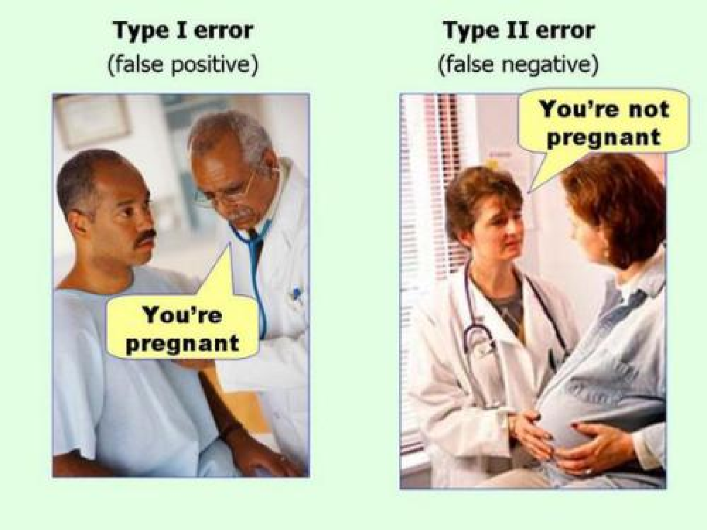
# Are the heights of local and non-local students different?

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

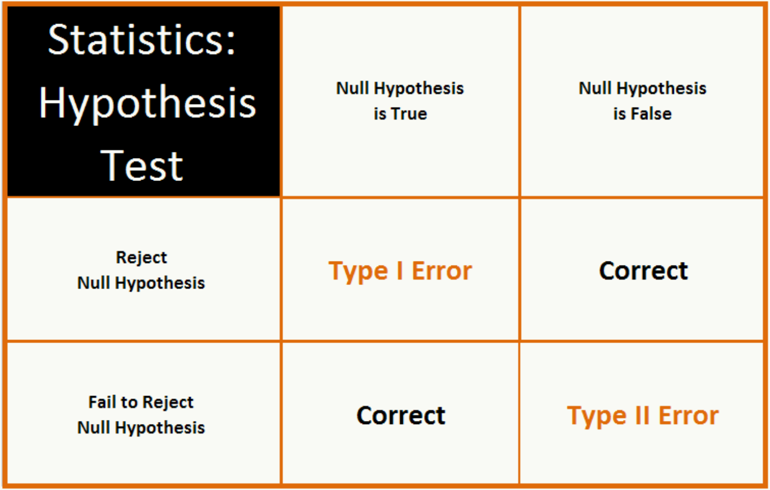
```

    Welch Two Sample t-test

data:  h.sevi and h.out
t = -3.159, df = 11.768, p-value = 0.00842
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -21.986075  -4.013925
sample estimates:
mean of x mean of y
    165.8     178.8
```

https://pollev.com/franciscorod726

# Rejecting hypotheses: two types of error



# 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] 1 0 1 0 1 0 0 1 1 0


	1-sample proportions test without continuity correction

data:  sum(coin) out of ntrials, null probability 0.5
X-squared = 0, df = 1, p-value = 1
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.2365931 0.7634069
sample estimates:
  p
0.5
```

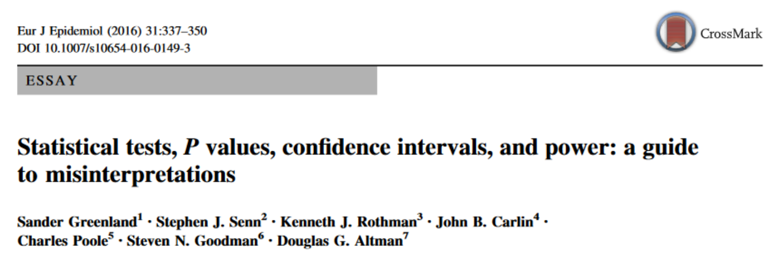
https://pollev.com/franciscorod726

# 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

CrossMark

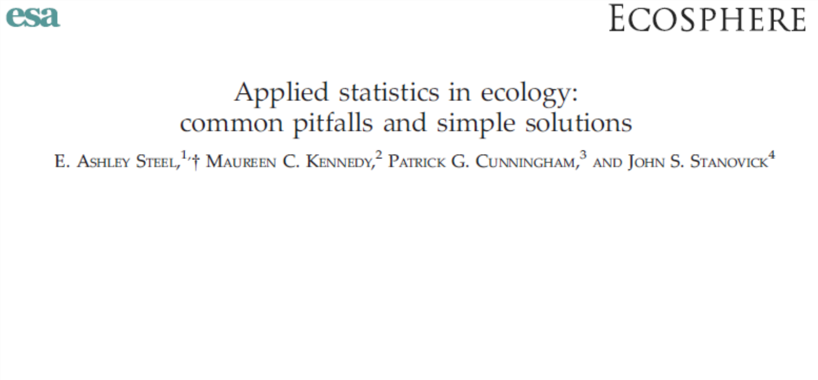
ESSAY

## 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

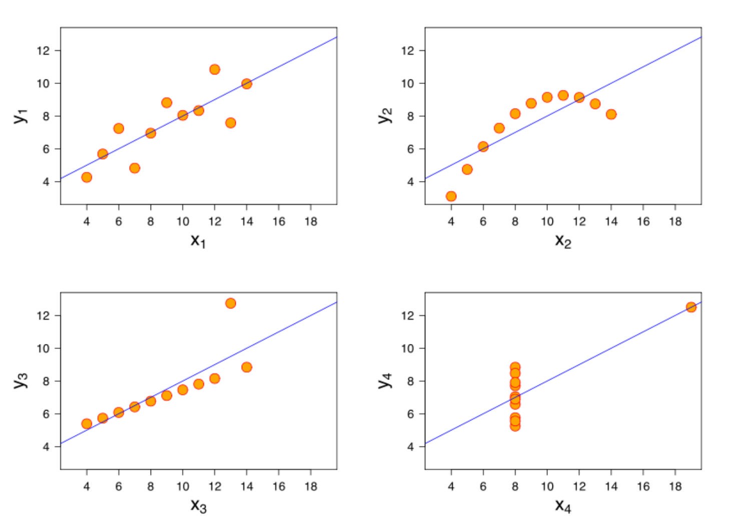
# First things first

- Always
- Always

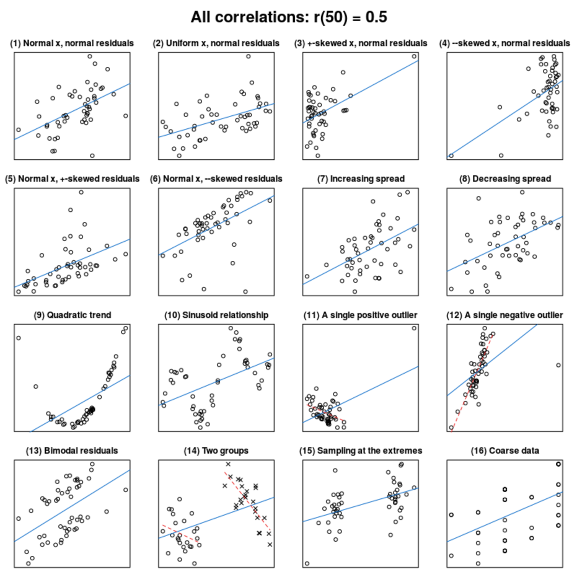
# First things first

- Always
- Always
- Always

# Plot data and models

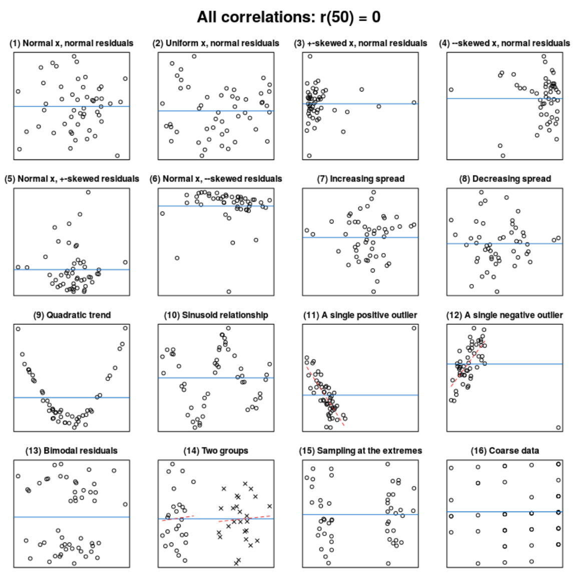


# Don’t use statistics blindly: *Visualise*



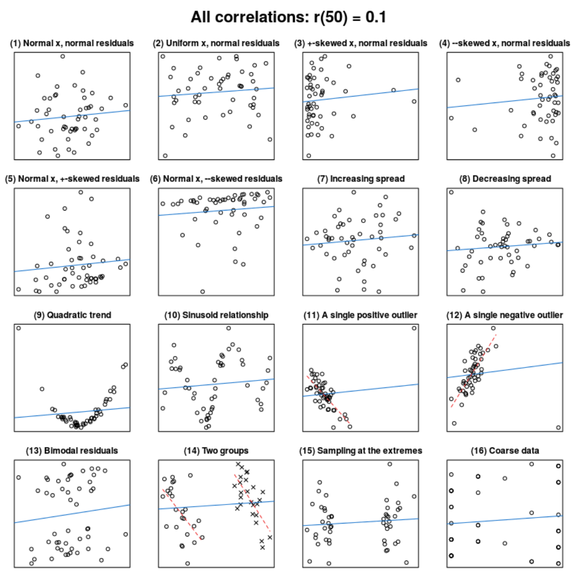
https://janhove.github.io/teaching/2016/11/21/what-

# Don’t use statistics blindly: *Visualise*



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

# Don’t use statistics blindly: *Visualise*



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

*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...

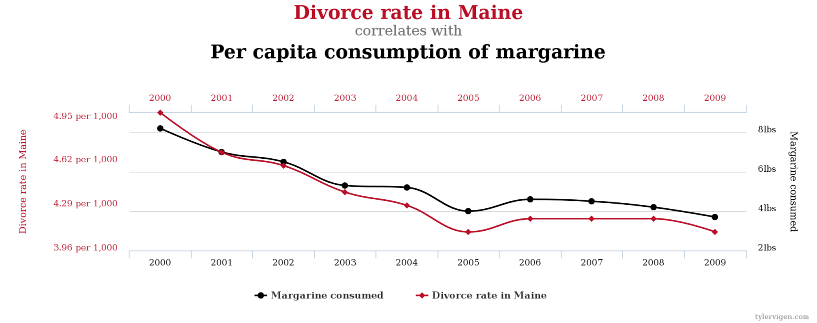
# Bigger flowers increase reproductive success

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

# Bigger flowers increase reproductive success

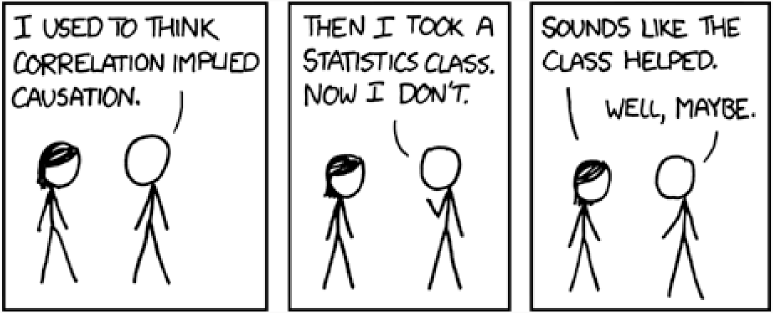
- We found that plants with big flowers produced 30% more seeds…
- **Do big flowers increase reproductive success?**
- `https://pollev.com/franciscorod726`

# Correlation vs Causation



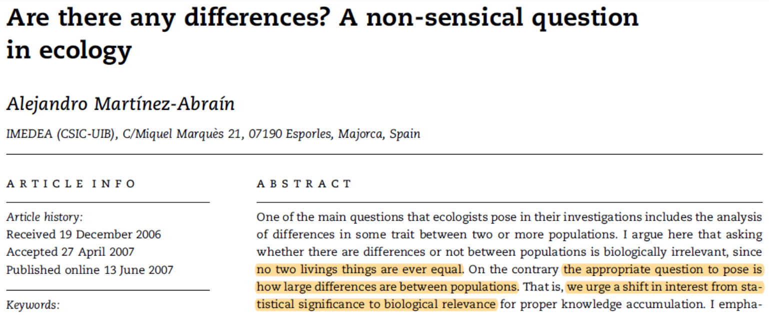
http://tylervigen.com/spurious-correlations

# Learning statistics through xkcd


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

# 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

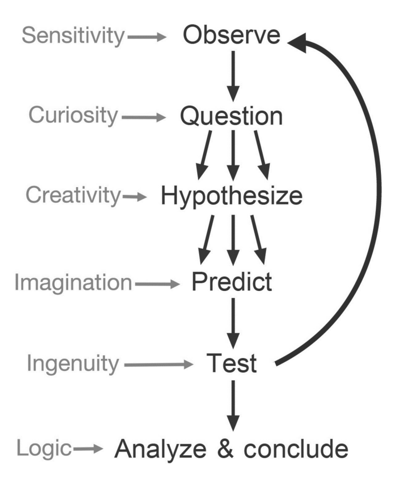
*Article history:*
Received 19 December 2006
Accepted 27 April 2007
Published online 13 June 2007

*Keywords:*

ABSTRACT

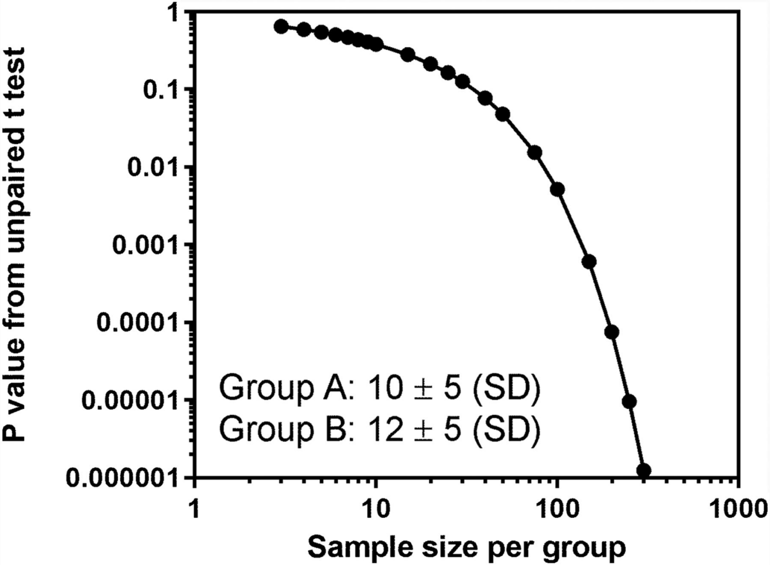
One of the main questions that ecologists pose in their investigations includes the analysis of differences in some trait between two or more populations. I argue here that asking whether there are differences or not between populations is biologically irrelevant, since no two livings things are ever 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 null model, compare meaningful models



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

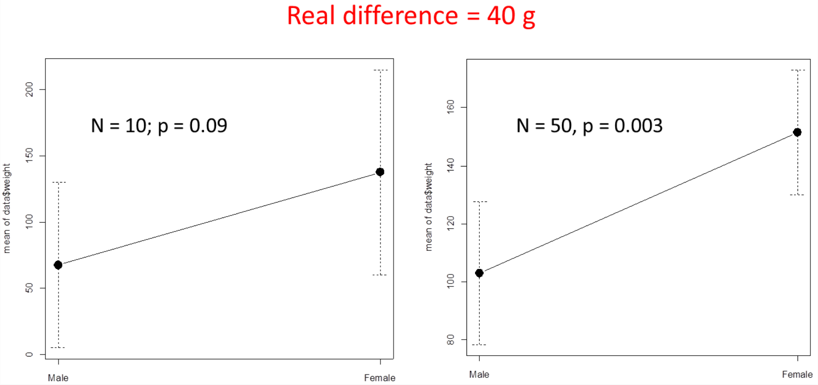
# P-value depends on sample size



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

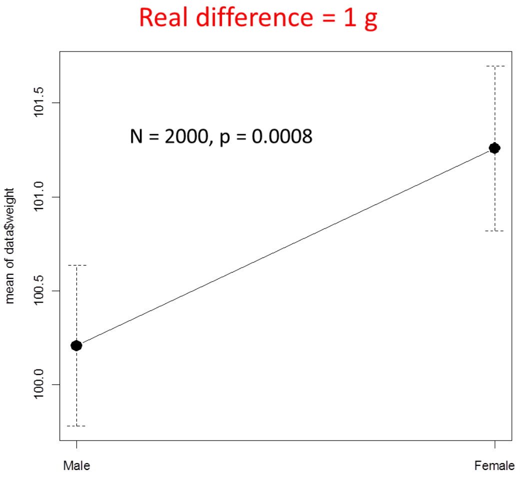
# P-value depends on sample size

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



# Statistically significant != biologically important

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



# Statistically significant != biologically important

- Statistically significant = unlikely to be zero

# Statistically significant != biologically important

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

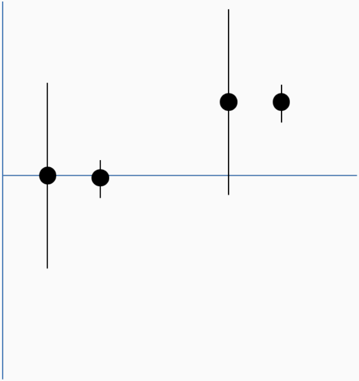
# Statistically significant != biologically important

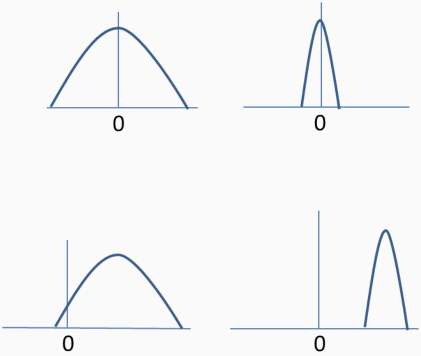
- Statistically significant = unlikely to be zero
- Good read: *significantly misleading*
- My suggestion: avoid significant/not significant (and maybe p-values too)

# 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*.

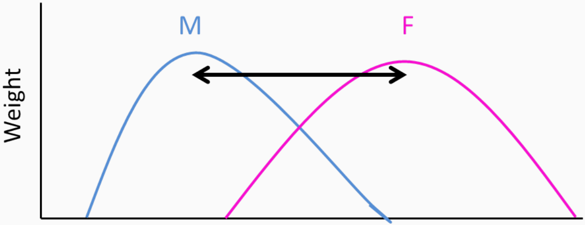
# ‘Not significant’ does NOT mean ‘there is no effect’



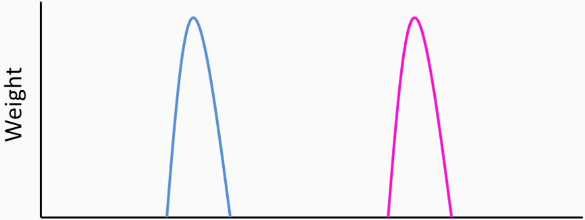


- Absence of evidence != Evidence of absence

# 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

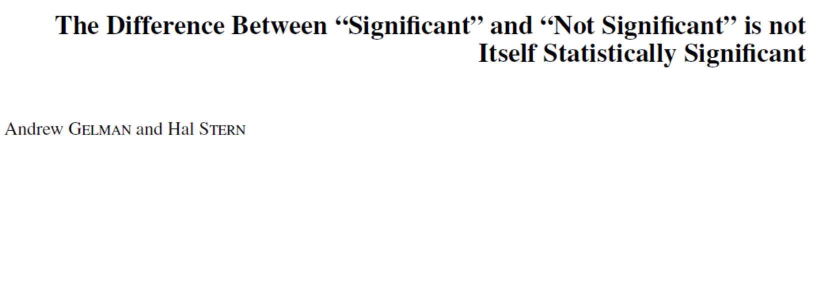
https://www.statisticsdonewrong.com/power.html#the-wrong-turn-on-red

# 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

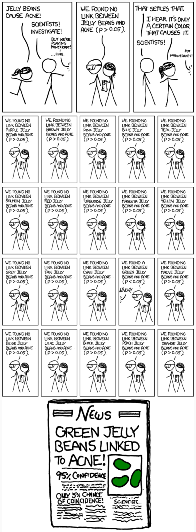


**The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant**

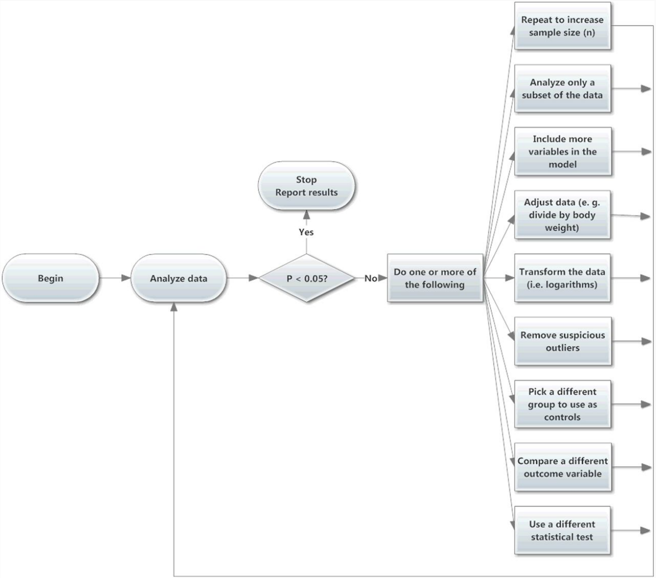
Andrew GELMAN and Hal STERN

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

# Multiple hypothesis testing


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...

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


Begin
Analyze data
P < 0.05?
Yes
Stop
Report results
No
Do one or more of the following
Repeat to increase sample size (n)
Analyze only a subset of the data
Include more variables in the model
Adjust data (e. g. divide by body weight)
Transform the data (i.e. logarithms)
Remove suspicious outliers
Pick a different group to use as controls
Compare a different outcome variable
Use a different statistical test

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

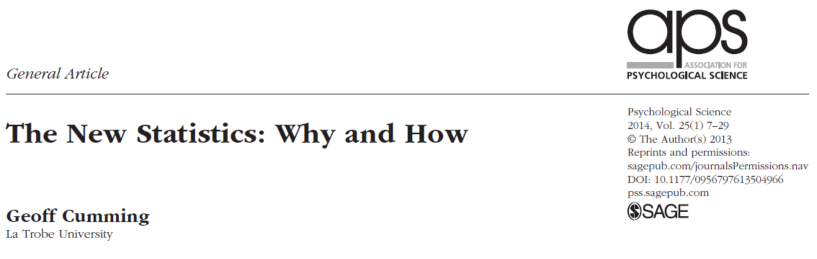
# ASA statement on p-values

- 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

## The New Statistics: Why and How

**Geoff Cumming**
La Trobe University

Psychological Science
2014, Vol. 25(1) 7–29
© The Author(s) 2013
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797613504966
pss.sagepub.com
SAGE

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