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 biology, 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

A R T I C L E I N F O

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

Keywords:

A B S T R A C T

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 observed data were produced by chance

https://pollev.com/franciscorod726

What is the p-value?

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

https://pollev.com/franciscorod726

What is the p-value?

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

https://pollev.com/franciscorod726

What is the p-value?

- The probability that the observed data were produced by chance
- The probability of getting results at least as extreme as observed if H0 was true
- The probability of null hypothesis being true
- The probability of alternative 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 every model assumption were correct*

P-value

- Very complicated concept: even statisticians fail to describe it well.
- Probability of observing data as or more extreme than these *if every model assumption were correct*
- What assumptions?

P-value

- Very complicated concept: even statisticians fail to describe it well.
- Probability of observing data as or more extreme than these *if every model assumption were correct*
- What assumptions?
    - Null hypothesis is true

P-value

- Very complicated concept: even statisticians fail to describe it well.
- Probability of observing data as or more extreme than these *if every model assumption were correct*
- What assumptions?
    - Null hypothesis is true
    - No uncontrolled sources of bias (measurement or programming error, p-hacking, etc)

A very small p-value does NOT mean H0 is automatically false

- A famous experiment found neutrinos faster than light

A very small p-value does NOT mean H0 is automatically false

- A famous experiment found neutrinos faster than light
- p-value < 10^{-7} -> reject null hypothesis of equal speed

A very small p-value does NOT mean H0 is automatically false

- A famous experiment found neutrinos faster than light
- p-value < 10^{-7} -> reject null hypothesis of equal speed
- In reality, measurement error (loose cable)

Testing a new fertiliser: Does it work?

10 treatment, 10 control plots

Crop biomass (treatment): $10kg \pm 2kg$

Crop biomass (control): $6kg \pm 2kg$

```
    Welch Two Sample t-test

data:  tr and ct
t = 4.1858, df = 17.872, p-value = 0.0005631
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.857902 5.606112
sample estimates:
mean of x mean of y
10.149251  6.417244
```

Testing a new fertiliser: Does it work?

10 treatment, 10 control plots

Crop biomass (treatment): $10kg \pm 2kg$

Crop biomass (control): $9kg \pm 2kg$

```
    Welch Two Sample t-test

data:  tr and ct
t = 0.82102, df = 17.872, p-value = 0.4225
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.142098  2.606112
sample estimates:
mean of x mean of y
10.149251  9.417244
```

How to interpret P-values

- Low P-value: data unlikely if every model assumption (including H0) were true.

How to interpret P-values

- Low P-value: data unlikely if every model assumption (including H0) were true.
- Large P-value: data not unusual if every model assumption (including H0) were true.

How to interpret P-values

- **Low P-value**: data unlikely if every model assumption (including H0) were true.
- **Large P-value**: data not unusual if every model assumption (including H0) were true.
- A very small P-value does not tell us which model assumption is incorrect:

How to interpret P-values

- **Low P-value**: data unlikely if every model assumption (including H0) were true.
- **Large P-value**: data not unusual if every model assumption (including H0) were true.
- A very small P-value does not tell us which model assumption is incorrect:
    - Could be that H0 is not true

How to interpret P-values

- **Low P-value**: data unlikely if every model assumption (including H0) were true.
- **Large P-value**: data not unusual if every model assumption (including H0) were true.
- A very small P-value does not tell us which model assumption is incorrect:
  - Could be that H0 is not true
  - But also that some auxiliary assumption is not true (e.g. sampling not random, measurement error, p-hacking...)

How to interpret P-values

- **Low P-value**: data unlikely if every model assumption (including H0) were true.
- **Large P-value**: data not unusual if every model assumption (including H0) were true.
- A very small P-value does not tell us which model assumption is incorrect:
    - Could be that H0 is not true
    - But also that some auxiliary assumption is not true (e.g. sampling not random, measurement error, p-hacking...)
- See Greenland et al 2016

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”?

Common practice:

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

Are differences “significant”?

Common practice:

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

Are differences "significant"?

Common practice:

- 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”?

Common practice:

- 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’d rather **avoid binary decisions** based on **arbitrary thresholds**?

Are differences “significant”?

Common practice:

- 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’d rather **avoid binary decisions** based on **arbitrary thresholds**?

Are differences “significant”?

Common practice:

- 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’d rather **avoid binary decisions** based on **arbitrary thresholds**?

MENU nature Subs

EDITORIAL · 20 MARCH 2019

It’s time to talk about ditching statistical significance

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

S (Surprise) values as alternative to p-values

- A p-value of 0.05 is as surprising as getting 4.3 heads in a row when flipping a coin.

Rafi & Greenland 2020

S (Surprise) values as alternative to p-values

- A p-value of 0.05 is as surprising as getting 4.3 heads in a row when flipping a coin.
- A p-value of 0.01 is as surprising as getting 6.6 heads in a row when flipping a coin.

Rafi & Greenland 2020

S (Surprise) values as alternative to p-values

- A p-value of 0.05 is as surprising as getting 4.3 heads in a row when flipping a coin.
- A p-value of 0.01 is as surprising as getting 6.6 heads in a row when flipping a coin.
- A p-value of 0.001 is as surprising as getting 10 heads in a row when flipping a coin.

Rafi & Greenland 2020

Are these two groups different?

```r
t.test(group.A, group.B)
```

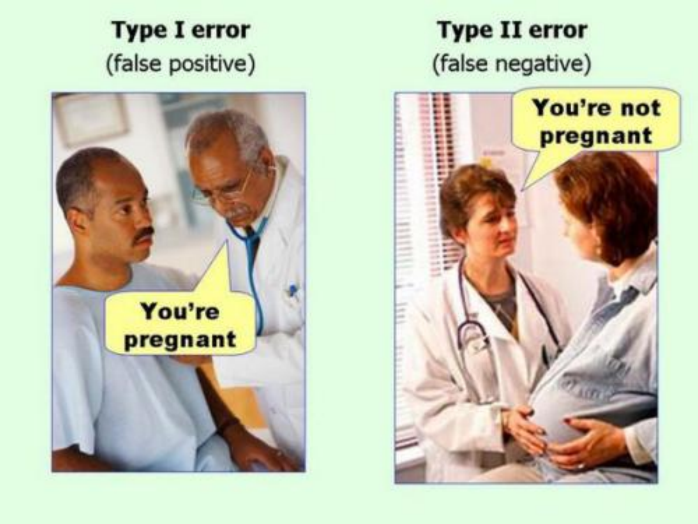
```

	Welch Two Sample t-test

data:  group.A and group.B
t = -5.0271, df = 12.464, p-value = 0.0002632
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -24.91029  -9.88971
sample estimates:
mean of x mean of y 
    162.8     180.2 
```

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).

Is this coin biased?

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


    1-sample proportions test with continuity correction

data:  sum(coin) out of ntrials, null probability 0.5
X-squared = 0.1, df = 1, p-value = 0.7518
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.2736697 0.8630694
sample estimates:
  p
0.6
```

https://pollev.com/franciscorod726

Understanding NHST

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

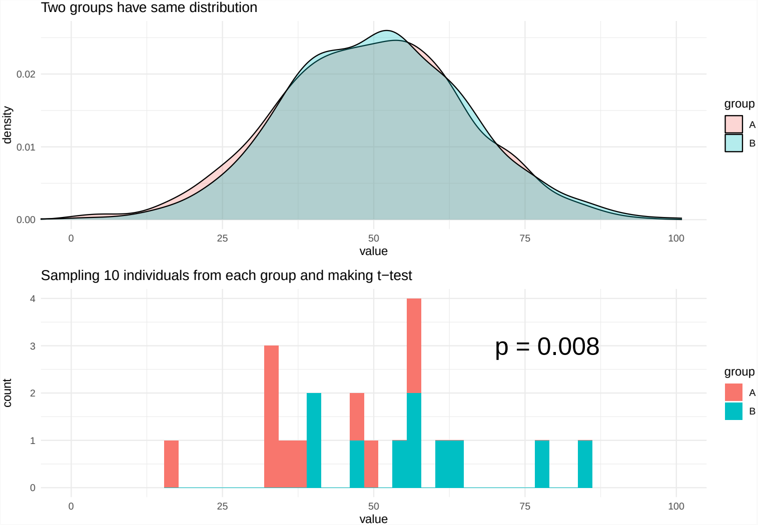
https://lakens.github.io/statistical_inferences/01-pvalue.html

NHST and p-values: common pitfalls

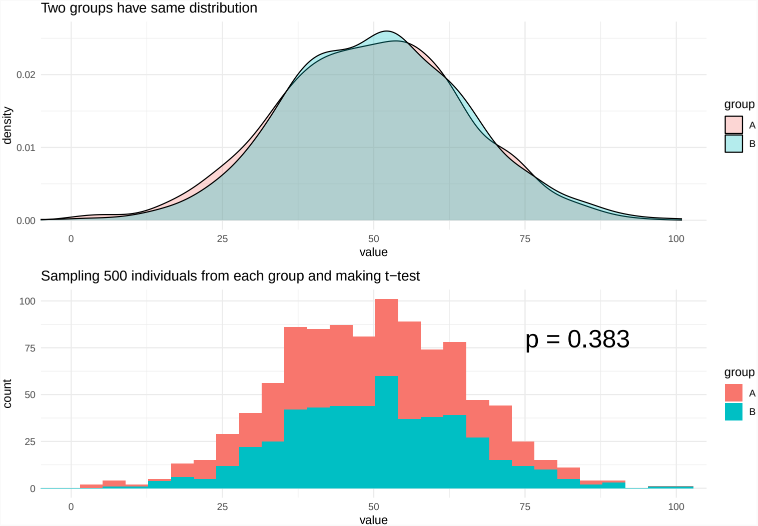
A significant p-value
does NOT mean
we found a true difference

A significant p-value does not mean we found a true difference

Particularly with low sample sizes



If sample size was larger...



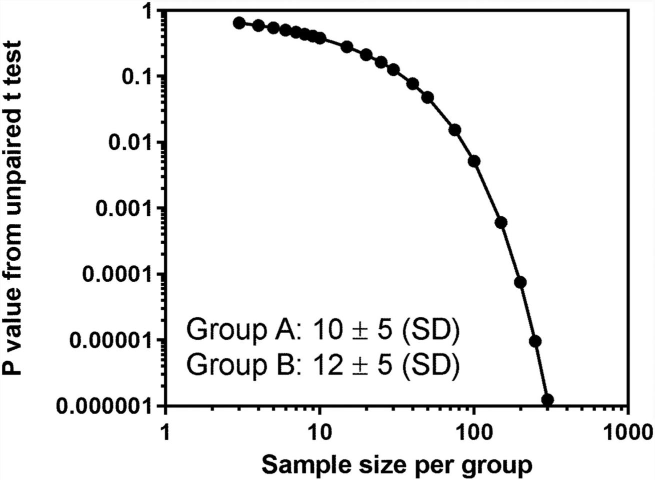
With low sample size (power),

significant p-values

are most likely overestimates

Loken & Gelman 2014, Vasisth et al. 2018

P-value depends on sample size

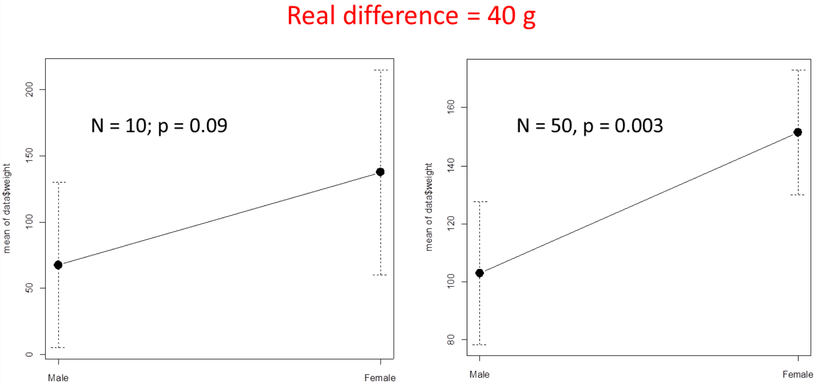


https://doi.org/10.1002/prp2.93

Try yourself here!

P-value depends on sample size

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



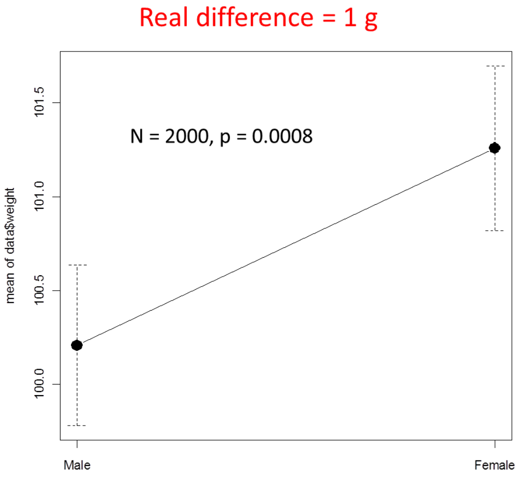
Statistically significant

!=

biologically important

Statistically significant != biologically important

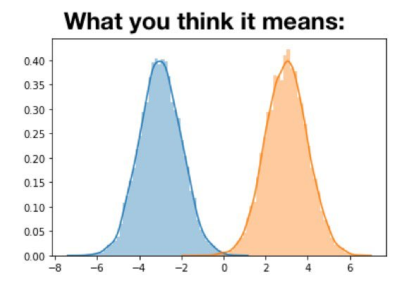
With big sample size, we can find **highly significant but biologically unimportant** differences.

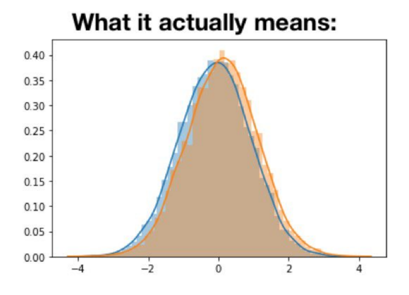


Statistically significant != biologically important

zara weinberg
@weinberz

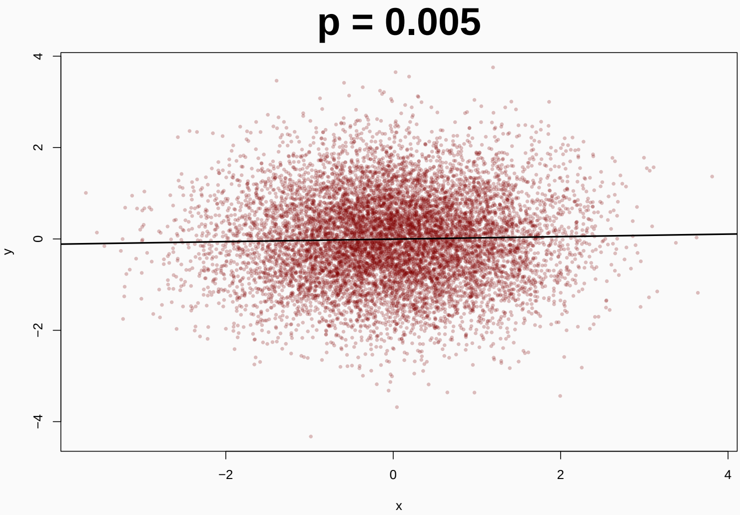
friendly reminder about $p < 0.0001$:





https://twitter.com/weinberz/status/1422405165236178947?s=20

Statistically significant != biologically important



Statistically significant != biologically important

- Statistically significant = unlikely to be zero

Statistically significant != biologically important

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

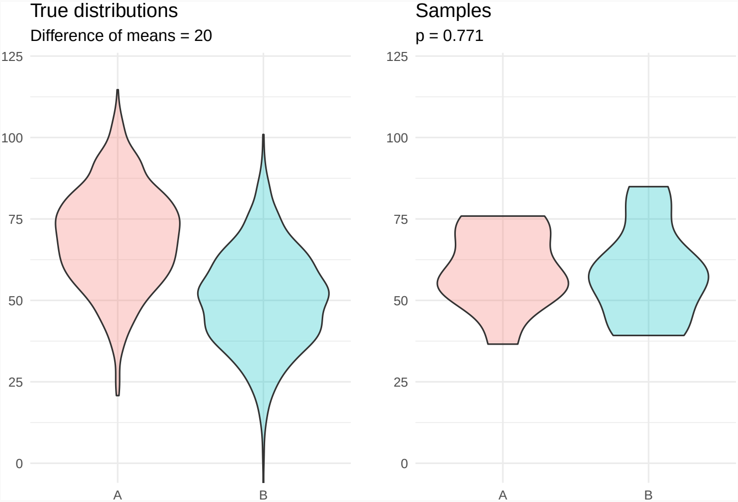
Statistically significant != biologically important

- Statistically significant = unlikely to be zero
- Good read: *significantly misleading*
- Beyond significant/not significant, look at **effect sizes and their uncertainty**.

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

‘Not significant’ does NOT mean ‘they are equal’

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



Failure to reject H0 != H0 is true

Absence of evidence != Evidence of absence

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
- Failure to reject H0 does NOT mean H0 is true!

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
- Failure to reject H0 does NOT mean H0 is true!
- Misinterpretation of underpowered study cost lives

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

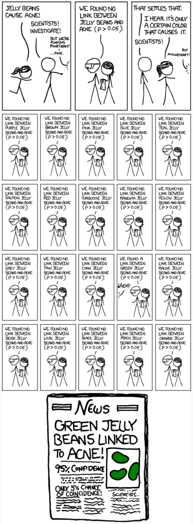
0.05 is an arbitrary threshold

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

Andrew GELMAN and Hal STERN

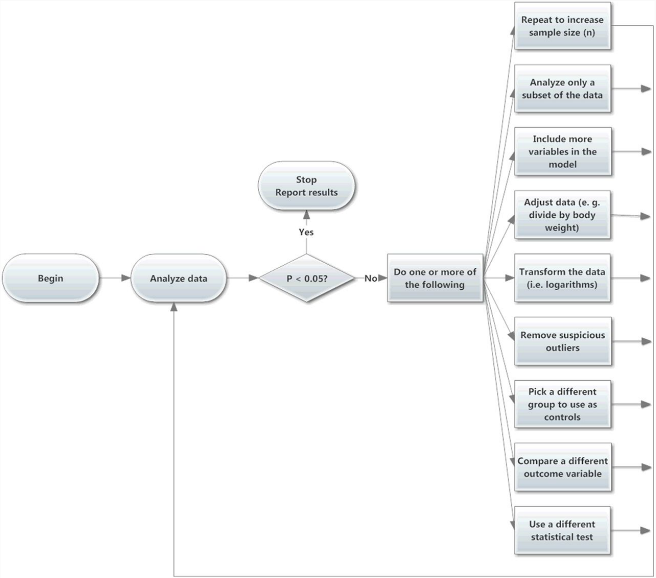
https://dx.doi.org/10.1198/000313006X152649

Multiple hypothesis testing



https://xkcd.com/882

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

p-hacking: try it yourself

https://www.shinyapps.org/apps/p-hacker/

https://shiny.psy.lmu.de/felix/ShinyPHack/

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

See also https://lakens.github.io/statistical_inferences/01-pvalue.html

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

See also https://lakens.github.io/statistical_inferences/01-pvalue.html

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

See also https://lakens.github.io/statistical_inferences/01-pvalue.html

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

See also https://lakens.github.io/statistical_inferences/01-pvalue.html

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

The New Statistics

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

General Article

The New Statistics: Why and How

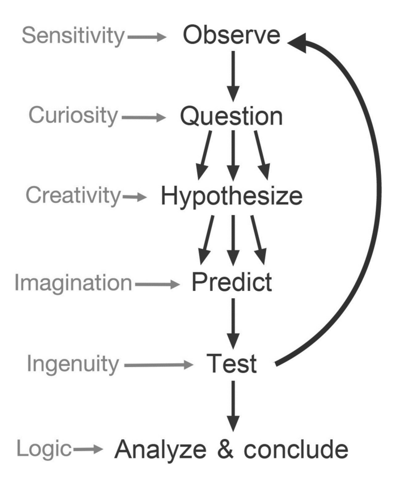
Geoff Cumming
La Trobe University

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

Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797613504966
pss.sagepub.com
SAGE

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

Instead of falsifying null model, compare meaningful models



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

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