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.

P value

- Tell me. . .

P value

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

P value

- Tell me. . .
- 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

- Tell me…
- 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

- Tell me. . .
- 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

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

```

    Welch Two Sample t-test

data:  h.sevi and h.out
t = -0.35784, df = 4.7983, p-value = 0.7357
alternative hypothesis: true difference in means is not equal to
95 percent confidence interval:
 -19.03344  14.43344
sample estimates:
mean of x mean of y 
    174.2     176.5 
```

Are heights different then?

Rejecting hypotheses: two types of error

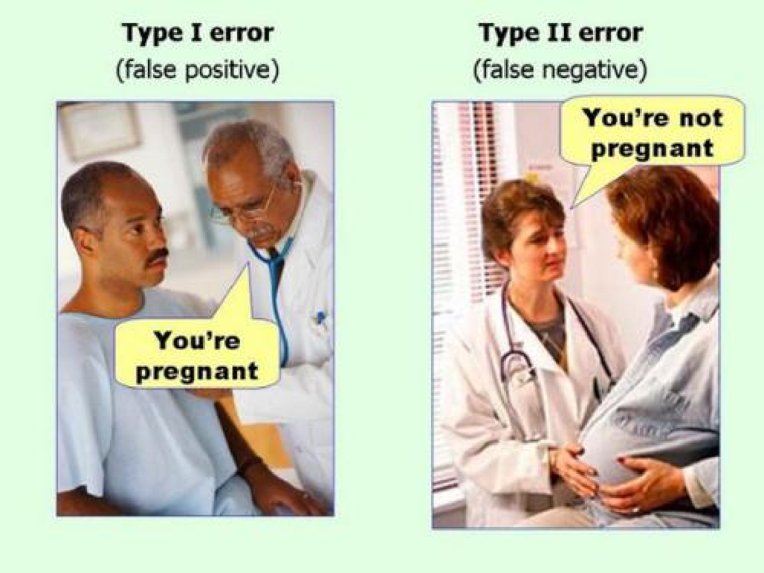
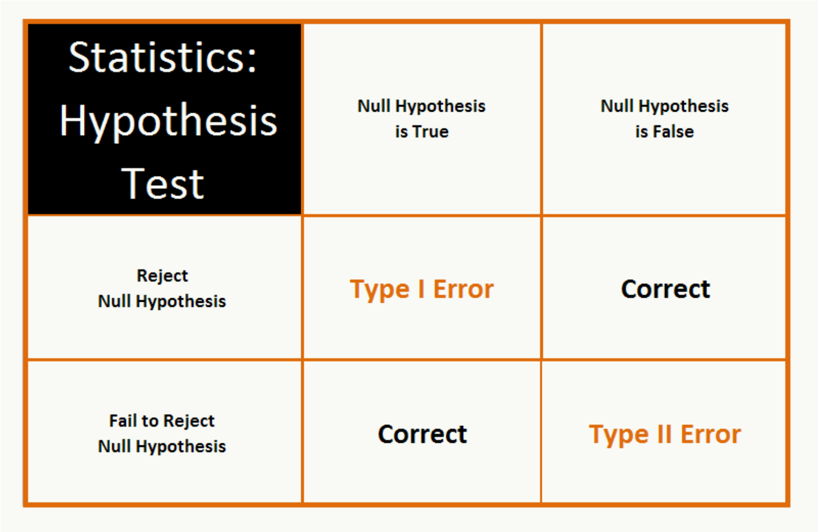


Figure 1:

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

Figure 2:

Understanding NHST

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

Example: biased coin

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

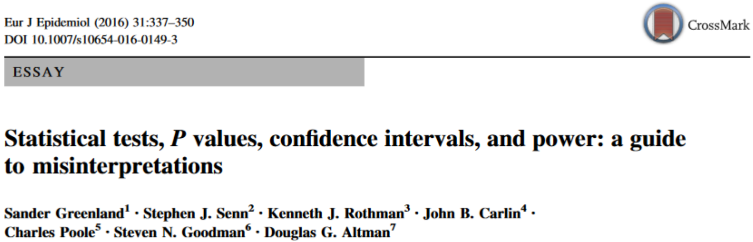
```
	1-sample proportions test with continuity correction

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:
 0.1369306 0.7263303
sample estimates:
  p 
0.4 
```

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

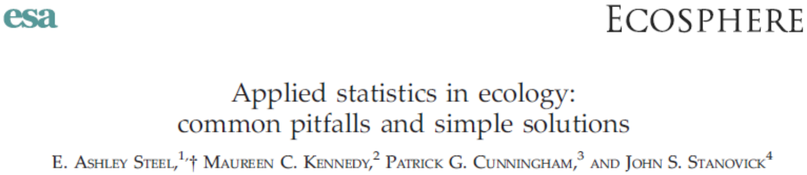
ESSAY

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 reading



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]

Figure 3:

http://dx.doi.org/10.1890/ES13-00160.1
Also http://www.statisticsdonewrong.com/

First things first

- Always

First things first

- Always
- Always

First things first

- Always
- Always
- Always

Plot data and models

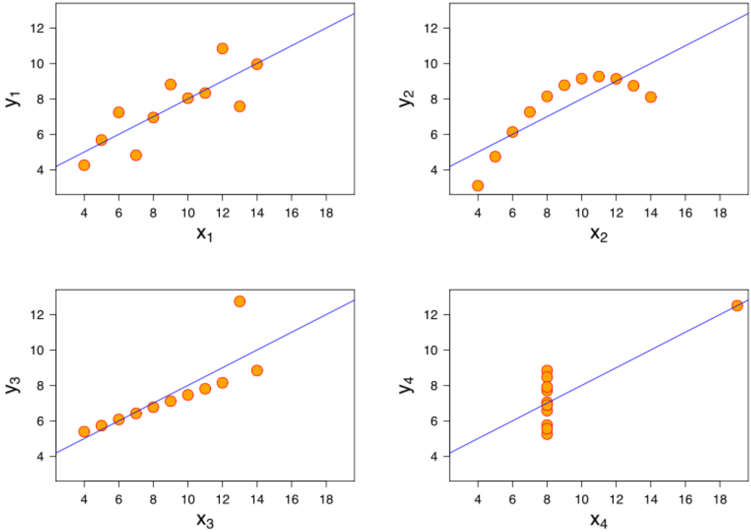
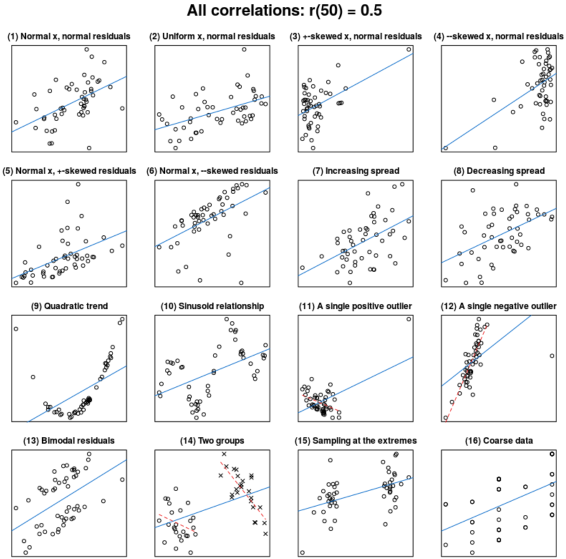


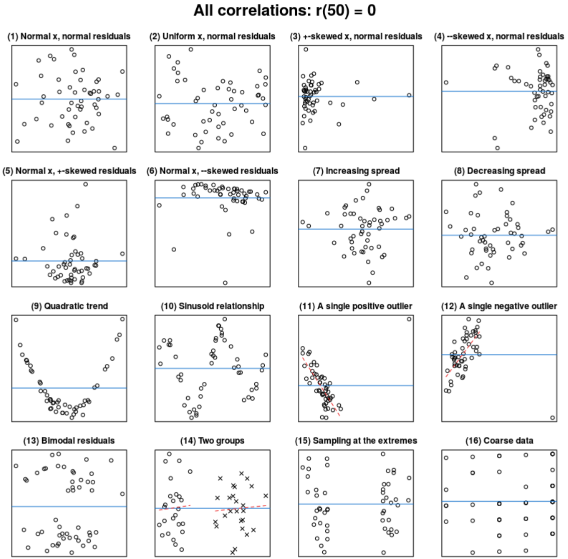
Figure 4:

Don't use statistics blindly: *Visualise*



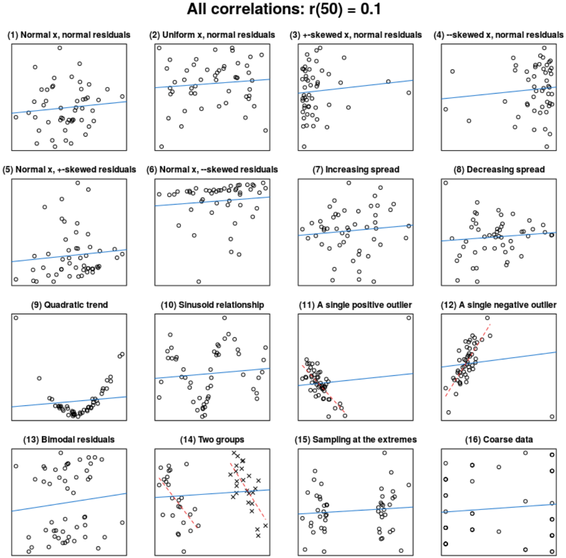
https:
//janhove.github.io/teaching/2016/11/21/what-correlations-look-like

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*

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

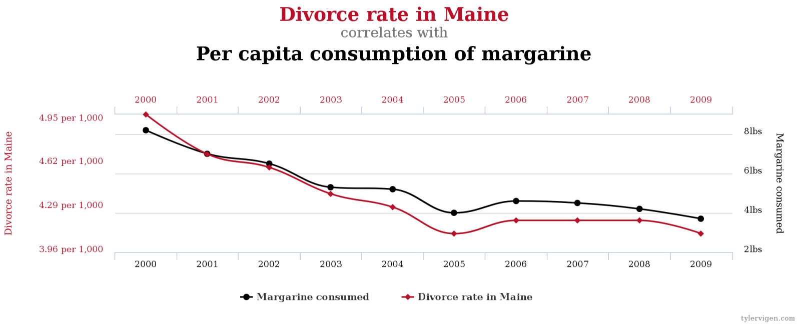
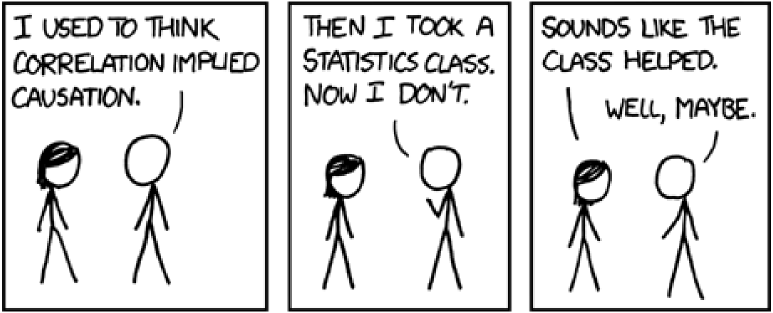


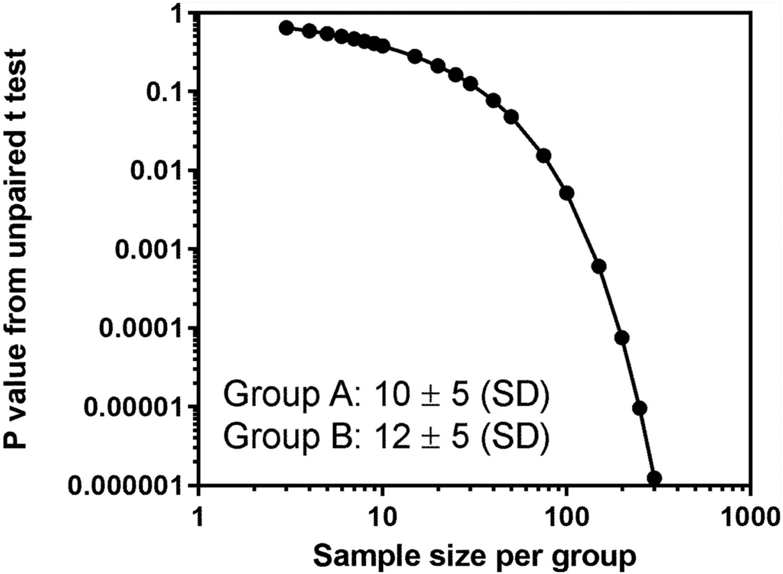
Figure 5:

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.

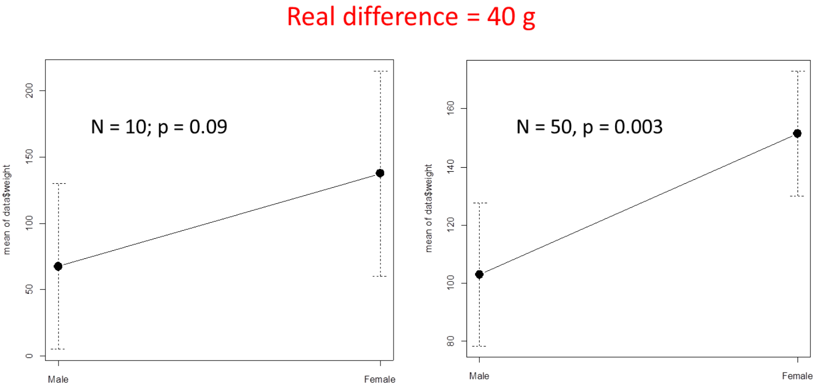
P-value depends on sample size



http://dx.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

N = 10; p = 0.09

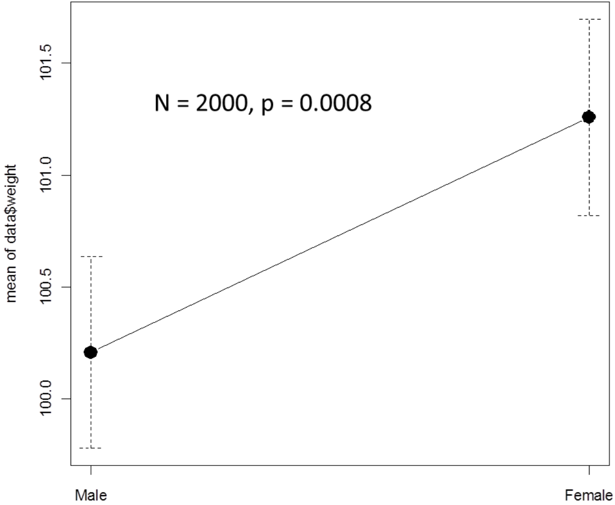
N = 50, p = 0.003

Figure 6:

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
- Suggested reading: *significantly misleading*

Statistically significant != biologically important

- Statistically significant = unlikely to be zero
- Suggested reading: *significantly misleading*
- Beyond significance, look at *effect sizes*.

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

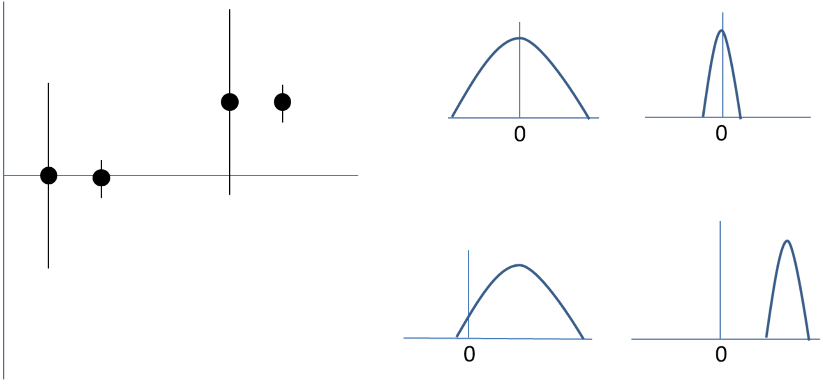


Figure 8:

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

Failure to reject H0 != H0 is true

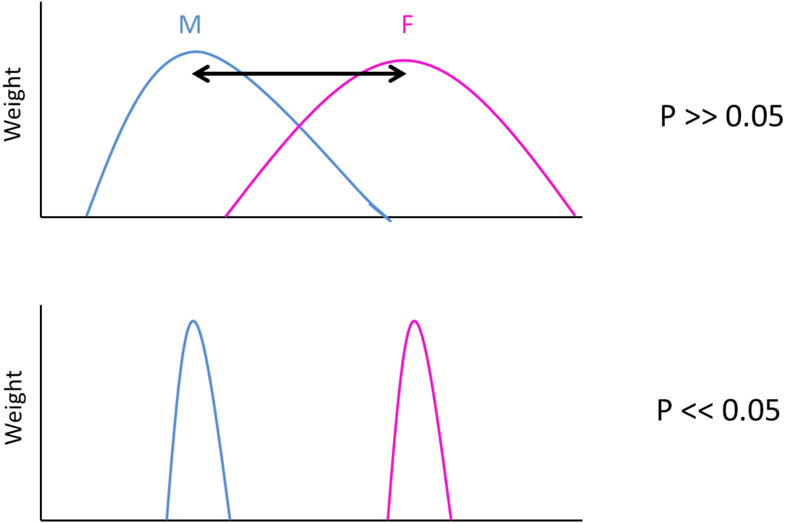
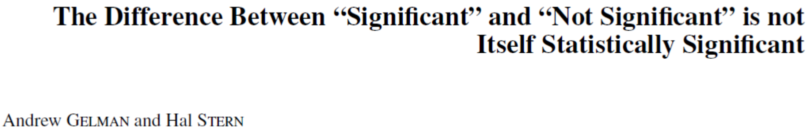


Figure 9:

0.05 is an arbitrary threshold



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

Andrew Gelman and Hal Stern

Figure 10:

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

Multiple hypothesis testing

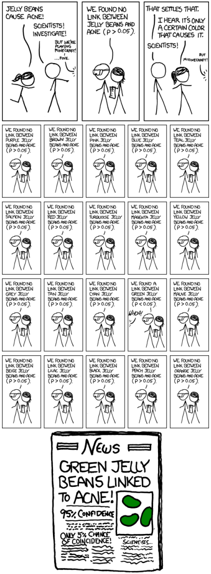
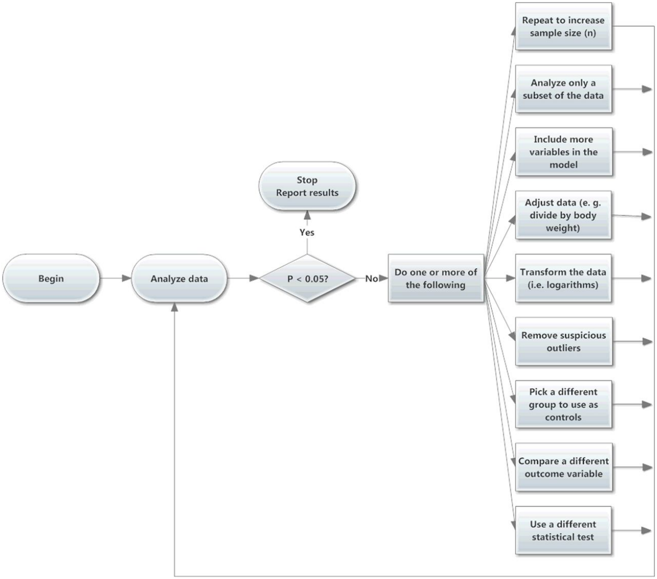

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
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...

Figure 11:

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

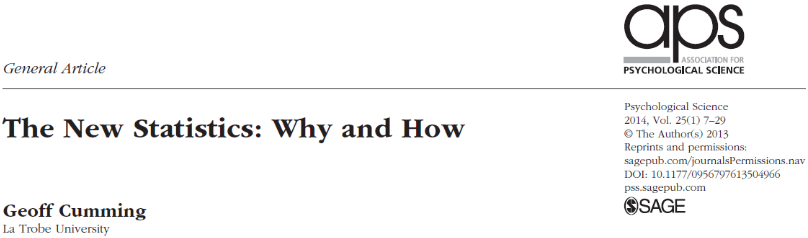
ASA statement on p-values

- 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.



General Article

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

aps
ASSOCIATION FOR
PSYCHOLOGICAL SCIENCE

Figure 12:

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