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

```
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
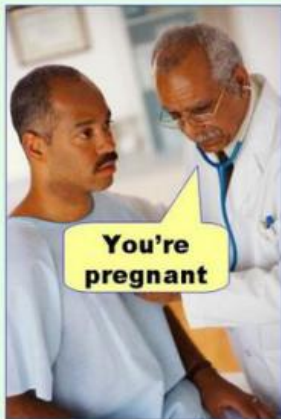
Figure 1:

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 0 0 0 1 1 1 1 0 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

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

Figure 3:

First things first

- Always

First things first

- Always
- Always

First things first
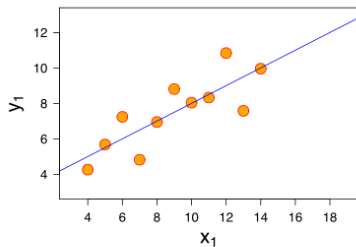
- Always
- Always
- Always

Plot data and models
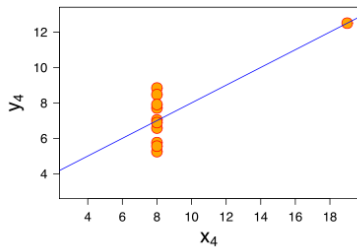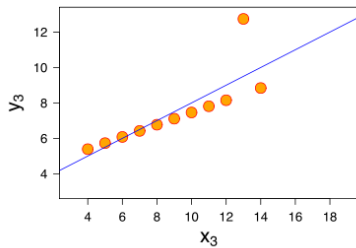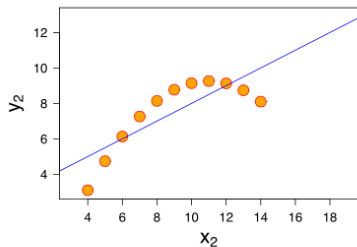
Figure 4:

Don't use statistics blindly: *Visualise*

All correlations: r(50) = 0.5

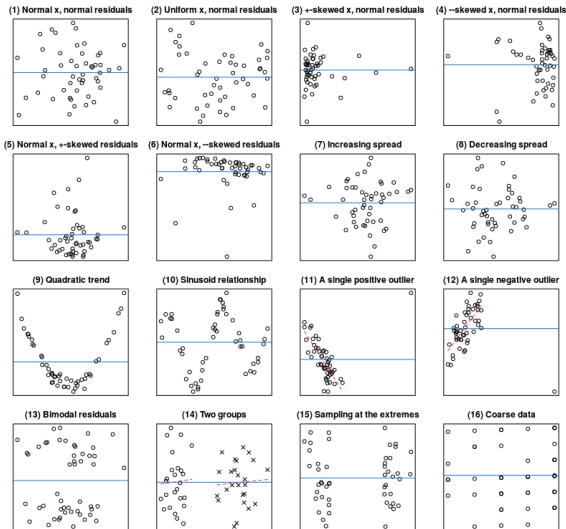# Don't use statistics blindly: *Visualise*

Don't use statistics blindly: *Visualise*

All correlations: r(50) = 0.1

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
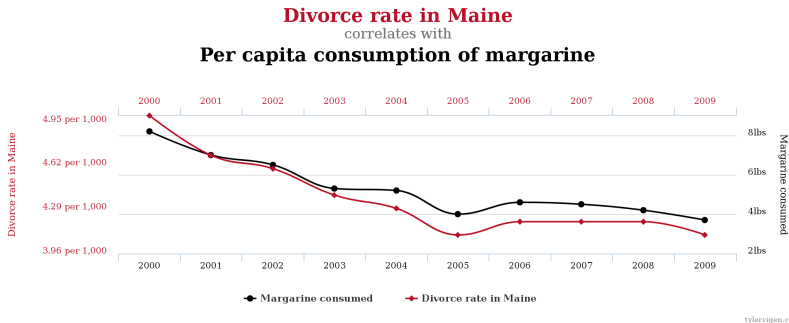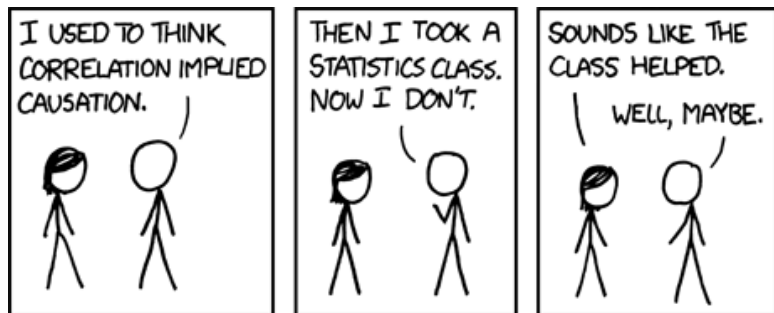

Figure 5:

http://tylervigen.com/spurious-correlations

Learning statistics through xkcd

P-value depends on sample size

▶ Same real difference is detected as significant or not depending on sample size:
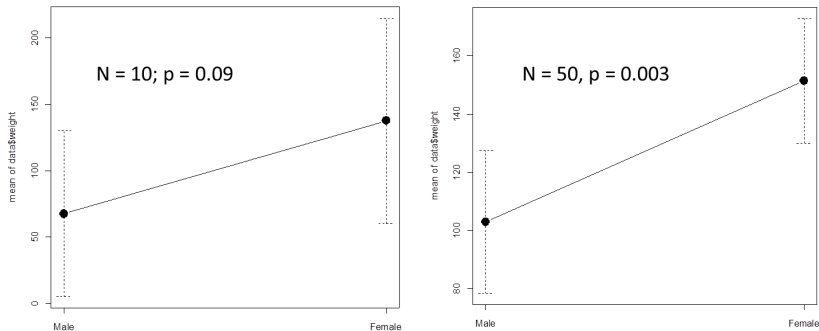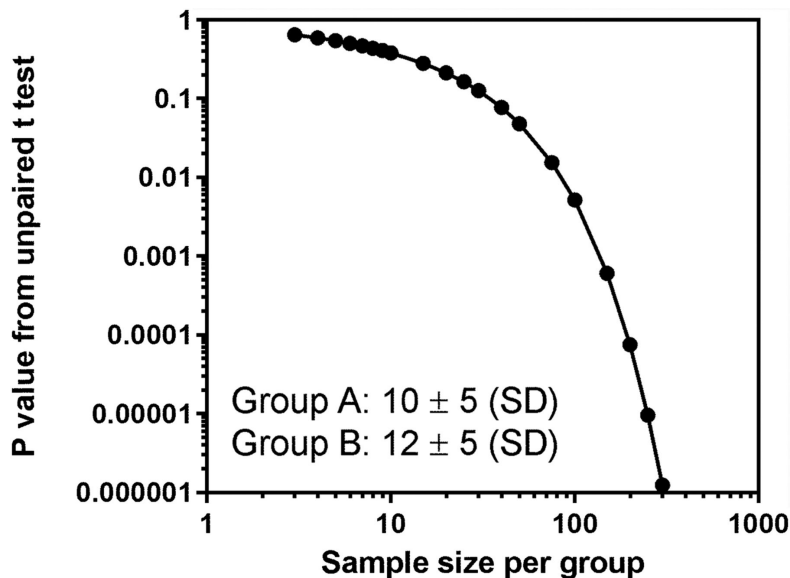
Real difference = 40 g

Figure 6:

P-value depends on sample size

Statistically significant != biologically important

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

Real difference = 1 g
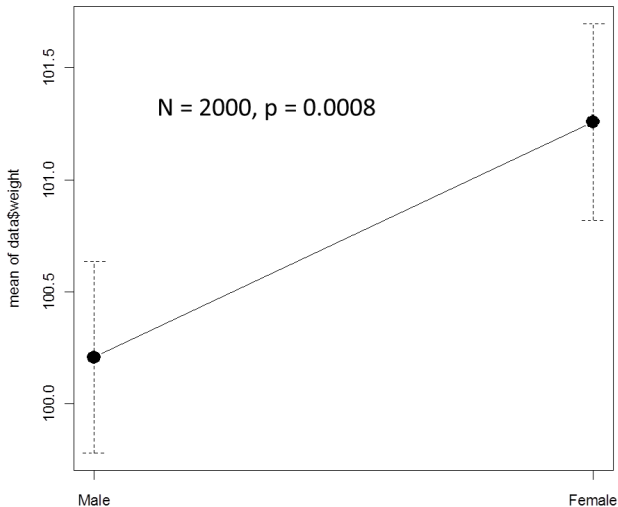
N = 2000, p = 0.0008

Statistically significant != biologically important

- Statistically significant = unlikely to be zero

Statistically significant != biologically important

- Statistically significant = unlikely to be zero
- Suggestion: Try to avoid 'significant' all together

Statistically significant != biologically important

- Statistically significant = unlikely to be zero
- Suggestion: Try to avoid 'significant' all together
- Suggested reading: *significantly misleading*

Statistically significant != biologically important

- Statistically significant = unlikely to be zero
- Suggestion: Try to avoid 'significant' all together
- Suggested reading: *significantly misleading*
- Beyond significance, look at *effect sizes*.

'Not significant' does NOT mean 'there is no effect'
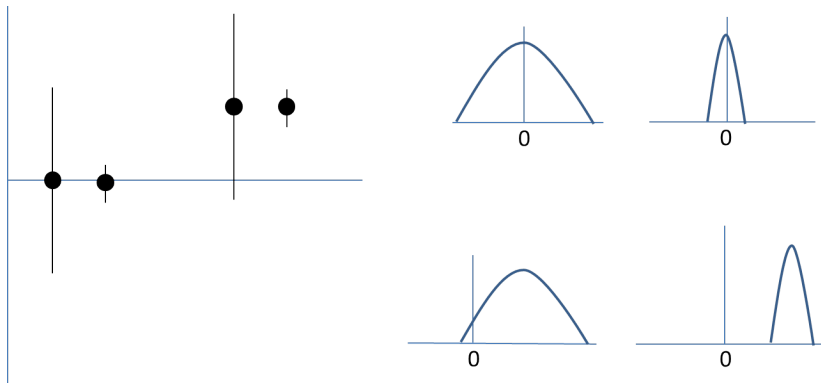
Figure 8:

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

Failure to reject H0 != H0 is true
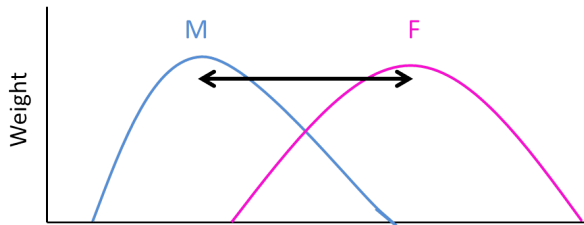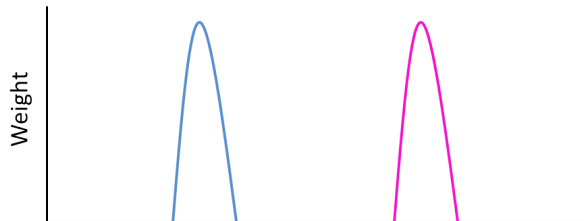
Figure 9:

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:

Multiple hypothesis testing
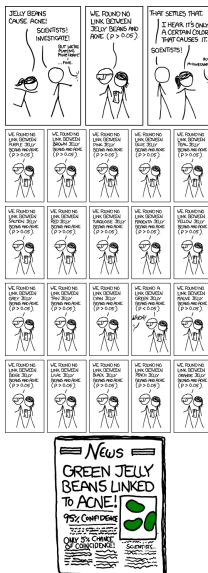
Figure 11:

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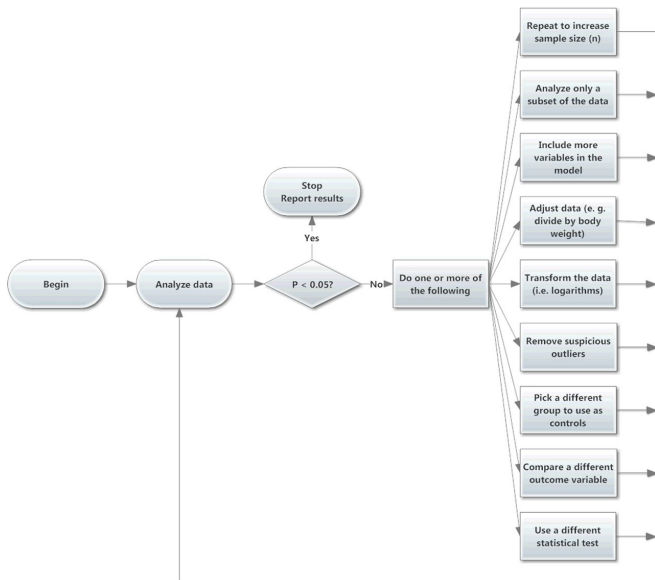
How to make your results significant: *p-hacking*

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.

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**.

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.

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.

The New Statistics

Aim for estimation of effects and their uncertainty.

The New Statistics: Why and How

Geoff Cumming
La Trobe University

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