

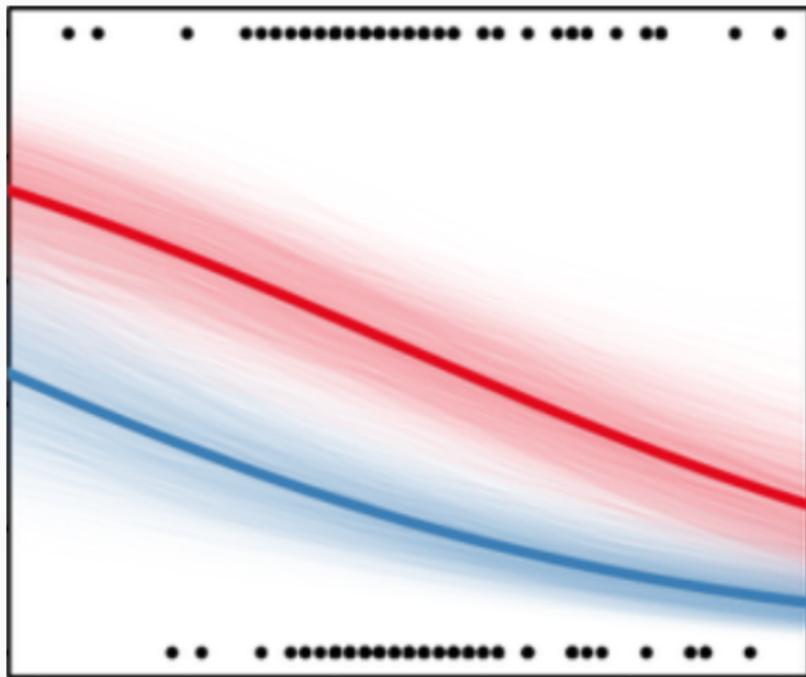
# An introduction to statistical inference

---

Francisco Rodríguez Sánchez

@frod\_san

<https://frodriguezsanchez.net>



# Why statistics?

---

To answer questions like...

- what's the probability that something occurs?

To answer questions like...

- what's the probability that something occurs?
- does X influence Y? How much?

To answer questions like...

- what's the probability that something occurs?
- does X influence Y? How much?
- can we predict Y knowing X, Z... How well?

# To ensure correct inferences

11	451	368	80	46	83	74	28	71	49	689	43
439	164	94	45	73	38	58	25	73	34	500	3
235	166	172	54	91	85	40	78	65	78	708	1
10	30	62	49	32	11	10	10	10	10	10	10
1.433	896	2.132	2.390	3.850	2.775	1.580	2.000	5.000	1.000	1.000	1.000
1.870	2.845	1.001	1.920	1.740	2.981	3.000	1.000	1.000	1.000	1.000	1.000
2.427	1.333	1.233	1.233	1.470	2.524	3.000	1.000	1.000	1.000	1.000	1.000
2.424	2.657	1.001	1.233	1.233	1.233	1.233	1.233	1.233	1.233	1.233	1.233
1.692	84	1.001	1.233	1.233	1.233	1.233	1.233	1.233	1.233	1.233	1.233
1.199	1.198	1.198	1.198	1.198	1.198	1.198	1.198	1.198	1.198	1.198	1.198
2.032	1.198	1.198	1.198	1.198	1.198	1.198	1.198	1.198	1.198	1.198	1.198
3	2.032	1.198	1.198	1.198	1.198	1.198	1.198	1.198	1.198	1.198	1.198
35	290	92	285	163	224	234	234	234	234	234	234
74	243	430	277	175	234	234	234	234	234	234	234
84	249	301	47	3.858	2.459	6.303	9	9	9	9	9

Inference



Bolker et al 2009 TREE:

'311 out of 537 GLMM analyses (58%) used these tools  
inappropriately'

To get answers to tough problems

How many seeds do trees produce?



# Inferring tree fecundity



# Course goals

- Understand statistical inference

# Course goals

- Understand statistical inference
- Avoid misconceptions

# Course goals

- Understand statistical inference
- Avoid misconceptions
- Promote good practices

*The purpose of models is not to fit data  
but to sharpen thinking*

Sam Karlin

# Topics

- Descriptive statistics

# Topics

- Descriptive statistics
- Graphics

# Topics

- Descriptive statistics
- Graphics
- Sampling

# Topics

- Descriptive statistics
- Graphics
- Sampling
- Experimental design

# Topics

- Descriptive statistics
- Graphics
- Sampling
- Experimental design
- Hypothesis testing

# Topics

- Descriptive statistics
- Graphics
- Sampling
- Experimental design
- Hypothesis testing
- Bayesian inference

# Topics

- Descriptive statistics
- Graphics
- Sampling
- Experimental design
- Hypothesis testing
- Bayesian inference
- Linear models & GLMs

# Topics

- Descriptive statistics
- Graphics
- Sampling
- Experimental design
- Hypothesis testing
- Bayesian inference
- Linear models & GLMs
- Model selection

# Descriptive statistics

---

Measure trunk diameter of 30 trees in your neighbourhood



## Read data

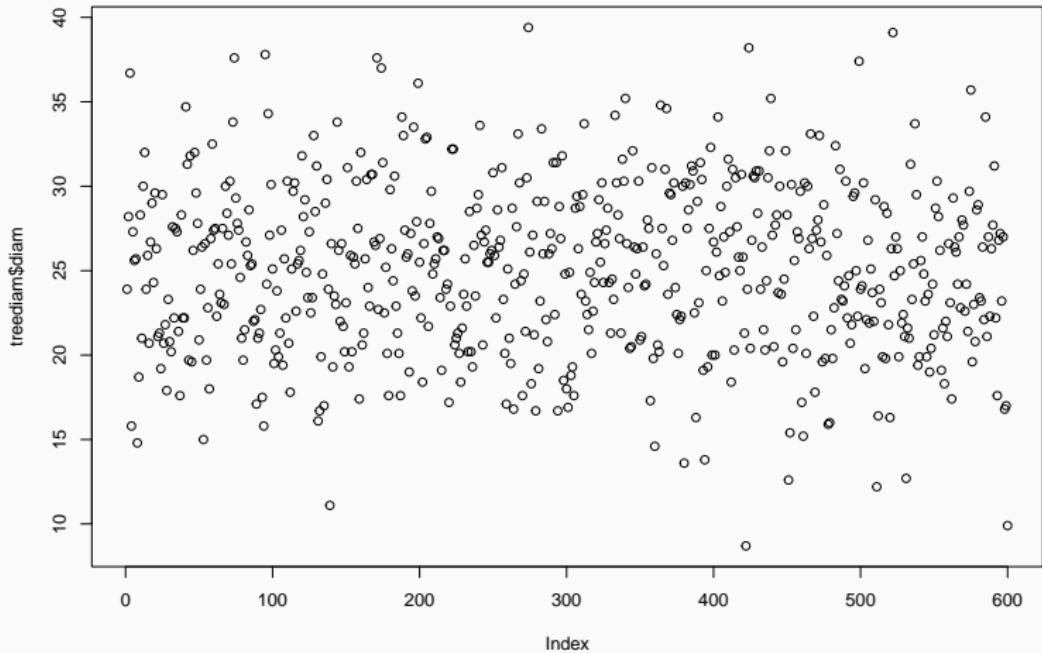
```
treediam <- read.csv("treediam.csv")
```

```
summary(treediam)
```

	site	tree	diam
Min.	: 1.00	Min. : 1.0	Min. : 8.70
1st Qu.	: 5.75	1st Qu.: 8.0	1st Qu.:21.40
Median	:10.50	Median :15.5	Median :25.25
Mean	:10.50	Mean :15.5	Mean :25.04
3rd Qu.	:15.25	3rd Qu.:23.0	3rd Qu.:28.40
Max.	:20.00	Max. :30.0	Max. :39.40

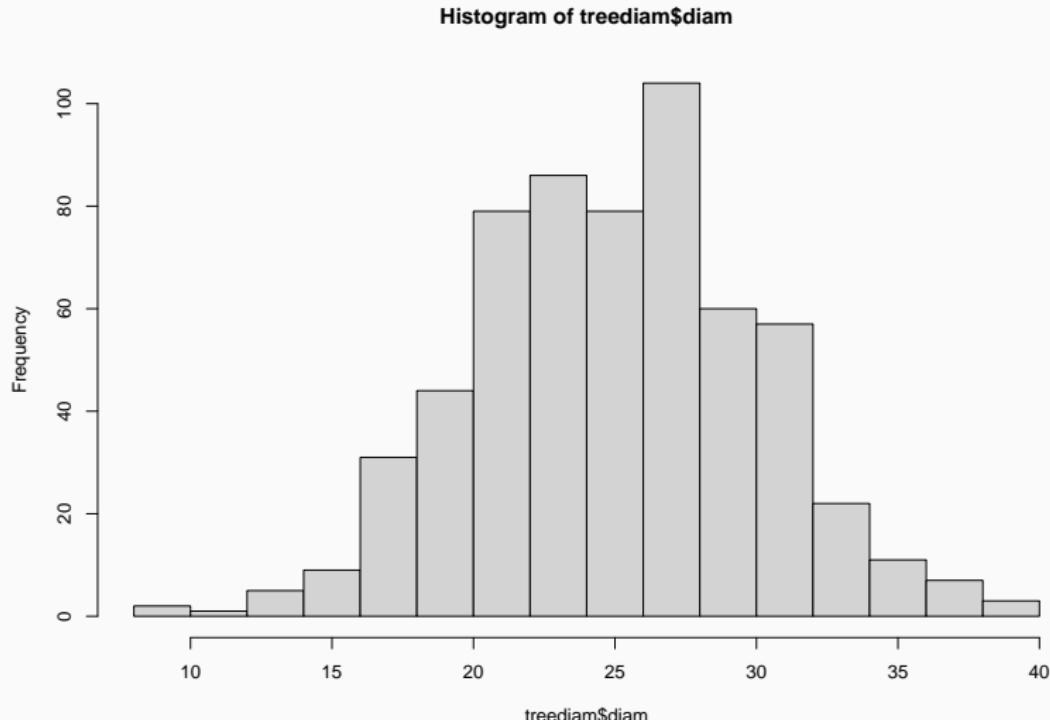
# Visualisation of tree diameters

```
plot(treediam$diam)
```



# Visualisation of tree diameters

```
hist(treediam$diam)
```



How well do these values  
represent actual tree diameters  
in your neighbourhood?

<https://pollev.com/franciscorod726>

- At what height did you measure?

- At what height did you measure?
- Did you include bark?

- At what height did you measure?
- Did you include bark?
- Did you measure with tape, caliper, by eye?

- At what height did you measure?
- Did you include bark?
- Did you measure with tape, caliper, by eye?
- When did you measure: dawn, midday, night?

- At what height did you measure?
- Did you include bark?
- Did you measure with tape, caliper, by eye?
- When did you measure: dawn, midday, night?
  - (trees may get thinner w/ high evapotranspiration)

- At what height did you measure?
- Did you include bark?
- Did you measure with tape, caliper, by eye?
- When did you measure: dawn, midday, night?
  - (trees may get thinner w/ high evapotranspiration)
- Where did you measure?

- At what height did you measure?
- Did you include bark?
- Did you measure with tape, caliper, by eye?
- When did you measure: dawn, midday, night?
  - (trees may get thinner w/ high evapotranspiration)
- Where did you measure?
  - (differences among streets, species, etc)



TRUTH



TRUTH



TRUTH

Data are hardly ever objective.

We decide **what to measure, when, where, and how**.

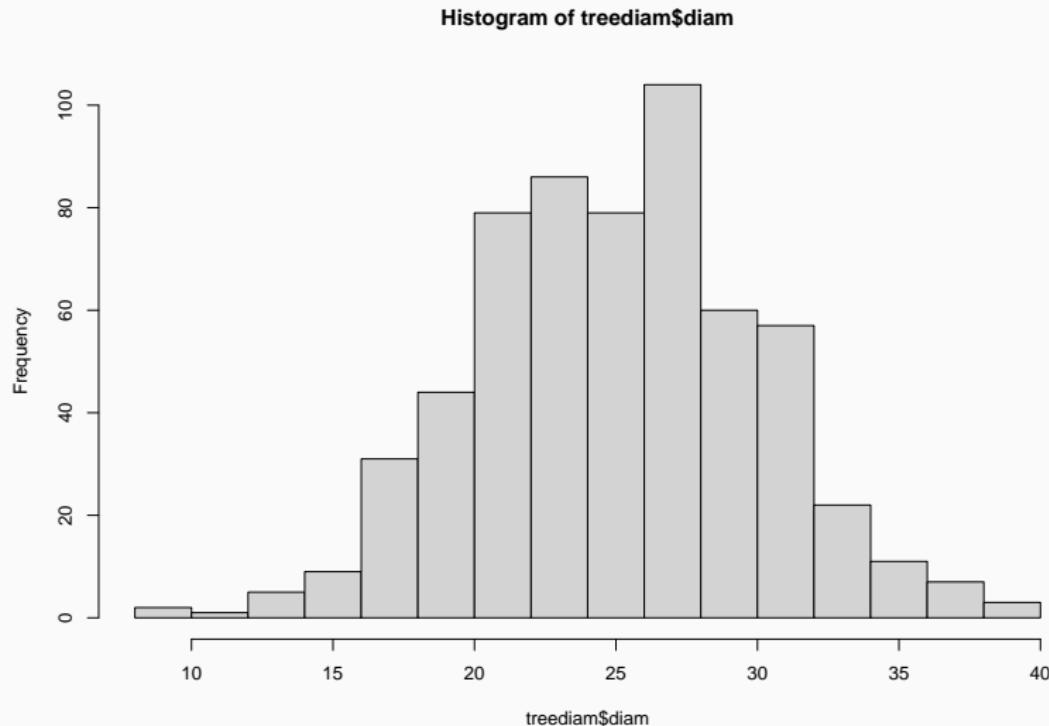
Always consider:

**How well do data reflect what we are trying to measure?**

## Describing your data

---

# How would you describe this distribution?



(Discuss with your partner)

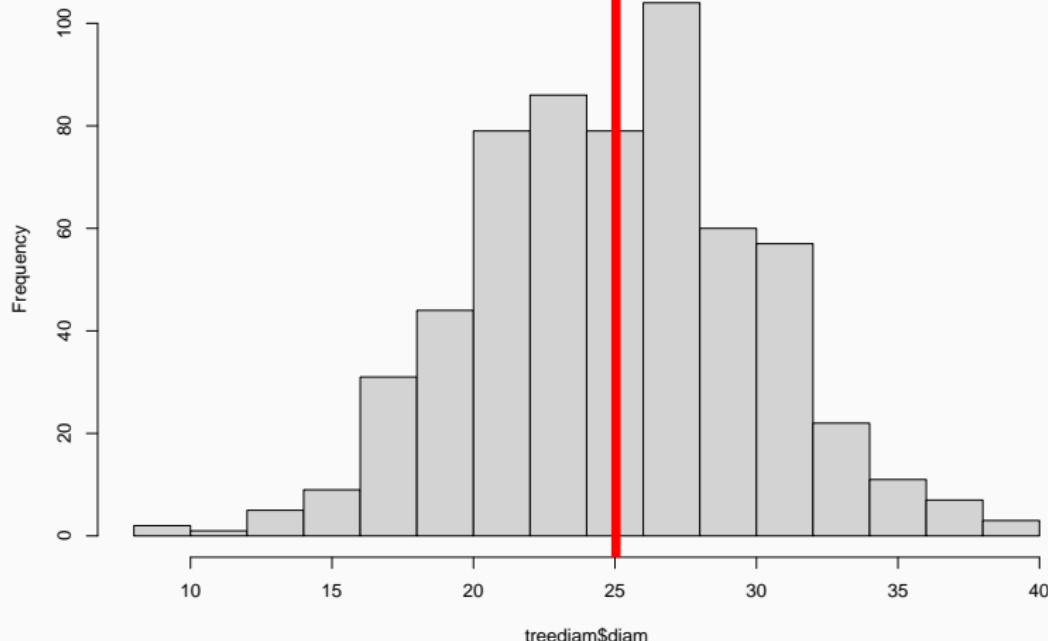
## Location / Central tendency

---

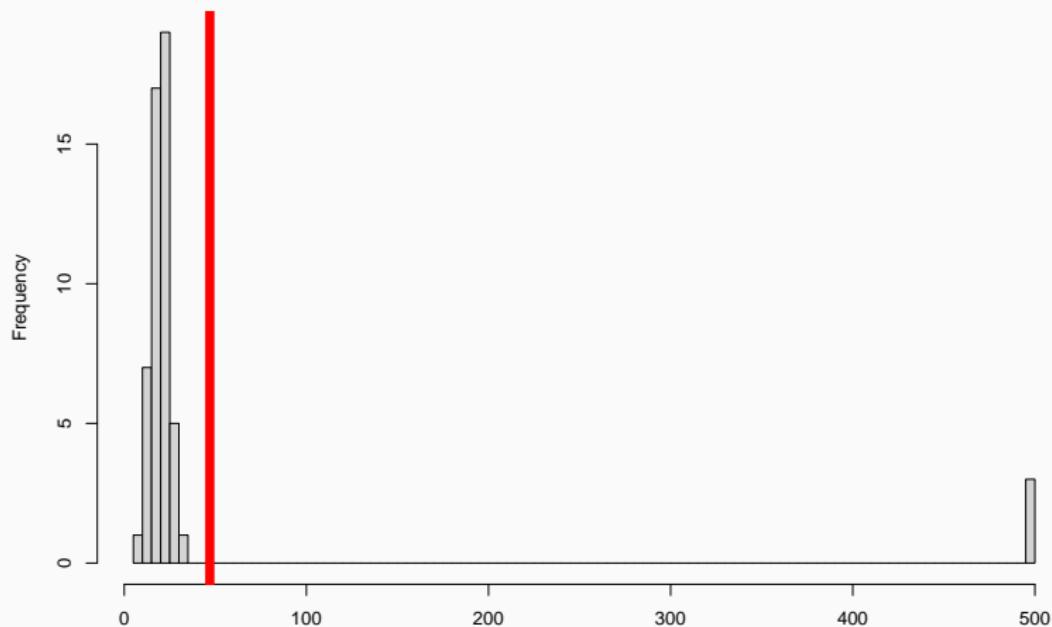
# Mean / Average

$$\text{mean} = \frac{d_1 + d_2 + d_3}{n}$$

Histogram of treediam\$diam

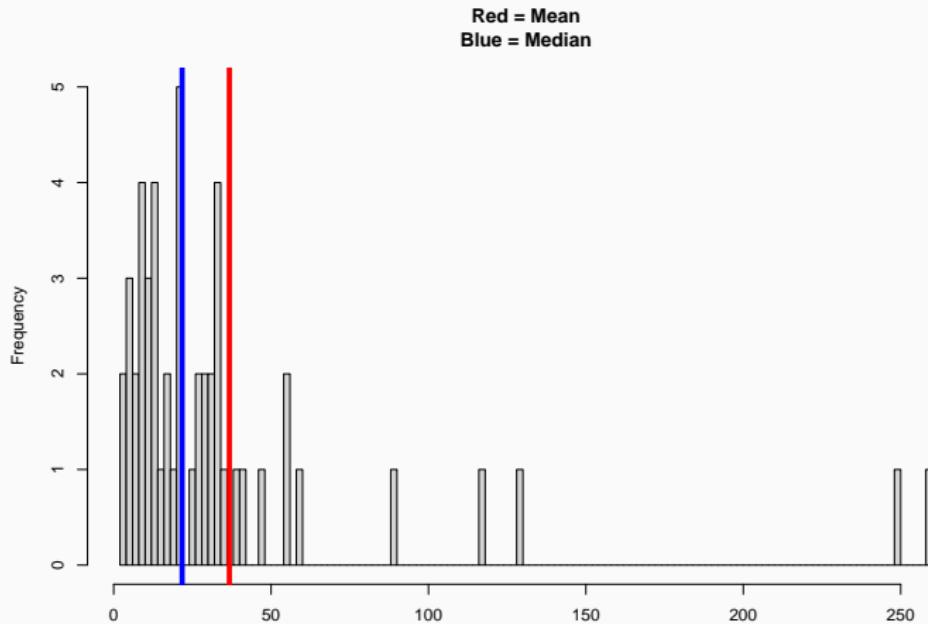


## Mean is sensitive to skew/outliers



# Median

50% percentile. Leaves half of the data values on each side



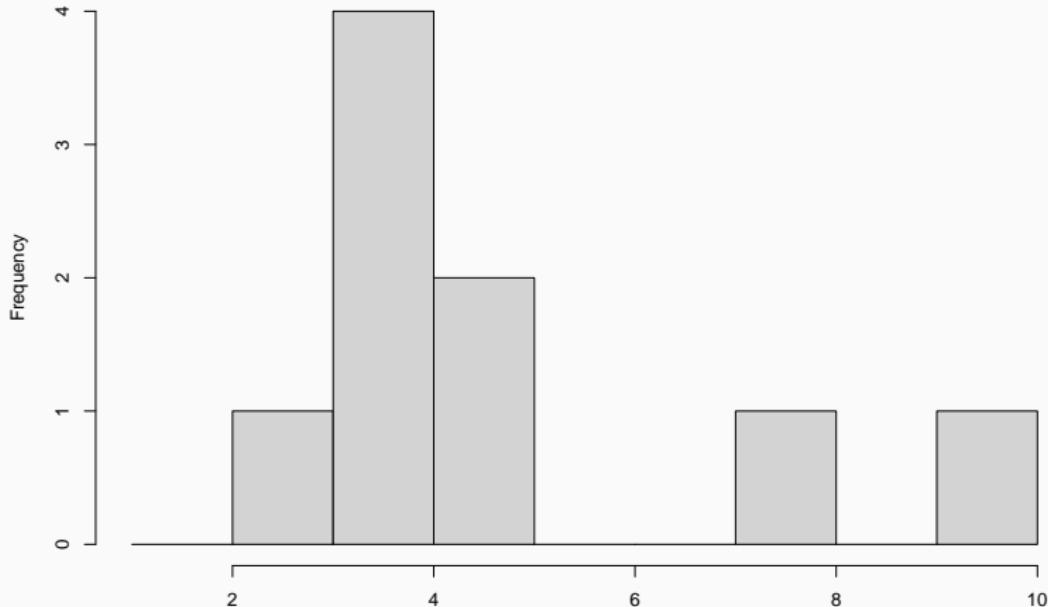
$$\text{Median of } c(2, 4, 6, 8, 10) = 6$$

$$\text{Median of } c(2, 4, 6, 8) = (4 + 6) / 2 = 5$$

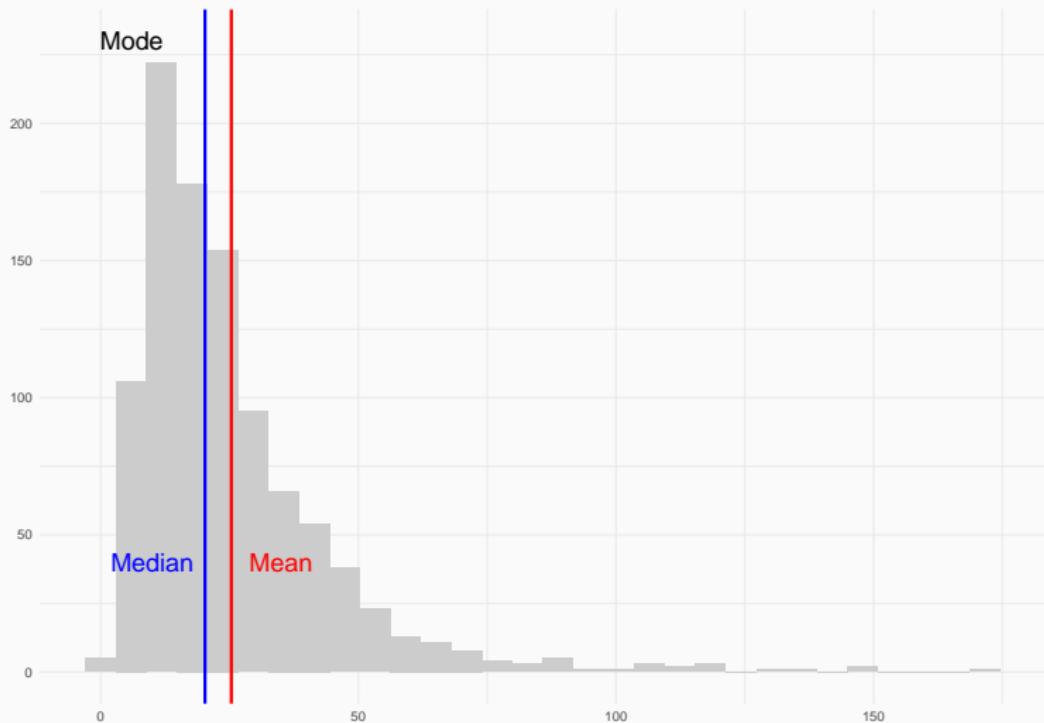
# Mode

Most frequent value

**Mode = 4**



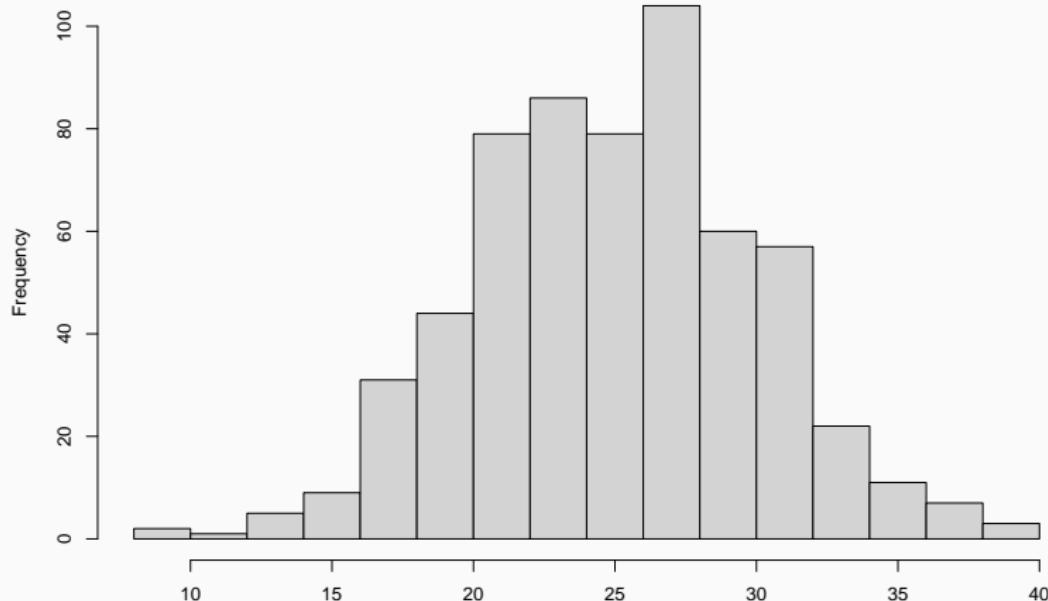
# Describing the location / central tendency



## Describing Variation / Spread

---

# Minimum, Maximum, Range



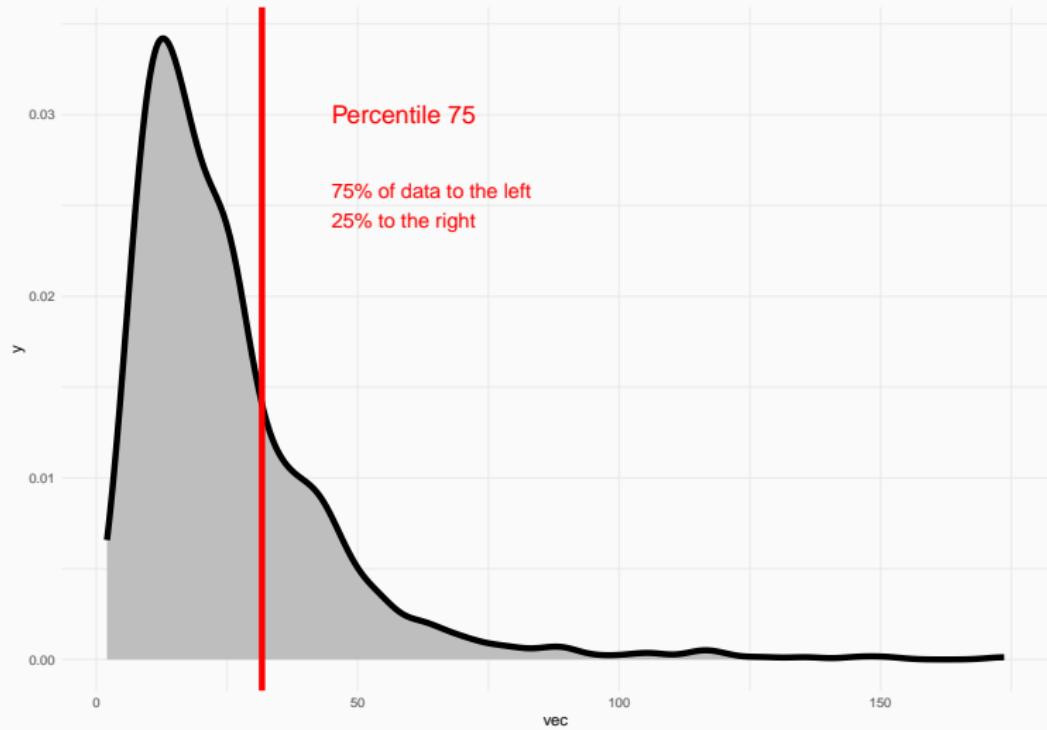
Minimum = 9.2

Maximum = 41.9

Range = 9.2, 41.9

# Quantiles

Quartiles, Percentiles...

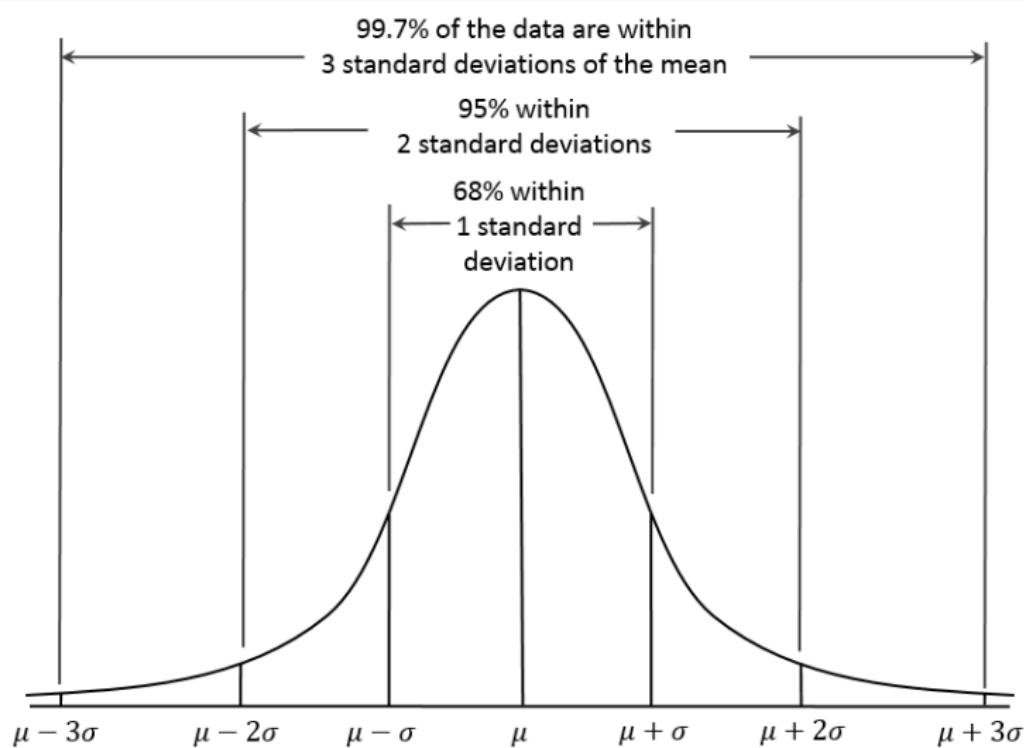


# Standard Deviation

Average distance between data points and the mean

$$SD = \sqrt{\frac{\sum (x - \mu)^2}{n - 1}}$$

# In a Normal distribution

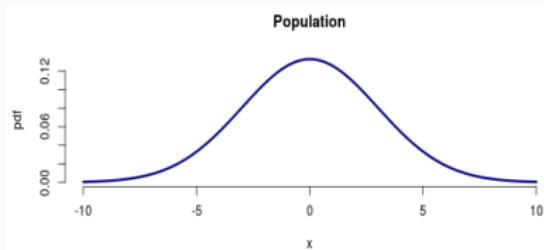


## Standard Error of the Mean

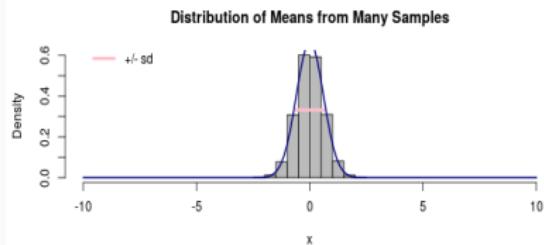
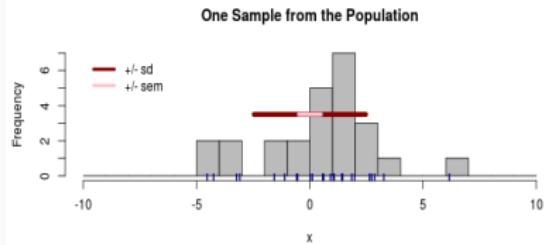
$$SEM = \frac{SD}{\sqrt{n}}$$

Estimates uncertainty (spread) of the parameter 'mean'

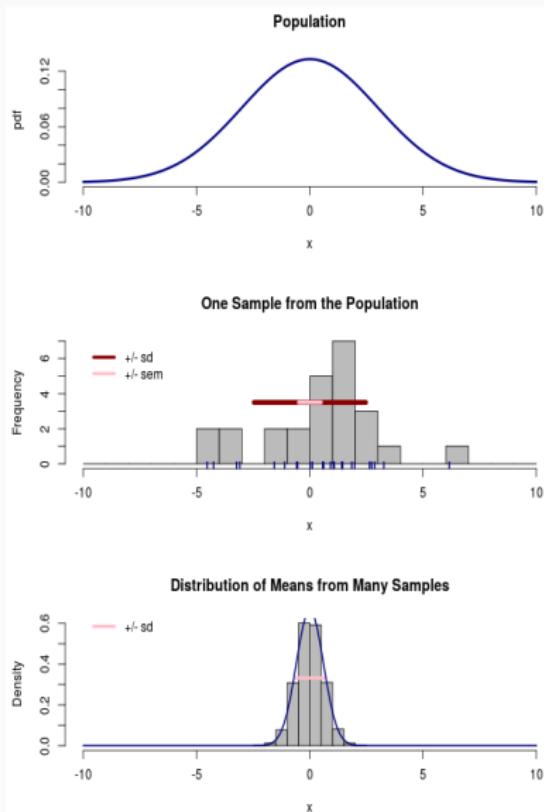
# Relationship between SD and SEM



- SD quantifies scatter in population

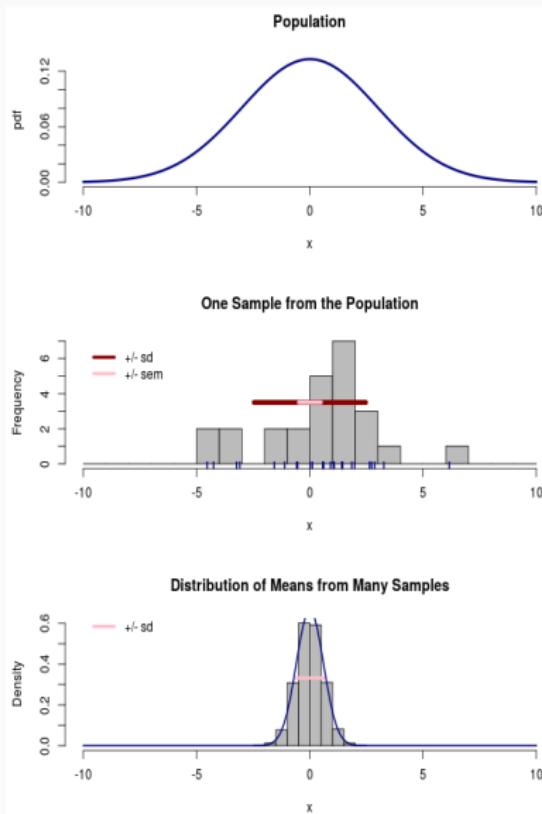


# Relationship between SD and SEM



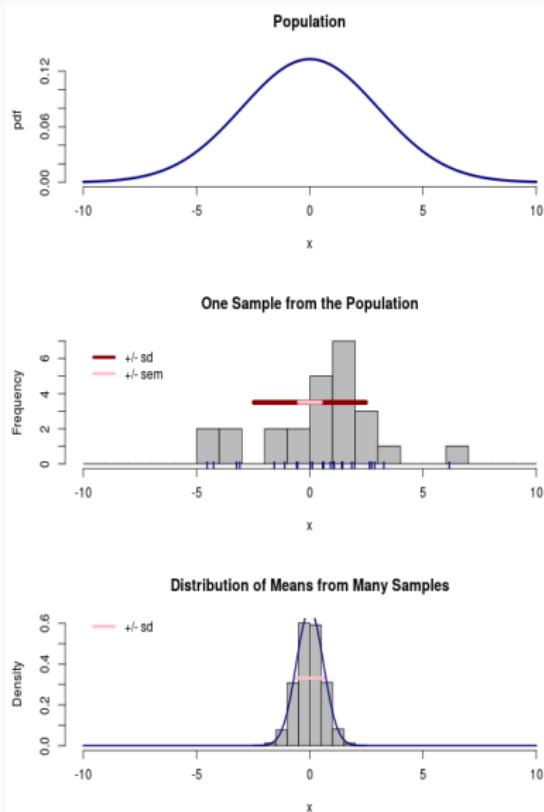
- SD quantifies scatter in population
- SEM quantifies uncertainty in parameter estimate (population mean)

# Relationship between SD and SEM



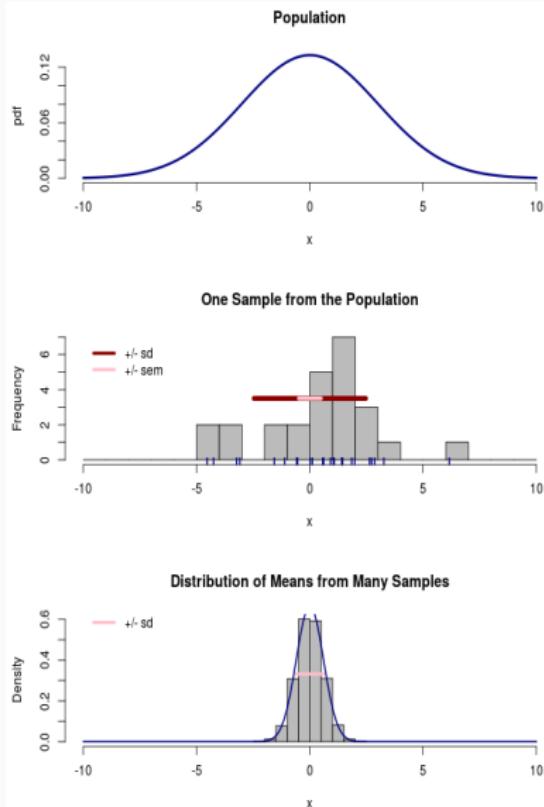
- SD quantifies scatter in population
- SEM quantifies uncertainty in parameter estimate (population mean)
- $\text{SEM} = \text{SD}/\sqrt{n}$

# Relationship between SD and SEM



- SD quantifies scatter in population
- SEM quantifies uncertainty in parameter estimate (population mean)
- $\text{SEM} = \text{SD}/\sqrt{n}$
- SEM decreases with sample size (mean better known), SD does not.

# Relationship between SD and SEM

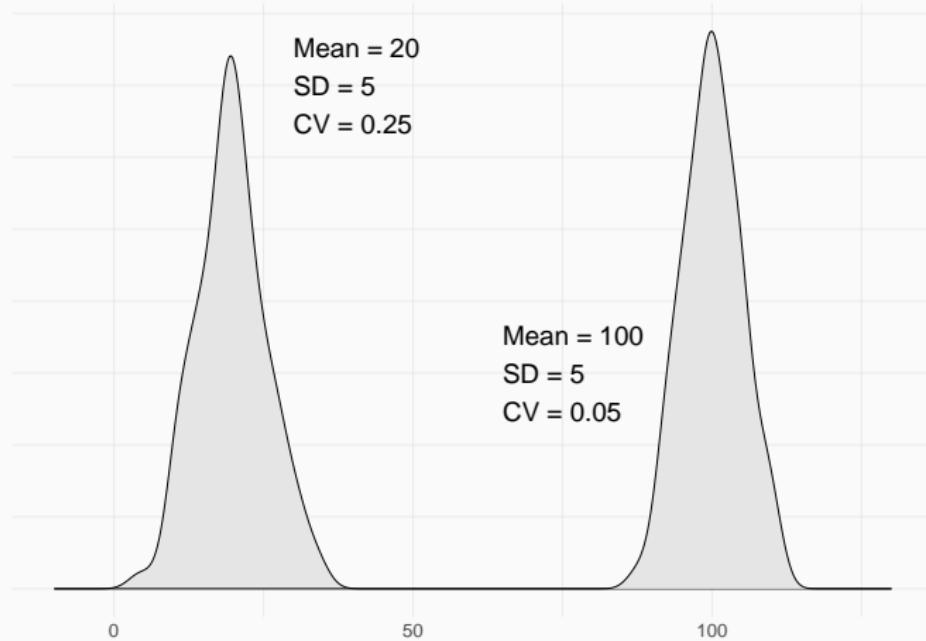


- SD quantifies scatter in population
- SEM quantifies uncertainty in parameter estimate (population mean)
- $\text{SEM} = \text{SD}/\sqrt{n}$
- SEM decreases with sample size (mean better known), SD does not.
- [https://gallery.shinyapps.io/sampling\\_and\\_stderr/](https://gallery.shinyapps.io/sampling_and_stderr/)

# Coefficient of Variation

Facilitates comparing spread of distributions with different means

$$CV = \frac{SD}{mean}$$

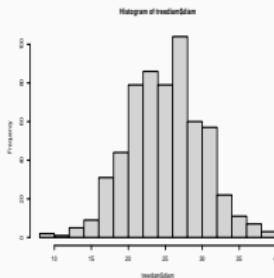


# Summarise a distribution

## Central tendency / location

- mean (average)

## Variation / Spread

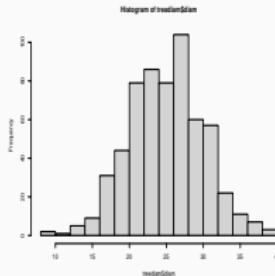


# Summarise a distribution

## Central tendency / location

- mean (average)
- median (50% percentile)

## Variation / Spread

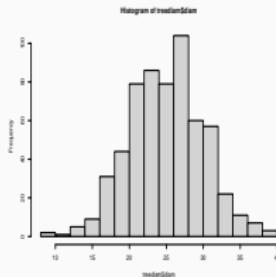


# Summarise a distribution

## Central tendency / location

- mean (average)
- median (50% percentile)
- mode (most frequent value)

## Variation / Spread



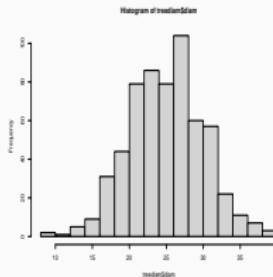
# Summarise a distribution

## Central tendency / location

- mean (average)
- median (50% percentile)
- mode (most frequent value)

## Variation / Spread

- min, max, range



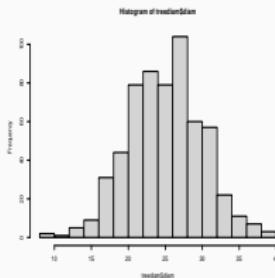
# Summarise a distribution

## Central tendency / location

- mean (average)
- median (50% percentile)
- mode (most frequent value)

## Variation / Spread

- min, max, range
- quantiles (quartiles, percentiles...)



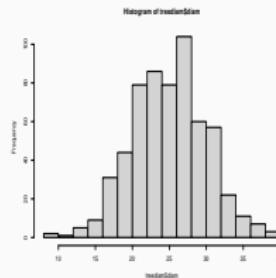
# Summarise a distribution

## Central tendency / location

- mean (average)
- median (50% percentile)
- mode (most frequent value)

## Variation / Spread

- min, max, range
- quantiles (quartiles, percentiles...)
- standard deviation



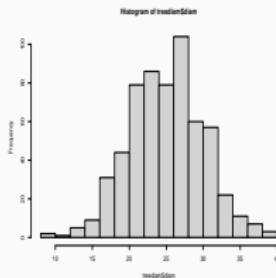
# Summarise a distribution

## Central tendency / location

- mean (average)
- median (50% percentile)
- mode (most frequent value)

## Variation / Spread

- min, max, range
- quantiles (quartiles, percentiles...)
- standard deviation
- standard error of the mean



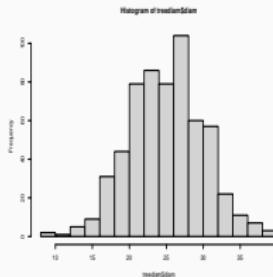
# Summarise a distribution

## Central tendency / location

- mean (average)
- median (50% percentile)
- mode (most frequent value)

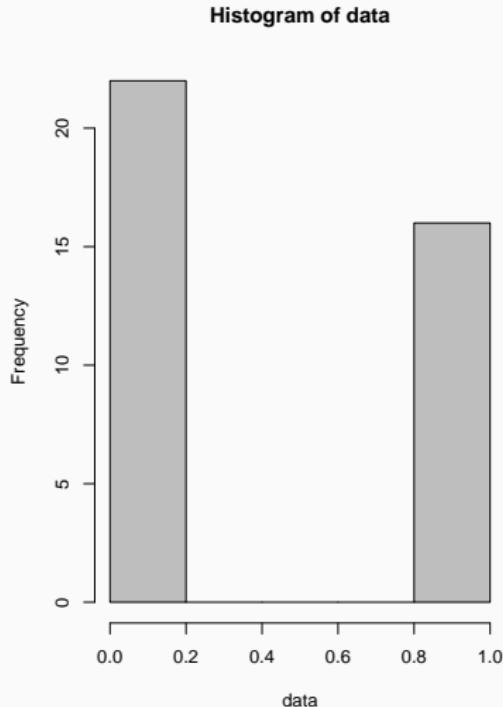
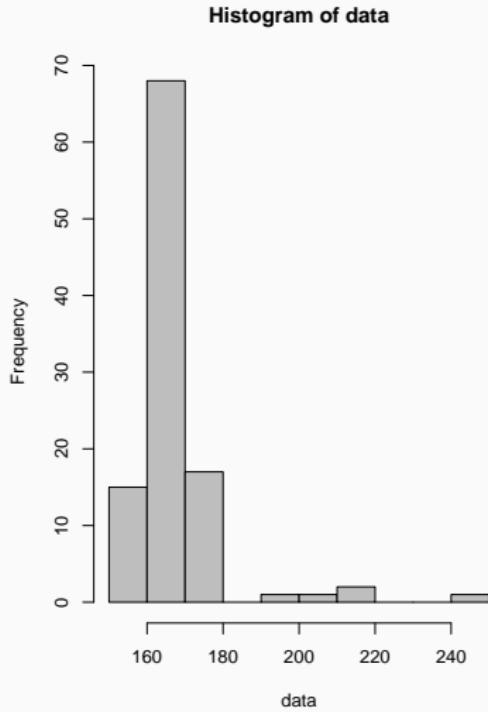
## Variation / Spread

- min, max, range
- quantiles (quartiles, percentiles...)
- standard deviation
- standard error of the mean
- coefficient of variation



# What statistical descriptors are best? (and why)

<https://pollev.com/franciscorod726>



## Sampling, confidence intervals, likelihood and Bayesian inference

---

# Inference: from samples to population

We rarely measure the whole **population**, but take **samples**.

Then we make inferences from sample to population.



If we sample 30 trees in our neighbourhood...

Can we extrapolate results to

- whole neighbourhood?
- whole city?
- whole country?
- the world?

What's the **suitable population** to make inferences given this sample?

## Understanding confidence intervals

- Summarise **uncertainty** in parameter estimates.

## Understanding confidence intervals

- Summarise **uncertainty** in parameter estimates.
- <https://rpsychologist.com/d3/CI/> (or [here](#))

## Understanding confidence intervals

- Summarise **uncertainty** in parameter estimates.
- <https://rpsychologist.com/d3/CI/> (or [here](#))
- A 95% CI is **NOT** 95% likely to contain the true parameter value!

## Understanding confidence intervals

- Summarise **uncertainty** in parameter estimates.
- <https://rpsychologist.com/d3/CI/> (or [here](#))
- A 95% CI is **NOT** 95% likely to contain the true parameter value!
- Instead, 95% of the CIs obtained with this sampling will contain the true value.

## Understanding confidence intervals

- Summarise **uncertainty** in parameter estimates.
- <https://rpsychologist.com/d3/CI/> (or [here](#))
- A 95% CI is **NOT** 95% likely to contain the true parameter value!
- Instead, 95% of the CIs obtained with this sampling will contain the true value.
- Like person who tells truth 95% of the time, but we can't tell if a particular statement is true.

## Understanding confidence intervals

- Summarise **uncertainty** in parameter estimates.
- <https://rpsychologist.com/d3/CI/> (or [here](#))
- A 95% CI is **NOT** 95% likely to contain the true parameter value!
- Instead, 95% of the CIs obtained with this sampling will contain the true value.
- Like person who tells truth 95% of the time, but we can't tell if a particular statement is true.
- It's a frequentist, long-run property.

## Understanding confidence intervals

- Summarise **uncertainty** in parameter estimates.
- <https://rpsychologist.com/d3/CI/> (or [here](#))
- A 95% CI is **NOT** 95% likely to contain the true parameter value!
- Instead, 95% of the CIs obtained with this sampling will contain the true value.
- Like person who tells truth 95% of the time, but we can't tell if a particular statement is true.
- It's a frequentist, long-run property.
- To read more: [Morey et al \(2015\)](#)

## What happens if we increase sample size?

<https://rpsychologist.com/d3/CI/>

- CI width *decreases*...

## What happens if we increase sample size?

<https://rpsychologist.com/d3/CI/>

- CI width *decreases*...
- but still 5% of CIs will NOT contain true mean!

If the 95% CI of X is (120, 150)...

- There is a 95% probability that X lies between 120 and 150

<https://pollev.com/franciscorod726>

If the 95% CI of X is (120, 150)...

- There is a 95% probability that X lies between 120 and 150
- We can be 95% confident that X lies between 120 and 150

<https://pollev.com/franciscorod726>

If the 95% CI of X is (120, 150)...

- There is a 95% probability that X lies between 120 and 150
- We can be 95% confident that X lies between 120 and 150
- If we repeated the experiment, 95% of the time X would fall between 120 and 150

<https://pollev.com/franciscorod726>

If the 95% CI of X is (120, 150)...

- There is a 95% probability that X lies between 120 and 150
- We can be 95% confident that X lies between 120 and 150
- If we repeated the experiment, 95% of the time X would fall between 120 and 150
- If we repeated the experiment, 95% of the CIs would contain the true value of X

<https://pollev.com/franciscorod726>

If the 95% CI of X is (120, 150)...

- There is a 95% probability that X lies between 120 and 150
- We can be 95% confident that X lies between 120 and 150
- If we repeated the experiment, 95% of the time X would fall between 120 and 150
- If we repeated the experiment, 95% of the CIs would contain the true value of X
- The probability that X is greater than 0 is at least 95%

<https://pollev.com/franciscorod726>

If the 95% CI of X is (120, 150)...

- There is a 95% probability that X lies between 120 and 150
- We can be 95% confident that X lies between 120 and 150
- If we repeated the experiment, 95% of the time X would fall between 120 and 150
- If we repeated the experiment, 95% of the CIs would contain the true value of X
- The probability that X is greater than 0 is at least 95%
- The probability that X equals 0 is smaller than 5%

<https://pollev.com/franciscorod726>

## Bayesian credible intervals

- Bayesian **credible** intervals do give the probability that true parameter value is contained within them.

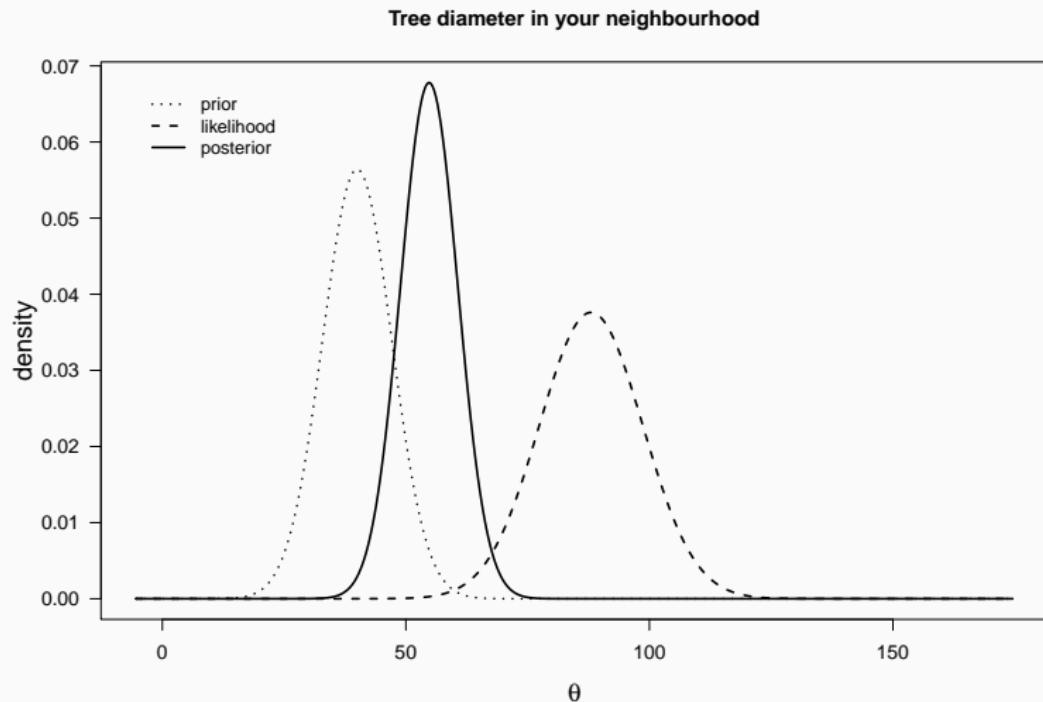
## Bayesian credible intervals

- Bayesian **credible** intervals do give the probability that true parameter value is contained within them.
- Frequentist CIs and Bayesian credible intervals can be similar, but not always.

# Bayesian inference: prior, posterior, and likelihood

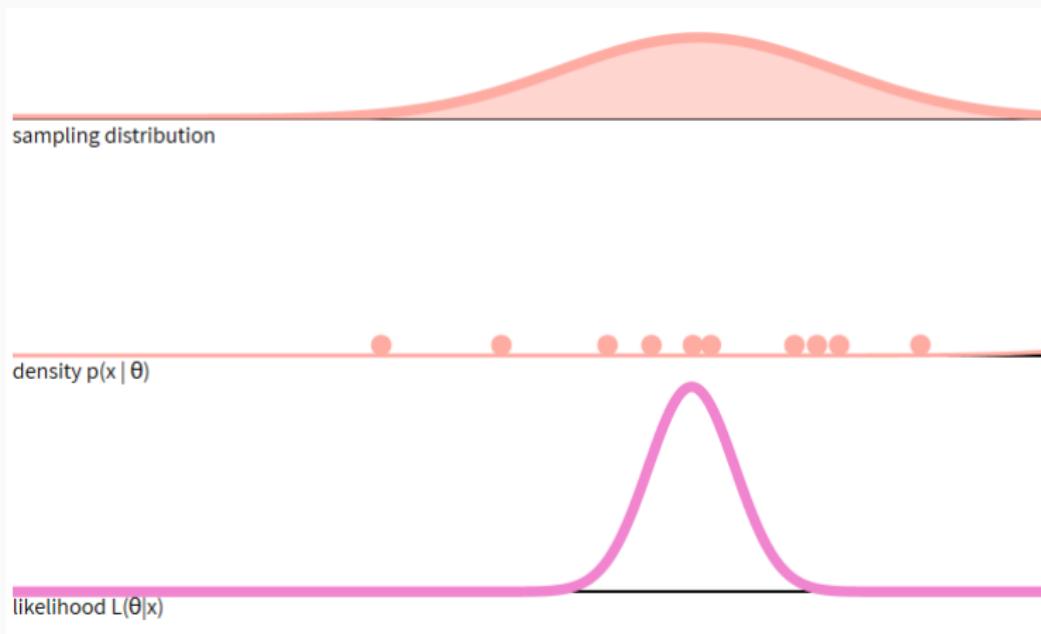
$$P(H|D) \propto P(D|H) \cdot P(H)$$

Posterior  $\propto$  Likelihood  $\cdot$  Prior

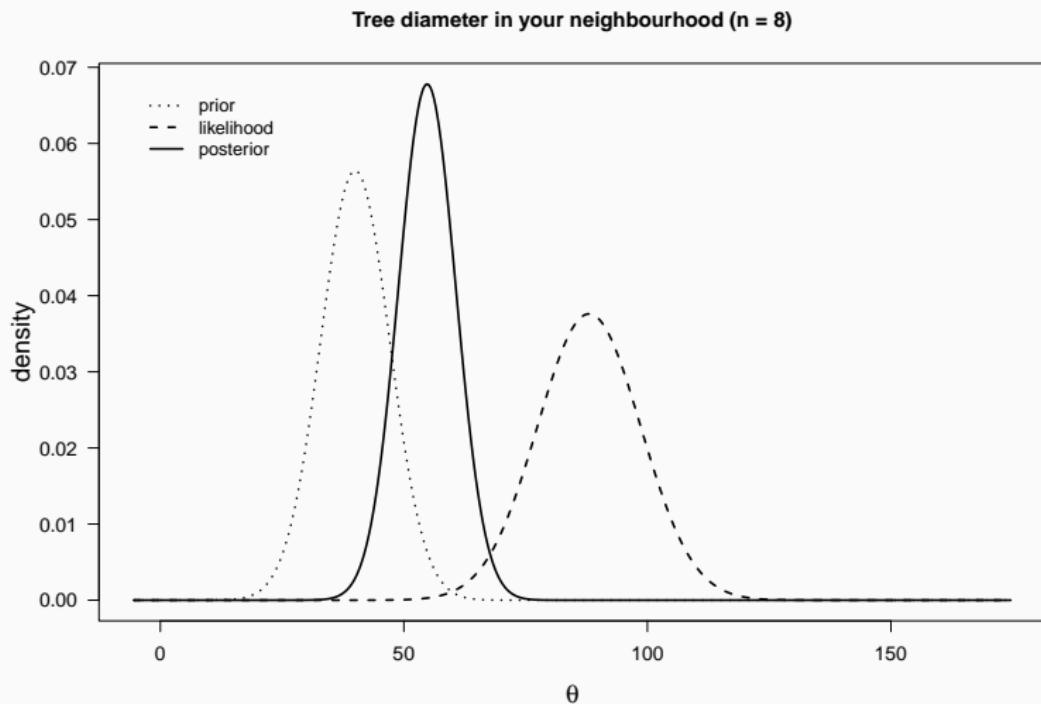


# What is the likelihood?

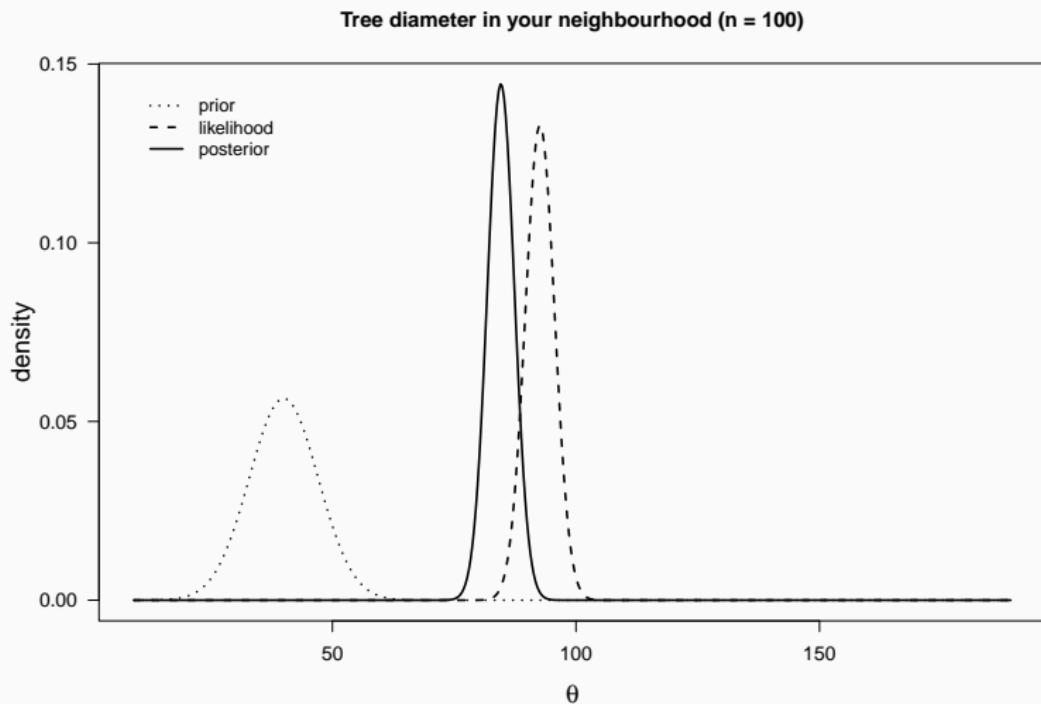
$$L(\theta|x) = P(x|\theta)$$



# Bayesian inference: prior and likelihood produce posterior



With increasing sample size, likelihood dominates prior



## More apps to introduce Bayesian inference

- Bayesian Demo

## More apps to introduce Bayesian inference

- Bayesian Demo
- Bayesian inference for a population mean

## More apps to introduce Bayesian inference

- Bayesian Demo
- Bayesian inference for a population mean
- Normal

## More apps to introduce Bayesian inference

- Bayesian Demo
- Bayesian inference for a population mean
- Normal
- Binomial

## More apps to introduce Bayesian inference

- Bayesian Demo
- Bayesian inference for a population mean
- Normal
- Binomial
- Own data

## More apps to introduce Bayesian inference

- Bayesian Demo
- Bayesian inference for a population mean
- Normal
- Binomial
- Own data
- Bayesian t-test

# Bayesian statistics in practice

- Integrate information (prior)

# Bayesian statistics in practice

- Integrate information (prior)
- Prior regularises unlikely estimates from data

# Bayesian statistics in practice

- Integrate information (prior)
- Prior regularises unlikely estimates from data
- Large dataset -> prior effect diminishes

# Bayesian statistics in practice

- Integrate information (prior)
- Prior regularises unlikely estimates from data
- Large dataset -> prior effect diminishes
- Uncertainty / Propagate errors

# Experimental design

---

# How would you evaluate fertilizer effect?

Discuss with partner (5')



## Experimental design principles

---

# Replication

---

# Replication!



# Replication

- Replication is key: we need several samples.

# Replication

- Replication is key: we need several samples.
- How many? As much as you can! See [Gelman & Carlin 2014](#).

H. Stern / A. Gelman

*The most important aspect of a statistical analysis is not what you do with the data, it's what data you use*

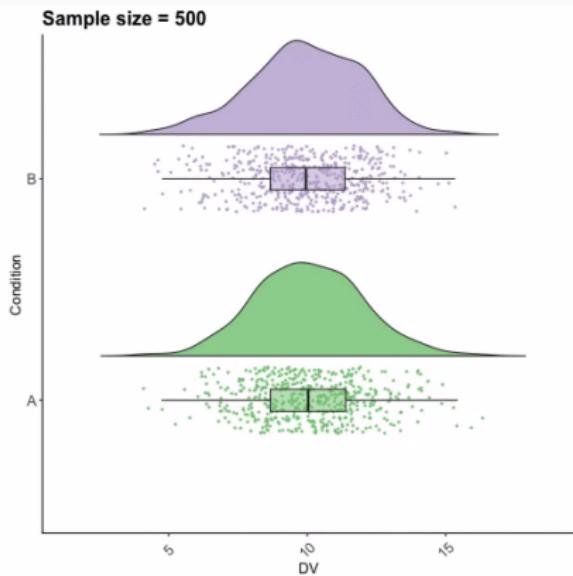
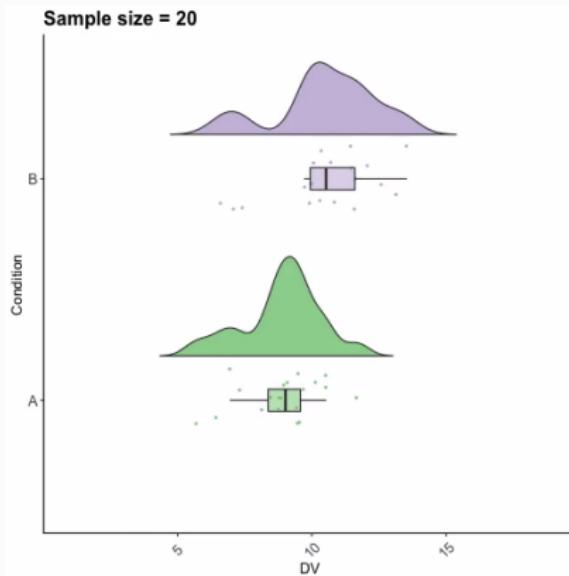
# The importance of sample size

- Many studies have **too low sample sizes**.

# The importance of sample size

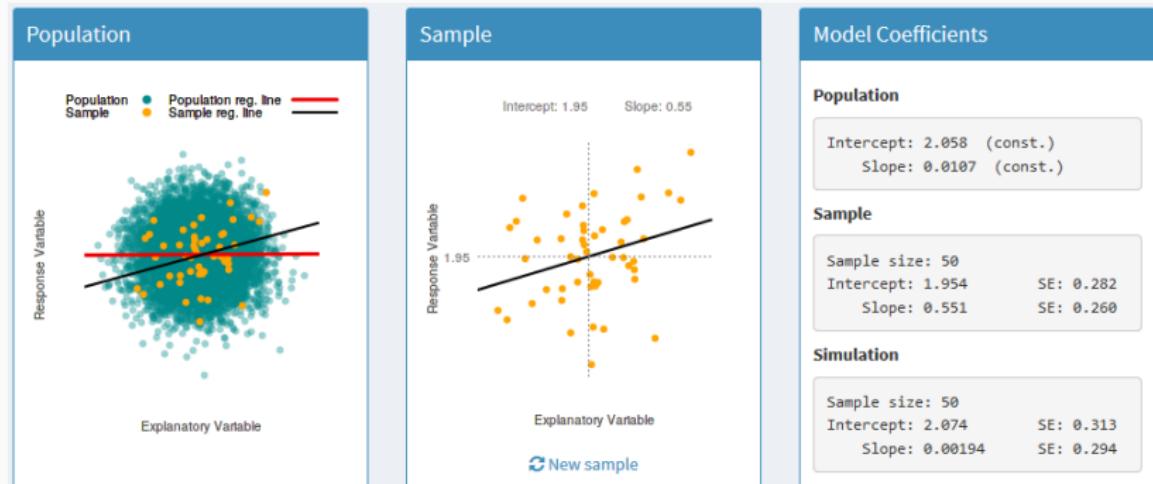
- Many studies have **too low sample sizes**.
- Low sample sizes miss subtle effects, but also **prone to bias**.

# Low sample sizes very sensitive to random noise



[https://twitter.com/ajstewart\\_lang/status/1020038488278945797](https://twitter.com/ajstewart_lang/status/1020038488278945797)

# Low sample sizes may bias inferences about population



<http://statisticalgate.com/regression-simulation/>

# Low sample sizes may bias inferences

See [The evolution of correlations](#)

Stopping rules

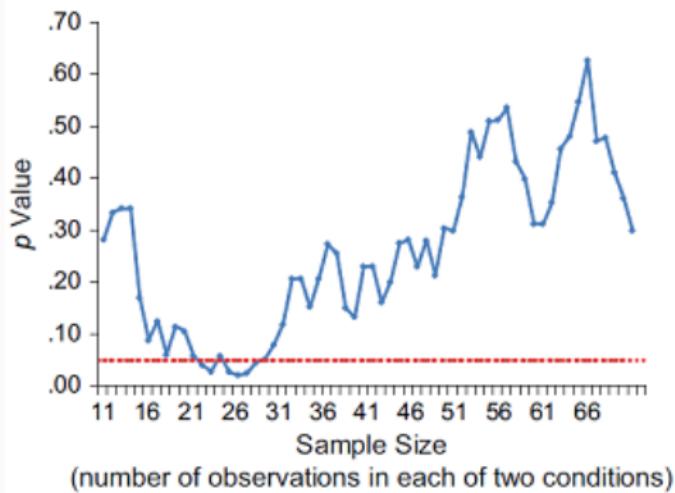


Fig. 2. Illustrative simulation of  $p$  values obtained by a researcher who continuously adds an observation to each of two conditions, conducting a t test after each addition. The dotted line highlights the conventional significance criterion of  $p \leq .05$ .

## Sample size estimation

- Plan model/statistical analysis **before** data collection.

## Sample size estimation

- Plan model/statistical analysis **before** data collection.
- **Do simulations.** Power/Sample size/Precision analyses  
(e.g. see papers like [this](#) & [this](#), or software like [this](#) & [this](#)).

## Sample size estimation

- Plan model/statistical analysis **before** data collection.
- **Do simulations.** Power/Sample size/Precision analyses (e.g. see papers like [this](#) & [this](#), or software like [this](#) & [this](#)).
- Plan to have at least 10-30 observations per predictor.

## Sample size estimation

- Plan model/statistical analysis **before** data collection.
- **Do simulations.** Power/Sample size/Precision analyses (e.g. see papers like [this](#) & [this](#), or software like [this](#) & [this](#)).
- Plan to have at least **10-30 observations per predictor**.
- Complex models (w/ many predictors, interactions etc) require **high** sample sizes.

# Sample size estimation

Calculating sample size for Gaussian (Normal) response model:

- expected mean: 30
- expected sd: 10
- 10 parameters (predictors)
- expected R-squared: 0.2

```
library(pmsampsize)
pmsampsize(type = "c", parameters = 10, intercept = 30, sd = 10, rsquared = 0.2)
```

NB: Assuming 0.05 acceptable difference in apparent & adjusted R-squared

NB: Assuming MMOE <= 1.1 in estimation of intercept & residual standard deviation

SPP - Subjects per Predictor Parameter

	Samp_size	Shrinkage	Parameter	Rsq	SPP
Criteria 1	313	0.900	10 0.2	31.3	
Criteria 2	161	0.827	10 0.2	16.1	
Criteria 3	244	0.876	10 0.2	24.4	
Criteria 4*	313	0.900	10 0.2	31.3	
Final	313	0.900	10 0.2	31.3	

Minimum sample size required for new model development based on user inputs = 313

\* 95% CI for intercept = (29.69, 30.31), for sample size n = 313

# Sample size estimation

Calculating sample size for binary response model:

- expected prevalence: 0.1
- 20 parameters (predictors)
- expected R-squared: 0.2

```
library(pmsampsize)
pmsampsize(type = "b", parameters = 20, prevalence = 0.1, rsquared = 0.2)
```

NB: Assuming 0.05 acceptable difference in apparent & adjusted R-squared

NB: Assuming 0.05 margin of error in estimation of intercept

NB: Events per Predictor Parameter (EPP) assumes prevalence = 0.1

	Samp_size	Shrinkage	Parameter	Rsq	Max_Rsq	EPP
Criteria 1	796	0.900	20	0.2	0.48	3.98
Criteria 2	738	0.893	20	0.2	0.48	3.69
Criteria 3	139	0.900	20	0.2	0.48	0.70
Final	796	0.900	20	0.2	0.48	3.98

Minimum sample size required for new model development based on user inputs = 796,  
with 80 events (assuming an outcome prevalence = 0.1) and an EPP = 3.98

# Randomization

---

# Randomization



# Randomization

- Haphazard  $\neq$  Random

# Randomization

- Haphazard  $\neq$  Random
- Stratify: randomize within groups (e.g. species, soil types)

## Controls

---

## Have controls

- Untreated individuals, plots... (assigned randomly, of course).

## Have controls

- Untreated individuals, plots... (assigned randomly, of course).
- Must differ only in treatment (i.e. homogeneous environment).

## Have controls

- Untreated individuals, plots... (assigned randomly, of course).
- Must differ only in treatment (i.e. homogeneous environment).
- Measure **before & after** treatment.

## Have controls

- Untreated individuals, plots... (assigned randomly, of course).
- Must differ only in treatment (i.e. homogeneous environment).
- Measure **before & after** treatment.
- Consider **blind designs** to avoid observer bias.

# Experimental design principles

1. Replication

# Experimental design principles

1. Replication
2. Randomization

# Experimental design principles

1. Replication
2. Randomization
3. Controls

## To read more

- Ruxton & Colegrave. Experimental Design for the Life Sciences.  
OUP

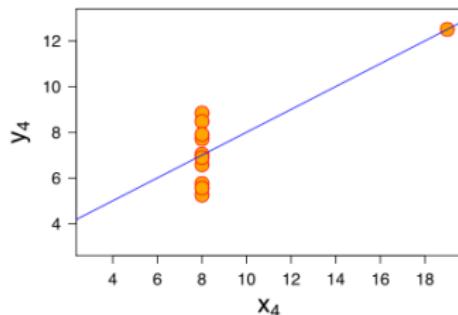
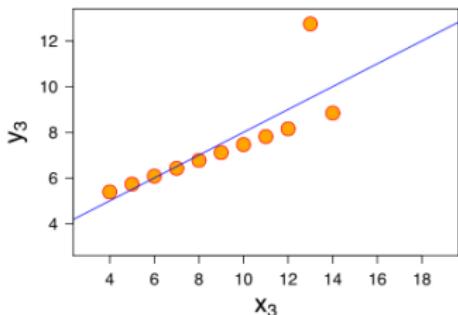
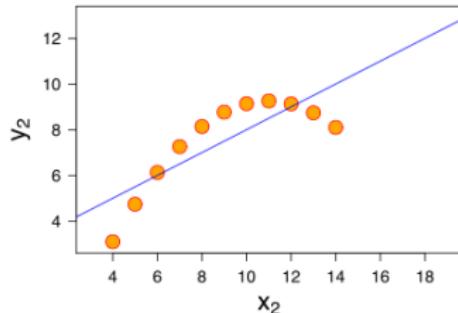
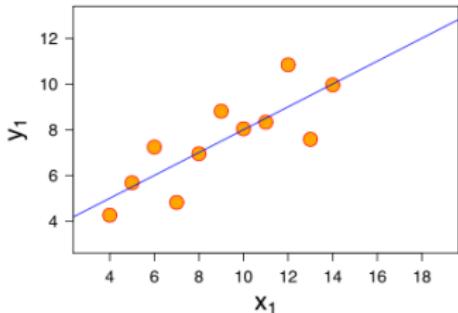
## The importance of visualisation

---

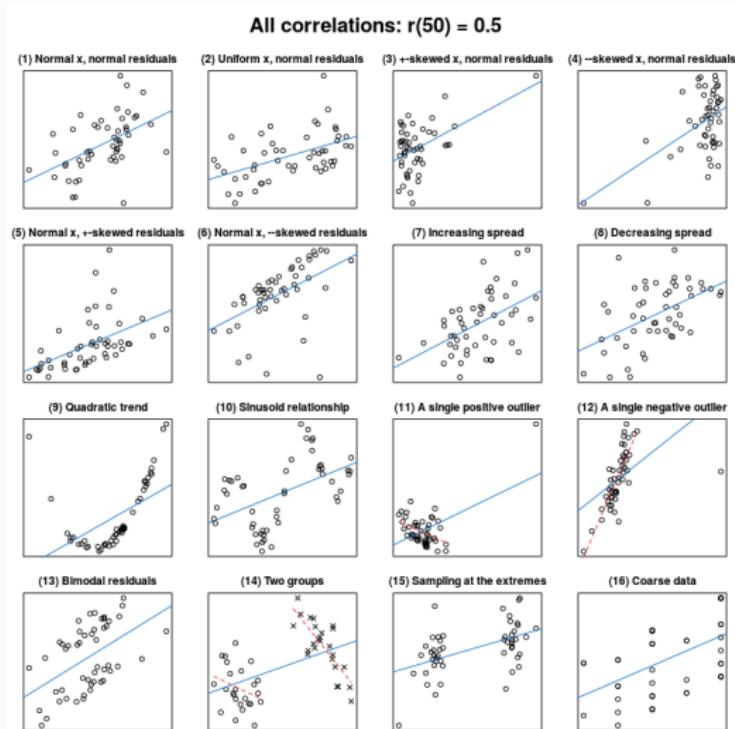
Visualisation of data & models is  
key

---

# Always plot data and models

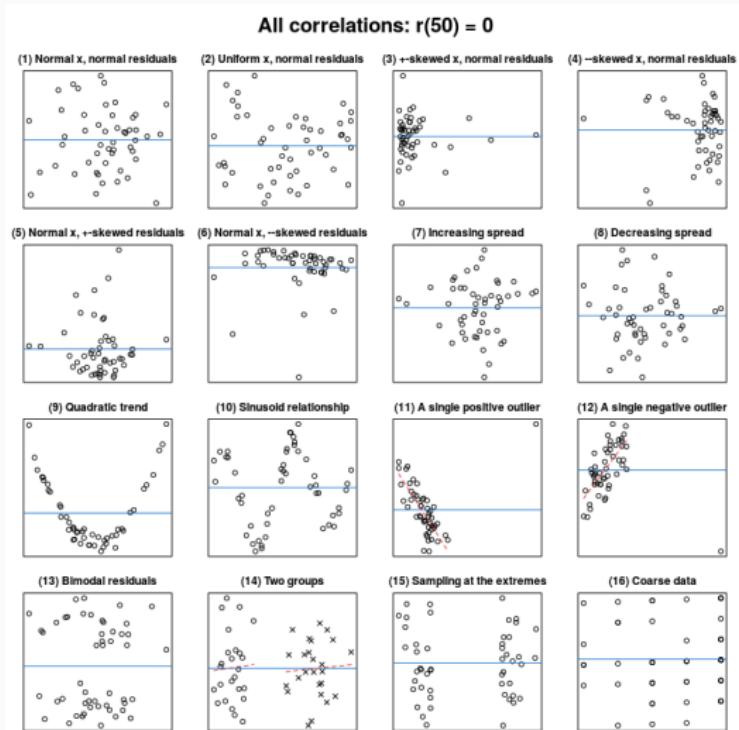


# Don't use statistics blindly: Visualise

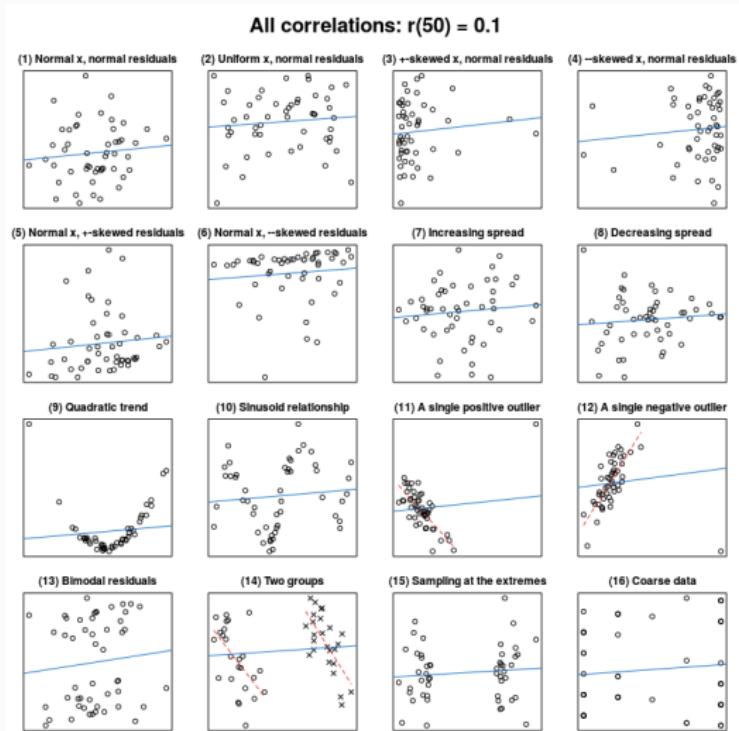


<https://janhove.github.io/teaching/2016/11/21/what-correlations-look-like>

# Don't use statistics blindly: Visualise



# 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

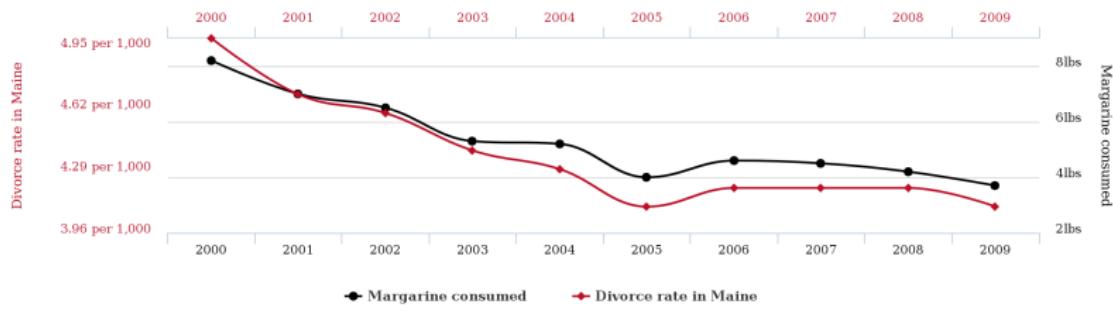
- We found that plants with big flowers produced 30% more seeds...
- Do big flowers increase reproductive success?

# 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

## Divorce rate in Maine correlates with Per capita consumption of margarine



<http://tylervigen.com/spurious-correlations>

# 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

---

### 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 living 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  $H_0$  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  $H_0$  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  $H_0$  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)

# How to interpret P-values

- **Low P-value:** data unlikely if every model assumption (including  $H_0$ ) were true.

## How to interpret P-values

- **Low P-value:** data unlikely if every model assumption (including  $H_0$ ) were true.
- **Large P-value:** data not unusual if every model assumption (including  $H_0$ ) were true.

## How to interpret P-values

- **Low P-value:** data unlikely if every model assumption (including  $H_0$ ) were true.
- **Large P-value:** data not unusual if every model assumption (including  $H_0$ ) 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  $H_0$ ) were true.
- **Large P-value:** data not unusual if every model assumption (including  $H_0$ ) were true.
- A very small P-value does not tell us which model assumption is incorrect:
  - Could be that  $H_0$  is not true

# How to interpret P-values

- **Low P-value:** data unlikely if every model assumption (including  $H_0$ ) were true.
- **Large P-value:** data not unusual if every model assumption (including  $H_0$ ) were true.
- A very small P-value does not tell us which model assumption is incorrect:
  - Could be that  $H_0$  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  $H_0$ ) were true.
- **Large P-value:** data not unusual if every model assumption (including  $H_0$ ) were true.
- A very small P-value does not tell us which model assumption is incorrect:
  - Could be that  $H_0$  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](#)

For example

- A famous experiment found neutrinos faster than light

For example

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

## For example

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

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**  $H_0$ .

# Are differences *significant*?

Common practice:

- If  $p < 0.05$ , we **reject**  $H_0$ .
- If  $p > 0.05$ , we **fail to reject**  $H_0$

# Are differences *significant*?

Common practice:

- If  $p < 0.05$ , we **reject**  $H_0$ .
- If  $p > 0.05$ , we **fail to reject**  $H_0$
- (which is **NOT** the same as ' $H_0$  is true')

# Are differences *significant*?

Common practice:

- If  $p < 0.05$ , we **reject**  $H_0$ .
- If  $p > 0.05$ , we **fail to reject**  $H_0$
- (which is **NOT** the same as ' $H_0$  is true')
- **CAUTION:** P-value is continuous. We must **avoid binary decisions** based on **arbitrary thresholds**.

# Are differences *significant*?

Common practice:

- If  $p < 0.05$ , we **reject**  $H_0$ .
- If  $p > 0.05$ , we **fail to reject**  $H_0$
- (which is **NOT** the same as ' $H_0$  is true')
- **CAUTION:** P-value is continuous. We must **avoid binary decisions** based on **arbitrary thresholds**.

# Are differences *significant*?

Common practice:

- If  $p < 0.05$ , we **reject**  $H_0$ .
- If  $p > 0.05$ , we **fail to reject**  $H_0$
- (which is **NOT** the same as ' $H_0$  is true')
- **CAUTION:** P-value is continuous. We must **avoid binary decisions** based on arbitrary thresholds.

The image shows a screenshot of a website header and a news article. The header is dark red with white text. On the left is a 'MENU' dropdown button. In the center is the 'nature' logo with the subtitle 'International journal of science'. On the right is a blue 'Subs' button. Below the header, a white rectangular box contains the word 'EDITORIAL' followed by a small dot and the date '20 MARCH 2019'. The main title of the article is 'It's time to talk about ditching statistical significance'.

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

# Are these two groups different?

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

Welch Two Sample t-test

```
data: group.A and group.B
```

```
t = -0.17796, df = 5.0294, p-value = 0.8657
```

```
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
```

```
-12.33558 10.73558
```

```
sample estimates:
```

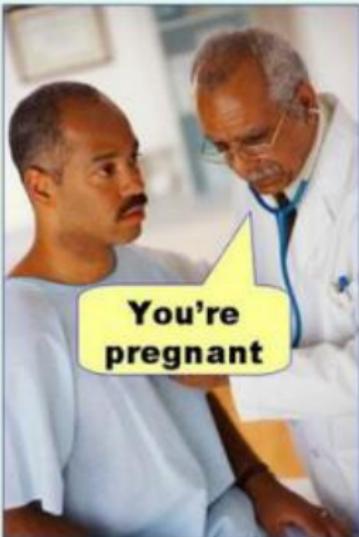
```
mean of x mean of y
```

```
175.6 176.4
```

<https://pollev.com/franciscorod726>

# 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 H<sub>0</sub> when it's false).

# Is this coin biased?

```
[1] 1 1 0 1 1 0 0 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/>

[http://daniellakens.blogspot.com/2017/12/  
understanding-common-misconceptions.html](http://daniellakens.blogspot.com/2017/12/understanding-common-misconceptions.html)

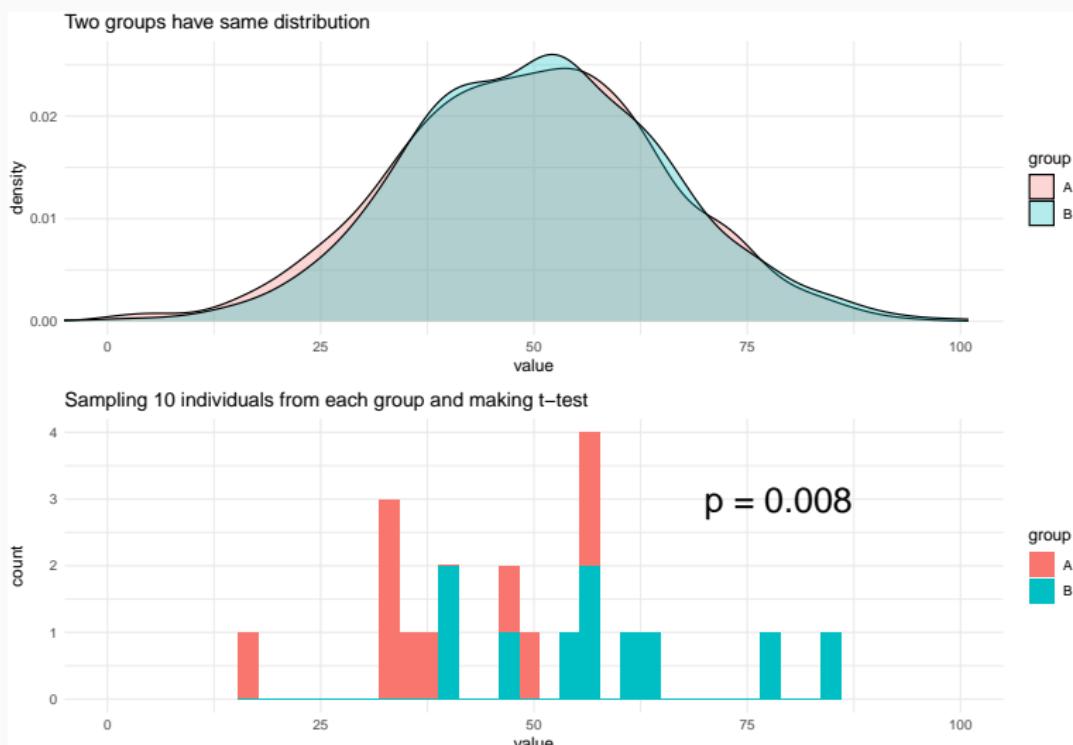
## NHST and p-values: common pitfalls

---

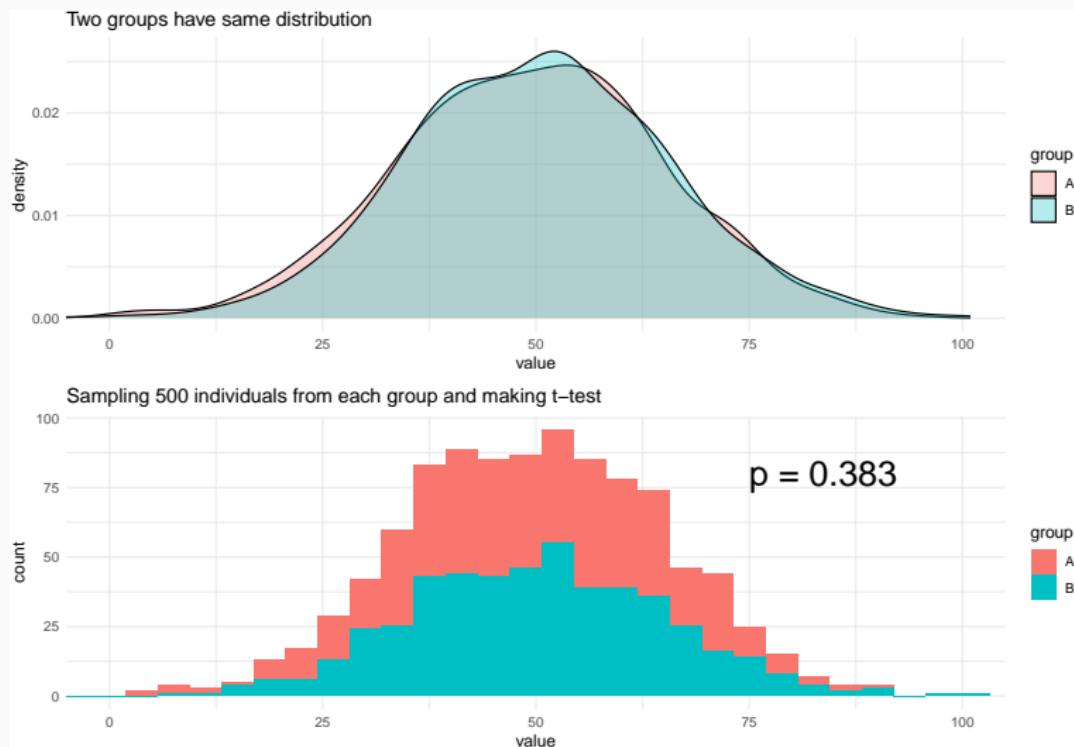
A significant p-value  
does NOT mean  
we found a true difference

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

Particularly with low sample sizes



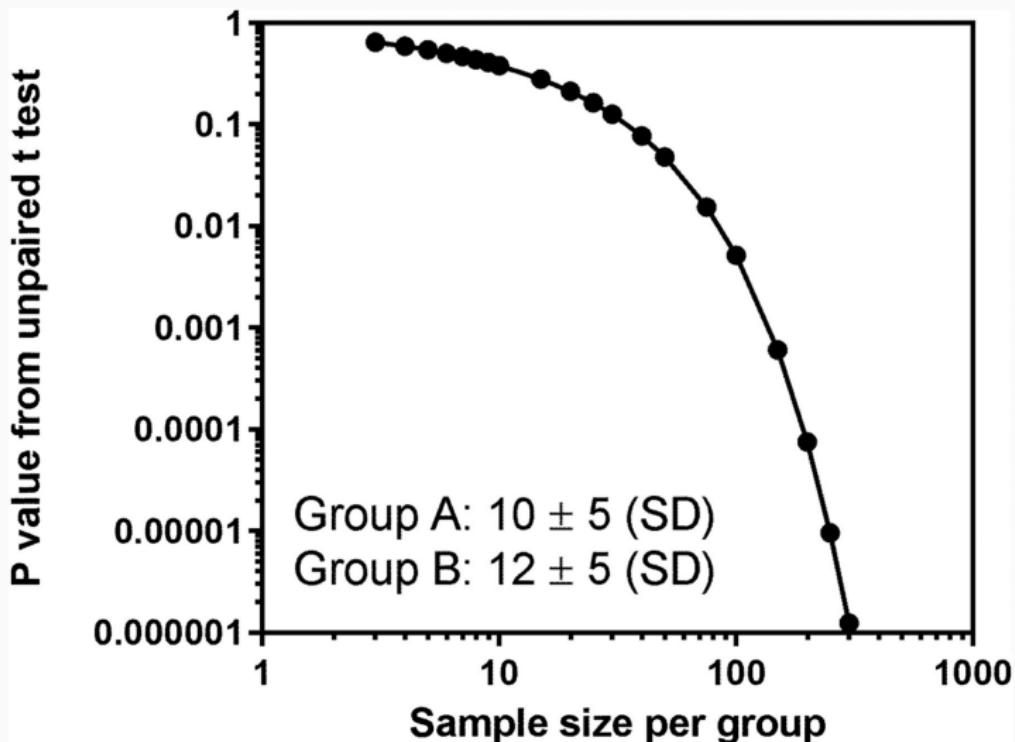
# If sample size was larger...



With low sample size (power),  
significant p-values  
are most likely overestimates

Loken & Gelman 2014, Vasisth et al. 2018

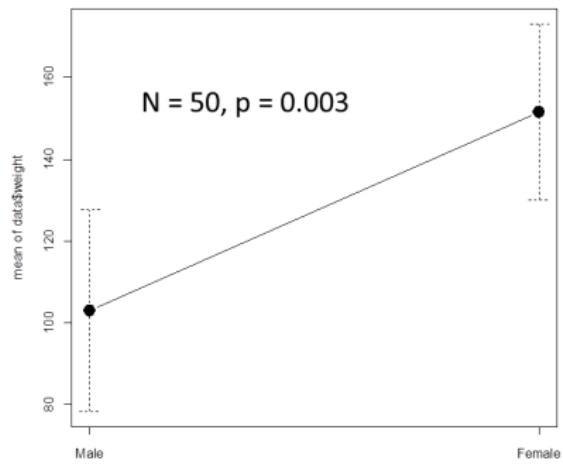
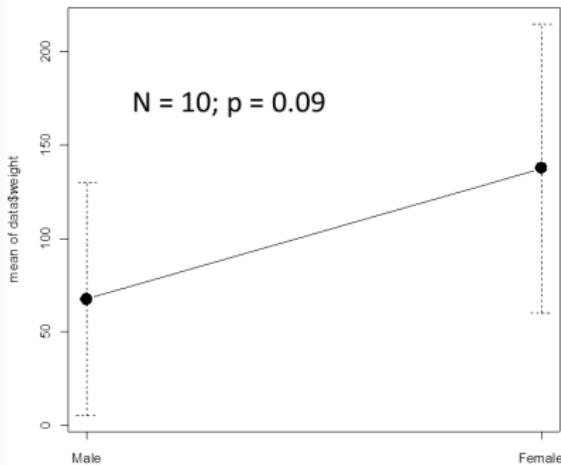
## P-value depends on sample size



# P-value depends on sample size

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

Real difference = 40 g



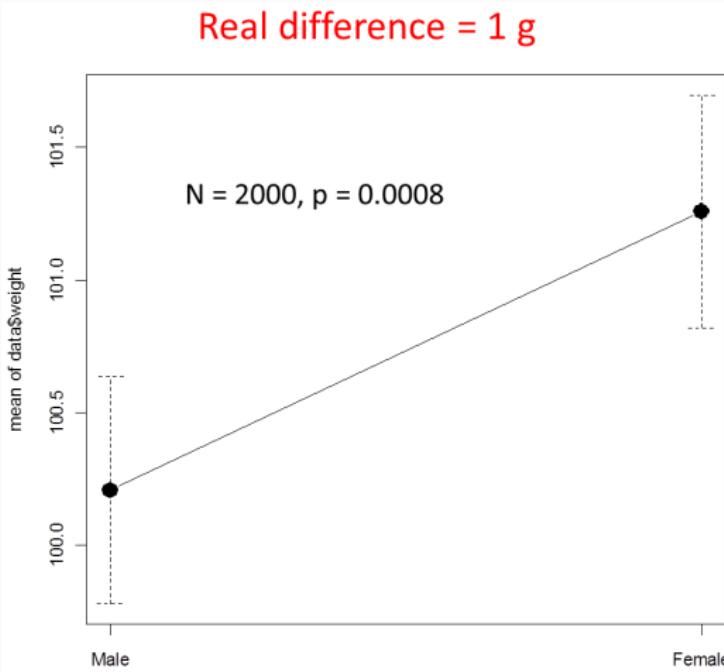
Statistically significant

$\neq$

biologically important

# Statistically significant != biologically important

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



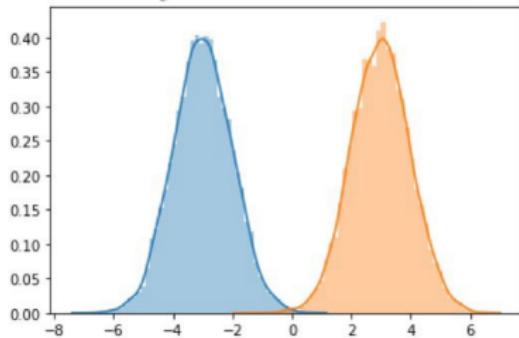
# Statistically significant != biologically important



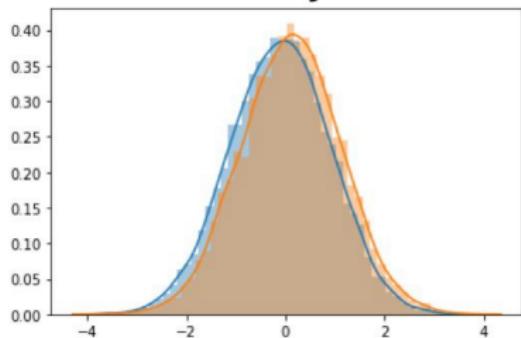
**zara weinberg**  
@weinberz

friendly reminder about  $p < 0.0001$ :

**What you think it means:**



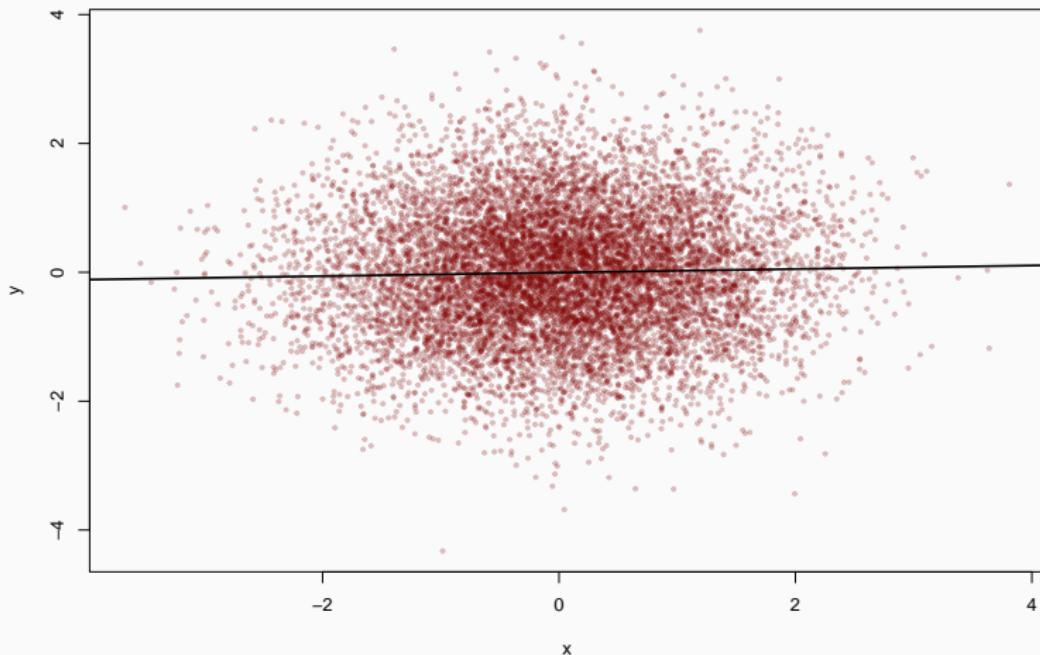
**What it actually means:**



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

Statistically significant != biologically important

**p = 0.005**



## Statistically significant != biologically important

- Statistically significant = unlikely to be zero

## Statistically significant != biologically important

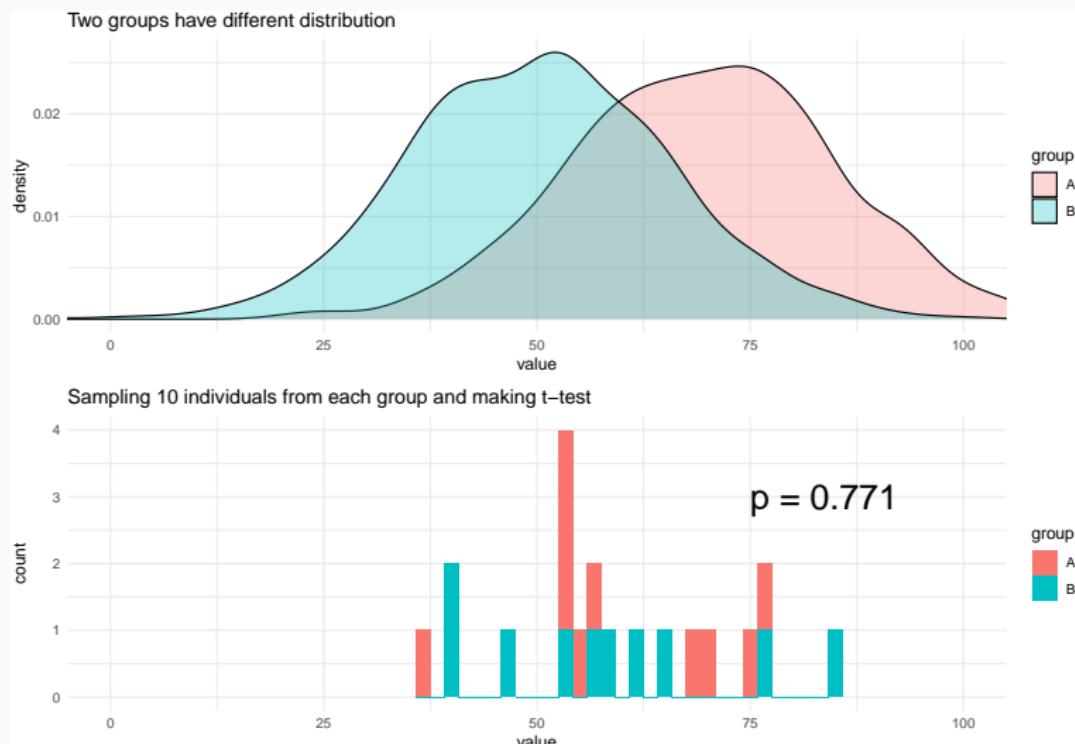
- Statistically significant = unlikely to be zero
- Good read: *significantly misleading*

## Statistically significant != biologically important

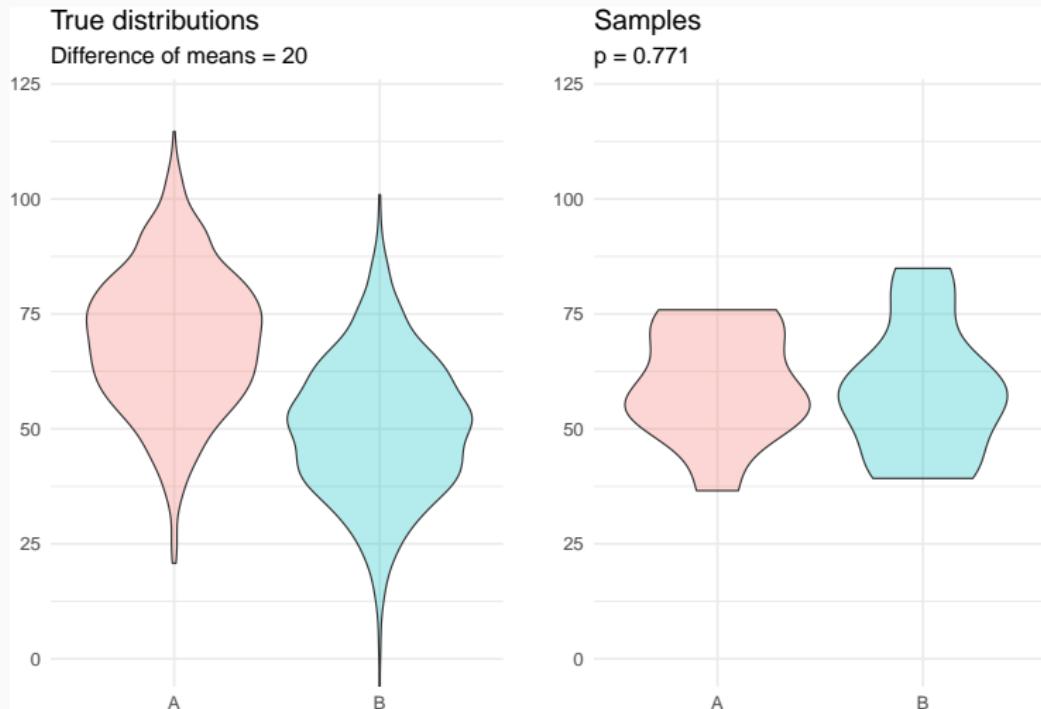
- Statistically significant = unlikely to be zero
- Good read: *significantly misleading*
- Beyond significant/not significant, look at **effect sizes and their uncertainty.**

‘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  $H_0$   $\neq H_0$  is true

Absence of evidence  $\neq$  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.statisticsonewrong.com/power.html#  
the-wrong-turn-on-red](https://www.statisticsonewrong.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.statisticsonewrong.com/power.html#  
the-wrong-turn-on-red](https://www.statisticsonewrong.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.statisticsonewrong.com/power.html#  
the-wrong-turn-on-red](https://www.statisticsonewrong.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 H<sub>0</sub> does NOT mean H<sub>0</sub> is true!



<https://www.statisticsonewrong.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 H<sub>0</sub> does NOT mean H<sub>0</sub> is true!
- Misinterpretation of underpowered study cost lives



<https://www.statisticsonewrong.com/power.html#the-wrong-turn-on-red>

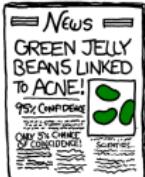
0.05 is an arbitrary threshold

**The Difference Between “Significant” and “Not Significant” is not  
Itself Statistically Significant**

Andrew GELMAN and Hal STERN

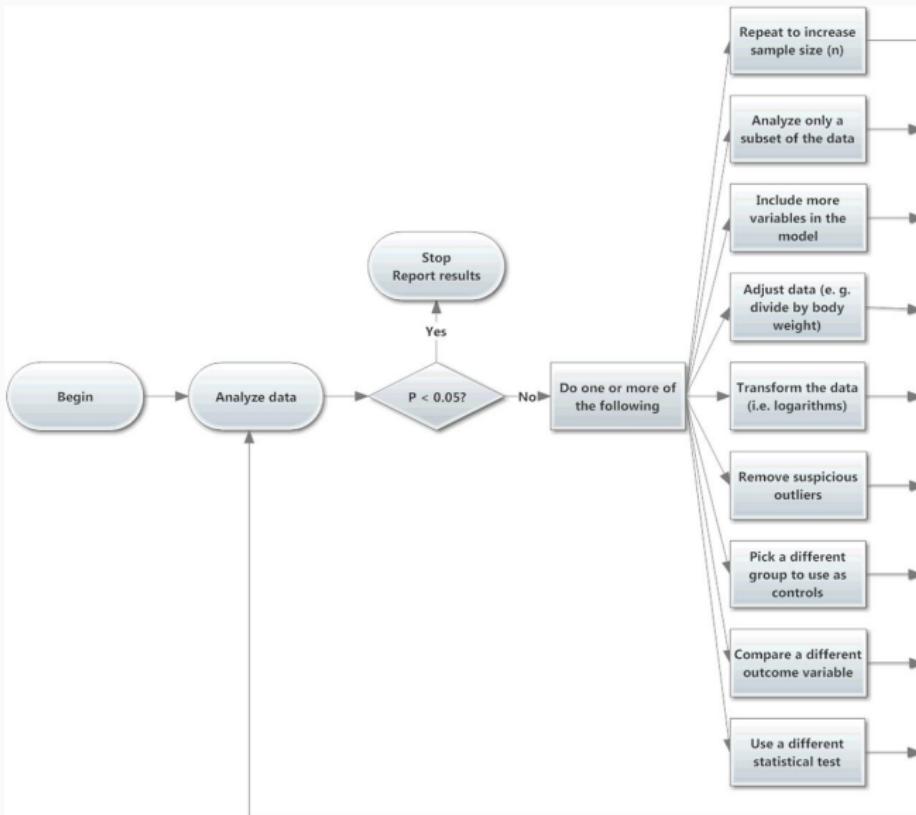
<http://dx.doi.org/10.1198/000313006X152649>

# Multiple hypothesis testing



<https://xkcd.com/882>

# 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](#)

## 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>

## 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>

## 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<sup>1</sup> · Stephen J. Senn<sup>2</sup> · Kenneth J. Rothman<sup>3</sup> · John B. Carlin<sup>4</sup> ·  
Charles Poole<sup>5</sup> · Steven N. Goodman<sup>6</sup> · Douglas G. Altman<sup>7</sup>

<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,<sup>1,†</sup> MAUREEN C. KENNEDY,<sup>2</sup> PATRICK G. CUNNINGHAM,<sup>3</sup> AND JOHN S. STANOVICK<sup>4</sup>

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

Also <http://www.statisticsdonewrong.com/>

# Good read



## Twenty tips for interpreting scientific claims

# The New Statistics

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



*General Article*

## The New Statistics: Why and How

**Geoff Cumming**

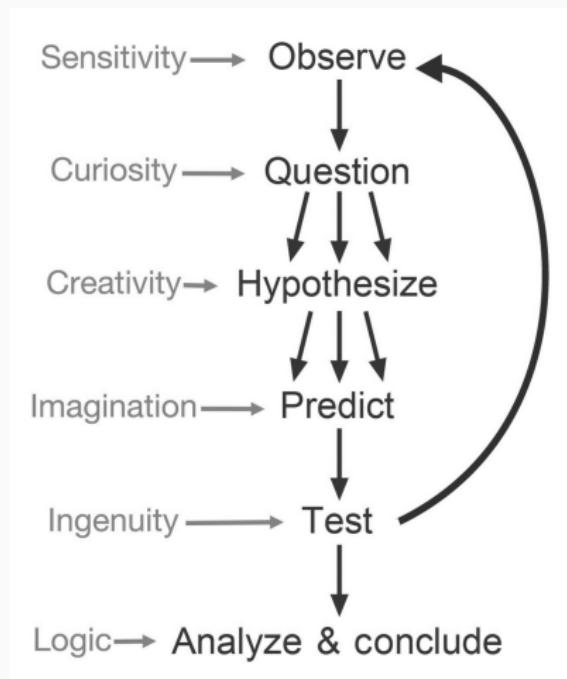
La Trobe University

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



<http://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