An introduction to statistical inference

Francisco Rodríguez-Sánchez

https://frodriguezsanchez.net

Why statistics?

To answer questions like...

- what's the probability that something occurs?

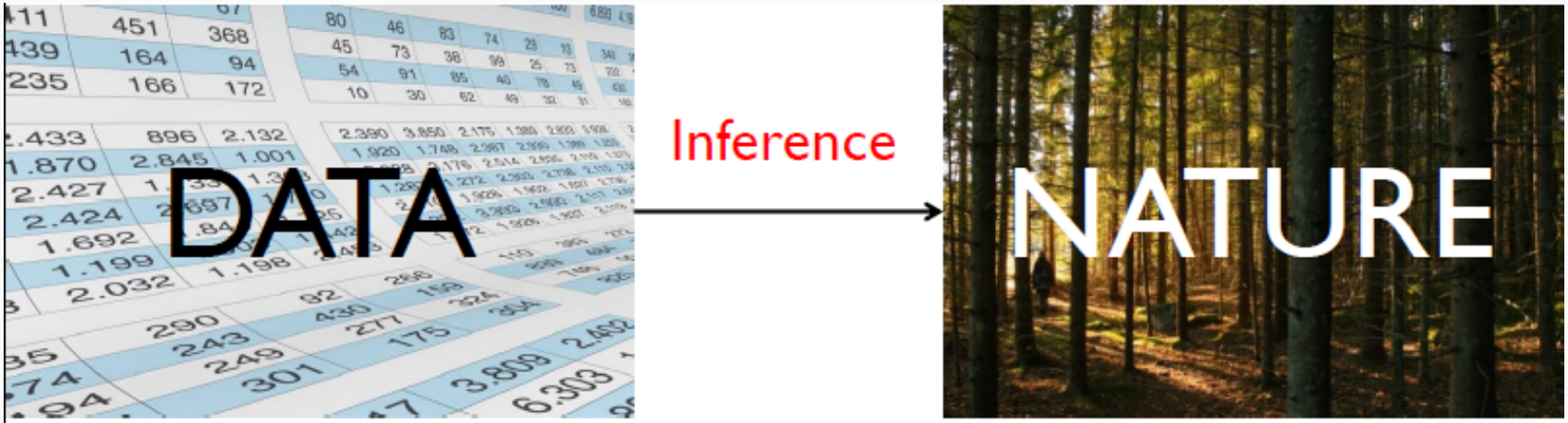
To answer questions like...

- 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



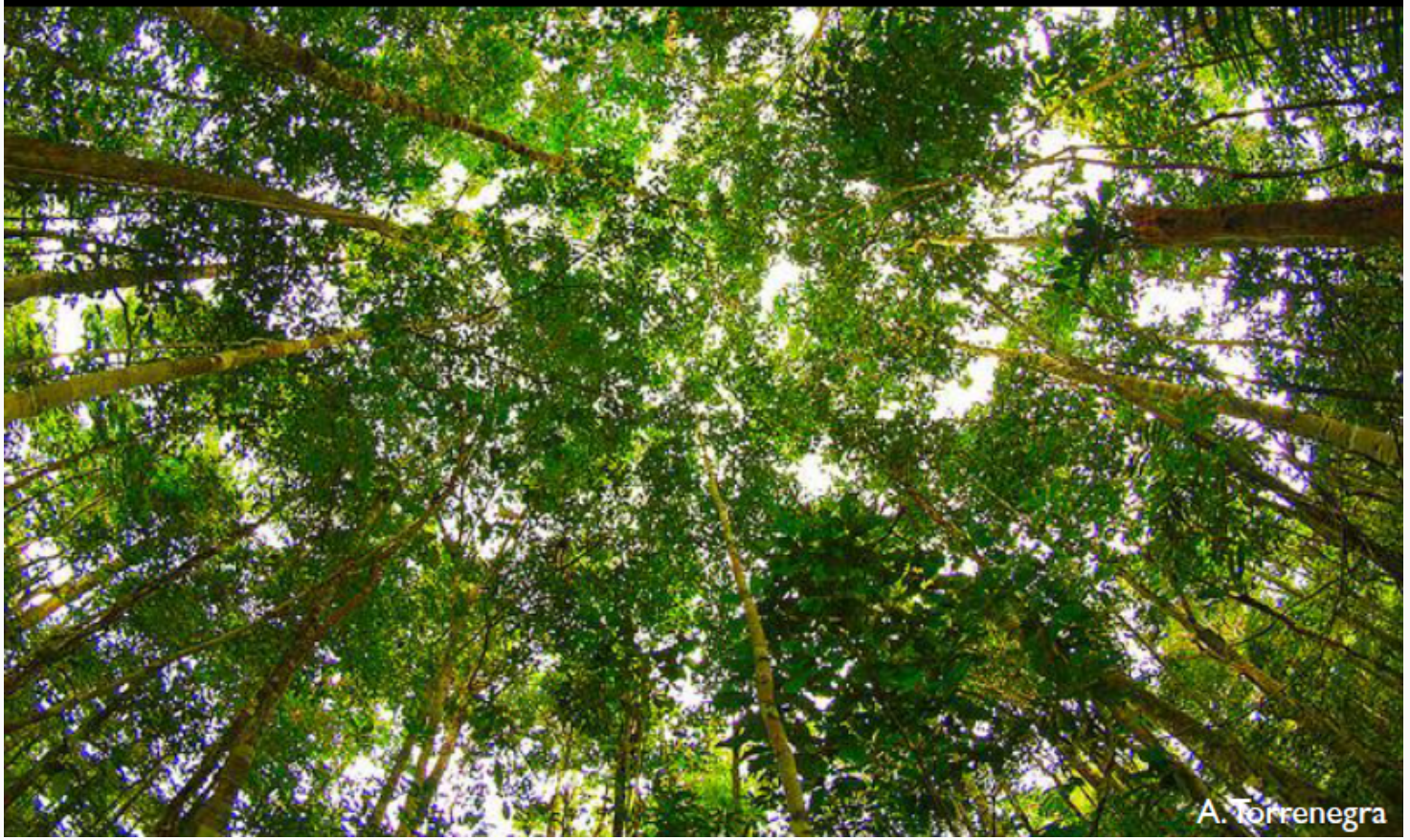
Bolker et al 2009 TREE:

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

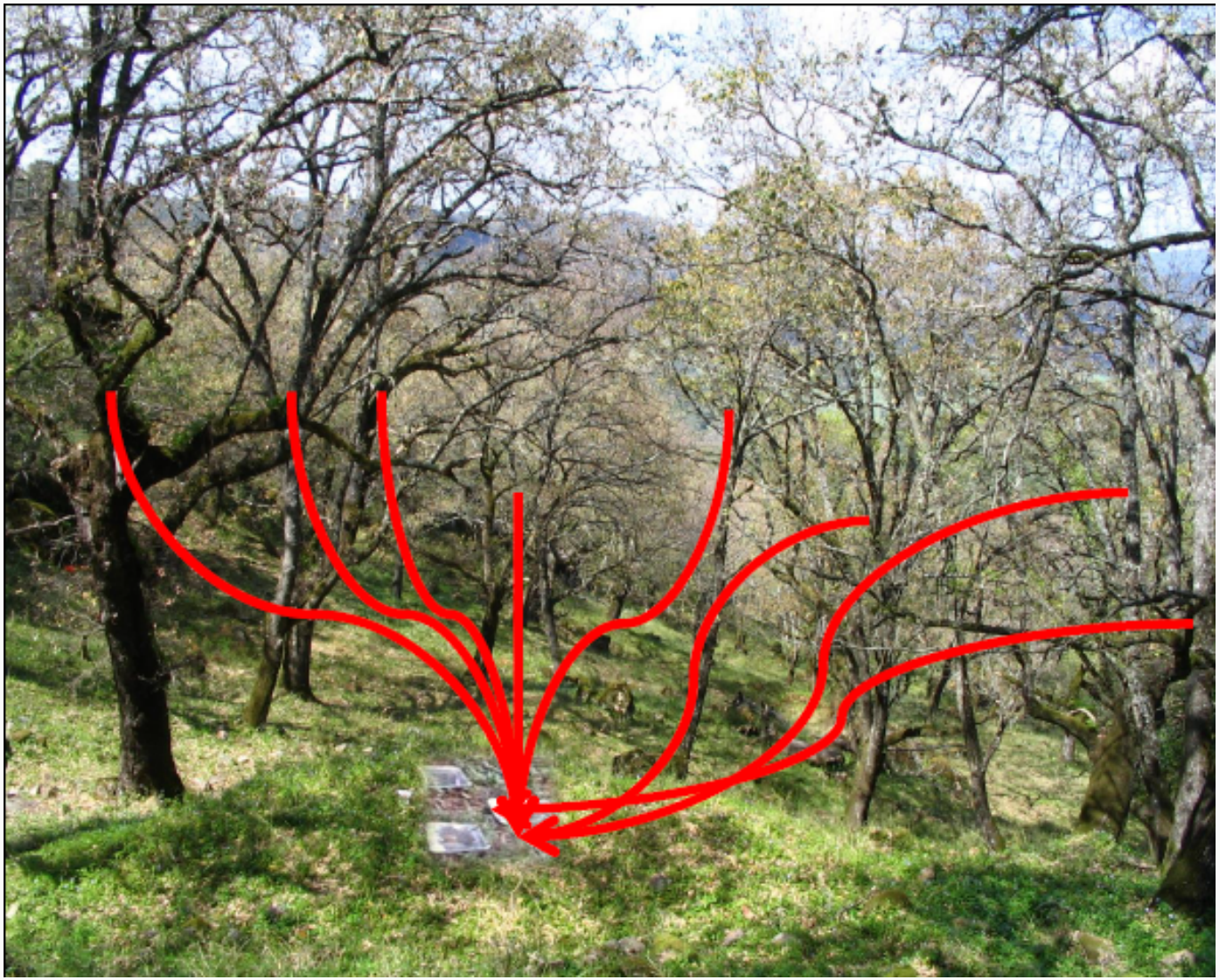
To get answers to tough problems

For example...

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

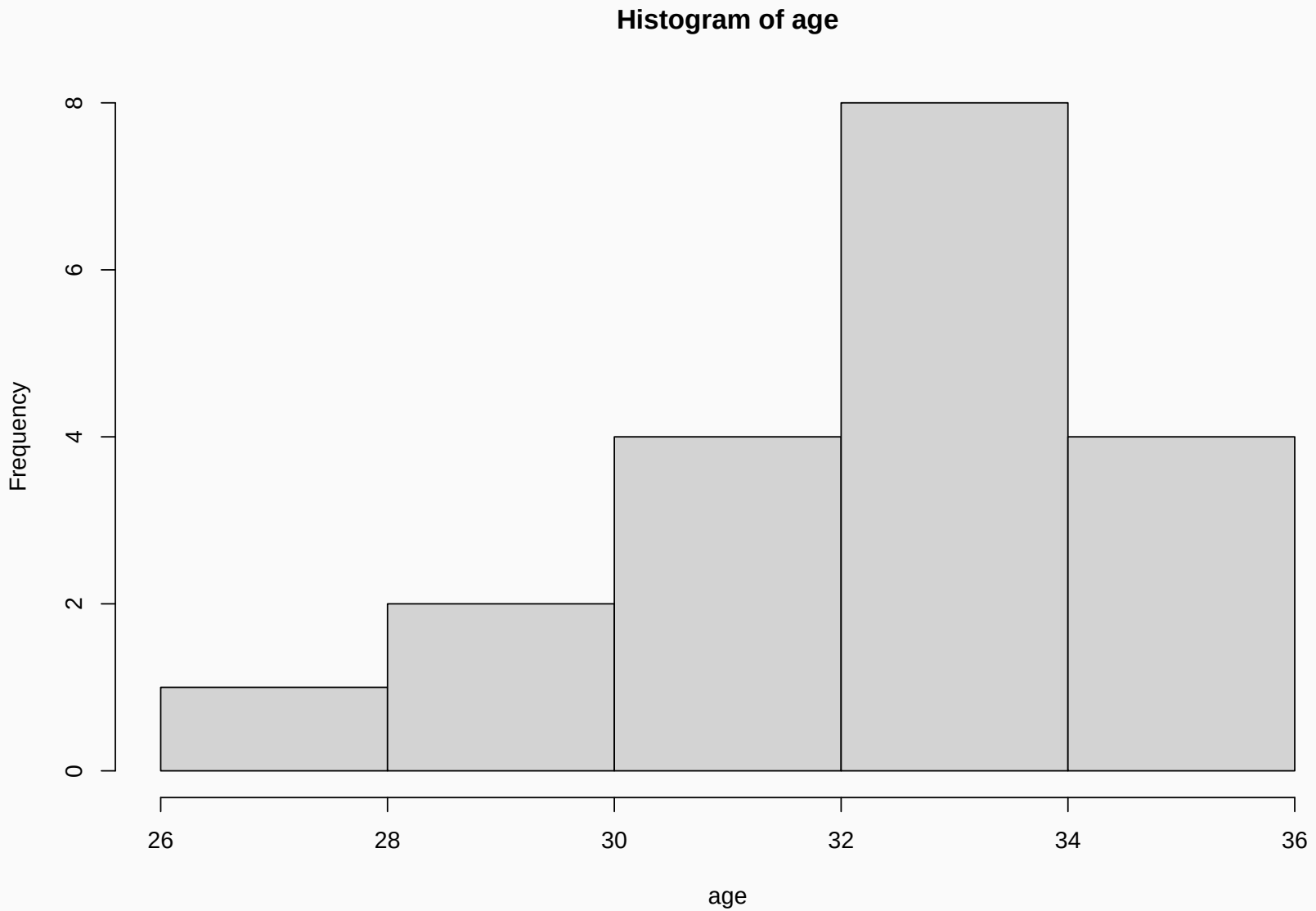
Topics

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

Descriptive statistics

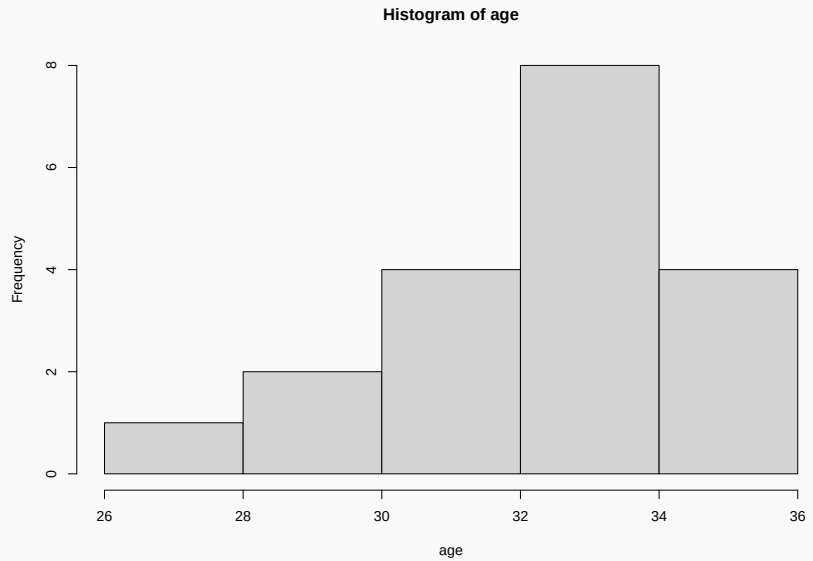
Guess my age

Graph your estimates



Do you think these data are good estimates of my age?

https://pollev.com/franciscorod726



Why / Why not?

Data are hardly ever objective.

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

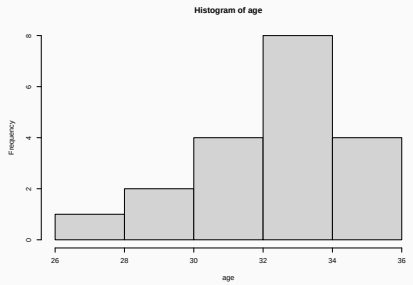
Always consider:

How well do these numbers reflect what we are trying to measure?

Summarise that distribution

Central tendency / location

- mean: $\frac{a_1 + a_2 + a_3}{n}$

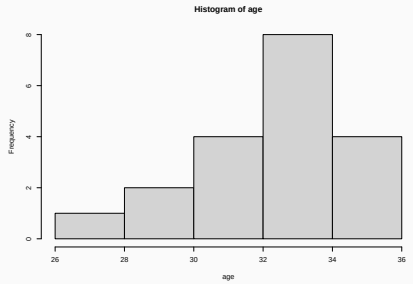


Variation / Spread

Summarise that distribution

Central tendency / location

- mean: $\frac{a_1 + a_2 + a_3}{n}$
- median (50% percentile)

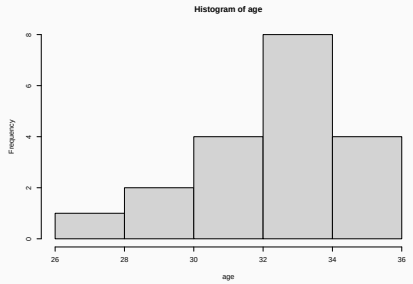


Variation / Spread

Summarise that distribution

Central tendency / location

- mean: $\frac{a_1 + a_2 + a_3}{n}$
- median (50% percentile)
- mode (most frequent value)

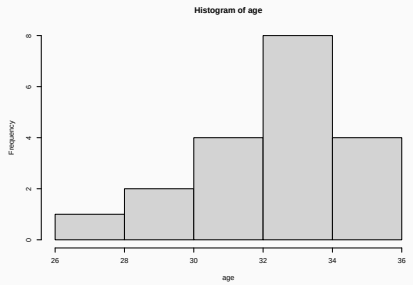


Variation / Spread

Summarise that distribution

Central tendency / location

- mean: $\frac{a_1 + a_2 + a_3}{n}$
- median (50% percentile)
- mode (most frequent value)



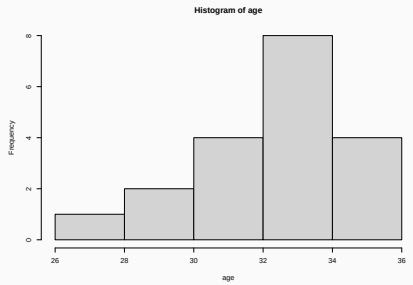
Variation / Spread

- min, max, range

Summarise that distribution

Central tendency / location

- mean: $\frac{a_1 + a_2 + a_3}{n}$
- median (50% percentile)
- mode (most frequent value)



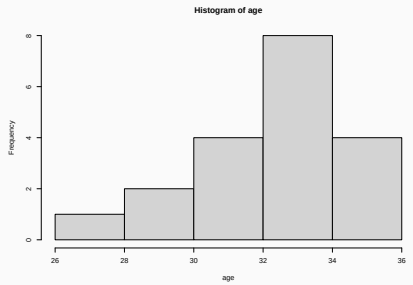
Variation / Spread

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

Summarise that distribution

Central tendency / location

- mean: $\frac{a_1 + a_2 + a_3}{n}$
- median (50% percentile)
- mode (most frequent value)



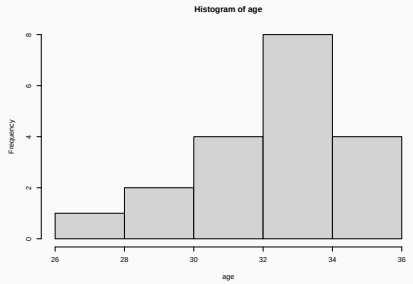
Variation / Spread

- min, max, range
- quantiles (quartiles, percentiles...)
- standard deviation: $SD = \sqrt{\frac{\sum (x - \mu)^2}{n - 1}}$

Summarise that distribution

Central tendency / location

- mean: $\frac{a_1 + a_2 + a_3}{n}$
- median (50% percentile)
- mode (most frequent value)



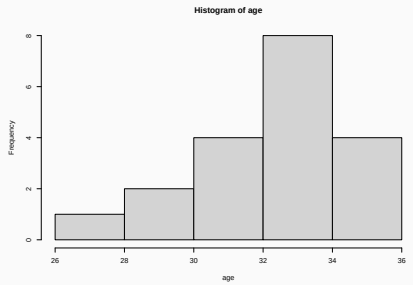
Variation / Spread

- min, max, range
- quantiles (quartiles, percentiles...)
- standard deviation: $SD = \sqrt{\frac{\sum (x - \mu)^2}{n - 1}}$
- standard error: $SEM = \frac{SD}{\sqrt{n}}$

Summarise that distribution

Central tendency / location

- mean: $\frac{a_1 + a_2 + a_3}{n}$
- median (50% percentile)
- mode (most frequent value)



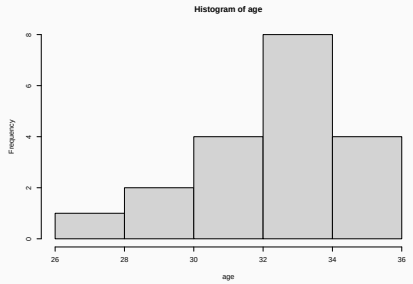
Variation / Spread

- min, max, range
- quantiles (quartiles, percentiles...)
- standard deviation: $SD = \sqrt{\frac{\sum (x - \mu)^2}{n - 1}}$
- standard error: $SEM = \frac{SD}{\sqrt{n}}$
- coefficient of variation (CV = SD / mean)

Summarise that distribution

Central tendency / location

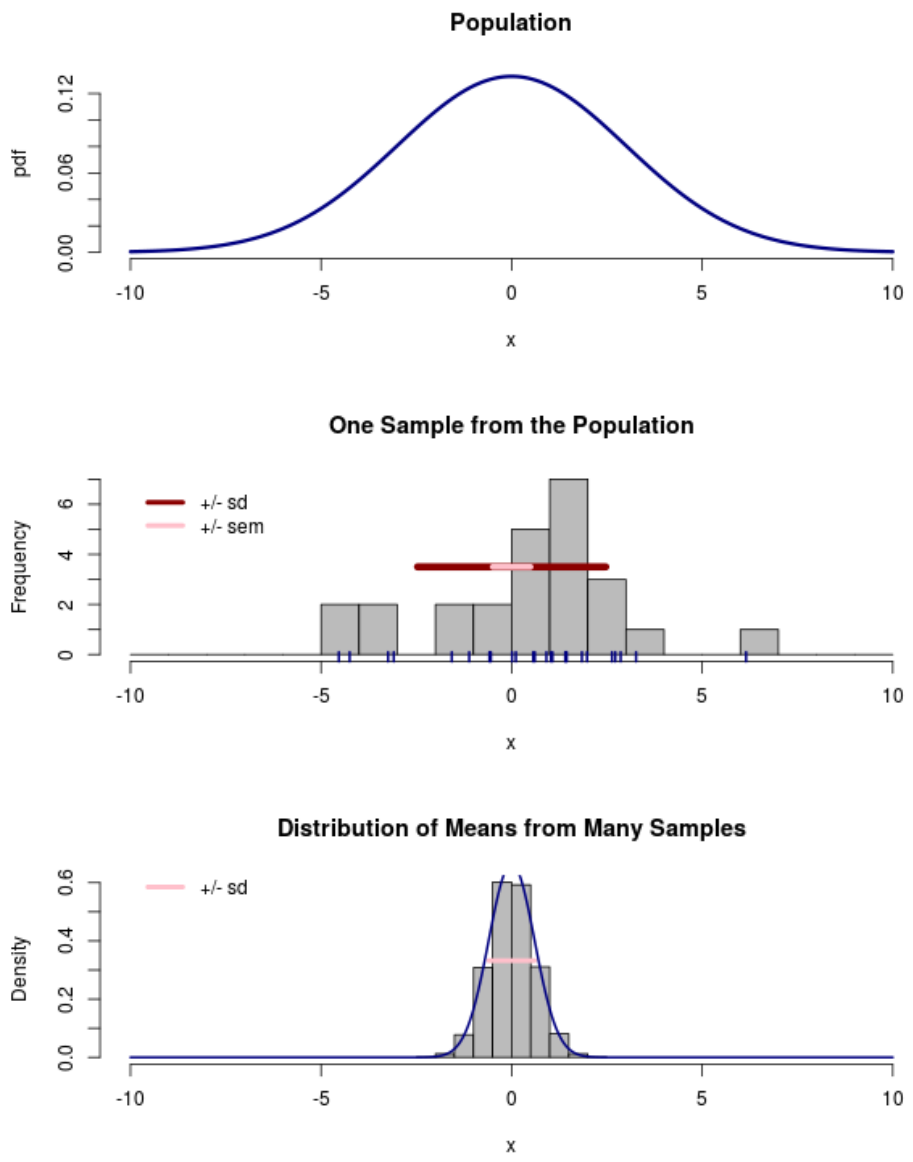
- mean: $\frac{a_1 + a_2 + a_3}{n}$
- median (50% percentile)
- mode (most frequent value)



Variation / Spread

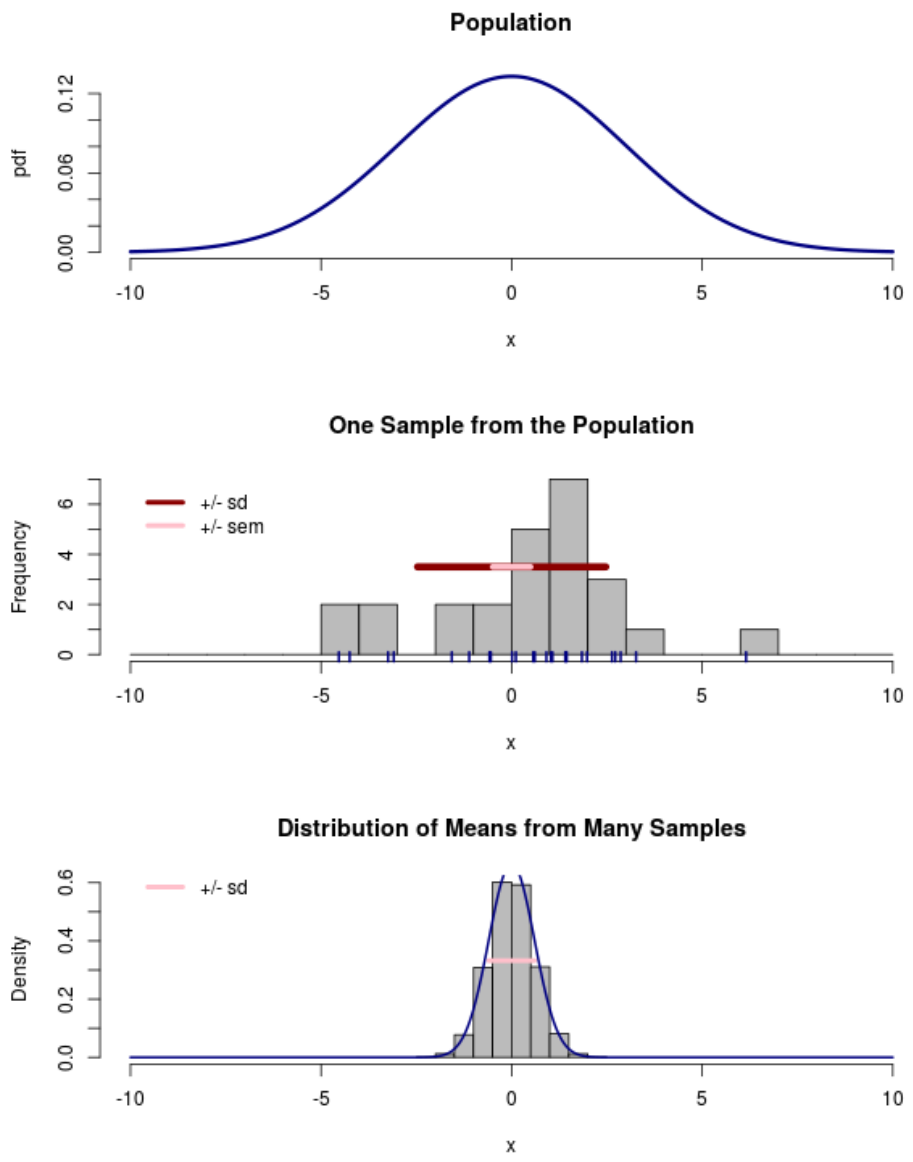
- min, max, range
- quantiles (quartiles, percentiles...)
- standard deviation: $SD = \sqrt{\frac{\sum (x - \mu)^2}{n - 1}}$
- standard error: $SEM = \frac{SD}{\sqrt{n}}$
- coefficient of variation (CV = SD / mean)
- confidence intervals

Relationship between SD and SEM



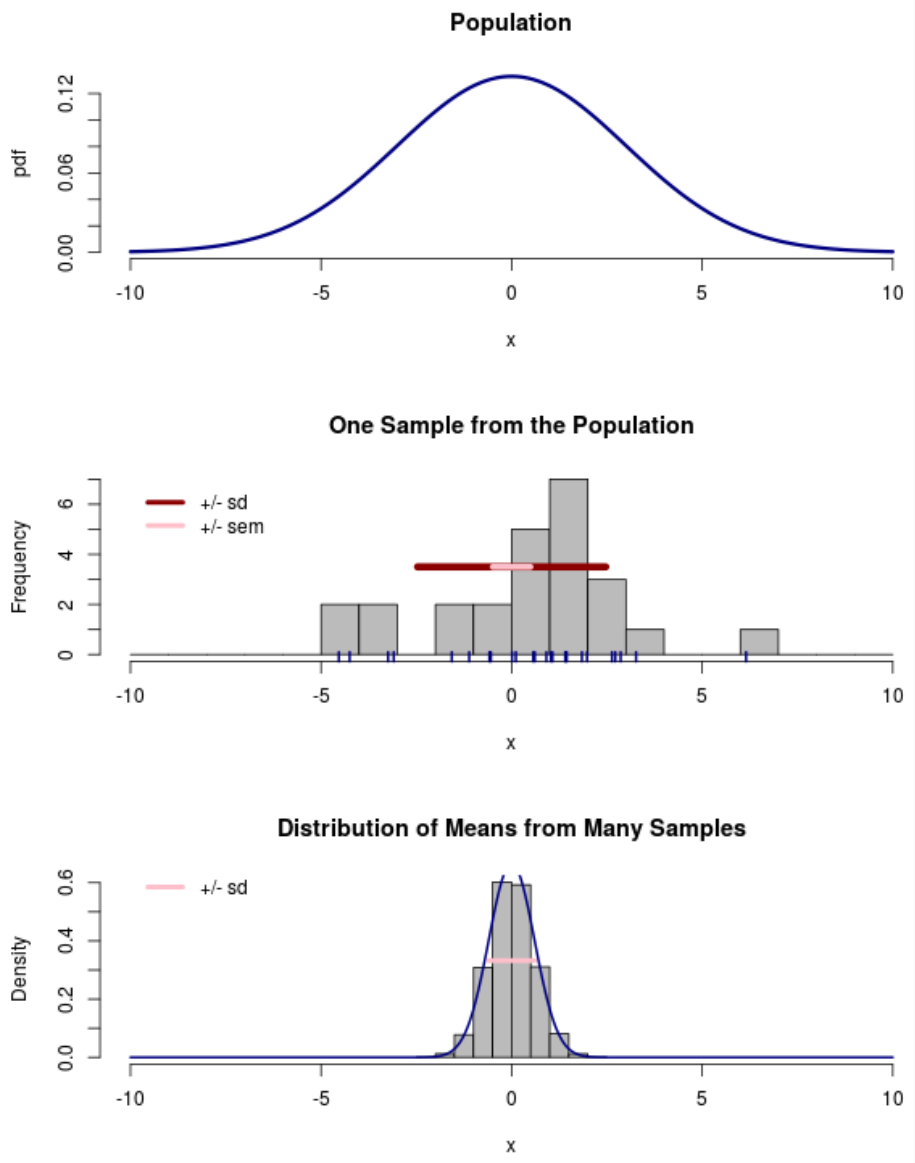
- SD quantifies scatter in population

Relationship between SD and SEM



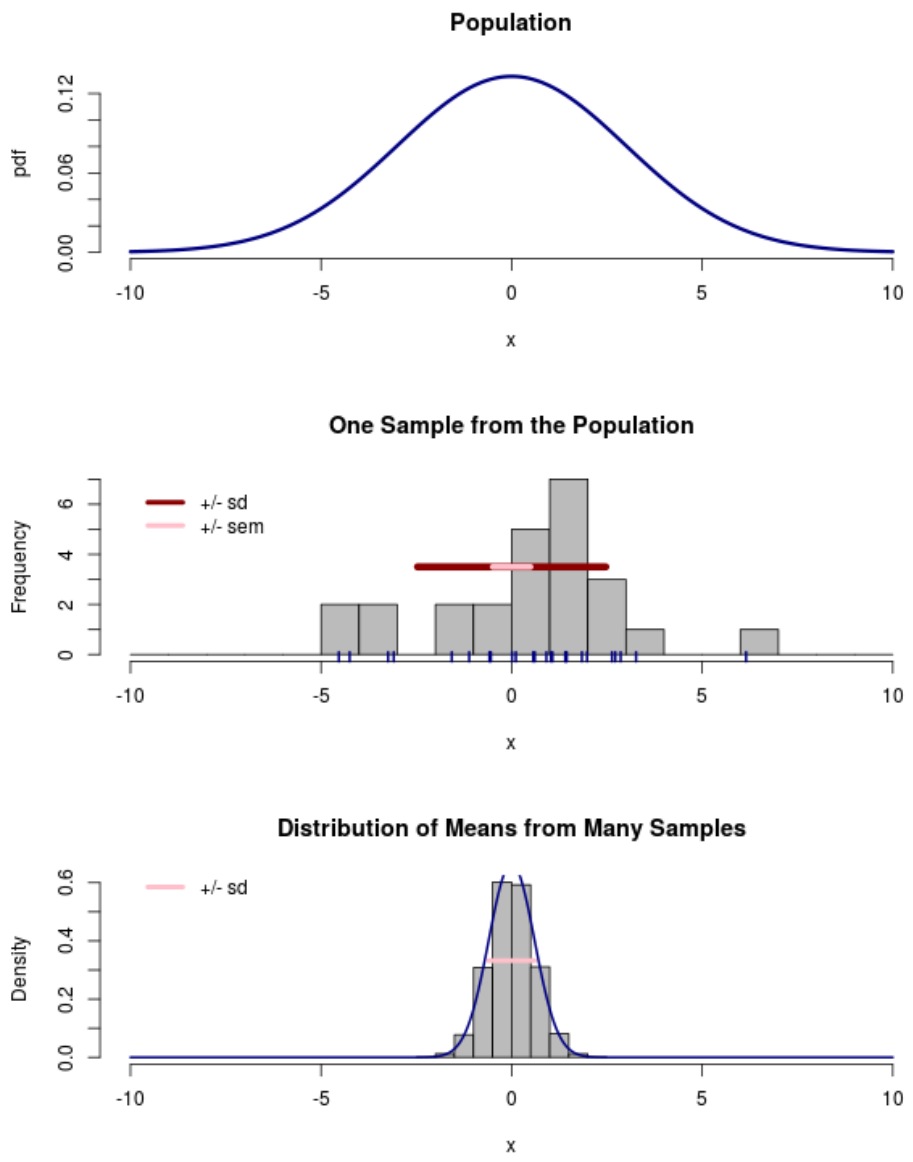
- SD quantifies scatter in population
- SEM quantifies uncertainty in parameter estimate (population mean)

Relationship between SD and SEM



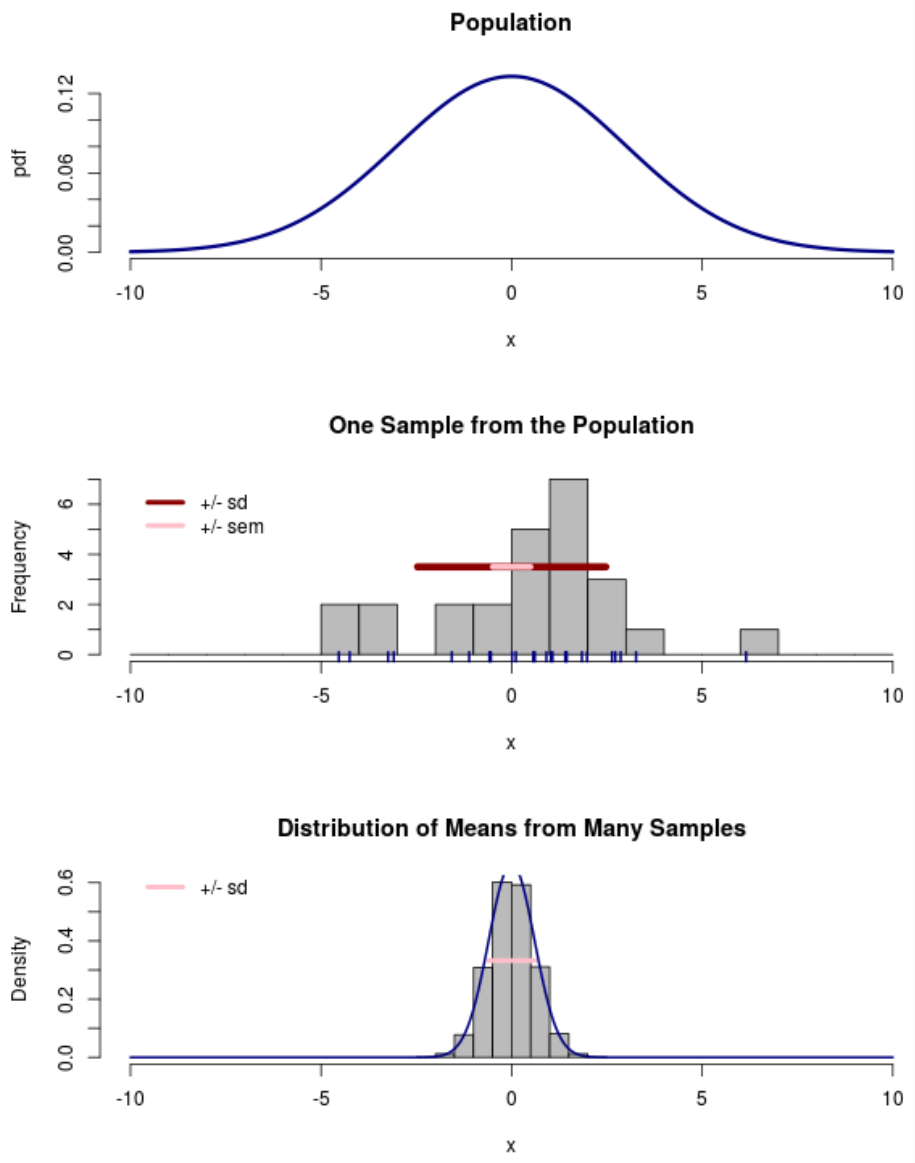
- SD quantifies scatter in population
- SEM quantifies uncertainty in parameter estimate (population mean)
- SEM = SD/sqrt(n)

Relationship between SD and SEM



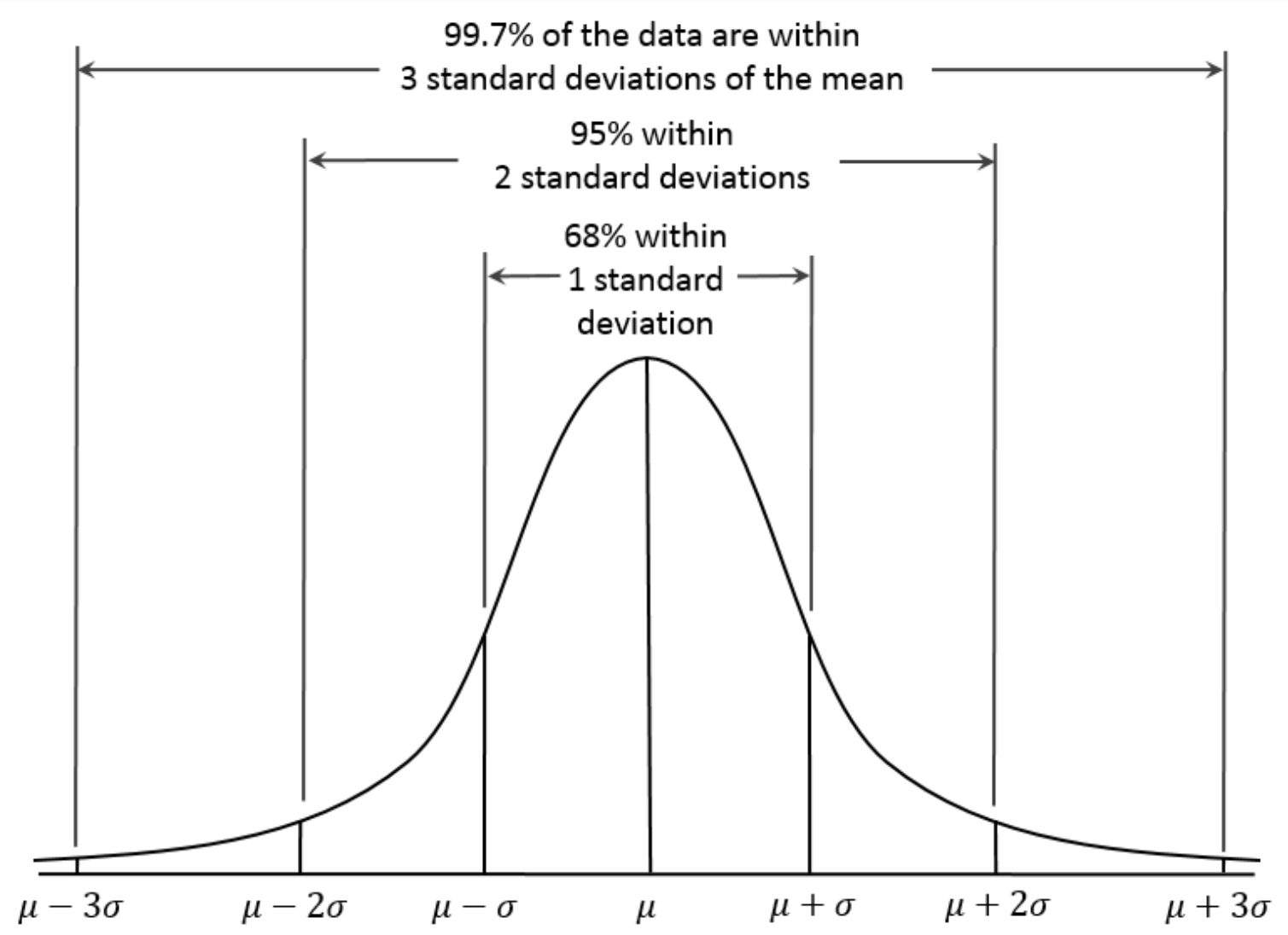
- SD quantifies scatter in population
- SEM quantifies uncertainty in parameter estimate (population mean)
- SEM = SD/sqrt(n)
- SEM decreases with sample size (mean better known), SD does not.

Relationship between SD and SEM



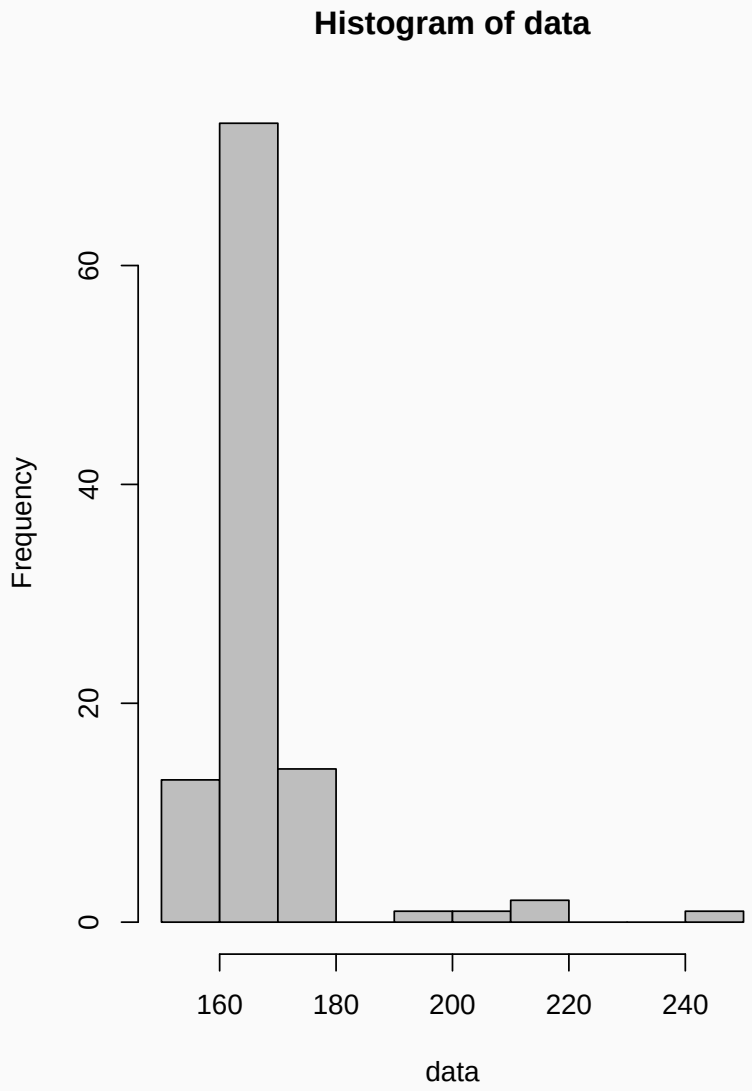
- SD quantifies scatter in population
- SEM quantifies uncertainty in parameter estimate (population mean)
- SEM = SD/sqrt(n)
- SEM decreases with sample size (mean better known), SD does not.
- https://gallery.shinyapps.io/sampling_and_stderr/

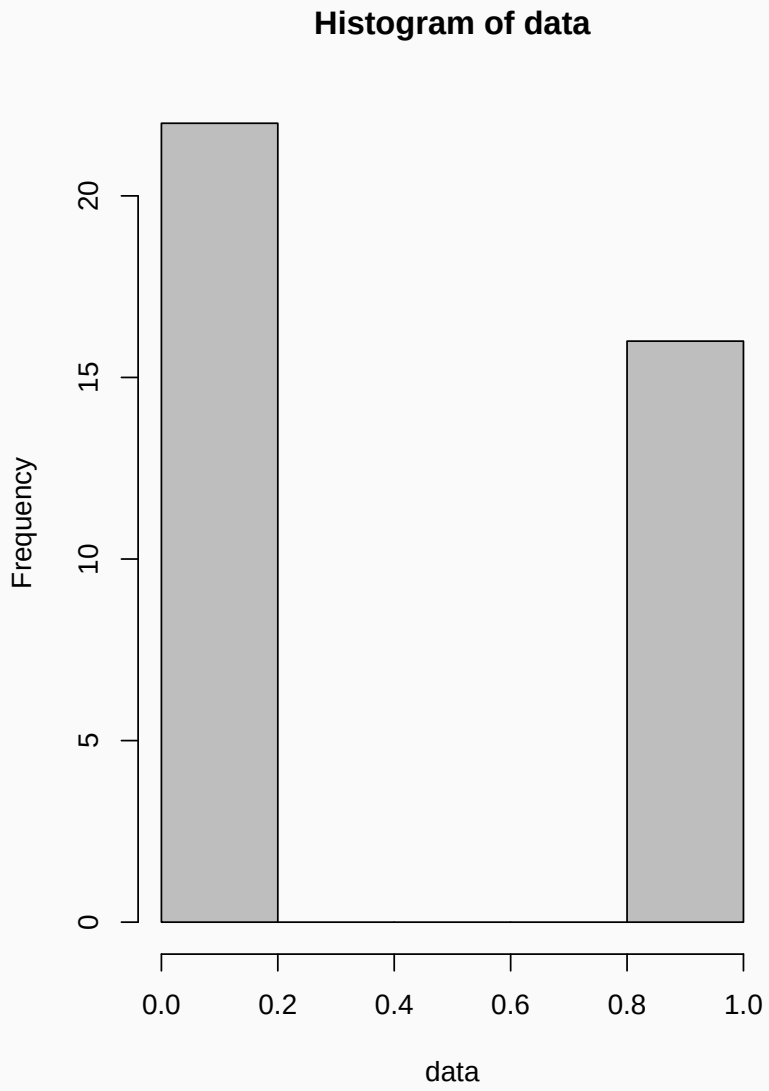
In a Normal distribution



What statistical descriptors are best? (and why)

https://pollev.com/franciscorod726





Sampling, confidence intervals, and Bayesian inference

Inference: from samples to population

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

What’s the average height in this group?

1. Write down your height and place of origin (Sevilla or other) in a piece of paper and put it in the bag.

What’s the average height in this group?

1. Write down your height and place of origin (Sevilla or other) in a piece of paper and put it in the bag.
2. Now everyone **sample** 5 individuals from the whole **population** of heights.

What's the average height in this group?

1. Write down your height and place of origin (Sevilla or other) in a piece of paper and put it in the bag.
2. Now everyone **sample** 5 individuals from the whole **population** of heights.
3. Calculate the mean and 95% CI for your sample (http://graphpad.com/quickcalcs/CImean1/).

What's the average height in this group?

1. Write down your height and place of origin (Sevilla or other) in a piece of paper and put it in the bag.
2. Now everyone **sample** 5 individuals from the whole **population** of heights.
3. Calculate the mean and 95% CI for your sample (http://graphpad.com/quickcalcs/CImean1/).
4. Draw on blackboard.

What's the average height in this group?

1. Write down your height and place of origin (Sevilla or other) in a piece of paper and put it in the bag.
2. Now everyone **sample** 5 individuals from the whole **population** of heights.
3. Calculate the mean and 95% CI for your sample (http://graphpad.com/quickcalcs/CImean1/).
4. Draw on blackboard.
5. Do all CIs contain true mean height?

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

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!

Bayesian credible intervals

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

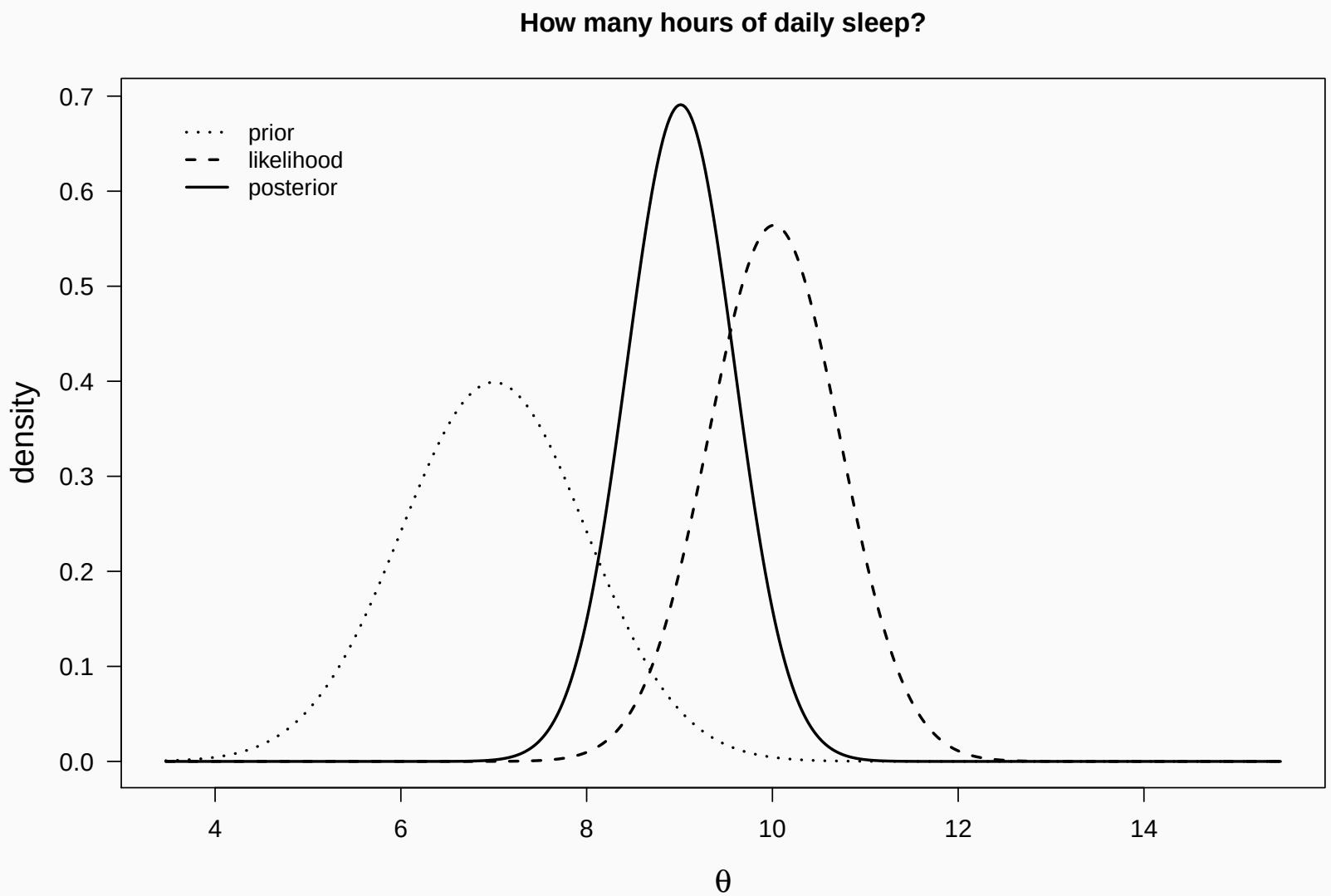
Bayesian credible intervals

- 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

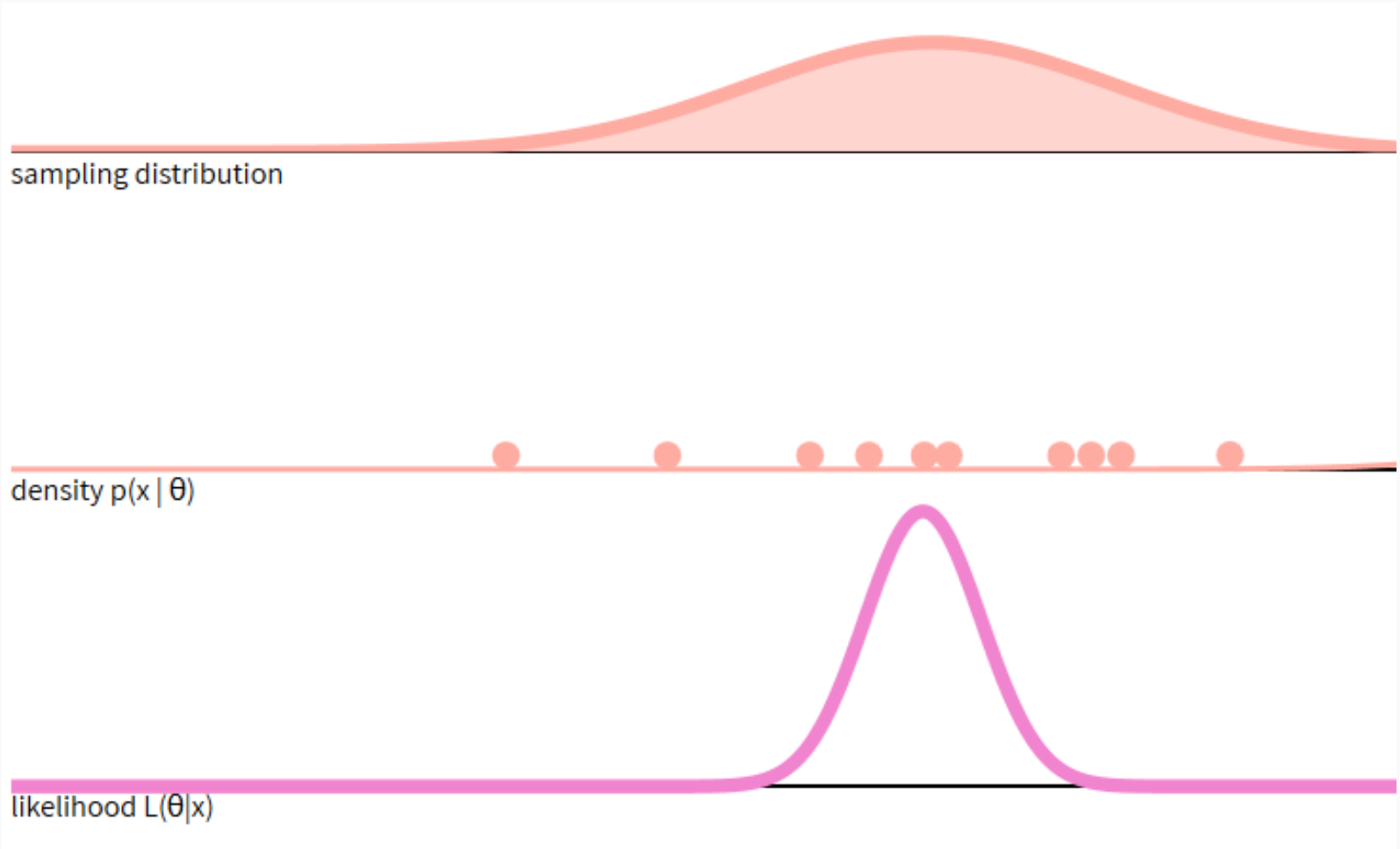
$P(H|D) \propto P(D|H) \cdot P(H)$

Posterior $\propto$ *Likelihood* $\cdot$ *Prior*



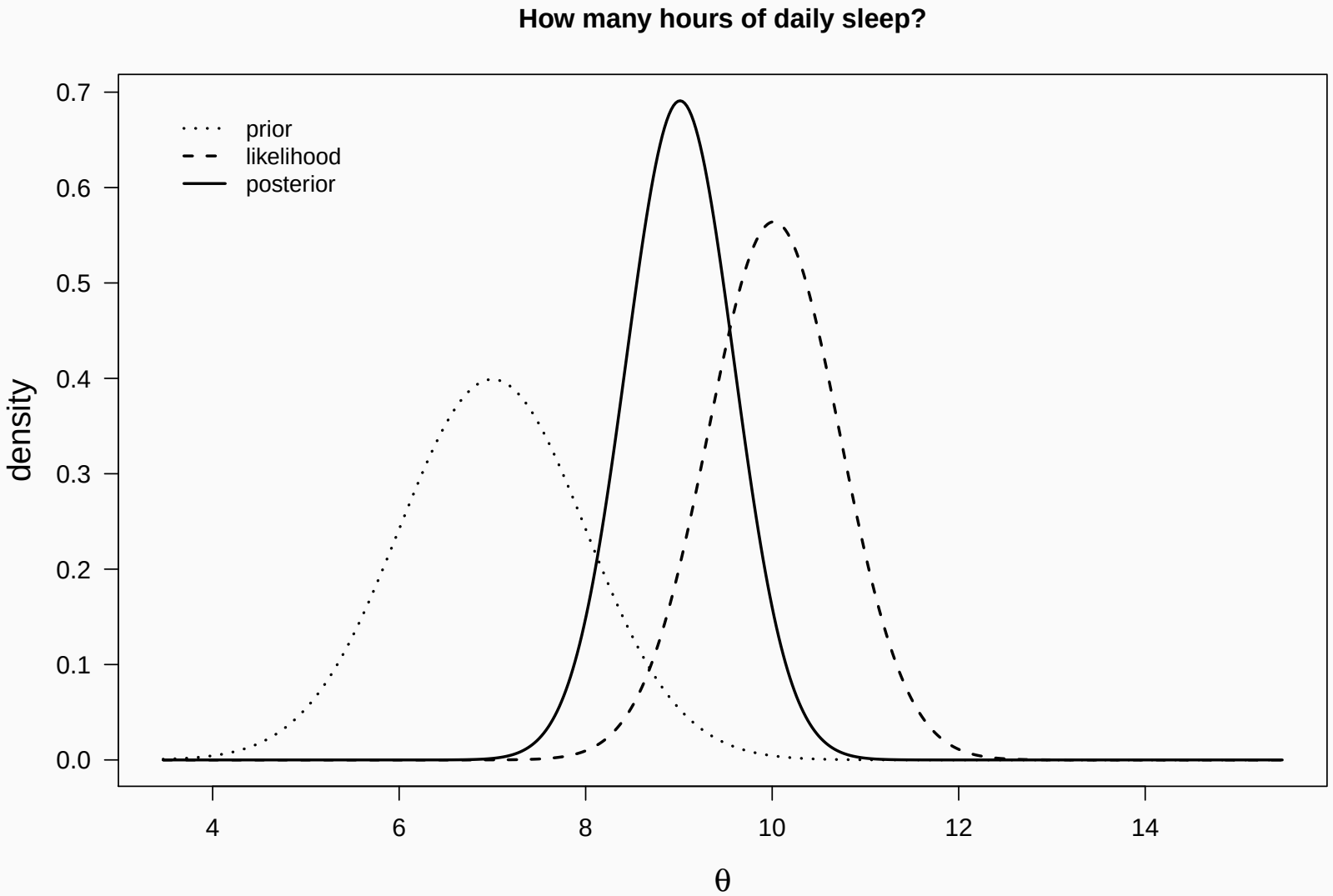
What is the likelihood?

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

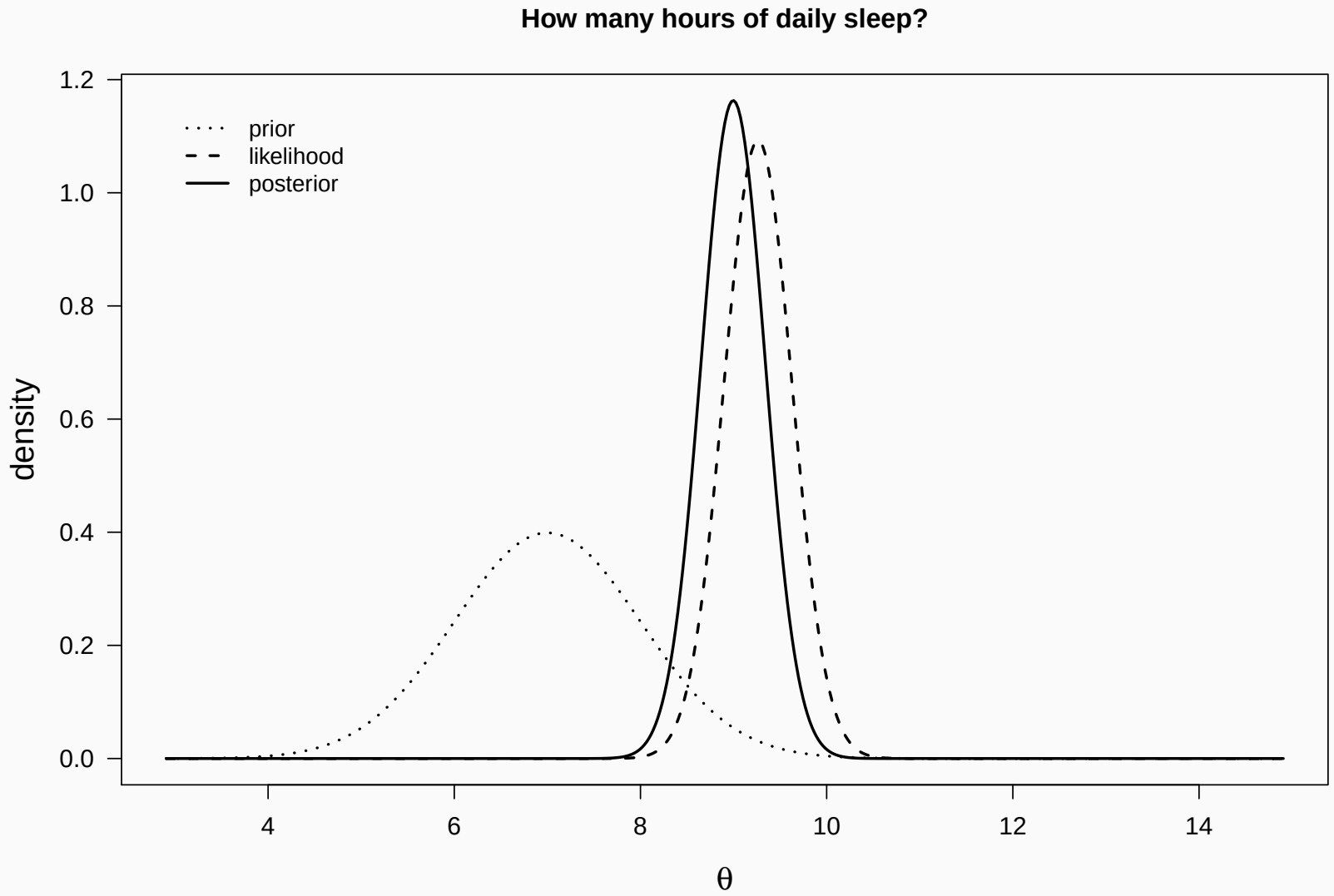


https://seeing-theory.brown.edu/bayesian-inference/index.html

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.

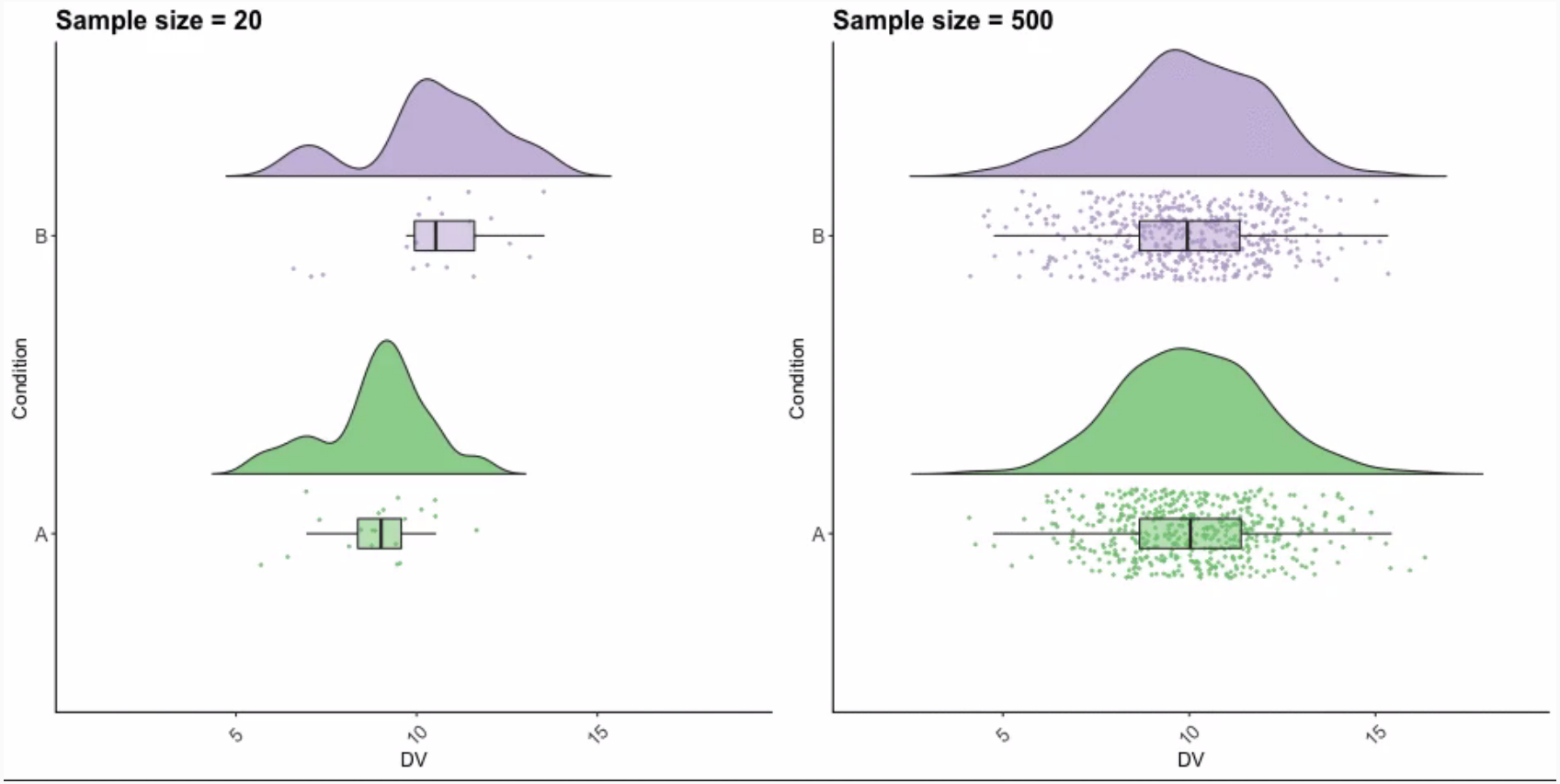
Replication

- Replication is key: we need several samples.
- How many? As much as you can! See Gelman & Carlin 2014.
- Traditionally, ecology studies have had **too low sample sizes**.

Replication

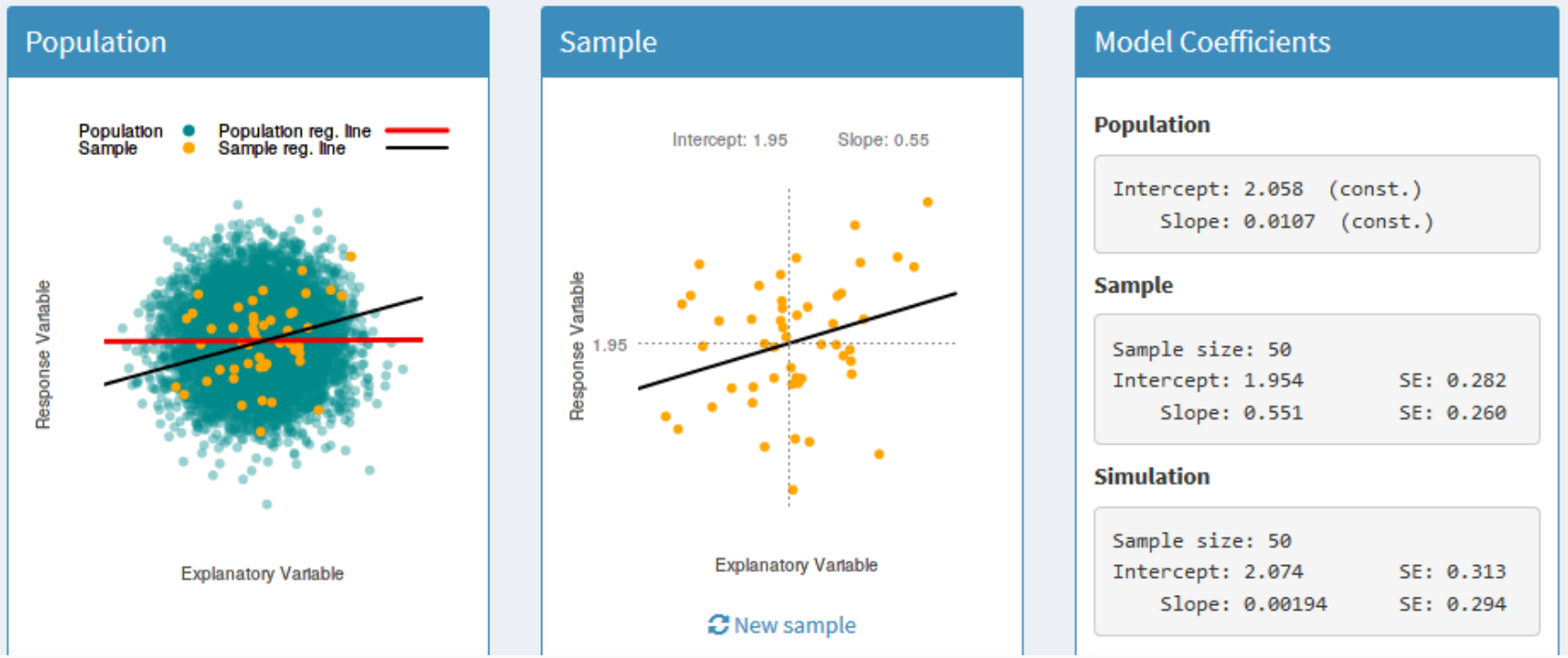
- Replication is key: we need several samples.
- How many? As much as you can! See Gelman & Carlin 2014.
- Traditionally, ecology studies have had **too low sample sizes**.
- Low sample sizes miss subtle effects, but also **prone to bias**.

Low sample sizes very sensitive to random noise



https://twitter.com/ajstewart_lang/status/1020038488278945797

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

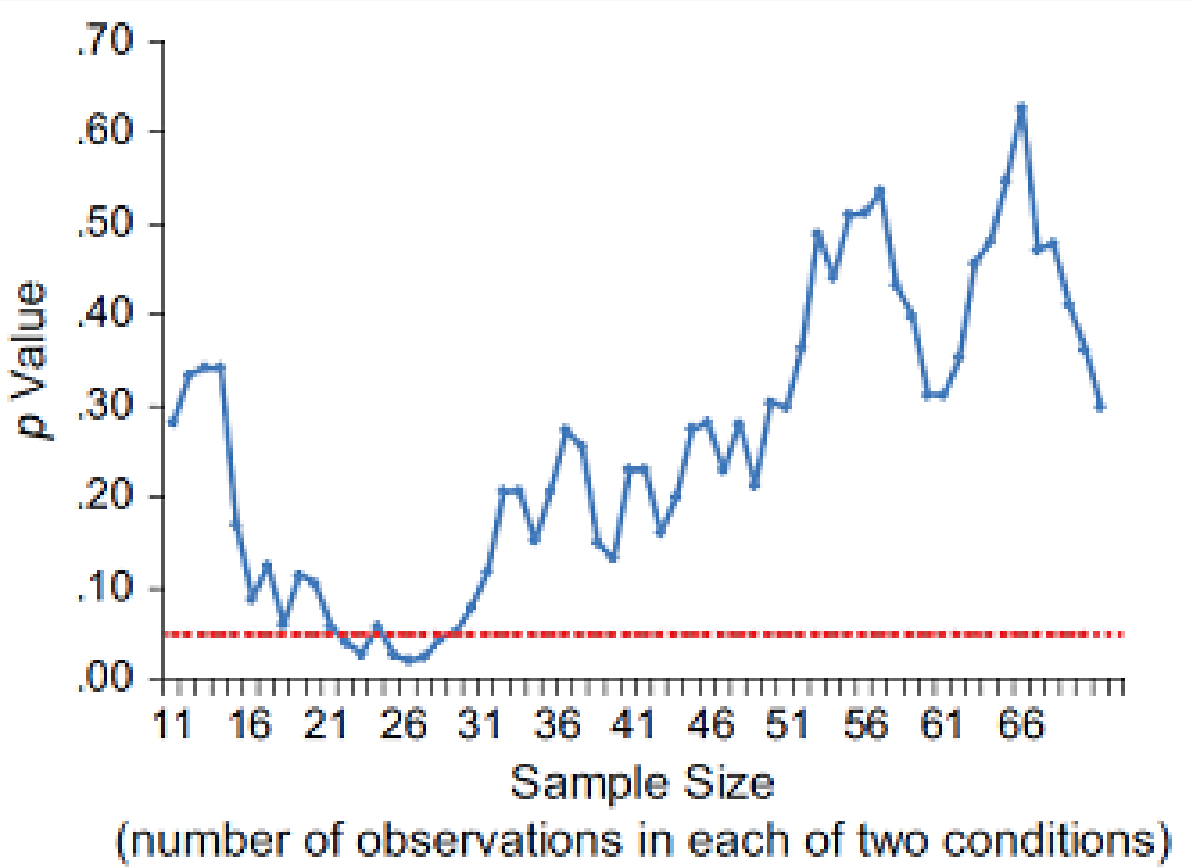


Fig. 2. Illustrative simulation of *p* values obtained by a researcher who continuously adds an observation to each of two conditions, conducting a *t* test after each addition. The dotted line highlights the conventional significance criterion of $p \leq .05$.

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.this, this or this).

Sample size estimation

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

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

Hypothesis testing

Do sleep hours differ between local and foreign students?

- Daily sleep hours by local students:

```
7 6 7 6 6
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    6.0     6.0     6.0     6.4     7.0     7.0 
```

```
9 9 9 9 11 8 6 10 8 8
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    6.0     8.0     9.0     8.7     9.0    11.0 
```

Do sleep hours differ between local and foreign students?

- Daily sleep hours by local students:

```
7 6 7 6 6
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    6.0     6.0     6.0     6.4     7.0     7.0 
```

- Daily sleep hours by non-local students:

```
9 9 9 9 11 8 6 10 8 8
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    6.0     8.0     9.0     8.7     9.0    11.0 
```

Do sleep hours differ between local and foreign students?

- Daily sleep hours by local students:

```
7 6 7 6 6
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    6.0     6.0     6.0     6.4     7.0     7.0
```

- Daily sleep hours by non-local students:

```
9 9 9 9 11 8 6 10 8 8
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    6.0     8.0     9.0     8.7     9.0    11.0
```

- We know what happens in **our samples**, but want to extrapolate to the whole **population**.

If we sample students in this class...

- Can we extrapolate results to

If we sample students in this class...

- Can we extrapolate results to
    - this class?

If we sample students in this class...

- Can we extrapolate results to
  - this class?
  - this university?

If we sample students in this class...

- Can we extrapolate results to
  - this class?
  - this university?
  - this city?

If we sample students in this class...

- Can we extrapolate results to
  - this class?
  - this university?
  - this city?
  - the world?

If we sample students in this class...

- Can we extrapolate results to
  - this class?
  - this university?
  - this city?
  - the world?
- What's the **suitable population** to make inferences given this sample?

NHST concepts

Null and alternative hypotheses

- Tell me...

Null and alternative hypotheses

- Tell me...
- **Null hypothesis**: there is no difference between groups.

Null and alternative hypotheses

- Tell me...
- **Null hypothesis**: there is no difference between groups.
- **Alternative hypothesis**: groups are different.

In ecology, everything is somewhat different

Are there any differences? A non-sensical question in ecology

Alejandro Martínez-Abraín

IMEDEA (CSIC-UIB), C/Miquel Marquès 21, 07190 Esporles, Majorca, Spain

ARTICLE INFO

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

Keywords:

ABSTRACT

One of the main questions that ecologists pose in their investigations includes the analysis of differences in some trait between two or more populations. I argue here that asking whether there are differences or not between populations is biologically irrelevant, since no two livings things are ever equal. On the contrary the appropriate question to pose is how large differences are between populations. That is, we urge a shift in interest from statistical significance to biological relevance for proper knowledge accumulation. I empha-

What is the p-value?

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

https://pollev.com/franciscorod726

What is the p-value?

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

https://pollev.com/franciscorod726

What is the p-value?

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

https://pollev.com/franciscorod726

P-value

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

P-value

- Very complicated concept: even statisticians fail to describe it well.
- Probability of observing data as or more extreme than these *if H0 was true*.

P-value

- Very complicated concept: even statisticians fail to describe it well.
- Probability of observing data as or more extreme than these *if H0 was true*.
- Low P-value: data unlikely if H0 was true.

P-value

- Very complicated concept: even statisticians fail to describe it well.
- Probability of observing data as or more extreme than these *if H0 was true*.
- Low P-value: data unlikely if H0 was true.
- Large P-value: data not unusual if H0 was true.

If p-value > 0.05

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

https://pollev.com/franciscorod726

If p-value > 0.05

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

https://pollev.com/franciscorod726

If p-value > 0.05

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

https://pollev.com/franciscorod726

If p-value > 0.05

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

https://pollev.com/franciscorod726

Are differences *significant*?

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

Are differences *significant*?

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

Are differences *significant*?

- If $p < 0.05$, we **reject** H0.
- If $p > 0.05$, we **fail to reject** H0
- (which is **NOT** the same as 'H0 is true')

Are differences *significant*?

- If p < 0.05, we **reject** H0.
- If p > 0.05, we **fail to reject** H0
- (which is **NOT** the same as 'H0 is true')
- **CAUTION:**

Are differences *significant*?

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

Are differences *significant*?

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

Are differences *significant*?

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

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

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

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

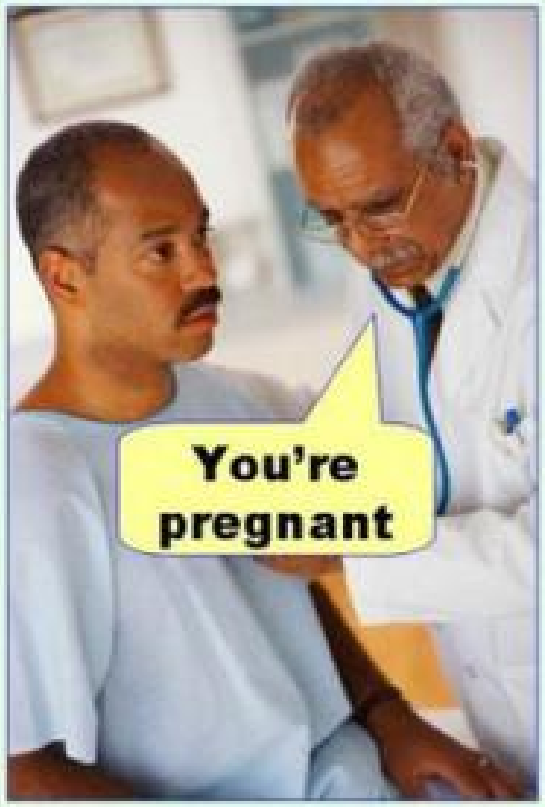
```
    Welch Two Sample t-test

data:  h.sevi and h.out
t = -2.9869, df = 5.8705, p-value = 0.02507
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -21.518508  -2.081492
sample estimates:
mean of x mean of y
    169.2     181.0
```

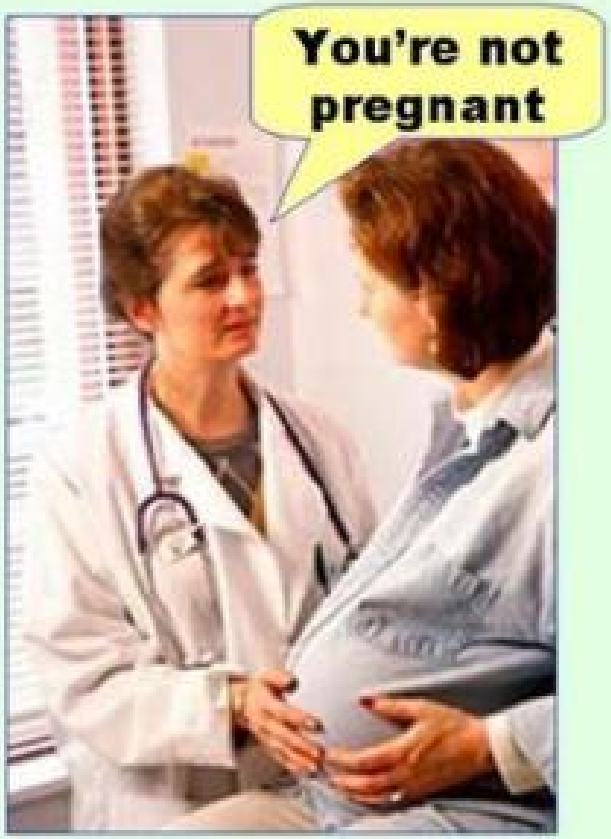
https://pollev.com/franciscorod726

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 H0 when it's false).

Understanding NHST

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

Example: biased coin

```
 [1] 0 1 0 1 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:
 0.2736697 0.8630694
sample estimates:
  p
0.6
```

https://pollev.com/franciscorod726

Correlation between variables

http://rpsychologist.com/d3/correlation/

Common pitfalls and good practice

A must read

Eur J Epidemiol (2016) 31:337–350
DOI 10.1007/s10654-016-0149-3

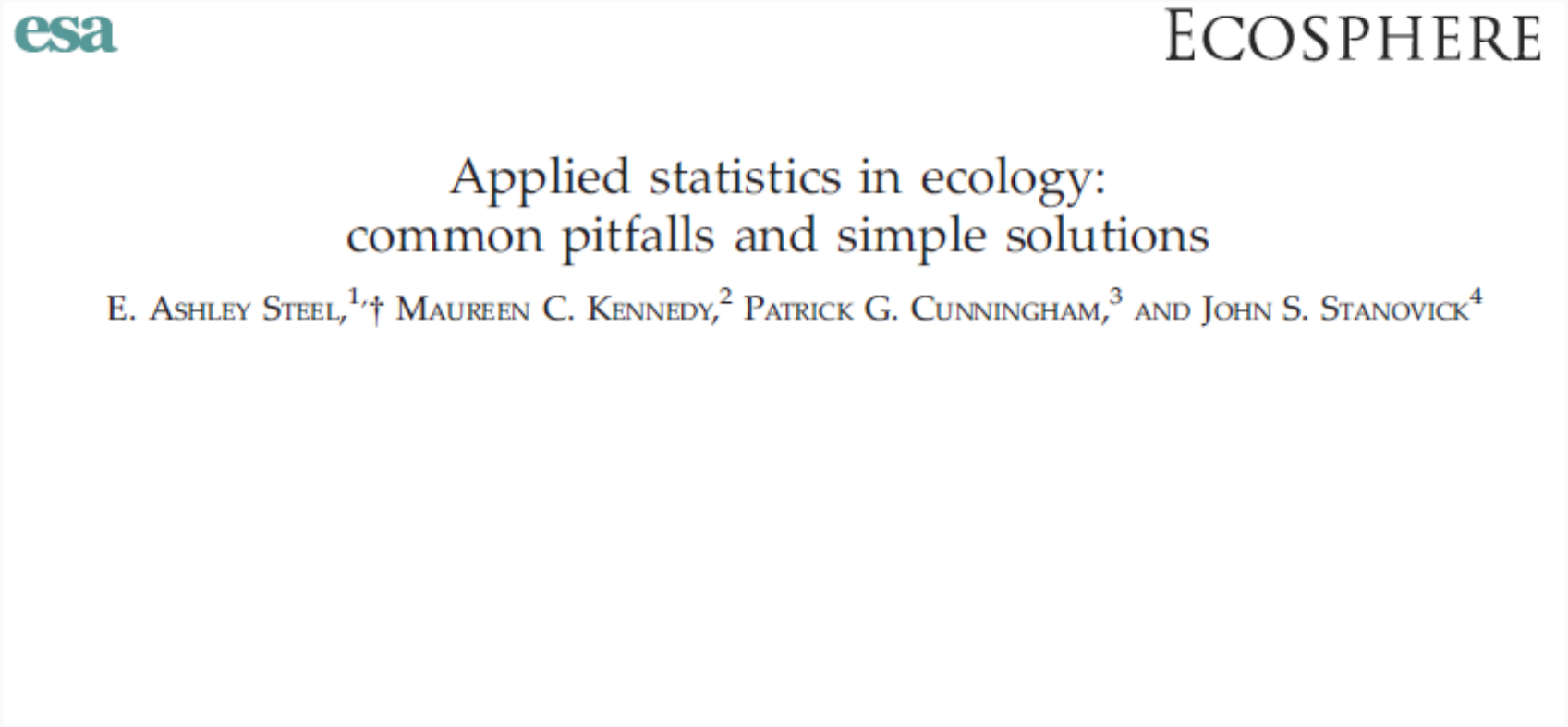
ESSAY

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



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

Also http://www.statisticsdonewrong.com/

Good read

Twenty tips for interpreting scientific claims

https://doi.org/10.1038/503335a

Visualisation of data and models is key

First things first

- Always

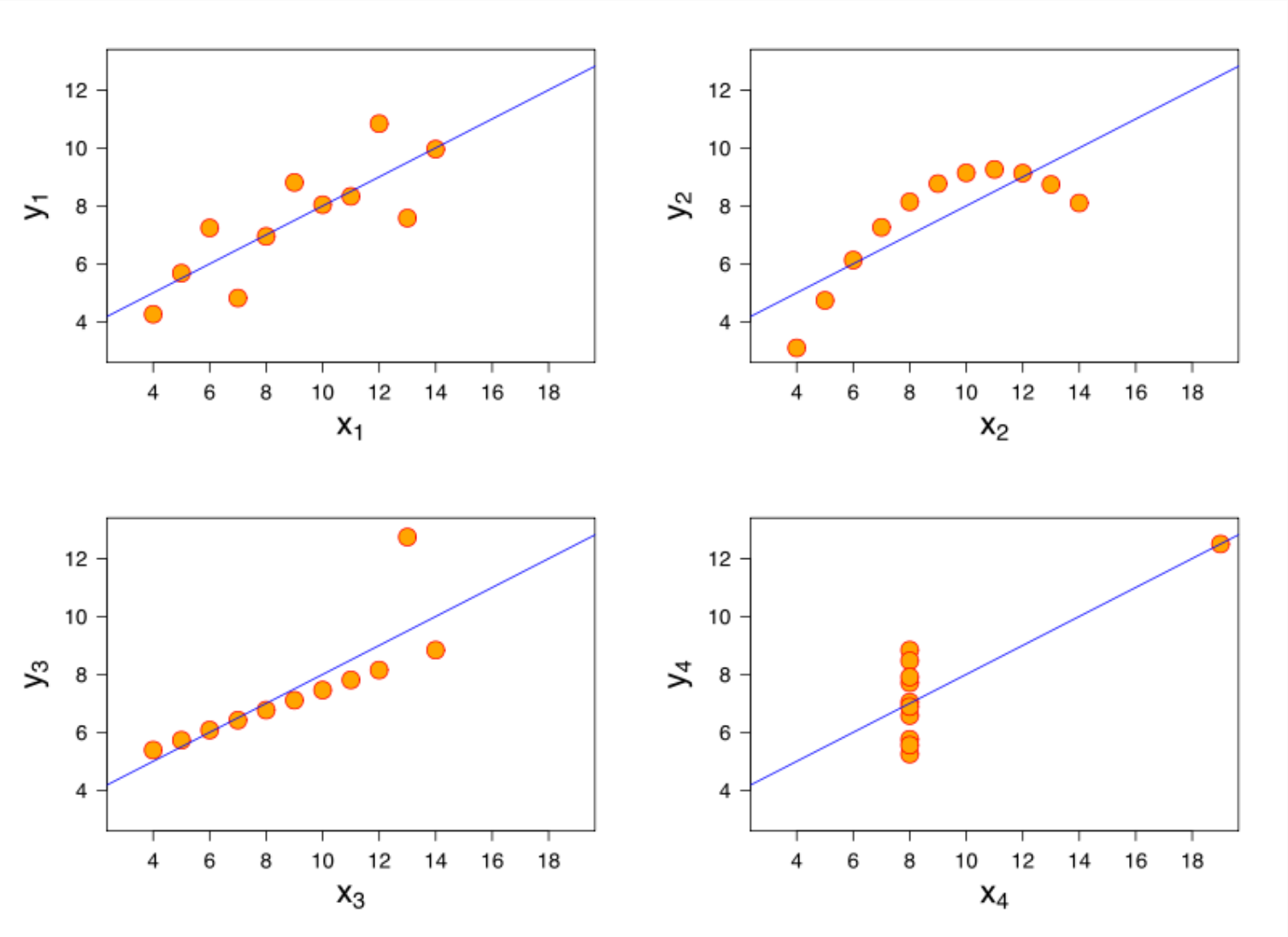
First things first

- Always
- Always

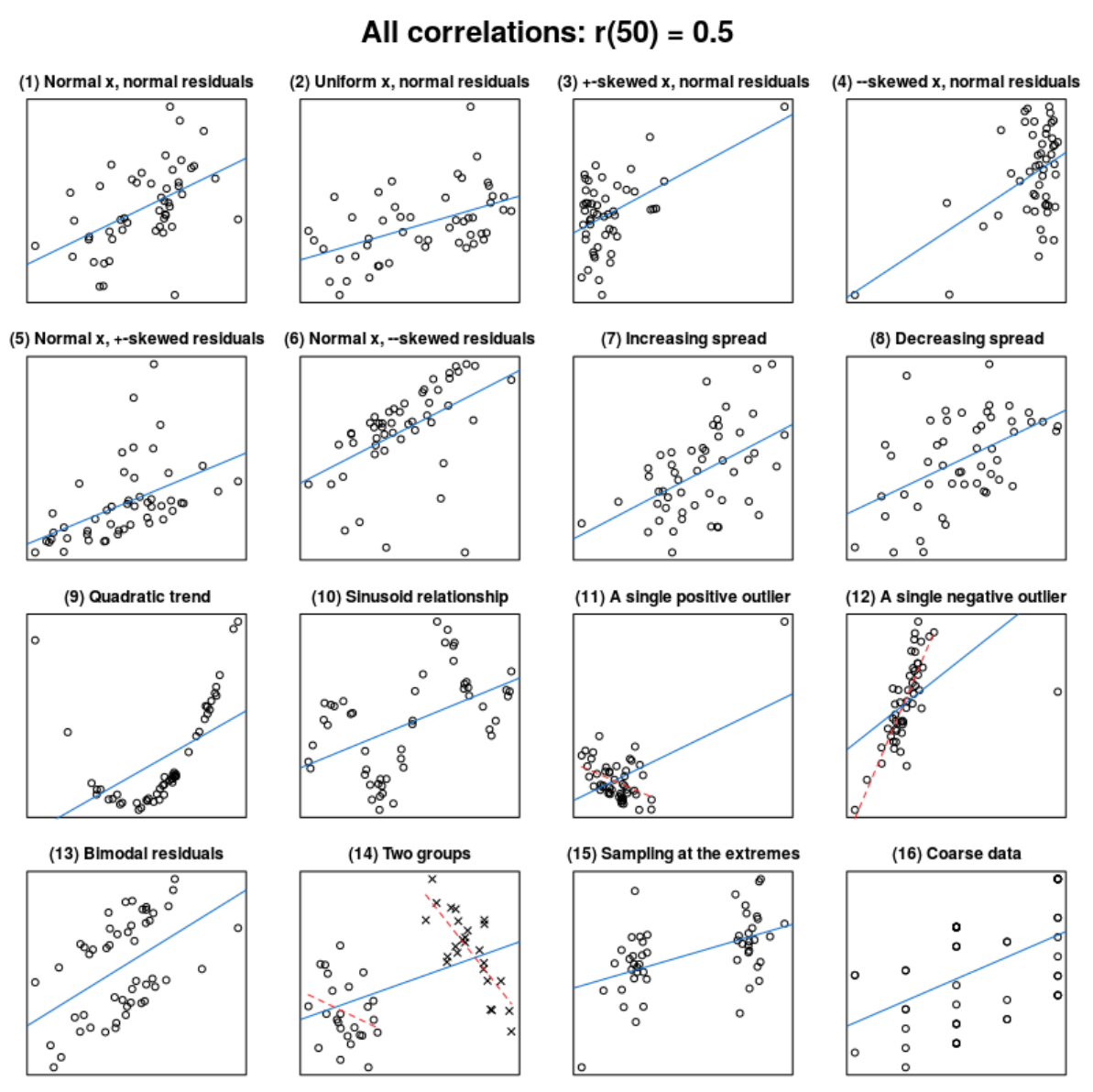
First things first

- Always
- Always
- Always

Plot data and models

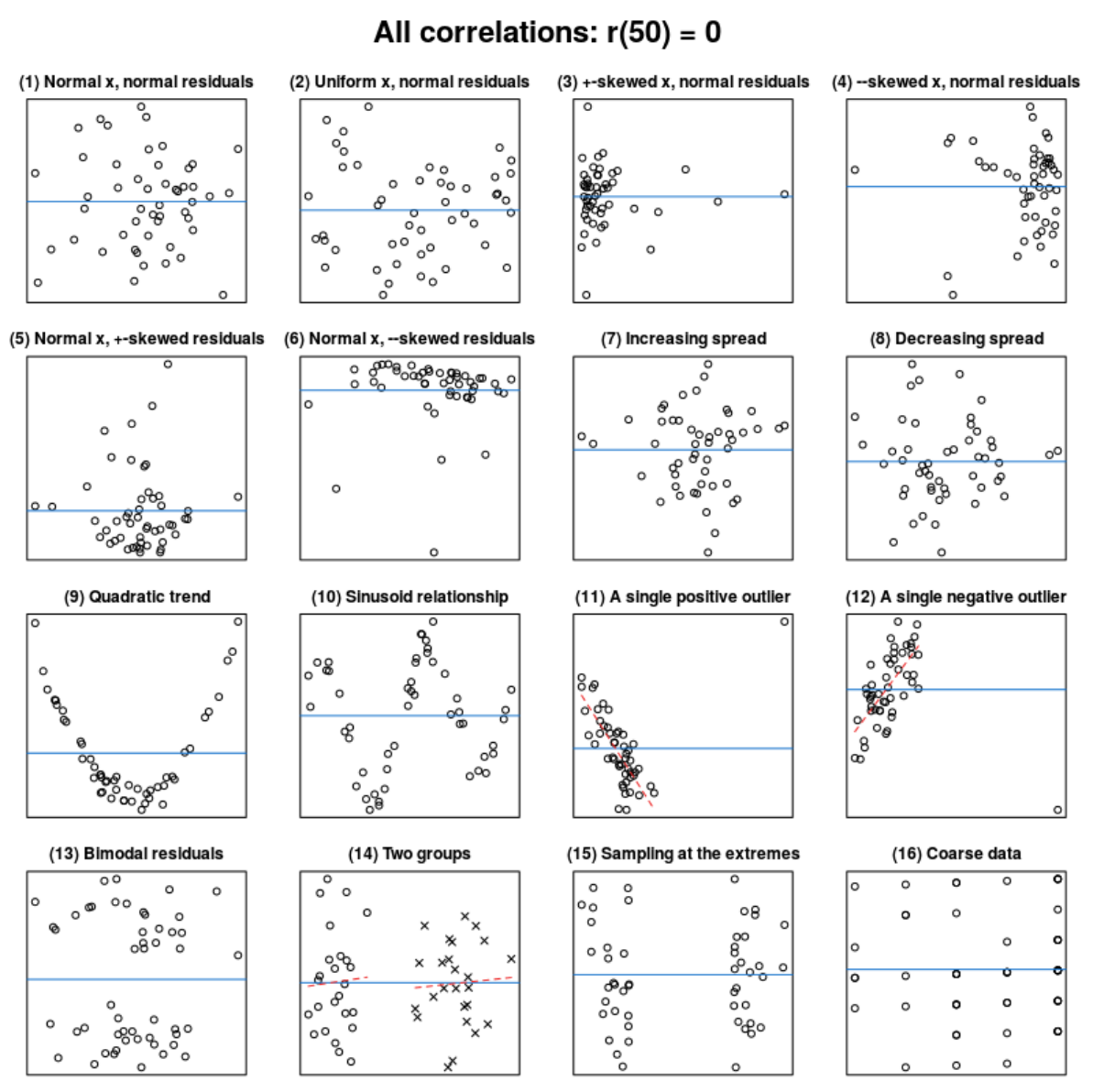


Don’t use statistics blindly: *Visualise*



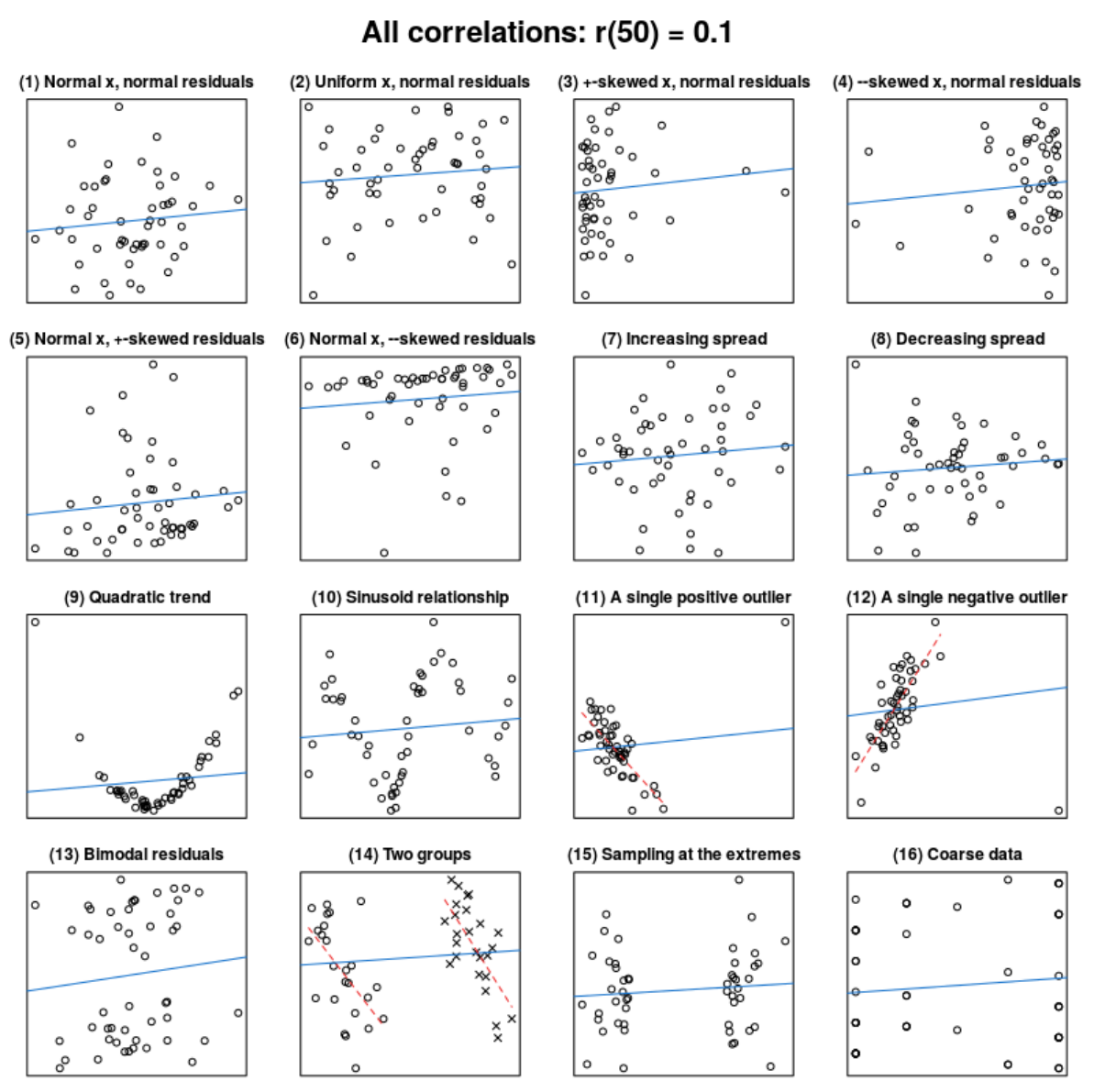
https://janhove.github.io/teaching/2016/11/21/what-

Don’t use statistics blindly: *Visualise*



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

Don’t use statistics blindly: *Visualise*



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

Plot. Check models. Plot. Check assumptions. Plot.

Lavine 2014 *Ecology*

Inference from observational studies

News: Hamburgers increase risk of heart attack

- In a sample of 10,000 people, it was found that people eating >2 hamburgers a week had 20% higher probability of heart attack.

News: Hamburgers increase risk of heart attack

- In a sample of 10,000 people, it was found that people eating >2 hamburgers a week had 20% higher probability of heart attack.
- **Do hamburgers increase heart attacks?**

News: Hamburgers increase risk of heart attack

- In a sample of 10,000 people, it was found that people eating >2 hamburgers a week had 20% higher probability of heart attack.
- **Do hamburgers increase heart attacks?**
- https://pollev.com/franciscorod726

Bigger flowers increase reproductive success

- We found that plants with big flowers produced 30% more seeds...

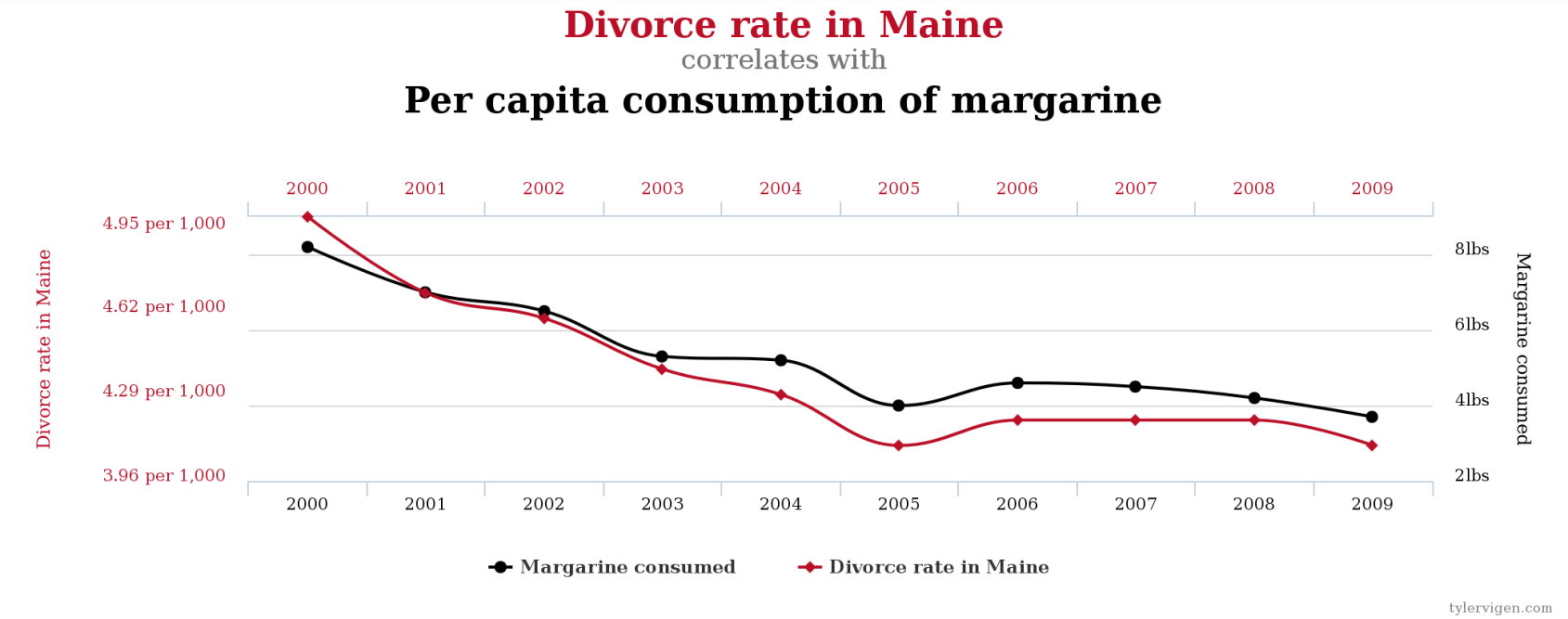
Bigger flowers increase reproductive success

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

Bigger flowers increase reproductive success

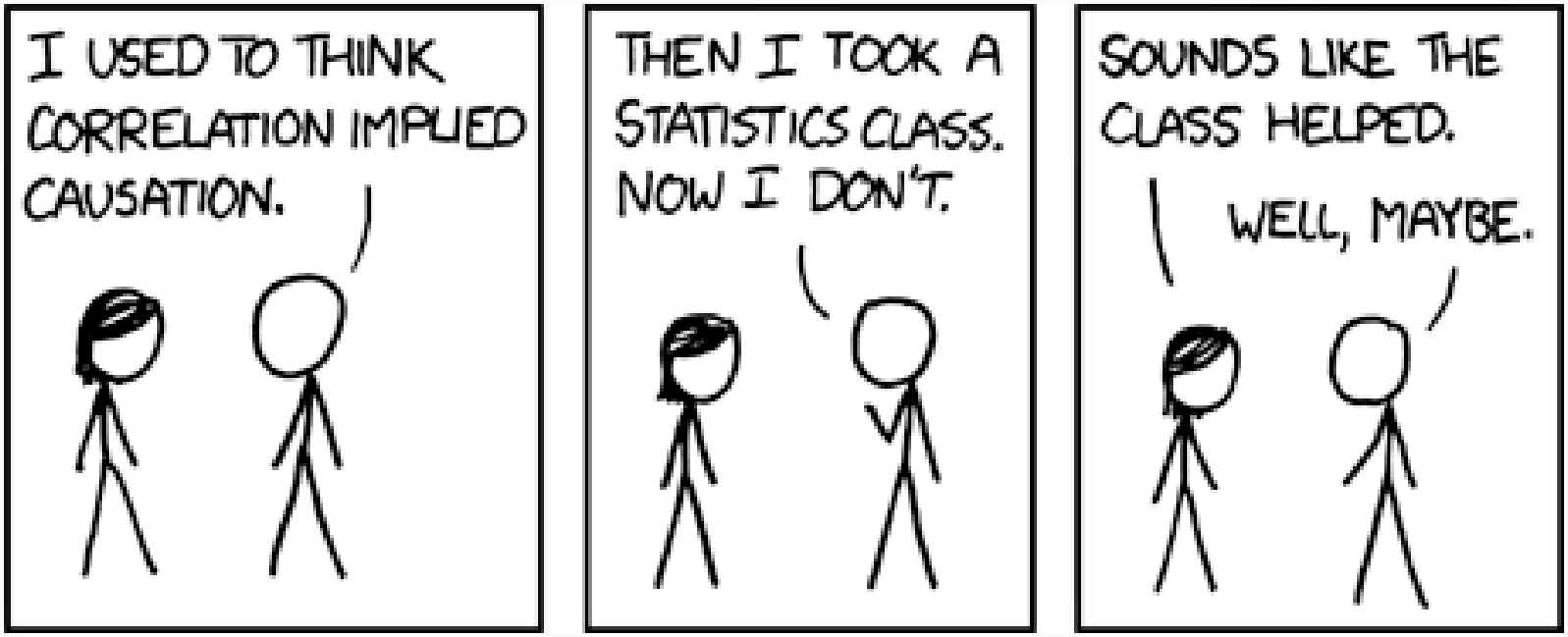
- We found that plants with big flowers produced 30% more seeds...
- **Do big flowers increase reproductive success?**
- https://pollev.com/franciscorod726

Correlation vs Causation



http://tylervigen.com/spurious-correlations

Learning statistics through xkcd



NHST and p-values

In ecology, everything is somewhat different

Are there any differences? A non-sensical question in ecology

Alejandro Martínez-Abraín

IMEDEA (CSIC-UIB), C/Miquel Marquès 21, 07190 Esporles, Majorca, Spain

ARTICLE INFO

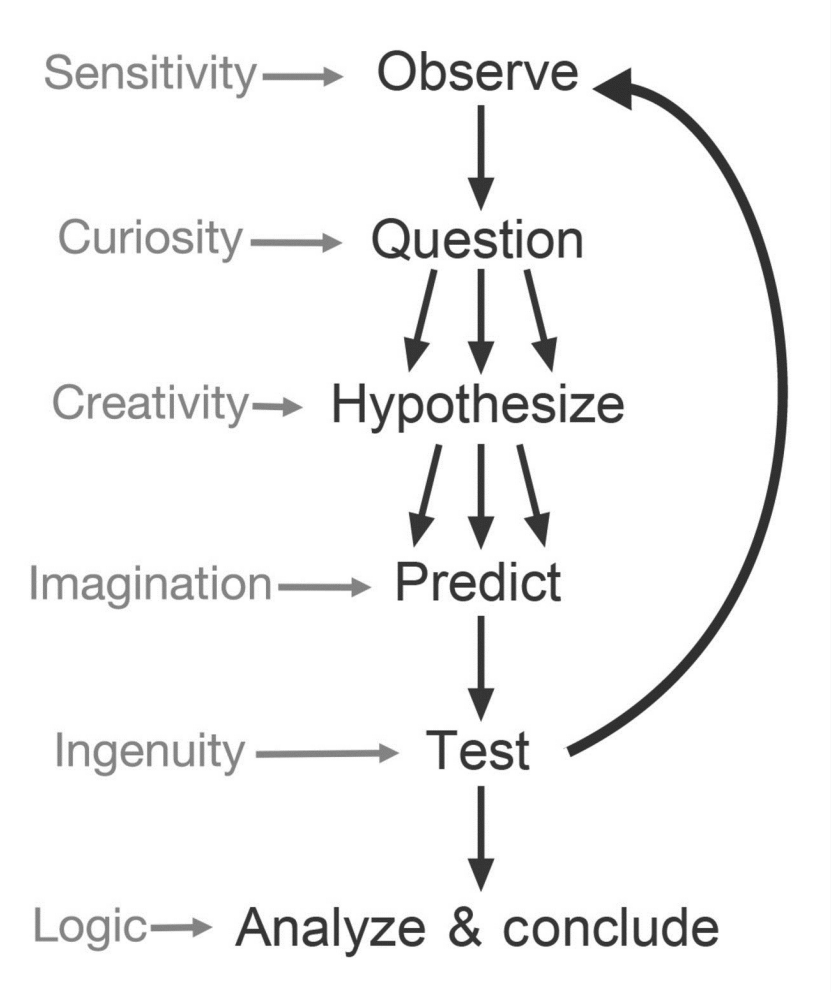
Article history:
Received 19 December 2006
Accepted 27 April 2007
Published online 13 June 2007

Keywords:

ABSTRACT

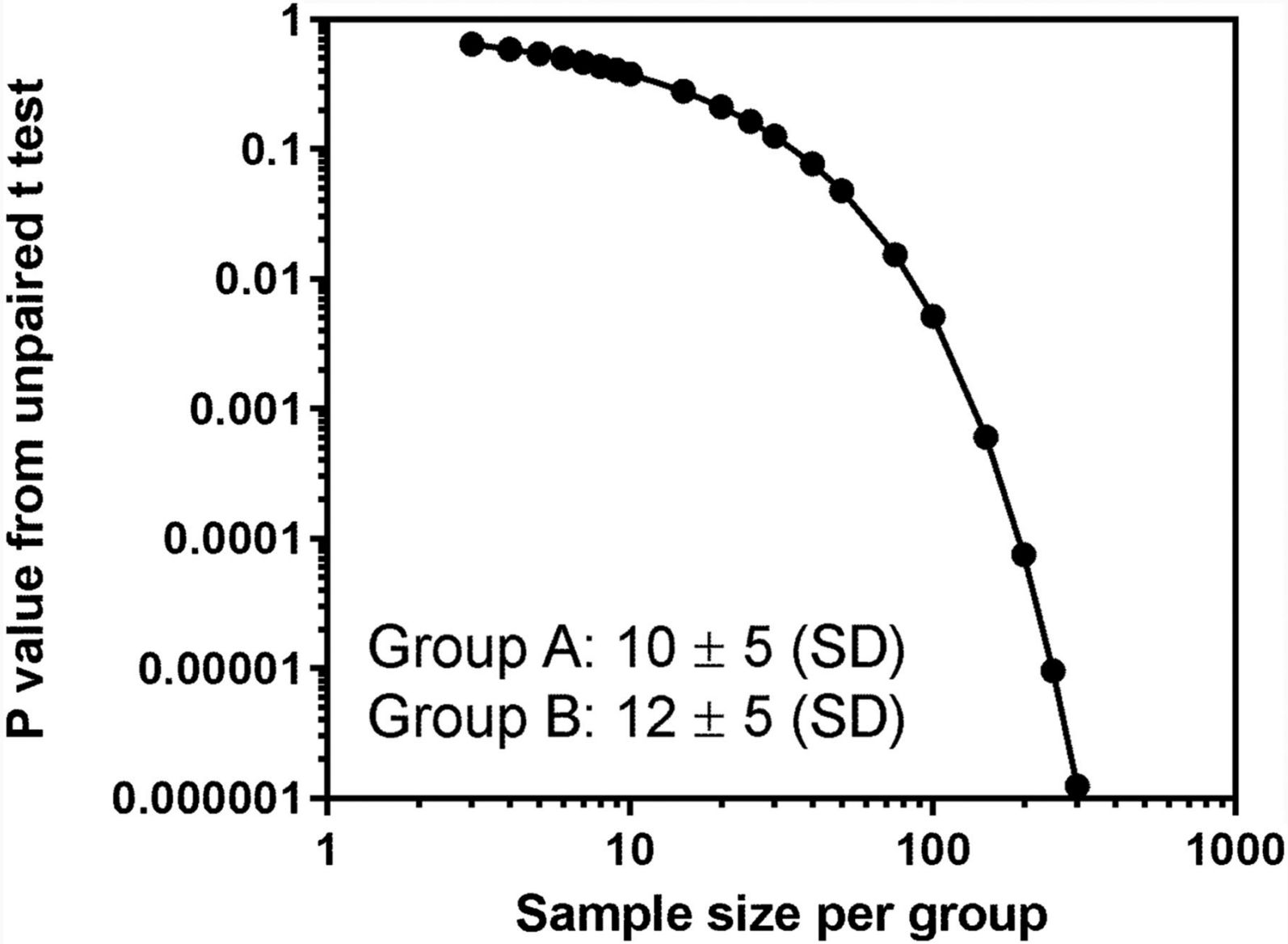
One of the main questions that ecologists pose in their investigations includes the analysis of differences in some trait between two or more populations. I argue here that asking whether there are differences or not between populations is biologically irrelevant, since no two livings things are ever equal. On the contrary the appropriate question to pose is how large differences are between populations. That is, we urge a shift in interest from statistical significance to biological relevance for proper knowledge accumulation. I empha-

Instead of falsifying null model, compare meaningful models



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

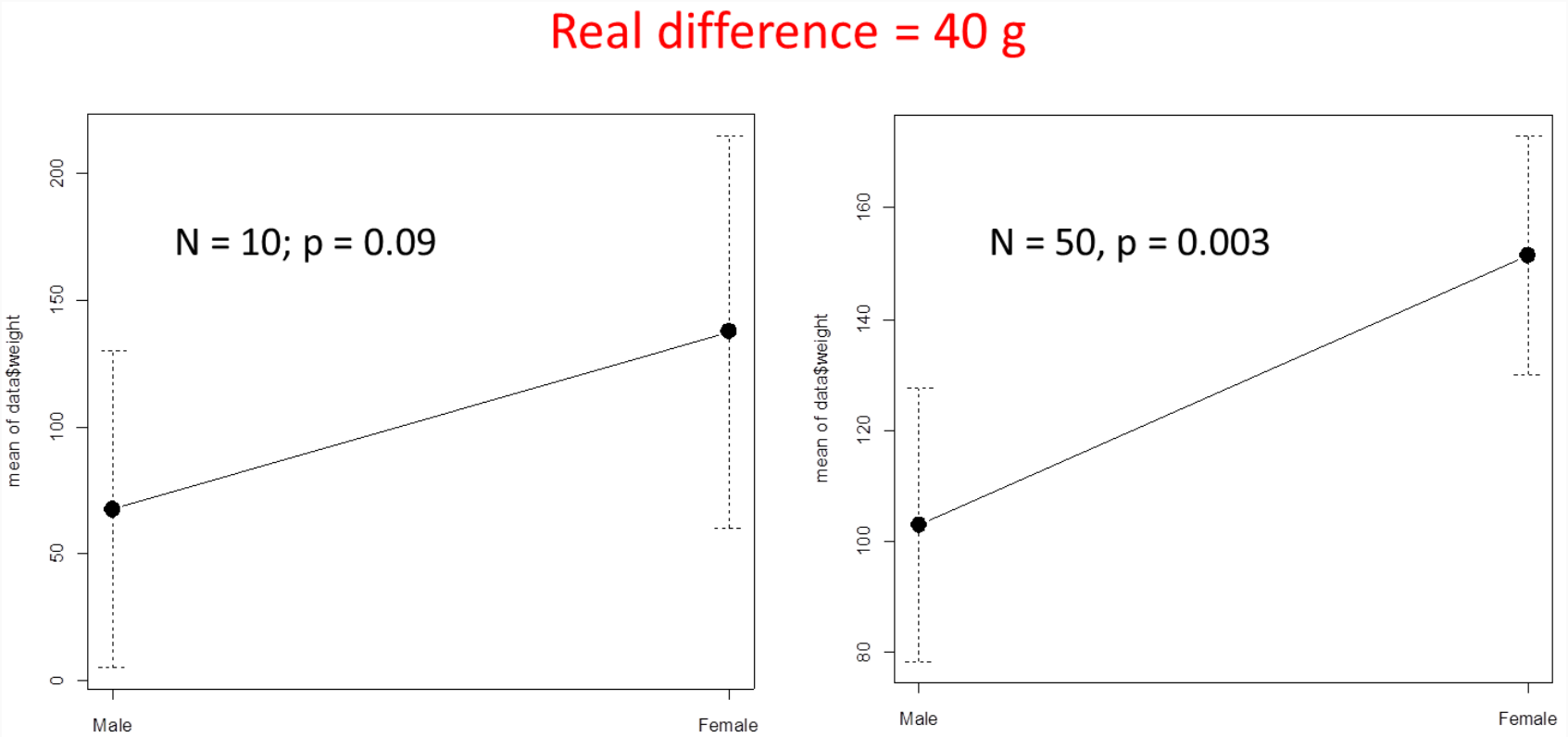
P-value depends on sample size



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

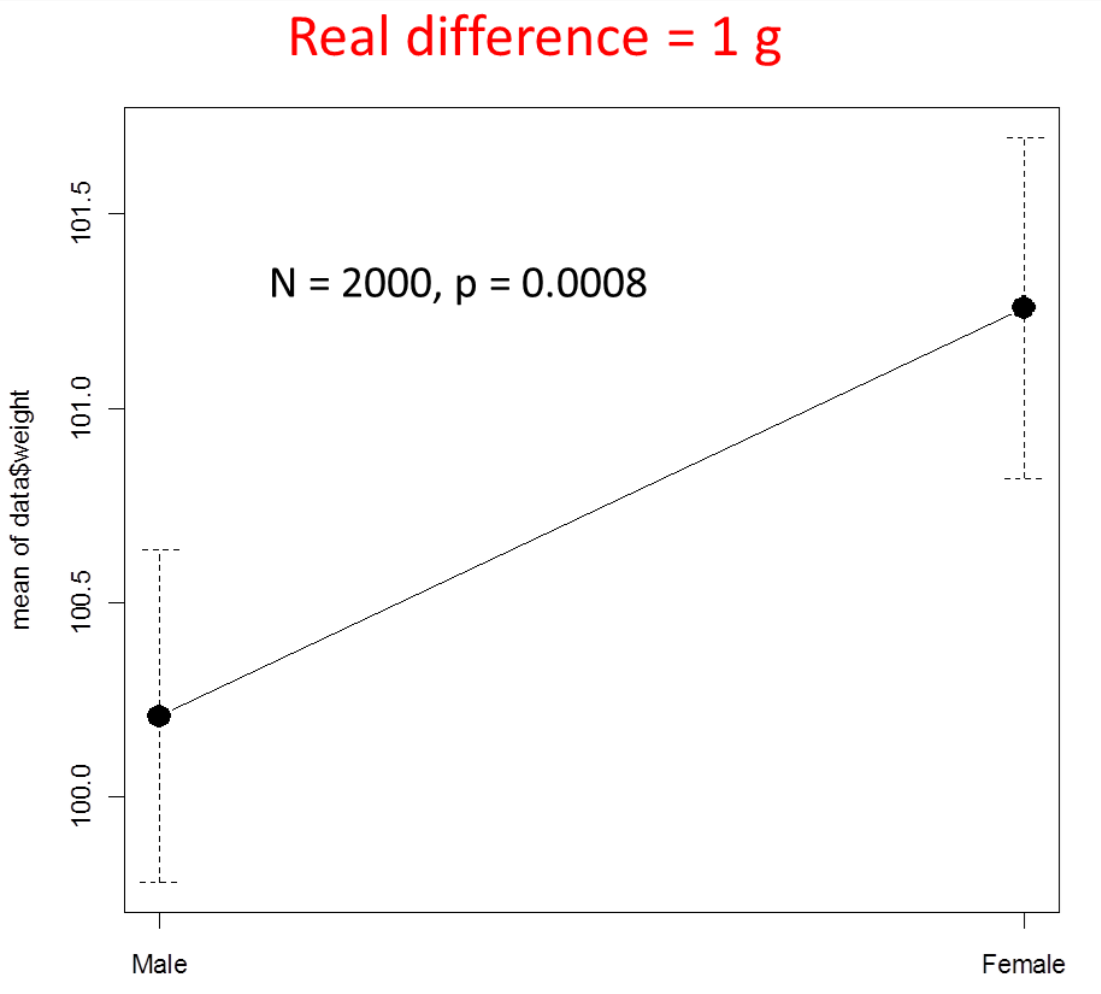
P-value depends on sample size

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



Statistically significant != biologically important

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



Statistically significant != biologically important

- Statistically significant = unlikely to be zero

Statistically significant != biologically important

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

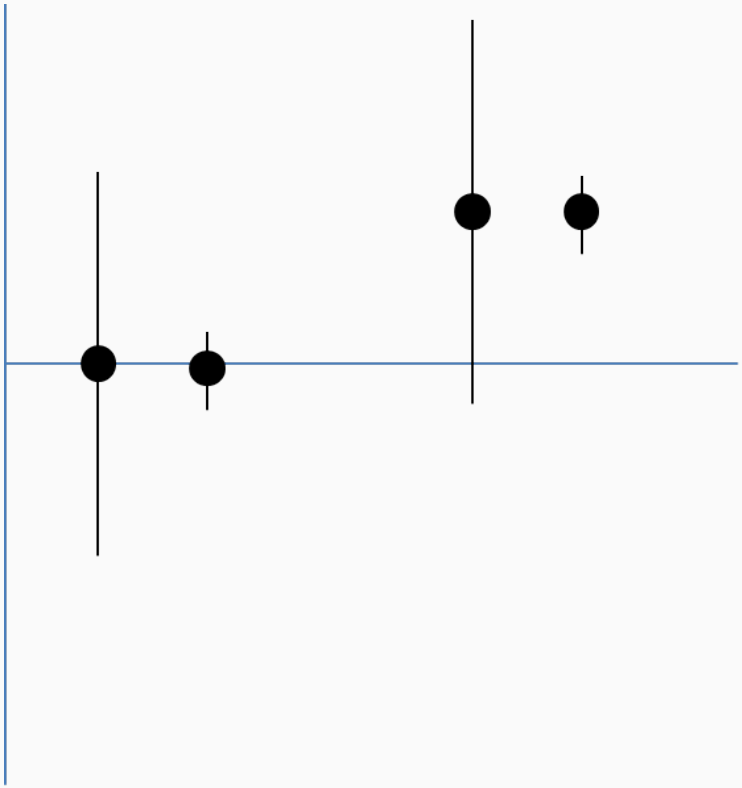
Statistically significant != biologically important

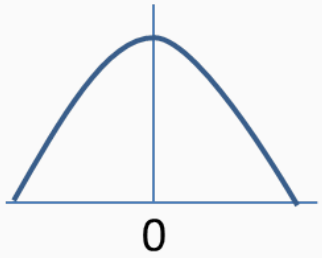
- Statistically significant = unlikely to be zero
- Good read: *significantly misleading*
- My suggestion: avoid significant/not significant (and maybe p-values too)

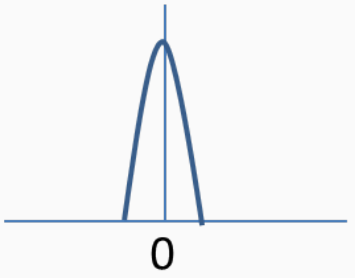
Statistically significant != biologically important

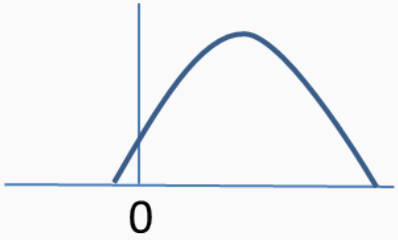
- Statistically significant = unlikely to be zero
- Good read: *significantly misleading*
- My suggestion: avoid significant/not significant (and maybe p-values too)
- Beyond significance, look at *effect sizes*.

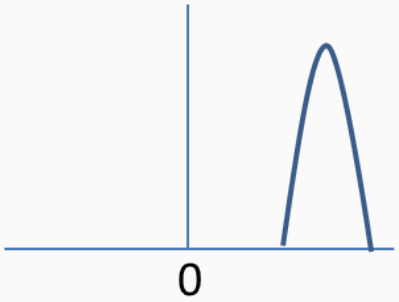
‘Not significant’ does NOT mean ‘there is no effect’





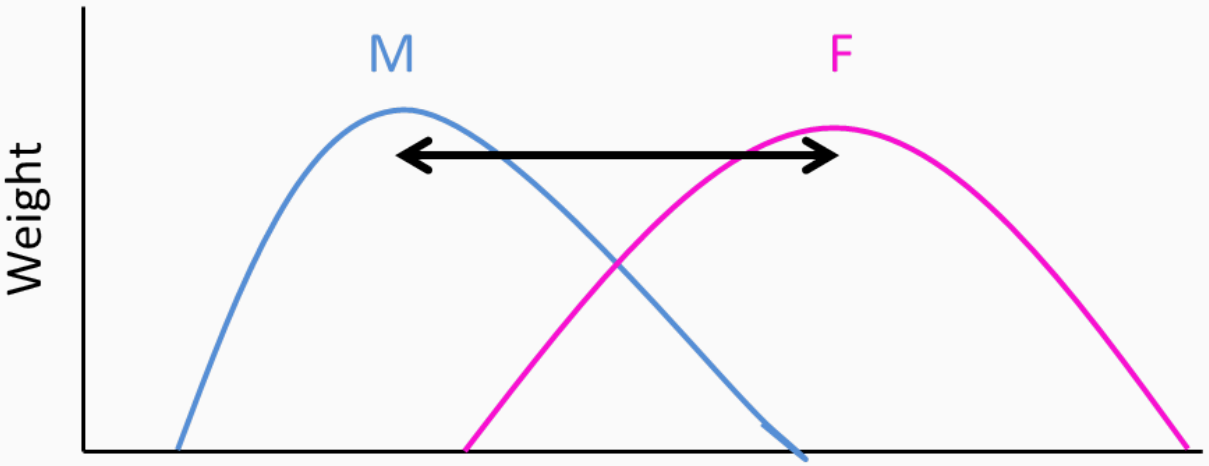




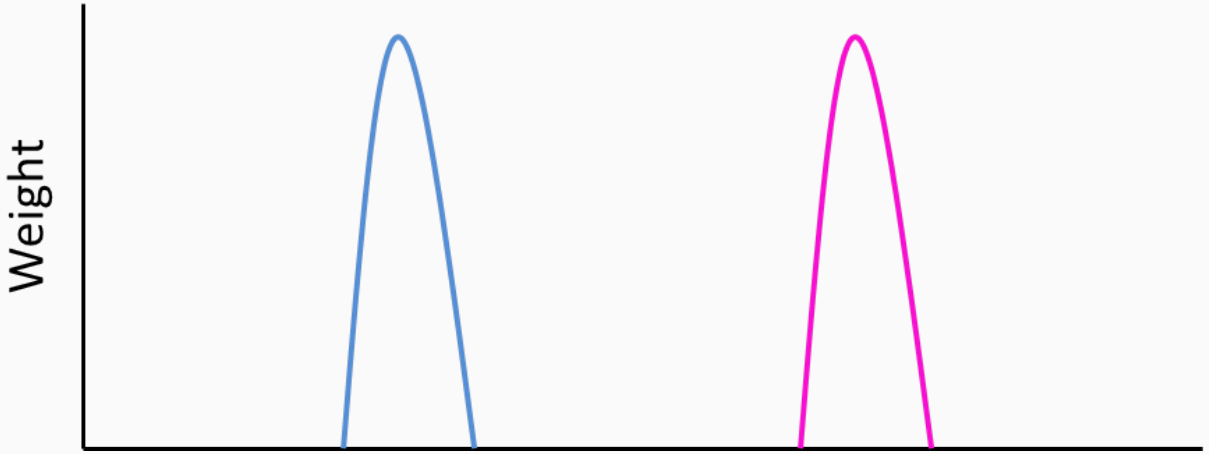


- Absence of evidence != Evidence of absence

Failure to reject H0 != H0 is true



P >> 0.05



P << 0.05

p-value > 0.05?

- “We were **unable to find evidence** against the hypothesis that A = B **with the current sample size**” (Harrell)

p-value > 0.05?

- "We were **unable to find evidence** against the hypothesis that A = B **with the current sample size**" (Harrell)
- "Differences between groups were **not statistically clear**" (Dushoff et al)

Is it safe to allow right turn with red lights?

- Right turn not allowed: 308 accidents

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

Is it safe to allow right turn with red lights?

- Right turn not allowed: 308 accidents
- Right turn allowed: 337 accidents

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

Is it safe to allow right turn with red lights?

- Right turn not allowed: 308 accidents
- Right turn allowed: 337 accidents
- No *significant* difference, hence safe

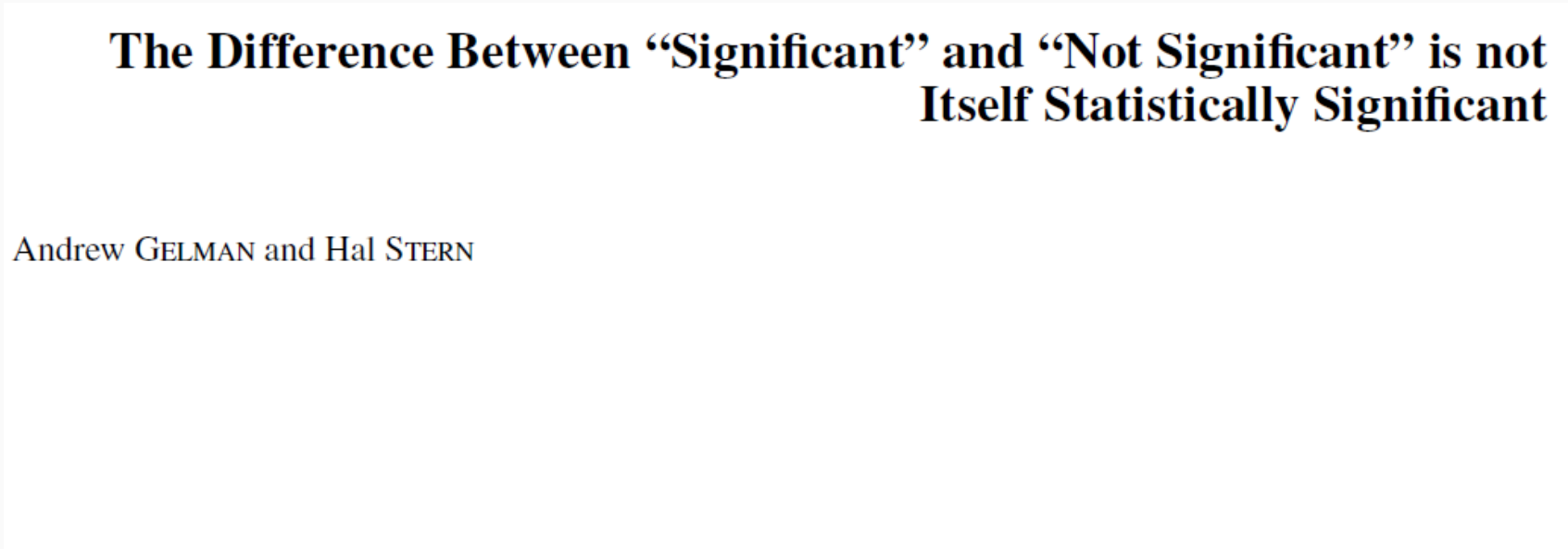
https://www.statisticsdonewrong.com/power.html#the-wrong-turn-on-red

Is it safe to allow right turn with red lights?

- Right turn not allowed: 308 accidents
- Right turn allowed: 337 accidents
- No *significant* difference, hence safe
- Misinterpretation of underpowered study cost lives

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

0.05 is an arbitrary threshold

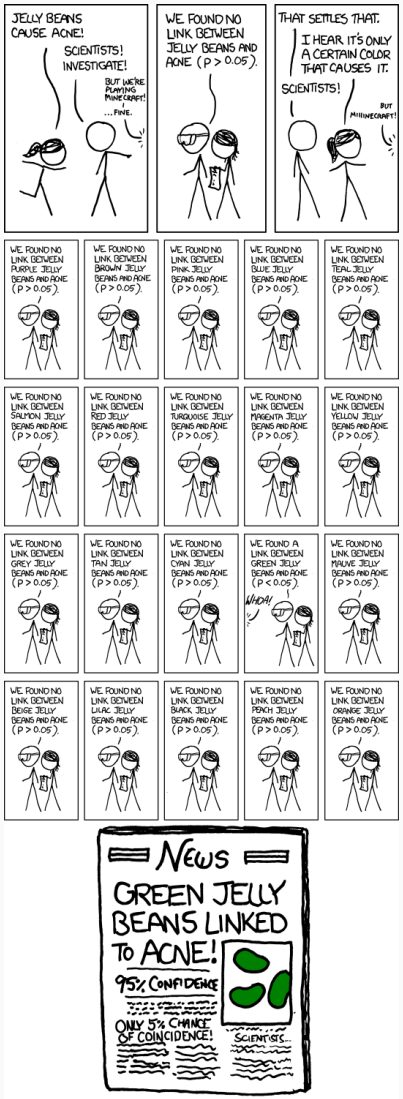


The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant

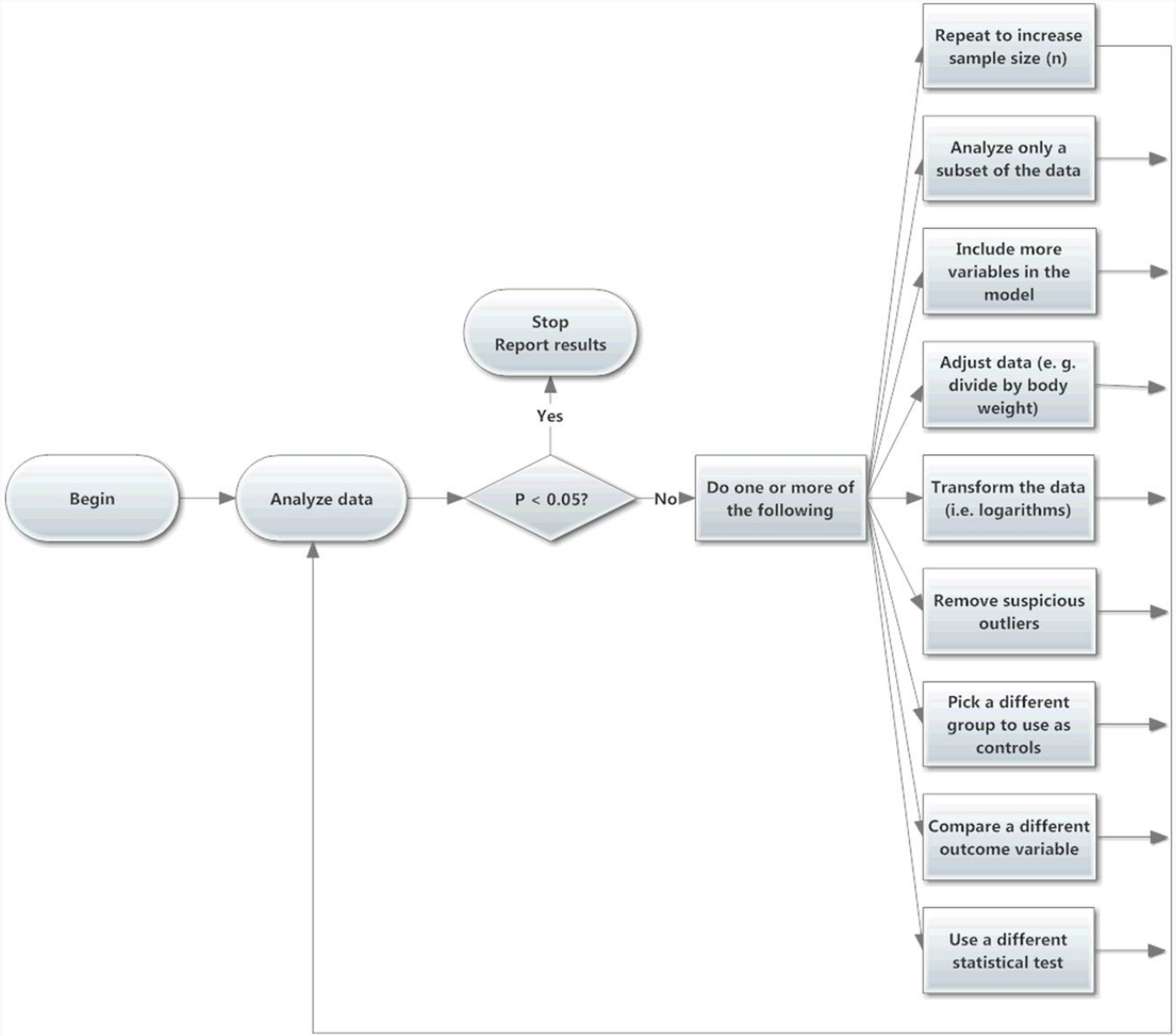
Andrew GELMAN and Hal STERN

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

Multiple hypothesis testing



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

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

https://www.youtube.com/watch?v=ZaNtz76dNSI

ASA statement on p-values

- P-values do not measure the **probability of hypothesis** being true, or the probability that the data were produced by **random chance** alone.

https://doi.org/10.1080/00031305.2016.1154108

ASA statement on p-values

- P-values do not measure the **probability of hypothesis** being true, or the probability that the data were produced by **random chance** alone.
- Scientific conclusions or policy decisions should NOT be based only on **whether a p-value passes a specific threshold**.

https://doi.org/10.1080/00031305.2016.1154108

ASA statement on p-values

- P-values do not measure the **probability of hypothesis** being true, or the probability that the data were produced by **random chance** alone.
- Scientific conclusions or policy decisions should NOT be based only on **whether a p-value passes a specific threshold**.
- P-value, or statistical significance, does not measure the **size of an effect** or the **importance** of a result.

https://doi.org/10.1080/00031305.2016.1154108

ASA statement on p-values

- P-values do not measure the **probability of hypothesis** being true, or the probability that the data were produced by **random chance** alone.
- Scientific conclusions or policy decisions should NOT be based only on **whether a p-value passes a specific threshold**.
- P-value, or statistical significance, does not measure the **size of an effect** or the **importance** of a result.
- By itself, a p-value does NOT provide a good **measure of evidence** regarding a model or hypothesis.

https://doi.org/10.1080/00031305.2016.1154108

The New Statistics

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

General Article

The New Statistics: Why and How

Geoff Cumming
La Trobe University

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

DOI: 10.1177/0956797613504966
pss.sagepub.com

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.

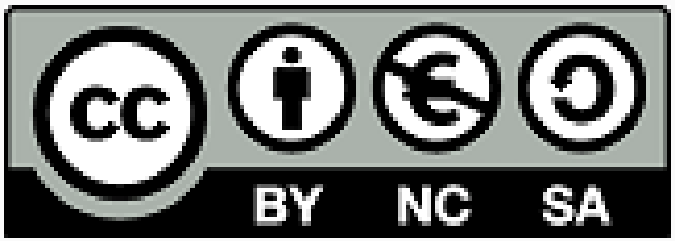
How many types of errors?

- **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

END



Source code and materials: https://github.com/Pakillo/stats-intro