Frequentist Inference

“The statistician cannot excuse himself from the duty of getting his head clear on the principles of scientific inference, but equally no other thinking man can avoid a like obligation.”
— Ronald Fisher

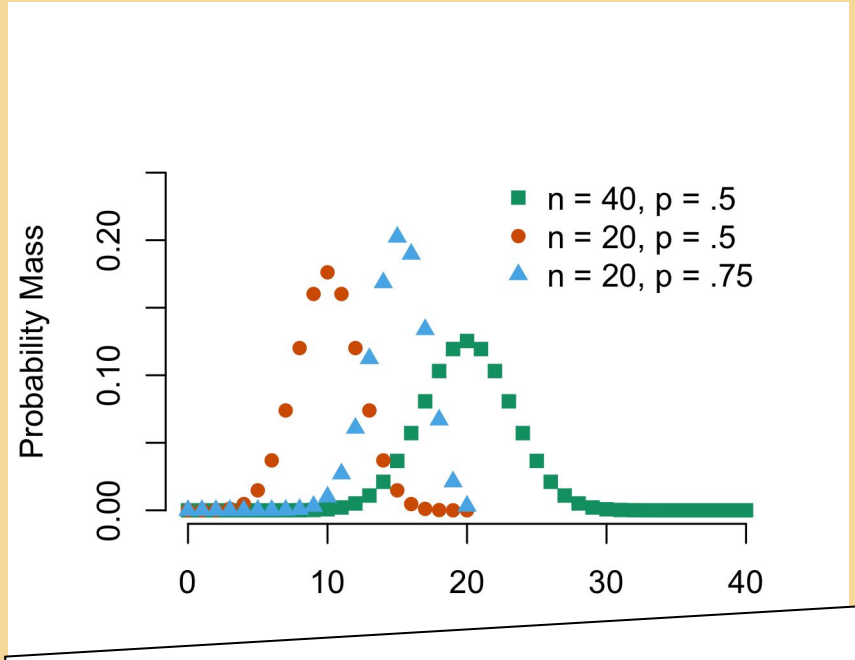
👨‍🏫 Statistical Reasoning Lecture #2
Alexander Savi, 2024
📚 Whitlock/Schluter, Ch. 3; Dienes, Ch. 3 from p. 60

Watercolors by Antonio Sánchez Chinchón 🎨

📢 Announcements

- Questions
  - Canvas discussion forum
  - Email only for private communication
    - Jonas for organization
    - Alexander for substance
- Questions about weekly assignment
  - How do you interpret the question?
  - What is the exact issue you're struggling with?
  - What have you tried to solve the issue?
- Personal course manual

🔁 Recap



🤔 Which distribution? Examples?

🔮 Overview

Topics

Probabilities & distributions
🔁 Frequentist inference
- Central limit theorem
- Null hypothesis significance testing
- Confidence intervals
- Student's *t*-test
- Correlation
- Simple linear regression

Multiple linear regression
Factorial ANOVA
Nonparametric inference
Bayesian inference

Learning goals

How can we use probability distributions to test a null hypothesis?

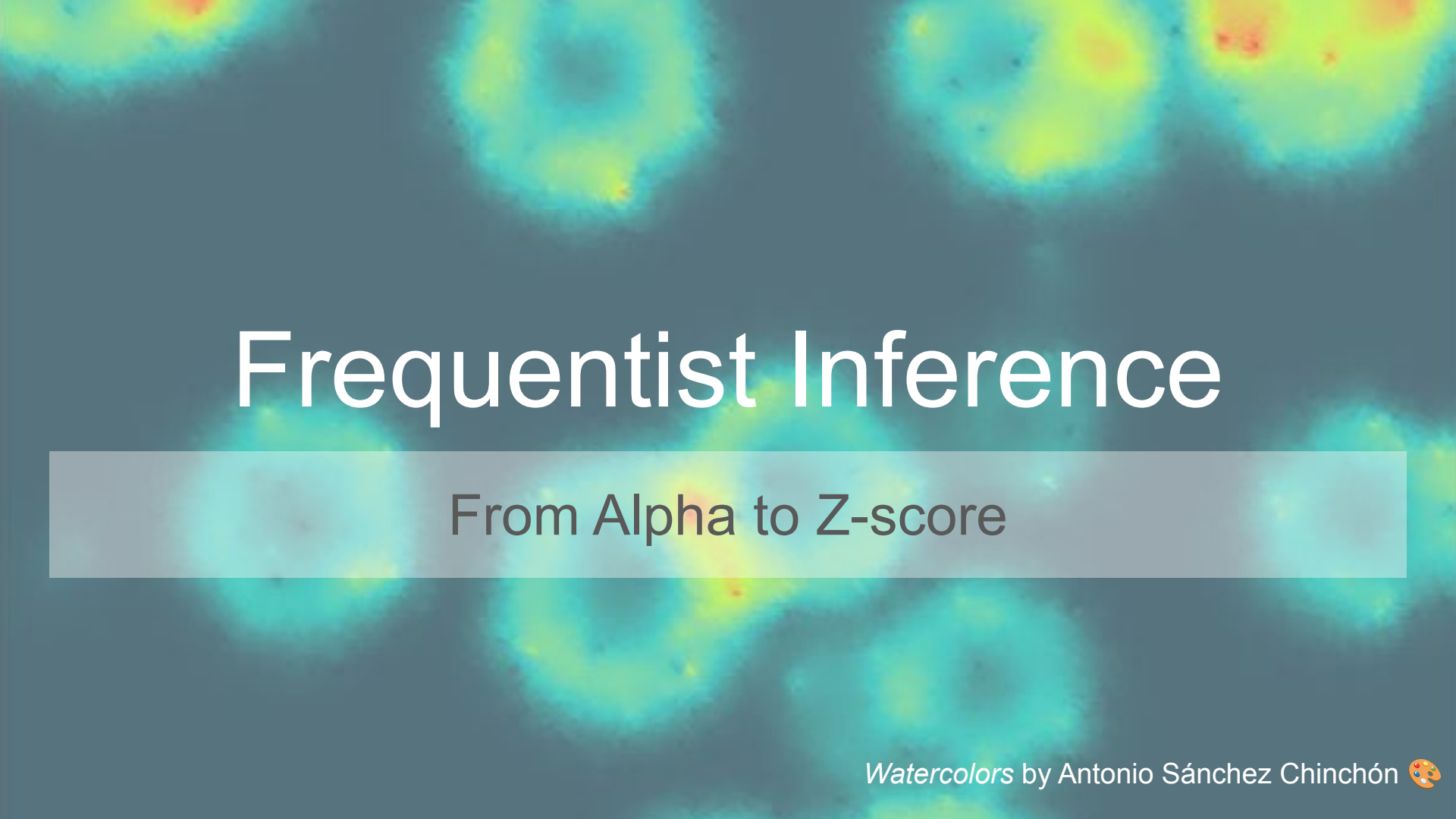
How can we determine how likely our observed data is, given a null hypothesis?

How can we make a decision about our null hypothesis? What affects this decision?

How can we quantify the uncertainty in our test result?

Frequentist Inference

From Alpha to Z-score



Watercolors by Antonio Sánchez Chinchón 🎨

Student perseverance

Photo by Jorn van Eck (UvA)

Q. Are UvA psychobiology students more perseverant than the average psychobiology student?

H. UvA psychobiology students are more perseverant than the average psychobiology student.

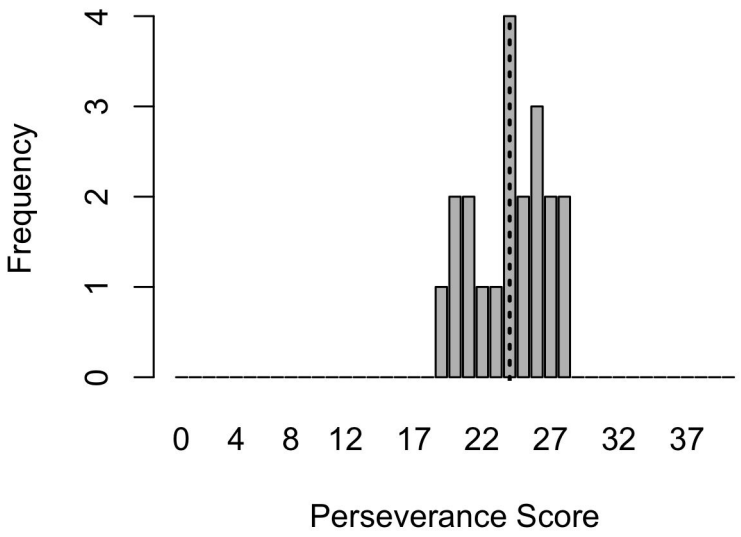
E. UvA psychobiology students score higher on the Student Perseverance Score Scale (SPSS) than the average psychobiology student.

SPSS. 40 two-choice items, ‘0’ = (very) low perseverance, ‘40’ = (very) high perseverance.

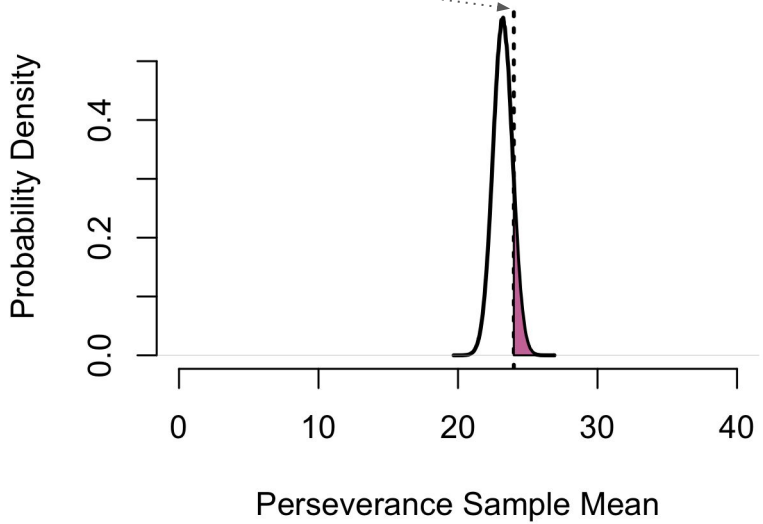
$\mu = 23.2$

Sample/sampling distribution

Sample distribution *(n = 20; $\bar{x}$ = 24)*



Sampling distribution *(n_s = 1.000.000; n = 20)*



Central limit theorem

Sample means are approximately normally distributed (if the sample size is large enough, i.e., n = 30). Even if the population is not normally distributed itself.

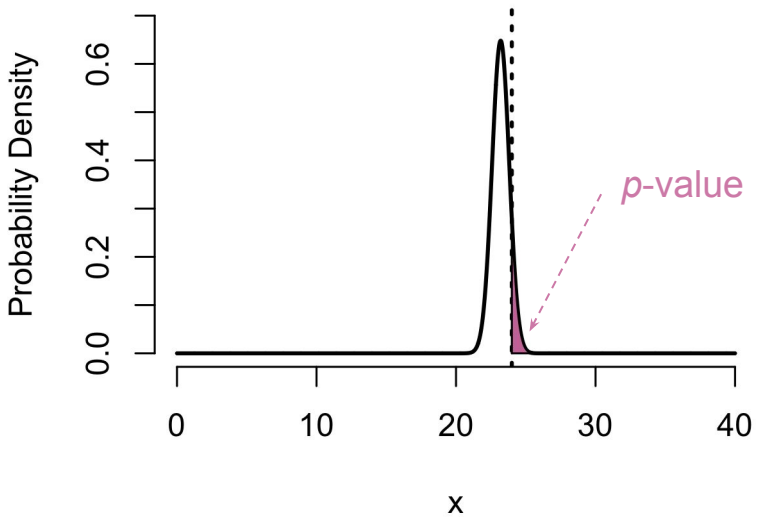
$\mu = 23.2$

Sample standard deviations are approximately normally distributed (if the sample is large enough).

$\sigma \rightarrow SE = s / \sqrt{(n)}$

```
1 - pnorm(q = 24, mean = 23.2, sd = .6)
```

Null distribution *($\mu = 23.2$; $\sigma = 0.6$)*



P-values

> The probability of the observed data (or of more extreme data points), given that the null hypothesis is true: p(D|H0).
> — Gigerenzer, 2004 🔒

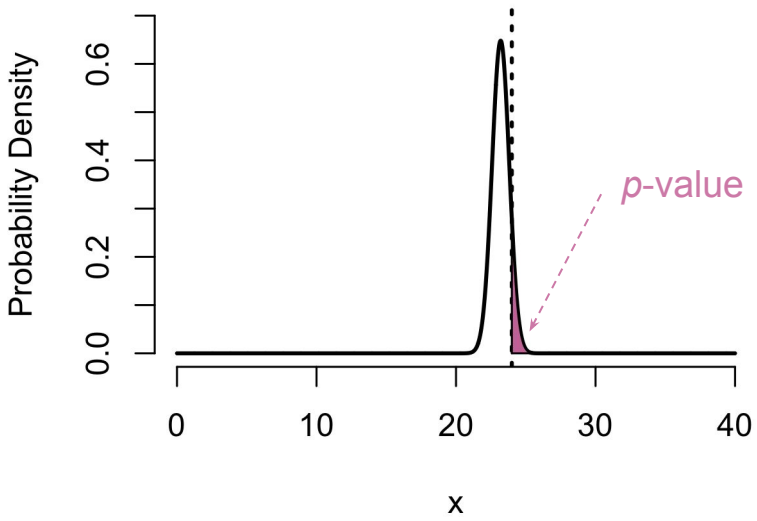
🤔 Is a *p*-value of .049 statistically significant?

🤔 How should you pick the alpha level?

🤔 How should you pick between a one-sided and two-sided test?

💡 Why .05?

Null distribution *(μ = 23.2; σ = 0.6)*



P-values in motion 🌊

Illustration by Randall Munroe (wtf)

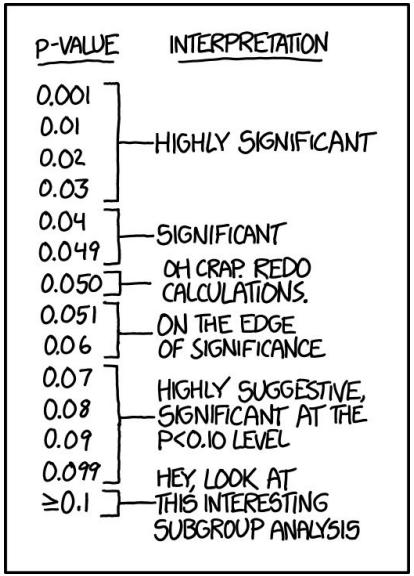


Illustration by Viktor Beekman

🔥 Scientific discussion

- Should we redefine or abandon statistical significance?
- How reliable are *p*-values?
- Mindless (Frequentist) statistics.

💡 Misinterpretations

- 25 misinterpretations of *p*-values, confidence intervals, and power.

Effect size

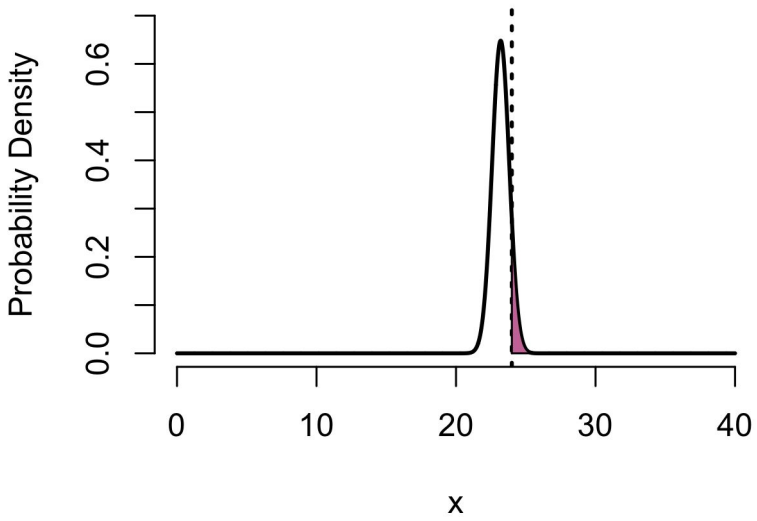
> In statistics, an effect size is a value measuring the strength of the relationship between two variables in a population, or a sample-based estimate of that quantity. — Wikipedia

🤔 How to express the effect size in our example?

Cohen's *d* = (sample mean − μ) / sd

💡 To standardize or not to standardize 🔒?

Null distribution *(μ = 23.2; σ = 0.6)*



Effect size in motion 🌊

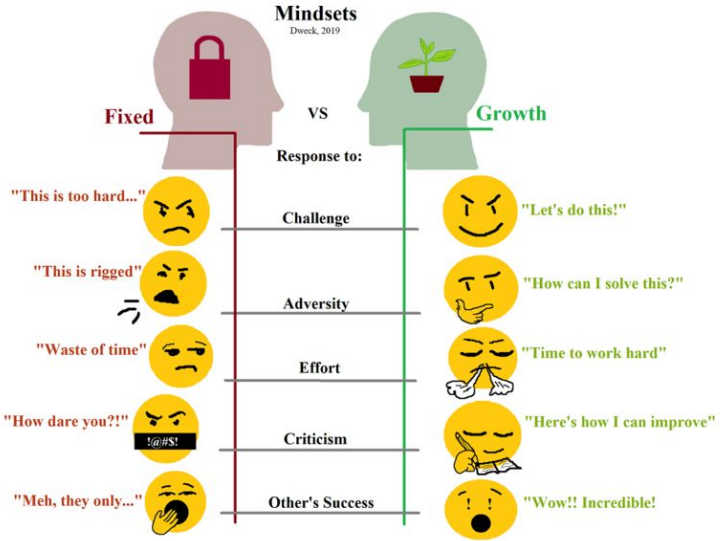


Illustration by Wikipedia

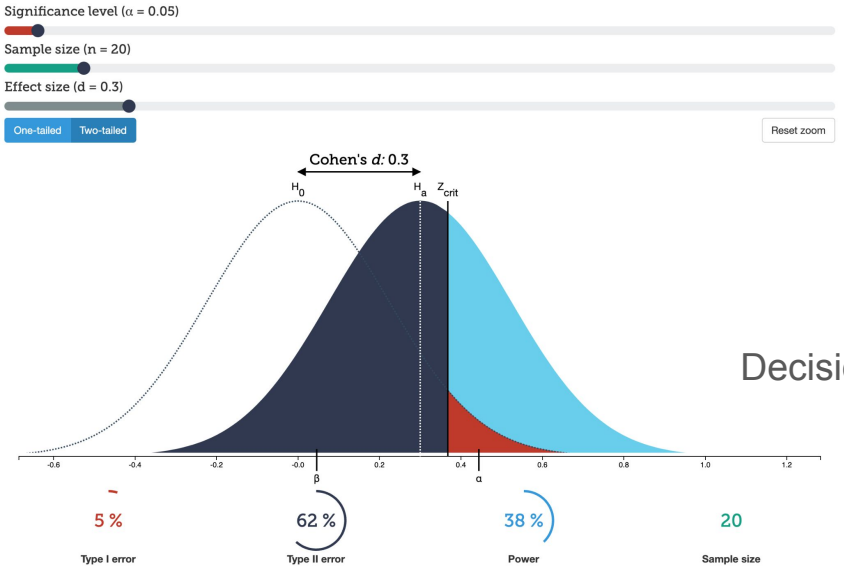
🔥 Scientific discussion

- Brooke Macnamara
- Carol Dweck (response)
- Andrew Gelman (response)

💡 Sense and nonsense

- Funder & Ozer, 2019

Error control



Interactive visualization by [Kristoffer Magnusson]

Decision	Null Hypothesis: True	Null Hypothesis: False
Don't Reject	… *(true negative)* $P(\neg Reject \mid H_0) = 1 - Alpha$	**Type II Error** *(false negative)* $P(\neg Reject \mid \neg H_0) = Beta\ (\beta)$
Reject	**Type I Error** *(false positive)* $P(Reject \mid H_0) =$ ***Alpha*** (α)	**Power** *(true positive)* $P(Reject \mid \neg H_0) = 1 - Beta$

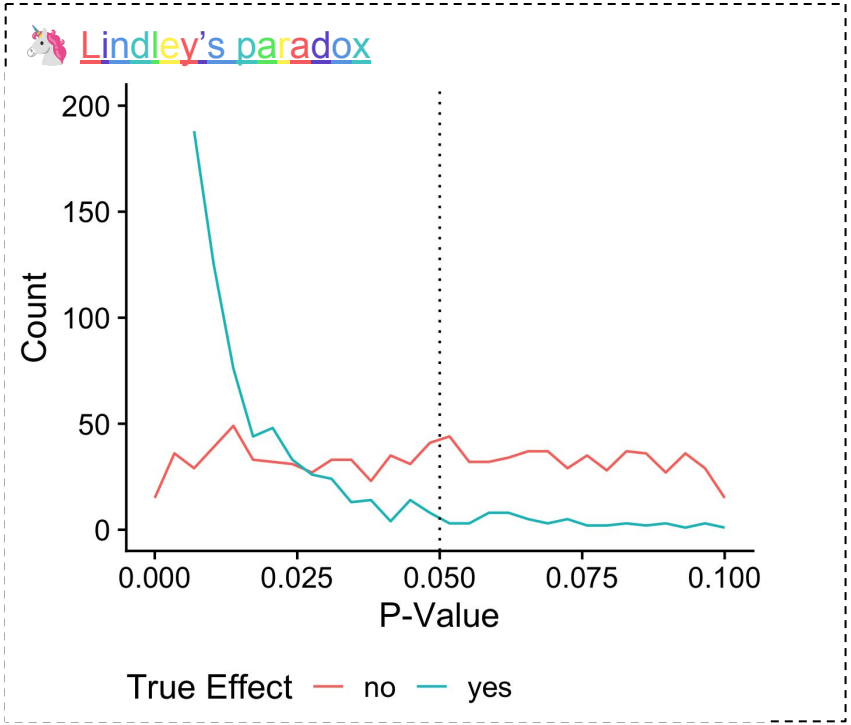
Error control

Type I error

- Multiple comparisons
- Optional stopping

Statistical power

- “80%”
- G*Power or Superpower (R)

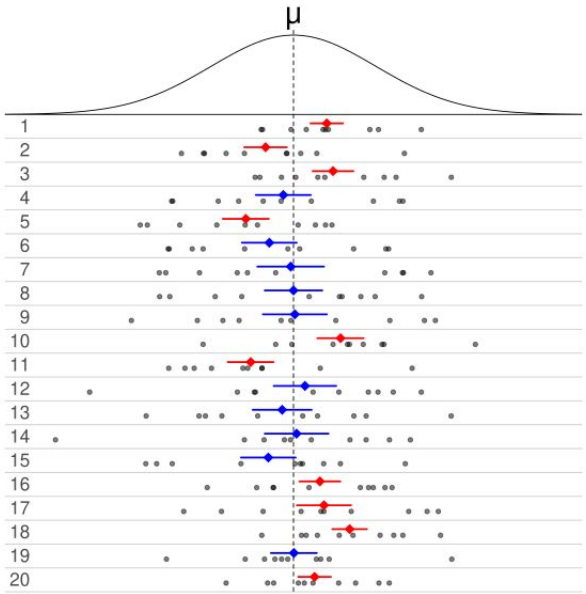


Error control in motion 🌊

💡

- Type II errors “don’t exist”, type I errors are “more important”
- Compute power before your experiment, not after (1, 2)
- Problem of type S/M errors in underpowered studies
- Type III/IV error
- File drawer problem

Confidence interval



Illustrations by Wikipedia (top, right)

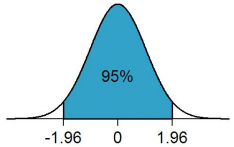
> “ If we would repeat the experiment, in ..% of the time, the true population mean will fall within the constructed interval.

sample mean ± SE × …

90% CI = 1.64

95% CI = 1.96

99% CI = 2.56




```
qnorm(p = .05 / 2)  # two-sided
```

Test statistics

$\bar{x}$

> “ A test statistic is a statistic (a quantity derived from the sample) used in statistical hypothesis testing. A hypothesis test is typically specified in terms of a test statistic, considered as a numerical summary of a data-set that reduces the data to one value that can be used to perform the hypothesis test. — Wikipedia

$\bar{x}$	Normal distribution
$Z = (\bar{x} - \mu) / \sigma$	Standard normal distribution
$t = (\bar{x} - \mu) / se$	Student’s *t* distribution
F	*F* distribution
H	χ^2 distribution

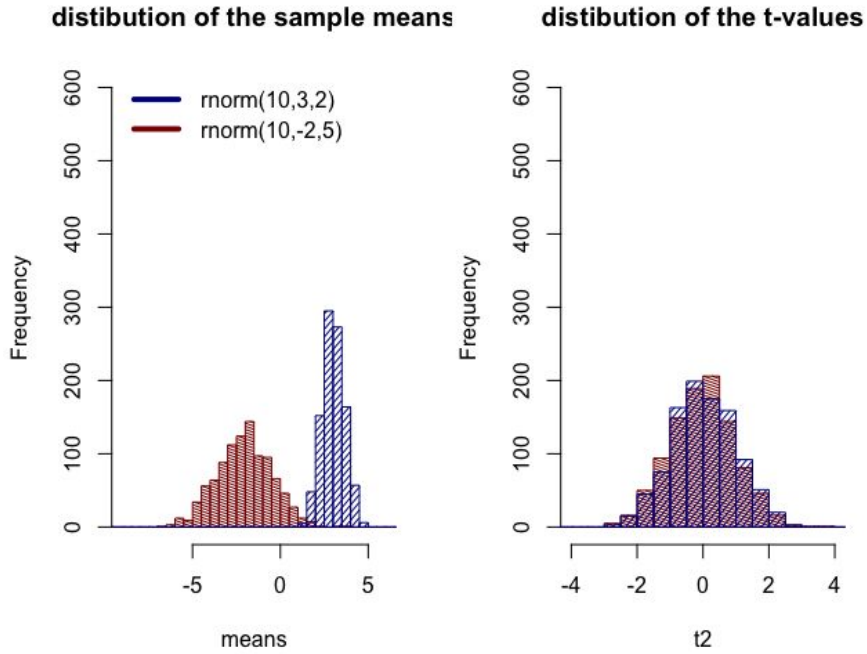
Student's *t*-statistic 🍻

$t = (\bar{x} - \mu) / se$

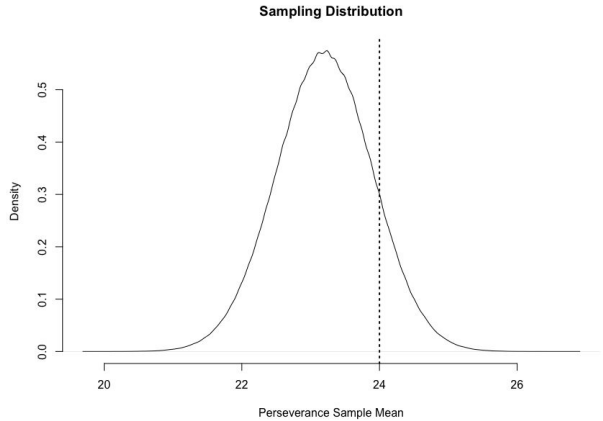
sample mean = 24
population mean = 23.2
sample sd = 2.75
n = 20

$t = (24 - 23.2) / (2.75 / \sqrt{(20)}) = 1.2996$

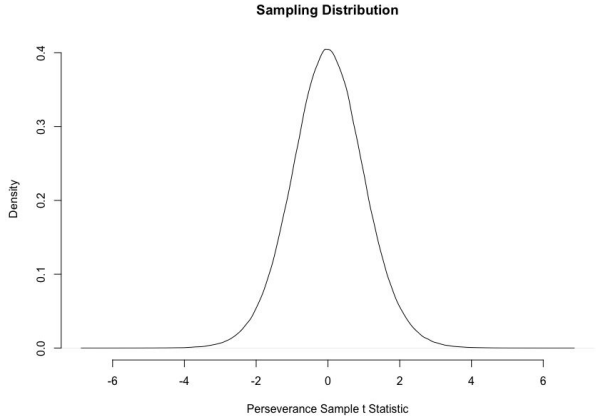
```
pt(1.2996, df = 20-1, lower.tail = FALSE)    # probability of t or higher

t.test(dat, mu = 23.2, alternative = "greater")    # one sample t-test
```



Student's *t*-distribution



Normal distribution
mu = population mean
sd = standard error of sample mean

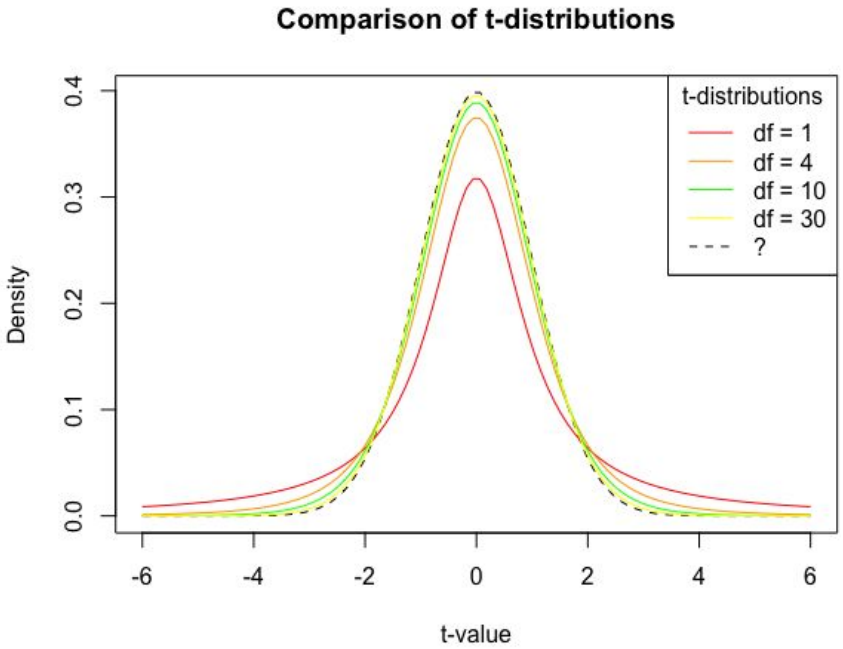


Student's *t*-distribution
df = n − 1

Student's *t*-distribution

🤔 What are the degrees of freedom at the dashed line?

🤔 If there is a difference in population means, is it easier to find a significant effect with a larger sample size?

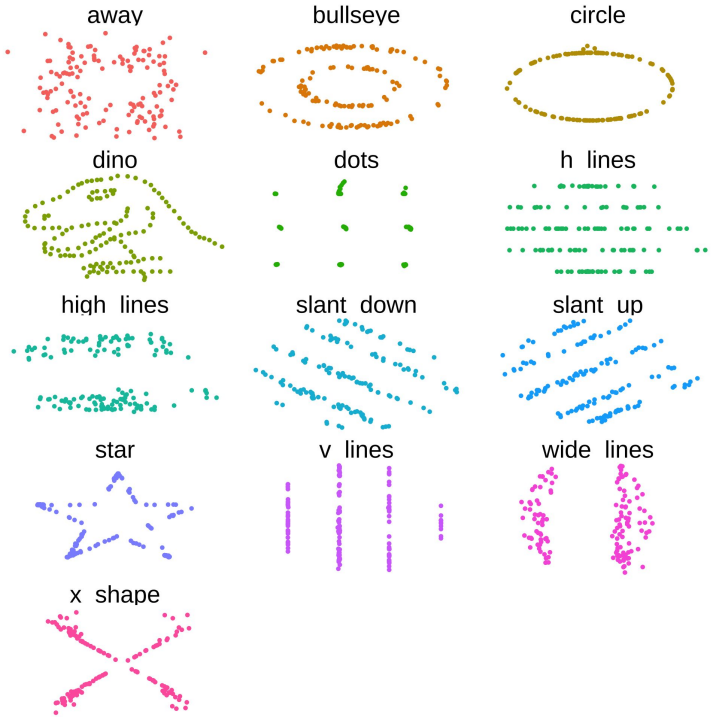


Test statistics in motion 🌊

🦄 Datasaurus dozen

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

Table by Wikipedia

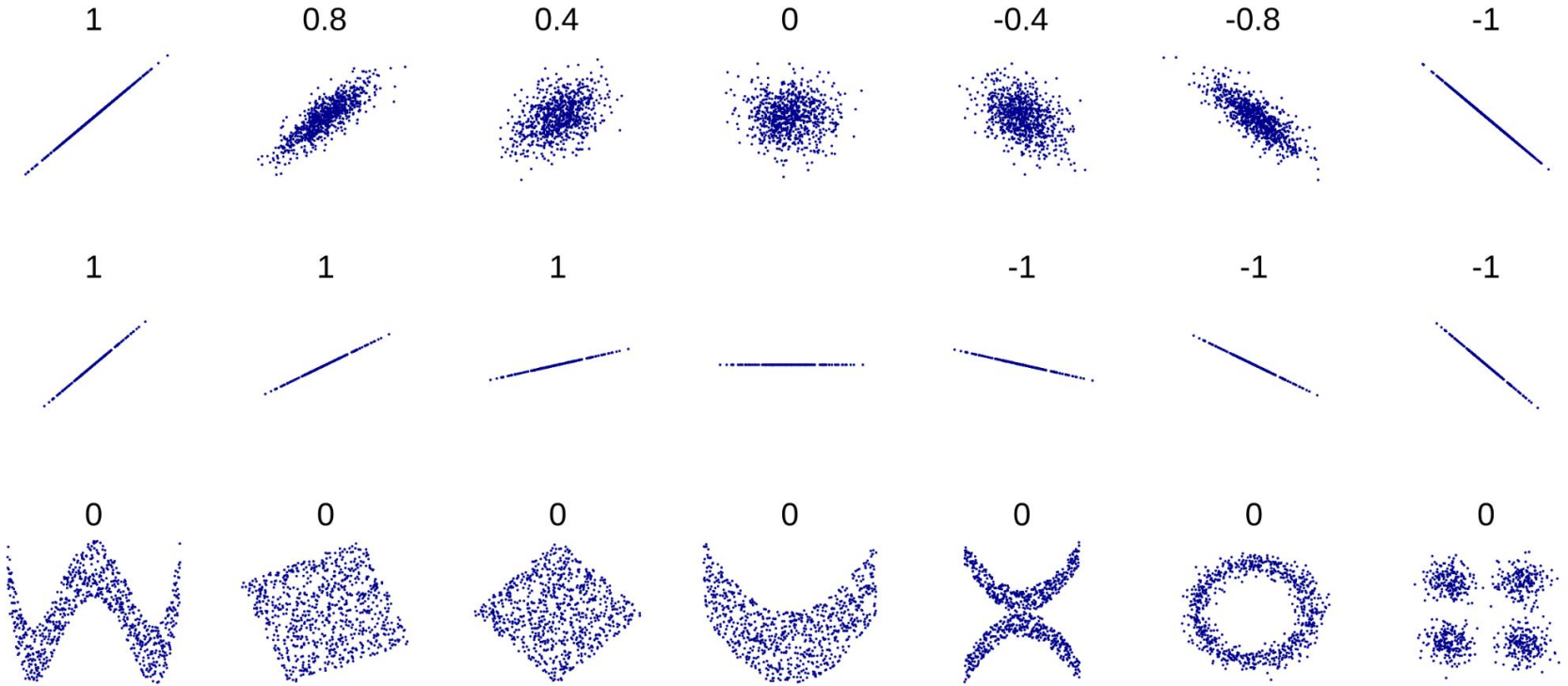


Multiple Linear Regresion

Preparation

Watercolors by Antonio Sánchez Chinchón 🎨

Correlation



Correlation

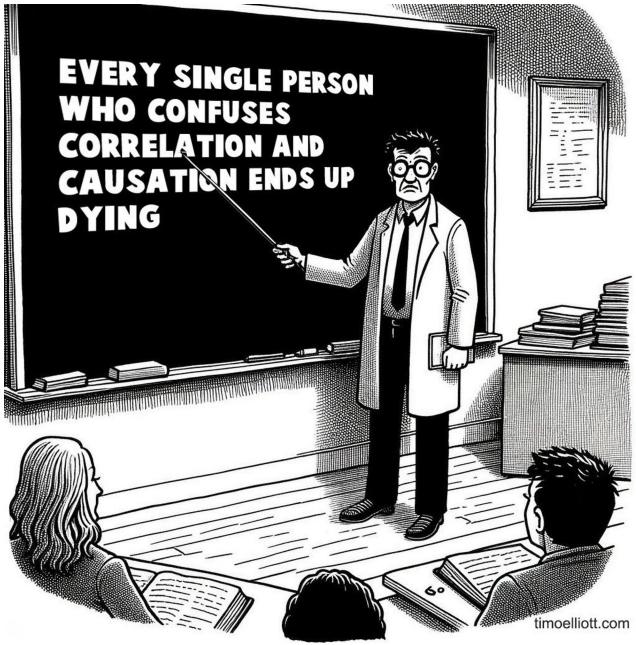


Illustration by Timo Elliott

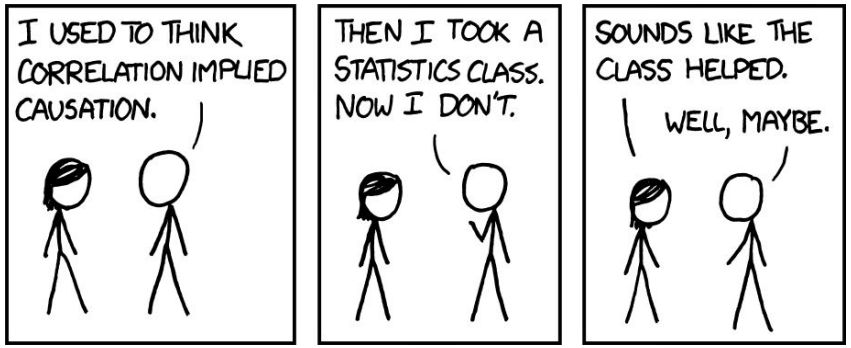
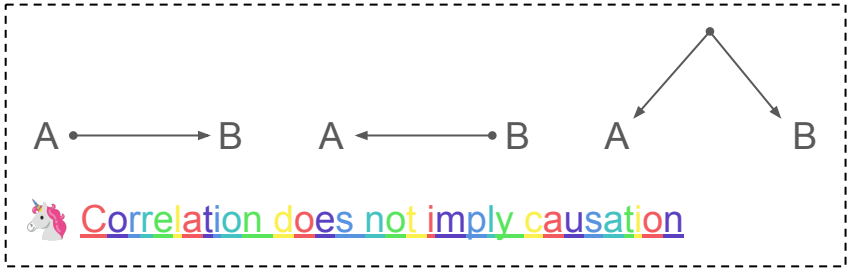
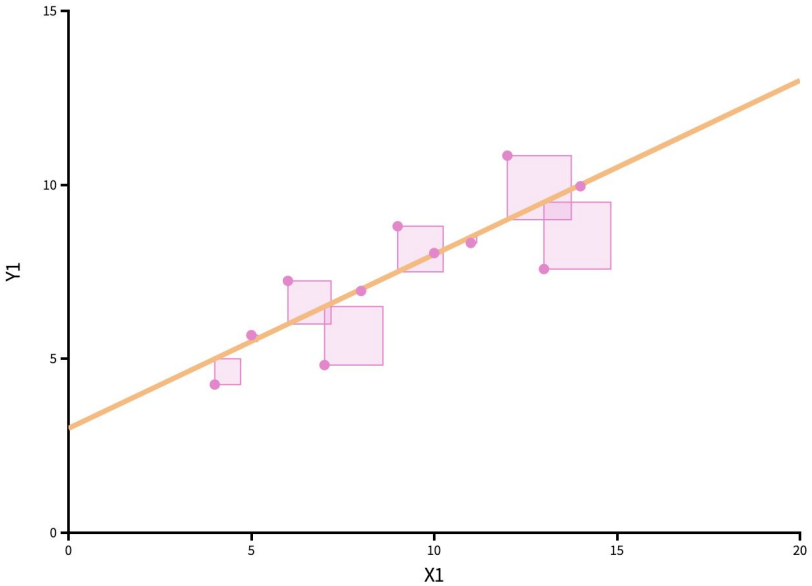


Illustration by Randall Munroe (wtf)



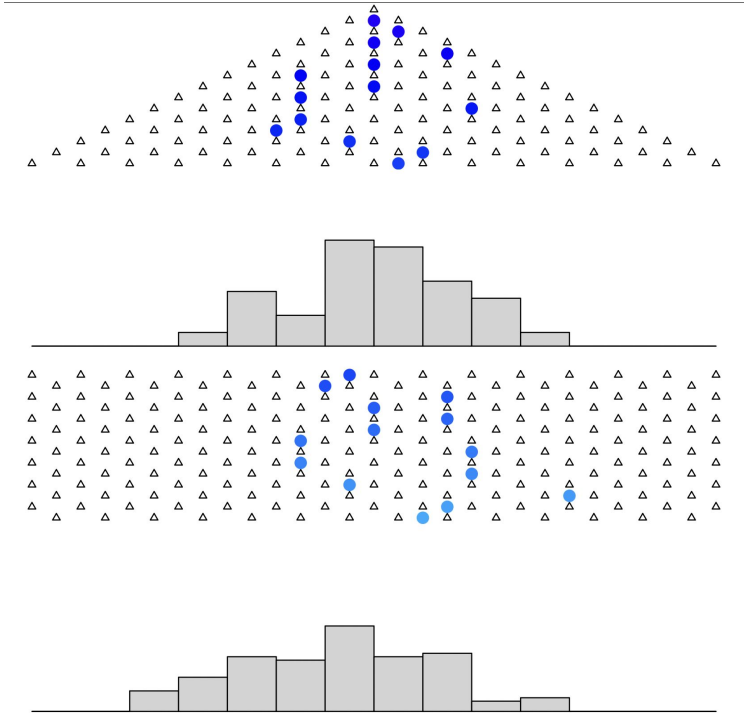
Regression analysis (for estimating relationships)

> *Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome' or 'response' variable, or a 'label' in machine learning parlance) and one or more independent variables (often called 'predictors', 'covariates', 'explanatory variables' or 'features').*
> — Wikipedia



Ordinary Least Squares (OLS)

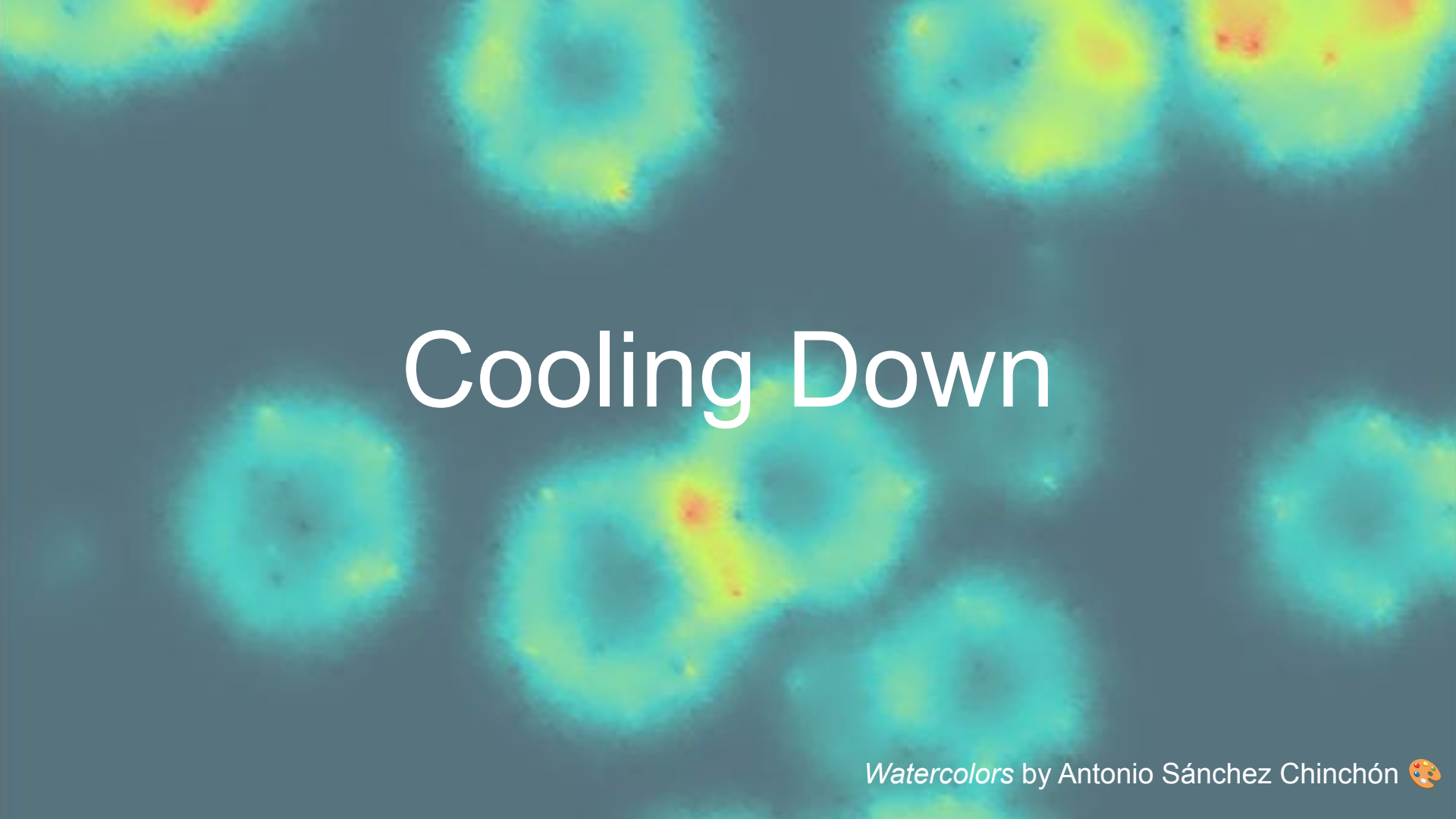
🦄 Regression to the mean



💡

```
# Galton board 2 (do not run in RStudio)
library("animation")
ani.options(nmax = 200 + 15 - 2, 2)
freq <- quincunx2(balls = 200, col.balls =
rainbow(200))
barplot(freq$top, space = 0) # top layers
barplot(freq$bottom, space = 0) # bottom layers
```

Cooling Down

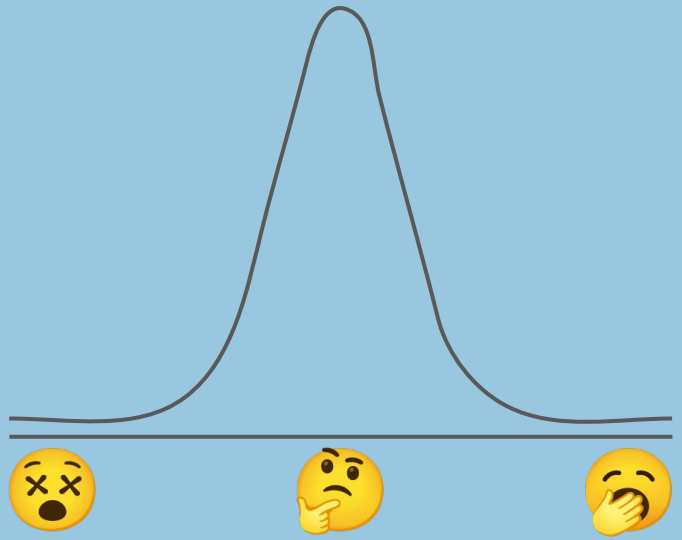


Watercolors by Antonio Sánchez Chinchón 🎨

🙋 Questionnaire

👍 👎

Frequentist Refresher



Student Puzzlement Scale

Takeaways

Illustration by Randall Munroe (wtf)

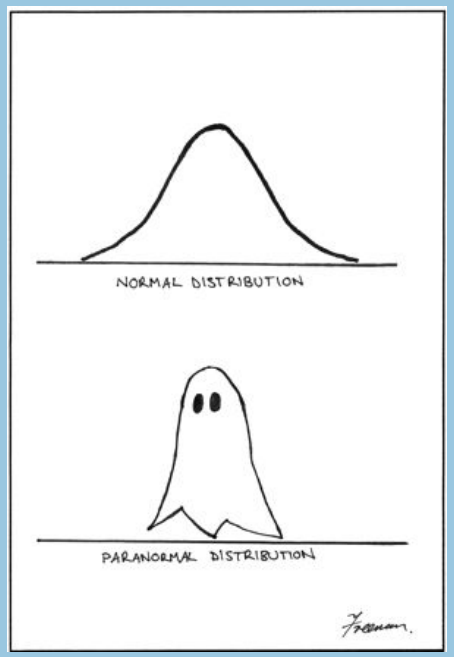


Illustration by Matthew Freeman

Takeaways

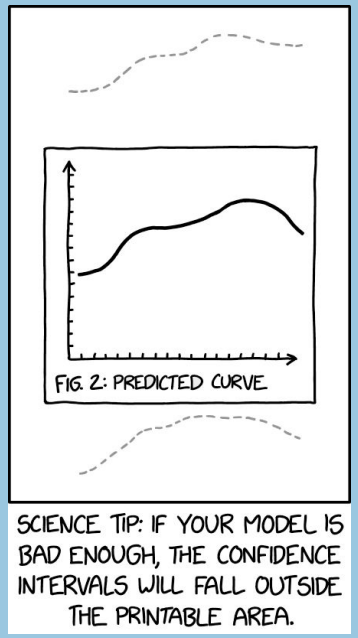


Illustration by Randall Munroe (wtf)

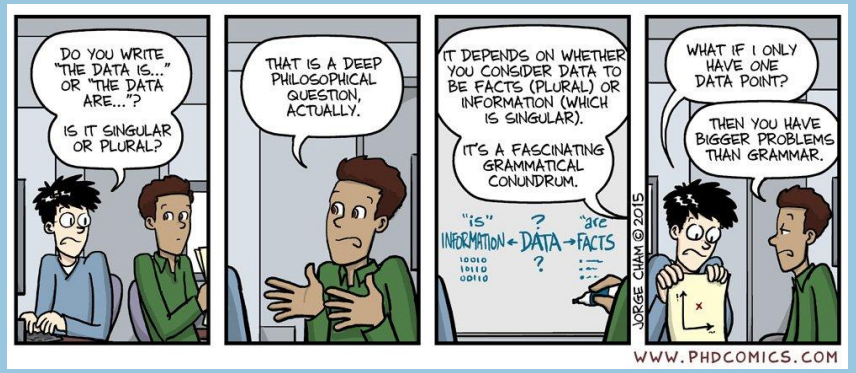
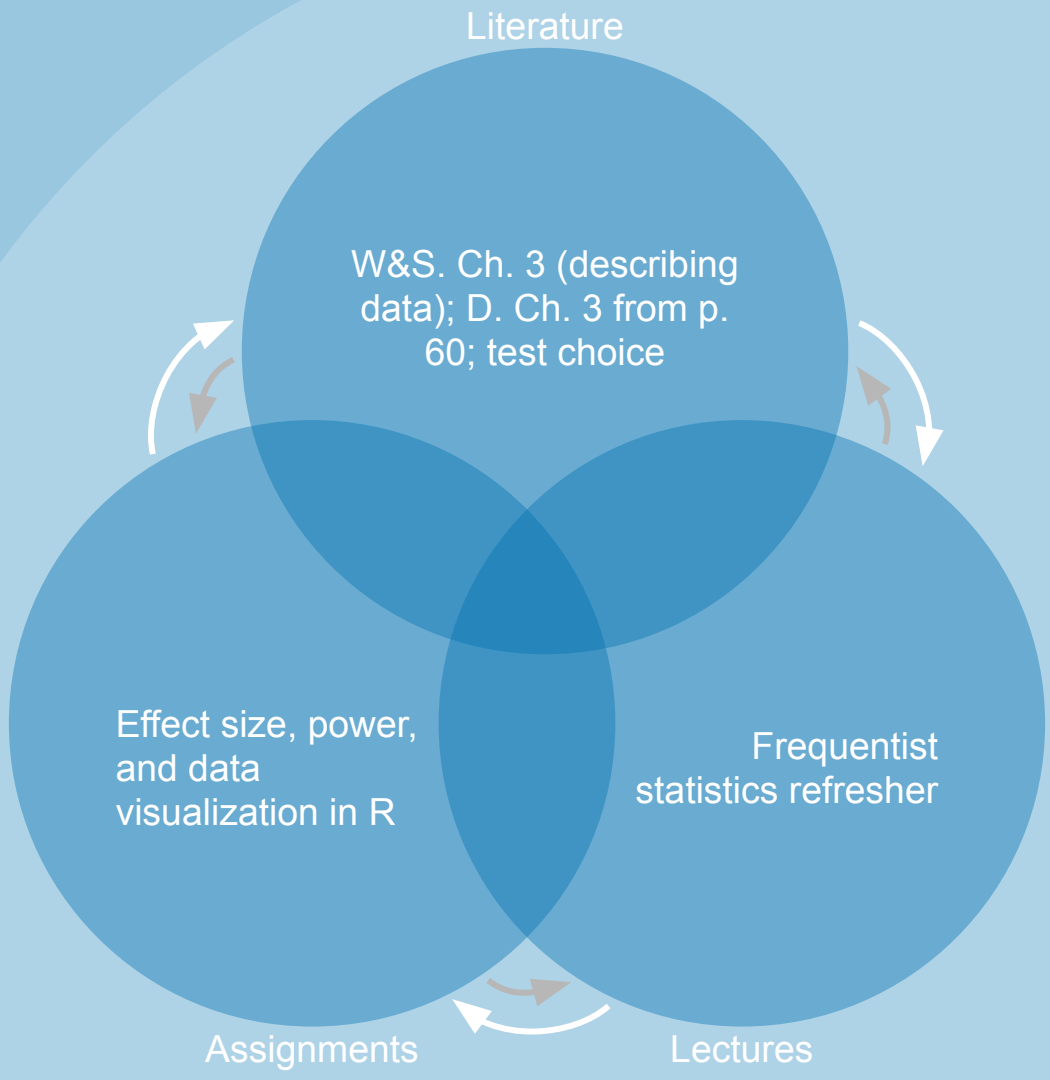


Illustration by Jorge Cham

🥇 Nail it

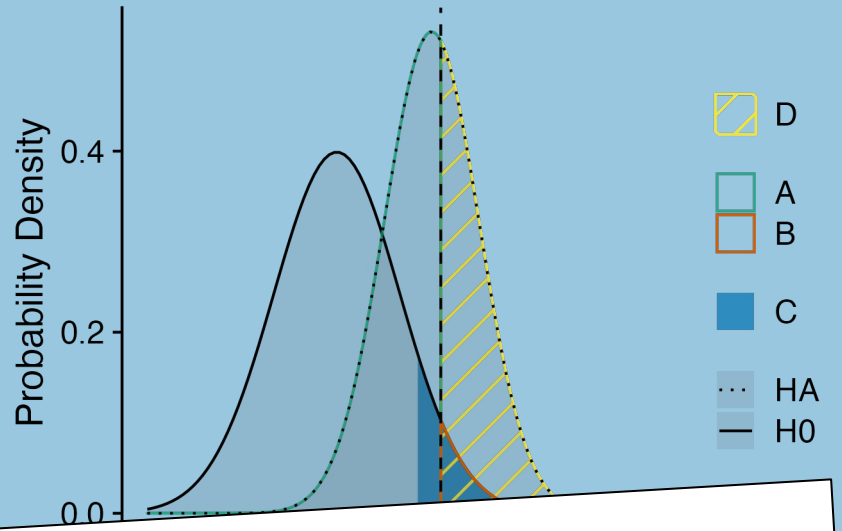

Literature
W&S. Ch. 3 (describing data); D. Ch. 3 from p. 60; test choice
Effect size, power, and data visualization in R
Frequentist statistics refresher
Assignments
Lectures

🥇 Exam(ple) question

In de figuur zie je de steekproevenverdelingen (sampling distributions) van een nulhypothese en een alternatieve hypothese. Er wordt eenzijdig getoetst. De verticale stippellijn geeft de kritieke waarde weer.

Welk oppervlakte beschrijft de kans op een type II fout?

A.
B.
C.
D.



🤔 What do the incorrect answer options represent?

👩‍💻 Take-home assignments

📆 Weekly assignment

🍻 Pub quiz

Create an *informative* four-choice question about the content of today's lecture.

An informative question has a large spread in responses across answer options.

Clarify answer options (which are (in)correct and why).

Illustration adapted from Snippets.com

🔮 Overview

Topics

Probabilities & distributions
🔁 Frequentist inference
  Central limit theorem
  Null hypothesis significance testing
  Confidence intervals
  Student's *t*-test
  Correlation
  Simple linear regression
Multiple linear regression
Factorial ANOVA
Nonparametric inference
Bayesian inference

Illustration by Jennifer Cheuk

🧰 Look here!

Central limit theorem

- CreatureCast (NYTimes)
- Visual introduction (3Blue1Brown)
- R simulation (animation package)

***P*-value**

- Interactive visualization (Kristoffer Magnusson)

Effect size

- Cohen's d web simulation (Kristoffer Magnusson)

Statistical power

- Sample size justification (Daniël Lakens)

Confidence interval

- Web simulation (Seeing Theory, adapted from Kristoffer Magnusson)
- R simulation (animation package)

Student's *t*-distribution

- Web simulation (Kristoffer Magnusson)

Correlation

- Guess the correlation (Omar Wagih, wiki)
- Spurious correlations (Tyler Vigen)
- Interactive visualization (Kristoffer Magnusson)

Regression analysis

- OLS visualization (Seeing Theory)
- Interactive OLS explanation (Explained Visually)
- Regression to the mean (DataCamp)

⚠️ Don’t look here!

Show the distribution of *p*-values when the null-hypothesis is true (e.g., two samples come from the same population).

Additional challenge I: what happens to the distribution if the null hypothesis is false?

Additional challenge II: argue why the distribution in the comic would be problematic.

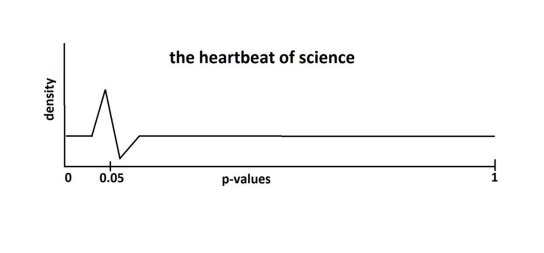


Illustration by František Bartoš

Hints (select and copy/paste the invisible text below to reveal it)

0.

1.

2.

3.

✍️ Colophon

Slides
alexandersavi.nl/teaching/

License