Resampling as a cure for traditional statistics

Matthew Brett

Traditional teaching doesn't work

> *But we've been trying to [teach the fundamentals of statistical analysis] for decades and it just hasn't worked. Instead we have run course after course that students hate. We've turned out generations of graduates who can remember sitting in labs pressing buttons in statistical software programmes like SPSS, but never really learned how to connect statistics to important issues in the real world.*

What do you get if you don't teach stats properly? Farage and Trump (Robert de Vries, Guardian, April 2019).

Statistics teachers think we need to change track

George Cobb (2015): "Mere renovation is too little too late: we need to rethink our undergraduate curriculum from the ground up".

See article and discussion at https://nhorton.people.amherst.edu/mererenovation.

The traditional approach is fast approaching a dead end

The Introductory Statistics Course: A Ptolemaic Curriculum

George W. Cobb
Mount Holyoke College

(Cobb 2007)

Data science!

Berkeley Division of Data Science and Information

About ▾   Academics ▾   Research ▾   News ▾   Events   Support Us

Home » Academics » Undergraduate Programs » Data Science Offerings » Data 8

Data 8: Foundations of Data Science

DATA SCIENCE OFFERINGS

- **Data 8** ▾
  - Data 8 FAQs
  - Data 8 Logistics
  - Data 8X
- Data 100
- Human Contexts and Ethics
- Connectors

Why now?

- Improvements in languages
- Greater experience of coding in science
- Marriage of coding and data analysis
- Data science!

What is data science? (link)

This session

- Introduction to R
- A mosquito problem

This is about 8 hours of teaching and practice

A mosquito problem

Beer Consumption Increases Human Attractiveness to Malaria Mosquitoes

Thierry Lefèvre ✉, Louis-Clément Gouagna, Kounbobr Roch Dabiré, Eric Elguero, Didier Fontenille, François Renaud, Carlo Costantini, Frédéric Thomas

With thanks to John Rauser: Statistics Without the Agonizing Pain

The data

Beer			
27	20	21	26
27	31	24	21
20	19	23	24
28	19	24	29
18	20	17	31
20	25	28	21
27			

Water			
21	22	15	12
21	16	19	15
22	24	19	23
13	22	20	24
18	20		

The t-test

Independent t-test formula

- Let A and B represent the two groups to compare.
- Let m_A and m_B represent the means of groups A and B, respectively.
- Let n_A and n_B represent the sizes of group A and B, respectively.

The **t test statistic value** to test whether the means are different can be calculated as follow :

$$t = \frac{m_A - m_B}{\sqrt{\frac{S^2}{n_A} + \frac{S^2}{n_B}}}$$

S^2 is an estimator of the common **variance** of the two samples. It can be calculated as follow :

$$S^2 = \frac{\sum (x - m_A)^2 + \sum (x - m_B)^2}{n_A + n_B - 2}$$

Once **t-test statistic value** is determined, you have to read in **t-test table** the **critical value of Student's t distribution** corresponding to the **significance level alpha** of your choice (5%). The **degrees of freedom** (df) used in this test are :
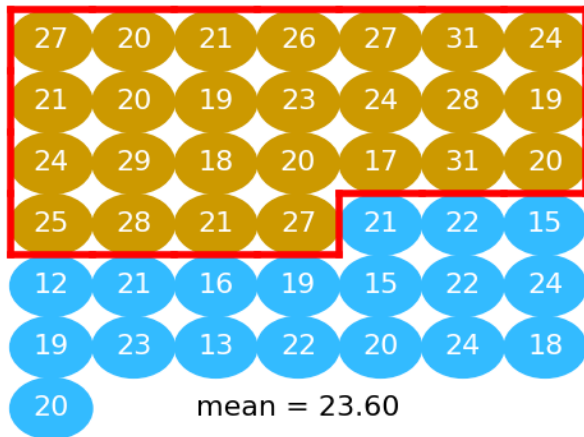
$$df = n_A + n_B - 2$$

The permutation way

- Calculate difference in means
- Pool
- Repeat many times:
    - Shuffle
    - Split
    - Recalculate difference in means
    - Store

On balls
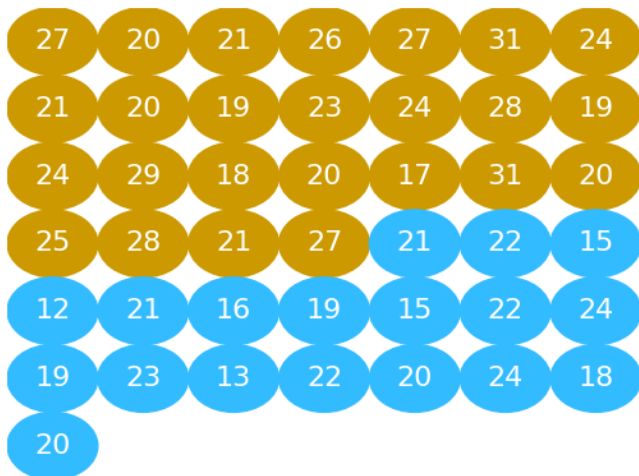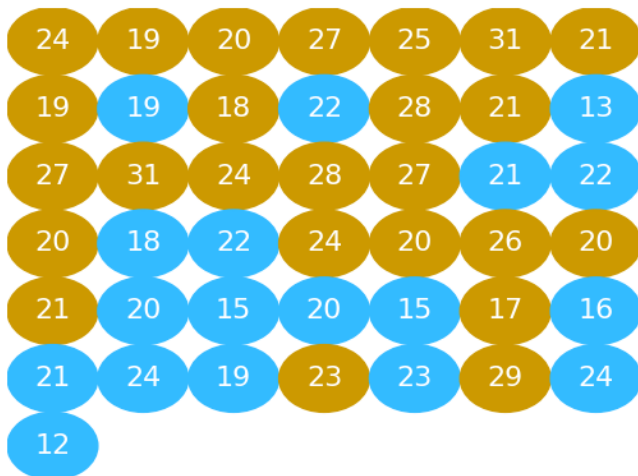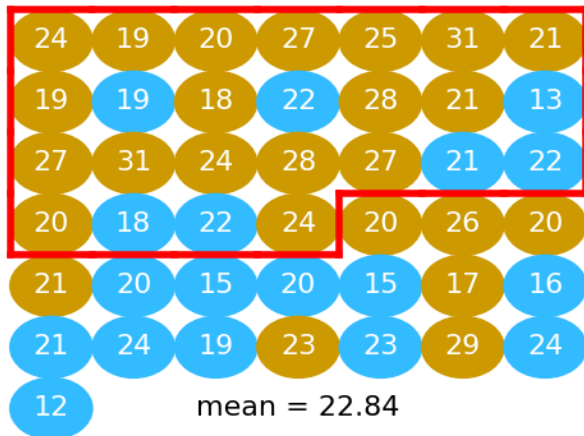
The difference in means

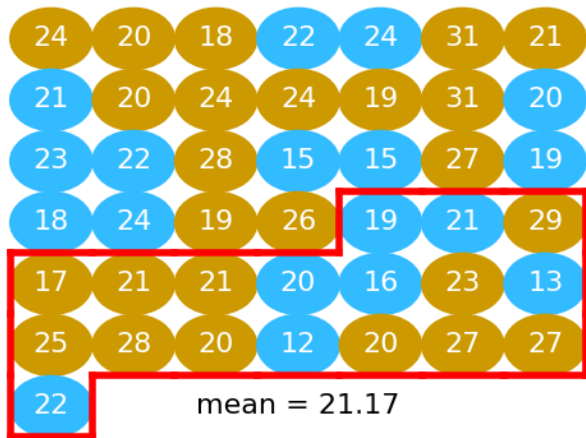mean = 23.60

Shuffle

A difference if the null is true

mean = 22.84

One difference on null: 22.84 - 20.28 = -1.26

And again

mean = 22.20

Another difference on null: 22.20 - 21.17 = 1.03

mean = 21.17

And so on, 10000 times

Sampling distribution of mean difference
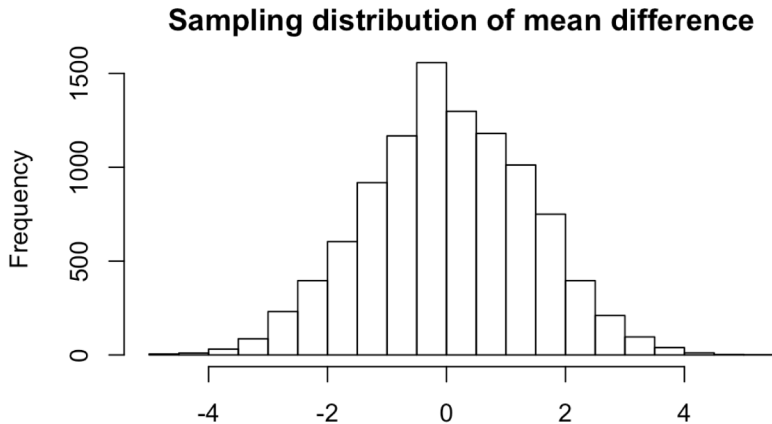
But how?

On to the notebook.