Resampling as a cure for traditional statistics

Matthew Brett

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

Published: March 4, 2010 • https://doi.org/10.1371/journal.pone.0009546

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 :

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

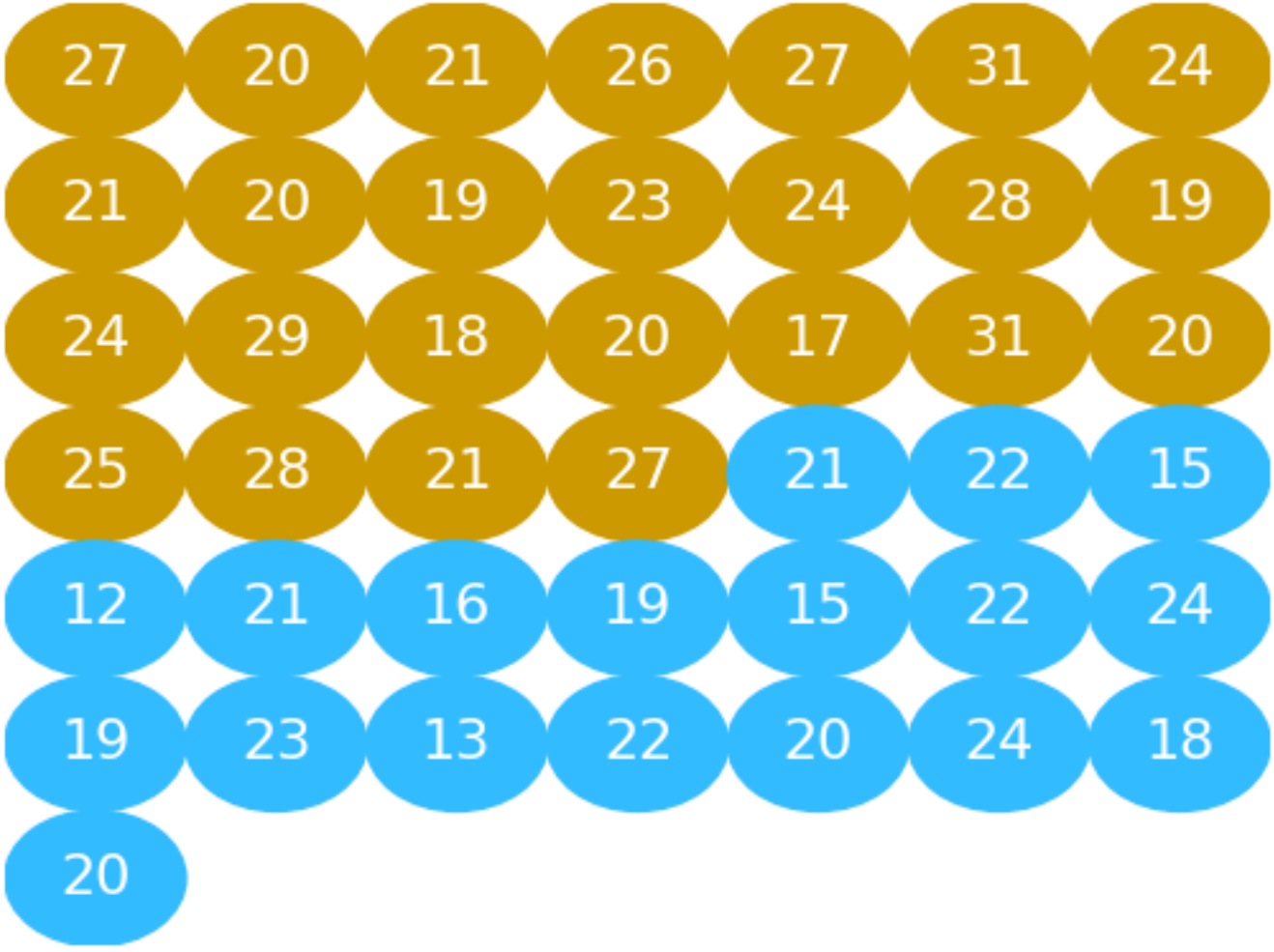
Once **t-test statistic value** is determined, you have to read in **t-test table** the **critical value of Student's t distribution** corresponding to the **significance level alpha** of your choice (5%). The **degrees of freedom** (df) used in this test are :

$$df = n_A + n_B - 2$$

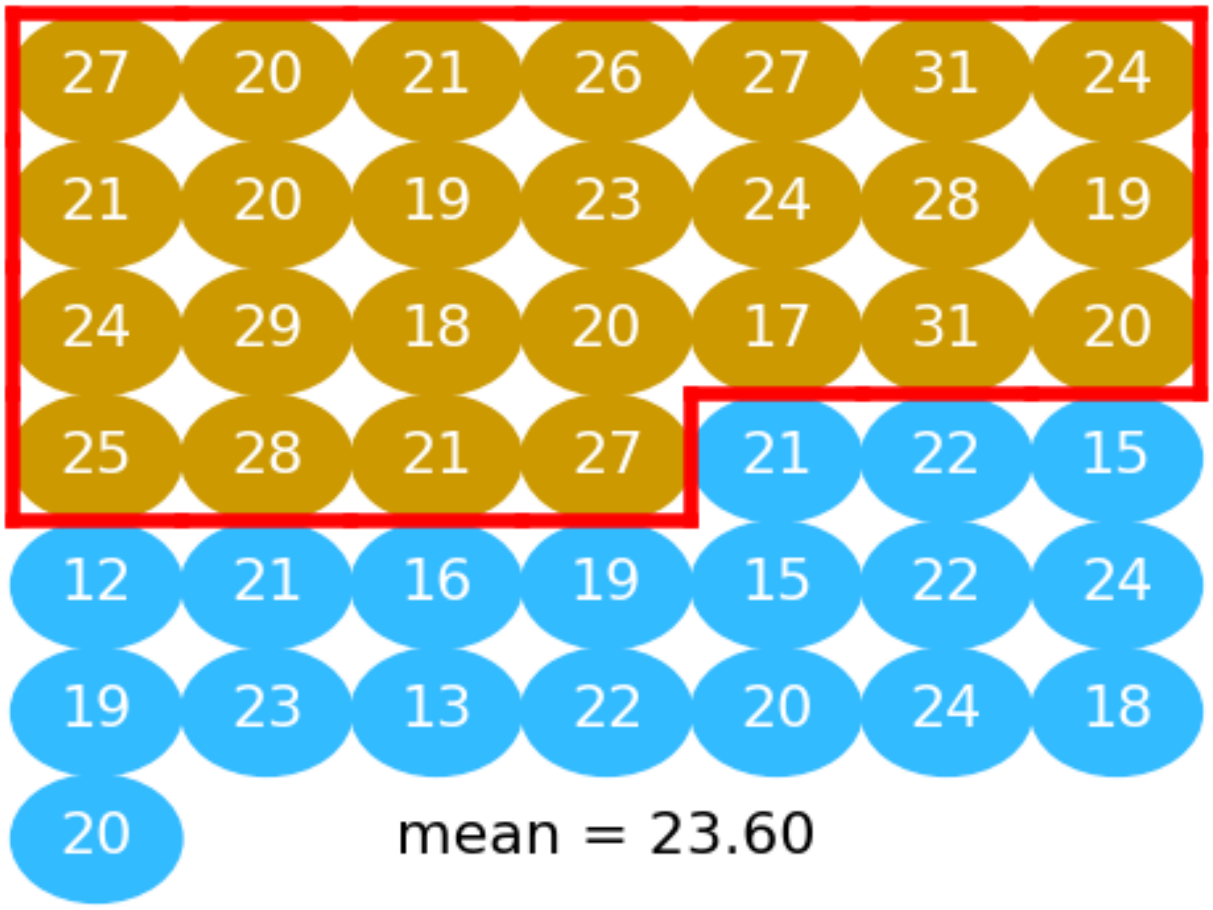
The permutation way

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

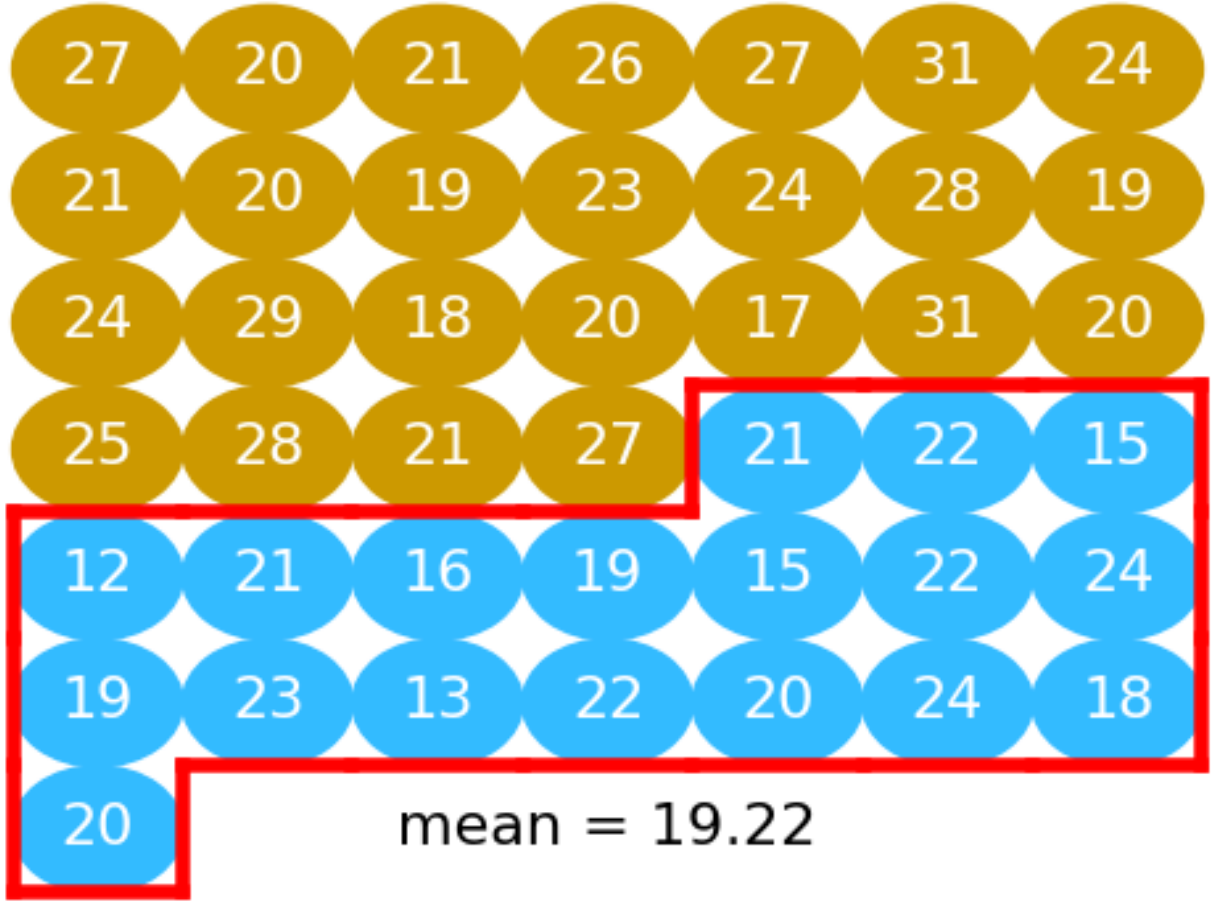
On balls


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

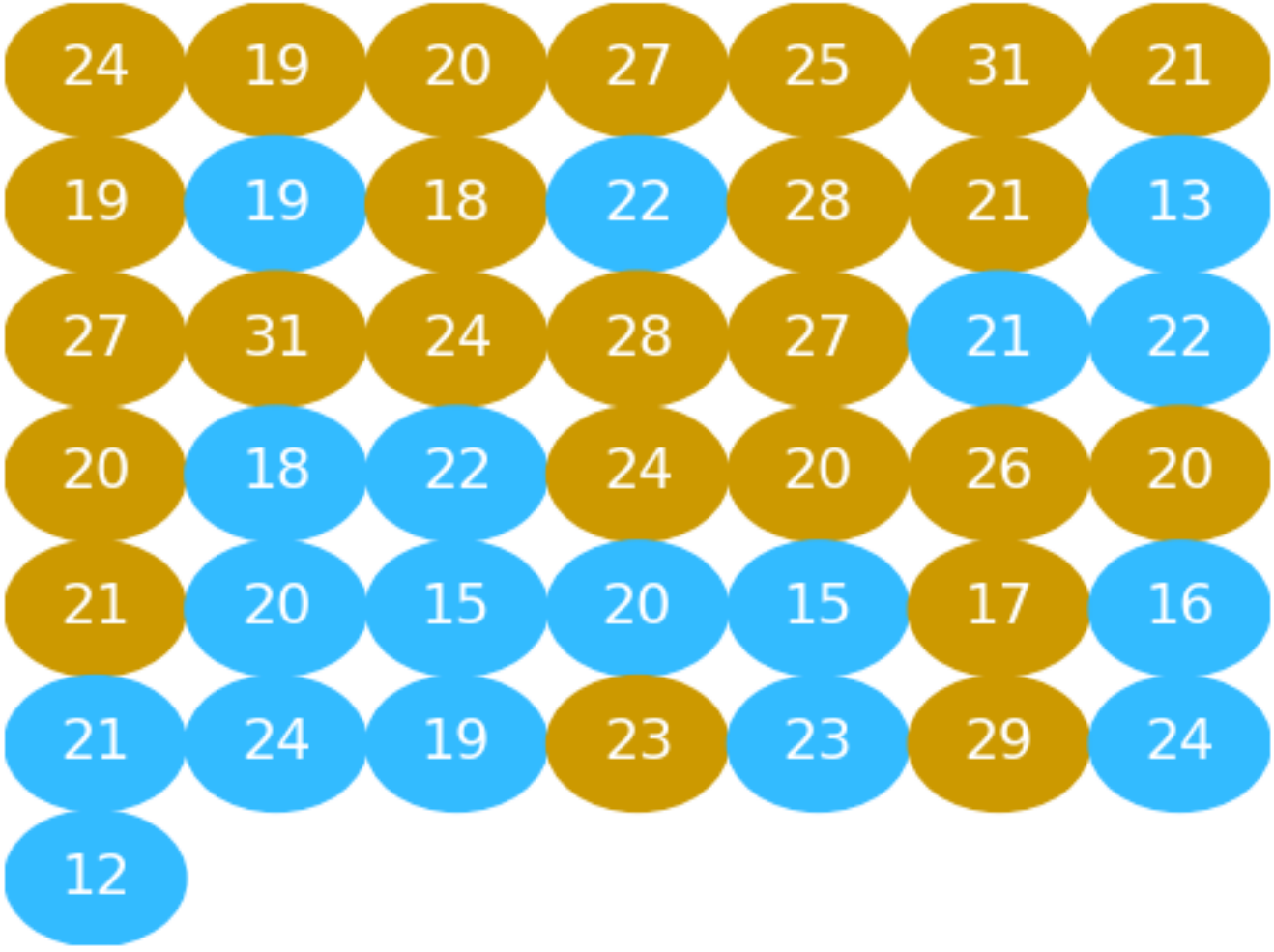
The difference in means



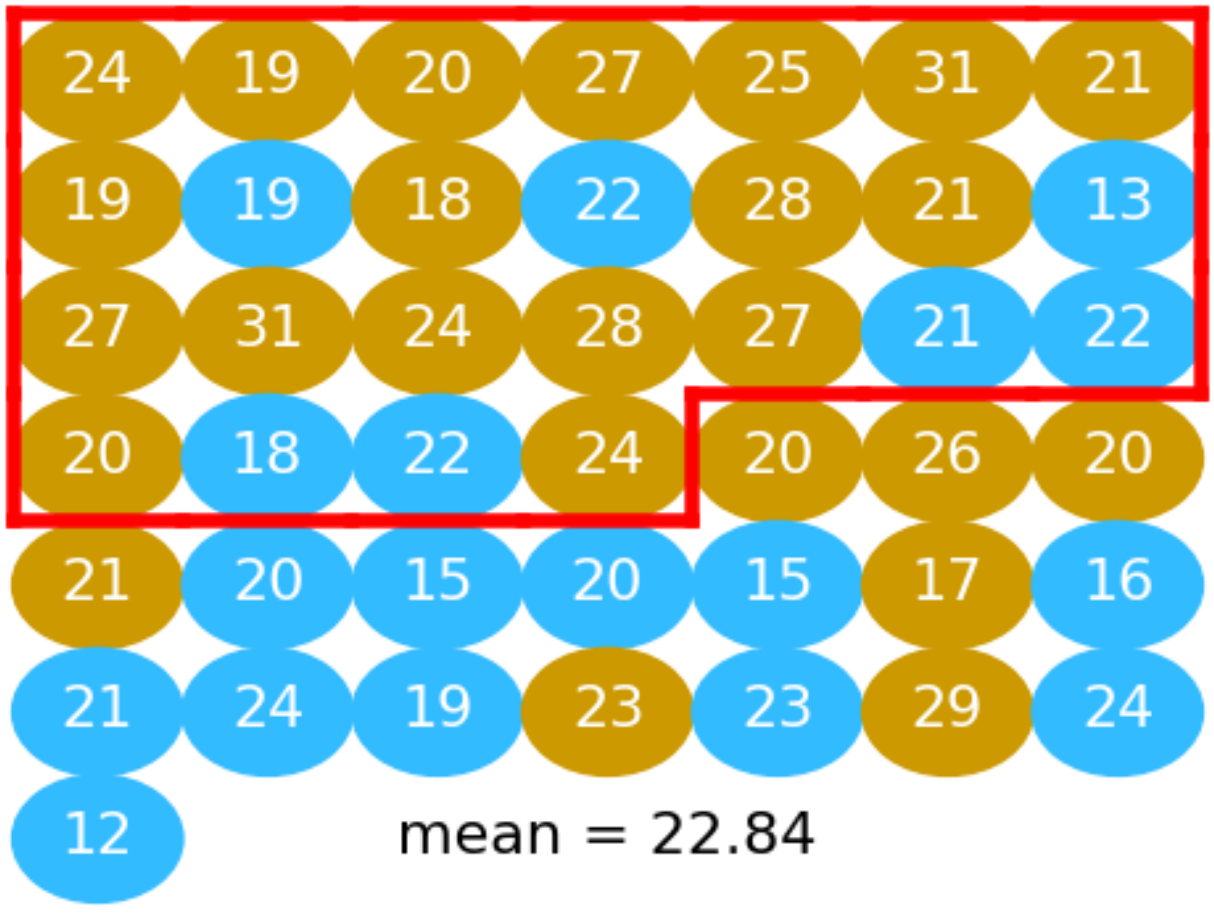
The difference in means: 23.60 - 19.22 4.38



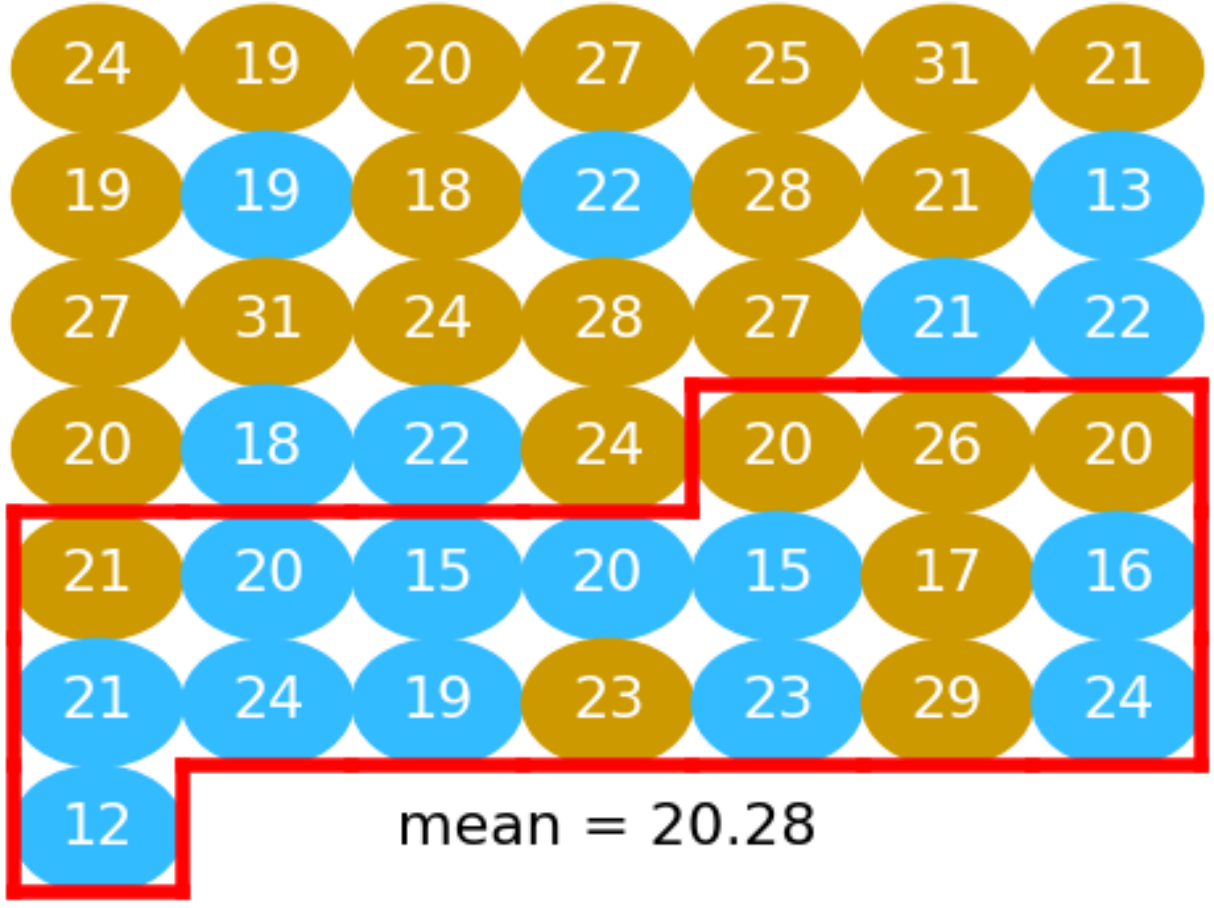
Shuffle


24 19 20 27 25 31 21
19 19 18 22 28 21 13
27 31 24 28 27 21 22
20 18 22 24 20 26 20
21 20 15 20 15 17 16
21 24 19 23 23 29 24
12

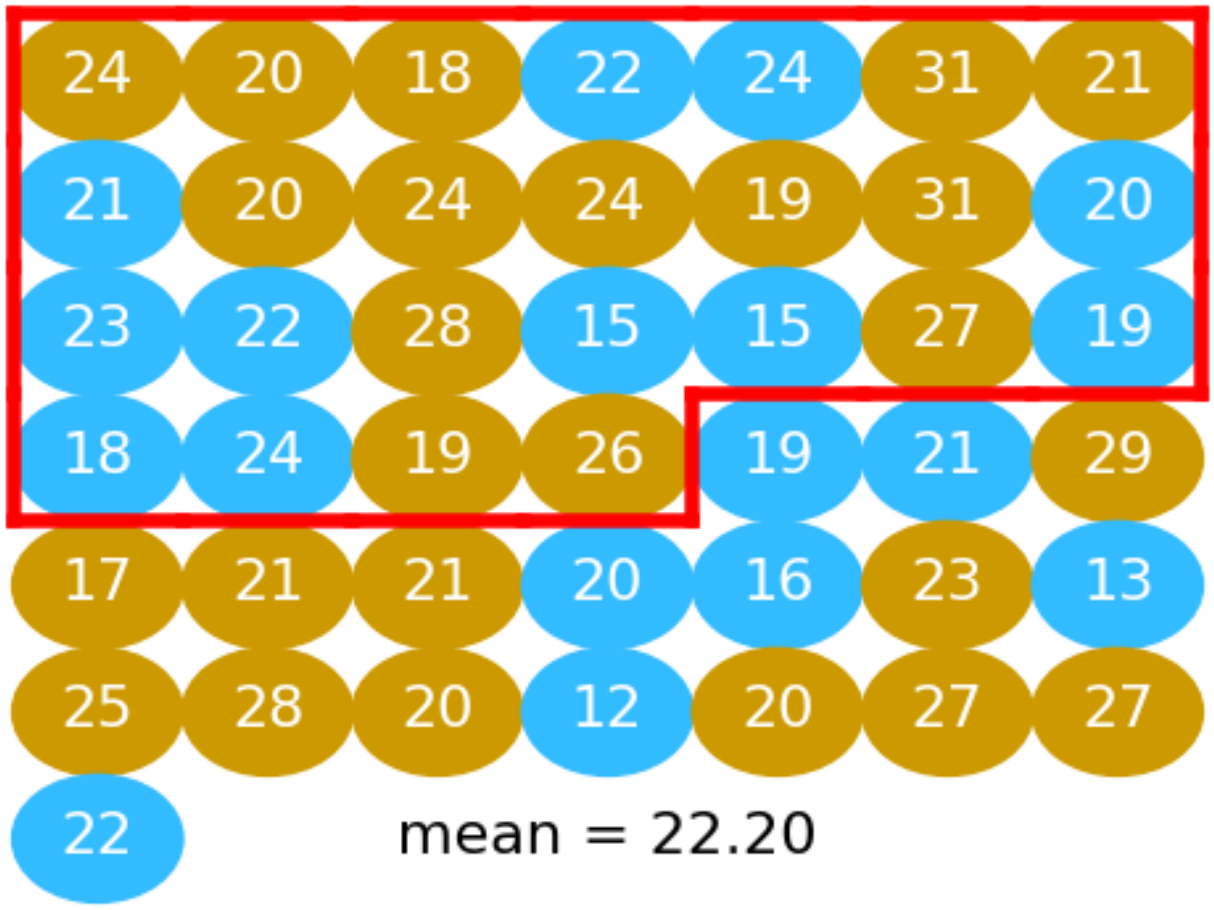
A difference if the null is true



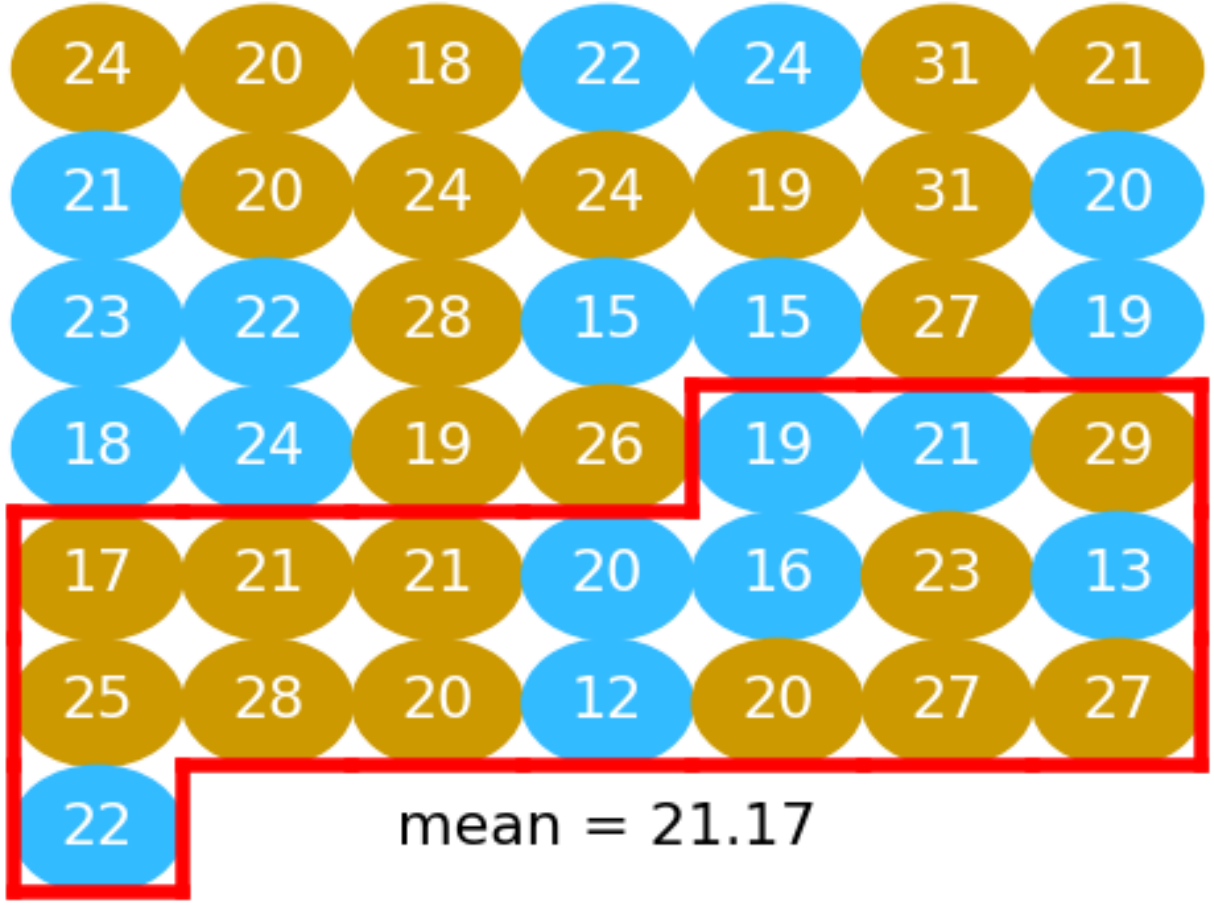
One difference on null: 22.84 - 20.28 = -1.26



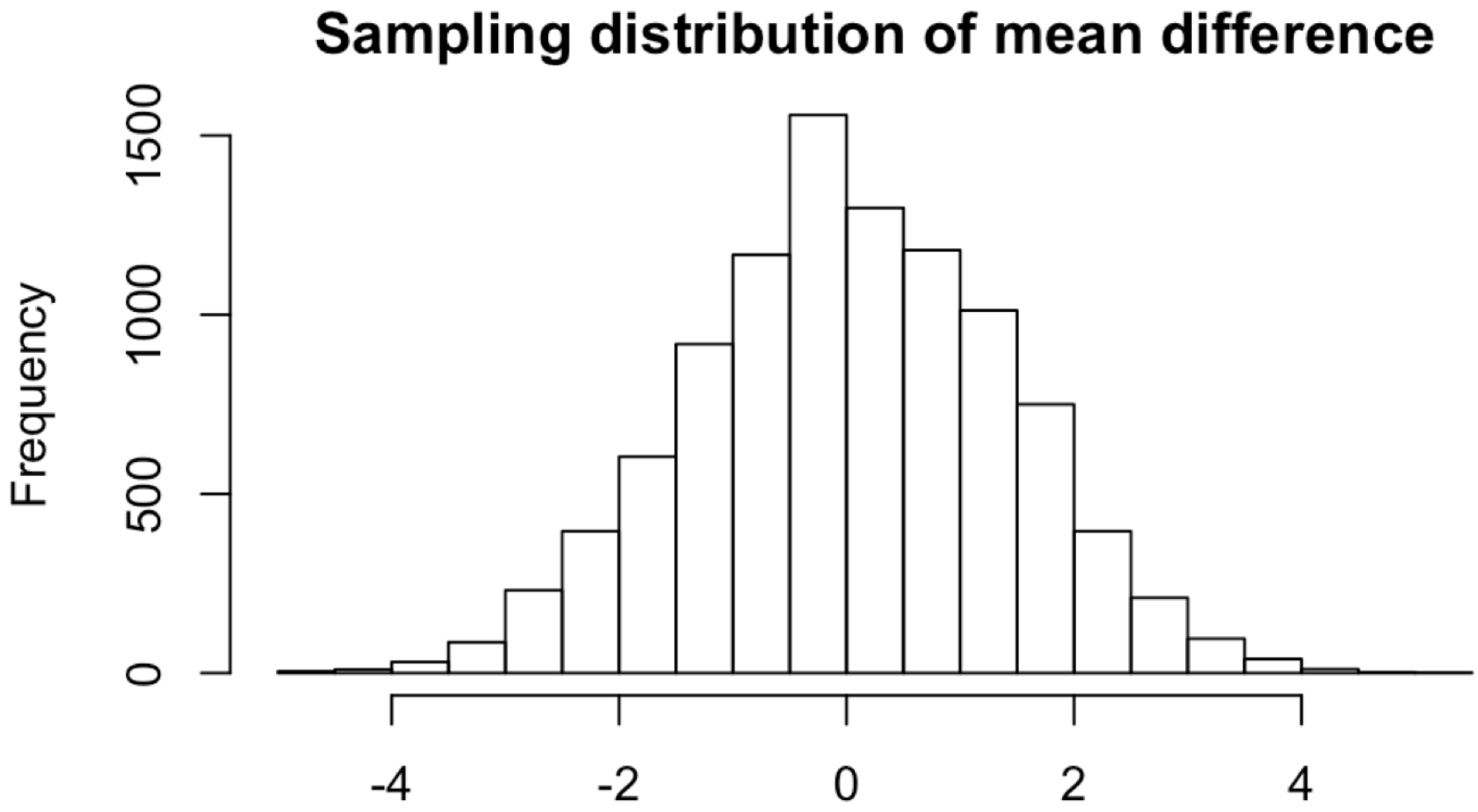
And again


24 20 18 22 24 31 21
21 20 24 24 19 31 20
23 22 28 15 15 27 19
18 24 19 26 19 21 29
17 21 21 20 16 23 13
25 28 20 12 20 27 27
22
mean = 22.20

Another difference on null: 22.20 - 21.17 = 1.03



And so on, 10000 times



But how?

On to the notebook.