

M1D7 Prelab Agenda

- 1) Statistics is also a skill
- 2) How do we communicate the spread of our data?
- 3) How do we determine whether differences in our data are statistically significant?

- 4) What belongs in our data summary?


↑ Posted by u/Evan2895 3 years ago

19.3k ↓ 9 out of 40 mice exposed to e-cigarette vapor developed a form of lung cancer, according to a new study. These findings have been criticized -- exposure wasn't similar to human vaping -- but the authors argue that e-cig vapor can cause DNA damage that leads to lung cancer over time.

[inverse.com/articl...](https://www.inverse.com/article/100000/e-cigarette-vapor-causes-lung-cancer-in-mice)

Health

1.5k Comments Share Save Hide Report 88% Upvoted



Of the 40 mice exposed to the nicotine vapor, nine of them (22.5 percent) developed adenocarcinoma, the most common form of lung cancer. None of the 18 mice who were exposed to the PG or VG got lung cancer, and only one of the 18 mice in the control group got cancer.



[Redacted] · 3 yr. ago

Especially since the control group had a mouse get cancer. Quite simply this study is trash. Small sample size inconclusive evidence and flawed practices.

↑ 2 ↓ Share Report Save



[Redacted] go

I would not consider 9 cancer mice to be evidence of carcinogenic properties, more likely to me natural causes of cancer. Also 40 is a small number to test imo.

↑ 2 ↓ Share Report Save



[Redacted] go

These studies are great because they'll grt exposure fork sensationalist journos without anyone ever mentioning the authors acknowledgements.

The science in raw form is honest but it's intent is dishonest, which is steering public discourse.

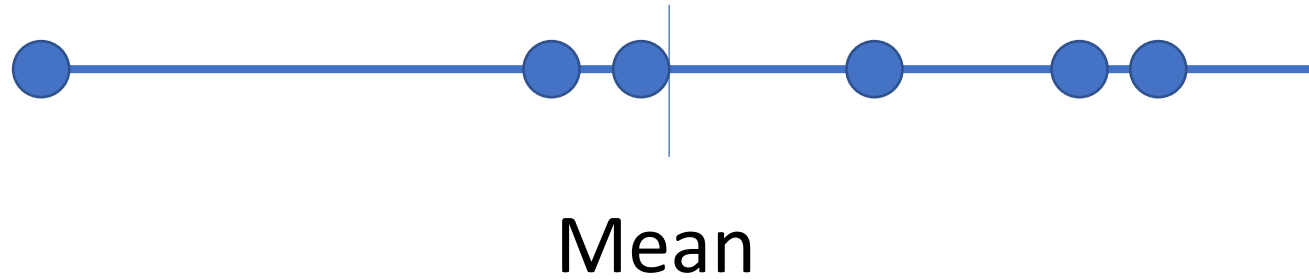
Statistics, much like imaging, is a skill

experiment cannot address the question of interest. We surveyed a random sample of published animal experiments from 2011 to 2016 where interventions were applied to parents and effects examined in the offspring, as regulatory authorities provide clear guidelines on replication with such designs. We found that only 22% of studies (95% CI = 17%–29%) replicated the correct entity–intervention pair and thus made valid statistical inferences. Nearly half of the studies (46%, 95% CI = 38%–53%) had pseudoreplication while 32% (95% CI = 26%–39%) provided insufficient information to make a judgement. Pseudoreplication artificially inflates the sample size, and thus the evidence for a scientific claim, resulting in false positives. We argue that distinguishing between biological units, experimental units, and observational units clarifies where replication should occur, describe the criteria for genuine replication, and provide concrete examples of in vitro, ex vivo, and in vivo experimental designs.

N = 200 papers

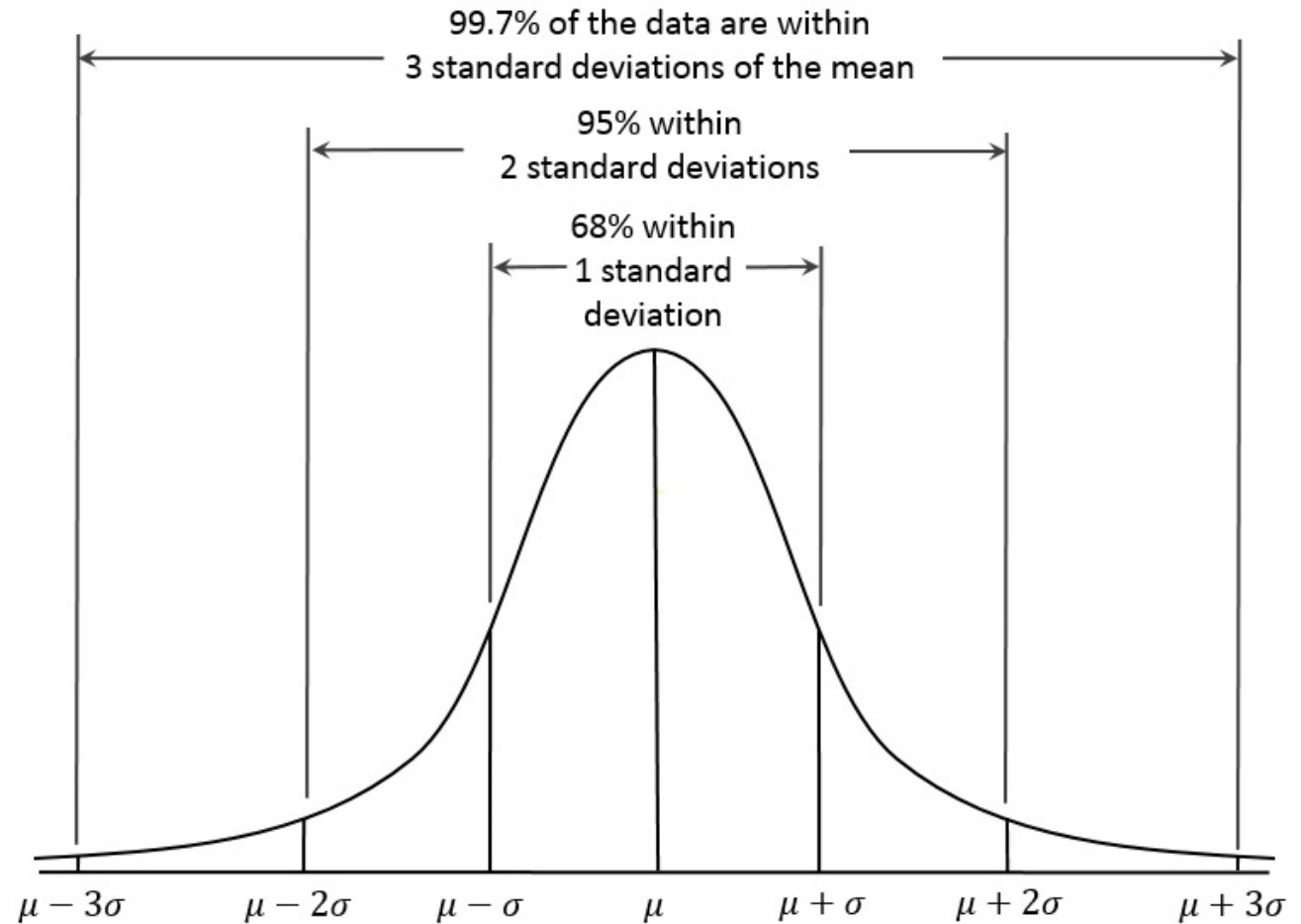
What exactly is 'N' in cell culture and animal experiments? - 2018

Real data varies around is mean



The Standard Deviation gives us a measure of how far observed values are from the mean

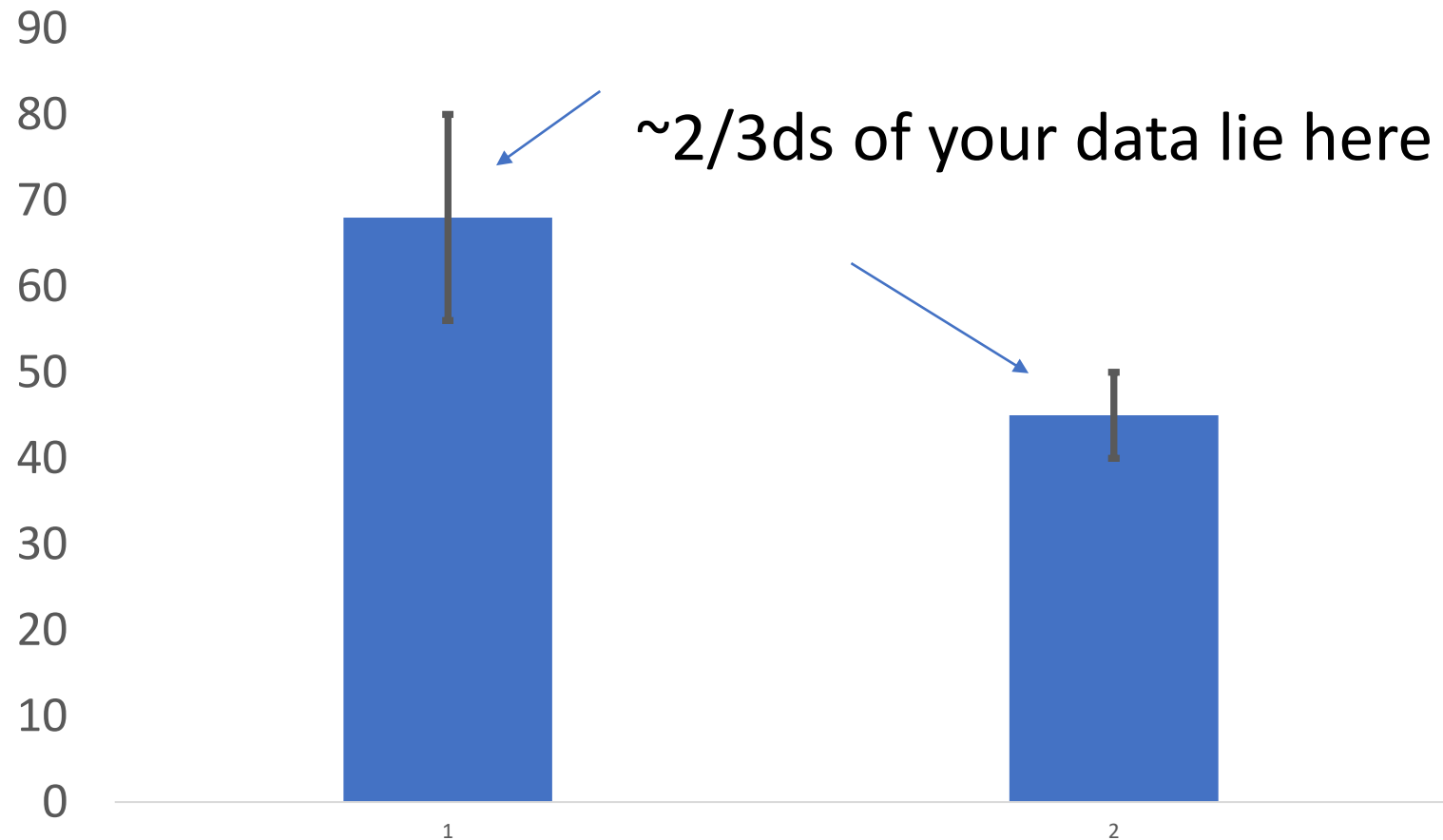
(valid only for
Normally
distributed data)



Standard Deviation gives us an idea of the spread of our data in our sample

Standard deviation gives us an *idea* of whether our differences in data are significant, but is NEVER conclusive

So how DO we assess significance?



Calculating Standard Deviation in Excel

VAR= STDEV(array1)

Sample formula =STDEV(A3:A12)

Calculating Standard Deviation in Python and Matlab

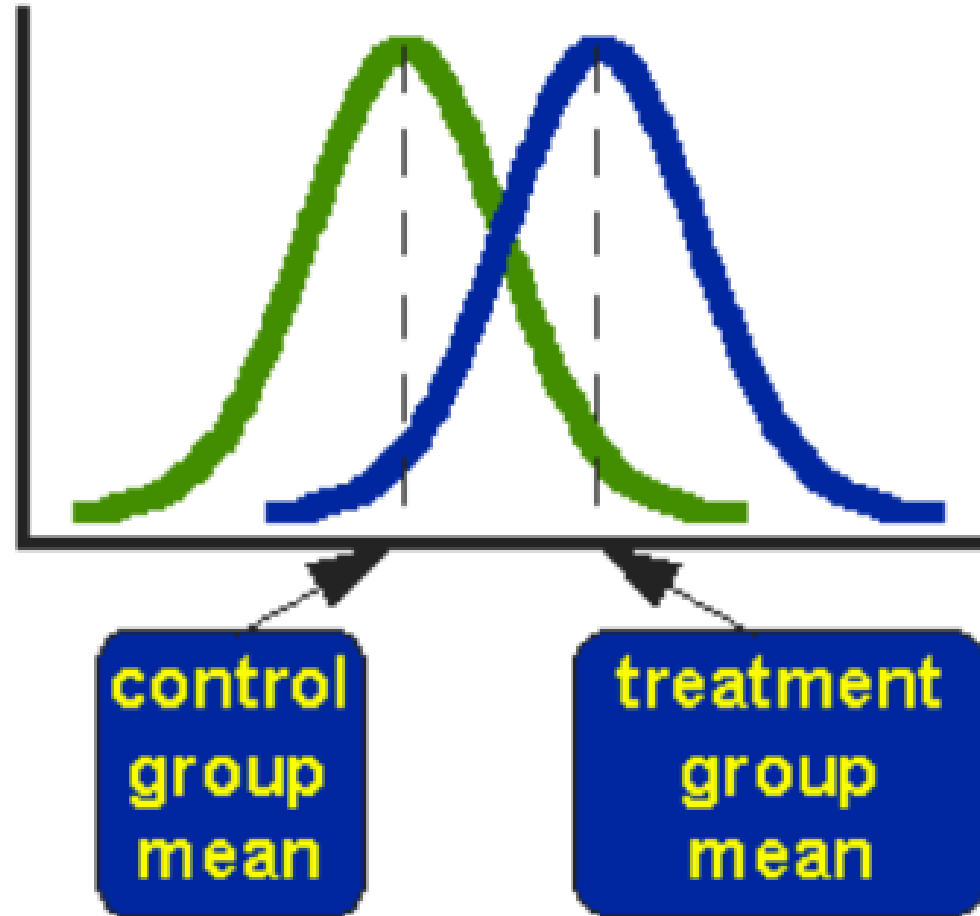
Var = stats.stdev(data)

- data = sequence, list, etc.

Var = std(A)

- A = array

Data Collection



Hypothesis: Students who do not routinely eat breakfast, and were given breakfast as an intervention, **have increased test scores** compared to their non-breakfast eating peers.

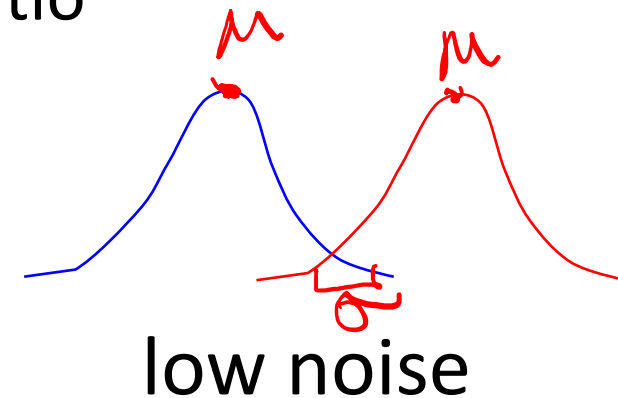
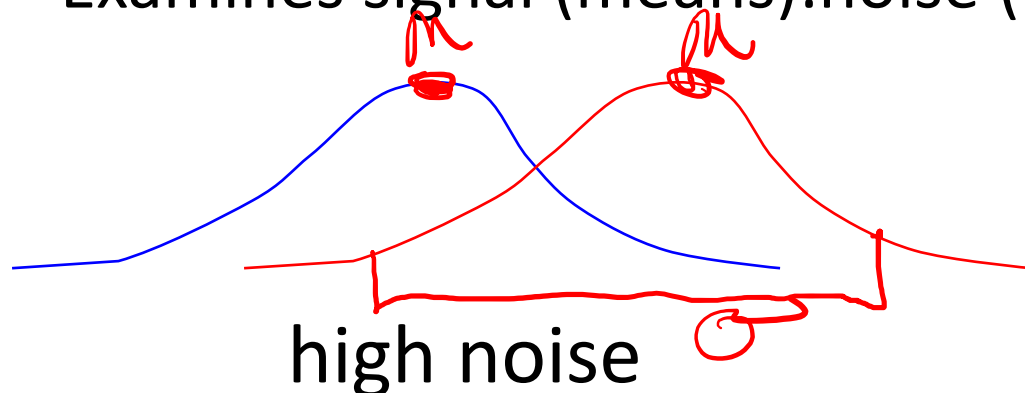
Hypothesis are INNOCENT
until PROVEN GUILTY

Null Hypothesis: Students who do not routinely eat breakfast, and were given breakfast as an intervention, **do NOT have increased test scores** compared to their non-breakfast eating peers.

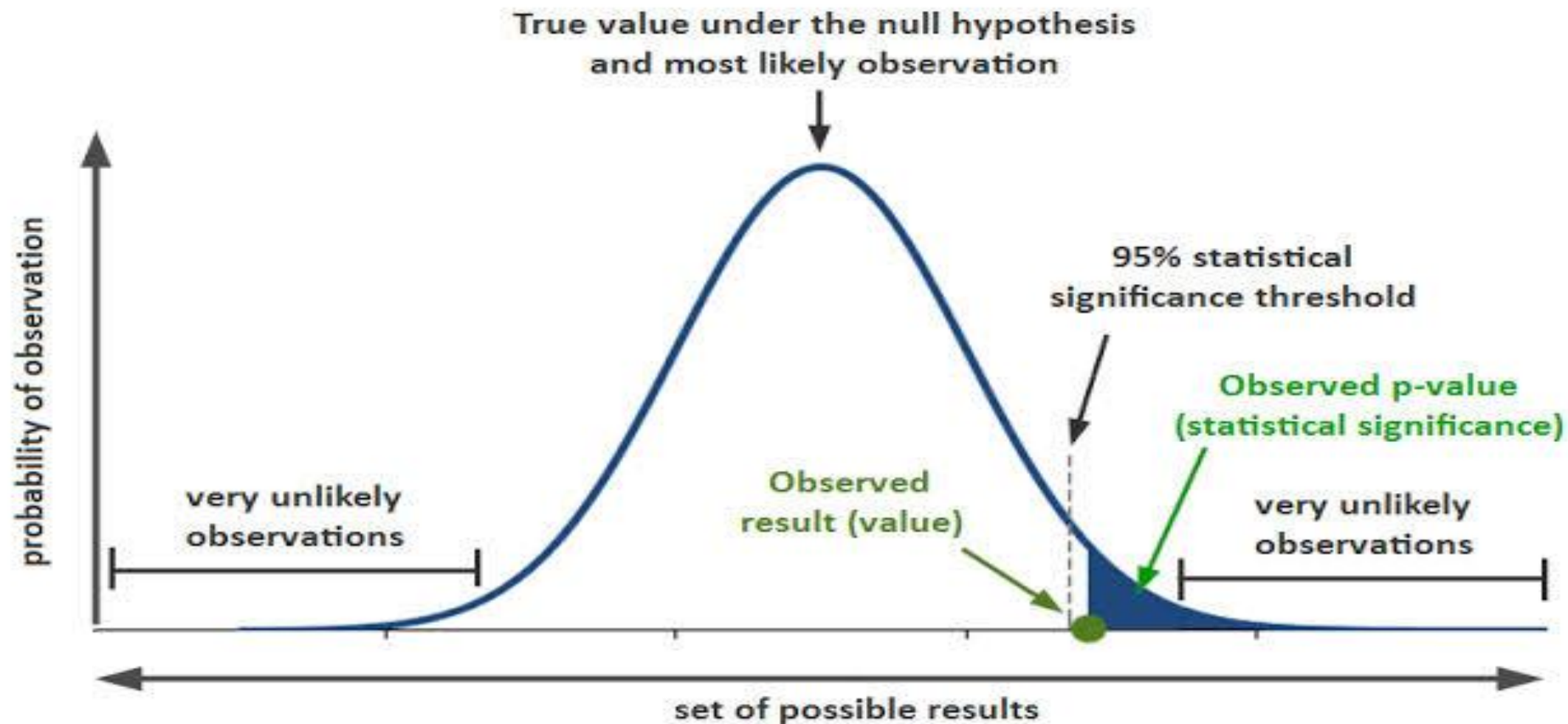
The T-test gives us a
probability of our Null
Hypothesis being TRUE

T Test: Determine if populations are significantly different by comparing the mean of two groups

- Assumption:
 - Smooth & symmetric distribution (continuous variable)
 - Data results in a normal distribution
 - Two populations being compared have similar variance
- At $p < 0.05$, there is less than a 5% chance that populations are the same (95% chance that populations are different)
- Examines signal (means):noise (variance) ratio



Probability & Statistical Significance Explained



Calculating Student's t in Excel

“The most agnostic approach”

$p = \text{T.TEST}(\text{array1}, \text{array2}, 2, 3)$

two-tailed



unequal variance

Sample formula =T.TEST(A2:A10, B2:B10, 2, 3)

Can only compare two data sets at a time

*Make sure it is clear on your plots/writing which conditions are being compared

T-Test in Python & Matlab

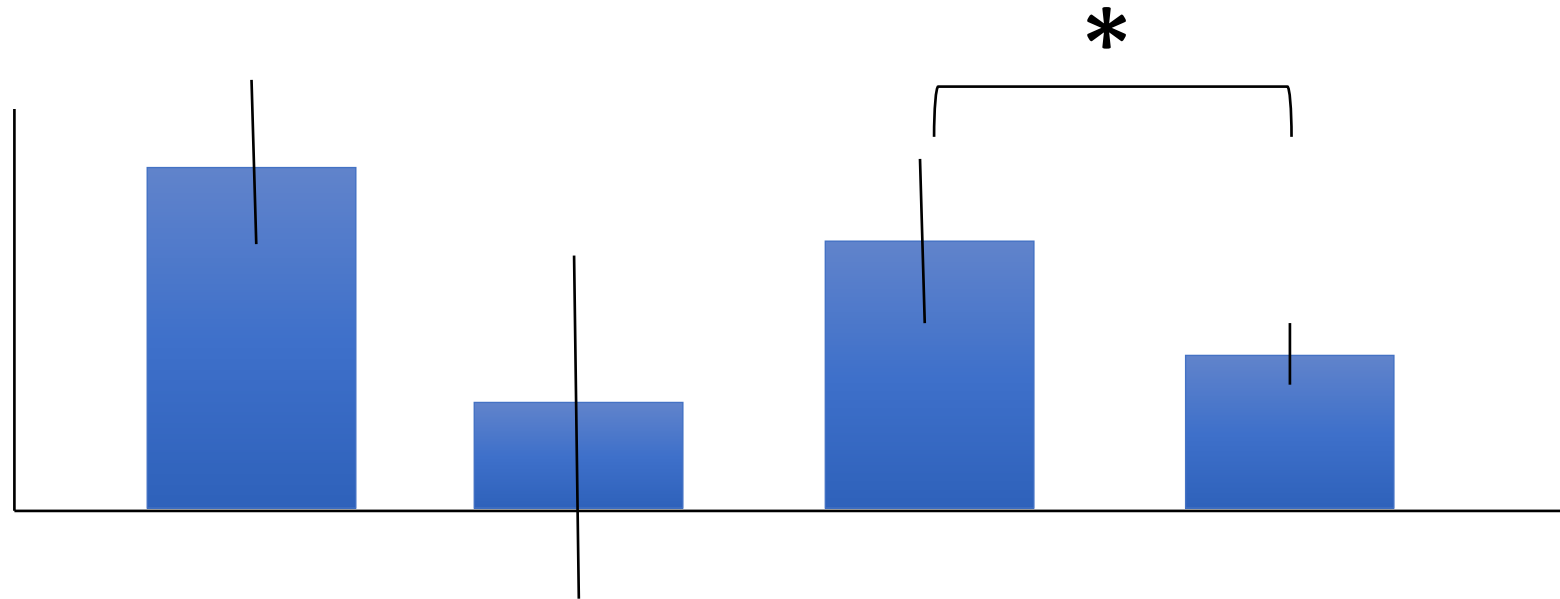
(stat, pvalue) = scipy.stats.ttest_ind(a, b, equal_var)

- a, b = separate lists containing each dataset
- equal_var
 - True assumes equal population variances
 - False assumes unequal → Welch's T-Test

[h, pvalue, ~, stats] = ttest2(data1, data2, 'Vartype', X 'Alpha', A)

- 'Vartype' = 'equal' or 'unequal' in place of X
- 'Alpha' = significance level, # in place of A

How will you use statistics in your data analysis?



What if the data are not statistically significant?

“The means are x and y , but the difference is not significant”

$p = 0.055$

Take Home Messages

- 1) Statistics are specific to every experiment
- 2) Use Standard Deviation for your error bars
- 3) Use a 2 tailed T-test with unequal variances for your significance test
- 4) M1D4 Lab Notebook due **Tomorrow, 10PM** as a PDF of your benchling entry (See the **Assignments Tab** in the wiki)

Today in class

- 1) Work on analyzing your data
 - 1) Refer to the wiki for guidance
- 2) Noreen will talk about what specifically goes in the data summary
- 3) Work on your data summary
- 4) Again, remember that M1D4 notebooks are due tomorrow, 10PM on canvas