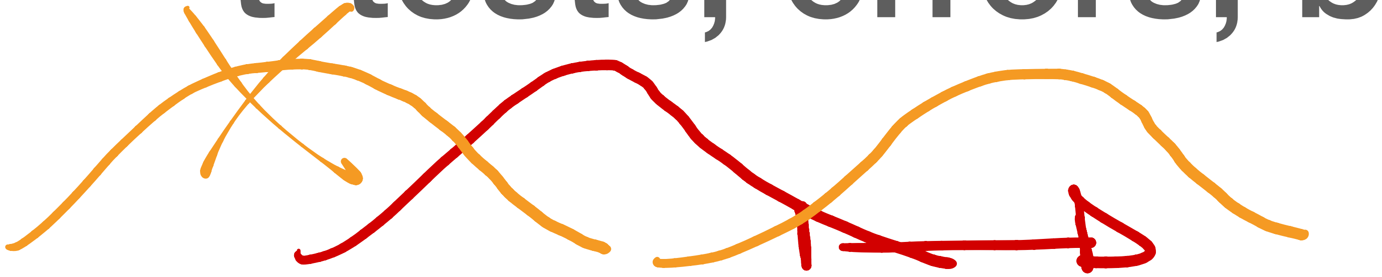
t-tests, errors, bias



What is your hypothesis and null hypothesis for the following question?

Question: Does the orange line experience longer delays than the red line?

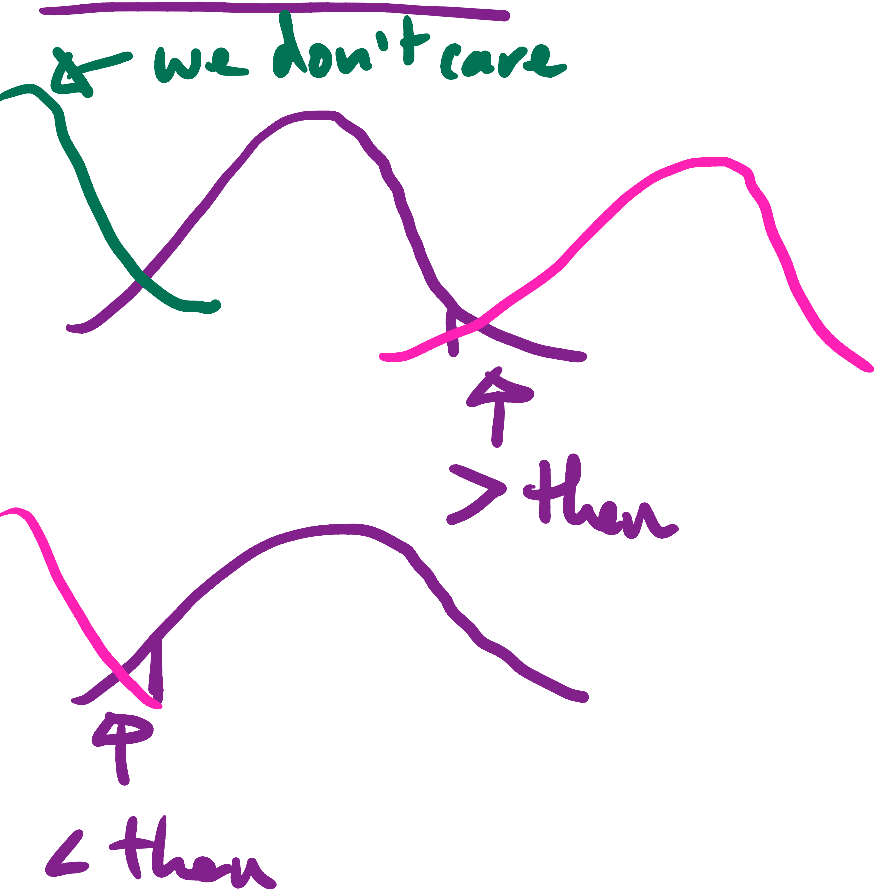
null: $H_0 : \mu_{orange} = \mu_{red}$

$\Rightarrow$ one-tailed t-test

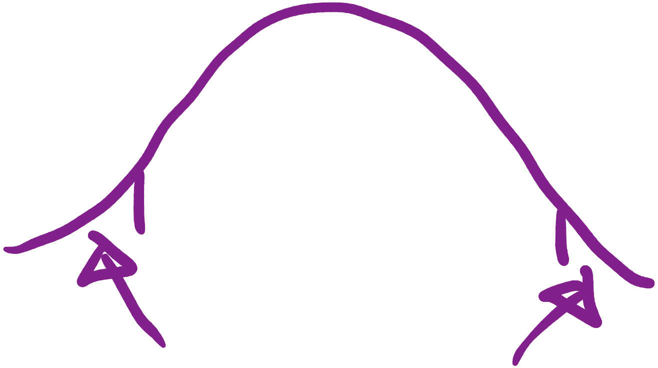
two-tailed t-test

$H_1 : \mu_{orange} > \mu_{red}$ vs. $\mu_{orange} \neq \mu_{red}$

One-tailed



two-tailed



we care about **any** difference in population

Student's t-test

→ two-tailed t-test

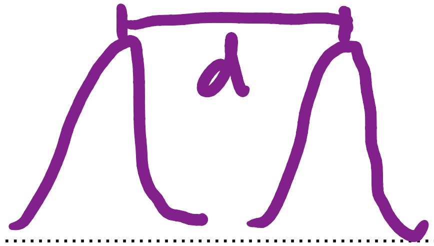
- Question: Is there a change in student test scores based on whether or not they listen to music beforehand?

- Hypothesis: $H_0 : \mu_{music} = \mu_{nomusic}$

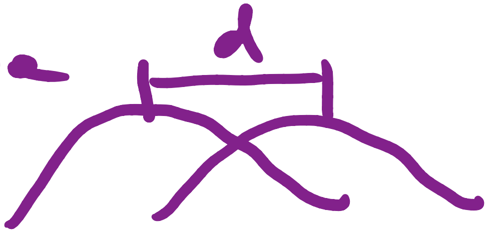
$H_1 : \mu_{music} \neq \mu_{nomusic}$

- Observations:
  - music: [97, 90, 91, 92]
  - no music: [95, 94, 89, 90]

Student's t-test



t-value > t-value



- To perform a t-test, we need to calculate two things:

#1

- t-value: $$\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{\underline{n_1}} + \frac{\sigma_2^2}{\underline{n_2}}}}$$

$$\frac{\text{signal : difference in means}}{\text{noise : spread of data - normalized by \# of observations}}$$

- $\bar{x}_1$ and $\bar{x}_2$ are the sample means
- σ_1^2 and σ_2^2 are the sample variances
- n_1 and n_2 are the number of observations in each sample

Student's t-test

- To perform a t-test, we need to calculate two things:

  - t-value: $$\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

#2

- Then, we'll use this to calculate the p-value using two factors

  ↳ % likelihood that the obseved event is due to random chance

  - degrees of freedom

    ↳ $(n_1 + n_2) - 2$

  - how many tails our test has

    ↳ affect how p is distributed

1) calc t-value

2) calc p-value from the t-value,
deg of freedom, tailedness

3) 0.03
p-threshold: 0.02 - do not reject
0.05 - reject the null
↑
chosen in advance

ICA Question 1: t-test using a spreadsheet

Using the observations given and a spreadsheet (excel or google sheets, for instance), calculate the **t-value** using the below equation:

- t-value: $\dfrac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$

0.2335

- $\bar{x}_1$ and $\bar{x}_2$ are the sample means
- σ_1^2 and σ_2^2 are the sample variances
- n_1 and n_2 are the number of observations in each sample

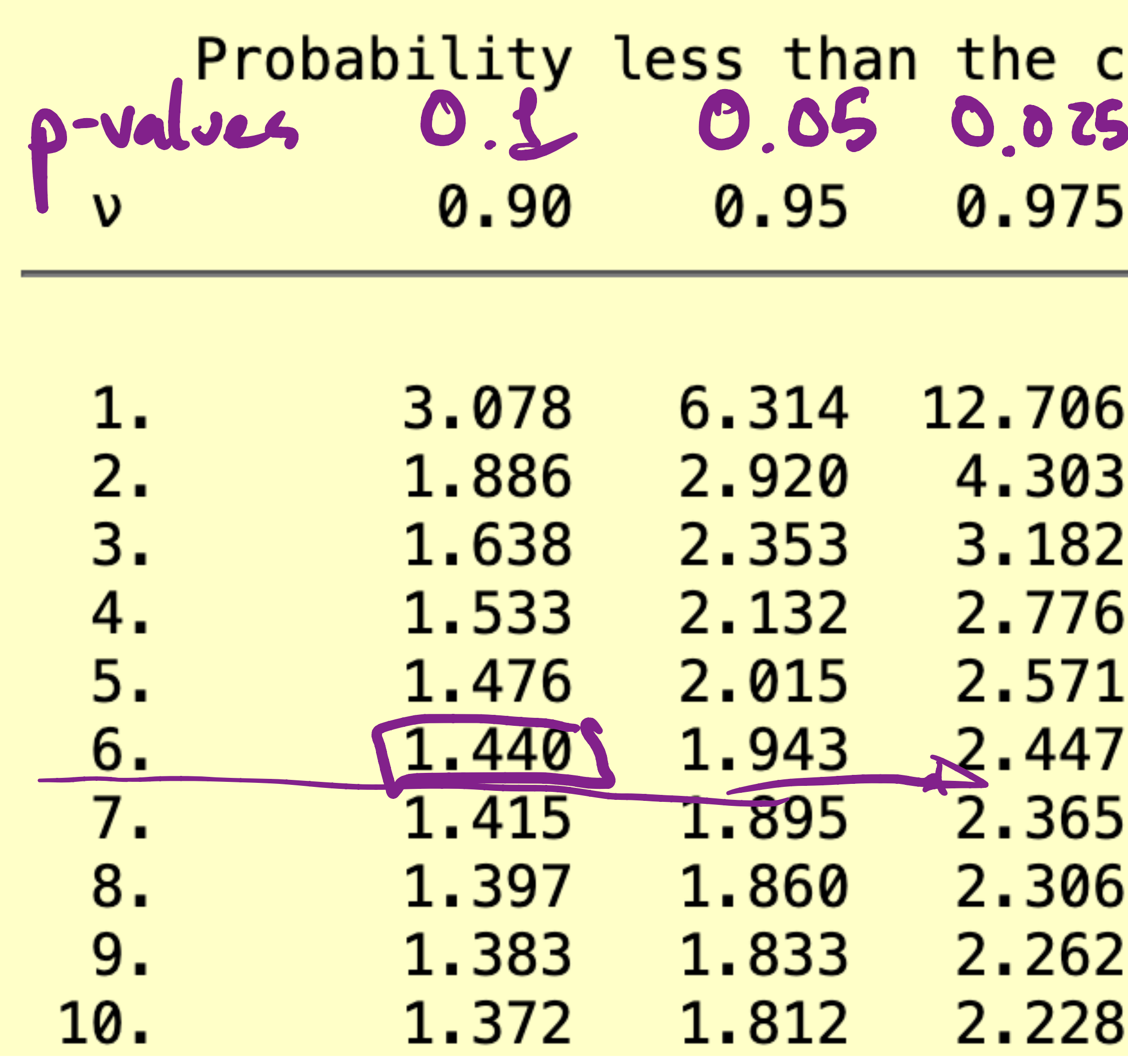
music	no music
97	95
90	94
91	89
92	90

ICA Question 1, continued: t-test using a spreadsheet

Using the observations given and a spreadsheet (excel or google sheets, for instance), calculate the **t-value.**

- Given your t-value, does this test allow us to reject the null hypothesis at a p-value of 0.10? $0.2335 < 1.440$ → No!
- (look your result up against a t-table, such as https://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm)

if t-value had been 1.5, we know p-value is between 0.1 & 0.05



	Probability less than the c		
p-values	0.1	0.05	0.025
ν	0.90	0.95	0.975
1.	3.078	6.314	12.706
2.	1.886	2.920	4.303
3.	1.638	2.353	3.182
4.	1.533	2.132	2.776
5.	1.476	2.015	2.571
6.	1.440	1.943	2.447
7.	1.415	1.895	2.365
8.	1.397	1.860	2.306
9.	1.383	1.833	2.262
10.	1.372	1.812	2.228

these are t-values

ICA Question 1, continued: t-test using a spreadsheet

- What happens to your values if your observations were:

p-value: 0.0027

reject H_0

music	no music
97	**85**
90	**84**
91	**79**
92	**80**

T-tables

1) go to the row w/ $n_1 + n_2 - 2$ degrees of freedom

2) find the value at your p-threshold

3) if your t-value is greater than the listed t-value, reject H_0

ICA Question 1, continued: t-test using a spreadsheet

- What happens to your values if your observations were:

means are equal

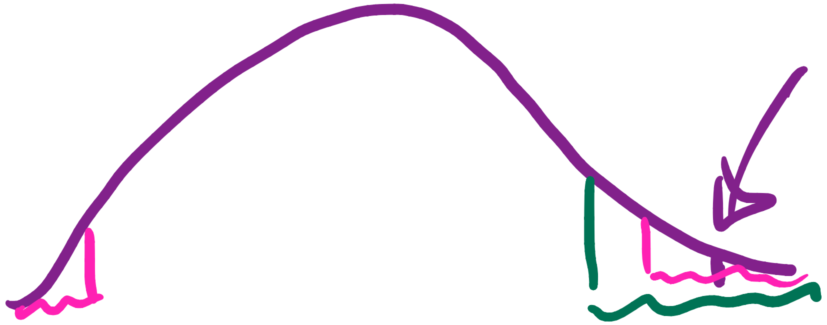
↳ t-value is 0

↳ p-value of 1

you will never reject H_0

music	no music
95	95
90	94
91	89
92	90

Reminders: tailed-ness and tests



96% of the way → 0.04 p-value

norm dist of t-values

↳ how far through the dist is this t-value → p-value

ICA Question 2: t-test using python

Using the observations given and **python**, calculate the **t-value** using the below equation:

- t-value: $\dfrac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$
- $\bar{x}_1$ and $\bar{x}_2$ are the sample means
- σ_1^2 and σ_2^2 are the sample variances
- n_1 and n_2 are the number of observations in each sample

music	no music
97	95
90	94
91	89
92	90

T-tests and errors

- Sometimes, a t-test will produce an error.

- Question: Is there a change in student test scores based on whether or not they listen to music beforehand?

p-val: 0.03 → reject H_0

$\mu_{music} \neq \mu_{no\ music}$

p-val: 0.223 → not reject H_0

T-tests and errors

- Question: Does fertilizer A have an effect on crop yield on my spinach farm?
- Field 1: no fertilizer
- Field 2: fertilizer A

Type 1 errors

- Type 1 error (false positive): test says that you have covid but you don't
- Type 1 error (for a t-test): Reject the null hypothesis when you shouldn't
    - Claim that field two has a higher yield when in fact it doesn't

↳ go spend $$$ on a fertilizer that doesn't work

Type 2 errors

- Type 2 error (false negative) : test says that you don't have covid but you actually do

- Type 2 error (for a t-test): Failing to reject the null hypothesis when you should reject it

    - Claim that both fields have the same yield when in fact they don't

↳ use default fert, end up w/ lower crop yield

↳ spend $$ testing other fert.

Family wise error

- Family-wise error (for t-tests): probability of making one or more false positives (type 1 errors) when performing multiple t-tests

  ↳ reject H_0

- We want to know whether or not using a certain fertilizer increases our crop yield on our spinach farm.

- Each week, we measure the crops in two fields and perform a t-test to determine whether no fertilizer or fertilizer is better.

ICA Question 3: errors on the dance floor

- If I have a Family Wise Error rate of 2.5% and I perform one test every week between my spinach fields:

- What is the meaning of a type 1 error for this context?

↳ claim that the fert. is effective when it isn't

- What is the probability that I get any type 1 errors over the course of one season (13 weeks)?

$$p(\text{no errors}) = .975^{13} = .7195$$

$$p(\text{any}) = 1 - p(\text{no}) \approx .28$$

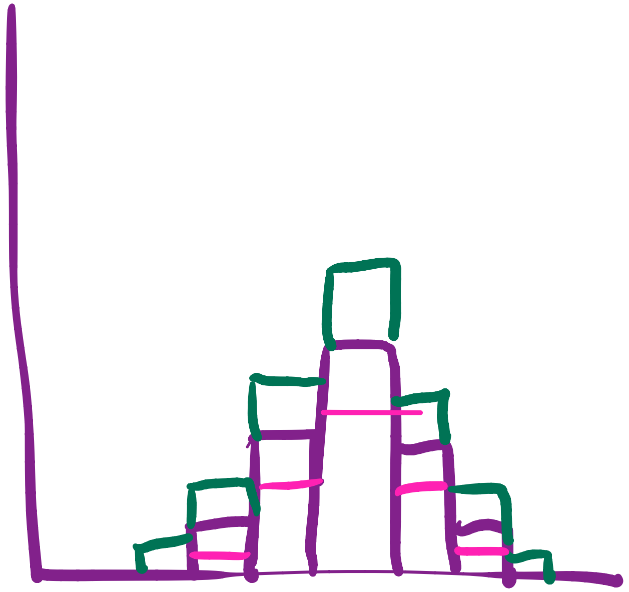
ICA Question 4: playing with significance testing

Go to https://rpsychologist.com/d3/nhst/. Select "Solve for?" -> "d"

What does alpha correspond to?

↳ p-value /threshold

What effect does changing the sample size have?



Tolerances & power & t-tests

- alpha: whatever you set the p value to be less than in order to reject the null hypothesis
  - (higher alpha means more likely to reject the null, more type 1 errors)
  - (lower alpha means more likely to not reject the null, more type 2 errors)
- power is likelihood of detecting the true effect if there is one

  ↳ not covered in this class → tell you sample size needed
- Effect size: how large is the difference between the two populations

t-tests and experimental bias

- Question: Do students at Northeastern enjoy computer science more than students at BU?
- Methodology:
  - Felix surveys students at Northeastern by standing between Snell Library & Engineering and stopping the first 100 students.
  - Felix surveys students at BU by standing in front of the Booth Theater and stopping the first 100 students.

- location → are there more STEM students in one place than the other
- not randomized
- time of day

t-tests and experimental bias

- Places to watch out for:
    - time
    - location
    - selection procedure

harking/p-hacking

- p-hacking is the term in the scientific community that refers to when researchers "go hunting" for statistically significant results after having already performed the experiments

- also known as "**harking**": **H**ypothesis **A**fter **R**esults **K**nown

- Recommended listening (podcast):
  - Maintenance Phase: "School Lunches, P-hacking and the Original "Pizzagate""

harking/p-hacking

- For example, we have have started with the question "Is there a change in student test scores based on whether or not they listen to music beforehand?"

- We surveyed students and asked them:

  - Whether or not they listened to music before a test

- did you study?
- did you eat breakfast?
- when did you get up?

⋮

many more questions

- calculate p-values for all combinations

harking/p-hacking

- For example, we started with the question "Is there a change in student test scores based on whether or not they listen to music beforehand?"

- We surveyed students found that the p-value of "music" vs. "no music" was 0.055, but our threshold was 0.05. What now?

↳ change threshold → 0.06

↳ survey a few more students

Schedule

Turn in **ICA 18** on Canvas (make sure that this is submitted by 2pm!) "test3"

Test 3 is in class on Thursday!

-> note that if an emergency arises that you *must* email me by the end of Thursday for make-up accommodations

Review (virtual) on Wednesday at **2pm** (link on Canvas/Piazza)

Mon	Tue	Wed	Thu	Fri	Sat	Sun
March 28th Lecture 18 - t-tests, errors, experimental bias	**Felix OH Calendly**	**Felix OH Calendly** **test 3 review @ 2pm**	**Felix OH Calendly** **Test 3**			**HW 7 due @ 11:59pm**
April 4th Lecture 19 - chi-square test, multiple comparison correction	**Felix OH Calendly**	**Felix OH Calendly**	**Felix OH Calendly** Lecture 20 - covariance, correlation			

More recommended resources on these topics

- Student's t-test: Youtube, Bozeman Science, Student's t-test
- t-table: https://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm
- p-values from t-scores in python: https://www.statology.org/p-value-from-t-score-python/
- type 1 and type 2 errors: https://www.scribbr.com/statistics/type-i-and-type-ii-errors/#:~:text=In%20statistics%2C%20a%20Type%20I%20error%20means%20rejecting%20the%20null,hypothesis%20when%20it%27s%20actually%20false
- p-hacking:
  - https://podcasts.apple.com/us/podcast/school-lunches-p-hacking-and-the-original-pizzagate/id1535408667?i=1000529447507
  - https://statisticalbullshit.com/2017/07/17/p-hacking/
  - https://www.upgrad.com/blog/what-is-p-hacking/