Comparing 2 Means

Extension to Comparing Two Means

Now that you have a foundation for data analysis (hypothesis testing, confidence intervals), the next several lectures are going to build on that foundation and apply it to different scenarios. But, we will always follow the same steps:

1. Define the Problem
2. Collect the data
3. Explore the data
4. Describe the population (including parameters) with a model
5. Make inference about population parameters (hypothesis test or confidence interval)

Research Objective

Research Question: Are the average number of semesters until graduation for students in Nursing greater than the average number of semesters until graduation for students in FHSS? In other words, do Nursing students take STAT 121 earlier in their college career than students in FHSS?

Population: All BYU students in Nursing or FHSS.

Parameter of Interest:

- We actually have two: μ_1 is the mean number of semesters until graduation for students in Nursing and μ_2 is the mean number of semesters until graduation for students in FHSS.

Sample: A convenience sample of 180 BYU students who are in 121 and completed the student survey AND who are either in Nursing or FHSS.

Are there any issues with this study setup?

More Problem Definitions

Response Variable (y): The number of semesters until graduation.

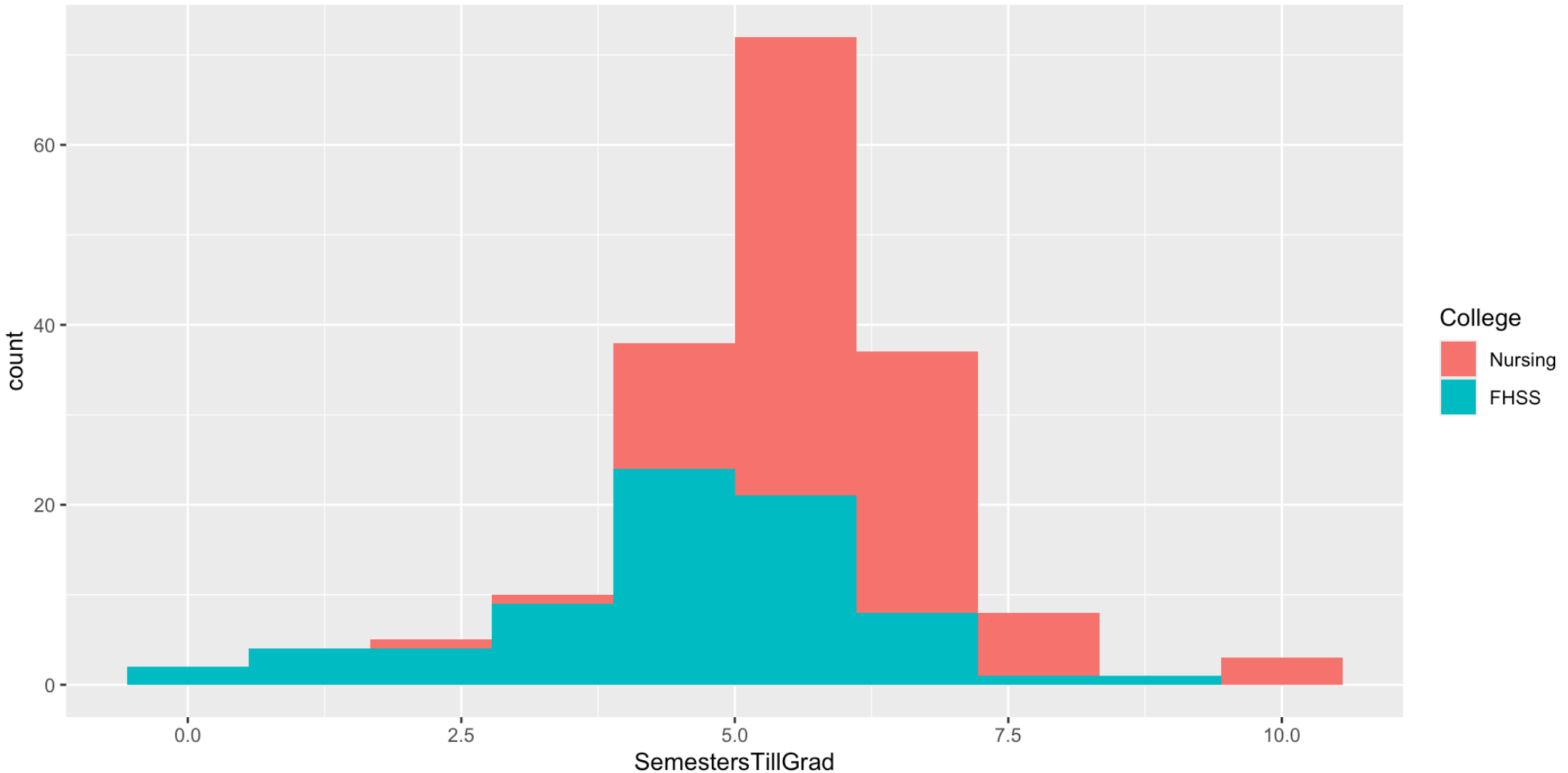
- This is a **continuous quantitative variable** meaning it can be any number (including decimals)

Explanatory Variable (x): The college (either Nursing or FHSS).

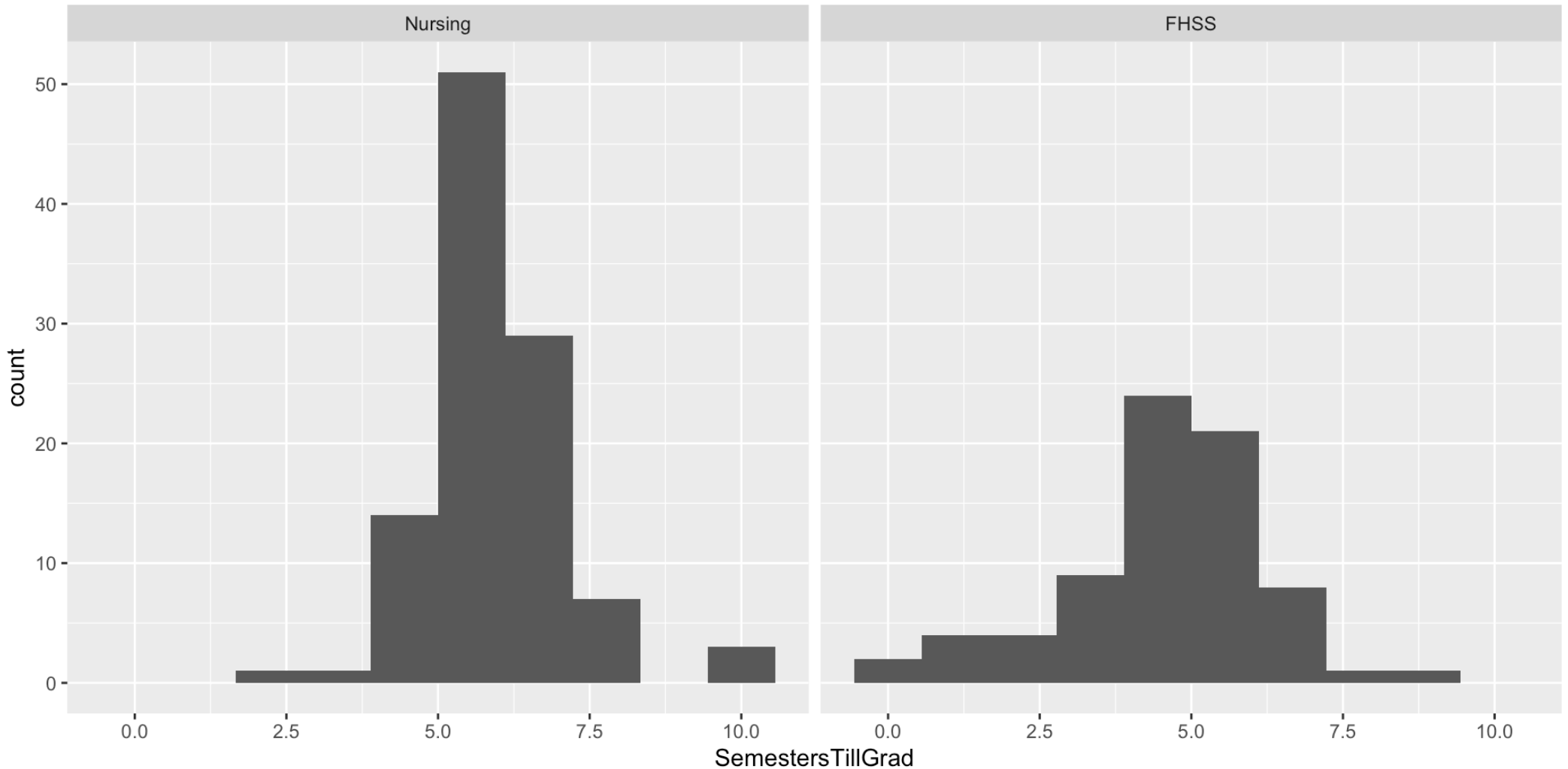
Exploratory Data Analysis (EDA)

Main goal: Compare the distribution of number of semesters until graduation in Nursing and FHSS.

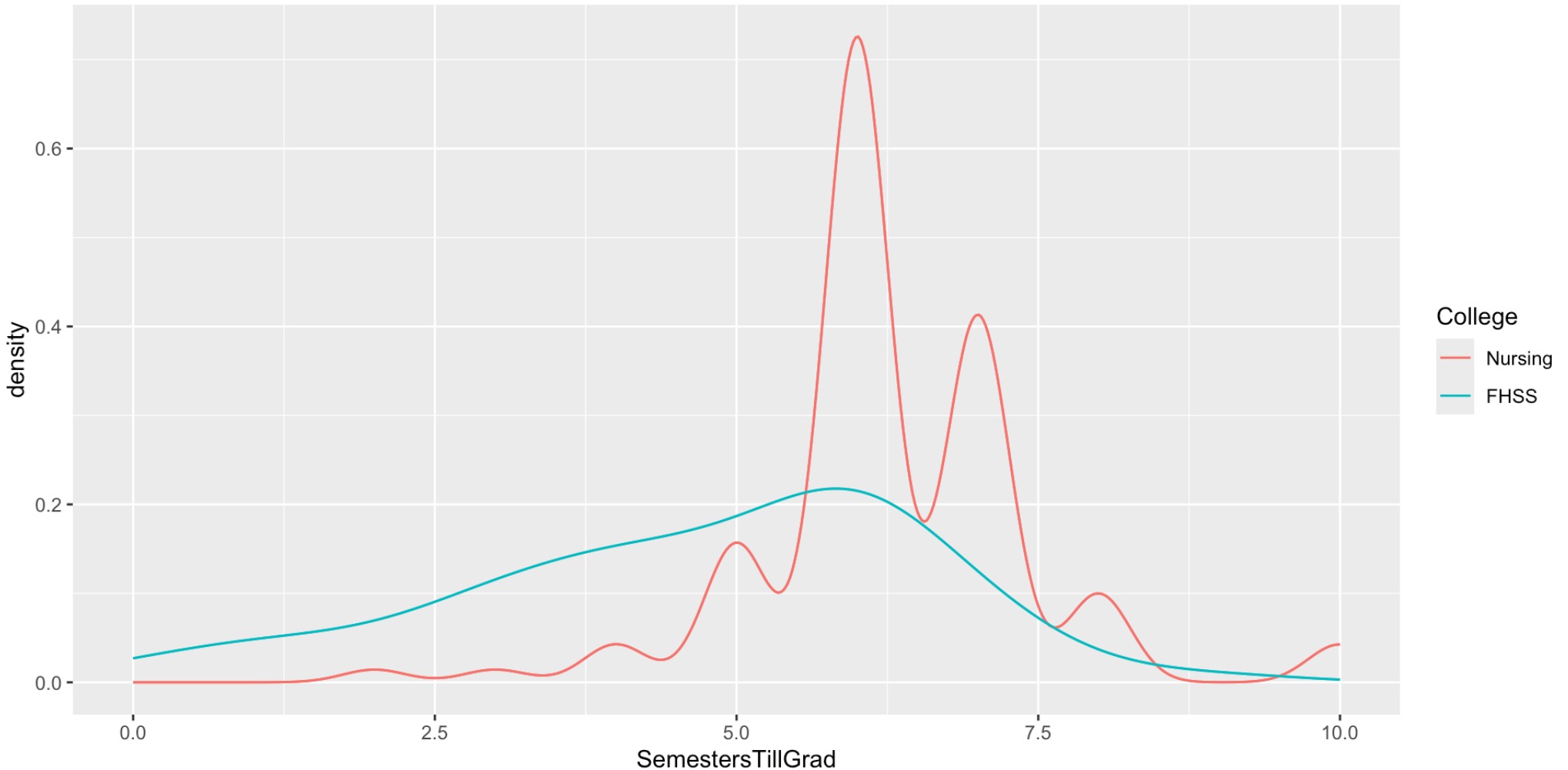
EDA Tool #1 - Histograms



EDA Tool #1 - Histograms

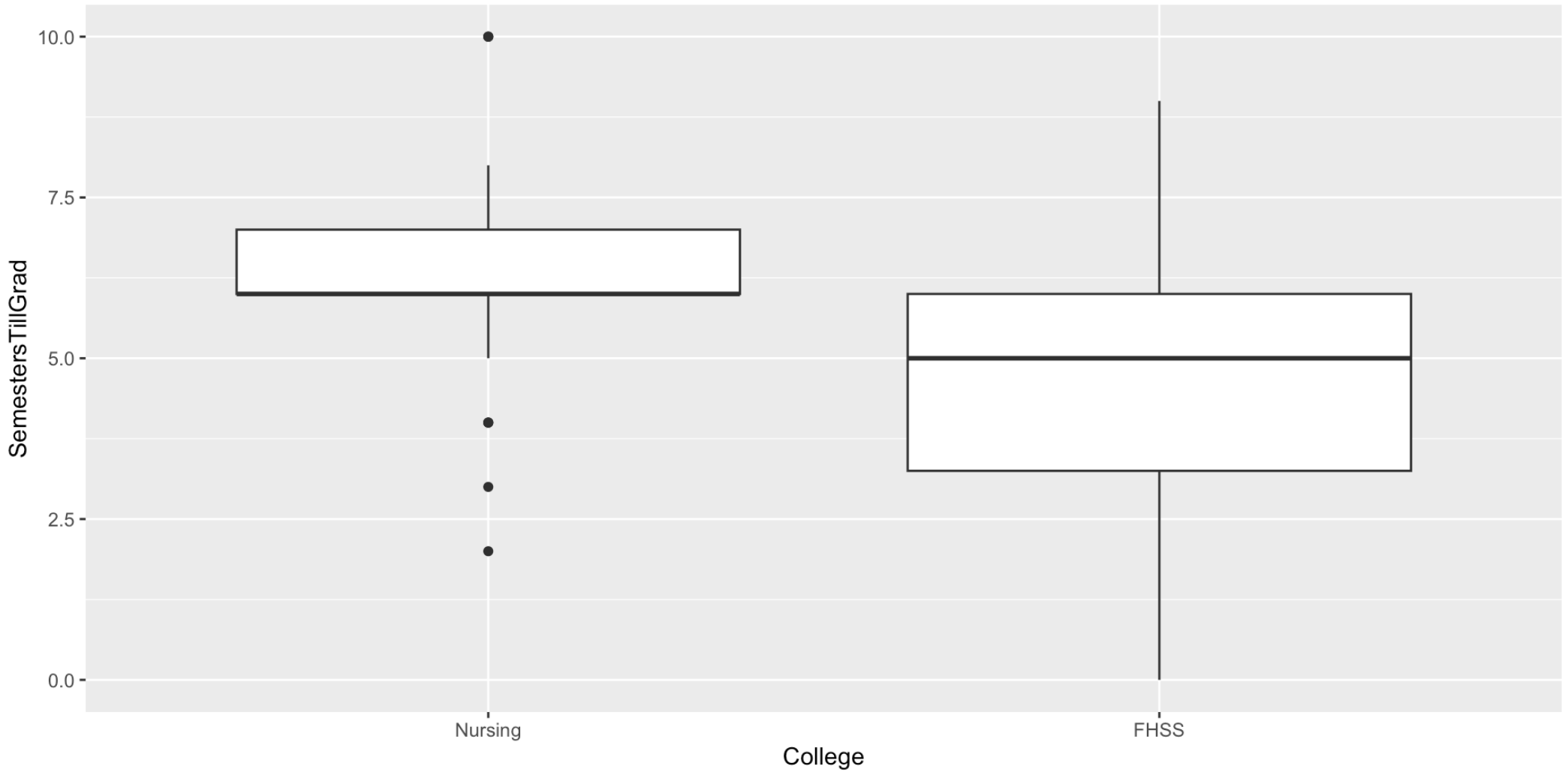


EDA Tool #2 - Density Plots



How would you describe shape, center, and spread?

EDA Tool #3 - Boxplots

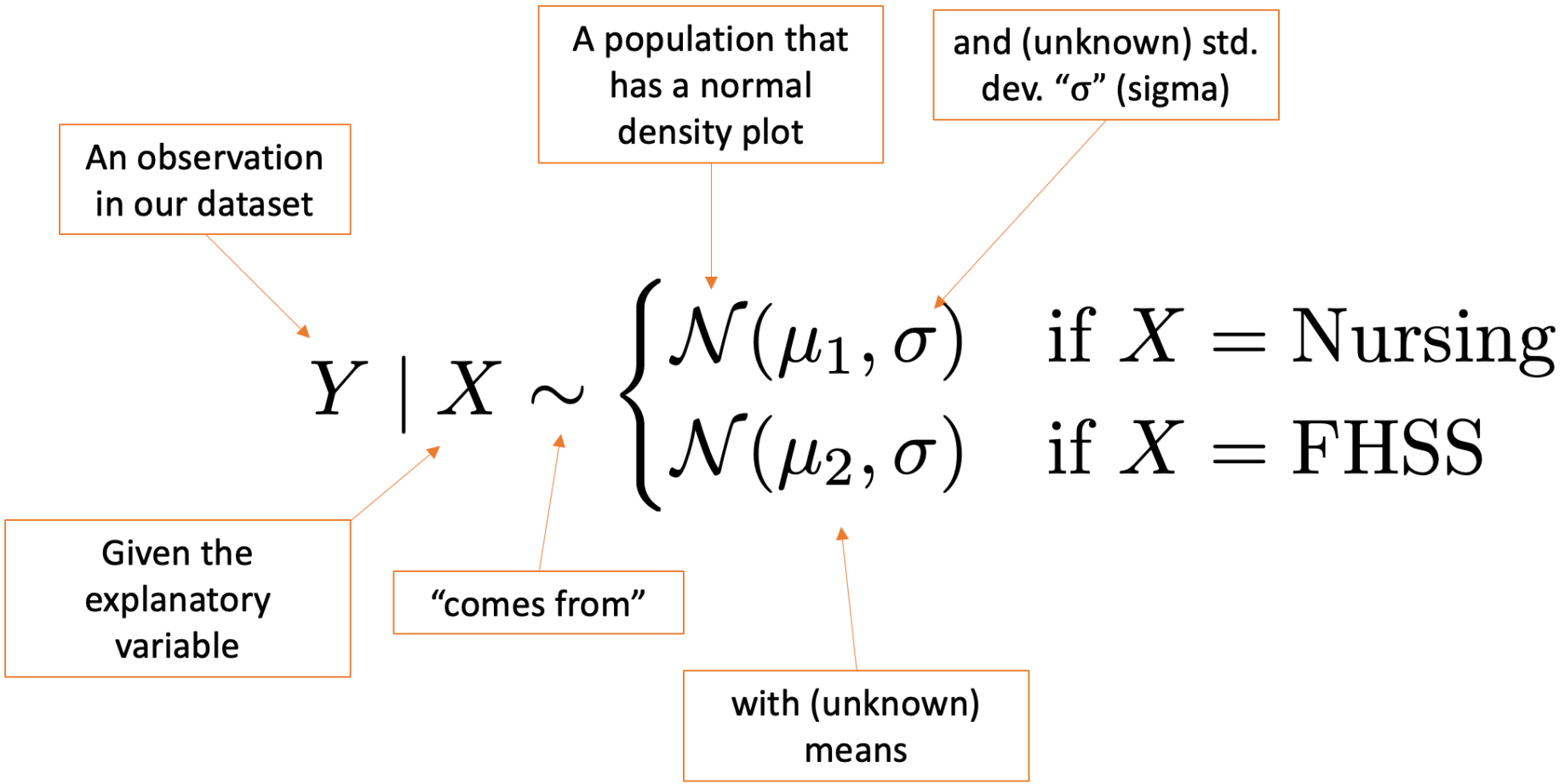


EDA Tool #4 - Numerical Summaries

Numerical Summary Comparison

College	n	Mean	SD	Min	Q1	M	Q3	Max	Skew
Nursing	106	6.29	1.17	2	6.00	6	7	10	0.13
FHSS	74	4.68	1.92	0	3.25	5	6	9	-0.51

Statistical Model



Statistical Model

Important notes about the model:

- Because we want to compare, we are primarily interested in $\mu_1 - \mu_2$.
- Skewness of both groups should be "close" to zero (remember rule of thumb is between -0.5 and 0.5).
- There is a common standard deviation (σ) between the two groups.
  - A good rule of thumb to check if this assumption is valid is that the $\max(s_1, s_2) / \min(s_1, s_2) < 2$.

Point Estimation

The parameters we want to estimate are

- $\mu_1 - \mu_2$
- σ

so we use

- $(\bar{y}_1 - \bar{y}_2) \rightarrow \mu_1 - \mu_2$

Point Estimation

How good of an estimate is $\bar{y}_1 - \bar{y}_2$ to $\mu_1 - \mu_2$?

Theorem: Law of Large Numbers

As the sample sizes (n_1 and n_2) get bigger, the probability that $\bar{y}_1 - \bar{y}_2$ gets closer and closer to $\mu_1 - \mu_2$ increases.

- Important note: how close $\bar{y}_1 - \bar{y}_2$ is to $\mu_1 - \mu_2$ depends on the smaller sample size. If one sample size is really small then $\bar{y}_1 - \bar{y}_2$ might be far away from $\mu_1 - \mu_2$ even if the other sample size is big.

Conditions for Performing Inference

Just like our other analyses, certain conditions need to be met in our data for an analysis to comparing two groups:

1. BOTH groups need to be normal
    - if any group is NOT normal then we need more than 30 observations in that group for the central limit theorem to ensure our inference is valid
2. Standard deviations of the two groups should be approximately the same
    - Use $\max(s_1, s_2)/\min(s_1, s_2) < 2$

- If either of these are NOT met, we should not do the analysis.

Hypothesis Testing

Recall the steps of hypothesis testing:

- Formulate hypotheses
- See if data matches (or doesn’t) match the hypotheses
- Draw conclusions about the parameter

Hypothesis Testing

Research Question: Is the average number of semesters until graduation for students in Nursing greater than the average number of semesters until graduation for students in FHSS?

How would you write the hypotheses?

$$H_0 :$$
$$H_a :$$

Hypothesis Testing

Research Question: Is the average number of semesters until graduation for students in Nursing (group “1”) greater than the average number of semesters until graduation for students in FHSS (group “2”)?

Two ways to write the hypotheses:

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 > \mu_2$$

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_a : \mu_1 - \mu_2 > 0$$

Hypothesis Testing

Step 2: See if the data matches the hypotheses.

- We need (1) a measure of how different what we observed in our sample is from what we expect to have observed if the null hypothesis is true and (2) if our observed difference is "big enough" to reject H_0.
- For (1), we calculate the t-statistic
- For (2), we compare our observed t-statistic to the sampling distribution of t with a p-value.
- Reminder: The sampling distribution of t tells us the values that t that we should see when sampling from the population IF the null hypothesis is true.

Hypothesis Testing - Step 2

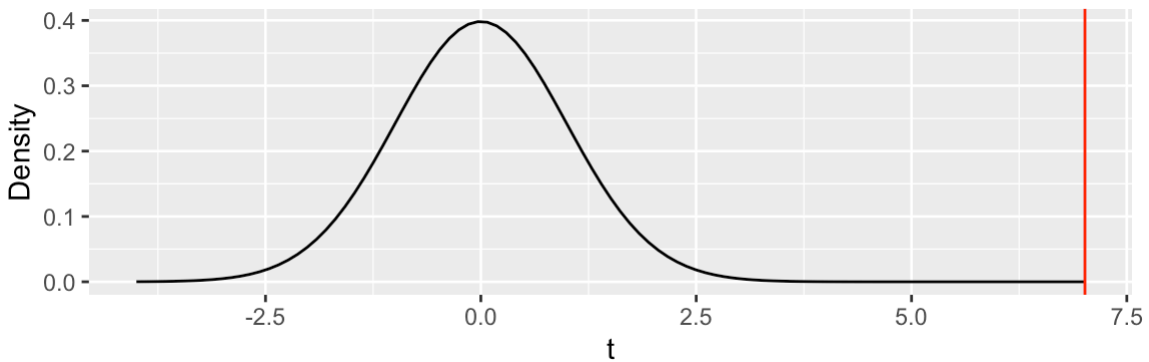
As before, we want to use *standardized* differences between $\bar{y}_1 - \bar{y}_2$ and $\mu_1 - \mu_2 = 0$ (by hypothesis) but, because we have two means, the formula changes to:

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 7.014$$

- Don't worry about the formula (we'll use the tool to calculate it for us)
- How do we interpret this standardized value?
- Our sample difference of $\bar{y}_1 - \bar{y}_2$ = 1.617 (recall Nursing = "group 1") is 7.014 standard errors away from the hypothesized difference of $\mu_1 - \mu_2 = 0$.

Hypothesis Testing

IF the conditions are met for inference and the null hypothesis is true, our t-statistic *should* fall somewhere in this sampling distribution curve:



$t = 7.014$
p-value $= 0$

Reminder:

- p-value is the probability of observing our t or one "more extreme" if the null is true.

Hypothesis Testing

Step 3: Draw a conclusion (use $\alpha = 0.05$)

Given that:

- $t = 7.014$
- p-value $= 0$

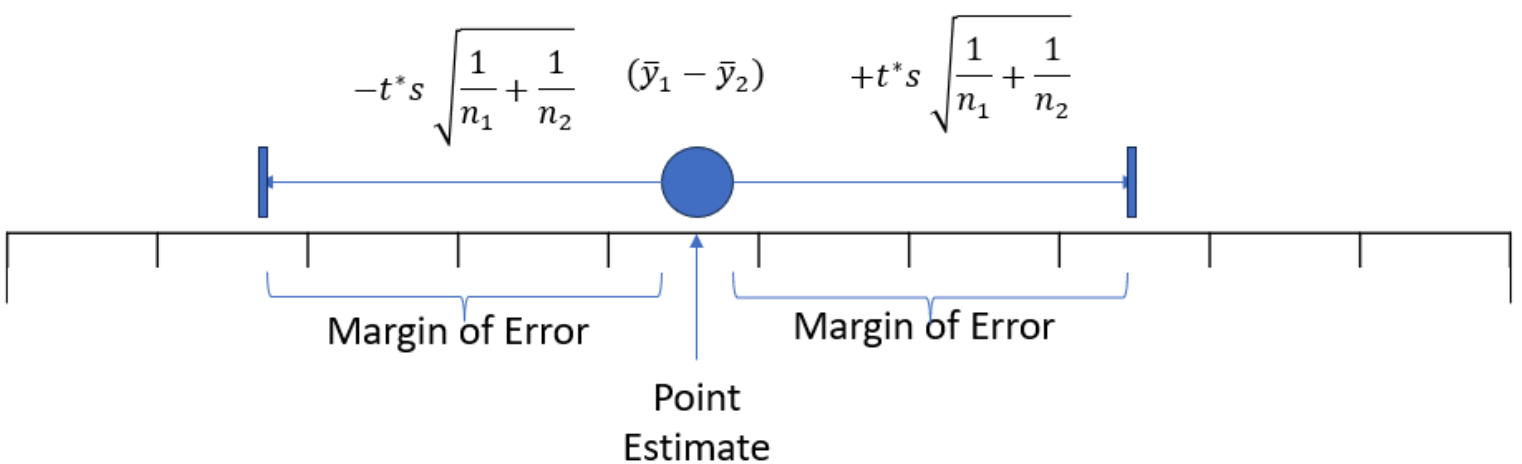
What should we conclude?

- Our data are inconsistent with the null hypothesis so we reject the null and conclude that the average number of semesters until graduation for students in Nursing is greater than the average number of semesters until graduation for students in FHSS.

Confidence Intervals

Hypothesis test conclusions can be vague so lets build a confidence interval. A C% confidence interval for $\mu_1 - \mu_2$ is given by the formula:

$$(\bar{y}_1 - \bar{y}_2) \pm t^{\star} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$



Confidence Intervals

A 95% confidence interval for $\mu_1 - \mu_2$ is (1.162, 2.072). How do we interpret this interval?

- We are 95% confident that the difference in the the average number of semesters until graduation for students in Nursing minus the the average number of semesters until graduation for students in FHSS is between 1.162 and 2.072.

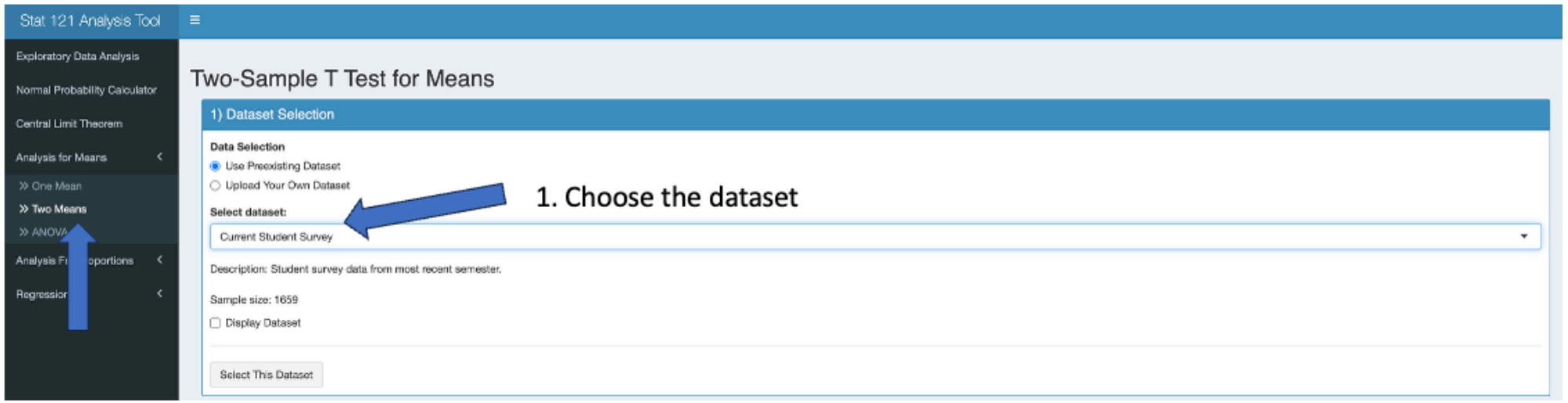
Things to Keep in Mind

1. Keep in mind key terms of hypothesis tests:
    - What would constitute Type 1 and Type 2 errors for our analysis of the number of semesters until graduation?
        - Type 1 = concluding Nursing has greater average time until graduation than FHSS when, in, fact they have the same time.
        - Type 2 = concluding Nursing has the same average time until graduation as FHSS when, in, fact they have greater time.
    - Are our results “statistically significant”?
        - Yes because we rejected H0
    - Are our results “practically significant”?
        - Maybe (which way would you argue?)

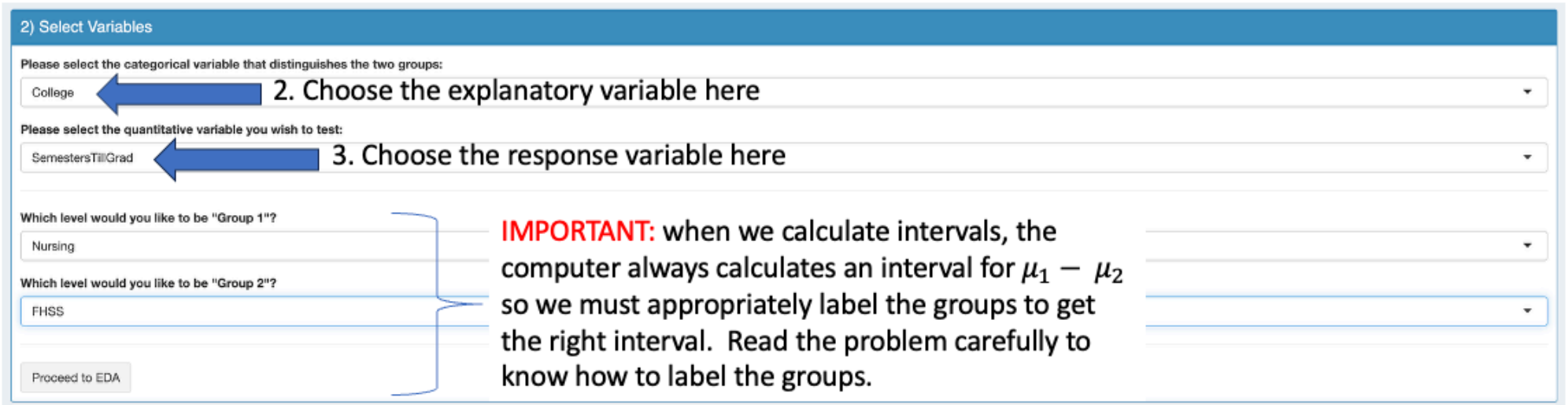
Things to Keep in Mind

2. Keep in mind key terms of confidence intervals:
    - Margin of error
        - Interval = Point Estimate $\pm$ Margin of Error
    - Effect of sample size on margin of error
        - As sample sizes go up, margin of error goes down.
    - Effect of confidence level on margin of error
        - As confidence level goes up, margin of error goes up.

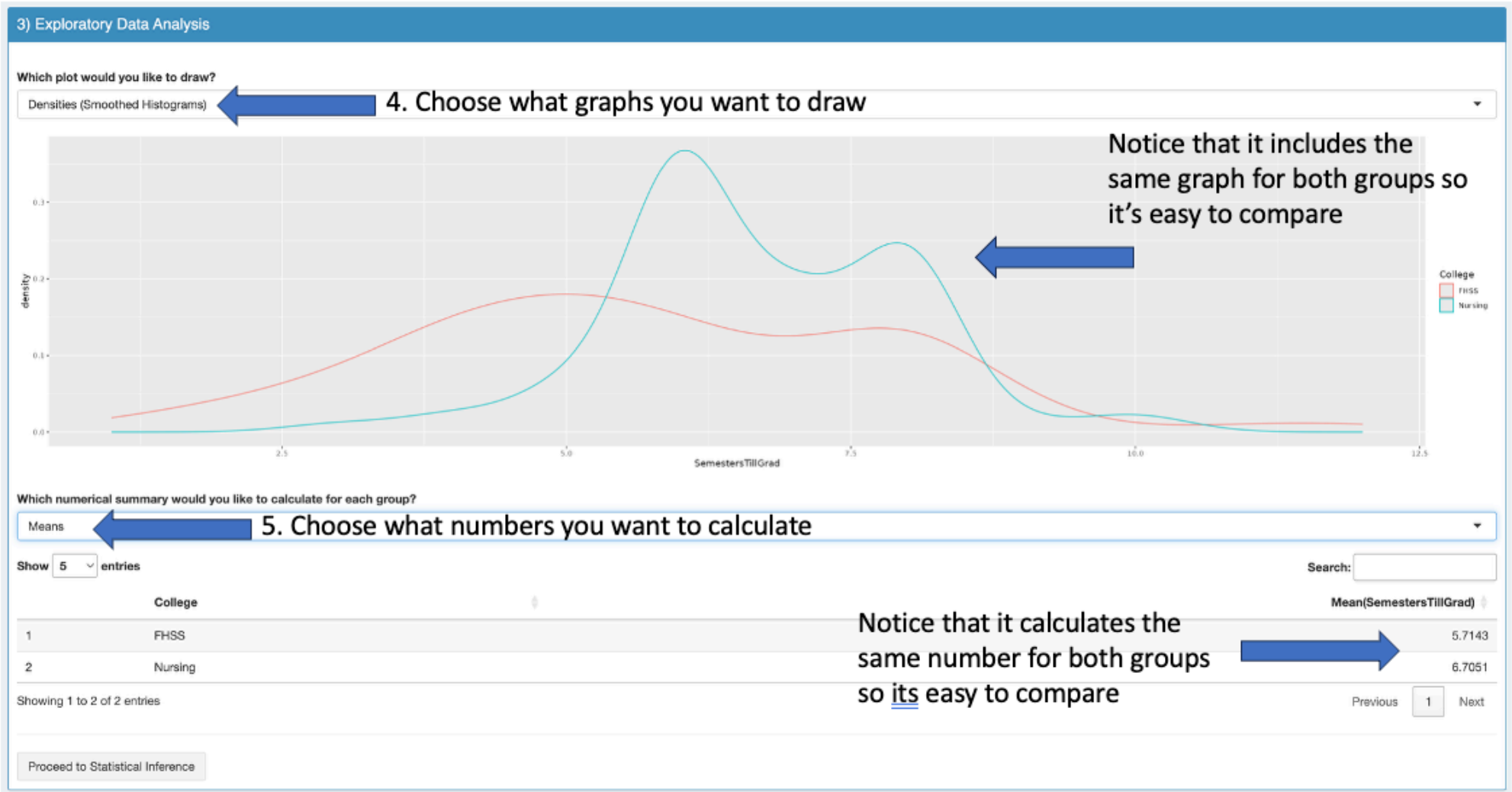
Using the Tool for EDAs



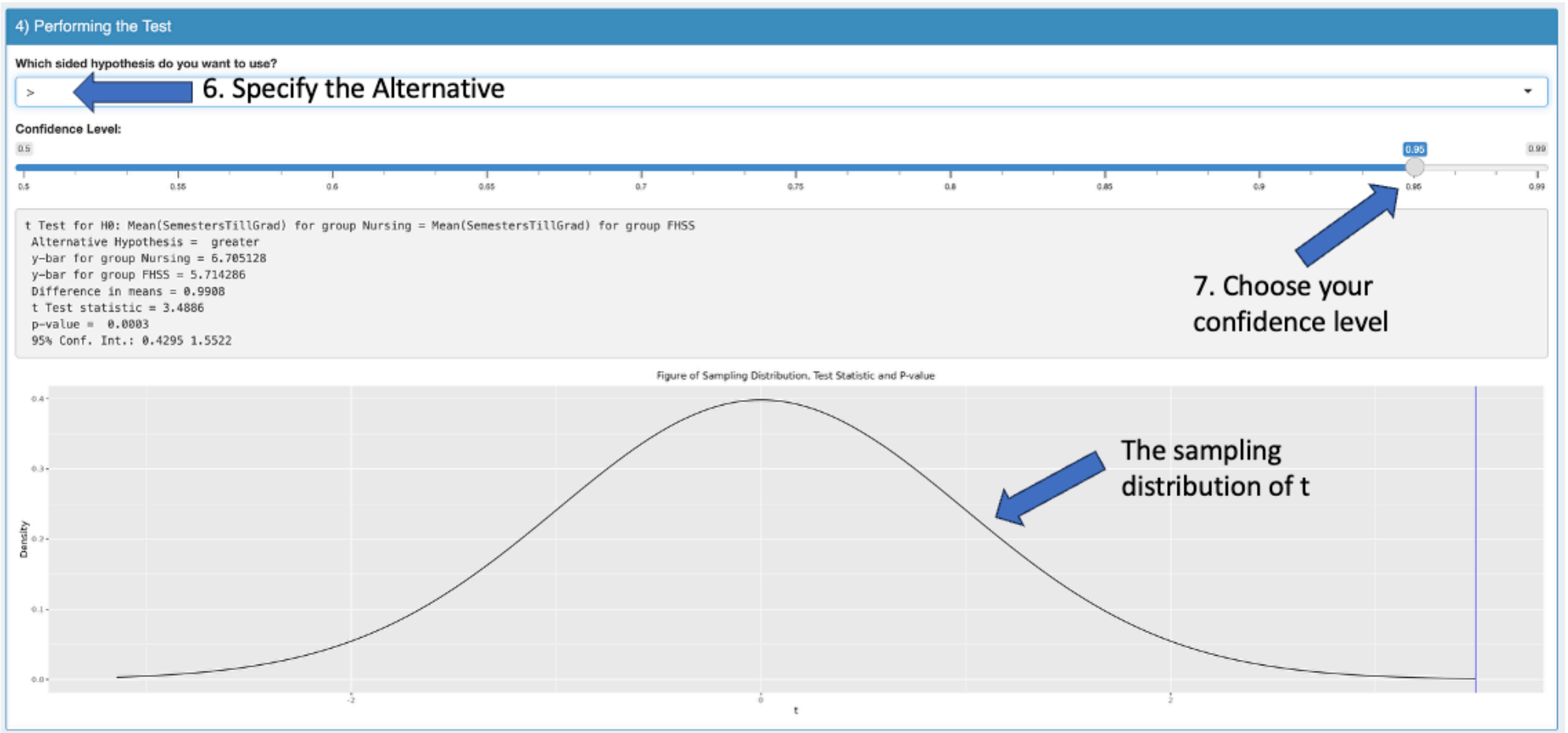
Using the Tool



Using the Tool for EDAs



Using the Tool



HW This Unit

1. Inmate stress - does putting inmates in isolation affect their mental health?
2. Going to college - are there differences in grades based on peoples interest in college?
3. Happiness - do different regions of the world have different happiness levels?

Key Terminology

- A/B Testing
- Sampling distribution of $\bar{y}_1 - \bar{y}_2$
- t-distribution
- Two means model
- Exploratory data analysis for two groups
- Interpreting a confidence interval
- Two-sample t-test
- Equal standard deviation between groups