# Open-ended Practice Test (similar to Quiz 3)

Originally these data were collected by two separate research groups. One group was a think tank (New Economic Foundation) interested in measuring the happiness of different countries (they compiled the data seen in **HappyPlanetIndex**). The other was the World Health Organization interested in how much and what kind of alcohol different countries drank for the purpose of understanding health (the data in **drink**). Here I’ve provided the code to join these two data frames together so that each country’s happiness is aligned with their drinking data.

library(fivethirtyeight)

drinks$Country <- drinks$country

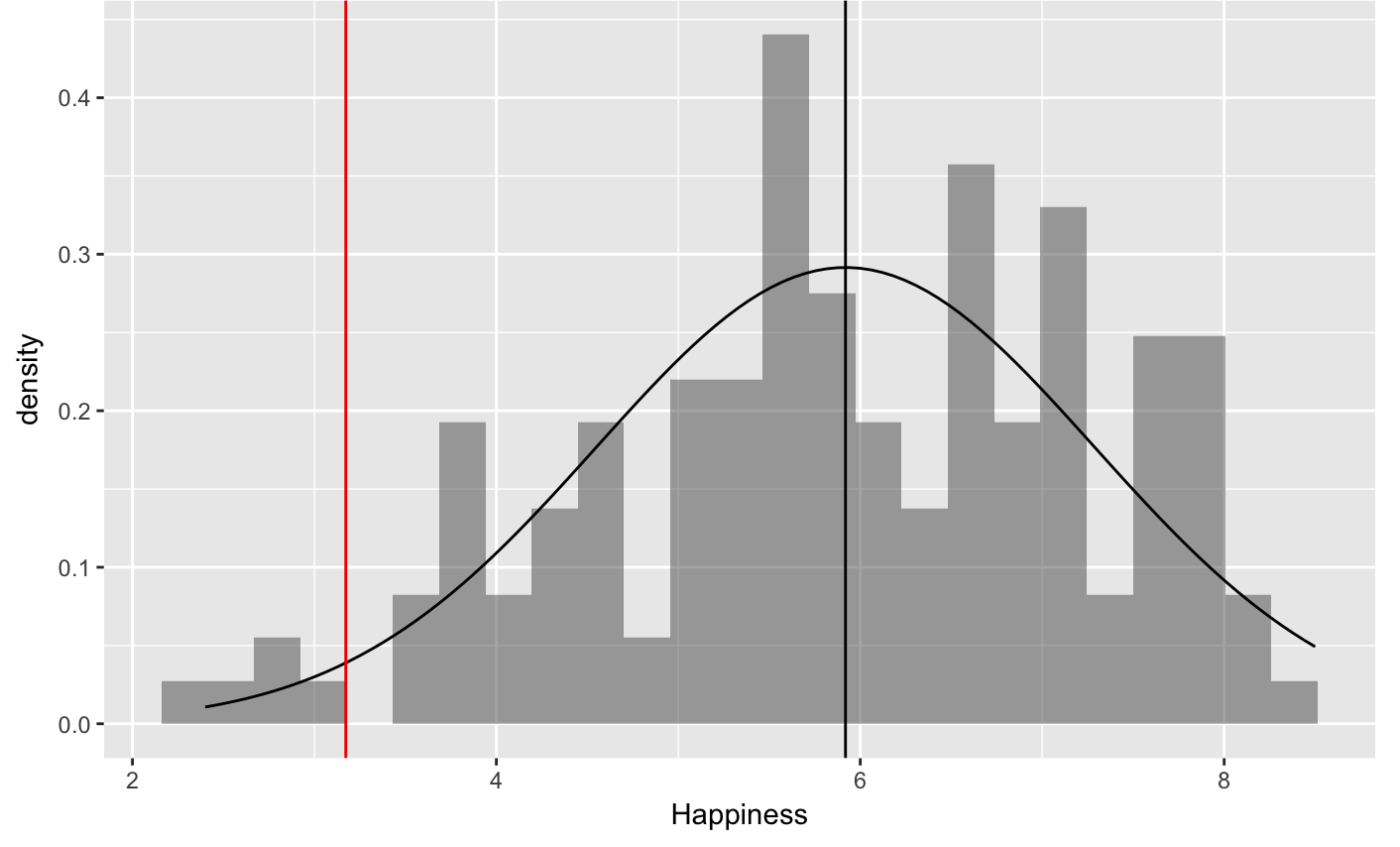
happydrinks <- left\_join(HappyPlanetIndex, drinks, by = "Country")

Here is more information about each data frame:

https://www.rdocumentation.org/packages/Lock5withR/versions/1.2.2/topics/HappyPlanetIndex

https://www.rdocumentation.org/packages/fivethirtyeight/versions/0.4.0/topics/drinks

1. How might someone use this data to explore whether drinking alcohol explains variation in happiness? Give an example of a pattern that would support their idea. (Note: this pattern does not have to be true in the data)
2. Which of the variables in the new data frame **happydrinks** are quantitative variables?
3. What are the cases in this data frame? How many are there?
4. This data frame is terribly labeled. Label **Region** appropriately using R code. The regions are numbered from 1 to 7 and correspond to these labels:
   * + 1. East Asia
       2. Former Communist Countries
       3. Latin America
       4. Middle East and North Africa
       5. South Asia
       6. Sub-Saharan Africa
       7. Western Nations
5. Write R code to create a histogram that shows you the variation in **Happiness**.
6. How happy are you today? Rate on a scale of 1 - 10 (10 being very happy). Can we put this data into this data frame as a new row? Why or why not?
7. Just from looking at the histogram, estimate the standard deviation using the histogram. Then calculate the standard deviation using R.
8. If you use the distribution of **Happiness** in the **happydrinks** data frame as a probability model, what is the likelihood that a country will have a happiness rating of 3 or lower?
9. Using the normal distribution to model error, what is the likelihood that a country will have a happiness rating of 3 or lower?
10. In the figure below, which part represents the probability that a country would have a **Happiness** score of 3 or lower (depicted in red)? Which part represents the z score for a happiness rating of 3?



1. Find the z-scores for all the countries. What’s Guatemala’s z-score? What does that number mean? What is Bangledesh’s z-score? What does that number mean?
2. Without looking at the data, write a story of the DGP arguing that one kind of alcohol (either wine, beer, or spirits) would make a country more or less happy. Write your story of the DGP as a word equation.
3. Write R code to create three different visualizations to see whether variation in wine, beer, or spirits drinking helps explain the variation in happiness.
4. Let’s look at spirits drinking. It almost seems like there are three types of spirit drinking countries -- those that don’t drink any, those that drink a medium amount of spirits, and those that drink a lot of spirits. Create a new categorical variable (call it **spirits3group**) that splits the countries up into these three groups. Label the groups. Write the R code here.
5. Create at least two different visualizations to see whether **spirits3group** explains any of the variation in **happiness** ratings.
6. There are 9 countries that do not have a value for spirit\_servings (it’s NA). Write R code to put only the countries that have a value for **spirit\_servings** into a new data frame called **happy** data frame.

From Question 17 and on, we will be using the **happy** data frame.

1. Re-make the visualizations without these NA countries. Does spirits (**spirit3group**) seem to explain some of the variation seen in happiness? How would you write this as a word equation?
2. Based on these visualizations, can we conclude that drinking spirits causes people to be more happy?
3. Make a histogram of **Happiness** by **spirit3group**. Put density plots over the histograms. Where on the density histogram would you look to see evidence of within group variation in **Happiness**?
4. Where on the density histogram would you look to see evidence of between group variation in **Happiness**?
5. In our explorations so far, what is the outcome variable and what is the explanatory variable?
6. Why wouldn’t we use tally() in order to explore variation in **Happiness**?
7. If we wanted to make a jitter plot with happiness and spirit groups, by convention, what should go on the y-axis?
8. Let’s say we are very interested in happiness of children in the world. What would be the problem with using this data set to explore this idea?
9. If you created a histogram of **Happiness** but added bins = 10, you would get a different looking histogram. Why is that?
10. If we are going to try and explain variation in Happiness with **spirit3group**, how would you write this model in GLM notation?
11. In the **spirit3group** model, what would does each part of the GLM notation represent? Y\_i, b\_0, etc.
12. Use R to find the best fitting estimates for the **spirit3group** model. What does each number mean? What part of the GLM notation would each number go with?
13. What would the empty model predict for each country? What would the **spirit3group** model predict for each country?
14. In the **spirit3group** model, how is the error from the model calculated? In the empty model, how is the error from the model calculated?
15. Is the **spirit3group** model better at explaining variation than if we had just used an empty model? How much error is there leftover after the empty model? How much error is leftover after the **spirit3group** model? Which one will have more error leftover?
16. Can a complex model (like **spirit3group** model) ever have more error left over than the empty model? Why or why not?
17. What code could you use to get the PRE of the **spirit3group** model? What does the PRE mean?
18. Why is the degrees of freedom for the **spirit3group** model equal to 2?
19. Do these results show that not drinking spirits causes some countries to be less happy?
20. Which of the following is a good reason to consider F instead of PRE to quantify how much variation in **Happiness** was explained by **spirit3group**?
21. If we created a model called **spirit12group**, would it do a better job of “explaining variation” than **spirit3group**? Which model would have a larger PRE? Which would have a larger F? (Run code to see.) Why does one model have the larger PRE and the other model have the larger F?
22. Could you write the GLM notation for the **spirit12group** model?
23. What is sum of squares model, sum of squares error, and sum of squares total?
24. Which would be a better explanatory variable for Happiness: **beer3group** (splitting the countries up into three equal sized groups based on beer\_servings) or **spirit3group**? Run analyses in R to make your case.
25. If you get the supernova table for a model explaining variation in happiness with **beer3group**, you will get the same sum of squares total as you did with the **spirit3group** model. Why?
26. Calculate the variance of **Happiness**. What does this represent?