Amir Arya
Kenneth Nguyen

Clinical Data Partner Activity

1. Define the following: categorical variable, discrete variable, continuous variable. Provide examples of each.

- Categorical variables are variables that assign category names to each value. They are stored into a factor in RStudio. Examples of this include alive/dead and cancer type.
- Discrete variables are whole-number variables that can't take on fractional/decimal values. Examples of this include age (years) and days since treatment.
- Continuous variables are variables that can take on fractional/decimal values. Examples of this include weight.

2. Look at the different column names of the clinical data frame. Choose one that is interesting to you and your partner. Ensure that there are not too many NAs in this column by using is.na(clinical$COLUMN_NAME). Remember that in coding, TRUE is equal to 1 and FALSE is equal to 0. You can then use the sum() function to find how many TRUEs exist. Which variable have you chosen?
We chose the "lymph_node_examined_count" variable.

3. Google your chosen variable. How is your variable measured or collected? Is your variable categorical, discrete, or continuous?
"Lymph_node_examined_count" is determined by the number of lymph nodes that are examined by a medical professional in the clinical setting. This is a discrete variable.

4. Find two research articles that mention your clinical variable. Provide the links and a brief description of the findings.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7607195/
This research article shows that in the clinical setting, too many lymph notes are examined for lower-stage breast cancer (T1) while too few lymph nodes are examined for higher-stage breast cancer (T2 and T3). This finding could help improve breast cancer diagnosis since it is currently difficult for medical professions to determine how many lymph nodes need to be dissected in order to fully study a patient's breast cancer (which can lead to non-treatment of ininspected lymph nodes or excessive invasive procedures for breast cancer patients).

https://onlinelibrary.wiley.com/doi/full/10.1111/1759-7714.13056
This article investigates the effect of lymph node examined count on accurate staging and survival of resected esophageal cancer. The study found that higher numbers of lymph nodes examined (LNEs), compared to accurate staging and survival, was linked to LN metastasis which led to better CSS, cancer specific survival.

5. Look at the different column names of the clinical.drug, clinical.rad, and clinical dataframes. Choose a variable from one of these data frames. Ensure there are not too many NAs (there will likely be more NAs in the drug and radiation dfs than in the patient data, don't worry about it too much). Which variable have you chosen? Provide a brief description of the variable.

We chose the variable “race_list” from the “clinical” dataframe. “race_list” is a categorical variable that identifies the race type of each patient in the dataset. We had to convert from a vector to a factor in order to plot the graphs.

6. Scientists generate hypotheses before experimenting or exploring data. Generate three hypotheses: (1) Relate your variables to each other, (2) Relate your first variable to survival in breast cancer, (3) Relate your second variable to survival in breast cancer.
Hypothesis 1. Disadvantaged groups are more likely to exhibit higher lymph node counts
Hypothesis 2. Patients with higher lymph node examined counts are less likely to survive their breast cancer (due to having more-developed cancer)
Hypothesis 3. The more disadvantaged groups of people in the data set would have lower survival rates when diagnosed with breast cancer.

7. Summarize what you learned from your graphs! What is the significance of these findings? (Answer this question after you finish your analyses)

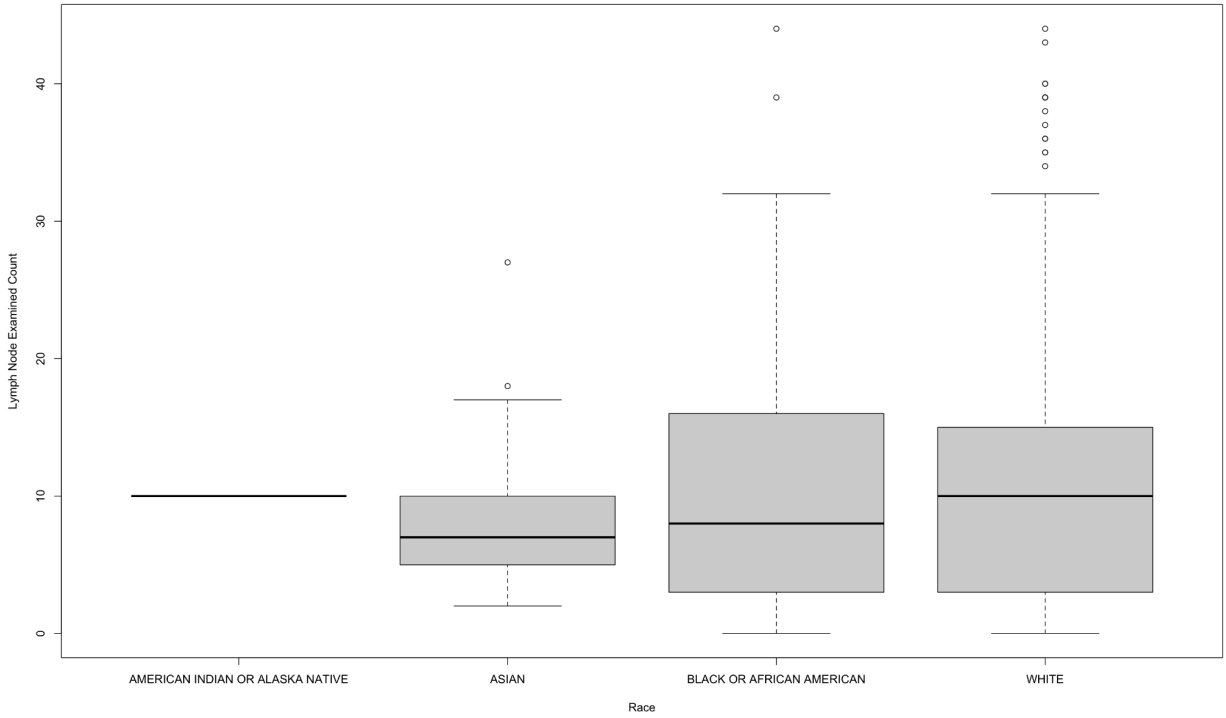
When it comes to lymph node examined counts across different groups of people, the data shows that there is not really a correlation between the disadvantaged groups having more lymph nodes examined since the boxplots have significant overlap. This disproves hypothesis 1 and suggests that lymph node counts do not differ significantly across different race groups.

In hypothesis 2, we inferred that those with higher lymph node examined counts were less likely to survive breast cancer. However, when analyzing the data, we see that our hypothesis was disproven ($p=0.46$) as the amount of lymph node examined count does not have a significant effect on the survivability rate. The rate of survival is consistent regardless of the lymph node examined count.

Comparing the data from the race list and survival data, we were not able to find evidence supporting hypothesis 3 ($p = 0.74$). While those who are thought to be more disadvantaged would historically be expected to exhibit lower survival rates, our data does not reflect this. In fact, we also noted that the white group visually appeared to have lower survival rates than other groups.

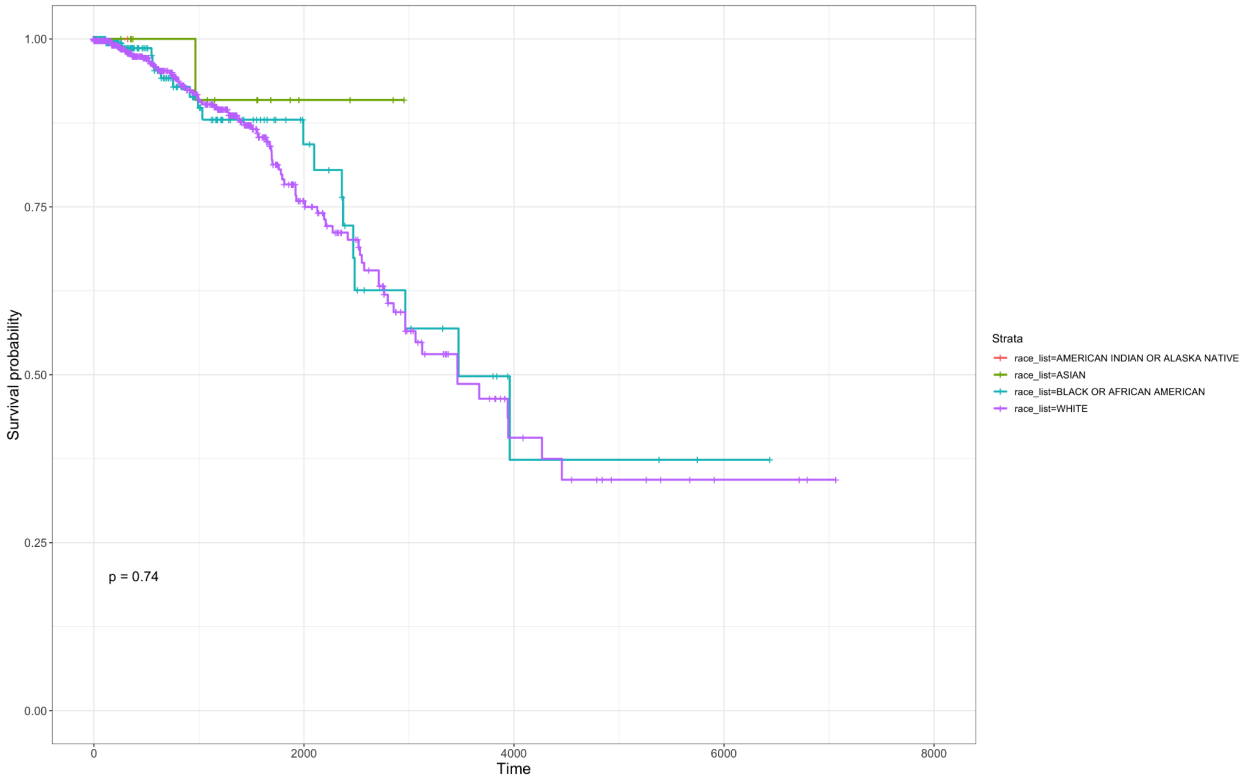
Overall, our graphs disproved all three of our hypotheses and demonstrated that there is no major correlation between the variables survival, lymph node expected count, and race.

Coding Activity:
1. Perform an analysis looking at the two variables that you chose. First brainstorm and sketch out a plot that contains both variables. Feel free to get creative, if you are struggling, feel free to ask for ideas! (Helpful functions/packages: plot(), hist(), boxplot(), pairs(), ggplot2 package + associated functions)
○ TIP: Sometimes it can be hard to plot a continuous variable with another variable. You can convert a continuous variable to a categorical one. For example, we previously defined age < 50yrs old as “Young” and age ≥ 50 yrs old as “Old.” Here we have converted age, a continuous variable, to young and old, a categorical variable.

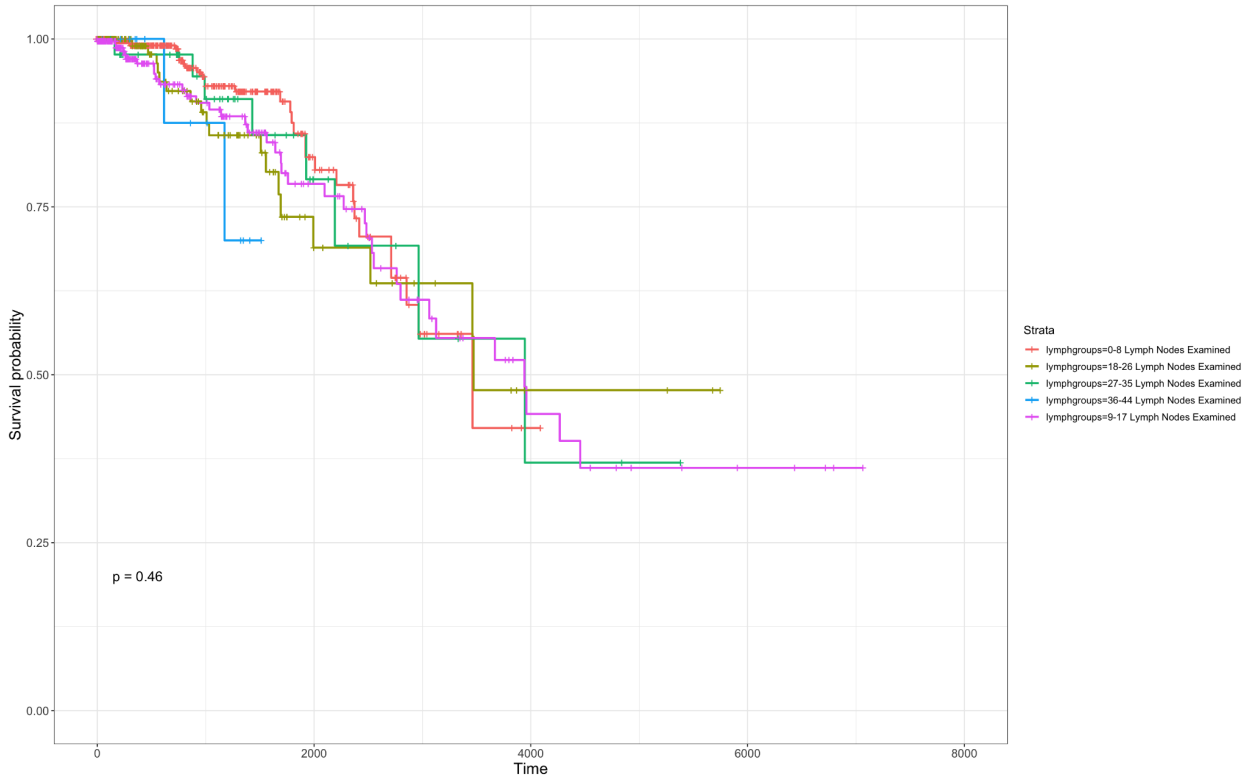


2. Perform a survival analysis, following the steps of the clinical data tutorial with the first variable.

○ As with the previous tip, the survival analysis needs a categorical variable. If you have a continuous variable, use an ifelse()statement to create a new column with a categorical version of the variable.



3. Repeat with the second variable. Note that for drug and radiation data, there might be many categories in one column. Try to keep the KM plot simple by limiting the data to ~5 stratification categories.
Groupings: 0-8 lymph nodes examined, 9-17 lymph nodes examined, 18-26 lymph nodes examined, 27-35 lymph nodes examined, and 36-44 lymph nodes examined



4. For an extra challenge (optional) perform a survival analysis where survival is stratified by both variables.
5. Save your plots and write any data frames you used to your local machine.