Written Activity

1. **Define the following: categorical variable, discrete variable, continuous variable. Provide examples of each.**
    a. **Categorical variable:** data types that can be separated into qualitative categories
        i. Tumor tissue site
        ii. Gender
        iii. Race
    b. **Discrete variable:** a countable amount in a limited/finite interval of time
        i. Appointment Date
        ii. A person's age in years
        iii. Lymph node count
    c. **Continuous variable:** infinite possible values, can take on any value within a range
        i. Age, days until last followup
2. **Look at the different column names of the `clinical` dataframe. 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?**
    a. From the clinical file we selected the $lymph_node_examined_count variable. There are 139 NA's in this variable. It describes the number of positive lymph nodes are in the patient.
3. **Google your chosen variable. How is your variable measured or collected? Is your variable categorical, discrete, or continuous?**
    a. Lymph Nodes are counted by setting a threshold for nodal status (how big the lymph node has to be to be counted) and then counted across the patient by pathologists. Additionally, sentinal node biopsies are conducted to detect if lymph nodes contain cancer.
    b. Discrete variable - set number of lymph nodes present at a time

4. **Find two research articles that mention your clinical variable. Provide the links and a brief description of the findings**.

    a. Lymph Nodes: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5506109/pdf/jtd-09-06-1531.pdf

    The value of positive lymph nodes ratio combined with negative lymph node count in prediction of breast cancer survival

        i. Lymph node ratio (LNR) helps physicians predict prognoses for invasive breast cancer. Negative lymph node (NLN) is another prediction factor. High NLN counts are associated with improved prognoses whereas LNR helps develop cutoff points for developing prognoses (ex. Categorizing patients into low, intermediate, and high risk). Combining the two provides physicians with a better way to tailor prognoses between populations that differ (ex. Cancer stage, race, age, etc.).

    b. Cancer Stage: https://www.sciencedirect.com/science/article/abs/pii/S107275150000257X

    Stage 0 to stage III breast cancer in young women

        i. The prognosis for breast cancer survival varies based on a variety of factors including how early the cancer is identified (what stage it is), age, race, and certain receptors. Patients diagnosed before age 36 are more likely to survive aggressive treatment than those who are 36 and older. However, diagnoses at a younger age are correlated with more aggressive cancers and higher rates of recurrence.

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.**

    a. From the clinical file we selected the $stage_event_pathologic_stage variable. There are 0 NA's in this variable. It describes what stage of breast cancer the breast cancer patient is at.

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.

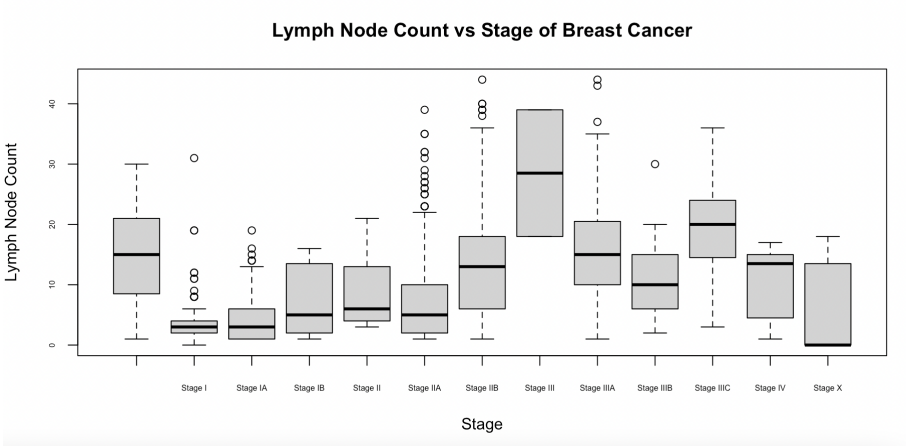
1. If a patient is at a higher breast cancer stage, they will have increased lymph node counts.
2. If a patient has a higher breast cancer stage, then they will have a decreased chance of survival.
3. If the patient has a higher lymph node count, they will have a decreased chance of survival.

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

a. Graph Comparing Lymph node Count:

i. Stage 3 has the highest median lymph node count. We can also see that on average all stages of breast cancer above stage 3 tend to have a higher lymph node count than those from earlier stages with the exception of stage 10 cancer.

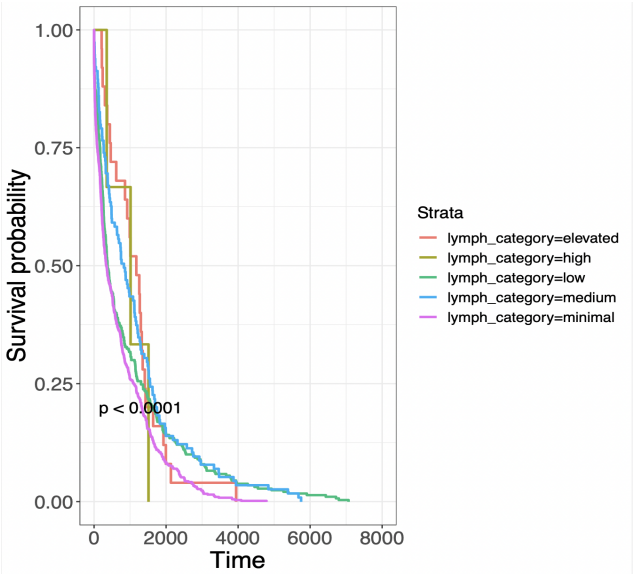
ii.



b. Survival Analysis of lymph Node count

i. As the number of lymph nodes increases, we see a decrease in how long the patient survives. Those with minimal and low counts survive far longer than those with High and elevated counts. This shows a correlation between the lymph node count and probability of survival over time.

ii.



c. Survival Analysis of Breast Cancer Stage

i. As the stage of breast cancer increases, the rate of survival decreases. Those with stage 3A,3C and 4 cancer live significantly shorter lives than those with stage 1,1A,1B, and 2 cancer. This shows a correlation between the breast cancer stage and probability of survival over time.

ii.